Chemometrics and Intelligent Laboratory Systems 267 (2025) 105541

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

# Bayesian optimization for interval selection in PLS models

Nicolás Hernández [a,*,1], Yoonsun Choi [b,1], Tom Fearn [b,1]

[a] School of Mathematical Sciences, Queen Mary University of London, Mile End Road E1 4NS, London, UK
[b] Department of Statistical Science, University College London, 1-19 Torrington Place WC1E 6BT, London, UK

ARTICLE INFO

*MSC:*
62J05
62K20
62P30

*Keywords:*
Interval selection
PLS
Near infrared
Bayesian optimization

ABSTRACT

We propose a novel Bayesian optimization framework for interval selection in Partial Least Squares (PLS) regression. Unlike traditional iPLS variants that rely on fixed or grid-based intervals, our approach adaptively searches over the discrete space of interval positions of a pre-defined width using a Gaussian Process surrogate model and an acquisition function. This enables the selection of one or more informative spectral regions without exhaustive enumeration or manual tuning. Through synthetic and real-world spectroscopic datasets, we demonstrate that the proposed method consistently identifies chemically relevant intervals, reduces model complexity, and improves predictive accuracy compared to full-spectrum PLS and stepwise interval selection techniques. A Monte Carlo study further confirms the robustness and convergence of the algorithm across varying signal complexities and uncertainty levels. This flexible, data-efficient approach offers an interpretable and computationally scalable alternative for chemometric applications.

## 1. Introduction

Chemometrics, particularly the analysis of spectral data, has driven the development of various algorithms for interval selection. Yet, many of these methods do not fully integrate Partial Least Squares (PLS) as a central modelling framework—despite PLS being one of the most widely used and effective techniques in spectroscopic data analysis; see Yun et al. [1] for a comprehensive review. PLS is especially suitable for high-dimensional data with multicollinearity among predictors, as it extracts latent components that maximize covariance between predictors and response variables. However, its global modelling approach may be suboptimal for interpretability or prediction in spectroscopy when the informative signal is confined to one or more specific spectral regions [2].

While PLS is a powerful tool for full-spectrum analysis, its predictive performance and interpretability can often be enhanced by a preceding variable selection step. The chemometrics literature provides a vast array of such methods, often classified based on their interaction with the modelling algorithm into filter, wrapper, and embedded techniques [3,4]. Common strategies include sequential searches that iteratively add or remove variables [5,6], and more sophisticated wrapper approaches based on intelligent optimization algorithms (IOA) such as Genetic Algorithms [7,8]. A further key distinction in spectroscopy is whether methods select individual variables (*Wavelength Point Selection*) or contiguous blocks of variables (*Wavelength Interval Selection*) -see Yun et al. [1] for a thorough review. Interval selection is often preferred as it respects the continuous nature of spectral bands and can improve model interpretation. Our work contributes to the domain of WIS by introducing a novel Bayesian Optimization framework to guide the search for informative intervals, aiming to overcome the limitations of simpler sequential or exhaustive search strategies.

To address this, the Interval Partial Least Squares (iPLS) method was proposed by Nørgaard et al. [2]. IPLS partitions the spectrum into fixed-width intervals and evaluates each subregion by fitting localized PLS models, allowing the identification of the most informative spectral ranges. This approach enhances both interpretability and prediction, particularly in settings where relevant chemical information is concentrated in narrow spectral bands. However, the standard iPLS implementations, such as those in Kucheryavskiy [9], are often limited to equal-width subintervals or a fixed number of intervals with automatically defined positions. While simple to implement, these constraints can disrupt the underlying correlation structure of the spectrum and limit the method's adaptability, especially in the absence of prior information. Moreover, the combinatorial space of possible subintervals grows rapidly with dimensionality, making exhaustive search computationally impractical.

