

Table S1. Summary of sample preparation, contamination levels, and dataset allocation.

Sample class	Contamination level (% w/w)	Sample preparation strategy	Number of independent samples	Calibration set	Validation set
Gluten-free flour control	0%	Gluten-free flour without allergen addition	66	53	13
Wheat contamination	1-25%	Two independent mixtures prepared at each contamination level (50 samples); additional two independent mixtures prepared at selected levels (16 samples)	66	53	13
Peanut contamination	1-25%	Two independent mixtures prepared at each contamination level (50 samples); additional two independent mixtures prepared at selected levels (16 samples)	66	53	13
Soybean contamination	1-25%	Two independent mixtures prepared at each contamination level (50 samples); additional two independent mixtures prepared at selected levels (16 samples)	66	53	13
Sesame contamination	1-25%	Two independent mixtures prepared at each contamination level (50 samples); additional two independent mixtures prepared at selected levels (16 samples)	66	53	13
Total	0-25%	—	330	265	65

Each contamination level ranged from 1% to 25% (w/w) at 1% intervals.

Selected contamination levels included 3%, 6%, 9%, 12%, 15%, 18%, 21%, and 24% (w/w).

Table S2. Confusion matrices of the SIMCA model for calibration/training and validation sets
(A) Calibration/training set (n = 53 samples per class)

Actual contamination class / Predicted contamination class	Control	Wheat	Peanut	Soybean	Sesame
Control	53	0	0	0	0
Wheat	0	53	0	0	0
Peanut	0	0	53	0	0
Soybean	0	0	0	53	0
Sesame	0	0	0	0	51

(B) Validation set (n = 13 samples per class)

Actual contamination class / Predicted contamination class	Control	Wheat	Peanut	Soybean	Sesame
Control	13	0	0	0	0
Wheat	0	13	0	0	0
Peanut	0	0	13	0	0
Soybean	0	0	0	13	0
Sesame	0	0	0	0	13

Table S3. Class-specific performance metrics of the SIMCA model for calibration/training and independent validation sets.

Dataset	Class	Sensitivity (%)	Specificity (%)	Precision (%)	Balanced accuracy (%)
Calibration/training	Control	100	100	100	100
Calibration/training	Wheat	100	100	100	100
Calibration/training	Peanut	100	100	100	100
Calibration/training	Soybean	100	100	100	100
Calibration/training	Sesame	100	100	100	100
Validation	Control	100	100	100	100
Validation	Wheat	100	100	100	100
Validation	Peanut	100	100	100	100
Validation	Soybean	100	100	100	100
Validation	Sesame	100	100	100	100

Table S4. Statistical ANOVA analysis of PLSR models with different spectral preprocessing strategies based on cross-validation after mean-centering.

Cross-validation method	Allergen	Models compared	ANOVA F-value	p-value ($\alpha = 0.05$)	Holm-adjusted post-hoc comparison
LOOCV	Wheat	2nd derivative; Normalization; 2nd derivative + normalization; 2nd derivative + smoothing	0.556	0.645	None
LOOCV	Peanut	2nd derivative; Normalization; 2nd derivative + normalization; 2nd derivative + smoothing	3.635	0.014	None
LOOCV	Soybean	2nd derivative; Normalization; 2nd derivative + normalization; 2nd derivative + smoothing	0.790	0.501	None
LOOCV	Sesame	2nd derivative; Normalization; 2nd derivative + normalization; 2nd derivative + smoothing	0.517	0.671	None
Leave-10-out CV	Wheat	2nd derivative; Normalization; 2nd derivative + normalization; 2nd derivative + smoothing	0.577	0.631	None
Leave-10-out CV	Peanut	2nd derivative; Normalization; 2nd derivative + normalization; 2nd derivative + smoothing	0.198	0.898	None
Leave-10-out CV	Soybean	2nd derivative; Normalization; 2nd derivative + normalization; 2nd derivative + smoothing	1.321	0.270	None
Leave-10-out CV	Sesame	2nd derivative; Normalization; 2nd derivative + normalization; 2nd derivative + smoothing	0.577	0.631	None

Note: The four PLSR models compared in each analysis used the following spectral preprocessing strategies, each applied after mean-centering: second derivative, normalization, second derivative combined with normalization, and second derivative combined with smoothing. Statistical comparisons were performed using one-way repeated-measures ANOVA based on sample-level cross-validated absolute prediction errors ($|\text{predicted} - \text{measured}|$). Statistical significance was evaluated at $\alpha = 0.05$. Pairwise comparisons were performed using Holm-adjusted correction for multiple comparisons. “None” indicates that no pairwise comparison remained statistically significant after correction.

Figure S1. Mahalanobis distance-based outlier diagnosis for the SIMCA model of sesame-containing samples. The vertical green line represents the critical Mahalanobis distance (MDcrit) at the 95% confidence level. Samples located beyond this boundary were identified as potential outliers.

