Variable selection algorithms for spectral data and their applications on quality evaluation of agricultural products

Mizuki TSUTA

Food Research Institute,
National Agriculture and Food Research Organization

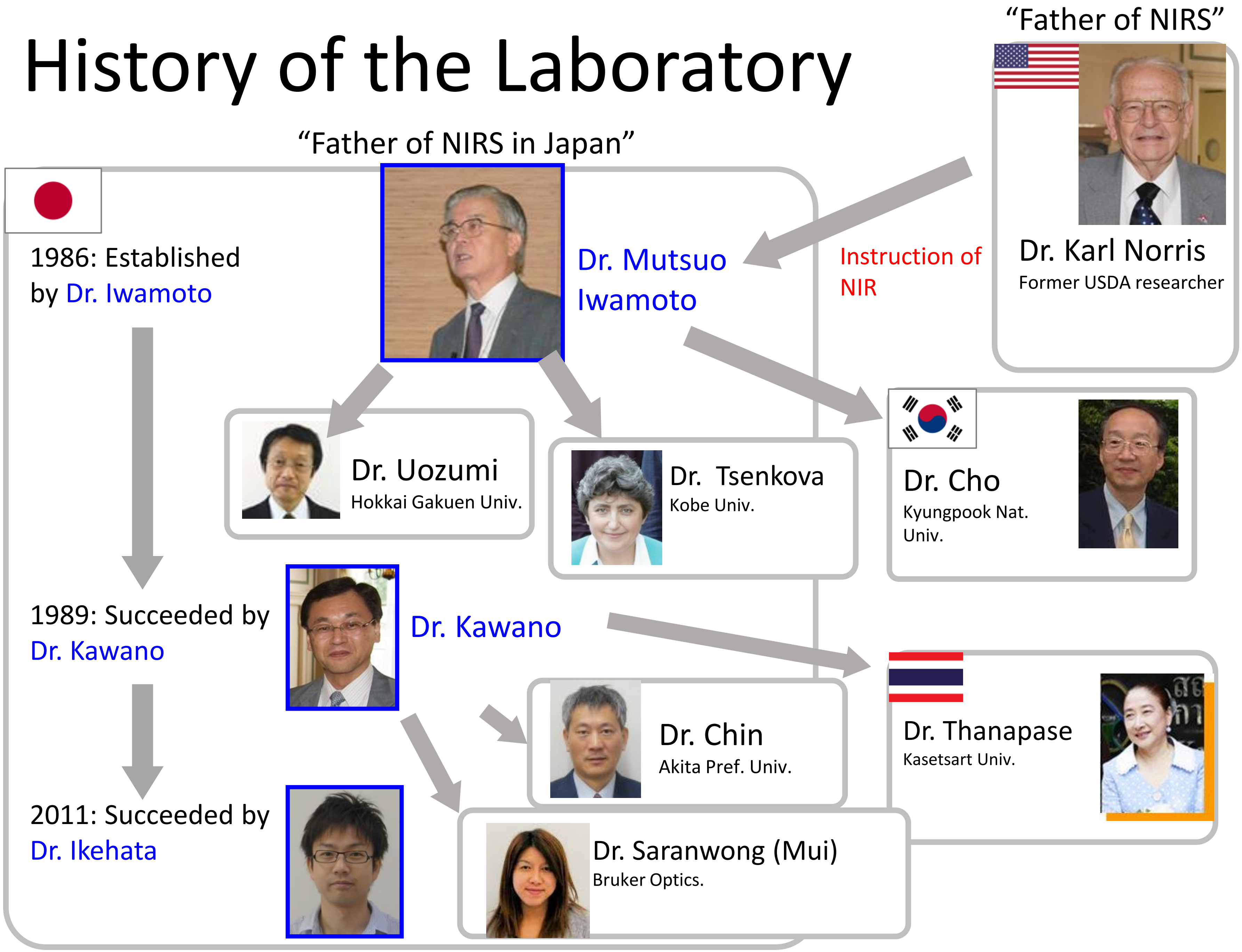
Agenda

- A bit of self-introduction
- Variable selection in spectral data analysis
  – why, how and problem –
- New variable selection algorithms
  - Stepwise selectivity ratio
  - Band-pass filter optimization
- Summary

A bit of self-introduction

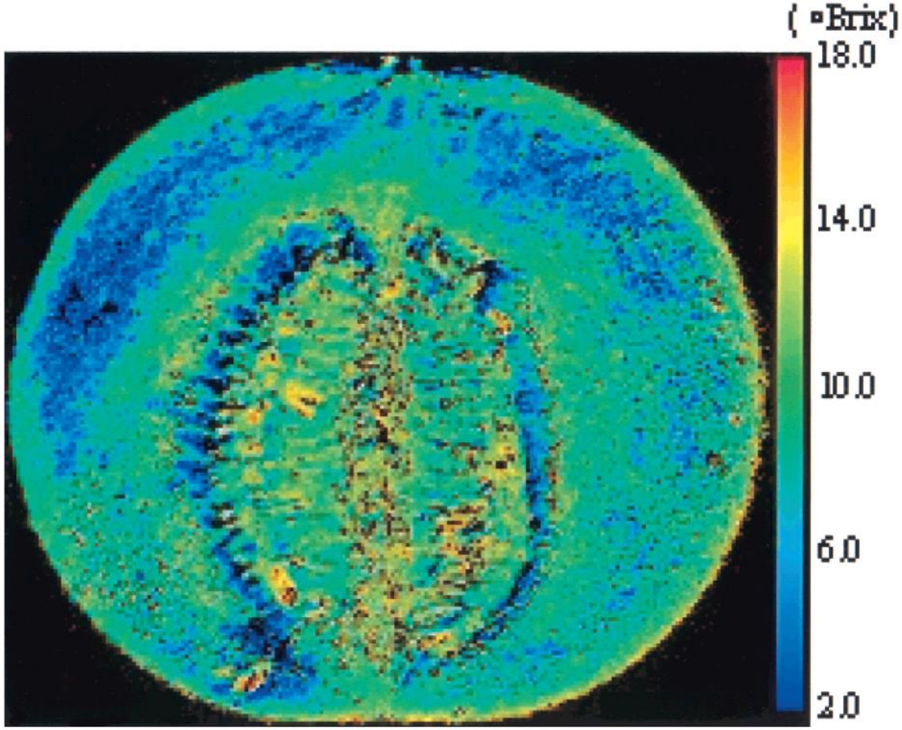
My laboratory

- Non-destructive evaluation unit (非破壊計測Unit)
- Near-infrared spectroscopy (NIRS), fluorescence fingerprint (aka excitation-emission matrix), spectral imaging, chemometrics…
- One of the most important laboratory for NIRS in Japan


History of the Laboratory
"Father of NIRS in Japan"
"Father of NIRS"
1986: Established
by Dr. Iwamoto
Dr. Mutsuo
Iwamoto
Instruction of
NIR
Dr. Karl Norris
Former USDA researcher
Dr. Uozumi
Hokkai Gakuen Univ.
Dr. Tsenkova
Kobe Univ.
Dr. Cho
Kyungpook Nat.
Univ.
1989: Succeeded by
Dr. Kawano
Dr. Kawano
Dr. Chin
Akita Pref. Univ.
Dr. Thanapase
Kasetsart Univ.
2011: Succeeded by
Dr. Ikehata
Dr. Saranwong (Mui)
Bruker Optics.

My research starting from imaging…

Melon



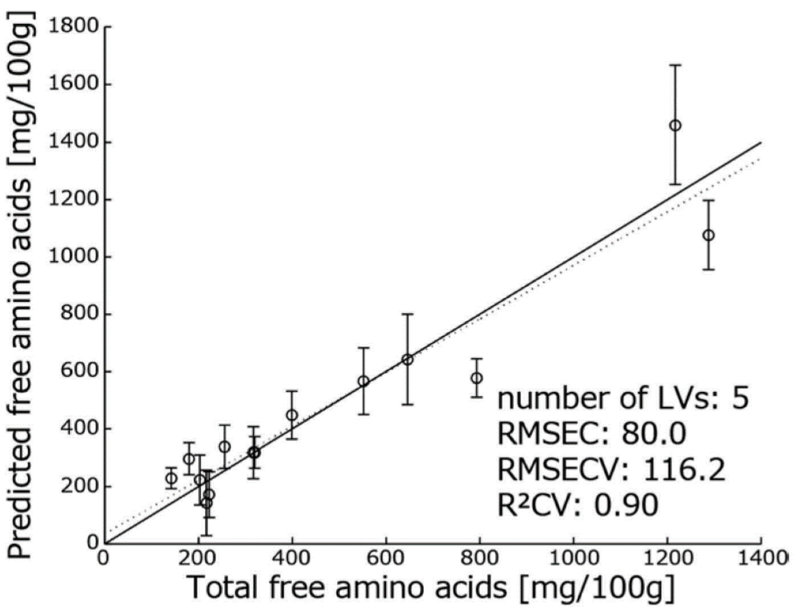
Soybean

Tsuta, M., et al. (2002). *Agricultural and Food Chemistry*, 50(1), 48-52.

Tsuta, M., et al. (2007). *Transactions of the ASABE*, 50(6), 2127-2136.

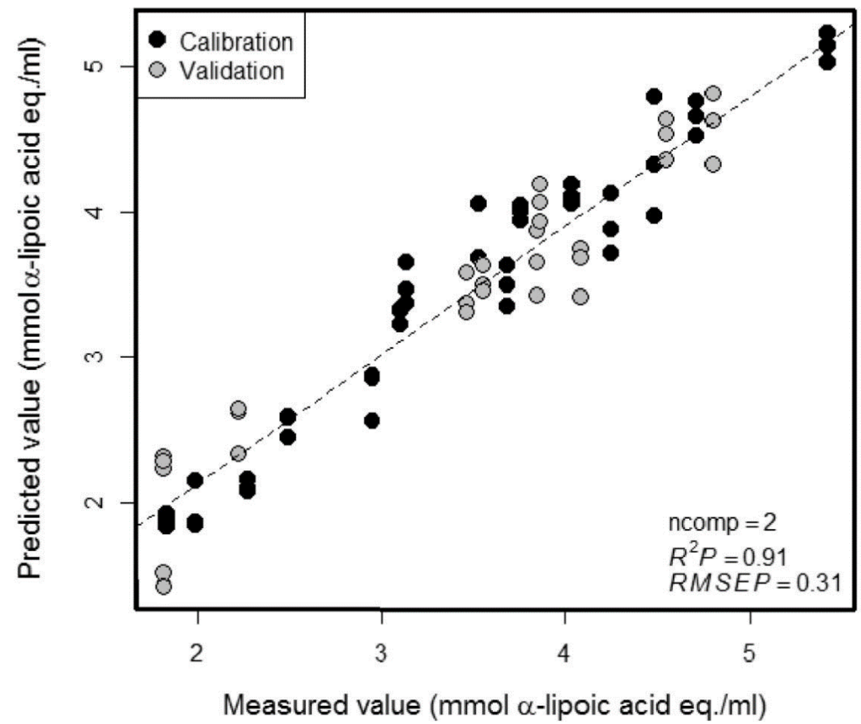
Fluorescence fingerprint...

Cheese



Kokawa, M., et al. (2015). *Food Science and Technology Research*, 21(4), 549-555.

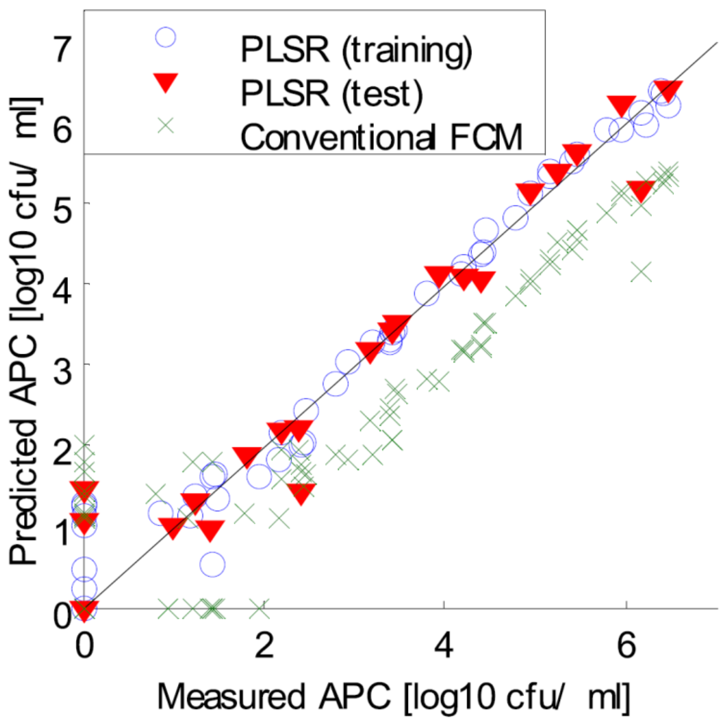
Peach juice



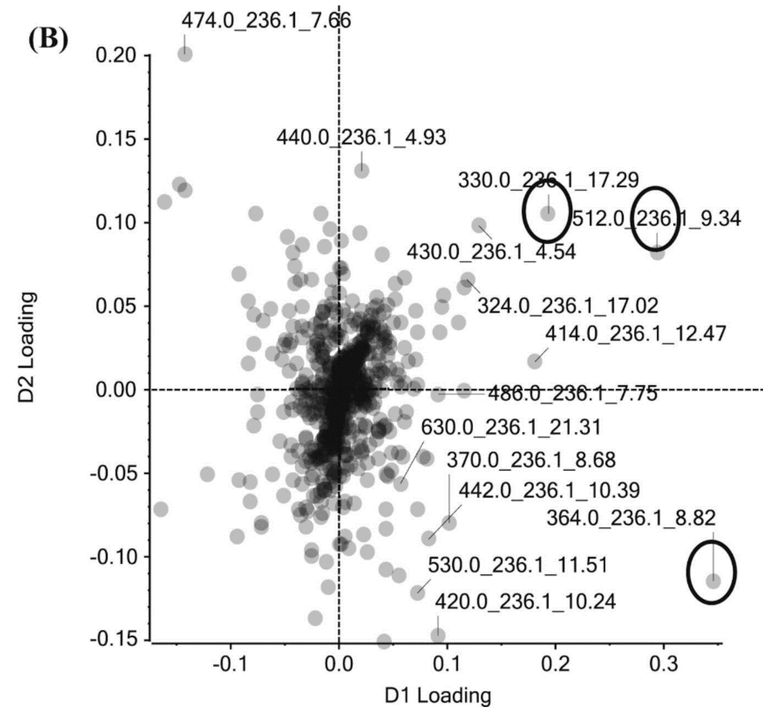
Trivittayasil, V., et al. (2017). *Food Chemistry*, 232, 523-530.

