Technical Touch Point:

SVM Regression For Online RAMAN

A study in reducing overfitting and improving signal to noise

PLS vs SVM(L1) regression

PLS approach employs a decomposition of X and Y as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{R}_1 = \sum \mathbf{t}_h \mathbf{p}'_h + \mathbf{R}_1 \qquad \mathbf{U} = \mathbf{T}\mathbf{P}^{\mathrm{T}}\mathbf{B}(\mathbf{Q}^{\mathrm{T}})^{-1} \qquad \mathbf{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \qquad cov(\mathbf{U}^{\mathrm{T}}\mathbf{T})$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{R}_2 = \sum \mathbf{u}_h \mathbf{q}'_h + \mathbf{R}_2 \qquad \mathbf{U} \triangleq \mathbf{B}\mathbf{T} \text{ or } \hat{\mathbf{u}}_h = \mathbf{b}_h \mathbf{t}_h \qquad \mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}$$

Moore-Penrose pseudo-inverse (regressor matrix in OLS), we can solve for U algebraically. This yields multiple solutions, so a relational restriction is applied to reduce the solution space and maximize covariance. The resulting equation estimates Y.

PLSR uses singular value decomposition over X = TPT to minimize the Frobenius norm (L2-like). Both OLS and PLS methods rely on L2 norms for regression. Though PLS reduces variable count, both methods are prone to overfitting, potentially introducing noise in the prediction of Y.

SVM(L1) approach employs a L1 Norm with a “slack” threshold:

PLS minimize residuals based on an L2-like Frobenius norm, focusing on square error reduction. SVR differs by aiming to minimize error only beyond a specific threshold (±ε). The objective function includes additional parameters (ξ) to handle this multi-objective optimization.

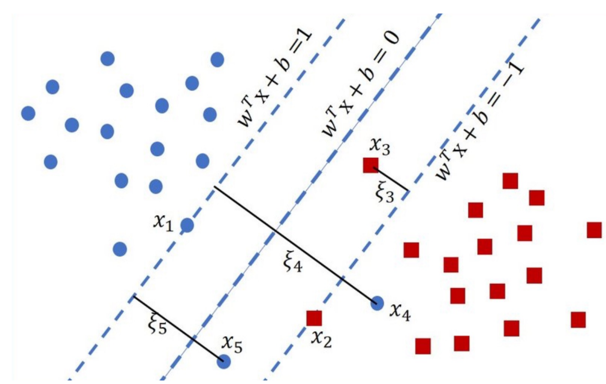
SVR establishes a hyperplane with a 'slack' boundary to accommodate deviations within the defined threshold.

The SVM objective:

