スペクトル・時系列データの前処理方法

～平滑化 (スムージング) と微分～

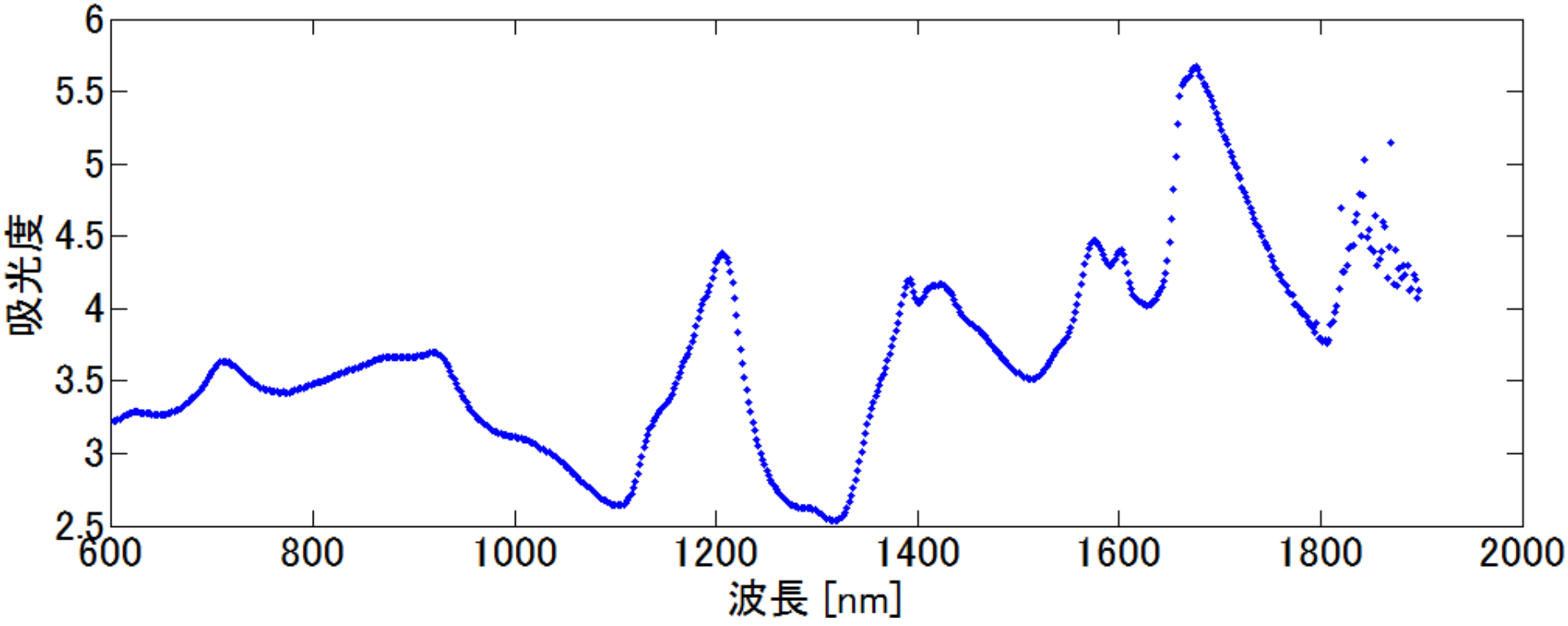
明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

