uDOP:

a new method of calibration transfer without standards

Jean-Michel Roger[1,2], Valeria Fonseca Diaz[2,3]

ITAP, Univ Montpellier, INRAE, Institut Agro, Montpellier, France

ChemHouse Research Group, Montpellier, France

KU Leuven Department of Biosystems, MeBioS division, Leuven, Belgium

jean-michel.roger@inrae.fr

Hypothesis and notations

- The data :

  $\mathbf{X}(n, p)$ the predictors

  $\mathbf{y}(n, 1)$ the response, to be estimated from $\mathbf{X}$

  We will specifically address the highly multivariate case, like spectra

- We only address the linear modelling, i.e. :

$$\hat{\mathbf{y}} = \mathbf{X}\,\mathbf{b}$$

- Two domains:

  > Source domain $\mathbf{X}_s$ where a model has been done

  > Target domain $\mathbf{X}_t$ where the model will be transfered

What's the problem?

Ageing of the devices

What's the problem?

Ageing of the

Changing the variety

What’s the problem?

Ageing of the

Changing the

Changing the origin

What’s the problem?

Ageing of the

Changing the

Changing the

Changing the measurement conditions

What's the problem?

Ageing of the

Changing the

Changing the

Changing the

Changing the sample preparation

What’s the problem?

Ageing of the

Changing the

Changing the

Changing the

Changing the

…

→

Provokes deviations in the measured spectra ($\delta\mathbf{x}$)

What's the problem?

Ageing of the

Changing the

Changing the

Changing the

Changing the

…

→ Provokes deviations in the measured spectra ($\delta\mathbf{x}$)

$\delta\mathbf{x}$ can be of various nature:

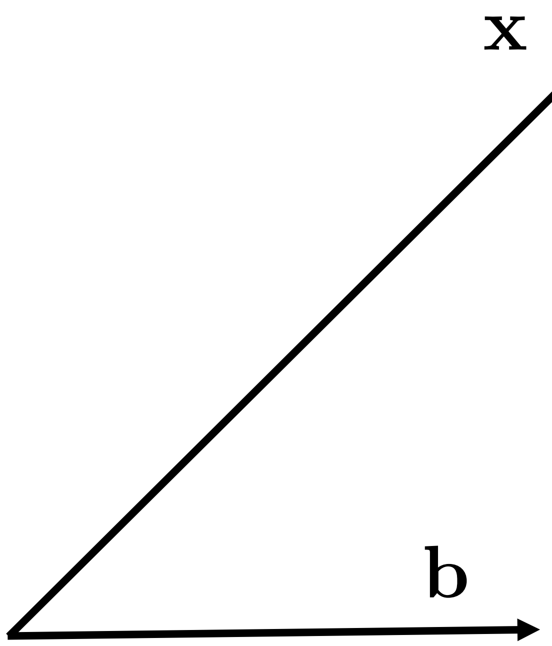
- Baselines
- Random noise
- Spikes
- Wavelength shifts and drifts
- New peaks
- etc.

