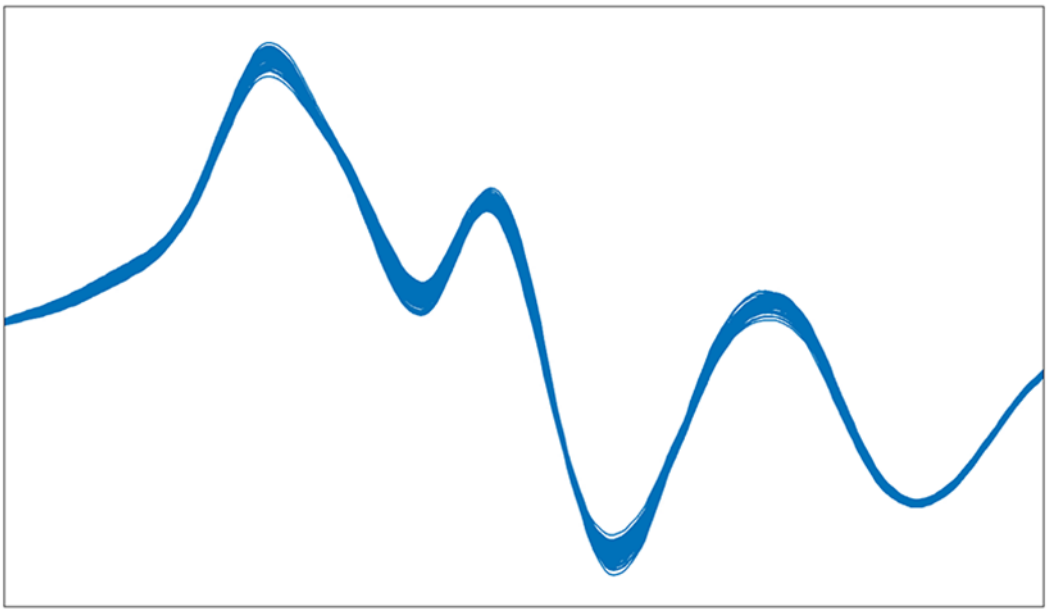
OMNIS NIR의 이론



매뉴얼

8.0600.8101KO / v9 / 2025-10-10

Metrohm

Metrohm AG
Ionenstrasse
CH-9100 Herisau
Switzerland
+41 71 353 85 85
info@metrohm.com
www.metrohm.com

OMNIS NIR의 이론

매뉴얼

8.0600.8101KO / v9 /
2025-10-10

Technical Communication
Metrohm AG
CH-9100 Herisau

본 문서는 저작권법의 보호를 받습니다. 모든 권리는 당사에 있습니다.

본 문서는 원본 문서입니다.

본 문서는 신중을 기하여 작성하였습니다. 하지만 오류를 완전히 배제할 수는 없습니다. 만약 본 문서에서 오류를 발견하신다면 위에 명시한 주소로 연락주시기 바랍니다.

면책조항

부적절한 보관, 부적절한 사용 등과 같이 Metrohm의 귀책사유가 아닌 다른 이유로 발생한 결함에 대해서는 품질보증이 제공되지 않음을 분명하게 밝히는 바입니다. 제품에서의 자체 변경(예를 들어 개조 또는 부착)에 대해 제조사는 그로 인해 발생하는 손해 및 후속 손해에 대해 어떠한 책임도 지지 않습니다. Metrohm 제품 문서에 명시된 지침 및 매뉴얼의 내용은 반드시 준수해야 합니다. 그렇지 않을 경우 Metrohm에서는 어떠한 보증도 제공하지 않습니다.

목차

1 개요 1

1.1 서문 ... 1

1.2 개념의 프레임 ... 1

1.3 문서 정보 ... 2

1.4 상세한 정보 ... 2

2 근적외선 및 스펙트럼 3

2.1 빛과 물질과의 상호작용 ... 3

2.2 수학적 기초 ... 6

2.2.1 Beer Lambert 법칙 ... 6

2.2.2 선형 회귀 ... 6

2.3 빛이 스펙트럼으로 변환되는 방식 ... 7

3 장비설정 11

3.1 파장 보정 ... 11

3.2 표준물질을 통한 점검 ... 12

3.2.1 OMNIS NIR Analyzer ... 13

3.2.2 2060 The NIR ... 14

3.3 장비 성능 테스트 ... 22

3.3.1 외부 장비 성능 테스트 (OMNIS NIR Analyzer) ... 24

4 모델 개발 27

4.1 물리적 샘플 ... 28

4.2 주요 구성 요소 분석 (PCA) ... 31

4.3 데이터 전처리 ... 35

4.3.1 데이터 전처리 ... 35

4.3.2 파장범위 ... 43

4.3.3 스펙트럼적인 특이값 ... 44

4.3.4 기준값 특이값 (수량화) ... 50

4.3.5 데이터 세트의 분할 ... 51

4.4 수량화 ... 53

4.4.1 PLS 회귀 ... 53

4.4.2 정량화 모델에 대한 검증 ... 55

4.4.3 OMNIS Model Developer (OMD) ... 62

4.4.4 기울기 수정/y 절편 수정 ... 63

4.5 동일시 및 검증 ... 65

4.5.1 지원 벡터 시스템 (SVM) ... 65

4.5.2 샘플의 제품 제휴 예측 ... 68

4.5.3 식별 모델의 검증 ... 69

4.6 인증 **71**
4.6.1 인증 모델의 계산 71
4.6.2 인증 모델의 유효성 검사 71

5 예측 **73**

5.1 수량화 **73**
5.1.1 특이값 및 결과 모니터링 73
5.2 식별 및 검증 **75**
5.3 인증 **76**

6 부록 **77**

6.1 선형 회귀의 예 **77**
6.2 PCA-알고리즘 **80**
6.3 PLS-알고리즘 **82**
6.4 Hotelling T^2 및 Q residuals **83**
6.5 스펙트럼적인 특이값 - 알고리즘 **84**
6.6 기준값 특이값 - 알고리즘 **86**

1 개요

1.1 서문

근적외선 분광법(NIR 분광법)은 넓은 스펙트럼에 대한 샘플에 적합한 비파괴적이고 빠르고 반응성이 있는 분석 방법입니다. 여러 parameter를 동시에 분석하고 재료의 물리적 및 화학적 속성을 모두 측정할 수 있습니다. 여기에는 분석 농도, 밀도, 입자 크기 또는 내부 점도가 포함됩니다.

이외에도 NIR 분광법은 알 수 없는 샘플의 동일시(OMNIS Software 4.0 이상) 및 샘플의 검증(OMNIS Software 4.2 이상)을 가능하게 합니다.

원격 샘플의 비파괴 측정은 품질 관리 및 프로세스 모니터링에 매우 중요합니다.

본 매뉴얼은 OMNIS Software에 구현된 근적외선 스펙트럼의 기록, 처리 및 분석 기술 및 알고리즘에 대해 설명합니다. 제2장에서는 측정 신호가 흡수도 스펙트럼으로 변경되는 방법을 간략히 설명합니다. 제3장에서는 장비 보정에 대한 내용이 있습니다. 제4장에서는 관심있는 parameter (정량화) 또는 제품 제휴(동일시)를 예측할 수 있는 모델의 개발을 설명합니다. 5장에서는 미지의 예측에 대해 다룹니다. 제6장은 다양한 알고리즘에 대한 설명과 함께 부록을 구성합니다.

1.2 개념의 프레임

제시된 처리는 다음 프레임에 통합됩니다 :

1. **보정, 표준화 및 성능 테스트**
   장비로 기록된 흡수도 스펙트럼의 이동성과 신뢰성을 보장합니다.
2. **모델 개발**
   정량적 parameter의 예측 또는 샘플 동일시에 대한 모델이 개발됩니다.
   개발은 알려진 관심있는 parameter 또는 제품 제휴를 가진 샘플에 기초합니다.
3. **샘플 분석**
   스펙트럼은 분석된 샘플에서 기록했습니다. 스펙트럼을 기반으로 정량화 모델은 정량적 예측을 제공하거나 또는 식별 모델은 샘플을 식별하거나 또는 검증합니다.
4. **모니터링**
   모델 및 장비를 모니터링하는 경우 시스템이 추가 분석에 적합함을 점검할 수 있습니다.

1.3 문서 정보

기호 설명

굵은 자모는 행렬을 나타내고 굵은 자모는 벡터를 나타냅니다. 행렬 또는 벡터의 전치는 높은 위치의 t, 예를 들어 $\mathbf{x}^t$로 특징지어집니다.

단계는 소문자로 표시됩니다. 곡절 부호(^)는 표시, 추정 및 (계산된) 값을 나타냅니다, 예를 들어 $\hat{y}$. 십자선은 평균값을 나타냅니다, 예를 들어 $\bar{y}$.

단계는 예를 들어 A 흡수도와 같은 파장 종속 변수를 나타낼 수 있습니다. 단계 표기법은 벡터 표기법과 상호 교환적으로 사용될 수 있습니다.

1.4 상세한 정보

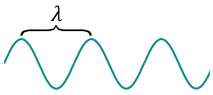
다음 사이트에서 제품에 대한 자세한 정보를 찾을 수 있습니다 :

- Metrohm 웹사이트 *https://www.metrohm.com* – 제품군에 대한 개요, PDF 형식의 문서, 부속품에 대한 정보 및 어플리케이션 정보.
- OMNIS Software의 도움말 *https://guide.metrohm.com* – 주제별로 필터링된 OMNIS Software 정보.

2 근적외선 및 스펙트럼

2.1 빛과 물질과의 상호작용

분광계는 샘플이 빛과 상호작용하는 방법을 측정합니다. 빛은 다양한 양으로 흡수되거나 확산될 수 있습니다. 상호작용은 빛의 성질, 특히 파장, 그리고 물질의 성질, 속성 분자 구조에 따라 달라집니다.



파장

빛은 전자기 방사선입니다. 빛은 전기장과 자기장이 흔들리는 파동처럼 공간을 움직입니다. 파도는 시간과 공간으로 퍼집니다. 파동은 파장 λ (예를 들어, 나노미터 = 10^{-9} 미터) 및 주파수(Hz)에 의해 특징지어집니다.

파장은 파장의 주파수에 반비례합니다. 주파수가 높은 파형(초당 더 많은 진동)은 더 짧은 파장을 가지고 그 반대의 경우도 마찬가지입니다. 이 관계 때문에 파동은 파장(nm) 또는 주파수(Hz)로 설명할 수 있습니다.

빛은 불연속적인 양자 단위, 즉 광자로 에너지를 교환할 수 있습니다. 단일 광자 E의 에너지는 주파수 f 또는 λ 파장에 따라 달라집니다 :

$$E = hf = h\frac{c}{\lambda}$$

여기서 h는 플랑크 상수에 해당하고 c는 광속에 해당합니다.

*그림1*은 전자기 방사선의 다양한 범위를 보여줍니다.

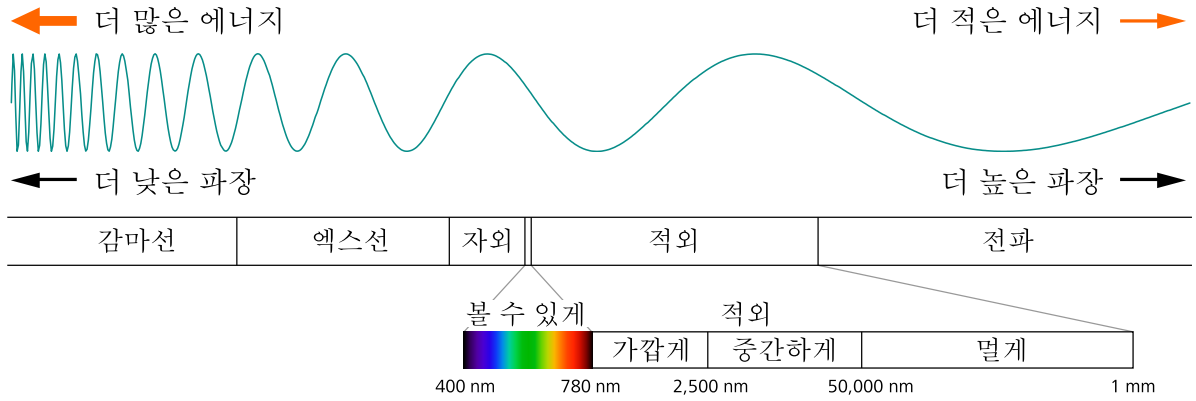


그림 1 전자기 방사선의 범위. 근적외선 범위(NIR)는 가시광선 범위 옆에 있습니다. NIR 780nm 에서 2,500nm 까지의 파장을 포함합니다.

방사선 근원

전자기 방사선의 다른 범위은 다른 방사선 근원을 가집니다. NIR 범위에서 근원은 **열 복사**입니다. 인간의 몸은 주변 환경보다 더 높은 온도를 가지고 있기 때문에 예를 들어, 적외선 카메라는 인간의 몸을 감지할 수 있습니다.

강도

전자기파의 진폭은 빛의 강도를 측정합니다. 진폭이 높을수록 강도가 높아집니다. 가시광선에서는 강도가 밝기로 인식됩니다.

빛과 물질의 상호작용

분자가 빛을 흡수하는 과정은 다음과 같습니다.

빛이 물질과 상호작용하는 방법은 전자기 방사선의 범위에 따라 달라집니다. 예를 들어, 에너지가 가시광선에서 분자로 전달될 때, 분자의 전자는 낮은 에너지 수준에서 더 높은 에너지 수준(전자 전환)으로 전환됩니다. **진동 전환**은 적외선 영역에서 발생합니다. 화학 결합, 기능 그룹 및 분자는 스트레칭 진동, 변형 진동 또는 회전 진동과 같은 다양한 방식으로 진동할 수 있습니다.

분자는 별도의 진동 조건만 사용해도 됩니다. 주변 온도에서 대부분의 분자는 기본 진동 상태(수준 0)에 있습니다. 기본 진동 상태에서 흥분 상태로의 전환은 다음 다이어그램의 이름을 따서 지명됩니다 :

진동 전환 ***i*** **➔** ***j***	**이름**
0 ➔ 1	기본 전환
0 ➔ 2	첫 번째 배음 전환
0 ➔ 3	두 번째 배음 전환

더 높은 진동 상태는 더 높은 에너지 수준에 해당합니다. 상태 *i*에서 흥분된 상태 *j*로 전환하려면 분자가 특정 과도 에너지 ΔE_{ij}를 흡수해야 합니다.

빛은 $E = hf$의 부분으로 에너지를 교환할 수 있고, 여기서 f는 빛의 주파수입니다. 빛의 흡수도는 광자 에너지 hf가 과도 에너지 ΔE_{ij}와 동일할 때 발생합니다.

허용되는 진동 상태는 무엇보다도 관련된 원자의 결합 강도 및 질량에 따라 달라집니다. 따라서 특정 타입의 결합은 특성 전환 에너지 또는 흡수된 파장과 연관될 수 있습니다.

빛을 흡수하기에 대해 더 많은 조건이 충족되어야 합니다. 진동 전환은 분자의 전기 쌍극 모멘트가 변경되도록 전하 분포를 이동시켜야 합니다. 에너지 흡수도 확률은 영향을 받는 화학적 결합을 따라 쌍극 모멘트의 변경 정도에 따라 달라집니다.

진동 전환은 극성 분자와 기능 그룹 모두에서 쌍극 모멘트를 변경할 수 있습니다. N_2와 같은 호모딘 핵 분자는 적외선을 흡수하지 않습니다.

흥분된 진동 상태의 지속 시간은 제한되어 있습니다. 분자가 낮은 진동 상태로 돌아가는 경우 에너지가 열로 변경됩니다.

NIR 스펙트럼적임 범위

기본 전환에 해당하는 파장은 중간 적외선 범위에 있습니다. 근적외선 범위는 오버톤 전환과 조합 밴드로 구성됩니다. *그림2*은 다양한 분자 및 기능 그룹에 의해 흡수되는 파장 밴드가 표시됩니다.

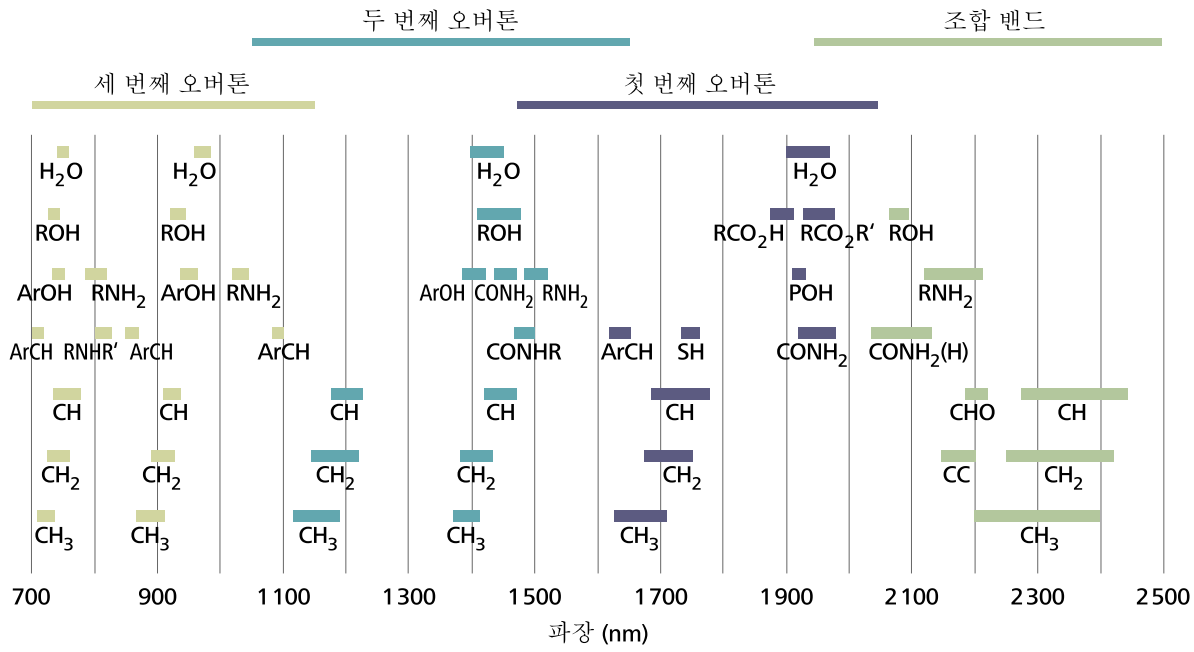


그림 2 NIR 흡수도 밴드

기본 전환은 가장 가능성고 높은 전환이며 가장 자주 발생합니다. 오버톤 전환 가능성은 더 낮습니다. 따라서 기본 전환은 오버톤 전환보다 더 많은 빛을 흡수합니다. 일반적으로 각 오버톤에 따라 흡수가 감소합니다. 따라서 오버톤은 높은 흡수 분자에 적합합니다.

두 개 이상의 기본 진동이 동시에 기본 진동의 결합된 주파수에 해당하는 단일 광 주파수로 트리거될 수 있습니다. 해당 흡수대를 **조합 밴드**라고 합니다. 일부 조합 밴드는 NIR 범위 즉, 1,900 및 2,500nm에 있습니다.

2.2 수학적 기초

2.2.1 Beer Lambert 법칙

Beer Lambert 법칙은 균질 샘플에 의한 빛의 흡수가 샘플 내 흡수 물질의 속성에 어떻게 의존하는지 설명합니다。

$$A = \varepsilon \cdot c \cdot l$$

여기서 A 흡수도, ε 흡수기의 몰 멸종 계수(L/mol/cm), c 흡수기의 농도 (mol/L) 및 l 샘플의 층 두께(cm)에 해당합니다。

몰 ε 멸종 계수는 물질이 흡수하는 양을 나타내는 상수입니다。 몰 멸종 계수는 특정 파장 *i*와 특정 물질 *j*에 한정되고 혼합물의 총 흡수는 혼합물에 포함된 모든 물질의 흡수도를 합한 것입니다：

$$A_i = \sum_{j=1}^{N} \varepsilon_{ij} c_j \, l$$

여기서 A_i는 파장 *i*의 흡수도, *N*은 혼합물의 물질 수, *i* 파장은 ε_{ij}의 몰 멸종 계수 및 물질 *j* 및 c_j의 농도 *j*에 해당합니다。

Beer Lambert 법칙은 흡수와 농도 사이의 선형 관계와 흡수와 몰 멸종 계수 사이의 선형 관계를 가정합니다。 이 선형 동작은 많은 상황에 적용됩니다。

Beer Lambert 법칙에 따르면, 분광학적 흡수 측정은 다음을 증명할 수 있습니다：

- 흡수기 농도의 변경。
  이것은 가장 일반적인 응용입니다。
- 몰 멸종 계수에 영향을 미치는 요인의 변동。
  용매의 온도, 점도, pH 값 또는 유전 상수는 몰 멸종 계수에 영향을 미칠 수 있습니다。 경우에 따라 분광학적 측정에 사용할 수 있습니다。

산란 효과는 Beer Lambert 법칙과 관련이 없습니다。 산란 효과는 예를 들어 입자 크기의 변동을 감지하는 데 사용될 수 있습다。

2.2.2 선형 회귀

한 파장

가장 간단한 경우에, 혼합물은 하나의 흡수기만 가지고 있습니다。 Beer Lambert 법칙에 따르면, 흡수도는 특정 파장에서 흡수기의 농도에 선형입니다。

그림 1에서 각 점은 알려진 흡수기 농도(x축)와 측정된 흡수도(y축)를 가진 샘플을 나타냅니다。 선형 회귀 분석에서는 회귀선 **A**가 생성됩니다。

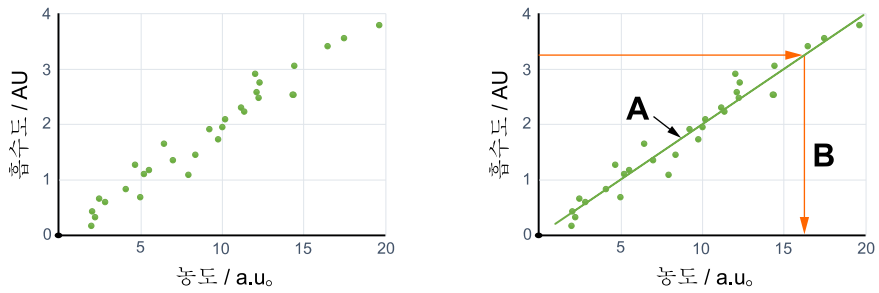


그림 3 흡광도 값과 농도 사이의 관계

알 수 없는 흡수기 농도의 샘플의 농도는 다음과 같이 측정할 수 있습니다 :

1. 지정된 파장에서 흡수도를 측정.
2. 회귀선을 사용하여 농도(**B**)를 측정합니다.
   회귀선은 관심있는 parameter 예를 들어: 흡수기 농도를 예측하기에 대한 간단한 정량화 모델입니다.

그러나 혼합물에 농도가 다른 여러 흡수기가 포함된 경우 이 진행 과정은 오류합니다.

여러 파장

단일 파장에서 흡수를 측정하는 대신 여러 파장에서 흡수를 측정하여 스펙트럼을 생성할 수 있습니다. 위와 마찬가지로 스펙트럼과 관심있는 parameter 간의 관계는 선형 회귀 분석을 통해 측정할 수 있습니다. 여러 파장의 경우 다중 선형 회귀 분석(MLR)이 필요합니다.

혼합물에 농도가 다른 여러 흡수기가 포함되어 있더라도 다중 선형 회귀 분석을 통해 관심있는 parameter를 예측할 수 있습니다 *선형 회귀의 예 (참조: 77 페이지, 6.1 장)*.

그러나 다중 선형 회귀 분석의 경우 파장 수보다 많은 샘플이 필요합니다. 분광적 예측의 경우 PCA 또는 PLS와 같은 다른 method를 사용할 수 있습니다.

2.3 빛이 스펙트럼으로 변환되는 방식

분광계(또는 분광도계)는 광원과 검출기 장비로 구성됩니다. 광원은 넓은 스펙트럼에 대한 파장을 가진 빛을 발산하고 즉, 다색광입니다. 빛은 샘플과 상호작용합니다. 그런 다음 분광계는 파장의 함수로 남아 있는 빛을 감지합니다.

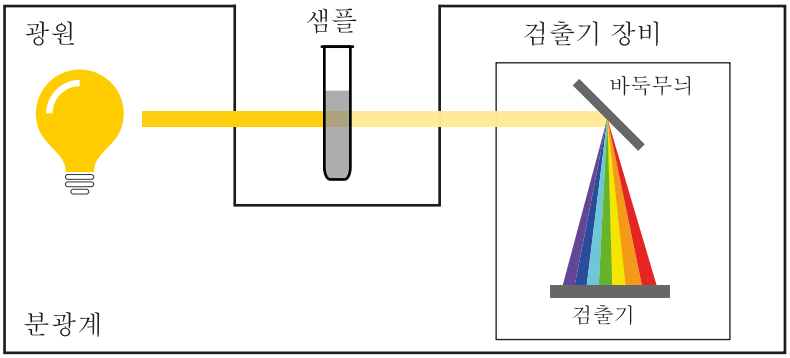


그림 4 광원과 검출기 장비가 있는 분광계.

분광계에서 빛은 그리드의 도움으로 개별 파장으로 분해됩니다. 검출기는 각 파장이 라인 센서의 다른 요소 또는 - 픽셀과 - 만나는 빛을 측정합니다.

스캔은 모든 픽셀에 대한 측정입니다. 각 픽셀은 빛의 강도에 비례하는 광전 신호를 생성합니다. 신호는 픽셀에 적용할 수 있습니다.

그림 5 픽셀의 함수로서 검출기 신호의 스펙트럼.

적분 시간

적분 시간은 검출기가 빛을 수집하는 시간입니다. 적분 시간이 높일수록 신호가 증가합니다.

적분 시간이 너무 길수록 검출기가 포화 상태가 되고 정보가 손실됩니다. 적분 시간이 너무 짧은 경우 신호/노이즈 비율이 감소합니다.

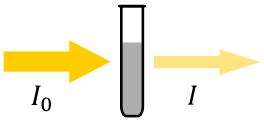
자동 적분 시간 은 최적의 노출을 보장하고 즉, 최적의 신호/노이즈 비율을 의미하고, 포화 상태가 발생하지 않습니다. 각 샘플 스캔과 각 기준 스캔 전에 여러 가지 측정이 수행됩니다. 적분 시간은 가장 높은 강도의 신호가 감지 가능한 범위의 약 90%에 도달하도록 설정됩니다.

적분 시간의 차이는 추가 계산에 주의합니다.

- **OMNIS NIR Analyzer**
  적분 시간은 항상 자동으로 설정됩니다.
- **2060 The NIR**
  적분 시간은 수동으로 또는 자동으로 설정할 수 있습니다.
  수동 적분 시간은 유사한 반복 측정의 경우 측정기간을 단축할 수 있습니다. 통합 시간이 너무 길면 검출기가 포화 상태가 되지 않도록 충분한 여유로 적분 시간을 설정해야 합니다 *(참조: 43페이지, "포화도")*.

흡수도

분광학 측정은 샘플이 얼마나 많은 빛을 흡수하거나 분산시키는지 측정합니다. 샘플은 광선에 노출됩니다. 검출기는 샘플과 상호작용한 후 광원에서 방출되는 빛과 나머지 빛을 측정합니다.



A 흡수도는 광빔이 샘플(I_0)과 상호작용하기 전의 광강도와, 광빔이 샘플(I)과 상호작용한 후의 광강도 사이의 관계에 대한 상용로그로 정의됩니다 :

$$A = \log_{10} \frac{I_0}{I}$$

흡수도가 1이는 경우 빛의 10%가 샘플을 통과하고, 흡수도가 2이는 경우 1%가 통과한다는 뜻입니다.

이 흡수도에는 단위가 없습니다.

참조 스캔 및 샘플 스캔

위의 수식에 따르면 흡수도 스펙트럼을 계산하려면 두 개의 스캔이 필요합니다. 스캔은 각 픽셀에 대해 광전 신호를 측정합니다 S:

- **기준 스캔**은 광선이 샘플과 상호작용하기 전에 신호 S_0 측정합니다.
- **샘플 스캔** 은 광선이 샘플과 상호작용하기 후에 신호 S 측정합니다.

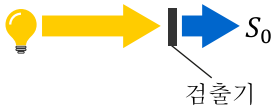
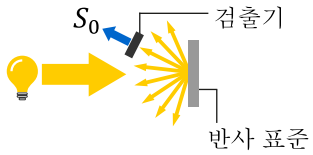
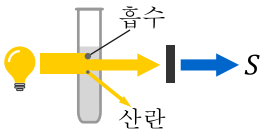
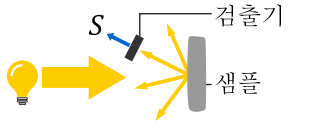
한 픽셀의 광전 신호는 픽셀 표면의 광도 평균값에 비례합니다. 따라서 $S_0/S = I_0/I$입니다. 따라서 광전 신호를 사용하여 흡수도를 계산할 수 있습니다 :

$$A = \log_{10} \frac{I_0}{I} = \log_{10} \frac{S_0}{S}$$

투과 및 반사

전송 모드에서 샘플을 투과하는 빛이 측정됩니다. S_0은 샘플이 없을 때 측정됩니다. S 샘플을 통과하는 빛에서 측정됩니다.

반사 모드에서 샘플에 의해 반사되는 빛을 측정됩니다. 참조로 샘플 대신 반사 표준이 사용됩니다. 반사 표준은 이상적으로 빛의 100%를 반사합니다. 반사된 빛의 일부는 검출기로 전달되고 신호 S_0을 제공합니다. S 신호는 빛을 반사하는 샘플과 동일한 방식으로 측정됩니다.

	투과	확산 반사
샘플 포함 없이		
샘플 포함		

계산된 A 흡광도는 검출기에 도달하지 않는 모든 빛을 나타냅니다. 따라서 A 에는 샘플에 의해 흡수된 빛뿐만 아니라 그것도 이것을 포함됩니다 :

- 검출기에서 산란되기 때문에 검출기에 도달할 수 없는 빛.
- 검출기에 잘못 분산된 빛.

흡수도 스펙트럼

흡수도 스펙트럼은 위의 수식에 따라 기준 스캔(신호 S_0)과 샘플 스캔(신호 S)을 사용하여 계산됩니다.

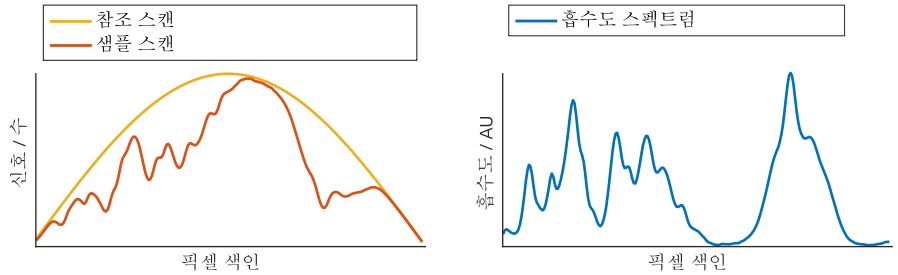


그림 6 기준 스캔 및 샘플 스캔(왼쪽) 및 계산된 흡수도 스펙트럼(오른쪽)은 픽셀 인덱스의 함수입니다.

상기 계산은 기준 스캔 및 샘플 스캔이 동일한 광학적 경로 또는 유사한 광학적 속성을 갖는 광학적 경로가 사용되는 것을 전제로 합니다. 프로세스 환경에서는 다단계 참조 접근 방식이 사용됩니다 *2060 The NIR (참조: 14 페이지, 3.2.2 장)*.

