새로운 2차원화 방법과 다차원 독립성분분석을 이용한 회분 공정모니터링 기술

(Batch process monitoring using another unfolding method and multiway independent component analysis)

유창규

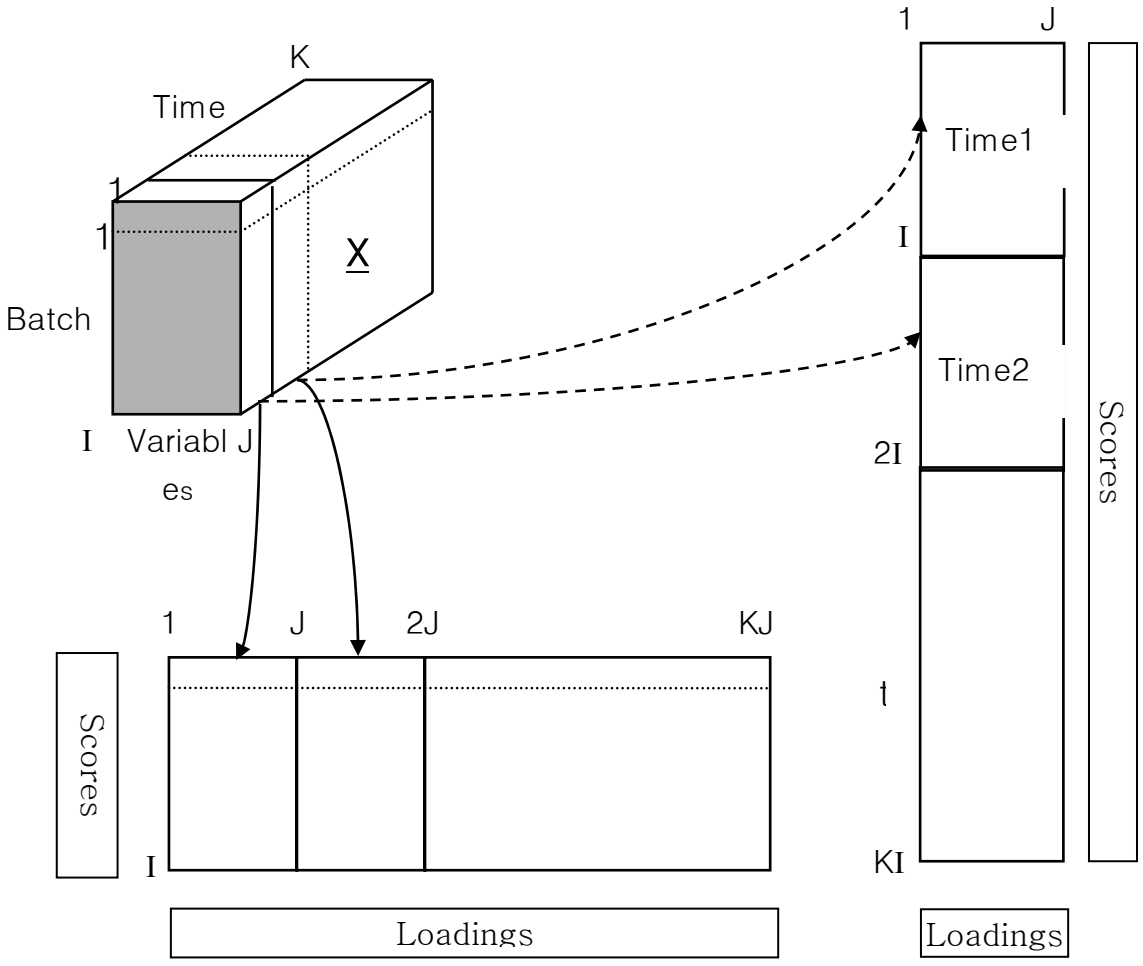
본 보고서는 저자가 2002년도에 제공한 해외 IP 정보를 연구 내용을 독자들의 편의를 위해 다시 재정리하고 회분 페니실린공정에 적용한 결과를 포함한다. 이 내용은 power point file로도 제공된다. 다음달의 연구 보고서는 기존의 MPCA와 같은 unfolding method를 ICA에 적용하여 두 가지 MPCA/MICA 방법의 비교와 공정 모니터링기술에 대한 정보를 제공할 예정이다.

