UNIVERSITEIT
GENT

BIOMATH
Department of Applied Mathematics,
Biometrics and Process Control

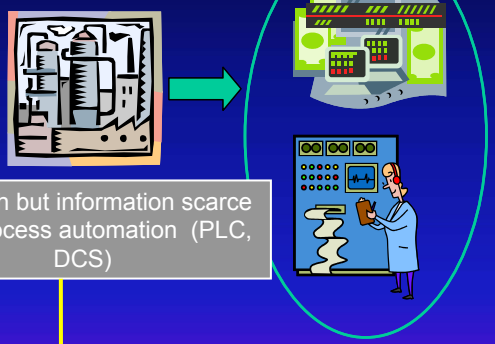
On-line Batch Monitoring Based on Multiway Independent Component Analysis

ChangKyoo Yoo
6-jan-2003

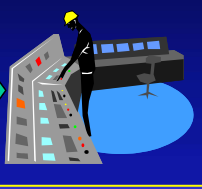
RUG-Biomath, Coupure 653, 9000 Gent, Belgium (e-mail ChangKyoo.Yoo@biomath.rug.ac.be)

Contents

Concept of Process Monitoring



An engineer has a chance to stop the problem before it grows into a full malady



Data rich but information scarce from process automation (PLC, DCS)

Obtaining stable condition
Improving process efficiency

Effective maintenance starts with an early detection of developing problems

How do we extract an useful information ? (Data mining)

Process monitoring using multivariate statistical analysis

BIOMATH

Integration of monitoring, prediction and control

Monitoring establishes confidence in prediction

MONITORING

Monitoring signals abnormal situations to controller

Statistical Correlation Model

measurements

CONTROL

Adjustments to batch

Predictions give controller ability to work beyond just abnormal situations

PREDICTION

BIOMATH

Characteristics of Batch Process

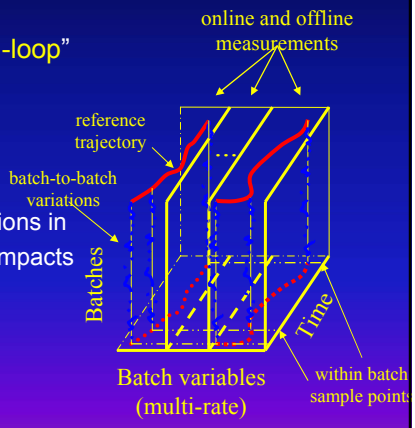
✓ Finite duration process: time variant and nonlinear behavior

✓ Batch operation is typically done in "open-loop" with respect to the product quality

- Quality variables sampled well *after* batch completed
- Monitoring provides a way to detect deviations in product quality to allow correction *before* it impacts the suitability of the product

✓ Batch-to-batch variation

✓ 3-way data (batch × variable × time)

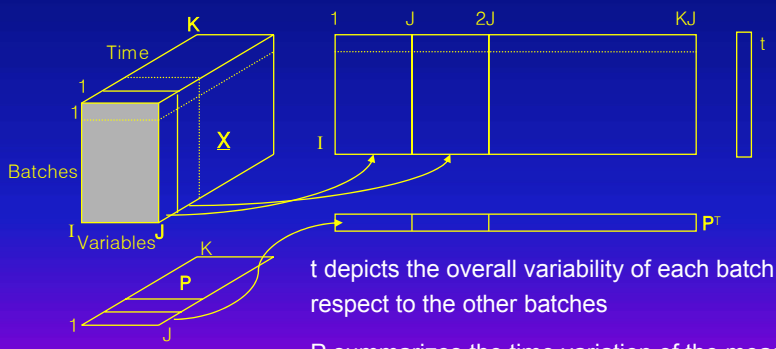


BIOMATH

Traditional Monitoring Method for Batch Process

✓ Multiway PCA (Nomikos and MacGregor, 1994)

- Extended version of PCA to handle multiway batch data
- 3-way data **X**(batch×variable×time) is unfolded into 2-dimensional matrix **X**(batch×(variable×time))



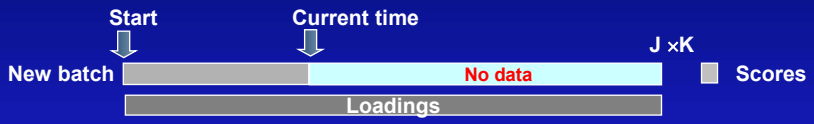
t depicts the overall variability of each batch with respect to the other batches

