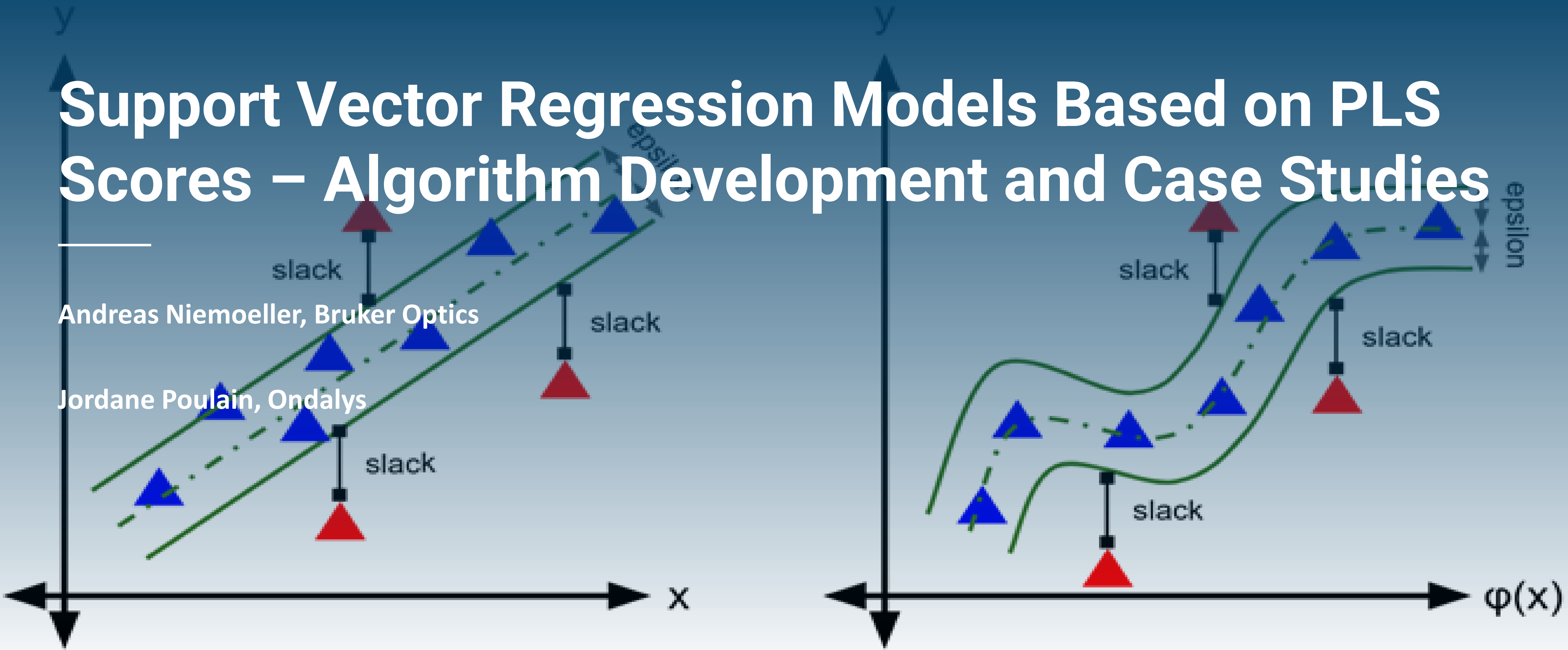
Support Vector Regression Models Based on PLS Scores – Algorithm Development and Case Studies

Andreas Niemoeller, Bruker Optics

Jordane Poulain, Ondalys



NIR and Modelling in Food, Feed and Agriculture

- Still increasing usage of NIR and other fast technologies
- NIR is an established and central element of quality control
- More applications and maintenance of existing methods
- new developing tasks: handheld and PAT
- Much more data available nowadays

- Networking of instruments helps to centralize expertise
- More calibration services available, especially in feed & agri

NIR and Modelling in Food, Feed and Agriculture

- Still increasing usage of NIR and other fast technologies
- NIR is an established and central element of quality control
- More applications and maintenance of existing methods
- new developing tasks: handheld and PAT
- Much more data available nowadays

- Networking of instruments helps to centralize expertise
- More calibration services available, especially in feed & agri

- Big companies have Center of Excellence, but efforts are huge
- Standard users have less time and expertise to cover the required tasks for modelling and validation

What do we need?

