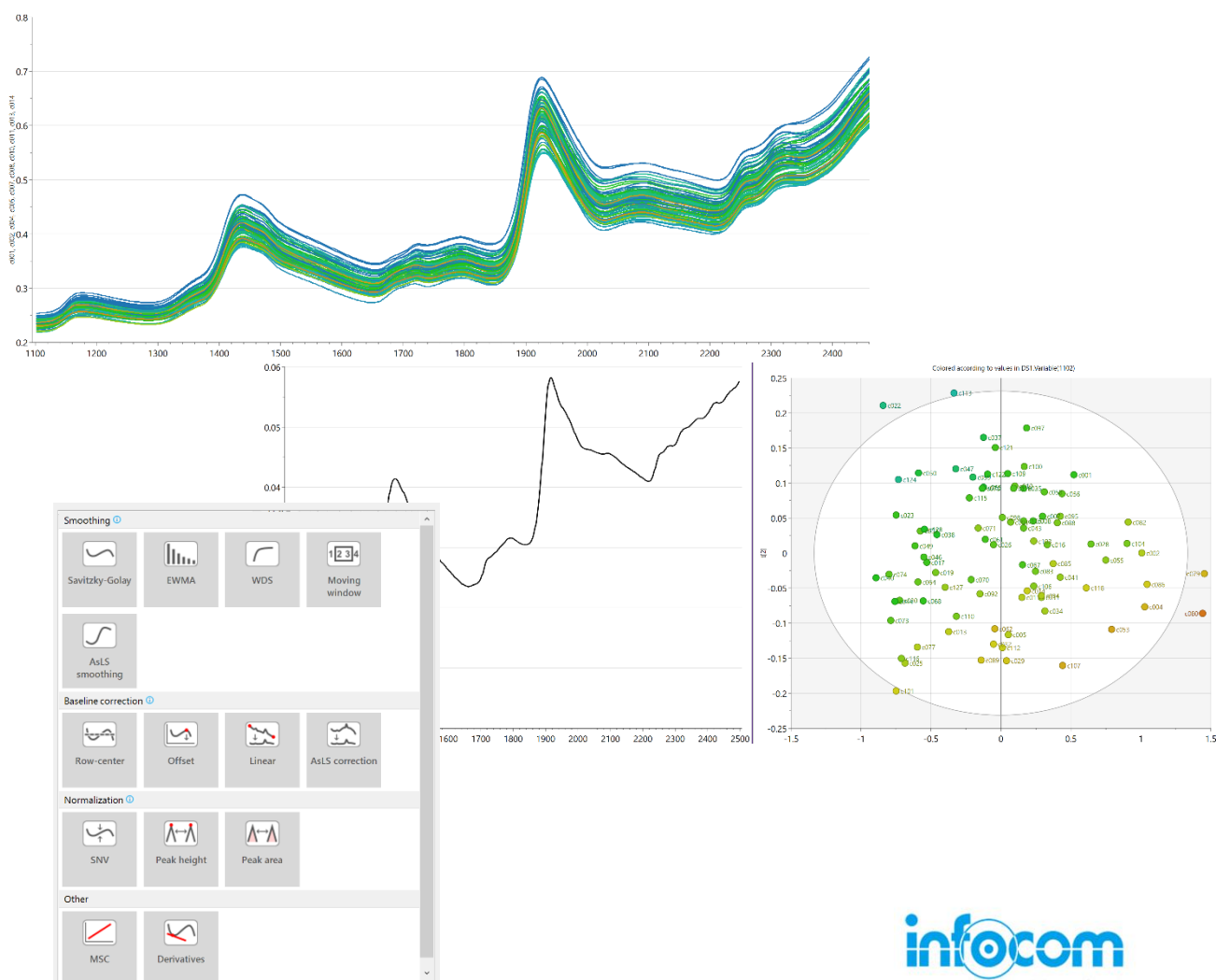


## SIMCA18 チュートリアル

### ～Spectroscopy データ (PCA,OPLS)編～



infocom

インフォコム株式会社  
ライフサイエンスグループ

備考：

本チュートリアルは SIMCA version 18.0. 0 に基づいて作成しております。そのため結果の一部（スコア、可視化部分も含む）や画面がお客様と異なる場合がございます。ご了承ください。

# 1. データセット

モデル構築のためのデータを読み込みます。

## データセットの準備

.dif 形式で用意します。

ファイル名：「Dataset - NIR Calibration samples.dif」

128 種類の異なるカラゲナン・サンプルに対する NIR（近赤外）スペクトルデータです。5 成分の混合物デザインが 6 段階で実施され、結果として 128 の粉末サンプルが得られました。

各サンプルについて、1100～2500 nm（699 の変数）の範囲で NIR スペクトルが取得されました。5 つの Y 変数は、各混合物におけるカラゲナンの種類（Lambda、Kappa、Iota、Mu および Nu）の相対量を表しています。

- ・ サンプル: 86 サンプルの NIR データ
- ・ 変数: スペクトルデータ（X 変数）、各カラゲナンタイプ相対量（Y 変数）
- ・ 目的: 1) スペクトルデータの概観とスコアプロット（PCA）  
2) スペクトルデータから各カラゲナンタイプの相対量を予測する回帰モデルの作成（OPLS）

備考：カラゲナン（carrageenan）について

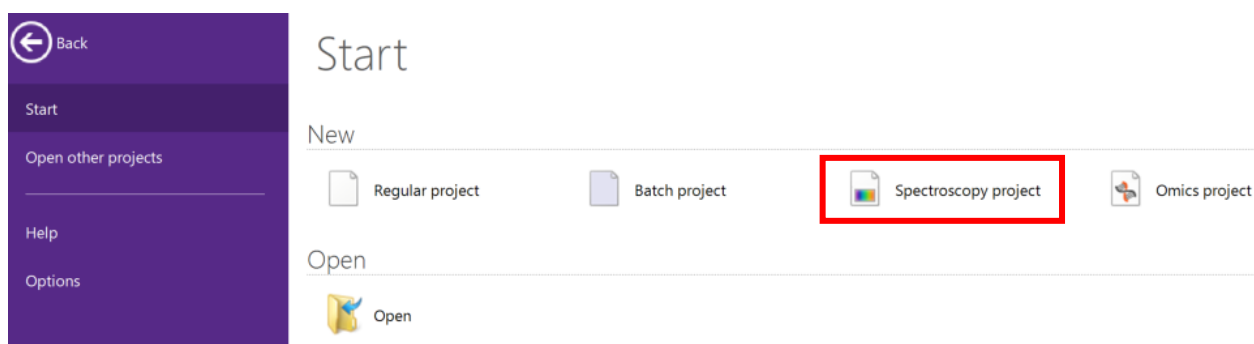
紅藻類から抽出される多糖類です。室温でゲルを形成し、食品その他の工業でゲル化剤、増粘剤、増粘安定剤などとして使われます。紅藻類の違いにより Lambda、Kappa、Iota などの種類があります。それぞれ粘性、ゲル化性などの物性が異なります。例えば、Lambda は Kappaphycus cottonii（オオキリンサイ属）から得られ、強く硬いゲルが作成されます。

## データセットのインポート

SIMCA18 を起動します。

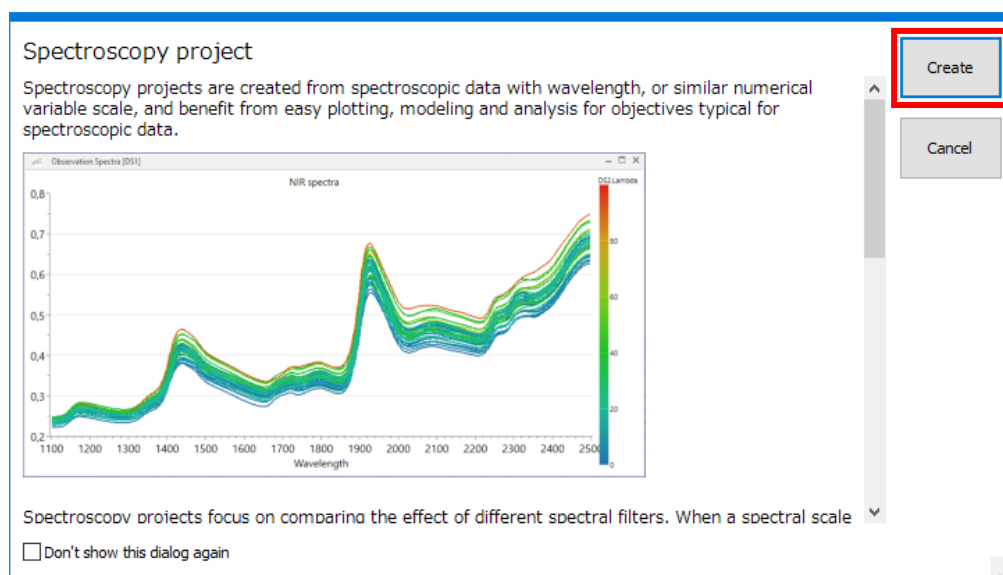
Start 画面が立ち上がるので、ここでは Spectroscopy project をクリックします。

SIMCA ではプロジェクト単位でデータの読み込みとモデルの構築を行います。

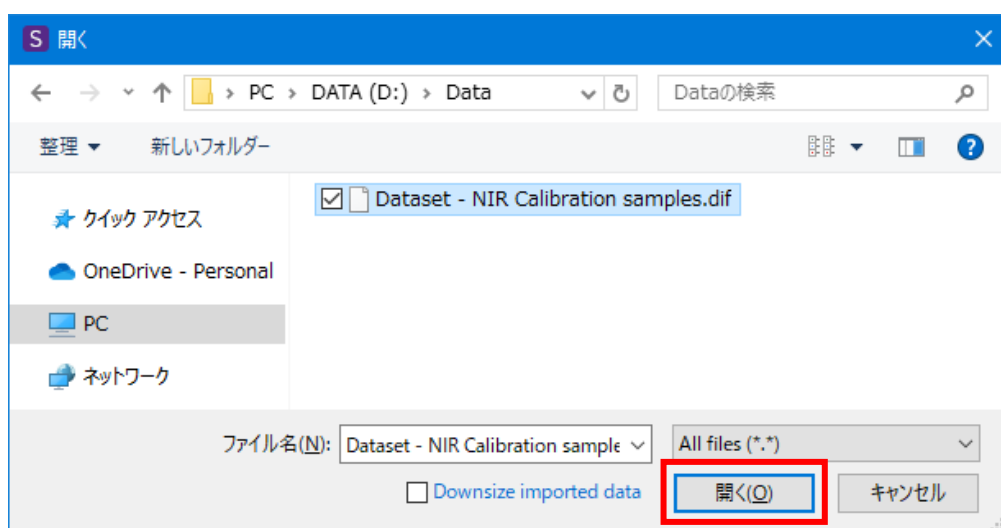


4 種のプロジェクトが存在しますが、選択に合わせてよく使う機能のメニュー表示（リボンメニュー）のみ変わります。

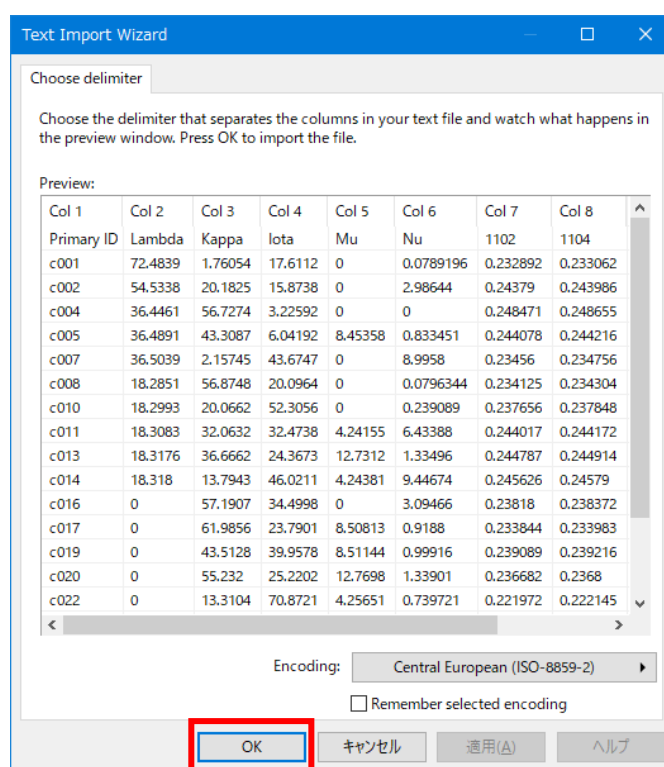
Create をクリックします。



用意していたデータセットを指定して、開きます。



Text Import Wizard が開きます。



内容を確認して、“OK”をクリックします。

Lambda から Nu までのカラムをクリックし、列全体を選択します。

Home > Y-variable をクリックすると、選ばれたカラムが Y 変数に設定されます。

The screenshot shows the SIMCA Import software interface. The 'Home' tab is active. In the 'Variable roles' section, the 'Y' role is highlighted with a red box. Below the toolbar, a table of data is displayed. The columns are labeled 'Primary ID', 'Primary ID', 'Lambda', 'Kappa', 'Iota', 'Mu', 'Nu', and then numerical columns from 7 to 14. The 'Lambda' through 'Nu' columns are highlighted with a red box, indicating they are selected as the Y-variable.