スペクトルデータの特徴

✓波長 (波数) が近いと、吸光度 (強度) の値も似ている

✓ノイズが含まれる

✓吸光度 (強度) の極大値 (ピーク) 以外のデータも重要



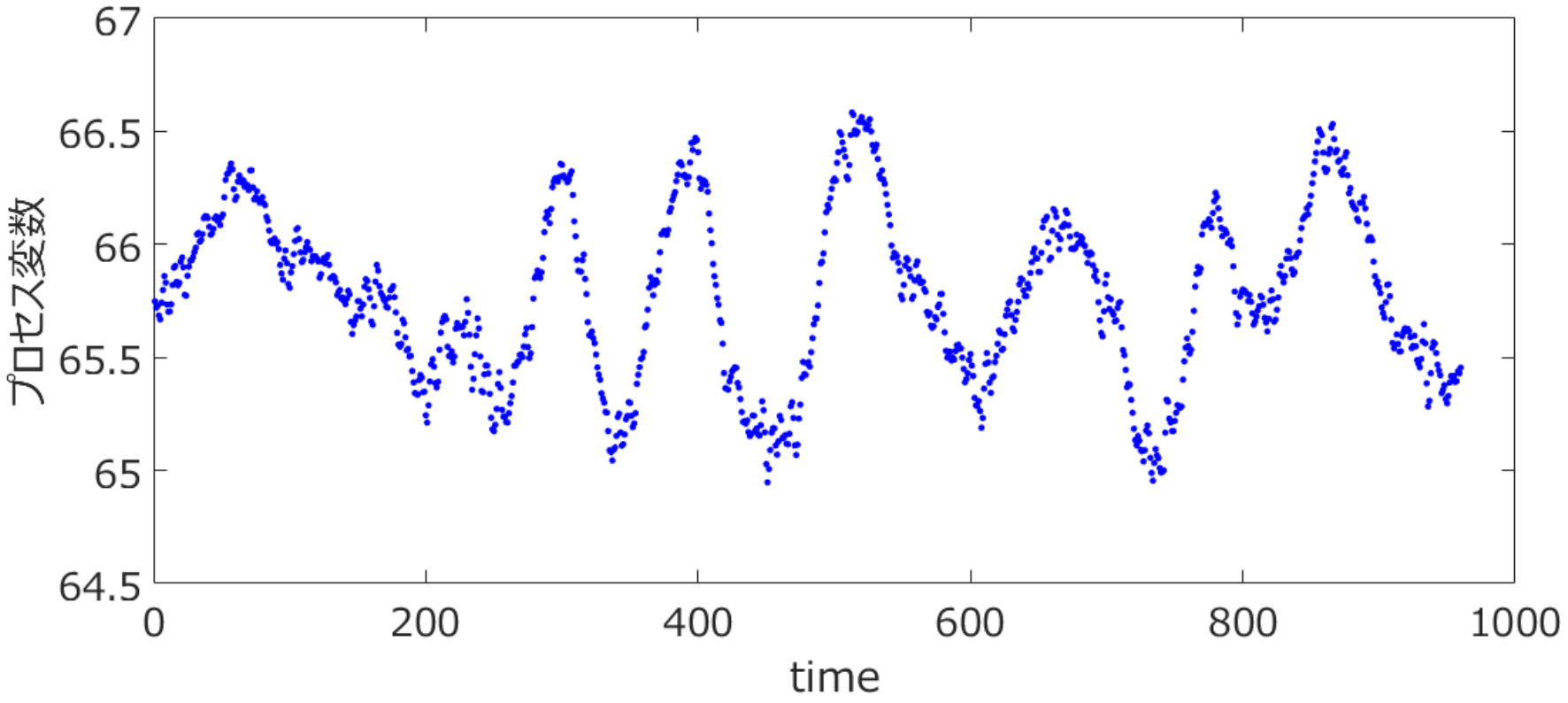
時系列データの特徴

✓時刻が近いと、プロセス変数の値も似ている

✓ノイズが含まれる

✓プロセス変数の極大値・極小値以外のデータも重要

✓時間が経つとデータが増える



スペクトル・時系列データ

✓スペクトル・時系列データの特徴は似ている

✓回帰分析・クラス分類の推定性能を向上させるための
データの前処理についても、同様の方法を適用できる

スペクトル・時系列データの前処理

✓平滑化 (スムージング)

