응용통계연구 제10권 2호
1997년 9월 pp. 227-239

부분최소제곱회귀(Partial Least Squares Regression) 이론과 분광학적 혈중 헤모글로빈 진단에의 응용[1)]

김 선 우[2)], 김 연 주[3)], 김 종 원[4)], 윤 길 원[3)]

요 약

분광학분야에서 측정되는 자료는 여러 파장에서 측정된 스펙트럼 행렬과 이 스펙트럼을 통하여 알고자하는 어떤 반응치들의 행렬 또는 벡터로 주어진다. 이 경우 측정자료에의 많은 잡음(noise)과 파장간의 상관관계가 내재한다. 부분최소제곱회귀 방법은 여러 개의 파장에서 측정된 자료를 모두 이용하는데 자료축약과정을 통하여 자료의 잡음 문제와 상관관계 문제를 해결하는 다변량통계방법이다. 본 연구에서는 이러한 자료에 적합한 부분최소제곱회귀의 이론을 알아보고 실제로 측정된 자료를 통하여 주어진 스펙트럼에 대한 반응치의 예측을 부분최소제곱회귀 방법을 이용하여 고찰하였다.

1. 서론

부분최소제곱회귀(Partial least squares regression; 이하 PLSR) 방법은 분석화학, 물리화학, 임상화학, 생산공정제어 등 많은 분야에서 유용한 통계적 방법으로 그 중요성이 증대되고 있다. 특히 최근 PLSR을 응용한 관심의 대상이 되는 분야로 분광학자료를 이용하여 혈중 성분의 농도를 진단하는 것을 들 수 있다. 혈액 속에 포함되어 있는 헤모글로빈 등 여러 가지 성분의 진단은 임상적으로나 의료 공학적으로 활용가치가 매우 크다. 비교적 많은 연구가 되어있고 또 제품도 출시되어 있는 것으로 혈중산소포화도나 빌리루빈 농도를 진단하는 것을 들 수 있는데, 이들은 측정 성분에 흡수가 민감한 파장과 그렇지 못한 파장의 흡수 비를 분석하는 단순한 이론에 바탕을 둔 것이 대부분이다(1-3). 이런 방법은 2개 ~ 4개의 제한된 파장에서 흡수도를 측정하여 보정(calibration)하는 것으로 혈액 속의 타 성분이 미치는 영향을 고려하기가 힘들고 또 측정하고자하는 성분의 혈중량이 작은 경우에는 사용하기가 어렵게 된다. 따라서 가시광선이나 근적외선을 포함하는 여러 파장에 대한 흡수 스펙트럼을 통계적으로 해석하여 혈중 성분의 농도를 보다 정확하게 추정하는 것이 필요하다. 분광학자료를 분석하는 많은 방법들이 측정된 흡수도와 농도간의 선형관계를 나타내는 Beer's law에 근거하여 제안되었는데 Classical least squares(CLS)(4-7), Inverse least squares(ILS)(8,9), 주성분회귀(Principal component

1) 본 연구는 '96년도 보건의료기술연구개발사업의 지원에 의하여 이루어진 것임.
2) (135-230) 서울시 강남구 일원동 50 삼성생명과학연구소 임상의학연구센터
3) (135-230) 서울시 강남구 일원동 50 삼성생명과학연구소 임상의공학연구센터
4) (135-230) 서울시 강남구 일원동 50 삼성서울병원 임상병리과

regression; 이하PCR)(10-12), PLSR(13-18) 등이 이에 속한다. 이 방법들 중 PCR, PLSR은 주성분분석과 같은 다변량분석법에 의해 유도되는 인자(factor)에 근거한 통계적 방법이며 PLSR은 CLS와 ILS의 장점을 갖고 있고(15) PCR과 비교해볼 때 더 적은 수의 인자로 PCR과 같은 정확도를 유지할 수 있으며 더 로버스트(robust)하다고 알려져 있다(19,20). 본 논문에서는 PLSR의 이론을 알아보고 이 연구를 위하여 7g/dl ~ 17g/dl의 다양한 농도의 헤모글로빈을 함유하는 혈액을 채취하여 500nm ~ 795nm 범위의 파장에서 흡수 스펙트럼을 측정한 자료로부터 헤모글로빈의 양을 진단하는데 통계적 방법인 PLSR의 이론을 적용하였다.

2. PLSR의 통계적 이론

보정이란 주어진 측정자료 **X**로부터 미지의 **y**에 대한 예측을 위하여 이미 측정된 **X**와 **y**의 자료 및 이에 대한 사전지식을 이용하여 $\mathbf{y} = f(\mathbf{X})$와 같은 **X**와 **y**의 수학적 관계식 또는 모형을 추정하는 과정을 말하며 예측(prediction)이란 보정과정으로부터 추정된 관계식 또는 모형으로 **y**를 예측하는 과정을 말한다. 여기서 보정과정에서 이용되는 자료 (**X**, **y**)를 보정자료(calibration set)이라 하고 다른 자료 **x**가 주어졌을 때 이에 대한 y를 예측하기 위해 주어진 자료 **x**를 예측자료(prediction set)이라 부른다. PLSR은 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 과 같은 회귀 식에서 **X**대신 인자라 부르는 서로 상관되어 있지 않은 열들로 구성된 행렬을 이용하는 방법으로 이 인자들은 주성분분석의 주성분과 유사하나 이 인자들이 **X**와 **y**의 정보를 모두 포함하도록 유도되고 추정된다.

