Journal of the Korean Institute of Industrial Engineers
Vol. 37, No. 4, pp. 400-407, December 2011.

고차원 스펙트라 데이터 분석을 위한 Adjusted Direct Orthogonal Signal Correction 기법

김신영 · 김성범†

고려대학교 산업경영공학과

Adjusted Direct Orthogonal Signal Correction For High-Dimensional Spectral Data

Sin Young Kim · Seoung Bum Kim

School of Industrial Management Engineering, Korea University

Modeling and analysis of high-dimensional spectral data provide an opportunity to uncover inherent patterns in various information-rich data. Orthogonal signal correction (OSC), a preprocessing technique has been widely used to remove unwanted variations of spectral data that do not contribute to prediction or classification. In the present study we propose a novel OSC algorithm, called adjusted direct OSC to improve visualization and the ability of classification. Experimental results with real mass spectral data from condom lubricants demonstrate the effectiveness of the proposed approach.

Keywords: Mass Spectra, Principal Component Analysis, Orthogonal Signal Correction, Visualization, Classification, Preprocessing

1. 서 론

오늘날 측정기기와 정보통신 기술의 눈부신 발달로 인해 방대한 양의 데이터가 쏟아져 나오고 있으며 이 데이터를 유용한 정보로 변환하는 작업은 매우 중요한 일로 인식되고 있다. 데이터는 숫자, 텍스트, 시그널, 이미지 등 다양한 형태로 생성되고 있으며 이를 분석하기 위한 기법들이 다양하게 존재한다(Witten *et al*., 2011; Shmueli *et al*., 2007). 특히 샘플의 특성이 시그널 형태로 표현된 데이터가 널리 생성되고 있으며 이와 같은 데이터의 패턴을 분석하는 기술이 개발되고 다양한 분야에 적용되고 있다. 의학 분야 에서는 사람의 혈액, 소변, 땀 등으로부터 스펙트라를 얻고 이들의 패턴을 분석하여 특정 질병 정도를 알아내고 사람의 생리적인 리듬 변화를 특성화시키기도 한다(Park *et al*., 2009). 또한 신약 개발 시 신약의 효과를 시그널을 통해 알 수 있으며 생약의 화학적 성분을 알아내는 데에도 사용되고 있다. 사람뿐 아니라 쥐나 모기 등 동물의 샘플을 통해 다양한 임상실험이 스펙트라 분석을 통해 행해졌다(Zijlstra *et al*., 2000). 범죄 분석학에도 질량 스펙트라의 분석이 사용되고 있는데, 그 예로 수사현장에서 채취한 약물, 합성 섬유, 석유화학 제품, 페인트, 볼펜 잉크 등의 식별 및 분류 등을 들 수 있다(Kher *et al*., 2006; Saferstein, 2005).

고차원 스펙트라 분석에 널리 사용되고 알고리즘으로는 주성분분석(Principal Component Analysis : PCA)과 Partial Least Squares(PLS) 등 특질추출법이 있다(Jang and Chang, 2006; Trygg and Wold, 2002). 그러나 일반적으로 스펙트라 데이터는 추출과정에서 실험실의 빛, 온도의 변화, 시간과 같은 노이즈의 영향을 많이 받기 때문에 PCA 및 PLS와 같은 특질추출법만으로는 신뢰성 있는 분석이 어렵다(Luypaert *et al*., 2006). 이러한 점을 극복하기 위하여 데이터 전처리 알고리즘인 Orthogonal Signal Correction(OSC)가 개발되었다(Wold *et al*., 1998). OSC는 데

† 연락저자 : 김성범 교수, 136-701 서울특별시 성북구 안암동 5가 1번지 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-929-5888, E-mail : sbkim1@korea.ac.kr
2011년 8월 31일 접수; 2011년 11월 13일 게재 확정.

이터에서 노이즈를 제거하여 분석 성능을 향상시키며 결과 해석을 용이하게 해 주는 데이터 전처리 기법 중 하나이다. 첫 OSC 알고리즘(Wold *et al*., 1998)이 발표된 이후 여러 방식의 OSC알고리즘(Fearn, 2000; Westerhuis *et al*., 2001; Trygg and Wold, 2002)이 발표되었으나, 대부분 Y가 연속형 이어야 한다는 한계를 보였다. 이러한 한계점을 극복하기 위해서 Direct Orthogonal Signal Correction(DOSC) 기법이 제안되었다(Westerhuis *et al*., 2001). DOSC는 클래스분류에 불필요한 노이즈를 제거하여 데이터의 클래스 간 격차를 벌려 분류결과를 향상시키며 더 나은 시각화 효과를 얻게 해 준다(Kim *et al*., 2008).

본 연구의 주목적은 스펙트라 전처리 알고리즘으로 유용하게 쓰이고 있는 기존의 DOSC 기법을 개선하여 고차원의 스펙트라 데이터의 시각화와 분류성능을 보다 향상시키는데 있다. 기존의 DOSC 방법은 일반적으로 노이즈 제거시 분류 정보를 담고 있는 변수의 유용한 정보 또한 제거하는 경향을 보이는데 반해 본 논문에서 제안하는 방법은 유용한 정보를 보존하여 스펙트라 데이터의 시각화와 예측 성능을 보다 향상 시키는 효과를 갖고 있다. 또한 본 논문에서는 제안하는 기법의 실제 적용 가능성을 살펴보기 위해 질량분석기로부터 얻은 실제 콘돔 윤활액의 스펙트럼 분석에 적용하여 그 효과를 입증하였다.

