총계오차판별기법 및 데이터 보정 연구흐름

포항공대 화학공학과 김 정 환

서울대학교 응용화학부 한 종 훈

본 강좌에서는 다양한 총계오차판별기법과 특징에 관하여 살펴보도록 하고, 데이터보정 및 총계오차판별에 관한 연구흐름을 소개하도록 한다.

1. 총계오차가 존재하는 경우의 데이터 보정

총계오차가 존재하는 경우, 총계오차를 제거하지 않고 데이터보정을 수행하지 않은 경우 어떤 영향을 미치는지를 예를 통하여 확인해 보자. 지난 회에 소개되었던 시스템에서 스트림 2 에 +4 의 총계오차가 존재하는 경우에 총계오차를 제거하지 않고 데이터보정을 수행한 경우와 총계오차를 제거하고 데이터보정을 수행한 경우의 결과는 <표 1>과 같다.

<표 1> 스트림 2 에 총계오차가 존재하는 경우의 데이터보정 수행결과

스트림	참값	측정값	모든 스트림 측정값을 사용하여 데이터 보정을 수행한 경우	스트림 2 를 제외하고 데이터 보정을 수행한 경우
1	100	101.91	100.89	100.23
2	64	*68.45* **(+4)**	65.83	64.53
3	36	34.65	35.05	35.71
4	64	64.20	65.83	64.53
5	36	36.44	35.05	35.71
6	100	98.88	100.89	100.23

두 결과를 비교해 보면 스트림 2 를 제외하고 데이터보정을 한 경우가 모든 스트림에 대하여 참값과의 오차가 작음을 알 수 있다. 비록 스트림 2 에 총계오차가 있었지만 스트림 2 를 포함한 데이터보정결과를 살펴보면 특별히 스트림 2 의 보정값의 오차가 데이터보정 수행 전에 존재했던 총계오차 값만큼 큰 것이 아니고

전체적으로 모든 스트림에서 오차가 증가했음을 알 수 있다. 이는 총계오차를 제거하지 않고 데이터보정을 하는 경우 총계오차가 데이터보정에 의하여 분산되기 때문이며 이러한 현상을 *smearing effect* 라고 부른다.

따라서, 데이터보정을 수행하기 이전에 총계오차를 판별하여 제거한 후 데이터보정을 하는 것이 필수적이다. 그렇다면 총계오차를 어떻게 판별할 것인가? 총계오차를 판별하기 위한 여러 가지 방법들이 제안되었는데, 대표적인 방법으로서는 Global Test, Constraint Test (Nodal Test), Measurement Test, Generalized Likelihood Ratio Test 등이 있다. 이러한 여러 가지 방법들은 모두 통계적 방법을 이용하여 데이터에 존재하는 특이점을 찾아내는 여러 기법을 이용하고 있다.

2. 총계오차판별 기법

총계오차는 크게 2 종류로 나눌 수 있는데 측정장치의 성능저하에 의한 것과 제약조건과 관련된 것이 있다. 전자에 해당하는 것으로는 센서의 바이어스, miscalibration, 고장 등이 있으며, 후자에 해당하는 것으로는 공정 내 장치에서의 알 수 없는 물질 및 에너지 손실이나 공정 파라미터의 부정확으로 인한 모델의 부정확성이 있다. 총계오차판별을 위해서 개발된 여러 방법들은 이러한 두 종류의 총계오차를 찾아내는 것이 목적이다.

총계오차의 판별을 위해서 주로 쓰이는 방법은 통계학에서 많이 사용하는 가설의 검정이다. 즉, 귀무가설 (null hypothesis)은 “총계오차가 존재하지 않는다” 이고, 대립가설 (Alternative hypothesis)은 “총계오차가 존재한다” 이다. 이 때 통계학에서 검정통계량 (test statistic)을 이용하여 가설의 참, 거짓을 결정하듯이 적정한 검정통계량을 이용하여 총계오차의 존재를 판별하게 된다. 물론, 오차가 존재가능한데, 총계오차가 존재하지 않는데 존재한다고 잘못된 알람을 울리게 되는 Type I 오차와 총계오차가 존재하는 데 존재하지 않는다고 판단을 하는 Type II 오차의 두 가지가 모두 가능하다. 대표적인 총계 오차판별기법에 대하여 알아보도록 하자.