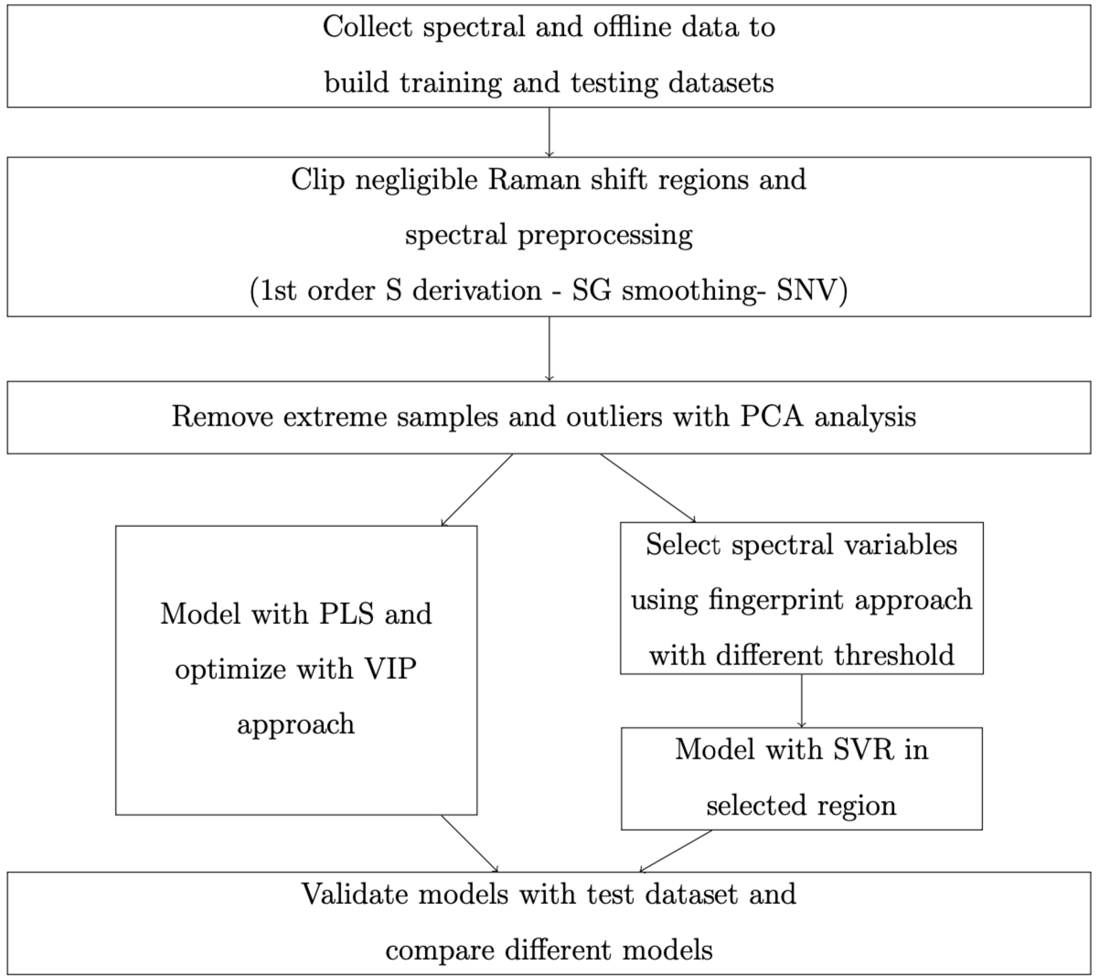
$$\Phi(w, \xi^*, \xi) = \frac{1}{2}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}) + C\left(\sum_{i=1}^{\ell} \xi_i^* + \sum_{i=1}^{\ell} \xi_i\right)$$