픽셀에서 파장까지

픽셀 단계가 파장 단계로 변경됩니다. 장비는 각 픽셀에 정확한 파장을 할당합니다, 예를 들어. :

픽셀 6 ➔ 파장 1,009.4 nm

각 픽셀의 정확한 파장은 파장 보정에 의해 측정됩니다 *파장 보정 (참조: 11 페이지, 3.1 장)*.

파장 단계를 변경합니다

스펙트럼은 보간법을 통해 표준 파장 척도로 변경됩니다 :

1,000.0 nm, 1,000.5 nm, 1,001.0 nm , …

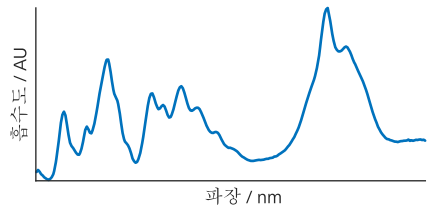


그림 7 파장 단계의 흡수도 스펙트럼.

3 장비설정

다음 단계들은 특정 샘플 측정 점검을 갖는 특정 샘플에 대해 동일한 스펙트럼이 얻어지도록 보장하는데, 이는 제품군 **OMNIS NIR Analyzer** 또는 다른 **2060 The NIR** 타입의 장비와 동일한 또는 다른 장비로 기록되었는지 여부에 관계없이 동일한 시간 및 동일한 또는 다른 장비로 기록했습니다.

스펙트럼의 x축과 y축을 모두 주의해야 합니다 :

- **x축**: 파장 보정 *파장 보정 (참조: 11 페이지, 3.1 장)*
- **y축**: 표준물질을 통한 점검 *표준물질을 통한 점검 (참조: 12 페이지, 3.2 장)*

또한 장비 성능이 요구 사항, 즉 요구 사항을 충족하는지 점검해야 합니다. :

- 장비로 스펙트럼을 측정하기 전에 **장비 성능 테스트** 성공적으로 수행되어야 합니다 *장비 성능 테스트 (참조: 22 페이지, 3.3 장)*.

표준

이 장비는 파장 보정 및 장비 성능 테스트에 대해 계측학적으로 추적 가능한 내부 **파장 표준** 을 사용합니다.

반사 모드를 사용할 때는 장비 타입에 따라 표준물질을 통한 점검 및 장비 성능 테스트에 대한 **반사 표준** 도 필요합니다 :

- **OMNIS NIR Analyzer**: 내부 반사 표준
- **2060 The NIR**: 외부 반사 표준

3.1 파장 보정

파장 보정은 파장 값, 즉 스펙트럼의 x축을 표준화합니다. 광센서 어레이의 각 픽셀을 파장에 할당합니다.

파장 보정은 계측학적으로 추적 가능한 내부 파장 표준을 사용합니다. 파장 표준은 정의된 피크 및 알려진 피크 위치를 가진 흡수도 스펙트럼을 가지고 있습니다.

CAL WL 명령은 다음 단계를 수행합니다 :

1. 파장 표준의 흡수도 스펙트럼은 내부 기준 경로를 사용하여 픽셀 단계에 기록됩니다.
2. 기록된 스펙트럼에서 피크 위치는 하위 픽셀 정밀도로 식별됩니다.
3. 픽셀 단계의 측정된 피크 위치와 파장 표준의 공칭 피크 위치를 사용하여 다항식 회귀 분석을 수행합니다.
4. 회귀 다항식은 각 픽셀을 해당 파장에 할당합니다.

회귀 계수가 장비에 저장됩니다 :

- **OMNIS NIR Analyzer**: 각 샘플 발표에 대해 회귀 계수 세트가 장비에 저장됩니다。 즉, OMNIS NIR Analyzer Liquid/Solid의 경우 두 실험 장비에서 파장 보정 및 검증을 수행해야 합니다。
- **2060 The NIR**: 회귀 계수는 장비에 따라 다릅니다。 모든 채널에 대해 동일한 세트가 사용됩니다。

파장 보정의 검증

파장 보정을 수행한 후에는 이 보정을 검증해야 합니다。 **VAL WL** 명령은 다음 단계를 수행합니다：

1. 파장 표준의 흡수도 스펙트럼이 기록됩니다。
2. 파장의 검증：
   a. 삽입된 스펙트럼에서 피크 위치는 파장 단계로 식별됩니다。
   b. 측정된 피크 위치와 알려진 피크 위치 사이의 파장 잔류가 계산됩니다。
   c. 테스트를 통과하려면 각 피크에 대해 파장 잔류가 공차 내에 있어야 합니다。
3. 밴드폭의 검증：
   a. 피크 폭은 삽입 스펙트럼에서 측정됩니다。
   b. 측정된 피크폭과 알려진 피크 폭 사이의 밴드폭 잔류가 계산됩니다。
   c. 테스트를 통과하려면 각 피크에 대해 밴드폭 잔류가 공차 내에 있어야 합니다。
4. 위의 모든 잔류가 공차 내에 있는 경우 검증의 전체 상태가 성공합니다。

장비로 스펙트럼을 측정하기 전에 검증이 성공적으로 수행되어야 합니다。

3.2 표준물질을 통한 점검

표준물질을 통한 점검은 스펙트럼의 y축인 흡수도 값을 표준화합니다。

흡수도의 측정

샘플의 A 흡수도를 계산하려면 신호 S_0 (기준 스캔) 및 S (샘플 스캔)가 필요합니다 *빛이 스펙트럼으로 변환되는 방식 (참조: 7 페이지, 2.3 장)*:

$$A = \log_{10} \frac{S_0}{S}$$

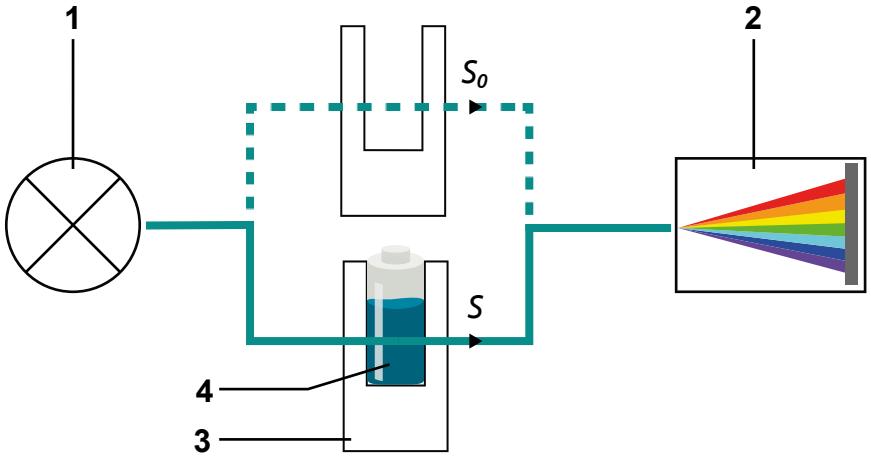


그림 8 전송 모드의 광학 경로 (예를 들어, 액체 샘플 발표).

*그림8*에서 빛은 샘플 홀더(**3**)를 통해 광원(**1**)에서 검출기(**2**)로 전달됩니다.

기준 신호 S_0은 샘플 없이 측정되고, S 신호는 샘플(**4**)을 사용하여 측정됩니다. 그렇지 않을 경우 양측 광학 경로에 대한 광학 속성은 동일하며 양측 신호에 대해 동일한 퍼센트의 댐핑 효과를 발생시킵니다. 이것은 상기 수식에서 결과에 아무런 영향을 미치지 않습니다.

방정식은 S S_0에 대한 참조로 설정합니다. 두 신호가 모두 똑같이 중요합니다. 두 신호 중 하나의 편차는 다른 흡수도 값과 궁극적으로 다른 스펙트럼으로 이어집니다.

S 및 S_0 모두 장비 및 주변 조건의 변동에 의해 영향을 받습니다. 이러한 영향이 서로 상쇄되도록 하려면 두 신호를 짧은 시간 내에 측정해야 합니다.

이 원칙의 구현은 장비 타입에 따라 달라집니다 :

- **OMNIS NIR Analyzer**
  샘플의 흡수 스펙트럼은 S 신호 및 S_0 신호를 통해 계산됩니다*OMNIS NIR Analyzer (참조: 13페이지, 3.2.1장)*.
- **2060 The NIR**
  프로세스 환경에서는 S 및 S_0의 측정에 동일한 광학 경로를 사용할 수 없습니다. 그래서 더 많은 조치를 필요합니다 *2060 The NIR (참조: 14페이지, 3.2.2장)*.

3.2.1 OMNIS NIR Analyzer

표준물질을 통한 점검은 신호 S_0과 S 측정하고 A 흡수도를 계산하여 수행됩니다.

샘플의 스펙트럼을 삽입합니다

실험 장비로 스펙트럼을 삽입할 수 있기 전에, 장비 성능 테스트 실험 장비에 대해 성공적으로 수행해야 합니다 *장비 성능 테스트 (참조: 22페이지, 3.3장)*.

1. 샘플은 샘플 발표에서 준비해야 합니다.
2. 흡수도 스펙트럼은 가장 최근에 기록된 기준 스펙트럼 S_0으로 계산됩니다. S_0의 현재 값을 얻으려면 **MEAS REF SPEC** 명령을 실행할 수 있습니다.
   고체 샘플 발표를 사용할 때 장비는 광학 경로에 반사 표준을 자동으로 삽입합니다. 이 반사 표준은 신호 S_0을 수정할 필요가 없습니다.
3. **MEAS SPEC** 명령을 사용하여 샘플을 측정합니다. 그러면 S 신호가 생성됩니다.
4. 소프트웨어가 샘플의 흡수도로서 A을 계산합니다 :

   $$A = \log_{10} \frac{S_0}{S}$$

   여기에서 S_0은 기준 경로에서 측정된 신호에 해당하며 S은 샘플을 통해 측정된 신호에 해당합니다.

3.2.2 2060 The NIR

타입 **2060 The NIR**의 장비는 외부 표준물질을 통한 점검을 필요합니다.

외부 참조 표준화

동일한 광학 속성을 갖는 광 경로를 통해 신호 S (샘플 포함) 및 신호 S_0 (샘플 포함되지 않음)을 반복적으로 측정하는 것은 시간이 많이 걸리고 오류가 발생하기 쉽습니다.

따라서 두 개의 시각로가 추가로 도입됩니다 *(참조: 14페이지, 그림 9)*:

- 장비의 **내부 참조**. 내부 참조 경로는 쉽게 측정할 수 있는 S_{ref} 신호를 제공합니다.
- 광섬유가 **보정 장비** 에 연결되는 또 다른 외부 광학 경로입니다. 이 광학 경로는 신호 S_{fiber}를 제공합니다.

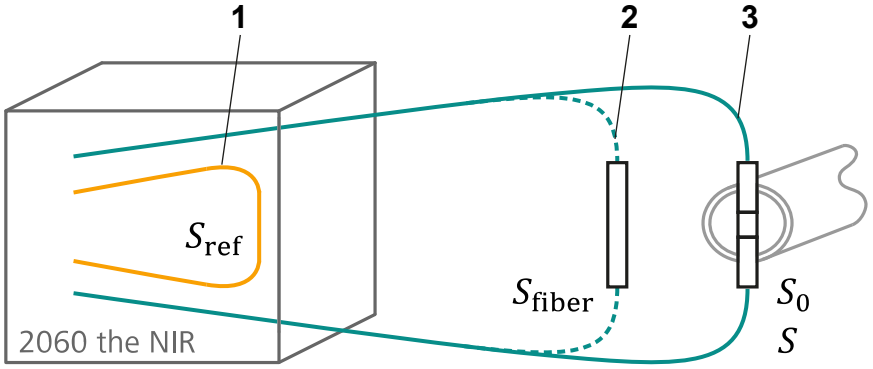


그림 9 전송 모드의 예로서 광학 경로: 내부 참조 경로(1)와 보정 장비(2)에 연결된 외부 광섬유 및 포함되거나 포함되지 않은 샘플(3)에 연결된 존데를 포함된 외부 광섬유. 광학 경로 2와 3은 서로 다른 방식으로 연결된 동일한 광섬유를 나타냅니다.

보정 장비는 광섬유를 고정하여 (**2**) 기준 경로를 형성합니다. 전송 모드에서 공기는 기준 역할을 하고 빛의 100%를 전송합니다. 반사 모드에서는

보정 장비도 반사 표준을 삽입합니다. 첫째, 빛의 100%를 반영하는 이상적인 반사 표준이 가정됩니다.

샘플의 A 흡수도는 신호 S_0 및 S에서 계산됩니다. 분자 및 분모에 두 개의 추가 신호 S_{ref} 및 S_{fiber}가 추가되면 결과는 변경되지 않습니다 :

$$A = \log_{10}\frac{S_0}{S} = \log_{10}\left(\frac{S_{ref}}{S}\cdot\frac{S_{fiber}}{S_{ref}}\cdot\frac{S_0}{S_{fiber}}\right)$$

이 방정식은 그것으로 변경할 수 있습니다 :

$$A = \log_{10}\left(\frac{S_{ref}}{S}\right) - \log_{10}\left(\frac{S_{ref}}{S_{fiber}}\right) - \log_{10}\left(\frac{S_{fiber}}{S_0}\right)$$

3개의 항은 흡수도 값을 나타내고 다음과 같이 설명할 수 있습니다 :

$$A = A_{total} - A_{fiber} - A_{window}$$

*그림10*은 S_{ref}, S_{fiber}, S_0 및 S 신호를 측정하는 방법을 보여줍니다.

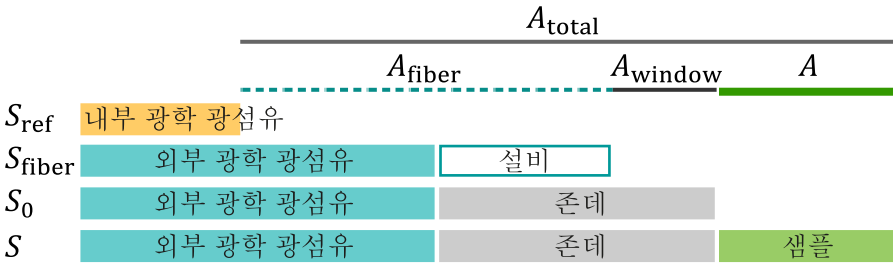


그림 10 외부 참조 표준화

A_{total}은 내부 광섬유와 관련된 외부 광섬유, 존데 및 샘플의 흡수도입니다.

A_{fiber}는 내부 광섬유와 관련된 외부 광섬유 또 보정 장비의 흡수도입니다.

A_{window}는 보정 장비를 뺀 존데의 흡수도입니다.

주변 변동의 제거

샘플의 A 흡수도를 측정하기에 대해 3개의 흡수도 값을 측정합니다. 위의 방정식에 따라 A 3 값이 계산됩니다 :

$$A = A_{total} - A_{fiber} - A_{window}$$

이를 통해 3가지 흡수도 값을 서로 다른 시간에 측정할 수 있습니다. 이렇게 하는 경우에 3 가지 측정 각각에 대해 장비 내에 변동이나 주변 조건의 변동을 쉽게 제거할 수 있습니다 :

- A_{total}은 각 샘플 측정에 대해 점검됩니다. 변동성을 제거하기에 대해 짧은 시간 내에 S_{ref} 및 S 측정해야 합니다.
- **광섬유 수정** A_{fiber}는 덜 자주 점검할 수 있습니다. 변동성을 제거하기에 대해 짧은 시간 내에 S_{ref} 및 S_{fiber}를 측정해야 합니다.

- **창 수정** A_{window}는 일반적으로 설치 후 한 번만 점검할 수 있습니다. 변동성을 제거하기에 대해 짧은 시간 내에 S_{fiber} 및 S_0을 측정해야 합니다.

창 수정은 언제 필요합니까 ?

보정 장비가 존데의 광학 속성을 완전히 재현하는 경우 A_{window}는 0입니다. 이 경우에 창 수정은 건너뛸 수 있습니다.

일반적으로 전송 모드에서는 창 수정이 필요합니다. 반사 모드에 대한 창 수정이 필요하지 않습니다. 그러나 다음 표에 나열된 예외가 있습니다 :

측정 모드	존데	표준물질을 통한 점검
투과	투과 쌍	유리 섬유 + 창
	투과 존데	
	단일 섬유가 포함된 투과 존데	
반사	반사 존데	유리 섬유
	MicroBundle이 포함된 단일 섬유가	유리 섬유 + 창

창 수정이 필요한지 여부를 측정하려면 보정 장비와 존데의 각 조합을 개별적으로 검사해야 합니다. 보정 장비가 존데의 광학 속성을 완전히 반영하지 *않는* 경우 창 수정이 필요합니다.

채널

하나 타입의 **2060 The NIR** 장비는 여러 채널을 제공합니다. 각 채널은 다른 광섬유 및 존데 구성에 연결할 수 있습니다. 따라서 표준물질을 통한 점검은 각 채널에 대해 별도로 수행해야 합니다.

모든 채널은 동일한 내부 참조 경로를 사용하고 즉, 동일한 신호 S_{ref}입니다. 멀티플렉서는 내부 참조와 다양한 측정 채널 사이에서 전환됩니다.

광섬유 수정을 수행합니다

최초 시운전 후에 또는 채널의 광섬유 구성이 변경된 경우에 광섬유 수정을 수행해야 합니다. 램프를 변경하거나 주변 조건을 극단적으로 변경하는 경우에 다시 표준화하는 것이 좋습니다.

이 처리에서 참조 자료가 사용됩니다 :

- 반사 모드에서는 참조 자료가 반사 표준입니다. 그것은 이상적인 반사 기준이 아니라고 가정합니다 (예를 들어 99 %). 반사 표준은 알려진 공칭 흡수도 스펙트럼 $A_{nominal}$을 가지고 있습니다.
- 전송 모드에서 공기는 기준 역할합니다. 공칭 흡수도 스펙트럼은 공기가 빛을 흡수하지 않는다고 가정하기 때문에 0선 ($A_{nominal}$ = 0)입니다.

*그림11*은 다음 절차가 나와 있습니다 :

1. 외부 광섬유는 보정 장비에 연결해야 합니다.
   반사모드에서 보정 장비는 반사 표준과 통합합니다.
2. **유리 섬유** 인터페이스가 있는 **REF STD** 명령은 다음 검색을 수행합니다 :
   a. 내부 참조 스캔은 S_{ref}에 대한 값을 반환합니다.
   b. 외부 스캔은 외부 광섬유, 보정 장비 및 참조 자료를 측정합니다. 그러면 S_{raw} 신호가 생성됩니다.
3. 소프트웨어가 A_{raw} (**측정된 원시 스펙트럼**) 계산됩니다 :

   $$A_{\text{raw}} = \log_{10} \frac{S_{\text{ref}}}{S_{\text{raw}}}$$

   여기서 A_{raw}는 내부 광학 경로와 관련된 외부 광섬유, 보정 장비 및 참조 자료의 흡수도에 해당합니다.
4. 참조 자료의 공칭 흡수도 스펙트럼, $A_{nominal}$은, 소프트웨어에 **기준 스펙트럼** 표시됩니다. 기준 스펙트럼은 A_{fiber}를 얻으려면 A_{raw}에서 빼야 합니다 :

   $$A_{\text{fiber}} = A_{\text{raw}} - A_{\text{nominal}}$$

   여기서 A_{fiber}는 내부 광학 경로와 관련된 광섬유 및 보정 장비의 흡수도에 해당합니다.
   주의상황: 전송 모드에서는 $A_{fiber} = 0$ 및 $A_{nominal} = A_{raw}$.
   A_{fiber}는 **수정 스펙트럼** 광섬유를 나타냅니다.
5. A_{fiber} 각 채널에 대해 광섬유 수정을 다시 수행할 때까지 변경되지 않은 상태로 유지됩니다.

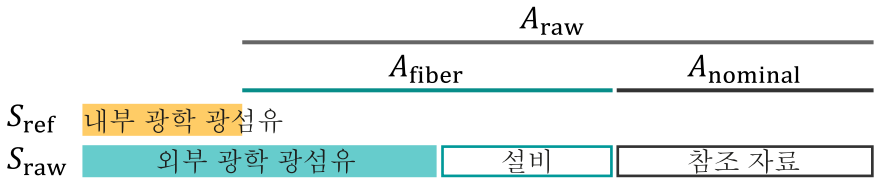


그림 11 광섬유 수정

광섬유 수정을 검증합니다

광섬유 수정은 동일한 측정 parameter와 동일한 보정 장비로 검증해야 합니다.

이 처리에서 참조 자료가 사용됩니다 :

- 반사 모드에서는 참조 자료가 반사 표준입니다. 그것은 이상적인 반사 기준이 아니라고 가정합니다 (예를 들어 99 %). 반사 표준은 알려진 공칭 흡수도 스펙트럼 $A_{nominal}$을 가지고 있습니다.
- 전송 모드에서 공기는 기준 역할합니다. 공칭 흡수도 스펙트럼은 공기가 빛을 흡수하지 않는다고 가정하기 때문에 0선 ($A_{nominal} = 0$)입니다.

*그림12*에서 validation residuals가 측정되는 방법을 보여 줍니다 :

1. 외부 광섬유는 보정 장비에 연결해야 합니다.
   반사모드에서 보정 장비는 반사 표준과 통합합니다.

2. **유리 섬유** 인터페이스가 있는 **VAL REF STD** 명령은 다음 검색을 수행합니다 :
   a. 내부 참조 스캔은 S_{ref}에 대한 값을 반환합니다.
   b. 각 채널의 외부 스캔은 외부 광섬유, 보정 장비 및 참조 자료를 측정합니다. 그러면 S_{raw} 신호가 생성됩니다.
3. 소프트웨어가 A_{raw} (**측정된 원시 스펙트럼**) 계산됩니다 :
   $$A_{raw} = \log_{10} \frac{S_{ref}}{S_{raw}}$$
4. A_{raw}는 광섬유 및 보정 장비의 흡수도를 제거하기에 대한 유리 섬유 수정 스펙트럼에 의해 수정됩니다 :
   $$A_{corrected} = A_{raw} - A_{fiber}$$
   $A_{corrected}$ 소프트웨어에 **측정되고 수정된 스펙트럼** 표시됩니다.
5. 이상적으로는 $A_{corrected}$ 및 **기준 스펙트럼** $A_{nominal}$ 일치해야 합니다. 둘 사이의 차이는 **Validation residuals** 계산됩니다 :
   $$A_{residual} = A_{corrected} - A_{nominal}$$
   주의상황: 전송 모드에서는 $A_{nominal}$ = 0 및 $A_{residual}$ = $A_{corrected}$.

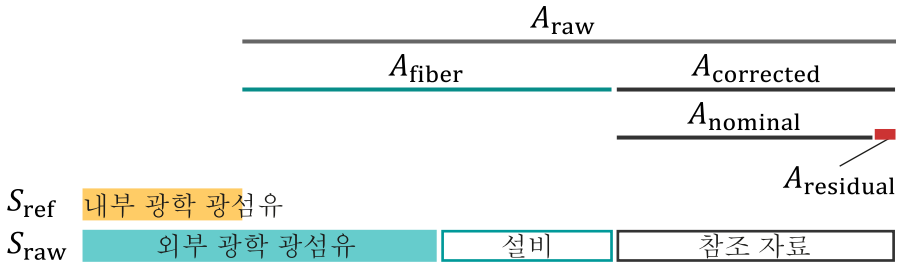


그림 12 광섬유 수정의 유효성을 검증에 대한 잔류

Validation residuals를 조사하기에 대해 파장범위가 여러 세그먼트로 분할됩니다. 각 세그먼트에 대해 파장에 대한 잔류의 제곱 평균값은 **RMS 잔류**(단위: mAU)를 생성합니다 :

$$A_{RMS} = \sqrt{\frac{\sum_{i=1}^{f} (A_{residual_i})^2}{f}}$$

여기서 f는 세그먼트 내 파장의 수이고 $A_{residual\ i}$는 파장 i의 잔류입니다.

각 세그먼트는 A_{RMS}에 대해 미리 정의된 공차를 준수해야 합니다. 모든 세그먼트의 공차를 준수하는 경우 전체 검증이 성공합니다.

각 채널의 스펙트럼을 장비로 측정하기 전에 검증이 성공적으로 수행되어야 합니다.

창 수정 수행

창 수정이 필요한 경우, 최초 시운전 후에 또는 채널의 존데 또는 광섬유 구성을 변경할 때마다 수행해야 합니다. 예를 들어 오염과 같은 존데의 변경은 다시 표준화하는 것이 바람직할 수 있습니다.

이 처리에서 참조 자료가 사용됩니다 :

- 반사 모드에서는 참조 자료가 반사 표준입니다。 그것은 이상적인 반사 기준이 아니라고 가정합니다 (예를 들어 99 %)。 반사 표준은 알려진 공칭 흡수도 스펙트럼 $A_{nominal}$을 가지고 있습니다。
- 전송 모드에서 공기는 기준 역할합니다。 공칭 흡수도 스펙트럼은 공기가 빛을 흡수하지 않는다고 가정하기 때문에 0선 ($A_{nominal}$ = 0)입니다。

*그림13*은 다음 절차가 나와 있습니다：

1. A_{fiber}에 대한 현재 값을 얻으려면 위에서 설명한 대로 광섬유 수정을 수행해야 합니다。
   주의사항: 창 수정 전에 광섬유 수정을 수행해야 합니다。
2. 그런 다음 외부 광섬유를 샘플 없이 존데에 연결해야 합니다。 필요한 경우에 반사 표준은 샘플을 대체합니다。
3. **창** 인터페이스가 있는 **REF STD** 명령은 다음 검색을 수행합니다：
   a. 내부 참조 스캔은 S_{ref}에 대한 값을 반환합니다。
   b. 각 채널의 외부 스캔은 외부 광섬유, 존데 및 참조 자료를 측정합니다。 그러면 S_{probe} 신호가 생성됩니다。
4. 흡수도 A_{probe}은 내부 광학 경로와 관련이 있습니다：
   $$A_{\text{probe}} = \log_{10} \frac{S_{\text{ref}}}{S_{\text{probe}}}$$
5. 소프트웨어가 A_{raw} (**측정된 원시 스펙트럼**) 계산됩니다：
   $$A_{\text{raw}} = A_{\text{probe}} - A_{\text{fiber}}$$
   여기서 A_{raw}는 보정 장비와 관련된 존데 및 참조 자료의 흡수도에 해당합니다。
6. 참조 자료의 공칭 흡수도 스펙트럼, $A_{nominal}$은, 소프트웨어에 **기준 스펙트럼** 표시됩니다。 기준 스펙트럼은 A_{raw}를 얻으려면 A_{window}에서 빼야 합니다：
   $$A_{\text{window}} = A_{\text{raw}} - A_{\text{nominal}}$$
   여기서 A_{window}는 보정 장비와 관련된 존데의 흡수도에 해당합니다。
   주의상황: 전송 모드에서는 $A_{nominal}$ = 0 및 A_{window} = A_{raw}。
   A_{window}는 **수정 스펙트럼** 창을 나타냅니다。
7. A_{window} 각 채널에 대해 창 수정을 다시 수행할 때까지 변경되지 않은 상태로 유지됩니다。

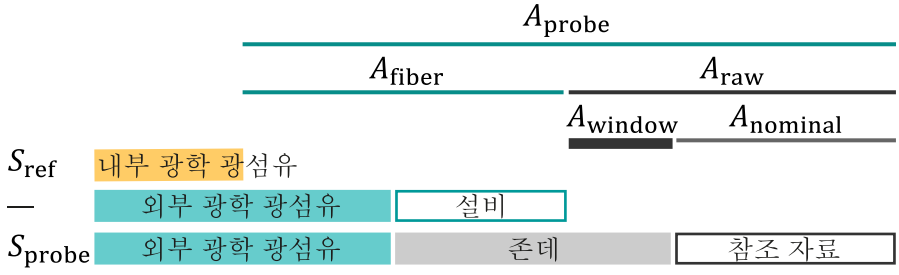


그림 13 창 수정

창 수정의 검증

창 수정은 동일한 측정 parameter와 동일한 보정 장비로 검증해야 합니다。

이 처리에서 참조 자료가 사용됩니다 :

- 반사 모드에서는 참조 자료가 반사 표준입니다。 그것은 이상적인 반사 기준이 아니라고 가정합니다 (예를 들어 99 %)。 반사 표준은 알려진 공칭 흡수도 스펙트럼 $A_{nominal}$을 가지고 있습니다。
- 전송 모드에서 공기는 기준 역할합니다。 공칭 흡수도 스펙트럼은 공기가 빛을 흡수하지 않는다고 가정하기 때문에 0선 ($A_{nominal}$ = 0)입니다。

*그림14*에서 validation residuals가 측정되는 방법을 보여 줍니다 :

1. 외부 광섬유를 샘플 없이 존데에 연결해야 합니다。 필요한 경우에 반사 표준은 샘플을 대체합니다。
2. **창** 인터페이스가 있는 **VAL REF STD** 명령은 다음 검색을 수행합니다 :
   a. 내부 참조 스캔은 S_{ref}에 대한 값을 반환합니다。
   b. 각 채널의 외부 스캔은 외부 광섬유, 존데 및 참조 자료를 측정합니다。 그러면 S_{probe} 신호가 생성됩니다。
3. 흡수도 A_{probe}은 내부 광학 경로와 관련이 있습니다 :
   $$A_{\mathrm{probe}} = \log_{10} \frac{S_{\mathrm{ref}}}{S_{\mathrm{probe}}}$$
4. 소프트웨어가 A_{raw} (**측정된 원시 스펙트럼**) 계산됩니다 :
   $$A_{\mathrm{raw}} = A_{\mathrm{probe}} - A_{\mathrm{fiber}}$$
5. 창 수정 스펙트럼을 A_{raw}에서 빼는 경우 보정 장비와 존데 간의 흡수도 차이가 제거됩니다 :
   $$A_{\mathrm{corrected}} = A_{\mathrm{raw}} - A_{\mathrm{window}}$$
   $A_{corrected}$ 소프트웨어에 **측정되고 수정된 스펙트럼** 표시됩니다。
6. 이상적으로는 $A_{corrected}$ 및 **기준 스펙트럼** $A_{nominal}$ 일치해야 합니다。 둘 사이의 차이는 **Validation residuals** 계산됩니다 :
   $$A_{\mathrm{residual}} = A_{\mathrm{corrected}} - A_{\mathrm{nominal}}$$
   주의상황: 전송 모드에서는 $A_{nominal}$ = 0 및 $A_{residual}$ = $A_{corrected}$。

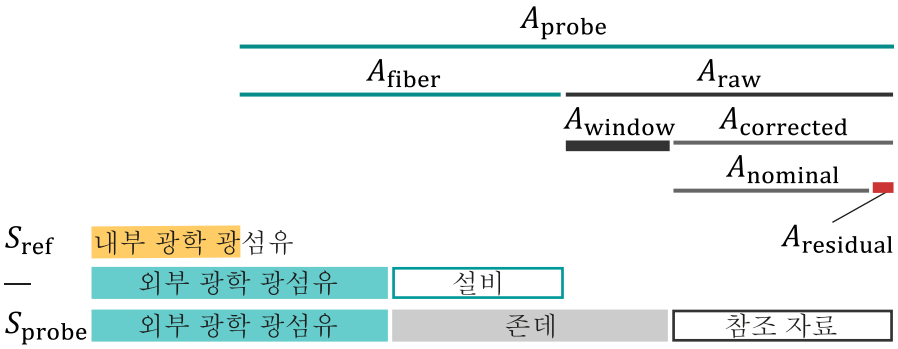


그림 14 창 수정의 유효성을 검증에 대한 잔류

Validation residuals를 조사하기에 대해 파장범위가 여러 세그먼트로 분할됩니다。 각 세그먼트에 대해 파장에 대한 잔류의 제곱 평균값은 **RMS 노이즈**(단위: mAU)를 생성합니다 :

$$A_{\mathrm{RMS}} = \sqrt{\frac{\sum_{i=1}^{f}(A_{\mathrm{residual}_i})^2}{f}}$$

여기서 f는 세그먼트 내 파장의 수이고 $A_{\mathrm{residual}\ i}$는 파장 i의 잔류입니다.