2.1. Global Test

Global Test 에서는 수식 (1)의 검정통계량을 이용한다. 이 식에서 r 은 잔차 (residual) 벡터이며 V 는 r 의 공분산행렬 (covariance matrix)로서 다음과 같이 정의된다. 대상시스템에 총계오차가 존재하지 않는다면 (즉, 귀무가설이 참이라면), 잔차벡터인 r 은 평균이 0 이고 분산이 V 인 정규분포를 따르게 된다.

$$\gamma = \mathrm{r}^T \mathrm{V}^{-1} \mathrm{r} \quad (1)$$

$$\mathbf{r} = \mathbf{Ay\text{-}c} \quad (2)$$

A: Constraint Matrix, **y**: vector of n measurement, **c**: constant nonzero vector

대상시스템이 선형시스템이며 물질수지를 고려하는 경우에 c 는 영벡터이며 대상시스템의 제약조건식들을 행렬식으로 표현하면, Ax = c 의 형태로 표현할 수 있다. 여기서, A 행렬의 각 행들은 대상시스템 내 각 제약조건들을 표현하게 되며, 각 장치에의 입력, 출력 관계에 의해서 행렬원소들은 +1, 0, -1 의 값을 갖는다. 검정통계량은 ν 의 자유도를 갖는 χ^2 분포를 따르므로 α 의 유의수준 (level of significance)에 대한 기각값은 $\chi^2_{1-\alpha,\nu}$ 이 되어, 이 값과 검정통계량값을 비교하여 총계오차의 존재여부를 결정하게 된다. Global Test 의 검정통계량은 대상 시스템의 모든 제약조건에 대하여 잔차를 계산하여 그 총합으로서 총계오차를 판별한다. 그러므로, 대상시스템에 총계오차가 존재한다는 사실만 밝힐 뿐 어디에 존재하는지 찾아낼 수는 없다.

지난 회에 소개되었던 시스템을 대상으로 Global Test 기법을 이용하여 총계오차를 판별해 보면 다음과 같다. 제약조건행렬 A 및 잔차, 공분산행렬은 각각 다음과 같다.

$$A = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 \end{bmatrix}, \quad r = \begin{bmatrix} -1.19 \\ 4.25 \\ -1.79 \\ 1.76 \end{bmatrix}, \quad V = \begin{bmatrix} 3 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 3 \end{bmatrix}$$

검정통계량 값은 16.694 가 되며, 5%의 유의수준을 고려할 때 (즉, 타입 I 에러, 즉 귀무가설인 총계오차가 없다가 맞는데 총계오차가 있다고 잘못된 판단을 내릴 확률이 5% 미만이 되는 경우), 자유도 4 인 카이제곱분포는 9.488 이므로 대상시스템에 총계오차가 존재한다는 결론을 내리게 된다.

2.2. Constraint Test (Nodal Test)

Global Test 에서는 각 제약조건들의 잔차총합을 이용하여 총계오차를 판별했다면 Constraint Test (Nodal Test)에서는 각 제약조건별로 잔차를 비교하여 총계오차의 존재여부를 결정짓는다. 따라서 검정통계량은 다음과 같다.

$$z_{r,i} = \frac{|r_i|}{\sqrt{V_{ii}}} \qquad i = 1,..,m \quad \text{……..……..……...} \quad (3)$$

검정통계량은 시스템에 총계오차가 존재하지 않는다는 가정하에서 표준정규분포를 따르므로 유의수준 α 에 대하여 $z_{1-\alpha/2}$ 의 기각값과 비교하여 총계오차를 판별하게 된다. 동일한 시스템에 대하여 Constraint Test 기법을 이용하면 검정통계량값이 [0.687, 3.0052, 1.2657, 1.0161]로서 5% 유의수준에 대한 양측검정으로의 기각값이 1.96 이므로 2 번째 제약조건에 대하여 총계오차로 판별이 됨을 알 수 있다.

