# N-CovSel

## a new strategy for feature selection in N-way data

A. Biancolillo[1], J.M. Roger[2,3] F. Marini[4]

[1] University of l'Aquila, Italy

[2] ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France

[3] ChemHouse Research Group, Montpellier, france

[4] University of Rome La Sapienza, Rome, Italy

# Outline

- Introduction
- Theory
- Application on simulated data
- Application on real data
- Conclusion

# Introduction

- Variable selection for multivariate data analysis:
  - For extracting meaningful features
  - For designing multispectral devices
- A lot of methods
  - Filters, Wrappers, Embedded

But a few address explicitly the case of N-way arrays

- VIP and SR for ND-arrays have been studied in:
  - S. Favilla, C. Durante, M.L. Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, Chemom. Intell. Lab. Syst. 129 (2013) 76–86. https://doi.org/https://doi.org/10.1016/j.chemolab.2013.05.013.
  - M. Cocchi, S. Favilla, M. Li Vigni, N. Cavallini, M. Cocchi, S. Favilla, M. Li Vigni, N. Cavallini, 14th Scandinavian Symposium on Chemometrics (SSC14), Sardegna (Italy), 14-17 June 2015, in: 14th Scand. Symp. Chemom. (SSC14), Sardegna (Italy), 14-17 June, 2015.

# Introduction

- Covariance Selection (Covsel)
  - Is an hybrid method (filter / embedded)
  - Well suited for the 2-D X and 2-D Y
- CovSel is well suited to regression and discrimination
- Covsel yields parsimonious variable selection
- CovSel has been extended to the multiblock framework: SO-CovSel

## N-CovSel proposes an extension of CovSel to the N-D array framework


- Roger, J. M., Palagos, B., Bertrand, D., & Fernandez-Ahumada, E. (2011). CovSel: Variable selection for highly multivariate and multi-response calibration: Application to IR spectroscopy. Chemometrics and Intelligent Laboratory Systems, 106(2), 216-223.
- Biancolillo, A., Marini, F., & Roger, J. M. (2020). SO-CovSel: A novel method for variable selection in a multiblock framework. Journal of Chemometrics, 34(2), e3120.

# Theory

- N-CovSel relies on 5 steps:

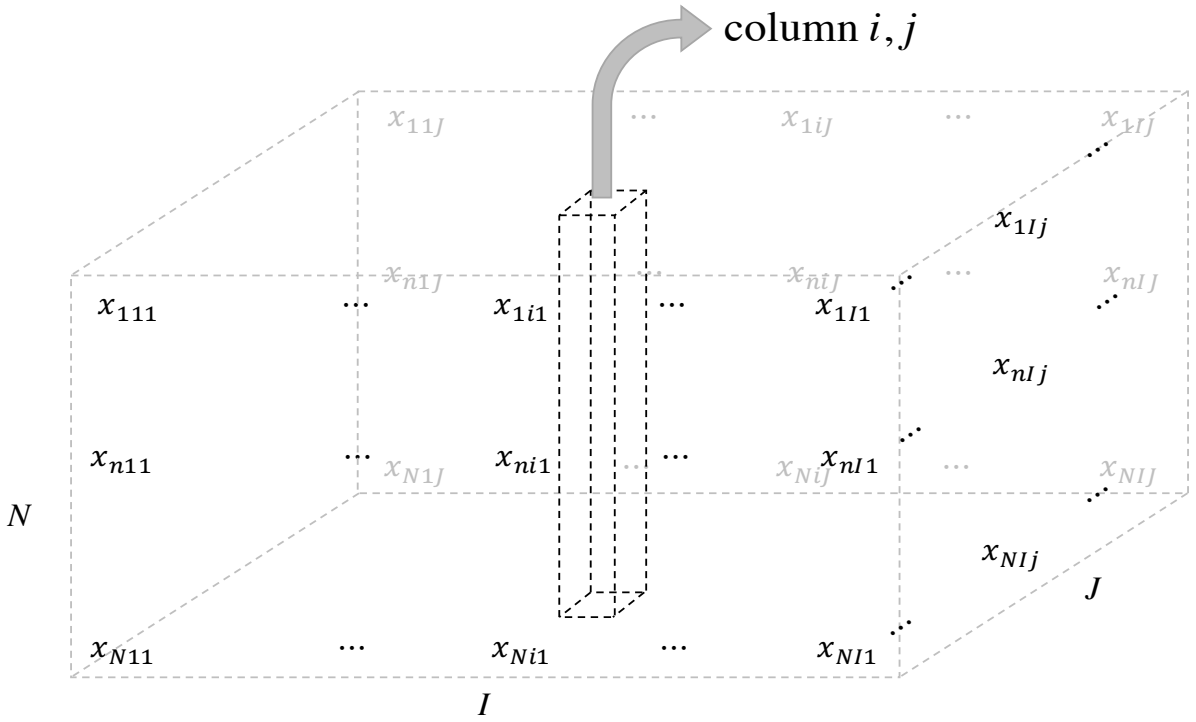
1. The structure of the features to be selected is defined
2. The number of features to be selected is defined
3. The feature of **X** with the highest squared covariance with **Y** is selected
4. **X** is deflated of the information present in the selected feature
5. Steps 3 and 4 are repeated according to the value defined in step 2.

# Theory, step 0

- Notations and operators:
  - **Transpose** : if A is a (I,J,K) array, A' is the (K,J,I) array so that $a'_{kji} = a_{ijk}$
  - **U-product** : if **A** is a (I,J,K) array and **B** is a (K,L,M) array, $\mathbf{A} \odot \mathbf{B}$ is obtained by:
    - Unfolding A and B to obtain $\mathbf{A_u}\ (I \times J, K)$ and $\mathbf{B_u}\ (K, L \times M)$
    - Compute the matrix $\mathbf{M} = \mathbf{A_u}\,\mathbf{B_u}$ $(I \times J, L \times M)$
    - Refold the result into $\mathbf{A} \circledcirc \mathbf{B}$ (*I*,*J*,*L*,*M*)
  - **Norm**: if **A** is a (I,J,K) array, $\mathrm{norm}(\mathbf{A}) = \sqrt{\sum a_{ijk}^2}$

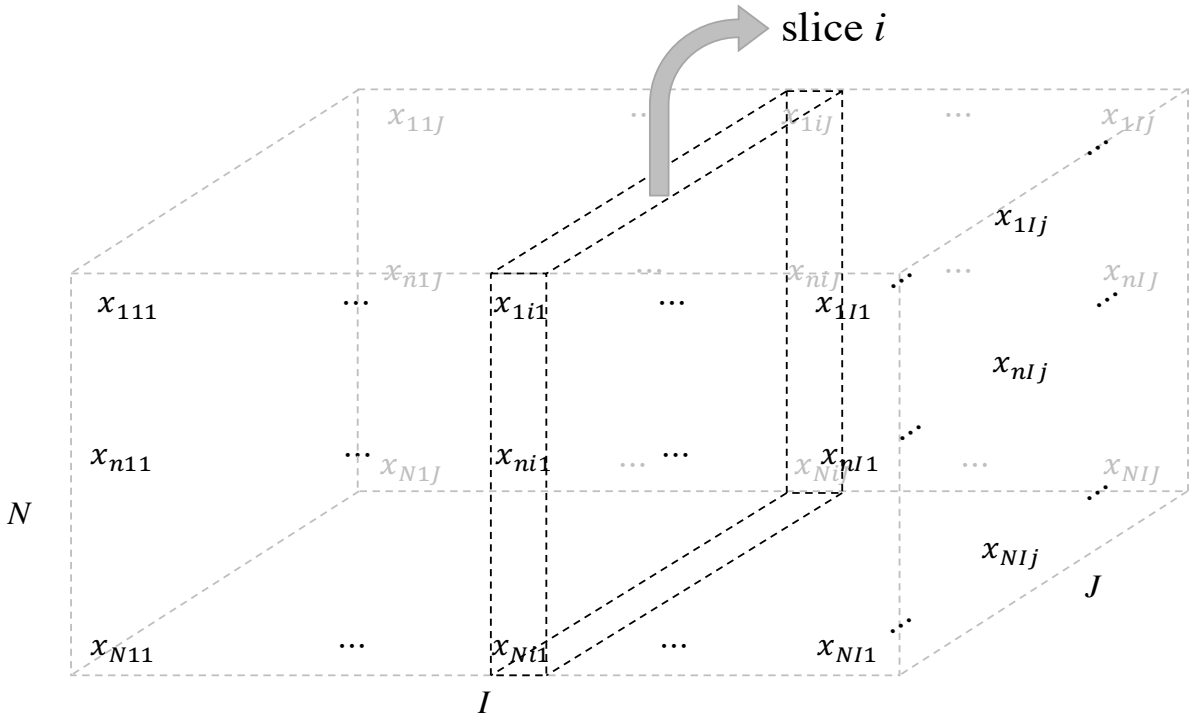
# Theory, step 1:
## Defining the type of features