- Easier handling
- Simpler procedures
- (Semi-) automated steps
- New algorithms

Algorithms: Increasing Demands but Decreasing Expertise

Algorithms			Machine Learning & Artificial Intelligence		
Linear		Non-linear		Non-linear & complex	
Univariate Algorithms	Multivariate Algorithms PLS, PCA, Local Regression	Kernel Methods SVM etc.	Artificial Neural Networks ANN	ML Convolutional Neural Networks CNN	AI Recurrent Neural Networks RNN

Increasing:
required data, model parameters,
model complexity, calculation time, performance,
need for validation of robustness and performance

Decreasing: number of users with chemometric knowledge and expertise in general

Algorithms: Bruker's Choice for New Software Tool

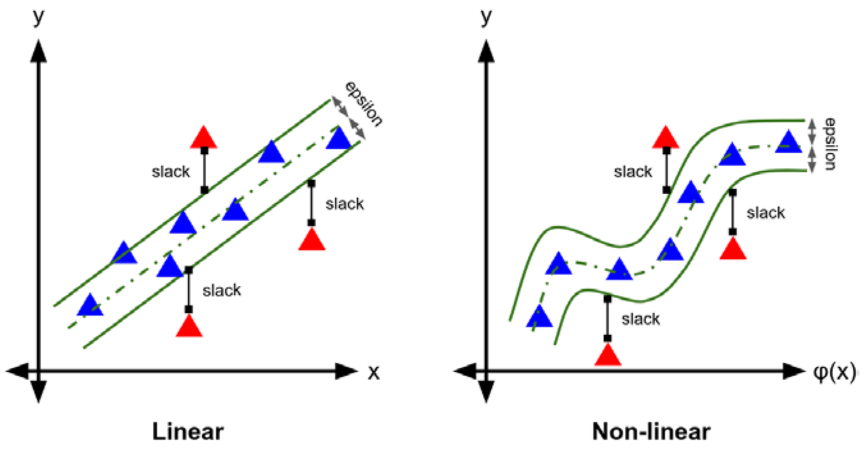
Algorithms			Machine Learning & Artificial Intelligence		
Linear		Non-linear		Non-linear & complex	
Univariate Algorithms	Multivariate Algorithms PLS Local Regression	Kernel Methods SVR	Artificial Neural Networks ANN	ML Convolutional Neural Networks CNN	AI Recurrent Neural Networks RNN

- New considered and tested approaches (Quant3)
  - Local Regression on transformed spectra
  - ε-SVR on PLS scores
- Features of Quant3
  - Automated handling of missing values
  - Kennard-Stone on reference values (managing missing values) or PLS scores
  - Improved Cal, TestSet and Validation set handling

Support Vector Regression (SVR)

Advantages

- Heterogeneous data sets can be handled over a broad calibration range or multiple product types
  - Possibly only one model instead of a few individual PLS models
- Handling of linear and non-linear data
- Deterministic models suitable even for Pharma industry (advantage over ANNs)
- Can work with same data set sizes as PLS
- Easy and fast prediction



Disadvantages

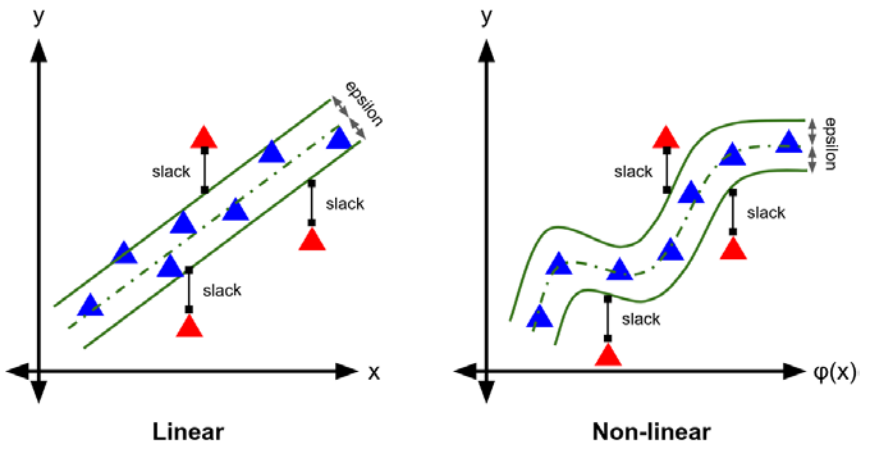
- Parameter grid search optimization takes time
- Absolute black box
  - New ways required to detect outliers

Support Vector Regression (SVR) on PLS Scores

Why PLS scores?