Primary ID	Primary ID	Lambda	Kappa	Iota	Mu	Nu	7	8	9	10	11	12	13	14
2	c001	72.4839	1.76054	17.6112	0	0.0789196	0.232892	0.233062	0.233211	0.233345	0.233474	0.233615	0.233812	0.234018
3	c002	54.5338	20.1825	15.8738	0	2.98644	0.24379	0.243986	0.244146	0.244303	0.244437	0.244598	0.244805	0.245022
4	c004	36.4461	56.7274	3.22592	0	0	0.248471	0.248655	0.248802	0.248933	0.249053	0.249187	0.249373	0.249579
5	c005	36.4891	43.3087	6.04192	8.45358	0.833451	0.244078	0.244216	0.244325	0.244407	0.244467	0.244554	0.244671	0.244807
6	c007	36.5039	2.15745	43.6747	0	8.9958	0.23456	0.234756	0.234912	0.235061	0.23519	0.235339	0.235539	0.235753
7	c008	18.2851	56.8748	20.0964	0	0.0796344	0.234125	0.234304	0.234449	0.234582	0.234703	0.234836	0.235026	0.23522
8	c010	18.2993	20.0662	52.3056	0	0.239089	0.237656	0.237848	0.238005	0.23816	0.238291	0.238445	0.238648	0.238863
9	c011	18.3083	32.0632	32.4738	4.24155	6.43388	0.244017	0.244172	0.244293	0.244391	0.244474	0.244568	0.244721	0.244888
10	c013	18.3176	36.6662	24.3673	12.7312	1.33496	0.244787	0.244914	0.245012	0.245084	0.245134	0.2452	0.245303	0.245425
11	c014	18.318	13.7943	46.0211	4.24381	9.44674	0.245626	0.24579	0.245913	0.24601	0.246111	0.246211	0.246367	0.246536
12	c016	0	57.1907	34.4998	0	3.09466	0.23818	0.238372	0.23853	0.238678	0.238801	0.238952	0.239151	0.239362
13	c017	0	61.9856	23.7901	8.50813	0.9188	0.233844	0.233983	0.234089	0.234172	0.234239	0.234326	0.234437	0.234567
14	c019	0	43.5128	39.9578	8.51144	0.99916	0.239089	0.239216	0.239319	0.239402	0.23946	0.239534	0.239634	0.239752
15	c020	0	55.232	25.2202	12.7698	1.33901	0.236682	0.2368	0.236884	0.236963	0.237015	0.237073	0.23718	0.237281

下図赤線部で囲んだ部分をスペクトルデータとして扱うために 1 列目を選択し、Home > Spectral-ID>をクリックして設定します。

The screenshot shows the SIMCA Import software interface. The 'Home' tab is active. In the 'Variable roles' section, the 'Spectral ID' role is highlighted with a red box. Below the toolbar, a table of data is displayed. The first column is highlighted with a red box, indicating it is selected as the Spectral ID. A dialog box titled 'Spectral ID Name' is open, showing the text 'Spectral ID' in the input field.

Primary ID	Primary ID	Lambda	Kappa	Iota	Mu	Nu	7	8	9	10	11	12	13	14
2	c001	72.4839	1.76054	17.6112	0	0.0789196	0.232892	0.233062	0.233211	0.233345	0.233474	0.233615	0.233812	0.234018
3	c002	54.5338	20.1825	15.8738	0	2.98644	0.24379	0.243986	0.244146	0.244303	0.244437	0.244598	0.244805	0.245022
4	c004	36.4461	56.7274	3.22592	0	0	0.248471	0.248655	0.248802	0.248933	0.249053	0.249187	0.249373	0.249579
5	c005	36.4891	43.3087	6.04192	8.45358	0.833451	0.244078	0.244216	0.244325	0.244407	0.244467	0.244554	0.244671	0.244807
6	c007	36.5039	2.15745	43.6747	0	8.9958	0.23456	0.234756	0.234912	0.235061	0.23519	0.235339	0.235539	0.235753
7	c008	18.2851	56.8748	20.0964	0	0.0796344	0.234125	0.234304	0.234449	0.234582	0.234703	0.234836	0.235026	0.23522
8	c010	18.2993	20.0662	52.3056	0	0.239089	0.237656	0.237848	0.238005	0.23816	0.238291	0.238445	0.238648	0.238863
9	c011	18.3083	32.0632	32.4738	4.24155	6.43388	0.244017	0.244172	0.244293	0.244391	0.244474	0.244568	0.244721	0.244888
10	c013	18.3176	36.6662	24.3673	12.7312	1.33496	0.244787	0.244914	0.245012	0.245084	0.245134	0.2452	0.245303	0.245425
11	c014	18.318	13.7943	46.0211	4.24381	9.44674	0.245626	0.24579	0.245913	0.24601	0.246111	0.246211	0.246367	0.246536
12	c016	0	57.1907	34.4998	0	3.09466	0.23818	0.238372	0.23853	0.238678	0.238801	0.238952	0.239151	0.239362
13	c017	0	61.9856	23.7901	8.50813	0.9188	0.233844	0.233983	0.234089	0.234172	0.234239	0.234326	0.234437	0.234567
14	c019	0	43.5128	39.9578	8.51144	0.99916	0.239089	0.239216	0.239319	0.239402	0.23946	0.239534	0.239634	0.239752
15	c020	0	55.232	25.2202	12.7698	1.33901	0.236682	0.2368	0.236884	0.236963	0.237015	0.237073	0.23718	0.237281

任意の Spectral-ID 名を入力します。

Lambda から Nu までが文字列であるため、SIMCA からエラーメッセージが发せられます。

Dataset - NIR Calibration samples (704 variables, 86 observations)

Go to Spectral ID must be numerical Rename... Exclude column **Split data**

Primary variable ID has not been specified. Unless specified, it will be automatically generated.

▶ Resolve all

しかしながら、解決策として SIMCA からデータを分ける Split data が提案されます。このまま Resolve all をクリックするとデータが分けられます。

Dataset - NIR Calibration samples (704 variables, 86 observations)

Go to Spectral ID must be numerical Rename... Exclude column **Split data**

Primary variable ID has not been specified. Unless specified, it will be automatically generated.

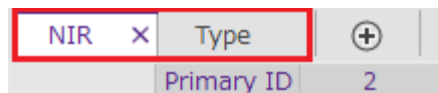
▶ **Resolve all**



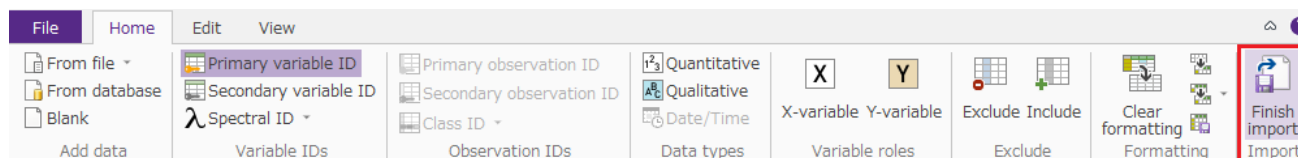
Primary ID	2	3	4	5	6	7	8
Spectral ID	Lambda	Kappa	Iota	Mu	Nu	1102	1104
3	c001	72.4839	1.76054	17.6112	0	0.0789196	0.232892
4	c002	54.5338	20.1825	15.8738	0	2.98644	0.24379
5	c004	36.4461	56.7274	3.22592	0	0	0.248471

Primary ID	Y	Y	Y	Y	Y	7	8
Spectral ID	Lambda	Kappa	Iota	Mu	Nu	1102	1104
2	c001	72.4839	1.76054	17.6112	0	0.0789196	0.232892
3	c002	54.5338	20.1825	15.8738	0	2.98644	0.24379
4	c004	36.4461	56.7274	3.22592	0	0	0.248471
5	c005	36.4891	43.3087	6.04192	8.45358	0.833451	0.244078

各データのシート名をダブルクリックして Dataset - NIR Calibration samples を NIR に、  
Dataset - NIR Calibration samples (2) を Type に変更します。



Finish import をクリックし、SIMCA project (.usp) ファイルの保存先を指定します。



Home > Dataset をクリックして、インポートしたデータセットを表示します。

Dataset - NIR Calibration samples.usp - SIMCA

File

Home

Data

Analyze

Predict

Plot/List

View

Tools

Developer

Add-Ins

Active model: N/A

Project

Dataset

Spectra

Calibration wizard

New

Edit

Delete

Statistics

Compare models

Model type

Autofit

Add

Remove

Summary of fit

Overview

Scores

Loadings

Hotelling's T2

DMod

Obs. vs. pred.

Coefficients

VIP

Create

Project

Workset

Model

Diagnostics & interpretation

Plot/List

NIR

1

2

3

4

5

6

7

1

1	Primary ID	1102	1104	1106	1108	1110	1112	1
2	Spectral ID	1102	1104	1106	1108	1110	1112	1
3	c001	0.232892	0.233062	0.233211	0.233345	0.233474	0.233615	
4	c002	0.24379	0.243986	0.244146	0.244303	0.244437	0.244598	
5	c004	0.248471	0.248655	0.248802	0.248933	0.249053	0.249187	
6	c005	0.244078	0.244216	0.244325	0.244407	0.244467	0.244554	
7	c007	0.23456	0.234756	0.234912	0.235061	0.23519	0.235339	
8	c008	0.234125	0.234304	0.234449	0.234582	0.234703	0.234836	
9	c010	0.237656	0.237848	0.238005	0.23816	0.238291	0.238445	
10	c011	0.244017	0.244172	0.244293	0.244391	0.244474	0.244568	
11	c013	0.244787	0.244914	0.245012	0.245084	0.245134	0.2452	
12	c014	0.245626	0.24579	0.245913	0.24601	0.246111	0.246211	
13	c016	0.23818	0.238372	0.23853	0.238678	0.238801	0.238952	
14	c017	0.233844	0.233983	0.234089	0.234172	0.234239	0.234326	
15	c019	0.239089	0.239216	0.239319	0.239402	0.23946	0.239534	

Type

1

2

3

4

5

6

1

1	Primary ID	Lambda	Kappa	Iota	Mu	Nu	1
2	c001	72.4839	1.76054	17.6112	0	0.0789196	
3	c002	54.5338	20.1825	15.8738	0	2.98644	
4	c004	36.4461	56.7274	3.22592	0	0	
5	c005	36.4891	43.3087	6.04192	8.45358	0.833451	
6	c007	36.5039	2.15745	43.6747	0	8.9958	
7	c008	18.2851	56.8748	20.0964	0	0.0796344	
8	c010	18.2993	20.0662	52.3056	0	0.239089	
9	c011	18.3083	32.0632	32.4738	4.24155	6.43388	
10	c013	18.3176	36.6662	24.3673	12.7312	1.33496	
11	c014	18.318	13.7943	46.0211	4.24381	9.44674	
12	c016	0	57.1907	34.4998	0	3.09466	
13	c017	0	61.9856	23.7901	8.50813	0.9188	
14	c019	0	43.5128	39.9578	8.51144	0.99916	
15	c020	0	55.232	25.2202	12.7698	1.33901	

以上でデータセットのインポートが完了となります。

💡 SIMCA では欠損値とゼロを分けて認識します。空欄は欠損値とみなされます。欠損値（Missing values）は  
ピンクのセルで表されます

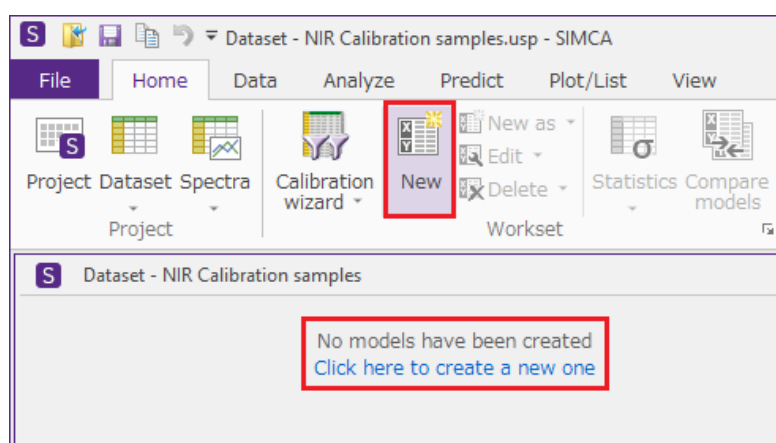
## 2.主成分分析(PCA)

