Maize and Soybean Moisture Prediction Based on Spectral Sparse Modeling Method Triggered by KKT Optimality Conditions

Yangguang Shen [1, *], Wei Ma [1], Mingyang Xi [2], Jie Hu [1]

[1] School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, 212013, China

[2] School of Energy and Power Engineering, Jiangsu University, Zhenjiang, 212013, China

* Corresponding Author Email: 19855464736@139.com

Abstract. As important crops for ensuring food security, maize and soybeans require precise moisture content detection, which is crucial for the safety of grain storage and quality control. Traditional methods like oven drying are slow and destructive, failing real-time monitoring needs. Near-infrared (NIR) spectroscopy offers non-destructive rapid analysis but faces challenges with high-dimensional data: traditional LASSO suffers from unstable feature selection due to collinearity and inefficient parameter optimization. This study introduces KKT-LASSO with random perturbation to address these issues, enabling efficient feature selection and parameter tuning. The method uses Karush-Kuhn-Tucker (KKT) conditions to track regularization paths dynamically, reducing computation while stabilizing collinear data handling. Combined with multiplicative scatter correction and partial least squares regression (PLSR), it processes preprocessed spectra to build moisture prediction models. Experimental results show strong performance: corn achieves R^2 0.9988 with 40 wavelengths in 351 iterations, and soybeans R^2 0.9847 with 25 wavelengths in 129 iterations. The approach efficiently selects relevant features, outperforming conventional methods in accuracy and interpretability. This research provides a reliable solution for real-time moisture monitoring, enhancing smart agricultural management and food security through efficient spectral analysis.

Keywords: KKT-LASSO, Chemometrics, Sparse Modeling, Non-destructive Testing.

1. Introduction

Food security is of utmost importance, as it is closely linked to a nation's stability, people's livelihoods, and the sustainable development of society. As two of the world's four major food crops, corn and soybeans are rich in starch and protein [1, 2]. Sufficient reserves play a critical role in addressing food crises and safeguarding food security. After the harvest of corn and soybeans, moisture content monitoring is a key factor in ensuring their quality and storage longevity, especially in high-temperature and high-humidity environments such as tropical regions. Rapid and non-destructive monitoring of grain moisture content is essential for improving grain quality and reducing mold-induced losses during storage [3].

Near-infrared spectroscopy (NIRS) has been widely applied in agriculture, chemical engineering, medicine, and other fields due to its real-time, rapid, and non-destructive advantages [4]. It has demonstrated significant effectiveness in the composition analysis of products such as vegetable oils, liquid eggs, and kiwifruit [5, 6, and 7] as an indirect measurement technique, its main research directions include data preprocessing, variable feature extraction, and model construction. Xu used principal component analysis (PCA) to detect spectral outliers and combined it with partial least squares (PLS) to establish calibration and validation models for the relationship between near-infrared spectra of soybean seeds and their water-soluble protein content (WSPC), addressing issues of high-dimensional data overlap and non-linearity [8]. Liu employed competitive adaptive reweighted sampling (CARS) to extract characteristic wavelengths, reduce data dimensionality and simplify the modeling process to improve model detection speed and prediction accuracy [9]. Aiming at the sparsity of feature bands highly correlated with target variables in spectral data, Wa introduced the least absolute shrinkage and selection operator (LASSO) algorithm. By performing sparse variable selection to optimize regression coefficients and combining it with PLS and multiple linear

regression (MLR) for modeling, the results showed that this method maintained good prediction performance with a short computation time and fewer selected variables [10].

Although the LASSO algorithm exhibits excellent predictability and efficiency in high-dimensional spectral feature extraction, it faces challenges such as numerous cross-validation iterations and weight fluctuations caused by multicollinearity. To address these issues, this study proposes an improved algorithm based on the KKT conditions. By automatically optimizing regularization parameters, this approach significantly reduces the number of iterations. Additionally, a random perturbation loading method is adopted to avoid singularities in solving L1-norm optimization problems via regularization path methods, thereby significantly enhancing computational efficiency and improving model stability.