P summarizes the time variation of the measurement variables around their average trajectories

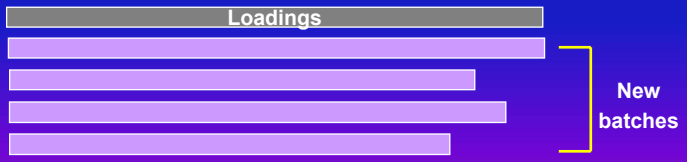
BIOMATH

Problems encountered for on-line monitoring of MPCA

1. $\mathbf{X}_{new}$ is not complete until the end of the batch operation.
   - The rest of the $\mathbf{X}_{new}$ matrix from the current time to the end of the batch is still undefined. (Filling method)



2. Batch length is not identical each batch.



BIOMATH

Motivation of this research

- PCA
  - Using only the information contained in the covariance matrix of the data
  - Not appropriate for non-stationary, dynamic, or non-Gaussian data.
    - ➡ Cause false alarms

➡ ICA

- Problems of conventional MPCA
  - Batch length should be identical.
  - Future observations should be anticipated for on-line monitoring
    - ➡ might cause bad performance

➡ Another unfolding

BIOMATH

What is Independent Component Analysis?

- ICA is a statistical method, the goal of which is to decompose given multivariate data into a linear sum of statistically independent components, that is, *the values of one variable do not convey any information about the other variable*.
- For example, given two-dimensional vector , $\boldsymbol{x} = [\, x_1\, x_2\,]^T$, ICA aims at finding the following decomposition

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} s_1 + \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix} s_2$$

$$\mathbf{x} = \mathbf{a}_1 s_1 + \mathbf{a}_2 s_2$$

where $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are basis vectors and s_1, s_2 are basis coefficients (sources)

Constraint: Basis coefficients s_1 and s_2 are statistically independent.

- We should find a_1, a_2, s_1 and s_2 from only x_1 and x_2

BIOMATH

How can we find source signals from only X?

$$\mathbf{X} = \mathbf{A}\,\mathbf{S}$$

X: measured variable, A: mixing matrix, S: source signal

s_i : statistically independent and var(s_i)=1

- If $\mathbf{W}=\mathbf{A}^{-1}$, we can exactly recover source signal from $\mathbf{S} = \mathbf{W}\,\mathbf{X}$.
- How can we find **W**(separating matrix) from only **X?**

➡ **W** is initialized and updated to maximize the non-Gaussianity of **S**

- More non-Gaussian, more independent !

❑ **Measure of non-Gaussianity**

- Kurtosis**:** $kurt(y) = E(y^4) - 3\,(E(y^2))^2$
  - kurtosis is zero for a Gaussian random variable
  - very sensitive to outliers
- Negentropy: information-theoretic quantity of entropy

  Hyvärinen(1998) developed a robust approximation equation of negentropy

BIOMATH

Detail procedures of ICA

1. **Centering** (mixed and independent source is zero-mean)
2. **Whitening** (remove all cross-correlations between X)
   Transform the observed vector **x** linearly so that $E\{\mathbf{z}(k)\mathbf{z}^T(k)\} = \mathbf{I}$