This $\delta\mathbf{x}$ challenges the reproducibility of the model

Robustness = Reproducibility of the model w.r.t. $\delta\mathbf{x}$

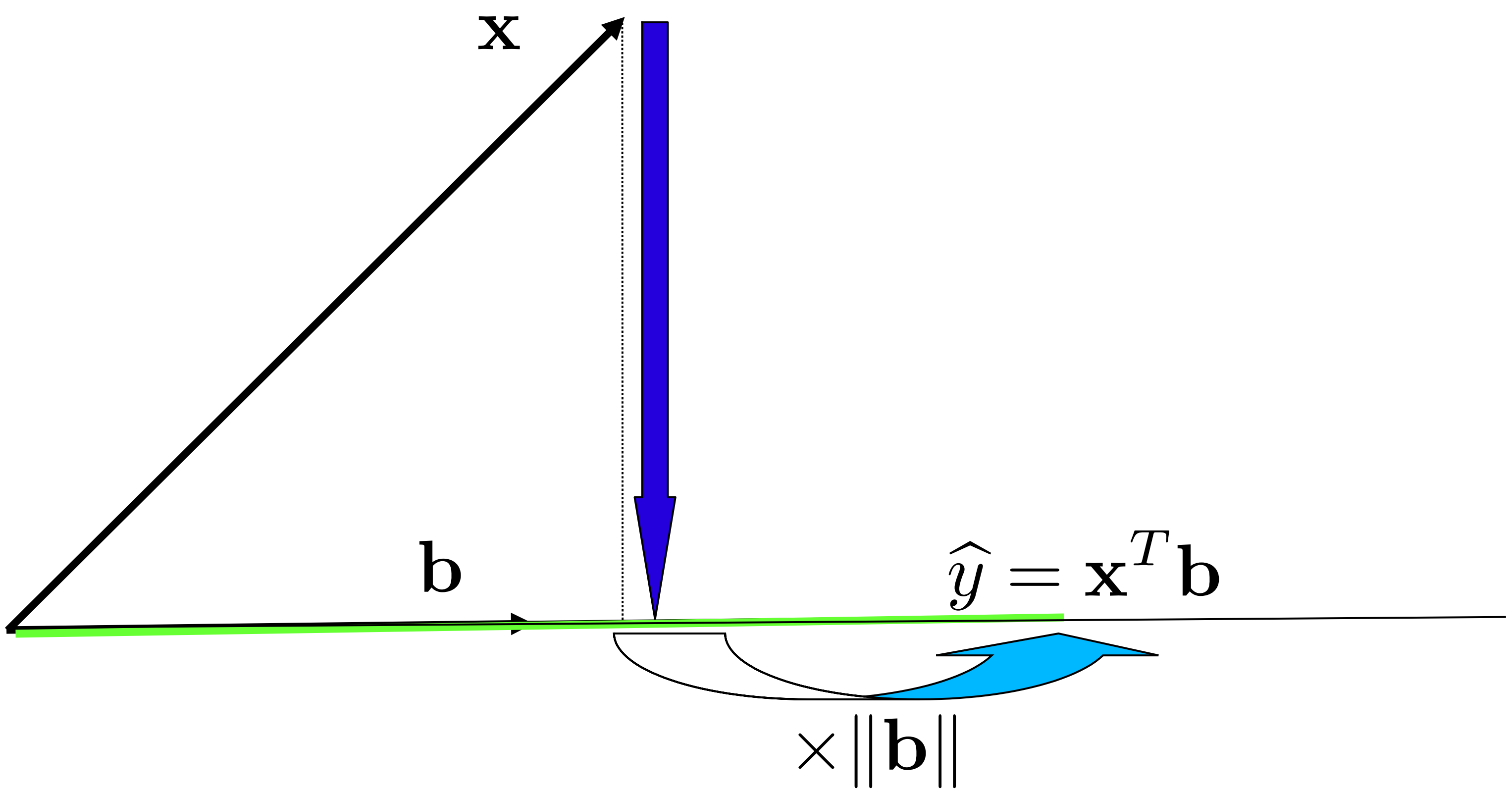
What's the problem?

Back to the problem: Let $\mathbf{b}$ be a model and $\mathbf{x}$ a spectrum (two vectors in $\mathbb{R}^p$)