- The structure of the features to be selected is defined



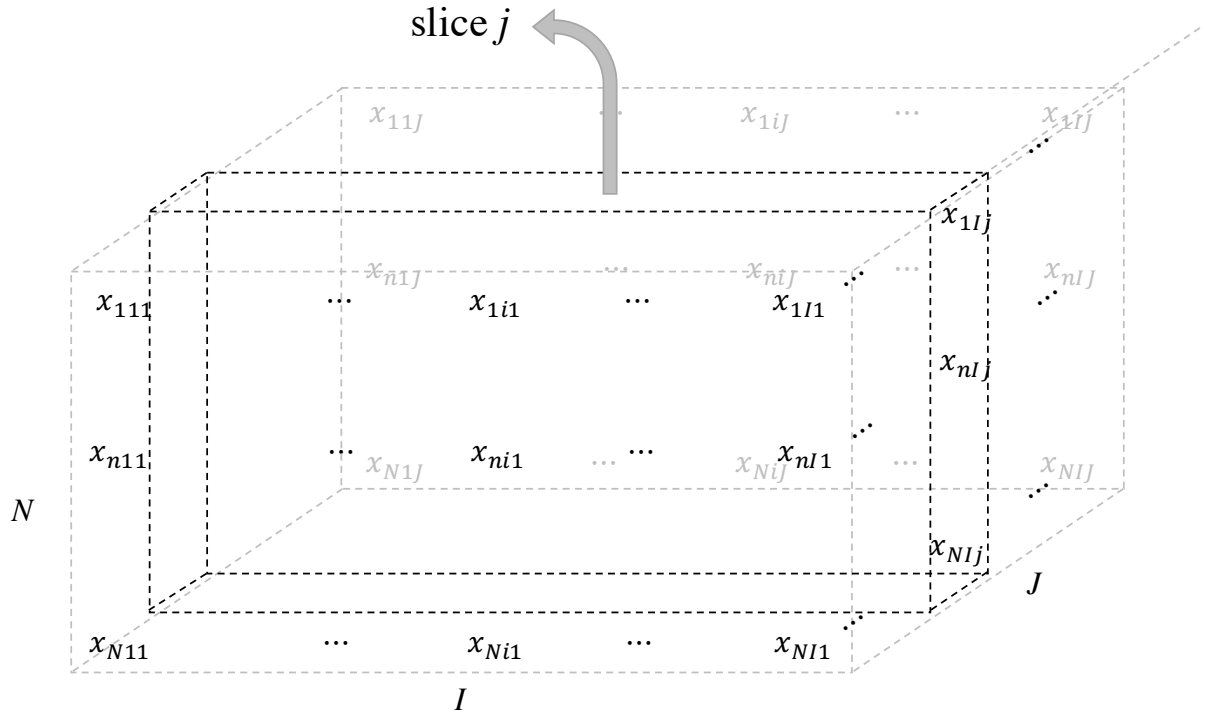
# Theory, step 1:
## Defining the type of features

- The structure of the features to be selected is defined



# Theory, step 1:
## Defining the type of features

- The structure of the features to be selected is defined




From a (N,I,J,K,L) array:

- A 1-D feature is given by 4 indexes I,j,k,l
- A 2-D feature by 3 indexes j,k,l or I,k,l or I,j,l of I,j,k
- A 3-D feature by 2 indexes I,j or I,k or I,l etc...
- A 4-D feature by 1 index I, j, k or l

# Theory, step 3:
## Selecting the feature

- The feature U of X with the highest squared covariance with Y is selected
  - For 1-D feature: $cov^2(\mathbf{u}, \mathbf{Y}) = \frac{1}{N}\mathbf{u}'\mathbf{Y}\mathbf{Y}'\mathbf{u}$
  - For 2-D feature: $cov^2(\mathbf{U}, \mathbf{Y}) = \frac{1}{N}\mathrm{norm}(\mathbf{U}'\mathbf{Y}\mathbf{Y}'\mathbf{U})$
  - For >2-D features $cov^2(\mathbf{U}, \mathbf{Y}) = \frac{1}{N}\mathrm{norm}(\mathbf{U}' \circledcirc \mathbf{Y} \circledcirc \mathbf{Y}' \circledcirc \mathbf{U})$
  - So $I_{sel} = \mathrm{ArgMax}_i\left(\mathrm{norm}(\mathbf{U_i}' \circledcirc \mathbf{Y} \circledcirc \mathbf{Y}' \circledcirc \mathbf{U_i})\right)$

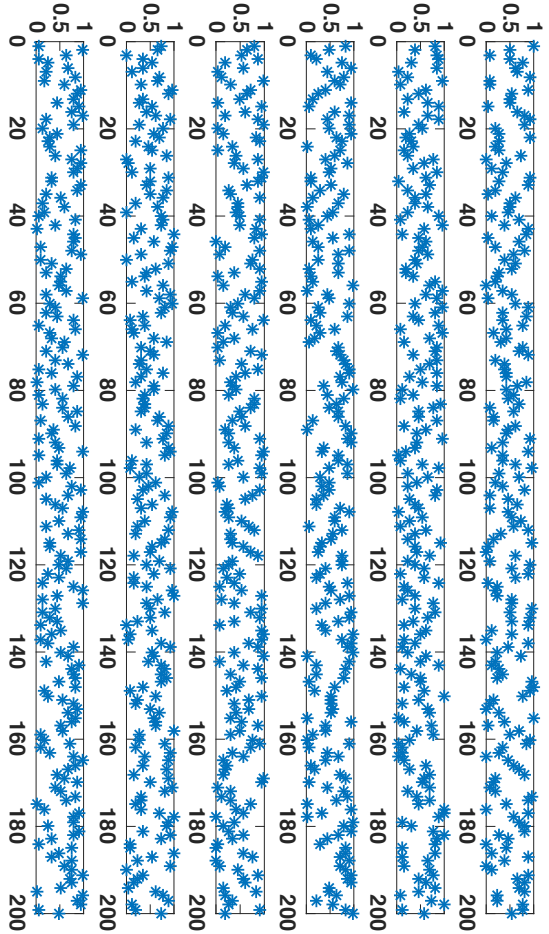
# Theory, step 4:
## Deflating the data

- Let assume that **X** is (*N,I,J,K,L,M*) and **U** is (*N,I,J*)

  - Unfold **U**: $\mathbf{u} = \text{reshape}(\mathbf{U}, N \times I \times J, 1)$ i.e. $\mathbf{u} = \text{vec}(\mathbf{U})$
  - Reshape **X**: $\mathbf{Z} = \text{reshape}(\mathbf{X}, N \times I \times J, K, L, M)$
  - Deflation: $\mathbf{Z}_\text{d} = \mathbf{Z} - (\mathbf{u}(\mathbf{u}^T\mathbf{u})^{-1}\mathbf{u}^T) \circledcirc \mathbf{Z}$
  - Reshape **X:** $\mathbf{X}_d = reshape(\mathbf{Z}_\text{d}, N, I, J, K, L, M)$

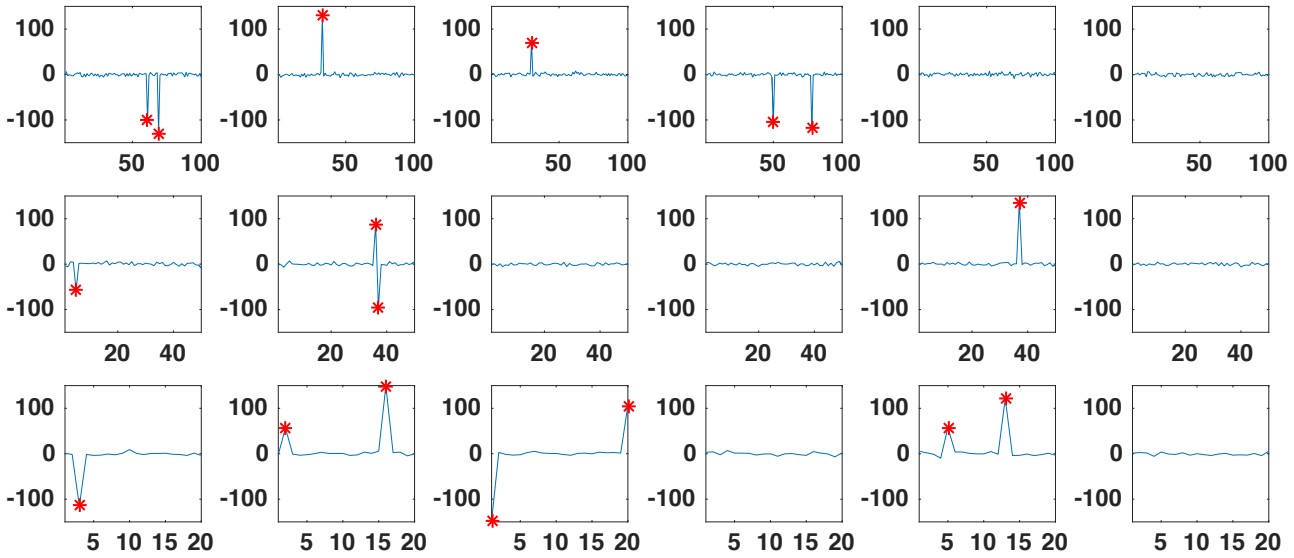
- If X is 2D and U 1D, we retrieve the classical deflation process

# Application on simulated data



6 vectors of 200 scores

*

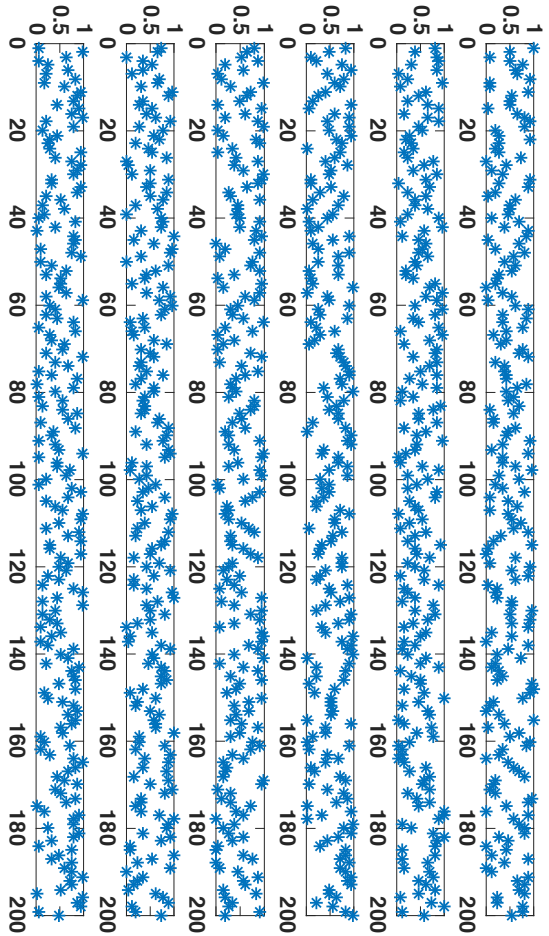


3 matrices of 6 loadings of size (100, 50, 20)
with 6, 4 and 7 variables sign. non null