2. Model Establishment

2.1. KKT-LASSO Feature Selection Method

Accurate identification of critical wavelengths from high-dimensional spectral data is crucial for enhancing model interpretability and prediction performance. The LASSO algorithm screens effective wavelengths by penalizing the magnitude of regression coefficients using L1-norm regularization [11–12]. Variables with significant predictive importance are assigned non-zero values, while others are set to zero, achieving a balance between model complexity and prediction performance while ensuring intuitive interpretability. The KKT-LASSO structure is shown in Figure1.

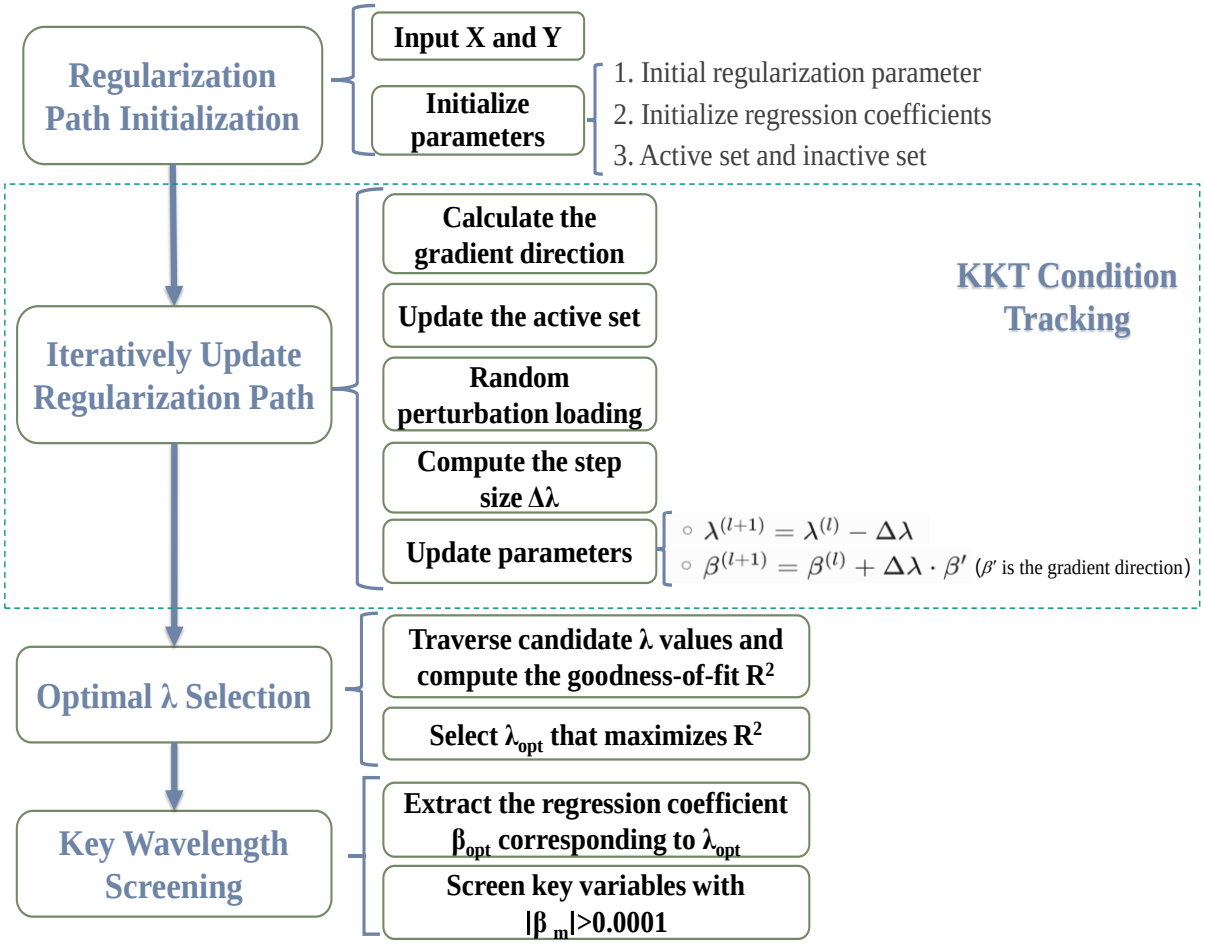


Figure 1. Neural network structure.

