Python を用いたケモメトリクス演習講習会受講報告書

1. はじめに

1.1 講習会の概略

講習会の目的は、「分光装置等から得られるデータセットを Python で解析処理、可視化する手法を実際に演習で取得する。」です。

成分分析を行うためには、取得した複雑なスペクトルデータを解析し検量線を作成します。よく使われる解析が「多変量解析」です。この分野をケモメトリクス（計量化学）と呼びます。多変量解析は計算が複雑になるため、一般的には専用のソフトウエアを使用して行いますが、今回は、Python で取り組む講習会です。

講師(田中成昭先生)の著作である「Python で始める機器分析データの解析とケモメトリクス」で進められました。内容を大目次で紹介すると、

1　機器分析の世界　/ 2 Python の基礎　/ 3　統計の基礎

4　データの前処理と可視化 /　5　ケモメトリクスの基礎

6　次元削減　/ 7　クラスタリング　/ 8　回帰

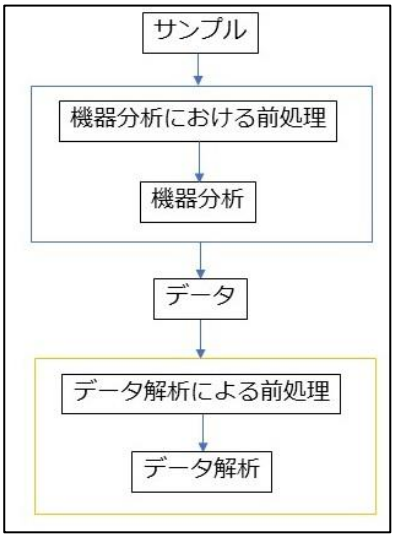
9　クラス分類 /　10　フィッティング /　11　2 次元相関分光法　　です。

今回は、１から４までの講習で、残りは次回に開催されるようです。

1.2 データ解析の考え方

機器分析の腕前と共に、データ解析の腕前も同等にして欲しいとのことです。

機器分析とデータ解析は、右の図で解説されました。



データ解析による前処理、とは、

ベースライン補正、検量線作成、等であり、

今回、講習した内容です。

データ解析関連のソフトについては、

・Excel

・Origin/Igor (グラフ作成ソフト)

・GRAMS (分光分析ソフト)　古い？

・Unscrambler/Pirouette　ケモメトリクス専用ソフト（100 万円程度）

・Python / R (無料)

・MATLAB　（100 万円程度）　　が挙げられます。

費用負担が難しい環境では、無料でできる Python / R が、有効と考えられ、

Python と R との比較では、Python は書籍、WEB　での情報量に優れているという利点があります。

2. 講習会の内容

2.1　Python によるケモメトリクスで使用するライブラリィ

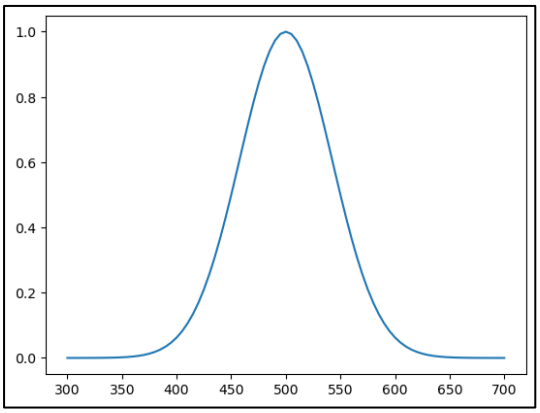
numpy	行列を扱う　計算が早い
pandas	エクセルのような整理ができる
matplotlib	グラフ作成ソフト
scipy	科学技術計算
scikit-learn	機械学習

2.2　データの概略

実測したデータではなく、関数で作成したガウスピーク波形で説明します。

縦軸が吸光度、横軸は波長と見てください。右がデータファイル構造です。

1 本のスペクトルですので、1 次配列（ベクトル）として扱います。



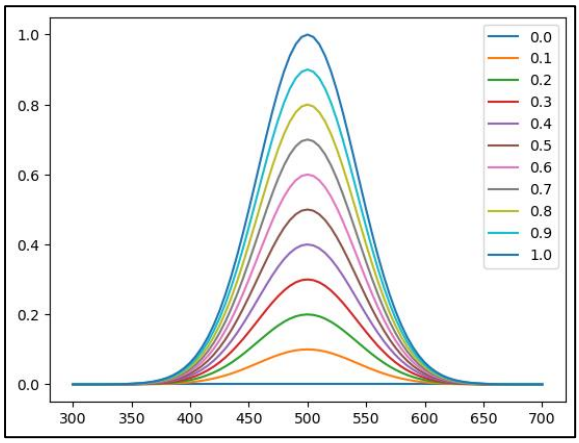
次に、濃度変化させた（11 種）複数のスペクトルが以下です。2 次元配列と呼ばれます。