각 세그먼트는 A_{RMS}에 대해 미리 정의된 공차를 준수해야 합니다. 모든 세그먼트의 공차를 준수하는 경우 전체 검증이 성공합니다.

샘플의 스펙트럼을 삽입합니다

장비에 스펙트럼을 삽입하기 전에 각 채널에서 장비 성능 테스트 성공적으로 수행해야 합니다 *장비 성능 테스트 (참조: 22페이지, 3.3 장)*.

샘플의 스펙트럼을 삽입하는 절차는 *그림15*에 나와 있습니다 :

1. 외부 광섬유는 존데에 연결해야 합니다. 한 샘플이 있어야 합니다.
2. 흡수도 스펙트럼은 가장 최근에 기록된 기준 스펙트럼 S_{ref}으로 계산됩니다. S_{ref}의 현재 값을 얻으려면 **MEAS REF SPEC** 명령을 실행할 수 있습니다.
3. **MEAS SPEC** 명령은 존데 및 광섬유를 포함하여 샘플을 측정합니다. 그러면 S 신호가 생성됩니다.
4. 소프트웨어는 A_{total}을 계산하고, 내부 광학 경로와 관련된 존데 및 광섬유 등 샘플의 흡수도 계산합니다 :

   $$A_{\mathrm{total}} = \log_{10}\frac{S_{\mathrm{ref}}}{S}$$

5. 그런 다음에 각 채널의 광섬유 수정 스펙트럼 A_{fiber} 및 창 수정 스펙트럼 A_{window}를 사용하여 위에서 설명한 대로 샘플의 흡수도를 계산합니다 :

   $$A = A_{\mathrm{total}} - A_{\mathrm{fiber}} - A_{\mathrm{window}}$$

   A 샘플의 스펙트럼을 나타냅니다.

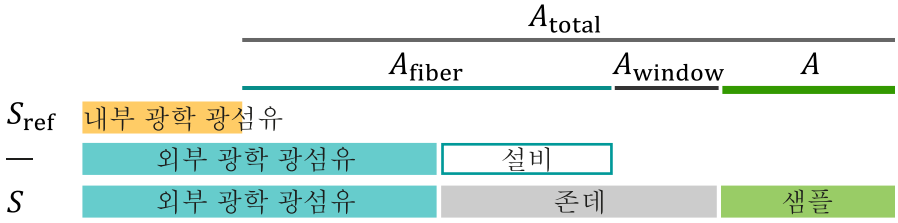


그림 15 샘플의 스펙트럼 기록

3.3 장비 성능 테스트

장비 성능 테스트은 내부 광학 경로 및 외부 광학 경로를 통해 안내할 수 있습니다.

- **OMNIS NIR Analyzer**
  - **내부 장비 성능 테스트** (기본): 내부 테스트에서는 각 샘플 발표의 기준 경로가 사용됩니다. 이 테스트에서는 파장 및 신호 노이즈가 테스트됩니다.
    테스트 전에, 파장 보정은 각 샘플 발표에 대해 성공적으로 수행되고 검증해야 합니다.
    장비를 이용해 각 샘플 발표에서 스펙트럼을 기록하기 전에 내부 테스트가 성공적으로 수행되어야 합니다.
  - **외부 장비 성능 테스트** (옵션): 외부 테스트는 USP <856>, Ph.Eur 2.2.40 및 JP 2.27과 같은 약전에 따른 검증을 지원합니다. 파장, 신호 노이즈 및 측광 선형성이 테스트됩니다*외부 장비 성능 테스트 (OMNIS NIR Analyzer) (참조: 24페이지, 3.3.1 장)*.
- **2060 The NIR**
  이 테스트에서는 내부 광학 경로 또는 외부 광학 경로 중 하나를 사용할 수 있습니다. 이 테스트에서는 파장 및 신호 노이즈가 테스트됩니다.
  테스트 전에 파장 보정 및 외부 기준 표준화는 각 채널에서 성공적으로 수행되고 검증해야 합니다.
  장비를 이용해 각 채널에서 스펙트럼을 기록하기 전에 장비 성능 테스트(이)가 성공적으로 수행되어야 합니다. 사전에 정의된 공차를 준수해야 합니다. 공차는 장비별 데이터(측정 모드, 섬유 타입 및 섬유 길이)에 지정된 각 채널의 광섬유 구성에 따라 달라집니다.

파장 테스트

파장 테스트는 파장 정확도와 파장 정밀도를 시험합니다. 이를 위해, 정의된 피크 및 알려진 피크 위치의 흡수도 스펙트럼을 갖는 파장 표준이 사용됩니다 :

- **내부**: 계측학적으로 추적 가능한 내부 파장 표준의 흡수도 스펙트럼은 내부 광학 경로를 통해 측정됩니다 :

  $$A_{\mathrm{WL}} = \log_{10}\left(\frac{S_{\mathrm{ref}}}{S_{\mathrm{ref,WL}}}\right)$$

  여기에서 A_{WL}은 파장 표준의 흡수도에 해당하며, S_{ref}는 내부 기준 경로에서 측정된 신호에 해당하고 $S_{\mathrm{ref,WL}}$은 내부 파장 표준을 갖는 내부 기준 경로에서 측정된 신호에 해당합니다.
- **외부**, 제품군 **OMNIS NIR Analyzer**의 장비에 적용됨 : 외부 파장 테스트는 옵션에 해당합니다*(참조: 24페이지, "외부 파장 테스트")*.

- **외부**, 타입 **2060 The NIR**의 장비에 적용됨 :
  외부 광섬유는 단일 섬유의 경우 보정 장치에 연결되어 있어야 하며 MicroBundle의 경우 반사 표준에 연결되어 있어야 합니다.
  계측학적으로 추적 가능한 내부 파장 표준의 흡수도 스펙트럼은 외부 광학 경로를 통해 측정됩니다 :

  $$A_{\mathrm{WL}} = \log_{10}\left(\frac{S_{\mathrm{ref}}}{S_{\mathrm{fiber,WL}}}\right) - A_{\mathrm{fiber}}$$

  여기에서 A_{WL}은 내부 파장 표준의 흡수도에 해당하며, S_{ref}는 내부 기준 경로에서 측정된 신호에 해당하고, $S_{\mathrm{fiber,WL}}$은 각 채널의 광학 경로에서 측정된 신호에 해당하며, 이때 섬유는 보정 장치 또는 반사 표준에 연결되고 내부 파장 표준은 광학 경로에 사용되며, A_{fiber}는 표준물질을 통한 점검에서 광섬유 수정 스펙트럼에 해당합니다.

A_{WL}은 반사 표준의 흡수도도 포함하는데, 이것은 다음 계산에서 중요하지 않습니다. 이상적인 경우 A_{WL}의 피크 위치는 파장 표준의 공칭 피크 위치입니다.

파장 정확도 및 파장 정밀도는 다음과 같이 테스트됩니다 :

1. 위에서 설명한 바와 같이 파장 표준의 일련의 흡수도 스펙트럼이 기록됩니다(A_{WL}).
2. 기록된 스펙트럼에서 피크 위치가 식별됩니다.
3. 각 피크 위치에 대해 다음과 같은 통계가 기록된 스펙트럼을 통해 계산됩니다 :
   a. 평균값 (단위: nm)
   b. 표준편차 (단위 nm)
4. **정확성**: 각 피크에서 평균 피크 위치와 공칭 피크 위치 사이의 차이는 미리 정의된 공차 내에 있어야 합니다.
   ➨ **주의상황**: 피크 위치는 테스트마다 약간 다를 수 있습니다. 이는 피크 위치가 온도 보정되기 때문입니다. 파장 보정에서도 유사한 온도 보정이 수행됩니다. 특정 온도에서 측정하는 경우 모든 장비에서 유사한 결과를 얻을 수 있습니다.
5. **정밀도**: 각 피크에 대해 표준편차는 미리 정의된 공차 내에 있어야 합니다.
6. 모든 피크의 공차를 준수하는 경우 파장시험의 전반적인 상태가 성공적입니다.

노이즈 테스트

신호 노이즈는 내부 또는 외부적으로 테스트할 수 있습니다 :

- **내부** : 노이즈는, 동일한 광학 경로에서 다른 측정의 흡수도를 기준으로 하는 내부 광학 경로의 흡수도로서 측정됩니다 :

  $$A_{\mathrm{noise}} = \log_{10}\left(\frac{S_{\mathrm{ref},1}}{S_{\mathrm{ref},2}}\right)$$

  여기에서 $S_{\mathrm{ref},1}$ 및 $S_{\mathrm{ref},2}$는 내부 기준 경로에서 측정된 신호에 해당합니다.

- **외부**, 제품군 **OMNIS NIR Analyzer**의 장비에 적용됨 : 외부 노이즈 테스트는 옵션에 해당합니다*(참조: 25 페이지, "외부 노이즈 테스트")*。
- **외부**, 타입 **2060 The NIR**의 장비에 적용됨 :
  외부 광섬유는 단일 섬유의 경우 보정 장치에 연결되어 있어야 하며 MicroBundle의 경우 반사 표준에 연결되어 있어야 합니다。
  노이즈는 측정된 흡수도 스펙트럼과 공칭 흡수도 스펙트럼 사이의 편차로서 측정됩니다 :

  $$A_{\text{noise}} = \log_{10}\left(\frac{S_{\text{ref}}}{S_{\text{fiber}}}\right) - A_{\text{fiber}} - A_{\text{nominal}}$$

  여기에서 S_{ref}는 내부 기준 경로에서 측정된 신호에 해당하고, S_{fiber}는 외부 광학 경로에서 측정된 신호에 해당하며, 이때 섬유는 보정 장치 또는 반사 표준에 연결되고, A_{fiber}는 표준물질을 통한 점검의 광섬유 보정 스펙트럼에 해당하고 A_{nominal}은 반사 표준의 공칭 흡수도 스펙트럼에 해당합니다。
  주의상황: 전송 모드에서는 $A_{\text{nominal}} = 0$。

이상적인 경우 $A_{\text{noise}} = 0$。

노이즈 테스트에서는 다음 단계가 수행됩니다 :

1. 일련의 노이즈 스펙트럼은 위에서 설명한 바와 같이 기록됩니다 (A_{noise})。
2. 이 노이즈 스펙트럼은 다양한 파장 세그먼트로 구분됩니다。
3. 각 노이즈 스펙트럼 및 각 세그먼트에 대해 3개의 값이 계산됩니다 :
   a. 측광 노이즈 (단위: mAU)
   b. 피크 대 피크 노이즈 (단위: mAU)
   c. 노이즈의 바탕선 바이어스 (단위: mAU)
4. 각 세그먼트에서 3개의 각 값에 대해 평균값이 기록된 노이즈 스펙트럼을 통해 계산됩니다。
5. 모든 평균값이 사전 정의된 공차 내에 있는 경우 노이즈 테스트의 전체 상태가 성공적입니다。

3.3.1 외부 장비 성능 테스트 (OMNIS NIR Analyzer)

제품군 **OMNIS NIR Analyzer**의 장비는 USP <856>, Ph.Eur 2.2.40 및 JP 2.27과 같은 약전에 따라 검증할 수 있습니다(ab OMNIS Software 버전 4.4 이상에 적용됨)。 이 테스트에서는 계측학적으로 추적 가능한 외부 기준 표준이 요구됩니다。 기준 표준은 주변 온도에서 기준 장비로 측정된 개별 공칭 흡수도 스펙트럼을 갖습니다。

외부 파장 테스트

파장 정확도 및 파장 정밀도는 다음과 같이 테스트됩니다 :

1. 외부 파장 표준(투과 또는 반사)은 해당 위치로 이동시켜야 합니다。

2. 외부 파장 표준의 일련의 흡수도 스펙트럼이 기록됩니다 :

$$A_{\mathrm{WL}} = \log_{10}\left(\frac{S_{\mathrm{ref}}}{S_{\mathrm{WL}}}\right)$$

여기에서 A_{WL}은 외부 파장 표준의 흡수도에 해당하며, S_{ref}는 내부 기준 경로에서 측정된 신호에 해당하고 $S_{\mathrm{ref,WL}}$은 외부 파장 표준을 통해 측정된 신호에 해당합니다.

3. 기록된 스펙트럼에서 피크 위치가 식별됩니다.
4. 각 피크 위치에 대해 다음과 같은 통계가 기록된 스펙트럼을 통해 계산됩니다 :
    a. 평균값 (단위: nm)
    b. 표준편차 (단위 nm)
5. **정확성**: 각 피크에서 평균 피크 위치와 공칭 피크 위치 사이의 차이는 미리 정의된 공차 내에 있어야 합니다.
6. **정밀도**: 각 피크에 대해 표준편차는 미리 정의된 공차 내에 있어야 합니다.
7. 모든 피크의 공차를 준수하는 경우 파장시험의 전반적인 상태가 성공적입니다.

외부 노이즈 테스트

이 신호 노이즈는 낮은 포톤 플럭스(로우 플럭스)에서 한 번 그리고 높은 포톤 플럭스(하이 플럭스)에서 한 번 테스트됩니다 :

1. 저 플럭스 테스트 또는 고온 플럭스 테스트를 위한 외부 기준 표준(투과 또는 반사)은 해당 위치로 이동시켜야 합니다.
2. 일련의 노이즈 스펙트럼은 기준 표준의 공칭 흡수도 스펙터럼과 측정된 흡수도 스펙트럼 사이의 편차로서 기록됩니다 :

$$A_{\mathrm{noise}} = \log_{10}\left(\frac{S_{\mathrm{ref}}}{S_{\mathrm{ND}}}\right) - A_{\mathrm{nominal}}$$

여기에서 S_{ref}는 내부 기준 경로에서 측정된 신호에 해당하고, S_{ND}는 외부 기준 표준을 통해 측정된 신호에 해당하며 A_{nominal}은 기준 표준의 공칭 스펙트럼에 해당합니다.

3. 이 노이즈 스펙트럼은 다양한 파장 세그먼트로 구분됩니다.
4. 각 노이즈 스펙트럼 및 각 세그먼트에 대해 3개의 값이 계산됩니다 :
    a. 측광 노이즈 (단위: mAU)
    b. 피크 대 피크 노이즈 (단위: mAU)
    c. 노이즈의 바탕선 바이어스 (단위: mAU)
5. 각 세그먼트에서 3개의 각 값에 대해 평균값이 기록된 노이즈 스펙트럼을 통해 계산됩니다.
6. 모든 평균값이 사전 정의된 공차 내에 있는 경우 노이즈 테스트의 전체 상태가 성공적입니다.

측광 선형성

이 테스트의 목적은 전체 파장범위에 걸쳐 반사도(또는 투과도)와 측정된 흡수도 사이에서의 선형 관계를 증명하는 것입니다 :

1. 서로 다른 반사도(또는 투과도)를 갖는 5개의 기준 표준의 흡수도 스펙트럼이 기록됩니다。
2. 반사도(또는 투과도)와 측정된 흡수도 사이의 선형 관계는 다수의 파장에서 선형 회귀를 통해 확인됩니다。
3. 모든 회귀선의 y축 절편 및 기울기가 사전 정의된 공차 내에 있는 경우 테스트의 전체 상태가 성공적입니다。

4 모델 개발

다음과 같은 유형의 모델이 구별됩니다 :

- **정량화 모델** 샘플의 삽입된 스펙트럼에 대한 관심있는 parameter (예를 들어 수분 함량)의 의존성을 설명합니다.
- **식별 모델** (OMNIS Software 버전 4.0부터) 샘플은 흡수된 스펙트럼에 따라 다른 제품(예를 들어 다양한 종류의 커피 콩)으로 분류됩니다. 제품은 특정 화학 물질 또는 특정 물리적 속성(예를 들어 입자 크기)을 가진 특정 화학 물질을 나타냅니다.

알 수 없는 샘플을 분석하기에 대해 샘플 스펙트럼이 삽입됩니다. 사용하기에 따라 스펙트럼은 다음과 같이 사용됩니다 :

- 수량화: 스펙트럼에 기초하여 정량화 모델은 예를 들어 새플의 수분 함량에 대한 예측을 생성됩니다.
- 동일시: 스펙트럼에 기초하여 식별 모델은 샘플을 예를 들어 아라비카 커피콩으로 식별합니다.
- 검증(OMNIS Software 버전 4.2): 스펙트럼을 근거로 예를 들어 식별 모델은 샘플이 아라비카 커피인지 아닌지 여부를 검증합니다.

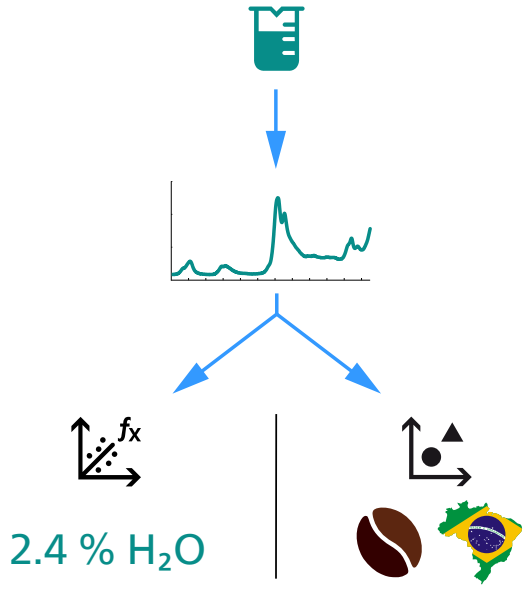


그림 16 수량화 (왼쪽 아래) 및 동일시 (오른쪽 아래).

모델 개발 및 샘플 분석에는 다음 단계가 포함됩니다 :

1. **샘플 추출**
   물리적 샘플을 채취하여 처리합니다 :
   a. 각 샘플에 대해 스펙트럼이 삽입됩니다。
   b. 수량화에 대해 관심있는 parameter(예를 들어 수분 함량)에 대한 기준값은 기준 방법(예를 들어 적정)을 사용하여 측정됩니다。 기준 측정은 정확하고 자세히 수행해야 합니다。
   c. 동일시에 대해 샘플의 제품 제휴를 알아야 합니다。
2. **모델 개발**
   모델 개발은 다음 단계를 포함하는 반복 과정에서 수행됩니다 :
   a. 데이터 레코드의 분할이 보정 데이터 세트, 유효성 검사 데이터 세트 및 특이값 데이터 세트로 수행합니다。
   b. 스펙트럼에 적절한 데이터 전처리 및 파장범위에 대한 적용。
   c. 보정 데이터 세트의 근거로 한 모델의 계산。
   d. 모델의 검증은 모델이 요구 사항을 충족하는지를 확인합니다。 검증은 주로 모델 개발에 사용되지 않은 유효성 검사 데이터 세트를 기반으로 합니다。
   수량화할 때 모델은 유효성 검사 데이터 세트의 스펙트럼에 대해 관심있는 parameter를 예측합니다。 그런 다음 계산된 값을 알려진 기준값과 비교합니다。
   동일시에 대해 모델은 스펙트럼을 서로 다른 제품에 할당합니다。 예측된 제품 제휴는 각각의 실제 제품 제휴와 비교됩니다。
3. **모니터링**
   모델을 모니터링하는 경우 시간이 지남에 따라 예측력이 저하되지 않습니다。 프로세스 또는 샘플의 모든 변경에 재검증이 필요합니다。

4.1 물리적 샘플

첫 번째 단계는 물리적 샘플을 수집하고 분석하는 것입니다。 견고한 모델을 개발하기에 대해 적절한 샘플 추출이 전제조건입니다。 몇 가지 측면을 주의해야 합니다。

변동폭

샘플은 미래에 예상되는 전형적인 샘플 변동을 포함해야 합니다。 모든 화학 성분과 입자 크기의 농도는 최소한 예상되는 변동폭을 포함해야 합니다。

샘플은 조건의 적절한 변동과 적절한 기간을 포함해야 합니다。 예를 들어 과정 변동, 계절 변동 또는 주변 조건 변동과 같은 모든 변동에 대해 주의해야 합니다。

샘플은 변동 폭에 걸쳐 고르게 분포되어야 합니다。 수량화에서 기준값의 범위가 포함됩니다。 예를 들어 기준값의 범위가 1%에서 10% 사이인 경우 샘플은 1%에서 10% 사이에 고르게 분포되어야 합니다。

보정 및 검증 샘플

일반적으로 2개의 샘플 세트가 사용됩니다 :

- 보정 세트: 모델 개발에 대해 사용되는 샘플.
- 검증 세트: 모델 검증에 대해 사용된 샘플.

두 세트 모두 예상 변동을 다루어야 합니다. 검증 세트의 경우 모델의 견고성을 테스트하기에 대해 독립적인 샘플을 수집하는 것이 좋습니다. 서로 다른 운영자, 서로 다른 공급업체 또는 다른 장비와 같은 시나리오를 주의해야 합니다.

보정 및 검증을 위해 독립적인 샘플을 수집할 수 없는 경우, 사용 가능한 샘플의 스펙트럼을 보정 데이터 세트와 유효성 검사 데이터 세트로 나눌 수 있습니다. 세트의 독립성을 품질보증하려면 자동 분할 알고리즘을 사용해야 합니다.

일단 모델이 개발되는 경우, 이상값 샘플 테스트하기에 대해 이상값 검출을 주의할 수 있습니다.

모든 샘플은 동일한 방법으로 취급해야 합니다. 스펙트럼을 삽입하기에 대해 동일한 하드웨어 구성과 동일한 측정 parameter를 가진 동일한 method을 사용해야 합니다. 수량화의 경우 동일한 측정 parameter와 함께 동일한 기준 방법을 사용해야 합니다.

시료 수량

조건, 화학 성분 또는 입자 크기의 변동이 많을수록 더 많은 샘플이 필요합니다.

수량화: 통계 분석이 제대로 작동하려면 최소 50개의 샘플이 필요하고, 보정 세트와 검증 세트에는 각각 최소 20개에서 25개의 샘플이 포함해야 합니다.

동일시 및 검증; 각 제품에 대해 샘플은 예상 변동을 포함해야 합니다. 제품의 샘플 개수는 다를 수 있고 최소 샘플 수는 3개입니다.

최소 10~20개의 샘플(변화 수에 따라 다름)을 사용하여 검증 세트 없이 첫 번째 모델을 개발할 수 있습니다. 교차 검증(정량화 모델에 대해) 또는 내부 검증(식별 모델에 대해)이 적절한 모델을 생성할 수 있음을 나타내는 경우, 최종 모델을 개발하기 위해 다른 보정 샘플 및 검증 샘플을 수집해야 합니다.

복제품

경우에 따라 특정 조건 또는 기준값 범위에 샘플이 거의 없습니다. 이를 보상하기에 대해 이러한 샘플을 복제하려는 유혹이 있을 수 있습니다. 하지만 그것은 문제가 있습니다. 샘플의 중복이 보정 세트와 검증 세트 모두에 있는 경우 성능지수는 오해의 소지가 있습니다(너무 낙관적). 동일한 세트의 중복도 피해야 합니다.

기준 방법(수량화)

수량화는 기준값을 측정하는 기준 방법을 사용합니다. 사용된 기준 방법에 대한 **실험실의 표준 오류(SEL)**는 정량화 모델 개발에 중요한 역할을 합니다. SEL은 중복 샘플의 측정 간 차이의 표준편차입니다.

SEL은 종종 NIR method의 예측의 표준 오차(SEP)에 기여하는 가장 큰 오류입니다 *(참조: 61 페이지, "SEP – 예측의 표준 오차")*。 SEL은 필요한 SEP의 0.7배, 가급적 0.5배를 초과해서는 안 됩니다。 기준값 범위는 SEL의 최소 3배, SEL의 5배 이상이어야 합니다。

SEL을 줄이는 한 가지 방법은 각 샘플에 대해 반복된 기준 측정을 수행하는 것입니다。 측정값의 평균은 샘플의 기준값으로 점검해야 합니다。 각 샘플에 대해 동일한 수의 기준 측정을 수행해야 합니다。 성능지수는 특정 수의 반복된 기준 측정과 관련하여 표시됩니다。 반복된 기준 측정 횟수가 다른 경우 성능지수의 추정치가 잘못될 수 있으므로 피해야 합니다。

샘플 온도

샘플 온도는 물 또는 기타 수소 교량을 포함하는 액체 스펙트럼에 상당한 영향을 미칩니다。 다른 극성 액체의 스펙트럼은 물, 습도 또는 용매를 포함하는 고체의 스펙트럼만큼 영향을 받을 수 있습니다。 그러한 샘플은 정의된 온도에서 측정해야 합니다。

특이값

일부 샘플은 나중에 특이값으로 식별됩니다。 특이값은 어떤 사유로 대부분의 샘플과 다른 샘플입니다。 모델에 부정적인 영향을 미치지 않도록 특이값으로 표시된 샘플은 모델 계산에 포함되지 않습니다。

특이값에 있어서 여러 가지 종류가 있습니다 :

- **스펙트럼적인 특이값**
  샘플의 스펙트럼이 대부분의 다른 스펙트럼과 다른 경우, 샘플은 스펙트럼적인 특이값으로 인식됩니다 *스펙트럼적인 특이값 (참조: 44 페이지, 4.3.3 장)*。
- **기준값 특이값 (수량화)**
  개별 기준값은 수량화 중에 이상 징후를 보일 수 있고 나중에 기준값 특이값으로 인식될 수 있습니다 *기준값 특이값 (수량화) (참조: 50 페이지, 4.3.4 장)*。
  OMD(OMNIS Model Developer)는 ASTM D8321-22에 따른 이상값 검출에 따라 해당 샘플을 특이값으로 인식합니다 *OMNIS Model Developer (OMD) (참조: 62 페이지, 4.4.3 장)*。

4.2 주요 구성 요소 분석 (PCA)

보정 샘플의 분광학적 데이터에는 많은 수의 변수(파장)가 포함되어 있습니다。 변수는 서로 밀접하게 관련되어 있습니다。 따라서 데이터는 매우 중복됩니다。 이러한 타입의 데이터를 처리하기에 대해 PCA 및 PLS와 같은 잠재 변수 모델이 사용됩니다。

주성분 분석(**PCA**, 영어로 *principal component analysis)*은 기준값을 주의하지 않고 스펙트럼에 초점을 맞추고 있습니다。

OMNIS Software는 모델 개발 중에 별도 특이값의 인식 및 자동 데이터 세트의 분할을 위해 PCA를 사용합니다。

준비 단계

다음 준비 단계가 필요합니다 :

1. **데이터 전처리**: OMNIS Software는 정의된 데이터 전처리를 스펙트럼에 적용합니다 *데이터 전처리 (참조: 35 페이지, 4.3.1 장)*。
2. **파장범위**: OMNIS Software는 정의된 파장 선택을를 스펙트럼에 적용합니다 *파장범위 (참조: 43 페이지, 4.3.2 장)*。
3. **평균값 중심화**: 각 파장에 대해 평균 흡수도 값이 계산되고 각 스펙트럼의 해당 값에서 차감됩니다。

첫 번째 주요 구성 요소

준비 단계에 따라 PCA는 스펙트럼 데이터의 정보를 다시 정렬하고 관련 데이터와 노이즈를 분리합니다。 이를 위해, PCA는 파장 변수들을 새로운 가변 공간, 소위 **주요 구성 요소** (**PC**, 영어로 Principal Components)으로 변경합니다。

PCA는 관련 정보를 다양한 파장 변수에서 몇 가지 주요 구성 요소로 변경합니다。 간단한 예를 들어 개념을 설명하기 위해 수천 개의 파장 변수 대신 2 개의 파장 변수만 있고 이러한 2 개 변수는 1개의 주요 구성 요소로 축소됩니다。

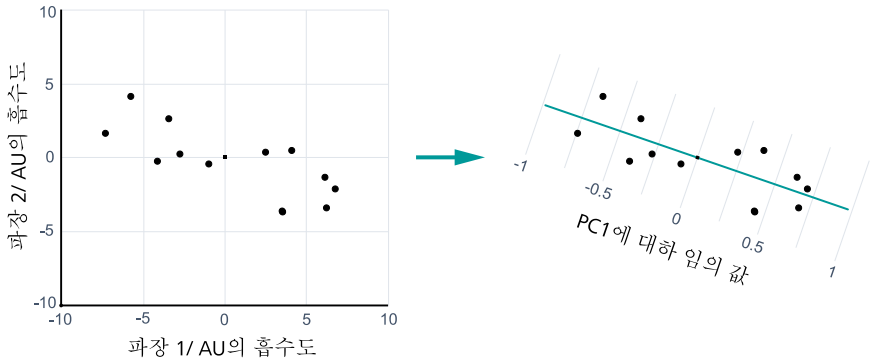


그림 17 2 차원 파장 공간(왼쪽)에서 스펙트럼을 나타내는 점。 1 차원 기본 구성 주요 구성(오른쪽)의 동일한 지점입니다。

왼쪽 *그림17*에서 수평축과 수직축은 2 개의 변수를 가진 원래 파장 공간을 형성합니다. 따라서 각 점은 2개의 파장만 있는 스펙트럼을 나타냅니다. 모든 파장 값의 평균값은 영점입니다.

오른쪽에는 최대 다양성을 설명하는 데이터가 주요 구성 요소 PC1을 구성합니다. 이 예에서는 PC1이 주 구성 요소 공간의 유일한 변수입니다. 따라서 2 개의 원래 변수가 1로 줄어듭니다.

Score 및 residuals

*그림18*은 스펙트럼 *i*를 특징짓는 크기를 보여줍니다 :

- 중심으로부터의 거리 s_i 주요 구성 요소 공간에서 측정됩니다. 주요 구성 요소 s_i 1개뿐인 예에서는 PC1 방향으로 측정됩니다. s_i 거리를 스펙트럼 *i*의 **score**라고 합니다.
- 주요 구성 요소 공간에서 스펙트럼까지의 오프셋 e_i입니다. e_i 거리를 스펙트럼 *i*의 **잔류**라고 합니다.

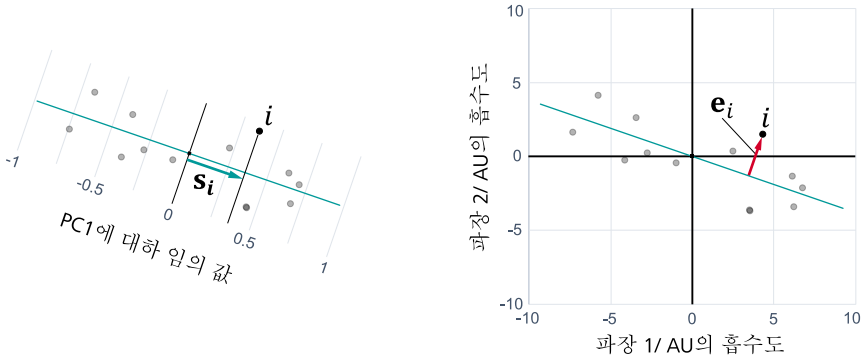


그림 18 Score (왼쪽) 및 잔류 (오른쪽)가 있는 스펙트럼 i.

s_i score는 주요 구성 요소 공간에서 측정됩니다. e_i 잔류는 원래 파장 공간에서 측정됩니다.

여러 주요 구성 요소로 변경

일반적으로 분광학적 데이터의 적절한 설명에 대해 두 개 이상의 주요 구성 요소가 필요합니다.

*그림19*에서 3 가지 원래 변수 x_1, x_2, x_3이 있습니다. 각 점은 3개의 파장을 가진 스펙트럼을 나타냅니다.