- スペクトル・時系列データを “均す (ならす)” ことでノイズを低減する
- やりすぎて極大値・極小値の情報が消えないように注意する

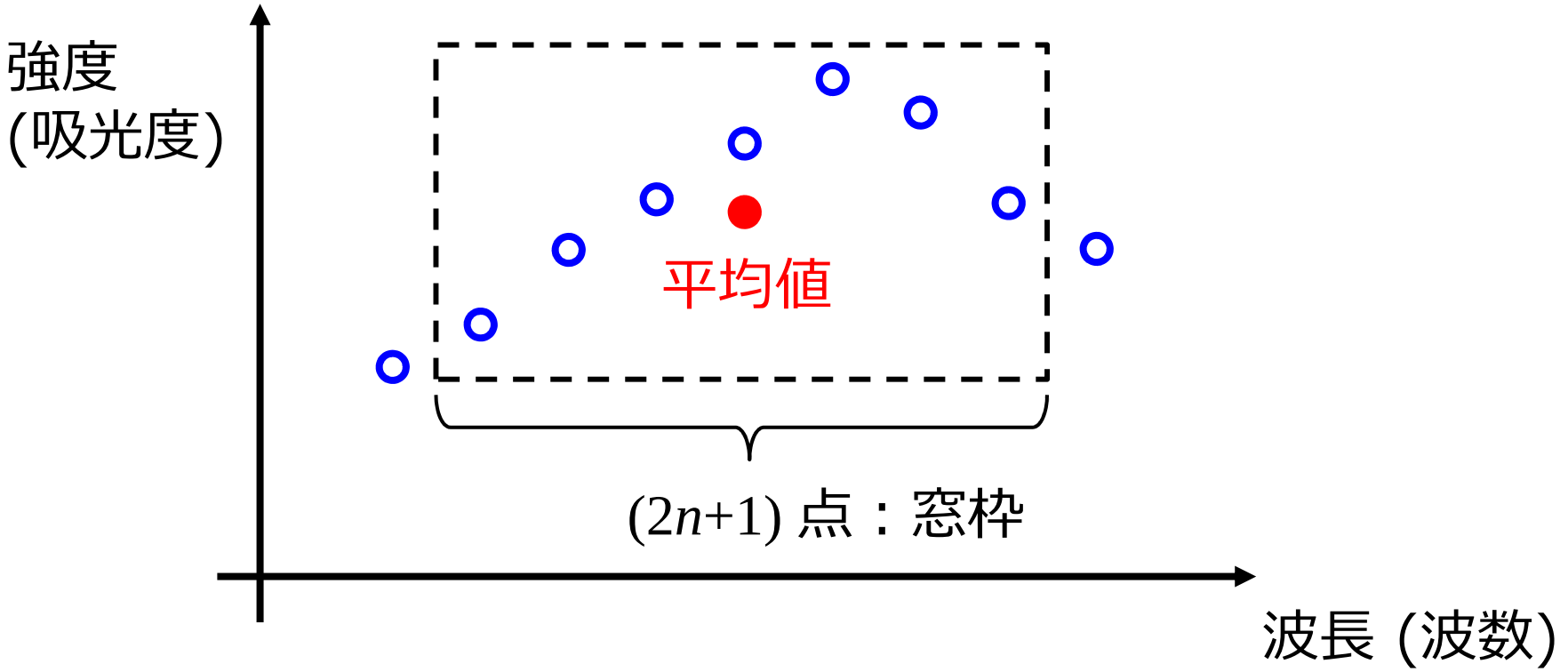
✓微分

- スペクトル・時系列データの傾きを計算することで、
  - ベースラインを補正する
  - 新しいスペクトル情報を抽出する
  - 時間変化を得る
- 一次微分、二次微分、三次微分、・・・
- 微分するとノイズが大きくなるので注意する

単純移動平均 (スペクトルデータ)

✓ある波長 (波数) の前後 n 点での強度 (吸光度) の平均値を、平滑化後の値にする

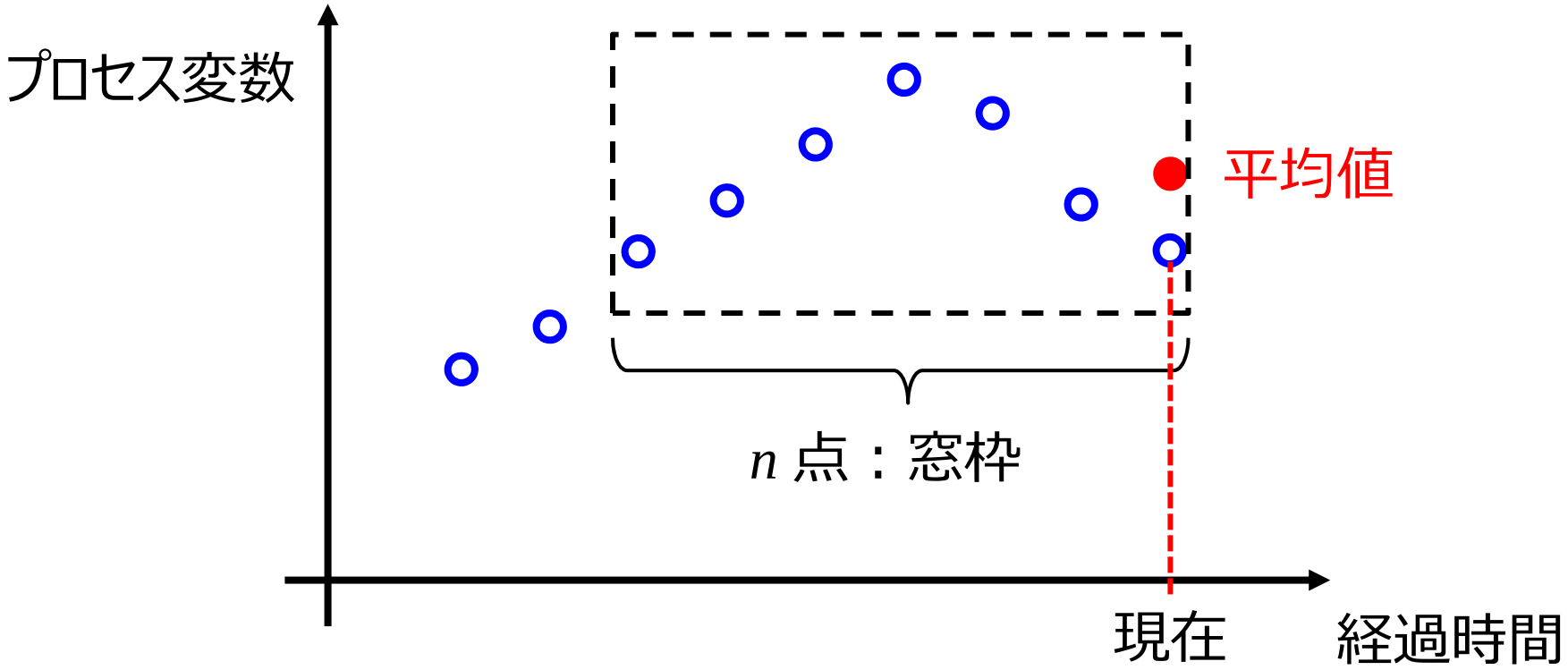
- 波長ごとに計算する
- $(2n+1)$ を 窓枠の数 と呼ぶ
- 端っこの波長については、$(2n+1)$ 点とれないこともある