2. Orthogonal Signal Correction

OSC는 시그널 데이터 분석 시 노이즈를 제거하기 위하여 행하는 데이터 전처리 기법으로 현재 많은 분야에서 널리 사용되고 있다(Luypaert *et al*., 2006; Saiz-Abajo *et al*., 2005; Svensson *et al*., 2002). OSC의 기본 개념은 다음과 같다. 입력변수 X와 출력변수 Y가 있다고 할 때, Y와 연관 없는 부분(즉 Y와 직교(orthogonal)인 부분)을 X에서 제거하는 것이다. X 데이터의 일부가 Y와 직교하면, Y와 연관이 없다고 할 수 있으며, Y와 연관이 없으면 분석에 불필요한 부분이라고 가정할 수 있다.

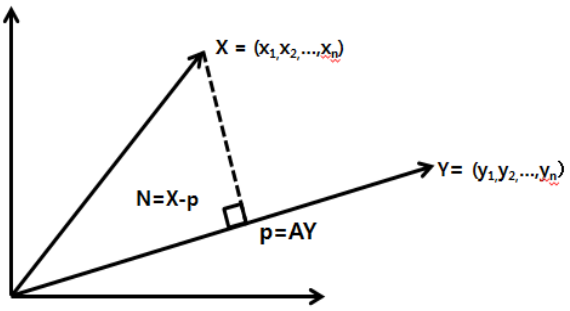

Figure 1. An Overview of an OSC Algorithm

위의 설명을 <Figure 1>에서 보여주고 있다. X는 n개의 관측치를 가진 데이터 평면을 나타내고, Y는 n개의 관측치를 가진 클래스 평면을 나타낸다. 여기서 X를 Y에 직각으로 사영하여 이 부분(p) 을 X에서 빼주면 Y와 관계없는 X부분 즉 노이즈(N)를 얻을 수 있게 된다.

최초 OSC 알고리즘(Wold *et al*., 1998)이 발표된 이후 여러 방식의 OSC알고리즘이 발표되었으나, Y와 연관 없는 부분을 제거하는 기본 개념은 모두 동일하다(Fearn, 2000; Westerhuis *et al*., 2001; Trygg and Wold, 2002). OSC에 관한 각 연구의 핵심은 X에서 Y와 연관 없는 부분, 즉 노이즈의 계산방식이다. 처음 발표된 OSC는 X에서 PCA를 이용하여 첫 번째 주성분을 추출한 후 이것을 Y에 사영하여 노이즈를 제거하였다(Wold *et al*., 1998). 해당 알고리즘은 노이즈 제거에 전반적으로 좋은 성능을 보였으나, 데이터에서 직접적으로 Y를 구하지 않고 주성분을 이용하여 간접적으로 구한다는 점과 Y가 연속형 변수이어야 한다는 한계를 보인다. 이후 노이즈를 구하는 방법에 있어 사영이 아닌 고유값 분해(Eigenvalue decomposition)를 이용한 OSC가 발표 되었으며(Fearn, 2000), 또한 PLS를 이용하여 p를 구한 후 사영을 통하여 간접적으로 노이즈를 구하는 OSC가 발표되었다(Trygg and Wold, 2002). 발표한 OSC는 대부분 노이즈 제거에 좋은 성능을 보였으나, Y가 연속형 변수이어야 한다는 한계를 가지고 있었다.

여러 OSC 알고리즘 중 Y의 형태에 상관없이 사용 가능한 알고리즘인 Direct Orthogonal Signal Correction (DOSC) 기법이 소개되었다(Westerhuis *et al*., 2001). 따라서 DOSC는 Y가 범주형(categorical)일 경우 널리 사용되고 있다(Kim *et al*., 2008). 본 연구에서 제안하고자 하는 기법이 DOSC의 확장임으로 DOSC기법에 관해 좀 더 자세히 설명을 하도록 하겠다. DOSC는 다른 OSC와 달리 X를 Y에 사영하여 노이즈를 제거한다(Westerhuis *et al*., 2001). 위에서 언급한 바와 같이 입력데이터인 X는 Y와 관련 있는 부분과 관련 없는 부분의 합으로 다음과 같은 식으로 표현할 수 있다.

$$X=\vec{p}+(X-\vec{p})=Y\frac{Y^TX}{Y^TY}+(X-Y\frac{Y^TX}{Y^TY}) \qquad (1)$$

여기서 Y와 관련 있는 부분은 p 이며, p는 Y에 임의의 계수를 곱한 AY라고 말할 수 있다. 그러나 계수인 A가 알려져 있지 않기 때문에 식 (1)과 같은 식으로 표현되며 Y와 관련 없는 부분은 X에서 p를 제거한 $(X-Y\frac{Y^TX}{Y^TY})=N$ 이다.