- Using just spectra for SVR is possible but takes time
- Different other latent variables work faster and better e.g., PCA, PLS or wavelet transformation
- NIR users know PLS and can review data and prepare models as today
- PLS scores represent nicely the component related variance
- PLS scores and loadings allow outlier detection for SVR predictions as today for PLS (Mahalanobis distance & spectral residuals)

➢ SVR can be a booster for users who work with PLS today



Comparison of PLS and ε-SVR on 2 Datasets

Algorithms

- PLS was performed with the Quant2 package of OPUS (Bruker Optics GmbH & Co. KG)
- ε-SVR is based on an implementation of LIBSVM* with RBF kernel

Sugar Factory Products

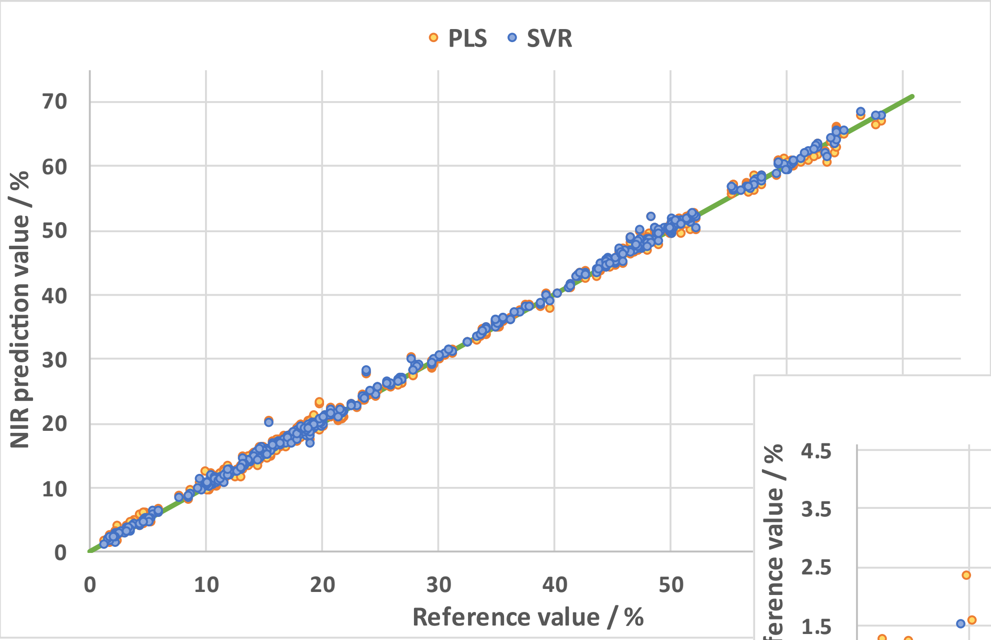
- Liquid products across the raw sugar production process: first juice, clarified juice, filtrate, last juice, lime juice, liquors, raw syrup
- Measurement in transflection, 2mm pathlength
  (integrating sphere, resolution 16cm^{-1}, 64 scans, 11500-4000cm^{-1})
- Parameters: °Brix, POL
- 900 calibration and 600 tuning spectra

