예제를 통한 PCA/PLS 의 동특성 묘사

다음의 두 가지 예제를 통해 동특성 묘사를 위한 일반적인 다변량 통계 기법 및 개선된 다변량 통계 기법의 적용과 그와 관련된 현안들을 알아보도록 한다.

예제 1. – Ku 등 [1]이 사용했던 예제

Ku 등은 다음과 같은 예제를 통해 dynamic PCA (DPCA)가 일반 PCA 가 동특성 묘사에서 우수하다고 보고하였다. Multivariate AR(1) process;

$$\mathbf{z}(k) = \begin{bmatrix} 0.118 & -0.191 \\ 0.847 & 0.264 \end{bmatrix} \mathbf{z}(k-1) + \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix} \mathbf{u}(k-1) \qquad (1)$$
$$\mathbf{y}(k) = \mathbf{z}(k) + \mathbf{v}(k)$$

이때 **u** 는 correlated 입력이며,

$$\mathbf{u}(k) = \begin{bmatrix} 0.811 & -0.226 \\ 0.477 & 0.415 \end{bmatrix} \mathbf{u}(k-1) + \begin{bmatrix} 0.193 & 0.689 \\ -0.320 & -0.749 \end{bmatrix} \mathbf{w}(k-1) \qquad (2)$$

w 와 **v** 는 평균값 0 과 분산이 각각 1 과 0.1 인 무작위 잡음이다. 이 예제에 PCA 와 DPCA 를 다시 적용한 결과를 표 1&2 에 나타내었다. PCA 와 DPCA 의 입력 행렬 **X** 는 각각 [$\mathbf{y}^T(k)$ $\mathbf{u}^T(k)$] 와 [$\mathbf{y}^T(k)$ $\mathbf{u}^T(k)$ $\mathbf{y}^T(k\text{-}1)$ $\mathbf{u}^T(k\text{-}1)$]으로 구성되었다. Ku 등의 결과에 따르면, 각 방법에 대한 PC 의 개수는 각각 PCA 가 2 개, DPCA 가 5 개였다. 표 1 에 보인 바와 같이, cumulative Q^2 는 PC 가 2 개일 때엔 0.347, 5 개 일때엔 0.895 이다. 게다가 모든 significant tests 를 통과하지 못하였다. PCA and DPCA 의 cumulative Q^2 값들을 비교해 보면, 5 개의 PC 를 갖는 DPCA 가 (Ku 등이 보고하였던) 2 개의 PC 를 갖는 PCA 보다 성능이 좋을 것이라는 것을 쉽게 알 수 있다. 하지만 3 개의 PC 를 갖는 PCA 와 DPCA 를 비교해 보면, 둘다 거의 동등한 성능을 보일 것이라는 것을 알 수 있다. 이 예제는 또한 다음의 중요한 사실을 잘 나타내고 있다: 동특성을 포함하는 데이터를 다룰 때는, PC 의 개수를 결정하는 새로운 방법이 필요하다. 이는 몇몇 논문들에서도 이미 언급한 사실이다. 한편, Russell 등[4]은 최근 Tennessee Eastman problem 을 이용한 정밀한 비교를 통해 DPCA 와 PCA 가 거의 동일한 monitoring and fault detection 성능을 보임을 증명하였다.

이론적으로 볼때, DPCA 는 동특성을 묘사하기 위한 PCA 의 확장이지만 위의 결과를 두고 볼때 성능의 비교는 아직은 분명치 않다. 그리고 PCA 로 동특성을 묘사할 때, stationarity 를 기반으로 한 여러 기준들(예를 들면 PC 개수를 결정하는 significance test) 또한 이에 맞게 적용되어야 한다.

Table 1. Ku 등에 의한 PCA 결과

# of PCA comp's	R2X(cum)	eigenvalues	Q2(cum)	significance
1	0.480	1.921	0.127	No
2	0.873	1.570	0.347	No
3	0.989	0.463	0.895	No

Table 2. Ku 등에 의한 PCA 결과

# of PCA comp's	R2X(cum)	eigenvalues	Q2(cum)	significance
1	0.445	3.56	0.195	Yes
2	0.829	3.071	0.589	Yes
3	0.929	0.799	0.731	Yes
4	0.990	0.489	0.950	No
5	1.000	0.079	0.998	No

예제 2. – 증류탑의 추론 모델.

다변량 통계 기법의 가장 성공적인 적용 분야가 증류탑의 추론모델 설계이다[2,3]. 이의 간단한 설명을 위해 일반적인 이성분 증류탑을 예로 든다. 입력 조성(feed composition, z_F), 탑상 조성(distillate composition, y_D) 그리고 탑저 조성(bottom product composition, x_B)이 일단 결정되면, 그에 따른 유일한 단 온도 profile, θ 또한 결정된다. 추론모델의 목적은 단 온도, θ 와 다른 측정 가능한 변수, φ 를 이용하여 최상의 품질 조성 추정값, $\hat{\mathbf{y}}$ 를 얻는 것이다. 추정기의 일반적인 형태는 다음과 같다:

$$\hat{\mathbf{y}} = \mathbf{K}([\theta, \varphi])^T \qquad (3)$$

여기서 $\hat{\mathbf{y}} = (\hat{y}_D \ \hat{x}_B)^T$, 그리고 $\mathbf{K}(\cdot)$는 PLS/PCR 모델의 계수이다. 앞서 언급한 목적은 결국 최적의 행렬 $\mathbf{K}$ 를 찾는 것이다.