X에서 N를 모두 제거한다면 학습 데이터(training data) 상에서는 가장 적합할 수 있으나, 실험 데이터(testing data)를 적용했을 경우 과적합(overfitting) 문제가 생길 수 있다(Westerhuis *et al*., 2001). 이 과적합 문제를 해결하기 위해서 DOSC 알고리즘에선 N에 PCA를 적용하여 주요 성분만을 사용한다. 여기서 몇 개의 주성분을 선택해야 하는가에 대한 문제가 생기는데 이를 OSC 알고리즘에서는 OSC 성분 개수를 정하는 문제로 부르고 있다. 너무 많은 OSC 성분개수를 사용할 경우 과적합의 위험이 있으며 너무 적은 수를 사용할 경우 Y와 관련 없는 부분이 충분히 제거되지 못하는 단점이 있다(Wold *et al*., 1998; Westerhuis *et al*., 2001). OSC 성분 개수를 정하는 방법으로는

대부분 시각화를 통하여 주관적으로 결정하고 있다. 본 연구에서는 OSC 성분의 개수를 변화해 가면서 OSC의 효과를 PCA score plot을 통하여 확인하며 적정 OSC 성분의 개수를 선정하였다. 노이즈 부분에서 선택된 주성분을 N_{pca}라 표현할 때 N_{pca}의 식은 다음과 같이 표현할 수 있다.

$$N_{pca} = Nw \tag{2}$$

N은 노이즈 부분을 나타내며, w는 PCA로부터 얻어진 계수이다. 주성분분석 알고리즘에 의하면, N_{pca}는 N의 차원을 축소하기 위하여 데이터를 공분산 행렬의 고유벡터방향으로 사영한 것이다. 따라서 학습데이터에서 N_{pca}를 제거하기 위해서 N_{pca}를 다시 X에 사영해야 한다. 이 과정을 다음 식으로 나타낼 수 있다.

$$P = X^t N_{pca} (N_{pca}^t N_{pca})^{-1} \tag{3}$$

최종적으로 보정된 데이터인 X^{DOSC}는 다음의 식을 통하여 구할 수 있다.

$$X^{DOSC} = X - N_{pca} P^t \tag{4}$$

3. Adjusted Direct Orthogonal Signal Correction

DOSC의 노이즈 제거 과정의 핵심은 X를 Y에 사영하여 Y와 관련 있는 부분과 관련 없는 부분의 합으로 분해하는데 있다. 식 (1)에서 보여주듯이 Y와 관련 있는 부분인 p는 X를 Y에 사영하여 구한다. 사영의 정의에 따라 p는 Y에 임의의 정수 행렬을 곱한 결과와 같으며 이는 <Figure 1>에서 자세히 보여주고 있다. 따라서 이 방법은 X와 Y의 관계가 선형임을 가정한다고 말할 수 있으며 이를 식으로 표현하면 다음과 같다.

$$X = p + N = AY + N \tag{5}$$

A는 임의의 정수 행렬이며 N은 식 (1)에서 정의한 Y와 관련 없는 부분인 노이즈 부분이다. 그러나 실질적으로 X와 Y의 관계가 언제나 선형일 수는 없다. 특히 Y가 연속형이 아닌 범주형 변수일 경우 X와 Y의 관계는 선형이 아니라는 연구결과를 볼 수 있다(Kuter *et al.*, 2005). X와 Y의 관계가 선형이 아닌 임의의 데이터를 표현하면 식 (6)과 같다.

$$X = CY^k + N, \ (k = 2, 3, \cdots) \tag{6}$$

식 (6)에서 노이즈 부분만을 추출하기 위해서는 X에서 Y에 대한 정보인 CY^k를 제거해야 한다. 그러나 기존 DOSC 알고리즘에서는 X를 Y에 사영한 AY를 구하여 이를 X로부터 빼줌으로써 노이즈를 얻는다. X와 Y가 선형관계가 아니라고 가정할 때 식 (6)의 양변에 AY를 빼주면 다음과 같은 식을 얻을 수 있다.

$$X - AY = CY^k - AY + N \tag{7}$$

이 경우 AY를 X에서 빼줌으로써 노이즈 부분만을 추출하려는 원 DOSC 알고리즘의 의도와는 달리 노이즈 부분 외에도 ($CY^k - AY$)부분이 여전히 남아있음을 확인할 수 있다. 즉, Y에 관한 정보가 여전히 존재한다는 점이다. 기존 여러 DOSC에 관한 연구 중 Svensson *et al.*(2002)과 Zhu *et al.*(2007)의 연구에서 DOSC는 노이즈를 제거하는 과정에서 유용한 비선형 정보 또한 제거하는 경향이 있음을 밝히고 있다.

따라서 본 논문에서는 기존 DOSC에서 제거되는 Y에 대한 유용한 정보를 보존하여 기존 DOSC의 성능을 보다 향상시키고자 한다. 제안하는 방법의 핵심은 N에서 Y에 대한 유용한 정보를 추출하여 X에 다시 더해주는 것이다. 이를 식으로 설명하면 다음과 같다. N의 주성분 중 Y에 관한 유용한 정보가 담긴 주성분을 adj_N_{pca}라고 표현할 때 그 식은 다음과 같이 표현할 수 있다.

$$adj\_N_{pca} = Nw \tag{8}$$

데이터에서 adj_N_{pca}를 더하기 위해서 adj_N_{pca}를 다시 X에 사영해야 하며 그 식은 다음과 같다.

$$adj\_P = X^t \cdot adj\_N_{pca} (adj\_N_{pca}^t \cdot adj\_N_{pca})^{-1} \tag{9}$$

최종적으로 보정된 데이터인 X^{adj_DOSC}는 다음의 식을 통해 구할 수 있다.

$$X^{adj\_DOSC} = X - N_{pca} \cdot P^t + adj\_N_{pca} \cdot adj\_P \tag{10}$$

즉 본 논문에서 제안하고 있는 기법을 사용하여 보정된 최종 데이터인 X^{adj_DOSC}는 X에서 Y와 연관 없는 부분을 제거하고 연관 있는 부분을 다시 더해서 얻을 수 있다.