먼저 n개의 변수에 대해 측정된 m개의 표본의 관측 값들로 구성된 (m × n) 자료행렬을 **X**라 하자. 일반적으로 자료행렬 **X**의 열 벡터들이 서로 상관되어 있고 자료 고유의 정보이외에 다른 잡음이 포함되어 있는 경우 자료축약과정을 통하여 자료의 분석 및 해석을 용이하게 할 수가 있다. **X**의 주성분행렬을 **T**라하면 주성분변환(principal component transformation)을 통하여 다음과 같이 **X**를 점수행렬(score matrix) **T**로 표현할 수 있다.

$$\mathbf{T} = \mathbf{XV}$$

X와 관련 자료벡터 **y**의 회귀관계식을 **T**를 이용하여 다음과 같은 모형으로 생각할 수 있다.

$$\mathbf{y} = \mathbf{Tq} + \mathbf{f} \tag{2.1}$$

그러면 **q**를 추정한 후 **y**는 다음과 같이 예측할 수 있다.

$$\hat{\mathbf{y}} = \mathbf{XV}\hat{\mathbf{q}}$$

위에서 **T**의 열 벡터(이하 인자라 부름)들이 서로 독립이 되도록 그리고 **X**의 계수(rank)보다 적은 수의 인자들만을 분석에 이용하여 행렬 **T**에 대한 **y**의 회귀식을 추정하면 자료의 공선성 문제로 야기될 수 있는 변수의 선택문제와 자료에 포함된 불필요한 잡음의 제거문제를 동시에 해결할 수가 있다.

그러나 **X**와 **y** 모두가 행렬 **T**의 회귀식으로 표현될 경우 다음과 같은 보정모형(calibration model)(2.2)을 생각할 수 있다.

$$\begin{aligned} \mathbf{X} &= \mathbf{TP}' + \mathbf{E} \\ \mathbf{y} &= \mathbf{Tq} + \mathbf{f} \end{aligned} \tag{2.2}$$

모형 (2.1)과는 달리 모형 (2.2)에서는 관계식 $\mathbf{T}=\mathbf{XV}$의 $\mathbf{V}$가 자료 $\mathbf{X}$와 $\mathbf{y}$로부터 추정된다.

적재행렬(loading matrix) $\mathbf{P}$와 적재벡터(loading vector) $\mathbf{q}$는 각각 $\mathbf{T}$에 대한 $\mathbf{X}$와 $\mathbf{y}$의 회귀계수행렬과 회귀계수벡터로 생각할 수 있으며 $\mathbf{E}$와 $\mathbf{f}$는 모형에 의해 설명되지 않는 오차항이다.

여기서 $\mathbf{V}$와 적절한 인자의 수는 보정과정으로부터 추정되고 결정되며 추정된 $\hat{\mathbf{V}}$에따라 $\hat{\mathbf{T}}$가 구해진다. 그러면 $\mathbf{P}$와 $\mathbf{q}$는 최소자승추정방법으로 다음과 같이 추정되며

$$\hat{\mathbf{P}}' = (\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}'\mathbf{X}$$

$$\hat{\mathbf{q}} = (\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}'\mathbf{y}$$

오차항 $\mathbf{E}$와 $\mathbf{f}$는 다음과 같이 계산된다.

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{T}}\hat{\mathbf{P}}'$$

$$\mathbf{f} = \mathbf{y} - \hat{\mathbf{T}}\hat{\mathbf{q}}$$

위의 모형에 근거하여 주어진 예측자료 $\mathbf{x}$에 대한 y의 예측과 예측과정에서 나타나는 $\mathbf{x}$의 오차항은 다음과 같다.

$$\hat{\mathbf{t}}' = \mathbf{x}'\hat{\mathbf{V}}$$

$$\mathbf{y} = \hat{\mathbf{t}}'\hat{\mathbf{q}}$$

$$\mathbf{e}' = \mathbf{x}' - \hat{\mathbf{t}}'\hat{\mathbf{P}}'$$

보정모형에서 인자의 수의 최대 값은 $\mathbf{X}$의 계수이며 인자의 수가 $\mathbf{X}$의 계수와 같으면 $\mathbf{X}$의 오차항 $\mathbf{E}$는 $\mathbf{0}$이 된다. 몇 개의 인자를 사용할 것인가는 모든 형태의 자료축약방법에서 거론되는 문제인데 이에 대한 기본적인 원칙은 $\mathbf{y}$를 좀 더 잘 예측하는데 기여하는 인자들만을 모형에 포함시키는 것이다.