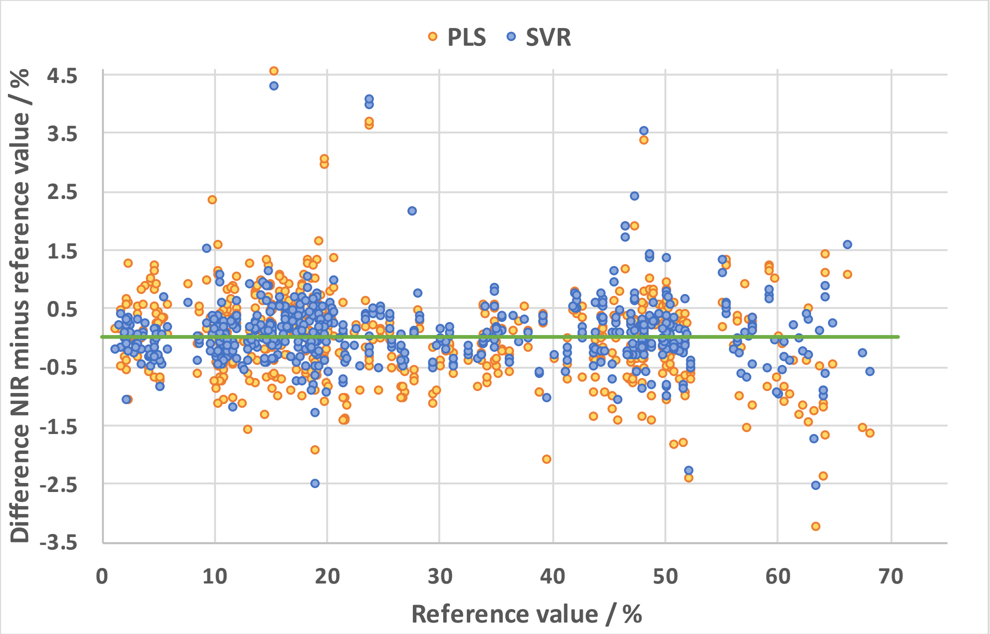
Feed dataset

- Various finished feeds: Poultry (Broiler, Chicken, Finisher, Layer, Turkey, Duck), Swine (Piglet, Swine) and Ruminant
- Measurement in reflection (integrating sphere, resolution 16cm^{-1}, 64 scans, 11500-4000cm^{-1})
- Parameters: protein, fat, moisture, ash
- 6537 calibration and 3119 tuning spectra

* LIBSVM -- A Library for Support Vector Machines, https://www.csie.ntu.edu.tw/~cjlin/libsvm/

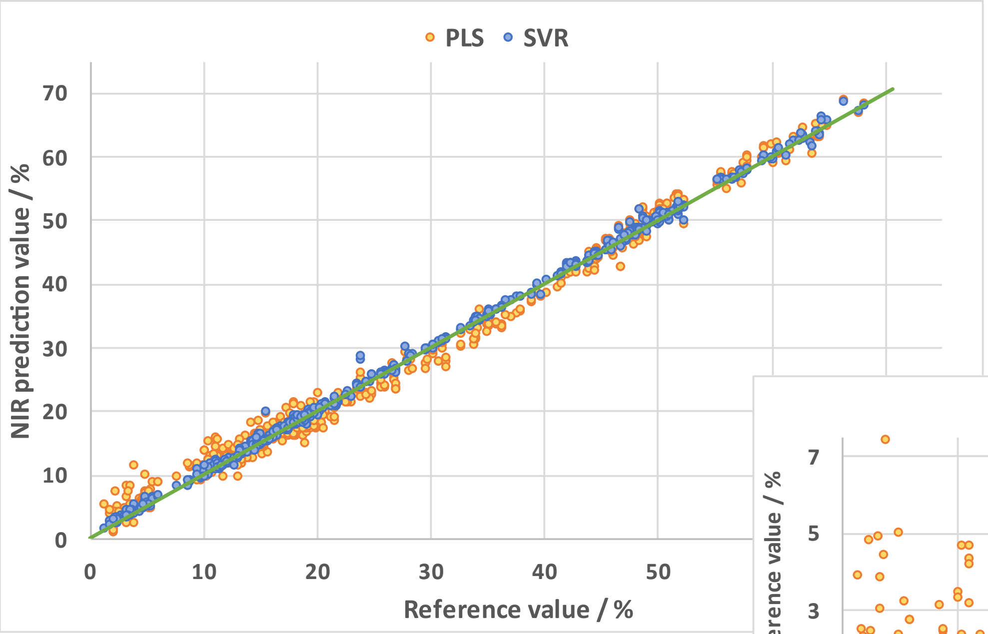
ε-SVR on PLS Scores: Sugar Products, Brix, Optimized PLS Model

