(IJACSA) International Journal of Advanced Computer Science and Applications,
Vol. 16, No. 11, 2025

# Iterative Partition Optimization: A Novel Approach for Feature Selection in NIR Spectroscopy

Phuong Nguyen Thi Hoang[1], Thinh Ngo Hung[2], Tuong Nguyen Huy[3], Hieu Nguyen Van[4]
Faculty of Information Technology, Pham Van Dong University, Quang Ngai 57000, Vietnam[1]
Faculty of Information Technology, The University of Danang-University of Science and Technology,
Da Nang 550000, Vietnam[2,3,4]

*Abstract*—**Machine learning for near-infrared (NIR) spectroscopy requires effective feature selection to address high dimensionality and multicollinearity. This study proposes Iterative Partition Optimization (IPO), a framework integrating Model Population Analysis with Weighted Binary Matrix Sampling through segment-wise optimization: partitioning spectra into segments, isolating one active segment while freezing others, and using adaptive weighted sampling that learns from best-performing sub-models. Validation across four diverse NIR datasets (n=54-523 samples, 100-700 wavelengths) demonstrates IPO's consistent performance improvement over conventional methods. For agricultural products (soy flour, wheat kernels), IPO achieved lower RMSECV while reducing wavelengths. In chemical analysis (diesel fuels, manure), the method maintained high prediction accuracy (RPD>3.0) using less than half the original variables. Notably in multi-component manure analysis, IPO improved predictions across seven chemical properties (N, NH4, P2O5, CaO, MgO, K2O, DM) while reducing spectral variables, consistently outperforming MCUVE in both accuracy and wavelength selection efficiency. These results establish IPO as an effective wavelength selection method for NIR spectroscopy, addressing multicollinearity while preserving spectral interpretation through optimized interval selection.**

***Keywords*—*Machine learning; feature extraction; near infrared spectroscopy; iterative partition optimization***

## I. Introduction

Computer science drives chemometrics for chemical data analysis [1], with rapid growth in pharmaceuticals, materials, and environmental monitoring [2]–[5]. NIR spectroscopy, while valuable, faces challenges in wavelength selection due to overlapping spectral features across hundreds or thousands of variables [6], [7].

While conventional approaches rely on single models to link inputs with outputs, Model Population Analysis (MPA) [8] transformed spectral analysis by examining patterns across thousands of sub-models simultaneously. MPA approaches spectral analysis by generating multiple sub-models through random sampling techniques, followed by statistical analysis of outputs to extract robust wavelength features - offering superior stability over single-model methods but potentially selecting redundant or correlated spectral regions [8], [9]. This innovative approach has spawned numerous advanced variable selection methods including Monte Carlo Uninformative Variable Elimination (MCUVE), Competitive Adaptive Reweighted Sampling (CARS), Random Frog, and Variable Iterative Space Shrinkage Approach (VISSA) [8]. These MPA-based methods have shown particular success in handling complex spectroscopic data where peak overlaps and baseline shifts confound traditional methods, with applications ranging from trace contaminant detection to drug-protein interaction prediction.

In spectral wavelength selection, PLS [6] remain fundamental for handling multicollinearity . Building on both PLS and MPA principles, MCUVE [6] enhanced feature selection by systematically evaluating wavelength importance through multiple PLS sub-models. While existing MPA-based methods like MCUVE excel at stability through ensemble modeling [6], they lack mechanisms to systematically optimize spectral partitioning boundaries and lack adaptive parameter tuning for different spectral characteristics. This often leads to suboptimal wavelength selection in regions with overlapping peaks or varying signal-to-noise ratios, as noted by study [10]. These challenges in high-resolution spectroscopic data highlight the need for more adaptive approaches.

To address these limitations, we propose Iterative Partition Optimization (IPO), extending MPA principles [11] through adaptive partition optimization designed explicitly for spectral complexity. This novel approach systematically evaluates and optimizes feature subsets, accounting for both local and global spectral relationships while maintaining robustness against measurement noise.

IPO introduces three core innovations in NIR spectral analysis:

- Iterative partitioning: Divides spectra into optimized intervals and refines boundaries based on predictive performance, allowing precise feature identification in complex regions.
- Dynamic selection: Balances exploration of new spectral regions with exploitation of known important features while avoiding local optima.
- Adaptive framework: Updates feature importance using both local and global spectral relationships, dynamically tuning parameters to each dataset's unique characteristics.

Validation across four diverse NIR datasets demonstrated IPO's effectiveness in both feature reduction and prediction accuracy. For agricultural products (soy flour, wheat kernels), IPO significantly reduced RMSECV while requiring only a subset of original wavelengths. In fuel analysis, the method achieved superior RPD values compared to MCUVE through more precise wavelength selection. For multi-component analysis of manure samples, IPO improved predictions across seven chemical properties while substantially reducing model complexity. These results establish IPO as an effective approach for spectral feature selection across different application domains.

## II. Related Work

### A. Machine Learning in NIR Spectral Analysis

Multivariate analysis methods have been extensively employed for NIR spectral data processing and interpretation, encompassing both supervised and unsupervised learning approaches [6], [12]. For exploratory data analysis and dimensionality reduction, Principal Component Analysis (PCA) stands out as the most widely adopted unsupervised technique, transforming high-dimensional spectral data into uncorrelated principal components while preserving essential spectral information. In supervised classification tasks, methods such as Linear Discriminant Analysis (LDA), Soft Independent Modeling of Class Analogy (SIMCA), and Partial Least Squares Discriminant Analysis (PLS-DA) have demonstrated robust performance in sample classification and adulteration detection [6], [12].

Particularly, PLS-based methods have gained prominence due to their ability to handle the multicollinearity inherent in NIR spectra while maximizing the covariance between predictor (X) and response (Y) matrices. PLS has been widely adopted in NIR spectral analysis across various applications. For instance, in food adulteration detection [13], PLS combined with VIP successfully identified 13 informative spectral bands for detecting adulterants in quinoa flour , improving the RPD from 4.04 to 7.71. In geographical origin tracing [14], PLS has evolved into variants like siPLS (synergy interval PLS) and integrated with genetic algorithms (GA), enabling accurate identification of fresh tea leaf origins with prediction accuracy up to 96.67%. However, traditional PLS still faces significant limitations when processing NIR spectral data. When working with quinoa flour, using the full spectral range resulted in an RPD of only 4.04, requiring VIP-based variable selection for performance enhancement. Similarly, in tea leaf analysis, basic PLS proved insufficient to handle data complexity, necessitating integration with techniques like siPLS and GA to improve accuracy. These studies demonstrate that while PLS serves as a foundation for developing new techniques in NIR spectral analysis, its basic form struggles with automatic selection of important spectral regions and handling complex correlations between variables in NIR spectral data.

### B. Feature Selection Approaches in NIR Spectroscopy

Determining the wavelength range of maximum absorbance for a substance is a fundamental step in understanding its interaction with light or electromagnetic waves [15], [16]. In NIR spectroscopy, this process involves selecting relevant spectral features, which significantly reduces input data dimensionality. Such reduction not only simplifies predictive models but also decreases computational costs, accelerates training, and enhances overall model efficiency [16], [17]. Moreover, effective wavelength selection mitigates overfitting in high-dimensional datasets and ensures the reproducibility of spectroscopic analyses across different instruments and experimental setups, making these methods more practical for industrial and research applications.

In the evolution of wavelength selection methods, traditional approaches like Competitive Adaptive Reweighted Sampling (CARS) and Random Forest (RF) focused on direct variable selection. However, MPA has emerged as a more sophisticated framework that examines patterns across multiple sub-models simultaneously. Among MPA-based approaches, MCUVE has shown particular promise when integrated with PLS for NIR spectral analysis.

