약제학회지(2007), 제 37 권 제 6 호
J. Kor. Pharm. Sci., Vol. 37, No. 6, 377-395 (2007)

근적외 스펙트럼을 이용한 정량분석용 최적 주성분회귀모델을 얻기 위한 알고리듬

조정환†

숙명여자대학교 약학대학

(2007년 12월 5일 접수 · 2007년 12월 14일 승인)

Algorithm for Finding the Best Principal Component Regression Models for Quantitative Analysis using NIR Spectra

JungHwan Cho†

College of Pharmacy, Sookmyung Women's University, Seoul, 140-742, Korea

(Received December 5, 2007 · Accepted December 14, 2007)

ABSTRACT – Near infrared(NIR) spectral data have been used for the noninvasive analysis of various biological samples. Nonetheless, absorption bands of NIR region are overlapped extensively. It is very difficult to select the proper wavelengths of spectral data, which give the best PCR(principal component regression) models for the analysis of constituents of biological samples. The NIR data were used after polynomial smoothing and differentiation of 1st order, using Savitzky-Golay filters. To find the best PCR models, all-possible combinations of available principal components from the given NIR spectral data were derived by in-house programs written in MATLAB codes. All of the extensively generated PCR models were compared in terms of SEC(standard error of calibration), R^2, SEP(standard error of prediction) and SECP(standard error of calibration and prediction) to find the best combination of principal components of the initial PCR models. The initial PCR models were found by SEC or Malinowski's indicator function and *a priori* selection of spectral points were examined in terms of correlation coefficients between NIR data at each wavelength and corresponding concentrations. For the test of the developed program, aqueous solutions of BSA(bovine serum albumin) and glucose were prepared and analyzed. As a result, the best PCR models were found using *a priori* selection of spectral points and the final model selection by SEP or SECP.

Key words – All-possible combinations, Principal component regression (PCR), Near infrared (NIR), Correlation coefficients

생체에서 얻어지는 시료를 직접적으로 분석하는 것은 환자의 질환 상태 및 예후를 판단하고 치료의 방향과 방법을 선택하는데 있어서 매우 중요한 일이다. 경우에 따라서는 질환의 종류나 진단과정에서 주어지는 환자의 불편함 또는 위험성 등을 최소화하기 위해서 시료 채취량을 최소화하는 것이 필요하기도 하고 비침습적(또는 비파괴적) 분석이 가능하다면 더욱 바람직하다고 할 수 있다. 이러한 목적에 사용될 수 있는 측정방법으로는 각종 전기화학적 방법과 생체막을 통해서 쉽게 투과될 수 있는 영역의 전자기파를 이용하는 분광분석적 방법[1]을 생각할 수 있다. 그 중 환자에 대한 생체분석에서 비침습적인 방법의 적용이 가장 용이할 것으로 생각되는 것은 근적외선 영역을 이용한 분광학적인 분석이라고 할 수 있다. 일반적으로 1100~2500 nm의 근적외선영역은 분자 결합의 다양한 진동 모드들과 그들의 배음, 결합음 등이 복잡하게 중첩되어 나타나서 개별적인 피크들에 대한 정성적인 확인이 거의 불가능하여 과거에는 쓸모없는 영역으로 인식되었으나 계량분석화학에 의한 다변량 자료처리법의 실제적 응용의 가능성이 확인되면서 다양한 적용 예들을 확인할 수 있게 되었다. 그러나 여전히 생체시료의 비침습적 분석이라는 측면에서는 많은 문제점이 나타나고 있는데 이는 상대적으로 미량이면서 구조적 특성에서 뚜렷한 흡수대를 가지기 어려운 물질이 대상인 경우에 더욱 심각하다. 특히 중첩되어 나타나는 여러 흡수대에 대한 회귀모델을 구성하는 경우에 회귀모델에 포함될 최적의 변수를 선정하는 것은 정확한 농도의 추정을 위해서 매우 중요하다고 할 수 있다. 이를 위해서 다중선형회귀 모델을 구성할 때 모델에 포함될 측정값들의 파장들을 선정하기 위한 다양한 변수선택의 방법이 있을 수 있지만 이 경우에는 채택되는 변수의 개수가 표준군의 시료의 개수에 의해 제한된다. 따라서 측정한 전체 파장 영역의 스펙트럼 정보 전체를 사용하지 못하고 일부 매우 제한된 개수의 파장들에서의 측정값 자료만을 대상

†본 논문에 관한 문의는 이 저자에게로
Tel : 02)710-9580, E-mail : jcho@sookmyung.ac.kr

으로 계산하게 되어 측정된 스펙트럼 정보를 제한적으로만 사용하는 것이 된다. 이에 대한 최선의 대안이 될 수 있는 방법은 인자분석(Factor Analysis)에 의해 자료에 포함된 정보컨텐츠를 최대한 유지하는 상호 독립적인 고유벡터들을 얻어, 이 적은 수의 고유벡터를 기준으로 재구성된 자료를 대상으로 회귀 모델을 구하는 것이다. 이런 종류의 방법으로 주성분회귀(Principal Component Regression) 모델을 사용하는 방법이 널리 사용되고 있다. 이 경우 회귀모델의 구성을 위하여 주어진 표준군의 시료들에 대해 얻어진 스펙트럼인 다변량 기기 측정 자료와 각 시료 중의 분석대상물질의 농도 사이의 회귀 모델을 구성하는 과정에서 다변량 기기 측정 자료에 대해 주성분분석(Principal Component Analysis, PCA)을 실시한다. 주성분분석의 결과로는 고유벡터들의 행렬인 로딩행렬(loading matrix)와 스코어들의 행렬인 스코어행렬(score matrix)가 얻어지는데 이 스코어행렬과 분석대상물질의 농도 사이에 다중선형회귀식을 얻는 방법을 주성분회귀법이라고 한다.[2] 주성분회귀분석은 단순한 다중선형회귀(Multiple Linear Regression, MLR) 모델에 비해서 안정적이고 더 나은 특성의 회귀식을 제시한다. 다중선형회귀의 경우에는 사용할 변수들의 개수가 표준군의 시료의 개수보다 작게 되도록 미리 변수의 개수를 극도로 제한할 수밖에 없으므로 넓은 파장 영역에서 측정이 이루어졌더라도 몇 개 파장에서의 측정값만을 사용하고 나머지 대부분을 모두 버리게 되지만, 주성분회귀법의 경우에는 주어진 모든 파장범위의 측정값 전체를 대상으로 하여 새로운 변수인 고유벡터들을 얻는 것이므로 더 나은 모델을 얻게 하는 특성이 있다. 한편 주성분회귀법의 경우에도 모델의 구성에 포함될 고유벡터들의 수를 정하고 그에 따라 스코어행렬이 정해지는데, 일반적으로 고유벡터들의 중요도를 표시하는 고유값들을 기준으로 고유값이 큰 것부터 순서대로 일정 개수의 고유벡터들을 순서대로 모두 회귀모델에 포함시킨다.[2-4] 그러나 이들 고유벡터들을 단순히 순서대로 선택하지 않고 적절한 방법으로 선별을 시행하는 것이 주성분회귀 모델을 향상시킨다는 보고들이 있다. 모델에 포함될 고유벡터들을 선정하기 위한 기존의 방법들은 대체적으로 제한된 연산능력의 범위 안에서 축차선택, 유전알고리듬 등의 활용을 통해서 제한된 횟수의 반복 연산하여 모델에 포함될 고유벡터들을 선정하고 있다.[4] 그러나 이런 기존의 방법들의 경우에는 제한된 경우의 수만큼의 모델들만을 검토하는 것이므로 최종적으로 얻어지는 주성분회귀모델이 진정한 최적의 모델인지에 대한 확신을 가지기 어렵다. 따라서 본 연구에서는 최적의 주성분회귀 모델을 찾기 위해서 1차적으로 고유값을 기준으로 그 크기 순서대로 연속적으로 고유벡터들을 선택한 다음, 이들 고유벡터들을 가지고 구성될 수 있는 고유벡터들의 모든 조합으로 주성분회귀모델들을 생성하고 이들 모델들을 적절한 기준에 따라 상호 비교하여 그 중 가장 최적의 것을 선택하는 알고리듬을 개발하였다. 이러한 알고리듬의 실제적인 적용을 가능하게 하기 위한 컴퓨터 프로그램을 개발하고 BSA 및 포도당을 함유한 모의 시료들의 근적외분광스펙트럼 자료를 대상으로 하여 이 프로그램의 적용 가능성을 검토하였다.

실험 방법

자료 처리

모든 계산은 MATLAB(Release 2007a, MathWorks, Inc., 미국) 언어의 문법에 따라 자체 제작한 프로그램으로 처리하였다.

실험 기기

근적외선 스펙트럼을 얻기 위해 사용된 기기는 NIRSystems 6500(FOSS, 미국)이며, 층장 1 mm 액체 큐벳을 사용하였다. 스펙트럼 자료의 획득은 기기와 함께 제공된 WinISI를 이용하였다. 이렇게 측정된 스펙트럼 파일은 그 파일구조의 분석을 통해서 자체 개발한 MATLAB 프로그램인 read_cal.m을 이용하여 MATLAB의 자료 파일인 MAT(행렬형식의 자료 파일)로 변환하여 사용하였다.

실험 재료

개발되는 알고리듬의 성능을 시험하기 위한 모의 생체 시료로서 혈액 중 많은 부분을 차지하고 있는 단백질 중 알부민과 혈당을 이루는 포도당을 사용하여 혼액을 만들었다. 알부민으로는 BSA(Bovine Serum Albumin, Sigma)을 사용하였다. 사용된 농도는 약 1.80~6.00 g/dL의 범위를 갖도록 하였고, 또한 저혈당에서 고혈당의 범위에 해당되는 농도를 구성하기 위하여 포도당(D-(+)-glucose, Sigma)을 써서 2.00~23.00 mmol/L의 범위를 갖도록 하였다. 사용된 시료 중의 BSA 및 포도당의 농도는 Table I에 표시된 바와 같다. 알고리듬의 적용성을 확인하기 위하여 조제된 시료들을 두 개의 군으로 나누어 하나는 회귀모델을 구성하기 위한 표준군(calibration group)으로 다른 하나는 모델의 적합성을 판단하기 위한 검정군(validation group)으로 하였다.