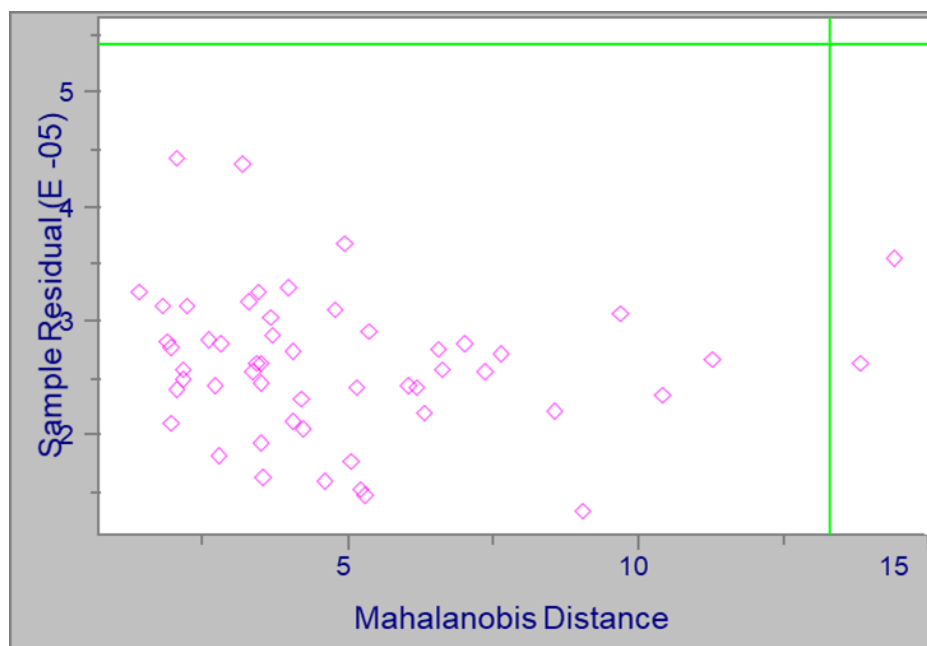


Figure S2. Selection of the optimal number of latent variables by leave-one-out cross-validation (LOOCV) and corresponding regression vectors for the soybean PLSR models using different spectral preprocessing strategies, each applied after mean-centering. For each preprocessing strategy, the three panels from left to right present the RMSECV plot against the number of latent variables (factors), the corresponding explained variance and statistical performance, and the regression vector, respectively. The optimal number of latent variables is highlighted in gray. Panels A–C correspond to second derivative preprocessing (35-point window), panels D–F to normalization, panels G–I to second derivative (35-point window) combined with normalization, and panels J–L to second derivative (35-point window) combined with smoothing (35-point window). Note: $r_{\text{Cal}}=1.00$ represents a rounded value; the actual value was slightly below 1.00.