+ noise

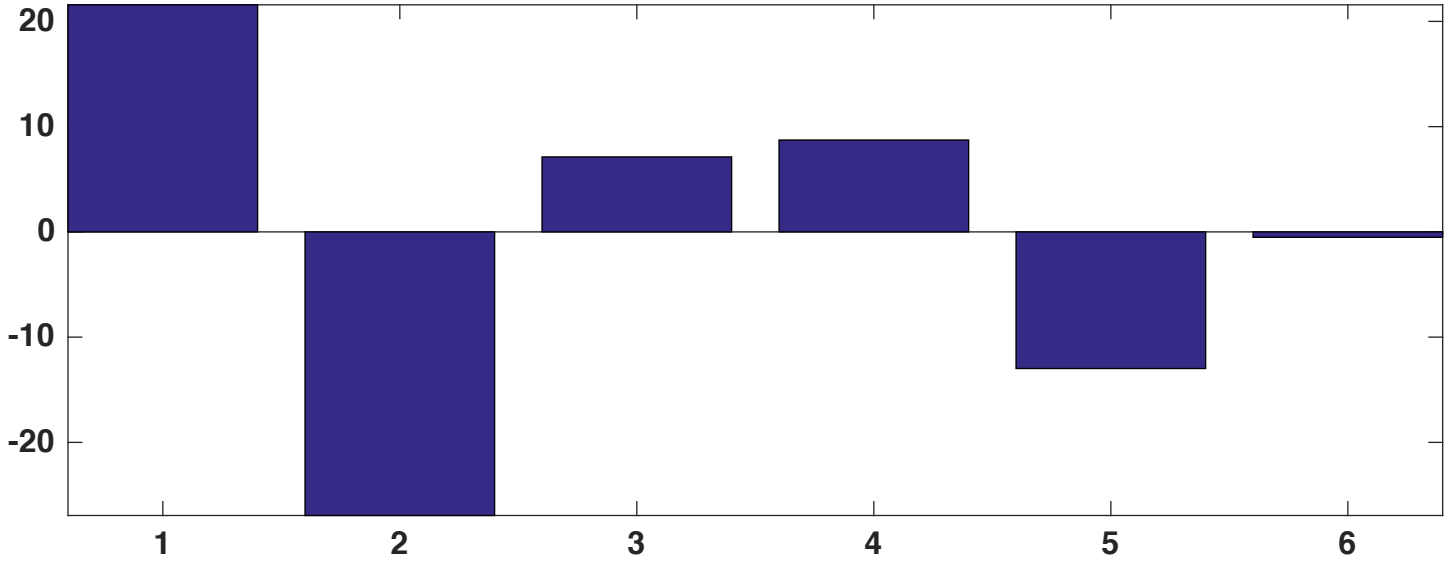
→ 4-D **X** (200,100,50,20)

# Application on simulated data



6 vectors of 200 scores

*



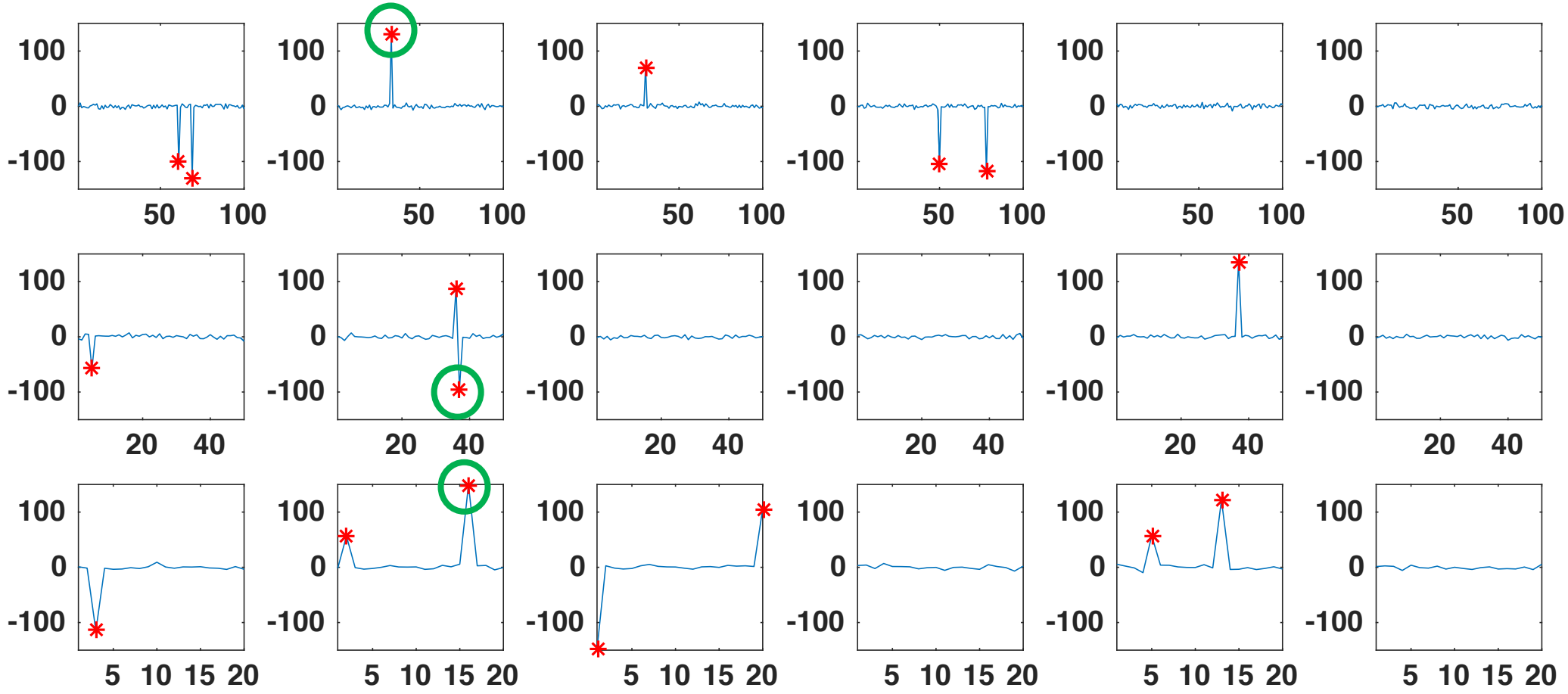

6 b-coefficients

+ noise

→ **Y** (200,1)