모델 구성에 사용될 초기 변수들의 선정

근적외선 스펙트럼은 1100~2498 nm까지 2 nm 간격으로 측정되었는데 이 중 뚜렷한 흡수대가 없거나 장파장 영역으로 흡광이 너무 커서 검출기 노이즈가 문제되는 영역 그리

Table I–*Concentration of BSA and Glucose Used for Modeling*

Calibration set			Validation set		
Sample No.	BSA (g/dL)	Glucose (mmol/L)	Sample No.	BSA (g/dL)	Glucose (mmol/L)
1	3.589	2.822	1	3.589	5.644
2	3.589	8.467	2	3.589	11.289
3	3.589	14.111	3	3.589	16.933
4	3.077	19.756	4	3.077	22.578
5	3.398	2.772	5	2.471	5.544
6	4.016	8.317	6	3.398	11.089
7	3.707	13.861	7	1.853	16.633
8	2.471	19.406	8	3.089	22.178
9	3.574	2.817	9	3.574	5.633
10	3.574	8.456	10	3.574	11.272
11	3.574	14.089	11	3.574	16.906
12	3.574	19.722	12	3.574	22.544
13	1.989	5.583	13	2.984	5.583
14	3.979	5.583	14	4.974	5.583
15	5.968	5.583	15	1.989	11.167
16	2.984	11.167	16	3.979	11.167
17	4.974	11.167	17	5.968	11.167
18	3.979	5.583	18	3.979	11.194
19	3.979	16.744	19	3.979	22.328

고 물에 의한 강한 흡광이 나타나는 영역 등을 제외하여 모델을 구성하기 위한 초기 변수로 1300~1900 nm와 2150~2350 nm의 범위의 흡광도값들을 선정하여 계산에 사용하였다. 이 경우 402개 파장에서의 측정값이 사용되었다. 또한 이 주어진 영역 중에서 각 파장에서의 측정값과 각 분석성분의 농도사이의 상관계수를 모두 구하고 그 상관계수를 기준으로 하여 최소한의 통계적 상관성이 있는 것으로 확인되는 파장만을 선택하여 그 이후의 계산에 사용한 경우와 모델링 결과를 비교하였다. 여기에서 사용된 상관계수는 피어슨(Pearson) 상관계수(r)로서 다음의 식으로 계산되어 사용되었다.[6]

$$r = \frac{\sum(x_i-x)(y_i-y)}{\sqrt{(x_i-\bar{x})^2}\sqrt{(y_i-\bar{y})^2}}$$

자료 전처리

스펙트럼 자료는 측정된 흡광도값 그대로에 대해 Savitzky와 Golay의 노이즈 제거 필터를 적용하여 노이즈를 감쇄시킨 스펙트럼 자료 및 같은 방법에 따른 미분 필터를 적용하여 노이즈 감쇄 및 1차 미분한 스펙트럼 자료를 계산에 사용하였다.[7]

주성분회귀모델을 위한 주성분분석에 사용된 자료는 2차원 행렬의 형태인데 이 행렬의 각 행은 표준군 또는 검정군의 각 시료들을 표시하고, 각 열은 계산에 사용된 스펙트럼상의 파장을 의미한다. 따라서 표준군 및 검정군은 각각 19개씩의 시료에 대해 402개 파장에서 측정된 흡광도 자료로 행렬이 구성되었으므로 19×402 크기의 행렬이 되었다. 이 행렬에 대해서 주성분분석에 앞서서 평균이동(mean centering) 또는 분산보정(variance scaling)을 실시하였다. 이들 자료 전처리법을 적용한 경우와 적용하지 않은 경우의 조합에 의해 4 가지의 자료 전처리 방법이 적용되었다. 평균이동 및 분산보정은 다음의 식으로 표시될 수 있다.[8]

$$\text{Mean centering: } d_{ik,mc} = d_{ik}-\bar{d}_k \qquad \bar{d}_k = \frac{\sum_{i=1}^{r} d_{ik}}{r}$$

$$\text{Variance scaling: } d_{ik,vs} = \frac{d_{ik}}{S_k} \qquad S_k = \left[\frac{1}{r-1}\sum(d_{ik}-\bar{d}_k)\right]^{1/2}$$

평균이동과 분산보정을 모두 실시하는 경우를 자동보정(autoscaling)이라 하고 이 경우에는 다음과 같이 식이 적용되었다.

$$\text{autoscaling: } d_{ik,as} = \frac{d_{ik}-\bar{d}_k}{S_k}$$

위 식들에서 i는 주어진 행렬에서 행의 번호이고, k는 열의 번호가 된다.

주성분회귀법(Principal Component Regression)

회귀분석에 주어진 독립변수자료행렬(**D**)와 농도벡터(**y**) 사이의 회귀모델을 얻을 때, 다중선형회귀법의 경우에는 **Db**=**y**의 모델이 구성되어 **b**를 최소자승법에 따라 다음의 식으로 얻게 된다.[5]

$$\mathbf{b}=(\mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}^T\mathbf{y}$$

그런데 주성분회귀법에서는 우선 **D**에 대한 주성분분석(Principal Component Analysis)를 하여 고유벡터들(eigenvectors)의 행렬(**C**)과 그에 대한 스코어(score) 행렬(**R**)를 얻어 **D**=**RC**의 관계로 **D**를 분해하는 계산을 한다. 이 때, 행렬 **C**의 각 고유벡터들은 연산의 정의상 단위길이를 가지고 상호간 직교하는 성질을 가진다. 또한 계산과정에서 각 고유벡터들에 대한 고유치들(eigenvalues)이 얻어지는데 이 고유치들의 크기를 기준으로 고유벡터의 순서가 결정된다. 이제 회귀모델은 **D**와 **y**사이에 구성되는 것이 아니라, **R**과 **y**사이에 구성되게 된다. 즉, **Rb**=**y**의 행렬식으로 표시되게 되고 이에 대한 **b**를 얻음으로써 회귀모델을 얻게 된다.[2,9]

$$\mathbf{b}=(\mathbf{R}\mathbf{R}^T)^{-1}\mathbf{R}^T\mathbf{y}$$

주성분분석은 NIPALS (Non-linear Iterative Partial Least Squares)법[7] 또는 SVD (Singular Value Decomposition)법[7]에 따라 시행했으며 이 연산의 결과는 주어진 자료행렬 **D**

에 대해 다음의 식으로 얻어진다. 계산되는 주성분의 개수를 미리 지정하는 경우에는 NIPALS법을 사용했으며, 모든 주성분을 얻고자 할 때는 SVD법을 사용하였다. NIPALS법은 MATLAB 언어체제에 따라 프로그램을 작성하였고, SVD법의 경우에는 MATLAB의 내장함수인 SVD 함수를 그대로 사용하였다. NIPALS법의 경우에는 동일한 주성분을 얻는 계산의 속도가 SVD에 비해 2배 정도 빠르기 때문에 사용하였다.

$$\mathbf{D}=\mathbf{USV}^{T}$$

이를 위에서 언급한 **R**과 **C** 및 고유치행렬(Λ, 고유치가 대각성분으로 나열된 대각행렬)에 대해 정리하면 다음의 식들에 표시된 관계와 같다.

$$\mathbf{R}=\mathbf{US} \qquad \mathbf{C}=\mathbf{V}^{T} \qquad \Lambda=\mathbf{S}^{2}$$

식에서 **S** 행렬의 대각선 성분은 크기 순서대로 나열된 고유치의 제곱근들이다. 따라서 Λ의 대각선 성분은 순서대로 나열된 고유치가 된다. **U**는 이들 관계를 이용하여 회귀계수 **b**를 얻는 식은 다음과 같이 표시될 수 있다.

$$\mathbf{b}=(\mathbf{RR}^{T})^{-1}\mathbf{R}^{T}\mathbf{y}=(\mathbf{USS}^{T}\mathbf{U}^{T})^{-1}\mathbf{S}^{T}\mathbf{U}^{T}\mathbf{y}=(\mathbf{U}\Lambda\mathbf{U}^{T})^{-1}\mathbf{S}^{T}\mathbf{U}^{T}\mathbf{y}$$

이와 같이 인자분석(Factor Analysis)에 근거한 회귀분석에서 인자(Factor)의 개수를 정할 때는 고유치의 크기가 큰 것부터 차례대로 고유벡터들을 나열하였을 때 처음부터 일정 개수까지의 고유벡터에 대한 스코어값들만을 선택하여 **R** 행렬의 크기를 한정하는 방법을 사용한다. 이 판정을 위해 몇 가지의 판단기준을 사용하는 것이 일반적이지만, 대체적으로 고유치가 가장 큰 것에서부터 통계적 기준에 따라 순서상 마지막으로 의미가 있는 고유치를 가지는 고유벡터의 스코어까지를 모두 사용한다. 즉, 선택된 고유벡터의 개수가 k ($\leq c$, c는 고유벡터의 최대개수로서 행렬 **D**의 열의 개수 및 행의 개수 중, 작은 쪽과 같음.) 라면, **R**의 크기는 n x k, **U**는 n x k, S는 k x k이고 S의 대각선 성분은 순서대로 $\sqrt{\lambda_1}$, $\sqrt{\lambda_2}$, ..., $\sqrt{\lambda_k}$이 된다. 일반적으로 n은 행렬 **D**의 행의 개수로 행렬 **D**의 구성에 사용된 시료의 개수와 같으며, c는 행렬 **D**의 열의 개수로 행렬 **D**의 구성을 위해서 측정된 파장의 개수에 해당된다.

그런데 이때 선택된 고유벡터에 대한 스코어 행렬의 모든 열들이 주어진 시료 중의 성분들의 농도를 정확하게 반영하는 것이 아닐 가능성이 있고, 이런 상황은 수용액인 경우에 물이 매질의 주요 물질이고 물의 흡광은 크게 나타나는데 비해서 근적외선 영역에서 특별히 강한 흡수패턴을 보이지 않는 물질들, 즉 예를 들어 단백질이나 포도당 등의 농도를 추정하는 경우에는 더욱 문제가 된다. 따라서 수용액 중의 이들 물질의 정확한 농도 예측을 가능하게 하는 회귀모델이 얻어지지 않는다. 따라서 1차적으로 얻어진 고유벡터들을 대상으로 좀 더 정확한 농도추정에 중요한 고유벡터들의 조합을 찾아내어 그 조합에 대한 스코어 행렬만을 기준으로 회귀모델을 만드는 것이 하나의 해결방법이 될 수 있다. 따라서 이 연구에서는 임의로 조제한 단백질 및 포도당의 혼합수용액에서 각 성분들의 농도를 가장 잘 예측할 수 있는 고유벡터들에 따른 스코어 행렬의 최적의 조합을 얻는 방법을 찾고자 하였다.