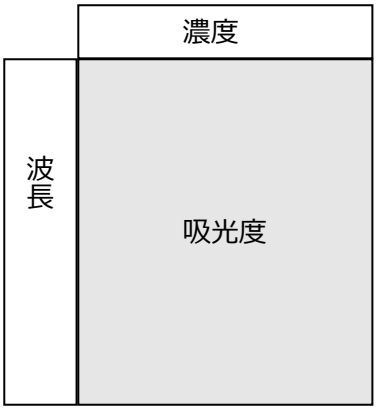
縦軸が吸光度、横軸は波長で、凡例としてあるのが、濃度です。

今回、配慮すべきは、装置から排出されるデータと、機械学習に使用するデータと、

プロットに使用するデータの間で、 x 軸、 y 軸が、異なります。

装置から排出されるデータと、プロット用のデータは以下となります。





プロット時の x 軸は、データの 1 列目になるので、波長を横軸にするなら、

上の右側のデータ構成になります。

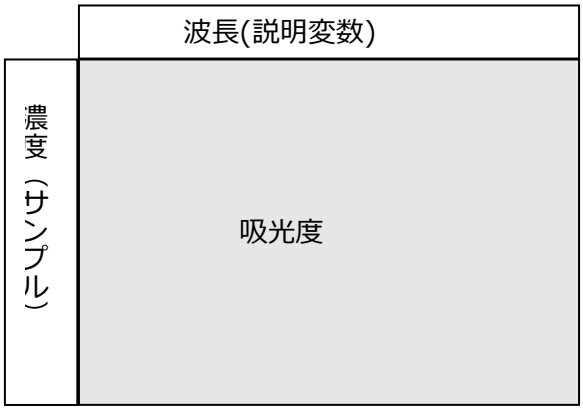
一方　機械学習には、以下のデータ構造となります。

この構造で、濃度を data.index

波長を data.columns 吸光度を

data.values とします。

（機械学習のアルゴリズムとのことです）



両データ構造の変換については、行列を転置する必要があります。（　.T を追加します）

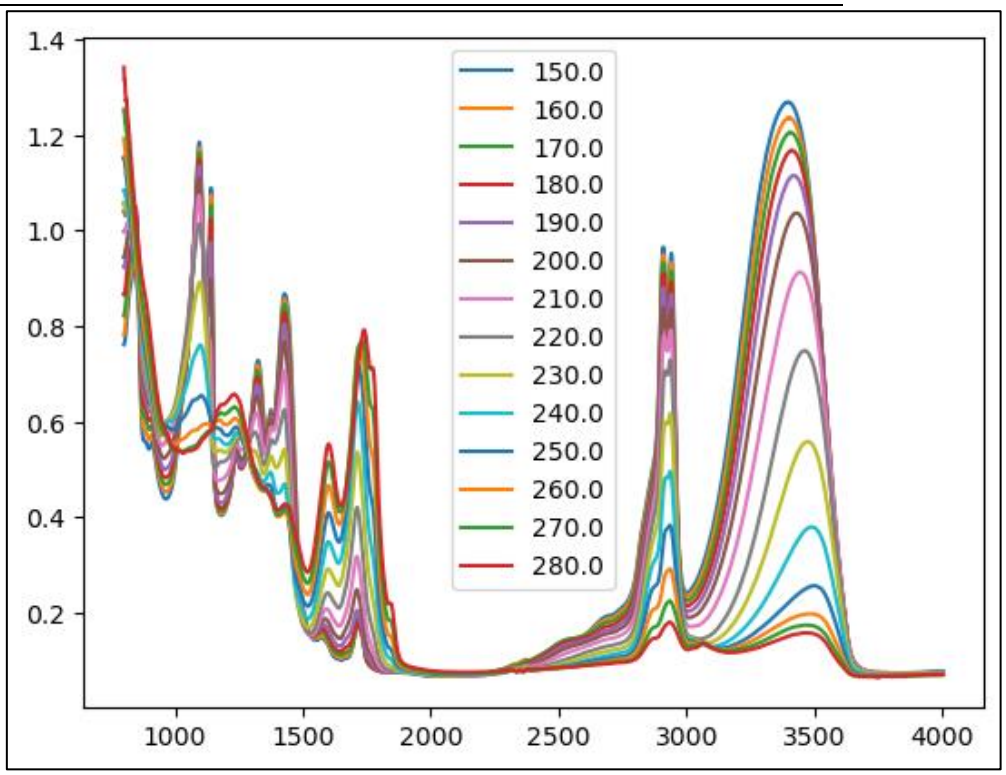
2.3　講習で学んだ事

今回は、データの前処理関連ですが、学んだ事は以下です。

（コードは記載していませんが、提供は可能です）

2.3.1　前処理１ (補完、ステップ、範囲指定、ベースライン補正、最大値、ピーク値)

元データ（PVA の温度依存赤外スペクトル）周波数と温度とスペクトル値



データ表

	800.33	801.29	802.26	803.22	804.18	805.15	806.11	807.08	808.04	809.01	...	3991.03	3991.99	3992.96	3993.92	3994.89	3995.85	3996.81	3997.78
150.0	0.7615	0.7634	0.7646	0.7652	0.7686	0.7706	0.7752	0.7787	0.7854	0.7913	...	0.0770	0.0771	0.0771	0.0771	0.0771	0.0771	0.0771	0.0771
160.0	0.7817	0.7825	0.7855	0.7878	0.7916	0.7955	0.8018	0.8057	0.8112	0.8160	...	0.0769	0.0769	0.0769	0.0769	0.0770	0.0770	0.0770	0.0770
170.0	0.8217	0.8227	0.8259	0.8301	0.8337	0.8345	0.8395	0.8445	0.8511	0.8548	...	0.0774	0.0774	0.0774	0.0774	0.0774	0.0774	0.0774	0.0775
180.0	0.8662	0.8652	0.8685	0.8713	0.8742	0.8754	0.8784	0.8816	0.8875	0.8912	...	0.0770	0.0770	0.0770	0.0770	0.0771	0.0771	0.0770	0.0771
190.0	0.9243	0.9222	0.9240	0.9254	0.9291	0.9311	0.9344	0.9374	0.9437	0.9459	...	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767
200.0	0.9439	0.9437	0.9457	0.9463	0.9490	0.9523	0.9571	0.9568	0.9602	0.9630	...	0.0779	0.0780	0.0780	0.0780	0.0780	0.0779	0.0779	0.0780
210.0	0.9978	0.9955	1.0007	1.0020	1.0029	1.0008	1.0018	1.0005	1.0034	1.0051	...	0.0769	0.0769	0.0770	0.0769	0.0769	0.0769	0.0769	0.0769
220.0	1.0414	1.0372	1.0389	1.0384	1.0364	1.0314	1.0349	1.0343	1.0358	1.0358	...	0.0753	0.0753	0.0753	0.0753	0.0753	0.0753	0.0753	0.0754
230.0	1.0594	1.0521	1.0573	1.0567	1.0542	1.0516	1.0564	1.0543	1.0523	1.0493	...	0.0745	0.0745	0.0745	0.0745	0.0745	0.0745	0.0745	0.0745
240.0	1.0832	1.0806	1.0840	1.0841	1.0837	1.0796	1.0760	1.0694	1.0688	1.0689	...	0.0733	0.0733	0.0733	0.0733	0.0733	0.0732	0.0733	0.0734
250.0	1.1529	1.1478	1.1495	1.1516	1.1488	1.1347	1.1329	1.1318	1.1236	1.1158	...	0.0724	0.0724	0.0724	0.0724	0.0724	0.0724	0.0724	0.0724
260.0	1.1928	1.1859	1.1824	1.1769	1.1755	1.1683	1.1629	1.1574	1.1550	1.1492	...	0.0720	0.0720	0.0720	0.0720	0.0720	0.0720	0.0720	0.0720
270.0	1.2536	1.2480	1.2479	1.2388	1.2345	1.2259	1.2239	1.2207	1.2168	1.2075	...	0.0704	0.0704	0.0704	0.0704	0.0703	0.0704	0.0704	0.0704
280.0	1.3414	1.3262	1.3157	1.3126	1.3160	1.2987	1.2910	1.2851	1.2787	1.2703	...	0.0716	0.0716	0.0716	0.0716	0.0716	0.0716	0.0716	0.0716

補完　→　周波数の数字を、１例ですが、800.3 から 801.0 に補完します。（装置間の違いの補完に利用可）補完は、線形補完、2 次/3 次スプライン補完の方法があります。

	801.0	802.0	803.0	804.0	805.0	806.0	807.0	808.0	809.0	810.0	...	3990.0	3991.0	3992.0	3993.0	3994.0
150.0	0.76289	0.764393	0.764833	0.768036	0.770143	0.774722	0.778298	0.785128	0.791235	0.79841	...	0.077001	0.076998	0.077101	0.0771	0.0771
160.0	0.781972	0.784718	0.787176	0.790884	0.794722	0.801182	0.805332	0.81099	0.815932	0.824541	...	0.076999	0.076902	0.0769	0.076899	0.076905
170.0	0.822153	0.824896	0.829118	0.833283	0.834157	0.838895	0.84401	0.850889	0.854743	0.86263	...	0.0774	0.0774	0.0774	0.0774	0.0774
180.0	0.864966	0.867548	0.870651	0.87379	0.875128	0.878051	0.881218	0.887303	0.891141	0.89922	...	0.077104	0.077002	0.077	0.077	0.077005
190.0	0.922339	0.923485	0.924899	0.928492	0.930734	0.934052	0.937007	0.94351	0.945873	0.949258	...	0.0767	0.0767	0.0767	0.0767	0.0767
200.0	0.943456	0.945254	0.946045	0.948436	0.951633	0.956809	0.956722	0.960067	0.962952	0.96957	...	0.078005	0.0779	0.078001	0.077999	0.078001
210.0	0.995197	0.99944	1.001812	1.002969	1.000954	1.00179	1.000465	1.003295	1.00507	1.009293	...	0.0769	0.0769	0.076901	0.077	0.076895
220.0	1.037702	1.038463	1.038649	1.037106	1.031653	1.034612	1.034307	1.035757	1.035798	1.035962	...	0.075404	0.075302	0.0753	0.0753	0.0753
230.0	1.052659	1.056042	1.057158	1.054788	1.051477	1.056098	1.054525	1.052411	1.049311	1.049645	...	0.0745	0.0745	0.0745	0.0745	0.0745
240.0	1.080582	1.083211	1.084171	1.084035	1.080225	1.076588	1.069754	1.068787	1.068898	1.068536	...	0.0733	0.0733	0.0733	0.0733	0.073301
250.0	1.148525	1.148789	1.151338	1.150327	1.136324	1.132856	1.132161	1.123987	1.115834	1.115836	...	0.0724	0.0724	0.0724	0.0724	0.0724
260.0	1.1875	1.183496	1.177828	1.176084	1.169407	1.16352	1.157717	1.155175	1.14923	1.149267	...	0.072104	0.072002	0.072	0.072	0.072
270.0	1.248803	1.248633	1.240695	1.235542	1.226905	1.224104	1.220968	1.217095	1.207558	1.206147	...	0.070401	0.0704	0.0704	0.070401	0.070394
280.0	1.330318	1.317914	1.312344	1.316767	1.301168	1.291645	1.285584	1.27905	1.270321	1.273717	...	0.071704	0.071602	0.0716	0.0716	0.0716