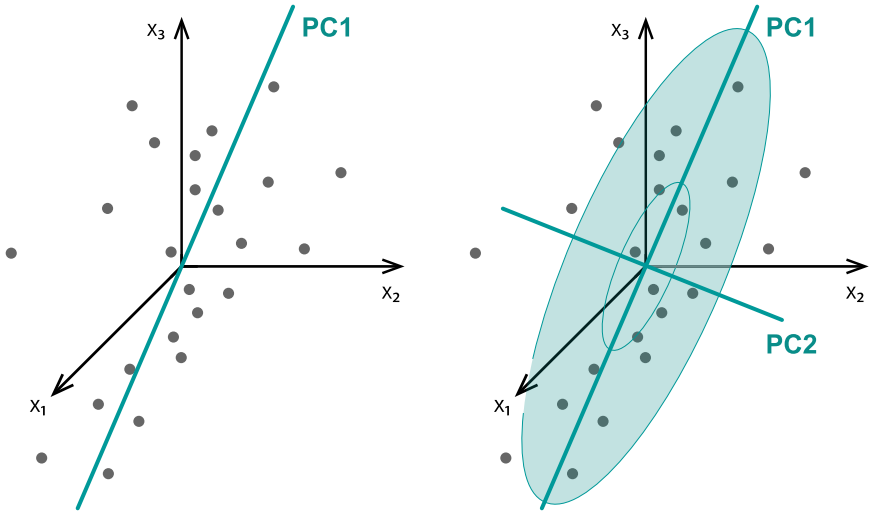


그림 19 원래 변수 3 개가 주요 구성 요소 1 개(왼쪽) 또는 주요 구성 2 개 (오른쪽)로 축소됩니다. PC1 과 PC2 는 2 차원 주요 구성 요소 공간을 형성합니다.

PC1의 첫 번째 주요 구성 요소는 최대 다양성을 설명하는 데이터를 통한 방향입니다.

두 번째 주요 구성 요소인 PC2는 최대 잔류 다양성을 설명하는 데이터를 통한 방향입니다. 다음 모든 주요 구성 요소에도 동일한 사항이 적용되고, 각 구성 요소는 최대 잔류 다양성을 설명합니다. 따라서 처음 몇 개의 주요 구성 요소는 데이터 다양성의 가장 큰 부분을 차지하고, 다른 구성 요소는 주로 노이즈를 포함하고 거부합니다. 이렇게 하면 변수 수를 줄일 수 있습니다.

PCA의 중요한 특징은 모든 주요 구성 요소가 (직각으로) **직교한다** 는 것입니다. 따라서 score는 무관합니다.

마할라노비스 거리

위와 같이 스펙트럼 *i* 의 score는 주요 구성 요소가 공간에서 측정되고 잔류는 원래 파장 공간에서 측정됩니다.

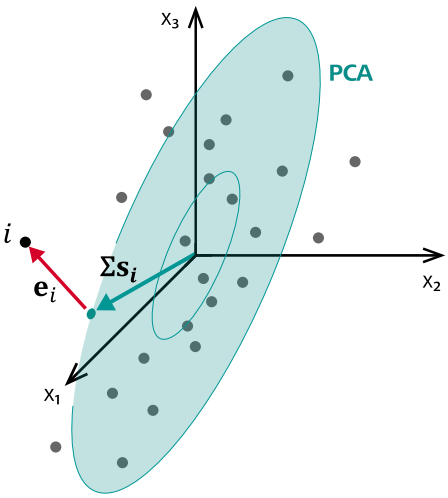
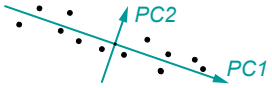


그림 20 스펙트럼 i의 score 및 잔류. 녹색 점은 주요 구성 요소 공간에 대한 점 i(스펙트럼 i)의 직교 투영입니다.

*그림20*에서 score 벡터 $\Sigma \mathbf{s}_i$가 PCA 모델의 중심에서 주요 구성 요소 공간에 스펙트럼을 직교적으로 투영하는 절대 거리(유클리드 거리)를 나타냅니다.



이 예에서 스펙트럼의 유클리드 거리는 PC2보다 PC1 방향으로 더 멀리 떨어져 있습니다. 분산은 **다양성**으로 측정할 수 있습니다. PC1의 다양성이 PC2보다 높습니다.

정규화된 score 벡터 $\mathbf{s}_i$는 **마할라노비스 거리** 라고 불리는 정규화된 거리를 의미합니다. 마할라노비스 거리는 다양성 주요 구성 요소 방향의 차이를 주의합니다. 각 방향은 동일한 가중치를 가집니다. 따라서 다양성이 낮은 한 방향의 작은 유클리드 거리는 다양성이 높은 한 방향의 큰 유클리드 거리만큼 셀 수 있습니다.

여러 파장의 스펙트럼 변환

다양한 파장 변수를 가진 스펙트럼을 주요 구성 요소로 변경하는 경우에도 동일한 컨셉이 적용됩니다. *그림21*에서 각 스펙트럼이 곡선(왼쪽)과 점(오른쪽)으로 표시됩니다.

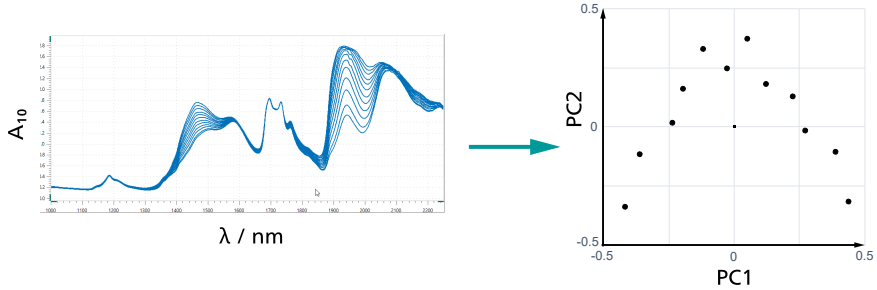


그림 21 스펙트럼 데이터를 주요 구성 요소 공간으로 변경합니다. 오른쪽의 score는 임의의 단위로 표시됩니다.

오른쪽 그림은 PC1과 PC2의 처음 2 가지 주요 구성 요소를 보여줍니다. PC3, PC4 등의 주요 구성 요소를 동일한 방식으로 시각화할 수 있습니다.

PCA 모델은 고정된 수의 주요 구성 요소를 사용합니다. 주요 구성 요소가 많을수록 더 관련성이 높은 스펙트럼적인 변동이 설명됩니다. 동시에 이 모델은 관련 없는 스펙트럼적인 변동 (노이즈)도 감지합니다. 균형 잡힌 타협이 필요합니다.

OMNIS Software에서 주요 구성 요소 분석을 수행하는 경우 PCA에 대해 주요 구성 요소의 수는 선언된 다양성이 최소 95%가 되도록 선택됩니다.

PCA-알고리즘

원본 데이터를 주요 구성 요소 공간으로 변환할 수 있는 몇 가지 방법이 있습니다. OMNIS Software는 단일 값 분할을 수행합니다 *PCA-알고리즘 (참조: 80 페이지, 6.2 장)*.

4.3 데이터 전처리

4.3.1 데이터 전처리

분광 모델은 흡수도 값과 관심있는 parameter (수량화) 또는 제품 제휴(동일시, 검증) 사이의 관계에 기초해 작동합니다. 스펙트럼의 **파라미터화**는 스펙트럼이 이 관계를 잘 표현하도록 보장합니다. 목표는 유용한 정보를 잃지 않고 관련 없는 다양성을 제거하는 것입니다. 아티팩트 및 비선형성이 수정됩니다. 올바르게 매개 변파라미터화하는 경우에 모델의 정확성과 견고성이 향상되고 예측의 반복성과 재현성이 향상됩니다.

파라미터화는 보정 데이터 세트, 유효성 검사 데이터 세트 및 특이값 데이터 세트뿐만 아니라 똑같은 모델과 함께 분석되는 모든 미래에 알 수 없는 샘플에도 적용됩니다.

파라미터화의 첫 번째 단계는 **데이터 전처리**입니다. 데이터 전처리는 지정된 순서대로 수행됩니다. 파라미터화의 두 번째 단계에서는 관련 파장 범위를 정의할 수 있습니다 *파장범위 (참조: 43 페이지, 4.3.2 장)*.

노이즈 감소

스펙트럼에서 신호 주변의 다양한 유형의 임의의 변동이 포함될 수 있습니다. 이러한 예로는 검출기와 장비의 전자 회로로 인한 고주파 노이즈 또는 스캔 측정 중 장비 드리프트로 인한 저주파 노이즈가 있습니다.

분광계는 일련의 단일 측정에서 평균화된 스펙트럼을 제공합니다. 이렇게 하면 고주파 노이즈가 크게 줄어듭니다. 스무딩 필터로 추가적인 소음 감소를 달성할 수 있습니다. 이러한 필터는 노이즈가 고주파이고 신호가 저주파라는 생각에 기초합니다. 주변 흡수도 값을 통해 신호를 근사하게 하고 평균값을 생성하여 노이즈를 줄입니다.

산란 수정

산란 샘플과의 상호작용에 의해 발생하는 빛의 방향 변경을 의미합니다. 검출기에 도달하지 못하는 산란광은 스펙트럼의 바탕선 변동을 유발합니다.

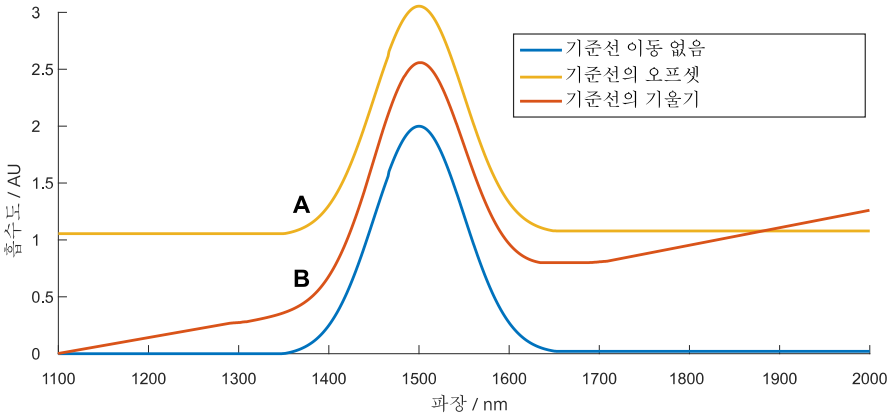


그림 22 바탕선 이동의 주요 타입은 바탕선의 오프셋과 바탕선의 기울기입니다.

다양한 타입의 바탕선 이동을 구별할 수 있습니다 :

- **바탕선의 오프셋** (스펙트럼 **A**)으로 이어지는 일정한 추가 요인입니다.
- **바탕선의 기울기** 를 유발하는 파장 의존 승수(스펙트럼 **B**)입니다.
- **바탕선의 사각 기울기**를 초래하는 이차 파장 의존 승수입니다. 더 높은 순서의 바탕선 기울기가 발생할 수 있습니다.
- **강화** 유발하는 흡수도 승수 요인。 그러나 강화는 무관합니다.

산란은 고체 샘플에서 가장 명백합니다。 결과적인 바탕선 이동을 사용하여 입자 크기 또는 기타 물리적 변경을 감지할 수 있습니다.

그러나 화학적 변동이 관심 있는 경우 적절한 전처리를 통해 바탕의 변위를 최소화해야 합니다.

데이터 전처리

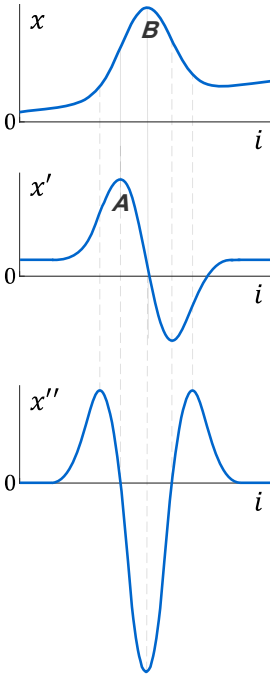
모든 후속 데이터 전처리는 단일 스펙트럼에 적용됩니다。 더 이상의 스펙트럼은 계산에 포함되지 않습니다.

데이터 전처리는 신호 값을 변경합니다。 Y축의 숫자는 더 이상 의미가 없습니다.

적용된 데이터 전처리는 선형 변경입니다。 따라서 Beer-Lambert 법칙은 여전히 유효합니다.

4.3.1.1 미분

Gap-Segment 필터 또는 Savitzky-Golay 필터를 사용하여 미분을 수행할 수 있습니다.



스펙트럼의 미분은 각 지점에서 적정곡선의 기울기 또는 기울기를 설명합니다. 기울기는 초기 스펙트럼의 변화경입니다.

스펙트럼에서 x_i는 파장 i의 흡수도입니다. 첫 번째 순서 x_i'의 미분은 파장 i에서 스펙트럼의 기울기를 나타냅니다. 초기 스펙트럼이 가장 가파른 경우 첫 번째 순서 미분은 최대(**A**)를 갖습니다. 출력 스펙트럼에 피크 (**B**)가 있는 경우 첫 번째 순서 미분은 0입니다.

첫 번째 순서 미분은 바탕선의 오프셋을 제거하고 바탕선의 오프셋에서 바탕선의 기울기를 변경합니다.

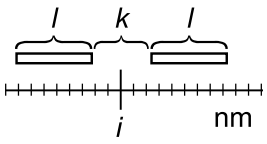
두 번째 순서 x_i''의 미분은 파장 i의 첫 번째 순서 미분의 기울기에 해당합니다. 원래 스펙트럼(**B**)의 양의 피크는 음의 피크가 되고 그 반대의 경우도 마찬가지입니다.

두 번째 순서 미분은 원래 스펙트럼에서 바탕선의 오프셋과 바탕선의 기울기를 제거합니다.

원래 스펙트럼에 상당한 양의 노이즈가 있는 경우 주의해야 합니다. 각 미분 모델은 신호 대 노이즈 비율을 상당히 악화시킵니다. 이러한 사유로 파생 모델은 Gap-Segment 필터 또는 Savitzky-Golay 필터의 평활 기능과 결합됩니다.

4.3.1.2 Gap-Segment

Gap-Segment 필터는 스펙트럼을 부드럽게 합니다. 선택적으로 Gap-Segment 필터는 첫 번째 또는 두 번째 순서에서 미분을 수행합니다. 계산은 미분 모델이 사용되는지 여부에 따라 달라집니다 :



- **도함수 차수 0**: 각 파장 i에 대해 Gap-Segment 필터는 세그먼트 크기 l 예를 들어 10nm를 사용하여 2 개 세그먼트의 평균값을 계산합니다. 2 개의 세그먼트는 예를 들어 5 nm 크기의 거리 k로 구분됩니다.
- **도함수 차수 1**: 첫 번째 순서 미분의 경우 2 세그먼트의 평균값이 별도로 계산됩니다. 그런 다음 두 평균값의 차이가 생성됩니다.
- **도함수 차수 2**: 두 번째 미분은 첫 번째 미분과 동일한 방법으로 계산할 수 있습니다.

스펙트럼의 시작과 끝에는 $l + k/2$ 파장이 계산되고, 스펙트럼을 벗어나는 세그먼트 파장의 경우 공값을 사용합니다.

스펙트럼의 시작과 끝에서, 제로 값은 스펙트럼 외부의 세그먼트 파장에 사용됩니다.

평활은 피크의 약간의 이동과 약간의 베이스를 수반할 수 있습니다.

파라미터 설정

더 큰 평활은 다음을 통해 달성됩니다 :

- 하위 순서에서 미분 ,
- 더 큰 세그먼트 크기 ,
- 더 큰 세그먼트 간격。

과도한 평활은 관련 다양성의 상실로 이어지고, 이는 모델의 예측 능력을 감소시킵니다。

4.3.1.3 Savitzky-Golay

Gap-Segment 필터와 마찬가지로 Savitzky-Golay 필터는 스펙트럼을 부드럽게 하고 선택적으로 일차 또는 이차 미분을 수행합니다。 그러나 Savitzky-Golay 필터는 다른 평활 방법을 사용합니다。

각 파장 *i*에 대해 Savitzky-Golay 필터는 해당 파장 범위에서 낮은 순서 다항식을 조정합니다。 파장 *i*의 다항식 값은 평활 값입니다。 미분의 경우 파생상품의 값이 사용됩니다。

가중 이웃 값의 합계는 모든 것을 한 번에 계산합니다 :

$$x_i = \sum_{j=-k/2}^{k/2} c_j x_{i+j}$$

여기서 k는 필터폭, 도함수 차수, 다항 차수 및 필터폭에 따라 달라지는 c_j 접힘 계수 및 표에서 찾을 수 있고, x_{i+j}는 파장 $i+j$의 초기 스펙트럼의 흡수도 값에 해당합니다。

스펙트럼의 시작과 끝에는 스펙트럼을 벗어나는 필터 파장에 대해 외삽된 값이 사용됩니다(수평 외삽)。

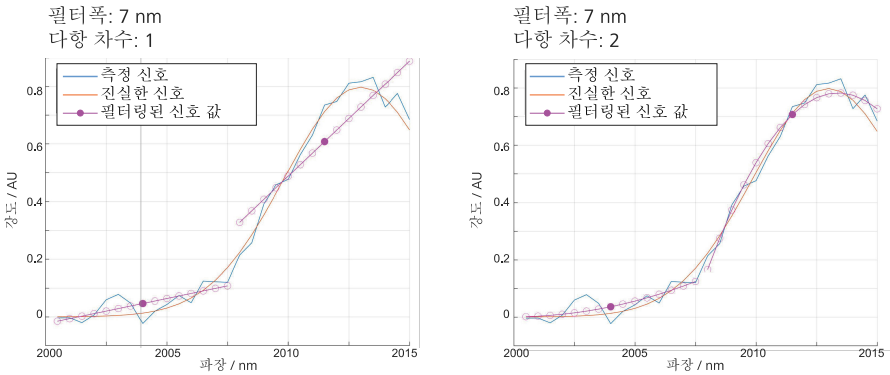


그림 23 다른 다항 차수의 Savitzky-Golay 필터

*그림 23*에는 Savitzky-Golay의 필터링을 표시됩니다。 필터폭은 7nm입니다. 2,004nm 및 2,011.5nm 파장의 다항식이 표시됩니다。 새로운 값은 채워진 점 또는 미분 모델을 사용하는 경우 배출관 모델입니다。 다른 모든 파장은 동일한 방식으로 처리됩니다。

필터폭은 각 다항식이 조정되는 파장범위를 정의합니다。 접힘은 각 파장의 양쪽에서 흡수 값의 영향을 감소시키는 방식으로 가중됩니다。

파라미터 설정

더 큰 평활은 다음을 통해 달성됩니다 :

- 하위 순서에서 미분 ,
- 더 큰 필터폭 ,
- 더 낮은 다항 차수。

과도한 평활은 관련 다양성의 상실로 이어지고, 이는 모델의 예측 능력을 감소시킵니다。

4.3.1.4 SNV - Standard Normal Variate

SNV는 단일 스펙트럼을 다양성 1 및 평균값 0으로 표준화합니다. SNV는 다음과 같이 정의된 파장범위 내에서 각 파장 *i*에 대한 흡수도 값 x_i를 표준화합니다 :

$$x_i = \frac{x_i - m}{s}$$

여기서 *m*은 지정된 파장범위 내의 모든 흡수 값의 평균값 및 *s* 표준편차에 해당합니다。

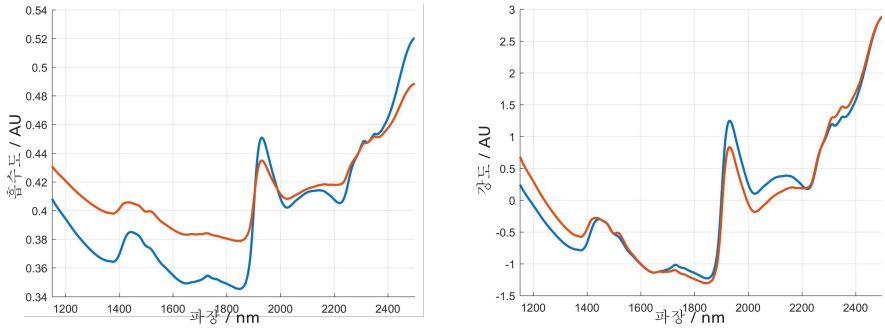


그림 24 흡수도 스펙트럼(왼쪽) 및 SNV 처리 스펙트럼(오른쪽)。

표준화는 스펙트럼 간의 다양성을 제거합니다。 이는 예를 들어 곡물 또는 분말 샘플 또는 흐릿한 매체의 다른 레이어 두께와 같이 관심 없는 속성에서 다양성이 발생하는 경우에 유용합니다。

주의상황: SNV 후에 미분을 사용하는 경우 제거된 다양성이 부분적으로 다시 나타날 수 있습니다。 따라서 SNV 전에 미분을 사용해야 합니다。 예외적으로 SNV가 미분을 수행하기 전에 수행되어야 하는 경우, 다음 순서를 주의할 수 있습니다: detrend, SNV, 배출관。 그러나 일반적인 순서: 미분, SNV, detrend *여러 데이터 전처리 단계의 순서 (참조: 42 페이지, 4.3.1.7 장)*。

parameter

- **파장범위**
  아티팩트가 특정 파장범위(예를 들어 포화도 또는 높은 노이즈)에 영향을 미치는 경우 이러한 범위를 제외할 수 있습니다.
  평균값 및 표준편차의 계산에는 오로지 정의된 파장범위만 고려됩니다. 이어지는 표준화는 정의된 파장범위뿐 아니라 중간 범위에서도 실시됩니다. 스펙트럼의 시작 위치에서 제외된 파장에 대해 인접한 시작 파장의 표준화된 값이 적용됩니다. 스펙트럼의 끝 위치에서 제외된 파장에 대해 인접한 끝 파장의 표준화된 값이 적용됩니다.
  필요 시 제외된 파장은 모델의 계산에서도 제외될 수 있습니다*파장범위 (참조: 43페이지, 4.3.2장)*.
  주의상황 : OMNIS Software 버전 4.6 이상에서는 복수의 파장범위를 정의할 수 있습니다.

4.3.1.5 detrend

Detrend는 최소 제곱 method를 사용하여 이차 다항식을 전체 스펙트럼에 맞게 조정합니다. 그런 다음 detrend는 스펙트럼에서 다항식을 뺍니다.

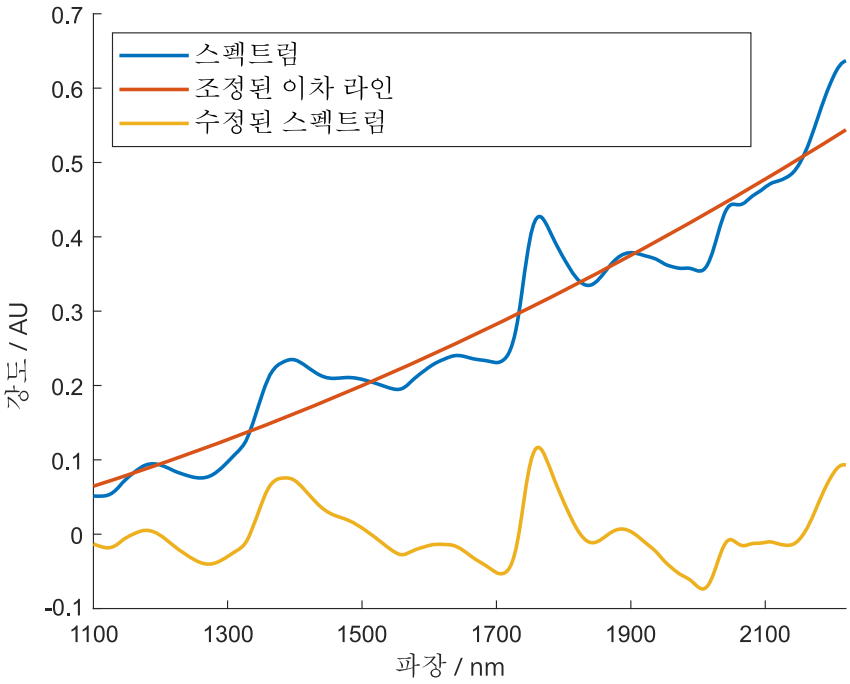


그림 25 Detrend는 청색 스펙트럼을 노란색 스펙트럼으로 변경합니다.

Detrend는 바탕선의 제곱 기울기까지 파장에 따른 산란 효과를 줄여줍니다.

위의 그림은 추세가 지배적인 스펙트럼(청색)을 보여줍니다. 이러한 지배적인 추세가 모든 스펙트럼에서 유사하는 경우, detrend는 잘 작동할 수 있습니다. 다른 경우에, detrend가 유용한 변경을 제거하는 경향이 있습니다. 그런 경우에 미분이 최선의 선택일 것은 가능성이 높습니다.

각 스펙트럼에 대해 별도의 다항식이 적용되기 때문에 추가적인 방해적 다양성이 발생할 수 있습니다. 일반적으로 SNV는 detrend 이전에 사용됩니다. 이는 다항식 계수의 보다 강력한 추정치로 이어집니다.

parameter

- **파장범위**
  아티팩트가 특정 파장범위(예를 들어 포화도 또는 높은 노이즈)에 영향을 미치는 경우 이러한 범위를 제외할 수 있습니다.
  다항식은 정의된 파장범위의 모든 강도값에 맞게 조절됩니다. 이어서 정의된 모든 파장범위에서 다항식이 스펙트럼에서 감산됩니다. 제외된 모든 파장에 대해 강도값이 영으로 설정됩니다.
  필요 시 제외된 파장은 모델의 계산에서도 제외될 수 있습니다*파장범위 (참조: 43 페이지, 4.3.2 장)*.
  주의상황 : OMNIS Software 버전 4.6 이상에서는 복수의 파장범위를 정의할 수 있습니다.

4.3.1.6 데이터 전처리에 대한 개요

전처리	목적	긍정적인 효과	부정적인 효과
Gap-Segment	평활 더 낮은 도함수 차수, 더 큰 세그먼트 크기 또는 더 큰 세그먼트 간격을 사용하여 더 큰 평활을 달성할 수 있습니다.	▪ 고주파 노이즈를 줄입니다.	▪ 과도한 평활은 관련 다양성의 상실로 이어집니다.
Gap-Segment가 있는 미분	바탕선 수정	▪ 첫 번째 순서 미분: 바탕선의 오프셋을 제거합니다. ▪ 두 번째 순서 미분: 바탕선의 오프셋과 바탕선의 기울기를 제거합니다.	▪ 노이즈를 증폭시킵니다. ▪ 스펙트럼의 모양을 변경합니다.
Savitzky-Golay	평활 더 낮은 도함수 차수, 더 큰 필터폭 또는 더 낮은 다항식을 사용하여 더 큰 평활을 달성할 수 있습니다.	▪ 고주파 노이즈를 줄입니다.	▪ 과도한 평활은 관련 다양성의 상실로 이어집니다.

전처리	목적	긍정적인 효과	부정적인 효과
Savitzky-Golay에 수행하는 미분	바탕선 수정	▪ 첫 번째 순서 미분: 바탕선의 오프셋을 제거합니다. ▪ 두 번째 순서 미분: 바탕선의 오프셋과 바탕선의 기울기를 제거합니다.	▪ 노이즈를 증폭시킵니다. ▪ 스펙트럼의 모양을 변경합니다.
SNV – Standard Normal Variate	산란 수정 *	▪ 바탕선의 오프셋을 제거합니다.	▪ 필요한 경우에 관련 다양성이 제거됩니다.
Detrend	산란 수정 *	▪ 바탕선의 오프셋을 제거합니다. ▪ 바탕선의 기울기와 바탕선의 제곱 기울기를 제거합니다.	▪ 필요한 경우에 관련 다양성이 제거됩니다. ▪ 관련 없는 다양성이 발생할 수 있습니다.

* 주의상황: 파장범위에서 포함된 아티팩트(예를 들어: 포화도 또는 높은 노이즈)는 제외해야 합니다.

4.3.1.7 여러 데이터 전처리 단계의 순서

여러 데이터 전처리 단계를 사용할 경우 순서가 중요할 수 있습니다. 기본 법칙은 다음과 같습니다.

바람직하게는 SNV **이전에** Gap Segment **또는** Savitzky Golay**를 사용하고**, Detrend **이전에** SNV**를 사용해야 합니다.**

첫 번째 순서 미분과 SNV가 있는 예: 첫 번째 순서 미분과 SNV가 있는 예: 첫 번째 순서 미분은 바탕선의 기울기를 바탕선의 오프셋으로 변경합니다. 이후 SNV는 이러한 오프셋을 제거합니다. 순서가 반전되는 경우 SNV는 바탕선의 기울기를 변경하지 않습니다. 이후 첫 번째 순서 미분은 바탕선의 오프셋으로 변경됩니다. 오프셋은 그대로 유지됩니다.

두 번째 순서 미분과 SNV에 대한 예: 어떤 경우에도 바탕선의 오프셋과 바탕선의 기울기가 제거됩니다. 그러나 두 번째 순서 미분과 SNV를 올바른 순서로 사용하는 경우 바탕선의 제곱 기울기를 제거할 수 있습니다. 두 번째 순서 미분은 바탕선의 제곱 기울기를 바탕선의 오프셋으로 변경합니다. 이후 SNV는 이러한 오프셋을 제거합니다. 순서가 반전되는 경우 SNV는 바탕선의 제곱 기울기를 변경하지 않습니다. 이후 두 번째 순서 미분은 바탕선의 오프셋으로 변경됩니다. 오프셋은 그대로 유지됩니다.

4.3.2 파장범위

데이터 전처리 *데이터 전처리 (참조: 35페이지, 4.3.1장)* 후에 두 번째 파라미터화 단계가 수행됩니다: 파장범위를 선택하는 경우 목적에 적합하지 않은 범위를 제외할 수 있습니다. 특히 노이즈화 했던 또는 포화 파장범위는 다음 계산에 영향을 미칠 수 있으므로 제외해야 합니다.

노이즈

노이즈는 소량의 빛만이 검출기에 도달할 때 높은 흡수도 수준에서 발생합니다. 다음 그림은 노이즈화 했던 범위를 보여줍니다.

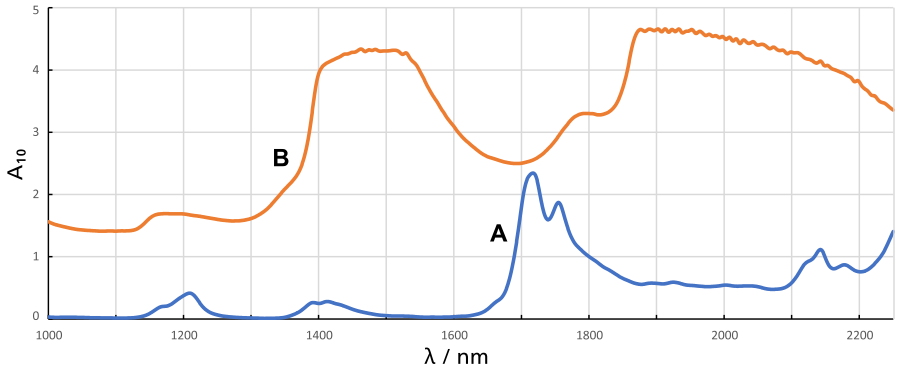


그림 26 노이즈화 했던 파장범위가 있는 예.

스펙트럼 **A**는 일반적인 피크의 형태를 보여줍니다. 스펙트럼 **B**에서 2개의 노이즈화 했던 범위가 있습니다: 1,400 - 1,550 nm 및 1,870 nm 이상. 노이즈화 했던 심한 범위는 벨 곡선이나 이들의 통합과 유사하지 않습니다.