자료행렬 $\mathbf{X}$에서 $\mathbf{y}$와 관련된 정보들중 많은 부분이 몇 개의 인자들로 표현될 수 있다. PLSR은 모형 (2.2)의 한 형태로 두 자료행렬 $\mathbf{X}$와 $\mathbf{y}$ 모두를 사용하여 $\mathbf{V}$를 추정하는 방법이다. 따라서 $\mathbf{X}$에 관한 주성분분석에서와 같이 $\mathbf{P'X}$의 분산을 최대화하는 $\mathbf{X}$의 적재행렬 $\mathbf{P}$를 구하는 대신 선형결합 $\mathbf{Xw}$와 $\mathbf{y}$간의 공분산을 최대화하는 가중적재벡터(loading weight vector), $\mathbf{w}$, 를 먼저 유도한다. 그러면 $\mathbf{V}=\mathbf{W}(\mathbf{P'W})^{-1}$과 같이 되며 아래의 알고리듬에서는 특별히 $\mathbf{V}$를 다시 계산하지 않고 보정과정과 예측과정을 수행한다. 보정과정에서는 자료의 변이를 가장 많이 설명하는 몇 개의 인자들만을 구하는 것으로 충분하기 때문에 자료의 변이를 가장 많이 설명하는 인자부터 시작하여 더 이상 자료의 유용한 정보를 포함하는 인자를 구할 수 없을 때까지 한 번에 한 개의 인자를 구하는 알고리듬이 유용하다. NIPLAS 알고리듬 (21)은 이 경우 사용할 수 있는 알고리듬인데 각 인자에 대해 먼저 구해진 인자들을 추정한 후 계산된 오차항으로부터 다시 적재벡터와 점수벡터를 구하는 반복절차가 수행된다.

상기의 이론을 바탕으로 PLSR의 보정을 위한 이 알고리듬은 다음과 같으며 각 인자에 대해 step 1부터 step 6을 수행한다.

step 0. 보정자료 $\mathbf{X}$와 $\mathbf{y}$에서 평균을 빼준다.(centering)

필요하다면 $\mathbf{X}$를 표준편차로 나누어준다.(scaling)

인자의 수 a를 1로 놓는다.

step 1. $\mathbf{w}_a'\mathbf{w}_a = 1$의 조건하에 선형결합 $\mathbf{Xw}_a$와 $\mathbf{y}$간의 공분산을 최대화하는 가중적재벡터 (loading weight vector) $\mathbf{w}_a$를 다음과 같이 구한다.

$$\hat{\mathbf{w}}_a = \mathbf{X}'\mathbf{y}/\mathbf{y}'\mathbf{y}$$

step 2. 구해진 $\hat{\mathbf{w}}_a$에 대한 **X**의 사영(projection)인 점수벡터 $\hat{\mathbf{t}}_a$는 다음과 같다.

$$\hat{\mathbf{t}}_a = \mathbf{X}\,\hat{\mathbf{w}}_a$$

step 3. 적재벡터 $\hat{\mathbf{p}}_a$을 구하기 위하여 $\hat{\mathbf{t}}_a$에 **X**를 회귀시키면 $\hat{\mathbf{p}}_a$는 다음과 같다.

$$\hat{\mathbf{p}}_a = \mathbf{X}'\hat{\mathbf{t}}_a/\ \hat{\mathbf{t}}_a'\hat{\mathbf{t}}_a$$

step 4. $\hat{q}_a$를 구하기 위하여 $\hat{\mathbf{t}}_a$ 에 **y**를 회귀시키면 $\hat{q}_a$는 다음과 같다.

$$\hat{q}_a = \mathbf{y}'\hat{\mathbf{t}}_a/\ \hat{\mathbf{t}}_a'\hat{\mathbf{t}}_a$$

step 5. **X**와 **y**의 오차항을 계산한다.

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{t}}_a\ \hat{\mathbf{p}}_a'$$

$$\mathbf{f} = \mathbf{y} - \hat{q}_a\ \hat{\mathbf{t}}_a$$

step 6. a를 하나 증가시키고 step 5에서 구해진 **X**의 오차항 **E**와 **y**의 오차항 f를 각각 **X**와 **y**로 대치한 후 step 1로 돌아간다.

이 알고리듬에서는 적절한 인자의 수만큼 위의 과정을 수행하는데 매번 구해지는 $\hat{\mathbf{w}}_a$는 서로 직교하게 된다. Step 5에서 구해진 점수벡터 $\hat{\mathbf{t}}_a$와 적재벡터 $\hat{\mathbf{p}}_a$의 곱과 곱 $\hat{\mathbf{t}}_a\ \hat{q}_a$는 각각 자료행렬 **X**와 **y**에 대하여 모형으로부터 추정된 PLSR 근사치로 인자의 수 a가 **X**의 계수일 때 **E**의 원소들은 모두 0이 된다.