主成分分析により、データの概観を捉えます。

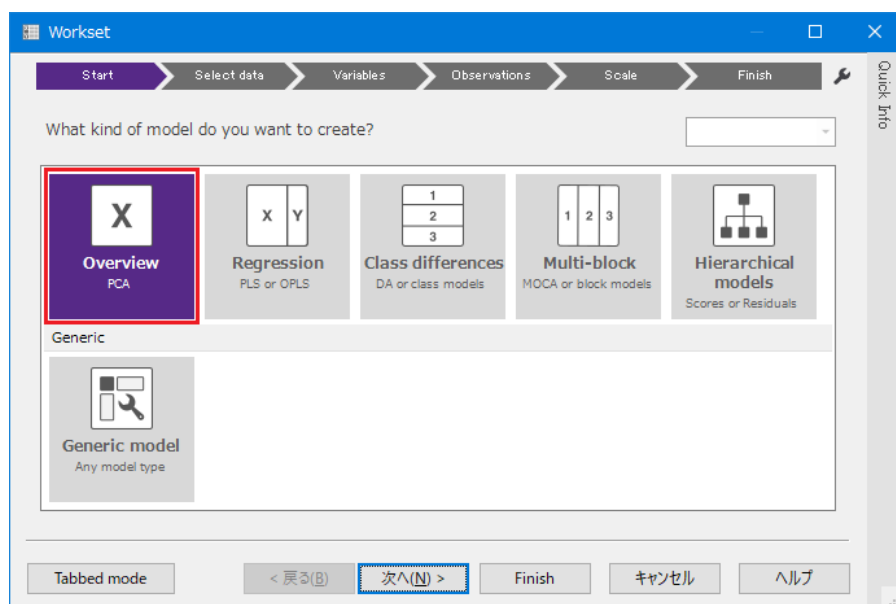
### モデルの作成

#### モデルを設定する

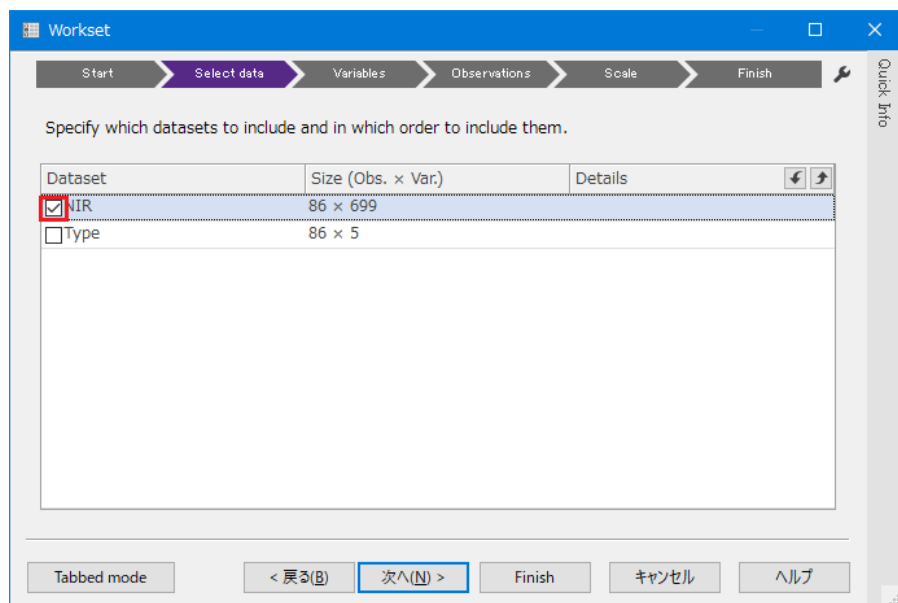
Home > New ボタンまたは “Click here to create a new one” をクリックします。



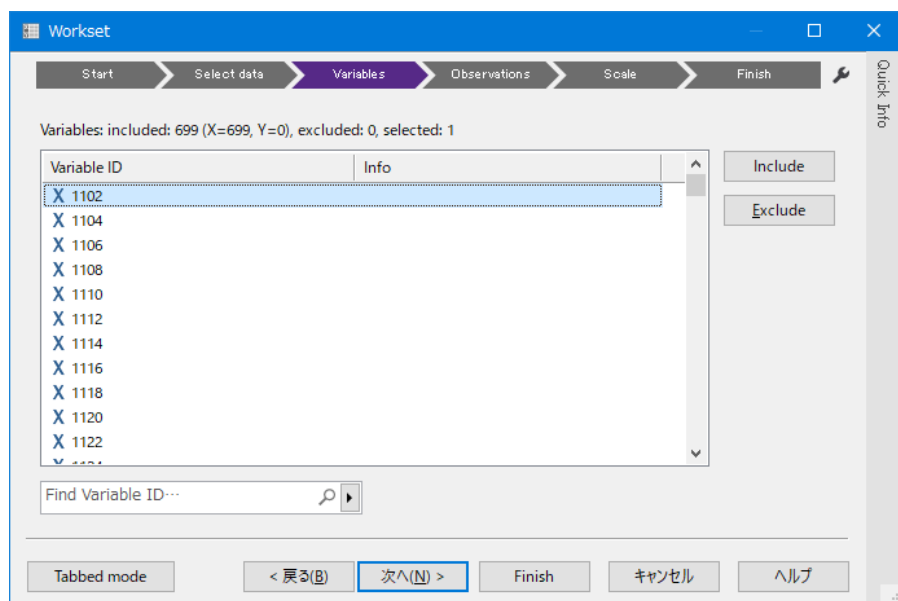
Overview (PCA) を選択し、次へをクリックします。



PCA の対象となるデータセットを選択します。NIR にチェックを入れ、次へをクリックします。

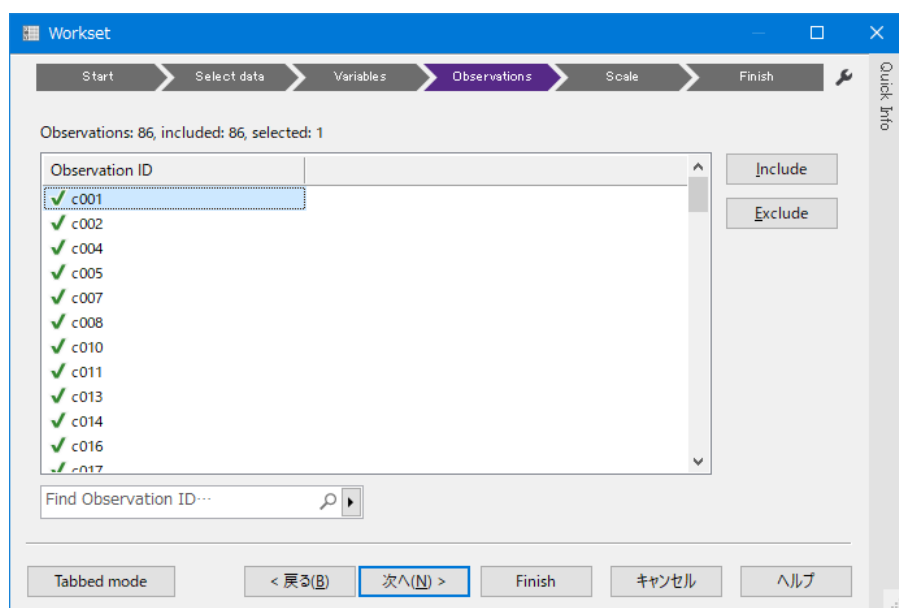


Variable（変数）の設定をします。不要な変数があれば Exclude をクリックして除外します。



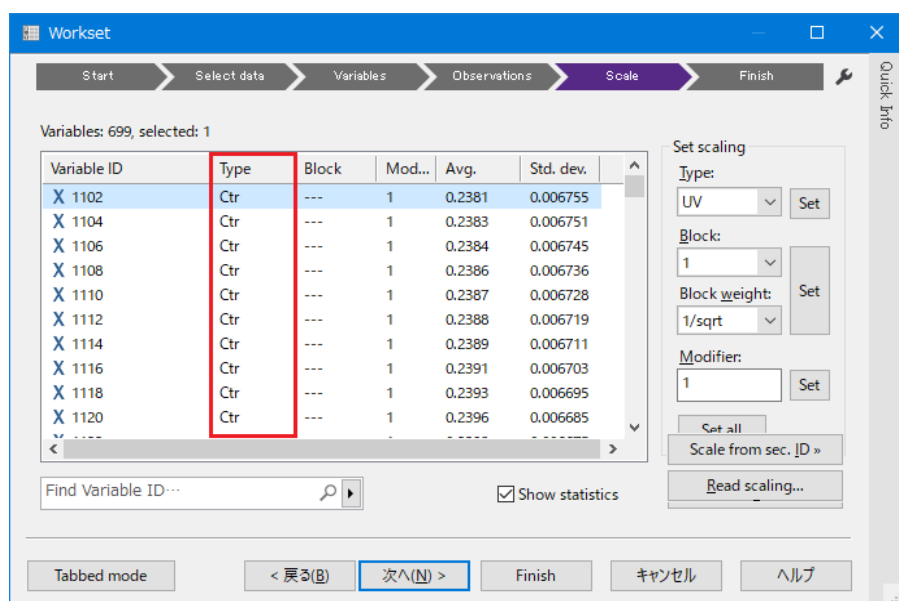
ここでは何も操作をせず、次へをクリックします。

Observation（観測）の設定をします。不要なサンプルがあれば Exclude をクリックして除外します。

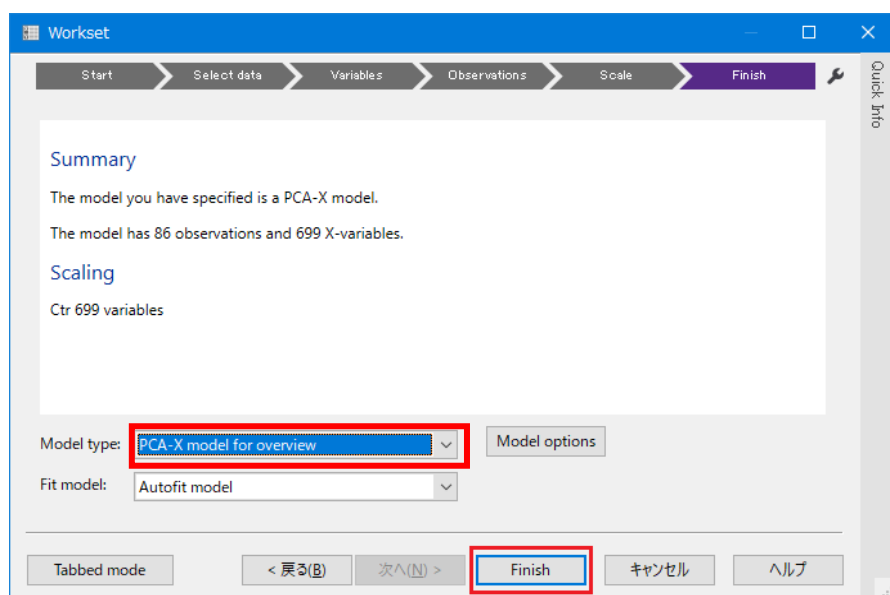


ここでも何も操作せず、次へをクリックします。

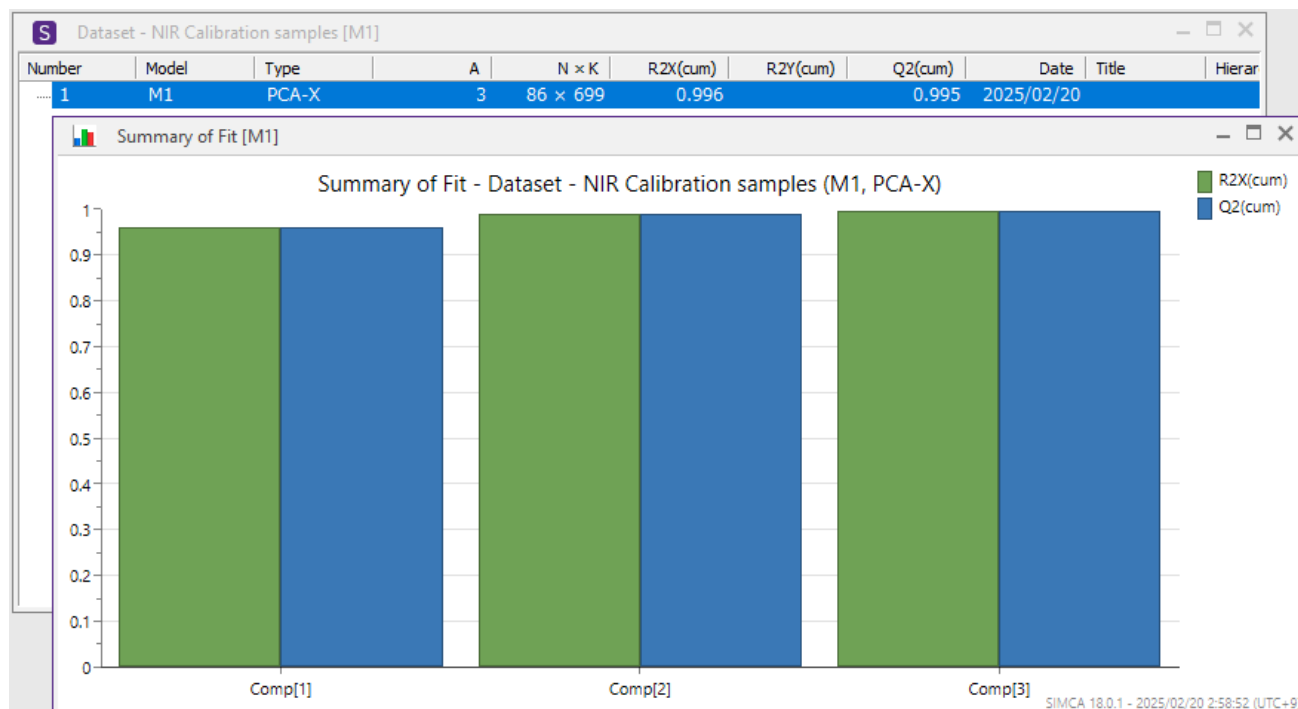
Scale（スケーリング：尺度を合わせる）の設定をします。スペクトルデータを使用する場合、分析でセンタリング（ctr）を使用して変数をスケーリングするのが一般的です。この場合、SIMCA はデフォルトでこれを選択しています。Type が Ctr になっていることを確認し、次へをクリックします。



Summary 画面では現在のモデルタイプ（model type）、観測値（observations）、変数（variables）、スケーリング（scaling）が表示され、最終確認を行います。Model types を“PCA-X model for overview”に設定し、Finish をクリックします。