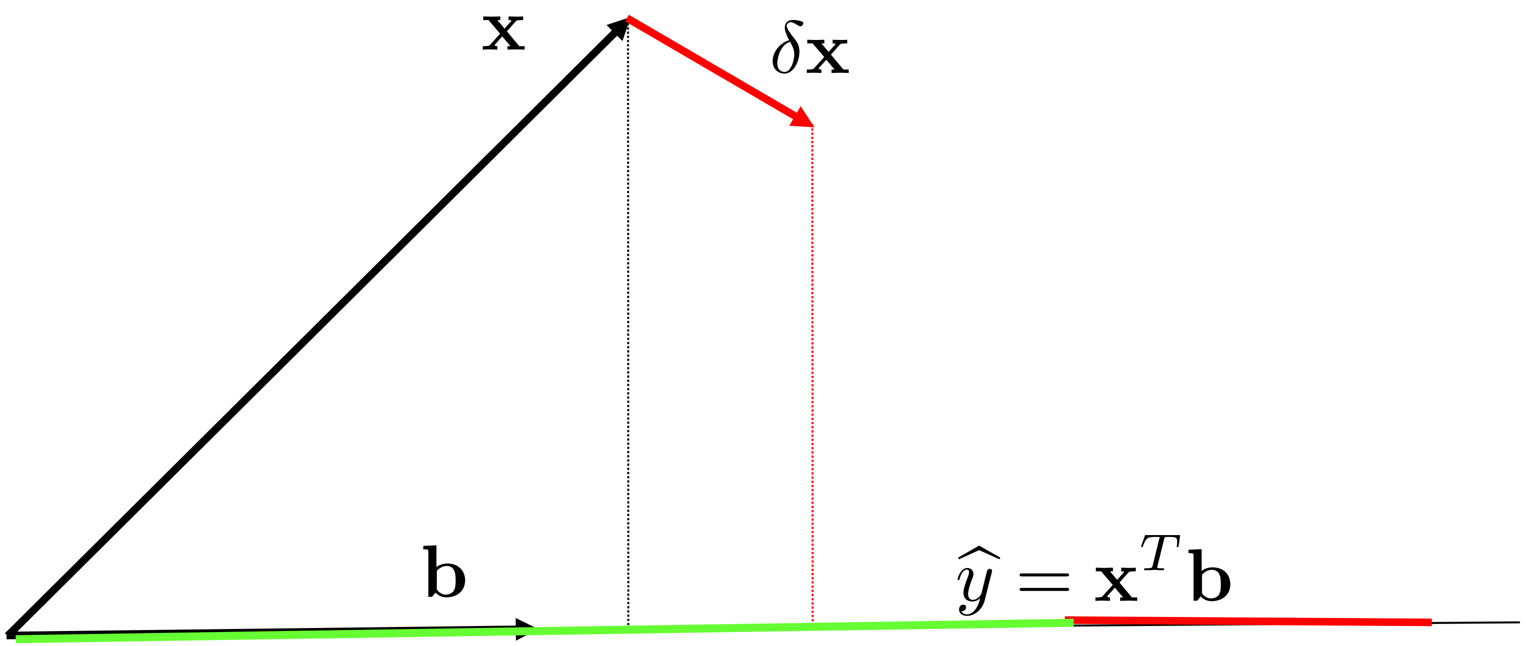
What's the problem?

Applying $\mathbf{b}$ on $\mathbf{x}$ relies on the inner product $\mathbf{x}^{\mathrm{T}}\mathbf{b}$



What's the problem?

Any deviation $\delta\mathbf{x}$ produces a error $\delta\hat{\mathbf{y}}$



$$\delta\hat{y} = \delta\mathbf{x}^T\mathbf{b}$$

$$|\delta\hat{y}| = \|\delta\mathbf{x}\| \times \|\mathbf{b}\| \times |\cos(\delta\mathbf{x}, \mathbf{b})|$$

How to improve the robustness

$$|\delta\widehat{y}| = \|\delta\mathbf{x}\| \times \|\mathbf{b}\| \times |\cos(\delta\mathbf{x}, \mathbf{b})|$$

- So, lowering $|\delta\hat{y}|$ can be done either by:

  - > lowering $\|\delta\mathbf{x}\|$ → controlling the perturbations, preprocessing, a priori correction

  - > lowering $\|\mathbf{b}\|$ → parsimonious modeling ($\|\mathbf{b}\|$ increases with #LV)
    variable selection ($\|\mathbf{b}\|$ increases with *p*)

  - > lowering $|\cos(\delta\mathbf{x}, \mathbf{b})|$ → implicit orthogonalization : model update, repfile
    explicit orthogonalization : EPO, TOP, DOP, etc.
    domain adaptation