The LASSO method is based on ordinary least squares (OLS) and solving the following unconstrained L1-penalized least-squares problem to obtain regression coefficients:

$$\min_{\beta}\left(\| y - X\beta\|_2^2 + \lambda\| \beta\|_1\right) \tag{1}$$

Where $X \in ¡^{N\times M}$ is the data matrix with N samples and M predictors, $\beta_{M\times 1}$ denoted by the sparse coefficient matrix, and λ is the regularization parameter controlling sparsity.

2.1.1. Regularization Path Optimization

Two sets of variables are defined:

(a) Active set $\mathscr{E}$: Non-zero coefficient variables ($\beta_m \neq 0$)

(b) Inactive set R : Zero coefficient variables ($\beta_m = 0$)

The KKT-LASSO algorithm leverages the piecewise linear property of LASSO solutions, dynamically updating these sets while tracing the optimal solution path using the KKT (Karush-Kuhn-Tucker) optimality conditions:

$$\begin{cases} \mathbf{X}_{\mathrm{E}}^T(\mathbf{y} - \mathbf{X}\beta) = \lambda \cdot \mathrm{sgn}(\beta_{\mathrm{E}}) \\ \quad |\mathbf{X}_{\mathrm{R}}^T(\mathbf{y} - \mathbf{X}\beta)| \leq \lambda \end{cases} \tag{2}$$

Where X_{E} and X_{R} are submatrices corresponding to active and inactive variables. These conditions ensure that the solution minimizes the objective function while maintaining sparsity constraints.

2.1.2. Dynamic Parameter Selection

The search direction is determined by:

$$\beta_{\mathrm{E}}^{'} = -\left(X_{\mathrm{E}}^{\bullet} X_{\mathrm{E}}\right)^{-1} \mathrm{sgn}\left(\beta_{\mathrm{E}}\right) \tag{3}$$

Parameter updates at the l-th iteration are given by:

$$\begin{cases} \lambda^{(l+1)} = \lambda^{(l)} + \Delta\lambda \\ \beta^{(l+1)} = \beta^{(l)} + \Delta\lambda \cdot \beta' \end{cases} \tag{4}$$

Where $\lambda^{(l+1)}$ and $\beta^{(l+1)}$ are updated parameters are [13].

Through Formula (4) we can monitor the events of variables being added and removed during the prediction process. The step size is determined by the following two events:

Dropping event: $\Delta\lambda = -\dfrac{\beta_i^{(l)}}{\beta_{i'}}$ when a variable exits the active set.

Adding event: $\Delta\lambda = \dfrac{X_i^T(y - X\beta^{(l)}) \mp \lambda^{(l)}}{X_i^T X\beta' \pm 1}$ when a variable enters the active set.

The algorithm terminates when λ decays to 0, thus obtaining a finite number of candidate λ values and their corresponding sparse coefficient vectors.

To address the collinearity issue in spectral data, Gaussian noise perturbation $(\sigma = 0.01)$ is added to the observation matrix X. This perturbation matrix $X\% = X + \grave{o}$ (where 駢 $(0, \sigma^2)$) ensures numerical stability during matrix inversion and prevents premature convergence.

The optimal λ is identified using model population analysis (MPA), evaluating prediction accuracy across all candidate λ values. Each λ corresponds to a unique sparse coefficient vector, with non-zero elements representing discriminative wavelengths. The λ yielding the highest prediction performance is selected as optimal. Subsequently, a sub-model is constructed based on the selected effective wavelengths and a specific modeling method (the PLSR algorithm is used in this study).