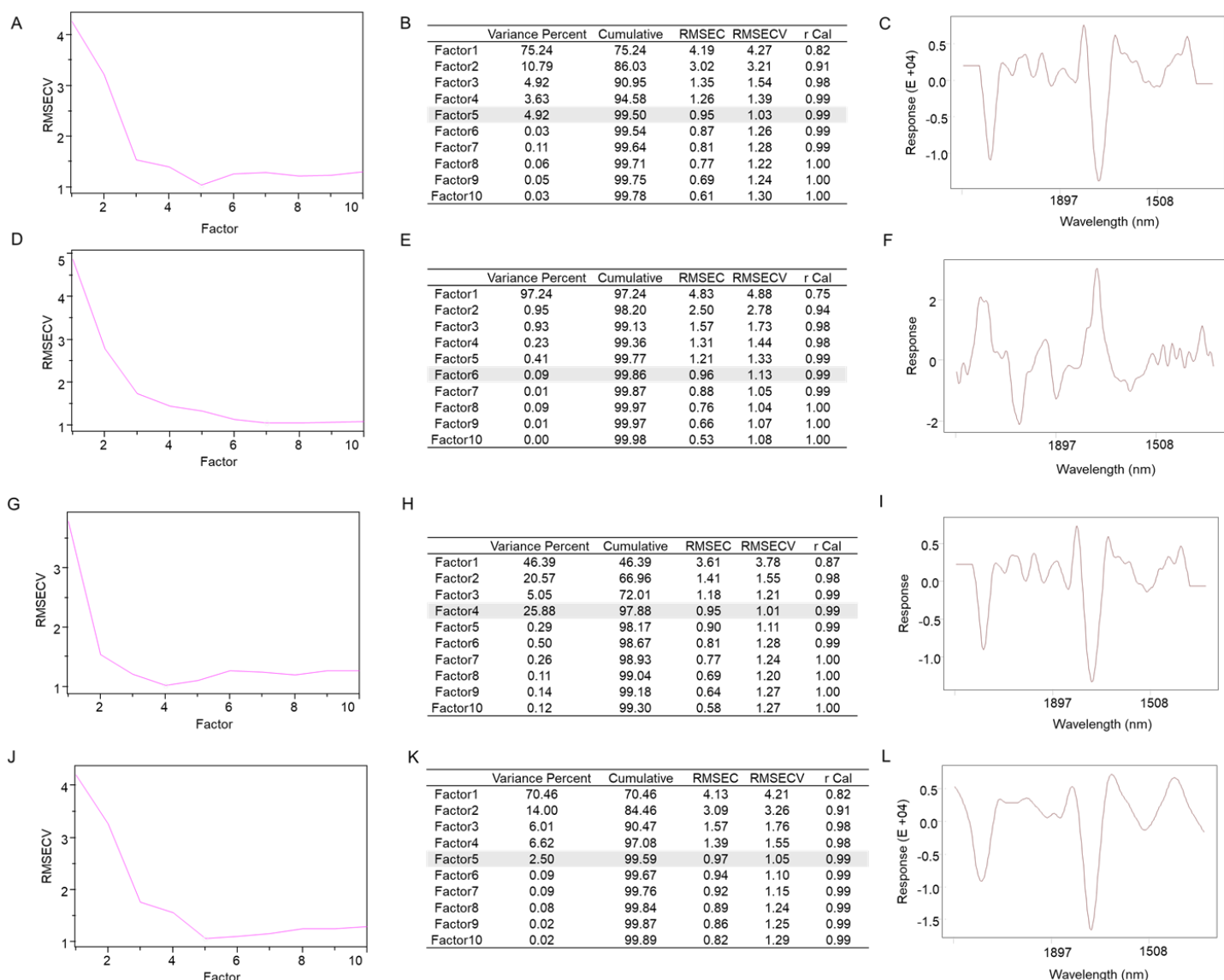


Figure S3. Selection of the optimal number of latent variables by leave-one-out cross-validation (LOOCV) and corresponding regression vectors for the peanut PLSR models using different spectral preprocessing strategies, each applied after mean-centering. For each preprocessing strategy, the three panels from left to right present the RMSECV plot against the number of latent variables (factors), the corresponding explained variance and statistical performance, and the regression vector, respectively. The optimal number of latent variables is highlighted in gray. Panels A–C correspond to second derivative preprocessing (35-point window), panels D–F to normalization, panels G–I to second derivative (35-point window) combined with normalization, and panels J–L to second derivative (35-point window) combined with smoothing (35-point window). Note: $r_{\text{Cal}}=1.00$ represents a rounded value; the actual value was slightly below 1.00.