화학산업체 뿐만 아니라 바이오, 제약, 반도체 등의 산업체에 사용되는 회분식 공정은 다품종 소량생산 공정이고 고부가가치 제품생산에 적합하기 때문에 산업계에서 중요한 역할을 담당해왔다. 회분식 공정은 원료를 반응기에 주입하는 단계와 시간에 따라 변하는 온도, 압력, 교반, 유량 등을 제어하면서 공정이 진행되는 단계, 마지막으로 일정 전환율에서 반응기로부터 제품을 뽑아내는 단계로 이루어져 있다. 일반적으로 회분식 공정은 배치마다 공정변수의 변화가 생기는데 이는 주로 평균추이를 벗어나는 변수들의 움직임, 원료 주입상의 에러, 불순물로 인해 생기는 외란 등으로 일어난다. 회분식 작업 중에 일어나는 이상 현상들은 한 배치 또는 일련의 여러 배치들에서 나오는 제품의 질을 크게 떨어뜨린다. 그럼에도 불구하고, 아직까지 대부분의 실제 산업체 현장에서는 한 회분 공정이 끝나면 제품의 품질을 따로 측정하여 그 공정이 잘 이뤄졌는지만 판단하는 오프라인 형식의 모니터링이 적용되고 있다. 따라서 제품의 품질이 나쁘게 나왔을 때 언제 어디에서 무엇 때문에 이상현상이 일어났는지 판단할 수 있는 방법이 없기에 다음의 회분 작업에서도 이상현상을 감지하지 못한 채 품질이 크게 떨어지는 제품을 생산하는 경우가 많다. 이에

회분식 공정이 끝나기 전에 이상현상을 감지하고 이를 제거할 수 있는 온라인 형식의 모니터링 및 진단 방법이 절실히 요구되고 있다.

최근에 들어서야 데이터 투영법을 이용한 온라인 형식의 회분식 공정 모니터링 기술이 개발되었다. 회분식 공정의 다변량 통계 투영 방법은 데이터를 압축하여 저차원 공간을 형성하고 이에 데이터를 투영하여 정보를 도출하는 방법이다. Nomikos와 Macgregor(1994)는 측정된 공정 변수들을 이용한 회분식 공정 감시를 위해 다차원 주 성분 분석(Multiway Principal Component Analysis)에 기반한 모니터링 방법을 처음으로 제시하였다. 그 후로 비선형 주성분 분석, 동적 주성분 분석 등 여러 다양한 주성분 분석을 이용한 회분식 공정이 개발되어 왔고 실제 산업체에 적용하여 성공한 사례도 나오고 있다. Nomikos와 Macgregor가 제안한 다차원 주성분 분석은 그림 1의 approach A에서 보듯이 3차원 배열의 회분식 데이터 $\underline{X}(I \times J \times K)$를 시간에 따라 잘라내어서 배치들과 변수의 데이터($I \times J$)를 시간에 따라 옆으로 2차원적으로 재배열하여 이 데이터($I \times KJ$)에 주성분 분석을 적용한다. 재배열한 2차원 데이터에서 열(column)의 평균들을 빼면 평균추이를 빼는 것과 같기 때문에 비선형적인 공정데이타의 성격을 미리 어느 정도 제거할 수 있다. 주성분 분석을 통해 나오는 스코어 벡터는 각 회분의 정보를 축약한 채로 가지게 되며 로딩 벡터는 시간에 따른 변수의 관계 정보를 축약한 채로 가지게 된다.하지만 이 방법은 온라인 모니터링에 적용 시 몇 가지 단점을 가지고 있다. 우선 재배열된 테스트 데이터의 길이가 모델을 구성할 때 구한 로딩 벡터의 길이와 같아야지만 투영된 스코어 값을 구하게 되는데 회분 시작부터 현재 시간까지의 데이터만 있기 때문에 이를 적용할 수 없으므로 회분이 끝날 때까지의 변수들의 값을 예측하여 계산을 해야 하므로 정확한 스코어 값을 구할 수 없다. 또한 각 회분의 시간이 같아야 하는데 실제의 경우 각 회분마다 공정 시간이 다르므로 이를 융합할 수 있는 방법이 따로 필요하게 된다. 이런 단점을 보완할 수 있는 재배열 방법은 그림 1의 Approach B의 경우가 되겠다. 시간

에 따라 잘라내어서 배치들과 변수의 데이터($I \times J$)를 시간에 따라 아래로 2차원적으로 재배열하여 이 데이터($IK \times J$)에 주성분 분석을 적용한다. 이 경우는 온라인 모니터링 적용 시 회분이 끝날 때까지의 값들을 예측할 필요가 없으며 각 회분의 시간이 달라도 된다. 하지만 이 방법으로 재배열한 뒤 열(column)의 평균을 뺀다 해도 평균추이를 빼는 것이 아니라 각 배치 및 시간에 따른 전체 평균을 빼게 되어 비선형적 성격의 데이터를 없앨 수 없다. 게다가 로딩 벡터는 시간에 따른 변수의 관계를 축약하는 게 아니라 전체 시간동안 변수의 평균관계를 축약하여 나타내기 때문에 시간에 따른 변수 관계의 변화를 제대로 파악할 수 없는 문제점이 생길 수 있다.



방법 A 방법 B

그림 1. 두 가지 재배열 방법

본 연구 기본 아이디어는 Approach B에 기반 한 방법을 택했으며 Approach A의 장점을 약간 접목시키고 여기에 주성분분석보다 독립요소분석(Independent Component Analysis)을 적용하는 모니터링 방법을 제시하였다.

