붓스트랩을 활용한 이상원인변수의 탐지 기법

강지훈 · 김성범†

고려대학교 산업경영공학부

Bootstrap-Based Fault Identification Method

Ji Hoon Kang · Seoung Bum Kim†

School of Industrial Management Engineering, Korea University, Seoul, Korea

Key Words : Hotelling's T^2, Decomposition, Bootstrap, Multivariate Process

Abstract

Multivariate control charts are widely used to monitor the performance of a multivariate process over time to maintain control of the process. Although existing multivariate control charts provide control limits to monitor the process and detect any extraordinary events, it is a challenge to identify the causes of an out-of-control alarm when the number of process variables is large. Several fault identification methods have been developed to address this issue. However, these methods require a normality assumption of the process data. In the present study, we propose a bootstrapped-based T^2 decomposition technique that does not require any distributional assumption. A simulation study was conducted to examine the properties of the proposed fault identification method under various scenarios and compare it with the existing parametric T^2 decomposition method. The simulation results showed that the proposed method produced better results than the existing one, especially in nonnormal situations.

1. 서 론

현재 많은 산업의 경쟁력 확보를 위해 가장 중요한 요소 중 하나는 고품질 제품의 생산이다. 기술수준이 같을 경우 생산과정에서 불량률을 줄여야 기업의 이윤은 증가될 수 있다. 또한 높은 품질의 제품은 이윤증가와 함께 고객의 만족도를 높임으로써 기업의 경쟁력 강화에 크게 기여한다. 따라서 많은 기업들이 품질력 향상을 위해 많은 노력을 기울이고 있는 실정이다. 공정을 통한 제품들이 모두 일관성 있게 동일한 규격으로 생산된다면 품질에 대한 관리는 필요 없겠지만 이는 현실적으로 불가능한 일이다.

그 이유는 생산공정에 변동(variability)이 존재하기 때문이다「김종덕과 장경, 2006」. 통계적 품질관리(statistical process control)는 통제 불가능한 우연적 요인(random cause)을 제외한 통제 가능한 이상 원인(assignable cause)을 통계적 기법을 활용하여 품질의 일관성을 향상시키는데 그 목적이 있다. 일반적으로 품질관리와 공정개선에 대표적으로 활용되는 기법인 관리도(control chart)는 공정을 실시간으로 모니터링 하여 불량현상을 조기에 발견하고 이에 대한 적절한 조치를 취함으로 공정이 지속적으로 정상 관리될 수 있도록 하는 기법이다「Sukchotrat et al., 2010」,「Woodall 과 Montgomery, 1999」. 관리도가 현장에서 널리 쓰이는 이유는 기법이 신뢰성 있는 통계적 이론에 근거하고 있고 결과가 해석하기 쉬운 도표로 표현되기 때문에 기법 도출에 관한 전문지식이 없는 일반 사용자들도 쉽게 이해할 수 있기 때문이다. 관리도는 모니터링 하고자 하는 변수의 개수에 따라 크게 단변량 관리도(univariate control chart)와 다변량 관리도(multivariate control

† 교신저자 sbkim1@korea.ac.kr

※ 이 논문은 한국연구재단의 지원 받아 수행된 연구임 (2011-0005372).

chart)로 나눌 수 있다. 단변량 관리도는 하나의 품질변수를 모니터링 하는 기법이고 다변량 관리도는 상관관계가 있는 여러 개의 변수를 동시에 모니터링하는 기법이다.

오늘날 생산 시스템이 복잡해지면서 생산 공정 자체가 여러 단계를 거치게 되었으며 각 단계에서도 여러 개의 관측값들을 동시에 모니터링 할 필요가 있게 되었다. 예를 들어, 반도체 공정의 경우, 웨이퍼(wafer) 생성부터 모듈완성 단계까지 수백 개의 세부 공정을 거치게 되고, 각 공정마다 수많은 센서를 통해 웨이퍼의 품질상태가 측정되는데 이를 단변량 관리도 기법을 통해 모니터링한다는 것은 매우 비효율적이다 「Jung et al., 2009」. 따라서 최근에는 상관관계를 가지고 있는 여러 개의 품질변수를 동시에 모니터링 할 수 있는 다변량 관리도의 연구가 활발히 이루어지고 있다 「Chongfuangprinya et al., 2010」.