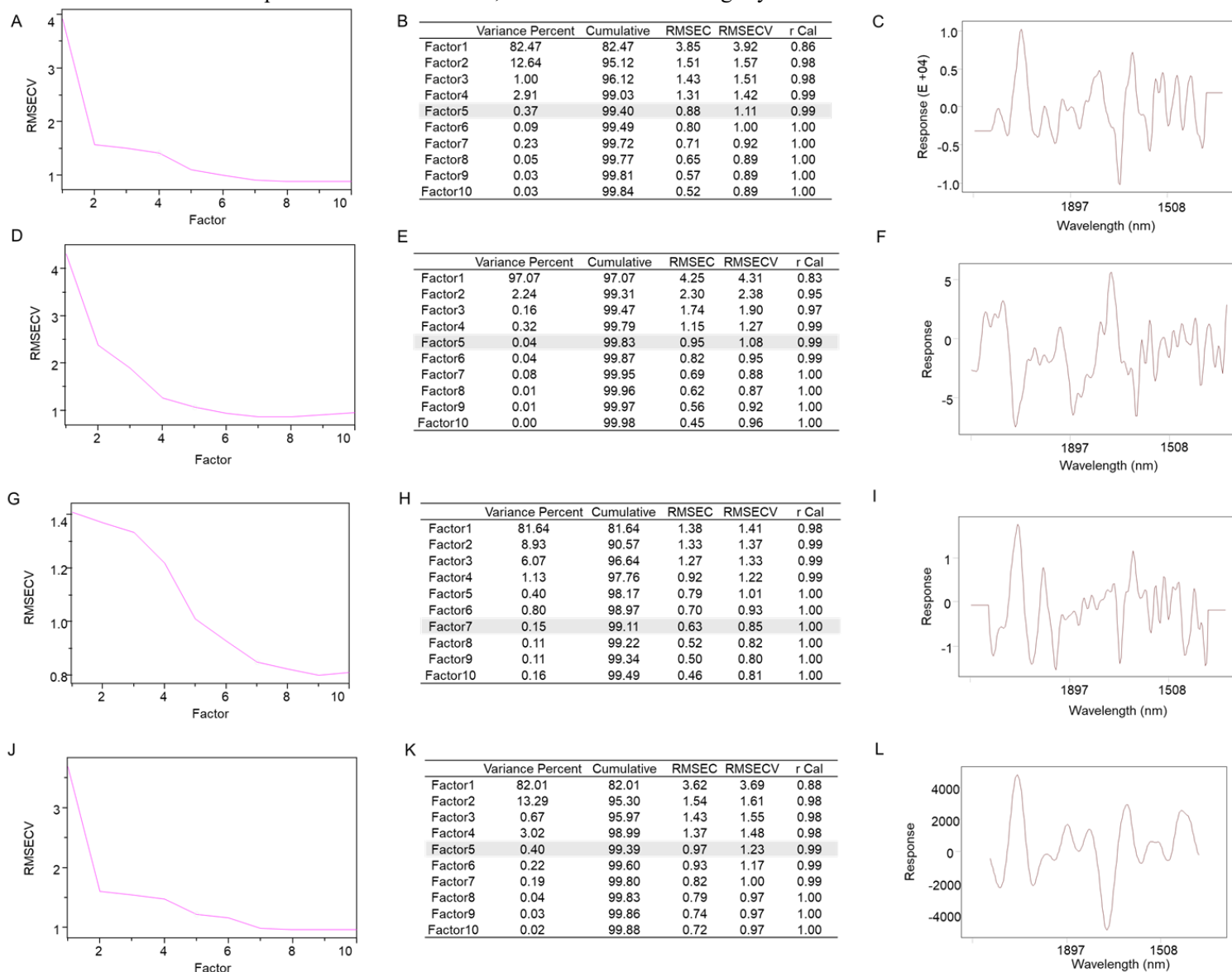


Figure S4. Selection of the optimal number of latent variables by leave-one-out cross-validation (LOOCV) and corresponding regression vectors for the wheat PLSR models using different spectral preprocessing strategies, each applied after mean-centering. For each preprocessing strategy, the three panels from left to right present the RMSECV plot against the number of latent variables (factors), the corresponding explained variance and statistical performance, and the regression vector, respectively. The optimal number of latent variables is highlighted in gray. Panels A–C correspond to second derivative preprocessing (35-point window), panels D–F to normalization, panels G–I to second derivative (35-point window) combined with normalization, and panels J–L to second derivative (35-point window) combined with smoothing (35-point window). Note: r Cal=1.00 represents a rounded value; the actual value was slightly below 1.00.