대부분의 공정에서 측정된 데이터들은 차원이 크고 각 변수간의 상관성이 크기 때문에 데이터 축약을 위해선 잠재해 있는 독립 요소를 찾아내야 한다. 독립요소분석은 이런 독립 요소를 찾아내는데 유용한 방법이 된다. 독립요소분석은 주성분분석의 확장된 개념으로 볼 수 있다. 주성분분석이 평균과 분산에만 기반하여 변수들의 상호 관계를 없앤 축을 찾아내어 분석하는 방법에 비하여 독립요소분석은 평균 및 분산뿐만 아니라 첨도(kurtosis), 비대칭도(skewness) 등 고차원의 통계학적 성질에 기반 하여 변수들의 상호 관계를 없애기 때문에 다변량 데이터에서 독립적인 요소만을 따로 뽑아낼 수 있다. 더구나 주성분 분석에 기반하여 모니터링을 하게 되는 경우 차트의 제한선을 설정할 때 주성분분석 후 나오는 스코어 값들이 정규분포를 따른다는 가정 하에 구하기 때문에 문제점이 있게 된다. 한편, 독립요소분석은 처음부터 다변량 데이터 속의 독립 요소들이 정규분포를 따르지 않는다는 가정을 한 채 찾기 때문에 실제 데이터를 처리하는 데 있어서 많은 장점들을 가질 수 있다. 그림 2은 주성분분석과 독립성분분석을 이용할 경우 각 방법의 복원신호의 차이를 보여준다.

독립요소분석은 아래 식과 같이 다변량 데이터 **X**(변수×샘플) 만을 가지고 분해행렬 **W**를 예측하여 독립요소 **S**를 구하게 된다.

$$\mathbf{S} = \mathbf{W}\mathbf{X} \tag{1}$$

물론 독립요소분석에서도 데이터 차원을 줄일 수 있는데 주성분분석처럼 각 스코어 벡터의 분산의 크기로 축의 중요성을 따지기가 힘들다. 독립요소분석에서 나오는 **S**의 행들은 모두 분산의 크기가 1 이 되며 중요한 요소를 찾아내기가 힘들다. 본 연구에선 **X**의 변화가 **W**를 통해 **S**에 전파됨을 이용하여 **W**의 L2 norm이 큰 행에 해

당하는 **S**의 행을 구하여 데이터를 축약하였다. 독립요소분석을 이용한 온라인 모니터링 방법은 다음과 같다. 우선 정상 상태의 데이터 $\mathbf{X}_{normal}$에서 독립요소분석 알고리즘을 이용하여 $\mathbf{S}_{normal} = \mathbf{W}\mathbf{X}_{normal}$ 의 식으로부터 $\mathbf{S}_{normal}$과 $\mathbf{W}$를 찾아낸다. 그런 다음 테스트 데이터를 정상 상태일 때 구한 $\mathbf{W}$에 투영함으로서 새로운 $\mathbf{S}_{new}$를 구하게 된다. 주성분분석을 이용한 모니터링에서 사용하는 T^2 와 SPE 차트처럼 독립요소분석을 이용한 모니터링에서는 다음과 같이 I^2 와 SPE를 구할 수 있다.

$$I^2(k) = \mathbf{s}_{new}(k)^T \mathbf{s}_{new}(k) \tag{2}$$

$$SPE(k) = \sum_{j=1}^{d} (x_j(k) - \hat{x}_j(k))^2 \tag{3}$$

식(2)와 (3)처럼 두가지 statistics를 구할 수 있다.

본 아이디어에서 제안한 회분 공정 모니터링 방법은 그림 3와 같다. 우선 3차원 배열의 회분 데이터 $\underline{X}(I \times J \times K)$를 Approach A로 재배열한 다음 열의 평균을 빼서 회분 데이터의 평균 추이를 빼어 비선형성을 어느 정도 제거한다. 그런 다음 이 데이터($I \times KJ$)에서 Approach B로 다시 재배열하여 독립요소분석을 적용한다. 이럴 경우 Approach A 사용 시 평균추이를 제거하는 장점을 살린 채 Approach B의 장점을 적용할 수 있다. 물론 이 아이디어는 MPCA에도 적용할 수 있다. 따라서 온라인 모니터링 적용 시 각 회분의 미래 값들을 예측할 필요가 없고 회분 시간의 길이가 달라도 된다. 더욱이 주성분분석보다 독립요소 분석을 사용하기 때문에 다변량 데이터에 내재되어 있는 요소를 보다 효율적으로 찾아냄으로써 모니터링에 적용 시 효과적으로 이상현상이 일어나는 경우들을 쉽게 감지할 수 있다. 또한 MICA모델 구축을 위해서 사용된 독립성분들의 I^2 메트릭 값뿐만 아니라 제외된 독립성분들에 대해 I^2를 적용하면 I_e^2 메트릭 값을 계산할 수 있다. *SPE*값이 에러의 평균값만의 변화를 탐지할 수 있는데 비해 I_e^2 메트릭 값은 평균화 분산의 변화를 탐지할 수 있는 능력이 있다. 한편 모니터링 차트의 제한선들은 kernel density estimation을 이용하

여 보다 정확하게 찾아내어 잘못된 alarm을 줄일 수 있다. 표 1은 다차원 독립성분 분석을 이용한 실시간 공정모니터링 방법의 순서가 정리되어있다.