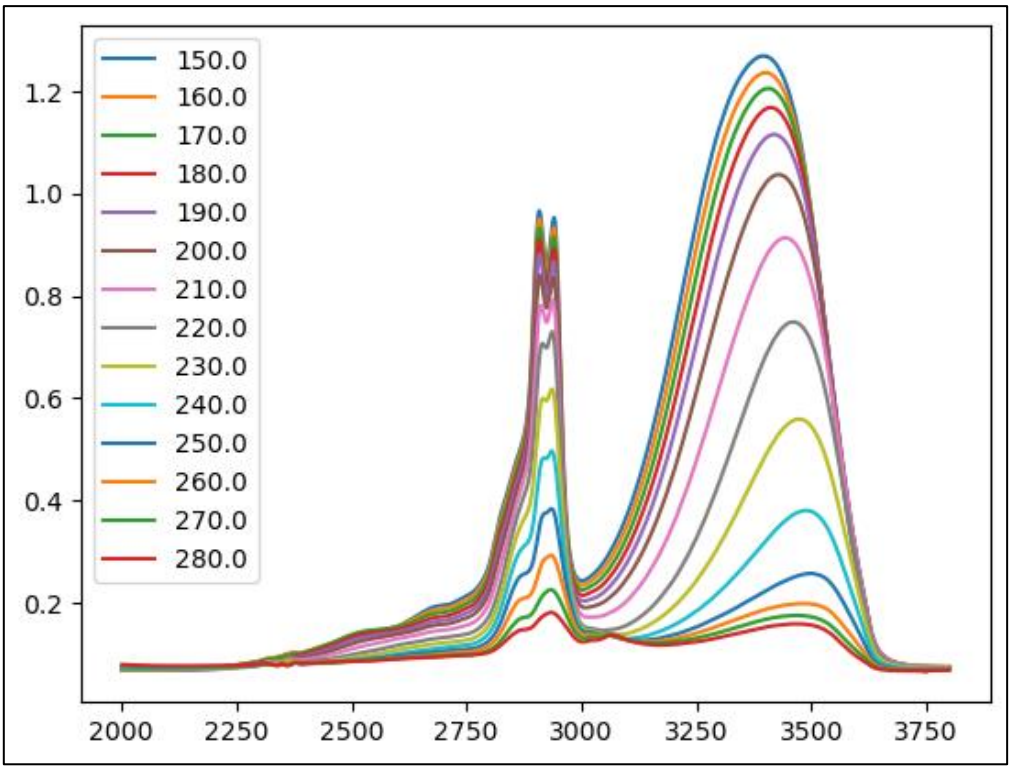
ステップ　→　温度の表示を 1 個飛ばしにします。

	800.33	801.29	802.26	803.22	804.18	805.15	806.11	807.08	808.04	809.01	...	3991.03	3991.99	3992.96	3993.92	3994.89	3995.85	3996.81	3997.78
150.0	0.7615	0.7634	0.7646	0.7652	0.7686	0.7706	0.7752	0.7787	0.7854	0.7913	...	0.0770	0.0771	0.0771	0.0771	0.0771	0.0771	0.0771	0.0771
170.0	0.8217	0.8227	0.8259	0.8301	0.8337	0.8345	0.8395	0.8445	0.8511	0.8548	...	0.0774	0.0774	0.0774	0.0774	0.0774	0.0774	0.0774	0.0775
190.0	0.9243	0.9222	0.9240	0.9254	0.9291	0.9311	0.9344	0.9374	0.9437	0.9459	...	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767
210.0	0.9978	0.9955	1.0007	1.0020	1.0029	1.0008	1.0018	1.0005	1.0034	1.0051	...	0.0769	0.0769	0.0770	0.0769	0.0769	0.0769	0.0769	0.0769
230.0	1.0594	1.0521	1.0573	1.0567	1.0542	1.0516	1.0564	1.0543	1.0523	1.0493	...	0.0745	0.0745	0.0745	0.0745	0.0745	0.0745	0.0745	0.0745
250.0	1.1529	1.1478	1.1495	1.1516	1.1488	1.1347	1.1329	1.1318	1.1236	1.1158	...	0.0724	0.0724	0.0724	0.0724	0.0724	0.0724	0.0724	0.0724
270.0	1.2536	1.2480	1.2479	1.2388	1.2345	1.2259	1.2239	1.2207	1.2168	1.2075	...	0.0704	0.0704	0.0704	0.0704	0.0703	0.0704	0.0704	0.0704