SIMCA が自動的に最適なモデルを計算し、モデルを作成します。ここでは、第 3 主成分まで計算されました。



緑はその成分でどれだけのデータを表しているかの寄与率(R2)、青は交差検定による寄与率(Q2)を表します。

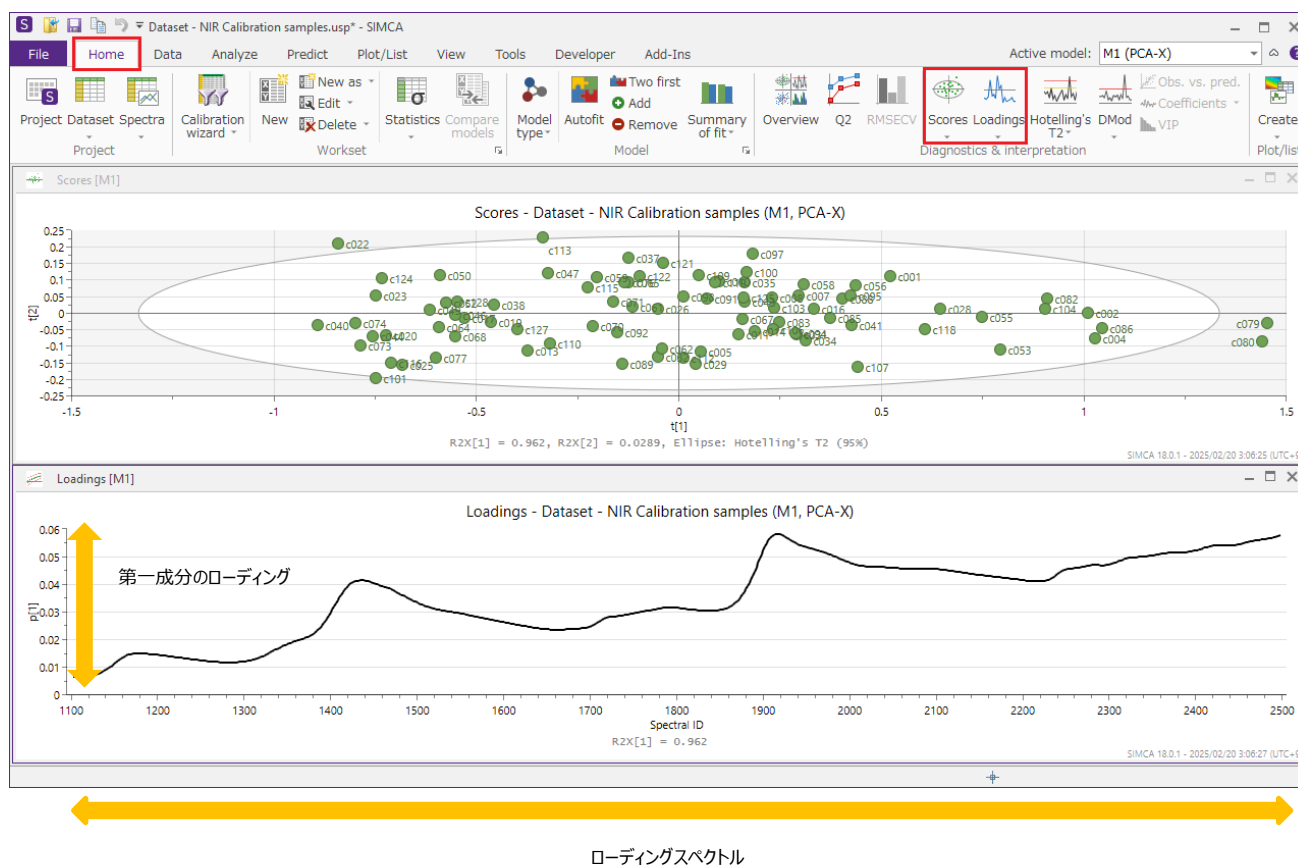
## 結果の解釈

スコアプロットとローディングプロットを中心に結果の解釈を行います。

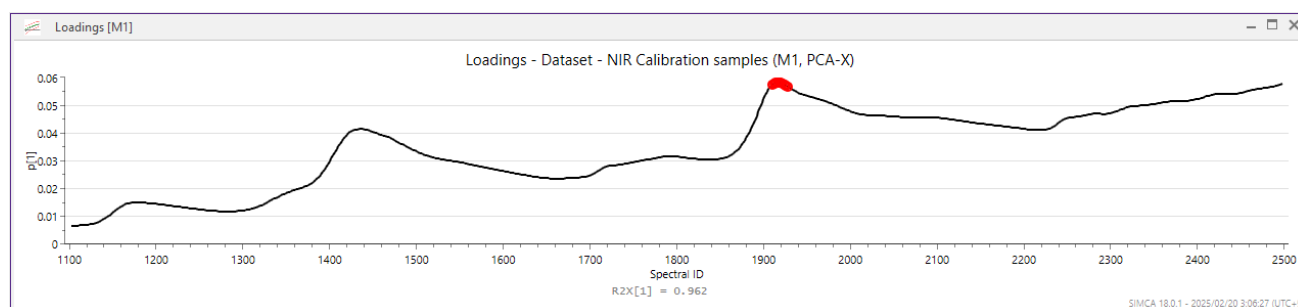
### スコアプロットとローディングプロット

Home > Scores と Loadings をクリックし、スコアプロットとローディングプロットを表示させます。

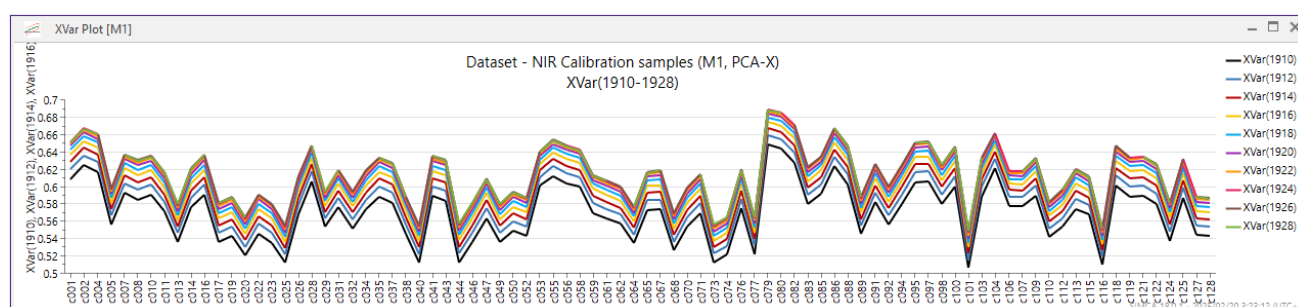
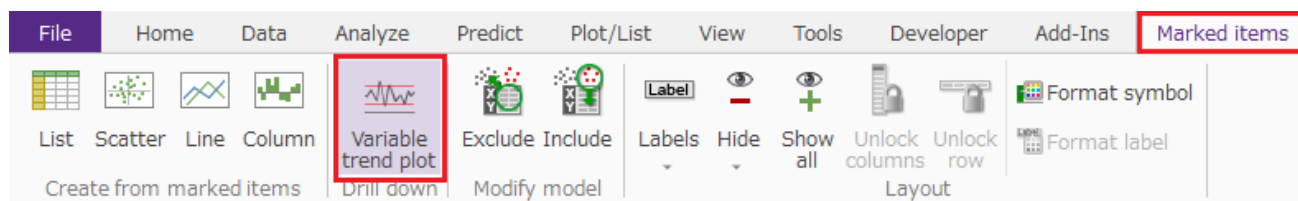
ローディングプロットの縦軸が第一成分のローディング、横軸がローディングスペクトル（1102～2498）を示します。



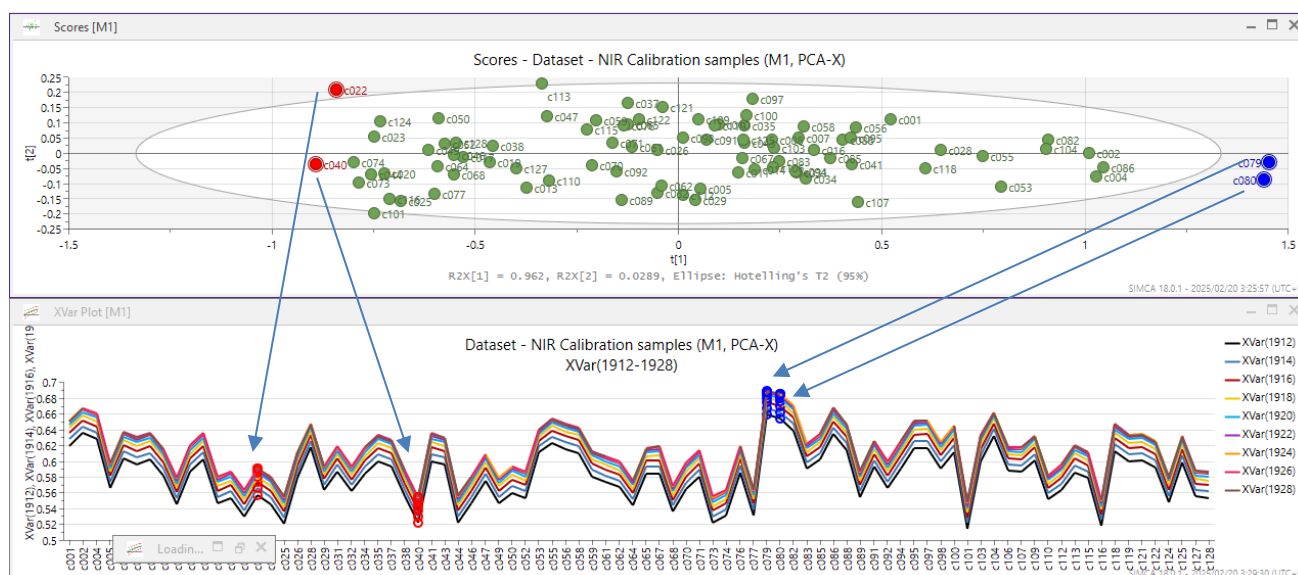
ローディングプロットの 1910～1930 付近のスペクトルを選択します（選択された部分は赤色になります）。



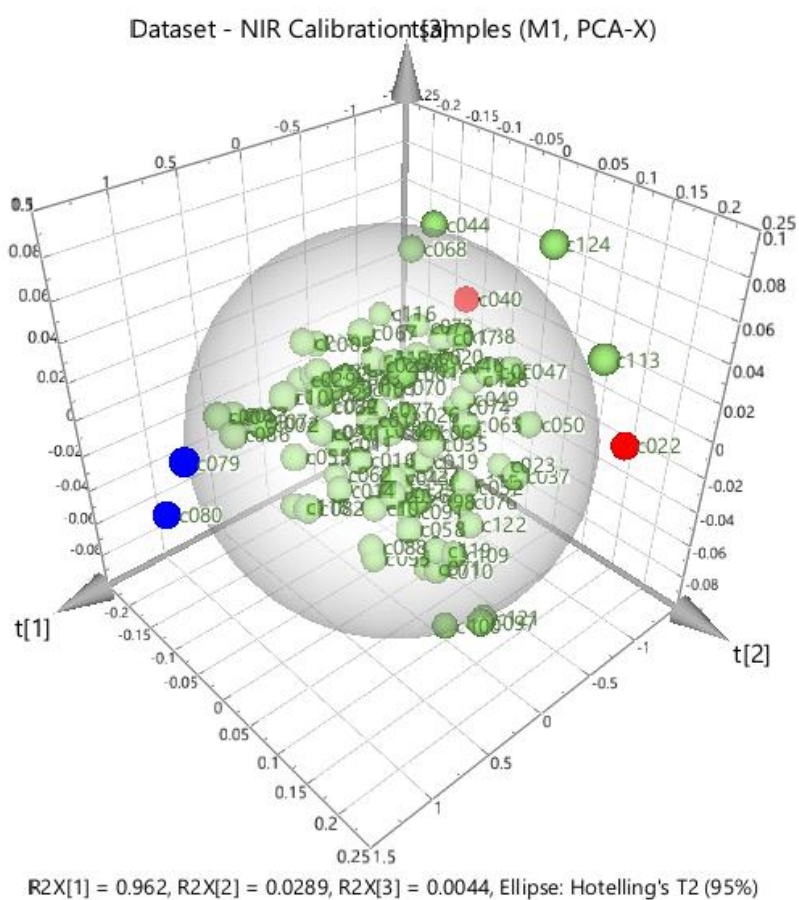
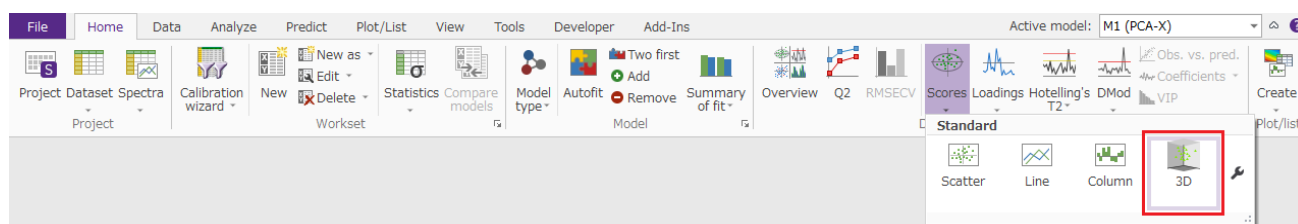
選択後、Marked Items>Variable trend plot をクリックし、XVar Plot を表示させます。



スコアプロットと XVar Plot を並べ、スコアプロットの両端のサンプルをそれぞれ選択します（例：右側で c079,c080 を選択した後に、c022 と c040 を選択します）。XVar Plot を確認すると右側で選ばれたものが、プロットの上部に、左側で選んだものが下部に示されます。