The effectiveness of these methods has been demonstrated through various practical applications. In pulpwood characterization [18], while Successive Projections Algorithm (SPA-PLS) used minimal variables (13-19 wavelengths) and Synergy Interval PLS (SiPLS) relied on fixed intervals, MCUVE-PLS achieved superior balance by selecting optimal sets of variables (607-639 wavelengths). This systematic evaluation led to better model transferability (RMSEP of 2.348 vs 2.299 for lignin) and more robust performance across different instruments. Similarly, in analyzing edible oil blends from binary to hexanary combinations [19], MCUVE optimized PLS models achieved high prediction accuracy ($R_P^2 > 0.93$, $RPD > 3$), demonstrating robust handling of multi-component spectral data.

Recent advances have focused on improving both model interpretability and predictive accuracy. Yu et al. [20] proposed a combined technique for diesel fuel analysis that offered greater reliability compared to conventional models like PLS and Elastic Net [21]. Their hybrid feature selection mechanism balanced statistical significance with domain relevance. In a different approach, Mamouei et al. employed genetic algorithms exclusively on NIR spectral data, though this method showed limitations in convergence and wavelength elimination efficiency [22].

Despite these advances, several challenges remain in wavelength selection methodology. MPA approaches like MCUVE still struggle to balance local and global spectral information, as they primarily focus on statistical stability without considering spatial relationships between adjacent wavelengths. Additionally, genetic algorithm approaches are heavily dependent on initialization parameters and fitness function design, which can significantly impact feature selection quality. Future research directions should focus on developing adaptive mechanisms that can optimize both aspects simultaneously while incorporating domain-specific knowledge for improved interpretability and reliability, particularly in applications involving complex and noisy spectral datasets.

## III. Theory and Algorithm

### A. Model Population Analysis

MPA is an innovative framework that seeks to enhance the performance of chemometric and bioinformatics algorithms by leveraging statistical insights derived from a population of sub-models [9], [11], [23]. The fundamental concept underlying MPA addresses the inherent uncertainty associated with individual models by statistically analyzing a large ensemble of sub-models generated from a single dataset through Monte Carlo Sampling (MCS). This framework is structured around three primary steps, as in Fig. 1:

- Step 1: Multiple sub-datasets are generated from the original dataset through MCS sampling techniques,

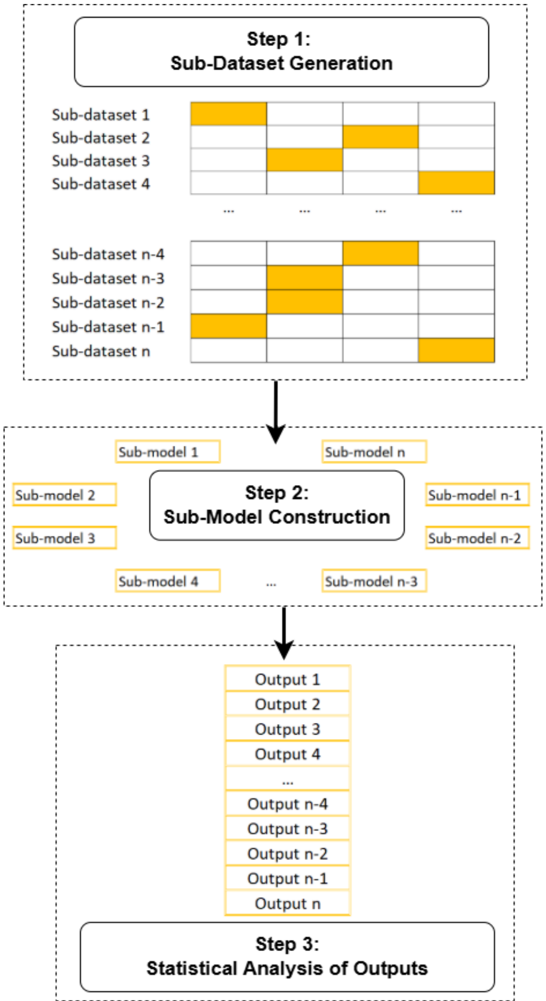


Fig. 1. MPA framework.

ensuring diverse yet representative subsets that maintain the original data characteristics and variability.

- Step 2: Individual sub-models are constructed from each sub-dataset, enabling localized analysis and pattern recognition, with their collective outputs forming a robust foundation for statistical evaluation.

- Step 3: Statistical analysis is performed on the aggregated sub-model outputs to identify stable features and evaluate variable importance, ultimately determining the most significant contributors to model performance.

The MPA framework places a strong emphasis on variable selection, particularly in the context of high-dimensional datasets where the risk of overfitting is pronounced. By systematically analyzing the outputs of a population of models, MPA provides a robust and reliable mechanism for identifying relevant variables while reducing redundancy. This approach ensures improved generalization and interpretability of predictive models, making it a valuable tool for addressing the challenges associated with complex datasets.

### B. Weighted Binary Matrix Sampling

Weighted Binary Matrix Sampling (WBMS) [24] represents an advancement of the traditional Binary Matrix Sampling (BMS) technique, developed to facilitate random sampling of variables in a structured and systematic manner. In the BMS approach, a binary matrix is constructed, wherein each row corresponds to a sampling instance and each column represents a variable. The entries in this matrix are binary (0 or 1), signifying whether a specific variable is included (1) or excluded (0) in a given sampling instance. This framework enables the generation of multiple sub-models from the dataset, thereby facilitating the evaluation of various variable combinations.

WBMS extends this methodology by incorporating a weighting mechanism into the binary matrix, enabling a more refined selection of variables based on their relative importance. The WBMS process can be described in the following steps:

*1) Binary matrix creation:* A binary matrix of dimensions is generated, where represents the number of sampling instances and denotes the number of variables. Each column of the matrix is populated with a predetermined proportion of 1s and 0s, ensuring uniform sampling frequency across all variables.

*2) Weight assignment:* In contrast to BMS, WBMS assigns weights to each variable based on their importance, as determined from prior modeling efforts. The weight for a given variable is calculated as the frequency with which the variable appears in the best-performing sub-models, divided by the total number of optimal models. This weighting mechanism ensures that variables with greater significance have an increased probability of being selected in subsequent sampling iterations.

*3) Sampling procedure:* The binary matrix is permuted according to the assigned weights, enabling a higher likelihood of selecting variables with greater weights. This informed sampling strategy prioritizes the inclusion of influential variables while still allowing exploration of less frequently selected variables, ensuring a comprehensive evaluation.

*4) Iterative optimization:* The WBMS process is iterative, with variable weights updated based on their performance in the generated sub-models. This iterative refinement process continues until the weights converge, resulting in an optimized selection of variables.

The primary advantage of WBMS lies in its capacity to adaptively reduce the variable space while ensuring the inclusion of variables that significantly enhance model performance. By emphasizing the combinatorial effects of variables, WBMS mitigates the risk of discarding potentially informative variables that may have limited individual predictive power but are critical when considered in conjunction with others.

### C. Iterative Partition Optimization

IPO is a novel method combining MPA and WBMS frameworks for spectroscopic feature selection. The method addresses high-dimensional data challenges by partitioning spectral features into manageable segments. IPO optimizes feature selection using the MPA framework while preserving

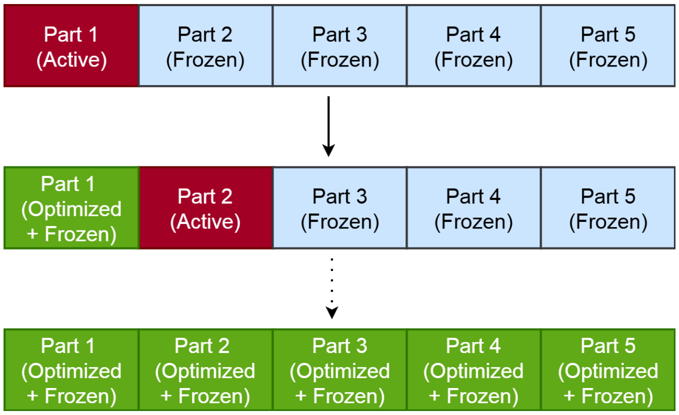


Fig. 2. Iterative optimization of spectral feature set with selective active part designation.

the dataset's global context. The initial phase divides spectral features into distinct segments, as in Fig. 2, enabling focused optimization and reduced computational complexity.