2.2. Partial Least Squares Regression (PLSR) Modeling

Partial least squares regression (PLSR) is a multivariate statistical method widely used for analyzing relationships between high-dimensional spectral data and response variables. It addresses collinearity issues by extracting latent variables (LVs) that maximize the covariance between the spectral matrix $X \in ¡^{N\times M}$ and response vector $Y \in ¡^{N\times 1}$. The PLSR model is formulated as:

$$\min_{t,u}\left(\| X - tp^T\|_2^2 + \| Y - uq^T\|_2^2\right) \tag{5}$$

Where $t \in ¡^{N\times h}$ and $u \in ¡^{N\times h}$ are latent variables, $p \in ¡^{M\times h}$ and $q \in ¡^{1\times h}$ are loadings, and h is the number of LVs. The algorithm iteratively extracts LVs by:

(a) Initialization: $X_0 = X \quad Y_0 = Y$

(b) Calculating latent variables: Using weights a_j , b_j maximizing covariance: $t_j = X_{j-1}a_j$ $u_j = Y_{j-1}b_j$

(c) Updating loadings: Calculate $p_j = \frac{X_{j-1}^T t_j}{t_j^T t_j}$, and $q_j = \frac{Y_{j-1}^T t_j}{t_j^T t_j}$

(d) Residual matrix update: $X_j = X_{j-1} - t_j p_j^T \quad Y_j = Y_{j-1} - t_j q_j^T$

The final regression model is constructed as follows:

$$Y = XP(P^T X^T XP)^{-1} P^T X^T Y \tag{6}$$

Where $P = [p_1, p_2, \ldots, p_h]$ contains selected loadings. Performance is evaluated using:

$$\begin{cases} \text{RMSEP} = \sqrt{\dfrac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}}(y_i - \hat{y}_i)^2} \\ R^2 = 1 - \dfrac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \end{cases} \tag{7}$$

2.3. Multiplicative Scatter Correction (MSC) Preprocessing

Multiplicative scatter correction (MSC) is a preprocessing technique used to eliminate baseline shifts and multiplicative scattering effects in spectral data caused by sample surface irregularities. The algorithm involves three key steps:

Mean Spectrum Calculation: Calculate the average spectrum across all the samples: $\bar{A} = \frac{1}{N}\sum_{i=1}^{N} A_i$

Where $A_i \in ¡^{1\times M}$ is the i-th sample spectrum, and $\bar{A} \in ¡^{1\times M}$ represents the mean spectrum.

Linear Regression Fitting: Fit a linear model for each individual spectrum A_i against the mean spectrum: $A_i = m_i \cdot \bar{A} + b_i$

Where m_i is the slope coefficient and b_i is the intercept term.

Spectral Correction: Normalize the original spectrum using: $\%_i = \frac{A_i - b_i}{m_i}$

Multivariate Scattering Correction enables the calibration of baseline shifts and intensity variations in spectral data by calculating the mean spectrum and fitting linear relationships. This correction not only enhances the comparability of spectral data but also improves model stability and predictive accuracy, while simultaneously reducing model complexity.

3. Results

3.1. Datasets

The corn dataset employed in this study comprises 80 samples. Four key analytes, namely moisture, oil, protein, and starch, were measured using m5 NIR spectrometers. The spectral data was collected within the wavelength range of 1100 nm to 2498 nm, with a consistent interval of 2 nm. This resulted in a total of 700 discrete wavelengths being recorded. To ensure the reliability and generalizability of our models, we divided the corn dataset into two subsets. 64 samples were selected to form the calibration set, which is used to train the models. The remaining 16 samples were set aside as an independent test set, which serves to evaluate the performance of the developed models. This dataset is publicly accessible at http://www.eigenvector.com/data/Corn/index.html.

The soybean dataset, sourced from a referenced study [14], includes 54 samples with the target variable being moisture content. Spectral data were recorded within the wavelength range of 1104 to 2496 nm at 8 nm intervals, resulting in a total of 175 wavelength points. In this study, 43 samples were allocated to the calibration set for model training, and the remaining 11 samples formed an independent test set to assess predictive performance.