자료벡터 **x**가 주어졌을 때 보정과정에서 추정된 모수들을 가지고 y를 예측하기 위한 과정은 다음과 같다.

step 0. 예측자료 **x**에서 보정자료의 평균을 빼준다.(centering)

보정과정에서 자료행렬 **X**를 표준편차로 나누어준 경우 **x**를 보정자료의 표준편차로 나누어준다.(scaling)

인자의 수 a를 1로 놓는다.

step 1. $\hat{\mathbf{t}}_a = \hat{\mathbf{w}}_a'\mathbf{x}$

step 2. $\mathbf{y}_a = \mathbf{y}_{a-1} + \hat{q}_a\,\hat{\mathbf{t}}_a$

step 3. $\mathbf{e}_a = \mathbf{e}_{a-1} - \hat{\mathbf{p}}_a\,\hat{\mathbf{t}}_a$

step 4. a를 하나 증가시키고 **x**를 step 3에서 구해진 $\mathbf{e}_a$로 대치한 후 step1로 되돌아간다.

여기서 $\hat{\mathbf{w}}_a$, $\hat{q}_a$ 및 $\hat{\mathbf{p}}_a$는 보정과정에서 추정된 추정치이며 $\mathbf{y}_0$는 보정자료에서 **y**의 원소들의 평균으로, $\mathbf{e}_0$는 **x**로 놓고 보정과정에서 결정된 인자의 수만큼 예측과정을 수행한다.

추정된 모형의 점수도(score plot)으로부터 자료들의 대략적인 패턴을 알 수 있고 적재도(loading plot)으로부터는 변수와 모형간의 대략적인 관계를 알 수가 있다. 즉 점수도에서 원점 근처에 있는 표본은 자료의 평균과 비슷한 형태를 가지며 적재도에서 원점근처에 위치하는 변수는 모형에 의해 설명되는 변동의 크기가 아주 적은 변수임을 알 수 있다.

모형을 과대 적합하지 않으면서 자료의 정보를 가능한 한 많이 포함하도록 인자의 수를 적절히 선택하여야한다. 이러한 목적으로 보정과정에서 교차타당성(cross-validation)방법을 적용할 수 있는데 즉 m개의 표본에서 (m-1)개의 표본만을 사용하여 보정한 후 나머지 한 개의 표본을 예측자료로 이용하여 y를 예측한다. 이러한 과정을 모든 자료에 대해 반복하면 m번의 보정과정을 수행하게 되는데 이때 각 표본에 대해 예측한 y와 이미 알고있는 y를 비교할 수가 있다. 어떤 한 인자의 수에 대해 적합된 PLSR 모형이 얼마나 잘 맞는가는 $\mathbf{e}_y'\mathbf{e}_y$ (PRESS; prediction error sum of squares)($\mathbf{e}_y = \mathbf{y} - \hat{\mathbf{y}}$)으로 측정될 수 있다. 따라서 최소의 PRESS 값을 주는 인자의 수(a^*)를 갖는 모형이 최적이 된다. 그러나 모형이 자료에 너무 과대 적합되지 않도록 하기위하여 a^*보다 더 작은 a를 갖는 모형에 대한 PRESS 값을 비교해 볼 필요가 있다. 다음과 같은 과정에 따라 a^*를 갖는 모형의 PRESS 값보다 유의하게 크지 않은 PRESS 값을 갖는 모형이 갖는 인자의 수를 구할 수 있다.

step 1. $F(a) = PRESS(a) / PRESS(a^*)$ $(a = 1, 2, ..., a^*)$

step 2. $F(a) < F_{\alpha;m,m}$을 만족하는 a중 가장 작은 a를 최적의 인자의 수로 결정하며 $F(a) < F_{\alpha;m,m}$을 만족하는 a가 없다면 a^*가 최적의 인자의 수로 결정된다. 여기서 $F_{\alpha;m,m}$는 자유도가 (m,m)인 F-분포의 우측 (100)α%에 대응되는 값이다.

3. 혈중 헤모글로빈 진단에 대한 PLSR의 적용

앞에서 고찰된 몇 가지 이론들을 실제로 측정된 분광학 자료에 적용해보자. 환자로부터 얻은 EDTA 혈액으로부터 SE8000 혈구계산기(Sysmax, Japan)를 이용하여 총 혈색소량(Total Hemoglobin : g/dl)을 결정한 다음, 혈액검체의 일부를 가지고 500nm ~ 795nm에 해당하는 영역에서 1nm의 해상도(resolution)으로 각 파장에 대한 흡수도(absorbance)를 측정하였다. 가시광선대역에서 흡수도가 큰 혈액의 측정을 위해 큐벧(cuvette)의 두께가 0.5mm에 해당하는 슬라이드(slide) 방식의 큐벧을 이용하였다. 스펙트럼은 Cary5E(Varian, Australia) 분광광도계로부터 얻었다. 혈액 스펙트럼 측정과 동일한 조건에서 물의 스펙트럼을 측정하였다. 혈액구성 성분중 대부분을 차지하는 물의 혈액 스펙트럼에 미치는 영향을 제거하기 위해 각 혈액 스펙트럼에서 물의 스펙트럼을 빼주었다. 이렇게 얻어진 혈액스펙트럼 70개가 분광학 자료로 이용되었다(N001 ~ N070). 이중 예측자료로 N041 ~ N050에 이르는 10개의 자료가, 나머지 60개의 자료는 보정자료로 이용되었다.