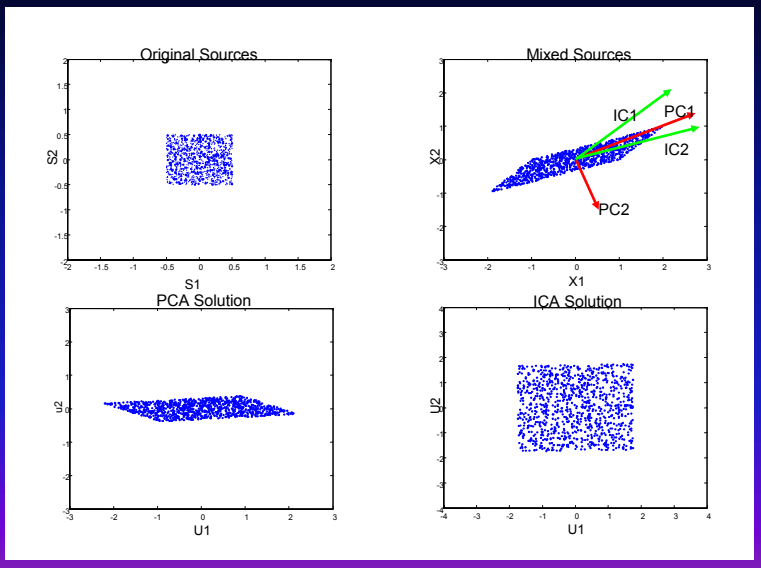
$$\mathbf{z}(k) = \mathbf{Q}\mathbf{x}(k) = \mathbf{Q}\mathbf{A}\mathbf{s}(k) = \mathbf{B}\mathbf{s}(k)$$
$$E\{\mathbf{z}(k)\mathbf{z}^T(k)\} = \mathbf{B}E\{\mathbf{s}(k)\mathbf{s}^T(k)\}\mathbf{B}^T = \mathbf{B}\mathbf{B}^T = \mathbf{I}$$
$$\mathbf{s}(k) = \mathbf{B}^T\mathbf{z}(k)$$

3. $\mathbf{B}^T$ is initialized and updated to maximize the negentropy of $\mathbf{s}(k)$
4. Since $\mathbf{s}(k) = \mathbf{W}\mathbf{x}(k)$ and $\mathbf{s}(k) = \mathbf{B}^T\mathbf{z}(k) = \mathbf{B}^T\mathbf{Q}\mathbf{x}(k)$, **W** can be obtained by $\mathbf{W} = \mathbf{B}^T\mathbf{Q}$.

Comparisons between PCA and ICA

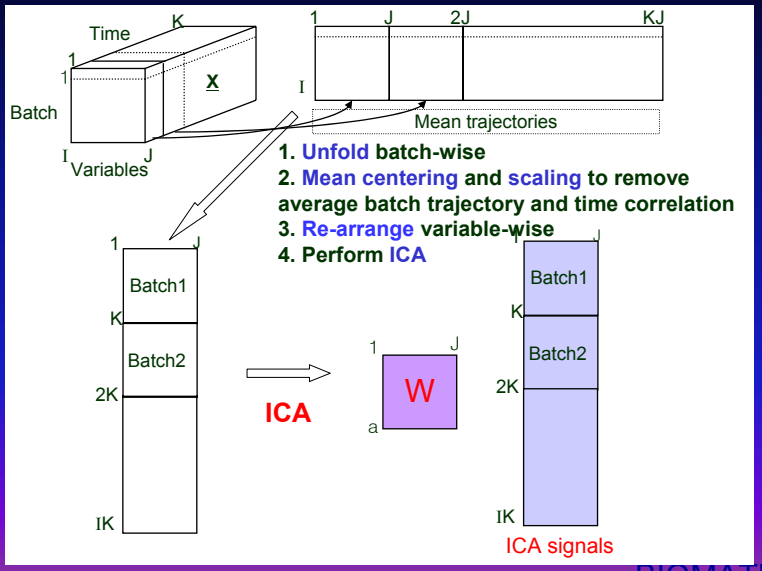
PCA	ICA
1. Second-order method (mean, variance)	1. Higher-order method (mean,variance, skewness, kurtosis, etc)
2. Using only the information contained in the covariance matrix of the data vector x	2. Use information on the distribution of x that is not contained in the covariance matrix
3. Assume Gaussian distribution of x	3. Assume non-Gaussian distribution of x
4. Computationally simple	4. More sophisticated techniques

Comparisons of ICA solution with PCA solution



BIOMATH

Proposed method: On-line batch monitoring based on MICA and unfolding



BIOMATH

New monitoring procedures (Part I)

1. Unfold $\underline{\mathbf{X}}(I \times J \times K)$ to $\mathbf{X}(I \times JK)$
2. The data $\mathbf{X}(I \times JK)$ are normalized using the mean and standard deviation of each variable at each time in the batch cycle over all batches.
3. Rearrange the scaled $\mathbf{X}(I \times JK)$ to $\mathbf{X}_{normal}(J \times IK)$
4. Whitening procedure: $\mathbf{Z}_{normal} = \mathbf{Q}\mathbf{X}_{normal}$
5. ICA procedure