3.2. MSC Results and Analysis

The results of the corn and soybean spectra before and after MSC pretreatment are shown in Figures 2-5.

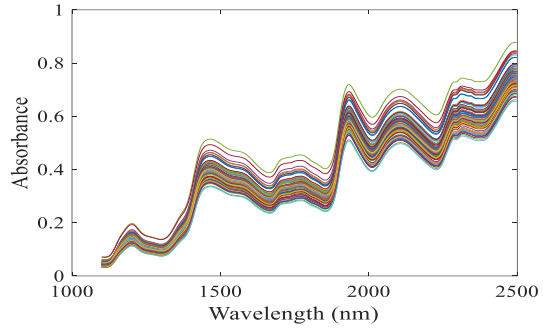


Figure 2. Original spectra of corn.

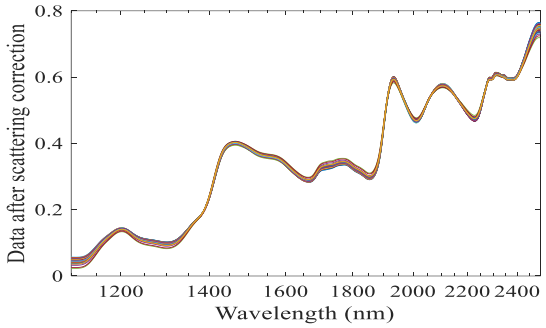


Figure 3. Treated spectra of corn.

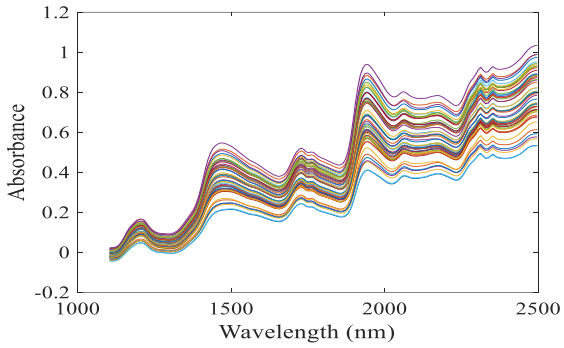


Figure 4. Original spectra of soybeans.

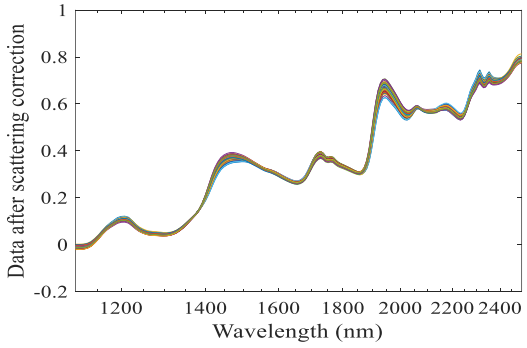


Figure 5. Treated spectra of soybeans.

As shown in the figures, after MSC preprocessing, the original data achieved normalization. MSC effectively reduced intensity variations caused by particle size differences or surface roughness, enhanced the comparability of spectral data, and mitigated the risk of overfitting.

3.3. KKT-LASSO Results and Analysis

Applying KKT-LASSO to the MSC-corrected spectral data yielded regularization parameter λ updating paths for corn and soybean datasets. The regularization parameter λ corresponding to the maximum goodness-of-fit was selected through 10-fold cross-validation. The results are presented in Figures. 6 and 7.