새로 제안하는 DOSC 또한 OSC 성분의 개수를 정해야 한다. 이때 기존 DOSC와는 달리 N에서 삭제 될 OSC 성분의 개수와 더해질 OSC 성분의 개수를 정해야 한다. 삭제될 OSC 성분의 개수의 경우 기존 OSC 알고리듬과 같이 너무 많은 성분개수를 사용할 경우 과적합의 위험이 있으며 너무 작은 수를 지정할 경우 노이즈가 제거되지 못할 수 있다(Wold *et al.*, 1998; Svensson *et al.*, 2002). 더해질 OSC 성분의 개수의 경우 너무 큰 수를 지정할 경우 과적합의 위험과 데이터에 노이즈가 포함될 위험이 있으며 너무 작은 수를 지정할 경우 N에 포함된 정보를 살리지 못할 위험이 있다. 본 연구에서는 삭제 그리고 더해질 OSC 성분 개수를 다양하게 고려하여 그 중 분류성능을 최대화하는

성분개수로 결정하였다.

4. 데이터 분석 결과

4.1 데이터

본 논문에 사용된 데이터는 질량분석계로부터 얻은 콘돔 윤활액의 스펙트라이다. 해당 데이터는 추후 범죄현장에서 미상의 윤활액 시그널 데이터를 확보할 경우 이 시그널이 어느 분류에 속한 데이터인지 예측하는데 도움을 주어 범죄수사에 유용한 정보를 얻기 위한 노력의 일환으로 수집되었다. 총 144개의 스펙트라를 추출하였으며 화학적 특성에 따라 <Table 1>과 같이 6개로 그룹화 하였다.

Table 1. Six classification of 144 condom lubricant spectra based on their chemical properties

용매	첨가제	개수
PDMS	No Detectable(High)	84
	Spermicidal	12
	No Detectable(Low)	12
PEG	No Detectable	24
	Benzocaine	6
	Multiple Distributions	6

<Table 1>은 추출된 콘돔 윤활액의 용매와 그에 첨가되는 첨가제에 따른 6개의 다른 그룹을 보여주고 있다. 본 스펙트라 데이터의 특징은 그룹 간 속해 있는 스펙트라의 수가 현저히 다르다는 점이다. <Figure 2>는 추출된 144개의 스펙트라를 그래프로 보여주고 있다.

<Figure 2>의 x축은 이온질량(m/z)을 나타내고 y축은 그들의 상대적인 존재량(intensity)을 나타내고 있다(Jung *et al.*, 2004). 본 스펙트라에는 총 24,817개의 이온질량이 있었으나 불필요한 부분을 제외하고 본 연구에서는 총 5,000개의 이온질량들을 특질로써 사용하였다.

본 연구에서 분석하고자 하는 고차원의 윤활액 스펙트라들이 <Table 1>에서 보여주는 화학적인 특성에 따라 분리가 잘 되는지 알아보기 위해 PCA 기법을 사용하였다. 먼저 주성분의 개수를 정하기 위해 각 주성분들이 데이터를 설명하는데 어느 정도 영향을 미치는지를 볼 수 있는 Scree plot을 통해 살펴보았다 <Figure 3>. Scree plot에 의하면 첫 3개의 주성분(PC1, PC2, PC3) 이 데이터 총 분산의 약 88%를 설명함을 보여주고 있으며 이를 토대로 3개의 주성분을 축으로 하여 데이터를 시각화 하였다<Figure 4>.

그림에서 보여주듯 PDMS에 속해있는 스펙트라와 PEG 그룹에 속해있는 스펙트라는 잘 구분됨을 알 수 있었지만 PEG와 PDMS 그룹 내에서는 그룹별 스펙트라가 뚜렷이 구별되지 않고 섞여 있음을 확인할 수 있었다. 본 연구에서는 OSC 기법을 통해 본 데이터와 같이 복잡하게 섞여있는 스펙트라의 분류성능을 향상시키고자 한다.