Flexibility in both interval width and location is particularly important in near-infrared (NIR) spectroscopy, where informative spectral regions may be narrow, noncontiguous, or located in chemically meaningful subregions. Rigid partitioning may overlook such features,

* Corresponding author.
*E-mail address:* n.hernandez@qmul.ac.uk (N. Hernández).
[1] Authors contributed equally to this work.

https://doi.org/10.1016/j.chemolab.2025.105541
Received 12 August 2025; Received in revised form 3 September 2025; Accepted 22 September 2025
Available online 10 October 2025

leading to suboptimal interpretation and performance. This motivates the need for an adaptive, data-driven approach that can efficiently explore a broader range of candidate intervals.

The main contribution of this paper is to introduce a novel interval selection algorithm for PLS modelling based on Bayesian optimization. This approach overcomes the limitations of fixed grid search by using a Gaussian Process (GP) surrogate model to approximate the performance landscape over the space of possible intervals. Starting from an initial random set of sampled intervals, the algorithm iteratively proposes new candidates by balancing exploration and exploitation using an acquisition function. This probabilistic modelling strategy enables a guided search through the discrete (but almost continuous) space of interval configurations, facilitating the discovery of high-performing regions without requiring exhaustive evaluation.

Through simulation experiments where the true informative intervals are known, we demonstrate that the method efficiently converges to the correct spectral regions. We further validate the algorithm on several real-world NIR spectroscopy datasets, showing that it achieves competitive or superior predictive performance compared to full-spectrum PLS and established iPLS variants, while maintaining strong interpretability and computational efficiency.

The remainder of this paper is organized as follows. Section 2 provides an overview of existing interval selection techniques in iPLS. Section 3 presents the proposed Bayesian optimization framework. Sections 4 and 5 report empirical results from simulation studies and real NIR datasets, respectively. A discussion and concluding remarks are given in Section 6.

## 2. Overview of interval specification techniques

In the realm of chemometrics and spectroscopic analysis, effective feature construction and/or feature selection are crucial for building robust predictive models. In the case of feature construction, PLS regression has gained prominence for its ability to handle high-dimensional data while uncovering latent structures [10]. Interval PLS combines feature selection with the traditional PLS methodology by evaluating specific spectral intervals, thereby underpinning a more detailed understanding of the data [2]. In this section we highlight the three main procedures for interval specification within the iPLS framework: Standard iPLS, Forward iPLS, and Backward iPLS. Each of these methods offers distinct strategies for identifying relevant spectral regions, and each has strengths and limitations in the context of feature selection. By examining these approaches, we can better understand how to optimize interval selection to enhance model performance and interpretability in complex datasets.

Standard iPLS, Nørgaard et al. [2] divides the entire spectral range into a series of non-overlapping equal sized intervals, specified by choosing either the width or the number of intervals, [9]. In each interval, a PLS model is developed to predict the target variable, and a metric, such as RMSE, is used to assess each interval's performance. Intervals that yield the highest performance are typically considered to contain the most relevant information.

There are two modification of the standard version: Forward iPLS (fiPLS) [11] and Backward iPLS (biPLS) [12]. FiPLS takes an incremental approach, sequentially adding intervals to the model based on their individual contribution to performance. The process begins with an empty set of intervals, and in each iteration, it selects the interval that results in the greatest improvement in model performance when added. This process continues until adding further intervals does not yield a substantial improvement. On the other hand, biPLS begins with the full set of intervals covering the entire spectral range. At each iteration, it removes the interval that contributes the least to model performance, evaluating the model after each removal. This process continues until removing further intervals would degrade model performance beyond an acceptable threshold.

While Standard iPLS provides an initial view of informative spectral regions, its approach is limited to analysing each interval independently, thus missing any potential synergies or interactions between different spectral regions. This univariate approach might overlook combinations of intervals that collectively contribute significantly to model accuracy. Additionally, Standard iPLS evaluates only single intervals, ignoring more complex configurations that could better capture relevant spectral features. Consequently, Standard iPLS can yield suboptimal results for complex datasets, where relevant information is distributed across multiple regions of the spectrum. As demonstrated in Munck et al. [13], evaluating combinations of two, three, or four intervals captures only a limited portion of the extensive solution space, resulting in suboptimal performance.