And chemomtrcis/ machine learning

Flow cytometer



LC/MS/MS



Tsuta, M., et al. (2014). *LWT-Food Science and Technology*, 55(2), 472-476.

Syukri, D., et al. (2018). *Food chemistry*, 269, 588-594.

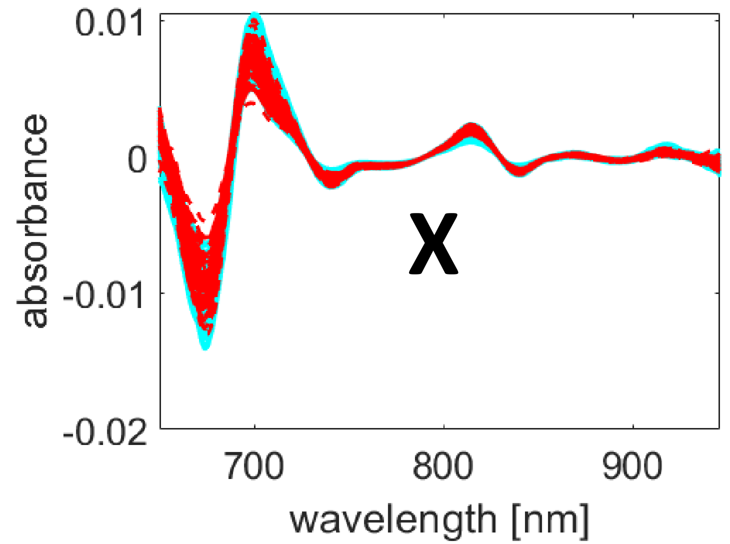
Variable selection

- why, how and problem -

Prediction model in spectroscopy

- Protein
- Brix
- Geographic origin
- …

y



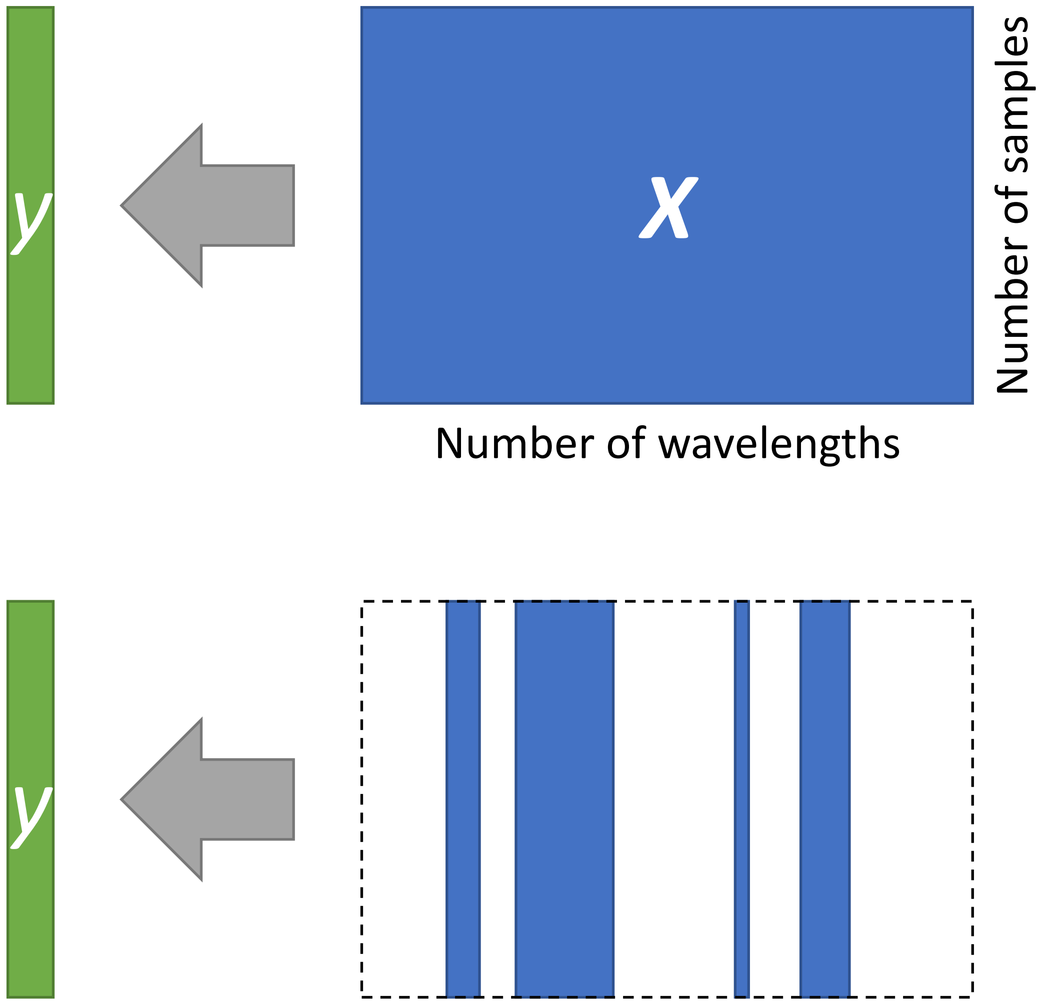
$$\hat{y} = f(\boldsymbol{X})$$

Objective variable

Explanatory variable

Prediction model

What is variable selection (VS) ?



Purposes of VS

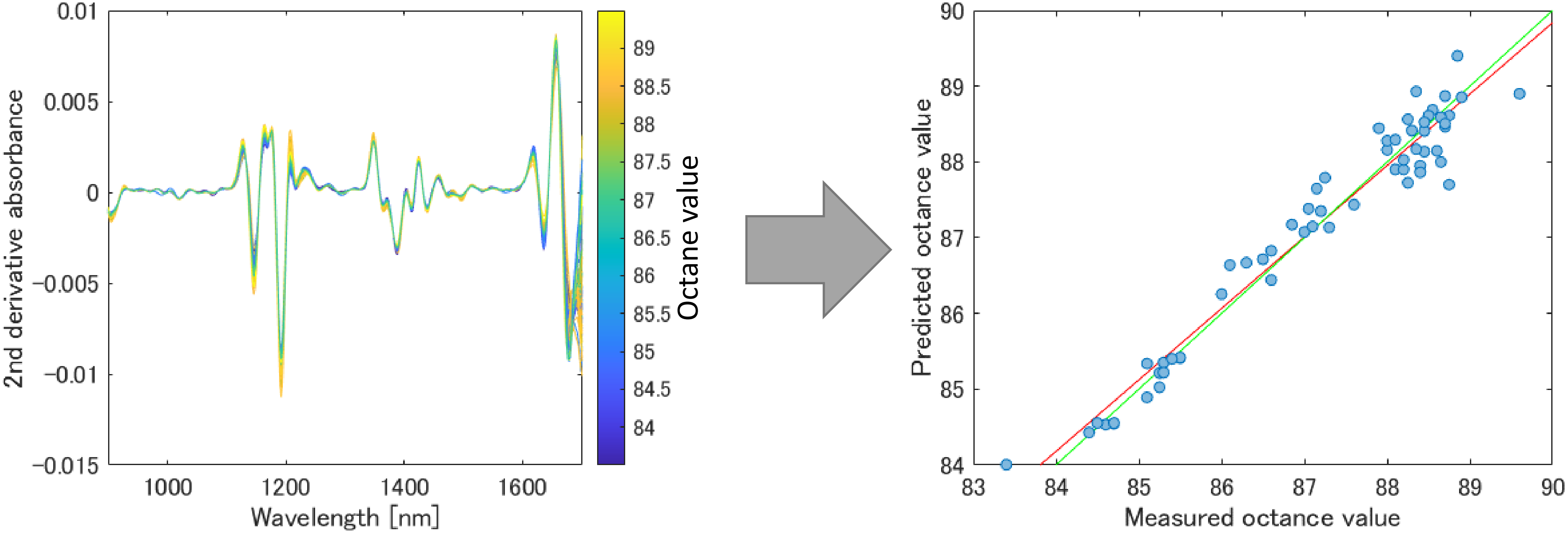
- Improvement of the model prediction
  - Removal of irrelevant, noisy or unreliable variables
- Better model interpretation
  - Focusing on variables contribute largely to the model
- Lower measurement costs
  - Shorter measurement time
  - Simpler, cheaper instruments

(Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, 24(11-12), 728-737.)

VS methods for partial least square (PLS) model

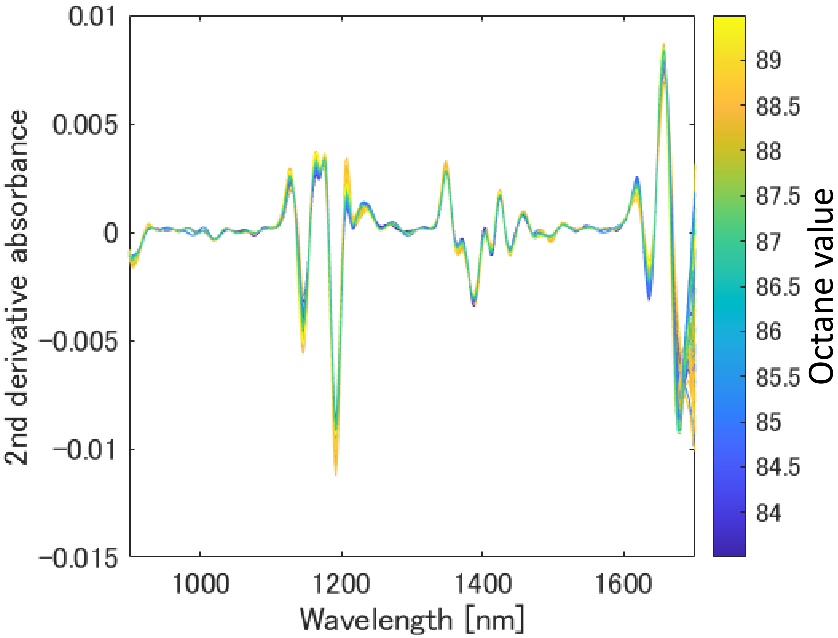
- Variable importance in projection (VIP)
- Selectivity ratio (SR)
- Interval PLS (iPLS)
- Genetic algorithms (GA)
- …

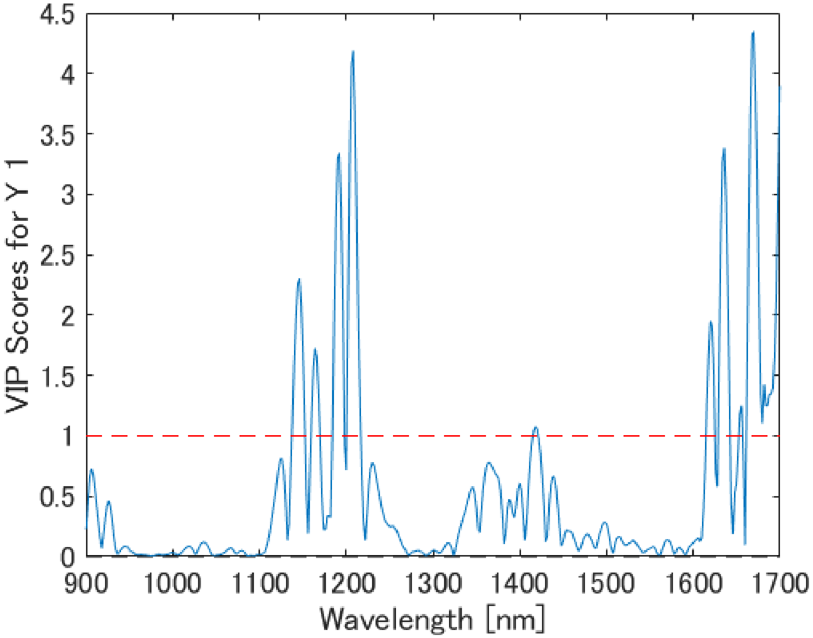
A gasoline NIR spectra case



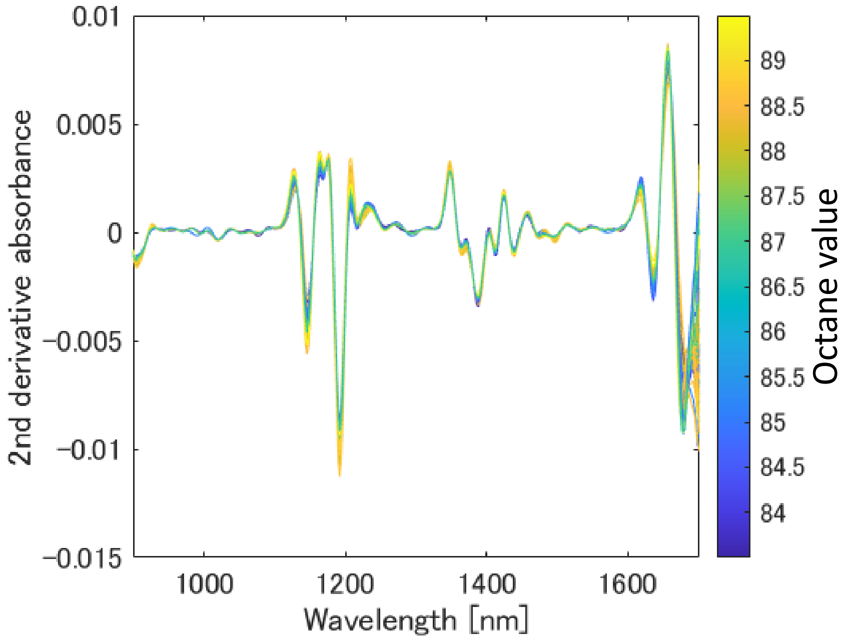
Kalivas, J. H. (1997). *Chemometrics and Intelligent Laboratory Systems*, *37*(2), 255-259.