   Obtain **W**, **B**, and $\mathbf{S}_{normal}$ from $\mathbf{S}_{normal} = \mathbf{W}\mathbf{X}_{normal} = \mathbf{B}^T\mathbf{Z}_{normal}$
6. Calculate the norm of the row vectors of **W** and separate **W** into the deterministic part and the excluded part based on the magnitude of norms. **B** and $\mathbf{S}_{normal}$ can be separated with the similar criterion.

$$\mathbf{W} \rightarrow \mathbf{W}_d, \mathbf{W}_e \qquad \mathbf{B} \rightarrow \mathbf{B}_d, \mathbf{B}_e \qquad \mathbf{S}_{normal} \rightarrow \begin{cases} \mathbf{S}_d = \mathbf{W}_d\mathbf{X}_{normal} \\ \mathbf{S}_e = \mathbf{W}_e\mathbf{X}_{normal} \end{cases}$$

BIOMATH

New monitoring procedures (Part II)

7. Calculate I^2, I_e^2 and *SPE* statistics

$$I^2(n) = \mathbf{s}_d(n)^T\mathbf{s}_d(n) \qquad I_e^2(n) = \mathbf{s}_e(n)^T\mathbf{s}_e(n) \qquad SPE(n) = \sum_{j=1}^{d}(x_j(n) - \hat{x}_j(n))^2$$

   where n is a value from 1 to IK and $\hat{\mathbf{X}} = \mathbf{Q}^{-1}\mathbf{B}_d\mathbf{S}_d = \mathbf{Q}^{-1}\mathbf{B}_d\mathbf{W}_d\mathbf{X}_{normal}$
8. Rearrange $I^2(1\times IK)$, $I_e^2(1\times IK)$ and $SPE(1\times IK)$ to $I^2(I\times K)$, $I_e^2(I\times K)$ and $SPE(I\times K)$, respectively.
8. Obtain control limits of I^2, I_e^2 and *SPE* metrics at each time using kernel density estimation

I^2, I_e^2 and *SPE* are not normally distributed

Kernel density estimation

Density estimation is the construction of an estimate of the density function from the observed data

BIOMATH

New monitoring procedures (Contribution Plot Part)

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

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

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

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

3. Variable contribution to *SPE* (k)

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

BIOMATH

On-line batch monitoring with MICA

Unfolding and mean trajectory removal under normal operation (NOC)

Modeling

Normalization and whitening

Obtain MICA model (W,S) and *SPE* from NOC

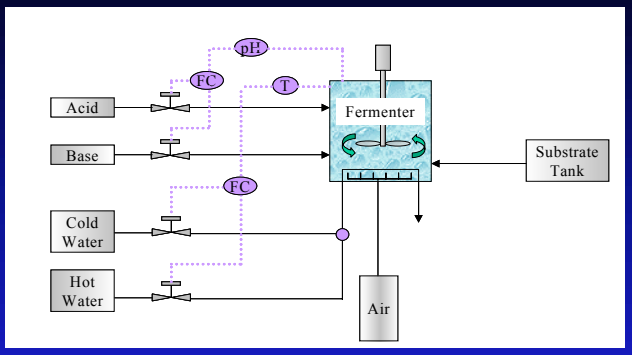
Determine the control chart limits of I^2 and I_e^2 and *SPE* statistics using kernel density estimation (KDE)

On-line monitoring

Project the new data into MICA model to calculate I^2 and I_e^2 and *SPE* values
Monitor if I_{new}^2, I_{newe}^2, and *SPE* values exceed the control limits

BIOMATH

Case Study (Fed-Batch Penicillin Production)



• Monitoring and control group (Ündey, C., Birol,B, and Çinar,A.) has developed a simulator (PenSim v2.0) that is capable of simulating concentrations of biomass, CO_2, hydrogen ion, penicillin, carbon source, oxygen and heat generation under various operating conditions.