Forward iPLS works by sequentially adding intervals based on their incremental contribution to model performance. However, this stepwise addition may produce suboptimal interval combinations by prematurely committing to intervals without a global assessment. Furthermore, models constructed using Forward iPLS tend to become overly complex, as the algorithm stops only when no further improvement is possible, potentially leading to overfitting and increased computational burden.

Backward iPLS, by contrast, begins with a full model and iteratively removes intervals. Although this approach aims to eliminate redundant intervals, it risks prematurely excluding relevant intervals that might have performed well in conjunction with others. The initial inclusion of all intervals makes Backward iPLS especially sensitive to noise, and as shown in Munck et al. [13], it is not ideal for tasks that require removing uninformative spectral regions. Its primary strength lies in identifying relevant, isolated regions, rather than detecting and excluding less significant ones.

## 3. Interval selection via Bayesian optimization

To overcome the some of the drawbacks of the techniques previously detailed we propose a novel algorithm for interval selection based within the framework of Bayesian Optimization (BO). BO is a powerful and flexible optimization technique, particularly suited to finding the global minimum (or maximum) of expensive functions that are costly or time-intensive to evaluate. Traditional optimization methods, like grid search or random sampling, become infeasible for such problems due to the exponential growth of the search space with increasing dimensions. BO addresses this challenge by efficiently utilizing information from past evaluations and uncertainty quantification techniques to guide the search towards promising regions in the parameter space, minimizing the number of function evaluations required to find an optimal solution. BO has proven effective in numerous applications, from hyperparameter tuning in machine learning models to real-world engineering problems, where each evaluation represents a significant time or monetary cost. For a thorough review on the topic see Shahriari et al. [14],Garnett [15],Wu et al. [16].

Formally, Bayesian optimization aims to find a global minimizer, $x^*$, of an objective function $f(x)$ over some domain $\mathcal{X}$. In our application, the objective function $f(x)$ is the Root Mean Squared Error (RMSE) of a PLS model, and the input $x$ represents the centre wavelength of a candidate spectral interval. The goal is thus to find the interval that minimizes the prediction error:

$$x^* = \arg\min_{x \in \mathcal{X}} f(x) \tag{1}$$

where $\mathcal{X}$ is the design space of interest. In global optimization, $\mathcal{X}$ is typically a compact subset of $\mathbb{R}^d$, but the Bayesian optimization framework can be extended to more complex search spaces, such as those with categorical or conditional inputs, or even combinatorial search spaces involving multiple categorical parameters. In this setup, we assume that the function $f$ does not have a closed-form representation but can

be evaluated at any arbitrary query point $x$ within the domain. Each evaluation produces noisy, stochastic outputs $y \in \mathbb{R}$ that satisfy:

$$\mathbb{E}[y \mid f(x)] = f(x),$$

meaning that we observe the function $f$ through unbiased noisy, pointwise observations $y$. In a sequential Bayesian optimization algorithm, at each iteration $j$ (after the initialization), a location $x_{j+1}$ is selected to query $f$, resulting in a new observation $y_{j+1}$. After $J$ queries, the algorithm recommends a final point $x_J$, which it estimates to be the most promising location based on the current surrogate model.

In the application to interval selection in spectroscopy, $f$ represents the performance of the PLS model under different intervals $[a, b] \in \mathcal{X}$, denoted by interval centre $x$, and the stochastic output $y$ could represent prediction accuracy measured via root mean squared error (RMSE). The Bayesian optimization framework is particularly useful in scenarios where evaluations of $f$ are expensive, where gradients of $f$ with respect to $x$ are unavailable, and where $f$ may be non-convex or multimodal. In these cases, it efficiently utilizes the history of the optimization to guide the search, making it highly data-efficient. This methodology needs two key components that we discuss in the next subsections.