単純移動平均 (時系列データ)

✓現在時刻の値を含めて、過去 n 点でのプロセス変数の平均値を、平滑化後の値にする (予測するときは 前後点 をとれないため)

- 時刻ごとに計算する
- n を 窓枠の数 と呼ぶ
- 初期時刻付近については、n 点とれないこともある



線形加重移動平均 (スペクトルデータ)

✓ある波長 (波数) の前後 n 点での強度 (吸光度) について、
対象の波長から離れるにつれて、線形に重みが小さくなる加重平均の
値を、平滑化後の値にする

- $(2n+1)$ を 窓枠の数 と呼ぶ

ある波長 i における強度を x_i とし、平滑化後の値を $x_{S,i}$ とすると、

$$x_{S,i} = \frac{x_{i-n} + 2x_{i-n+1} + \cdots + (n-1)x_{i-1} + nx_i + (n-1)x_{i+1} + \cdots + 2x_{i+n-1} + x_{i+n}}{1 + 2 + \cdots + (n-1) + n + (n-1) + \cdots + 2 + 1}$$

線形加重移動平均 (時系列データ)

✓現在時刻の値を含めて、過去 n 点でのプロセス変数の値について、現在時刻から離れるにつれて、線形に重みが小さくなる加重平均の値を、平滑化後の値にする

- $(2n+1)$ を 窓枠の数 と呼ぶ

ある時刻 t におけるプロセス変数の値を x_t とし、
平滑化後の値を $x_{S,t}$ とすると、

$$x_{S,t} = \frac{\sum_{j=1}^{n}\left\{(n-j+1)\,x_{t-j+1}\right\}}{\sum_{j=1}^{n}(n-j+1)}$$

指数加重移動平均 (スペクトルデータ)

✓ある波長 (波数) の前後 n 点での強度 (吸光度) について、
対象の波長から離れるにつれて、指数関数的に重みが小さくなる
加重平均の値を、平滑化後の値にする

- 波長からある程度離れると、重みはほぼ 0 になるため、
窓枠をある程度大きくしておけば、細かい数字は気にしなくてよい

ある波長 i における強度を x_i とし、平滑化後の値を $x_{S,i}$ とすると、

$$x_{S,i} = \frac{\cdots + \alpha^2 x_{i-2} + \alpha x_{i-1} + x_i + \alpha x_{i+1} + \alpha^2 x_{i+2} + \cdots}{\cdots + \alpha^2 + \alpha + 1 + \alpha + \alpha^2 + \cdots}$$

α を 平滑化係数 とよぶ

指数加重移動平均 (時系列データ)

✓現在時刻の値を含めて、過去 n 点でのプロセス変数の値について、現在時刻から離れるにつれて、指数関数的に重みが小さくなる加重平均の値を、平滑化後の値にする

- 波長からある程度離れると、重みはほぼ 0 になるため、窓枠をある程度大きくしておけば、細かい数字は気にしなくてよい

ある時刻 t におけるプロセス変数の値を x_t とし、
平滑化後の値を $x_{\mathrm{S},t}$ とすると、

$$x_{\mathrm{S},t} = \alpha\left\{x_t + (1-\alpha)\,x_{t-1} + (1-\alpha)^2\,x_{t-2} + \cdots\right\}$$