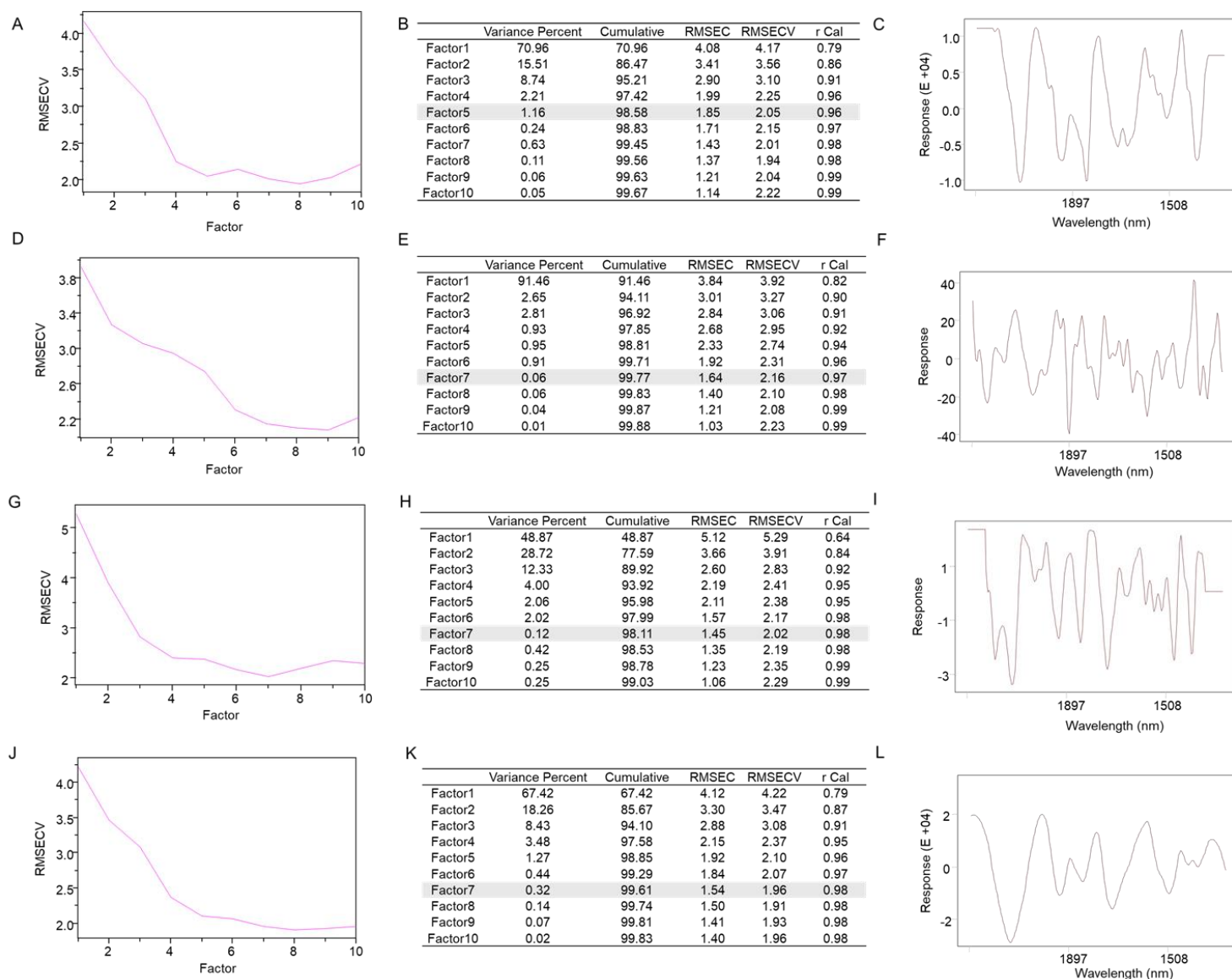


Figure S5. Selection of the optimal number of latent variables by leave-one-out cross-validation (LOOCV) and corresponding regression vectors for the sesame PLSR models using different spectral preprocessing strategies, each applied after mean-centering. For each preprocessing strategy, the three panels from left to right present the RMSECV plot against the number of latent variables (factors), the corresponding explained variance and statistical performance, and the regression vector, respectively. The optimal number of latent variables is highlighted in gray. Panels A–C correspond to second derivative preprocessing (35-point window), panels D–F to normalization, panels G–I to second derivative (35-point window) combined with normalization, and panels J–L to second derivative (35-point window) combined with smoothing (35-point window). Note: r Cal=1.00 represents a rounded value; the actual value was slightly below 1.00.

