부적합률의 다중변화점분석을 위한 베이지안절차

김 경 숙* · 김 희 정* · 박 정 수* · 손 영 숙*†

* 전남대학교 자연과학대학 통계학과

Bayesian Procedure for the Multiple Change Point Analysis of Fraction Nonconforming

Kyungsook Kim* · Heejeong Kim* · Jeong soo Park* · Young Sook Son*†

Department of Statistics, College of Natural Sciences, Chonnam National University

Keywords : fraction nonconforming, change point, Bayes factor, Gibbs sampling, Metropolis-Hastings algorithm.

Abstract

In this paper, we propose Bayesian procedure for the multiple change points analysis in a sequence of fractions nonconforming. We first compute the Bayes factor for detecting the existence of no change, a single change or multiple changes. The Gibbs sampler with the Metropolis-Hastings subchain is run to estimate parameters of the change point model, once the number of change points is identified. Finally, we apply the results developed in this paper to both a real and simulated data.

1. 서 론

본 논문에서는 품질의 특성값으로서 부적합율(fraction nonconforming)에 관심을 갖고, 공정의 상태가 어느 시점을 기준으로 하여 부적합율이 기존의 수준과 달라졌는지, 달라졌다면 어떠한 유형으로, 그리고 어느 정도로 달라졌는지를 밝혀내기 위한 베이지안 변화점분석법을 논의하고자 한다.

김경숙 등 (2006)에서는 공정 상태 하에서 허용되는 부적합률의 최대수준(p_0)을 기준으로 봤을 때 관측치열에서의 부적합률(p)의 수준이 그대로 유지되고 있는지($p = p_0$), 최대 허용수준보다 더 높아졌는지($p > p_0$), 또는 더 감소되었는지 ($p < p_0$)를 검정할 수 있는 베이지안 다중검정 절차가 제안되었다. 이 방법은 일정 시기 내의 관측치열에 대해 전체적인 수준을 검토하므로 시간흐름에 따른 변화추이는 간과되는 제한점이 있다.

본 논문에서는 관측치열 내에서 기준 시점을 처음부터 한 시점씩 뒤로 밀쳐가면서 앞부분의 부적합률 수준과 뒷부분의 부적합률 수준이 변화되었는지를 비교하여 내재하는 모든 변화점들을 찾아내고, 자료에 적합한 모형이 식별된 후에는 모수추정 과정을 통해 모형적합을 수행한다.

본 논문의 2절에서는 모형선택을 위한 검정도구인 베이즈인자(Bayes factor)와 검정모형 및 변화점의 사후확률을 소개하고, 3절에서는 선택된 모형에 내재된 모수들의 추정을 위한 깁스샘플러(Gibbs sampler) 및 메트로폴리스-헤스팅스(Metropolis-Hastings: M-H) 알고리즘을 구축한다. 본 논문에서 제안된 기법에 대한 검증을 위하여 모의실험 및 실제 자료분석을 수행하였다. 모의실험결과는 지면관계상 생략하고 논문발표시 제시한다.

† 교신저자 ysson@chonnam.ac.kr

※ 본 연구는 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음.

2. 모형선택

부적합률이 p인 공정 상태에서 생산된 무한개의 제품들 가운데 어느 시점 $t(t=1,2,...,T)$에서 n_t개의 표본(시료)을 추출하여 그 중 부적합품으로 판정되는 개수를 X_t라 하자. 이 때 확률변수 X_t는 $b(n_t,p)$인 이항분포를 따르는 것으로 가정할 수 있고 확률분포는 다음과 같이 정의된다.

$$f(x_t|n_t,p)=\binom{n_t}{x_t}p^{x_t}(1-p)^{n_t-x_t},$$
$$x_t=0,1,2,...,n_t,\ t=1,2,...,T.$$

2.1 변화점 검정모형의 설정

T개 시점에서 관측된 자료를 $\boldsymbol{x}=\{x_1,x_2,...,x_T\}$라고 할 때, 어느 시점 $r(r=1,2,...,T-1)$에서 부적합률(p)에 변화가 있었는지를 검정하기 위해 무변화모형(M_0) 대 변화모형(M_1)으로서 다음과 같이 설정한다.

$$M_0:\ X_t\sim b(n_t,p_0),\ t=1,2,...,T,$$

$$M_1:\begin{cases}X_t\sim b(n_t,p_0),\ t=1,2,...,r,\\ X_t\sim b(n_t,p_1),\ t=r+1,r+2,...,T,\end{cases}$$

여기서 $r\in\{1,2,...,T-1\}$은 변화시점을 의미하고, p_0와 p_1은 변화점을 중심으로 하여 전후의 자료들에 대한 모부적합률을 의미한다.