α を 平滑化係数 とよぶ

微分

✓隣の波長・時刻における値との差分をとることで、一次微分

✓一次微分の値について、隣の波長・時刻における値との差分をとることで、二次微分

✓・・・

Savitzky-Golay (SG) 法 [1,2]

✓データの平滑化と微分とを同時に行う方法

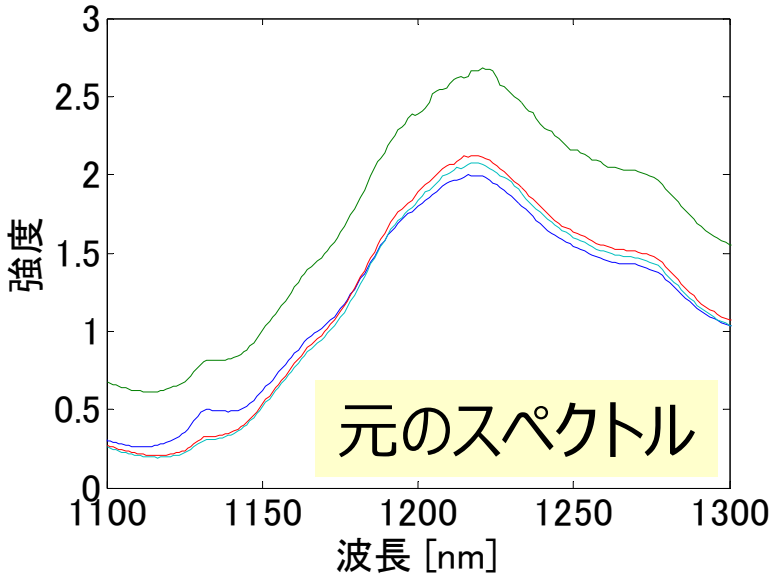
- 窓枠のデータを多項式で近似して、多項式の計算値を平滑化後の値とする
- 多項式の微分係数を微分後の値とする
  - 波長や時刻ごとに計算