BIOMATH

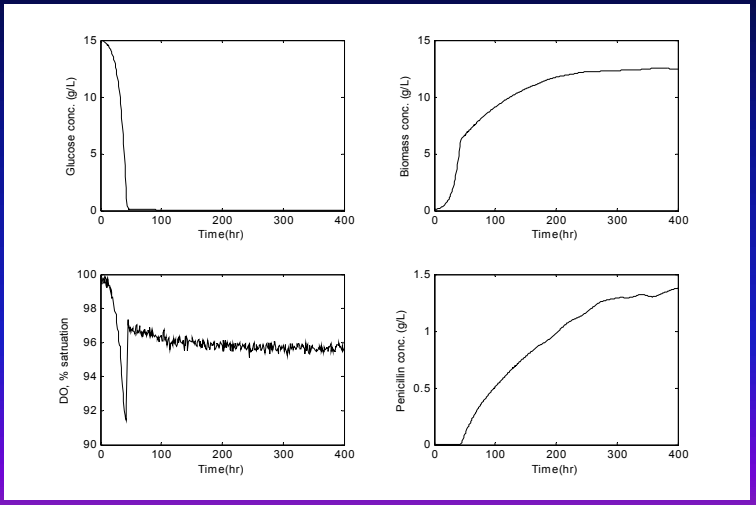
Variables of Penicillin Process

• Total 11 Variables with underline are used for batch monitoring

- ✓ Load variables: aeration rate, agitator power, substrate feed rate and substrate feed temperature
- ✓ Manipulated variables: acid/base and heating/cooling water flow rates
- ✓ Internal state variables: culture volume, generated heat, carbon dioxide, dissolved oxygen, biomass, penicillin and substrate feed concentrations
- ✓ Controlled variables: bioreactor pH and temperature

BIOMATH

Normal batch trajectory



BIOMATH

Case I: Equal batch length

- A total of 67 batches based on the normal operation are generated for the development of MPCA and MICA model providing the necessary information for on-line monitoring
  - ✓ The duration of each batch is 400hr. (45hr: pre-culture state, 355hr: fed-batch stage)
  - ✓ The sampling interval: 1 hr
  - ✓ Small variations were added to simulation input data for process common variations
  - ✓ Measurement noises were also added to 11 variables used in monitoring.
- Disturbance: substrate feed rate is linearly decreased with slope –0.002 from time 100hr to time 250hr (slow process drift).

BIOMATH

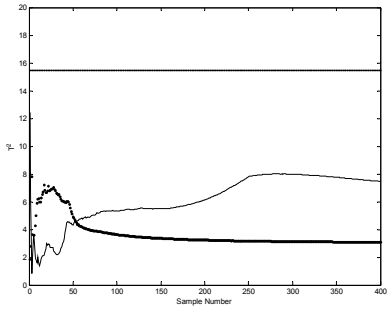
MPCA Monitoring Results (Equal batch length)

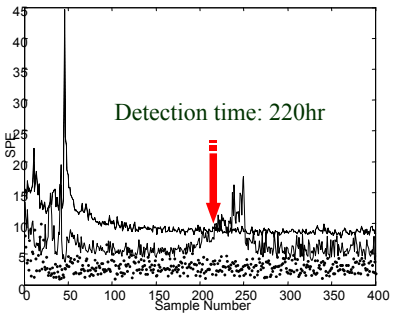
✓4 PCs (explain 62.2% of the variation) are used.

✓ To fill in the future values, the ability of PCA that handles missing data is used (Filling method 3 of Nomikos and MacGregor, 1994).

✓ The dotted points represent the normal batch.

✓ Detection time: 220hr (delayed about 120hr after occurring a fault)





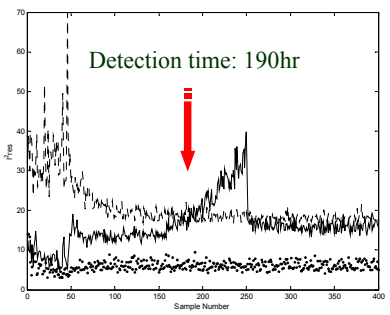
BIOMATH

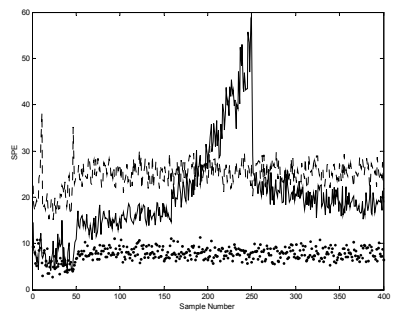
MICA Monitoring Results (Equal batch length)

✓ 3 ICs are used for deterministic part.