<그림 1>은 보정과정에서 이용된 60개 표본의 평균 스펙트럼과 예측에 사용된 10개 표본의 스펙트럼을 나타낸다.

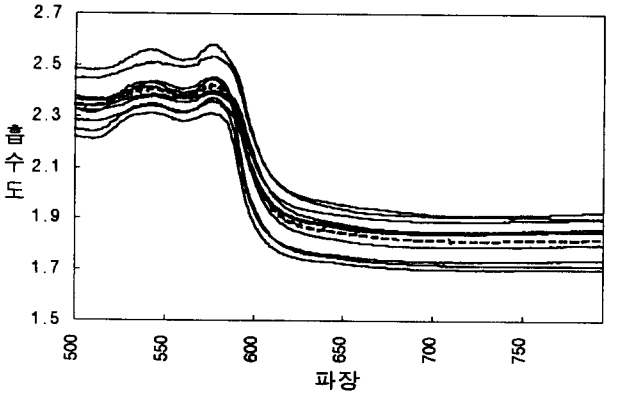
PLSR의 효율적인 수행을 위하여 보정과정과 예측과정에서 NIPALS 알고리듬이 이용되었고 적절한 인자의 수를 결정하기 위하여 교차타당성기법이 적용되었다. 교차타당성 적용결과 각 인자의 수(a)에 대한 PRESS를 계산한 결과는 <표 1>과 <그림 2>과 같다. 여기서 최소의 PRESS 값을 갖는 모형은 인자가 5개인 모형이다. 그러나 $F(a) < F_{0.05;60,60}$ (=1.534)를 만족하는 가장 작은 인자의 수는 4이다. 따라서 4개의 인자를 갖는 모형이 가장 적절한 보정 모형으로 판단되었다. 그림 <3-1>에서 그림 <3-7>은 인자의 수가 하나씩 증가해 감에 따라 만들어

지는 PLSR 모형의 가중적재벡터와 적재벡터를 나타낸다. 인자가 3개일때까지의 모형의 가중적재벡터, $\mathbf{w}_a$,와 적재벡터, $\mathbf{p}_a$,는 차이를 보이지만 인자의 수가 4개 또는 5개를 갖는 모형의 가중적재벡터와 적재벡터는 잘 일치함을 알 수 있다. 또한 처음 4개 인자의 적재스펙트럼(loading spectrum)은 부드러운 곡선의 형태인 반면 그 이상의 인자의 수를 갖는 모형은 임의잡음(random noise)를 반영하고 있음을 알 수 있는데 이는 교차타당성 결과와 일치한다.

또한 PLSR외에 예측자료의 스펙트럼을 가지고 이에 대한 혈중 헤모글로빈을 예측하기 위한 방법으로 CLS, ILS와 PCR을 생각할 수 있는데 ILS는 표본의 수가 적어도 파장의 수만큼 되어야 적용가능하기 때문에 제외하고 CLS와 PCR을 자료에 적용하였다.

CLS, PCR 및 PLSR의 적용결과는 <표 2>와 같다. SEP(Standard Error of Prediction)은 예측자료의 실제농도를 알고 있을 때 산출 가능한 값인데 예측자료의 실제농도와 모형에 의해서 예측된 농도의 차이의 제곱의 합을 예측자료의 크기로 나눈 값의 제곱근으로 적용된 방법 또는 모형이 얼마나 정확한 예측치를 주는가에 대한 기준으로 이용된다. SEP 값을 보면 인자에 근거한 통계적 방법인 PCR과 PLSR이 CLS보다 매우 우월함을 알 수 있다. 특히 CLS는 Beer's law에 적합한 자료인 경우에 적용하면 좋은 방법인데 <그림 1>에서 볼 수 있듯이 이용된 자료는 흡수도와 농도의 선형관계가 항상 성립되지는 않기 때문에 부정확한 결과를 산출하였다. 또한 PCR과 PLSR의 적합한 모형은 모두 인자의 수가 4개이나 PLSR이 PCR보다 더 정확한 예측치를 제공함을 알 수 있다.

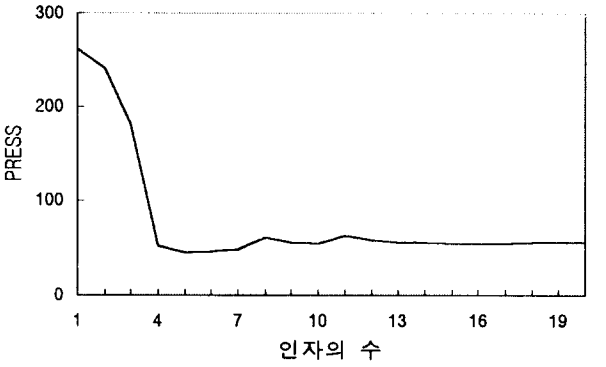
<그림 4>는 추정된 PLSR 모형으로부터 예측자료의 헤모글로빈 농도를 예측한 값과 실제 농도 값을 나타내는데 실제 값과 예측 값이 정확히 같지는 않으나 임상적으로 볼 때 어느 정도 잘 일치함을 알 수 있다.