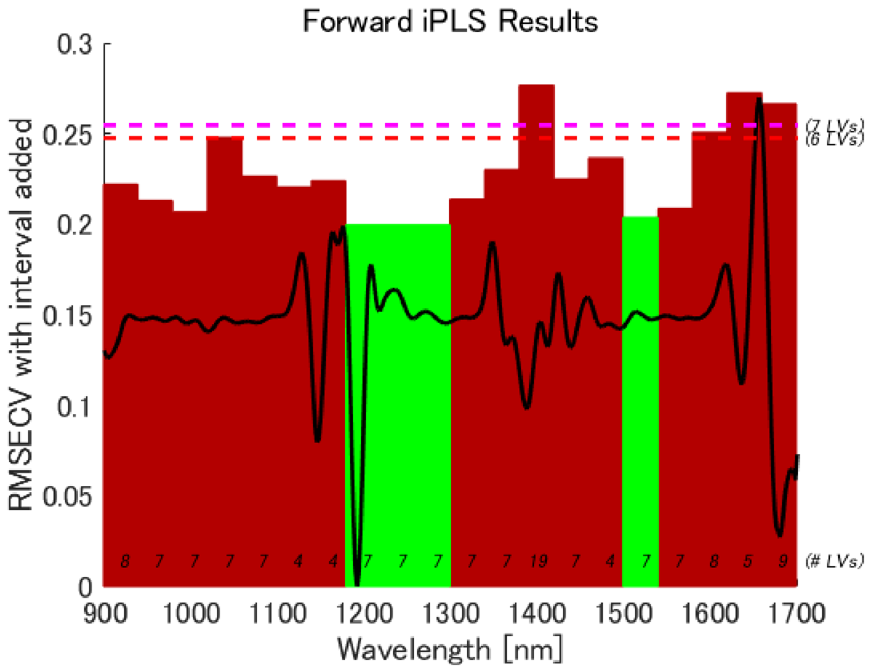
A gasoline NIR spectra case: VIP



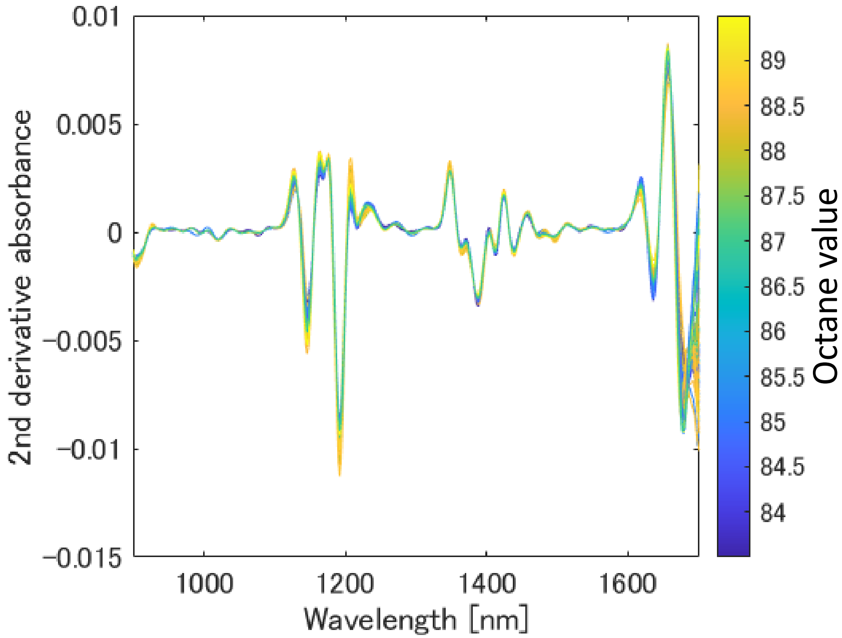


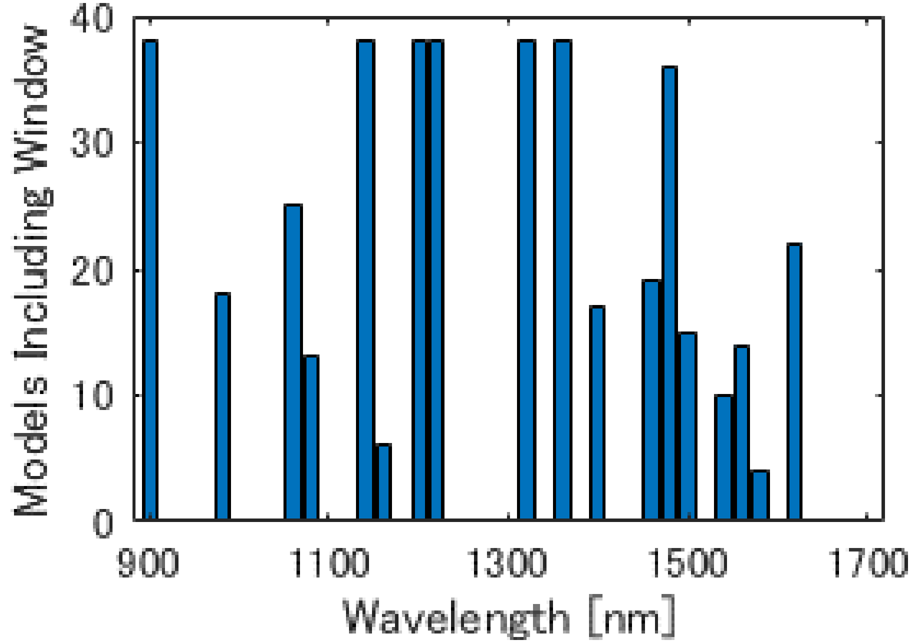
A gasoline NIR spectra case: iPLS





A gasoline NIR spectra case: GA





Problem: hyperparameters

- VIP and SR
  - Threshold (VIP=1 in many cases, but why? As for SR?)
- iPLS
  - Interval size (width in nm)
  - Number of interval to be used in the model
- GA
  - Genome size (width in nm)
  - Number of population (models)
  - Number of generations

- Arbitrary and unstable results
- Trial and errors

New VS algorithm 1
- stepwise selectivity ratio -

Objective

- VS with NO hyperparameter
  - No trial and errors
  - Always same result
- Candidate algorithm for modification
  - VIP or SR
    - They have only one hyperparameter (threshold)
  - SR has been reported* to yield less false positives

*Rajalahti, T. et al. (2009). *Analytical Chemistry*, 81(7), 2581-2590./ Farrés, M. et al. (2015). *Journal of Chemometrics*, 29(10), 528-536.

Selectivity Ratio (SR)

- Proposed by Rajalahti et al.[*] for biomarker discovery from mass spectra data
- "The ratio between explained and residual variance of the spectral variables on the target-projected component"
- The higher the SR value, the more important variable

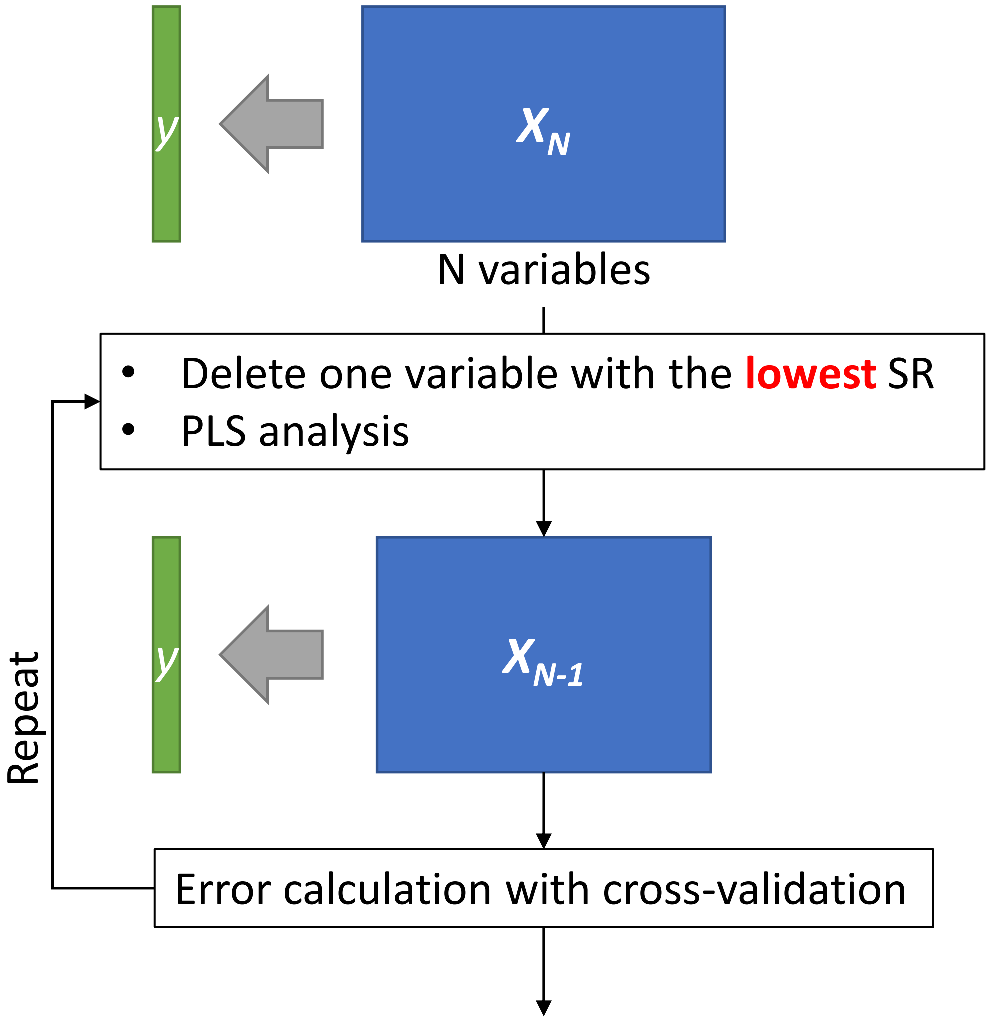
$$\mathrm{SR}_i = v_{\mathrm{expl},i}/v_{\mathrm{res},i} \qquad i = 1, 2, 3, \ldots$$

*Rajalahti, T., et al. (2009). *Chemometrics and Intelligent Laboratory Systems*, 95(1), 35-48.

Selection criteria w/o threshold

- Highest or lowest SR value as a criterion
- Only one variable chosen with the highest SR value
- One variable excluded with the lowest SR value
- What if we **repeat** the variable excluding procedure?

Stepwise SR: procedure



Choose the number of variables with the **lowest** error

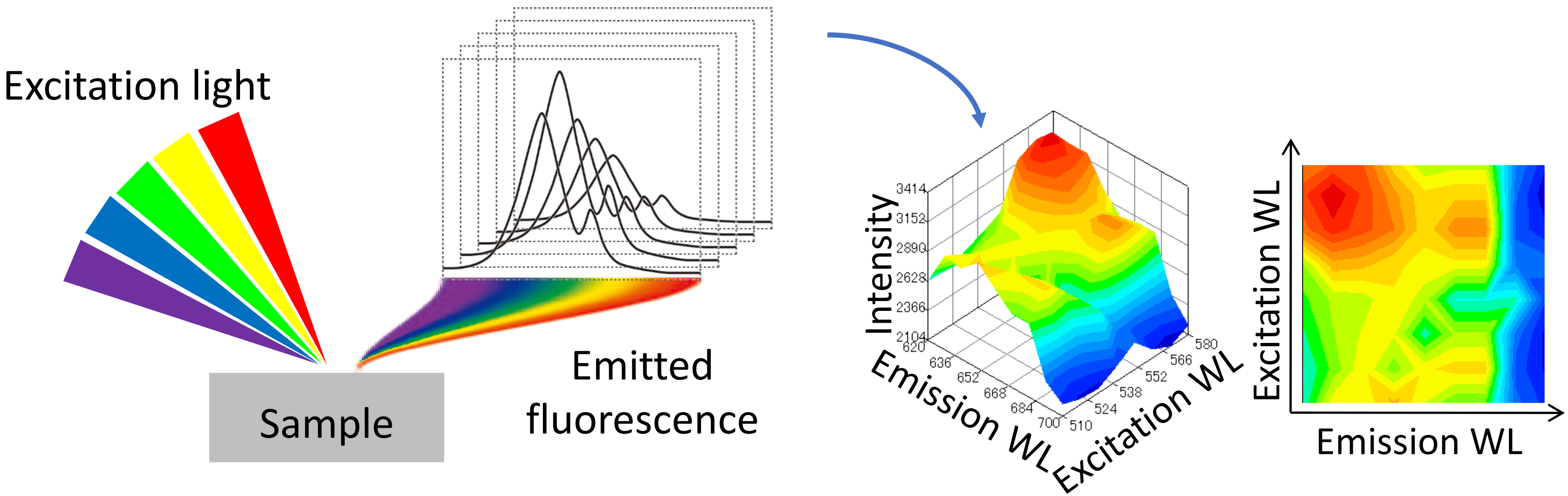
A case study: apple fluorescence fingerprint