2.2 베이즈 인자 정의

모형선택을 위한 베이지안 기법은 모수에 대한 사전정보와 관측자료에서 얻는 정보를 결합시켜 구성되는 베이즈인자라는 검정도구를 사용하여 수행된다.

먼저, 모수에 대한 사전분포로서 p_0와 p_1에 대해서는 각각 $beta(\alpha_0,\beta_0)$와 $beta(\alpha_1,\beta_1)$분포를, 그리고 변화점 r에 대해서는 이산 균일분포를 다음과 같이 가정하였다.

$$\pi_0(p_0)=\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}p_0^{\alpha_0-1}(1-p_0)^{\beta_0-1},$$
$$\pi_1(p_1)=\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}p_1^{\alpha_1-1}(1-p_1)^{\beta_1-1},$$
$$\pi_1(r)=\frac{1}{T-1},\quad r=1,2,...,T-1.$$

여기서 $0<p_0,p_1<1,\ p_0\neq p_1,\ \alpha_0,\beta_0>0,\ \alpha_1,\beta_1>0$. 또한 본 논문에서 모든 사전분포들은 서로 독립이라 가정한다.

다음으로, 각 모형 M_0,M_1 하의 관측 자료로부터는 다음과 같은 각각의 우도함수 $L_0(p_0|\boldsymbol{x})$와 $L_1(p_0,p_1,r|\boldsymbol{x})$를 통해 정보를 얻는다.

$$L_0(p_0|\boldsymbol{x})=\left\{\prod_{t=1}^{T}\binom{n_t}{x_t}\right\}p_0^{\sum_{t=1}^{T}x_t}(1-p_0)^{\sum_{t=1}^{T}n_t-\sum_{t=1}^{T}x_t},$$

$$L_1(p_0,p_1,r|\boldsymbol{x})=\left\{\prod_{t=1}^{T}\binom{n_t}{x_t}\right\}p_0^{\sum_{t=1}^{r}x_t}(1-p_0)^{\sum_{t=1}^{r}n_t-\sum_{t=1}^{r}x_t}\times p_1^{\sum_{t=r+1}^{T}x_t}(1-p_1)^{\sum_{t=r+1}^{T}n_t-\sum_{t=r+1}^{T}x_t}.$$

모형 M_0에 대한 모형 M_1의 베이즈인자는 각 모형에 대한 사전예측분포(혹은 주변확률분포라 불림)인 $m_0(\boldsymbol{x})$와 $m_1(\boldsymbol{x})$의 비(ratio)로서 정의된다.

$$B_{10}=\frac{m_1(\boldsymbol{x})}{m_0(\boldsymbol{x})},$$

여기서

$$m_0(\boldsymbol{x})=\int_0^1\pi_0(p_0)L_0(p_0|\boldsymbol{x})dp_0$$
$$=C(\boldsymbol{x})\frac{\Gamma\left(\alpha_0+\sum_{t=1}^{T}x_t\right)\Gamma\left(\beta_0+\sum_{t=1}^{T}n_t-\sum_{t=1}^{T}x_t\right)}{\Gamma\left(\alpha_0+\beta_0+\sum_{t=1}^{T}n_t\right)},$$

$$m_1(\boldsymbol{x})=\sum_{r=1}^{T-1}m_1(r,\boldsymbol{x}),$$

$$m_1(r,\boldsymbol{x})$$
$$=\int_0^1\int_0^1\pi_1(r)\pi_1(p_0,p_1)L_1(p_0,p_1,r|\boldsymbol{x})dp_0dp_1$$

$$= C(\boldsymbol{x})\frac{1}{T-1}\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}$$