포화도

많은 양의 빛이 검출기에 도달하는 경우 즉, 낮은 흡수도에 경우 검출기가 포화도가 될 수 있습니다.

- **OMNIS NIR Analyzer**
  적분 시간은 항상 자동으로 설정됩니다. 이렇게 하면 포화도가 방지되고 노이즈가 최소화됩니다 *(참조: 8페이지, "적분 시간")*.
- **2060 The NIR**
  자동 적분 시간이 활성화되는 경우 포화도가 발생하지 않습니다.
  수동 적분 시간이 활성화되는 경우 너무 긴 적분 시간으로 인해 검출기가 포화될 수 있습니다. 포화된 범위는 낮은 흡수도 수준에서 발생하지만 항상 시각적으로 쉽게 식별할 수 있는 것은 아닙니다. 따라서 수동 적분 시간은 충분한 여유를 두고 설정해야 합니다 *(참조: 8페이지, "적분 시간")*.

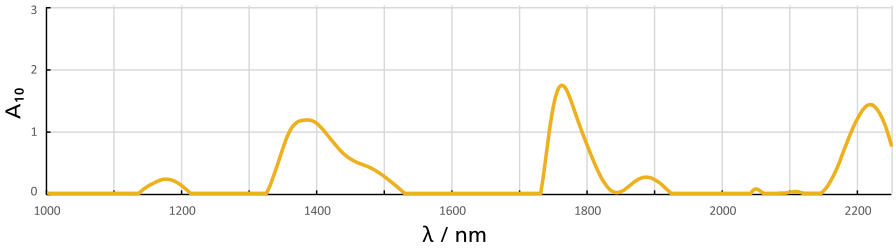


그림 27 포화된 파장범위가 있는 예.

기타 사유

파장범위를 포함하거나 제외하는 다른 사유가 있습니다. 선택은 관심있는 parameter와 해당 흡수도 밴드에 대한 지식에 기초할 수 있습니다 *빛과 물질과의 상호작용 (참조: 3 페이지, 2.1 장)*. 그러나 관련 정보는 전처리 타입에 따라 다른 파장범위로 이동할 수 있는 점을 주의해야 합니다.

데이터 전처리 중에 스펙트럼의 시작과 끝에 이상이 발생한 경우 해당 파장은 제외될 수 있습니다.

화학적 성분 또는 주변 조건의 변동은 특정 파장범위에 영향을 미칠 수 있습니다. 이러한 파장범위를 제외하는 경우 모델의 견고성이 향상될 수 있습니다.

잘 형성된 파장범위를 제외할 때는 주의해야 합니다. 정보가 없는 것처럼 보이는 범위는 실제로 숨겨진 중요한 정보를 제공할 수 있습니다. 그것들은 특이값을 감지하거나 간섭 흡수도 밴드를 처리하는 데 도움이 될 수 있습니다. 실제로, 다변량 측정을 수행하는 주된 사유는 간섭 흡수도 밴드입니다 *선형 회귀의 예 (참조: 77 페이지, 6.1 장)*.

4.3.3 스펙트럼적인 특이값

대부분의 다른 스펙트럼과 다른 스펙트럼인 경우에 스펙트럼적인 특이값이라고 합니다.

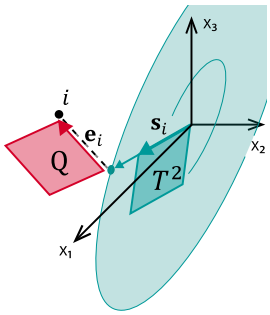
특이값은 주의 깊게 검사해야 합니다. 예를 들어 오염된 샘플이나 측정 오류가 원인인 경우 특이값이 모델을 왜곡할 수 있습니다. 이 경우에 특이값은 모델을 계산할 때 주의하지 않아야 합니다.

반면에 특이값은 다른 스펙트럼에서 잘 다루지 않는 속성을 나타낼 수 있습니다. 이 경우에 특이값이 모델을 개선합니다. 특이값이 유효한 샘플로 보이는 경우 보정 샘플이 변동폭에 걸쳐 균일하게 분포되어 있는지 점검해야 합니다.

특이값 감지에 대한 적절한 조치는 hotelling T^2 및 Q residuals입니다.

Hotelling T² 및 Q residuals

분광 데이터를 주요 구성 요소 공간으로 변경할 때 스펙트럼은 score와 잔류로 특징지어질 수 있습니다 *주요 구성 요소 분석 (PCA) (참조: 31 페이지, 4.2 장)*. 잠재 변수의 공간으로 변경하는 경우에도 마찬가지입니다 *PLS 회귀 (참조: 53 페이지, 4.4.1 장)*.



예: 3차원 파장 공간(x_1, x_2, x_3)이 2차원 공간(녹색)으로 변경됩니다. 점 *i*는 스펙트럼 *i*를 나타내고 3차원 공간에서 2차원 공간으로 투영됩니다. 그것으로 이것을 결과합니다 :

- 2차원 공간 내의 score 벡터 $\Sigma \mathbf{s}_i$ 또는 마할라노비스 거리를 나타내는 정규화된 score 벡터 $\mathbf{s}_i$ 입니다.
- 3차원 공간 내의 잔류 $\mathbf{e}_i$ 입니다.

다음 크기는 $\mathbf{s}_i$ 및 $\mathbf{e}_i$에서 도출할 수 있습니다 *Hotelling T² 및 Q residuals (참조: 83 페이지, 6.4 장)*:

- **Hotelling T^2** 또는 단순히 **T^2**는 모델 중심에서 주요 구성 요소 공간 또는 잠재 변수 공간에 스펙트럼의 직교 투영까지의 제곱 정규화된 거리인 마할라노비스 거리입니다.
  스펙트럼의 모든 score가 평균값과 같은 경우 T^2는 0이고 스펙트럼은 모델의 중심에 있습니다. 그 모델은 중심 가까이에 가장 적합합니다. 중심에서 멀리 떨어진 곳에서 모델이 제대로 적합하지 않을 수 있습니다. T^2-값이 높습니다. 높은 T^2 값은 극단적인 스펙트럼을 나타내고, 예를 들어 화학 성분의 극단적인 구성을 가진 샘플을 나타냅니다.
- **Q-잔류**는 스펙트럼에서 주요 구성 요소 공간 또는 잠재 변수 공간까지의 제곱 직교 거리입니다.
  Q residuals 모델에 의해 설명되지 않는 변동을 나타냅니다. Q 잔류가 높은 경우 예를 들어 측정된 샘플에 다른 물질이 포함된 경우 스펙트럼이 모델에 맞지 않을 수 있음을 나타냅니다.

스펙트럼적인 특이값의 감지

스펙트럼적인 특이값의 검출은 모집단과 다른 스펙트럼을 식별합니다.

1. 파라미터 지정은 다음과 같이 고려됩니다 :
   a. OMNIS Software 버전 4.2 이상: 사용자는 파라미터 지정(데이터 전처리 및 파장 선택)을 사용하는지 또는 사용되지 않는지 여부를 결정합니다.
   파라미터 지정의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다.
   b. OMNIS Software 버전 3.3에서 OMNIS Software 버전 4.1까지: 사용자는 데이터 전처리가 고려되는지 또는 고려되지 않는지 여부를 결정합니다. 파장 선택 및 데이터 전처리의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다.
   c. OMNIS Software 버전 3.2까지: 데이터 전처리는 이상값 검출 시점에까지 정의된 바와 같이 고려됩니다. 파장 선택 및 데이터 전처리의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다.
2. 스펙트럼적인 특이값의 검출은 모든 평균값 중심 스펙트럼의 PCA 모델에 기초합니다 *주요 구성 요소 분석 (PCA) (참조: 31 페이지, 4.2 장)*。 주요 구성 요소의 수는 선언된 다양성이 최소 95%가 되도록 선택됩니다。

3. Hotelling T^2 및 스펙트럼적인 Q residuals에 대한 값은 스펙트럼 특이값을 감지하는 데 사용됩니다. 알고리즘은 측정된 스펙트럼의 hotelling T^2 또는 Q 잔류가 임의의 또는 체계적인 변경의 결과인지 평가합니다.
알고리즘에 대한 설명은 부록에서 확인할 수 있습니다 *스펙트럼적인 특이값 - 알고리즘 (참조: 84 페이지, 6.5 장)*.

4.3.3.1 influence plot

influence plot은 스펙트럼의 기본적 속성을 나타내고 스펙트럼 특이값을 분석하는 데 도움이 됩니다.

influence plot의 기초는 PCA 모델 *주요 구성 요소 분석 (PCA) (참조: 31 페이지, 4.2 장)* 또는 PLS 모델입니다 :

- **수량화**: influence plot은 선택적으로 **PCA** 또는 **PLS**에 기반합니다. PCA와 마찬가지로 PLS 회귀 분석도 분광학적 데이터를 더 적은 변수로 줄입니다. 이때 PLS는 기준값도 고려합니다.
  PCA의 주요 구성 요소는 PLS에서 **잠재 변수**라고 합니다.
  PLS 회귀 (참조: 53 페이지, 4.4.1 장)
- **동일시**: **PCA** 기반 influence plot을 사용할 수 있습니다(OMNIS Software 버전 4.3 이상).

스펙트럼적인 특이값의 타입

Influence plot는 각 스펙트럼에 대해 hotelling T^2 및 Q 잔류의 값을 시각화합니다 *(참조: 44 페이지, "Hotelling T² 및 Q residuals")*. Hotelling T^2 및 Q residuals는 다양한 타입의 스펙트럼적인 특이값을 보여줍니다 :

- **Hotelling T^2 특이값**은 지레의 자루 특이값이라고도 합니다 (영어로 Leverage outlier): 높은 T^2는 주요 구성 요소 공간(PCA) 또는 잠재 변수 (PLS) 공간에 대한 스펙트럼의 투영이 모델 중심에서 멀리 떨어져 있음을 의미합니다.
- **Q residuals 특이값**: Q 잔류가 큰 경우 스펙트럼이 모델에 의해 잘못 설명된다는 것을 의미합니다.

*그림28*에서 여러 가지 스펙트럼을 서로 다른 보기로 보여줍니다 :

- 왼쪽 influence plot: Q residuals는 모델에 의해 설명되지 않은 변동을 설명하는 반면에, hotelings T^2는 모델 자체의 변동을 주의합니다.
  점선은 정의된 특이값 레벨에 대한 **임계값** 또는 **한계값**을 나타냅니다 *스펙트럼적인 특이값 - 알고리즘 (참조: 84 페이지, 6.5 장)*.
  특이값 레벨이 높을수록 한계값이 낮아지므로 더 많은 포인트가 한계값 밖에 놓일 수 있습니다.
- 오른쪽: 예를 들어, 잠재 변수가 있는 2차원 공간으로 변경되는 3개의 변수 x_1, x_2, x_3이 있는 샘플 원래 공간입니다.
  점 A에서 D까지에 대해 수준까지의 직교 거리(점선)와 잠재 변수 공간에서의 모델링된 점(녹색점)을 나타냅니다.

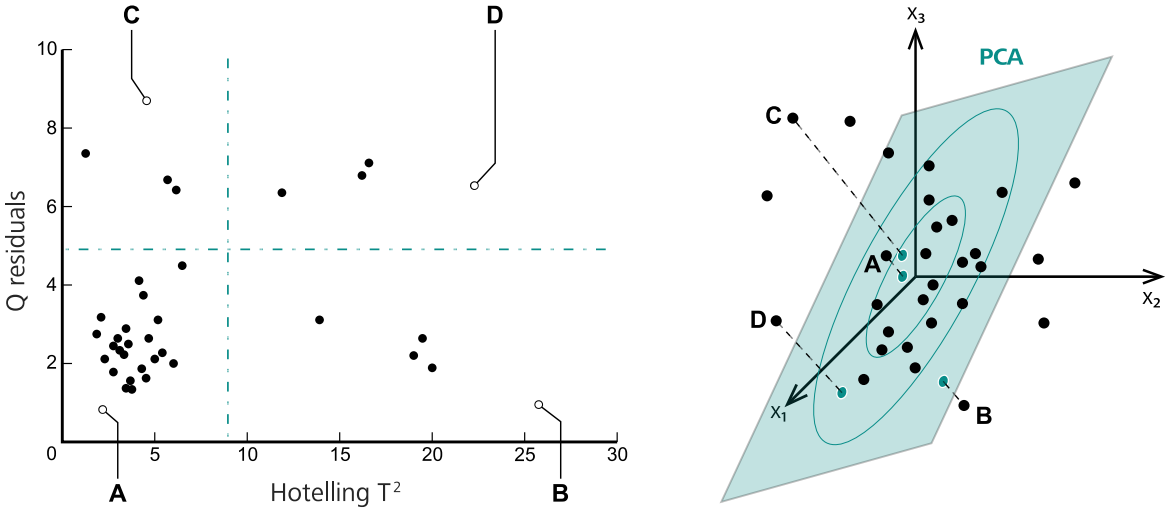


그림 28 Influence plot (왼쪽), 원래 공간 및 잠재 변수(오른쪽)의 공간。 각 스펙트럼은 왼쪽 그림에서 점과 오른쪽 그림에서 점으로 표시됩니다。

보기 두 개에서 모두 서로 다른 특성을 가진 4개의 점이 강조 표시됩니다：

- 스펙트럼 A는 낮은 score 및 낮은 잔류를 가지고 있습니다。 그것은 모델 중심 가까이에 있고 모델로부터 잘 설명됩니다。
- 스펙트럼 B는 hotelling T^2 특이값입니다。 그것은 중심에서 멀리 떨어져 있지만 모델에 의해 잘 설명됩니다。
- 스펙트럼 C는 Q residuals 특이값입니다。 그것은 중심에서 멀리 떨어져 있고 모델에 의해 잘 설명되지 않습니다。
- 스펙트럼 D는 hotelling T^2 특이값 및 Q residuals 특이값입니다。 그것은 중심에서 멀리 떨어져 있고 모델에 의해 부분적으로만 설명됩니다。

Influence plot 다양한 스펙트럼이 모델에 어떻게 영향을 미치는지 보여줍니다。 모든 잠재 변수가 중심을 통과하기 때문에 중심 가까이에 있는 스펙트럼(예를 들어 스펙트럼 A)은 잠재 변수의 방향을 변경할 가능성이 거의 가 없습니다。 지레 값이 없으십니다。 거리가 중심에서 멀어질수록 지레 값과 모델에 영향을 미칠 수 있는 가능성이 커집니다。 일부 스펙트럼은 실제로 모델을 방향(스펙트럼 B)으로 끌어당기는 데 성공하는 반면, 다른 스펙트럼은 특정 범위(스펙트럼 D)에서만 또는 전혀 그렇지 않습니다(스펙트럼 C)。

모든 스펙트럼에 기초한 모델과 비교하여 스펙트럼 B가 없는 모델의 계산은 스펙트럼 D가 없는 모델보다 모델을 더 강하게 변경시키고 스펙트럼 C가 없는 모델보다 더 강하게 변경시킵니다。 스펙트럼 B는 정량화 모델에 큰 영향을 미칠 가능성이 있습니다 - 좋든 나쁘든。 Influence plot의 오른쪽 하단 사분면에 있는 잠재적 특이값을 삭제할지 여부를 결정하기 위해서는 특별한 주의가 필요합니다。

이상적으로는 모델이 많은 스펙트럼의 다양성을 포착해야 합니다。 그 모델은 단지 몇 개의 스펙트럼으로 특징지어지는 것은 바람직하지 않습니

다。 위의 그림에서 일부 스펙트럼은 중심과 대부분의 다른 스펙트럼과의 먼 거리가 있습니다。 그것은 의심스럽습니다。 그 모델은 몇 개의 스펙트럼에 의해 영향을 받습니다。 이것들은 조사되어야 할 잠재적인 특이값입니다。 또한 샘플이 변동폭에 걸쳐 고르게 분포되어 있는지 점검해야 합니다。

PCA 및 PLS influence plot

PCA influence plot 스펙트럼에만 의존합니다。 PLS influence plot 스펙트럼과 기준값에 따라 달라집니다。

다음 표는 PCA influence plot 및 PLS influence plot에 대한 다양한 설정의 영향을 보여줍니다。

	PCA influence-Plot	**PLS influence-Plot**
스펙트럼	기본 PCA 모델은 보정 데이터 세트, 유효성 검사 데이터 세트 및 특이값 데이터 세트의 모든 스펙트럼을 기반으로 합니다。	기본 PLS 모델은 보정 데이터 세트의 모든 스펙트럼을 기반으로 합니다。 이 PLS 모델을 기반으로 스펙트럼의 T^2 및 Q 잔류 값이 3 데이터 세트 모두에서 계산되고 플롯에 표시됩니다。
파라미터화	선택한 데이터 전처리 및 파장범위를 고려합니다。 주의상황: 이상값 검출은 PCA를 기반으로 하며 사용자 설정에 따른 파라미터 지정 및 OMNIS Software 버전을 고려합니다*(참조: 45페이지, "스펙트럼적인 특이값의 감지")*。	선택한 데이터 전처리 및 파장범위를 고려합니다。 주의상황: 예측 시 특이값의 평가는 PLS를 기반으로 하며 데이터 전처리 및 파장범위를 고려합니다。
변수의 수	최소 **95%**의 지정된 다양성에 도달하는 주요 구성 요소의 수를 사용합니다。	현재 선택한 잠재 변수의 수를 사용합니다。
특이값 레벨 및 임계값	현재 선택한 특이값 레벨을 사용하여 임계값(점선)을 계산하고 표시합니다。 동일시: 최근 실시된 데이터 세트의 분할이 특이값 측정 없이 실시된 경우 influence plot은 5%의 특이값 레벨을 사용합니다。 주의상황: 특이값 레벨을 높인 경우 임계값이 낮아지고 모델 개발에서 더 많은 특이값이 발생합니다。	현재 선택한 특이값 레벨을 사용하여 임계값(점선)을 계산하고 표시합니다。 주의상황: 특이값 레벨을 높인 경우 임계값이 낮아지고 예측에서 더 많은 특이값이 발생합니다。

	PCA influence-Plot	PLS influence-Plot
기준값 (수량화)	기준값은 PCA 모델에 영향을 미치지 않습니다. 그러나 각 스펙트럼에는 해당 기준값 특이값이 있을 수 있으므로 특이값으로 표시할 수 있습니다.	기준값은 PLS 모델 및 PLS influence plot에 영향을 미칩니다. 또한 각 스펙트럼에는 해당 기준값 특이값이 있을 수 있으므로 특이값으로 표시할 수 있습니다.

특이값의 분석

잠재적 특이값을 분석할 때 다음 요인을 주의해야 합니다 :

- Hotelling T^2 특이값은 다른 샘플과 비교하여 화학 성분의 극단적인 구성을 가진 샘플을 나타냅니다.
- 예를 들어 Q residuals 특이값은 오염된 샘플 또는 스펙트럼 삽입 오류를 주의할 수 있습니다.

잠재적 특이값은 주의 깊게 검사해야 합니다. 실제 특이값은 스펙트럼 목록에서 제거해야 합니다. 유효한 샘플은 보관해야 합니다. 데이터 세트가 다시 분할되는 경우에 이상값 검출은 첫 번째 실행에서 찾을 수 없는 잠재적 특이값을 찾을 수 있습니다. 한 가지 가능한 사유는 새로운 PCA 모델이 95 %의 명시된 다양성에 도달하기 위해 더 적은 주요 구성 요소를 필요로 하기 때문입니다. 새로 발견된 특이값이 유효한 샘플로 판명될 경우 이러한 특이값은 유지해야 합니다. 이 경우에 이상값 검출 없이 자동 분할을 반복할 수 있습니다.

4.3.3.2 score plot

score plot의 기초는 PCA 모델 또는 PLS 모델입니다 :

- **수량화**: score plot(OMNIS Software 버전 3.0 이상)은 **PLS**을 기반으로 합니다*PLS 회귀 (참조: 53 페이지, 4.4.1 장)*。
- **동일시**: score plot(OMNIS Software 버전 4.3 이상)은 **PCA**을 기반으로 합니다*주요 구성 요소 분석 (PCA) (참조: 31 페이지, 4.2 장)*。

각 스펙트럼은 각 주요 구성 요소 또는 잠재 변수에 대해 하나의 score 값을 갖습니다. score plot에서 각 스펙트럼은 하나의 점으로 표시됩니다. 예를 들어 x축은 첫 번째 잠재 변수의 score를 나타내고, 예를 들어 y축은 두 번째 잠재 변수의 score를 나타냅니다. 또한 각 잠재 변수 쌍을 표시할 수 있습니다.

각 파장 변수의 흡수도 값이 평균값으로 중심되었기 때문에 각 잠재 변수의 score도 평균값으로 중심화됩니다. Score plot (0/0)에 가까운 점은 표시된 두 잠재 변수에 대한 평균 스펙트럼을 나타냅니다. 인접한 점은 유사한 스펙트럼을 나타내고, 더 멀리 떨어져 있는 점은 표시된 두 잠재 변수와 관련하여 서로 다른 스펙트럼을 나타냅니다.

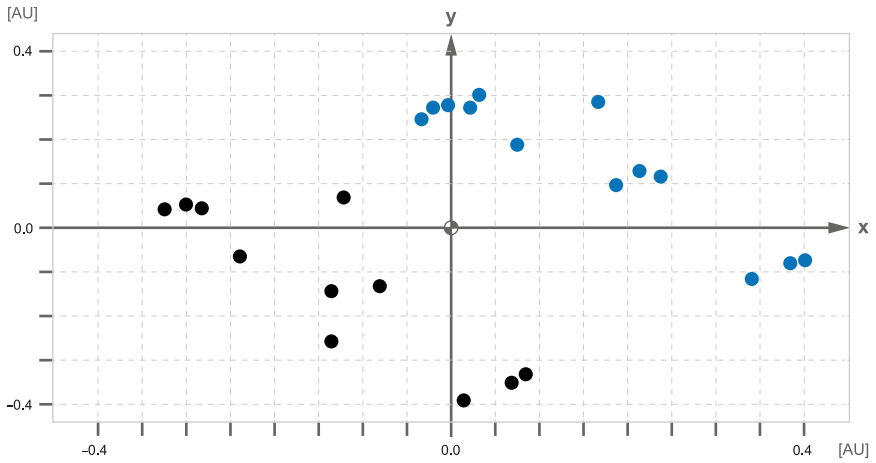


그림 29 잠재 변수 1(x 축) 및 잠재 변수 2(y 축)의 score plot。 AU = 임의의 단위。

그림29은 서로 다른 조건에서 측정된 2 데이터 세트의 예를 보여줍니다。 Scores는 정규화되었고, 각 잠재 변수는 동일한 가중치를 가집니다。

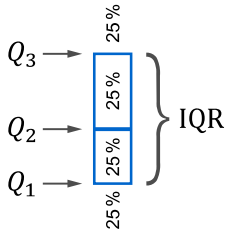
스펙트럼의 모든 잠재적 변수 또는 주요 구성 요소의 score**는** PLS influence plot**의** x**축에 표시되는 단일 값** (hotelling T^2)**으로 결합할 수 있습니다。**

4.3.4 기준값 특이값 (수량화)

정량화 모델의 경우 스펙트럼적인 특이값뿐만 아니라 기준값 특이값도 측정됩니다。 기준값 특이값은 기준값의 이상을 나타냅니다。

일반적으로 기준값 특이값은 잘못 전송된 숫자입니다. 예를 들어, 143은 14.3이 아니라 15.9가 51.9가 아닙니다。 이상값 검출은 경험적 접근법에 근거하여 그러한 전송 또는 전사 오류를 식별합니다。 추가 조사를 위해 명백한 오류만 표시됩니다。

상자 그림



기준값 특이값은 상자 그림을 기반으로 하는 method를 사용하여 측정됩니다。 **상자 그림**은 기준값을 오름차순으로 정렬합니다。 사분위는 데이터 세트를 4개 부분으로 나눕니다。 각 부품에서 기준값의 25 %가 포함됩니다。

첫 번째 분기 Q_1은 25%의 최저치를 나머지와 구분합니다. Q_2는 중위수이고 나머지 50 %에서 가장 낮은 값을 분리합니다. 세 번째 분기 Q_3는 75 %의 최저치를 나머지와 구분합니다. 수직 사각형은 데이터의 평균 50 %를 나타냅니다. 즉, 사분위수 간격(IQR, 영어로 interquartile range)입니다。

IQR 상자 외부에 있는 데이터는 잠재적 특이값으로 간주되고 작은 동그라미로 표시할 수 있습니다。 브레이크아웃의 하한 및 상한은 종종 IQR의 1.5배로 정의됩니다 :

$$[Q_1 - 1.5\ \mathrm{IQR}\,;\, Q_3 + 1.5\ \mathrm{IQR}]$$

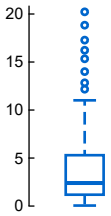
여기서 Q_1은 첫 번째 분기에 해당하고, Q_3는 세 번째 분기에 해당하고, IQR은 분기 간 거리 ($Q_3 - Q_1$)에 해당합니다.

상자 그림을 완성하기 위해 상자 위아래에 있는 휘스커는 잠재적 특이값으로 표시되지 않는 가장 먼 지점까지 도달합니다.

경사 분포에 대한 조정

일반적인 상자 그림은 데이터의 거의 대칭 분포를 가정합니다. 기울어진 분포의 경우 일반적으로 많은 정규 기준값이 잠재적 특이값으로 표시됩니다. 이러한 사유로 OMNIS Software는 분포의 기울기를 주의하는 수정된 버전의 상자 그림을 사용합니다.

다음 예에서는 분포가 더 높은 기준 값으로 이동합니다. 그 결과 높은 값을 가진 8개의 특이값이 발생합니다. 특이값 한계값을 조정한 후에는 2개만 남습니다. 다른 쪽에는 값이 낮은 새 특이값이 나타납니다.

경사 분포	조정된 특이값 한계값
	

적응에 대한 설명은 부록에서 확인할 수 있습니다 *기준값 특이값 - 알고리즘 (참조: 86페이지, 6.6장)*.

4.3.5 데이터 세트의 분할

데이터 세트는 스펙트럼 및 기준값(수량화) 또는 스펙트럼 및 관련 제품 이름(동일시, 검증)으로 구성됩니다. 데이터 세트는 모델을 개발하고 검증하는 데 사용됩니다. 그래서 데이터 세트를 보정 데이터 세트, 유효성 검사 데이터 세트 및 특이값 데이터 세트로 분할해야 됩니다. 분할은 수동으로 또는 자동으로 수행할 수 있습니다.

예를 들어, 충분한 스펙트럼을 사용할 수 있다고 가정할 때, 유효성 검사 데이터 세트에 총 데이터의 20-30 %를 사용할 수 있습니다.

자동 분할 알고리즘

자동 데이터 세트의 분할은 보정 데이터 세트와 유효성 검사 데이터 세트가 전체를 대표하고 서로 독립적이라는 것을 보장합니다. 목표는 PCA 공간에서 거의 동일한 범위를 포괄하고 유사한 통계 속성을 가진 두 개의 세트로 데이터를 분할하는 것입니다. 사용된 알고리즘은 약간 수정된 R. D. Snee, *Validation of Regression Models: Methods and Examples*, Technometrics 19권, 4번 (1977년 11월), 415–428쪽 이중 알고리즘입니다.

복제품

데이터 세트에서 복제본이 없어야 합니다。 이 알고리즘은 복제품이나 유사 복제품을 제거하지 않고 동일한 레코드의 특정 샘플에 복제품을 추가하도록 강제하지도 않습니다。

1. 파라미터 지정은 다음과 같이 고려됩니다 :
    a. OMNIS Software 버전 4.2 이상: 사용자는 파라미터 지정(데이터 전처리 및 파장 선택)을 사용하는지 또는 사용되지 않는지 여부를 결정합니다。
    파라미터 지정의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다。
    b. OMNIS Software 버전 3.3에서 OMNIS Software 버전 4.1까지: 사용자는 데이터 전처리가 고려되는지 또는 고려되지 않는지 여부를 결정합니다。 파장 선택 및 데이터 전처리의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다。
    c. OMNIS Software 버전 3.2까지: 데이터 전처리는 이상값 검출 시점에까지 정의된 바와 같이 고려됩니다。 파장 선택 및 데이터 전처리의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다。
2. 분할은 모든 평균값 중심 스펙트럼의 PCA 모델에 기초합니다。 주요 구성 요소의 수는 선언된 다양성이 최소 95%가 되도록 선택됩니다。
3. 가능한 모든 스펙트럼 쌍 사이의 거리는 PCA score에서 계산됩니다。
4. 가장 멀리 떨어진 2개 스펙트럼은 보정 데이터 세트에 할당됩니다。
5. 나머지 스펙트럼에서 가장 멀리 떨어진 2개의 스펙트럼이 유효성 검사 데이터 세트에 할당됩니다。
6. 나머지 스펙트럼은 이미 보정 데이터 세트에 포함된 스펙트럼에서 가장 멀리 떨어진 스펙트럼을 보정 데이터 세트에 할당합니다。
7. 나머지 스펙트럼은 이미 유효성 검사 데이터 세트에 포함된 스펙트럼에서 가장 멀리 떨어진 스펙트럼을 유효성 검사 데이터 세트에 할당합니다。
8. 데이터 세트 중 하나가 지정된 크기에 도달할 때까지 변경이 계속됩니다。 나머지 스펙트럼은 다른 데이터 세트에 할당됩니다。

4.4 수량화

4.4.1 PLS 회귀

OMNIS Software는 PLS 회귀를 사용하여 정량화 모델을 계산합니다.

PCA와 마찬가지로 부분 **최소 제곱 회귀** (**PLS 회귀**, 영어로 *partial least squares regression*)은 분광학적 데이터를 더 적은 변수로 줄입니다. 하지만 :

- PCA의 주요 구성 요소는 PLS에서 **잠재 변수**라고 합니다. 잠재 변수와 주 구성 요소의 방향은 일반적으로 유사하지만 동일하지는 않습니다.
- 스펙트럼 외에도 PLS 회귀 분석에서 **기준값**도 포함됩니다. 따라서 기준값과 충분한 상관 관계를 달성하고 노이즈를 줄이기 위해 잠복 변수가 덜 필요합니다.

PLS는 광범위하고 고도로 중복된 스펙트럼 분석 데이터에서 기본 **잠재 변수**를 추출합니다. 잠재 변수는 직접 측정되지 않기 때문에 숨겨진 변수라고도 합니다. 잠재 변수는 데이터 변동의 가장 큰 부분을 설명하고 동시에 기준값을 잘 모델링합니다.

준비 단계

PLS 회귀의 준비 단계는 PCA의 준비 단계와 유사합니다 :

1. **파라미터화**: OMNIS Software는 정의된 데이터 전처리 및 정의된 파장 선택을 스펙트럼에 적용합니다.
2. **평균값 중심화**: 각 파장에 대해 평균 흡수도 값이 계산되고 각 스펙트럼의 해당 값에서 차감됩니다.
   기준값도 평균값으로 설정됩니다.

잠재 변수로 변경

준비 단계 후 PLS 회귀 분석에서 기준값을 주의하여 스펙트럼적인 데이터를 잠재적 변수의 공간으로 변경합니다. *그림30*은 그림은 처음 2 개의 잠재 변수 LV1 및 LV2가 나와 있습니다. 마찬가지로 잠재 변수 LV3, LV4 등을 시각화할 수 있습니다.