- 1-methylcyclopropene (1-MCP)
  - Inhibitor of ethylene perception
  - Freshness preserving agent for fruits including apple
- Need for 1-MCP treatment discrimination
  - Cannot see the difference by naked eye
  - 1-MCP not approved in some apple importing countries
  - Individual fruit suitable for long storage or not
- Conventional analysis method
  - GC-FID
  - Destructive, time-consuming and laborious

Fluorescence Fingerprint (FF)

= Excitation Emission Matrix (EEM)

Set of fluorescence spectra at consecutive wavelengths (WL)

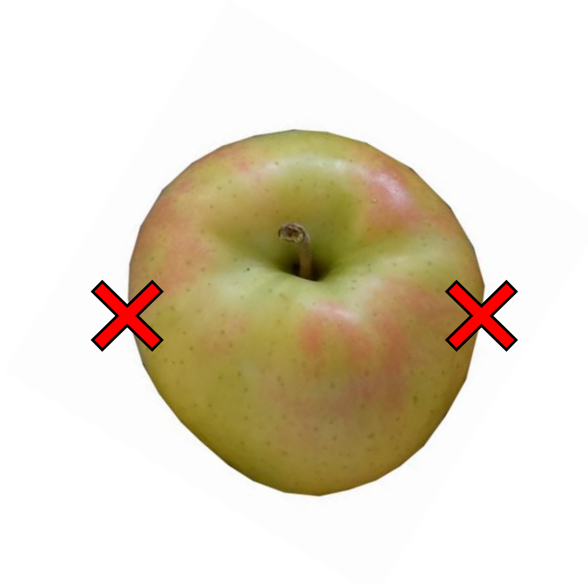


- slight differences in fluorescence characteristics is detectable
- non-destructive observation is possible

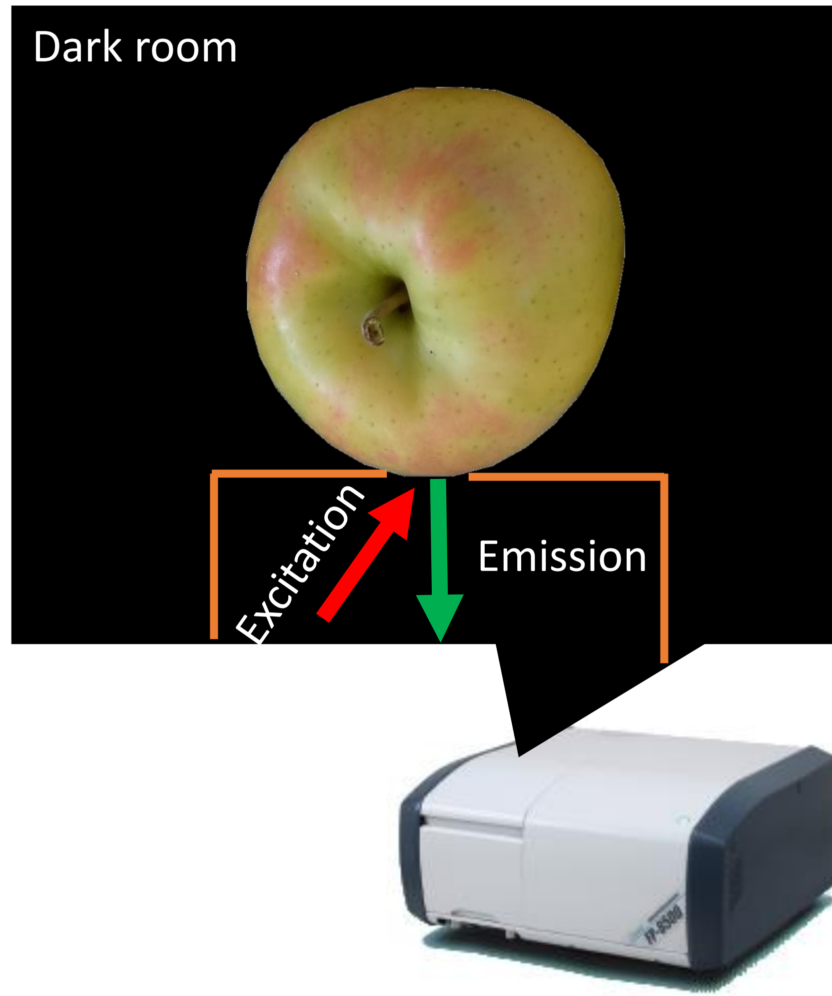
1-MCP treatment classification?

Trivittayasil, V., et al. (2018). *Chemometrics and Intelligent Laboratory Systems*, 175, 30-36.

Methods

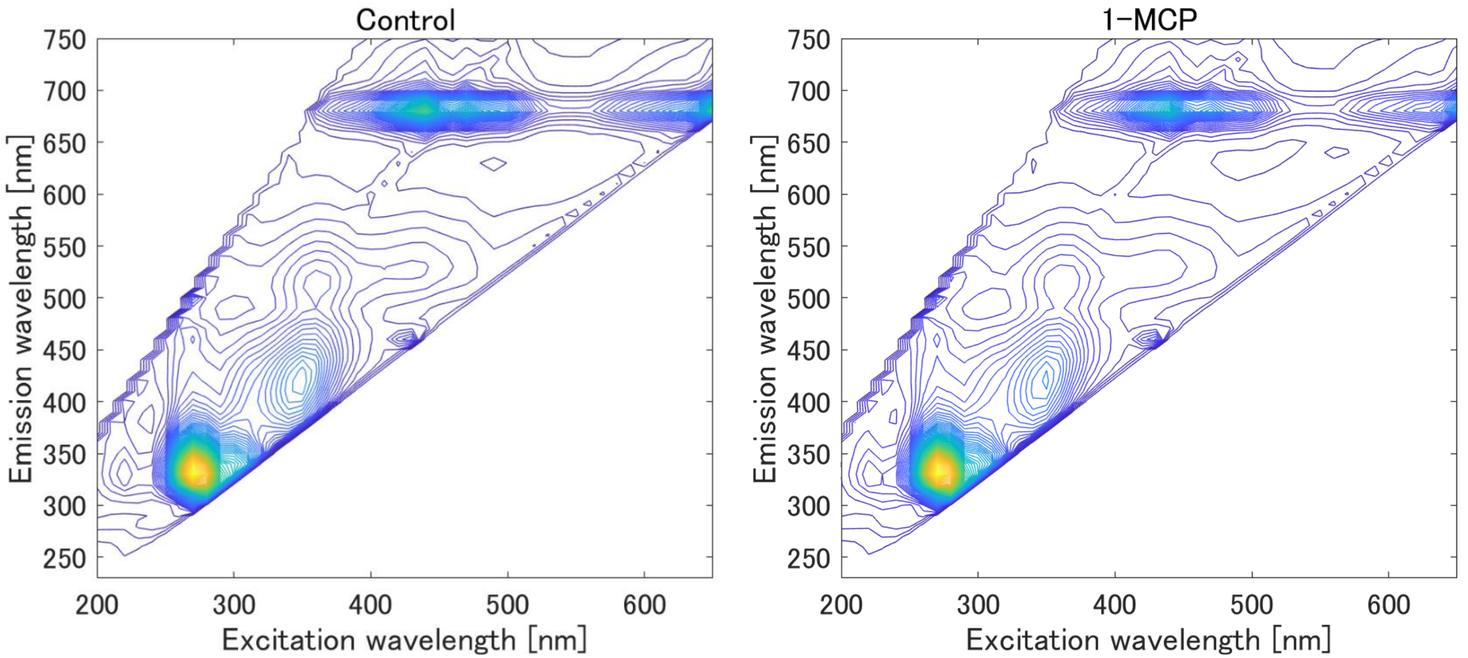


- 442 Fruits
- Fuji and Orin cultivars
- Control and 1-MCP
- 2 measurement points on the equator



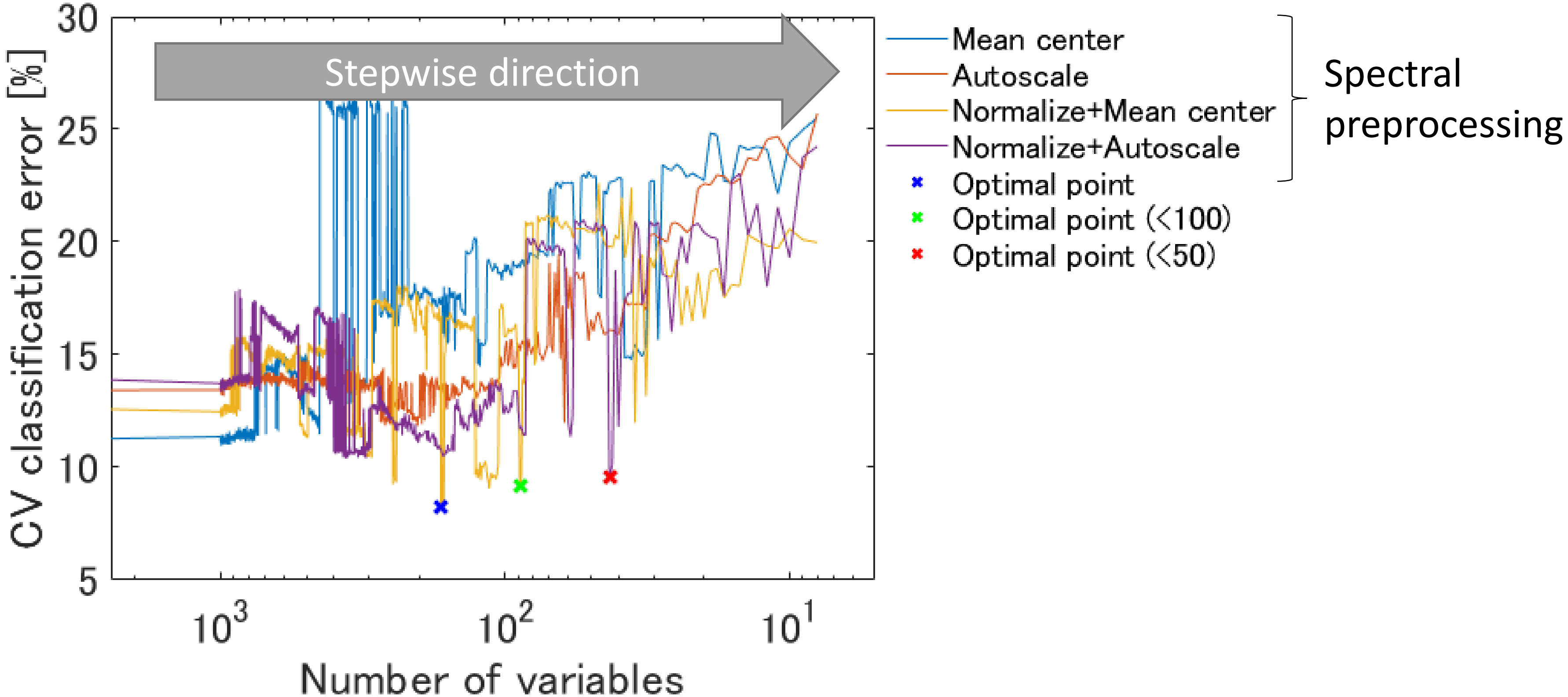
FP8500 fluorescence spectrophotometer (JASCO)
EFA-833 epi-fluorescence unit (JASCO)

Sample FF



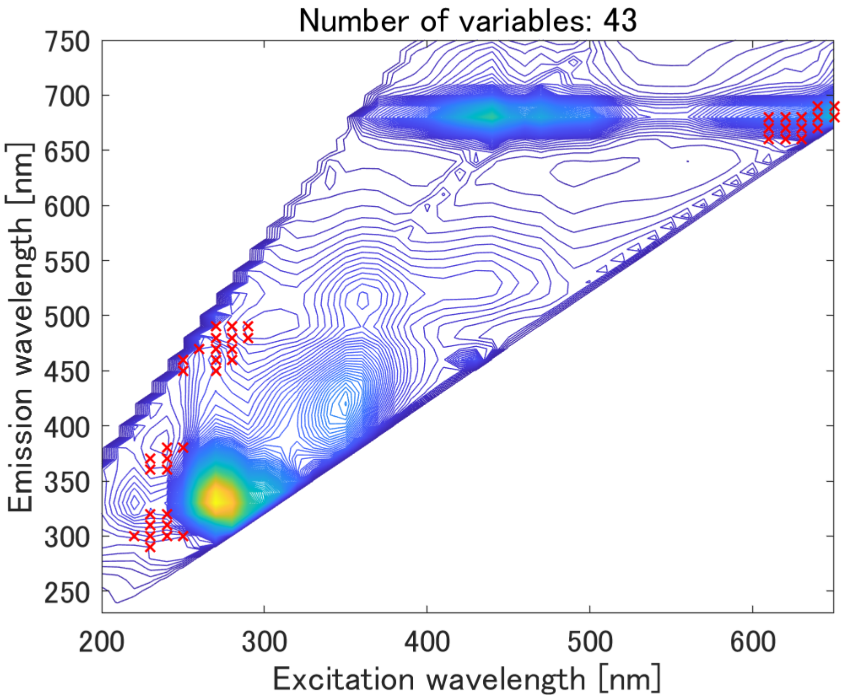
- Wavelength conditions
  - Excitation: 200-650 nm, 10 nm intervals
  - Emission: 230-750 nm, 10 nm intervals
  - Total 2438 wavelengths
- No clear difference between control and 1-MCP

Stepwise SR result



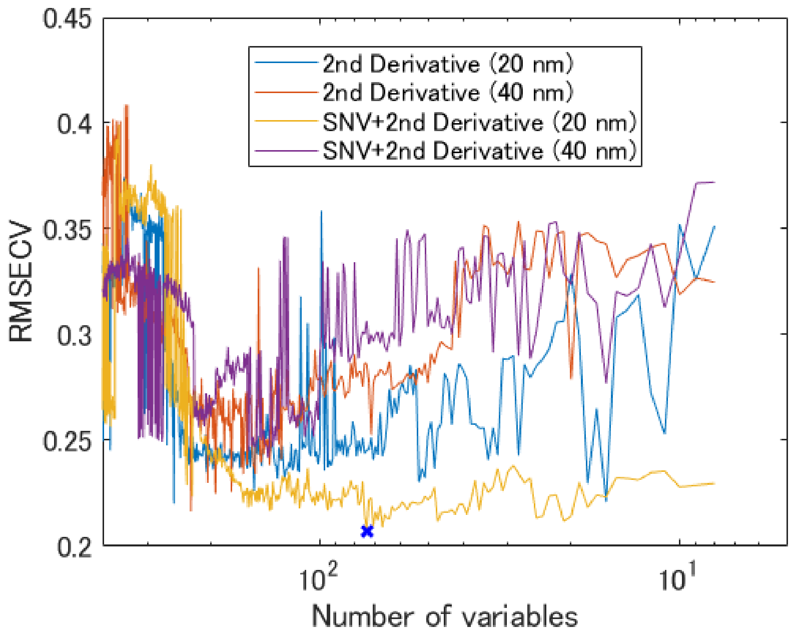
- Several points with lower CV error than the original model
- Choices according to requirements (# variables etc.)