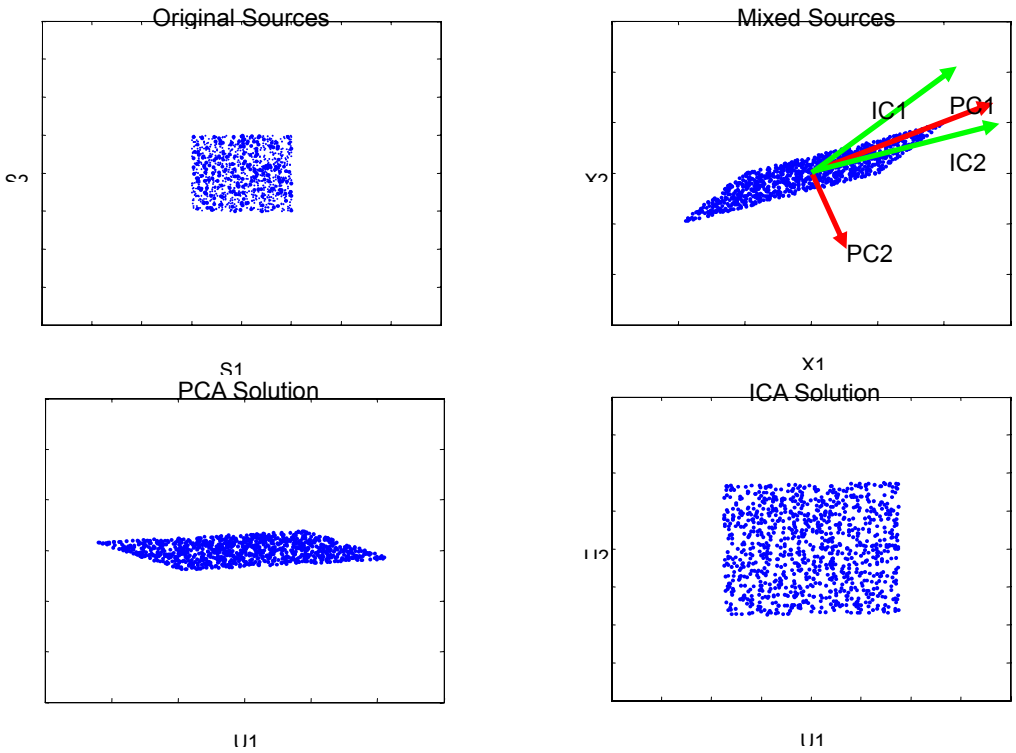


그림 2.주성분 분석과 독립성분분석의 복원 신호의 차이점

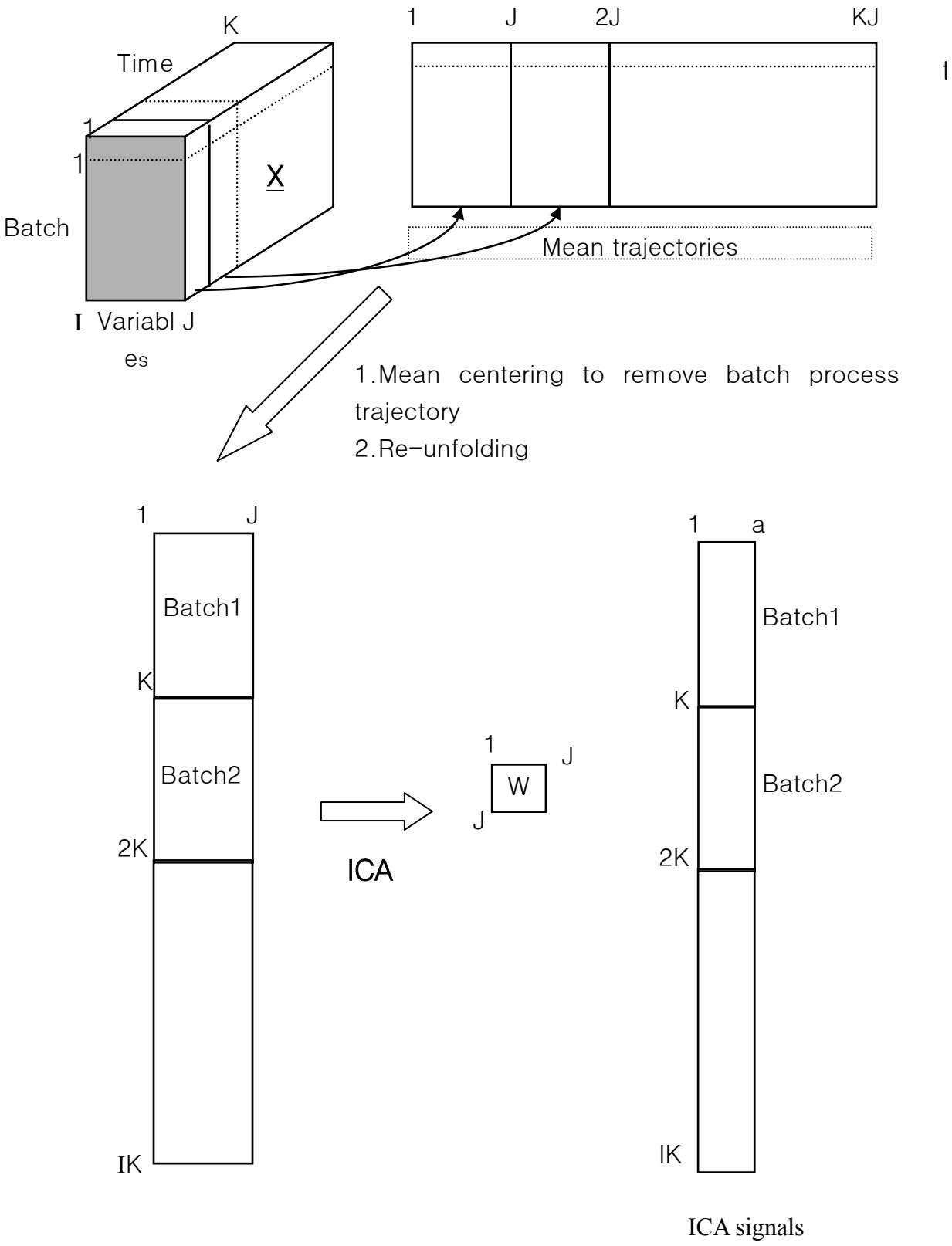


그림 3. 독립요소분석을 이용하는 새로운 회분식 모니터링 과정

표 1. MICA on-line monitoring procedures

A. 정상상태 공정 모델링

1. 배치 공정데이타 $\underline{\mathbf{X}}(I \times J \times K)$를 2차원 배열로($\mathbf{X}(I \times JK)$) 펼침

2. 공정 데이터 $\mathbf{X}(I \times JK)$는 모든 배치에서 각시간에서 각 변수의 평균과 표준편차를 이용하여 표준화

3. 표준화된 2차원 배열을 $\mathbf{X}(I \times JK)$ 제안된 각 변수에 대한 배열로 변환 $\mathbf{X}_{normal}(J \times IK)$

4. Whitening 공정

$$\mathbf{Z}_{normal} = \mathbf{Q}\mathbf{X}_{normal}$$

5. 독립성분분석(ICA) 실행

Obtain $\mathbf{W}$, $\mathbf{B}$, and $\mathbf{S}_{normal}$ from $\mathbf{S}_{normal} = \mathbf{W}\mathbf{X}_{normal} = \mathbf{B}^T\mathbf{Z}_{normal}$.

6. $\mathbf{W}$ 각 행의 크기를 계산하고 크기에 따라 구해진 독립성분의 순서를 정한 후 $\mathbf{W}$를 중요부분과 노이즈 부분으로 분리, $\mathbf{B}$ and $\mathbf{S}_{normal}$도 같은 방법으로 분리 가능