# Application on simulated data

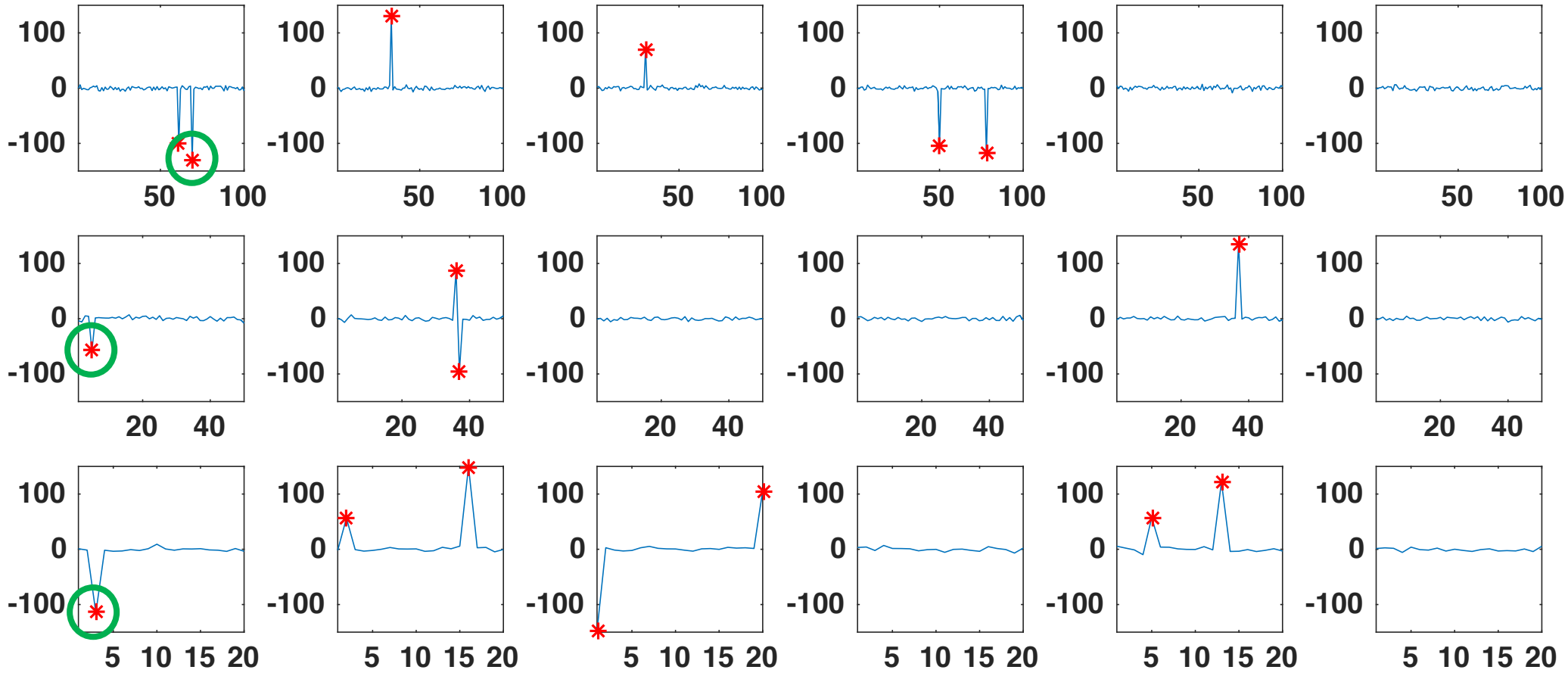
## Searching for 1-D features




First round -> 33, 37, 16

# Application on simulated data

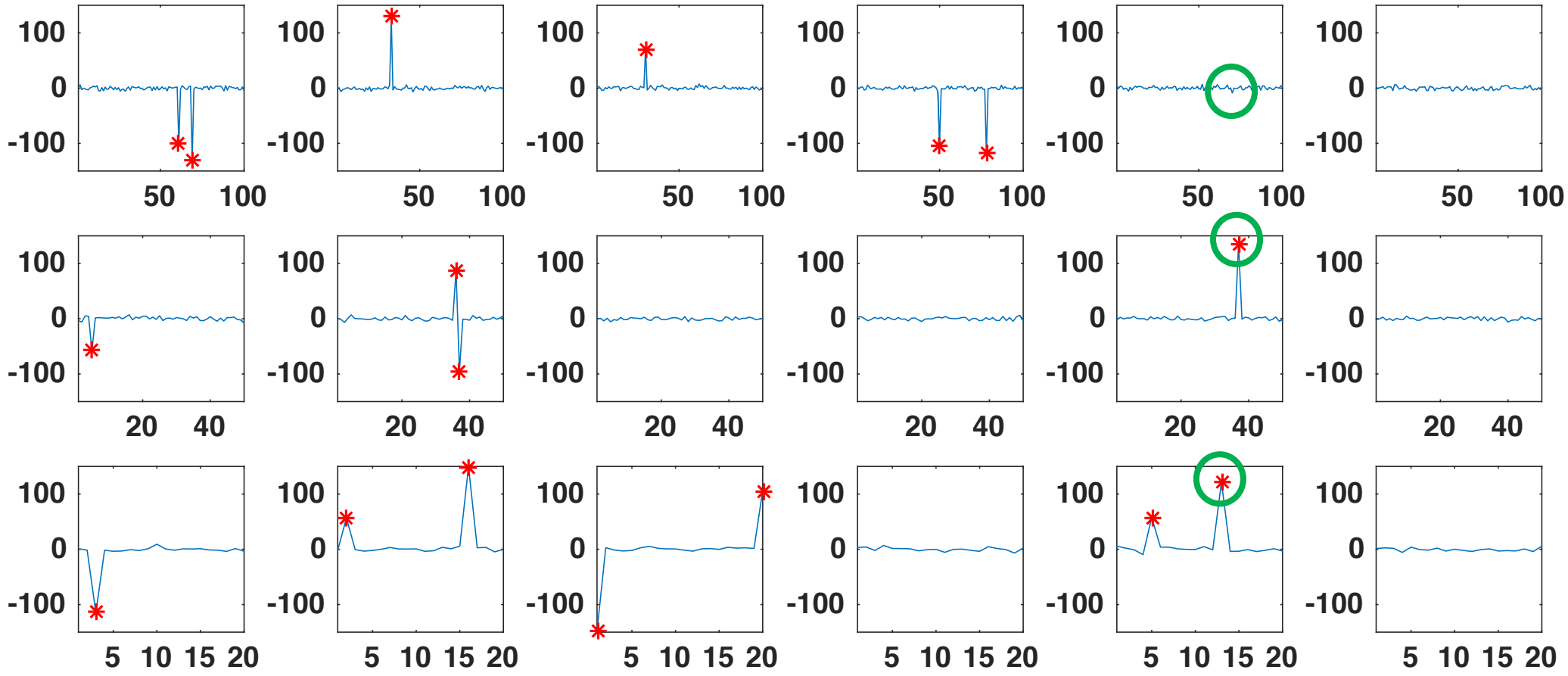
## Searching for 1-D features




Second round -> 69, 5, 3

# Application on simulated data

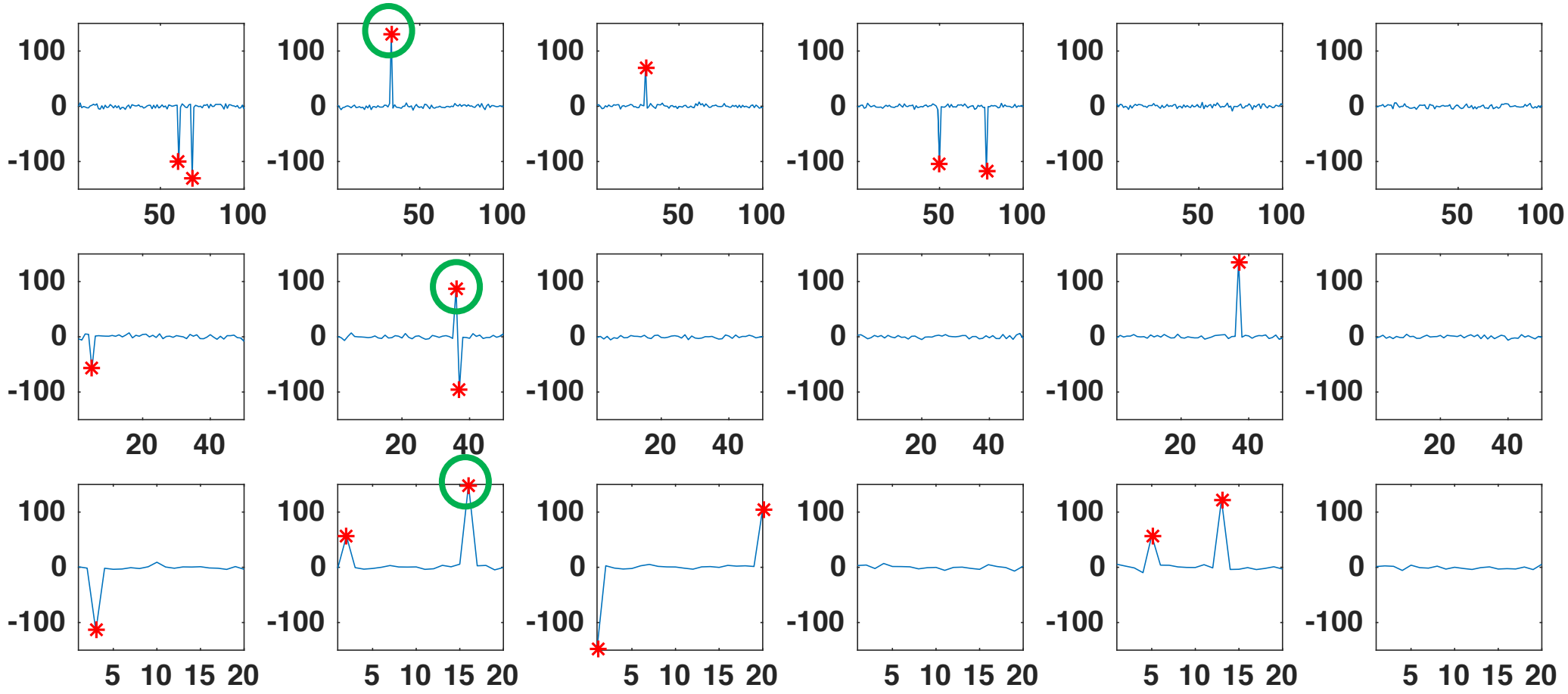
## Searching for 1-D features




Third round -> 70, 37, 13