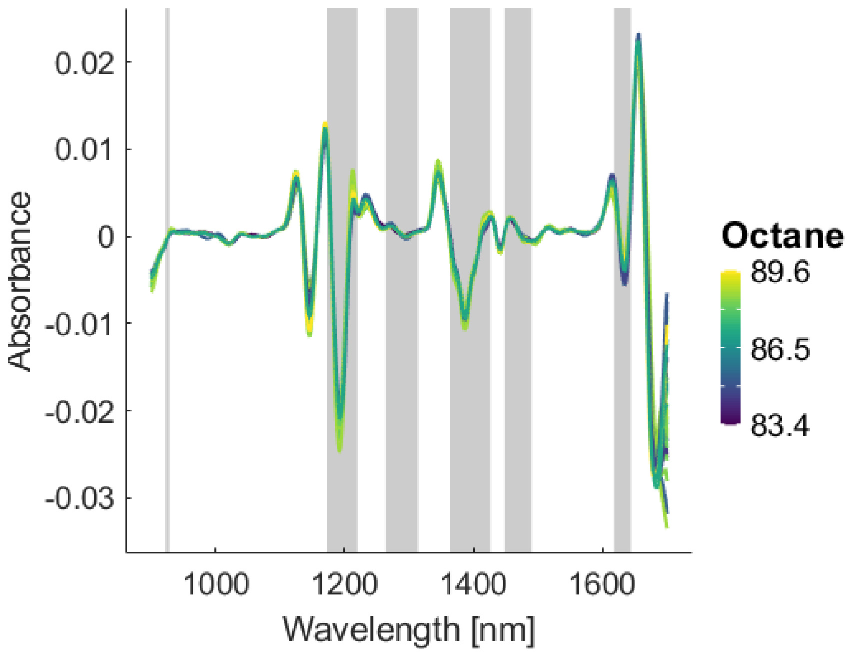
Selected wavelength conditions



- Number of variables: 2438->43
- Classification error on independent test set: 12.5%->10.1%

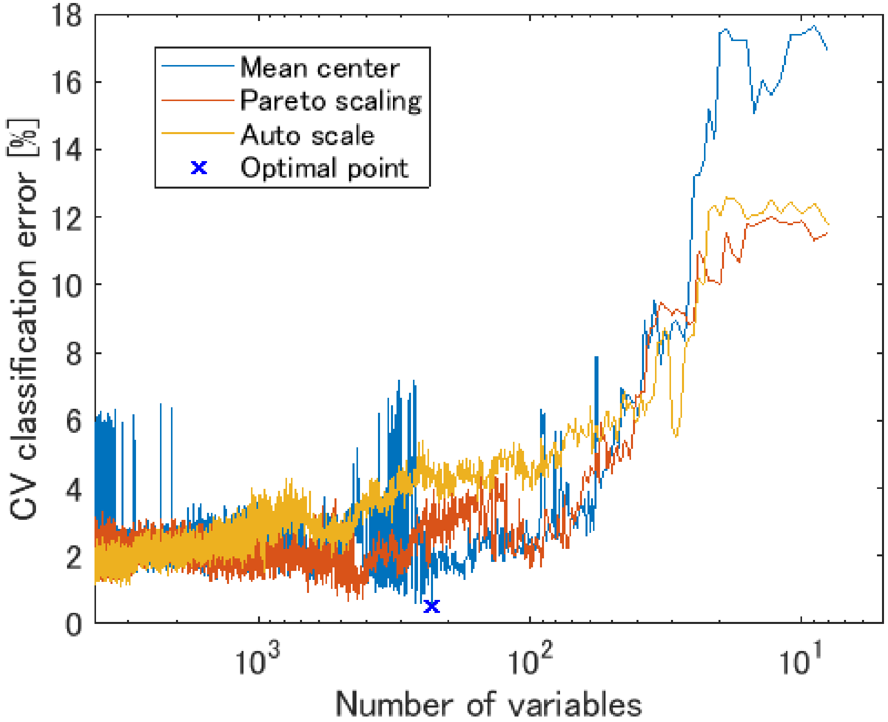
Other cases #1: gasoline NIR





- Number of variables: 401->74
- Root mean squared error of cross-validation (RMSECV): 0.264->0.207

Other cases #2: cancer proteomics



- Number of variables: 4000->230
- Classification error of cross-validation: 2.08%->0.50%

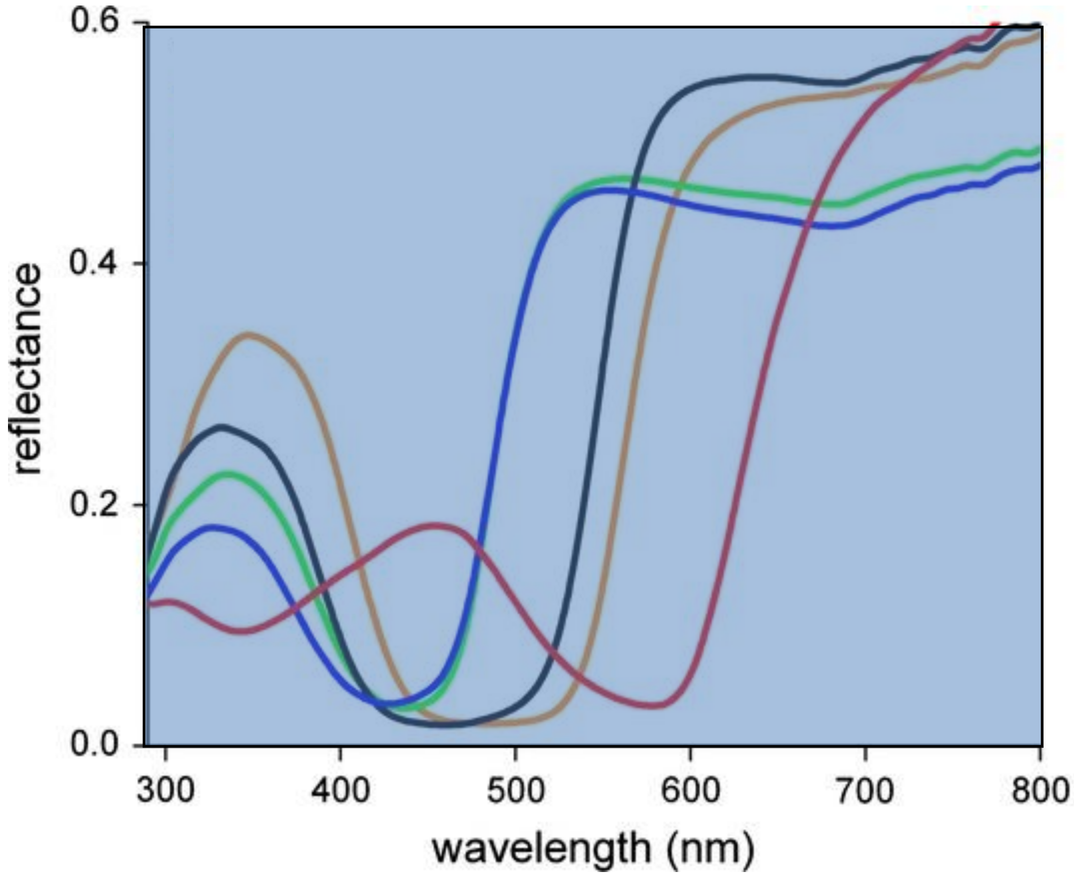
Petricoin III, E. F., et al. (2002). *The lancet*, 359(9306), 572-577.

Stepwise SR: summary

- No hyperparameter and no trial and error required
- Can be applied to spectral as well as discrete data such as –omics data
- Effective for the improvement of the model prediction power
- Model interpretation can be easier with smaller number of variables
- Remaining problem – dozens of variables are still too many for simple instruments such as band-pass filter based spectrometers

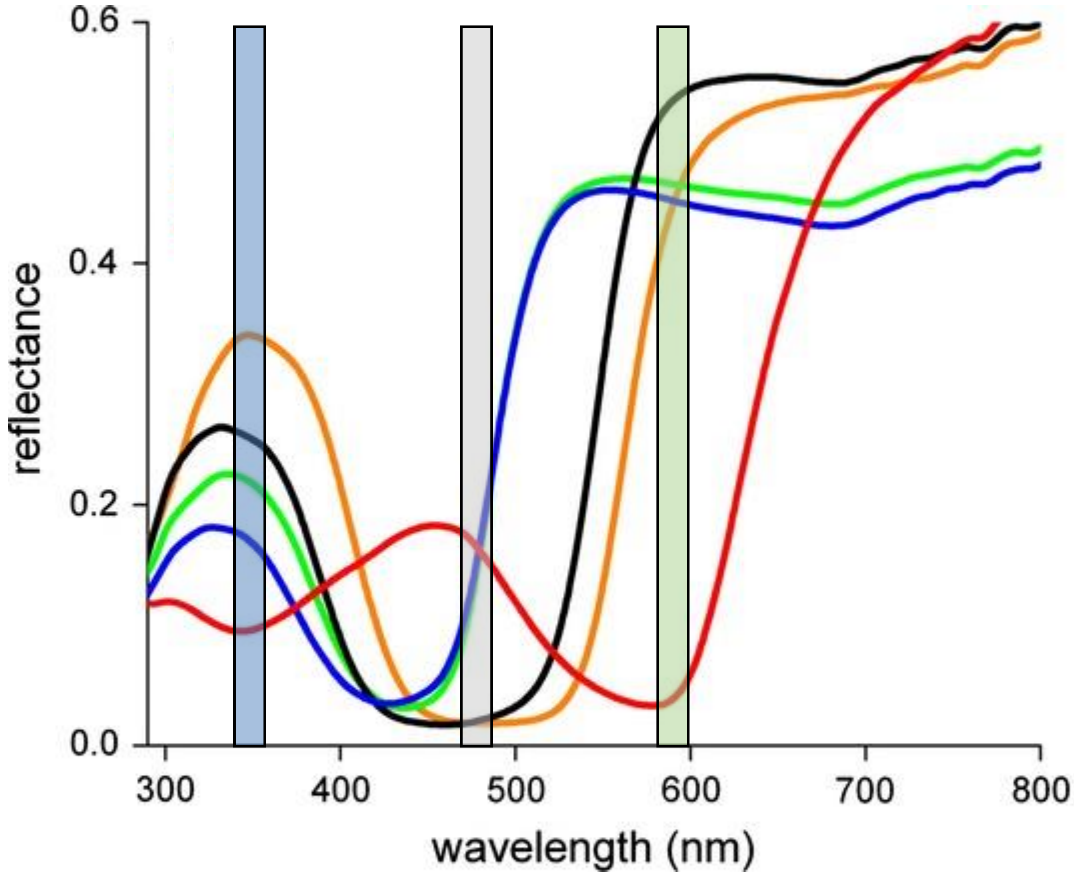
New VS algorithm 2
- band-pass filter optimization -

Model with all wavelengths (PLS etc.)