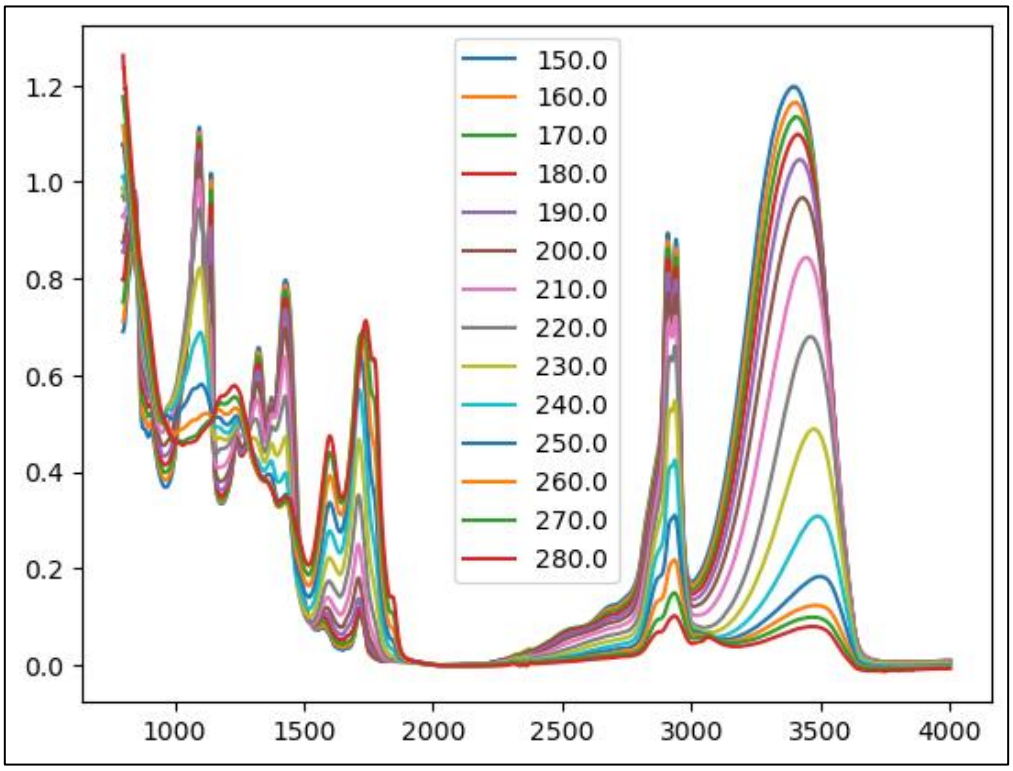
ステップ　→　周波数の表示を 1 個飛ばしにします。

	800.33	802.26	804.18	806.11	808.04	809.97	811.90	813.83	815.75	817.68	...	3982.35	3984.28	3986.21	3988.14	3990.06	3991.99	3993.92	3995.85
150.0	0.7615	0.7646	0.7686	0.7752	0.7854	0.7982	0.8128	0.8280	0.8440	0.8560	...	0.0767	0.0768	0.0769	0.0770	0.0770	0.0771	0.0771	0.0771
160.0	0.7817	0.7855	0.7916	0.8018	0.8112	0.8243	0.8384	0.8553	0.8693	0.8793	...	0.0767	0.0767	0.0769	0.0769	0.0770	0.0769	0.0769	0.0770
170.0	0.8217	0.8259	0.8337	0.8395	0.8511	0.8624	0.8749	0.8885	0.9057	0.9137	...	0.0772	0.0773	0.0773	0.0774	0.0774	0.0774	0.0774	0.0774
180.0	0.8662	0.8685	0.8742	0.8784	0.8875	0.8990	0.9124	0.9263	0.9390	0.9446	...	0.0769	0.0769	0.0770	0.0770	0.0771	0.0770	0.0770	0.0771
190.0	0.9243	0.9240	0.9291	0.9344	0.9437	0.9492	0.9574	0.9704	0.9775	0.9840	...	0.0766	0.0766	0.0767	0.0767	0.0767	0.0767	0.0767	0.0767
200.0	0.9439	0.9457	0.9490	0.9571	0.9602	0.9694	0.9768	0.9850	0.9901	0.9954	...	0.0778	0.0778	0.0779	0.0779	0.0780	0.0780	0.0780	0.0779
210.0	0.9978	1.0007	1.0029	1.0018	1.0034	1.0092	1.0116	1.0149	1.0182	1.0178	...	0.0768	0.0768	0.0769	0.0769	0.0769	0.0769	0.0769	0.0769
220.0	1.0414	1.0389	1.0364	1.0349	1.0358	1.0360	1.0350	1.0339	1.0322	1.0250	...	0.0752	0.0753	0.0753	0.0753	0.0754	0.0753	0.0753	0.0753
230.0	1.0594	1.0573	1.0542	1.0564	1.0523	1.0497	1.0442	1.0435	1.0371	1.0310	...	0.0744	0.0744	0.0745	0.0745	0.0745	0.0745	0.0745	0.0745
240.0	1.0832	1.0840	1.0837	1.0760	1.0688	1.0686	1.0644	1.0568	1.0523	1.0427	...	0.0732	0.0732	0.0733	0.0733	0.0733	0.0733	0.0733	0.0732
250.0	1.1529	1.1495	1.1488	1.1329	1.1236	1.1159	1.1077	1.0963	1.0867	1.0781	...	0.0723	0.0723	0.0723	0.0724	0.0724	0.0724	0.0724	0.0724
260.0	1.1928	1.1824	1.1755	1.1629	1.1550	1.1493	1.1424	1.1309	1.1224	1.1134	...	0.0720	0.0720	0.0720	0.0720	0.0721	0.0720	0.0720	0.0720
270.0	1.2536	1.2479	1.2345	1.2239	1.2168	1.2062	1.2022	1.1859	1.1749	1.1606	...	0.0703	0.0703	0.0703	0.0703	0.0704	0.0704	0.0704	0.0704
280.0	1.3414	1.3157	1.3160	1.2910	1.2787	1.2737	1.2643	1.2476	1.2357	1.2226	...	0.0716	0.0715	0.0716	0.0716	0.0717	0.0716	0.0716	0.0716