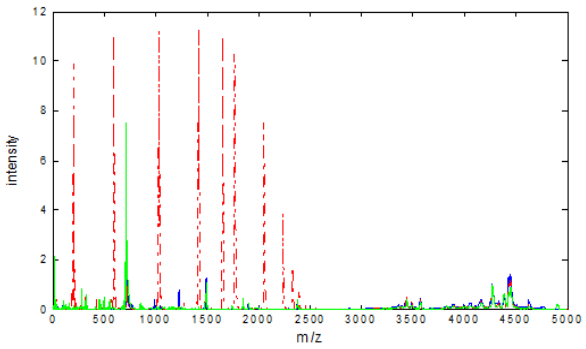

Figure 2. 144 NMR spectra of condom lubricants obtained by mass spectrometer

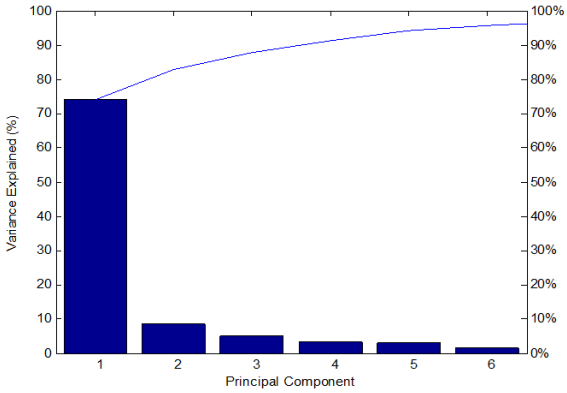

Figure 3. Scree plot showing the amount of variability explained by each PC

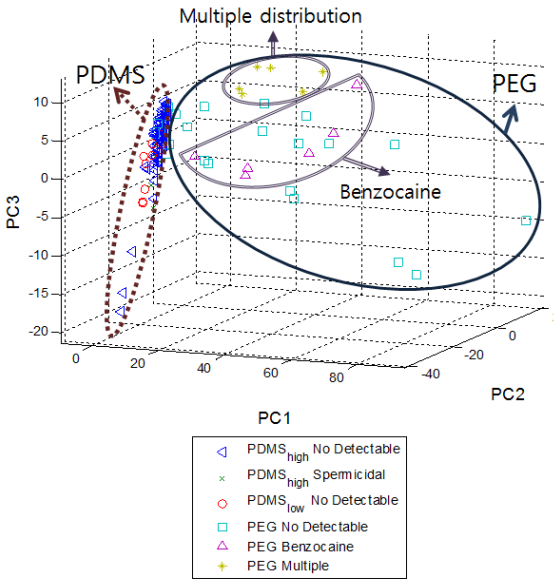

Figure 4. PCA score plot (PC1 vs. PC2 vs. PC3) of 144 condom lubricant spectra

4.2 Adjusted Direct Orthogonal Signal Correction

4.2.1 기존 DOSC 적용 결과

그림 <Figure 5>는 OSC 성분 개수를 달리한 DOSC 결과를 3차원 PCA Score plot으로 표현한 결과이다. <Figure 5>에서 성분 개수가 증가함에 따라 화학성분에 따른 6그룹의 스펙트라가 뚜렷이 군집됨을 확인할 수 있다. OSC 성분을 1개 사용할 경우 기존에 구분할 수 없었던 스펙트라 그룹(PEG No Detectable)이 새롭게 구분되기 시작하였고 성분을 2개 사용했을 때는 기존에 구분되지 않았던 또 다른 스펙트라 그룹(PDMS High No Detectable)이 새로이 구분되기 시작하였다. 성분을 3개 사용했을 때 모든 클래스의 구분이 뚜렷해 졌음을 볼 수 있었다. 성분을 4개 사용 할 경우에는 성분을 3개 사용했을 때와 큰 차이가 없었음을 보였다. 너무 많은 수의 성분을 사용함으로써 발생하는 과적합의 위험을 피하기 위해 본 연구에서는 3개의 OSC 성분을 사용하였다. DOSC 처리 전 PCA Score plot <Figure 4>과 비교해보았을 때 DOSC 데이터 처리를 거친 후 각 스펙트라 그룹간의 구분이 시각적으로 명확해 졌음을 확인할 수 있다.

4.2.2 기존 DOSC 기법의 한계

<Figure 5>에서 시각적으로 볼 수 있듯이, DOSC는 데이터에서 OSC 성분을 제거하여 노이즈를 제거한다. DOSC의 알고리즘 내부에서 제거되는 OSC 성분은 N에 PCA를 적용한 결과 얻어지는 주성분 중 첫 번째, 두 번째, 세 번째 주성분들이다. 이 성분들이 확실히 노이즈인지 확인하기 위하여 이들을 주성분을 축으로 한 그래프로 나타내어 보았다 <Figure 6>.

<Figure 6>을 보면 그룹 간 전혀 구분이 없는 것으로 보아 노이즈로 확인할 수 있었다. 하지만 N의 마지막의 3개의 주성분을 그래프로 표현하면<Figure 7>, 144개의 스펙트라가 6개의 점으로 보일 만큼 완벽하게 분류하고 있음을 볼 수 있다.

이는 기존 DOSC가 가정하고 있는 노이즈 부분에서 Y와 관계가 있는 부분이 존재하고 있음을 보여주고 있다. 본 논문에서 제안하고 있는 Adjusted DOSC는 기존 DOSC에서 삭제되는 Y에 관한 유용한 정보를 더해주어 기존 DOSC를 개선한다.

4.3 제안하는 Adjusted DOSC 기법 결과

4.3.1 OSC 성분의 개수 결정 및 PCA를 통한 DOSC와 비교

앞부분에서 언급하였든 Adjusted DOSC를 행하기 위해서는 빼주어야 할 성분과 더해주어야 할 성분 두 가지 모두를 정해야 한다. 본 연구에서는 두 종류의 OSC 성분 개수를 바꾸어보고 K-인접이웃 분류 알고리즘(K-nearest Neighbor; KNN)을 이용하여 가장 낮은 오분류율을 보이는 성분을 사용하였다.