다변량 관리도의 기법으로는 Hotelling's T^2 관리도, 다변량 지수가중이동평균관리도 (multivariate exponential weighted moving average chart), 다변량 누적합관리도 (multivariate cumulative sum chart)등이 있으며 이 중 대표적으로 사용되는 기법이 Hotelling's T^2 관리도이다. T^2 관리도는 다음 장 (Section 2)에서 설명하도록 하겠다. 다변량 관리도는 모니터링하고자 하는 여러 개의 변수를 하나의 통계량으로 요약하여 관리도에 표현함으로 다수의 품질변수들을 동시에 모니터링하기에 효과적인 기법이다. 하지만, 관리도로부터 이상판정을 받을 경우 해당 통계량은 여러 개의 변수를 포함하고 있기 때문에 어떤 변수(들)가 공정에 이상을 일으켰는지에 대한 판단이 어렵다는 한계점을 갖고 있다. 즉, 이상원인변수에 대한 검출이 어렵다는 단점이 있다 「Mason 과 Young, 2002」. 이러한 다변량 관리도의 이상원인변수 진단문제를 해결하기 위해서 몇몇 연구들이 수행되었다. Mason et al.(1995)은 변수 간의 조건부 관계를 통한 개별 변수의 기여도를 측정하는 MTY 분해(decomposition) 기법을 제안하였다. 이 기법의 기본 아이디어는 데이터가 정규분포를 따른다는 가정 하에 관측치의 전체적인 T^2(overall T^2 value)값이 카이제곱분포를 따를 뿐 아니라 분해된 개별 성분 역시 카이제곱분포를 따른다는 사실을 이용하여 이상원인변수를 찾아내는 것이다. 하지만 데이터가 정규분포를 따르지 않을 경우 사용하기 어렵다는 한계와 연산과정이 복잡하고 시간이 많이 걸린다는 것이 단점으로 작용한다 「Mason et al., 1997」. Runger(1996)는 이상관측치에 영향을 미치는 변수들에 대한 사전지식이 있는 경우 효과적으로 사용할 수 있는 기법인 U^2 통계량 기법을 제안하였다. 전체 변수를 통해 계산된 T^2 값에서 공정이상에 기여하지 않는다고 정의된 변수들로 이루어진 부분집단(subgroup)의 영향력을 제거한 T^2 값, 즉, U^2 값을 얻음으로써 공정이상원인변수를 파악하였다. 방법자체가 이해하기 쉽고 계산양이 적어 주요 이상원인변수를 찾는데 효과적이라 할 수 있지만, 관리이탈에 기여하는 변수 군에 대한 사전정보를 알아야 한다는 한계점을 갖고 있다. 또한, 이 방법 역시 개별 이상원인변수를 찾을 시 확률분포를 이용하는데 이는 전체데이터가 정규분포를 따른다는 가정이 있어야 한다. 최근 Li et al.(2008)은 변수들의 인과관계를 고려한 Causation-based T^2 분해 기법을 제안하였다. 이 기법의 특징은 베이지안 네트워크(Bayesian network)를 활용하여 변수 간에 존재하는 인과관계에 대한 정보를 얻고 이를 이용하여 T^2 값을 분해하는 것이다. Causation-based T^2 분해 기법은 MTY's 분해 방법과 비교했을 때 보다 좋은 성능을 보이고 있지만 (Li et al., 2008) 일반 사용자들이 이해하기 어려워 실제 문제에 널리 보급되기에는 한계가 있다. 또한, 이 방법 역시 데이터가 정규분포를 따름을 가정하고 있어 비정규분포를 따르고 있는 데이터에는 적용하기가 어렵다.

위에서 언급한 이상진단방법은 데이터가 정규분포를 따른다는 가정 하에 적용될 수 있는데 오늘날 복잡한 시스템으로부터 생성된 대부분의 데이터는 정규분포를 따르지 않고 있다 「Song et al., 2003」. 따라서 본 논문의 주목적은 데이터가 정규분포를 따르지 않는 경우 효과적으로 이상진단을 할 수 있는 기법을 개발하는데 있다. 본 연구에서는 널리 쓰이고 있는 T^2 분해 기법의 장점을 유지하며 기존 기법들의 한계점인 확률분포 가정이 필요 없는 강건한(robust) T^2 분해 기법을 제안하고자 한다.

2장에서는 Hotelling's T^2에 대한 전반적인 개념에 대해 서술 하였고, 3장에서 다변량 관리도 해석의 한계를 해결하기 위한 T^2 분해 기법에 대해 살펴보았다. 4장에서는 본 논문에서 제안하고 있는 붓스트랩 기반 T^2 분해 기법에 대해서 기술하였다. 5장에서는 기존 기법과 제안하는 기법의 성능 비교를 위해 수행한 시뮬레이션 실험 결과에 대해 정리하였고, 마지막으로 6장에서는 본 연구 결과에 대해 요약 하였다.

2. Hotelling's T^2 Control Chart

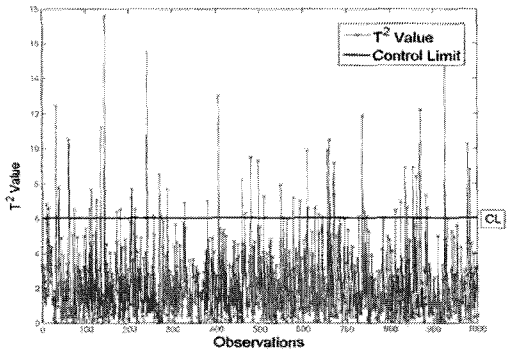
모니터링 하고자 하는 변수가 하나인 경우 관측 값이 정상인지 비정상인지를 구분하기 위한 기본적인 아이디어는 해당 관측 값과 나머지 정상관측치들과의 거리를 이용하는 것이다. 즉, 거리가 멀수록 관측값은 비정상일 가능성이 높은 것이다. 하지만 변수가 2개 이상인 다변량 공정에서는 단순 거리뿐 아니라 변수들간의 상관관계까지 고려해야 한다. 물론 다변량공정을 모니터링 하기 위해 각각의 변수를 단변량관리도를 이용해 모니터링 할 수 있으나 이는 변수들 사이의 상관관계를 전혀 고려하고 있지 않기 때문에 상관관계가 존재하는 대부분의 다변량공정에서는 정확한 이상탐지를 할 수 없다「Mason and Young, 2002」. 또한 변수의 수가 많을 경우 효과적인 해석이 어려울 수 있다. 이러한 문제점들을 해결하기 위해 여러 변수를 하나의 monitoring statistic으로 나타내는 다변량관리도가 개발되었다. 그중 대표적인 방법이 T^2 control chart인데 이 방법은 아래 식(1)에서 보는 바와 같이 관측치와 기존데이터의 단순거리뿐 아니라 변수간의 상관관계를 고려한 값 (T2)을 monitoring statistic로 사용하고 있다.