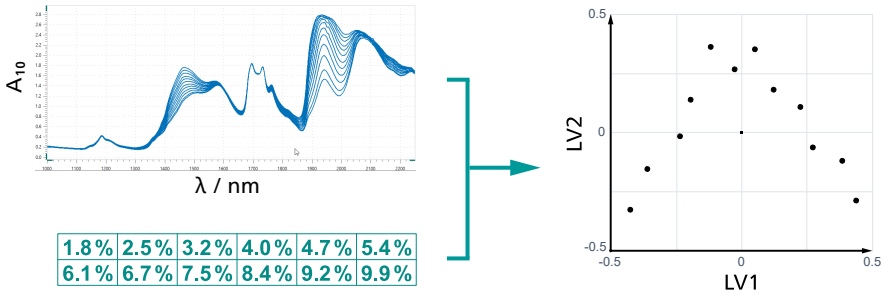
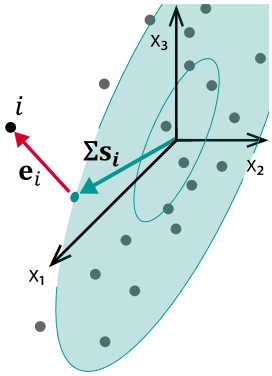


그림 30 스펙트럼 및 기준값을 잠재적 변수의 공간으로 변경합니다. 오른쪽의 score는 임의의 단위로 표시됩니다.

첫 번째 잠재 변수 LV1은 스펙트럼적인 데이터의 다양성을 가장 잘 설명하고 기준값과 가능한 가장 높은 상관관계를 보여줍니다. 이후의 모든 잠재 변수 LV2, LV3 등은 나머지 다양성을 가장 잘 설명하고 기준값과 가능한 가장 높은 상관 관계를 보여줍니다. 따라서 처음 몇 개의 잠재 변수는 다양성성의 가장 큰 부분을 설명하고 상관 관계를 최대화하는 반면, 다른 변수는 주로 노이즈를 포함하거나 거부할 수 있습니다.



Score 및 residuals

PLS는 PCA와 크기가 유사합니다 :

- **Score**: score는 잠재 변수 공간에서 측정됩니다. 잠재 변수의 각 방향에 대한 샘플 *i*의 직교 투영은 중심으로부터의 유클리드 거리를 나타내는 score 벡터 $\Sigma \mathbf{s}_i$를 생성합니다.
  마할라노비스 거리 $\mathbf{s}_i$는 각 방향에 동일한 가중치를 부여하는 정규화된 score 벡터입니다.
- **잔류**: 잔류 벡터 $\mathbf{e}_i$는 원래 파장 공간에서 측정한 샘플 *i*와 잠재 변수 공간 사이의 오프셋입니다.

PLS-알고리즘

PLS 알고리즘은 스펙트럼과 기준값 사이의 공분산을 극대화합니다 *PLS-알고리즘 (참조: 82 페이지, 6.3 장)*.

4.4.1.1 잠재 변수의 수

정량화 모델에서 잠재 변수 수를 선택하는 것은 모델의 예측 능력에 매우 중요합니다. 잠재 변수의 수가 너무 적은 경우 관련 스펙트럼적인 변동은 삽입되지 않습니다. 이를 **과소 조정**이라고 하고 정확한 예측으로 이어집니다.

잠재 변수 수가 너무 많은 경우 보정 샘플이 너무 잘 모델링됩니다. 이 모델은 관련이 없는 스펙트럼적인 변동(노이즈)을 감지합니다. 이를 **과적합**이라고 하고 알 수 없는 샘플의 변동, 불안정 및 부정확한 예측으로 이어집니다.

최적의 수의 잠재 변수를 찾으려면 다음 목표 사이에서 균형을 이루어야 합니다 :

- SEP는 최소 값에 충분히 가깝습니다.
  유효성 검사 데이터 세트가 없는 경우에 SECV가 사용됩니다.
- 정량화 모델은 가능한 한 적은 잠재적 변수를 사용해야 합니다. 의심스러운 경우에 더 낮은 수를 사용해야 합니다.
- 유효성 검사 데이터 세트에 대한 상관관계 그래프는 최적에 가깝습니다. 이상적으로는 기울기가 1에 가깝고 y축 절편이 0에 가깝고 데이터 포인트는 최소 산란입니다.
  유효성 검사 데이터 세트가 없는 경우 교차 검증 값이 사용됩니다 *(참조: 55 페이지, "교차 검증")*.

성능지수는 사용 가능한 보정 샘플 및 검증 샘플에 기초한 추정치일 뿐입니다. 일반적으로 잠재 변수가 적은 정량화 모델이 더 강력합니다.

4.4.1.2 loading plot

Loading plot OMNIS Software에서 PLS 모델을 기반으로 합니다 *PLS-알고리즘 (참조: 82 페이지, 6.3 장)*。

PLS loading은 원래 파장 변수(모수화 포함)가 각 잠재 변수의 형성에 어떻게 기여하는지 보여줍니다。 Loading이 양수인지 음수인지는 중요하지 않습니다。

Loading은 첫 번째 잠재 변수가 기준 parameter에 가장 적합한 다양성을 기록하도록 계산됩니다。 이후의 모든 잠재 변수는 기준 parameter에 가장 중요한 나머지 다양성을 기록합니다。 따라서 0과 크게 다른 PLS loading은 기준 parameter를 모델링하는 데 적합한 파장을 나타냅니다。

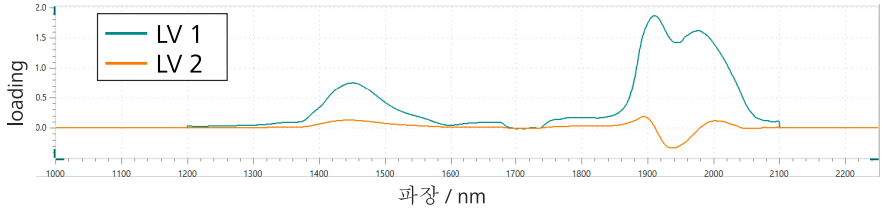


그림 31 잠재 변수 LV1 및 LV2 에 대한 loading plot。

*그림31*에서 파장 선택이 1.200nm ~ 2.100nm 범위로 제한했습니다。 따라서 이 범위 밖에서 loading이 발생하지 않습니다。

4.4.2 정량화 모델에 대한 검증

검증 중에 모델이 성능과 견고성에 대한 요구 사항을 충족하는지 점검합니다。 이를 위해 예측 오류는 가능한 한 현실적으로 평가해야 합니다。

정량화 모델은 일반적으로 다음과 같이 검증됩니다 :

1. 제한된 수의 샘플에서 시작하여 유효성 검사 데이터 세트가 없는 모델은 교차 검증을 통해 개발 및 테스트됩니다 (아래 참조)。
2. 샘플 수가 많을수록 데이터 세트는 모델 개발을 위한 보정 데이터 세트와 모델 검증에 대한 유효성 검사 데이터 세트로 분할됩니다。
3. 유효성 검사 데이터 세트를 위한 샘플은 다른 날에 수집되고 측정되며 필요한 경우 다른 장비로 다른 인원을 통해 수집되고 측정됩니다。

교차 검증

교차 검증은 보정 데이터 세트에만 의존합니다。 이는 각 보정 샘플 없이 생성된 임시 모델을 사용하여 각 보정 샘플에 대한 추정치를 제공합니다。

교차 검증은 다음 방법 중 하나로 다중수행 공정에 종류를 사용합니다 :

- **Leave-One-Out**
  Leave-One-Out 교차 검증 (LOO-CV)를 사용하는 경우 각 라운드에서 샘플 1개가 재설정되고 나머지 샘플은 모델을 생성합니다。 그런 다음에 이 모델은 반환된 샘플에 대한 관심있는 parameter를 예측합니다。 이 예측은 샘플의 추정치 역할을 합니다。
  각 샘플이 한 번 재설정될 때까지 사이클이 계속됩니다。

- **K 폴드**
  K 폴드 교차 유효성 검사 방식에서 보정 데이터 세트가 가능한 한 동일한 크기의 *k* 블록으로 분할됩니다.
  각 라운드에서 1개의 블록이 재설정되고 나머지 블록에서 모델이 생성됩니다. 그런 다음에 이 모델은 반환된 샘플에 대한 관심있는 parameter를 예측합니다.
  각 블록이 한 번 재설정될 때까지 사이클이 계속됩니다.
  블록은 여러 가지 방법으로 선택할 수 있습니다 :
  - **Fixed Blocks (DUPLEX)** (OMNIS Software 버전 3.2 이상): 블록은 이중 알고리즘을 사용하여 재현할 수 있도록 선택됩니다. 각 예측은 각 샘플에 대한 추정치로 직접 사용됩니다.
  - **Random**: 블록은 임의의로 선택됩니다. 위에서 설명한 절차는 여러 번 반복됩니다. 보정 데이터 세트는 매번 *k* 블록으로 다르게 분할됩니다. 마지막으로 각 샘플에 대해 여러 가지 추정치가 있습니다. 그들의 평균값은 각 샘플의 추정치 역할을 합니다.

일반적으로 Leave-One-Out 선호됩니다. 대규모 보정 데이터 세트의 경우 교차 유효성 검사 방식 K 폴드를 사용하여 계산 시간을 절약할 수 있습니다. *k*의 일반적인 값은 5입니다.

4.4.2.1 상관관계 그래프

상관관계 그래프는 기준값과 계산된 값 사이의 상관 관계를 시각화합니다. 이를 통해 정량화 모델을 한 눈에 평가할 수 있습니다.

계산된 값은 다음과 같이 측정됩니다 :

- 유효성 검사 데이터 세트 및 특이값 데이터 세트: 정량화 모델을 통한 예측
- 보정 데이터 세트: 교차 검증 추정치

상관관계 그래프에서 각 샘플은 점으로 표시됩니다. 샘플의 계산된 값은 x축에서 찾을 수 있고, 기준값은 y축에서 찾을 수 있습니다. 회귀선은 변수 사이의 체계적인 관계를 보여줍니다. 이상적으로는 회귀 분석의 기울기가 1이고 y축 절편이 0이고 모든 점이 직선에 있습니다. 즉, 각 샘플에 대해 계산된 값은 기준값과 동일합니다.

이상적인 경우로부터의 편차는 체계적 오류와 임의의 오류를 구별하는 데 도움이 됩니다. 회귀선의 위치는 체계적인 오류를 나타냅니다. 회귀선으로부터의 점 거리는 임의의 오류를 나타냅니다.

다음 상관관계 그래프 **A**는 양호한 상관 관계를 보여줍니다. 다른 다이어그램은 아래에 설명된 다양한 종류의 오류를 보여 줍니다.

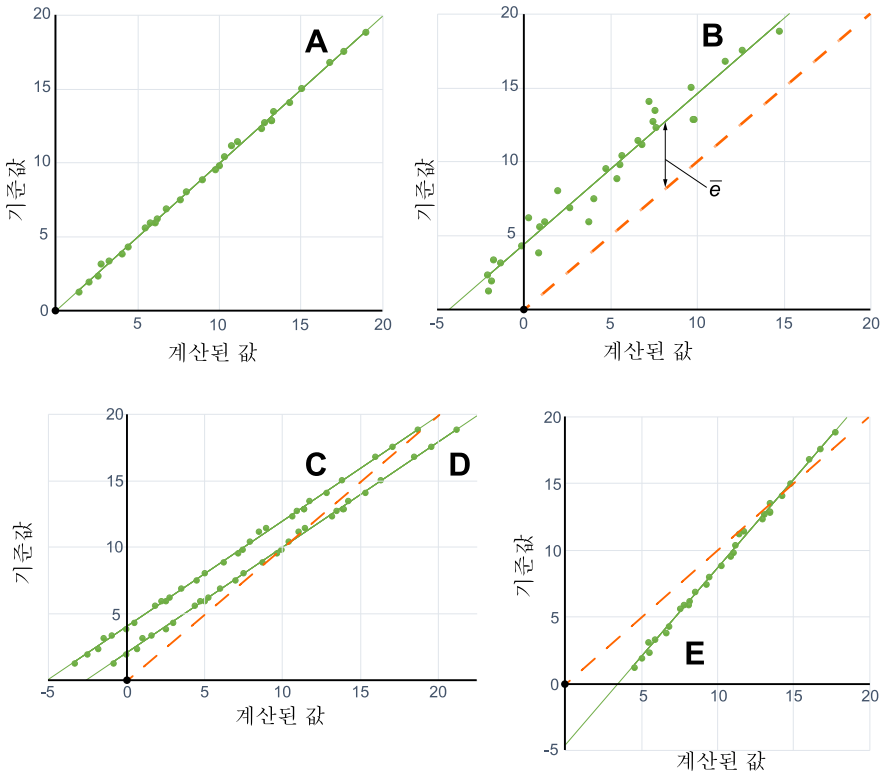


그림 32 상관관계 그래프。 각 점은 샘플을 나타내고 기준값과 계산된 값을 표시합니다。 점선은 이상적인 45° 직선을 나타냅니다。

체계적인 오류

체계적인 오류는 항상 발생하고 특정 사용에 대해 반복할 수 있는 오류입니다。 체계적인 오류는 수정할 수 있습니다。 그것들은 회귀선의 바이어스 $\bar{e}$와 기울기 b로 정량화됩니다 :

$$y = b\hat{y} + \bar{e}$$

기울기가 1이고 바이어스가 0이는 경우 체계적인 오류가 없습니다。

바이어스는 기준값과 계산된 값 사이의 평균 오류입니다 :

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \bar{y} - \bar{\hat{y}}$$

여기서 n 샘플의 수, e_i는 i 번째 오류인데, y_i는 i 번째 기준값, $\hat{y}_i$ 아이콘은 i 계산된 값, $\bar{y}$ 아이콘은 준값의 평균 및 $\bar{\hat{y}}$ 아이콘은계산된 값의 평균값에 해당합니다。

직선 **B**와 **C**는 양의 바이어스를 가지고 있고 직선 **E**는 음의 바이어스를 가지고 있습니다。

반대되는 징후가 있는 오류는 서로를 상쇄합니다。 따라서 직선 **D**의 바이어스는 0에 가깝습니다。

회귀 직선의 **기울기**는 :

$$b = \frac{s_{\hat{y}y}}{s_{\hat{y}}^2}$$

여기에서 $s_{\hat{y}y}$ 아이콘은 기준값과 계산된 값 사이의 공분산에 해당하며 아이콘은 분산 그리고 $s_{\hat{y}}^2$ 아이콘은 계산된 값에 해당합니다.

기울기는 속성 기반 오류로 간주할 수 있습니다 :

- $b > 1$ (직선 **E**): 계산된 값이 높을수록 바이어스에 기여하는 오류가 더 높습니다 (긍정적).
- $b < 1$ (직선 **C** 및 **D**): 계산된 값이 높을수록 편차에 기여하는 오차가 작습니다 (부정적).
- $b = 1$ (직선 **A** 및 **B**): 바이어스를 일으키는 오류는 일정합니다.

Y축이 있는 회귀선의 **y축 절편**은 $\bar{y} - b\bar{\hat{y}}$입니다.

기울기와 y축 절편은 기준값을 종속 변수(y축)로 사용하고 계산된 값은 독립 변수(x축)로 계산됩니다.

임의의 오류

모든 점이 회귀선에 직접 있는 경우 임의의 오류는 없습니다. 분산될수록 임의의 오류가 더 높습니다.

상관관계 그래프 **B**에서 랜덤 오류는 다른 오류보다 큽니다.

오류 종류의 시각화

위 그림의 직선은 다음과 같은 종류의 오류를 보여줍니다 :

	체계적인 오류			임의의 오류
직선	**바이어스**	**기울기**	**Y축 절편**	
A	~ 0	~ 1	~ 0	소형
B	> 0	~ 1	> 0	대형
C	> 0	< 1	> 0	소형
D	~ 0	< 1	> 0	소형
E	< 0	> 1	< 0	소형

4.4.2.2 성능지수

성능지수는 기준값과 계산된 값 사이의 일치도를 수치로 표시합니다. 계산된 값은 정량화 모델에 의해 측정됩니다.

R^2 – 결정계수

R^2 결정 계수 (영어로 coefficient of determination)는 정량화 모델의 적합성을 측정합니다. 특정 데이터 세트에 대해 이것이 정량화 모델에 의해 선언된 기준 값 변동의 비율입니다 :

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

여기서 SS_{reg}는 회귀제곱합(계산된 값의 다양성, 즉 설명된 다양성) 해당하고 SS_{tot} 총 제곱합(기준값 분산) 해당하고 잔류제곱합의 SS_{res}(잔류 다양성, 즉 설명할 수 없는 다양성) 해당하고 y_i i 번째 샘플의 기준값을 해당하고 $\hat{y}_i$ i 번째 샘플의 계산된 값을 해당하고 $\bar{y}$ 기준값의 평균값을 해당합니다.

R^2 값은 1의 일부입니다. R^2는 계산된 1 값이 기준값과 완벽하게 일치함을 의미합니다. R^2는 0.9 기준값은 분산의 90 %가 계산된 값으로 설명되고 10 %는 설명되지 않음을 의미합니다.

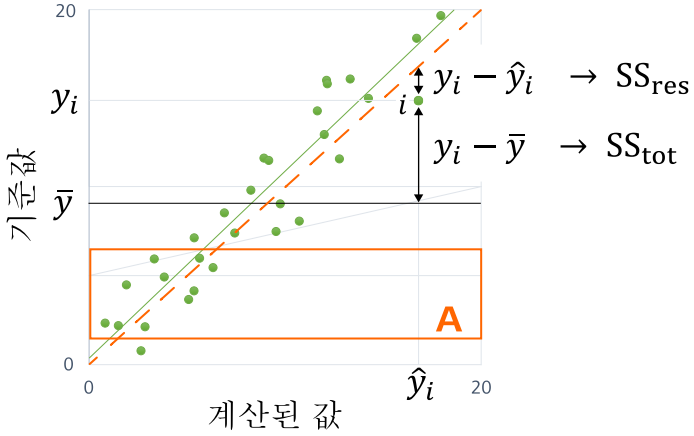


그림 33 R^2 계산에 대한 성분. 점선은 45°입니다.

위의 그림은 PLS 회귀에서 샘플 i와 그 잔류를 사용한 상관관계 그래프를 보여 줍니다. 잔류 변동은 SS_{res}에 포함되고, 기준값 분산은 SS_{tot}에 포함됩니다.

높은 R^2 값은 사용 가능한 정량화 모델 또는 정확한 예측을 보장하지 않습니다. R^2의 크기는 기준값의 변동에 따라 직접 달라집니다. 기준값의 범위가 더 작은 회귀(범위 A)은 잔류 변동과 거의 동일하지만 기준값의 변동은 더 작습니다. 결과 R^2 값이 더 낮습니다. 따라서 R^2가 높은 사유는 비현실적으로 큰 기준값 범위일 수 있습니다. 반면에 제조 공정의 데이터는 제한된 값 범위로 인해 R^2 값이 낮아질 수 있습니다. 예측 가능성을 평가하기에 대해 표준 오류를 사용해야 합니다.

절대 R^2 값은 주의하게 고려해야 합니다. 더 중요한 것은 각 추가 잠재 변수에 대한 변화의 정도입니다 *잠재 변수의 수 (참조: 54페이지, 4.4.1.1 장)*.

계산에 사용되는 값에 따라 다른 R^2 값이 생성됩니다 :

- R^2C (OMNIS Software에 표시되지 않음): 보정 데이터 세트의 스펙트럼 계산된 값으로 계산됩니다.
- **R^2CV**: 보정 데이터 세트의 스펙트럼 교차 검증 추정치를 사용하여 계산한다 *(참조: 55페이지, "교차 검증")*.

- **R²P**: 유효성 검사 데이터 세트에서 스펙트럼의 계산된 값을 사용하여 계산합니다。

계산을 위해 OMNIS Software는 피어슨 샘플 상관 계수 $r_{y,\hat{y}}$의 제곱을 사용합니다：

교차 검증의 결정계수：

$$\mathrm{R^2CV} = r^2_{y,\hat{y}_{\mathrm{cv}}} = \frac{(\sum_{i=1}^{n}(y_i - \bar{y})\,(\hat{y}_{\mathrm{cv}_i} - \bar{\hat{y}}_{\mathrm{cv}}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(\hat{y}_{\mathrm{cv}_i} - \bar{\hat{y}}_{\mathrm{cv}})^2}$$

여기서 y_i 아이콘이 *i* 번째 샘플의 기준값, $\bar{y}$ 아이콘이 기준값의 평균값, $\hat{y}_{\mathrm{cv}_i}$ 아이콘이 *i* 번째 샘플의 교차 검증 추정치, 아이콘이 $\bar{\hat{y}}_{\mathrm{cv}}$의 평균값, 그리고 *n* 보정 데이터 세트의 샘플 수를 해당합니다。 보정 데이터 세트의 각 샘플에서 교차 검증의 정확한 추정치가 있습니다。

예측의 결정계수：

$$\mathrm{R^2P} = r^2_{v,\hat{v}} = \frac{(\sum_{i=1}^{v}(v_i - \bar{v})\,(\hat{v}_i - \bar{\hat{v}}))^2}{\sum_{i=1}^{v}(v_i - \bar{v})^2 \cdot \sum_{i=1}^{v}(\hat{v}_i - \bar{\hat{v}})^2}$$

여기서 v_i는 *i* 번째 검증 샘플의 기준값, $\bar{v}$ 기준값의 평균값, *i* i번째 검증 샘플의 $\hat{v}_i$ 계산된 값, $\bar{\hat{v}}$ 계산된 값의 평균값 및 v 검증 샘플의 수를 해당합니다。

SEC - 보정의 표준 오차

보정의 표준 오차 (SEC)는 보정 데이터 세트를 기반으로 합니다。 SEC는 이론적으로 최고의 예측 정확도에 대한 추정치로 간주될 수 있습니다。 SEC는 부분 최소 제곱 회귀(PLS)의 잔류의 표준 편차입니다：

$$\mathrm{SEC} = \sqrt{\frac{\mathbf{e^t e}}{n - k - 1}}$$

여기서 **e**는 모델에 의해 설명되지 않은 보정 데이터 세트의 모든 기준값 변동을 포함하는 잔류 벡터이고, *n* 보정 샘플의 수 및 *k* 잠재 변수 수입니다。 분모 *n-k*–1은 잔류 벡터 **e**의 자유도 수입니다。

즉: SEC는 보정 데이터 세트의 샘플에 대한 기준값과 계산된 값 사이의 차이의 표준 편차입니다：

$$\mathrm{SEC} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - k - 1}}$$

여기서 y_i *i* 번째 보정 샘플의 기준값, $\hat{y}_i$ *i* 번째 보정 샘플의 계산된 값, *n* 보정 샘플의 수, *k* 잠재 변수 수입니다。

SEC는 때때로 RMSEC라고도 합니다。 SEC에서 임의의 오류와 체계적인 오류(기울기 및 바이어스)가 포함됩니다。

SECV - 교차 검증의 표준 오차

교차 검증의 표준 오차 (SECV)는 보정 데이터 세트를 기반으로 합니다. SECV는 보정 데이터 세트와 교차 유효성 검사 방식를 기반으로 예측 정확도를 추정합니다 *(참조: 55페이지, "교차 검증")*。 SECV는 초기 모델 평가 또는 최적의 잠재 변수 수를 결정하는 데 사용할 수 있습니다.

SECV는 보정 데이터 세트의 샘플에 대한 기준값과 교차 검증 추정치 간의 차이의 표준편차입니다.

$$\mathrm{SECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{\mathrm{cv}_i})^2}{n}}$$

여기서 y_i i 번째 샘플의 기준값, $\hat{y}_{\mathrm{cv}_i}$ i 번째 샘플의 교차 검증 추정치 및 n 보정 데이터 세트의 샘플 수입니다.

SECV에서 임의의 오류와 체계적인 오류(기울기 및 바이어스)가 모두 포함됩니다.
다른 접근법은 RMSECV라고 하는 수정되지 않은 SECV와 수정되지 않은 SECV에 대해 별도의 값을 사용합니다。 낮은 바이어스에 대해 이러한 값은 유사합니다.

SEP – 예측의 표준 오차

예측의 표준 오차 (SEP)는 유효성 검사 데이터 세트를 기반으로 합니다. 따라서 SEP는 예측 정확도에 대한 가장 현실적인 추정치를 제공합니다.

SEP는 유효성 검사 데이터 세트의 샘플에 대한 기준값과 계산된 값 사이의 차이의 표준 편차입니다 :

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{v}(v_i - \hat{v}_i)^2}{v}}$$

여기서 v_i는 i 번째 검증 샘플의 기준값, 검증 샘플의 $\hat{v}_i$는 i 번째 계산된 값 및 v 검증 샘플의 수에 해당합니다.

SEP에서 임의의 오류와 체계적인 오류(기울기 및 바이어스)가 모두 포함됩니다.
다른 접근법은 RMSEP라고 하는 수정되지 않은 SEP와 수정되지 않은 SEP에 대해 별도의 값을 사용합니다。 낮은 바이어스에 대해 이러한 값은 유사합니다.

성능지수의 해석

성능지수는 추정치입니다。 예: SEP 값은 기존 샘플에 기초한 표준편차의 *추정치*이고 자체 표준편차를 가지고 있습니다。 검증 샘플의 수가 많을수록 추정치의 신뢰성이 높아집니다.

SEP는 SECV 및 SEC와 유사해야 합니다. 너무 큰 차이는 과도한 조정을 나타낼 수 있습니다 *잠재 변수의 수 (참조: 54페이지, 4.4.1.1 장)*. 경험칙은 그 차이가 20 %를 초과해서는 안 된다는 것입니다.

또한 기준 방법에 대한 실험실(SEL)의 표준 오류와 관련하여 표준 오류를 주의해야 합니다. SEP가 SEL보다 1.4배에서 2.0배 높은 경우 NIR method의 정밀도가 허용됩니다. 그럼에도 불구하고, 더 큰 SEP는 요구 사항을 충족하는 한 받아들여질 수 있습니다.

기준 방법에 대한 SEL보다 낮은 SEC 또는 SECV는 과다 적합을 나타냅니다.

위의 SEL과의 관계에서 SEL은 각 샘플에 대해 정확한 반복 기준 측정 횟수에 기초한다고 가정합니다 *(참조: 29페이지, "기준 방법(수량화)")*.

4.4.3 OMNIS Model Developer (OMD)

정량화 모델을 개발하는 것은 어렵고 시간이 많이 소요되고 어느 정도의 전문 지식이 필요합니다. Omnis Model Developer (OMD, OMNIS Software 버전 4.0 이상)는 개발을 자동화하고 최적화된 정량화 모델을 제공합니다.

작동 방식

적절한 샘플 추출이 전제조건입니다 *물리적 샘플 (참조: 28페이지, 4.1 장)*. 스펙트럼 및 해당 기준값으로 구성된 데이터 세트가 OMD에 대한 입력으로 사용됩니다.

OMD는 데이터 전처리를 주의하지 않고 5%의 특이값 레벨과 부록에 나열된 알고리즘을 사용하여 스펙트럼적인 특이값을 측정합니다 *(참조: 84페이지, "모델 개발에서 스펙트럼적인 특이값의 감지")*.

데이터 세트의 분할은 이상값 검출 후에 남은 스펙트럼 수에 따라 달라집니다 :

스펙트럼 수	교차 유효성 검사 방식	유효성 검사 데이터 세트
> 99	K 폴드 (블록 5개, DUPLEX)	스펙트럼의 25 %
30–99	K 폴드 (블록 5개, DUPLEX)	—
< 30	Leave-One-Out	—

분할 후 ASTM D8321-22에 따라 보정 데이터 세트에서 추가 특이값이 측정됩니다.

모델의 평가 및 분류는 다양한 주요 수치에 기초합니다. OMD는 데이터 전처리, 파장 선택 및 잠재 변수 수를 최적화하여 과도한 조정의 위험과 과소 조정의 위험 사이의 균형을 유지합니다.

결과

OMD의 결과는 예측 강도에 따라 정렬된 모델 목록입니다. **예측력**은 모델 복잡성, 성능지수 및 데이터 세트 크기에 따라 계산됩니다.

이 목록은 원하는 모델을 쉽게 선택할 수 있도록 색상 코딩되어 있습니다 :

- 녹색: 좋은 예견력。
  샘플 수가 충분히 많은 경우 동일한 타입의 알려지지 않은 모든 샘플에서 모델이 잘 작동합니다。 성능지수는 향후 오류에 대한 신뢰할 수 있는 추정치를 제공합니다。
- 노란색: 평균 예측력。
  샘플 수가 충분히 많은 경우 모델이 잘 작동할 것으로 예상됩니다。 성능지수는 향후 샘플에 대해 지나치게 낙관적일 수 있습니다。 별도의 검증이 권장됩니다。
- 적색: 예측력이 부족합니다。
  그 모델에는 심각한 단점이 있습니다。 그것은 사용해서는 안 됩니다。

동일한 색상의 모델은 낮은 예측 오류와 소수의 잠재적 변수 간의 균형 잡힌 절충을 선호하는 정보 기준에 따라 정렬됩니다。

파라미터 지정 최적화

전체 모델을 자동으로 생성하는 대신 파라미터 지정만 최적화하는 것도 가능합니다。 현재 설정(예를 들어 데이터 세트의 분할, 교차 유효성 검사 방식)은 그대로 유지되지만 최적화에 영향을 미치지 않습니다。

4.4.4 기울기 수정/y 절편 수정

기울기 수정/y 절편 수정은 사용하는 경우 정량화 모델을 적용할 때 체계적인 오류 (바이어스, 기울기)를 수정할 수 있습니다。

보정 데이터 세트의 체계적인 오류의 가능한 원인은 다음과 같습니다 :

- 정량화 모델의 체계적 오류。 예를 들어 특이값 또는 샘플 수가 충분하지 않습니다。
- 분광학적 측정 방법의 체계적 오류。
- 기준 측정 절차의 체계적 오류。

유효성 검사 데이터 세트에 체계적인 오류가 있는 경우 다른 원인이 고려될 수 있습니다 :

- 예를 들어 장비의 분광학적 측정 방법의 변경。
- 예를 들어 새 실험실, 새 장비 등 기준 측정 절차의 변경。
- 예를 들어 취급, 보관 또는 운반 중에 샘플의 변경。

바이어스 수정, 특히 기울기 수정/y 절편 수정을 신중하게 적용해야 합니다。

체계적 오류가 유의하지 않은 경우 수정하지 않으면 좋겠습니다。 오류가 유의한 경우 철저히 조사해야 합니다。 가능한 경우 오류의 원인을 해결해야 합니다。 정당한 이유로 오류를 수정할 수 없는 경우 바이어스 수정 또는 기울기 수정/y 절편 수정을 적용할 수 있습니다。

신뢰할 수 있는 바이어스 추정치를 얻으려면 최소 20개의 샘플이 필요합니다。 신뢰할 수 있는 기울기 추정치를 얻으려면 최소 30개의 샘플이 필요합니다。

바이어스 수정

다음 상관관계 그래프는 바이어스 수정을 보여 줍니다。 원래 회귀선 **F**의 기울기는 변경되지 않습니다。 보정 후(회귀선 **G**) 긍정적 오류와 부정적 오류는 서로 상쇄됩니다。

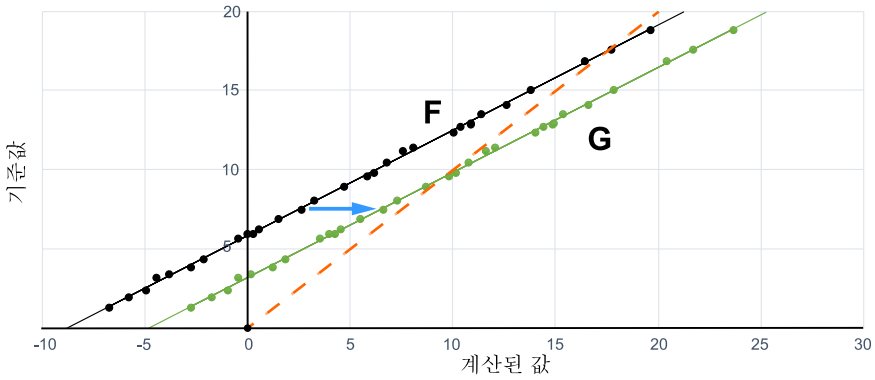


그림 34 바이어스 수정

기울기 수정/y 절편 수정

다음 상관관계 그래프는 기울기 수정/y 절편 수정을 보여 줍니다。 원래 회귀선 **H**는 기울기와 y축 절편에 의해 보정됩니다。 따라서 기울기와 바이어스가 모두 보정됩니다(회귀선 **K**)。

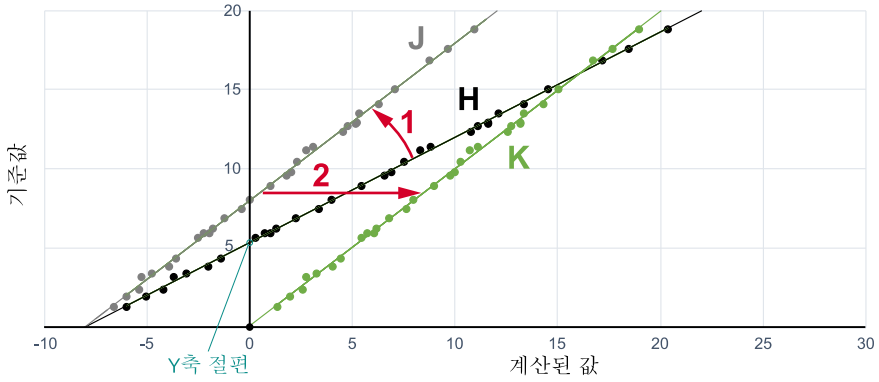


그림 35 기울기 수정/y 절편 수정

SEP

수정 데이터 세트에서 샘플은 기울기 수정/y 절편 수정의 기초를 형성합니다。 이 샘플을 근거로 OMNIS Software에서는 다음 **예측의 표준 오차 (SEP)**를 계산합니다。 분모는 각각 해당 자유도를 고려합니다 :

수정의 타입	SEP
수정 안됨	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n}}$
바이어스 수정	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-1}}$
기울기 수정/y 절편 수정	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-2}}$

여기에서 v_i는 수정 데이터 세트에서 *i*번째 샘플의 기준값에 해당하고 $\hat{v}_i$는 수정 데이터 세트에서 *i*번째 샘플의 예측된 값에 해당하며 *n*은 수정 데이터 세트에서 샘플의 수에 해당합니다.