주성분회귀 모델의 주성분 변수의 모든 조합의 생성

주성분회귀법에 따른 농도예측모델에 포함될 가장 최적의 고유벡터들의 조합에 따른 스코어 행렬을 얻기 위한 검색을 위한 초기 모델을 얻기 위하여 1차적으로 SEC(Standard Error of Calibration) 또는 IND(Malinowski's Indicator Function)를 기준으로 고유값이 큰 것에서 작은 것의 순서로 그 사이의 모든 고유벡터들로 구성된 모델을 구성하였다. 이 초기 모델에 포함된 고유벡터들에 대해 그들의 모든 조합(All-possible combination)을 만들어서 모델들에 대한 적합성 검정(Goodness-of-fit test)으로 비교하는 방법을 채택하였다.

SEC를 기준으로 하는 경우에는 SEC가 가장 작게 되는 고유벡터의 개수 k를 선택하였다.[7]

$$SEC=\left[\frac{1}{n-k-1}\sum_{i=1}^{n}(\hat{y}_{i,ref}-y_{i,ref})^{2}\right]^{1/2}$$

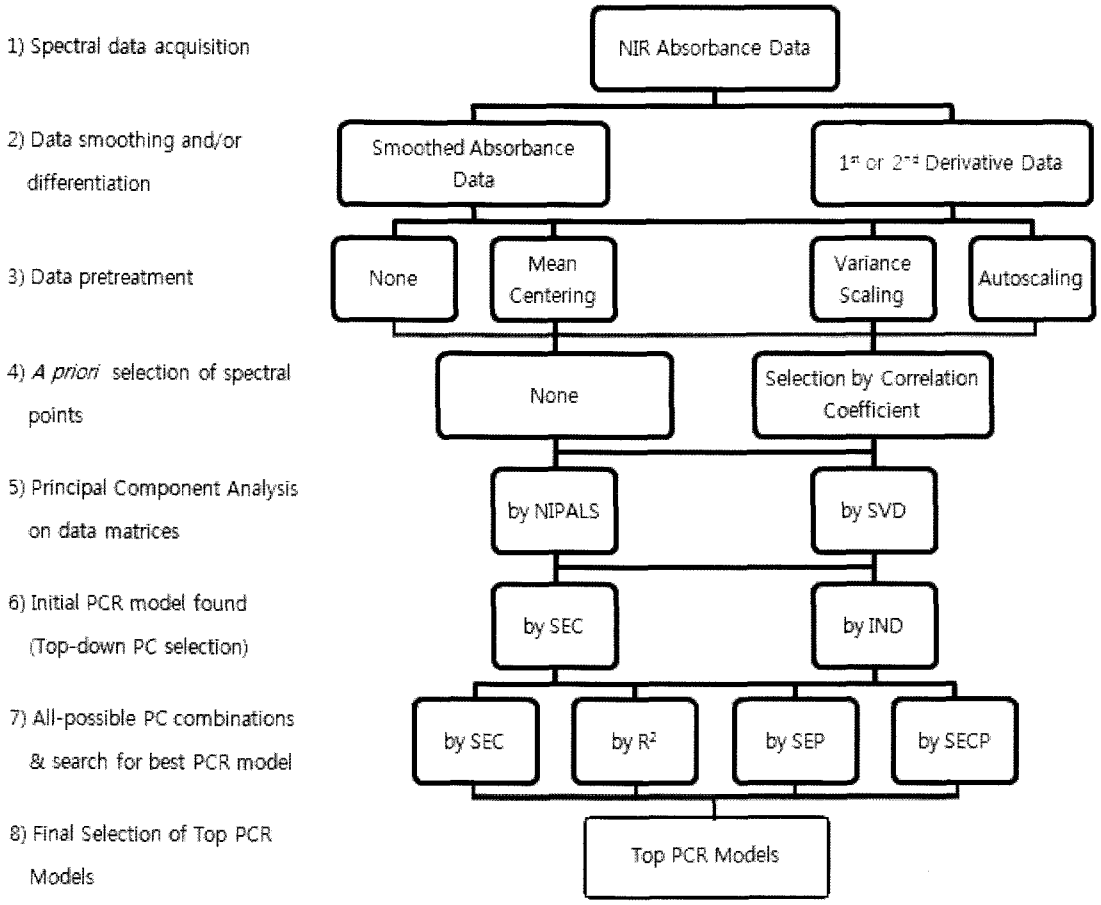
위의 식에서 n은 표준군에 사용한 자료의 수, k는 선택된 고유벡터의 수이며, $y_{i,ref}$는 표준군에 포함된 시료 중의 각 성분의 아는 농도이고, $\hat{y}_{i,ref}$는 얻어진 p개의 고유벡터를 이용한 주성분회귀 모델에 의해서 추정된 각 성분의 농도이다.

IND는 다음에 주어진 식으로 계산하였다.[7]

$$RE=\sqrt{\frac{\sum_{i=k+1}^{j=c}\lambda_i}{n(c-k)}} \qquad IND=\frac{RE}{(c-k)^{2}}$$

식에서 λ_i는 i 번째 고유벡터의 고유값이고 n은 모델에 포함된 시료의 개수, c는 주어진 변수의 총개수, k는 선택된 고유벡터의 개수를 의미한다. k가 최적인 경우에 IND는 최소값이 되므로 IND가 최소값이 되는 경우의 k를 기준으로 첫 번째 고유벡터에서부터 고유값 크기의 차례대로 k 번째 고유벡터까지를 초기 모델에 포함되는 변수로 하였다.

SEC 또는 IND를 기준으로 선택되는 변수로서의 고유벡터들은 그 고유값들을 기준으로 첫 번째 것부터 최종 선택된 것까지 빠짐없이 모두 선택되는 것이며 이는 단순히 주



Scheme I–Procedure of Data Processing. For each row of the scheme (in the cases of row number 2 to 7) with two or more options, one of them was selected for each run of best PCR model search and all of the combinations of those options have been studied. The meanings of acronyms are as following: NIR=Near Infrared, Autoscaling=Meaning centering+variance scaling, NIPALS=Non-linear Iterative Partial Least Squares, SVD=Singular Value Decomposition, IND=Malinowski's Indicator Function, SEC=Standard Error of Calibration, R^2=multiple coefficient of determination, SEP=Standard Error of Prediction, SECP=Standard Error of Calibration and Prediction, PCR=Principal Component Regression.

어진 독립변수들 중에서 주어진 자료가 가지는 총변이를 기준으로 하여 가장 많은 변이를 설명하는 인자들이 선택되는 것이므로 검정군에 대한 예측에 있어서 최적의 조합이라 할 수는 없다. 따라서 이 최초 선택된 변수로서의 고유벡터들 중에서 검정군에 대한 실제 예측에 있어서 최적의 결과를 가져오는 고유벡터들만을 선택하고자 했다.

모든 주어진 변수들의 모든 조합을 생성하는 이 방법은 주어진 변수의 개수가 너무 많지 않고 각 모델의 유용성을 판단할 수 있는 적절한 진단값을 일일이 계산할 수 있는 경우에 가능한 검토 방법이다. 조합 가능한 독립변수의 개수를 k라 하고, 선택할 변수의 개수를 p라고 할 때 가능한 조합의 수는 ${}_kC_p$ $(=k!/(k-p)!/p!)$로 표현할 수 있다. 예를 들어 ${}_{16}C_8$의 경우에는 12,870개의 조합이 생성된다. 또한 이러한 변수의 조합을 p가 1에서부터 k에 이르기까지 모두 생성하여 비교하여야 한다. 따라서 이 방법을 실험에 사용된 402개 파장에서의 측정값들 모두를 대상 변수로 하여 적용하는 것은 너무나 많은 후보 모델들에 대한 연산 결과의 검토가 필요하므로 연산의 시간이 문제가 된다.[6] 그러나 변수의 조합을 생성하기 전에 변수들을 원래의 스펙트럼에서의 측정된 파장들에 의한 변수 402개 모두를 그대로 사용하지 않고 주성분분석에 의해서 원래의 스펙트럼이 가진 정보의 컨텐츠는 거의 그대로 유지하면서도 일단 최소한의 고유벡터들로 변환하여 그에 따른 스코어 행렬(**R**)로 바꾸면, 연산의 대상이 되는 변수의 개수 즉, k가 현저하게 줄어들게 된다. 본 시험에서 사용한 19×402 크기의 행렬이 주성분분석에 따른 스코어행렬의 구성으로 19×16의 크기와 같이 열의 개수가 16 또는 그 이하로 현저하게 작아지게 되므로 이들 변수들의 조합의 개수 또한 현저하게 적어지게 된다. 또한 최근의 강력한 연산기능의 컴퓨터에서 실질적으로 의미 있는 수준에서 계산을 완료할 수 있게 되었고, MATLAB 언어를 활

용하여 적절한 프로그램을 개발함으로써 그 연산을 실질적으로 가능하게 하였다.

대상이 되는 k개의 사용가능한 변수들로부터 p개의 변수를 포함하는 모든 가능한 조합인 APC (All-possible combination, combo)를 얻기 위한 알고리듬을 pseudocode로 표시하면 다음과 같으며 이를 MATLAB 프로그램으로 자체 개발하여 사용하였다. 이 프로그램을 이용하여 p가 1에서부터 k일 때까지의 모든 가능한 변수들의 조합을 구성하였다.

```
calculate kCp=k!/(k-p)!/p!
prepare combo_set {matrix of k by kCp}
prepare vidx vector of 1 by p
set cidx=0
set idx=0
```

Table II–*PCR Models Found Best using Absorbance Spectra of 402 Given Spectral Points. Initial Sets of Factors (Eigenvectors) for Initial PCR Models Were Found by Means of SEC or Malinowski's IND Function and then Final Sets of Factors Were Found by Means of SEC, R^2, SEP or SECP*