$$T^2=(X-\overline{X})^T S^{-1}(X-\overline{X}) \quad (1)$$

여기서 $\overline{X}$, S는 각각 정상 관측값 들로부터 계산한 표본의 평균과 분산을 나타낸다. 정상 관측 값으로부터 얻은 T^2 값을 토대로, 미래 관측치들의 정상 및 비정상 여부를 판단하게 되는데 그 판단을 결정하는 기준으로 관리한계선 (control limit)을 이용한다. T^2 관리도의 관리한계선은 데이터가 다변량 정규분포를 따른다는 가정 하에 다음의 식으로 계산할 수 있다.

$$CL=\frac{p(m+1)(m-1)}{m^2-mp}F_{\alpha,p,m-p} \quad (2)$$

여기서 p는 변수의 개수, m은 총 관측치의 개수를 나타낸다. 즉, 관리도의 관리한계선은 사용자가 임의로 정하는 1종 오류(α)가 주어졌을 때 자유도가 p와 m인 F분포로써 결정된다. 보통 관리한계선은 상한선과 하한선이 존재하는데, T^2 관리도의 경우 T^2 통계량은 언제나 양의 값이고 따라서 관리하한선은 항상 0이 되기 때문에 통상 따로 표시하지 않는다. <Figure 1>에 보여주듯 모니터링 하고자 하는 관측치의 값이 식 (2)로부터 구한 한계선보다 크면 해당 관측치는 공정이상판정을 받게 된다. 하지만 앞에서 언급했듯이 T^2 값은 모든 변수들의 선형 결합된 형태(식(1))이기 때문에 공정이상을 발견하였어도 과연 어떤 변수(들)가 이상을 일으켰는지 대한 정확한 해석이 어렵다는 문제를 갖고 있다「Wierda, 1994」.



<Figure 1> Example of Hotelling's T^2 Control Chart

3. T^2 Decomposition Technique

T^2 관리도의 이상원인변수의 검출에 대한 한계점을 극복하기 위해 사용되는 대표적인 기법이 T^2 분해 기법이다 「Montgomery, 2005」. 이 기법의 기본적인 아이디어는 이상관측치가 발견되면 해당 T^2 값을 각 변수별로 분해(decompose)하여 분해한 값이 큰 변수(들)를 찾아 이상을 발생시킨 주변수라고 판정하는 것이다. 아래 식 (3)으로부터 얻어지는 d_i 값들이 이상관측치에 대한 각 변수별 분해 값이다.

$$d_i = T^2 - T^2_{(i)}, \qquad i=1,2,3,\ldots,p \quad (3)$$

여기서 $T^2_{(i)}$ 값은 i 번째 변수를 제외하고 계산한 T^2 값이다. 따라서 d_i 값은 모든 변수를 사용하여 구한 T^2 값과 i 번째 변수를 제외하고 계산한 T^2 값의 차이가 된다. 이는 i 번째 변수가 전체 T^2 값에 얼마나 영향을 주었는지를 의미하여 영향력이 클수록 d_i 값은 커지게 됨을 알 수 있다「Runger et al., 1996」.

요약하여 말하자면, 이상관측치가 발견되면 각 변수별로 d_i 값을 계산하고 d_i 값이 큰 변수를 이상치에 대한 주변수라고 판정하는 것이다. 여기서 d_i 값이 얼마나 커야 주변수라고 판정할지에 대한 기준이 필요하게 된

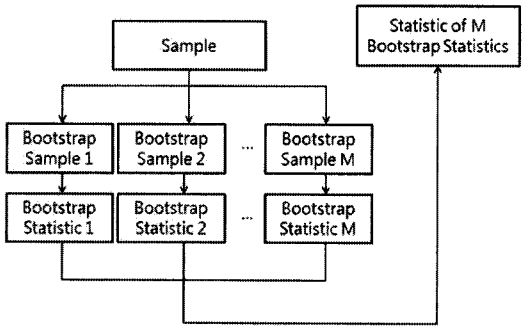
다. Runger et al.(1996)는 데이터가 정규분포를 따를 때, 각 변수별 d_i 값 들이 자유도가 1인 카이자승(χ^2) 분포를 따름을 알아내었다. 이를 이용하여 d_i 값에 대한 임계치를 $\chi^2_{(\alpha,1)}$값으로 결정하였다. 다시 말해 d_i 값이 $\chi^2_{(\alpha,1)}$보다 크면 해당 변수는 이상을 일으킨 주된 변수로써 판정되는 것이다.

이 방법은 T^2 가 갖고 있는 고유의 통계적 성질을 이용하였고 하나의 전역적(global) 임계값을 사용함으로써 실제 적용에 용이한 점이 있지만, 데이터가 반드시 정규분포를 따라야 한다는 한계점을 가지고 있다. 이 한계점을 극복하기 위해 본 논문에서는 정규성 가정(normality assumption)으로부터 자유로운 붓스트랩(bootstrap) 기반 분해 방법을 제안하고자 한다.

4. Bootstrap-Based Fault Identification Method

대부분 기존 관리도의 관리한계선은 관리되어야 할 품질특성치가 정규분포를 따른다는 가정 하에서 설정된 것이다. 하지만 실제 공정 데이터들은 정규분포뿐만 아니라 다양한 형태로 존재하며(Song et al., 2003), 이러한 경우, 정규분포를 가정하고 설계된 기존의 모수기반 기법은 오류를 야기 할 수 있다.