$$\mathbf{W} \rightarrow \mathbf{W}_d, \mathbf{W}_e$$

$$\mathbf{B} \rightarrow \mathbf{B}_d, \mathbf{B}_e$$

$$\mathbf{S}_{normal} \rightarrow \mathbf{S}_d, \mathbf{S}_e$$

7. 3가지 I^2, I_e^2, and SPE metrics을 계산

$$I^2(n) = \mathbf{s}_d(n)^T \mathbf{s}_d(n)$$

$$I_e^2(n) = \mathbf{s}_e(n)^T \mathbf{s}_e(n)$$

$$SPE(n) = \sum_{j=1}^{d} (x_j(n) - \hat{x}_j(n))^2$$

여기서 n은 1부터 IK값을 가지고 $\hat{\mathbf{X}} = \mathbf{Q}^{-1}\mathbf{B}_d\mathbf{S}_d = \mathbf{Q}^{-1}\mathbf{B}_d\mathbf{W}_d\mathbf{X}_{normal}$로 복원됨

8. $I^2(1\times IK)$, $I_e{}^2(1\times IK)$ and $SPE(1\times IK)$ 벡터를 각각 $I^2(I\times K)$, $I_e{}^2(I\times K)$, $SPE(I\times K)$로 재배열

9. 커널밀도추정을 통한 I^2, $I_e{}^2$ and SPE metrics의 제어 한계치를 계산

B. 실시간 공정 모니터링

1. 새로운 배치 데이터의 시간 k까지의 데이터 $\mathbf{X}_{test}(k\times J)$를 $\mathbf{x}_{test}{}^T(1\times Jk)$로 펼친 후 모델링에 사용된 같은 스케일로 표준화

2. 표준화된 $\mathbf{x}^T{}_{new}(1\times Jk)$을 $\mathbf{X}_{new}(J\times k)$로 재배열..