2.3. Measurement Test

Measurement Test 는 측정값에 대한 조정값을 검정통계량으로 이용하는 방법이다. 즉, 검정통계량은 다음과 같다.

$$z_{a,j} = \frac{|a_j|}{\sqrt{\overline{W}_{jj}}} \qquad j = 1,..,n \quad \text{……..……..……...} \quad (4)$$

$a = y - \bar{x}$, where y: 측정값, $\bar{x}$: 보정값 ……..……... (5)

$\overline{W} = \mathrm{cov}(a)$ ……...……... (6)

시스템에 총계오차가 존재하지 않는 경우,

$a \sim N(0, \overline{W})$ ……...……... (7)

이므로 유의수준에 해당하는 표준정규분포값과 각 유량의 조정값을 비교하여 총계오차를 판별하게 된다. 동일한 시스템에 대하여 Measurement Test 기법을 이용하면 각 조정값에 대한 검정통계량값이 [1.2533, 3.2047, 0.494, 2.0004, 1.6983, 2.4577] 이며, 따라서 유의수준 5%의 양측 검정값 1.96 과 비교해 보면 2,4,6 번의 조정값에서 총계오차가 발생하고 있음을 알 수 있다. 이런 경우에, 3 개의 센서에서 모두 총계오차가 존재하여 이런 결과가 나왔다고 생각할 수도 있겠지만 앞에서 설명한 smearing effect 에 의해서 여러 곳에서 총계오차가 발생했다고 보는 것이 보다 타당하다. 그러므로, 가장 총계오차가 크게 나타나는 2 번째 measurement 에만 총계오차가 있다고 판단하여 이를 제외하고 데이터보정을 수행한 후에 검정통계량을 확인해 보면 총계오차가 없음을 확인할 수 있고, 2 번째 센서에 총계오차가 있었음을 알 수 있다. 이와 같이 Measurement Test 는 앞의 두 가지 방법과 비교하여 총계오차가 존재할 경우 총계오차 존재 위치를 찾을 수 있다는 특징이 있으며 일반적으로 많이 사용되는 방법이다.

이외에도 Maximum likelihood ratio principle 을 이용하는 GLR (Generalized Likelihood Ratio) Test 및 주성분분석을 이용하는 방법 등 다양한 방법이 있다.

3. 데이터 보정 및 총계오차 판별 연구흐름

데이터보정 및 총계오차판별에 관한 연구는 1960 년대 초반부터 연구되기 시작하였으며 약 200 편 이상의 연구논문들이 발표되었다. 주요연구 흐름에 관하여 정리하면 다음과 같다.

- ✓ Kuehn & Davidson (1961) 모든 변수들이 측정 가능한 경우의 선형물질수지식에 대한 해를 유도
- ✓ Vaclavek (1968,1976); Observability 와 Redundancy 등의 데이터 보정에서 다루는 중요한 개념들이 소개
- ✓ Almasy & Sztano (1975); Global Test 및 Measurement Test 를 이용한 총계오차판별법 제안
- ✓ Mah et al (1976); Constraint Test 를 이용한 총계오차판별법 제안
- ✓ Crowe et al (1983); 데이터보정 문제에서 제약조건 내에 존재하는 미측정변수를 제거하기 위하여 Projection Matrix 를 이용하는 방법 제안
- ✓ Knepper & Gorman (1980); parameter estimiation 과 iterative 기법을 이용하여 비선형 데이터보정문제를 푸는 방법 제안
- ✓ Narasimhan & Mah (1987); GLR (Generalized Likelihood Ratio)를 이용한 총계오차판별 기법 제안
- ✓ Tjoa & Biegler (1991); SQP (Successive Quadratic Programming)기법을 이용하여 데이터보정 및 총계오차판별문제를 푸는 방법 제안
- ✓ Almasy (1991); 칼만 필터를 이용하여 선형 시스템에 대한 동적 데이터보정 문제를 푸는 것을 제안
- ✓ Tong & Crowe (1995); 주성분 분석법을 이용한 총계오차 판별 기법 제안

다음 회에서는 총계오차판별 및 데이터보정의 현장적용사례 및 데이터보정 시스템에 관하여 살펴보도록 하겠다.