붓스트랩은 모집단의 확률분포가 알려져 있지 않은 경우 표본의 통계량을 추정하는 방법이다. 「Efron 과 Tibshirani, 1993」. <Figure 2>는 붓스트랩 기법을 그림으로 설명한 것이다. 각각의 표본을 B번(최소 1000번 이상) 복원 추출하고, 각각 얻어진 붓스트랩 표본 하나하나의 개별 통계량 값을 구한 다음, 이를 이용하여 통계량을 추정하게 된다.



<Figure 2> Explanation of Bootstrap Method

붓스트랩을 이용한 이상치 진단방법을 요약해보면 다음과 같다.

1. 정상관측치들의 T^2 를 식(1)을 통해 계산하고, 연산된 T^2 를 식(3)를 통해 분해하여 d_i 행렬(matrix)을 얻는다.

2. d_i 행렬에서 각 변수별로 B번(최소 1,000번) 복원추출 하여 붓스트랩 표본을 얻는다.

3. 각 변수별 붓스트랩 표본로부터$100\times(1-\alpha)$ 분위수 값을 얻는다. 여기서 α는 1종 오류로써 0과 1사이에서 존재하며 보통 사용자가 정하게 된다.

4. 위 단계에서 구한 $100\times(1-\alpha)$ 분위수 값들의 평균을 취함으로써 임계치를 얻는다.

5. 각 변수별로 산출된 p개의 임계치를 미래에 이상관측치가 발생하였을 경우 각 변수별 분해값과 비교하여 이상치에 주원인이 되는 변수를 찾는다. 즉, 임계치보다 큰 d_i를 찾고 해당 변수를 주원인변수로 판정한다.

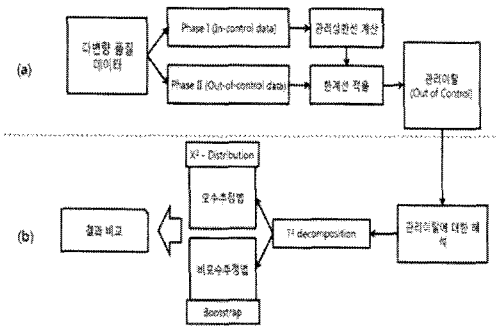
본 논문에서 제안하고 있는 붓스트랩 기반 T^2 분해 방법은 기존의 T^2 분해 기법에 비해 다음과 같은 장점이 있다. 비모수 추정법인 붓스트랩을 사용함으로 모집단의 확률분포에 대해 사전정보가 없어도 사용할 수 있다. 또한, 기존 T^2 분해 방법은 모든 분해값에 대해 전역적인 임계값을 적용하는데 반해 제안하고 있는 방법은 데이터의 변수가 p개 인 경우, 각 변수의 분포에 맞는 p개의 임계값을 갖게 되어, 보다 정확한 이상치의 원인변수 탐지가 가능해 진다는 장점이 있다.

5. Simulation

5.1 시뮬레이션 설계

시뮬레이션을 통해 생성된 데이터를 이용하여 기존 모수 기반 T^2 분해 방법과 제안하고 있는 붓스트랩 기반 T^2 분해 방법의 성능을 비교하였다. 실험결과의 일반화를 위하여 일련의 과정을 1,000회 실시하였으며 이로부터 계산된 평균값을 성능에 대한 척도로써 사용하였다. <Figure 3>은 전체 시뮬레이션과정을 그림으로 보여주고 있다. <Figure 3>의 (a)는 데이터를 생성한 후 정상데이터만을 가지고 관리상한선을 결정하는

Phase I 단계와 관리상한선을 이용 관측치를 정상과 비정상으로 구분하는 Phase II 단계를 보여주고 있다. Phase II 단계에서 이상치로 판단된 관측값은 <Figure 3>의 (b)에서 보여주듯 T^2 분해 기법에 의해 변수 별 분해되어 기존모수 기반 분해 방법과 제안하고 있는 붓스트랩 기반 이상원인변수 탐지기법에 의해 이상에 주 원인이 되는 변수(들)를 찾게 된다.