$$\times\frac{\Gamma\left(\alpha_0+\sum_{t=1}^{r}x_t\right)\Gamma\left(\beta_0+\sum_{t=1}^{r}n_t-\sum_{t=1}^{r}x_t\right)}{\Gamma\left(\alpha_0+\beta_0+\sum_{t=1}^{r}n_t\right)}$$

$$\times\frac{\Gamma\left(\alpha_1+\sum_{r+1}^{T}x_t\right)\Gamma\left(\beta_1+\sum_{r+1}^{T}n_t-\sum_{r+1}^{T}x_t\right)}{\Gamma\left(\alpha_1+\beta_1+\sum_{r+1}^{T}n_t\right)},$$

$$C(\boldsymbol{x})=\left\{\prod_{t=1}^{T}\binom{n_t}{x_t}\right\}\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}.$$

2.3 모형의 사후확률 계산

경쟁모형들 가운데 어느 하나가 자료에 가장 적합한 것으로 선택이 되는 기준은 주어진 자료에 대한 사후확률이 가장 큰 것이다. 각 모형에 대한 사후확률의 계산식은 다음과 같다.

$$P(M_0|\boldsymbol{x})=\frac{q_0}{q_0+q_1B_{10}},$$

$$P(M_1|\boldsymbol{x})=1-P(M_0|\boldsymbol{x})$$

여기서, q_0와 q_1은 각 모형에 대한 사전확률로서 모수에 대한 사전정보가 거의 없는 경우에는 보통 동일한 값을 부여한다. 이 때 모형의 사후확률은 다음과 같이 다시 표현될 수 있다.

$$P(M_0|\boldsymbol{x})=\frac{m_0(\boldsymbol{x})}{m_0(\boldsymbol{x})+m_1(\boldsymbol{x})}.$$

2.4 변화점의 사후확률 계산

검정결과로서 변화모형(M_1)이 선택되었다면 변화점의 위치 또한 규명되어야 한다. 이는 다음과 같은 식을 통해 변화점으로서 가능한 모든 각 후보 시점에 대해 사후확률을 계산하고, 그 가운데 가장 큰 값에 대응하는 시점을 변화점(r)으로서 채택한다.

$$f(r|\boldsymbol{x})=\frac{m_1(r,\boldsymbol{x})}{m_1(\boldsymbol{x})},\quad r=1,2,...,T-1.$$

여기서 채택된 변화점(r)은 모수추정 단계에서 초기값으로서 사용된다.

2.5 최적 모형선택

이제 변화점이 하나 이상 존재하는 경우에 변화점을 찾는 절차에 대하여 설명해 보기로 하자.

먼저, 전체 관측치열에 대해 하나의 변화점(r_1)이 있는지 검정하고 변화모형(M_1)이 선택되면 다음으로는, 첫 번째 변화점 r_1을 중심으로 분리한 전후의 새로운 관측치열에 대해 각각 검정을 수행한다. 이러한 과정을 새로이 정의된 관측치열에 대한 모형선택 결과로서 모두 무변화모형(M_0)이 선택될 때까지 반복한다. 이렇게 모형선택 과정을 반복적으로 수행하여 다중변화점 모형(multiple change point model)을 식별할 수 있다.

만약 관측치열에 가장 적합한 모형으로서 K개의 변화점들($r_1,r_2,...,r_K$)이 있는 것으로 탐지되었다면 다중 변화점모형 M_K는 일반적으로 다음과 같은 모형식으로 표현될 수 있다.

$$M_K:\ X_t\sim\begin{cases}b(n_t,p_0), & t=1,2,...,r_1,\\ b(n_t,p_1), & t=r_1+1,r_1+2,...,r_2,\\ \vdots\\ b(n_t,p_{K-1}), & t=r_{K-1}+1,r_{K-1}+2,...,r_K,\\ b(n_t,p_K), & t=r_K+1,r_K+2,...,T.\end{cases}$$

3. 모수추정

2절의 모형선택 단계에서 관측치열에 최적인 모형으로서 변화점이 K개인 변화모형이 선택되었다면, 다음 단계로는 K개의 변화점들 및 각 변화점을 전후로 하여 새로 정의되는 관측치열들에 대한 $(K+1)$개의 부적합률들($p_0,p_1,...,p_K$)을 추정하여야 한다. 이를 위한 절차로서 먼저 모수들의 결합사후확률분포로부터 완전조건부 사후확률분포를 구한 후, 각 조건부 사후확률분포들을 이용하여 M-H 알고리즘을 포함하는 깁스샘플러를 구성할 것이다.