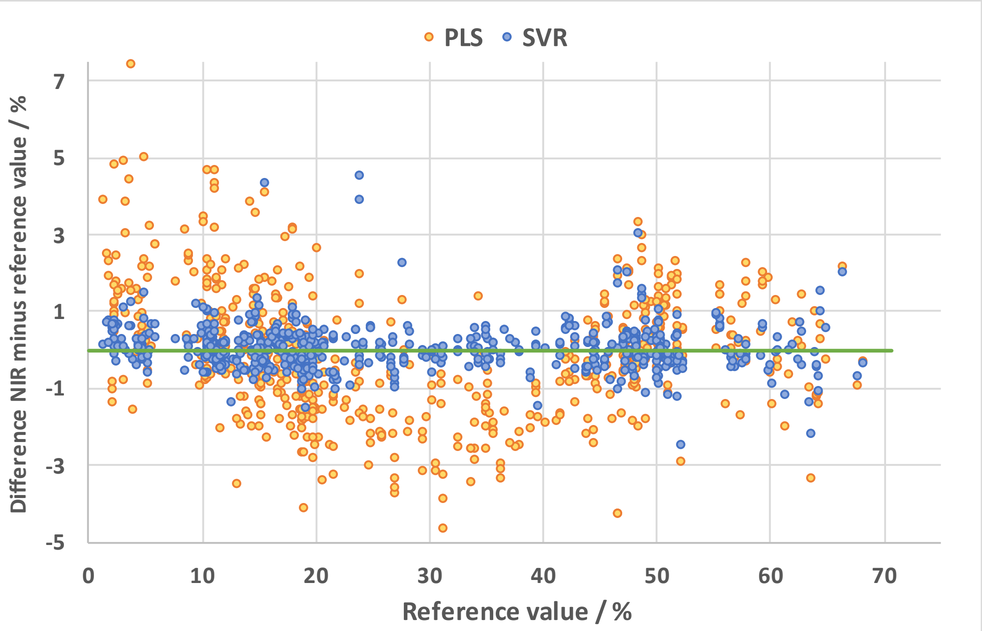




Mode	ε	C	Gamma	r	RMSEP	#SV
PLS				0.999	0.785	9
RBF	0.1	2000	21	0.999	0.593	696
RBF	0.1	1500	21	0.999	0.593	704
RBF	0.1	2500	21	0.999	0.593	702
RBF	0.1	1500	26	0.999	0.593	704
RBF	0.1	2000	26	0.999	0.594	700
RBF	0.1	2500	16	0.999	0.594	698
RBF	0.3	2000	26	0.999	0.594	431
RBF	0.1	2500	26	0.999	0.595	689
RBF	0.1	2000	16	0.999	0.595	711
RBF	0.1	1000	26	0.999	0.595	708
RBF	0.3	2500	26	0.999	0.595	433
RBF	0.3	1500	26	0.999	0.595	427
RBF	0.1	1000	21	0.999	0.596	717
RBF	0.1	1500	16	0.999	0.596	717
RBF	0.3	2500	21	0.999	0.597	427
RBF	0.3	2000	21	0.999	0.597	435
RBF	0.1	2500	11	0.999	0.598	715
RBF	0.3	1500	21	0.999	0.599	431
RBF	0.3	1000	26	0.999	0.599	438
RBF	0.1	2000	11	0.999	0.599	724