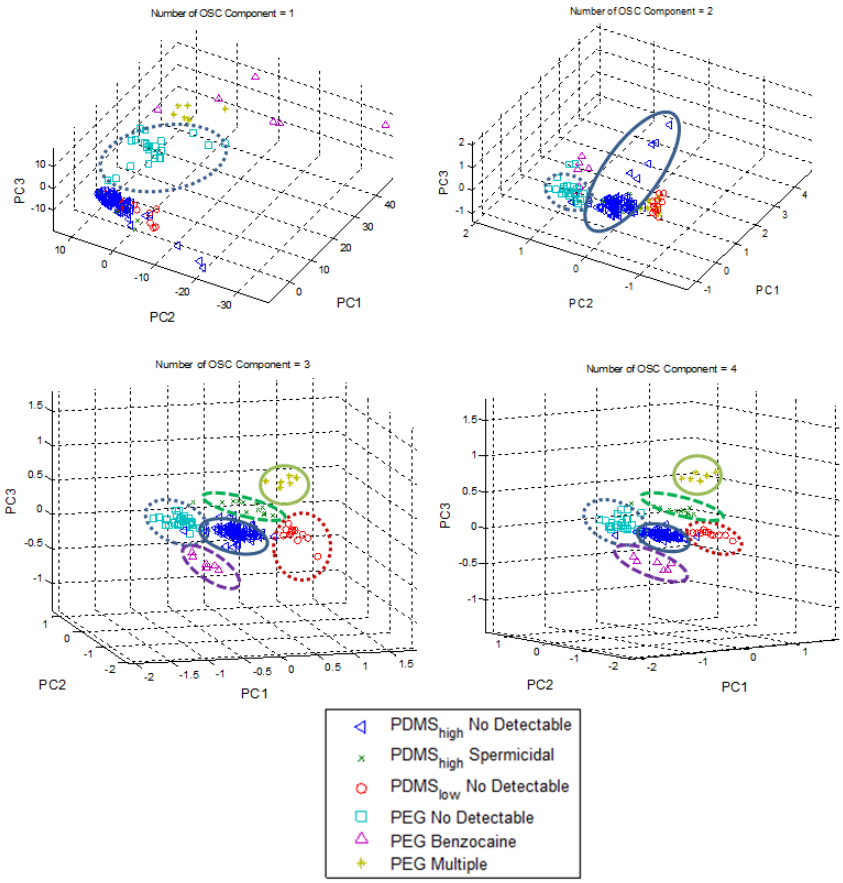

Figure 5. PCA score plots (PC1 vs. PC2 vs. PC3) of DOSC-processed data with different number of OSC components

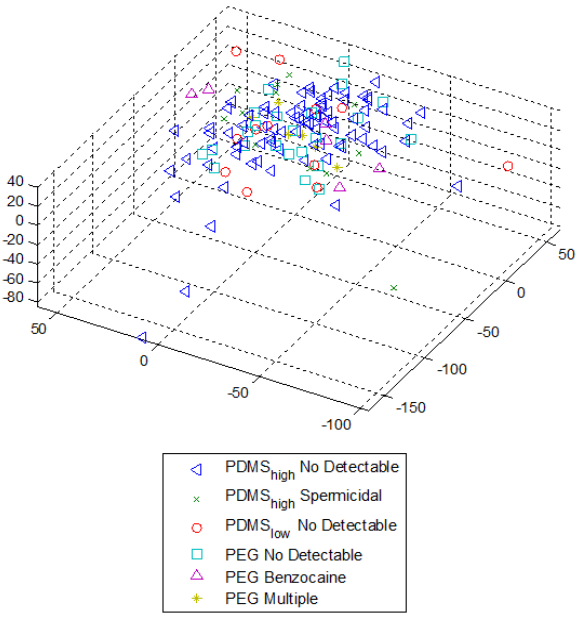

Figure 6. PCA score plot based on the first three PCs of N

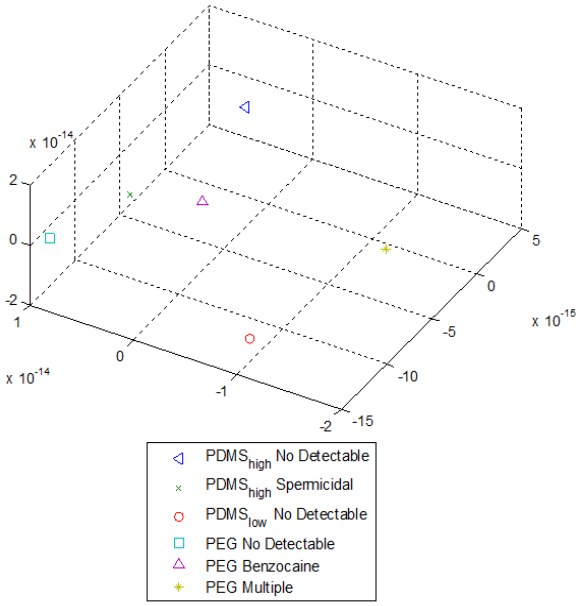

Figure 7. PCA score plot based on the last three PCs of N

KNN은 학습데이터를 모두 사용하지 않고 분류 혹은 예측하고자 하는 관측치와 가까운 k개의 데이터를 사용하는 알고리즘이다. KNN은 노이즈가 심한 데이터에 로버스트(robust) 하며 학습 데이터가 클 경우 정확한 예측력을 보이고 있다(Kim *et al.*, 2008). KNN 알고리즘을 사용하기 위해서는 두 가지 고려할 사항이 있는 있는데 이웃 점의 개수(k)와 거리종류이다. 보통 k의 결정은 일정 범위 안에서 k의 값을 변화시켜 분류성능이 최대화되는 k로 결정한다(Tan *et al.*, 2007). 거리함수를 결정하는 객관적 기준은 없으며 분석자의 주관에 따라 가장 데이터를 잘 설명하는 거리함수를 사용한다.