$$minimize\ [\ \Phi(\boldsymbol{w}, \xi^*, \xi)\ ]$$

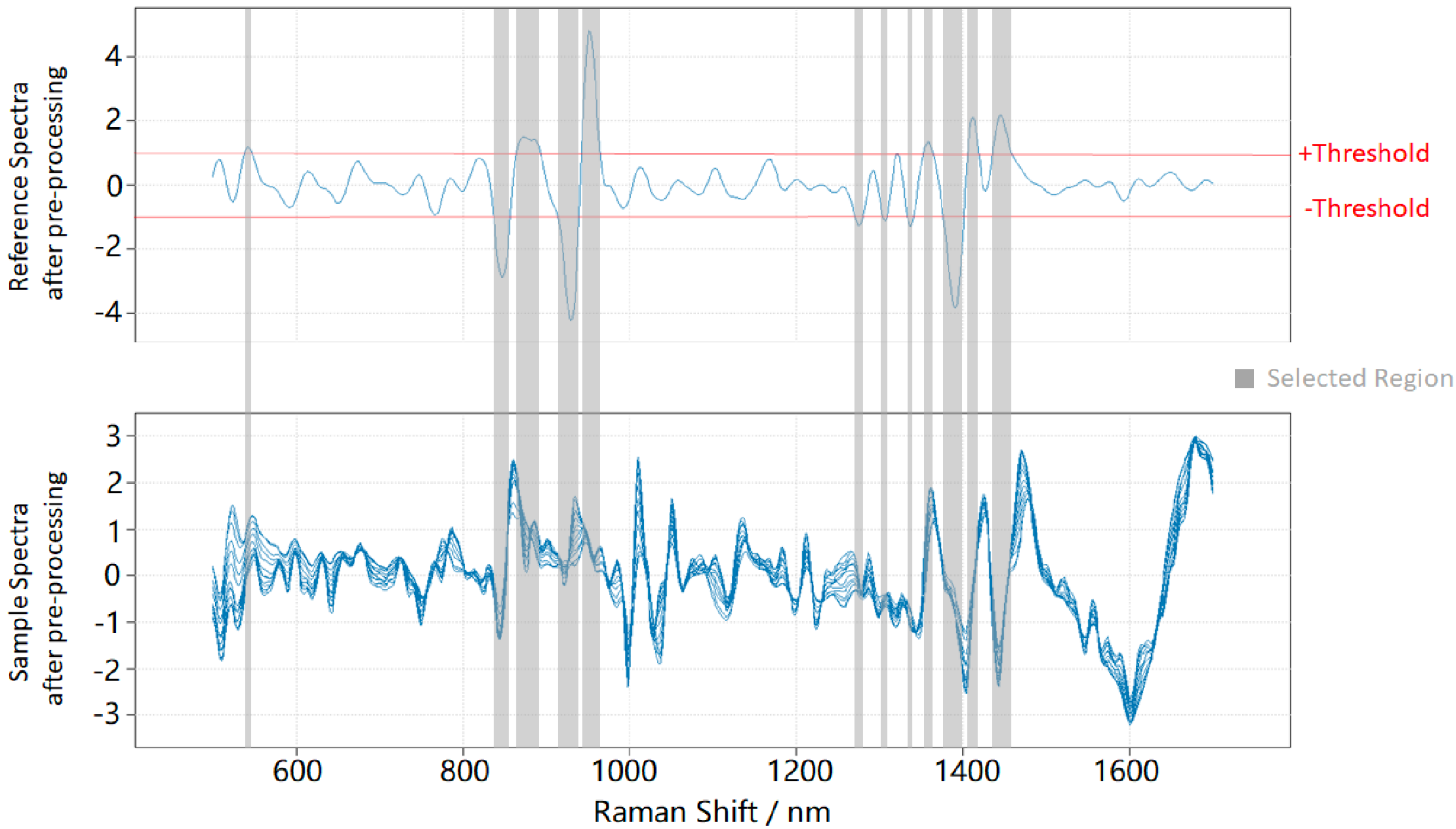
$$\begin{aligned} y_i - (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i) - b &\le \varepsilon + \xi_i^*, & i = 1, \ldots, \ell \\ (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i) + b - y_i &\le \varepsilon + \xi_i, & i = 1, \ldots, \ell \\ \xi_i^* &\ge 0, & i = 1, \ldots, \ell \\ \xi_i &\ge 0, & i = 1, \ldots, \ell \end{aligned}$$