Mean Centering	Variance Scaling	Initial Selection	Compound	Initial Factors	Final Selection	Final Factors	R^2	SEC	SEP	SECP
No	No	by SEC	BSA	13	by SEC	12	1.0000	0.0143	0.0227	0.0208
					by R^2	13	1.0000	0.0144	0.0244	0.0224
					by SEP	9	1.0000	0.0215	0.0178	0.0192
					by SECP	11	1.0000	0.0209	0.0179	0.0188
			Glucose	16	by SEC	15	0.9999	0.2542	0.8995	0.8244
					by R^2	16	0.9999	0.2914	0.8908	0.8348
					by SEP	11	0.9986	0.6842	0.7138	0.7052
					by SECP	13	0.9993	0.5559	0.7290	0.6914
		by IND	BSA	8	by SEC	8	1.0000	0.0217	0.0212	0.0214
					by R^2	8	1.0000	0.0217	0.0212	0.0214
					by SEP	8	1.0000	0.0217	0.0212	0.0214
					by SECP	8	1.0000	0.0217	0.0212	0.0214
			Glucose	8	by SEC	7	0.9980	0.6556	0.9483	0.8471
					by R^2	8	0.9981	0.6782	0.9191	0.8389
					by SEP	8	0.9981	0.6782	0.9191	0.8389
					by SECP	8	0.9981	0.6782	0.9191	0.8389
	Yes	by SEC	BSA	14	by SEC	12	1.0000	0.0121	0.0224	0.0202
					by R^2	14	1.0000	0.0142	0.0222	0.0208
					by SEP	10	1.0000	0.0191	0.0139	0.0157
					by SECP	9	1.0000	0.0171	0.0144	0.0154
			Glucose	16	by SEC	11	0.9999	0.2161	0.9362	0.7941
					by R^2	16	0.9999	0.3164	0.8591	0.8069
					by SEP	14	0.9998	0.3423	0.7927	0.7224
					by SECP	11	0.9998	0.2773	0.8180	0.7026
		by IND	BSA	7	by SEC	6	1.0000	0.0193	0.0179	0.0184
					by R^2	7	1.0000	0.0200	0.0177	0.0186
					by SEP	7	1.0000	0.0200	0.0177	0.0186
					by SECP	6	1.0000	0.0193	0.0179	0.0184
			Glucose	7	by SEC	7	0.9995	0.3369	0.9133	0.7451
					by R^2	7	0.9995	0.3369	0.9133	0.7451
					by SEP	7	0.9995	0.3369	0.9133	0.7451
					by SECP	7	0.9995	0.3369	0.9133	0.7451

Table II–*Continued*

Yes	No	by SEC	BSA	13	by SEC	12	1.0000	0.0149	0.0209	0.0196
					by R^2	13	1.0000	0.0149	0.0222	0.0209
					by SEP	9	1.0000	0.0192	0.0150	0.0165
					by SECP	9	1.0000	0.0192	0.0150	0.0165
			Glucose	16	by SEC	15	0.9999	0.2925	1.0754	1.0052
					by R^2	16	0.9999	0.3518	1.0650	1.0188
					by SEP	12	0.9935	1.6921	0.7269	1.0435
					by SECP	12	0.9988	0.7263	0.7567	0.7495
		by IND	BSA	7	by SEC	7	1.0000	0.0227	0.0215	0.0219
					by R^2	7	1.0000	0.0227	0.0215	0.0219
					by SEP	7	1.0000	0.0227	0.0215	0.0219
					by SECP	7	1.0000	0.0227	0.0215	0.0219
			Glucose	7	by SEC	7	0.9979	0.7168	0.9813	0.8935
					by R^2	7	0.9979	0.7168	0.9813	0.8935
					by SEP	7	0.9979	0.7168	0.9813	0.8935
					by SECP	7	0.9979	0.7168	0.9813	0.8935
	Yes	by SEC	BSA	12	by SEC	11	1.0000	0.0134	0.0199	0.0184
					by R^2	12	1.0000	0.0144	0.0196	0.0185
					by SEP	9	1.0000	0.0180	0.0148	0.0159
					by SECP	9	1.0000	0.0163	0.0153	0.0156
			Glucose	16	by SEC	12	0.9999	0.2482	0.9330	0.8224
					by R^2	16	0.9999	0.3266	0.8595	0.8237
					by SEP	11	0.9994	0.4632	0.8110	0.7338
					by SECP	11	0.9998	0.2776	0.8239	0.7189
		by IND	BSA	7	by SEC	7	1.0000	0.0196	0.0176	0.0184
					by R^2	7	1.0000	0.0196	0.0176	0.0184
					by SEP	7	1.0000	0.0196	0.0176	0.0184
					by SECP	7	1.0000	0.0196	0.0176	0.0184
			Glucose	7	by SEC	7	0.9993	0.4189	0.9406	0.7904
					by R^2	7	0.9993	0.4189	0.9406	0.7904
					by SEP	7	0.9993	0.4189	0.9406	0.7904
					by SECP	7	0.9993	0.4189	0.9406	0.7904

```
set vidx(1)=-1

repeat while idx>-1
  set vidx(idx+1)=vidx(idx+1)+1
  repeat while vidx(idx+1)<k-idx+p-1
    if idx<p-1
      set idx=idx+1
      set vidx(idx+1)=vidx(idx)
    else
      set combo_set(i+1, cidx+1)
          =vidx(i+1)+1{for i=0 to p-1}
      set cidx=cidx+1
    end
    set vidx(idx+1)=vidx(idx+1)+1
  end
  set idx=idx-1
end
```

모델의 평가방법

최초의 주성분회귀식에 주어진 변수(주성분, 고유벡터) k개 중에서 p개를 선택하되 이 p를 1에서 k까지 변화시킬 때 얻어지는 모든 조합에 의한 주성분회귀식의 후보들 중에서 최상의 모델을 결정하기 위하여 그 각 모델에 대한 적합성 검정(Goodness-of-fit)을 실시하였다. 적합성 검정을 위하여 얻어진 모든 후보 주성분회귀식들의 SEC(Standard Error of Calibration)값, 결정계수(R^2), SEP(Standard Error of Prediction)값 및 SECP(Standard Error of Calibration and Prediction)값을 계산하여 이들의 값을 비교하였다.[4-5),10)] SEC, 결정계수, SEP, SECP는 아래의 식을 이용하여 계산하였다.

$$SEC = \left[\frac{1}{n-p-1} \sum_{i=1}^{n} (\hat{y}_{i,ref} - y_{i,ref})^2 \right]^{1/2}$$

Table III—*PCR Models Found Best using 1st Derivative Spectra of 402 Given Spectral Points. Initial Sets of Factors (Eigenvectors) for Initial PCR Models Were Found by Means of SEC or Malinowski's IND Function and then Final Sets of Factors Were Found by Means of SEC, R^2, SEP or SECP*

Mean Centering	Variance Scaling	Initial Selection	Compound	Initial Factors	Final Selection	Final Factors	R^2	SEC	SEP	SECP
No	No	by SEC	BSA	7	by SEC	7	1.0000	0.0183	0.0534	0.0433
					by R^2	7	1.0000	0.0183	0.0534	0.0433
					by SEP	6	1.0000	0.0305	0.0425	0.0381
					by SECP	6	1.0000	0.0305	0.0425	0.0381
			Glucose	16	by SEC	16	0.9995	0.6512	3.2593	3.0385
					by R^2	16	0.9995	0.6512	3.2594	3.0385
					by SEP	12	0.9854	2.3434	1.6275	1.8477
					by SECP	13	0.9940	1.6256	1.7893	1.7514
		by IND	BSA	8	by SEC	7	1.0000	0.0183	0.0534	0.0433
					by R^2	8	1.0000	0.0191	0.0534	0.0440
					by SEP	7	1.0000	0.0318	0.0425	0.0387
					by SECP	6	1.0000	0.0305	0.0425	0.0381
			Glucose	8	by SEC	8	0.9595	3.1140	6.7923	5.7249
					by R^2	8	0.9595	3.1140	6.7923	5.7249
					by SEP	3	0.7537	6.3699	6.0062	6.1751
					by SECP	6	0.9440	3.3684	6.2573	5.2780
	Yes	by SEC	BSA	12	by SEC	11	1.0000	0.0150	0.0295	0.0261
					by R^2	12	1.0000	0.0160	0.0295	0.0265
					by SEP	9	1.0000	0.0202	0.0264	0.0244
					by SECP	9	1.0000	0.0163	0.0265	0.0235
			Glucose	16	by SEC	16	0.9999	0.3275	1.4931	1.3928
					by R^2	16	0.9999	0.3275	1.4931	1.3928
					by SEP	15	0.9997	0.4490	1.4393	1.3215
					by SECP	13	0.9994	0.5133	1.4651	1.3018
		by IND	BSA	6	by SEC	6	1.0000	0.0230	0.0302	0.0275
					by R^2	6	1.0000	0.0230	0.0302	0.0275
					by SEP	6	1.0000	0.0230	0.0302	0.0275
					by SECP	6	1.0000	0.0230	0.0302	0.0275
			Glucose	6	by SEC	4	0.9601	2.6474	2.8513	2.7632
					by R^2	6	0.9607	2.8235	2.7007	2.7512
					by SEP	6	0.9607	2.8235	2.7007	2.7512
					by SECP	5	0.9605	2.7282	2.7019	2.7131

Table III–*Continued*

No	by SEC	BSA	13	by SEC	9	1.0000	0.0149	0.0249	0.0222
				by R^2	13	1.0000	0.0194	0.0239	0.0230
				by SEP	13	1.0000	0.0194	0.0239	0.0230
				by SECP	9	1.0000	0.0149	0.0249	0.0222
		Glucose	16	by SEC	16	1.0000	0.0782	3.3711	3.2066
				by R^2	16	1.0000	0.0782	3.3710	3.2066
				by SEP	10	0.9792	2.6165	1.7365	2.0373
				by SECP	10	0.9792	2.6165	1.7365	2.0373
	by IND	BSA	7	by SEC	7	1.0000	0.0256	0.0552	0.0466
				by R^2	7	1.0000	0.0256	0.0552	0.0466
				by SEP	6	0.9999	0.0417	0.0500	0.0469
				by SECP	7	1.0000	0.0256	0.0552	0.0466
		Glucose	7	by SEC	2	0.8741	4.5547	6.9061	5.9476
				by R^2	7	0.8879	5.1835	6.6400	6.1462
				by SEP	4	0.8634	5.0706	5.9408	5.5882
				by SECP	3	0.8625	4.9164	6.0425	5.5738
Yes	by SEC	BSA	16	by SEC	10	1.0000	0.0130	0.0318	0.0276
				by R^2	16	1.0000	0.0227	0.0310	0.0303
				by SEP	11	1.0000	0.0202	0.0219	0.0215
				by SECP	9	1.0000	0.0178	0.0221	0.0208
		Glucose	15	by SEC	15	0.9999	0.3144	1.4829	1.3830
				by R^2	15	0.9999	0.3144	1.4829	1.3830
				by SEP	14	0.9997	0.4664	1.4282	1.3126
				by SECP	13	0.9996	0.4477	1.4354	1.2934
Yes	by IND	BSA	6	by SEC	5	1.0000	0.0235	0.0308	0.0280
				by R^2	6	1.0000	0.0242	0.0308	0.0284
				by SEP	6	1.0000	0.0242	0.0308	0.0284
				by SECP	5	1.0000	0.0235	0.0308	0.0280
		Glucose	6	by SEC	4	0.9545	2.9285	2.7629	2.8343
				by R^2	6	0.9576	3.0535	2.6614	2.8197
				by SEP	6	0.9576	3.0535	2.6614	2.8197
				by SECP	5	0.9575	2.9355	2.6625	2.7766