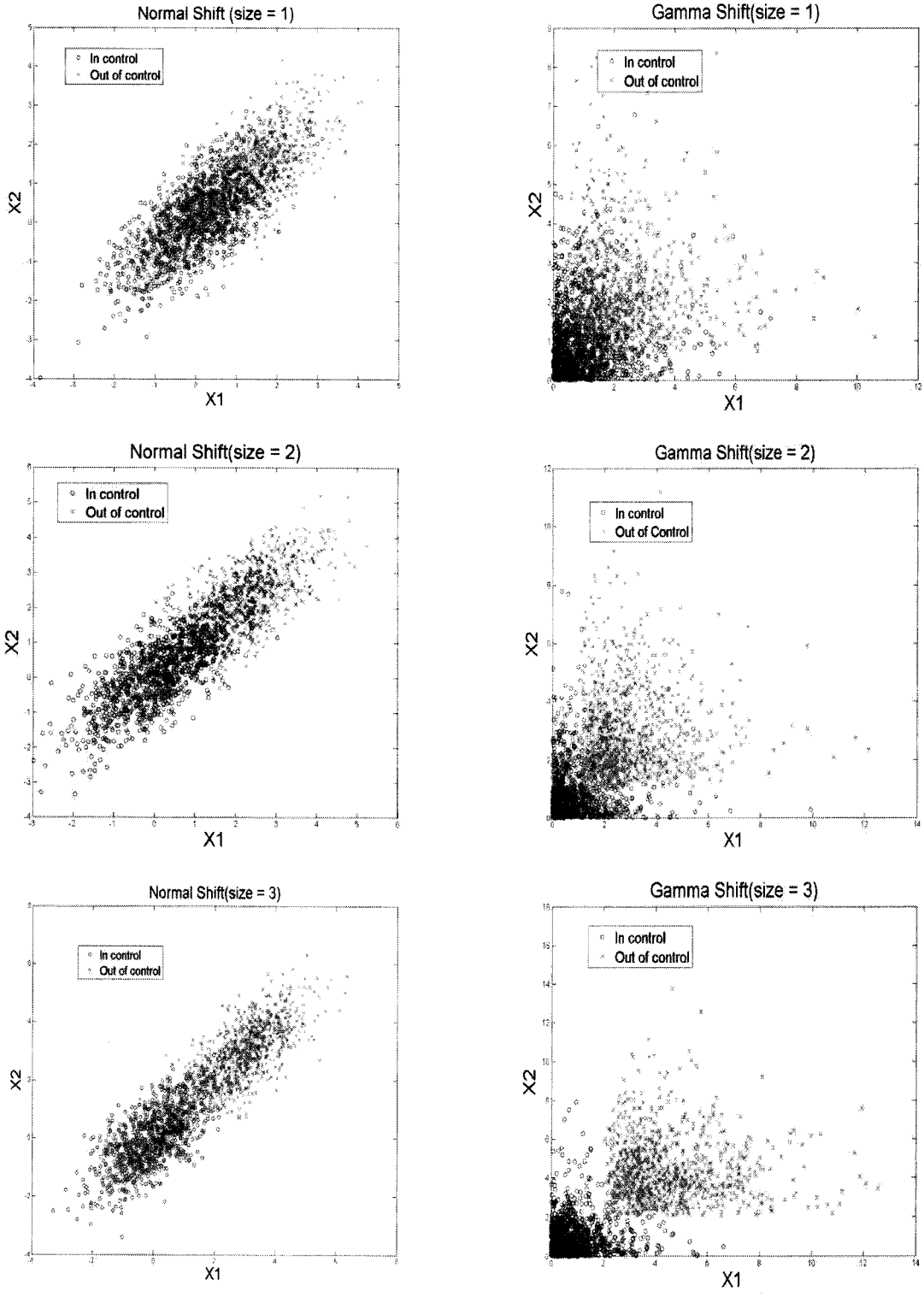
<Figure 3> Simulation Procedure

물레이션 데이터는 정규분포와 비정규분포인 감마분포에 의해 생성되었다. 실험데이터는 Phase I 단계를 위하여 1,000개, Phase II 단계를 위하여 100개를 생성하였다. 실험의 주된 목표가 평균 변화(mean shift)에 기인한 공정이상에 기여하는 변수를 올바르게 찾는 것이기 때문에 Phase I 데이터는 전혀 변화가 없는 즉, 모든 변수가 각각 평균이 0인 다변량 분포로 생성하였으며, Phase II의 경우, 임의적으로 몇몇 변수에 평균에 변화를 주어 공정이상에 영향을 주는 변수로 지정하였다. Table 1에서 보여주듯 다양한 공정이상상태를 구현하기 위해 변수의 개수를 10개, 5개, 3개로 나누어 데이터를 생성하였고 각각의 경우 평균 변화된 변수의 개수도 다양하게 실험을 하였다. 또한 변화정도도 작은 경우, 중간 경우, 큰 경우로 나누어 실험을 하였다.

예를 들어, Table 1에서 case 1의 경우에는 총 10개의 변수를 생성하고 이중 첫 번째 변수(X1)의 평균이 2만큼 변화된 비정상관측치(out-of-control observation)를 생성하게 된다. 이와 마찬가지로, case 26의 경우는 3개의 변수를 고려하였고 3변수 모두 평균이 1만큼 이동된 데이터가 생성되게 된다.

<Table 1> Simulation Scenarios for Out-of-Control Observations

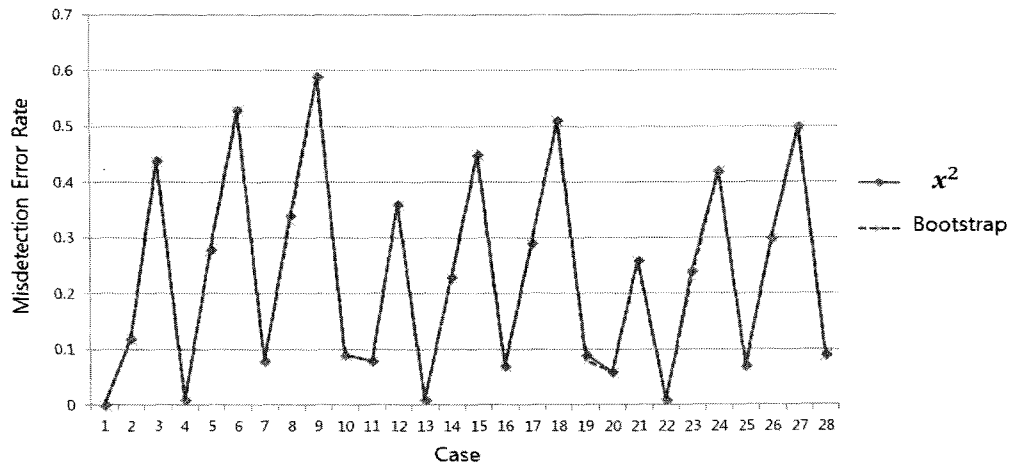
Case	변수의 개수	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1		2	0	0	0	0	0	0	0	0	0
2		1	0	0	0	0	0	0	0	0	0
3		3	0	0	0	0	0	0	0	0	0
4		2	2	2	0	0	0	0	0	0	0
5	10개	1	1	1	0	0	0	0	0	0	0
6		3	3	3	0	0	0	0	0	0	0
7		2	2	2	2	2	0	0	0	0	0
8		1	1	1	1	1	0	0	0	0	0
9		3	3	3	3	3	0	0	0	0	0
10		2	0	0	0	0					
11		1	0	0	0	0					
12		3	0	0	0	0					
13		2	2	0	0	0					
14	5개	1	1	0	0	0					
15		3	3	0	0	0					
16		2	2	2	0	0					
17		1	1	1	0	0					
18		3	3	3	0	0					
19		2	0	0							
20		1	0	0							
21		3	0	0							
22		2	2	0							
23	3개	1	1	0							
24		3	3	0							
25		2	2	2							
26		1	1	1							
27		3	3	3							