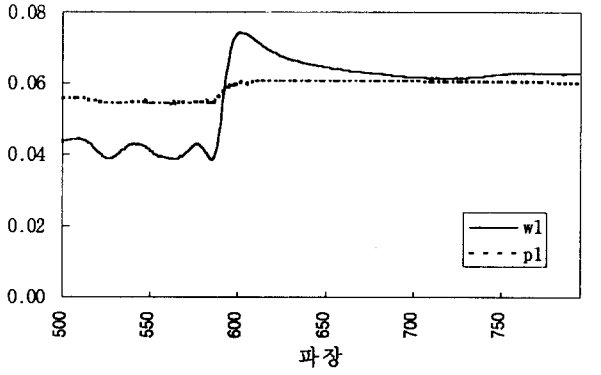
<그림 1> 예측자료의 스펙트럼과 보정자료의 평균 스펙트럼

< 표 1 > 인자의 수와 PRESS

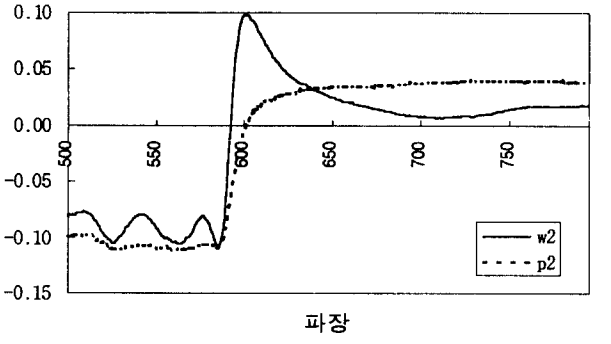
factor의 수 (a)	1	2	3	4	5	6	7	8	9	10
PRESS	261.23	240.17	181.18	51.35	44.22	45.29	48.43	60.77	55.16	54.41
factor의 수 (a)	11	12	13	14	15	16	17	18	19	20
PRESS	63.27	58.11	55.69	55.72	54.26	54.14	54.62	55.02	55.29	55.43