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_{i,ref}-\bar{y}_{ref})^2}{\sum_{i=1}^{n}(y_{i,ref}-\bar{y}_{ref})^2}$$

$$SEP = \left[\frac{1}{\nu}\sum_{i=1}^{\nu}(\hat{y}_{i,ref}-y_{i,ref})^2\right]^{1/2}$$

$$SECP = \left[\frac{1}{n+\nu-p-1}\left(\sum_{i=1}^{n}(\hat{y}_{i,ref}-y_{i,ref})^2+\sum_{i=1}^{\nu}(\hat{y}_{i,val}-y_{i,val})^2\right)\right]^{1/2}$$

위의 식에서 n은 표준군에 사용한 자료의 수, k는 선택된 변수의 수, ν는 검정군에 사용한 자료의 수를 나타낸다. SEC, SEP 및 SECP의 경우는 모두 그 값이 최소인 경우가 주성분회귀식에 포함될 변수로서의 주어진 조건에서 가장 좋은 회귀모델을 보이는 것이라고 할 수 있다. SEC와 SEP가 모두 최소가 되지 못하는 경우에는 SEP값을 가장 작게 하는 경우나 SEC와 SEP를 결합한 것에 해당되는 SECP를 가장 작게 하는 경우를 주어진 조건에서의 최적 회귀모델로 간주하였다. 한편 결정계수 R^2의 경우는 1에 가까울수록 적

절한 모델로 간주하였다.

분석에 사용된 시료들의 근적외스펙트럼을 얻어 노이즈 제거 필터 및 미분 필터를 적용하고 자료에 대한 전처리로서 평균이동 또는 분산보정을 실시하고, 피어슨 상관계수를 이용한 파장에 대한 사전선택을 적용한 다음, 최초 주성분회귀모델을 얻고 그 고유벡터들의 모든 가능한 조합들을 얻고 이들의 성능을 상호비교하여 최적의 주성분회귀모델을 얻는 과정을 Scheme I에 정리하였다.

결과 및 고찰

모델링에 사용할 스펙트럼 영역의 선정, 노이즈 감쇄 및 1차 미분 스펙트럼

스펙트럼 자료는 2 nm 간격으로 측정되었는데 뚜렷한 흡수대가 없는 1100~1300 nm 영역, 장파장 영역으로 흡광이 너무 커서 검출기 노이즈가 문제되는 2350~2500 nm 영역, 그리고 수용액으로 측정되어 물에 의한 자체적인 강한 흡광이 나타나는 1900~2150 nm 영역 등을 제외하여 모델을 구성하기 위한 초기 변수로 1300~1900 nm와 2150~2350 nm의 범위에서의 흡광도값들을 선택하여 하나의 시료에 대해서 총 402개 파장에서의 측정값이 이후의 모든 계산에 사용되었다.

피어슨 상관계수를 이용한 모델링에 사용할 변수의 사전 선택

주어진 402개 파장에서의 흡광도 측정값들 또는 1차 미분값들 전체를 대상으로 주성분분석을 하여 얻어진 스코어 행렬을 기준으로 주성분회귀분석을 시행한 경우의 회귀 모델의 적합성 검정값들은 Table II 및 Table III에 표시하였다. 또한 주어진 402개 파장에서의 흡광도 측정값들 또는 1차 미분값들 전체를 사용하지 않고 그 개별 파장에서의 흡광도 측정값들 또는 1차 미분값들과 표준군 시료들의 농도 사이에 피어슨 상관계수를 구해서 모델링에 사용할 파장들에서 대한 사전 선택을 시행하였다. 최소한의 상관성이 있는 것으로 일반적으로 간주되는 기준인 상관계수의 절대값이 0.2 이상으로 나타난 파장들에서의 흡광도 측정값들 또는 1차 미분값들을 선택하였는데 이처럼 모델링 계산에 사용된 파장들의 개수는 Table IV에 표시한 바와 같이 흡광도 측정값을 기준으로 한 경우에는 BSA에 대해서는 308개, 포도당에 대해서는 281개이었다. 또한 Table V에 표시한 바와 같이 1차 미분값을 기준으로 한 경우에는 347개, 포도당에 대해서는 25개이었다.

변수 조합 알고리듬 적용 결과 예시

일차적으로 구성된 주성분회귀모델의 고유벡터들을 대상으로 그들로부터 구성 가능한 모든 조합의 변수 세트들을 얻기 위해서 위에 pseudocode의 형태로 표시된 알고리듬을 MATLAB 프로그램으로 개발하여 사용하였다. 이를 이용하여 생성한 변수 조합의 사례로 16개의 변수 중 2개, 7개 및 14개를 선택하는 경우는 다음과 같다. 숫자는 변수의 번호를 의미하며 수직 방향이 형성된 각각의 변수 조합 세트들을 의미한다. 이를 이용하여 주어진 변수에 대한 모든 조합을 빠짐없이 생성하였다.

16 개의 변수 중 2 개 선택의 경우, $_{16}C_2 = 120$ 세트의 변수 조합 생성

1	1	1	1	1	1	1	1	…	11	11	11	11	11	12	12	12	12	13	13	13	14	14	15
2	3	4	5	6	7	8	9	…	12	13	14	15	16	13	14	15	16	14	15	16	15	16	16

16 개의 변수 중 7 개 선택의 경우, $_{16}C_7 = 11440$ 세트의 변수 조합 생성

1	1	1	1	1	1	1	1	…	8	8	8	8	8	8	8	9	9	9	9	9	9	9	10
2	2	2	2	2	2	2	2	…	10	10	10	10	10	10	11	10	10	10	10	10	10	11	11
3	3	3	3	3	3	3	3	…	11	11	11	11	11	12	12	11	11	11	11	11	12	12	12
4	4	4	4	4	4	4	4	…	12	12	12	12	13	13	13	12	12	12	12	13	13	13	13
5	5	5	5	5	5	5	5	…	13	13	13	14	14	14	14	13	13	13	14	14	14	14	14
6	6	6	6	6	6	6	6	…	14	14	15	15	15	15	15	14	14	15	15	15	15	15	15
7	8	9	10	11	12	13	14	…	15	16	16	16	16	16	16	15	16	16	16	16	16	16	16

16 개의 변수 중 14 개 선택의 경우, $_{16}C_{14} = 120$ 세트의 변수 조합 생성

1	1	1	1	1	1	1	1	…	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3
2	2	2	2	2	2	2	2	…	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4
3	3	3	3	3	3	3	3	…	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5
4	4	4	4	4	4	4	4	…	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6
5	5	5	5	5	5	5	5	…	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7
6	6	6	6	6	6	6	6	…	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8
7	7	7	7	7	7	7	7	…	8	8	8	8	8	8	8	9	9	9	9	9	9	9	9
8	8	8	8	8	8	8	8	…	9	9	9	9	9	9	9	10	10	10	10	10	10	10	10
9	9	9	9	9	9	9	9	…	10	10	10	10	10	10	11	11	11	11	11	11	11	11	11
10	10	10	10	10	10	10	10	…	11	11	11	11	11	12	12	12	12	12	12	12	12	12	12
11	11	11	11	11	11	11	11	…	12	12	12	12	13	13	13	13	13	13	13	13	13	13	13
12	12	12	12	12	12	13	13	…	13	13	13	14	14	14	14	14	14	14	14	14	14	14	14
13	13	13	14	14	15	14	14	…	14	14	15	15	15	15	15	15	15	15	15	15	15	15	15
14	15	16	15	16	16	15	16	…	15	16	16	16	16	16	16	16	16	16	16	16	16	16	16

흡광도 측정값 전체를 이용한 최적 주성분회귀모델의 구성

주어진 402개의 파장에서의 흡광도 측정값들 전체를 이용한 최적 주성분회귀모델의 구성의 결과는 Table II에 표시된 바와 같다. 평균이동 및 분산보정을 한 경우와 하지 않은 경우를 구분하여 최적 주성분회귀모델을 검색하였으며, 고유값을 기준으로 차례대로 선택된 초기 모델의 고유벡터의 개수를 결정하기 위한 방법으로 SEC 및 IND를 사용한 경우로 나누어 최종 최적 주성분회귀모델을 검색하였다. Table II에 표시된 바와 같이 BSA에 대해서 평균이동은 사용하지 않고, 분산보정은 적용한 자료에 대해서 SEC를 기준으로 1번부터 14번의 고유벡터에 의한 스코어 행렬로 구성한 초기 주성분회귀모델이 얻어졌고, 이 14개의 고유벡터들 중에서 1개부터 14개의 고유벡터를 선택하는 모든 조합에 대해서 SEC, 결정계수, SEP, SECP의 적합성 검정값들을 기준으로 최종적으로 최적의 주성분회귀모델을 검색하였다. 그 결과 SEP

를 판단기준으로 한 경우에는 10개의 고유벡터가, SECP를 판단기준으로 한 경우에는 9개의 고유벡터가 선택된 주성분회귀모델이 최적의 농도 추정 모델이 됨을 알 수 있었다. 포도당에 대해서는 평균이동 및 분산보정을 모두 적용하지 않은 자료에 대해서 SEC를 기준으로 1번부터 16번의 고유벡터에 의한 스코어 행렬로 구성한 초기 주성분회귀모델이 얻어졌고, 이 16개의 고유벡터들 중에서 1개부터 16 개의 고유벡터를 선택하는 모든 조합에 대해서 SEC, 결정계수, SEP, SECP의 적합성 검정값들을 기준으로 최종적으로 최적의 주성분회귀모델을 검색하였다. 그 결과 SEP를 판단기준으로 한 경우에는 11개의 고유벡터가, SECP를 판단기준으로 한 경우에는 13개의 고유벡터가 선택된 주성분회귀모델

Table IV–*PCR Models Found Best using Absorbance Spectra with a Priori Selection of Spectral Points (308 and 281 Spectral Points for BSA and Glucose, Respectively) by Correlation Coefficients. Initial Sets of Factors (Eigenvectors) for Initial PCR Models Were Found by Means of SEC or Malinowski's IND Function and then Final Sets of Factors Were Found by Means of SEC, R^2, SEP or SECP*