3. $\mathbf{S}_{newd} = \mathbf{W}_d\mathbf{X}_{new}$, $\mathbf{S}_{newe} = \mathbf{W}_e\mathbf{X}_{new}$로부터 $\mathbf{S}_{newd}$와 $\mathbf{S}_{newe}$계산

5. 계산된 $\mathbf{S}_{newd}$와 $\mathbf{S}_{newe}$값으로부터 각 시간의 $I_{newd}{}^2(k)$, $I_{newe}{}^2(k)$, $SPE(k)$ 계산

$$I_{newd}{}^2(k) = \mathbf{s}_{newd}(k)^T\mathbf{s}_{newd}(k),\ I_{newe}{}^2(k) = \mathbf{s}_{newe}(k)^T\mathbf{s}_{newe}(k)$$

$$SPE(k) = \sum_{j=1}^{d}(x_{newj}(k) - \hat{x}_{newj}(k))^2$$

여기서 $\hat{\mathbf{X}} = \mathbf{Q}^{-1}\mathbf{B}_d\mathbf{S}_{newd} = \mathbf{Q}^{-1}\mathbf{B}_d\mathbf{W}_d\mathbf{X}_{new}$

6. 계산된 $I_{newd}{}^2(k)$, $I_{newe}{}^2(k)$, $SPE(k)$과 정상상태 모델과정에서 계산된 제어한계치와 비교하여 공정의 이상여부 판단

C. 기여도 분석 (Contribution Plot)

1. Variable contribution for $I_{newd}{}^2(k)$

$$\mathbf{x}_{cd}(k) = \frac{\mathbf{Q}^{-1}\mathbf{B}_d\mathbf{s}_{newd}(k)}{\left\|\mathbf{Q}^{-1}\mathbf{B}_d\mathbf{s}_{newd}(k)\right\|}\left\|\mathbf{s}_{newd}(k)\right\|$$

2. Variable contribution for $I_{newe}{}^2(k)$

$$\mathbf{x}_{ce}(k) = \frac{\mathbf{Q}^{-1}\mathbf{B}_e\mathbf{s}_{newe}(k)}{\left\|\mathbf{Q}^{-1}\mathbf{B}_e\mathbf{s}_{newe}(k)\right\|}\left\|\mathbf{s}_{newe}(k)\right\|$$

3. Variable contribution for $SPE(k)$

$$\mathbf{x}_{cspe}(k) = \mathbf{x}(k) - \hat{\mathbf{x}}(k)$$

Results and Discussions

제안된 방법은 반회분식 페니실링 생산공정의 시뮬레이션 벤치마크의 공정 모니터링에 응용되었다. 그림 4는 회분식 발효공정으로 이루어진 페니실린 공정을 나타내며 표2는 공정모니터링에 사용된 11개의 공정변수들을 나타낸다. 전체 67 배치가 MPCA/MICA 정상상태 모델 구축을 위해 사용되었다. 각 배치의 시간은 400시간이면 그 중 45시간은 pre-culture stage이고 나머지 355시간은 fed-batch stage이다. MPCA모델은 4개의 주성분을 선택했으며 전체 변화의 62.2%를 설명한다. MPCA의 온라인 모니터링을 위해 Macgregor가 제시한 PCA를 이용하여 missing data를 다루는 세번째 filling method를 사용하였다. MPCA와 비교를 위핸 MICA모델은 4개의 독립성분을 선택했으며 MPCA/MICA모델 둘 다 99% control limit를 사용하여 테스트 하였다. 비정상 배치를 위해 왜란(disturbance)이 도입되었다. 여기서는 기질 공급속도(substrate feed rate)가 시간 100*hr*부터 시간 250*hr*까지 기울기 –0.002를 가지고 선형으로 감소하였다. 그림 5와 6은 각각의 모니터링 결과이다. 작은 크기의 왜란의 경우 MPCA는

feature extraction방법이 Gaussian distribution assumption에 기초하고 Schwart chart에 기초하므로 작은 크기의 외란를 탐지하기는 어려운 특징을 지닌다. 그림 5에서 보면 온라인 MPCA방법은 작은 크기의 외란을 탐지 못한다. 반면에 온라인 MICA방법은 non-Gaussian distribution이라는 feature extraction방법에 기초하기 때문에 공정의 주영향 인자를 더 잘 뽑아낸 모델이 기초한다. 그림 6은 MICA가 정확히 외란을 탐지하는 결과를 보여주고 있다. 그림 7은 외란의 원인을 찾기위해 70시간에서 MICA모델의 I^2, I_e^2 과 SPE charts각각에 대한 contribution plots을 나타낸다.