This structured approach differentiates IPO from conventional methods by emphasizing localized optimization within segments, creating a scalable and robust solution for spectroscopic applications. The segmentation process maintains global data relationships while allowing targeted analysis of each partition, effectively balancing local optimization with overall spectral context.

In each iteration, one segment (illustrated as the red cell in Fig. 2) is designated as the active component for optimization, while the remaining segments (depicted as blue cells in Fig. 2) remain frozen. This isolation of the active segment ensures that the algorithm can dedicate computational resources to refine its features. The WBMS technique is then employed to generate multiple subsets of intervals within the active segment. Each subset undergoes evaluation using PLS regression, a widely recognized technique for handling multicollinear and high-dimensional data, with its performance assessed through RMSECV. This process is illustrate through Algorithm 1. To ensure consistency and avoid bias, the fixed segments are combined with subsets derived from the active segment to construct complete feature sets for PLS modeling. The top-performing subsets are identified based on their RMSECV scores, and the results are used to dynamically update the sampling weights. This dynamic adjustment enables the algorithm to progressively focus on variables that contribute significantly to the predictive performance of the model.

The iterative process, comprising subset generation, evaluation, and weight adjustment, continues until no significant improvement in RMSECV is observed for the active segment. At this point, the selected intervals for the optimized segment are frozen to ensure stability, and the algorithm proceeds to the next segment for refinement. This stepwise optimization ensures that the IPO algorithm systematically identifies robust and informative features across the entire spectral dataset, while mitigating risks such as overfitting and redundancy.

By iteratively refining each segment, the IPO method achieves a globally optimized model with enhanced predictive performance. The methodology not only addresses the dimensionality challenges inherent in spectroscopic datasets but also integrates flexibility by incorporating dynamic weight adjustments and iterative feature refinement. Ultimately, the IPO method offers significant advantages, including improved model interpretability, reduced computational cost, and the ability to capture interactions between spectral features. Fig. 3 provides a detailed visual representation of the optimization process, illustrating the methodological steps and enhancing the understanding of the proposed approach. The IPO process is further elaborated in detail in Algorithm 2.

Algorithm 1: Feature Subset Optimization

**Input:**
$all\_indices$: A list of indices corresponding to a specific feature column in $X$
$frozen\_parts$: A concatenation of other parts except current active part
**Output:** $final\_subset$: A list of indices selected after optimization.

$weight \leftarrow$ Initialize weights as $0.5$
$subsets \leftarrow$ List[List[]]
$previous\_rmsecv\_avg \leftarrow \infty$
**while** True **do**
  {Step 1: Generate subsets using WBMS}
  $binary\_matrix \leftarrow$
    WEIGHTEDBINARYMATRIX $(all\_indices, M, weight)$
  **for** $i \leftarrow 1$ **to** $M$ **do**
    {Step 2: Compute RMSECV for each subset}
    $subsets[i] \leftarrow$
      SELECTBINARYSUBSET $(binary\_matrix[i])$
    $rmsecvs[i] \leftarrow$ COMPUTERMSECV$(X, y,$
      $frozen\_parts, subsets[i])$
  **end for**
  {Step 3: Select top $\alpha$ percent of subsets}
  $sorted\_indices \leftarrow$ SORTINDICES$(rmsecvs)[: \alpha \cdot M]$
  $rmsecv\_avg \leftarrow$
    MEAN$([rmsecvs[i] \mid i \in sorted_indices])$
  {Step 4: Check and update the average RMSECV}
  **if** $rmsecv\_avg \geq previous\_rmsecv\_avg$ **then**
    **break**
  **end if**
  $previous\_rmsecv\_avg \leftarrow rmsecv\_avg$
  {Step 5: Update weights}
  $weight \leftarrow$
    UPDATEWEIGHTS$(sorted_indices, binary_matrix)$
**end while**
{Step 6: Extract informative features for current part}
$final\_subset \leftarrow$ Subset with minimum RMSECV
**return** $final\_subset$

## IV. EXPERIMENT AND DISCUSSION

### A. Hyperparameter of IPO

In this study, several parameter configurations were employed to ensure the effective implementation of the proposed methodologies. For the PLSR, a parameter combination of

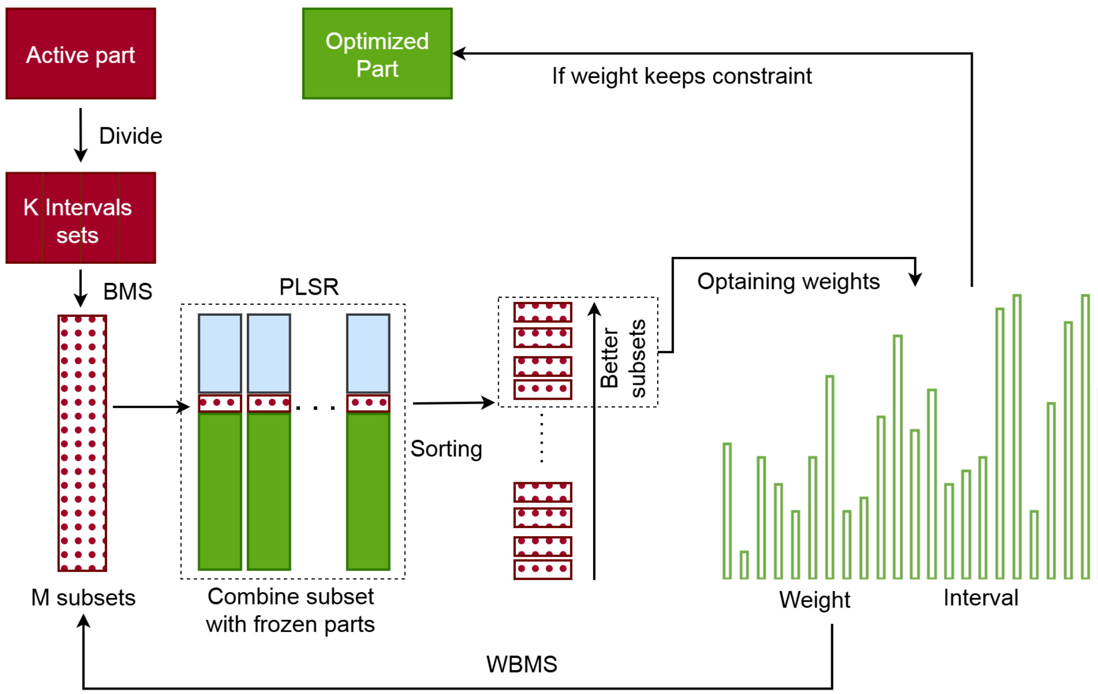


Fig. 3. The core idea for optimization in each active part.

---

Algorithm 2: Iterative Partition Optimization

---

**Input:**
$X$: Feature data of samples
$y$: Target data of samples
$M$: Number of subsets
$num\_parts$: Number of parts
$intervals\_per\_part$: Intervals per part
$\alpha$: Top selection ratio
**Output:** $informative\_parts$: List of Informative features

{Step 1: Divide features into parts}
$parts \leftarrow$ DIVIDEDATAINTOPARTS$(X, num_parts)$
$informative\_parts \leftarrow List[List[] * num\_parts]$
**for** $part\_idx \leftarrow 1$ **to** $num\_parts$ **do**
  {Step 2: Define frozen parts}
  $frozen\_parts \leftarrow$ Concatenate frozen parts
  {Step 3: Split current part into intervals}
  $intervals \leftarrow$ DIVIDEPARTINTOINTERVALS $(parts[part\_idx],$
    $intervals\_per\_part)$
  $all\_indices \leftarrow$ FLATTENINDICES$(intervals)$
  $final\_subset \leftarrow$ ALGORITHM 1 $(all\_indices, frozen\_parts)$
  {Step 4: Extract informative features for current part}
  $informative\_parts[part\_idx] \leftarrow$
    Flatten$(final_subset)$
**end for**
**return** $informative\_parts$

---

10 was utilized. The WBMS approach was configured with the number of subsets, $M$, set to 1000, and the proportion of top-performing subsets, $\alpha$, defined as 0.05. Additionally, the notation IPO $(N, K)$ denotes the application of the IPO algorithm to the dataset, where $N$ represents the number of parts, and $K$ refers to the number of intervals within each part. The specific values of $N$ and $K$ are detailed in Tables I, II, III, and IV.