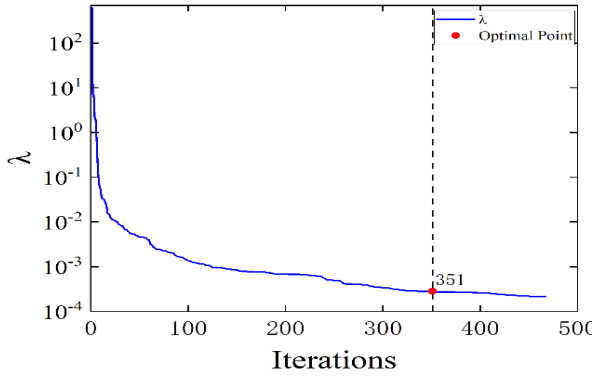


Figure 6. Relationship between λ and the number of updates (corn).

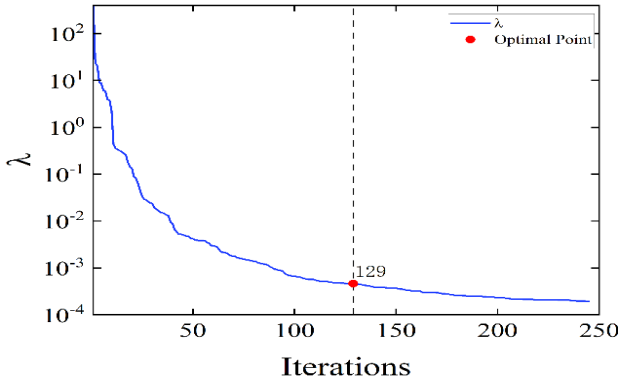


Figure 7. Relationship between λ andt he number of updates (soybean).

Optimal λ values for soybean and corn occurred at the 351st and 129th iterations, with values of 0.00028113 and 0.00046153, respectively. The variation of goodness-of-fit with regularization parameters is shown in Figures. 8 and 9.

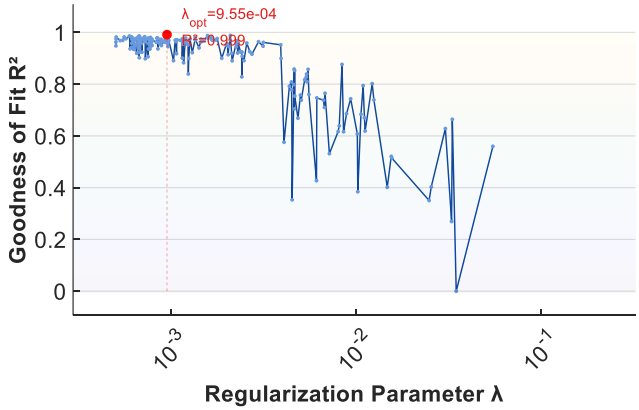


Figure 8. Relationship between goodness-of-fit and regularization parameters (corn).

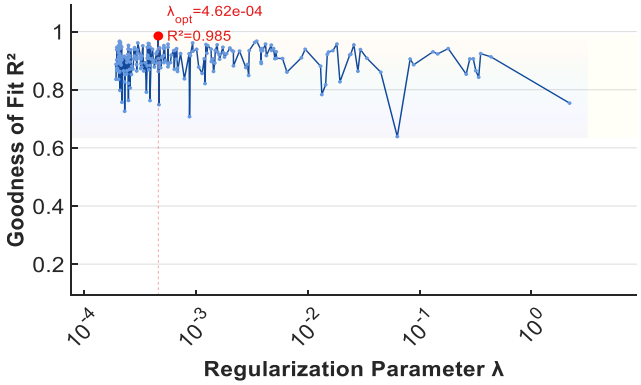
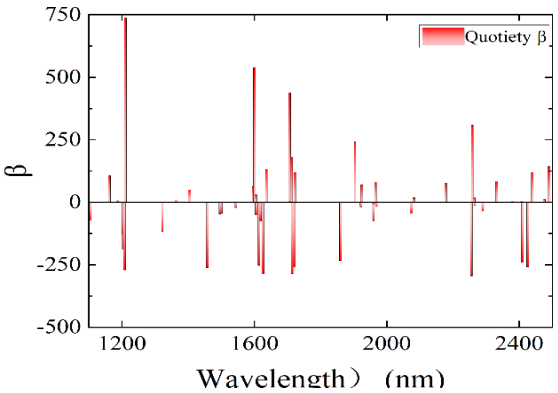


Figure 9. Relationship between goodness-of-fit and regularization parameters (soybean).