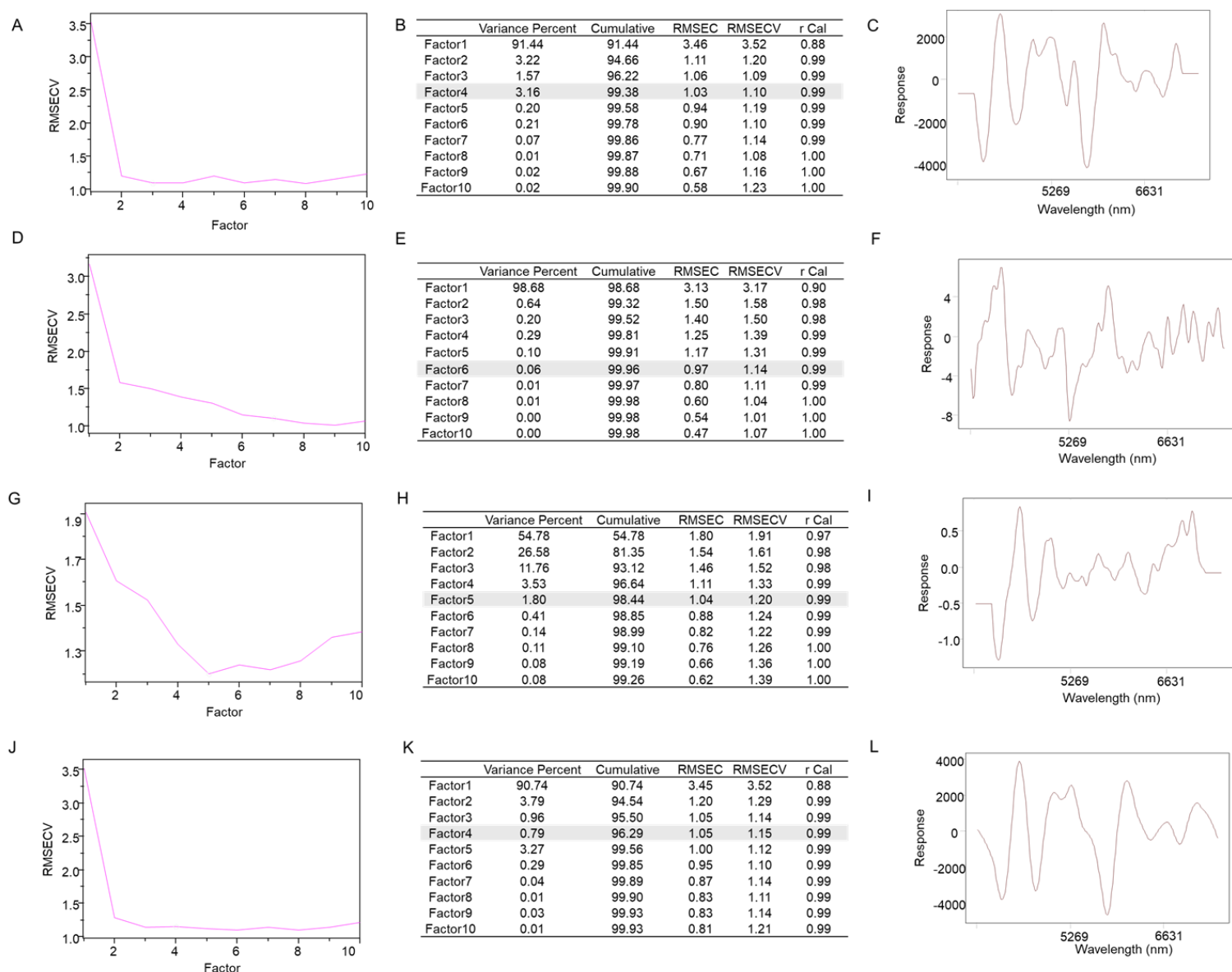


Figure S6. Selection of the optimal number of latent variables by leave-10-out cross-validation and corresponding regression vectors for the soybean PLSR models using different spectral preprocessing strategies, each applied after mean-centering. For each preprocessing strategy, the three panels from left to right present the RMSECV plot against the number of latent variables (factors), the corresponding explained variance and statistical performance, and the regression vector, respectively. The optimal number of latent variables is highlighted in gray. Panels A–C correspond to second derivative preprocessing (35 points), panels D–F to normalization, panels G–I to second derivative combined with normalization (35 points), and panels J–L to second derivative combined with smoothing (35 points). Note: $r_{\text{Cal}}=1.00$ represents a rounded value; the actual value was slightly below 1.00.