Mean Centering	Variance Scaling	Initial Selection	Compound	Initial Factors	Final Selection	Final Factors	R^2	SEC	SEP	SECP
No	No	by SEC	BSA (308 spectral points)	16	by SEC	12	1.0000	0.0139	0.0381	0.0334
					by R^2	16	1.0000	0.0190	0.0321	0.0307
					by SEP	13	1.0000	0.0235	0.0197	0.0207
					by SECP	13	1.0000	0.0235	0.0197	0.0207
			Glucose (281 spectral points)	16	by SEC	14	0.9999	0.2346	0.5215	0.4762
					by R^2	16	0.9999	0.2989	0.5540	0.5265
					by SEP	11	0.9997	0.3139	0.3751	0.3581
					by SECP	11	0.9997	0.3139	0.3751	0.3581
		by IND	BSA	8	by SEC	7	1.0000	0.0227	0.0251	0.0242
					by R^2	8	1.0000	0.0235	0.0247	0.0242
					by SEP	8	1.0000	0.0235	0.0247	0.0242
					by SECP	7	1.0000	0.0227	0.0251	0.0242
			Glucose	10	by SEC	9	0.9993	0.4265	0.8720	0.7489
					by R^2	10	0.9993	0.4389	0.8009	0.7051
					by SEP	9	0.9992	0.4513	0.7009	0.6261
					by SECP	9	0.9992	0.4513	0.7009	0.6261
	Yes	by SEC	BSA	16	by SEC	12	1.0000	0.0128	0.0325	0.0286
					by R^2	16	1.0000	0.0188	0.0319	0.0304
					by SEP	12	1.0000	0.0210	0.0164	0.0177
					by SECP	9	1.0000	0.0178	0.0167	0.0171
			Glucose	16	by SEC	13	1.0000	0.1283	0.4690	0.4136
					by R^2	16	1.0000	0.1680	0.4778	0.4484
					by SEP	13	0.9997	0.3408	0.3981	0.3851
					by SECP	12	0.9998	0.2775	0.4057	0.3755
		by IND	BSA	8	by SEC	6	1.0000	0.0200	0.0246	0.0228
					by R^2	8	1.0000	0.0215	0.0242	0.0232
					by SEP	6	1.0000	0.0239	0.0239	0.0239
					by SECP	6	1.0000	0.0200	0.0246	0.0228
			Glucose	10	by SEC	8	0.9994	0.3781	0.7430	0.6341
					by R^2	10	0.9994	0.4161	0.7408	0.6543
					by SEP	9	0.9994	0.3955	0.7262	0.6320
					by SECP	9	0.9994	0.3955	0.7262	0.6320

Table IV–*Continued*

			BSA	16	by SEC	13	1.0000	0.0123	0.0347	0.0314
					by R^2	16	1.0000	0.0190	0.0325	0.0315
					by SEP	14	1.0000	0.0242	0.0170	0.0184
		by SEC			by SECP	13	1.0000	0.0217	0.0174	0.0183
			Glucose	16	by SEC	14	1.0000	0.1759	0.4869	0.4486
					by R^2	16	1.0000	0.1960	0.4736	0.4545
					by SEP	13	0.9998	0.3094	0.3250	0.3218
	No				by SECP	12	0.9998	0.2885	0.3287	0.3195
			BSA	8	by SEC	7	1.0000	0.0233	0.0251	0.0245
					by R^2	8	1.0000	0.0244	0.0251	0.0249
					by SEP	8	1.0000	0.0244	0.0251	0.0249
		by IND			by SECP	7	1.0000	0.0233	0.0251	0.0245
			Glucose	10	by SEC	8	0.9993	0.4278	0.8658	0.7445
					by R^2	10	0.9994	0.4606	0.7866	0.7058
					by SEP	9	0.9991	0.5220	0.6484	0.6106
Yes					by SECP	9	0.9991	0.5220	0.6484	0.6106
			BSA	16	by SEC	12	1.0000	0.0142	0.0242	0.0222
					by R^2	16	1.0000	0.0225	0.0290	0.0284
					by SEP	10	1.0000	0.0208	0.0137	0.0161
		by SEC			by SECP	11	1.0000	0.0187	0.0144	0.0157
			Glucose	16	by SEC	13	1.0000	0.1441	0.4908	0.4416
					by R^2	16	1.0000	0.1990	0.4776	0.4584
					by SEP	13	0.9997	0.3884	0.3918	0.3910
	Yes				by SECP	11	0.9997	0.3335	0.4057	0.3876
			BSA	8	by SEC	7	1.0000	0.0216	0.0248	0.0237
					by R^2	8	1.0000	0.0227	0.0245	0.0239
					by SEP	7	1.0000	0.0236	0.0230	0.0232
		by IND			by SECP	6	1.0000	0.0226	0.0232	0.0230
			Glucose	10	by SEC	8	0.9994	0.4003	0.7162	0.6255
					by R^2	10	0.9994	0.4357	0.7162	0.6459
					by SEP	9	0.9994	0.4172	0.6738	0.6033
					by SECP	9	0.9994	0.4172	0.6738	0.6033

이 최적의 농도 추정 모델이 됨을 알 수 있었다.

1차 미분값 전체를 이용한 최적 주성분회귀모델의 구성

주어진 402개의 파장에서의 1차 미분값 전체를 이용한 최적 주성분회귀모델의 구성의 결과는 Table III에 표시된 바와 같다. BSA에 대해서 평균이동 및 분산보정을 모두 적용한 자료에 대해서 SEC를 기준으로 1번부터 16번의 고유벡터에 의한 스코어 행렬로 구성한 초기 주성분회귀모델이 얻어졌고, 이 16개의 고유벡터들 중에서 1개부터 16개의 고유벡터를 선택하는 모든 조합에 대해서 최적의 최종 주성분회귀모델을 검색하였다. 그 결과 SEP를 판단기준으로 한 경우에는 11개의 고유벡터가, SECP를 판단기준으로 한 경우에는 9개의 고유벡터가 선택된 주성분회귀모델이 최적의 농도 추정 모델이었다. 포도당에 대해서는 평균이동 및 분산보정을 모두 적용한 자료에 대해서 SEC를 기준으로 1번부터 15번의 고유벡터에 의한 스코어 행렬로 구성한 초기 주성분회귀모델이 얻어졌고, 이 15개의 고유벡터들 중에서 1개부터 15개의 고유벡터를 선택하는 모든 조합에 대해서 최적의 최

종 주성분회귀모델을 검색하였다. 그 결과 SEP를 판단기준으로 한 경우에는 14개의 고유벡터가, SECP를 판단기준으로 한 경우에는 13개의 고유벡터가 선택된 주성분회귀모델이 최적의 농도 추정 모델이 됨을 알 수 있었다.

흡광도와 농도사이의 피어슨 상관계수 기준 파장 선별에 의한 최적 주성분회귀모델 구성

주어진 402개의 파장에서의 흡광도 측정값들 중에서 피어슨 상관계수를 기준으로 하여 그 절대값이 0.2 이상을 보이는 파장에서의 흡광도 측정값만을 사전 선별하여 최적 주성분회귀모델을 구성한 결과는 Table IV에 표시된 바와 같다. BSA의 경우에는 평균이동 및 분산보정을 모두 적용한 자료

Table V–*PCR Models Found Best using 1st Derivative Spectra with a Priori Selection of Spectral Points (347 and 25 Spectral Points for BSA and Glucose, Respectively) by Correlation Coefficients. Initial Sets of Factors (Eigenvectors) for Initial PCR Models Were Found by Means of SEC or Malinowski's IND Function and then Final Sets of Factors Were Found by Means of SEC, R^2, SEP or SECP*

Mean Centering	Variance Scaling	Initial Selection	Compound	Initial Factors	Final Selection	Final Factors	R^2	SEC	SEP	SECP
No	No	by SEC	BSA	13	by SEC	9	1.0000	0.0149	0.0256	0.0225
					by R^2	13	1.0000	0.0186	0.0248	0.0235
					by SEP	9	1.0000	0.0275	0.0226	0.0244
					by SECP	10	1.0000	0.0155	0.0232	0.0211
			Glucose	16	by SEC	14	0.9972	1.2175	5.7411	5.1383
					by R^2	16	0.9975	1.4865	5.6235	5.2548
					by SEP	12	0.9103	5.8125	2.9704	3.9425
					by SECP	10	0.9671	3.1036	3.2747	3.2207
		by IND	BSA	8	by SEC	8	1.0000	0.0191	0.0527	0.0435
					by R^2	8	1.0000	0.0191	0.0527	0.0435
					by SEP	7	1.0000	0.0280	0.0472	0.0408
					by SECP	7	1.0000	0.0280	0.0472	0.0408
			Glucose	8	by SEC	6	0.9479	3.2514	6.1944	5.2036
					by R^2	8	0.9521	3.3890	5.7781	5.0355
					by SEP	6	0.8737	5.0603	5.1085	5.0890
					by SECP	6	0.9416	3.4401	5.1857	4.5579
	Yes	by SEC	BSA	16	by SEC	13	1.0000	0.0115	0.0353	0.0313
					by R^2	16	1.0000	0.0152	0.0332	0.0313
					by SEP	12	1.0000	0.0253	0.0221	0.0230
					by SECP	10	1.0000	0.0197	0.0230	0.0220
			Glucose	16	by SEC	12	0.9976	0.9559	5.0827	4.3731
					by R^2	16	0.9979	1.3588	5.3408	4.9886
					by SEP	10	0.9312	4.4899	3.4654	3.8247
					by SECP	9	0.9311	4.2608	3.5251	3.7949
		by IND	BSA	6	by SEC	6	1.0000	0.0207	0.0254	0.0236
					by R^2	6	1.0000	0.0207	0.0254	0.0236
					by SEP	6	1.0000	0.0207	0.0254	0.0236
					by SECP	6	1.0000	0.0207	0.0254	0.0236
			Glucose	8	by SEC	6	0.9272	3.8433	5.5679	4.9405
					by R^2	8	0.9292	4.1200	5.7270	5.1958
					by SEP	5	0.8547	5.2305	4.3039	4.7193
					by SECP	6	0.9195	4.0411	5.0771	4.6839