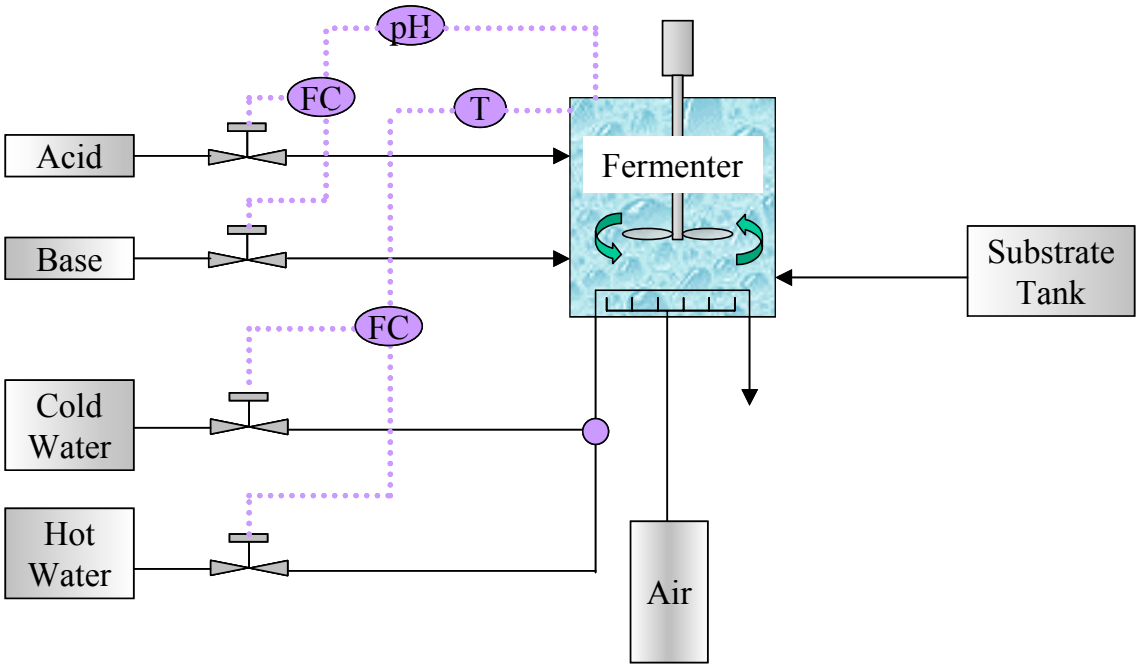


그림 **4.** Flowsheet of the penicillin fermentation process

표2. Variables used in the monitoring of the benchmark model

No.	Variables
1	Aeration rate(L/h)
2	Agitator power(W)
3	Substrate feed rate(L/h)
4	Substrate Feed Temperature (K)
5	Dissolved Oxygen Concentration (g/L)
6	Culture Volume (L)
7	Carbon Dioxide Concentration (g/L)
8	pH
9	Fermentor Temperature (K)
10	Generated Heat (kcal)
11	Cooling Water Flow Rate (L/h)