### 3.1. Gaussian processes as surrogate model

The first component is a probabilistic surrogate model, which includes a prior distribution that reflects our initial beliefs about the unknown objective function's behaviour, along with an observation model that describes the data generation process. In this sense Gaussian Processes (GPs) are widely used in BO as the surrogate model to approximate the unknown objective function $f(x)$, due to their flexibility and ability to provide both predictions and uncertainty estimates for the objective function across the search space. As a non-parametric model, a GP defines a distribution over functions and is fully characterized by a mean function and a covariance function (or kernel). The core assumption is that any finite set of function values will follow a multivariate normal distribution, which provides a flexible way to model complex functions without being restricted to a specific functional form.

The GP surrogate model provides a predictive mean, which acts as an estimate of the objective function at a given point, and a predictive variance, which quantifies the uncertainty in that prediction. The choice of kernel in the GP plays a pivotal role in defining the smoothness and continuity assumptions of the model; common choices include the Radial Basis Function (RBF) kernel and the Matérn kernel, [17], both of which allow the model to adapt to a wide variety of objective landscapes.

The GP is defined as a distribution over functions, specified by a mean function $m(x)$ and a kernel $k(x, x')$. Here, $x$ and $x'$ represent any two points in the input space (e.g., two different interval centres), and the kernel quantifies the similarity between the function's outputs at these points. A high kernel value implies that the RMSEs of PLS models built on these two nearby intervals are expected to be strongly correlated. The GP is then formally written as $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$. Given $j$ observations $D_j = \{(x_i, y_i)\}_{i=1}^{j}$, the posterior distribution for the function value $f(x_{j+1})$ at a new point $x_{j+1}$

$$p(f(x_{j+1}) | D_j, x_{j+1}) \sim \mathcal{N}(\mu_j(x_{j+1}), \sigma_j^2(x_{j+1}))$$

where $\mu_j(x_{j+1})$ and $\sigma_j^2(x_{j+1})$ are the posterior mean and variance, respectively, whose forms are conditioned on the data.

### 3.2. The acquisition function

The second component is the acquisition function, which guides the search for the optimal solution. It determines the next point in the search space to evaluate by balancing exploration and exploitation of the probabilistic surrogate model. Exploitation means sampling new interval centres in regions where our GP surrogate model predicts a low RMSE is likely. Exploration means sampling in regions where the GP is most uncertain about the RMSE, which are typically areas far from any previously evaluated intervals.

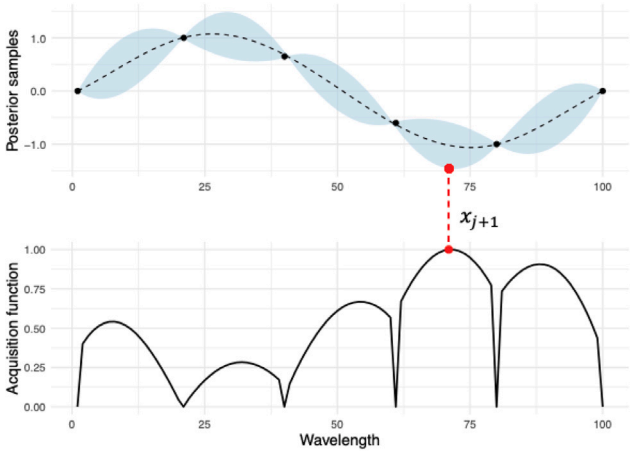


**Fig. 1. Example:** Bayesian optimization framework for interval selection. Upper panel: the posterior process conditioned on six exact observations, with highlighted regions (light-blue) indicating areas of *exploitation* (high predicted performance) and *exploration* (high uncertainty). Lower panel its corresponding acquisition function (bottom). Next sampling location $x_{j+1}$(•).