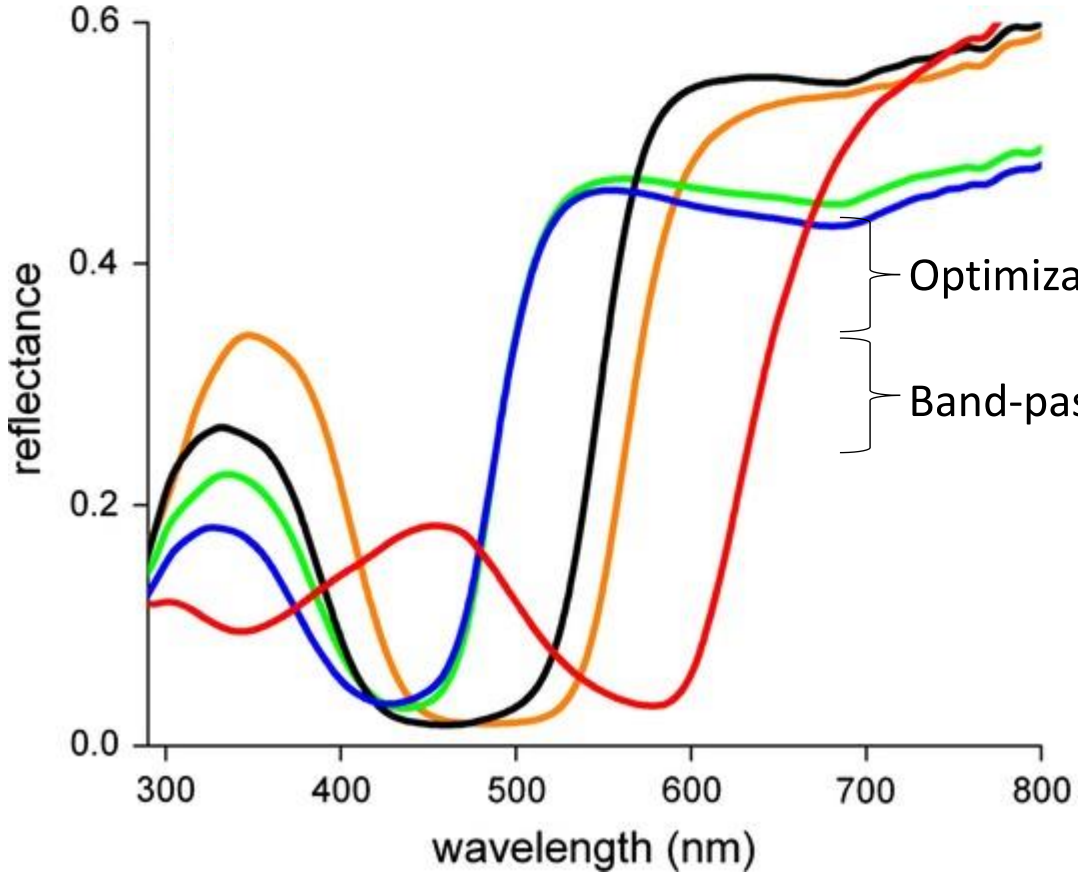
- Info: **Large**
- Accuracy：**High**
- Cost：**High**
- Imaging：**Difficult**

Model with few wavelengths (MLR etc.)



- Info: **Small**
- Accuracy：**Low**
- Cost：**Low**
- Imaging：**Easy**

The best of both worlds?



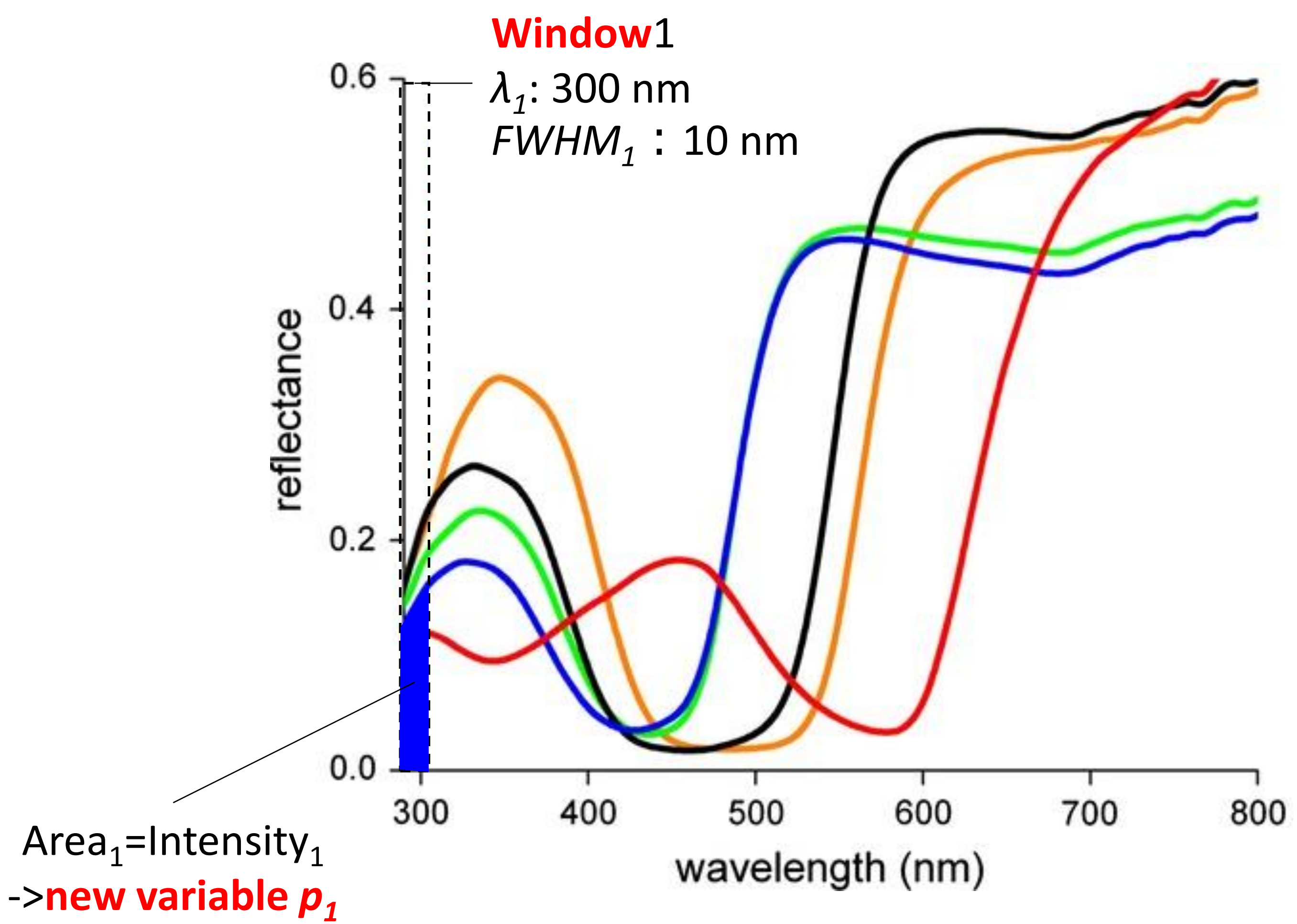
- Information: **Large**
- Accuracy: **High**
- Cost: **Low**
- Imaging: **Easy**

Optimization of characteristics

Band-pass filter (BPF)

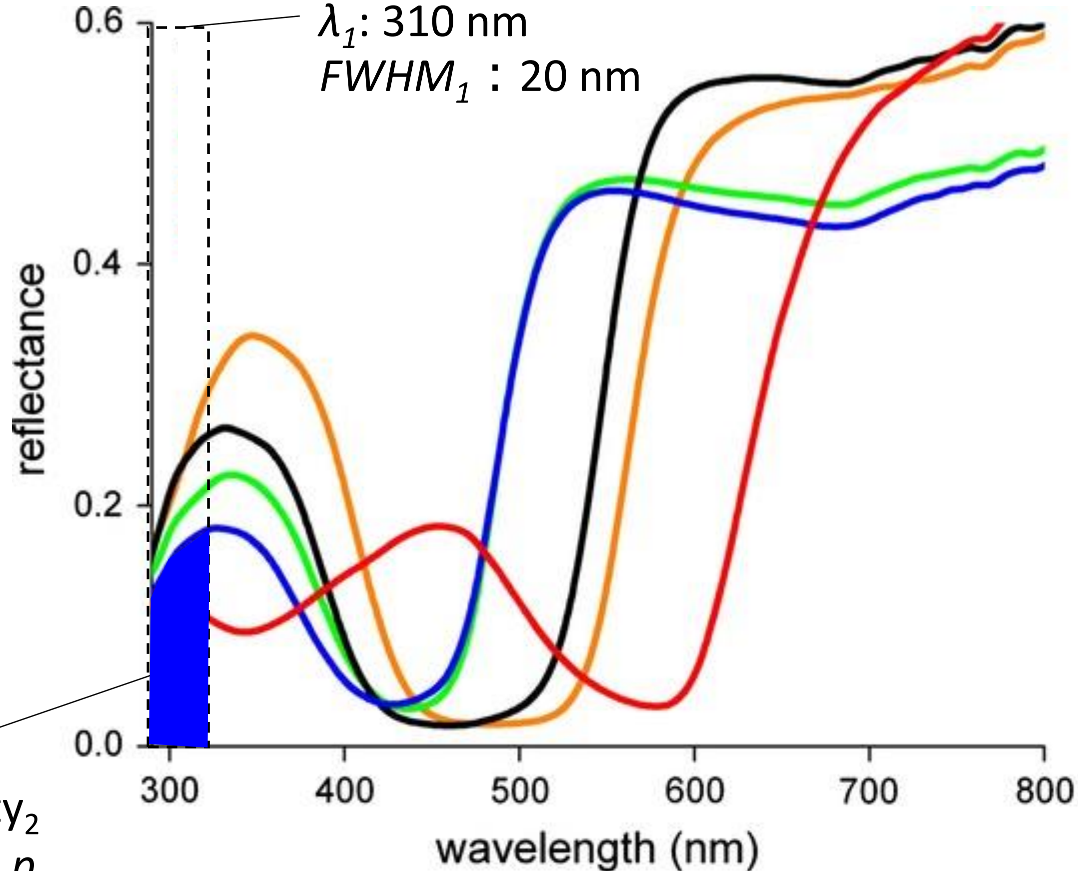
Nakauchi et al., *Optics Express* **20**, 2, 986 (2012)/ 蔦他, 日本食品科学工学会誌, **59**, 3, 139 (2012)

Step 1: Calculation of light intensity through band-pass filters



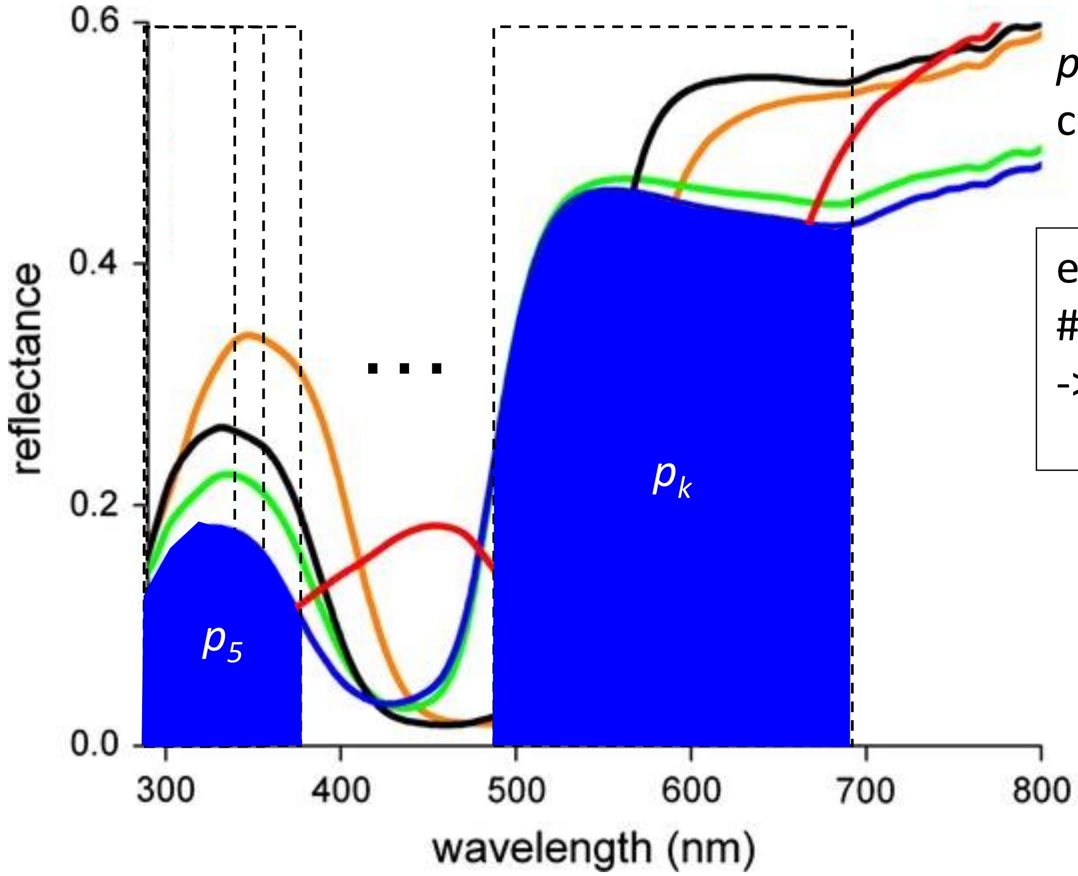
Step 1: Calculation of light intensity through band-pass filters

Window2
λ_1: 310 nm
$FWHM_1$: 20 nm



$Area_2 = Intensity_2$
->new variable p_2

Step 1: Calculation of light intensity through band-pass filters



p_k for all possible combinations

e.g.
#Wavelength=100
->#New variable=$_{100}C_2$
=**4950**

Step 2: Model construction

- Multiple linear regression (MLR) or linear discriminant analysis (LDA) model
  - *Predicted value*$=a_0+a_i \times p_i + a_j \times p_j$…
- Choose few variables from new variables (=windows) effective for regression/ classification
- Brute-force search
  - 2 windows->$_{4950}C_2$=12,248,775 combinations
  - 3 windows->$_{4950}C_3$=2^{10} combinations…