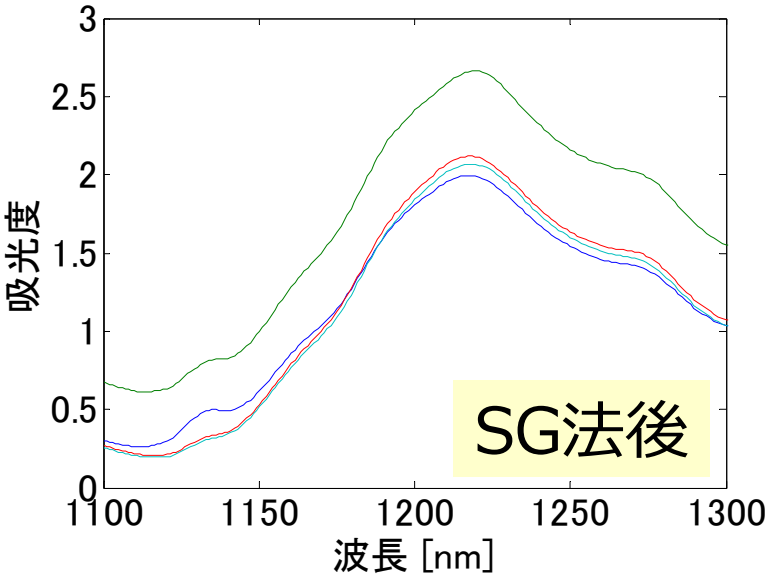
✓スペクトル解析の分野における前処理の方法として一般的

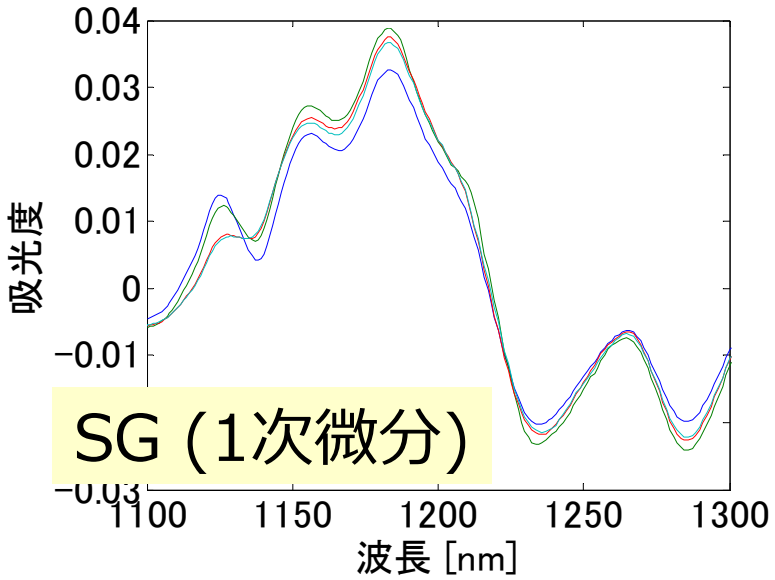
✓時系列データに用いられる例はあまりないが、効果は確認済み [3,4]

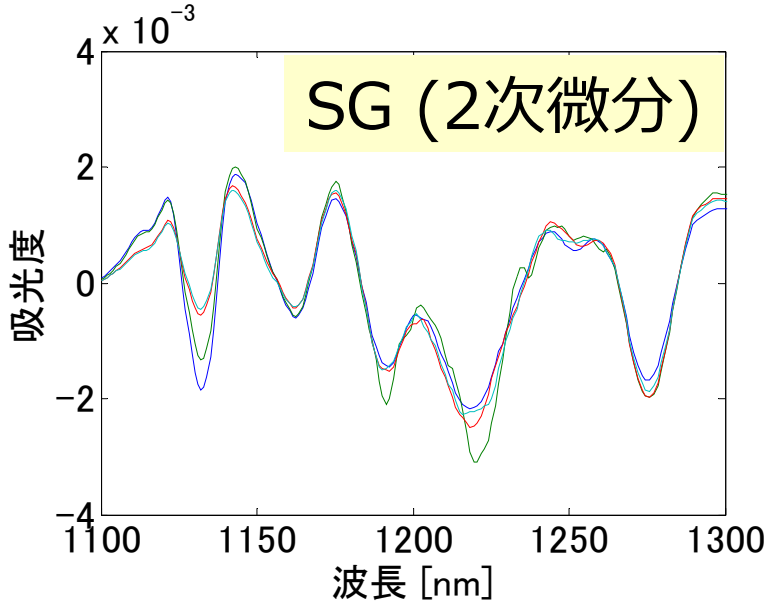
[1] A. Savitzky, M.J.E. Golay, Anal. Chem. 36, 1627-1639, 1964.
[2] 吉村 季織, 高柳 正夫, Journal of Computer Chemistry, Japan, 11, 149-158, 2012
[3] H. Kaneko, K. Funatsu, Ind. Eng. Chem. Res., 54, 12630-12638, 2015.
[4] H. Kaneko, K. Funatsu, J. Chem. Eng. Jpn., 50, 422-429, 2017