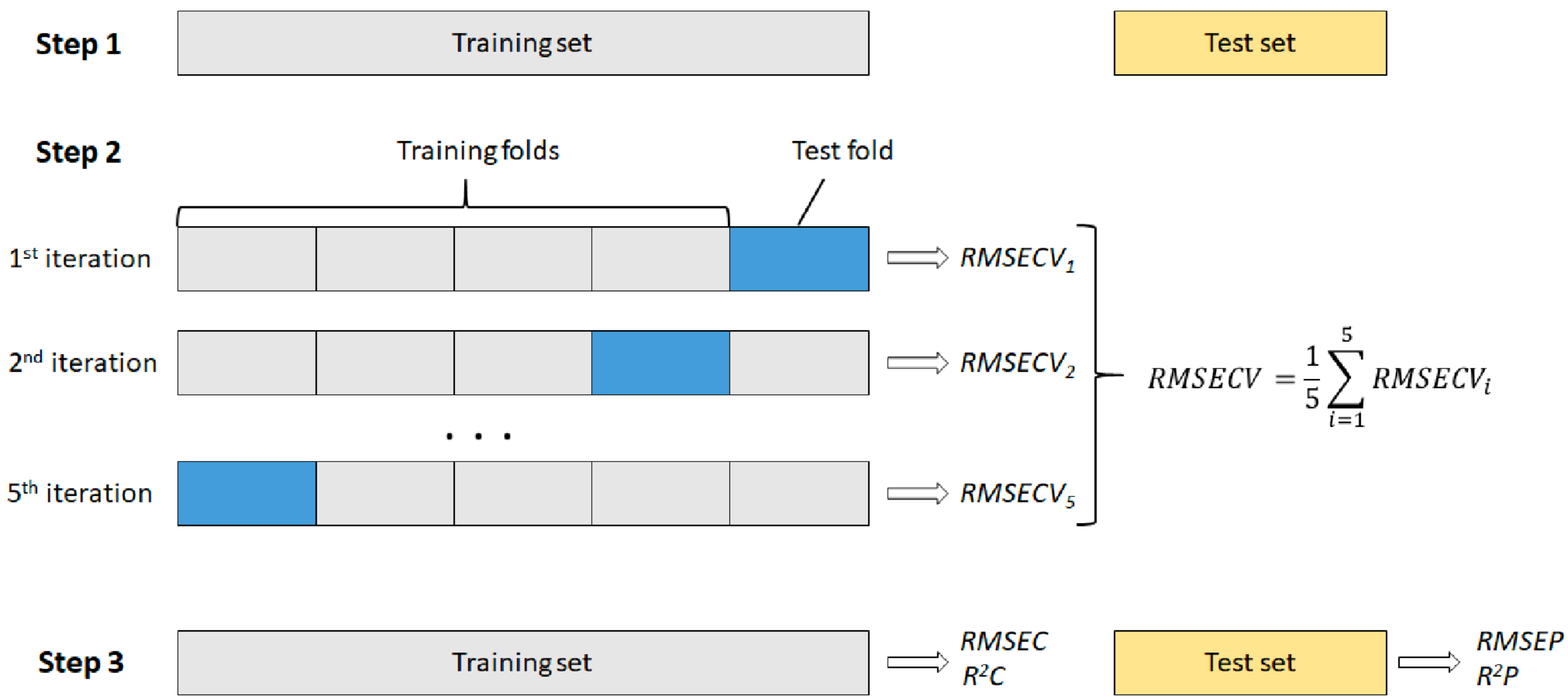
Raman PLS vs SVM Regression Procedure



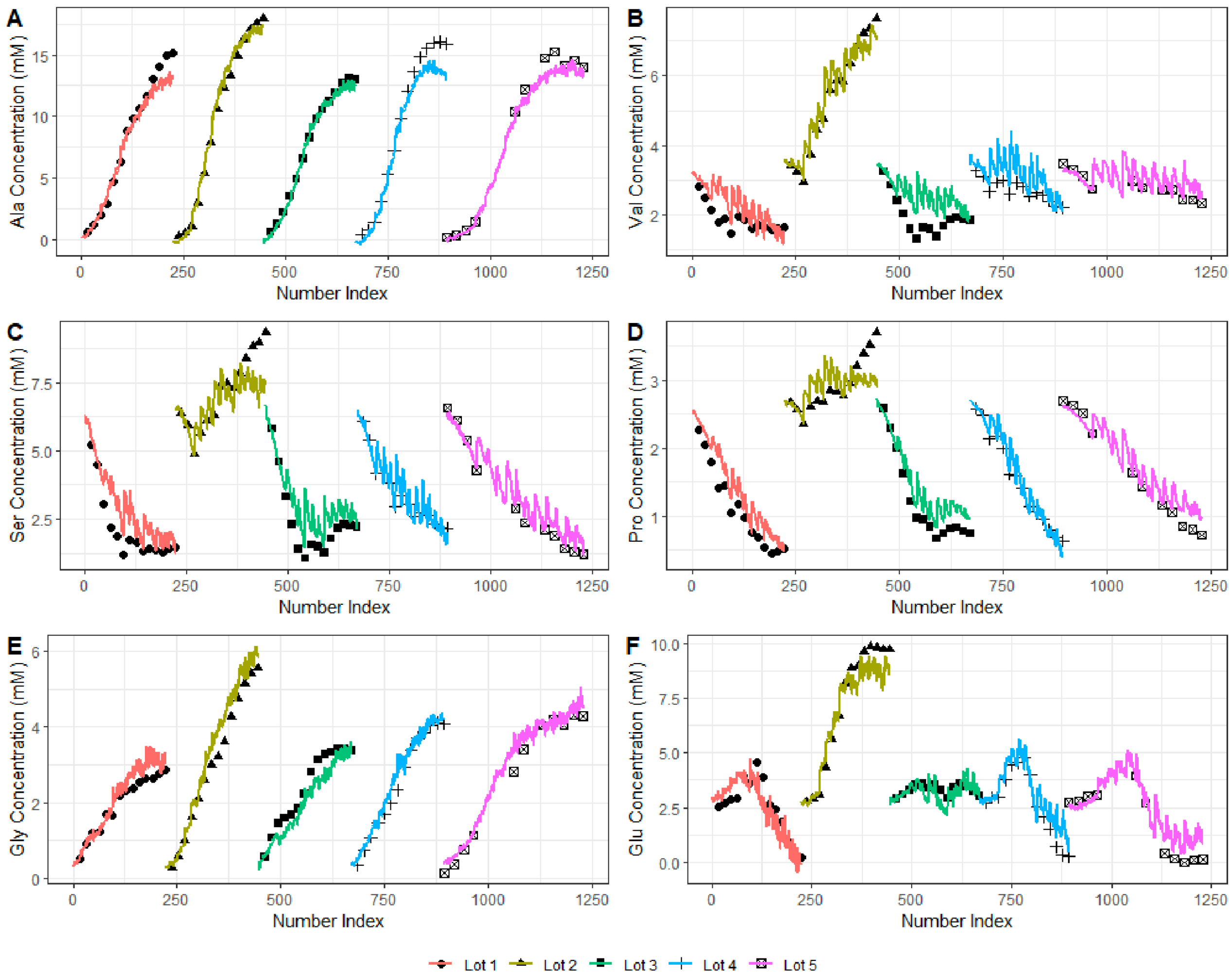
Raman Signal Windowing for SVM Regression



Raman SVM Regression: Training and Test Sets



Raman SVM Regression: Regression Performance



Outcome PLS vs SVM Regression: Comparable and potentially superior !

Analytes		Alanine	Valine	Serine	Proline	Glycine	Glutamate
Ranges	(mM)	0.15 – 19.99	1.79 – 8.24	1.00–9.90	0.59 – 4.34	0.12 – 6.92	0.10 – 9.33
SVR	Variable No.	477	89	177	506	143	215
	R^2C (%)	99.5	97.6	98.1	98.4	97.3	98.9
	R^2P (%)	97.8	94.1	90	93.6	95.9	92.5
	RMSEC (mM)	0.45	0.18	0.23	0.09	0.26	0.22
	RMSEP (mM)	0.82	0.33	0.6	0.2	0.3	0.6
PLS	Variable No.	359	316	359	339	361	295
	R^2C (%)	99.5	97.9	97.7	98.6	99	93.5
	R^2P (%)	98.5	97.7	92.2	96.4	91.9	69.3
	RMSEC (mM)	0.44	0.17	0.25	0.09	0.16	0.54
	RMSEP (mM)	0.7	0.23	0.65	0.17	0.39	1.4