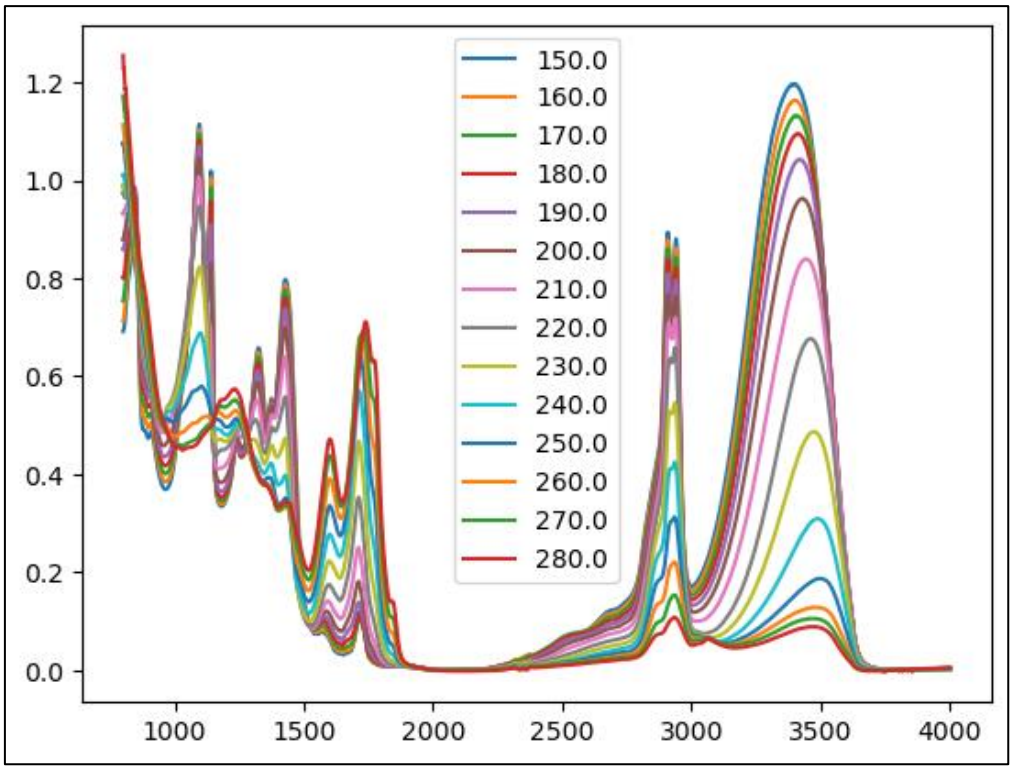
横軸の範囲指定　→　2000　から　3800　の範囲にします。



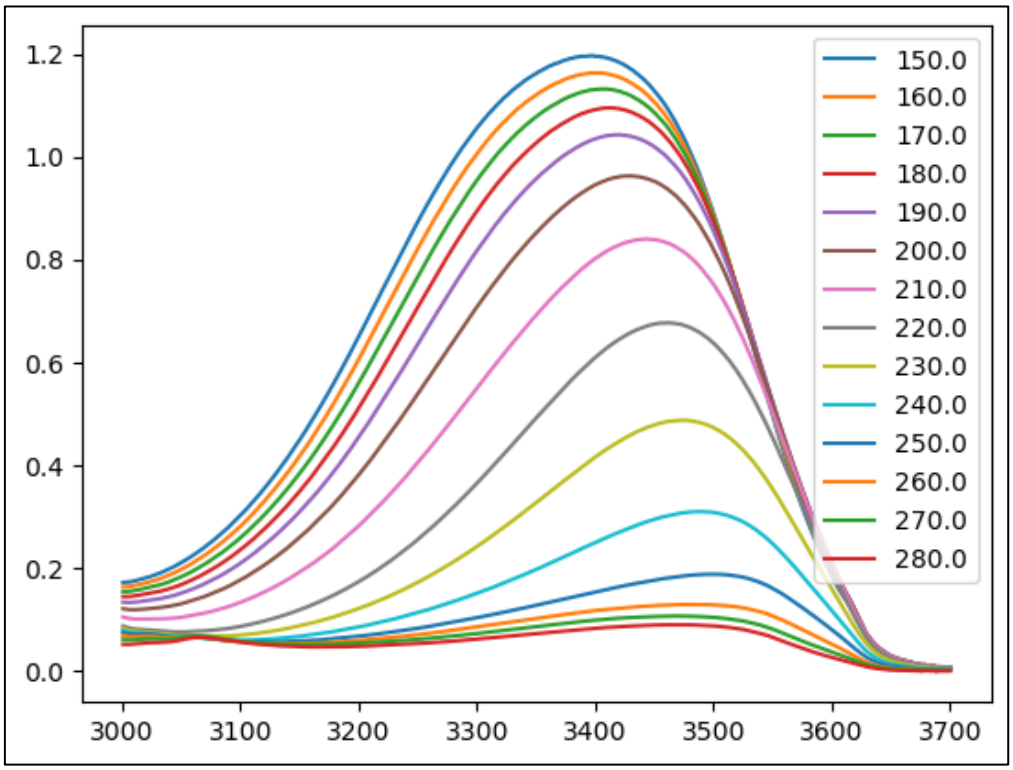
ベースライン補正（1 点で）→　2000 のところで、補正します。



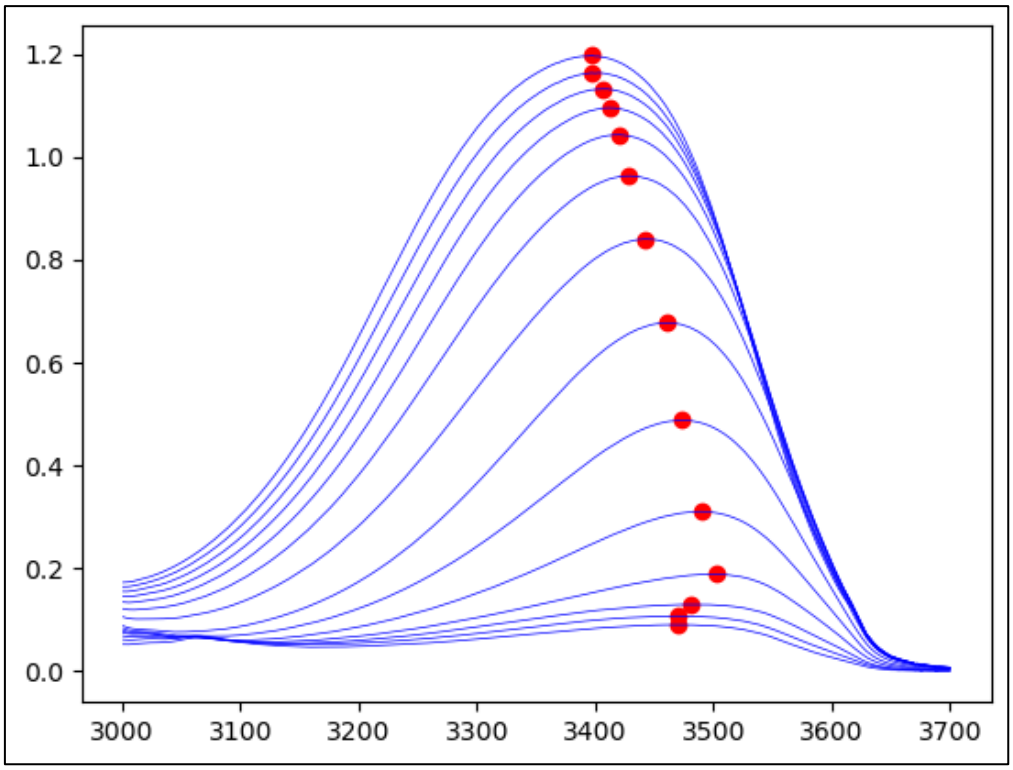
ベースライン補正（2 点で）→　2000,3800 のところで、線分により、補正します。



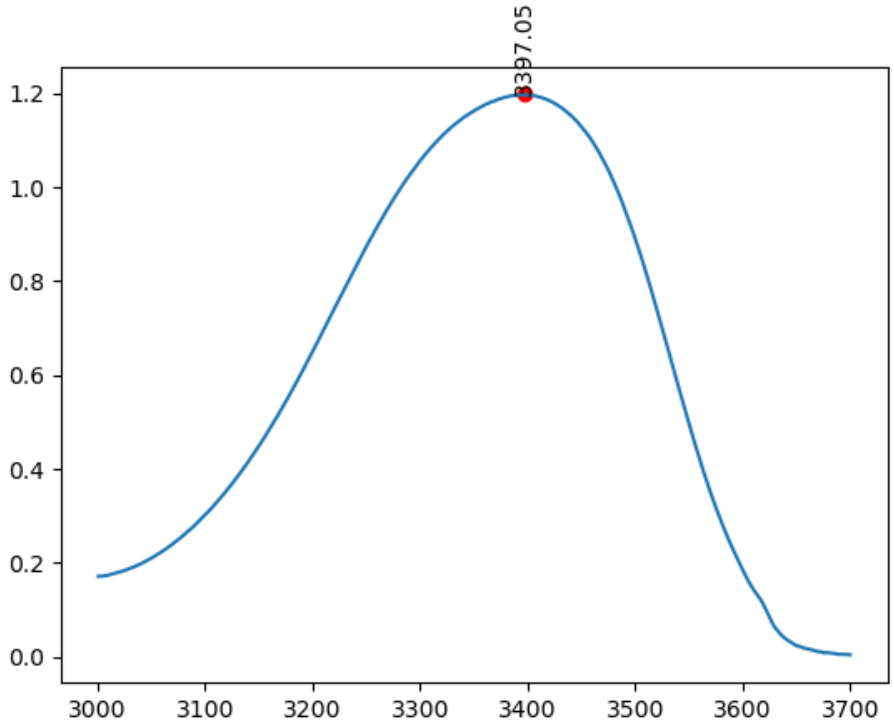
横軸の範囲指定　→　3000　から　3700　の範囲にします。



最大値検出　→　値の特定も可能です。



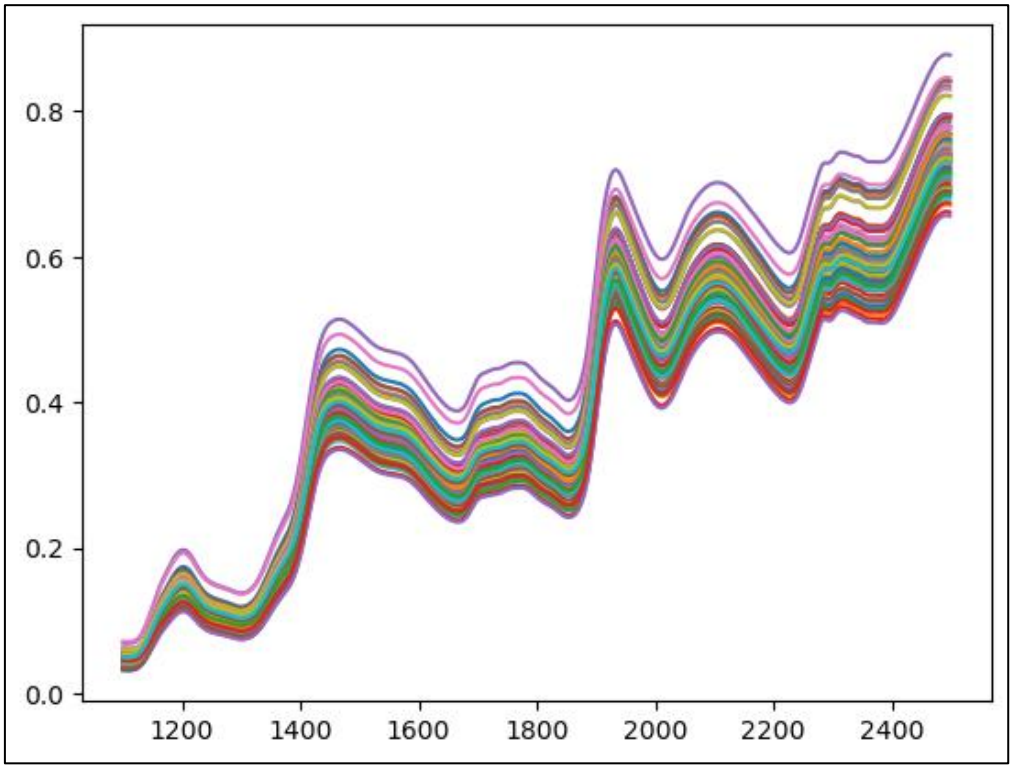
ピーク値検出　→　値の特定も可能です。ピーク検出の閾値を 0.5 にしています。