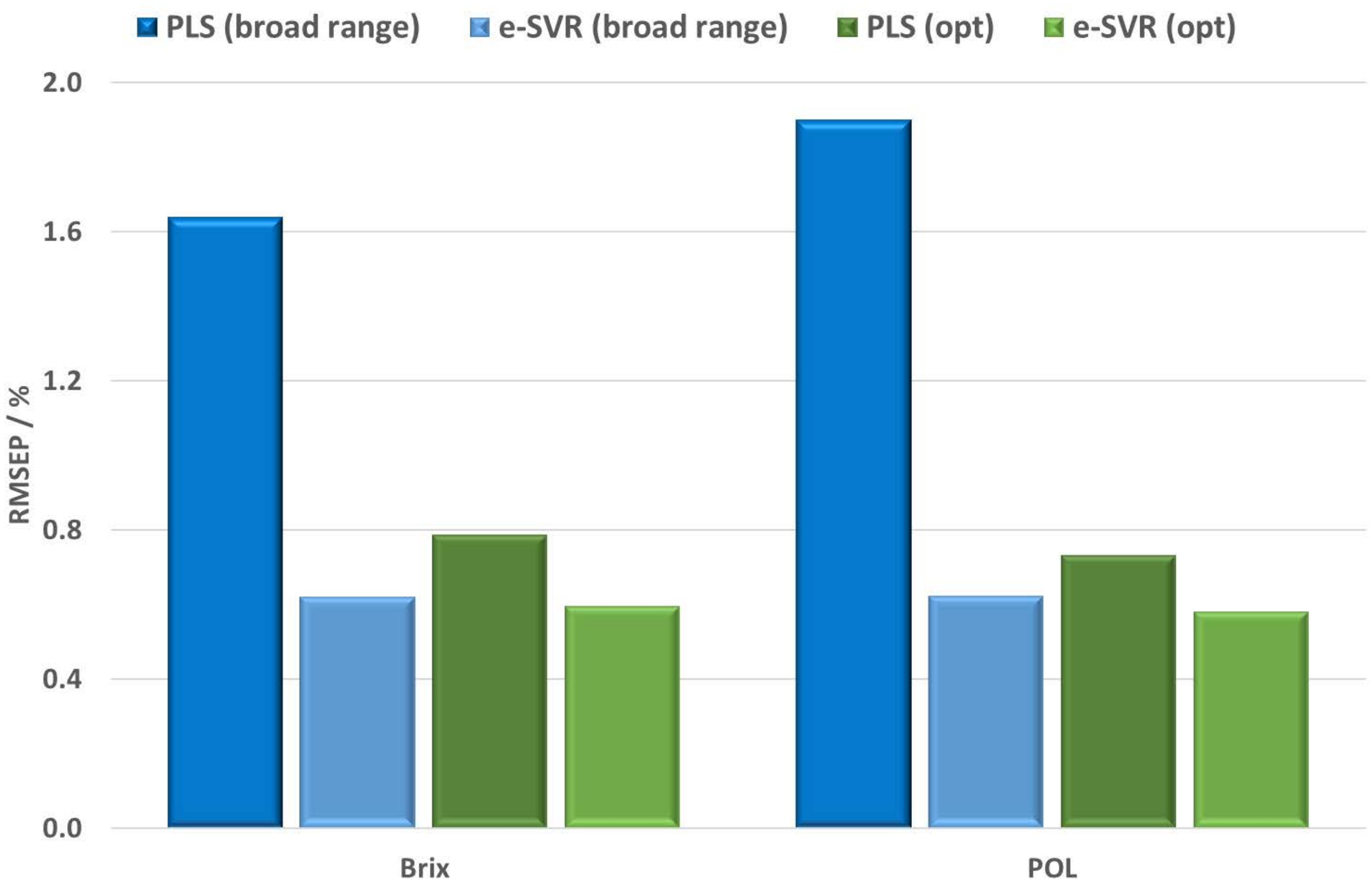
ε-SVR on PLS Scores: Sugar Products, Brix, Broad Spectral Range