<Figure 4> Various Mean Shift in Normal and Gamma Distributions

<Table 2> Comparison of The Decomposition Result between χ^2 Threshold and Bootstrap threshold in Normal Cases

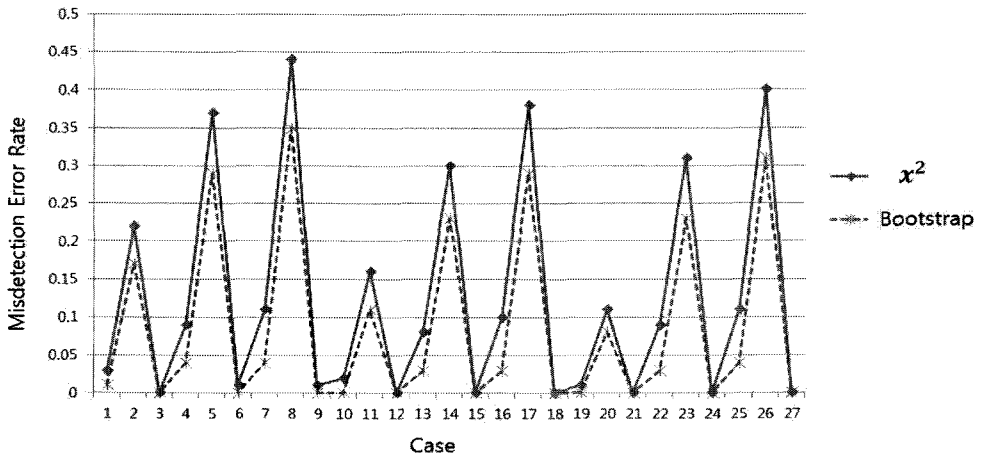
	χ^2 threshold		Bootstrap threshold	
	Faulty detection error	Misdetection error	Faulty detection error	Misdetection error
case1	0.17(0.01)	0.12(0.005)	0.17(0.001)	0.12(0.005)
case2	0.22(0.002)	0.44(0.01)	0.22(0.003)	0.44(0.01)
case3	0.13(0.001)	0.01(0.01)	0.13(0.001)	0.01(0.001)
case4	0.12(0.001)	0.28(0.003)	0.12(0.001)	0.28(0.03)
case5	0.18(0.0008)	0.53(0.001)	0.19(0.0008)	0.53(0.001)
case6	0.10(0.001)	0.08(0.002)	0.10(0.001)	0.08(0.002)
case7	0.11(0.0005)	0.34(0.001)	0.11(0.0005)	0.33(0.001)
case8	0.16(0.002)	0.59(0.003)	0.16(0.002)	0.59(0.003)
case9	0.11(0.001)	0.09(0.002)	0.11(0.001)	0.09(0.002)
case10	0.17(0.0009)	0.08(0.001)	0.17(0.0009)	0.08(0.001)
case11	0.26(0.005)	0.36(0.01)	0.26(0.005)	0.36(0.01)
case12	0.13(0.001)	0.01(0.001)	0.13(0.001)	0.01(0.001)
case13	0.13(0.002)	0.23(0.003)	0.13(0.002)	0.23(0.003)
case14	0.21(0.004)	0.45(0.005)	0.21(0.004)	0.45(0.006)
case15	0.11(0.001)	0.07(0.002)	0.11(0.001)	0.07(0.002)
case16	0.11(0.002)	0.29(0.003)	0.11(0.002)	0.29(0.003)
case17	0.19(0.001)	0.51(0.0014)	0.19(0.001)	0.51(0.001)
case18	0.10(0.0007)	0.09(0.0007)	0.10(0.0007)	0.08(0.0007)
case19	0.17(0.001)	0.06(0.001)	0.17(0.001)	0.06(0.001)
case20	0.28(0.002)	0.26(0.003)	0.28(0.002)	0.26(0.003)
case21	0.12(0.002)	0.01(0.001)	0.12(0.002)	0.01(0.01)
case22	0.12(0.004)	0.24(0.003)	0.12(0.004)	0.23(0.003)
case23	0.22(0.002)	0.42(0.001)	0.22(0.002)	0.42(0.001)
case24	0.10(0.001)	0.07(0.0007)	0.10(0.001)	0.07(0.0007)
case25	NA	0.30(0.003)	NA	0.30(0.003)
case26	NA	0.50(0.001)	NA	0.50(0.001)
case27	NA	0.09(0.0007)	NA	0.09(0.0007)
Average error	0.15	0.24	0.15	0.24