### B. Datasets

In this study, a selection of widely recognized and publicly accessible benchmark NIR spectral datasets was employed to rigorously evaluate the performance and robustness of the proposed methodologies across diverse applications. conditions. Four benchmark NIR spectral datasets were utilized to evaluate the proposed methodologies.

- The soy oil dataset comprised 54 soy flour samples with spectral measurements across 175 wavelengths (1104–2496 nm, 8 nm intervals), divided into 43 calibration and 11 test samples for oil content prediction [25].
- The diesel fuels dataset included 246 samples (20 high-leverage) measured across 401 wavelengths (750–1550 nm, 2 nm intervals). After outlier removal within the summer fuels class, 133 samples were allocated for calibration and 113 for testing, targeting BP50 prediction [26].
- The wheat kernels dataset contained 523 samples from three locations, measured at 100 wavelengths (850–1050 nm). The dataset was partitioned into 415

calibration and 108 test samples for protein content prediction [27].

- The manure dataset (JIGO8R) consisted of 332 poultry and cattle manure samples from mainland France and Reunion Island. Samples were analyzed using three spectrometers in both fresh homogenized and dried ground states, targeting seven chemical properties: DM, $NH_4$, N, $P_2O_5$, CaO, MgO, and $K_2O$ [28].

These datasets collectively provide a robust and diverse basis for evaluating the effectiveness of the proposed methodologies in different spectroscopic applications.

### C. Metrics

To evaluate the performance of the IPO method, three widely used regression metrics were employed: Root Mean Square Error (RMSE), the coefficient of determination ($R^2$), and the Residual Predictive Deviation (RPD). These metrics were selected for their capacity to provide comprehensive insights into model accuracy, explanatory power, and predictive robustness, ensuring consistency and comparability with established approaches such as Partial Least Squares (PLS) regression.

The evaluation framework employs three standard metrics in spectroscopic analysis:

- Root Mean Square Error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

Characterizes prediction deviation through:

RMSECV: Cross-validation error for generalization assessment

RMSEC: Training error for model fit evaluation

RMSEP: Test error for predictive performance validation

- Coefficient of Determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0, 1] \tag{2}$$

Quantifies explained variance ratio, with $R^2 \to 1$ indicating optimal fit.

- Ratio of Performance to Deviation:

$$RPD = \frac{\sigma_y}{RMSE} \tag{3}$$

Establishes analytical quality:

$$RPD \text{ threshold} : \begin{cases} < 1.5 : \text{Insufficient} \\ [1.5, 2.0) : \text{Approximate} \\ [2.0, 2.5) : \text{Quantitative} \\ \geq 2.5 : \text{High-precision} \end{cases} \tag{4}$$

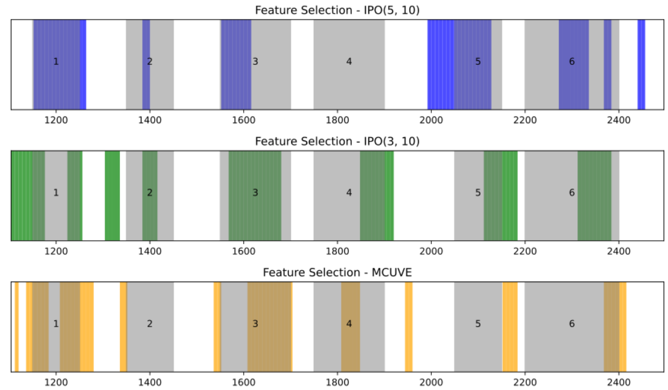


Fig. 4. Selected wavelength intervals for different optimization methods on the Soy oil dataset.

Variables $(y_i, \hat{y}_i, \bar{y}, \sigma_y)$ denote observed, predicted, mean, and standard deviation values. This framework provides systematic validation of model performance across prediction accuracy, explained variance, and analytical precision.

### D. Experiment Hardware

The experiments were conducted on a high-performance laptop tailored for machine learning tasks. The system featured a 13th Generation Intel Core i7-13620H processor, providing powerful mobile computing capabilities. The laptop was equipped with 40GB of RAM, ensuring ample memory for handling large datasets and complex model training. Graphics acceleration was provided by an NVIDIA GeForce RTX 4070 mobile GPU with 8GB of GDDR6 memory.

### E. Experimental Results

In this experiment, we also employ Monte Carlo Uninformative Variable Elimination as a baseline for comparison with our proposed method. MCUVE is a widely used feature selection method designed to identify and eliminate irrelevant or uninformative variables from a dataset. The method involves repeated random sampling (Monte Carlo simulations) to assess the stability and relevance of each variable. Variables with low importance or unstable contributions across multiple iterations are removed, thereby enhancing model interpretability and reducing computational complexity.

*1) Soy oil dataset:* In Table I, the IPO method demonstrates a clear advantage over the MCUVE and baseline PLS methods in predictive accuracy and model robustness for the soy oil dataset. Specifically, IPO (5, 10) achieves the lowest RMSECV (0.9226) and the highest $R^2$ value (0.85), indicating superior calibration and validation performance. Similarly, IPO (3, 10) achieves competitive results with a marginally higher RMSECV (0.9266) but compensates with the lowest RMSEC (0.6773) and RMSEP (1.0090), showcasing its consistency in external validation.

The wavelength intervals selected by different optimization methods, as shown in Fig. 4, provide insight into why IPO outperforms other methods. The regions identified by IPO align well with the chemical structure of soy oil, emphasizing

TABLE I. Comparison of PLS Model Performance using Different Optimization Methods on the Soy Oil Dataset

| Model | Optimization Method | Number of Variables | Parameter Combination | RMSECV | $R^2$ | RPD | RMSEC | RMSEP |
|---|---|---|---|---|---|---|---|---|
| PLS | / | 175 | PCs = 10 | 1.0643 | 0.81 | 2.24 | 0.7021 | 1.1471 |
| | MCUVE | 61 | PCs = 10 | 1.0465 | 0.81 | 2.55 | 0.6975 | 1.0557 |
| | IPO(5, 10) | 60 | PCs = 10 | **0.9226** | **0.85** | **2.67** | 0.6819 | 1.0393 |
| | IPO(3, 10) | 70 | PCs = 10 | 0.9266 | 0.84 | 2.61 | **0.6773** | **1.0090** |

its capacity to isolate chemically relevant spectral features. Specifically:

- Region 1 (1150–1250 nm): Corresponds to the second overtone of the C–H stretching vibration, which is crucial for capturing aliphatic hydrocarbon signals in soy oil.
- Region 3 (1550–1700 nm): Linked to the first overtone of the C–H stretching vibration, contributing to the quantification of unsaturated and saturated fatty acids.
- Region 6 (2200–2400 nm): Reflects a combination of C–H vibrations and other molecular interactions, providing critical spectral information for accurate predictions.

In IPO (3, 10), the inclusion of additional regions, such as:

- Region 2 (1350–1450 nm): Associated with the first overtone of the O–H stretching mode, which reflects water and hydroxyl groups in soy oil.
- Region 4 (1750–1900 nm): Also related to O–H bonds and associated with minor functional groups in fatty acid structures.

These additional regions likely enhance the model's capacity to handle subtle spectral variations, improving its overall predictive performance.