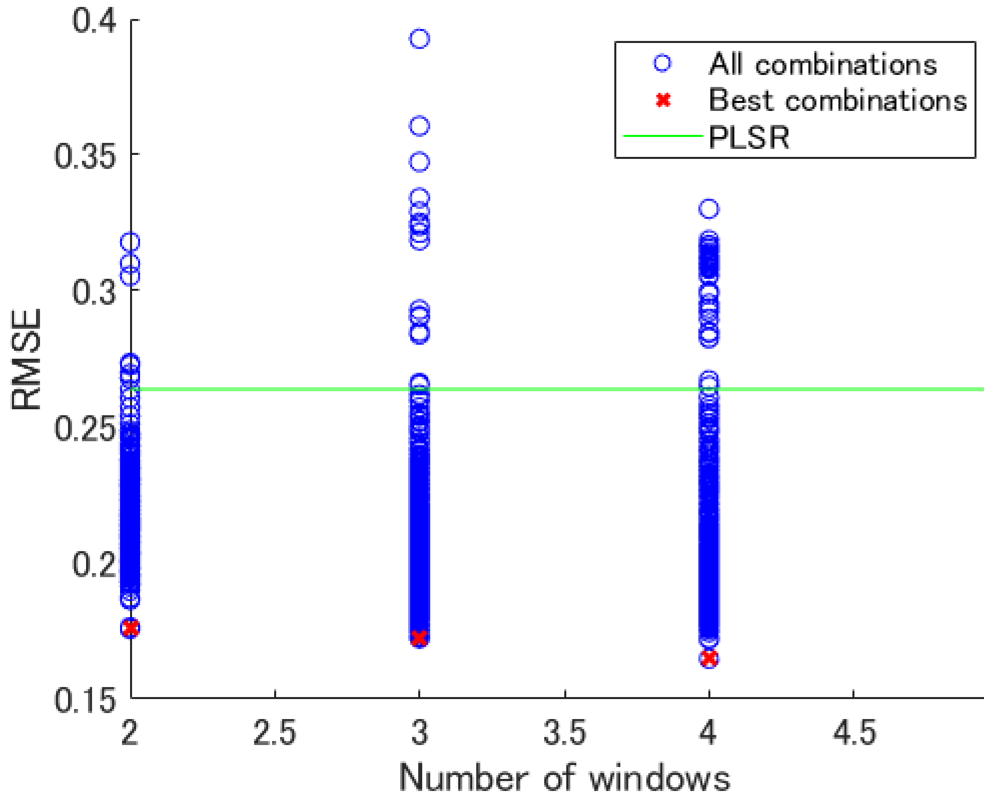
A solution: stepwise variable selection

- Create a MLR/ LDA model with one variable
- Add other variables one by one until certain criteria is satisfied
  - Criteria: F-Statistic, p-value, Akaike's Information Criterion…
- Repeat the procedure with different initial variable to cover the all possible combinations
  - e. g. 4950 new variables -> 4950 initial variables, 4950 different models
- Record λ, *FWHM* and prediction accuracy of each model

Step 3: Optimization of BPF

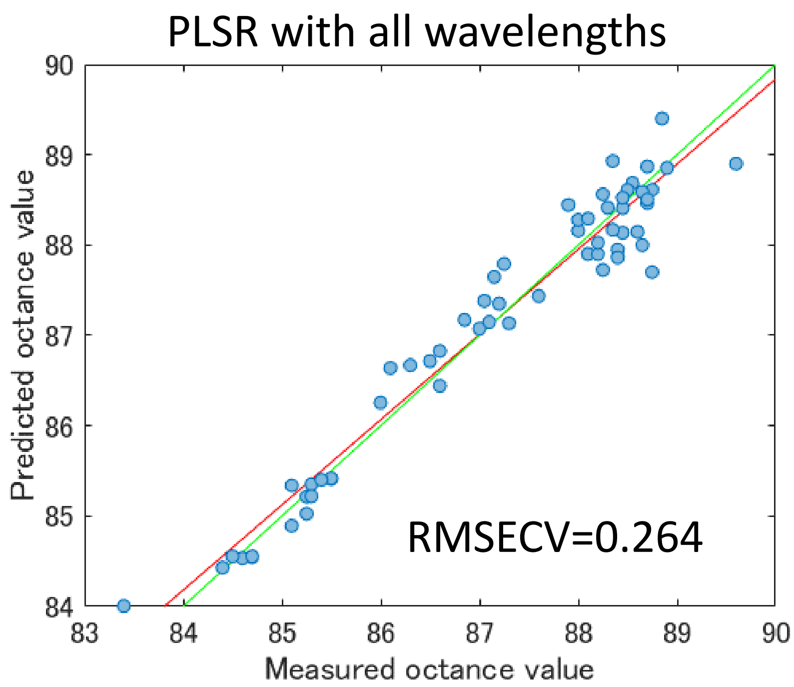
- Decide how many windows to be used in the application
- Choose the model with the highest prediction accuracy with the desired number of windows
- Trade-off between the cost and accuracy
  - Number of windows = number of BPFs
  - More BPFs -> higher accuracy, higher cost
  - Less BPFs -> lower accuracy, lower cost

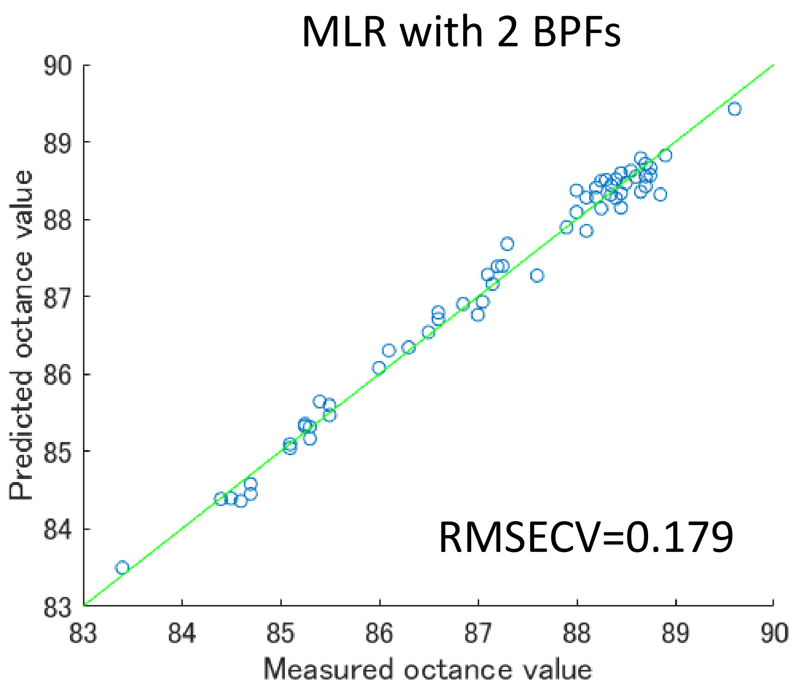
A gasoline NIR spectra case



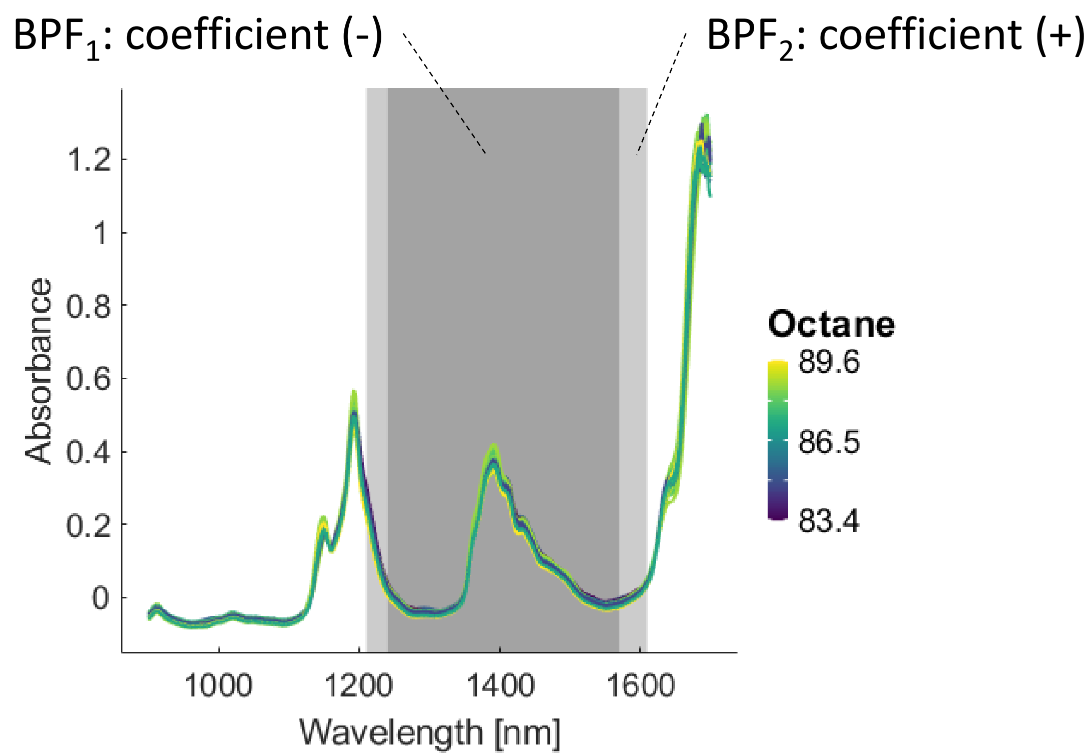
Lower RMSE than PLSR regardless of number of windows

A gasoline NIR spectra case: prediction results



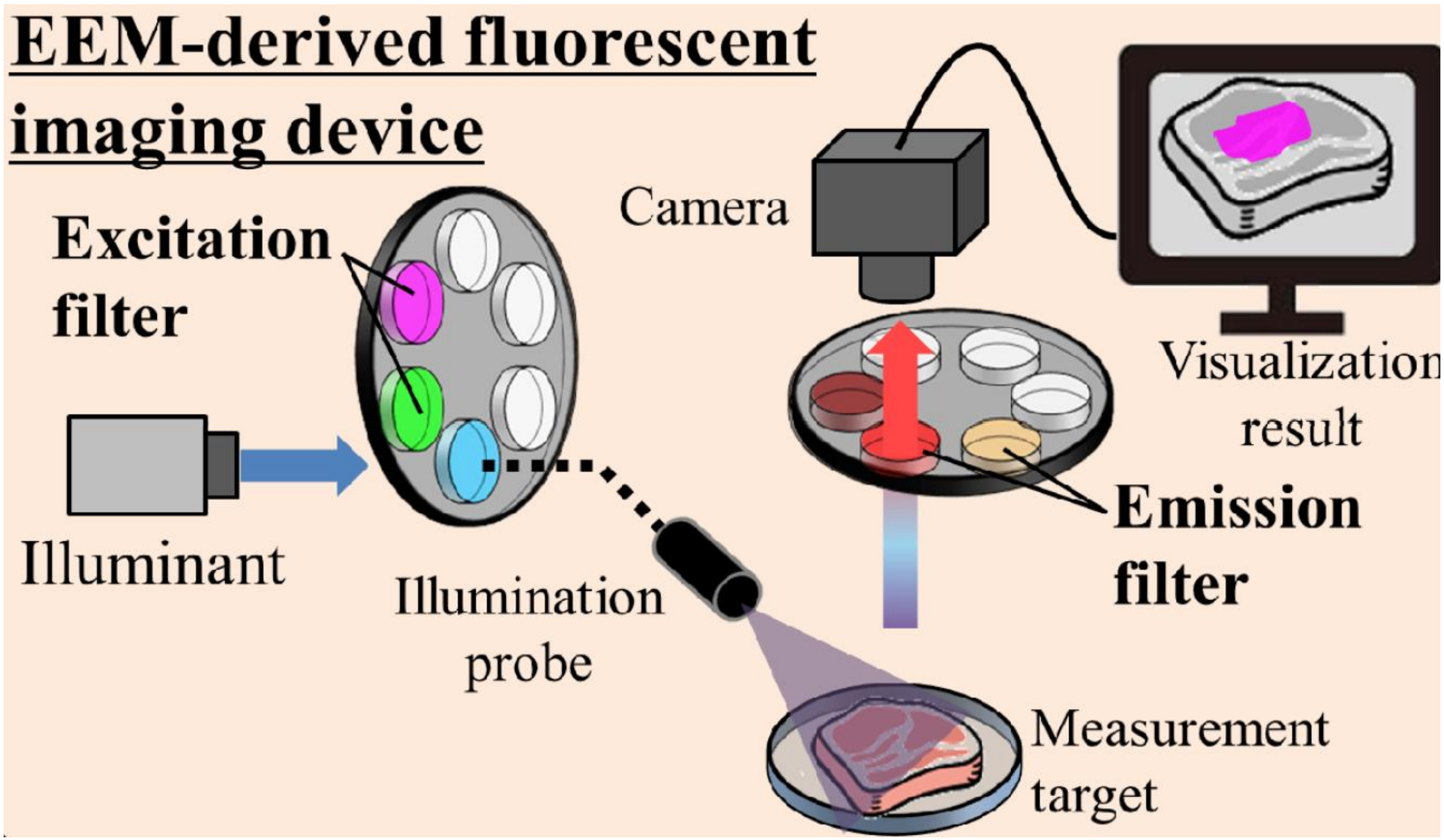


A gasoline NIR spectra case: position of BPFs



- Two BPFs overlapping each other
- Difference between these outputs used
  - Similar to derivatives in NIRS?