SEP에서 임의의 오류와 체계적인 오류(기울기 및 바이어스)가 모두 포함됩니다.

4.5 동일시 및 검증

4.5.1 지원 벡터 시스템 (SVM)

식별 모델 (OMNIS Software 버전 4.0 이상)은 지원 벡터 시스템 (SVM)을 사용하여 서로 다른 제품을 분류합니다. 지원 벡터 시스템은 모니터링되는 기계 학습 알고리즘입니다. 보정 샘플을 사용하여 제품에 새 샘플을 할당하는 방법을 배웁니다.

단순성을 위해 2개 제품 사이의 분류는 아래에 설명되어 있습니다. 컨셉은 확장 가능하고, 최종 모델은 임의 수의 제품 사이에서 분류됩니다.

선형 분류

*그림 36 (왼쪽)*은 2 개의 변수가 있는 입력 데이터를 보여 줍니다.

단순성을 위해 파라미터가 지정된 스펙트럼은 2차원 가변 공간에 표시됩니다. 각 점은 스펙트럼을 나타내고, 색상은 제품의 특성을 나타냅니다.

그 제품들은 선형적으로 분리될 수 있습니다. SVM 알고리즘은 제품 사이에 초평면을 생성합니다(오른쪽 그림).

초평면은 3차원 공간에서 임의의 크기의 공간으로 수준을 일반화하는 것입니다. 초평면의 치수는 주변 공간의 크기보다 한 단계 작습니다. 2차원 공간의 초평면은 선입니다.

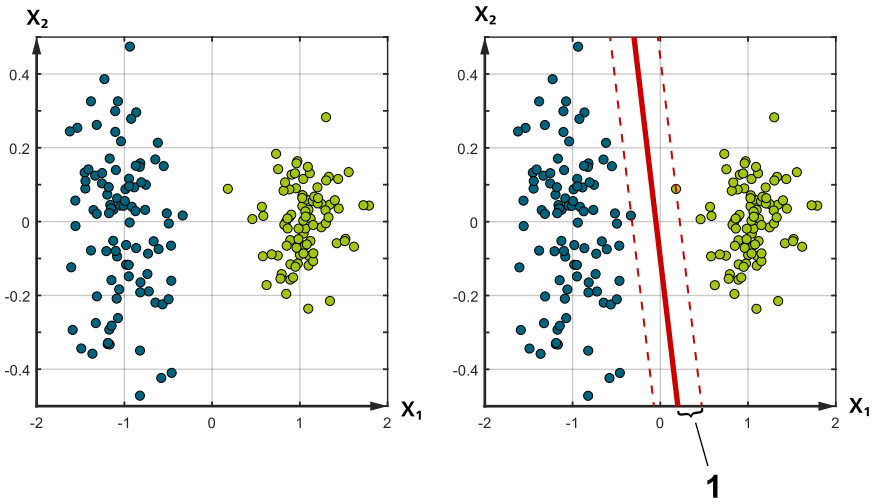


그림 36 입력 데이터(왼쪽) 및 SVM에 의해 생성된 하이퍼 레벨(오른쪽 적색 선)입니다. 값은 임의의 단위로 표시됩니다.

SVM 알고리즘은 하이퍼레벨과 각 페이지의 가장 가까운 지점 사이의 간격(**1**)을 최대화합니다. 새 스펙트럼은 동일한 공간에 표시할 수 있고 초평면의 어느 쪽에 속하느냐에 따라 제품에 할당될 수 있습니다.

초평면 정의의 경우 SVM 알고리즘은 상대 제품의 점에만 가장 가까운 점만 주의합니다. 이러한 점 또는 벡터는 초평면의 형성을 지원하고 지원 벡터라고 합니다.

점을 선형으로 분리할 수 없는 경우 - 예를 들어 특이값 때문에 - 그럼에도 불구하고 선형 분류 초평면을 측정할 수 있습니다. 이 경우 최적화 알고리즘은 각 편의 초평면에서 지원 벡터로 범위를 확장하는 것과 모든 지점이 초평면의 오른쪽임을 품질보증하는 것 사이에서 절충을 찾습니다. 정규화 파라미터는 타협을 조절하고, 따라서 초평면의 최종 위치를 조절합니다.

비선형 분류

*그림 37 (왼쪽)*에서 제품은 선형적으로 분리할 수 없습니다. 제품을 분리하려면 비선형 분류기가 필요합니다.

선형 또는 비선형 커널 기능은 데이터를 더 높은 차원의 특성 공간으로 변경합니다. 변경은 속성 공간의 데이터가 초평면에 의해 선형적으로 분리될 수 있도록 수행됩니다.

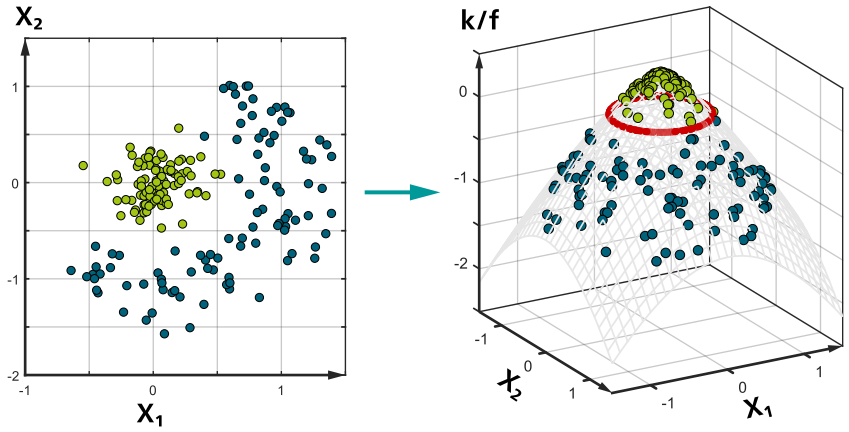


그림 37 선형적으로 분리할 수 없는 제품(왼쪽)。 추가 크기 k/f(커널 기능)는 분리를 용이하게 합니다(오른쪽)。 값은 임의의 단위로 표시됩니다。

이 그림에서 데이터는 2차원 공간에서 3차원 공간으로 변경됩니다。 한 제품의 점은 원래 수준 위로 상승하는 반면 다른 제품의 점은 아래로 이동합니다。 제품 사이의 선형 초평면은 원래 면과 매우 유사한 2차원 면입니다。 2차원으로 볼 때 결정 한계는 비선형 선이고, 이 경우 빨간색의 원형 선입니다。

여기서도 초평면은 분류기 역할을 합니다。 제품에 새 샘플을 할당할 수 있고, 이 샘플은 스펙트럼이 초평면의 어느 쪽에 속하느냐에 따라 달라집니다。

OMNIS Software는 기본 방사형 커널을 사용하여 데이터를 특성 공간으로 변경합니다。 커널은 비선형성을 조절하는 확장 파라미터를 사용합니다。

파라미터 선택

초평면의 위치를 조절하는 정규화 파라미터와 확장 파라미터에 적합한 값을 선택해야 합니다。

두 parameter 모두 지원 벡터 시스템의 일반화에 영향을 미칩니다. 즉, 보정 스펙트럼에서 새로운 알려지지 않은 스펙트럼으로 얼마나 잘 일반화할 수 있는지에 영향을 미칩니다。 예를 들어 확장 파라미터가 높은 비선형성을 제공하는 경우 초평면이 보정 스펙트럼에 너무 많이 적응할 수 있습니다(과도한 조정)。 확장 파라미터가 낮은 수준의 비선형성을 제공하는 경우, 초평면이 보정 스펙트럼에 충분히 적응하지 못할 수 있습니다 (과소 조정)。

그리드 검색은 양호한 일반화를 위해 사용됩니다。 이 알고리즘은 가능한 한 적은 시도만으로 최상의 파라미터 조합을 찾습니다 :

1. 사전 정의된 파라미터 조합이 사용됩니다。
2. 일부 선택된 파라미터 조합의 경우 SVM은 최대 정확도로 서로 다른 제품의 보정 스펙트럼을 구분하는 분류 법칙을 학습합니다。
3. 각 분류 법칙에 대해 교차 유효성 검사를 사용하여 제품 할당의 성공을 추정합니다。

4. 교차 유효성 검사의 결과에 따라 추가 파라미터 조합이 선택됩니다. SVM은 해당 분류 법칙을 다시 학습하고 교차 유효성 검사 방식은 분류 정확도를 추정합니다.
   몇 번 반복한 후 최종 파라미터 조합이 측정됩니다.
5. SVM은 최종 파라미터 조합과 모든 보정 스펙트럼을 사용하여 최종 분류 규칙을 평가합니다.

확률

식별 모델은 분류 규칙을 근거로 특정 샘플이 특정 제품에 속하는지 또는 속하지 않는지의 확률을 계산합니다. 확률은 제품 전체 및 모델 파라미터까지의 거리에 따라 달라집니다.

이렇게 계산된 확률을 통해 제품에 대한 샘플의 할당을 제어할 수 있습니다*(참조: 68페이지, "샘플의 할당 (OMNIS Software 버전 4.4 이상)")*.

잘못된 포지티브 식별을 최소화하기 위해 OMNIS Software 버전 4.4 이상에서는 각 제품에 추가적으로 인증 모델이 교육됩니다*인증 (참조: 71 페이지, 4.6 장)*.

4.5.2 샘플의 제품 제휴 예측

특정 샘플의 식별을 위해 식별 모델에서는 제품당 하나의 확률을 제시합니다*(참조: 68페이지, "확률")*.

각 확률은 0%와 100% 사이에 있는 하나의 개별 값에 해당합니다. 이 값은 제품들에 대해 100%로 합산되지는 않습니다.
이 값은 서로 상대적인 것으로 간주해야 하며 이런 접근 방식은 다양한 제품의 비교를 가능하게 합니다.

샘플의 할당 (OMNIS Software 버전 4.4 이상)

평가는 설정 가능한 **확률 스레드홀드** 및 개별 제품에 대한 인증을 통해 이루어집니다 :

1. 그 확률이 확률 스레드홀드를 초과하는 각 제품에 대해 샘플 인증이 상응하는 인증 모델을 통해 수행됩니다. 인증에 실패하는 경우 해당 제품에 대한 확률은 영으로 설정됩니다.
2. 단계 1에서 변형된 확률을 통한 평가 :
   a. 확률 스레드홀드를 초과하는 확률이 존재하지 않는 경우 식별에 실패한 것입니다(식별 상태 **식별하지 못 했기**).
   b. 하나의 개별 확률이 확률 스레드홀드를 초과하는 경우 샘플이 성공적으로 식별된 것이며 이 샘플은 상응하는 제품에 할당됩니다(식별 상태 **식별했기**).
   c. 복수의 확률이 확률 스레드홀드를 초과하는 경우 이런 예측은 다의적으로 해석되며 식별에 실패한 것입니다(식별 상태 **다의적임**).

*표 1*에서는 서로 다른 확률 스레드홀드가 사용된 평가 예시가 설명되어 있습니다. 이 예시에서 제품 A 및 C 양측에 대한 인증이 성공적으로 진행됩니다.

표 1 서로 다른 확률 스레드홀드가 사용된 평가 예시 (OMNIS Software 버전 4.4 이상)

확률	확률 스레드홀드	인증	식별 결과
제품 A: 87% 제품 B: 71% 제품 C: 68% 제품 D: 30%	90%	–	식별하지 못 했기
	80%	제품 A: 성공적 → 87%	제품 A
	70%	제품 A: 성공적 → 87% 제품 B: 실패함 → 0%	제품 A
	60%	제품 A: 성공적 → 87% 제품 B: 실패함 → 0% 제품 C: 성공적 → 68%	다의적임

샘플의 할당 (OMNIS Software 버전 4.0~4.3)

샘플의 확률은 설정 가능한 **확률 스레드홀드**를 통해 평가됩니다 :

- 확률 스레드홀드를 초과하는 확률이 존재하지 않는 경우 식별에 실패한 것입니다(식별 상태 **식별하지 못 했기**).
- 하나의 개별 확률이 확률 스레드홀드를 초과하는 경우 샘플이 성공적으로 식별된 것이며 이 샘플은 상응하는 제품에 할당됩니다(식별 상태 **식별했기**).
- 복수의 확률이 확률 스레드홀드를 초과하는 경우 이런 예측은 다의적으로 해석되며 식별에 실패한 것입니다(식별 상태 **다의적임**).

*표 2*에서는 서로 다른 확률 스레드홀드가 사용된 평가 예시가 설명되어 있습니다.

표 2 서로 다른 확률 스레드홀드가 사용된 평가 예시 (OMNIS Software 버전 4.0~4.3)

확률	확률 스레드홀드	식별 결과
제품 A: 87%	90%	식별하지 못 했기
제품 B: 72%	80%	제품 A
제품 C: 68%	70%	다의적임

4.5.3 식별 모델의 검증

식별 모델은 일반적으로 다음과 같이 검증됩니다 :

1. 제한된 수의 샘플에서 시작하여 유효성 검사 데이터 세트가 없는 모델은 샘플을 통해 보정 데이터 세트에서 테스트됩니다.
2. 샘플 수가 충분한 경우에 데이터 세트를 보정 데이터 세트와 유효성 검사 데이터 세트로 나눌 수 있습니다. 유효성 검사 데이터 세트의 샘플은 모델 개발에 사용되지 않습니다.

3. 외부 유효성 검사 데이터 세트를 위한 샘플은 다른 날에 수집되고 측정되며 필요한 경우 다른 장비로 다른 인원을 통해 수집되고 측정됩니다。

각 샘플에 대해 예측된 제품 제휴를 실제 제품 제휴와 비교합니다。 일치하는 경우 예측이 참이고(= 성공) 그렇지 않은 경우 거짓(= 실패)입니다。

검증

OMNIS Software는 식별 모델의 경우 다음 크기를 표시하여 모델이 얼마나 잘 작동하는지 나타냅니다。 이상적으로는 모든 지표가 100%입니다。

성공적으로 %(총) 모델의 전체 *정확성*을 측정합니다。 백분율은 이 질문에 대해 대답합니다: 모델이 정확하게 식별할 수 있는 샘플은 몇 개입니까 ?

$$\textit{성공적 \% (총)} = \frac{\textit{올바른 분류화}}{\textit{모든 분류화}}$$

i ***성공적 %(총)*에 각 샘플에 동일한 중량이 부여됩니다。 따라서 샘플이 많은 제품은 샘플이 적은 제품보다 백분율에 더 큰 영향을 미칩니다。**

*성공적 %*에 대한 유사한 수치를 각 제품에 사용할 수 있습니다。

동일시를 개선합니다

다음 작업을 통해 모델을 개선할 수 있습니다 :

- **확률 임계값의 조정**
  - 다수의 예측이 다의적이거나 또는 다수의 0.0% 확률이 발생하는 경우 확률 임계값을 높일 수 있습니;다。
  - 확률 임계값에 도달하지 않아서 다수의 샘플이 식별되지 않은 경우 확률 임계값을 낮출 수 있습니다。
- **파라미터화의 조정**
  더 적합한 데이터 전처리 및 파장을 선택합니다。
- **모델 계층 구조의 사용하기**
  모델 계층 구조는 식별 모델의 계층적 구조를 가능하게 합니다。
  예: 4가지 다른 제품을 가진 식별 모델은 과당과 포도당을 쉽게 구별할 수 없습니다。 만약 과당과 포도당이 하나의 제품분류로 결합된 경우 이 모델은 설탕과 다른 두 제품을 구별할 수 있습니다。 만약 샘플이 설탕으로 식별된 경우 다른 모델은 과당과 포도당을 구별합니다。 이 모델은 더 전문적이기 때문에 유사한 제품을 더 쉽게 구별할 수 있습니다。

식별되지 않은 샘플

샘플이 왜곡으로 식별되지 않은 경우 :

- 샘플 및 시료 처리의 비정상 여부를 점검합니다。
- 확률 임계값을 점검하고 필요 시 낮춥니다。
- 데이터 전처리 및 파장 선택을 점검합니다。

- 식별되지 않은 샘플의 스펙트럼을 score plot에서 점검합니다 :
  - 스펙트럼이 이미 모델에 포함되지 않은 경우 : 스펙트럼을 각 제품의 유효성 검사 데이터 세트에 추가합니다.
  - score plot에서 점검할 스펙트럼의 score를 보정 데이터 세트의 스펙트럼 score와 비교합니다.

  샘플이 outlier가 아니지만 그 변형의 대표성이 보정 데이터 세트에서 부족한 경우 보정 데이터 세트를 상응하게 확대해야 합니다.

4.6 인증

인증 모델(OMNIS Software 버전 4.4 이상)은 이 샘플 그룹을 다른 샘플 그룹과 구분합니다. 예를 들어 이 모델은, 사용 가능한 샘플(포지티브 샘플)을 사용할 수 없는 샘플(네거티브 샘플)에서 구분하는 용도에 적합합니다.

4.6.1 인증 모델의 계산

인증 모델의 계산은 식별 모델에서와 유사하게 진행됩니다*지원 벡터 시스템 (SVM) (참조: 65 페이지, 4.5.1 장)*. 하지만 인증을 위한 보정 데이터 세트에는 단 하나의 샘플 종류만 포함됩니다(포지티브 샘플).

지원 벡터 시스템(SVM)은 입력 데이터를 고차원 공간으로 변환합니다. 정규화 파라미터는 초평면의 위치를 결정하며 반면 스케일링 파라미터는 비선형성의 정도를 결정합니다. 그리드 검색은 적합한 예측을 검출합니다. 이 예측의 결정 한계는 인증 모델의 기초를 형성합니다.

4.6.2 인증 모델의 유효성 검사

진행 과정

일반적으로 인증 모델은 단계적으로 개발되고 유효성이 검사됩니다. 이 과정에서 샘플의 수는 점진적으로 증가됩니다 :

1. 포지티브 유효성 검사 데이터 세트 :
   a. 포지티브 샘플의 수가 제한된 경우 처음에는 포지티브 유효성 검사 데이터 세트가 생성되지 않습니다.
   b. 충분한 수의 포지티브 샘플을 사용할 수 있게 되면 자동 데이터 세트의 분할을 통해 포지티브 유효성 검사 데이터 세트가 생성됩니다.

   네거티브 유효성 검사 데이터 세트 :
   a. 수집된 네거티브 샘플은 네거티브 유효성 검사 데이터 세트에 할당됩니다.
   b. 스펙트럼 특이값을 인식하고 네거티브 유효성 검사 데이터 세트에 할당하기 위해, 네거티브 스펙트럼의 자동 측정을 사용할 수 있습니다*스펙트럼적인 특이값 - 알고리즘 (참조: 84 페이지, 6.5 장)*.

2. 포지티브 및 네거티브 유효성 검사 데이터 세트를 위한 샘플은 다른 날에 수집되고 측정되며 필요한 경우 다른 장비로 다른 인원을 통해 수집되고 측정됩니다。

포지티브 및 네거티브 유효성 검사 데이터 세트의 샘플은 모델의 계산에 사용되지 않습니다。

유효성 검사

각 샘플에 대해 인증 모델에서 하나의 결과가 측정됩니다(포지티브 또는 네거티브)。 보정 데이터 세트의 샘플 및 유효성 검사 데이터 세트의 샘플에 대해 포지티브 결과가 요구됩니다。 네거티브 유효성 검사 데이터 세트의 샘플에 대해 네거티브 결과가 요구됩니다。 이 결과가 요구에 부합하지 않는 경우 예측이 참이고(= 성공) 그렇지 않은 경우 거짓(= 실패)입니다。

OMNIS Software에는 인증 모델에 대해 다음 값이 표시됩니다：

*성공적 % (총)*는 모델의 전체 정확성을 측정합니다。

$$\textit{성공적 \% (총)} = \frac{\textit{올바른 예측}}{\textit{모든 예측}}$$

*성공적 %*의 유사한 수치는 각 데이터 세트에 대해 사용할 수 있습니다。 이상적으로는 모든 지표가 100%입니다。

인증 개선

적합한 파라미터 지정(데이터 전처리 및 파장 선택)은 인증 모델을 개선할 수 있습니다。

인증되지 않은 샘플

샘플이 왜곡으로 인증되지 않은 경우：

- 샘플 및 시료 처리의 비정상 여부를 점검합니다。
- 데이터 전처리 및 파장 선택을 점검합니다。
- 인증되지 않은 샘플의 스펙트럼을 score plot에서 점검합니다：
  - 스펙트럼이 이미 모델에 포함되지 않은 경우： 스펙트럼을 포지티브 유효성 검사 데이터 세트에 추가합니다。
  - score plot에서 점검할 스펙트럼의 score를 보정 데이터 세트의 스펙트럼 score와 비교합니다。

  샘플이 outlier가 아니지만 그 변형의 대표성이 보정 데이터 세트에서 부족한 경우 보정 데이터 세트를 상응하게 확대해야 합니다。

5 예측

5.1 수량화

알 수 없는 관심있는 parameter를 가진 샘플의 예측은 다음과 같이 진행됩니다 :

1. 샘플의 스펙트럼을 삽입됩니다。
2. 정량화 모델은 보정 데이터 세트의 스펙트럼과 동일한 데이터 전처리 및 파장범위를 적용합니다。
3. 결과 스펙트럼에 기초하여 모델은 관심있는 parameter를 예측합니다。

정량화 모델을 계산할 때 보정 스펙트럼과 기준값이 평균으로 중심화 되었습니다。 이것은 물론 위에서 설명한 절차에서 고려됩니다。

5.1.1 특이값 및 결과 모니터링

정량적 관심있는 parameter의 예측 시 다양한 유형의 특이값이 인식되고 결과가 모니터링될 수 있습니다 :

특이값/모니터링		원인 (예시)
스펙트럼적인 특이값	Hotelling T^2	화학적 성분의 농도가 보정 샘플의 농도 범위 밖에 있습니다。
	Q residuals	이 샘플에 보정 샘플에 존재하지 않는 구성요소가 포함되어 있습니다。
Nearest Neighbor 특이값 (옵션, OMNIS Software 버전 4.2 이상)		보정 샘플의 이 조합에서는 불가능한 샘플 유형이 이 샘플에 포함되어 있습니다。
결과 모니터링 (옵션)		관심있는 parameter의 결과값이 고객이 선택한 수용 기준 밖에 있습니다。

스펙트럼적인 특이값

스펙트럼 특이값은 다음과 같이 측정됩니다 :

1. 소프트웨어는 정량화 모델(PLS 모델)을 사용하여 스펙트럼에 대한 hotelling T^2 및 Q residuals의 값을 계산합니다。
2. 스펙트럼의 T^2 또는 Q residuals 값이 모델에 의해 계산된 상응하는 임계 값보다 큰 경우 샘플은 적용된 모델을 기반으로 특이값으로 표시됩니다*(참조: 85 페이지, "예측 (수량화)에서 특이값 평가")*。

*T*² 및 Q residuals 값은 OMNIS Software에서 변수로 사용할 수 있습니다. 이것은 정량화 모델의 PLS influence plot 값과 비교할 수 있습니다. 점선은 임계값을 나타냅니다.

Nearest Neighbor 특이값

(OMNIS Software 4.2 버전 이상)

이상적인 경우 보정 샘플은 샘플 유형의 모든 가능한 조합을 커버합니다. 실제 현장에서는 몇몇 조합이 더욱 빈번하게 발생되며 다른 조합은 거의 나타나지 않습니다. 이에 상응하게 보정 샘플은 잠재 변수 공간에서 불균일하게 분포되어 있습니다. 몇몇 영역에서는 다수의 보정 샘플이 존재하고 그 사이에는 틈새가 형성되어 있습니다.

알 수 없는 샘플의 스펙트럼이 보정 샘플의 틈새에 존재하는 경우 예측 결과가 유효하지 않거나 또는 부정확할 수 있습니다. 이런 경우를 인식하기 위해, 알 수 없는 샘플 *i*와 각 보정 샘플 *u* 사이의 거리 *D*가 계산됩니다 :

$$D = \sqrt{(\mathbf{s_i} - \mathbf{s_u})^{\mathbf{t}} (\mathbf{s_i} - \mathbf{s_u})}$$

여기에서 $\mathbf{s_i}$는 알 수 없는 샘플 *i*의 score에 해당하며 $\mathbf{s_u}$는 보정 샘플 *u*의 score에 해당합니다. score는 표준화된 직교 상태입니다.

가장 짧은 거리는 다음 보정 샘플에 대한 거리이며 **Nearest Neighbor Distance** (NND)로 불립니다 :

NND 값이 특정한 NND 한계값을 초과하는 경우 알 수 없는 샘플은 Nearest Neighbor 특이값으로 불립니다.

NND 한계값은 다음과 같이 측정됩니다 :

1. 각각의 보정 샘플에 대해 하나의 NND 값이 측정됩니다. 이 값은 다음에 놓인 다른 보정 샘플에 대한 거리에 해당합니다.
2. 모든 보정 샘플의 최대 NND 값은 NND 한계값입니다.

알 수 없는 샘플의 NND 값 및 NND 한계값은 OMNIS Software에서 변수로 사용할 수 있습니다.

결과 모니터링

OMNIS Software에서는 예측 결과의 범위에 대한 경고 한계 및 조절 한계를 정의하기 위해 결과 모니터링을 사용할 수 있습니다. 옵션으로서, 정의된 한계의 위반 시 수행되는 작업을 정의할 수 있습니다.

5.2 식별 및 검증

샘플의 동일시 또는 검증 시 다음과 같이 진행됩니다 :

1. 샘플 스펙트럼이 기록됩니다.
2. 식별 모델은 보정 데이터 세트의 스펙트럼과 동일한 데이터 전처리 및 파장 선택을 적용합니다.
3. 이 식별 모델은 *샘플의 제품 제휴 예측 (참조: 68 페이지, 4.5.2 장)*의 스펙트럼을 평가합니다.
4. 식별 모델이 모델 계층 구조의 일부이고 다른 식별 모델이 측정된 제품과 링크되어 있는 경우 이 모델이 실행됩니다. 하나 또는 복수의 정량화 모델이 측정된 제품과 링크되어 있는 경우 이 모델이 실행됩니다.
5. 식별 결과 또는 검증 결과가 표시되고 모델 계층 구조에서는 상황에 따라 정량화 결과가 표시됩니다.

동일시 상태

- 식별했기
  동일시 성공했습니다.
- 다의적임
  여러 제품이 가능성의 스레시홀드를 초과합니다. 동일시를 실패했습니다.
- 식별하지 못 했기
  가능성의 스레시홀드를 초과하는 제품은 없습니다. 동일시를 실패했습니다.

검증 상태

- 성공적
  샘플이 성공적으로 식별되었으며 결과는 예측된 제품과 일치합니다.
- 실패함
  검증에 실패했습니다.

5.3 인증

샘플의 인증 시 다음과 같이 진행합니다 :

1. 샘플 스펙트럼이 기록됩니다。
2. 인증 모델은 보정 데이터 세트의 스펙트럼과 동일한 데이터 전처리 및 파장 선택을 적용합니다。
3. 모델은 결과 스펙트럼을 사용하여 샘플을 인증합니다。
4. 인증 결과가 표시됩니다。

인증 상태

- 성공적
- 실패함

6 부록

6.1 선형 회귀의 예

일변량 선형 회귀

가장 간단한 경우에 혼합물은 하나의 흡수기만을 가지고 있고 스펙트럼은 하나의 피크만을 가지고 있습니다. 흡수기의 농도가 다른 샘플은 흡수도 값이 다른 피크를 가집니다 *Beer Lambert 법칙 (참조: 6 페이지, 2.2.1 장)*.

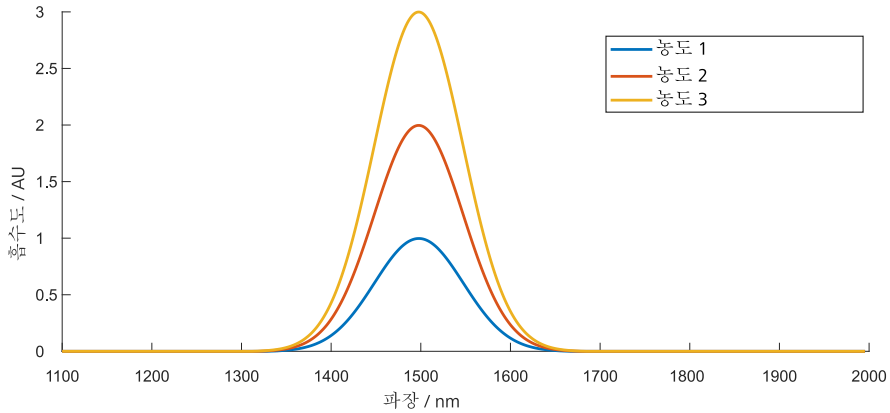


그림 38 1,500 nm의 피크에서 3개의 샘플에 대한 모델링 데이터.

1,500 nm에서 측정한 3개의 흡수도 값은 흡수기의 농도에 반하여 적용할 수 있습니다.