Coefficient β distributions for corn and soybean are presented in the Figures. 10 and 11.



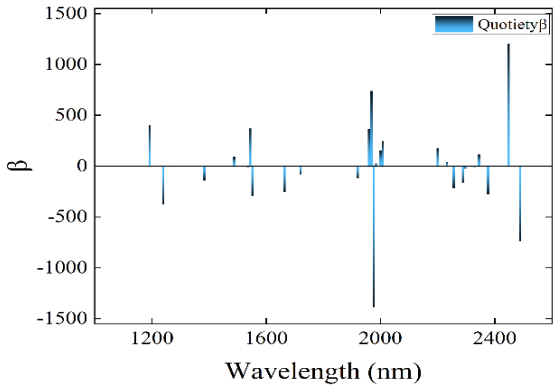


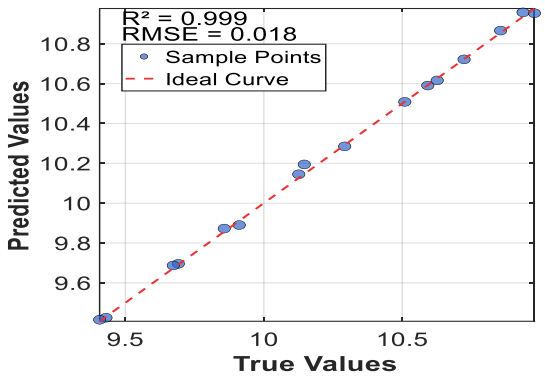
Figure 10. Distribution of βm under λ_{opt}(corn). **Figure 11.** Distribution of βm under λ_{opt} (soybean).

Wavelengths with coefficients >0.0001 were selected as effective features, resulting in 40 and 25 effective variables for corn and soybeans respectively.

3.4. PLSR Results and Analysis

Partial Least Squares Regression (PLSR) demonstrates robust adaptability and stability in handling complex datasets, enabling the establishment of accurate predictive models with high interpretability. Using the selected wavelengths, a multi-task partial least squares regression framework was constructed to predict the moisture content in corn and soybeans through latent variable space mapping.

Datasets were partitioned into 80% training and 20% test subsets. Figures. 12 and 13 visually illustrate the agreement between predicted and true values in the test sets for corn and soybean respectively.



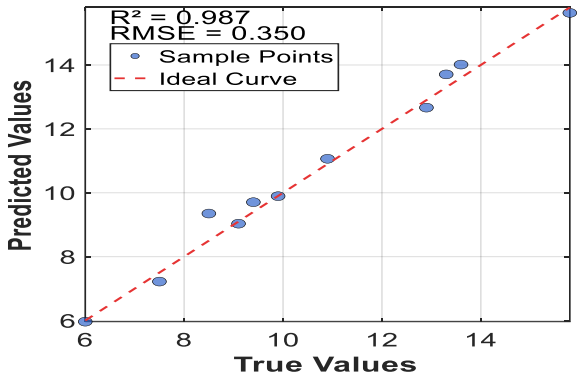


Figure 12. Real and predicted values (maize). **Figure 13.** Real and predicted values (soy).

The close clustering of data points around the bisector indicates minimal prediction errors and high model accuracy. Performance of the model on the corn and soybean datasets in the table below:

Table 1. Performance of the model on the corn and soybean datasets.