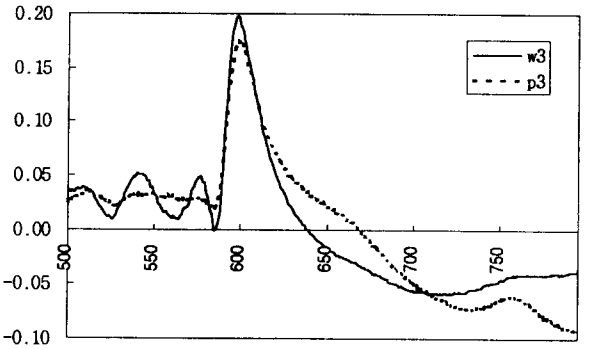
<그림 2> 인자의 수와 PRESS



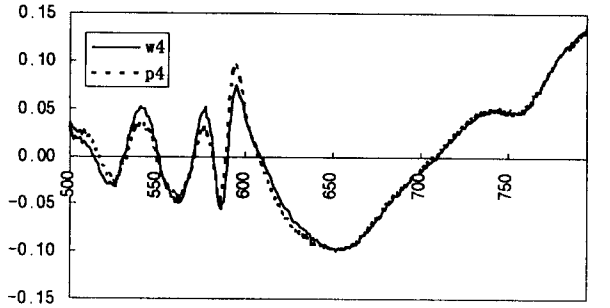
<그림 3-1> 가중적재벡터 w1, 적재벡터 p1



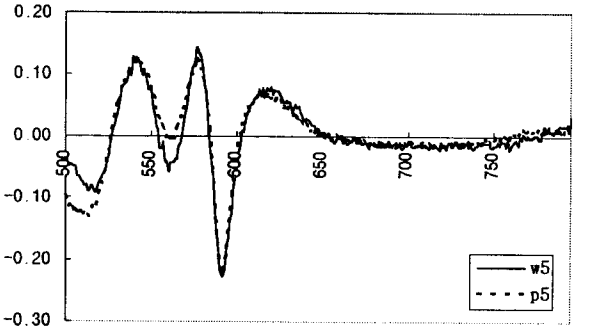
<그림 3-2> 가중적재벡터 w2, 적재벡터 p2



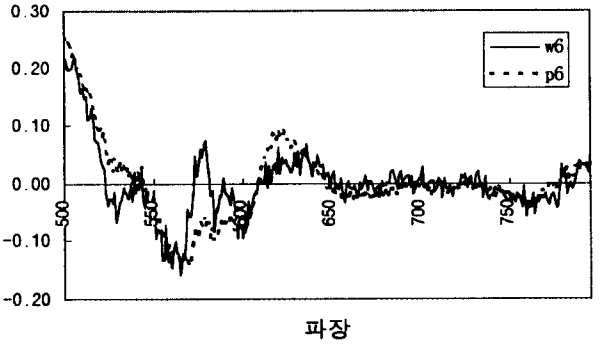
<그림 3-3> 가중적재벡터 w3, 적재벡터 p3



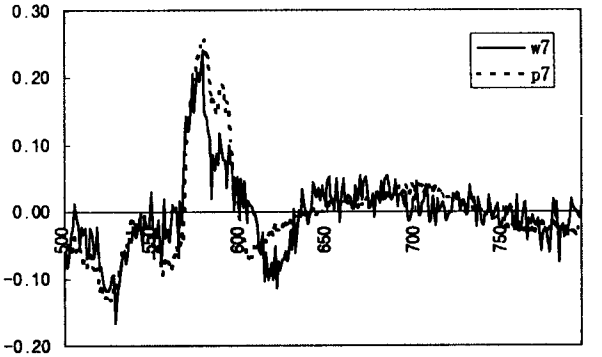
<그림 3-4> 가중적재벡터 w4, 적재벡터 p4



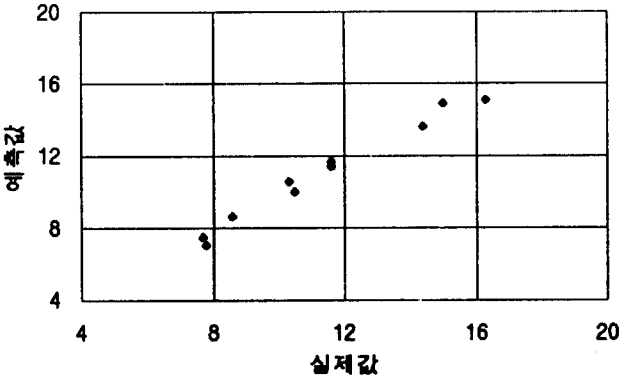
<그림 3-5> 가중적재벡터 w5, 적재벡터 p5



<그림 3-6> 가중적재벡터 w6, 적재벡터 p6



<그림 3-7> 가중적재벡터 w7, 적재벡터 p7



<그림 4> 헤모글로빈농도 실제값과 예측값

< 표 2 > CLS, PCR, PLSR의 적용 결과

실제농도	7.7	7.8	8.6	10.3	10.5	11.6	11.6	14.4	15	16.3	SEP
PLSR	7.51	7.02	8.64	10.59	9.99	11.41	11.67	13.63	14.91	15.09	0.56
PCR	6.83	6.54	8.13	10.73	10.15	11.46	11.86	13.44	14.51	14.79	0.80
CLS	11.25	10.91	11.10	11.67	12.18	11.63	11.81	11.37	12.25	11.77	2.66

4. 결론

본 논문에서 고려한 PLSR 방법은 분광학자료와 같이 잡음이 많고 변수간 상관관계가 내재하는 자료행렬의 분석에 매우 유용한 방법으로 생각된다. PLSR은 여러 개의 변수로 측정된 자료와 이 자료에 대한 반응자료로부터 몇 개의 인자를 유도하여 유도된 인자들과 측정된 자료 및 반응자료와의 관계식을 추정하고 주어진 자료에 대한 미지의 반응 값을 예측하는 다변량분석방법이다. 적합한 모형이 갖는 인자의 수는 교차타당성 기법을 적용하여 결정되었다. 각 인자 수에 대한 모형의 가중적재벡터와 적재벡터의 그림으로부터 자료에 대해 각 모형이 자료에 어떻게 적합되는가를 알 수 있다.

본 논문에서 실제로 혈중 헤모글로빈 농도의 예측에 CLS, PCR 및 PLSR을 적용해본결과 인자에 근거한 통계적 방법을 적용함이 적합함을 알 수 있었으며 이러한 방법들중 PLSR 방법이 가장 정확한 결과를 제공하였고 그 결과는 만족할만하였다.

참고문헌

[1] Hamza, M. and Hamza, H. (1988). Laser transcutaneous bilirubin meter, *SPIE*, Vol. 907 (Laser surgery: Characterization and therapeutics), 146-149

[2] Mendelson, Y. (1992). Pulse oximetry: Theory and applications for noninvasive monitoring, *Clinical Chemistry*, Vol. 38, No. 9, 1601-1607

[3] Kitai, T., Tanaka, A., Tokuka, A., Tanaka, K., Yamaoka, Y., Ozawa, K., and Hirao, K. (1993). Quantitative detection of hemoglobin saturation in the liver with near-infrared spectroscopy, *Hepatology*, Vol. 18, No. 4, 926-936

[4] Antoon, M. K., Koening, J. H., and Koenig, J. L. (1977). Least-squares curve-fitting of Fourier transform infrared spectra with applications to polymer systems, *Applied Spectroscopy*, Vol. 31, No. 6, 518-524