2.3.2　前処理２(センタリング、スケーリング、SNV,MSC, Savitzky-Golay フィルター)

機械学習に適したデータ（特徴量が明確）に変換する方法の説明です。

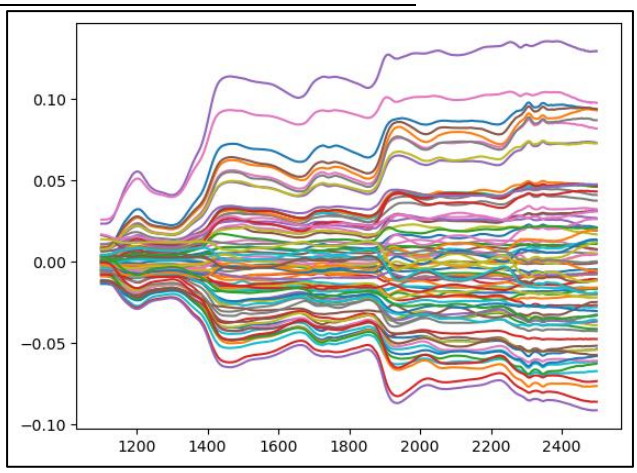
元データ（トウモロコシの近赤外スペクトル）周波数と異なるサンプルとスペクトル値



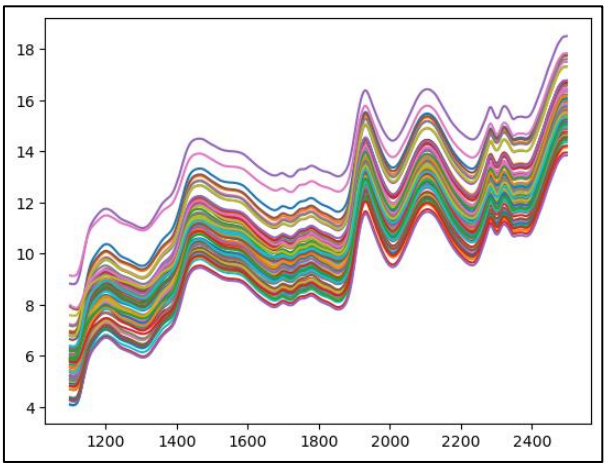
データ表

	1100	1102	1104	1106	1108	1110	1112	1114	1116	1118	...	2480	2482	2484	2486	2488	
0	0.044495	0.044383	0.044258	0.044212	0.044184	0.044229	0.044323	0.044451	0.044668	0.045067	...	0.727260	0.728491	0.729454	0.730516	0.731137	0
1	0.046504	0.046349	0.046230	0.046205	0.046183	0.046192	0.046329	0.046497	0.046735	0.047097	...	0.723239	0.724576	0.725750	0.726709	0.727484	0
2	0.046958	0.046817	0.046663	0.046601	0.046599	0.046639	0.046701	0.046817	0.047045	0.047419	...	0.702499	0.703823	0.704844	0.705664	0.706383	0
3	0.045461	0.045321	0.045205	0.045159	0.045152	0.045188	0.045300	0.045463	0.045664	0.046025	...	0.696089	0.697367	0.698669	0.699479	0.700075	0
4	0.053948	0.053786	0.053650	0.053613	0.053576	0.053623	0.053759	0.053915	0.054199	0.054589	...	0.770826	0.772279	0.773220	0.774120	0.774652	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
75	0.049494	0.049362	0.049267	0.049176	0.049150	0.049147	0.049206	0.049354	0.049572	0.049935	...	0.731541	0.733183	0.734272	0.735387	0.735956	0
76	0.071578	0.071420	0.071306	0.071222	0.071218	0.071220	0.071329	0.071560	0.071839	0.072276	...	0.841469	0.843240	0.844080	0.845105	0.845669	0
77	0.049717	0.049573	0.049471	0.049391	0.049386	0.049382	0.049458	0.049639	0.049883	0.050283	...	0.743541	0.744987	0.746214	0.747167	0.747683	0
78	0.059438	0.059321	0.059222	0.059123	0.059093	0.059085	0.059157	0.059281	0.059488	0.059826	...	0.731438	0.732894	0.734180	0.735227	0.735765	0
79	0.050085	0.049928	0.049829	0.049758	0.049746	0.049741	0.049814	0.050011	0.050265	0.050677	...	0.724219	0.725795	0.726721	0.727717	0.728331	0