Table V–*Continued*

					by SEC	8	1.0000	0.0194	0.0521	0.0437
				BSA	by R^2	8	1.0000	0.0194	0.0521	0.0437
				8	by SEP	7	1.0000	0.0288	0.0466	0.0410
			by SEC		by SECP	7	1.0000	0.0288	0.0466	0.0410
					by SEC	13	0.9951	1.6108	4.2091	3.8166
				Glucose	by R^2	16	0.9974	1.8448	4.6569	4.4660
				16	by SEP	10	0.9068	5.5427	3.1683	4.0207
					by SECP	8	0.9813	2.2209	3.3936	3.0407
	No				by SEC	7	1.0000	0.0256	0.0563	0.0474
				BSA	by R^2	7	1.0000	0.0256	0.0563	0.0474
				7	by SEP	6	1.0000	0.0324	0.0509	0.0447
			by IND		by SECP	6	1.0000	0.0324	0.0509	0.0447
					by SEC	6	0.9547	3.1540	5.2511	4.5553
				Glucose	by R^2	8	0.9572	3.3573	5.2590	4.6911
				8	by SEP	5	0.8730	5.0744	4.1753	4.5620
					by SECP	5	0.9472	3.2724	4.7390	4.2053
					by SEC	9	1.0000	0.0122	0.0304	0.0260
				BSA	by R^2	16	1.0000	0.0222	0.0309	0.0302
				16	by SEP	12	1.0000	0.0245	0.0205	0.0215
			by SEC		by SECP	8	1.0000	0.0196	0.0210	0.0205
					by SEC	11	0.9972	1.0319	4.5420	3.9195
				Glucose	by R^2	16	0.9980	1.6377	4.3733	4.1904
				16	by SEP	12	0.9452	4.9080	3.2913	3.7435
					by SECP	11	0.9731	3.1836	3.6521	3.5321
		Yes			by SEC	5	1.0000	0.0201	0.0262	0.0239
				BSA	by R^2	6	1.0000	0.0207	0.0261	0.0242
				6	by SEP	6	1.0000	0.0207	0.0261	0.0242
			by IND		by SECP	5	1.0000	0.0201	0.0262	0.0239
					by SEC	6	0.9319	3.8666	4.7693	4.4417
				Glucose	by R^2	8	0.9344	4.1574	5.0634	4.7705
				8	by SEP	5	0.8803	4.9256	3.7429	4.2631
					by SECP	5	0.9250	3.8994	4.4349	4.2256

에 대해서 SEC를 기준으로 1번부터 16번의 고유벡터에 의한 스코어 행렬로 구성한 초기 주성분회귀모델이 얻어졌고, 이 16개의 고유벡터들 중에서 1개부터 16개의 고유벡터를 선택하는 모든 조합에 대해서 최적의 최종 주성분회귀모델을 검색한 결과 SEP를 기준으로 한 경우에는 10개의 고유벡터가, SECP를 판단기준으로 한 경우에는 11개의 고유벡터가 선택된 주성분회귀모델이 최적의 농도 추정 모델이었다. 포도당에 대해서는 평균이동은 적용하고 분산보정을 적용하지 않은 자료에 대해서 SEC를 기준으로 얻어진 16 개의 고유벡터들 중에서 1개부터 16개의 고유벡터를 선택하는 모든 조합에 대해서 최적의 최종 주성분회귀모델을 검색한 결과 SEP를 기준으로 한 경우에는 13개의 고유벡터가, SECP를 판단기준으로 한 경우에는 12개의 고유벡터가 선택된 주성분회귀모델이 최적의 농도 추정 모델이었다.

미분값와 농도사이의 피어슨 상관계수 기준 파장 선별에 의한 최적 주성분회귀모델 구성

주어진 402개의 파장에서의 미분값들 중에서 피어슨 상

Table VI–*The Best Estimation Results of BSA Concentration using Absorbance Spectra of 402 Given Spectral Points. Data Preprocessing Options Are: Mean-centering = No; Variance Scaling = Yes. Prepared and Estimated Concentrations Are in g/dL. Reduction in SEP or SECP is Shown in Percentages*

Modeling	All factors		Factor selection by SEP		Factor selection by SECP	
Used Factors	1~14		1~7, 10, 11, 12		1, 2, 3, 5~8, 10, 11	
Prepared	Estimated	Residual	Estimated	Residual	Estimated	Residual
3.589	3.544	−0.045	3.579	−0.010	3.575	−0.014
3.589	3.606	0.017	3.610	0.021	3.609	0.020
3.589	3.583	−0.006	3.568	−0.021	3.569	−0.020
3.077	3.127	0.050	3.075	−0.002	3.080	0.003
2.471	2.498	0.027	2.484	0.013	2.486	0.015
3.398	3.388	−0.011	3.395	−0.003	3.395	−0.003
1.853	1.838	−0.015	1.839	−0.014	1.839	−0.014
3.089	3.090	0.001	3.103	0.014	3.097	0.008
3.574	3.587	0.013	3.578	0.004	3.588	0.014
3.574	3.567	−0.007	3.560	−0.014	3.568	−0.006
3.574	3.543	−0.031	3.548	−0.026	3.549	−0.025
3.574	3.567	−0.008	3.570	−0.004	3.567	−0.007
2.984	2.961	−0.023	2.966	−0.018	2.962	−0.022
4.974	4.988	0.014	4.975	0.001	4.977	0.003
1.989	1.989	0.000	2.002	0.013	1.999	0.010
3.979	3.955	−0.024	3.966	−0.013	3.962	−0.017
5.968	5.952	−0.016	5.972	0.004	5.967	−0.001
3.979	3.953	−0.026	3.957	−0.022	3.954	−0.025
3.979	3.978	−0.002	3.972	−0.007	3.976	−0.003
SEP	0.0222		0.0139 37.4% Reduction		0.0144	
SECP	0.0208		0.0158		0.0154 25.9% Reduction	

관계수를 기준으로 하여 그 절대값이 0.2 이상을 보이는 파장에서의 미분값만을 사전 선별하여 최적 주성분회귀모델을 구성한 결과는 Table V에 표시된 바와 같다. BSA의 경우에는 평균이동 및 분산보정을 모두 적용한 자료에 대해서 SEC를 기준으로 1번부터 16번의 고유벡터에 의한 스코어 행렬로 구성한 초기 주성분회귀모델이 얻어졌고, 이 16개의 고유벡터들 중에서 1개부터 16개의 고유벡터를 선택하는 모든 조합에 대해서 최적의 최종 주성분회귀모델을 검색한 결과 SEP를 기준으로 한 경우에는 12개의 고유벡터가, SECP를 판단기준으로 한 경우에는 8개의 고유벡터가 선택된 주성분회귀모델이 최적의 농도 추정 모델이었다. 포도당에 대해서는 평균이동은 적용하고 분산보정을 적용하지 않은 자료에 대해서 SEC를 기준으로 얻어진 16개의 고유벡터들 중에서 1개부터 16개의 고유벡터를 선택하는 모든 조합에 대해서 최적의 최종 주성분회귀모델을 검색한 결과 SEP를 기준으로 한 경우에는 12개의 고유벡터가 선택된 주성분회귀모델이 최적의 농도 추정 모델이었고, 평균이동은 적용하고 분산보정은 적용하지 않은 경우에 대해서 SECP를 기준으로 한 경우에 최종적으로 8개의 고유벡터가 선택된 주성분회귀모델이 최적의 농도 추정 모델이었다.

최적 주성분회귀모델에 따른 검정군 농도 추정 결과

위에 언급한 여러 연산 조건에 따른 모든 가능한 고유벡터들의 조합에 의한 주성분회귀모델들의 최적조건을 비교한 결과가 정리된 Table II에서 Table IV의 결과를 종합하면 1차 미분값을 대상으로 한 모델의 경우에는 그 최적의 주성분모델들이라 하여도 흡광도 측정값을 대상으로 한 모델에 비하여 SEP 및 SECP를 기준으로 했을 때 더 나은 결과를

Table VII–*The Best Estimation Results of BSA Concentration using Absorbance Spectra with a Priori Selection of Spectral Points by Correlation Coefficient. Data Preprocessing Options Are: Mean-centering = Yes; Variance Scaling = Yes. Prepared and Estimated Concentrations Are in g/dL. Reduction in SEP or SECP Is Shown in Percentages*

Modeling	All factors		Factor selection by SEP		Factor selection by SECP	
Used Factors	1~16		1~5, 7, 10, 12, 13, 15		1~5, 7, 8, 10, 12, 13, 15	
Prepared	Estimated	Residual	Estimated	Residual	Estimated	Residual
3.589	3.552	−0.037	3.592	0.003	3.590	0.001
3.589	3.617	0.028	3.611	0.022	3.609	0.020
3.589	3.587	−0.002	3.567	−0.022	3.576	−0.014
3.077	3.165	0.088	3.077	0.000	3.086	0.009
2.471	2.477	0.006	2.473	0.002	2.477	0.006
3.398	3.417	0.019	3.402	0.004	3.404	0.006
1.853	1.836	−0.017	1.841	−0.012	1.841	−0.012
3.089	3.118	0.029	3.118	0.029	3.107	0.018
3.574	3.610	0.036	3.577	0.003	3.585	0.011
3.574	3.574	0.000	3.559	−0.015	3.581	0.007
3.574	3.550	−0.024	3.556	−0.018	3.560	−0.014
3.574	3.565	−0.009	3.579	0.005	3.580	0.006
2.984	2.961	−0.023	2.968	−0.016	2.957	−0.027
4.974	4.980	0.006	4.984	0.010	4.991	0.017
1.989	2.004	0.015	1.996	0.007	1.981	−0.008
3.979	3.963	−0.016	3.965	−0.014	3.958	−0.021
5.968	5.981	0.013	5.979	0.011	5.977	0.009
3.979	3.982	0.003	3.961	−0.018	3.953	−0.026
3.979	4.016	0.037	3.982	0.003	3.987	0.008
SEP	0.0290		0.0137 52.8% Reduction		0.0144	
SECP	0.0284		0.0161		0.0157 44.7% Reduction	