SG法の例

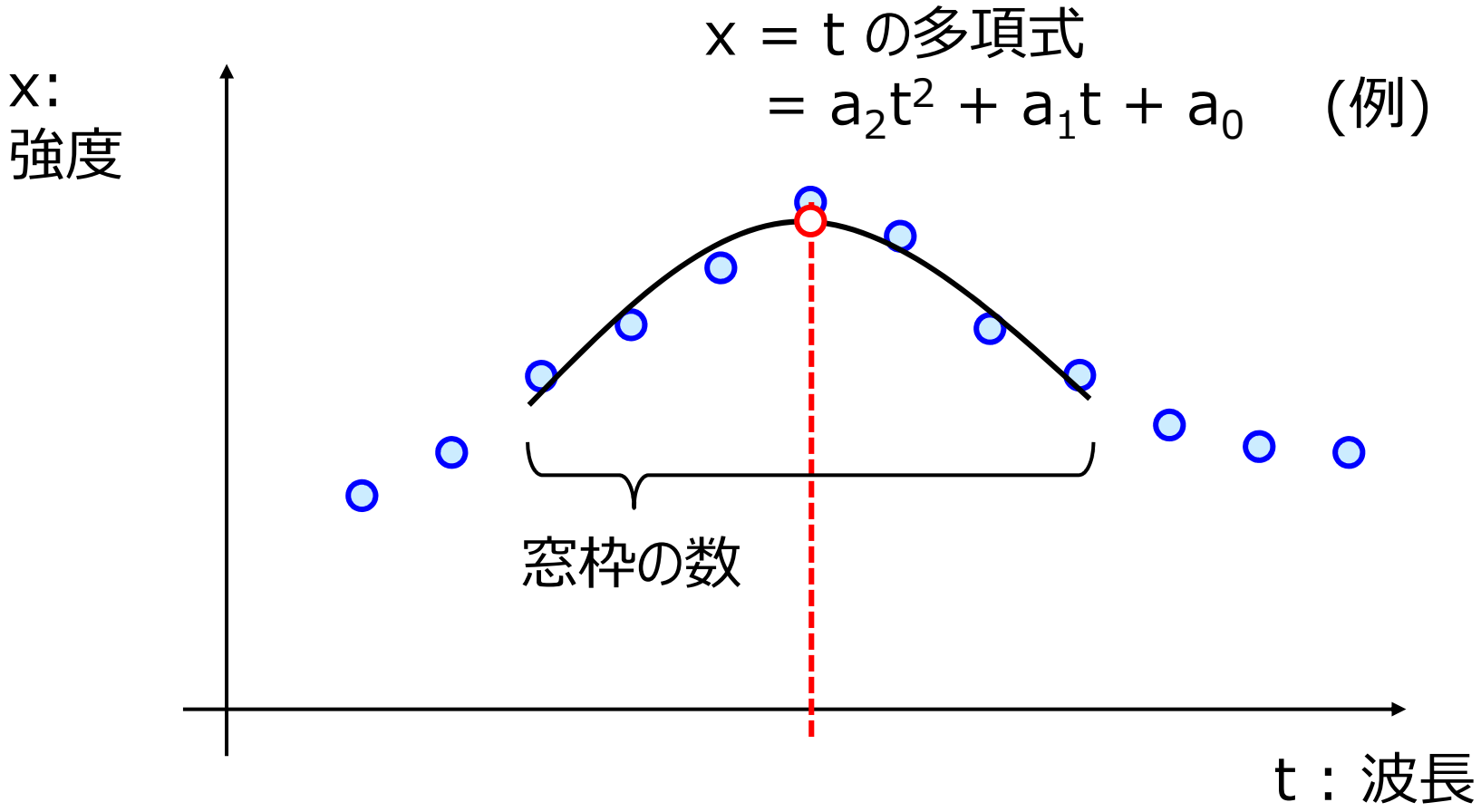

強度
波長 [nm]
元のスペクトル


吸光度
波長 [nm]
SG法後


吸光度
波長 [nm]
SG (1次微分)


x 10^-3
吸光度
波長 [nm]
SG (2次微分)

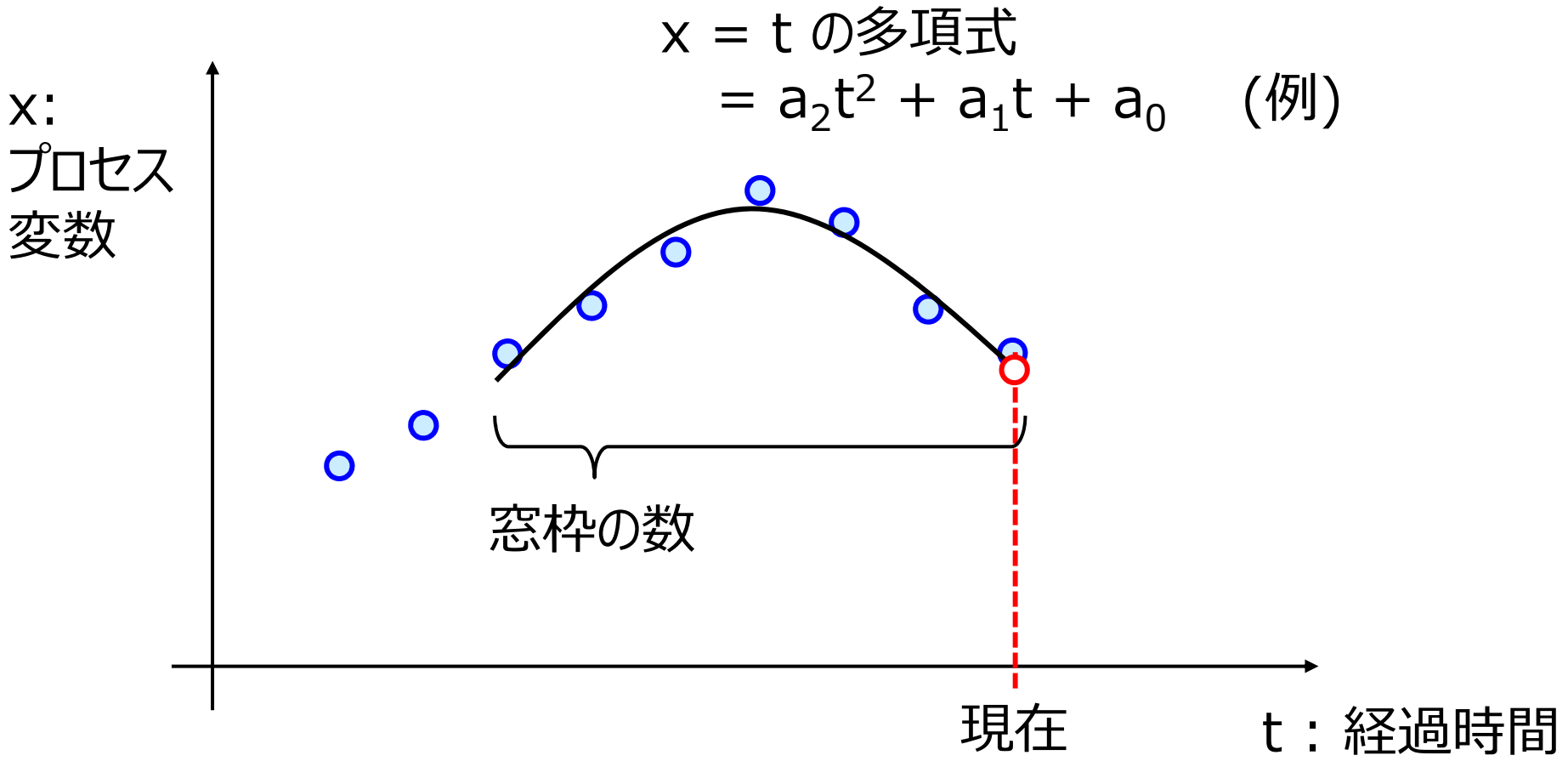
SG法 (スペクトルデータ)