Orthogonalisation based methods

Orthogonalisation based methods

- The aim of OP-based methods is to explicitly lower $|\cos(\delta\mathbf{x}, \mathbf{b})|$
- The idea consists in removing from $\mathbf{X}$ the space spanned by $\delta\mathbf{x}$
  - > A collection of $\delta\mathbf{x}$ is put into a matrix $\mathbf{D}$
  - > A basis $\mathbf{P}$ of the subspace spanned by $\delta\mathbf{x}$ is estimated by an SVD on $\mathbf{D}$
  - > $\mathbf{X}$ is projected orthogonally to $\mathbf{P}$: $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \mathbf{P}(\mathbf{P}^{\mathbf{T}}\mathbf{P})^{-1}\mathbf{P}^{\mathrm{T}})$
  - > A model is calibrated on $\tilde{\mathbf{X}}$; the resulting $\mathbf{b}$ is orthogonal to any $\delta\mathbf{x}$
- All the OP-based methods differ in the way $\mathbf{D}$ is built.

Orthogonalisation based methods

- Different ways of collecting the $\delta \mathbf{x}$
  - > With an experimental design: EPO
  - > With standard samples: TOP
  - > With repetitions: EROS
  - > Using some $\mathbf{y}_t$ values: DOP
- How to do when we have:
  - > Only spectra of the source and target domains
  - > $\mathbf{X_s}(n_s, p)$ and $\mathbf{X_t}(n_t, p)$, $n_s \neq n_t$

Unsupervised DOP

- Initial question: how to implement DOP without $\mathbf{y}_t$?

- Algorithm:

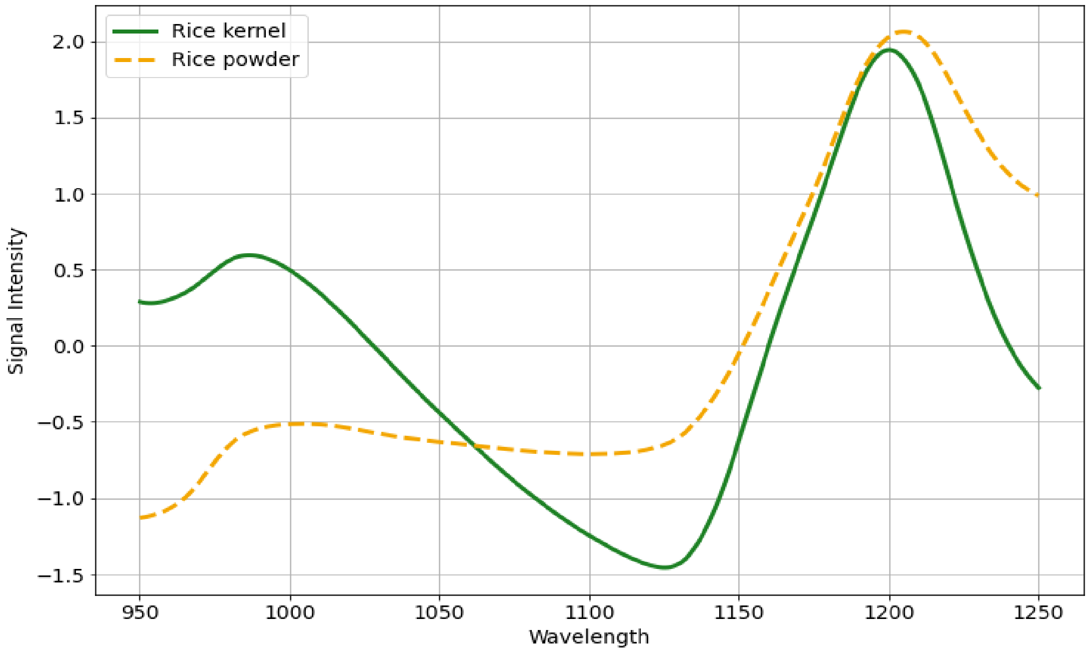
$$\mathbf{X}_s = \mathbf{U}_s\,\Delta_s \mathbf{V}_s^T$$

$$\mathbf{X}_t = \mathbf{U}_t\,\Delta_t \mathbf{V}_t^T = \mathbf{X}_t \mathbf{V}_t \mathbf{V}_t^T$$