# Application on simulated data

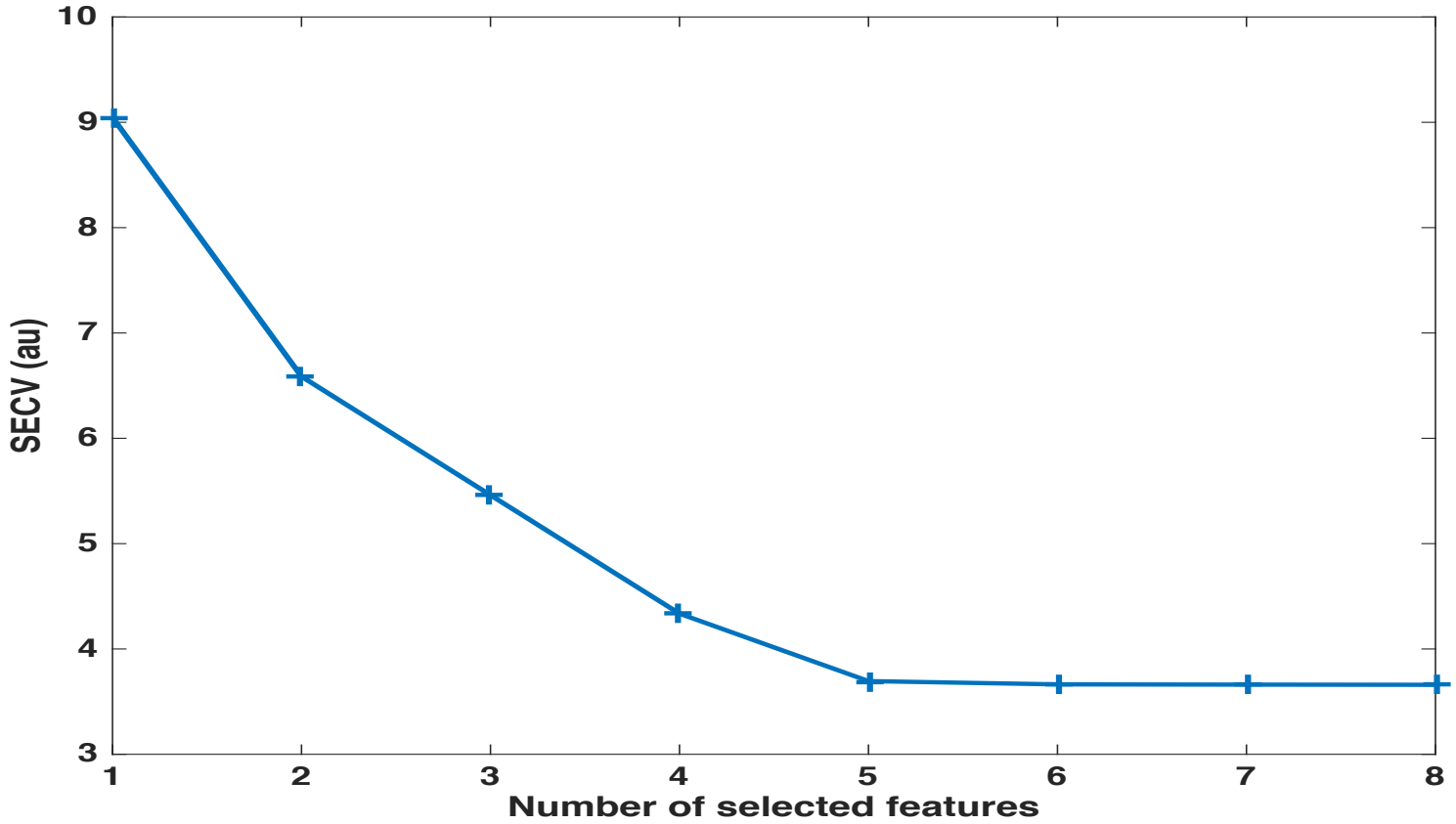
## Searching for 1-D features




Fourth round -> 33, 36, 16

# Application on simulated data

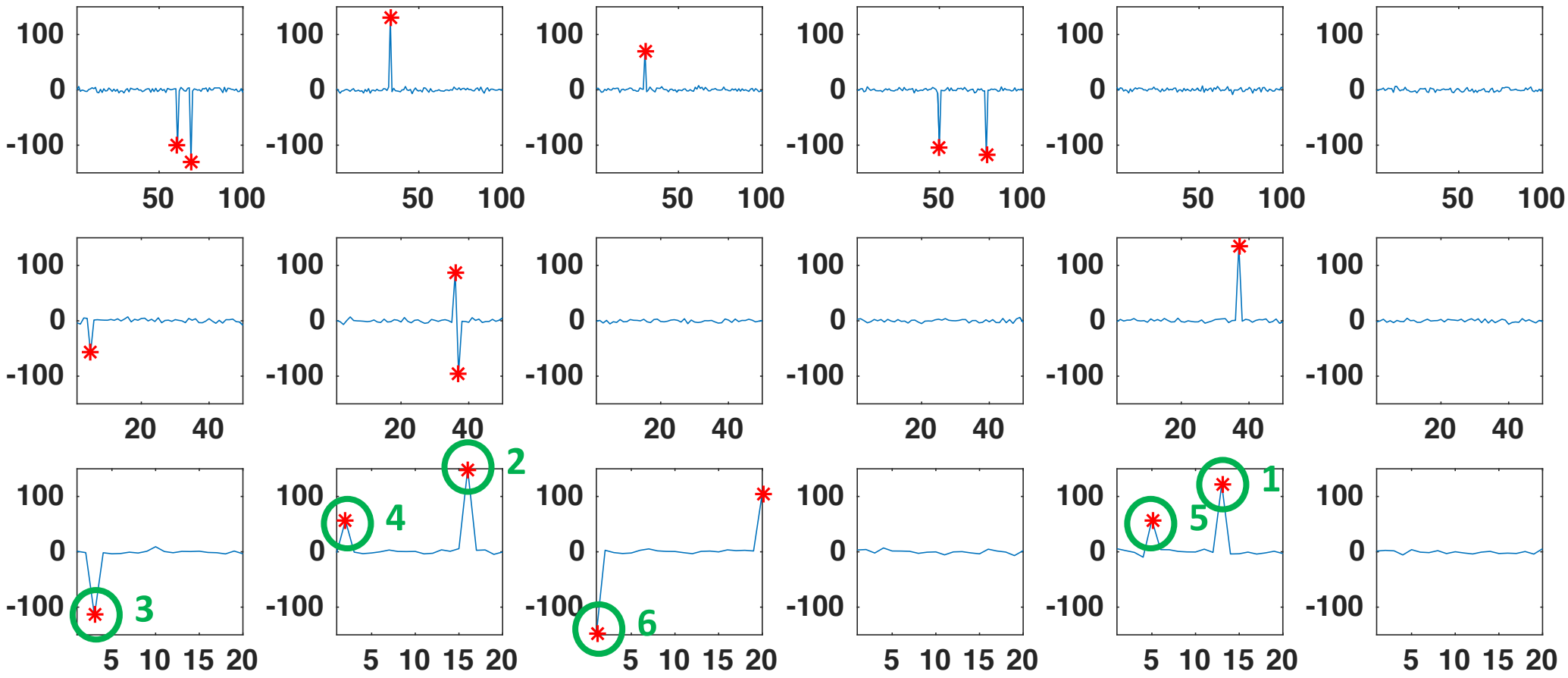
## Searching for 1-D features




Cross-validation of a PLS calibrated on the selected columns

# Application on simulated data

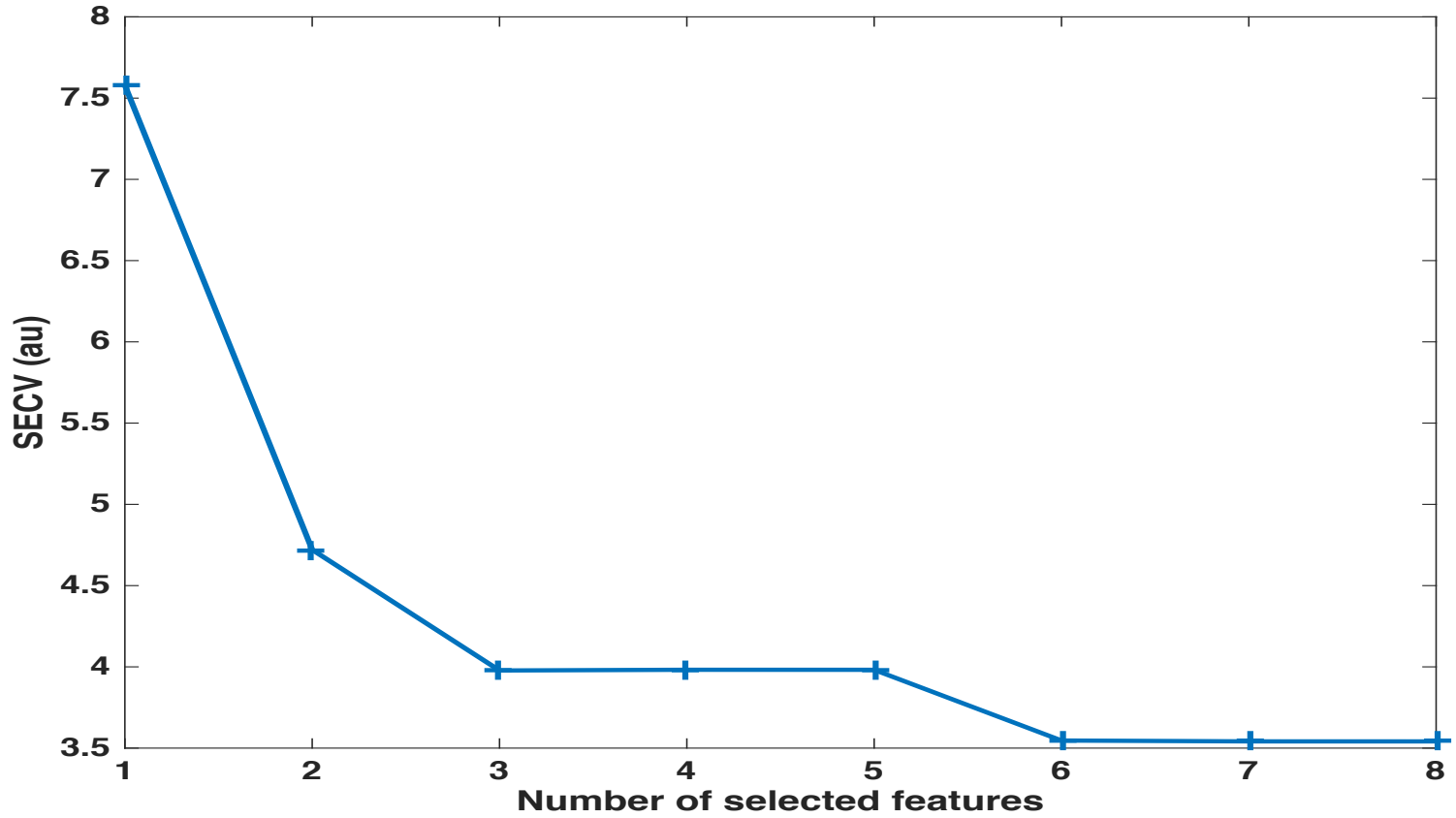
## Searching for 3-D features, along mode 4