The main objective of the acquisition function is to maximize the expected improvement over the current best observations. This is expressed mathematically as:

$$x_{j+1} = \arg\max_{x \in \mathcal{X}} \alpha(x; D_j) \tag{2}$$

where $D_j$ is the set of observations up to iteration $j$, and $\alpha(x; D_j)$ denotes the acquisition function.

In BO literature can be found several functional forms for the acquisition function (see Garnett [15] for further details). However as our loss function is the RMSE, and our optimization problem is a minimization one we consider the Lower Credible Bound at a point $x$ as the acquisition function defined as follows:

$$\alpha(x; D_j) = \kappa\sigma(x) - \mu(x) \tag{3}$$

where $\mu(x)$ and $\sigma(x)$ are the predicted mean and standard deviation (or uncertainty) of the Gaussian Process at point $x$ respectively; $\kappa$ is a positive constant (hyperparameter) that controls the trade-off between exploration and exploitation. This trade-off parameter balances between exploring areas of high uncertainty and exploiting areas of low predicted value. Adjusting $\kappa$ allows for flexible control of the exploration–exploitation dynamics in the optimization process.

Fig. 1 illustrates the core principle of Bayesian optimization (BO): selecting the next query point by optimizing an acquisition function. The upper plot shows the Gaussian Process posterior fitted to six observations, with uncertainty ($1.5\sigma$) indicated by the shaded region. The lower plot displays the acquisition function used to propose the next sampling location, $x_{j+1}$, which balances exploration and exploitation. Here, the $x$-axis represents *wavelength*, a key input variable in spectroscopy that determines the energy of light interacting with a material. By modelling the response across different wavelengths, BO can efficiently identify the most informative regions of the spectrum for further sampling.

### 3.3. Proposed methodology

Based on our review of the literature and empirical experience, it has become clear that selecting appropriate intervals for PLS models in the chemometric field is sometimes an important feature to achieving

accurate prediction outcomes. The selection of these intervals, particularly those that contribute to improved predictive performance, requires a flexible approach. To address this and leverage the Bayesian Optimization framework we propose a novel interval selection methodology tailored to the specific challenge of wavelength selection in spectral analysis, following the steps in Algorithm 1, (see Appendix A).

The proposed algorithm aims to identify optimal spectral intervals that minimize prediction error in PLS model. It begins by training a benchmark PLS model using the full spectrum to establish a baseline error. To initialize the search, an initial set of interval centres is randomly sampled. In practice, we advise tightening the sampling domain at the extremes to help bound uncertainty at the edges of the wavelength range. PLS models are then fitted to each of these initial points to compute their Root Mean Squared Error of Prediction (RMSEP). A GP regression model (surrogate model) is then fitted to the observed RMSEPs, and its predictive mean and uncertainty are used to define an acquisition function. Local maxima of this function are extracted and filtered based on whether their lower uncertainty bound lies below a threshold-initially set to the full-spectrum RMSEP. These filtered points are stored as candidates to be evaluated in the next iteration using new PLS models (Bayesian Optimization procedure). The resulting RMSEPs are added to the dataset, and the GP model is updated. This process repeats, allowing the method to progressively focus on informative regions while balancing exploration and exploitation. The algorithm stops when no new candidates improve over previous ones, returning a set of intervals with high predictive performance and interpretability. In practice, we found that a budget of 10–15 iterations typically provides convergence in most settings, as shown in Section 4.