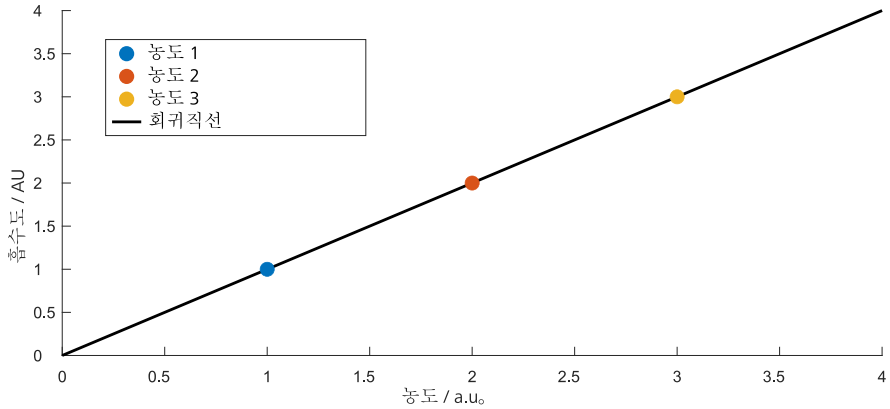


그림 39 흡광도 값과 농도 사이의 관계.

Beer-Lambert 법칙에 따르면, 흡수도 값과 농도 사이의 관계는 선형입니다. 3개 샘플 세트를 사용하여 선형 회귀를 수행할 수 있습니다.

변수는 1개만 있으므로 (파장 1개) 일변량 선형 회귀입니다. 이 회귀를 정량화 모델로 사용할 수 있습니다. 농도를 알 수 없는 샘플의 경우 A 흡수도는 1,500 nm에서 측정됩니다. 회귀 직선으로부터 흡수기의 해당 c 농도는 다음과 같이 나타납니다 :

$$c = bA$$

계수 b는 회귀선의 기울기와 동일하고 일정합니다.

모든 샘플은 동일한 몰 멸종 계수를 가진 동일한 흡수기를 포함해야 합니다. 또한 모든 흡수도 측정은 동일한 층 두께로 수행해야 합니다.

다변량 선형 회귀

실제 혼합물은 둘 이상의 흡수기를 포함합니다. 삽입된 스펙트럼은 모든 흡수기 스펙트럼의 합계입니다 *Beer Lambert 법칙 (참조: 6페이지, 2.2.1 장)*.

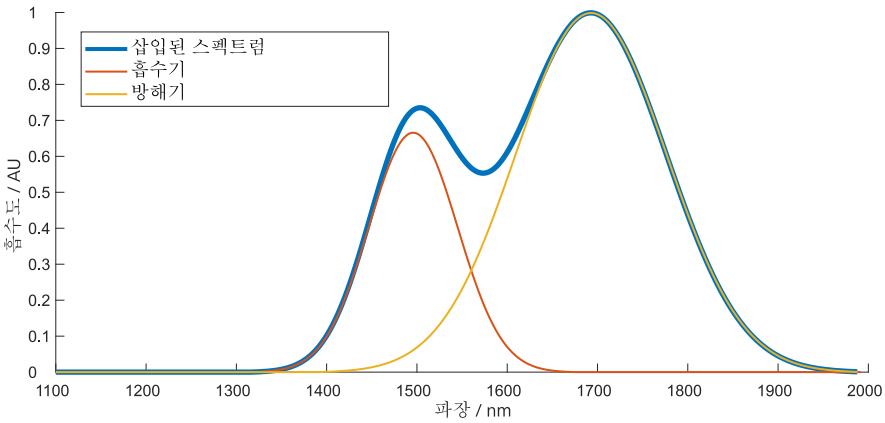


그림 40 2 가지 성분으로 모델링된 데이터. 흡수기 (적색 선)를 정량화해야 합니다.

삽입된 스펙트럼 (청색 선)은 순수 흡수기 스펙트럼과 겹치는 간섭 스펙트럼의 합계입니다. 1,500 nm에서 측정된 흡수도 값은 흡수기의 흡수도뿐만 아니라 간섭기의 흡수도입니다. 관심있는 parameter는 1,500 nm에서 단일 측정으로 정량화할 수 없습니다. 간섭이 존재하는지 여부 및 측정이 신뢰할 수 있는지 여부도 알 수 없습니다.

예를 들어 2개의 파장이 1,500 nm 및 1,700 nm에서 측정되면 어떻게 됩니까 ? 파장 1, A_1에서 측정한 흡수도는 순수 흡수기 신호 A_1^{a} (색인 a = 흡수기)와 순수 간섭기 신호 A_1^{f} (색인 f = 간섭기)의 합계입니다. 파장 2, A_2에서 측정한 흡수도에도 동일하게 적용됩니다 :

$$A_1 = A_1^{\mathrm{a}} + A_1^{\mathrm{f}} = \varepsilon_1^{\mathrm{a}} c_{\mathrm{a}} + \varepsilon_1^{\mathrm{f}} c_{\mathrm{f}}$$
$$A_2 = A_2^{\mathrm{a}} + A_2^{\mathrm{f}} = \varepsilon_2^{\mathrm{a}} c_{\mathrm{a}} + \varepsilon_2^{\mathrm{f}} c_{\mathrm{f}}$$

여기서 $\varepsilon_1^{\mathrm{a}}$와 $\varepsilon_1^{\mathrm{f}}$는 각각 흡수기 또는 간섭에 대한 파장 1의 몰 멸종 계수에 해당하고, c_{a}와 c_{f}는 각각 흡수기 또는 간섭의 농도에 해당합니다.

위의 방정식에서 층두께 l Beer-Lambert 법칙에서 제외됩니다. 이것은 나중에 대수학을 조금 더 쉽게 합니다. 물론 층두께는 모든 샘플에서 동일해야 합니다. 따라서 방정식의 흡수도는 cm당 흡수도입니다.

방정식은 다음과 같이 행렬 형식으로 작성할 수 있습니다 :

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \varepsilon_1^{\mathrm{a}} & \varepsilon_1^{\mathrm{f}} \\ \varepsilon_2^{\mathrm{a}} & \varepsilon_2^{\mathrm{f}} \end{bmatrix} \begin{bmatrix} c_{\mathrm{a}} \\ c_{\mathrm{f}} \end{bmatrix}$$

따라서 :

$$\begin{bmatrix} c_{\mathrm{a}} \\ c_{\mathrm{f}} \end{bmatrix} = \begin{bmatrix} \varepsilon_1^{\mathrm{a}} & \varepsilon_1^{\mathrm{f}} \\ \varepsilon_2^{\mathrm{a}} & \varepsilon_2^{\mathrm{f}} \end{bmatrix}^{-1} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

흡수기의 농도에 대한 해결책이 있습니다 :

$$c = c_{\mathrm{a}} = \frac{\varepsilon_2^{\mathrm{f}}}{\varepsilon_1^{\mathrm{a}}\varepsilon_2^{\mathrm{f}} - \varepsilon_1^{\mathrm{f}}\varepsilon_2^{\mathrm{a}}} A_1 + \frac{-\varepsilon_1^{\mathrm{f}}}{\varepsilon_1^{\mathrm{a}}\varepsilon_2^{\mathrm{f}} - \varepsilon_1^{\mathrm{f}}\varepsilon_2^{\mathrm{a}}} A_2$$

따라서 두 파장의 흡수도를 측정하고 각 흡수도를 상수로 곱하여 간섭이 있더라도 흡수기의 농도를 계산할 수 있습니다。

상수는 몰 멸종 계수를 나타내고 표에서 찾을 수 있습니다。 그러나 이것은 현실에서 일어나지 않습니다。 대신에 PLS 회귀와 같은 다변량 선형 회귀를 통해 선형 방정식을 보정하고 해결합니다。 따라서 상수를 회귀 계수 b_1 및 b_2라고 합니다 :

$$c = b_1 A_1 + b_2 A_2$$

2개 이상의 흡수기

위에 표시된 것처럼 1개의 흡수기를 사용하는 경우에 1개의 파장에서 흡수도를 측정할 수 있습니다。 흡수기 2개를 사용하는 경우에 2개의 파장에서 흡수도를 측정할 수 있습니다。

이것은 일반화될 수 있습니다。 더 많은 흡수기는 다른 파장 *i*에서 더 많은 A_i 흡수도 값을 필요로 합니다. 그것은 여전히 선형적인 관계입니다 :

$$c = b_1 A_1 + b_2 A_2 + \cdots + b_n A_n$$

정량화 모델의 개발하기

위의 방정식이 알 수 없는 샘플의 농도를 예측하기 전에 계수 b_1, b_2 등을 측정해야 합니다。 이를 위해 보정 단계가 필요합니다。 관심있는 parameter의 농도가 다른 여러 샘플이 측정됩니다。

나중에 PCA 및 PLS에 사용되는 용어에 따라 *c*는 *y* 및 *A*로 *x*로 대체할 수 있습니다。 그런 다음 각 보정 샘플에 대해 명시된 위의 방정식은 다음과 같습니다 :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,f} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,f} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,f} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

여기서 *n*은 샘플 수, *f* 파장 수, y_1은 기준 방법(예: 적정)으로 측정한 샘플 1의 기준값, $x_{1,1}$ 파장 1에서 측정한 샘플 1의 흡수도 및 $\{x_{1,1}\ldots x_{1,f}\}$ 스펙트럼의 샘플 1 *f* 파장 1에 측정하고 해당합니다. $b_1\ldots b_n$은 회귀 계수에 해당하

고, $e_1 \ldots e_n$은 회귀 계수가 측정 데이터를 얼마나 잘 모델링하는지 나타내는 오류 항입니다.

이것은 더 작은 행렬 형태로 나타냅니다 :

$$\mathbf{y} = \mathbf{X^t p} + \mathbf{e}$$

X는 행렬 $f \times n$으로 정의됩니다. $\mathbf{X^t}$는 변환된 행렬 **X**입니다. 즉, 행과 열이 교환되어 위의 행렬 $n \times f$ 얻습니다. 예측 벡터 **p**는 위의 회귀 계수 **b**에 해당합니다.

예측 벡터 **p**를 사용하는 경우 스펙트럼 **x** 를 기준으로 새 샘플의 관심있는 parameter를 예측할 수 있습니다. $\hat{y}$ 계산된 값은 :

$$\hat{y} = \mathbf{x^t p}$$

다변량 선형 회귀의 목적은 최소 오차항으로 이어지는 회귀 계수를 측정하는 것입니다. 그러나 다중 선형 회귀(MLR)에서 파장 수보다 많은 수의 보정 샘플이 필요합니다. 또 다른 장애는 변수 사이의 높은 상관 관계입니다.

분광적 예측의 경우 다른 method를 사용할 수 있습니다. PCA를 사용하는 경우 데이터 볼륨을 크게 줄이고 상관 관계를 완전히 제거할 수 있습니다. PLS 회귀에서 샘플의 기준값도 주의됩니다.

6.2 PCA-알고리즘

주요 구성 요소 분석 (**PCA**, 영어로 *principal component analysis*)은 자동 데이터 세트의 분할 및 스펙트럼적인 특이값 감지에 사용됩니다 *주요 구성 요소 분석 (PCA) (참조: 31 페이지, 4.2 장)*.

OMNIS Software는 매개 변수화된 스펙트럼과 중간값 스펙트럼의 주요 구성 요소 분석에 대해 **단일 값 분해** (**SVD**, 영어로 *singular value decomposition*)를 수행합니다. 원래 데이터 행렬 **X** (보정 데이터 세트의 스펙트럼)를 3개의 행렬로 나누고 n개의 주요 구성 요소를 계산합니다 :

$$\mathbf{X} = \mathbf{L\Sigma S^t}$$

X는 원래 스펙트럼 데이터 (*f* 파장과 *n* 샘플이 있는 행렬 $f \times n$), **L** loading (행렬 $f \times n$), $\mathbf{\Sigma}$ 단일 값 (행렬 $n \times n$), **S**는 score (행렬 $n \times n$)에 해당합니다. 방정식은 그래픽 형식으로 표시할 수 있습니다 *(참조: 81 페이지, 그림 41)*.

Loading **L**은 파장 공간의 축을 주요 구성 요소 공간의 축에 투영합니다. **L**의 열 벡터는 주축 또는 주방향입니다.

Score **S**는 주축에 대한 원래 스펙트럼의 투영입니다. **S**의 각 행 벡터는 특정 스펙트럼을 포함합니다.

$\mathbf{\Sigma}$ 단일 값 σ_i의 대각선 행렬입니다. 단일 값은 각 주요 구성 요소에 의해 선언된 다양성을 설명합니다. 규약에 따르면, 그것들은 $\sigma_1 > \sigma_2 > \ldots > \sigma_n$으로 분류됩니다.

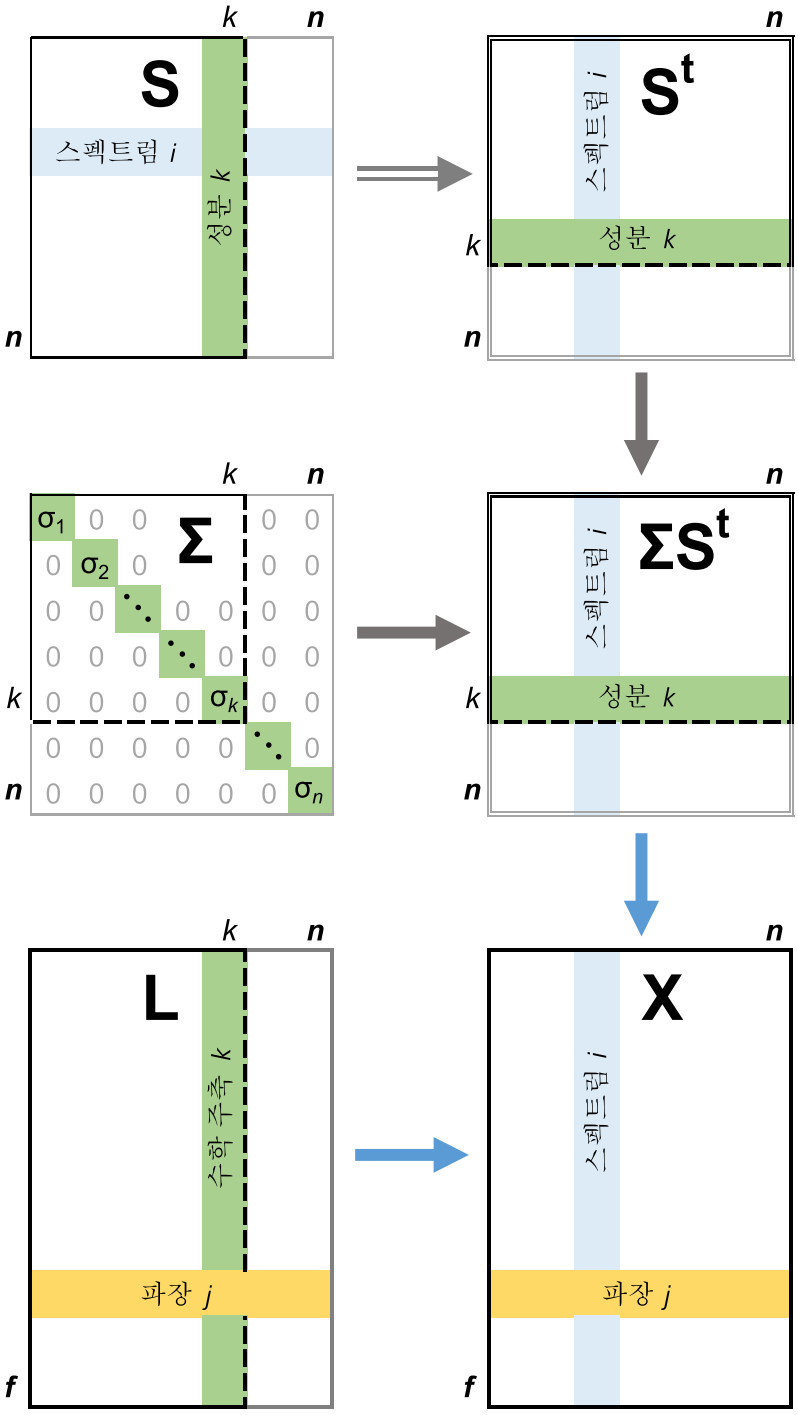


그림 41 *그래픽 형식의 단일 값 분해 방정식. 점선은 주요 구성 요소가 k인 모델의 짧은 행렬을 나타냅니다。 N-k의 생략된 주요 구성 요소의 정보는 잔류 행렬(그림에 표시되지 않은 행렬 f × n)로 들어갑니다。*

잔류 행렬

PCA 모델은 계산된 *n* 주요 구성 요소 중 처음 몇 개만 사용합니다. *그림 에서41*에서 *n*개의 주요 구성 요소 중 첫 번째 *k*가 나와 있습니다. 원래 데이터 **X** 는 모델에 의해 설명되는 데이터가 아니라 모델에 의해 설명되는 데이터로 나눌 수 있습니다 :

$$\mathbf{X} = \mathbf{L_a}\mathbf{\Sigma_a}\mathbf{S_a^t} + \mathbf{E}$$

여기서 **X**는 원래 스펙트럼 데이터(행렬 $f \times n$), $\mathbf{L_a}$ (행렬 $f \times k$)의 첫 번째 k 열 **L**, $\mathbf{\Sigma}_a$ (대각 행렬 $k \times k$)의 첫 번째 k 단일 값, $\mathbf{S_a}$ (주요 구성 요소가 없는 행렬 $n \times k$ 및 k 포함)의 첫 번째 k 열 **S**, 및 **E**의 잔류 행렬 (행렬 $f \times n$)에 해당하고, 모델에 의해 설명될 수 없는 **X** 의 모든 스펙트럼적인 변동을 포함하고 모델에 의해 설명될 수 없습니다.

일반적으로 $k \ll n \ll f$.입니다. 예를 들어: k = 3개의 주요 구성 요소, n = 100 개의 샘플 및 f = 2,500 개의 파장.

잔류 행렬 **E**의 각 $\mathbf{e}_i$ 열에서 **잔류**라고 하는 PCA 공간에 대한 스펙트럼 *i*의 직교 거리가 표시됩니다. 모델이 더 많은 주요 구성 요소를 사용할수록 잔류가 더 작습니다.

6.3 PLS-알고리즘

부분 최소 제곱 회귀 (**PLS 회귀**, 영어로 *partial least squares regression*) 정량화 모델을 계산하는 데 사용됩니다 *PLS 회귀 (참조: 53페이지, 4.4.1 장)*.

PLS 두 개의 데이터 블록이 포함됩니다 :

- 파라미터화된 평균 중심 행렬 **X** (스펙트럼)입니다.
- 평균 중심 **y** 벡터(기준값)입니다.

PLS는 행렬 **X**를 2개의 행렬로 분해합니다 :

$$\mathbf{X} = \mathbf{L}\mathbf{S}^\mathbf{t} + \mathbf{Z}$$

여기서 **X** (*f* 파장과 *n* 샘플이 있는 행렬 $f \times n$), **L**은 전처리 및 평균 중심 스펙트럼, L은 loading (접착된 변수가 있는 행렬 $f \times k$ 또 k 잠재 변수 포함), **S** score (행렬 $n \times k$), **Z**는 모든 **X** 스펙트럼적인 변동을 포함하는 잔류 행렬 (행렬 $f \times n$)에 해당합니다. 모델에 의해 설명될 수 없습니다.

PCA의 경우 score 행렬 **S**가 **X**의 다양성을 설명하는 반면 PLS의 경우 score 행렬 **S**는 **X**와 **y** 사이의 공다양성을 설명합니다. PLS는 score에 의해 선언된 공분산을 최대화합니다. 즉, score는 **X**의 다양성을 가장 잘 설명할 뿐만 아니라 기준값과 가능한 가장 높은 상관 관계를 나타냅니다.

X와 **y** 사이의 공분산을 극대화하기에 대해 PLS 알고리즘은 **X**와 **y** 사이의 데이터를 교환합니다. 따라서 **X**와 **y**는 하나의 통합된 시스템으로 병합됩니다. 여기서 score **S**는 회귀 계수 **b**를 얻기에 대해 기준값 **y**에 대해 회귀됩니다 :

$$\mathbf{y} = \mathbf{Sb} + \mathbf{e}$$

여기서 **e**는 모델에 의해 설명될 수 없는 **y**의 모든 참조 값 변동을 포함하는 잔류 벡터입니다.

예측

회귀 계수 **b**에서 **p** 예측 벡터를 측정할 수 있습니다. 새로운 샘플의 분석 매개변수를 예측하기에 대해 $\hat{y}$ 예측 벡터 **p**와 전처리된 평균 중심 스펙트럼 **x**가 사용됩니다 :

$$\hat{y} = \mathbf{x}^{\mathbf{t}}\mathbf{p}$$

OMNIS Software는 SIMPLS 알고리즘과 단일 기준값(PLS-1)을 사용하여 PLS를 구현합니다.

6.4 Hotelling T² 및 Q residuals

Hotelling T^2 및 Q residuals은 PCA 또는 PLS 모델의 스펙트럼을 특징짓습니다. 특히, 그들은 가능한 특이값을 동일시에 대해 도움이 됩니다 *(참조: 44페이지, "Hotelling T² 및 Q residuals")*.

Hotelling T²

마할라노비스 거리는 모델의 중심에서 스펙트럼의 편차의 크기를 측정하는 것입니다. 거리가 정규화됩니다. 각 주요 구성 요소 또는 잠재 변수는 동일한 중량을 가집니다.

스펙트럼 또는 score가 정규 분포를 따른다고 가정하는 경우, 마할라노비스의 제곱된 거리인 MD^2는 Hoteling의 T^2 분포를 따릅니다 :

$$\mathrm{MD}^2 \sim T^2$$

스펙트럼 *i*의 제곱 마할라노비스 거리는 첫 번째 *k* 주요 구성 요소 또는 잠재 변수에 대한 정규화된 score의 제곱합입니다 :

$$\mathrm{MD}_i^2 = \mathbf{s_i}\mathbf{s_i^t} = \sum_{a=1}^{k} s_{i,a}^2$$

여기서 $\mathbf{s_i}$는 감소화된 행렬의 **S**, $s_{i,a}$의 *i* 번째 행은 스펙트럼 *i* 주요 구성 요소 (또는 잠재 변수)에 대한 정규화된 score와 *a*, 및 *k* 사용된 주 구성 요소 또는 잠재 변수의 수에 해당합니다.

MD^2는 단순히 T^2라고 불릴 수 있습니다. $T^2 = 0$의 스펙트럼은 중간값 모델의 중심에 투영되고 즉, 모든 score가 중심에 있습니다. 스펙트럼의 초평면에 대한 투영이 중심에서 멀어질수록 T^2 값이 더 높습니다. 그 모델은 중심 가까이에 가장 적합합니다. 중심에서 멀리 떨어진 곳에서 모델이 제대로 적합하지 않을 수 있습니다.

예를 들어 **특이값 레벨** 하는 가설검정의 경우 5 %의 유의 수준을 설정할 수 있습니다 *스펙트럼적인 특이값 - 알고리즘 (참조: 84페이지, 6.5장)*.

Q 잔류

스펙트럼 *i*의 Q 잔류는 PCA 또는 PLS 모델에서 스펙트럼까지의 제곱 거리입니다 :

$$Q_i = \mathbf{e}_i^t \mathbf{e}_i = \sum_{j=1}^{f} e_{i,j}^2$$

여기서 *i* 번째 열은 스펙트럼 $\mathbf{e}_i$에 대한 잔류 행렬 **E**, $e_{i,j}$ 잔류 *i*, 및 *f*의 파장 수에 해당합니다.

Q residuals 모델에 의해 설명되지 않는 변동을 나타냅니다. Q 잔류가 높은 경우 예를 들어 측정된 샘플에 다른 물질이 포함된 경우 스펙트럼이 모델에 일치할 수 있음을 나타냅니다.

예를 들어 **특이값 레벨** 하는 가설검정의 경우 5 %의 유의 수준을 설정할 수 있습니다 *스펙트럼적인 특이값 - 알고리즘 (참조: 84 페이지, 6.5 장)*.

6.5 스펙트럼적인 특이값 - 알고리즘

스펙트럼적인 특이값의 검출은 모집단과 다른 스펙트럼을 식별합니다 *(참조: 45 페이지, "스펙트럼적인 특이값의 감지")*. 알고리즘은 측정된 스펙트럼의 hotelling T^2에 대한 값 또는 Q 잔류에 대한 값 임의의 또는 체계적인 변경의 결과인지 평가합니다.

모델 개발에서 스펙트럼적인 특이값의 감지

1. 파라미터 지정은 다음과 같이 고려됩니다 :
   a. OMNIS Software 버전 4.2 이상: 사용자는 파라미터 지정(데이터 전처리 및 파장 선택)을 사용하는지 또는 사용되지 않는지 여부를 결정합니다.
   파라미터 지정의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다.
   b. OMNIS Software 버전 3.3에서 OMNIS Software 버전 4.1까지: 사용자는 데이터 전처리가 고려되는지 또는 고려되지 않는지 여부를 결정합니다. 파장 선택 및 데이터 전처리의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다.
   c. OMNIS Software 버전 3.2까지: 데이터 전처리는 이상값 검출 시점에까지 정의된 바와 같이 고려됩니다. 파장 선택 및 데이터 전처리의 나중 변경은 데이터 세트의 분할에 영향을 미치지 않습니다.
2. 스펙트럼 목록에서 스펙트럼적인 특이값의 검출은 모든 평균값 중심 스펙트럼의 PCA 모델에 기초합니다 *주요 구성 요소 분석 (PCA) (참조: 31 페이지, 4.2 장)*. 테스트할 스펙트럼도 PCA 모델에 포함됩니다. 주요 구성 요소의 수는 선언된 다양성이 최소 95%가 되도록 선택됩니다.

3. **Hotelling T^2 특이값**:
   H. Hotelling, *The Generalization of Student's Ratio*, The Annals of Mathematical Statistics 2권, 3번 (1931년 8월), 360–378쪽에 따른 가설검정은 hotelling T^2 값을 *(참조: 83페이지, "Hotelling T^2")* 기반으로 합니다.
   a. 귀무 가설은 검사할 스펙트럼의 T^2 값이 PCA 모델의 T^2 값과 일치한다는 것입니다. 귀무 가설이 참인 경우 T^2는 Hotelling의 T^2 분포를 따릅니다. 분포는 확장된 F 분포라고 할 수 있습니다 :

   $$T^2 \sim \frac{k(n-1)}{n-k} F_{k,n-k}$$

   여기서 k는 주요 구성 요소의 수, n은 스펙트럼과 $F_{k,n-k}$의 수, k 및 $n-k$는 parameter 가진 F 분포입니다.
   b. 조정 가능한 특이값 레벨은 귀무 가설이 참인 경우 (타입 I 오류라고 함) 거부 가능성을 조절합니다. 표준값은 5%입니다.
   c. 이 분포와 유의 수준을 기준으로 T^2에 대한 임계값이 계산됩니다.
   d. 스펙트럼의 T^2 값이 임계 값보다 큰 경우 귀무 가설이 거부되고 스펙트럼이 잠재적 특이값으로 식별됩니다.
4. **Q residuals 특이값**
   J. E. Jackson 및 G. S. Mudholkar, *Control Procedures for Residuals Associated With Principal Component Analysis*, Technometrics 21권, 3번 (1979년 8월), 341–349쪽에 따른 가설검정은 Q residuals를 기반으로 *(참조: 84페이지, "Q 잔류")* 수행됩니다.
   a. 귀무 가설은 검사할 스펙트럼의 Q 잔류이 PCA 모델의 Q 잔류 값과 일치한다는 것입니다. 귀무 가설이 참인 경우 거의 정규 분포 Q 잔류 테스트 통계를 생성할 수 있습니다.
   b. 조정 가능한 특이값 레벨은 귀무 가설이 참인 경우 (타입 I 오류라고 함) 거부 가능성을 조절합니다. 표준값은 5%입니다.
   c. 이 테스트 통계와 유의 수준을 기준으로 Q에 대한 임계값이 계산됩니다.
   d. 스펙트럼의 Q 잔류 값이 임계 값보다 큰 경우 귀무 가설이 거부되고 스펙트럼이 잠재적 특이값으로 식별됩니다.

예측 (수량화)에서 특이값 평가

수량화에 예측 동안 특이값 평가를 사용할 수 있습니다 :

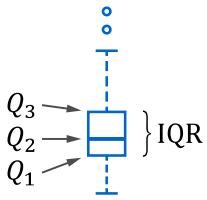
1. 준비 단계 :
   a. 정량화 모델은 위와 동일한 알고리즘을 사용합니다. 그러나 그 근거는 보정 데이터 세트의 모든 평균 중심 스펙트럼을 가진 정량화 모델 (PLS 모델)입니다. 데이터 전처리, 파장범위 및 잠재 변수 수는 정량화 모델에 명시된 대로 주의됩니다.
   b. 정량화 모델은 정의된 유의 수준을 기반으로 T^2 및 Q(PLS influence-Plot의 점선)에 대한 임계 값을 계산합니다. 임계값은 정량화 모델에 저장됩니다.
2. 예측에서 스펙트럼의 T^2 및 Q 값은 정량화 모델을 기반으로 계산됩니다.

3. 스펙트럼의 T^2 또는 Q 값이 임계 값보다 큰 경우 샘플은 적용된 정량화 모델과 관련하여 잠재적 특이값으로 간주됩니다.

6.6 기준값 특이값 - 알고리즘

상자 그림을 사용하는 경우 기준값의 특이값을 감지할 수 있습니다 *기준값 특이값 (수량화) (참조: 50페이지, 4.3.4장)*。

분포의 기울기를 주의하기에 대해 특이값 한계값은 다음 계산에 의해 조정됩니다。



Medcouple (MC)은 기준값의 기울기를 측정합니다。 계산은 상자 그림 Q_2의 중위수로 시작합니다。 기준값의 상단과 하단에서 모든 가능한 쌍 (영어로 *couple*)을 사용하여 함수가 계산됩니다。 결과의 매체는 Medcouple입니다：

$$\mathrm{MC} = \underset{y_i \le Q_2 \le y_j}{\mathrm{med}} \frac{(y_j - Q_2) - (Q_2 - y_i)}{y_j - y_i}$$

여기서 Q_2는 상자 그림의 중앙선을 정의하는 두 번째 사분면에 해당하고, y_i, y_j는 기준값 쌍을 나타냅니다。

Medcouple은 항상 –1에서 1 사이입니다. 대칭 분포의 경우 MC = 0입니다. MC > 0의 기울기 분포는 더 높은 기준값에 대해 왜곡되고, MC < 0은 더 낮은 기준 값에 대해 왜곡됩니다。

조정된 특이값 한계값의 계산은 분포가 어느 쪽으로 이동하느냐에 따라 달라집니다：

$$\mathrm{MC} \ge 0\text{: } [Q_1 - 1.5\, e^{-4\mathrm{MC}}\, \mathrm{IQR}\,;\, Q_3 + 1.5\, e^{3\mathrm{MC}}\, \mathrm{IQR}]$$
$$\mathrm{MC} < 0\text{: } [Q_1 - 1.5\, e^{-3\mathrm{MC}}\, \mathrm{IQR}\,;\, Q_3 + 1.5\, e^{4\mathrm{MC}}\, \mathrm{IQR}]$$

대칭 분포 (MC = 0)에서 한계값과 상자 사이의 거리는 1.5 IQR입니다。

지수 함수는 M. Hubert 및 E. Vandervieren, *An adjusted boxplot for skewed distributions*, Computational Statistics & Data Analysis 52권, 12번 (2008년 8월), 5186–5201쪽에 의해 경험적으로 증명된 것과 같이 다양한 경사도에서 서로 다른 분포에서 정확하고 강력한 이상값 검출을 가능하게 합니다。

표시된 특이값의 예상 백분율은 약 1 %이고 정규 분포에 대한 표준 상자 그림의 백분율과 매우 유사합니다。 주의상황: 이 비율은 특이값 레벨과 무관합니다。