Rounds 1 to 6: 13, 16, 3, 2, 5, 1

# Application on simulated data

## Searching for 3-D features, along mode 4




Cross-validation of a N-PLS calibrated on the selected cubes

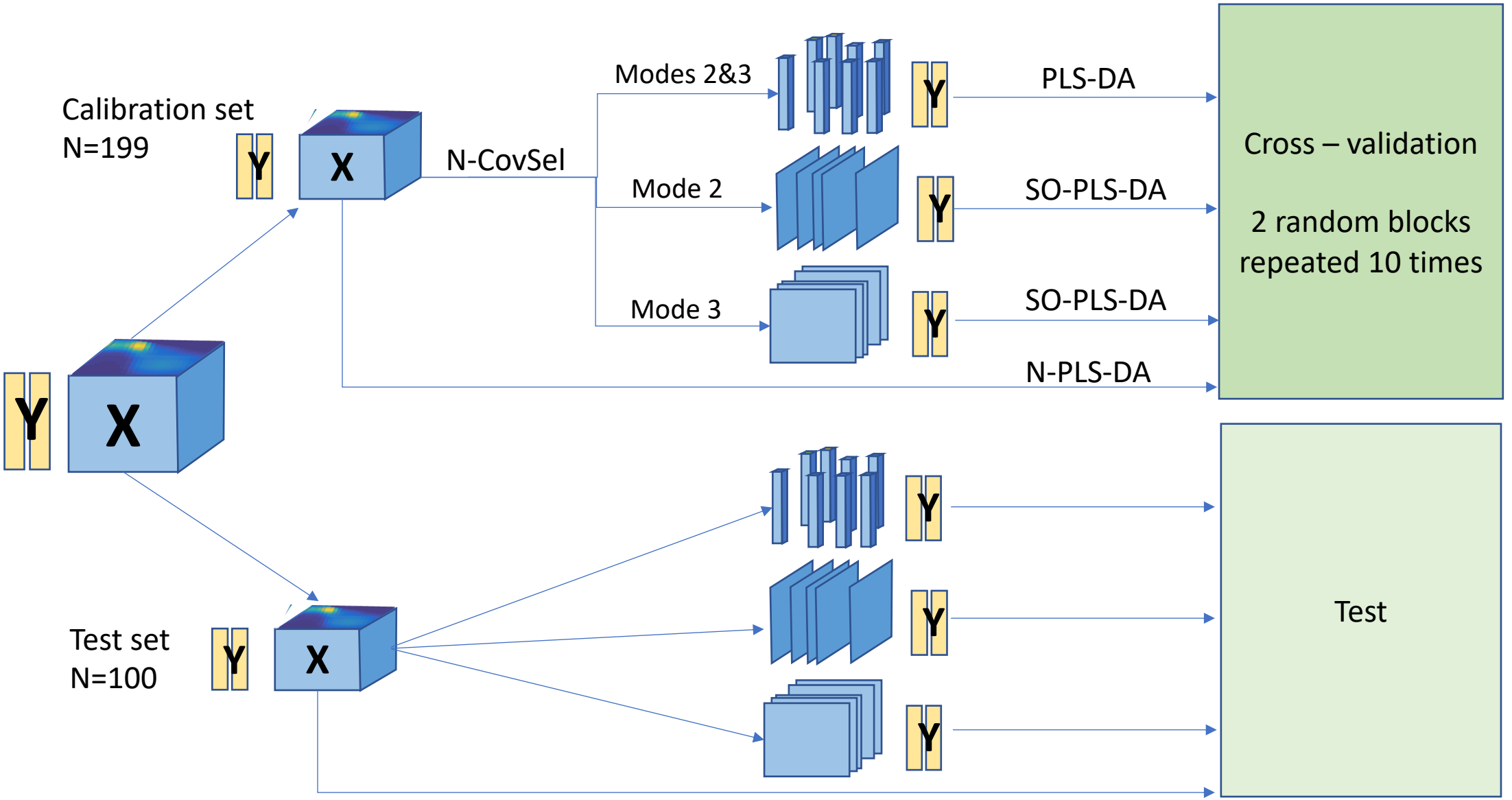
# Application on real data

- 3-D fluorescence spectra of human blood serum (X) vs cancer cases (Y)
  - N = 299 samples, 225 controls, 74 cases
  - I = 301 emission wavelength, from 300 to 600 nm
  - J = 41 excitation wavelengths, from 250 to 450 nm

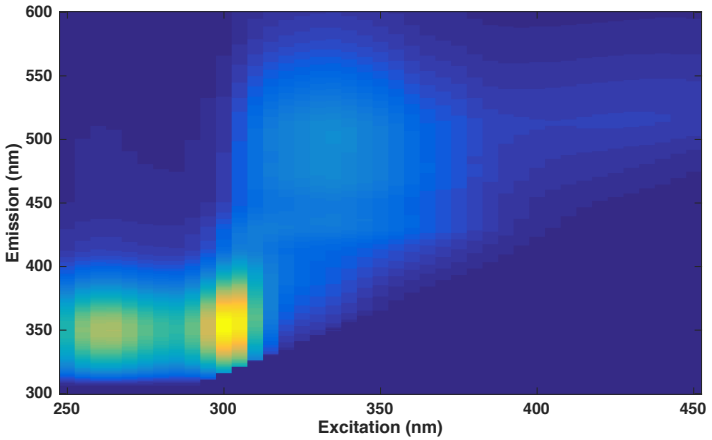
- Available online at: http://www.models.life.ku.dk/anders-cancer

Lawaetz, A.; Bro, R.; Kamstrup-Nielsen, M.; Christensen, I.; Jørgensen, L.; Nielsen, H. Fluorescence spectroscopy as a potential metabonomic tool for early detection of colorectal cancer. Metabolomics 8 (supplement 1): 111-121 (2012).