Home>Scores>3D でスコアプロットを立体表示させることも可能です。



# 3.OPLS

OPLS では Y 変数を設定し、X と Y の回帰式を作成します。

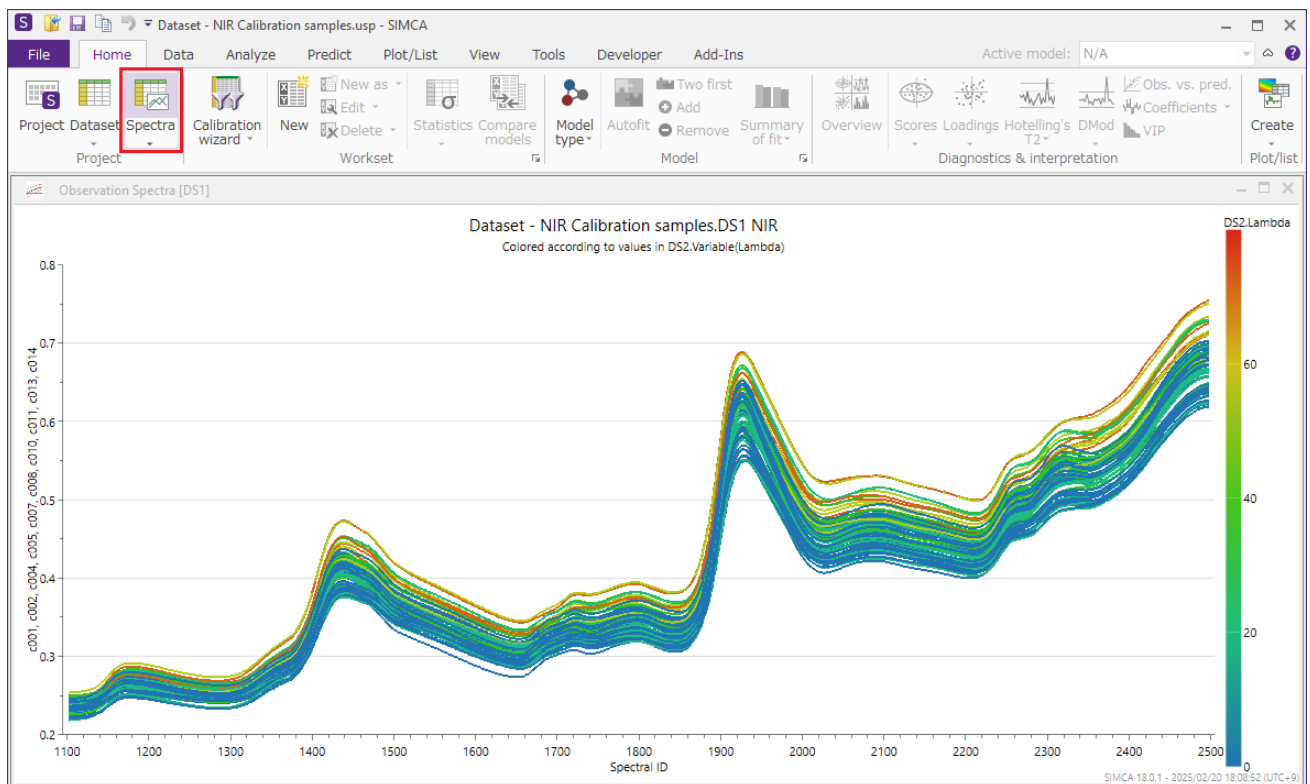
目的変数 Y に連動する、説明変数 X を探すだけではなく、OPLS モデルを使って Y 変数を予測も可能です。

本項ではスペクトルデータの補正（Calibration）を実施し、OPLS モデルを作成します。

## スペクトルデータの補正からの OPLS モデル作成

### スペクトルデータの確認

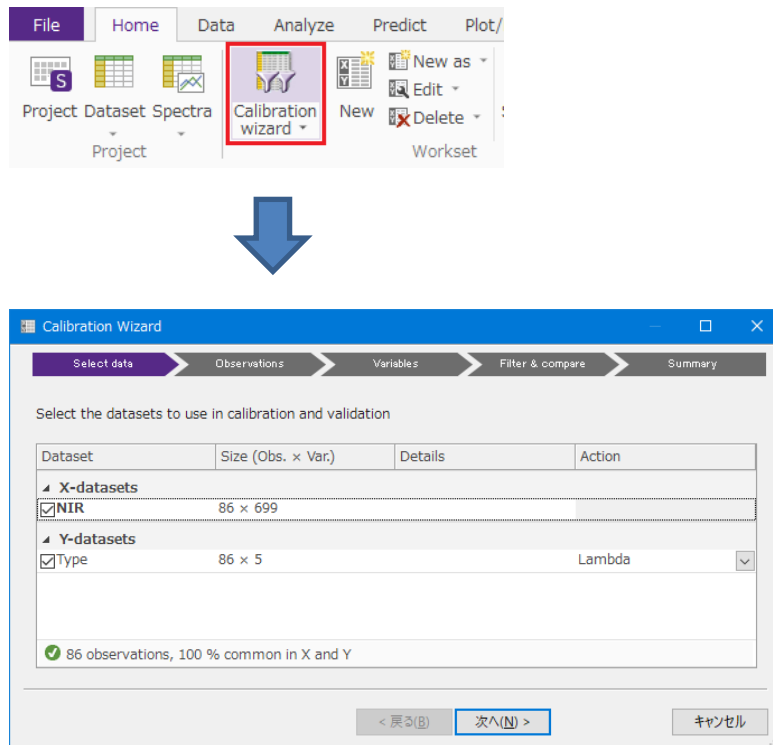
Home > Spectra> DS1 : NIR を選択し、スペクトルデータを確認します。



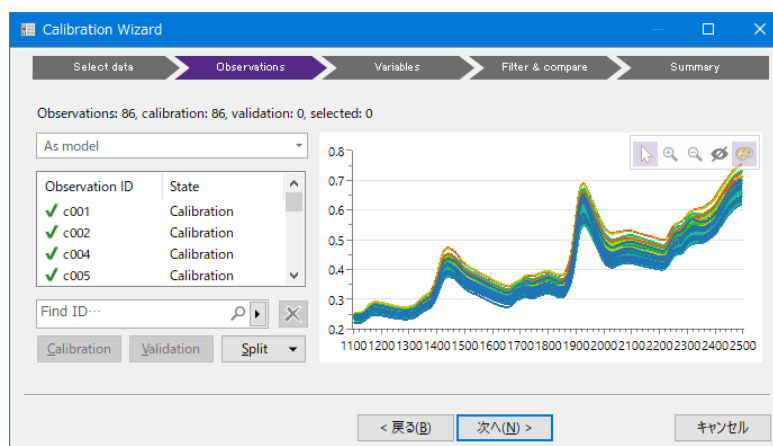
このスペクトルデータは未補正でブロードな状態です。そこで補正をかけベースラインを整えます。

## 補正設定

Home > Calibration wizard> create new session を選択し、Calibration wizard を開きます。  
X-datasets と Y-Datasets でそれぞれ NIR と TYPE にチェックがあることを確認して次へをクリックします。

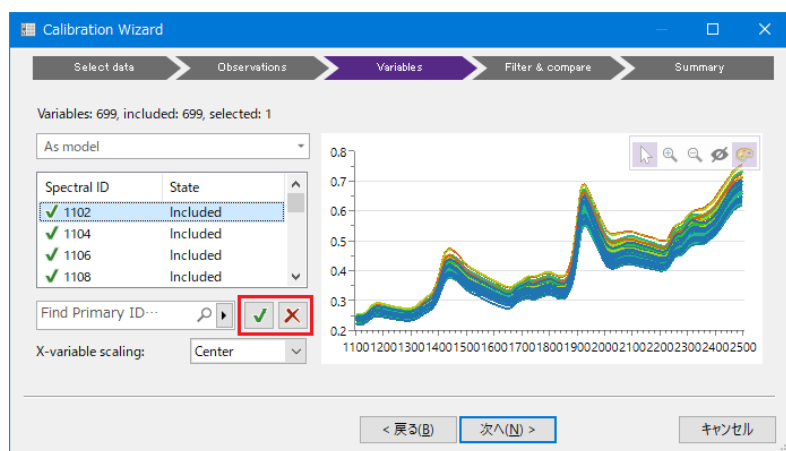


Observation（観測）の設定をします。



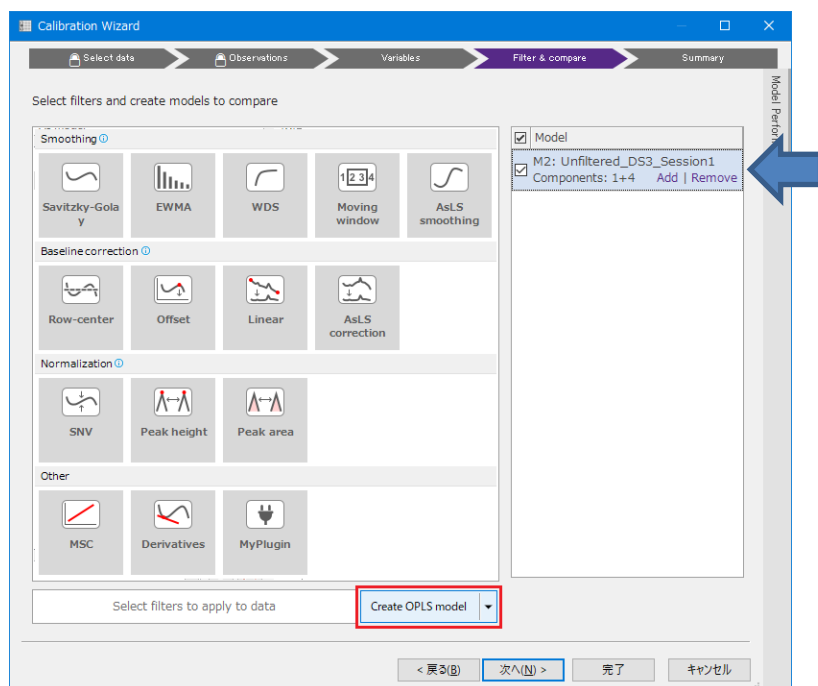
ここでは何も操作をせず、次へをクリックします。

Variable（変数）の設定をします。不要なサンプルがあれば赤色×アイコンをクリックして削除します（Exclude）。  
戻す場合(Include)は、緑色✓アイコンをクリックします。

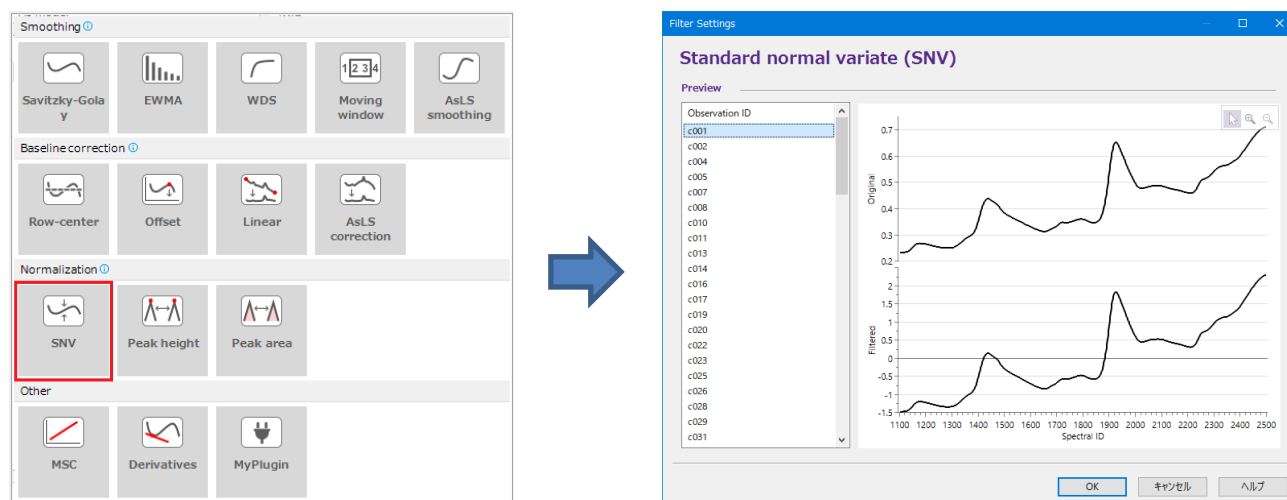


ここでは何も操作をせず、次へをクリックします。

補正の種類を選択します。基本的に『種類（Filter）選択→Filter 前後の確認→回帰モデルの作成（ここでは OPLS）→次の種類（Filter）選択』の流れになります。まずは Filter 未選択のまま Create OPLS model をクリックして、補正のかかっていない状態で OPLS モデルを作成します。モデルが作成されると右側の Model リストに追加されます。