The MCUVE method, while effective in selecting regions 1, 3, and 4, exhibits limitations in other regions (2, 5, and 6). The features it selects in these regions either deviate slightly from the chemically significant intervals or are sparse, introducing noise and reducing predictive accuracy. For instance, the lack of robust selection in regions 5 (2050–2150 nm) and 6 (2200–2400 nm) could result in missing key spectral signatures of C–H and O–H vibrations, impairing the overall model reliability.

Additionally, the ability of IPO to maintain a compact set of selected variables (60–70) compared to MCUVE (61) and baseline PLS (175) highlights its efficiency in reducing model complexity while preserving essential information. This efficiency is reflected in the improved RPD values (2.67 for IPO (5, 10) and 2.61 for IPO (3, 10)), compared to 2.55 for MCUVE and 2.24 for baseline PLS. The higher RPD values indicate better model precision and reliability in practical applications.

Overall, these results highlight the superiority of IPO in handling complex spectral datasets. By accurately isolating chemically significant features and minimizing irrelevant intervals, IPO enhances the calibration and prediction processes. This capability is particularly valuable for spectral datasets like soy oil, where overlapping spectral features and noise

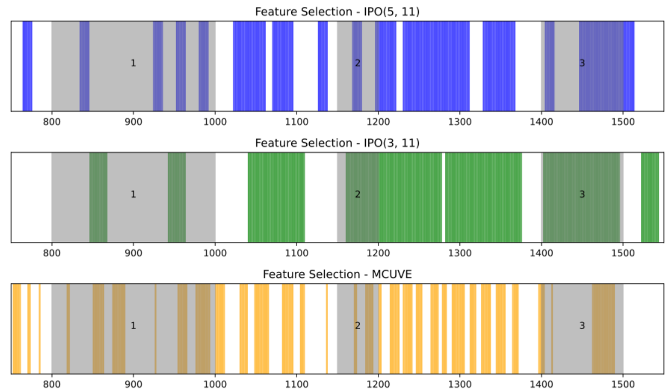


Fig. 5. Selected wavelength intervals for different optimization methods on the Diesel fuels dataset.

can compromise model performance when using traditional methods.

*2) Diesel fuels dataset:* In Table II, a detailed comparison highlights the superior performance of the IPO (5, 11) method across all metrics. Specifically, IPO (5, 11) achieves the lowest RMSECV (2.8759), the highest $R^2$ (0.97), and the highest RPD (5.76), demonstrating its robustness in both predictive accuracy and explanatory power. These results underscore the efficacy of the IPO method in selecting meaningful variables while maintaining a compact model structure. Notably, even the IPO (3, 11) configuration yields strong results, closely approaching the performance of IPO (5, 11), and outperforms the MCUVE method in all metrics. This suggests that the interval-based optimization in IPO effectively captures key spectral features with a minimal increase in the number of variables.

Fig. 5 illustrates the selected wavelength intervals for the diesel fuels dataset using various optimization methods. The property of interest in this dataset is the boiling point, which is influenced by multiple chemical bonds. Specifically, the regions between 800 and 1000 nm (labeled as region 1) and 1400 to 1500 nm (region 2) are associated with O–H bonds, while the region from 1150 to 1200 nm (region 3) is linked to C–H bonds. The observed selection patterns highlight the critical role of these regions in quantifying boiling point variations. However, the distribution of selected variables across the entire spectral range emphasizes the importance of capturing complementary information from diverse regions, as no single region provides sufficient data for accurate quantitation.

It is noteworthy that all optimization methods consistently identify spectral features in regions 1, 2, and 3, as well as common regions between 1050–1100 nm and 1200–1350 nm. This overlap suggests that these regions contain fundamental

TABLE II. COMPARISON OF PLS MODEL PERFORMANCE USING DIFFERENT OPTIMIZATION METHODS ON THE DIESEL FUELS DATASET

| Model | Optimization Method | Number of Variables | Parameter Combination | RMSECV | $R^2$ | RPD | RMSEC | RMSEP |
|---|---|---|---|---|---|---|---|---|
| PLS | / | 401 | PCs = 10 | 3.6189 | 0.93 | 4.58 | 2.8353 | 3.6404 |
| | MCUVE | 186 | PCs = 10 | 3.0852 | 0.95 | 5.38 | 2.6036 | 3.1779 |
| | IPO(5, 11) | 203 | PCs = 10 | **2.8759** | **0.97** | **5.76** | **2.5727** | **3.0961** |
| | IPO(3, 11) | 228 | PCs = 10 | 3.0511 | 0.96 | 5.40 | 2.6421 | 3.1629 |

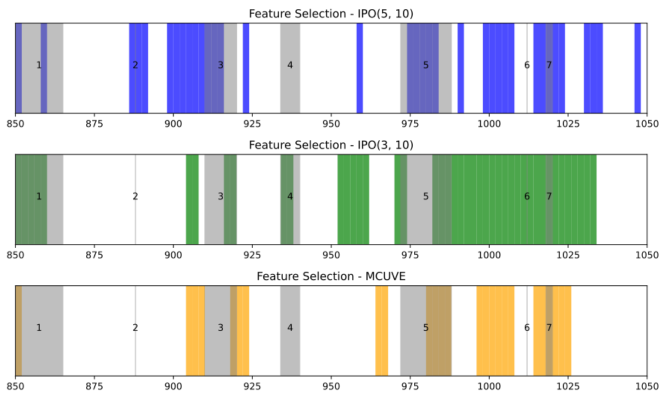


Fig. 6. Selected wavelength intervals for different optimization methods on the Wheat kernel dataset.

information relevant to boiling point prediction, likely due to their association with chemical bonds contributing to the property of interest. Despite this shared focus, the IPO configurations demonstrate a denser selection of features within region 3 compared to MCUVE. This denser coverage enhances the model's ability to leverage critical spectral information, resulting in improved predictive performance as reflected by the lower RMSECV and RMSEP values achieved by IPO.

Furthermore, the broader distribution of selected variables across the entire spectral range by IPO highlights its strength in capturing the synergistic effects of multiple spectral regions. This comprehensive selection strategy likely contributes to its higher RPD values, indicating greater robustness and reliability in predictive modeling. Overall, the results confirm the superior capability of IPO in optimizing variable selection for accurate and reliable prediction, particularly in challenging datasets like the diesel fuels dataset, where complex interactions between spectral features influence the target property.

*3) Wheat kernels dataset:* In Table III, the IPO method demonstrates superior performance compared to MCUVE and the baseline PLS model across all metrics. The IPO(5, 10) variation achieves the lowest RMSECV (0.5301) and RMSEP (0.5902), along with the highest RPD (3.02), highlighting its robustness in both calibration and external validation. Similarly, IPO(3, 10) shows competitive performance, with slightly higher RMSECV (0.5348) and RMSEP (0.6998) but the lowest RMSEC (0.5146), indicating better internal fit.

The wavelengths selected by different methods, visualized in Fig. 6, provide further insights into the differences in performance. Notably:

- Region 1 (850–865 nm): Assigned to the third overtones of C–H bonds, this region captures essential spectral information related to the hydrocarbon content in wheat kernels.
- Region 5 (972–988 nm): Represents the second overtones of O–H bonds, which are critical for detecting water and hydroxyl groups in the sample.
- Region 2 (888 nm): Associated with the second overtones of C–H bonds, this region provides additional structural information about hydrocarbons.
- Region 6 (1012 nm): Reflects the second overtones of N–H bonds, which are essential for characterizing protein content.
- Regions 3 (910–920 nm), 4 (934–940 nm), and 7 (1018–1020 nm): While these regions are difficult to assign definitively, they may still contribute to capturing subtle variations in the spectral profile of wheat kernels, which could be related to minor protein or moisture interactions.

The selection patterns of the two IPO configurations reveal their complementary nature in feature extraction. IPO (5, 10) focuses on regions (1), (2), (3), (5), and (7), whereas IPO (3, 10) emphasizes regions (1), (3), (4), (5), (6), and (7). The inclusion of additional regions by IPO (3, 10), such as (4) and (6), may help the model account for complex structural variations in proteins, enhancing its predictive reliability. This variation in selected regions demonstrates IPO's flexibility in adapting to dataset complexities while maintaining robust predictive performance.