여러가지 추정 방법(PCR estimator, Kalman filter, Browsilow estimator)에 대한 평가가 이미 보고되었다 [2,3]. 이 평가에서, PLS 에 기반한 *정적* 추론 모델이 *동적* Kalman filter 와 가의 동일한 성능을 보임이 밝혀졌다. 이는 증류탑의 특성에 기인한 것으로 각

단의 온도는 그 전단에 대한 정보를 포함하고 있기 때문이다. 여기서는 PLS 모델이 FIR model 과 동일하다는 것을 보임으로서 이를 증명하고자 한다.

식 (3)의 PLS 모델을 자세히 나타내면,

$$[\hat{y}_D(k)\ \hat{x}_B(k)]^T = \mathbf{K}[T_1(k)\ \ T_2(k)\cdots T_n(k)\ \ T_{n+1}(k)]^T . \quad (4)$$

그리고 단 온도는 내부 흐름 (L and V) 의 함수로 표현할 수 있다 (단 열역학적 평형임을 가정):

$$T_i(k) = f(L_{i-1}(k), V_{i+1}(k)). \quad (5)$$

흐름의 동특성이 충분히 빠르고 1 차의 전달함수로 표현될 수 있으면, 내부 흐름은 다음과 같이 외부 흐름으로 표현이 가능하다.

$$L_i(k) = G_i(q^{-1})L_{i-1}(k), \quad V_i(k) = H_i(q^{-1})V_{i+1}(k) \quad , \quad (6)$$

$$\begin{aligned}\frac{L_i(k)}{L_{i-1}(k)} &= \frac{a}{1-bq^{-1}} = a(1+bq^{-1}+b^2q^{-2}+\cdots) \\ &\cong c_0 + c_1q^{-1}\end{aligned} , \quad (7)$$

$$\begin{aligned}L_{i-1}(k) &= \prod_{j=1}^{i-1} G_j(q^{-1})L_0(k) \\ V_i(k) &= \prod_{j=1+1}^{n} H_j(q^{-1})V_0(k)\end{aligned} , \quad (8)$$

그러면, 식 (5)는다음과 같이 다시 표현되고

$$\begin{aligned}T_i(k) &= f\left(\prod_{j=1}^{i-1} G_j(q^{-1})L_0(k), \prod_{j=i+1`}^{n} H_j(q^{-1})V_0(k)\right) \\ &= f([L_0(k)\ L_0(k-1)\cdots L_0(k-i+1)\ V_0(k)\ V_0(k-1)\cdots V_0(k-n-i)]^T)\end{aligned} \quad (9)$$

시간 k 에서 온도 vector θ 는

$$\begin{bmatrix} T_1(k) \\ \vdots \\ T_i(k) \\ \vdots \\ T_{n+1}(k) \end{bmatrix} = \mathbf{f}\begin{pmatrix} L_0(k) & V_0(k)\cdots V_0(k-i+1)\cdots V_0(k-n+1) \\ & \vdots \\ L_0(k)\ L_0(k-1)\cdots L_0(k-i+1) & V_0(k)\ V_0(k-1)\cdots V_0(k-n-i) \\ & \vdots \\ L_0(k)\cdots L_0(k-i+1)\cdots L_0(k-n+1) & V_0(k) \end{pmatrix} \quad (10)$$

이다. 즉 각 단의 온도는 과거의 외부 흐름의 함수이다. 만약 **f** 가 선형 함수이면, 식 (3)의 PLS 모델은 단수와 같은 크기의 ‘window’를 갖는 FIR 모델과 동일하다:

$$\hat{\mathbf{y}}(k) = \mathbf{K}\theta(k) \cong \sum_{i=0}^{n-1}\begin{pmatrix} h_{11}(i) & h_{12}(i) \\ h_{21}(i) & h_{22}(i)\end{pmatrix}\begin{pmatrix} L_0(k-i) \\ V_0(k-i)\end{pmatrix} \quad (11)$$

위의 가정들은 그리 무리한 가정이 아니며, 쉬운 유도와 간단한 최종식을 위해서 도입되었다. 이 같은 이유로 Mejdell 등이 보고한 바와 같이 '정적' PLS 모델이 증류탑의 동특성을 잘 묘사할 수 있다.

참고문헌

1. Ku, W., Storer, R.H. and Georgakis, C., Chemometrics and Intelligent Laboratory Systems, 30 179(1995).
2. Mejdell, T and Skogestad, S., Ind. Eng. Chem. Res., 33, 2555(1991).
3. Mejdell, T and Skogestad, S., AICHE J., 39(10), 1641(1993).
4. Russel, E.L., Chiang, L.H. and Braatz R.D., Chemometrics and Intelligent Laboratory Systems, 51 81(2000).