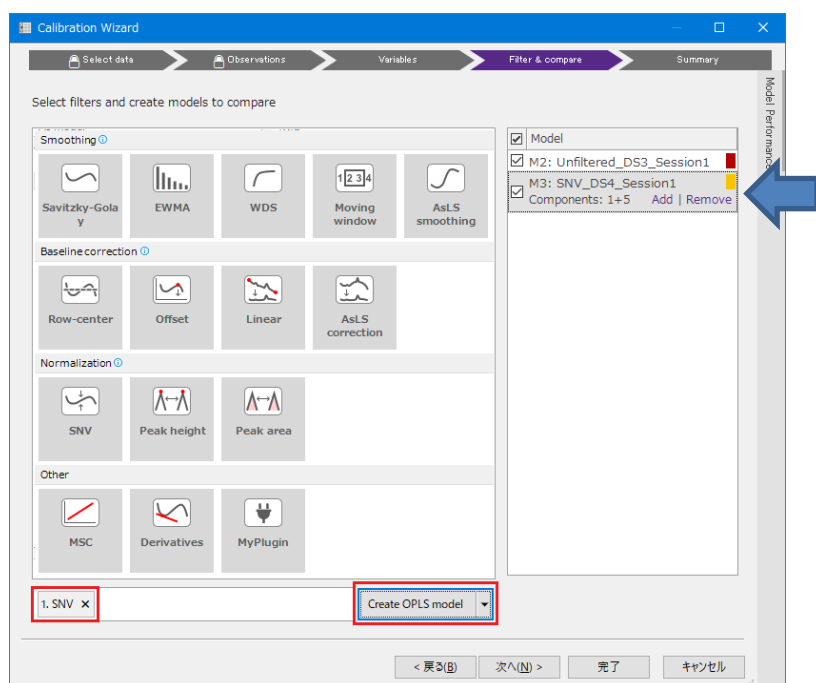


SNV（Standard Normal Variate）を選択します。選択すると補正前後(Original/Filtered)が表示されます。

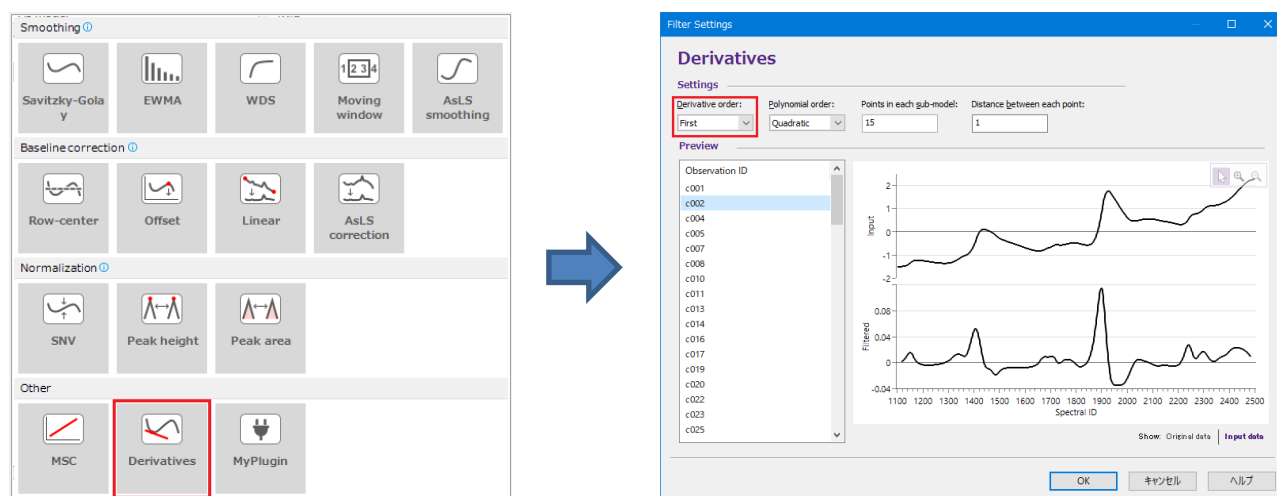


確認後に OK をクリックします。

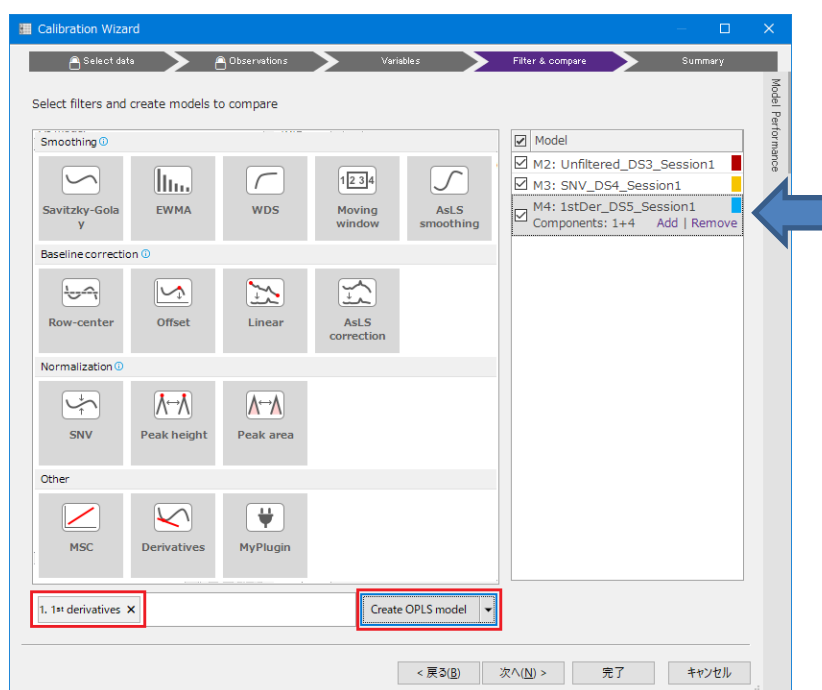
SNV が選択されていることを確認し Create OPLS model をクリックして、OPLS モデルを作成します。モデルが作成されると右側の Model リストに追加されます。



Derivative を選択します。今回は一次微分を行うので Derivative order:First を選択します。  
 選択すると補正前後(Original/Filtered)が表示されます。確認後に OK をクリックします。

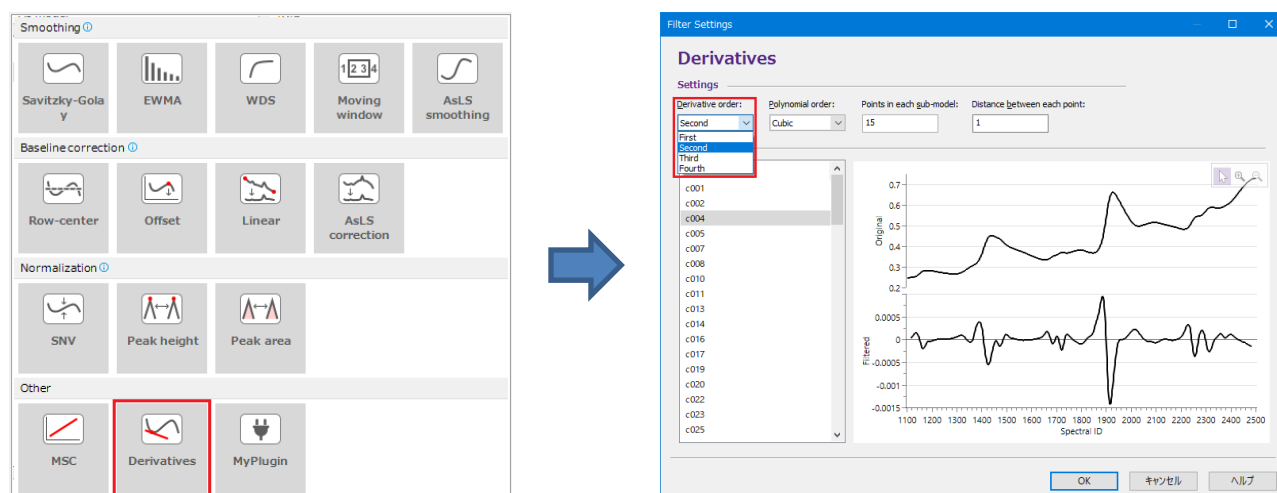


1<sup>st</sup> derivative が選択されていることを確認し Create OPLS model をクリックして、OPLS モデルを作成します。  
 モデルが作成されると右側の Model リストに追加されます。

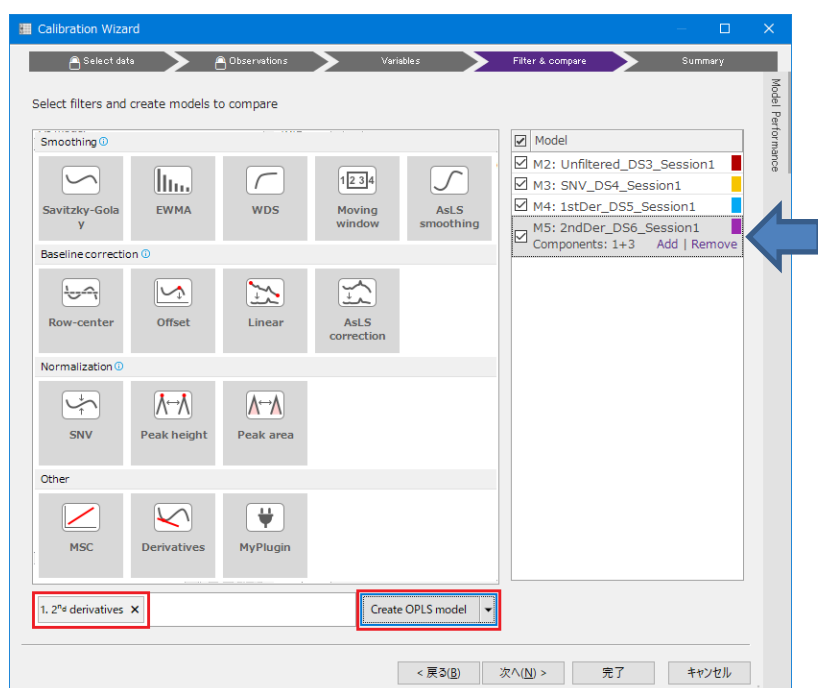


💡 先に行った Filter が残っている場合があります、その場合は×をクリックして削除します。

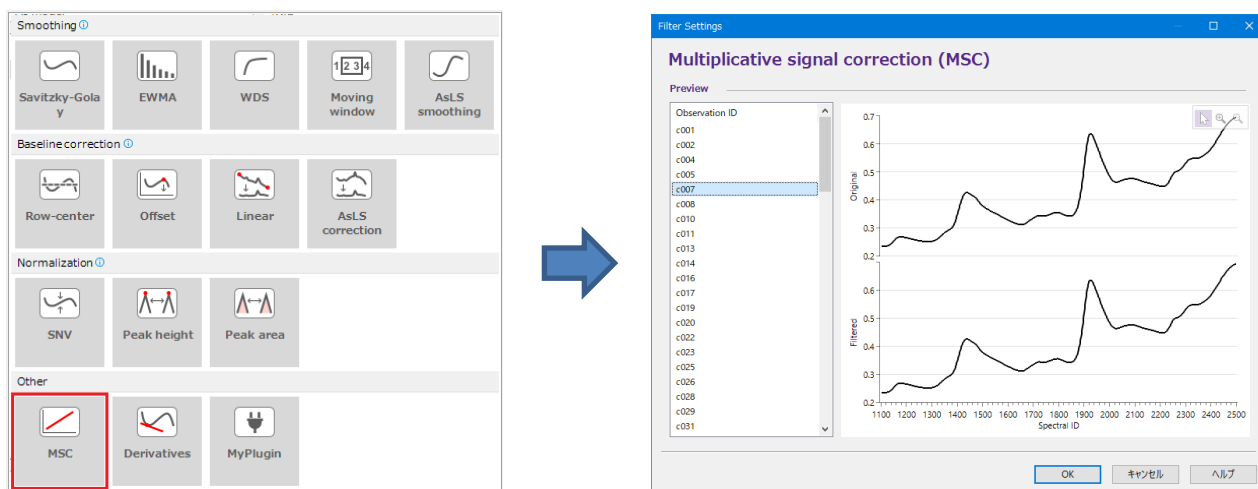
Derivative を選択します。今回は二次微分を行うので Derivative order:Second を選択します。  
選択すると補正前後(Original/Filtered)が表示されます。確認後に OK をクリックします。



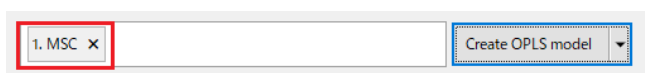
2<sup>nd</sup> derivative が選択されていることを確認し Create OPLS model をクリックして、OPLS モデルを作成します。  
モデルが作成されると右側の Model リストに追加されます。



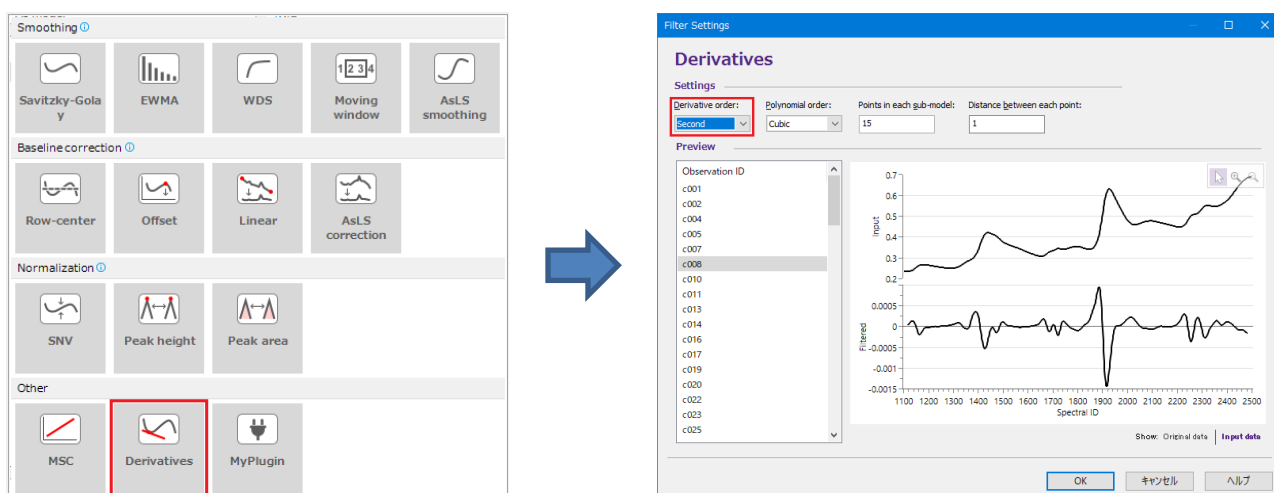
MSC (Multiplicative Scatter Correction) を選択します。選択すると補正前後(Original/Filtered)が表示されます。確認後に OK をクリックします