<Figure 5> Graphical Summary of Misdetection Error Rates in Normal Case

<Table 3> Comparison of The Decomposition Result between χ^2 Threshold and Bootstrap threshold in Gamma Cases

	χ^2 threshold		Bootstrap threshold	
	Faulty detection error	Misdetection error	Faulty detection error	Misdetection error
case1	0.10(0.001)	0.03(0.002)	0.13(0.001)	0.01(0.001)
case2	0.13(0.0005)	0.22(0.002)	0.15(0.0006)	0.17(0.002)
case3	0.08(0.0003)	0.00(0)	0.11(0.0004)	0.00(0)
case4	0.07(0.001)	0.09(0.003)	0.10(0.001)	0.04(0.002)
case5	0.09(0.001)	0.37(0.004)	0.12(0.001)	0.29(0.004)
case6	0.07(0.001)	0.01(0.0003)	0.11(0.002)	0.00(0.00001)
case7	0.07(0.001)	0.11(0.003)	0.10(0.002)	0.04(0.0005)
case8	0.08(0.001)	0.44(0.003)	0.10(0.001)	0.35(0.003)
case9	0.08(0.001)	0.01(0.0004)	0.12(0.002)	0.00(0.0001)
case10	0.09(0.001)	0.02(0.001)	0.12(0.002)	0.00(0.001)
case11	0.14(0.003)	0.16(0.006)	0.17(0.003)	0.11(0.006)
case12	0.08(0.001)	0.00(0)	0.10(0.002)	0.00(0)
case13	0.07(0.001)	0.08(0.003)	0.10(0.002)	0.03(0.001)
case14	0.11(0.002)	0.3(0.004)	0.13(0.002)	0.23(0.004)
case15	0.07(0)	0.00(0.00001)	0.10(0.002)	0.00(0)
case16	0.07(0.0002)	0.10(0.001)	0.10(0.001)	0.03(0.0005)
case17	0.09(0.01)	0.38(0.005)	0.11(0.001)	0.29(0.007)
case18	0.07(0.0005)	0.00(0.0001)	0.10(0.002)	0.00(0)
case19	0.09(0.003)	0.01(0.002)	0.12(0.01)	0.00(0.0002)
case20	0.15(0.002)	0.11(0.005)	0.18(0.01)	0.08(0.001)
case21	0.07(0.01)	0.00(0)	0.10(0.003)	0.00(0)
case22	0.07(0.002)	0.09(0.003)	0.10(0.007)	0.03(0.002)
case23	0.10(0.01)	0.31(0.004)	0.12(0.01)	0.23(0.002)
case24	0.07(0.001)	0.00(0.00002)	0.10(0.002)	0.00(0)
case25	NA	0.11(0.001)	NA	0.04(0.001)
case26	NA	0.40(0.003)	NA	0.31(0.003)
case27	NA	0.00(0.0001)	NA	0.00(0.000001)
Average error	0.09	0.12	0.11	0.08



<Figure 6> Graphical Summary of Misdetection Error Rates in Gamma Case

본 시뮬레이션에서는 변수의 개수를 3개, 5개, 10개로 생성하였지만 이에 대한 평균 변화를 시각적으로 표현하기 어렵게 때문에 <Figure 4>에서 2개의 변수만을 가지고 각각 정규분포와 감마분포의 평균 변화의 크기에 따른 데이터의 형태를 나타내 보았다. 예상한대로 평균 변화의 크기가 작은 경우 정상관측치와 비정상관측치의 차이가 거의 없음을 보여주고 있고 변화의 크기가 점점 커짐에 따라 정상과 비정상 관측치간에 차이가 극명해 짐을 볼 수 있다.

이상원인변수 탐지기법의 성능은 평균이 변화된 변수를 얼마나 정확히 탐지하는지로 평가할 수 있을 것이다. 여기서 실제 변화된 변수를 정확히 탐지한다는 의미는 두 가지 관점에서 생각해 볼 수 있다. 첫째는 실제 변화된 변수 중 몇 개를 올바르게 찾는 가 이고 둘째는 실제 변화가 되지 않은 변수를 얼마나 제대로 찾지 않는지에 대한 여부이다. 본 논문에서는 이 두 척도를 계산하기 위해 다음의 misdetection error rate와 faulty detection error rate를 이용하였다 「Li et al., 2008」.

- Misdetection error rate = P(임계치 보다 낮은 di를 갖는 변수들 | 실제 평균이 변화된 변수들)
- Faulty detection error rate= P(임계치를 넘는 di를 갖는 변수들 | 실제 평균이 변화되지 않은 변수들)

5.2 시뮬레이션 결과

본 연구에서 제안한 붓스트랩 기반 T^2 분해 기법과 기존 모수 기반 T^2 분해 기법을 앞장에서 설명한 시뮬레이션 데이터를 통해 비교한 결과를 <Table 2>과 <Table 3>에 정리하였다. 그리고 <Figure 5>와 <Figure 6>를 통해 Table에 제시된 성능 비교결과를 시각적으로 요약하였다. 두 방법의 객관적인 비교를 위해 본 논문에서는 faulty detection error rate를 비슷하게 조절한 상태에서 misdetection error rate를 비교하였다. 즉, 각 case별 비슷한 faulty detection error rate하에서 misdetection error rate가 낮은 방법이 이상진단 면에서 더 우수하다고 할 수 있겠다. <Table 2>는 관측치가 다변량 정규 분포를 따르는 경우로써 기존 방법과 제안한 방법과의 차이가 거의 없음을 보여주고 있다. Table내 괄호 안의 값들은 1,000번 실험에 대한 표준편차를 의미하며, 표 하단의 NA는 3개의 변수 모두를 평균 변화시켜 생성한 케이스이기 때문에 평균이 변하지 않은 변수를 기여변수로 규정짓는 error가 존재하지 않음을 의미한다.