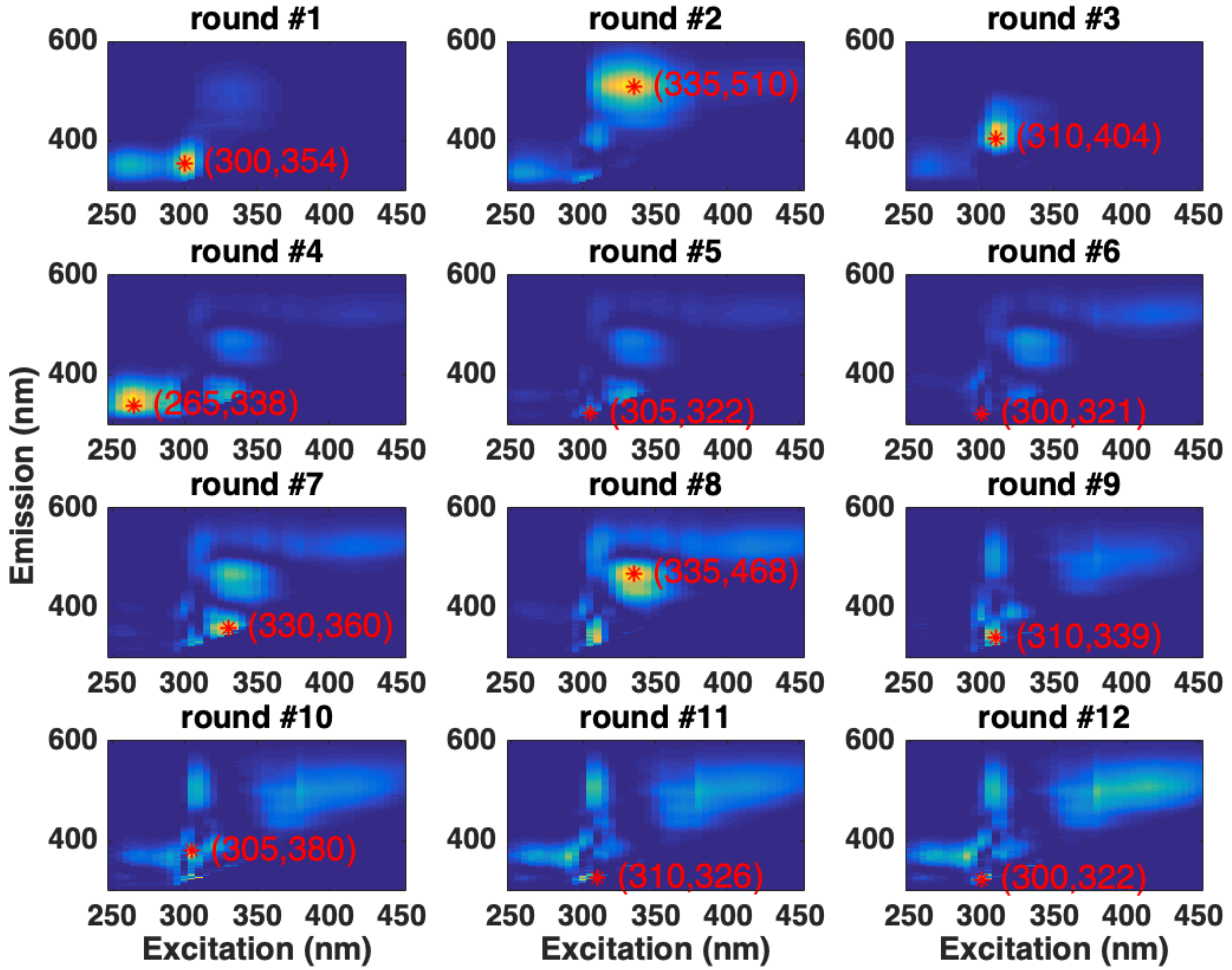
# Application on real data



# Application on real data

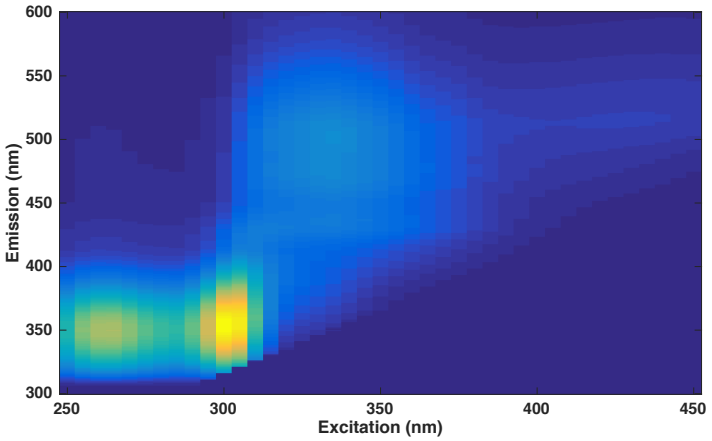



Mean spectrum




Evolution of the $Cov^2$ map along N-CovSel column selection process

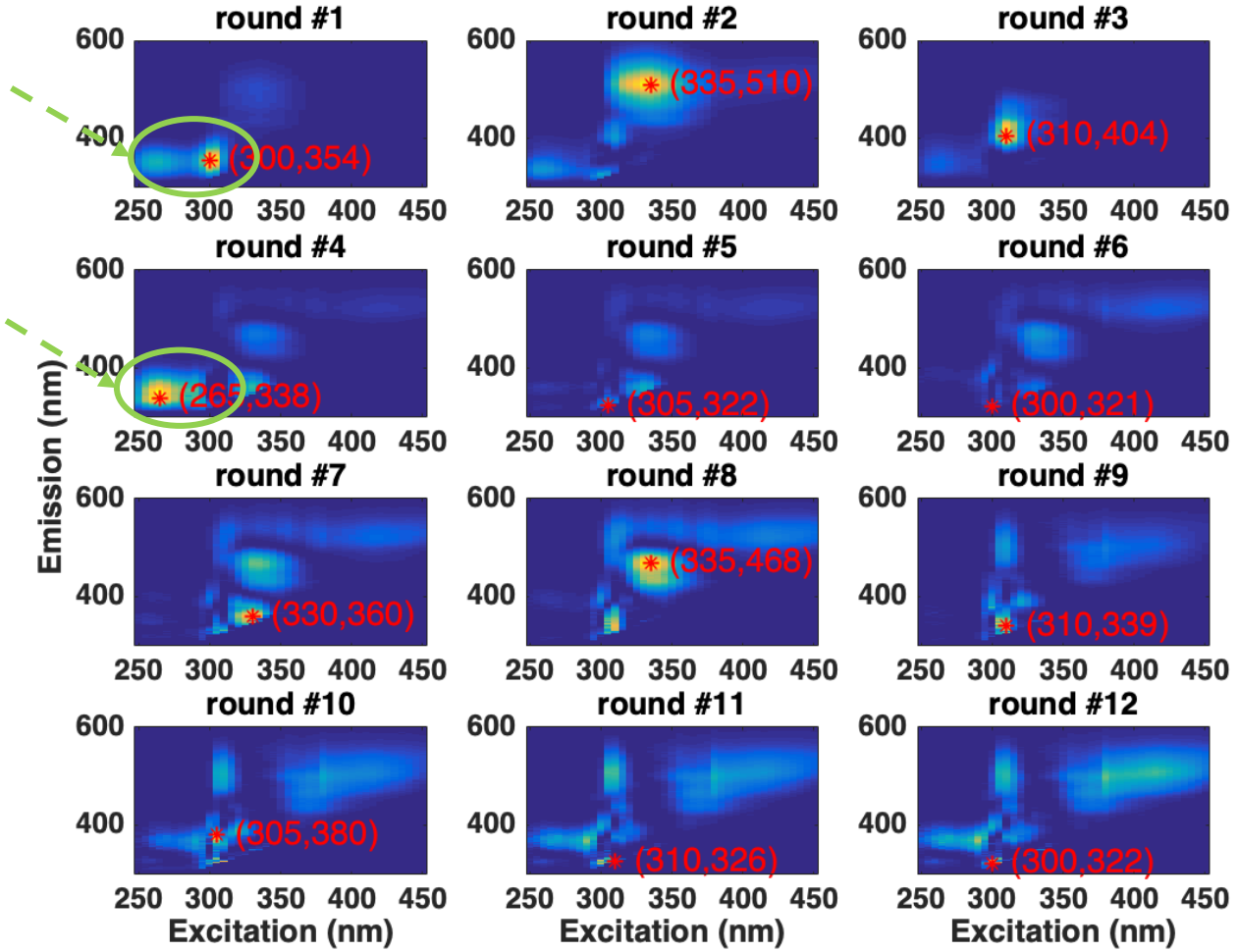
# Application on real data




Mean spectrum

Tryptophan

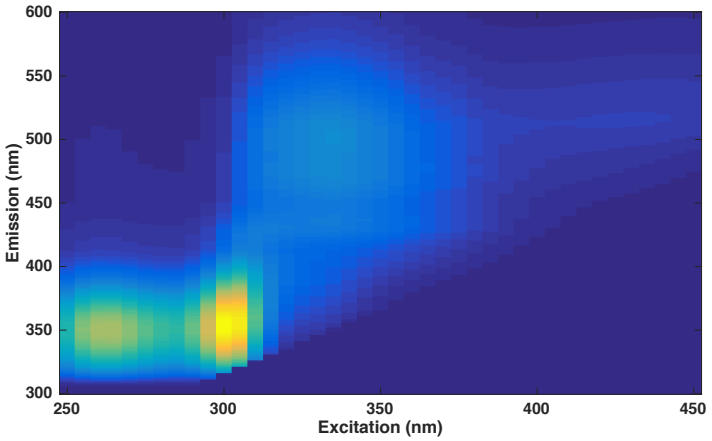
Tyrosine




Evolution of the $Cov^2$ map along N-CovSel column selection process

Lawaetz, A.; Bro, R.; Kamstrup-Nielsen, M.; Christensen, I.; Jørgensen, L.; Nielsen, H. Fluorescence spectroscopy as a potential metabonomic tool for early detection of colorectal cancer. Metabolomics 8 (supplement 1): 111-121 (2012).