MSC が残っていることを確認し、Derivative を選択します。

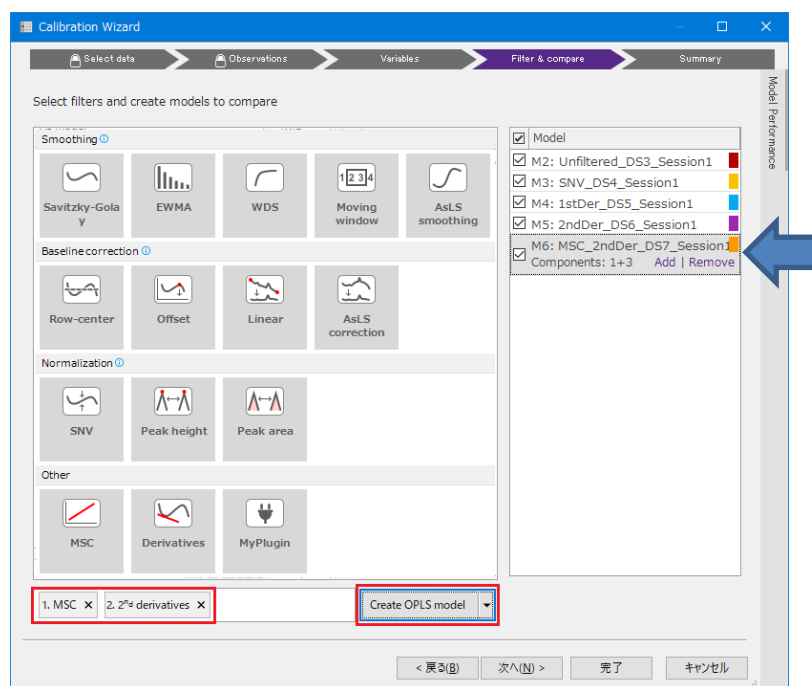


今回は二次微分を行うので Derivative order:Second を選択します。選択すると補正前後(Original/Filtered)が表示されます。確認後に OK をクリックします。



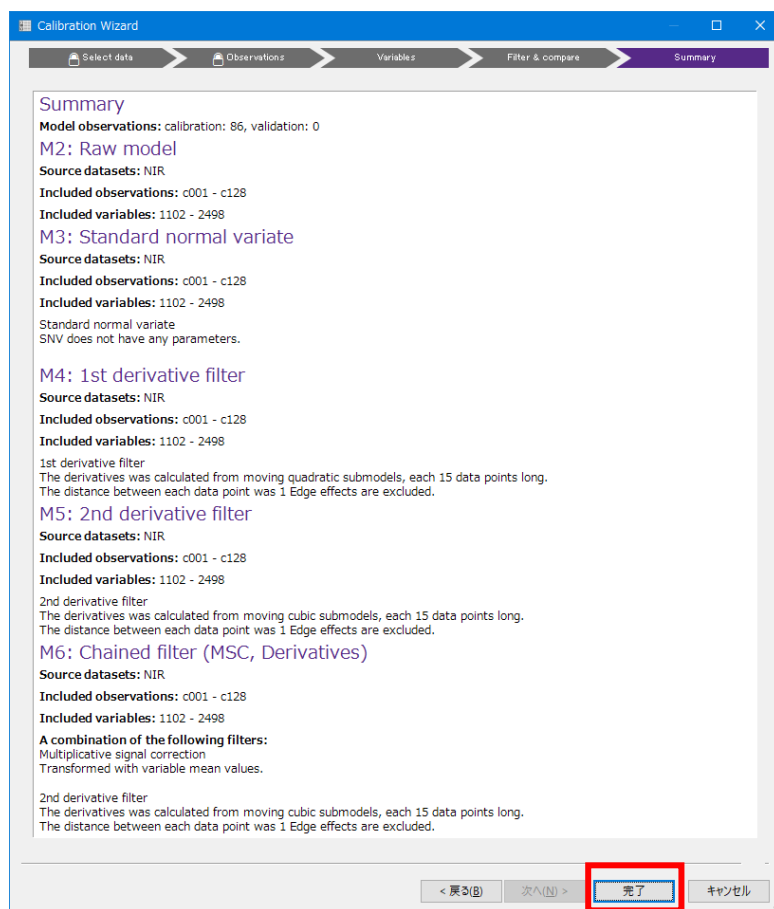
💡 SIMCA では複数の Filter を組み合わせて実行することが可能です。

MSCと2<sup>nd</sup> derivative が選択されていることを確認し Create OPLS model をクリックして、OPLS モデルを作成します。モデルが作成されると右側の Model リストに追加されます。



次へをクリックします。

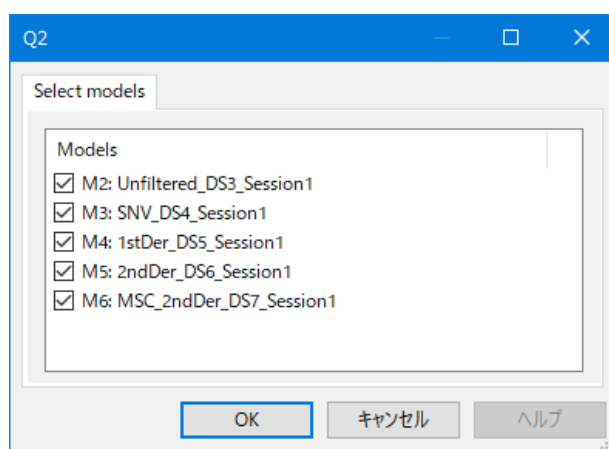
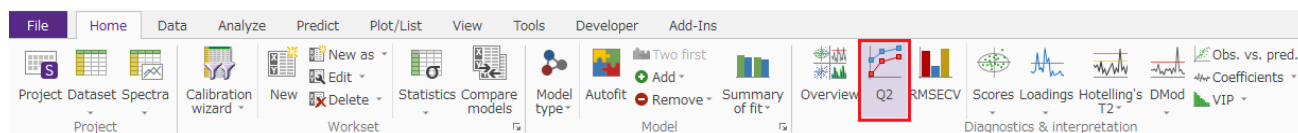
作成したモデルの確認をし、完了をクリックします。



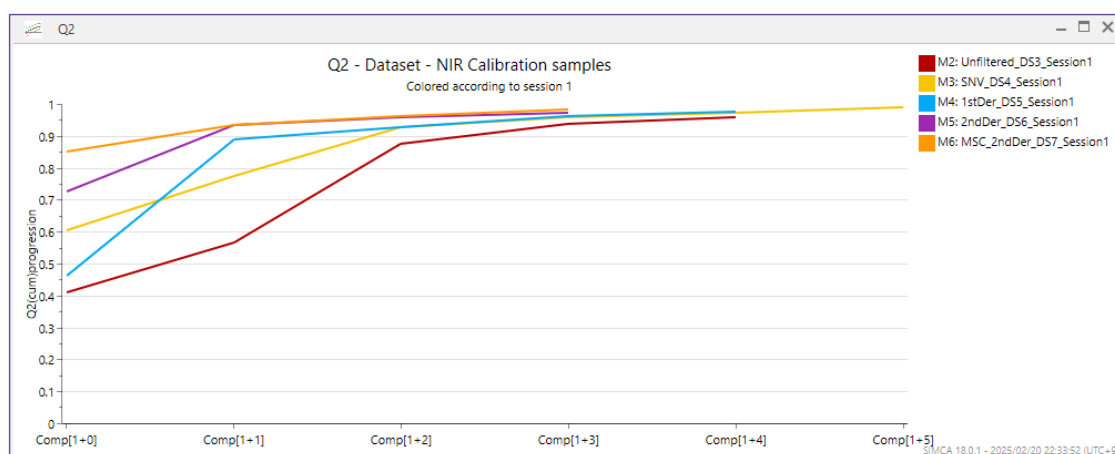
# モデルの比較と予測

## モデルの比較

Home > Q2 をクリックし、比較するモデルを確認した後に OK をクリックします。

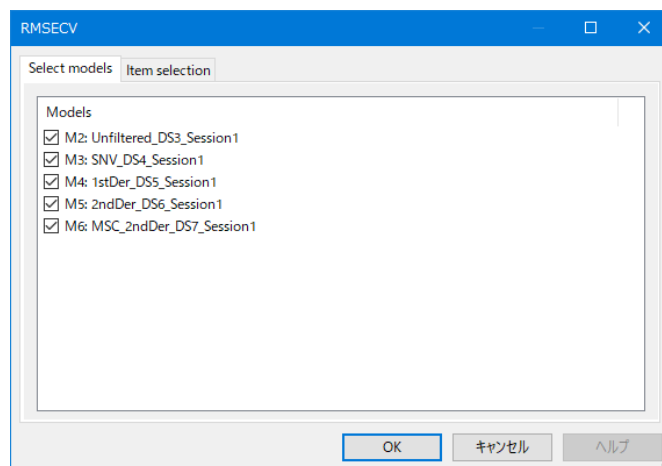
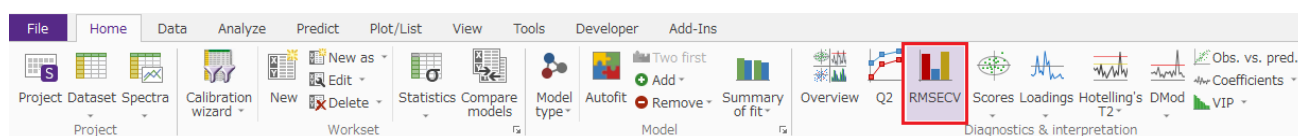


Q2 の結果です。Q2 が高いものほど精度が良いモデルですが、コンポーネント（成分）が少ないモデルを選びます。今回の比較では M6 が良モデルの候補になります。

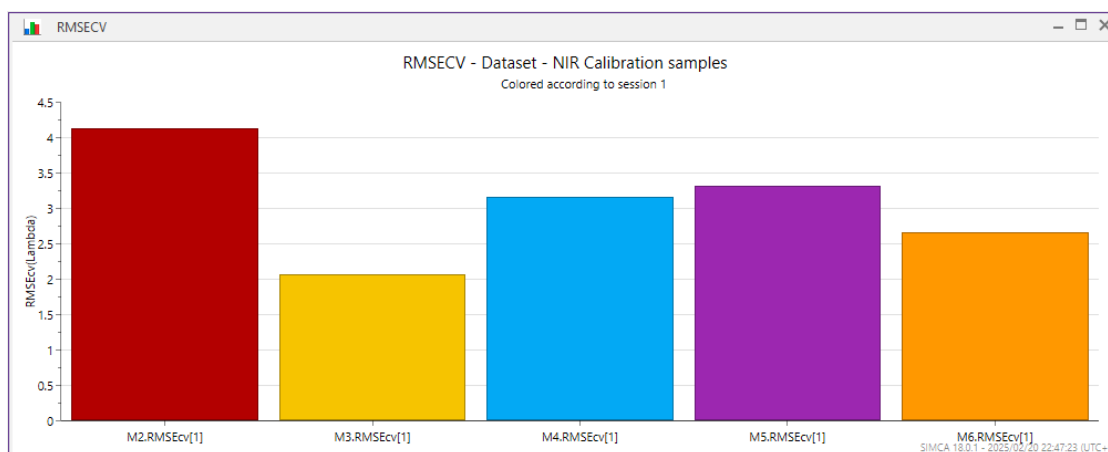


💡 コンポーネント（成分）増えモデルが複雑化すると過剰適合（オーバーフィッティング）が起こりやすくなります。

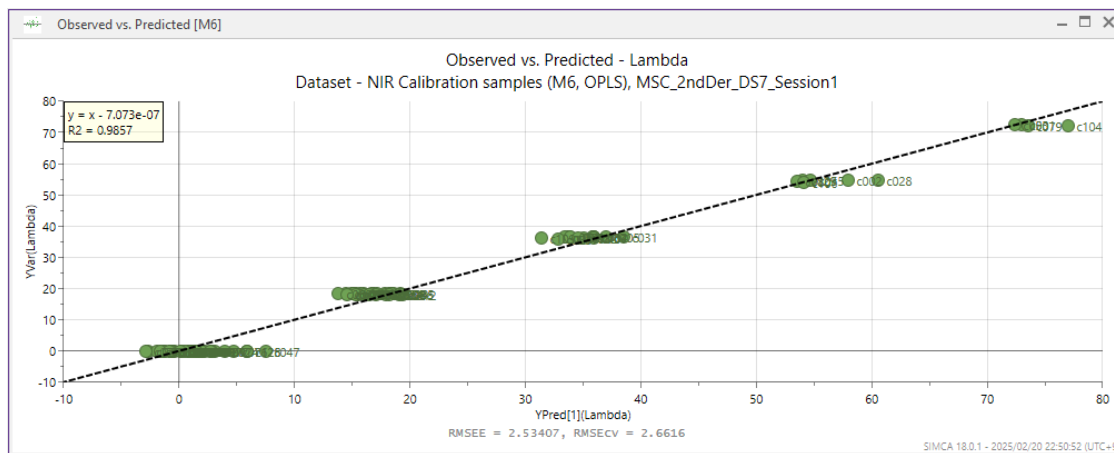
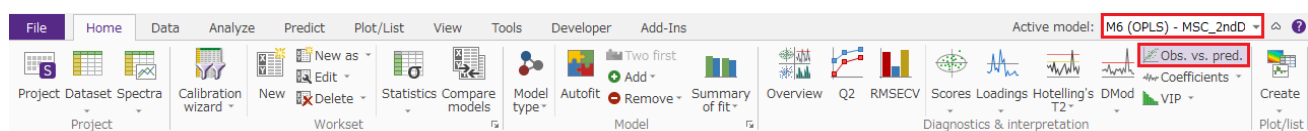
Home > RMSECV をクリックし、比較するモデルを確認した後に OK をクリックします。