In contrast, the MCUVE method shows limited effectiveness in feature selection. While it successfully identifies regions (5) and (7), its selection in regions (1), (2), (3), and (4) either deviates slightly from the target intervals or includes insufficient spectral features. These suboptimal selections likely introduce noise, reducing the model's ability to generalize and resulting in a higher RMSEP (0.6800).

The results also highlight the efficiency of IPO in reducing the number of variables while maintaining or improving predictive accuracy. IPO configurations select 48–51 variables, a significant reduction compared to the 100 variables used in the baseline PLS model, which still underperforms in metrics such as RMSECV (0.5597) and RPD (2.84). The compact variable sets identified by IPO contribute to better model interpretability and reduced computational complexity, both critical factors in practical applications.

The chemical assignments of the selected wavelengths underline the relationship between the spectral features and the chemical structure of wheat kernels. Regions associated with O–H and N–H bonds provide information about moisture and protein content, while those linked to C–H bonds reflect hydrocarbon structures. These spectral markers are essential

TABLE III. Comparison of PLS Model Performance using Different Optimization Methods on the Wheat Kernel Dataset

| Model | Optimization Method | Number of Variables | Parameter Combination | RMSECV | $R^2$ | RPD | RMSEC | RMSEP |
|---|---|---|---|---|---|---|---|---|
| PLS | / | 100 | PCs = 10 | 0.5597 | 0.88 | 2.84 | 0.5411 | 0.7847 |
| | MCUVE | 35 | PCs = 10 | 0.5442 | 0.88 | 2.96 | 0.5335 | 0.6800 |
| | IPO(5, 10) | 48 | PCs = 10 | **0.5301** | 0.88 | **3.02** | 0.5171 | **0.5902** |
| | IPO(3, 10) | 51 | PCs = 10 | 0.5348 | 0.88 | 2.99 | **0.5146** | 0.6998 |

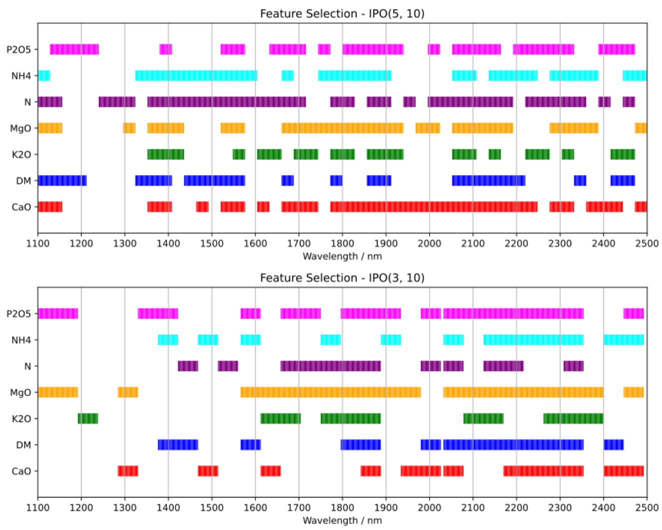


Fig. 7. Selected wavelength intervals across 2 configuration of IPO on the manure dataset.

for accurately predicting wheat kernel properties, and IPO's ability to consistently isolate these regions explains its superior performance.

Overall, the results confirm that IPO not only enhances predictive accuracy but also demonstrates versatility in adapting to complex spectral datasets like wheat kernels. Its ability to select chemically relevant features with precision positions it as a powerful tool for quantitative analysis in food science and other related domains.

*4) Manure dataset:* In Table IV, the results across all substances consistently demonstrate the effectiveness of IPO in improving the predictive performance of PLS models by isolating informative spectral intervals and excluding irrelevant regions. This selective approach not only reduces the number of variables but also enhances the robustness and accuracy of the models.

For substances such as nitrogen ($N$) and ammonium ($NH_4$), the IPO (3, 10) and IPO (5, 10) configurations exhibit superior performance compared to MCUVE and the baseline PLS. Both configurations effectively target NIR spectral regions associated with the chemical properties of nitrogen, capturing the first and second overtones of N-H bonds at 1500–1530 nm and 2168–2180 nm, respectively. Additionally, the third overtone, corresponding to the N-H stretching vibration in the 2050–2060 nm range, is accurately identified. This ability to focus on chemically significant wavelengths allows IPO to achieve a higher RPD and lower RMSECV, affirming its utility in quantitative analysis of nitrogen-related compounds.

For $P_2O_5$, the IPO (5, 10) method achieves a marked improvement, reducing the RMSECV to 0.1565 while achieving an RPD of 3.68, the highest among all methods. This indicates a strong predictive capability, with the selected variables effectively capturing spectral regions pertinent to $P_2O_5$ concentrations. The dense coverage of critical wavelengths by IPO ensures that the combined effect of spectral features is leveraged to enhance model accuracy.

When examining calcium oxide ($CaO$), the PLS model demonstrates limited predictive capability, yielding an RMSECV of 0.3284 and an RPD of only 1.69. This suboptimal performance is likely due to the mismatch between the theoretical absorption range of $CaO$ (550–650 nm) and the spectral range of the dataset (1100–2498 nm). Despite this limitation, the IPO (3, 10) configuration achieves notable improvements, reducing the RMSECV to 0.2808 and increasing the RPD to 1.98. The enhanced performance underscores the robustness of IPO in handling challenging datasets, even when the spectral range does not perfectly align with the chemical properties of the target substance.

For magnesium oxide ($MgO$), IPO (5, 10) and IPO (3, 10) again demonstrate superior performance, achieving lower RMSECV values of 0.0648 and 0.0668, respectively, compared to the MCUVE and baseline methods. The RPD values also indicate better model reliability, with IPO(5, 10) achieving an RPD of 2.80, the highest among all tested methods. This suggests that the interval optimization provided by IPO allows for the identification of subtle yet critical spectral regions that contribute to accurate $MgO$ prediction.

Fig. 7 highlights the selected wavelength intervals by IPO (3, 10) and IPO (5, 10). For all substances, the intervals align well with known spectral features corresponding to the chemical bonds and overtones relevant to the target properties. The figure also emphasizes the ability of IPO to select denser and more comprehensive intervals compared to MCUVE, leading to better model calibration and prediction.

For dry matter (DM) and potassium oxide ($K_2O$), IPO again outperforms other methods in terms of both RMSECV and RPD. For DM, IPO (5, 10) achieves an RPD of 7.93, indicative of excellent predictive performance. For $K_2O$, the IPO (3, 10) configuration achieves an RMSECV of 0.2189, significantly outperforming MCUVE and the baseline PLS model. These improvements highlight IPO's effectiveness in isolating key spectral intervals while maintaining a balanced model complexity.