# Application on real data

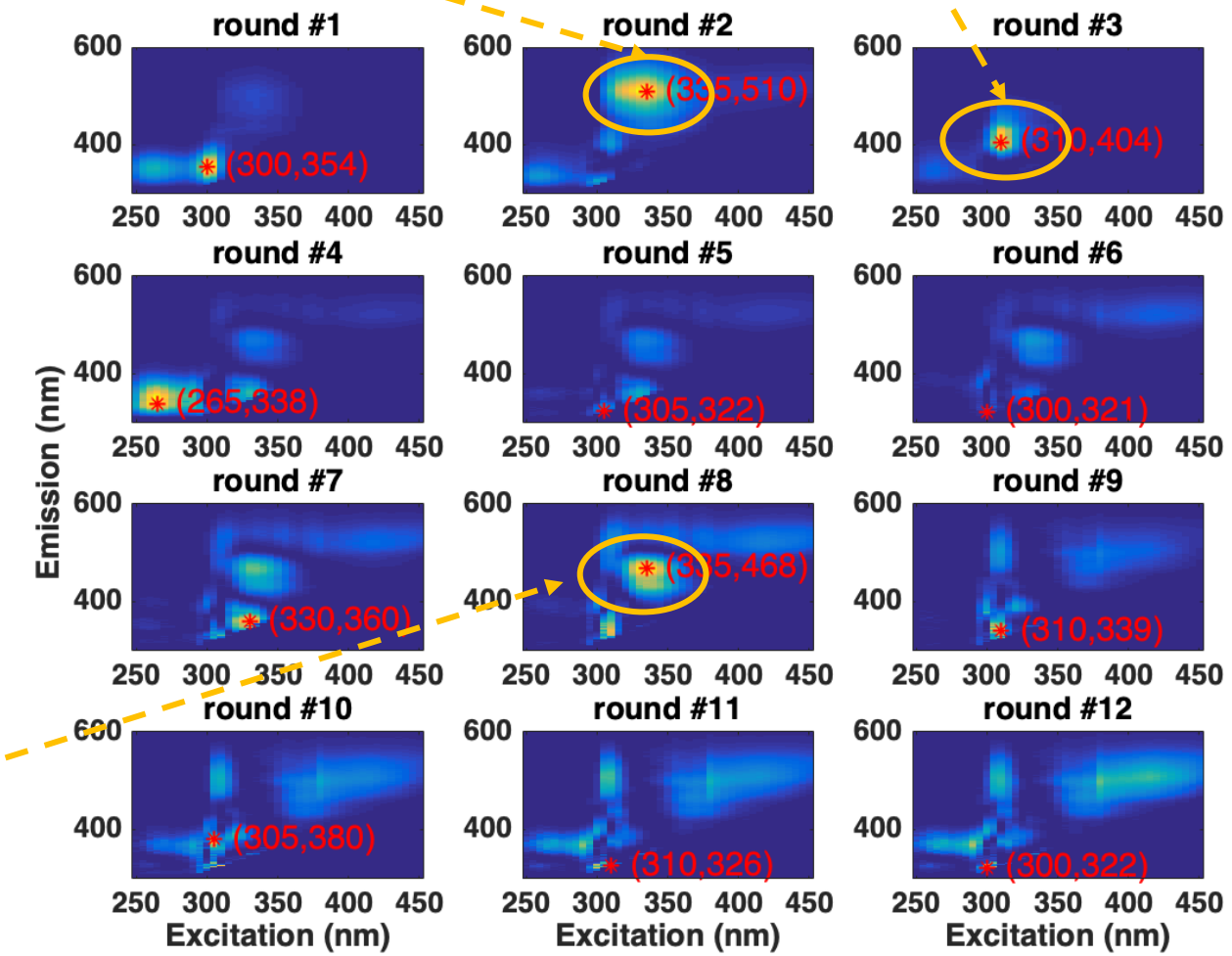



Mean spectrum

Ryboflavin ?
B vitamin are associated to intestinal diseases

3-Hydroxyanthranilic acid ?
Intermediate in tryptophan metabolism

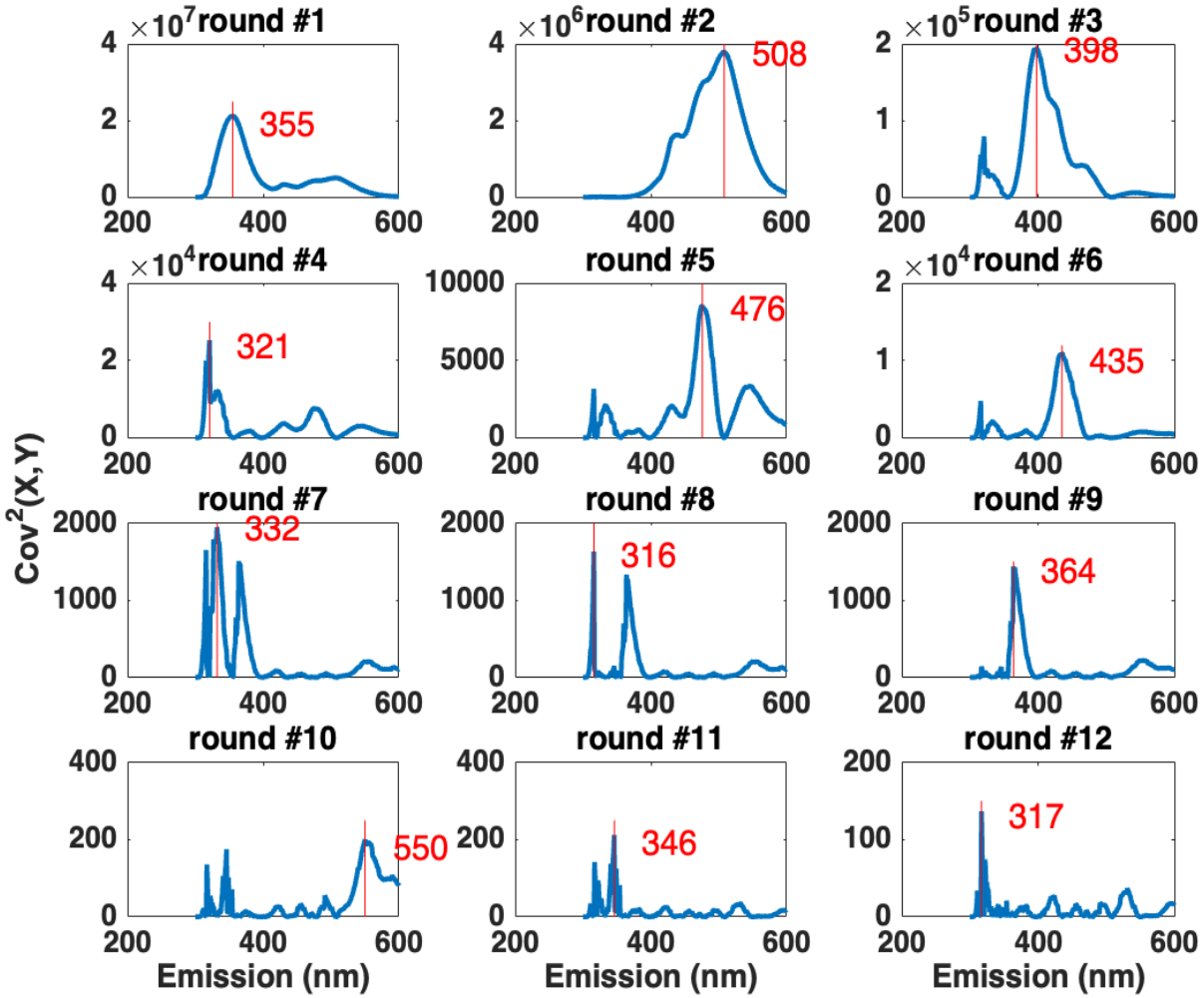
Free NAD(P)H




Evolution of the $Cov^2$ map along N-CovSel column selection process

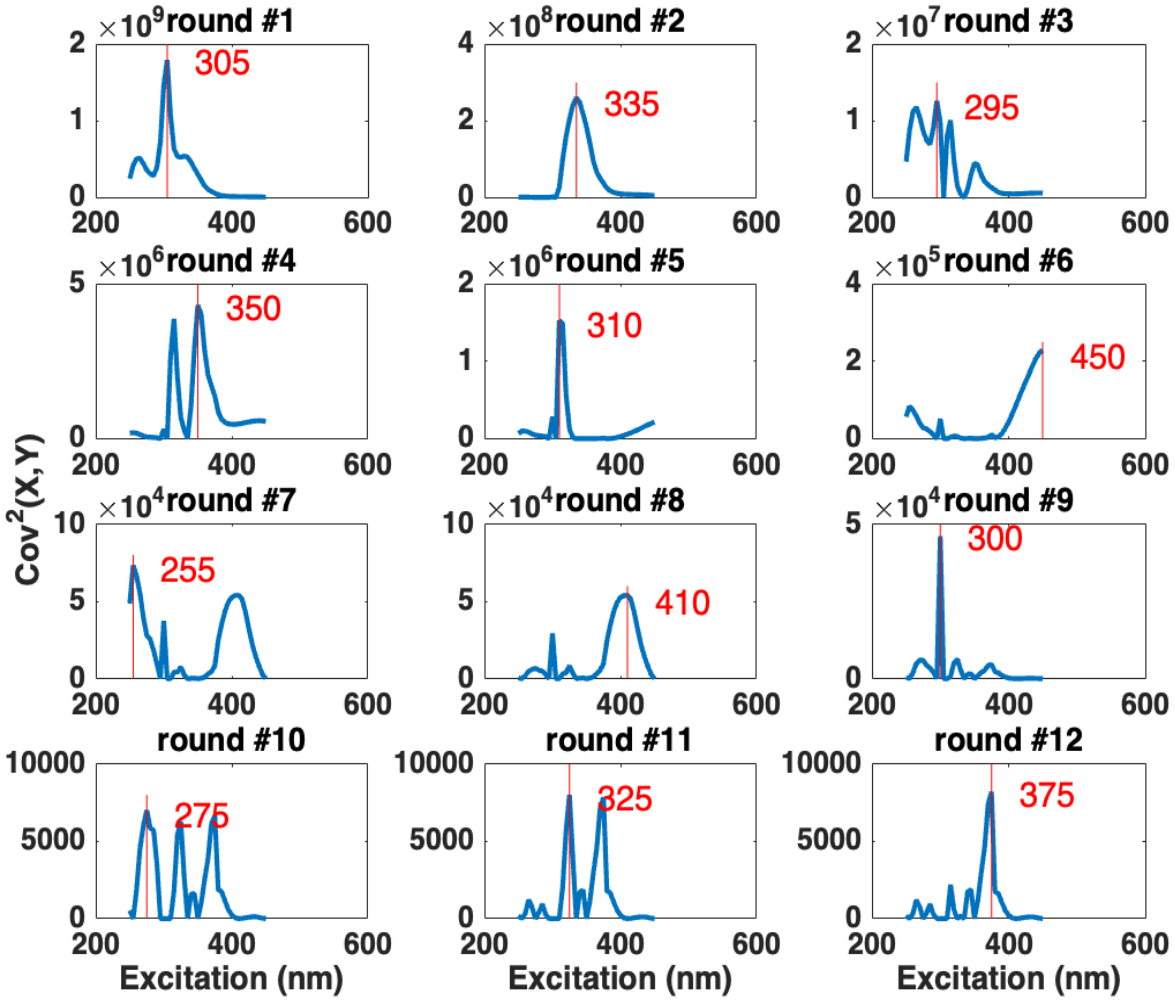
Wolfbeis, O. S., & Leiner, M. (1985). Mapping of the total fluorescence of human blood serum as a new method for its characterization. *Analytica Chimica Acta*, *167*, 203-215.

# Application on real data




Evolution of the Cov2 curve along N-CovSel emission slice selection process

# Application on real data




Evolution of the Cov2 curve along N-CovSel excitation slice selection process

# Application on real data

| Type of Feature | Features selected (nm) | ACA* in CV (%) | ACA in test (%) |
|---|---|---|---|
| Excitation x Emission | 354/510, 404/338, 322/321, 360/468, 339/380, 300/335, 310/265, 305/300, 330/335, 310/305 | 66.3 | 70.7 |
| Emission | 355, 508 | 68.4 | 62.0 |
| Excitation | 305,335,295,350,310,450 | 67.2 | 66.7 |
| | 305 | 64.5 | 69.3 |

| | LVs | ACA in CV (%) | ACA in test (%) |
|---|---|---|---|
| N-PLS-DA | 5 | 67.6 | 67.3 |

N-CovSel results are comparable to those obtained with a N-PLS-DA on the whole cube
Parsimonious selections of slices (emission or excitation) are available

*: Average Classification Accuracy

# Conclusion

- N-CovSel extends the Covariance based selection to the N-way
- It allows us to select any sub-dimensions of the N-D Array
- It provides parsimonious selections, allowing:
  - Compound identification
  - Simplification of the measurement
  - Sensor design

- Further work has to be carried out:
  - To speed up the algorithm
  - To study alternative modes of deflation