RMSECV の結果です。RMSECV が低いものほど精度が良いモデルですが、M3、M6 が良モデルの候補になります。

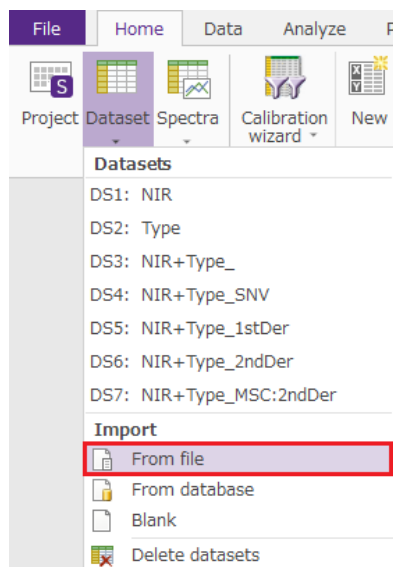


Home> Obs. vs. pred をクリックし、回帰モデル（M6）を確認します。



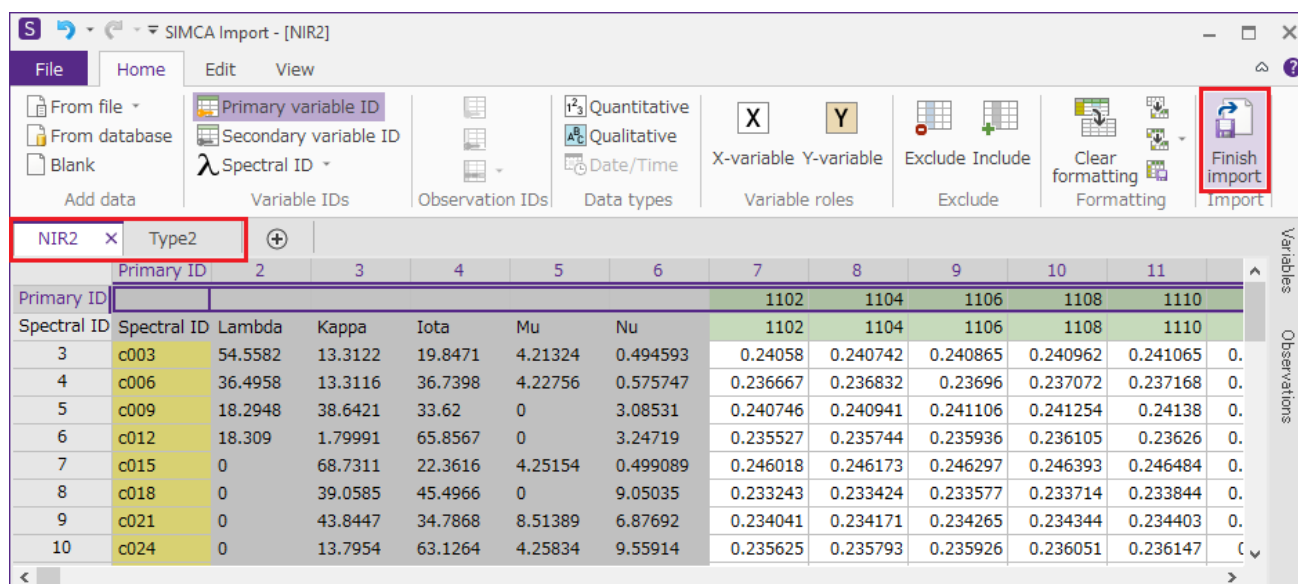
## OPLS 予測

Home > Dataset>Import : From file をクリックし、Dataset - NIR new samples.dif ファイルを選択します。

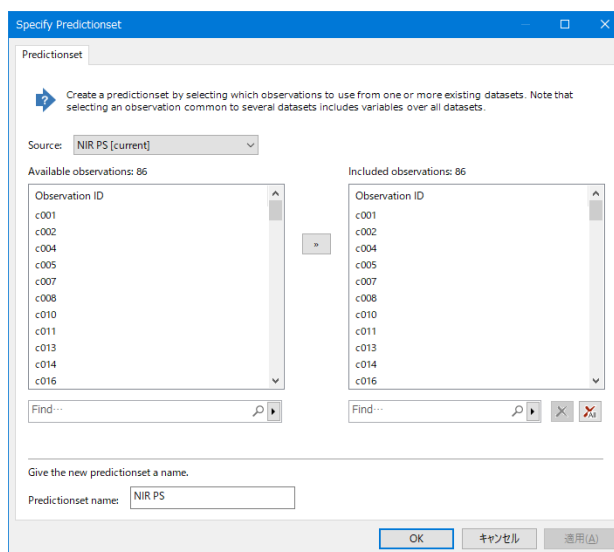
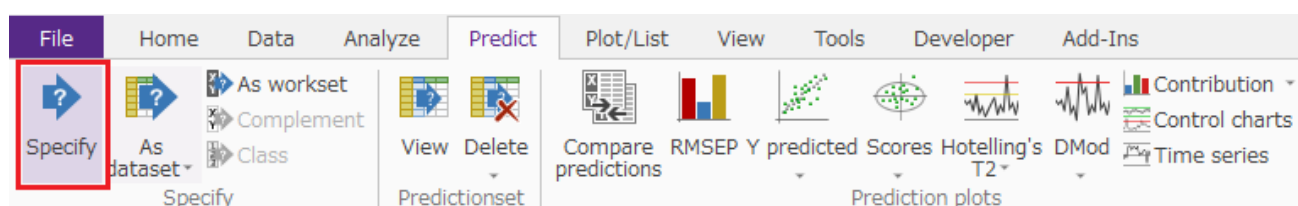


データセットのインポートと同じ方法でファイルを読み込みます（5～8 ページ）

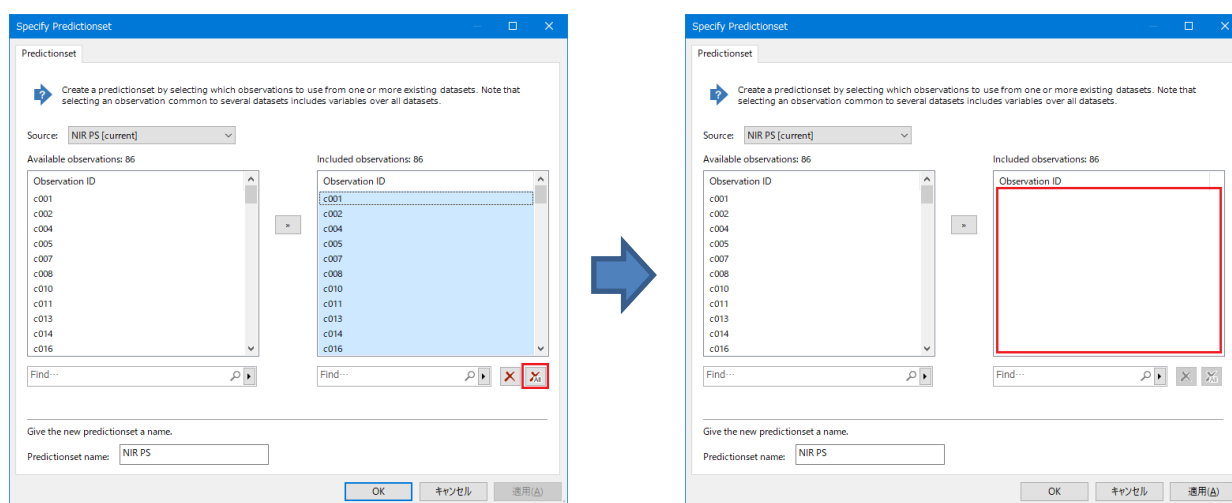
各データのシート名をクリックして Dataset - NIR new samples を NIR2 に、Dataset - NIR new samples (2) を Type2 に変更します。



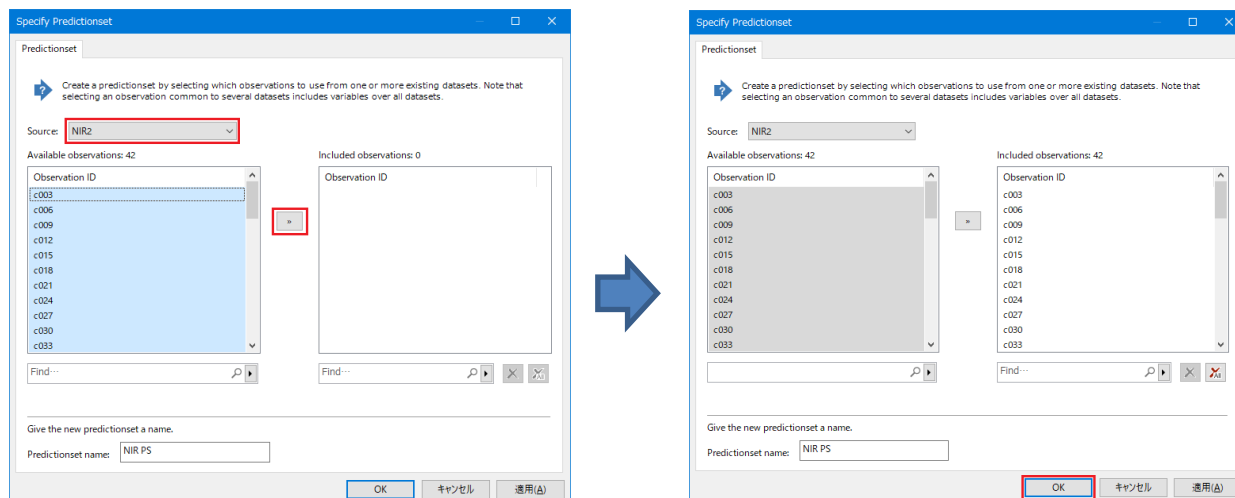
Predict > Specify をクリックし、Specify Predictionsetを開きます。



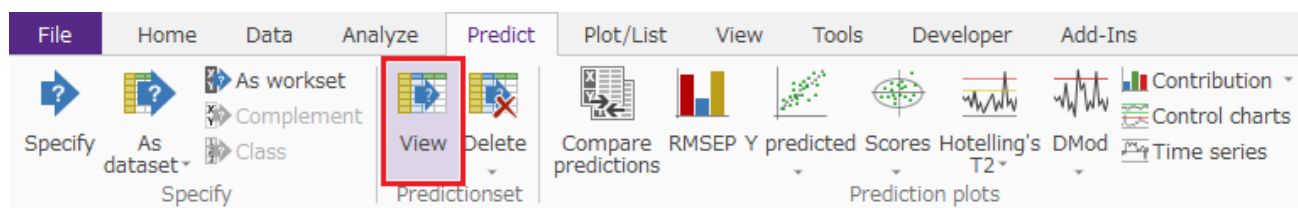
Predictionset には、モデル構築時に使用したデータが残っているので Included observation の赤色×アイコンをクリックして予測対象から外します。



Source> NIR2 を選択し Available observation 上で Ctrl+A で全選択後、画面中央の「>>」アイコンをクリックして Included observation に加えて、OK をクリックします。



Predict > View をクリックし、Predictionset を開きます。YPredPS に予測値が表示されます。



Predictionset - NIR_PS [M6]						
1	2	3	4	5	6	7
Primary ID	Lambda	YPredPS[1](Lambda)	Set	PModXPS+[1]	DModXPS+[1](Norm), Weighted residuals	C.I. tPS[1]
c003	54.5582	54.5622	TS	0.9879	0.797729	6.44264e-05
c006	36.4958	32.639	TS	0.876341	0.909271	4.13935e-05
c009	18.2948	16.7673	TS	1	0.532546	2.24714e-05
c012	18.309	12.6054	TS	0.012454	1.33812	3.57693e-05
c015	0	-0.630079	TS	0.99997	0.654012	3.98104e-05
c018	0	2.98582	TS	0.00331752	1.30331	5.44163e-05
c021	0	1.953	TS	0.95013	0.854664	3.10893e-05
c024	0	2.943	TS	0.638095	0.964391	3.99643e-05
c027	90.3	101.486	TS	0.495144	1.02751	0.000127105
c030	36.4601	34.6585	TS	0.992437	0.784289	3.87469e-05
c033	36.5106	37.0552	TS	0.208218	1.12446	5.82382e-05
c036	18.3083	20.3825	TS	0.13859	1.13595	2.21066e-05
c039	18.3165	16.7561	TS	0.999997	0.64864	2.53609e-05
c042	0	-1.29458	TS	0.551737	1.02702	5.85875e-05
c045	0	1.51779	TS	0.456908	1.06718	3.35426e-05
c048	0	4.29745	TS	0.254375	1.02748	4.13746e-05
c051	0	2.75365	TS	0.994397	0.768537	3.3639e-05
c054	54.5261	55.6686	TS	1	0.571787	5.61848e-05
c057	36.4846	36.4889	TS	0.798515	0.900423	2.87193e-05