3.1 완전조건부 사후분포

M_K모형에서 모수에 대한 결합사후분포는 모

수 $(r_1,r_2,...,r_K, p_0,p_1,...,p_K)$에 대한 사전분포와 관측치열에 대한 우도함수를 결합시킴으로서 얻어진다.

본 논문에서는 모수에 대한 사전분포로서 각 $p_k(k=0,1,...,K)$에 대해서는 $beta(\alpha_k,\beta_k)$분포를, $r_k(k=1,2,...,K)$에 대해서는 균일분포를 가정하였다. 이 때 각 r_k의 범위는 $r_{k-1}<r_k<r_{k+1}$(여기서 $r_0\equiv 1, r_{K+1}\equiv T$로 정의함)로 제한한다. 이를 바탕으로 결합사후분포는 다음과 같이 정의된다.

$$\pi(r_1,r_2,...,r_K, p_0,p_1,...,p_K|\boldsymbol{x}) \propto \pi_k(r_1,r_2,...,r_K) \times \prod_{k=0}^{K} p_k^{\alpha_k+\sum_{r_k+1}^{r_{k+1}}x_t-1}(1-p_k)^{\beta_k+\sum_{r_k+1}^{r_{k+1}}n_t-\sum_{r_k+1}^{r_{k+1}}x_t-1}$$

이와 같은 결합사후분포로부터 모수에 대한 완전조건부 사후분포는 각각 다음과 같이 유도된다.

$$[p_k \mid r_1,r_2,...,r_K, p_i, i\neq k, i=0,1,...,K] \sim beta\left(\alpha_k+\sum_{r_k+1}^{r_{k+1}}x_t,\ \beta_k+\sum_{r_k+1}^{r_{k+1}}n_t-\sum_{r_k+1}^{r_{k+1}}x_t\right)$$

$$[r_1,r_2,...,r_K \mid p_0,p_1,...,p_K,\boldsymbol{x}] \sim h(r_1,r_2,...,r_K) = \pi_k(r_1,r_2,...,r_K) \times \prod_{k=0}^{K} p_k^{\alpha_k+\sum_{r_k+1}^{r_{k+1}}x_t-1}(1-p_k)^{\beta_k+\sum_{r_k+1}^{r_{k+1}}n_t-\sum_{r_k+1}^{r_{k+1}}x_t-1}$$

3.2 깁스샘플러의 구성

위의 완전조건부 사후분포를 이용하여 깁스샘플러를 구성한다. 각 변화점 $r_k(k=1,2,...,K)$에 대한 추정을 위해 M-H 알고리즘이 적용된다.

Gibbs Sampling Algorithm

[단계1: 초기화단계] 변화점과 부적합률 모수에 대한 초기값 $\underline{r}^{(0)}=\{r_1^{(0)},r_2^{(0)},...,r_K^{(0)}\}$, $\underline{p}^{(0)}=\{p_0^{(0)},p_1^{(0)},...,p_K^{(0)}\}$를 정한다. $\underline{r}^{(0)}$는 모형선택 단계에서 결정되고, $p_k^{(0)}=\sum_{r_k^{(0)}+1}^{r_{k+1}^{(0)}}x_t \Big/ \sum_{r_k^{(0)}+1}^{r_{k+1}^{(0)}}n_t$, $k=0,1,...,K$로 정한다.

[단계2: 반복단계] 다음 단계를 $I(i=1,2,...,I)$번 반복수행한다.