<Table 3>는 관측치가 비정규분포인 다변량 감마분포를 따르고 있을 때 기존방법과 제안하고 있는 방법의 이상 원인 진단 성능을 비교하고 있다. 표에서 보듯이 본 연구에서 제시하고 있는 붓스트랩 기반 T^2 분해 방법이 기존 방법에 비해 대부분의 경우 misdetection error rate가 작음을 알 수 있다. 특히, 변화의 크기가 작은 경우 결과 차이가 더욱 확연한 것으로 보아 제안하는 기법은 작은 변동을 야기하는 변수도 잘 탐지함을 알 수 있다.

6. 결 론

본 논문에서는 대표적인 비모수 추정법인 붓스트랩 기법을 활용하여 데이터의 확률분포에 대한 사전정보가 없는 경우에 공정이상 관측치에 대한 해석을 용이하게 하는 방법을 제시하였다. T^2 값을 통해 관측치가 이상으로 규정되면, 어떤 변수에 의한 요인으로 문제가 발생했는지에 대한 정보가 필요하게 된다. 기존의 χ^2 통계량 값을 이용하는 모수 추정법은 관측치들이 정규분포를 따르고 있어야 한다는 가정이 있어야 사용할 수 있었으나 제안하는 기법은 보다 유연한 임계치의 설정을 통해 관측치들의 확률분포에 관계없이 사용할 수 있다.

다양한 경우를 반영한 시뮬레이션 실험을 통해 제안하고 있는 이상원인변수 탐지기법이 데이터가 정규분포를 따를 경우에는 기존방법과 유사한 결과를, 데이터가 비정규 분포를 따를 경우는 더 정확한 결과를 얻음을 보여주었다.

비단 T^2 뿐만 아니라 다변량 관측 통계량이 쓰이는 모든 관리도에서의 해석문제는 공정관리에 있어서 매우 중요한 사안이다. 본 연구에서 제시한 붓스트랩 기반 추정을 통한 영향 변수의 탐색기법은 다양한 분야의 다변량 관리도의 해석에 효과적으로 활용될 수 있을 것이라 기대한다.

참고문헌

[1] 김종덕, 장경 (2006), 「통계적 품질관리」, 자유아카데미.

[2] Chongfuangprinya, P., Kim, S. B., Park, S. K. and Sukchotrat, T.(2010), "Integration of support vector machines and control charts for multivariate process monitoring," *Journal of Statistical Computation and Simulation*, In Press.

[3] Efron, B. and Tibshirani, R.(1993), *An Introduction to the Bootstrap*, Chapman & Hall/CRC, Boca Raton, FL.

[4] Hotelling, H.(1947), "Multivariate Quality Control." In C. Eisenhart, M. W. Hastay, and W. A. Wallis, eds. *Techniques of Statistical Analysis.* McGraw-Hill, New York, NY.

[5] Jung, H. C., Kang, C. W. and Kang, H. W. (2009), "Dynamic Yield Improvement Model Using Neural Networks," *Journal of the Society of Korea Industrial and Systems Engineering*, Vol. 32, No.2, pp. 132-139.

[6] Li, J., Jin., and Shi, J.(2008), "Causation-Based T^2 Decomposition for Multivariate Process Monitoring and Diagnosis," *Journal of Quality Technology*, Vol.40 No.1, 4, pp. 6-58.

[7] Mason, R. L., Tracy, N. D. and Young, J. C.(1995), "Decomposition of T^2 for multivariate control chart interpretation," *Journal of Quality Technology*, Vol.27 No.2, pp. 99-108.

[8] Mason, R. L. and Young, J. C.(2002), *Multivariate Statistical Process Control With Industrial Applications*, American Statistical Association and the Society for Industrial and Applied Mathematics, Philadelphia, PA.

[9] Montgomery, D. C.(2005), *Introduction to Statistical Quality Control*, 5th ed., Wiley, New York, NY.

[10] Runger, G. C.(1996), "Projections and the U^2 multivariate control chart," *Journal of Quality Technology*, Vol.28 No.3, pp. 313-319.

[11] Runger, G. C., Alt, F. B. and Montgomery, D. C.(1996), "Contributors to a multivariate statistical process control chart signal," *Communications in Statistics: Theory and Methods*, Vol.25 No.10, pp. 2203-2213.

[12] Shewhart, W. A. (1939), *Statistical Methods from the viewpoint of Quality Control*, Republished in 1986 by Dover Publications, New York, NY.

[13] Sukchotrat, T., Kim, S. B. and Tsung, F. (2010), "One-class classification-based controlcharts for multivariate process monitoring," *IIE Transactions*, Vol. 42, pp. 107-120.

[14] Song, S. I., Cho, Y. C. and Park, H. K.(2003), "Robust Control Chart using Bootstrap Method," *Journal of the Society of Korea Industrial and Systems Engineering*, Vol. 26, No.3, pp. 39-49.

[15] Wierda, S. J.(1994), "Multivariate statistical process control: recent results and directions for future research," *Statistica Neerlandica*, Vol.48 No.2, 147-168.

[16] Woodall, W. H. and Montgomery, D.C.(1999), "Research issues and ideas in statistical process control," *Journal of Quality Technology*, Vol.31 No.4, 376-386.

2011년 2월 9일 접수, 2011년 4월 20일 수정, 2011년 4월 28일 채택