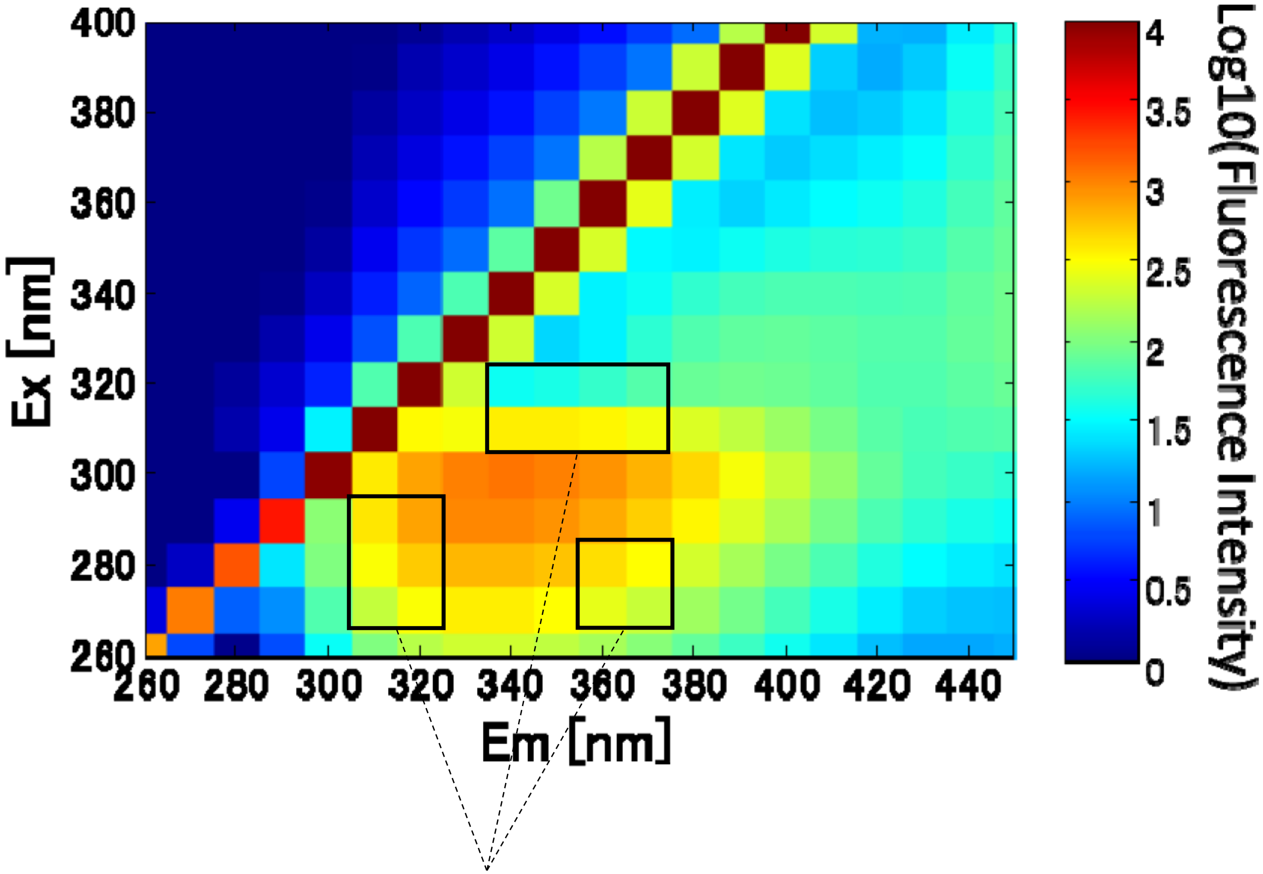
Extension to FF imaging



Sample: pork meat (0-72 h storage@15℃)
Objective: viable bacteria (colony forming unit: CFU)

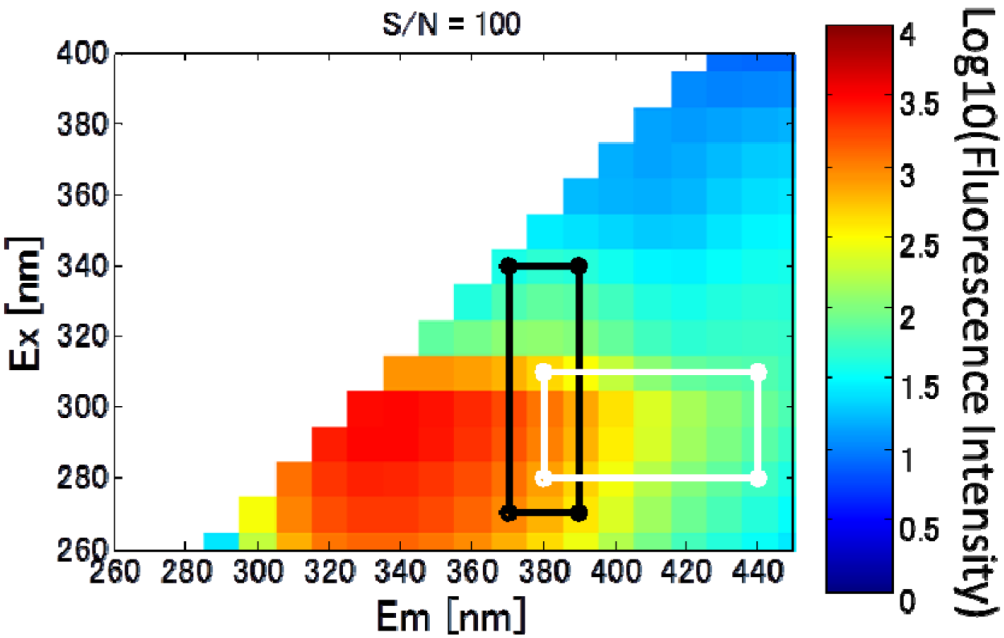
Nishino, K. et al. (2013). *Optics express*, 21(10), 12579-12591.

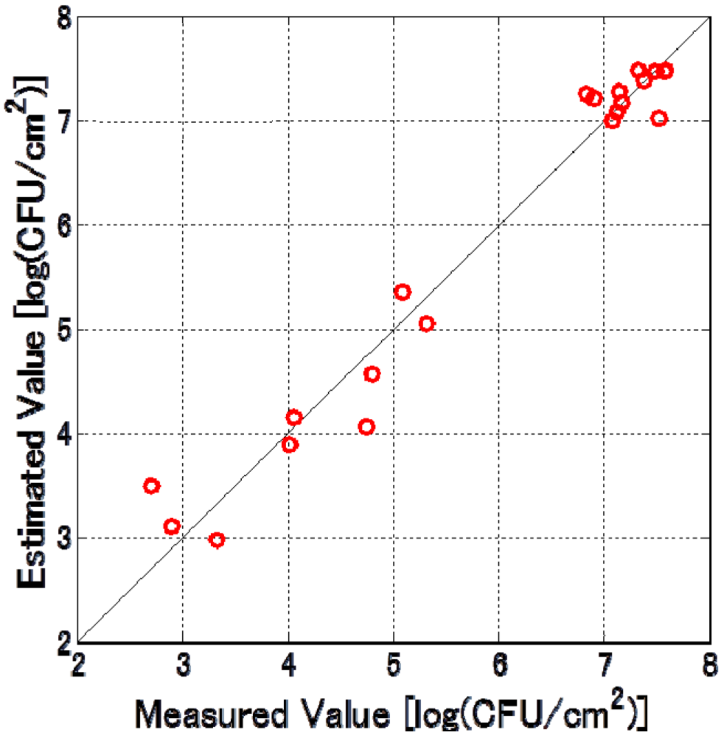
Window search on FF



Sum of intensity=BPF output
->new variable *p*

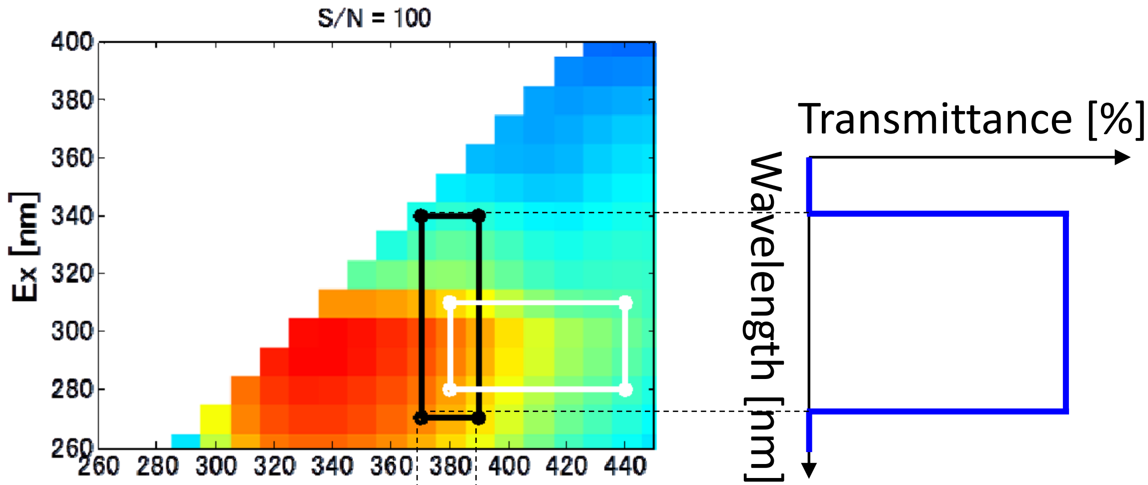
Optimization results

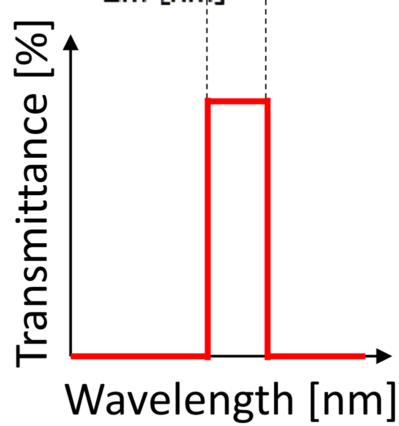


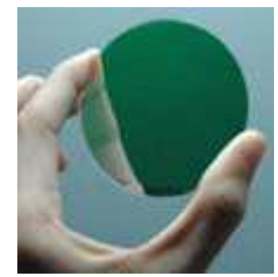


- Squared error of prediction (SEP)
  - PLSR with whole wavelength range: 0.957
  - MLR with two BPFs: 0.805

Customized BPFs based on optimization



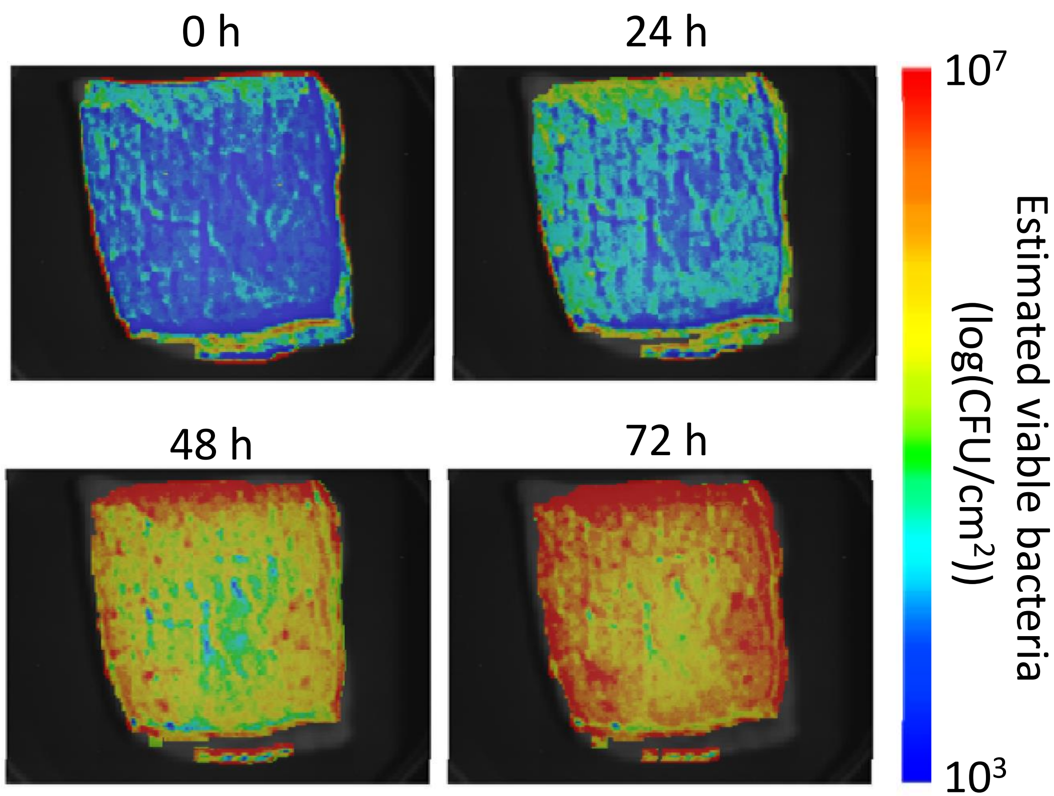




Excitation band-pass filter

Emission band-pass filter

Visualization with customized BPFs



BPF optimization: summary

- Three steps
  - creation of new variables
  - model development
  - selection of optimal variables
- Can be applied to 2D (NIR etc.) and 3D (FF etc.) spectral data
- Can be better than PLSR using whole wavelength range
- Customized BPFs can be developed for imaging

To take home…

Two new VS algorithms

- Stepwise SR
  - No hyperparmeters. You can run it once and will get the same results every time.
  - Good for model accuracy and interpretability improvement.
  - Maybe not enough for BPF based instrument design.
- BPF optimization
  - A bit complicated with 3 steps and high computational load.
  - We can get better accuracy than normal PLSR with only 2-3 BPFs.
  - Imaging hardware can be realized based on the optimization results.

Thank you for your kind attention!