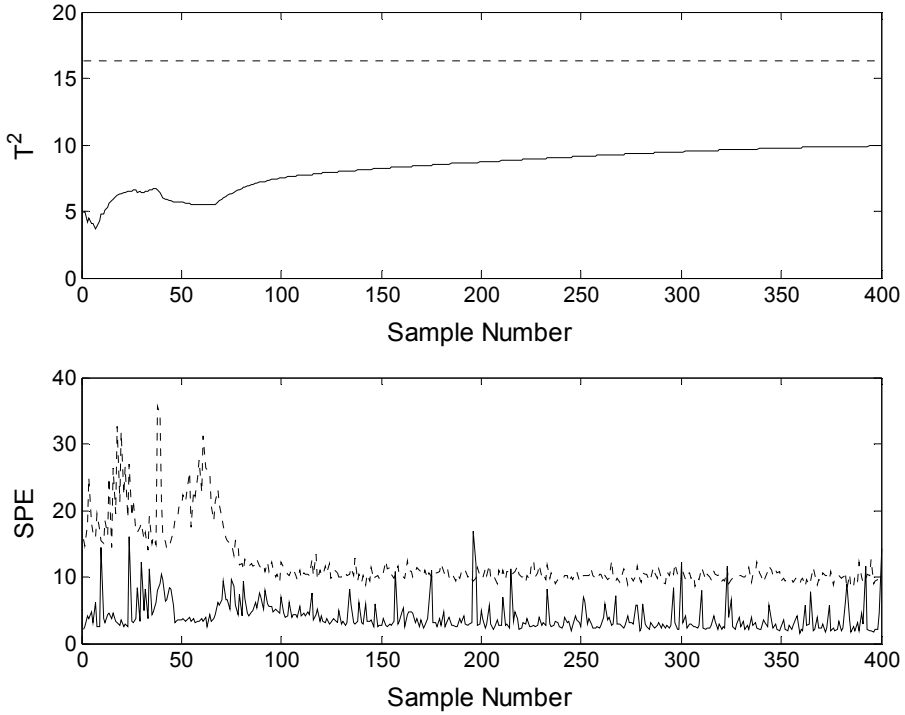


그림 5. T^2 and *SPE* charts for on-line monitoring using MPCA

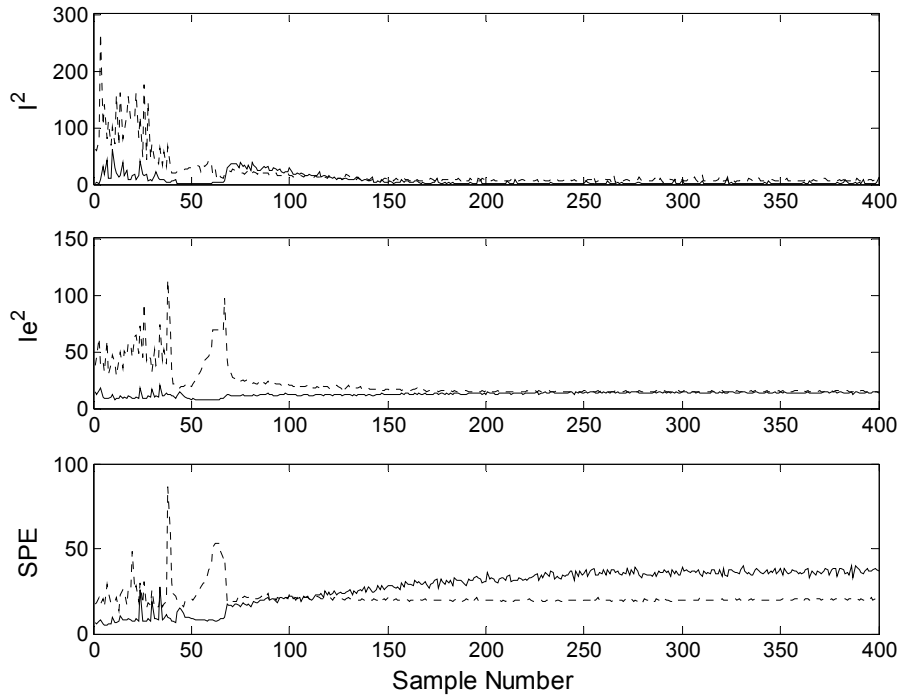


그림 **6.** I^2, I_e^2 and *SPE* charts for on-line monitoring using the proposed method

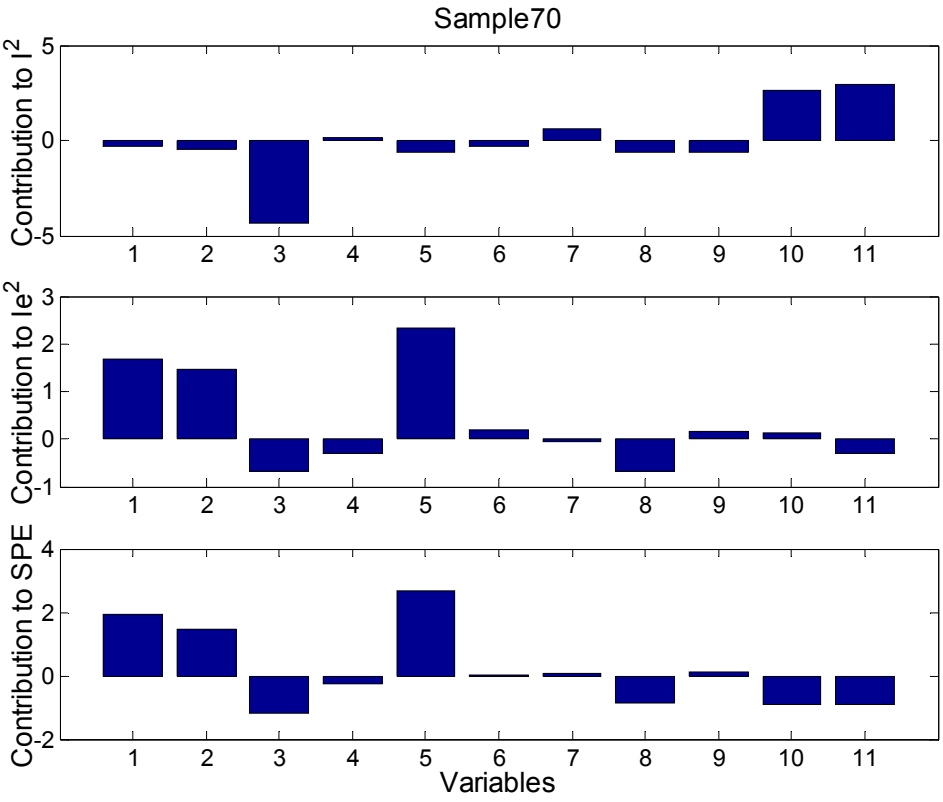


그림 **7.** Contribution plots for I^2, I_e^2 and *SPE* charts at time 70