$$\widehat{\mathbf{X}}_t = \mathbf{X}_t \mathbf{V}_t \mathbf{V}_s^T$$

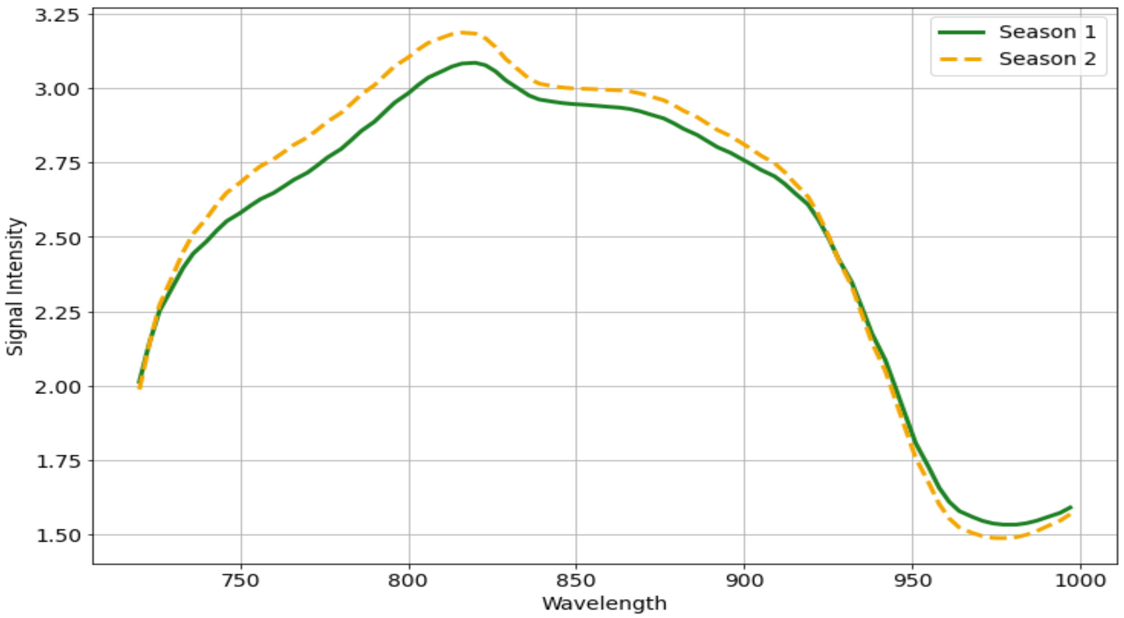
$$\mathbf{D} = \mathbf{X}_t - \widehat{\mathbf{X}}_t$$