✓ Detection time: 190hr (earlier than MPCA monitoring charts by 30hr)

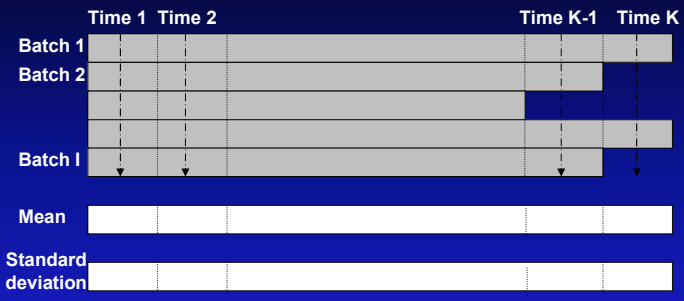
✓ MICA without estimating the future values still yields better results.





BIOMATH

Case II: Unequal batch length



✓ total 67 batches are generated.

✓ 20 batches(time 400), 10 batches(time399), 10 batches(time398), 5 batches(time397), 5 batches(time396), 5 batches(time395), 5 batches(time394), 5 batches(time393), 2 batches(time392),

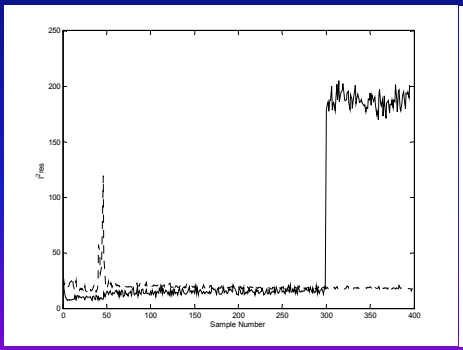
✓ Test batch (time 395) having a step disturbance is also generated.

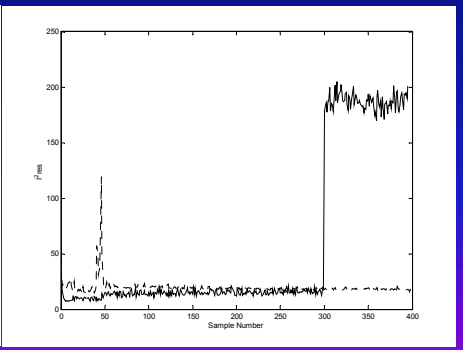
BIOMATH

MICA results (Unequal batch length)

❑ Disturbance: step decreasing the agitation power by 10% at 300hr, where it is lasted until the end of batch.

❑ The disturbance is detected well by *Ie*2 and *SPE* charts from 300hr without delay.

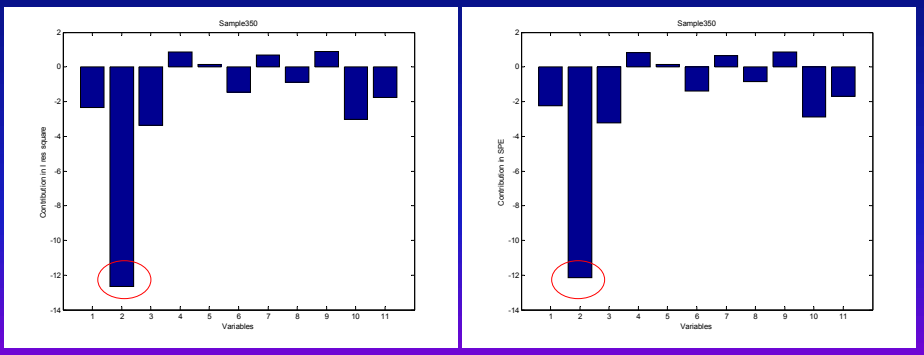




BIOMATH

Contribution plot of MICA (Unequal batch length)

- From the contribution plot for I_e^2 and *SPE* at sample 350, we can conclude variables 2 (agitation power) causes the large deviation primarily.



BIOMATH

Conclusion

- In comparison to PCA, ICA can reveal underlying factors from multivariate statistical data including non-Gaussian.
- MICA can detect a disturbance earlier than MPCA.
  - ✓ MICA need not anticipate the future values.
  - ✓ MICA is useful when the batch length is different from each batch.
- MICA can be easily applied in most batch and semi-batch processes (i.e. SBR)
- Integration of monitoring, prediction, control and optimization)

BIOMATH