Overall, the results reinforce the conclusion that IPO provides significant advantages in spectral feature selection, particularly for substances where the spectral data encompass diverse or overlapping chemical properties. By leveraging interval optimization, IPO ensures robust model performance

TABLE IV. COMPARISON OF PLS MODEL PERFORMANCE USING DIFFERENT OPTIMIZATION METHODS ON THE MANURE DATASET

| Model | Substance | Optimization Method | Number of Variables | Parameter Combination | RMSECV | $R^2$ | RPD | RMSEC | RMSEP |
|---|---|---|---|---|---|---|---|---|---|
| PLS | $P_2O_5$ | / | 700 | PCs = 10 | 0.1957 | **0.93** | 2.96 | 0.1601 | 0.1668 |
| | | MCUVE | 275 | PCs = 10 | 0.1871 | 0.89 | 3.09 | 0.1643 | **0.1571** |
| | | IPO(5, 10) | 406 | PCs = 10 | **0.1565** | 0.92 | **3.68** | **0.1383** | **0.1393** |
| | | IPO(3, 10) | 437 | PCs = 10 | 0.1662 | 0.91 | 3.47 | 0.1423 | 0.1475 |
| | $NH_4$ | / | - | PCs = 10 | 0.0781 | **0.95** | 3.58 | 0.0682 | 0.0955 |
| | | MCUVE | 395 | PCs = 10 | 0.0771 | 0.92 | 3.61 | 0.0669 | 0.0926 |
| | | IPO(5, 10) | 420 | PCs = 10 | **0.0737** | 0.93 | **3.80** | 0.0668 | **0.0904** |
| | | IPO(3, 10) | 299 | PCs = 10 | 0.0751 | 0.92 | 3.71 | **0.0661** | 0.0909 |
| | $N$ | / | - | PCs = 10 | 0.2388 | 0.94 | 4.69 | 0.2257 | 0.2347 |
| | | MCUVE | 191 | PCs = 10 | 0.2421 | 0.95 | 4.59 | 0.2313 | 0.2311 |
| | | IPO(5, 10) | 518 | PCs = 10 | 0.2313 | 0.95 | 4.80 | 0.2211 | 0.2320 |
| | | IPO(3, 10) | 276 | PCs = 10 | **0.2052** | **0.96** | **5.45** | **0.1902** | **0.2174** |
| | $MgO$ | / | - | PCs = 10 | 0.0711 | 0.83 | 2.61 | 0.0525 | 0.0866 |
| | | MCUVE | 165 | PCs = 10 | 0.0732 | 0.83 | 2.56 | 0.0536 | 0.0872 |
| | | IPO(5, 10) | 420 | PCs = 10 | **0.0648** | **0.86** | **2.80** | **0.0491** | **0.0823** |
| | | IPO(3, 10) | 483 | PCs = 10 | 0.0668 | **0.86** | 2.73 | 0.0498 | 0.0864 |
| | $K_2O$ | / | - | PCs = 10 | 0.2682 | 0.81 | 2.38 | 0.2141 | 0.2949 |
| | | MCUVE | 406 | PCs = 10 | 0.2556 | 0.82 | 2.51 | 0.2109 | 0.2747 |
| | | IPO(5, 10) | 294 | PCs = 10 | 0.2286 | 0.86 | 2.80 | 0.1974 | 0.2488 |
| | | IPO(3, 10) | 253 | PCs = 10 | **0.2189** | **0.87** | **2.89** | **0.1879** | **0.2248** |
| | $DM$ | / | - | PCs = 10 | 2.7993 | 0.97 | 7.22 | 2.5836 | 2.6829 |
| | | MCUVE | 279 | PCs = 10 | 2.7711 | **0.98** | 7.30 | 2.6079 | 2.7265 |
| | | IPO(5, 10) | 350 | PCs = 10 | **2.5368** | **0.98** | **7.93** | **2.3711** | **2.4525** |
| | | IPO(3, 10) | 322 | PCs = 10 | 2.5752 | **0.98** | 7.90 | 2.3948 | 2.4584 |
| | $CaO$ | / | - | PCs = 10 | 0.3284 | **0.78** | 1.69 | 0.2505 | 0.3315 |
| | | MCUVE | 88 | PCs = 10 | 0.2931 | 0.68 | 1.85 | 0.2412 | 0.3132 |
| | | IPO(5, 10) | 476 | PCs = 10 | 0.2887 | 0.70 | 1.92 | 0.2402 | 0.3068 |
| | | IPO(3, 10) | 299 | PCs = 10 | **0.2808** | 0.72 | **1.98** | **0.2397** | **0.2924** |

across a wide range of substances and spectral conditions, outperforming traditional methods such as MCUVE and baseline PLS in both predictive accuracy and reliability.

### F. Discussion

*1) The influence of parameters on the IPO results:* The results highlight the impact of partitioning parameters on the performance of the IPO method. Dividing the spectral features into five parts consistently yields better predictive performance compared to dividing them into three parts. This improvement can be attributed to the smaller size of feature intervals in the five-part configuration, which allows the algorithm to identify compact, informative intervals. These intervals may otherwise be overshadowed by larger, uninformative regions in the three-part division.

However, certain feature intervals were only selected when using the three-part configuration. This suggests that larger intervals may occasionally capture features that, while not significant individually, become influential when grouped together. This observation reveals a trade-off between locally optimized feature selection (favoring smaller partitions) and globally optimized feature selection (favoring larger partitions).

This interplay between local and global selection strategies highlights the potential value of ensemble learning, where predictions are based on combining results from different partition configurations. Such an approach could leverage the strengths of both local and global feature selection to enhance overall predictive performance.

*2) Versatility and robustness of IPO across diverse datasets:* The IPO method has demonstrated strong versatility and robustness across diverse spectroscopic datasets, each characterized by unique chemical and spectral complexities. For instance, in the soy oil dataset, IPO effectively selected spectral regions associated with key chemical bonds, such as the overtones and combinations of C–H and O–H vibrations, resulting in significantly improved prediction accuracy compared to traditional methods like MCUVE.

Similarly, in the wheat kernel dataset, IPO identified a broad range of spectral intervals corresponding to critical chemical features, such as the overtones of C–H, O–H, and N–H bonds. This ability to capture essential regions while reducing noise underscores IPO's adaptability to datasets with varying degrees of spectral intricacy.

Moreover, IPO's parameter configurability, such as the number of partitions and interval sizes, allows the method to tailor feature selection to the specific characteristics of each dataset. This flexibility enables IPO to consistently achieve high predictive performance, affirming its utility as a versatile and effective tool in chemometrics and bioinformatics.

*3) Differences in feature selection methods:* The results also underscore significant differences between IPO and MCUVE in feature selection. Across all datasets, IPO reliably identifies key spectral regions, while MCUVE tends to select features slightly outside the critical intervals or captures too few features. These limitations in MCUVE's performance are attributed to its reliance on excluding uninformative features without fully addressing multicollinearity, a common challenge in spectroscopic data.

In contrast, IPO's iterative refinement process considers the interaction effects among variables. This "combination effect" ensures that important spectral features, along with their adjacent features, are more likely to be selected, as they collectively contribute to predictive accuracy. By systematically subdividing the dataset and examining smaller intervals, IPO enhances its ability to identify critical features while minimizing the impact of noise.

These distinctions highlight IPO's superior capacity to extract relevant features and enhance predictive outcomes. Its ability to balance local and global feature selection through its customizable parameters makes it a powerful alternative to traditional methods like MCUVE.

## V. Conclusion

Variable selection in near-infrared spectroscopy presents persistent challenges due to high dimensionality and complex spectral correlations. This study proposes the Iterative Partition Optimization framework, which systematically addresses these challenges through a novel integration of localized wavelength optimization and global spectral context preservation. The method's effectiveness was rigorously evaluated using four diverse spectroscopic datasets encompassing agricultural products, petroleum derivatives, and biological samples. Quantitative assessments demonstrated IPO's consistent superiority in both dimensional reduction and predictive accuracy, superior performance across four diverse datasets from agriculture to industry.

For the soy oil dataset, IPO (5,10) achieved the lowest RMSECV of 0.9226 and the highest $R^2$ of 0.85, while IPO (3,10) delivered a competitive performance with the lowest RMSEC (0.6773) and RMSEP (1.0090). In the wheat kernel dataset, IPO (5,10) outperformed other methods with an RMSECV of 0.5301 and an RMSEP of 0.5902, and IPO (3,10) excelled in internal calibration by achieving the lowest RMSEC (0.5146). For the diesel fuels dataset, IPO (5,11) attained the best overall performance with an RMSECV of 2.8759, an $R^2$ of 0.97, and an RPD of 5.76, while IPO (3,11) also demonstrated strong results, underscoring its efficiency in capturing critical spectral features. Moreover, the manure dataset highlights the method's notable gains in predicting nitrogen compounds, $P_2O_5$, $MgO$, and $K_2O$. The ability of IPO to isolate chemically significant spectral intervals, corresponding to key N-H, O-H, and C-H bond overtones, ensures robust predictive performance even under challenging spectral conditions. Compared to MCUVE, IPO demonstrated superior advantages in comprehensive detection of important spectral regions, effective handling of suboptimal spectral ranges, and balanced variable reduction with accuracy. Notably, IPO (5,x) configurations consistently outperformed IPO (3,x) in prediction efficiency across all datasets, suggesting that finer partitioning enables more precise feature identification. However, three-part configurations occasionally captured unique spectral interactions missed by finer divisions, indicating potential value in ensemble approaches. Overall, these results confirm that IPO is a powerful and reliable tool for spectral analysis. Its capacity to identify and leverage informative spectral intervals not only enhances predictive accuracy and model robustness but also improves interpretability and computational efficiency, making it highly applicable in diverse domains such as food science, fuel characterization, and agricultural quality assessment.