[단계2-1] 각 p_k의 조건부 사후분포로부터 $p_k^{(i)}$의 난수값을 발생시킨다.

$$p_k^{(i)} \sim beta\left(\alpha_k+\sum_{r_k^{(i-1)}+1}^{r_{k+1}^{(i-1)}}x_t, \beta_k+\sum_{r_k^{(i-1)}+1}^{r_{k+1}^{(i-1)}}n_t-\sum_{r_k^{(i-1)}+1}^{r_{k+1}^{(i-1)}}x_t\right), \quad k=0,1,2,...,K.$$

[단계2-2] ***M-H 알고리즘***을 이용하여 $\underline{r}_k^{(i)}=\{r_1^{(i)},r_2^{(i)},...,r_K^{(i)}\}$의 난수값을 얻는다.

[단계2-2-1: 초기화단계] $\underline{r}_{(m)}=\{r_{1(m)},r_{2(m)},...,r_{K(m)}\}$의 초기값 $\underline{r}_{(0)}=\{r_{1(0)},r_{2(0)},...,r_{K(0)}\}$에는 $\underline{r}^{(i-1)}=\{r_1^{(i-1)},r_2^{(i-1)},...,r_K^{(i-1)}\}$ 값을 대입하고, $\underline{r}^{(i)}=(r_1^{(i)},r_2^{(i)},...,r_K^{(i)})$에 대한 적절한 조건부 전이확률함수 $g(\cdot\mid\cdot)$를 설정한다.

[단계2-2-2: 반복단계] 다음의 과정을 $M(m=1,2,...,M)$번 반복수행한다.

1. $g(\underline{r}^*\mid\underline{r}_{(m-1)})$로부터 $\underline{r}^*=(r_1^*,r_2^*,...,r_K^*)$를 발생시킨다.
$r_{1(m-1)}-c<r_1^*<r_2^*<\cdots<r_k^*<r_{k(m-1)}+c$,
(c는 임의의 상수임)
2. 전이확률 α를 계산한다.
$$\alpha=\min\left\{1,\frac{h(\underline{r}^*)\,g(\underline{r}_{(m-1)}\mid\underline{r}^*)}{h(\underline{r}_{(m-1)})\,g(\underline{r}^*\mid\underline{r}_{(m-1)})}\right\}.$$
3. $U(0,1)$분포로부터 난수 u를 발생시킨다.
4. $\underline{r}_{(m)}=\begin{cases}\underline{r}^*, & u\le\alpha \text{ 인 경우,}\\ \underline{r}_{(m-1)}, & u>\alpha \text{ 인 경우.}\end{cases}$
5. 최대 M번 반복하여 얻은 $\underline{r}_{(M)}$ 값을 $\underline{r}^{(i)}$ 값으로 대체시킨 후, [단계2]를 반복수행한다.

4. 실제 자료 분석

얼린 오렌지주스 농축액을 담는 6온스(oz) 캔

이 생산되는 공정에서 캔의 접합부위에 이상이 생긴 경우에 이는 부적합품으로 분류된다. 공정상태 관리를 위한 자료를 얻기 위해 가동되고 있는 기존의 공정상태에서 30분 간격으로 50개(n)씩을 30회(T) 반복하여 1차 표본추출하였다. 이로부터 공정의 부적합률을 추정한 결과 $\bar{p}_0 = .231$이었고, 자료들 중 매우 특이한 두 개(15, 23번째)를 제외하더라도 $\bar{p}_0 = .215$로서 부적합률은 상당히 높은 수준이었다. 이에 대해 적절한 조치를 취한 후 동일한 방법으로 64회 반복하여 2차 표본을 얻었다. 이에 대한 부적합률의 추정치는 $\bar{p}_1 = .1108$로서 1차 표본검사에서 보다 크게 감소된 변화 양상을 보였다(자료출처: Montgomery, 2001).

본 논문에서는 1·2차 자료를 합하여 제안된 절차를 통해 부적합률에 대한 분석을 수행하였고 <표1>에 결과를 제시하였다. 변화점(r)으로는 33번째에서 하나가 나타났으며, 변화점을 전후로 부적합률의 초기 추정값은 각각 $\bar{p}_0 = .227$, $\bar{p}_1 = .107$ 이었고, 이를 기초로 하여 모수 p_0, p_1, r에 대한 사후분포가 추정되었다. 각 자료에 대한 <그림1>를 살펴보면 공정을 조절한 시기인 $T = 30$에서 보다는 실제적으로 $T = 33$에서 변화가 더 크게 발생한 특징을 찾아내고 있다.