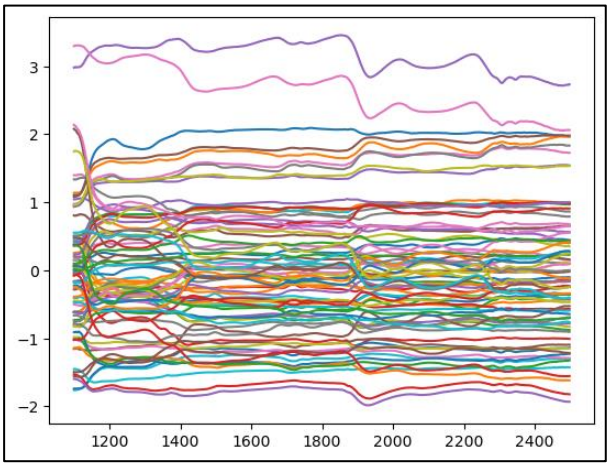
センタリング（平均値から差分の値）処理



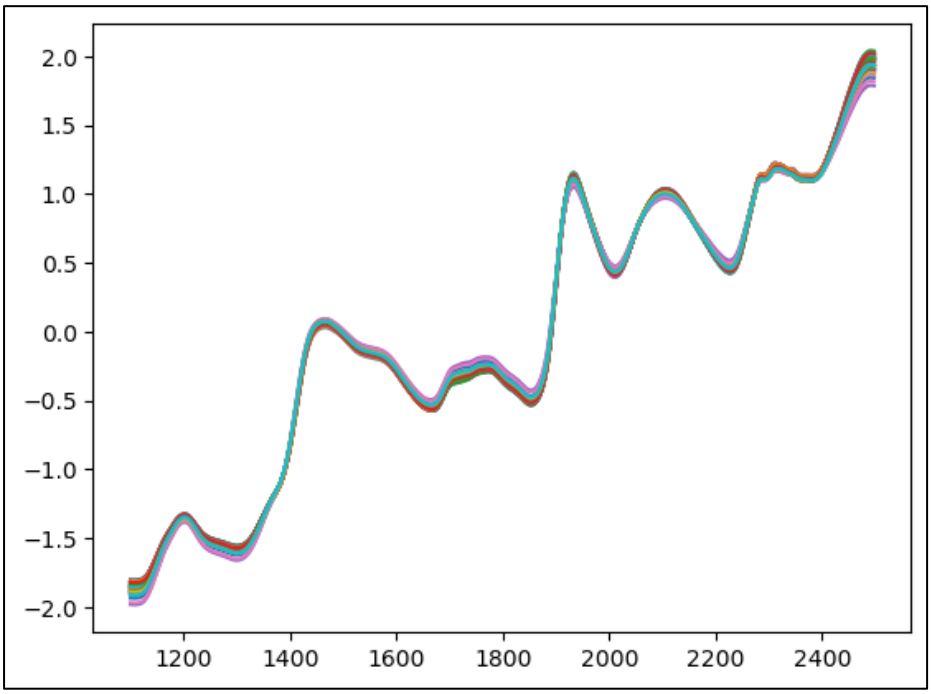
スケーリング（標準偏差で割る）処理



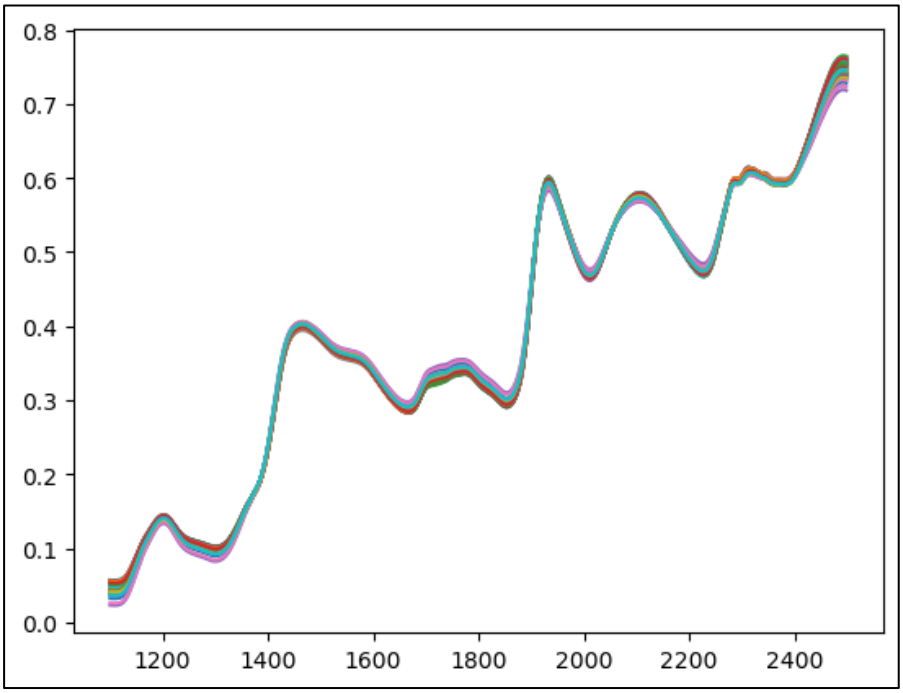
オートスケーリング（センタリングして、標準偏差で割る）処理



Standard Normal Variate(SNV) 処理　→　各サンプルで平均を０、標準偏差を１にする。拡散反射スペクトルのように、散乱の影響でベースラインと信号強度のどちらも変化してしまうときに有効なベースライン補正。

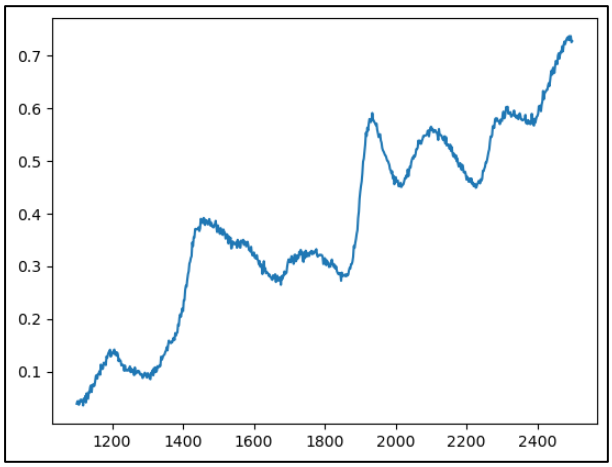


Multiplicative Scatter Correction (MSC)　→　SNV と同様に散乱の影響でベースラインと信号強度のどちらも変化してしまう拡散反射スペクトルのベースライン補正に用いられる。

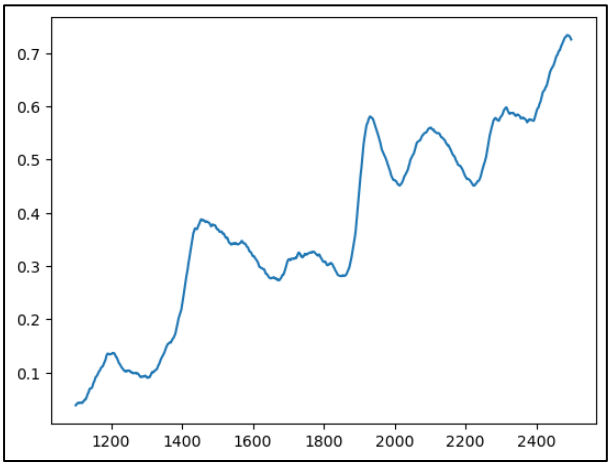


Savitzky-Golay フィルター → window : 窓幅、polynom : 多項式の次数、order : 微分の次数 を設定して作成します。

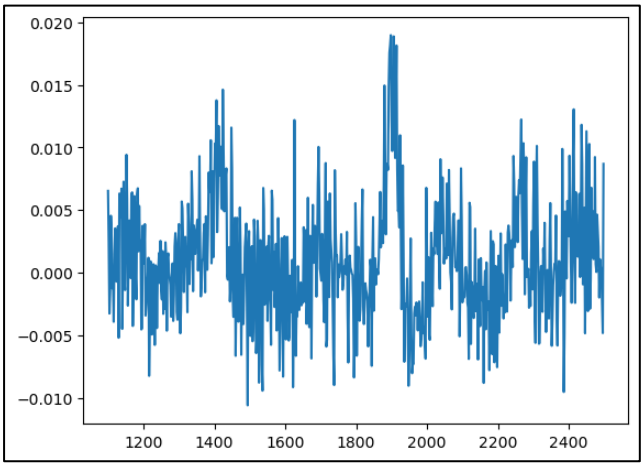
window = 3 polynom =2 order =0 で作成



window = 10 polynom =2 order =0 で作成 (曲線は滑になる)

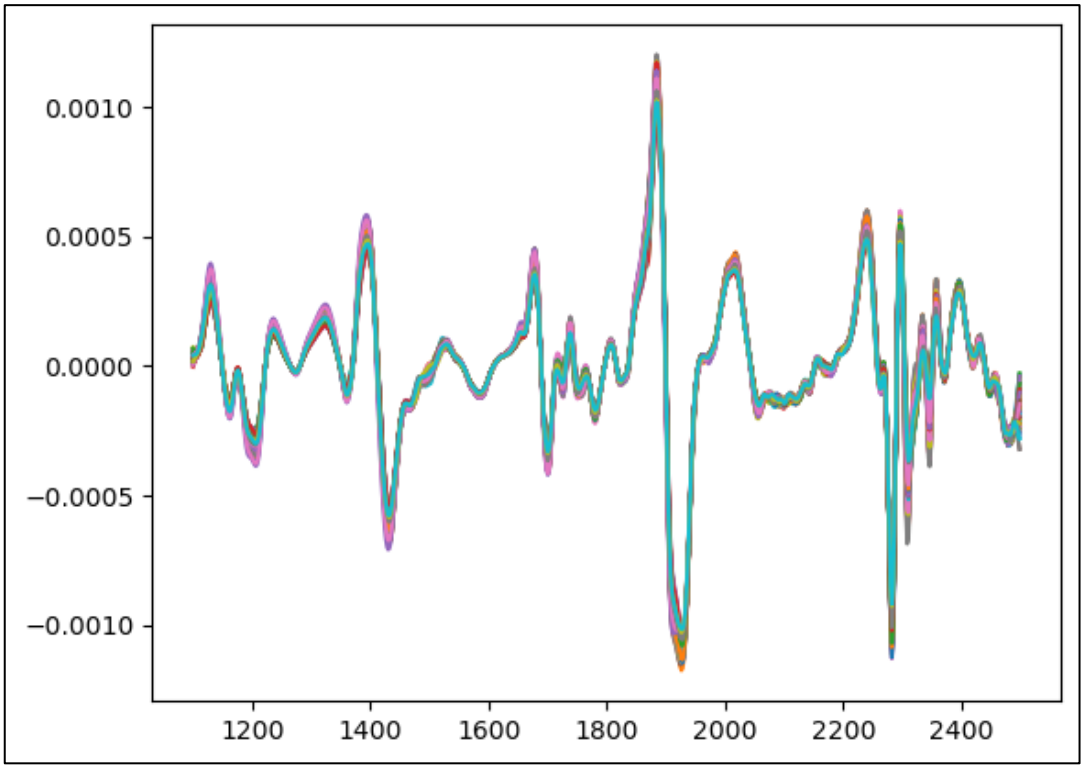


window = 3 polynom =2 order =1 で作成 (分かりにくくなる)