uDOP case studies



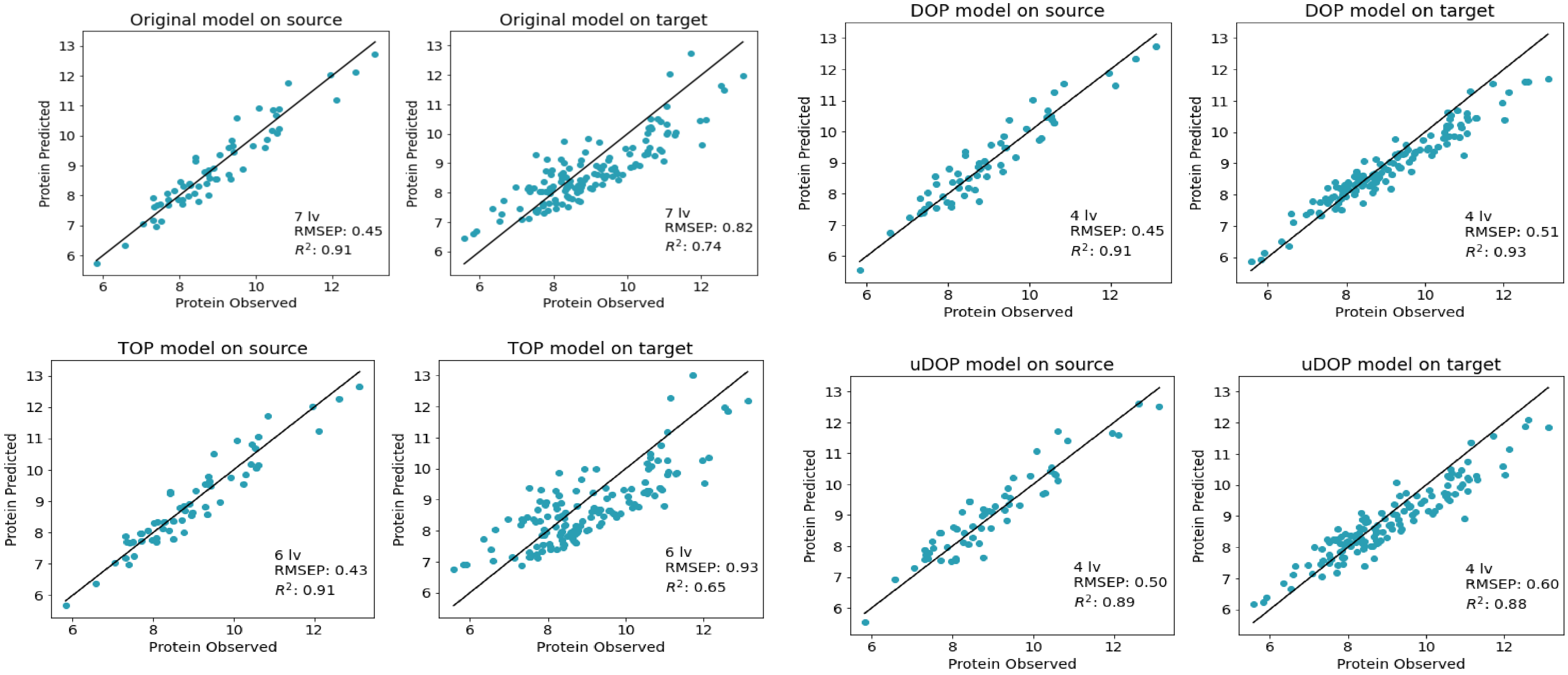
Rice: grain vs powder

Pears: season 1 vs season 2



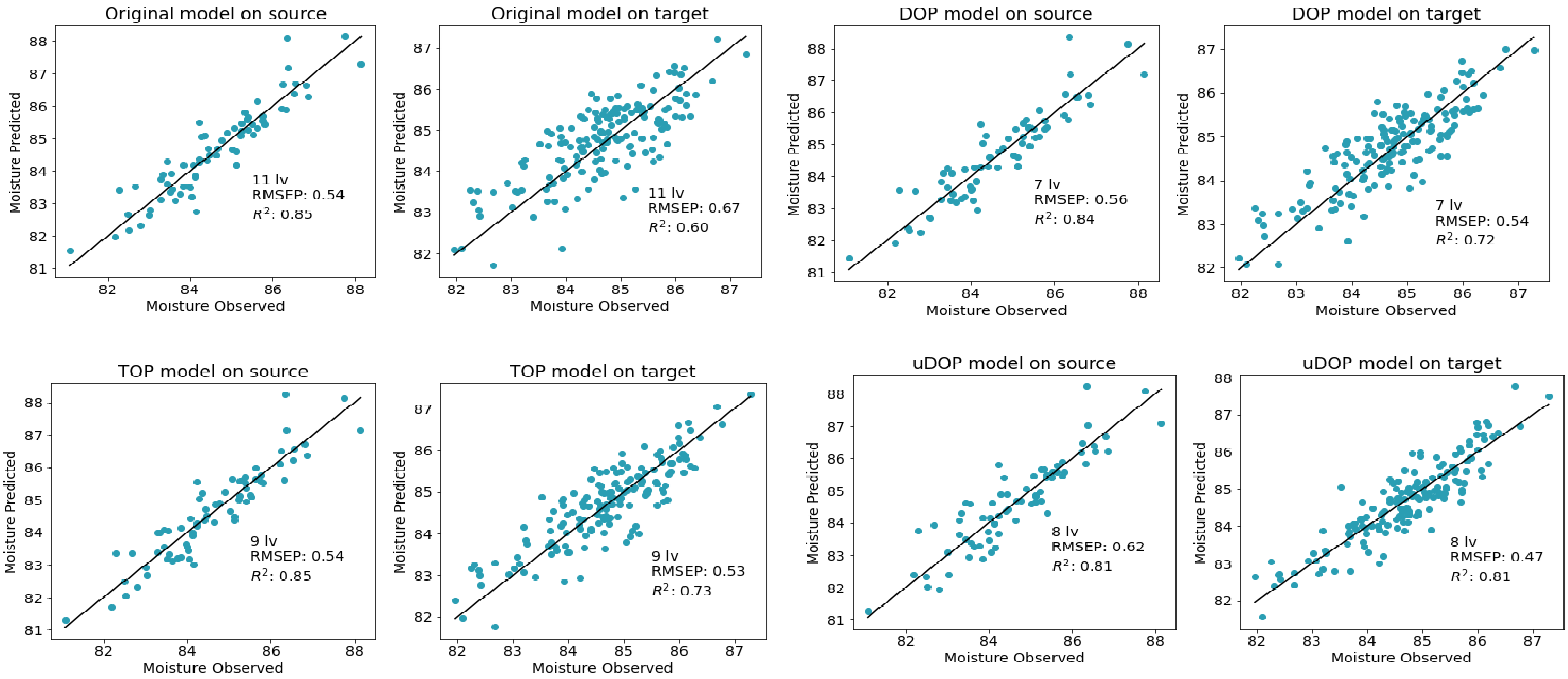
uDOP case studies

Rice



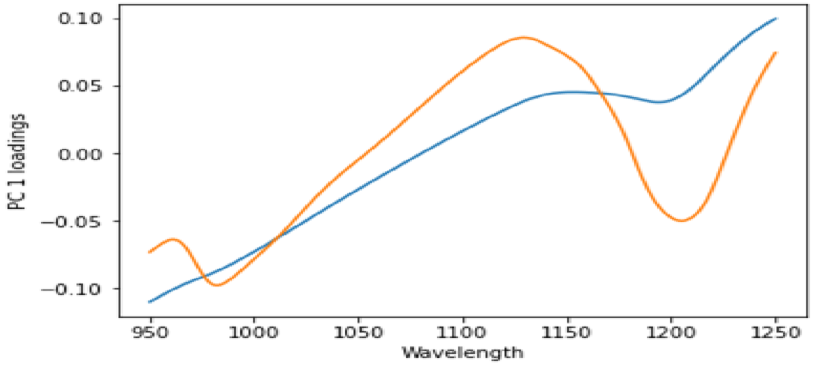
uDOP case studies

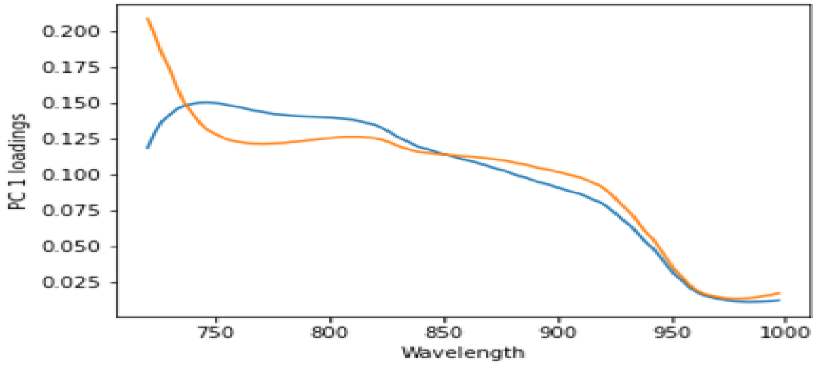