나타내지 못했다. 따라서 시험에 사용된 자료에 대해서는 미분값이 아닌 흡광도 측정값을 그대로 사용하는 것이 적절한 것으로 판단하였다. 각 조건에서 최적의 주성분회귀모델에 의한 검정군 시료의 농도 추정 결과를 Table VI에서 Table IX에 정리하였다. Table VI는 주어진 402개의 파장 전체를 대상으로 주성분회귀모델을 얻어 BSA의 농도를 추정한 것인데 SEP를 기준으로 주성분회귀모델에 포함될 고유벡터를 선택한 결과 37.4%의 SEP의 감소가 있었다. 한편 SECP를 기준으로 고유벡터를 선택한 경우에는 25.9%의 SECP의 감소가 있었다. Table VII는 피어슨 상관계수의 절대값이 0.2 이상인 파장만을 선택한 다음 주성분회귀모델을 얻어 BSA의 농도를 추정한 것인데 SEP를 기준으로 한 경우에는 52.8%의 SEP의 감소가 있었다. SECP를 기준으로 한 경우에는 44.7%의 SECP의 감소가 있었다. Table VIII는 주어진 402개의 파장 전체를 대상으로 주성분회귀모델을 얻어 포도당의 농도를 추정한 것인데 SEP를 기준으로 주성분회귀모델에 포함될 고유벡터를 선택한 결과 19.9%의 SEP의 감소가 있었다. SECP를 기준으로 한 경우에는 17.2%의 SECP의 감소가 있었다. Table IX는 피어슨 상관계수의 절대값이 0.2 이상인 파장만을 선택한 다음 주성분회귀모델을 얻어 포도당의 농도를 추정한 것인데 SEP를 기준으로 한 경우에는 31.4%의 SEP의 감소가 있었다. SECP를 기준으로 한 경우에는 29.7%의 SECP의 감소가 있었다. 주성분회귀모델의 구성에 있어서 고유값의 크기를 기준으로 가장 고유값이 큰 고유벡터부터 순서대로 일련의 모든 고유벡터를 모두 그대로 사용하는 방법 대신, 1차적으로 선정된 고유벡터들에 대해서 가능한 모든 고유벡터들의 조합을 얻었다. 얻어진 조합들 중에서 SEP 또는 SECP를 기준으로 최적의 고유벡

Table VIII–*The Best Estimation Results of Glucose Concentration using Absorbance Spectra of 402 Given Spectral Points. Data Preprocessing Options Are: Mean-centering = No; Variance Scaling = No. Prepared and Estimated Concentrations Are in mmol/L. Reduction in SEP or SECP Is Shown in Percentages*

Modeling	All factors		Factor selection by SEP		Factor selection by SECP	
Used Factors	1~16		1~8, 13, 15, 16		1~10, 12, 13, 15	
Prepared	Estimated	Residual	Estimated	Residual	Estimated	Residual
5.644	5.628	−0.017	5.389	−0.256	5.509	−0.136
11.289	11.324	0.035	11.230	−0.059	11.318	0.029
16.933	18.052	1.118	17.718	0.785	17.592	0.658
22.578	24.927	2.349	22.630	0.052	23.709	1.131
5.544	7.018	1.474	6.851	1.306	6.886	1.342
11.089	12.498	1.409	12.816	1.727	12.795	1.706
16.633	17.854	1.220	17.666	1.033	17.465	0.832
22.178	22.000	−0.178	21.970	−0.208	21.761	−0.417
5.633	5.681	0.048	5.030	−0.603	5.448	−0.186
11.272	11.943	0.670	11.658	0.386	11.603	0.330
16.906	16.773	−0.133	16.685	−0.221	16.916	0.010
22.544	22.295	−0.250	22.180	−0.364	22.442	−0.102
5.583	4.903	−0.680	4.544	−1.040	4.620	−0.964
5.583	5.453	−0.131	5.117	−0.466	5.227	−0.356
11.167	11.595	0.429	11.537	0.370	11.245	0.078
11.167	11.317	0.150	11.381	0.214	10.855	−0.312
11.167	10.641	−0.526	11.213	0.047	10.654	−0.512
11.194	10.309	−0.886	10.312	−0.883	10.325	−0.870
22.328	21.737	−0.591	21.840	−0.488	21.651	−0.677
SEP	0.8908		0.7138 19.9% Reduction		0.7290	
SECP	0.8348		0.7052		0.6914 17.2% Reduction	

터 조합을 검색하였다. 이렇게 한 경우 각 Table에서 보인바와 같이 상당한 수준의 SEP 또는 SECP의 감소가 나타나서 고유벡터에 대한 선별을 하지 않은 경우에 비해 더욱 나은 예측 모델을 구성할 수 있음을 알 수 있었다. 또한 계산을 위해서 주어진 파장들을 그대로 모두 사용하지 않고 피어슨 상관계수를 기준으로 사전 선별을 한 후에 주성분회귀모델을 구성하고 이로부터 고유벡터의 조합들을 구성한 경우에 더욱 현저하게 SEP 또는 SECP가 감소하여 농도의 예측에 있어서 더욱 나은 주성분회귀모델이 구성될 수 있음을 확인할 수 있었다. 일반적으로 채택되는 변수 선택의 방법은 제한 개수의 모델에 대한 평가에 불과하여 실제 최적의 모델이기 보다는 우연히 얻은 차선의 모델일 가능성이 있다. 그러나 본 연구에서 얻어진 결과는 주어진 최초의 주성분회귀모델에서 파생될 수 있는 고유벡터들의 모든 조합을 얻고 이들에 대한 비교 평가에 의해서 얻은 것이므로 주어진 상황에서 얻을 수 있는 최적의 모델이 얻어진 것으로 볼 수 있다.

최적의 주성분회귀 모델을 얻기 위한 방법을 정리하면 다음과 같다. 우선 주어진 측정 자료에 대한 미분은 일반적으로 미분 차수가 높아질수록 해상도는 증가하지만 노이즈의 확대 문제가 발생하므로 본 연구에 사용된 자료의 경우처럼 전체 신호의 크기에 비해 목적하는 성분의 신호가 작은 경우에는 미분이 더 나은 결과를 주지 않았다. 따라서 미분하지 않은 원래의 흡광도값들을 대상으로 최적의 주성분회귀모델을 검색하였다. NIPALS 또는 SVD에 의한 인자분석을 하기에 앞서서 자료에 대한 전처리법으로 평균이동 및 분산보정을 실시하는 문제는 주어진 자료의 특성에 따라 달라지는 것이지만, 이 연구에서 사용된 자료의 경우에는 대체적으로 평균이동 및 분산보정 모두를 실시하거나 그 중 하나를

Table IX–*The Best Estimation Results of Glucose Concentration using Absorbance Spectra with a Priori Selection of Spectra Points by Correlation Coefficient. Data Preprocessing Options Are: Mean-centering = Yes; Variance Scaling = No. Prepared and Estimated Concentrations Are in mmol/L*

Modeling	All factors		Factor selection by SEP		Factor selection by SECP	
Used Factors	1~16		1, 3~10, 12, 13, 14, 16		1, 3~10, 13, 14, 16	
Prepared	Estimated	Residual	Estimated	Residual	Estimated	Residual
5.644	5.834	0.189	5.834	0.190	5.801	0.157
11.289	10.392	−0.897	10.764	−0.525	10.747	−0.542
16.933	17.074	0.141	17.073	0.140	17.056	0.123
22.578	22.800	0.222	22.537	−0.040	22.539	−0.039
5.544	5.181	−0.363	5.164	−0.380	5.171	−0.374
11.089	10.988	−0.101	10.979	−0.110	10.964	−0.125
16.633	16.528	−0.106	16.782	0.149	16.767	0.133
22.178	21.753	−0.425	22.154	−0.024	22.167	−0.011
5.633	6.717	1.083	6.432	0.798	6.434	0.800
11.272	11.291	0.019	11.022	−0.250	10.981	−0.291
16.906	16.285	−0.620	16.463	−0.443	16.476	−0.430
22.544	22.321	−0.223	22.502	−0.043	22.542	−0.002
5.583	4.996	−0.587	5.286	−0.297	5.291	−0.293
5.583	5.619	0.036	5.733	0.149	5.772	0.189
11.167	12.115	0.948	11.819	0.653	11.814	0.647
11.167	11.203	0.036	10.944	−0.223	10.916	−0.251
11.167	11.428	0.261	11.090	−0.077	11.075	−0.092
11.194	10.981	−0.214	11.141	−0.054	11.106	−0.088
22.328	22.549	0.221	22.283	−0.045	22.200	−0.128
SEP	0.4736		0.3250 31.4% Reduction		0.3287	
SECP	0.4545		0.3218		0.3195 29.7% Reduction	

시행한 경우가 그렇지 않은 경우에 비해 최종 SEP 또는 SECP를 기준으로 더 나은 모델을 제시하는 것으로 보여 진다. 또한 각 개별 파장에서의 측정값과 농도사이의 상관계수를 구하여 상관계수의 절대값이 0.2 이상인 파장에서의 측정값만을 선택하여 인자분석을 실시하는 경우가 그렇지 않은 경우에 비해서 현저하게 향상된 회귀모델을 얻게 함을 확인할 수 있었다. 또한 일차적으로 얻어진 주성분회귀 모델의 구성에 사용된 고유벡터들에서 얻을 수 있는 모든 조합의 세트들에 대해서 SEP 및 SECP를 기준으로 최적의 회귀식을 검색한 결과 기존의 방법인 고유값이 큰 것부터 차례대로 선택하여 최종 고유벡터의 개수만을 선정하는 경우보다 SEP 또는 SECP를 기준으로 볼 때 현저하게 나은 주성분회귀 모델이 얻어짐을 확인할 수 있었다.

감사의 말씀

이 연구는 숙명여자대학교 약학연구소 2005년도 특별연구비 지원에 의하여 이루어졌음을 밝혀둡니다.

참고문헌

1) T. Koschkinsky and L. Heinemann. Sensors for glucose monitoring: technical and clinical aspects, *Diabetes Metab. Res. Rev.*, **17**, 113-123 (2001).
2) R. Kramer, *Chemometric Techniques for Quantitative Analysis*, Marcel Dekker, Inc., 99p (1998).
3) E. R. Malinowski, Determination of the number of factors and the experimental error in a data matrix, *Anal. Chem.*, **49**, 612-617 (1977).
4) K. Faber, B. R. Kowalski, Critical evaluation of two F-tests

for selecting the number of factors in abstract factor analysis, *Anal. Chim. Acta*, **337**, 57-71 (1997).
5) U. Depczynski, V. J. Frost and K. Molt, Genetic algorithm applied to the selection of factors in principal component regression, *Anal. Chim. Acta*, **420**, 217-227 (2000).
6) N. R. Draper and H. Smith, *Applied Regression Analysis, 3rd Ed.*, John Wiley & Sons, Inc., 41p (1998).
7) A. Savitzky and M. J. E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.*, **36**, 1627-1639 (1964).
8) E. R. Malinowski, *Factor Analysis in Chemistry, 2nd Ed.*, John Wiley & Sons, Inc., (1991).
9) K. Faber and B. R. Kowalski, Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler, *Chemom. Intell. Lab. Syst.*, **34**, 283-292 (1996).
10) M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, J. Pagès, NIR calibration in non-linear systems: different PLS apporoaches and artificial neural networks, *Chemom. Intell. Lab. Syst.*, **50**, 75-82 (2000).