Savitzky-Golay フィルターによる 2 次微分という方法があります。

→0 次微分と 1 時微分を 2 回繰り返します。見えやすくなりました。



今回の研修は、ここまででした。

　データの前処理ができたので、次から、回帰、クラス分類等も進むとのことです。

ケモメトリクスに限らず、機械学習は、モデルの構築は標準化されているので、計算の開始は容易ですが、データの前処理を適切にしないと、うまくいかないですし、

説明変数の選択も大きい要因となります。

　次回の研修を楽しみにしています。

Savitzky-Golay フィルター　について、平滑化は理解できそうですが、2 次微分については、丁寧な説明がないように思います。

最後のページで、スペクトルの微分について、補足しました。

スペクトルの微分

実データではなく、ガウスピーク波形でのスペクトルの疑似としています。

・ピークの分離を微分で行います。

右図において、青の破線と橙色の破線のスペクトルがあるとして、

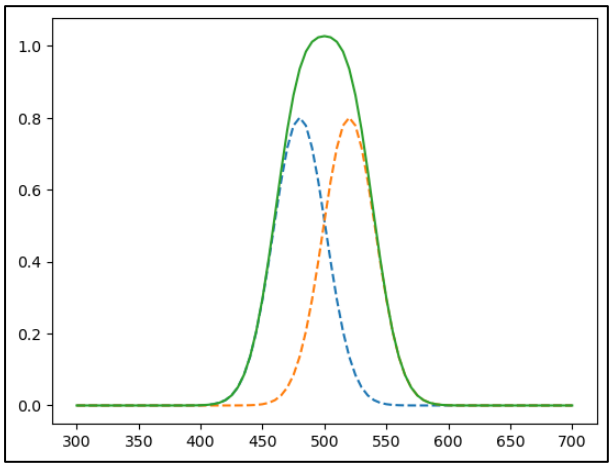
近接しているので、合計されて、

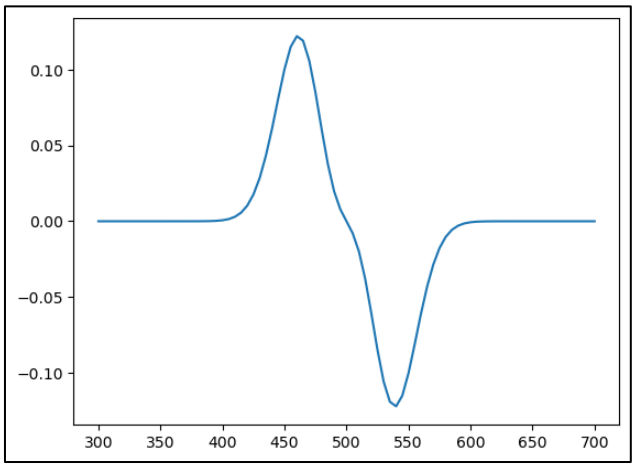
緑色の実線で表示された例です。

青破線は、中心が 480 で半値全幅は 50 です。

橙破線は中心が 520 で半値全幅は 50 です

右が 1 次微分した例です。





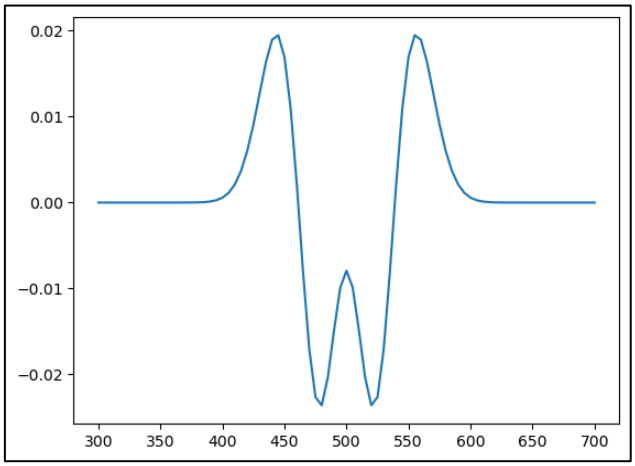
さらに、 2 次微分します。

480　と　520 のピークが

見えてきます。

450 と 550 の付近に、正の信号があらわれますが、元データにはないピークです。

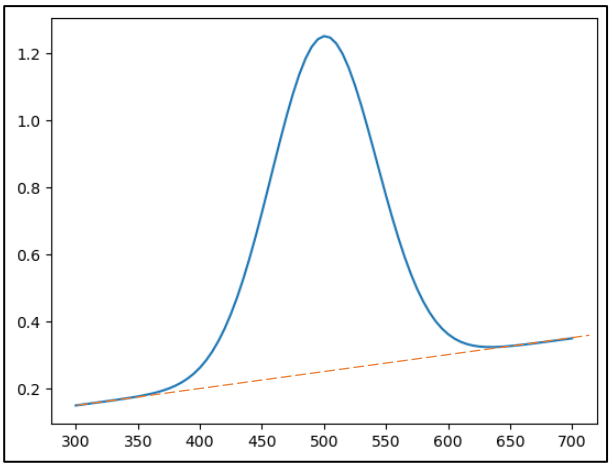
解説本は、正の信号には注意せよとのことでした。



・ベースライン補正を微分で行います。

中心が 480 で半値全幅は 50 のガウスピーク波形に、橙線破線のベースラインを加えて、右上がりのスペクトルを作成しました。

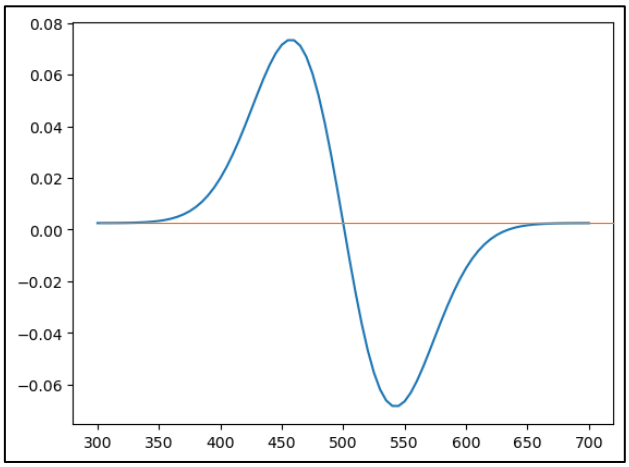
ベースラインは、y = 0.0005x という 1 次関数です。（ベースラインは直線という前提です）



右図は、1 次微分したものです。

ベースラインは、微分して、

y = 0.0005 という定数になります。



右図は、2 次微分したものです。

定数の微分ですので、 0 になり、

ベースライン補正しています。

また、元データは、1 個のピークですので、2 次微分すると、

マイナス方向にピークが見えます。