OSC 성분 개수를 알아내기 위하여 KNN의 k는 2를 사용하였으며, 거리함수로는 유클리드거리를 사용하였다. OSC 성분은 더해주는 성분과(added OSC) 빼주는 성분(deleted OSC) 모두 1에서 5까지 변화시켜 보았다. 전체 데이터의 80%를 학습데이터로 20%를 평가데이터로 랜덤하게 분할하고 각각을 KNN에 적용하여 평가데이터에 대한 오분류율을 구하였다. 각 경우에 대하여(총 25경우) 위 작업을 1000번 반복수행한 평균 오분류율이 <Table 2>에 나타나있다.

Table 2. Misclassification rates from KNN for various added and deleted OSC components

		Added OSC components				
		1	2	3	4	5
Deleted OSC components	1	0.180	0.175	0.175	0.174	0.170
	2	0.188	0.192	0.188	0.187	0.182
	3	0.191	0.190	0.188	0.185	0.183
	4	0.189	0.192	0.191	0.188	0.189
	5	0.190	0.191	0.191	0.192	0.185

<Table 2>에서 보여주고 있는 결과는 Adjusted DOSC 사용시 삭제될 OSC 성분개수가 1이고 보강될 OSC 성분 개수가 5일 때 KNN 결과가 가장 좋음을 보여주고 있다. 따라서 본 연구에서는 1개의 OSC 성분을 제거하고 5개의 OSC 성분을 첨가한 Adjusted DOSC를 사용하였다. 이 결과를 PCA로 나타내면 <Figure 8>과 같다. 기존 DOSC 결과를 보여주고 있는 <Figure 5>과 비교 했을 때 각 스펙트라 그룹간의 구분이 시각적으로 보다 명확해 졌음을 확인할 수 있다.

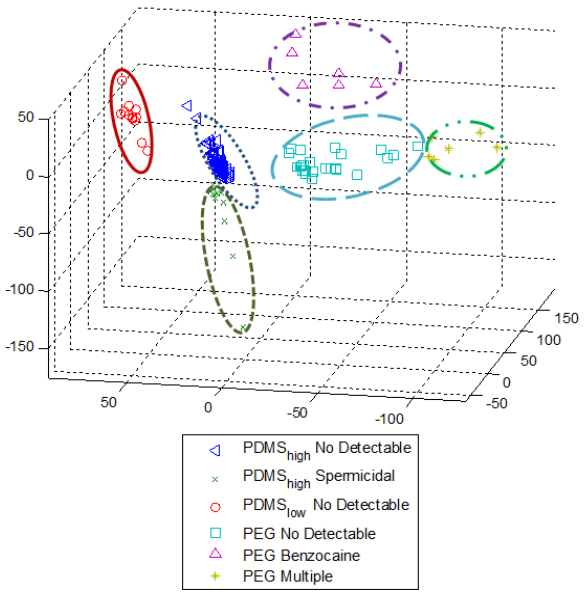

Figure 8. PCA score plot (PC1 vs. PC2 vs. PC3) of condom lubricant spectra processed by an Adjusted DOSC algorithm

4.3.2 분류알고리즘을 통한 DOSC와의 비교

시각화뿐 아니라 Adjusted DOSC의 분류능력이 기존 DOSC를 사용했을 경우 보다 향상되었음을 증명하기 위하여 데이터를 (1) 원본 데이터(Original), (2) DOSC 처리 후 데이터(after DOSC), 그리고 (3) Adjusted DOSC 처리 후 데이터(after Adjusted DOSC) 3개로 구분하여 분류실험을 하였으며 그 방법은 다음과 같다. 위에서 언급한 바와 같아 DOSC 성분의 개수는 3개로 Adjusted DOSC 성분은 보강성분 = 5, 삭제성분 = 1로 하였다.

원본 데이터, DOSC 처리 후 데이터, Adjusted DOSC 처리 후 데이터를 각각 80%의 학습데이터와 20%의 평가데이터로 랜덤하게 분할하고 각각의 데이터에 KNN 알고리즘을 적용하여 평가데이터에 대한 오분류율을 구하였다. 위 작업을 1000번 반복수행한 평균 오분류율이 <Table 3>에 나타나있다. 여기서 강조하고 싶은 것은 Adjusted OSC의 성분 개수를 결정할 때 사용하였던 평가데이터와 KNN 분류 시 사용된 평가데이터와는 다르다는 점이다. 이는 각각의 실험 시 랜덤하게 학습데이터와 평가데이터를 나누었기 때문이다.

Table 3. Misclassification rates obtained by KNN for the original, DOSC processed, and Adjusted processed data sets(standard errors are shown inside the parentheses)

	Original Data	DOSC Processed Data	Adjusted DOSC Processed Data
KNN(k = 2)	0.26(0.08)	0.19(0.07)	0.17(0.07)
KNN(k = 4)	0.31(0.08)	0.21(0.07)	0.17(0.07)
KNN(k = 8)	0.29(0.08)	0.24(0.08)	0.21(0.08)
KNN(k = 16)	0.33(0.09)	0.29(0.08)	0.26(0.07)

분류알고리즘으로 KNN을 사용하였고, 유클리드 거리를 사용하였다. KNN의 특성상 k가 짝수일 경우 클래스를 배정하는 과정에서 각각 클래스에 속하는 관측치의 수가 동일하게 나올 수 있다. 이러한 경우 클래스의 결정을 랜덤하게 (2개의 클래의 경우 각각 0.5의 확률로 결정) 결정하는 알고리즘을 사용 하였다. k의 개수의 영향을 보기 위해 4개의 다른 k를 사용 하여 결과를 비교하여 보았다.