In the future, balancing localized and global feature selection will be a focus, as increasing the number of partitions emphasizes local optimization, while fewer partitions favor a more global perspective. Future work may also explore ensemble techniques to integrate insights from multiple variations of feature partitioning, providing a more comprehensive and adaptable feature selection framework. These enhancements could improve the robustness and versatility of the methodology, enabling its application to more complex datasets.

## Acknowledgment

This study is funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2024.ThS.38. This work is supported by The University of Danang- University of Science and Technology, and The People's Committee, Da Nang.

## References

[1] H. Parastar and R. Tauler, "Big (bio) chemical data mining using chemometric methods: a need for chemists," *Angewandte Chemie*, vol. 134, no. 44, p. e201801134, 2022.

[2] Z. Qiu, X. Chen, D. Xie, Y. Ren, Y. Wang, Z. Yang, M. Guo, Y. Song, J. Guo, Y. Feng *et al.*, "Identification and detection of frozen-thawed muscle foods based on spectroscopy and machine learning: A review," *Trends in Food Science and Technology*, p. 104797, 2024.

[3] S. K. Sen, E. D. Green, C. M. Hutter, M. Craven, T. Ideker, and V. Di Francesco, "Opportunities for basic, clinical, and bioethics research at the intersection of machine learning and genomics," *Cell Genomics*, vol. 4, no. 1, 2024.

[4] P. Vizza, F. Aracri, P. H. Guzzi, M. Gaspari, P. Veltri, and G. Tradigo, "Machine learning pipeline to analyze clinical and proteomics data: experiences on a prostate cancer case," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 93, 2024.

[5] C. O. Okoye, H. Jiang, M. Nazar, X.-L. Tan, and J. Jiang, "Redefining modern food analysis: significance of omics analytical techniques integration, chemometrics and bioinformatics," *TrAC Trends in Analytical Chemistry*, p. 117706, 2024.

[6] J. Fu, H.-D. Yu, Z. Chen, and Y.-H. Yun, "A review on hybrid strategy-based wavelength selection methods in analysis of near-infrared spectral data," *Infrared Physics & Technology*, vol. 125, p. 104231, 2022.

[7] H. Yildirim, "The multicollinearity effect on the performance of machine learning algorithms: case examples in healthcare modelling," *Journal of Engineering Smart Systems*, vol. 12, no. 3, 2024.

[8] Y. Huang, "Chemometric methods in analytical spectroscopy technology," in *Chemometric Methods in Analytical Spectroscopy Technology*. Springer, 2022, pp. 1–29.

[9] B. Deng, H. Lu, C. Tan, J. Deng, and Y. Yin, "Model population analysis in model evaluation," *Chemometrics and Intelligent Laboratory Systems*, vol. 172, pp. 223–228, 2018.

[10] R. M. High, "Evaluation of volatile flavour compounds in blue cheeses by linear and non-linear chemometric approaches," Ph.D. dissertation, University of Otago, 2021.

[11] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, and D.-S. Cao, "Model population analysis for variable selection," *Journal of Chemometrics*, vol. 24, no. 7–8, pp. 418–423, 2010.

[12] H. N. Moghaddam, Z. Tamiji, M. A. Lakeh, M. R. Khoshayand, and M. H. Mahmoodi, "Multivariate analysis of food fraud: A review of nir based instruments in tandem with chemometrics," *Journal of Food Composition and Analysis*, vol. 107, p. 104343, 2022.

[13] Z. Wang, Q. Wu, and M. Kamruzzaman, "Portable nir spectroscopy and pls based variable selection for adulteration detection in quinoa flour," *Food Control*, vol. 138, p. 108970, 2022.

[14] S. Wang, L. Feng, P. Liu, A. Gui, S. Gao, J. Teng, F. Ye, Z. Jiang, X. Wang, J. Xue, Z. Song, and P. Zheng, "An efficient method for tracing the geographic origin of enshi yulu fresh tea leaves based on near infrared spectroscopy combined with synergy interval pls and genetic algorithm," *LWT*, vol. 203, p. 116372, 2024.

[15] D. L. Pavia, G. M. Lampman, G. S. Kriz, J. R. Vyvyan *et al.*, *Introduction to Spectroscopy*. Stamford, CT: Cengage Learning, 2015.

[16] W. Zhang, L. C. Kasun, Q. J. Wang, Y. Zheng, and Z. Lin, "A review of machine learning for near-infrared spectroscopy," *Sensors*, vol. 22, no. 24, p. 9764, 2022.

[17] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[18] H. Wang, Z. Xiong, Y. Hu, Z. Liu, and L. Liang, "Application of wavelength selection combined with ds algorithm for model transfer between nir instruments," *Journal of Renewable Materials*, vol. 11, no. 6, pp. 2713–2727, 2023.

[19] H. Zhang, X. Hu, L. Liu, J. Wei, and X. Bian, "Near infrared spectroscopy combined with chemometrics for quantitative analysis of corn oil in edible blend oil," *Spectrochimica Acta Part A*, vol. 270, p. 120841, 2022.

[20] H. Yu, Y. Li, W. Du, M. Yang, X. Peng, X. Wang, and J. Long, "A novel interpretable ensemble learning method for nir-based rapid characterization of petroleum products," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[21] H. N. Le Huy, H. H. Minh, T. N. Van, and H. N. Van, "Keyphrase extraction model: a new design and application on tourism information," *Informatica*, vol. 45, no. 4, 2021.

[22] M. Mamouei, K. Budidha, N. Baishya, M. Qassem, and P. Kyriacou, "Comparison of wavelength selection methods for in-vitro estimation of lactate: a new unconstrained, genetic algorithm-based wavelength selection," *Scientific Reports*, vol. 10, no. 1, p. 16905, 2020.

[23] H.-D. Li, Y.-Z. Liang, and Q.-S. Xu, "Model population analysis for statistical model comparison," in *Chemometrics: Practical Applications*, 2012, pp. 1–20.

[24] B.-C. Deng, Y.-H. Yun, Y.-Z. Liang, and L.-Z. Yi, "A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling," *Analyst*, vol. 139, no. 19, pp. 4836–4845, 2014.

[25] M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena *et al.*, "Transfer of calibration function in near-infrared spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 27, no. 2, pp. 189–203, 1995.

[26] Eigenvector Research Inc. and Southwest Research Institute, "Near infrared spectra of diesel fuels," https://www.eigenvector.com/data/SWRI/index.html, 2005.

[27] D. K. Pedersen, H. Martens, J. P. Nielsen, and S. B. Engelsen, "Near-infrared absorption and scattering separated by extended inverted signal correction (eisc): Analysis of near-infrared transmittance spectra of single wheat seeds," *Applied Spectroscopy*, vol. 56, no. 9, pp. 1206–1212, 2002.

[28] F. Gogé, L. Thuriès, Y. Fouad, N. Damay, F. Davrieux, G. Moussard, C. Le Roux, S. Trupin-Maudemain, M. Valé, and T. Morvan, "Dataset of chemical and near-infrared spectroscopy measurements of fresh and dried poultry and cattle manure," 2020, version V1.