- ✓ 多項式の次数
- ✓ 窓枠の数

を事前に決めなければならない

SG法 (時系列データ)



- ✓ 多項式の次数
- ✓ 窓枠の数

を事前に決めなければならない

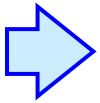
手法・ハイパーパラメータ・微分次数はどうする？[16]

✓４つの手法とハイパーパラメータの値の候補

- 単純移動平均：窓枠の数 (5, 11, 21, 31, …, 201)
- 線形加重移動平均：窓枠の数 (5, 11, 21, 31, …, 201)
- 指数加重移動平均：平滑化係数 (0.01, 0.02, …, 1)
- SG法：　多項式の次数 (1, 2, 3, 4)
  窓枠の数 (5, 11, 21, 31, …, 201)

✓微分次数 (場合によってはその組み合わせ)

をどのように決めるか？



① モデルの検証により選択する
② ノイズの正規分布性により選択する

① モデルの検証による選択

✓各手法・各ハイパーパラメータの値・各微分係数の値で、
回帰分析・クラス分類のモデルの検証を行い、
最も検証結果のよい組み合わせを選択する

- たとえば、
  - クロスバリデーション推定値の r^2 が最も大きい組み合わせ
  - バリデーションデータの r^2 が最も大きい組み合わせ

- モデルの検証：http://datachemeng.com/modelvalidation/

① モデルの検証による選択 特徴

✓メリット

- モデルの検証の仕方によっては、推定性能の高いモデルを構築できる手法・ハイパーパラメータの値・微分係数 を選択可能

✓デメリット

- 教師ありデータが必要
- モデリングを何回も行わなくてはならない (時間がかかる)

② ノイズの正規分布性による選択

- ✓平滑化前後の値を引くことで、平滑化によって “均(なら)された” ノイズの値を計算できる
- ✓ノイズは正規分布であると仮定すると、平滑化によって減少したノイズの分布も正規分布に従う必要がある
- ✓コルモゴロフ–スミルノフ検定などの正規分布性の検定により、ノイズが正規分布に従う手法・ハイパーパラメータの組み合わせを選択
- ✓選択された手法・ハイパーパラメータの組の中で、標準偏差が最も大きい (= ノイズが最も減少した) 組を選択
- ✓詳しくは下の論文を参照のこと

H. Kaneko, K. Funatsu, J. Chem. Eng. Jpn., 50, 422-429, 2017

② ノイズの正規分布性による選択 特徴

✓メリット

- 教師データ不要
- モデリング不要 (時間がかからない)

✓デメリット

- 微分次数は選択できない
- 選択の際、モデルの推定性能は考慮されていない