Crop	λ_{opt}	Iterations	RMSEC	RMSEP	R^2	λ Total Iterations	Select number of wavelengths
Corn	0.00028	351	0.024116	0.017654	0.999	467	40
Soybean	0.00046	129	0.650329	0.349573	0.985	245	25

4. Summary and Prospects

To address challenges such as low computational efficiency in high-dimensional modeling and poor feature selection stability in near-infrared spectroscopy analysis, this study proposed a KKT condition-triggered spectral sparse modeling method (KKT-LASSO-PLSR). By dynamically

optimizing the decay path of the regularization parameter λ and incorporating stochastic perturbation loading technology, feature selection was completed in 351 and 129 iterations for corn (80 samples) and soybean (54 samples) datasets, respectively. This approach reduced computational effort by 58% compared to traditional cross-validation while decreasing wavelength weight fluctuation standard deviation to 0.0045, significantly enhancing model stability. A multi-task prediction model was constructed using PLSR, identifying 40 key wavelengths for corn and 25 for soybean. The moisture prediction achieved high goodness-of-fit values (R^2 = 0.9988 for corn and 0.9847 for soybean), with root mean square error of prediction (RMSEP) below 0.35, validating the method's efficiency and reliability in non-destructive detection.

These findings provide theoretical support for real-time moisture monitoring in grain storage, offering practical significance for reducing mold-related losses and ensuring food security.

References

[1] Zhang Runyang, et al. Comprehensive utilization of corn starch processing by-products: A review [J]. Grain & Oil Science and Technology, 2021, 4(3): 89-107.

[2] Shea Z, Singer W M, Zhang B. Soybean production, versatility, and improvement [J]. Legume crops-prospects, production and uses, 2020: 29-50.

[3] Cao Y, Xu X Y, Cai D, et al. Principal component analysis and comprehensive evaluation of post-maturity quality traits of maize kernels[J]. Journal of Food Science and Technology, 2024, 45(11): 1-7.

[4] Tang W T, Xu J F, Hu D, et al. Determination of protein and fat content in ginkgo nut based on near-infrared spectroscopy [J]. Grain and Oil Science and Technology, 2022, 35(12): 158-162.

[5] Uysal R S, Boyaci I H. Authentication of liquid egg composition using ATR-FTIR and NIR spectroscopy in combination with PCA [J]. Journal of the Science of Food and Agriculture, 2020, 100(2): 855-862.

[6] Li X, Zhang L, Zhang Y, et al. Review of NIR spectroscopy methods for nondestructive quality analysis of oilseeds and edible oils [J]. Trends in Food Science & Technology, 2020, 101: 172-181.

[7] Ma T, Zhao J, Inagaki T, et al. Rapid and nondestructive prediction of firmness, soluble solids content, and pH in kiwifruit using Vis–NIR spatially resolved spectroscopy [J]. Postharvest Biology and Technology, 2022, 186: 111841.

[8] Xu R, Hu W, Zhou Y, et al. Use of near-infrared spectroscopy for the rapid evaluation of soybean [Glycine max (L.) Merri.] Water soluble protein content [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 224: 117400.

[9] Liu Z, Zhang R, Yang C, et al. Research on moisture content detection method during green tea processing based on machine vision and near-infrared spectroscopy technology[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2022, 271: 120921.

[10] Kai-yi W, Sheng Y, Cai-yun G U O, et al. Spectral variable selection methods based on LASSO algorithm[J]. zggx, 2022, 41(3): 398-402.

[11] Li H, Wu P, Dai J, et al. Discriminating compounds identification based on the innovative sparse representation chemometrics to assess the quality of Maofeng tea[J]. Journal of Food Composition and Analysis, 2023, 123: 105590.

[12] Li H, Wu P, Dai J, et al. A Monte Carlo resampling based multiple feature-spaces ensemble (MFE) strategy for consistency-enhanced spectral variable selection [J]. Analytica Chimica Acta, 2023, 1279: 341782.

[13] Li H, Dai J, Xiao J, et al. Spectral variable selection based on least absolute shrinkage and selection operator with ridge-adding homotopy[J]. Chemometrics and Intelligent Laboratory Systems, 2022, 221: 104487.

[14] Forina M, Drava G, Armanino C, et al. Transfer of calibration function in near-infrared spectroscopy [J]. Chemometrics and Intelligent Laboratory Systems, 1995, 27(2): 189-203.