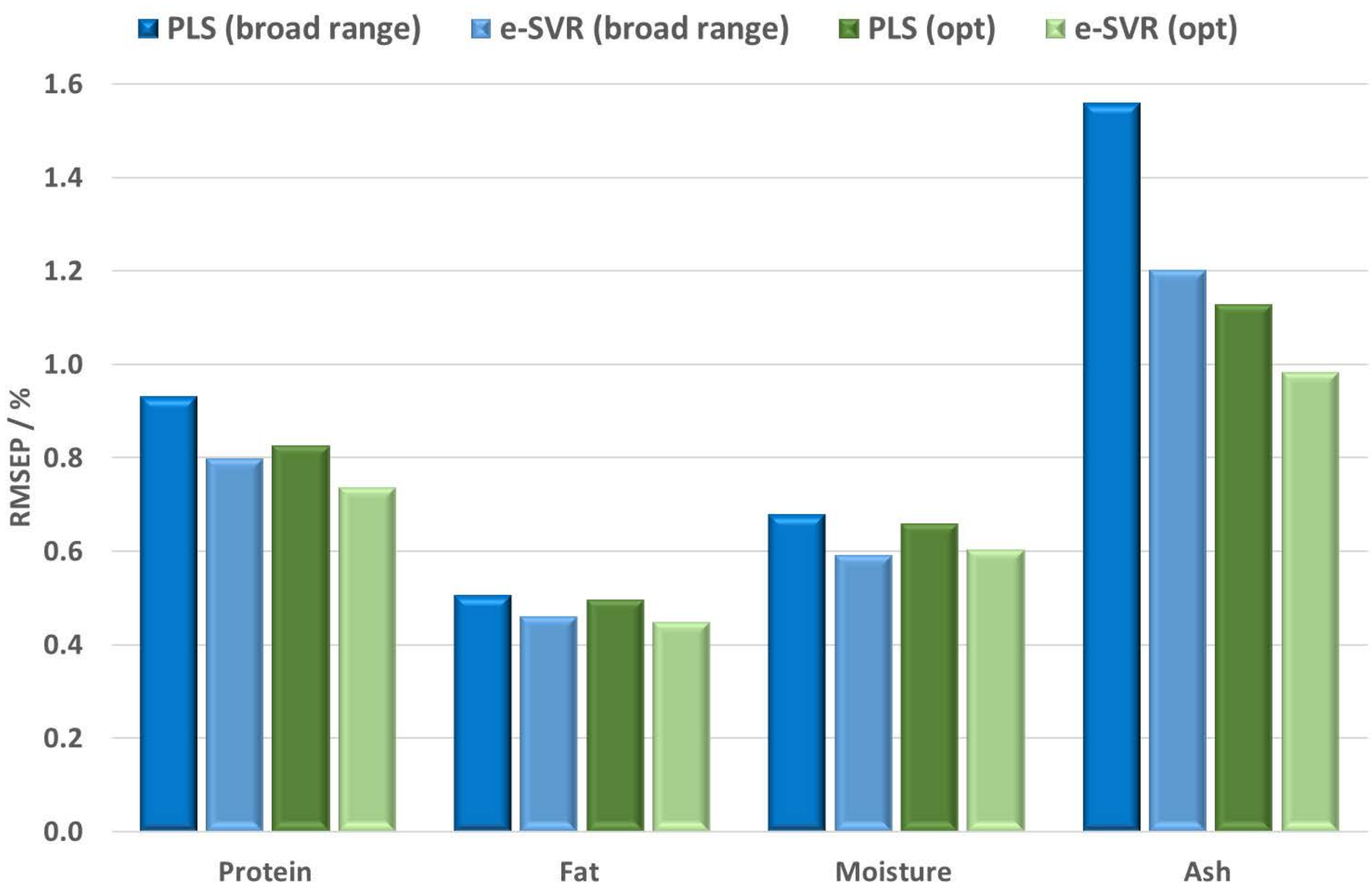


Mode	ε	C	Gamma	r	RMSEP	#SV
PLS				**0.9989**	**1.640**	**6**
RBF	0.1	500	6	0.9994	0.614	682
RBF	0.1	1000	6	0.9994	0.614	662
RBF	0.3	500	6	0.9994	0.618	415
RBF	0.1	1500	6	0.9994	0.621	672
RBF	0.3	1000	6	0.9994	0.623	411
RBF	0.1	2000	6	0.9994	0.627	682
RBF	0.3	500	11	0.9994	0.631	434
RBF	0.1	500	11	0.9994	0.631	684
RBF	0.1	2500	6	0.9994	0.632	677
RBF	0.3	1500	6	0.9994	0.634	418
RBF	0.1	1000	11	0.9994	0.641	676
RBF	0.5	500	6	0.9994	0.641	285
RBF	0.3	2000	6	0.9994	0.645	428
RBF	0.1	2500	1	0.9994	0.648	730
RBF	0.1	500	16	0.9994	0.651	687
RBF	0.5	1000	6	0.9993	0.654	309
RBF	0.3	2500	6	0.9993	0.655	438
RBF	0.1	2000	1	0.9993	0.657	737
RBF	0.1	1500	11	0.9993	0.658	689
RBF	0.3	1000	11	0.9993	0.659	446

ε-SVR on PLS Scores: Sugar Products, Brix and POL



ε-SVR on PLS Scores: 3 Feed Types (Poultry, Swine, Ruminant)



Conclusions

- ε-SVR model based on PLS scores is a good alternative to PLS
  - ➢ Do not need to be completely optimized in terms of data pre-treatment and spectral regions to achieve good performance contrary to PLS
- Even with data sets that can be easily calibrated with PLS today, this results in a simpler approach with very good results
- For complex data sets with high spectral variances, combinations of different products or raw materials and very large value ranges, the advantage for SVR with PLS scores is even greater in terms of effort and results
- ε-SVR based on PLS scores will be part of a new Bruker calibration package: Quant3

References

- LIBSVM -- A Library for Support Vector Machines, https://www.csie.ntu.edu.tw/~cjlin/libsvm/
- How to Apply Support Vector Machines (SVM) to predict spectroscopic data - Comparison of SVM and PLS regression models, White paper, Ondalys, http://www.ondalys.fr, 2020

Acknowledgements

- This scientific study was made by Bruker Optics in collaboration with Ondalys (A white paper is on the way)

Ondalys
4 rue Georges Besse
34 830 Clapiers
France

+33 (0)4 67 67 97 87
contact@ondalys.fr
www.ondalys.fr

BRUKER

Innovation with Integrity