<Table 3>에서 보여주듯 모든 경우의 데이터에 대해서 k=2일 때 가장 좋은 결과를 얻을 수 있었다. 또한 모든 k에서 제안한 Adjusted DOSC가 기존 DOSC 보다 더 나은 결과를 보임을 알 수 있었다. 결론적으로 위의 시각화와 분류실험을 통하여 본 논문에서 제안하고 있는 Adjusted DOSC의 성능이 기존의 DOSC 보다 향상되었음을 알 수 있다.

5. 결 론

스펙트라 분석은 매우 다양한 분야에서 행해지고 있지만, 데이터의 형태가 매우 고차원이고 노이즈의 영향을 많이 받기 때문에 분석이 매우 어렵다. 때문에 본 연구에서는 스펙트라 전처리 알고리즘으로 유용하게 쓰이고 있는 기존의 DOSC 기법을 개선하는 Adjusted DOSC를 제안하였다. 기존 DOSC는 노이즈를 제거하는 과정에서 유용한 Y의 정보를 제거하는 데 반해 제안한 Adjusted DOSC는 보정을 통해 유용한 Y 정보를 보충해 주고 있다. 본 연구에서는 실제 콘돔 윤활액 스펙트라 데이터를 이용하여 DOSC가 노이즈 제거 과정에서 유용한 Y 정보를 제거함을 알아내고 이 유용한 정보를 다시 데이터에 더하는 보정작업을 추가하였다. 그 결과 Adjusted DOSC는 기존 DOSC 보다 더 나은 시각화 효과를 보였으며 분류 알고리즘에서도 보다 더 나은 성능을 보였다.

참고문헌

Fearn, T. (2000), On orthogonal signal correction, *Chemometrics and Intelligent Laboratory Systems*, **50**, 47-52.

Jang, W. and Chang, W. (2006), A wavelet based feature selection method to improve classification of large signal-type data, *Journal of the Korea Institute of Industrial Engineers*, **32**, 133-140.

Jung, Y. S., Lee, W. I., and Lee, W. M. (2004), I*nstrumental analysis, basic and experiment,* Info-Tech Corea, Seoul, Korea.

Kher, A., Mulholland, M., Green, M., and Reedy, B. (2006), Forensic classification of ballpoint pen inks using high performance liquid chromatography and infrared spectroscopy with principal components analysis and linear discriminant analysis, *Vibrational Spectroscopy*, **40**, 270-277.

Kim, S. B., Chen, V. C. P., Park, Y., Ziegler, T. R., and Jones, D. P. (2008), Controlling the false discovery rate for feature selection in high-resolution NMR spectra, *Statistical Analysis and Data Mining*, **1**, 57-66.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005), *Applied Linear Statistical Models(5th edition)*, McGraw-Hill/Irwin, New York, USA.

Luypaert, J., Heuerding, S., Massart, D. L., and Vander, Y. V. (2006), Direct orthogonal signal correction as data pretreatment in the classification of clinical lots of creams from near infrared spectroscopy data, *Analytica Chimica Acta*, **582**, 191-189.

Park, Y., Kim, S. B., Wang, B, Blanco, R. A., Le, N.-A., Wu, S., Accardi, C. J., Alexander, R. W., Ziegler, T. R., and Jones, D. P. (2009), Individual variation in macronutrient regulation measured by proton magnetic resonance spectroscopy of human plasma, *AmJ Physiol Regul Integr Comp Physiol*, **297**, 202-209.

Saferstein, R. (2005), *An Introduction to Forensic Science*, Hanrimwon, Seoul, Korea.

Saiz-Abajo, M. J., Gonzalez-Saiz, J. M., and Pizarro, C. (2005), Orthogonal signal correction applied to the classification of wine and molasses vinegar samples by near-infrared spectroscopy. Feasibility study for the detection and quantification of adulterated vinegar samples, *Anal Bioanal. Chem*, **382**(2), 412-420.

Shmueli, G., Patel, N. R., and Bruce, P. C. (2007), *Data Mining for Business Intelligence*, Wiley, Hoboken, NJ.

Svensson, O., Kourti, T., and MacGregor, J. F. (2002), An investigation of orthognal signal correction algorithms and their characteristic, *Journal of Chemometrics*, **16**, 176-188.

Tan, P. N., Steinbach, M., and Kumar, V. (2007), *Introduction to data-mining*, Infinity books, Seoul, Korea.

Trygg, J. and Wold, S. (2002), Orthogonal projection to latent structures(O-PLS), *Journal of Chemometrics*, **16**, 119-128.

Westerhuis, J. A., Jong, S., and Smilde, A. K. (2001), Direct orthogonal signal correction, *Chemometrics and Intelligent Laboratory Systems*, 56, 13-25.

Witten, I. H., Frank, E., and Hall, M. A. (2011), *Data Mining*, Morgan Kaufmann, Burlington, MA.

Wold, S., Antti, H., Lindgren, F., and Ohman, J. (1998), Orthogonal signal correction of near-infrared spectra, *Chemometrics and Intelligent Laboratory Systems*, **44**, 175-185.

Zhu, D., Ji, B., Meng, C., Shi, B., Tu, Z., and Qing, Z. (2007), The application of direct orthogonal signal correction for linear and non-linear multivariate calibration, *Chemometrics and Intelligent Laboratory Systems*, **90**, 108-115.

Zijlstra, W. G., Buursma, A., and Assendelft, O. W. (2000), *Visible and Near Infrared Absorption Spectra of Human and Animal Hemoglobin*, *Determination and Application*, VSP, Amsterdam, The Netherlands.