[5] Haaland, D. M. and Easterling, R. G. (1980). Improved sensitivity of infrared spectroscopy by the application of least squares methods, *Applied Spectroscopy*, Vol. 34, No. 5, 539-546

[6] Haaland, D. M. and Easterling, R. G. (1982). Application of new least-squares methods

for the quantitative infrared analysis of multicomponent samples, *Applied Spectroscopy*, Vol. 36, No. 6, 665-672

[7] Haaland, D. M. and Easterling, R. G., and Vopicka, D. A. (1985). Multivariate least-squares methods applied to the quantitative spectral-analysis of multicomponent samples, *Applied Spectroscopy*, Vol. 39, 73-83

[8] Brown, C. W., Lynch, P. F., Obremski, R. J. and Lavery, D. S. (1982). Matrix representations and criteria for selecting analytical wavelengths for multicomponent spectroscopic analysis, *Analytical Chemistry*, Vol. 54, 1472-1479

[9] Kisner, H. J., Brown, C. W. and Kavarnos, B. J. (1983). Multiple analytical frequencies and standards for the least-squares spectrometric analysis of serum lipids, *Analytical Chemistry*, Vol. 55, 1703-1707

[10] Fredericks, P. M., Lee, J. B., Osborn, P. R. and Swinkels, D. A. (1985). Materials characterization using 인자 analysis of FT-IR spectra. Part 1: Results, *Journal of Applied Spectroscopy*, Vol. 39, No.2, 303-310

[11] Fredericks, P. M., Lee, J. B., Osborn, P. R. and Swinkels, D. A. (1985). Materials characterization using factor analysis of FT-IR spectra. Part 2: Mathematical and statistical considerations, *Journal of Applied Spectroscopy*, Vol. 39, No. 2, 311-316

[12] Brown, C. W., Obremski, R. J. and Anderson, P. (1986). Infrared quantitative analysis in the Fourier domain: Processing vector representations, *Applied Spectroscopy*, Vol. 40, No. 6, 734-742

[13] Lindberg, W, Persson, J. and Wold, S. (1983). Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate, *Analytical Chemistry*, Vol. 55, 643-648

[14] Geladi, P. and Kowalski, B. R. (1986). Paritial least-squares regression: A tutorial, *Analytica Chimica Acta*, Vol. 185, 1-7

[15] Haaland, D. M. and Thomas, E. V. (1988). Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, *Analytical Chemistry*, Vol. 60, 1193-1202

[16] Haaland, D. M. and Thomas, E. V. (1988). Partial least-squares methods for spectral analyses. 2. Application to simulated and glass spectral data, *Analytical Chemistry*, Vol. 60, 1202-1208

[17] Heise, H. M., Marbach, R., Janatsch, G. and Kruse-Jarres, J. D. (1989). Multivariate determination of glucose in whole blood by attenuated total reflection infrared spectroscopy, *Analytical Chemistry*, Vol. 61, 2009-2015

[18] Bhandare, P., Mendelson, Y., Peura, R. A., Janatsch, G., Kruse-Jarres, J. D., Marbach, R. and Heise, H. M. (1993). Multivariate determination of glucose in whole blood using partial least-squares and artificial neural networks based on mid-infrared spectroscopy, *Applied Spectroscopy*, Vol. 47, No. 8, 1214-1221

[19] Wold, S., Ruhe, A., Wold, H. and Dunn III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM Journal of Science and Statistical Computations*, Vol. 5, 735-743

[20] Otto, M. and Wegscheider, W. (1985). Spectrophotometric multicomponent analysis applied to trace element determinations, *Analytical Chemistry*, Vol. 57, 63-69

[21] Wold, H. (1966). *Multivariate Analysis*, Academic, New York

Partial Least Squares Regression theory and Application in Spectroscopic Diagnosis of Total Hemoglobin in Whole blood[5)]

Seonwoo Kim[6)], Yoen-Joo Kim[7)], Jong-Won Kim[8)], Gilwon Yoon[7)]

Abstract

PLSR is a powerful multivariate statistical tool that has been successfully applied to the quantitative analyses of data in spectroscopy, chemistry, and industrial process control. Data in spectroscopy is represented by spectrum matrix measured in many wavelengths. Problems of many kinds of noise in data and intercorrelations between wavelengths are quite common in such data. PLSR utilizes whole data set measured in many wavelengths to the analysis, and handles such problems through data compression method. We investigated the PLSR theory, and applied this method to the real data for spectroscopic diagnosis of Total Hemoglobin in whole blood.

5) This study was supported by a grant (#HMP-95-G-1-4) of the '96 Good Health R&D Project, Ministry of Health & Welfare, R. O. K.

6) Clinical research center, Samsung Biomedical Research Institute, Seoul, 135-230, Korea

7) Biomedical engineering center, Samsung Biomedical Research Institute, Seoul, 135-230, Korea

8) Department of clinical pathology, Samsung Medical Center, Seoul, 135-230, Korea