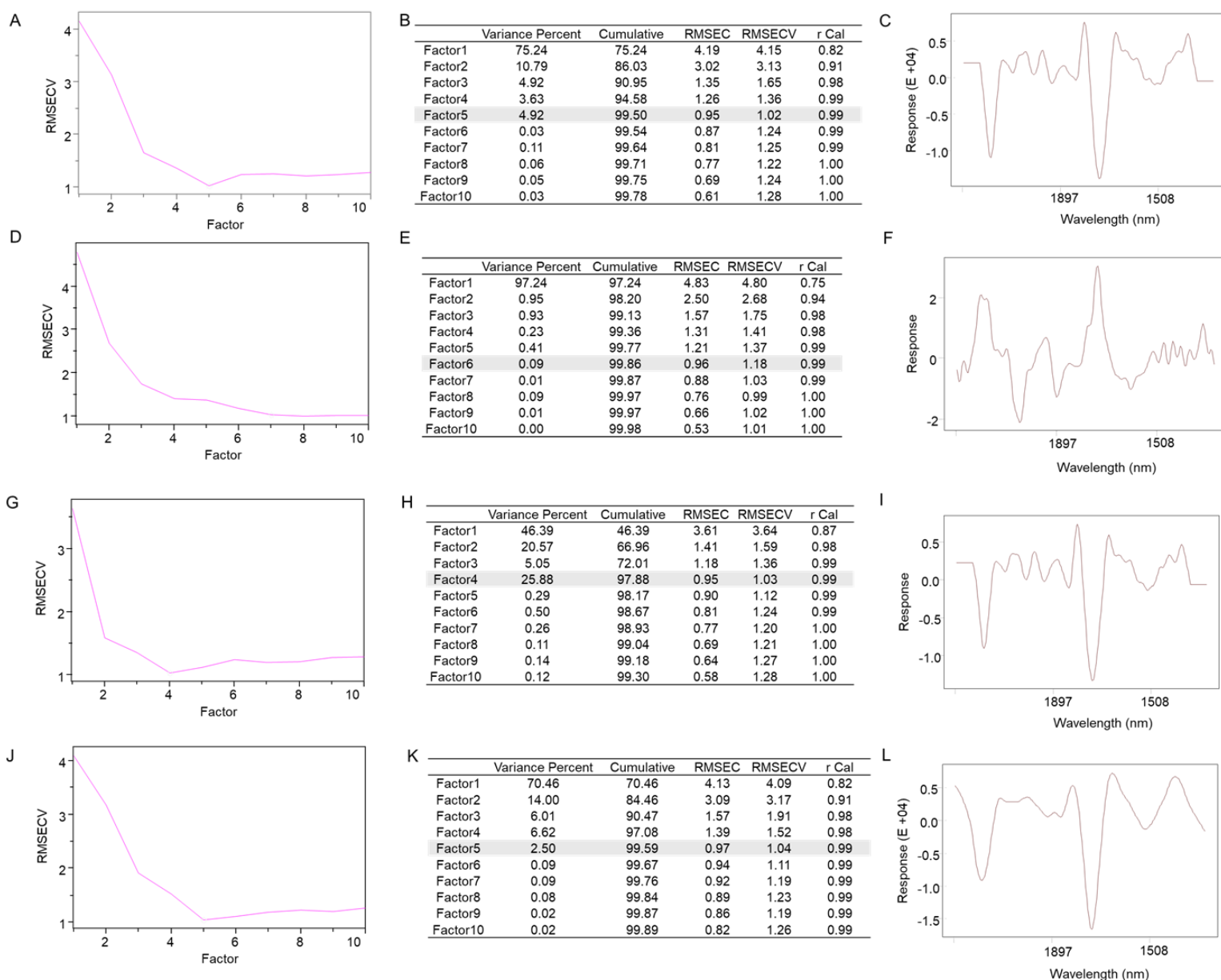


Figure S7. Selection of the optimal number of latent variables by leave-10-out cross-validation and corresponding regression vectors for the peanut PLSR models using different spectral preprocessing strategies, each applied after mean-centering. For each preprocessing strategy, the three panels from left to right present the RMSECV plot against the number of latent variables (factors), the corresponding explained variance and statistical performance, and the regression vector, respectively. The optimal number of latent variables is highlighted in gray. Panels A–C correspond to second derivative preprocessing (35 points), panels D–F to normalization, panels G–I to second derivative combined with normalization (35 points), and panels J–L to second derivative combined with smoothing (35 points). Note: $r_{\text{Cal}}=1.00$ represents a rounded value; the actual value was slightly below 1.00.

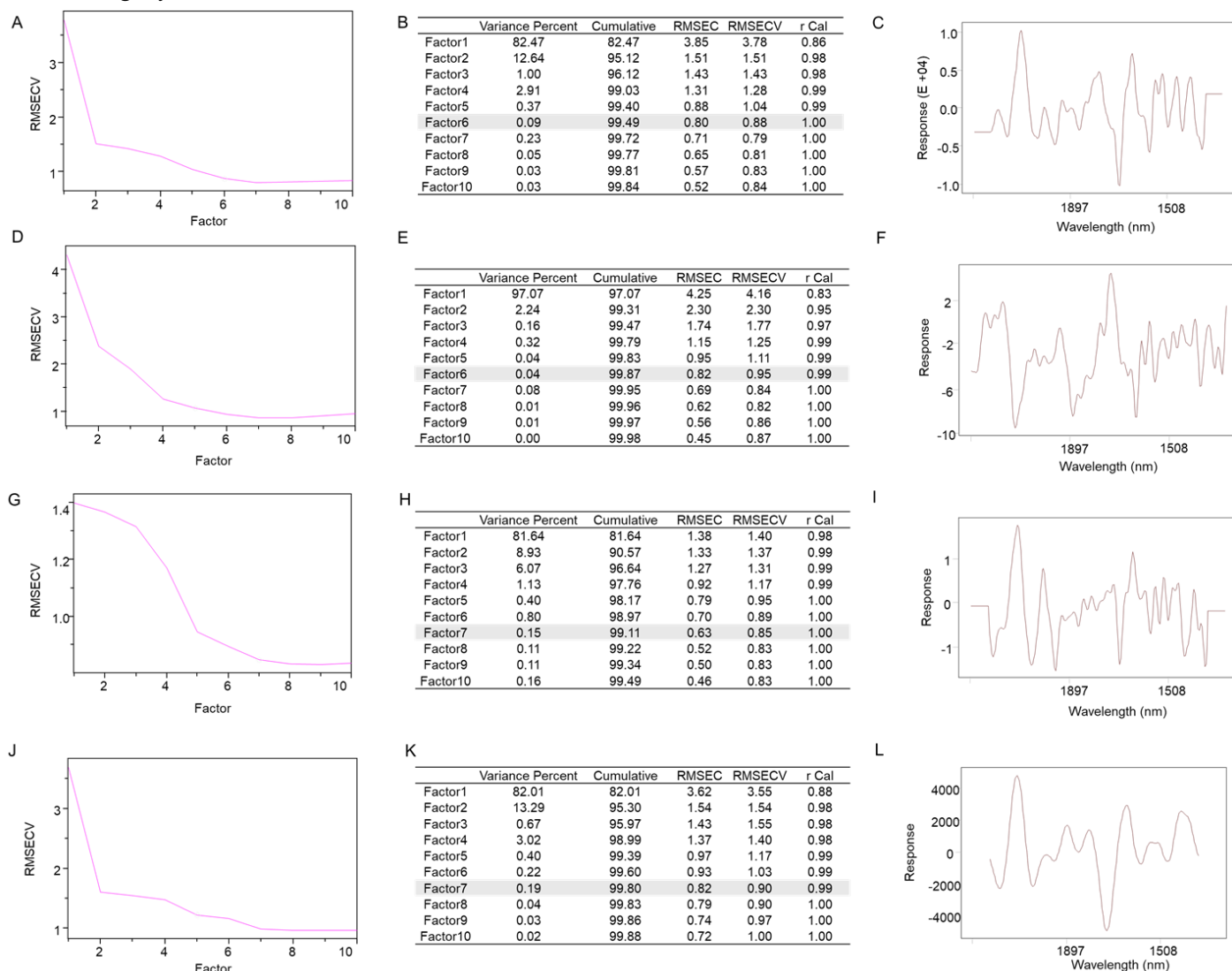


Figure S8. Selection of the optimal number of latent variables by leave-10-out cross-validation and corresponding regression vectors for the wheat PLSR models using different spectral preprocessing strategies, each applied after mean-centering. For each preprocessing strategy, the three panels from left to right present the RMSECV plot against the number of latent variables (factors), the corresponding explained variance and statistical performance, and the regression vector, respectively. The optimal number of latent variables is highlighted in gray. Panels A–C correspond to second derivative preprocessing (35 points), panels D–F to normalization, panels G–I to second derivative combined with normalization (35 points), and panels J–L to second derivative combined with smoothing (35 points). Note: r Cal=1.00 represents a rounded value; the actual value was slightly below 1.00.