Pears



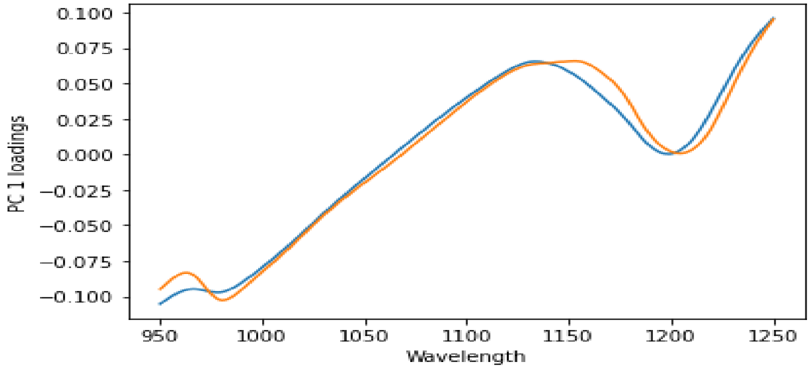
uDOP case studies

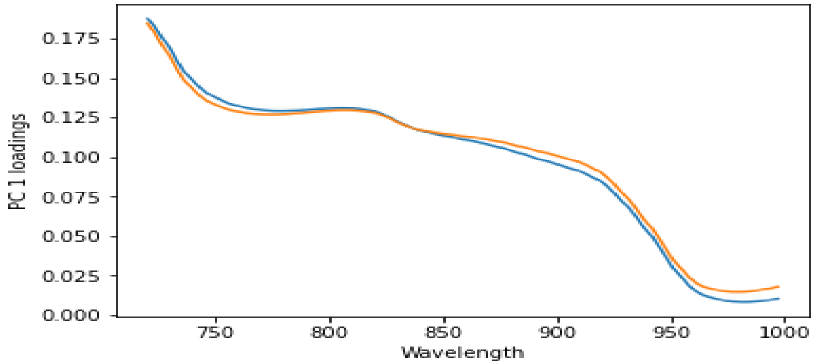
No orthogonalization



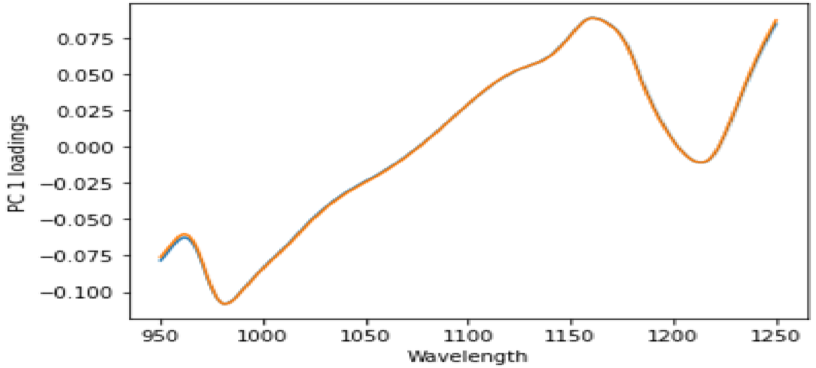


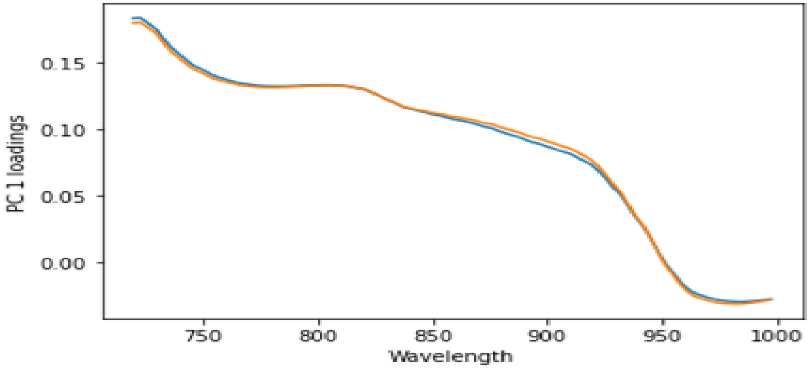
uDOP
1 comp





uDOP
2 comp





Conclusion

- Calibration transfer and maintenance is a widely studied issue
- A lot of methods, for a lot of situations
- uDOP offers a solution when only source and target spectra are available
- It eliminates the differences between the covariance structures