5. 결론 및 토의

본 논문에서는 먼저 생산공정에서 생산되어져 나오는 일련의 제품들에 대해 일정수준의 부적합률이 어느 시점에서 변화했는지, 몇 개의 변화점이 있는지를 탐지할 수 있는 베이지안 모형선택 기법이 제안되었다. 또한, 관측치열에 변화점(들)이 존재하는 경우에는 그 시점(들)을 추정하고 이를 중심으로 관측치열을 분리하여 각 부분 관측치열에 대한 모부적합률을 추정함으로서 자료에 적합한 모형을 찾아낼 수 있는 베이지안 모수추정 기법이 제안되었다. 모수추정 단계에서는 깁스샘플링 및 M-H 알고리즘이 사용되었다. 제안된 방법은 모의실험 및 실제 자료에서 적합한 결과를 보임으로서 이론적 타당성이 입증되었다.

참 고 문 헌

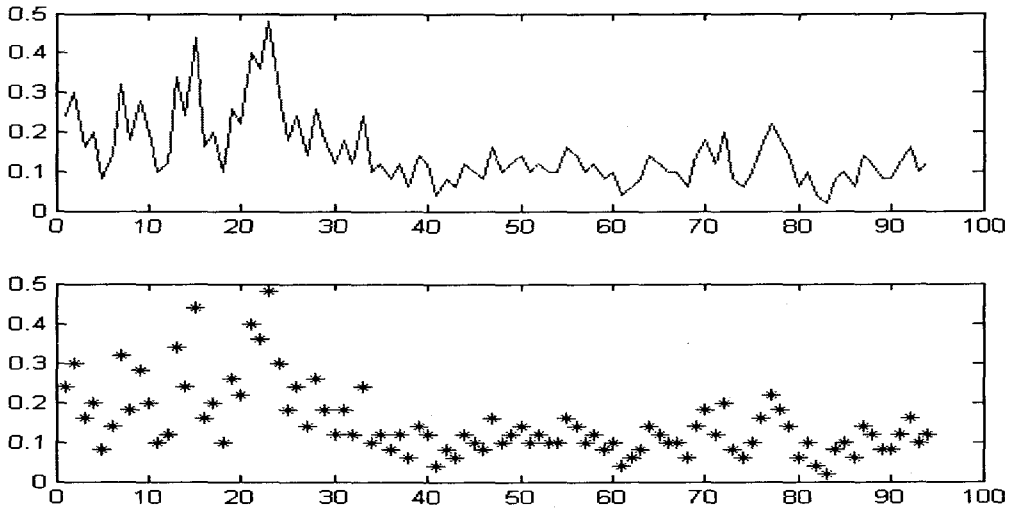
[1] 김경숙, 김희정, 나명환, 손영숙 (2006), "부적합률의 다중검정을 위한 베이지안절차", 품질경영학회지에 투고된 논문.

[2] Andrew B. Gelman, John S. Carlin, Hal S. Stern and Donald B. Rubin. (2000), Bayesian data analysis, Campman & Hall/CRC.

[3] Montgomery, D. C.(2001), Introduction to Statistical Quality Control, fourth edition, John Wiley & Sons, Inc.

[4] The MATH WORK Inc.(2002), MATLAB/Statistics Toolbox, Version 6.5, Natick, MA.

* 변화점 r의 초기값에서 ()안의 값은 $r = 33$일 때의 사후확률로서 가능한 모든 후보 시점들 ($r = 1,2,...,93$)의 사후확률 가운데 최대값에 해당함.



<그림4> 오렌지주스 캔의 부적합률자료

○ 표시는 추정된 변화점 위치에 있는 자료를 나타냄.

<표4> 오렌지주스 캔의 부적합률 자료에 대한 모형선택 및 모수추정 결과

모형의 사후확률		모수	초기값	사후분포		
$P(M_0 \mid x)$	$P(M_1 \mid x)$			mean	S.D.	median
.0000	1.0000	p_0	.227	.228	.010	.228
		p_1	.107	.109	.004	.108
		r	33 (.360)	33.150	.386	33.000