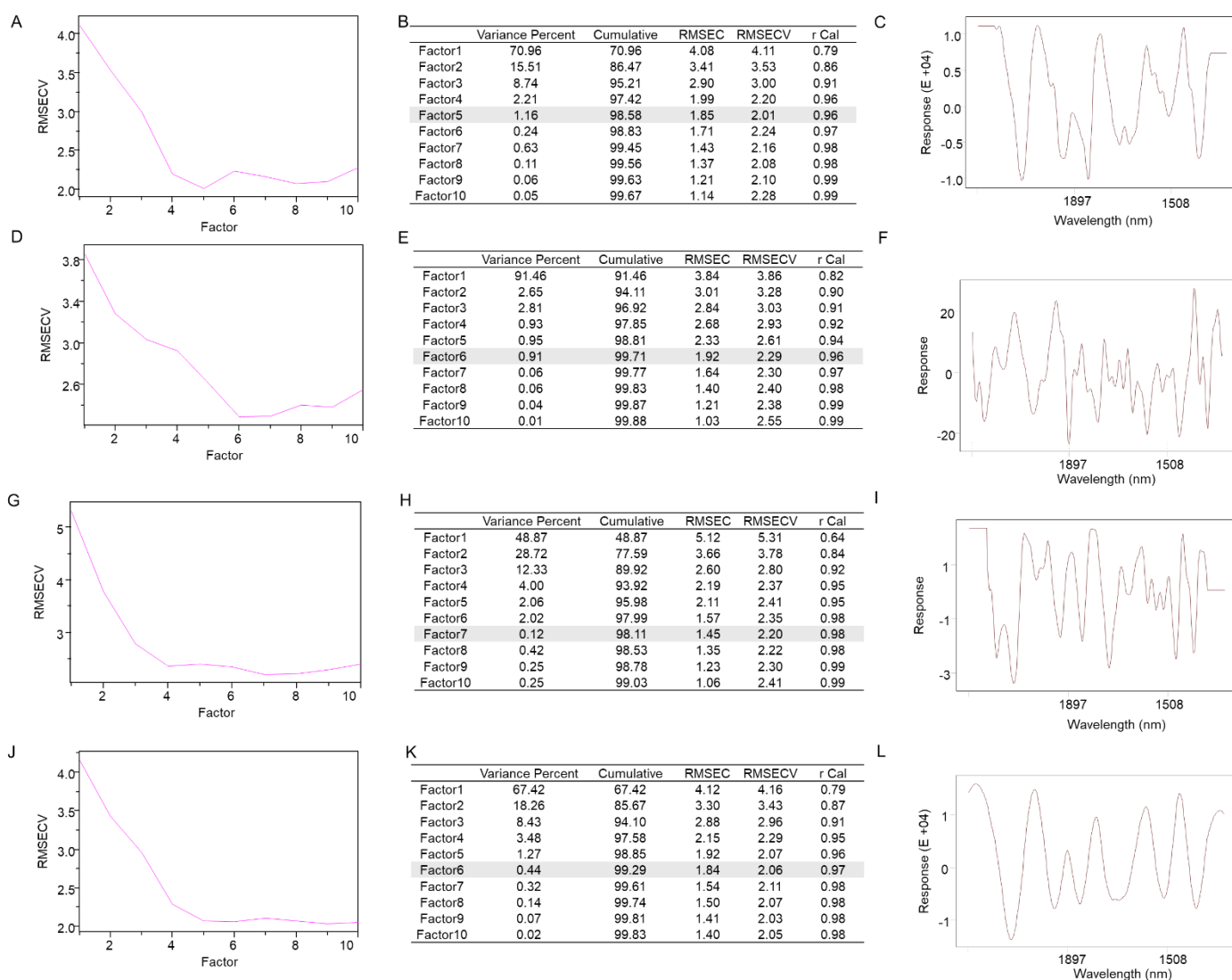


Figure S9. Selection of the optimal number of latent variables by leave-10-out cross-validation and corresponding regression vectors for the sesame PLSR models using different spectral preprocessing strategies, each applied after mean-centering. For each preprocessing strategy, the three panels from left to right present the RMSECV plot against the number of latent variables (factors), the corresponding explained variance and statistical performance, and the regression vector, respectively. The optimal number of latent variables is highlighted in gray. Panels A–C correspond to second derivative preprocessing (35 points), panels D–F to normalization, panels G–I to second derivative combined with normalization (35 points), and panels J–L to second derivative combined with smoothing (35 points). Note: r Cal=1.00 represents a rounded value; the actual value was slightly below 1.00.