At this point, several key considerations regarding the proposed methodology should be highlighted. First, it is important to clarify that any initial division of data into training and testing sets (for which we recommend a random split or the Kennard-Stone (KS) algorithm) is a one-time pre-processing step performed before the optimization loop begins. The subsequent BO algorithm then operates exclusively on the training portion, using cross-validation to evaluate candidate intervals without repeated splitting. Next, defining the kernel function is a crucial step, as it embeds prior assumptions about the objective function and directly impacts model performance. In this work, we use the Matérn kernel with $\nu = 3/2$. This choice is motivated by its flexibility and widespread adoption as a robust default in the BO literature [18]. Unlike the infinitely smooth Radial Basis Function (RBF) kernel, the Matérn kernel allows for controlling the smoothness assumption of the surrogate model via the parameter $\nu$. The $\nu = 3/2$ setting, which assumes the function is once-differentiable, represents a more realistic and less restrictive prior for real-world objective functions whose smoothness is unknown [19] (Ch. 2). Our preliminary sensitivity analysis confirmed its suitability for this application.

For the acquisition function parameter $\kappa$, we consider several scenarios $\kappa = \{1.5, 2, 3\}$, however a common choice is to set $\kappa = 2 \approx 1.96$ which is the normal (two-tailed) quantile value at 95%. Regarding each interval Partial Least Squares (PLS) model, it is important to emphasize the optimization of the number of components based on the training data to ensure the model captures relevant patterns. Additionally, it is advisable to tighten the extremes of the domain, which helps bound uncertainty at the edges of the wavelength range.

In this study we have used Root Mean Square Error (RMSE) as the criterion to optimize. This is a standard metric for assessing predictive performance in regression tasks. There are, of course, other desirable properties one might wish to consider at the same time, interpretability being the most obvious one. The twin difficulties here would be quantifying this property in a numerical fashion and establishing the trade off between RMSE and interpretability in any given application. Setting up a formal framework for this is well beyond the scope of this paper, but it would be relatively easy for a user who felt unhappy with the interpretability of any suggested solution to add constraints that forced either the inclusion or exclusion of particular intervals on the grounds of interpretability or any other domain-specific criterion.

### 3.4. Modelling multiple intervals

The proposed approach can be naturally extended to the selection of multiple intervals, enabling the model to capture more complex spectral patterns. By considering combinations of informative regions rather than a single interval, the method gains flexibility and improves its ability to identify relevant spectral features across broader or discontinuous wavelength ranges. This extension enhances both the interpretability and predictive performance of the resulting PLS models [20].

The selection of multiple intervals can be approached using two distinct strategies: *sequential* or *simultaneous* optimization. In the sequential approach, the process is formalized as a greedy, multi-stage optimization. To clarify the procedure, let the objective function $f(x_1, \ldots, x_k)$ represent the model's prediction error (e.g., RMSEP) for a set of $k$ intervals centred at $x_1, \ldots, x_k$. The initial stage identifies the single most informative interval by finding the centre, $x_1^*$, that minimizes the error for a single interval:

$$x_1^* = \arg min_{x_1} f(x_1)$$

Following this, the second stage finds the best complementary interval by performing a new optimization where the first interval's location is fixed. This second search solves for $x_2^*$ by optimizing the two-interval objective function:

$$x_2^* = \arg min_{x_2} f(x_1^*, x_2)$$

This process reduces the dimensionality of each search, making it computationally simpler than a full simultaneous search. However, this greedy strategy is not guaranteed to find the globally optimal combination of intervals and may instead converge to a local optimum.

The *simultaneous* approach conceptually transforms the task from a one-dimensional line search into a multi-dimensional optimization problem. For selecting two intervals, this creates a 2D search space where the axes represent the centre positions of the first $(x_1)$ and second $(x_2)$ intervals, respectively. Every coordinate $(x_1, x_2)$ in this space corresponds to a unique *pair* of intervals, and the objective function, $f(x_1, x_2)$, can be visualized as a complex performance landscape over this 2D plane. The goal of the Bayesian Optimization is to find the global minimum-the lowest "valley"-in this RMSEP landscape. To do this efficiently, we employ a multi-input Gaussian Process surrogate model:

$$f(x_1, x_2) \sim \mathcal{GP}(m(x_1, x_2), k((x_1, x_2), (x_1', x_2')))$$

This model's key component is the multi-input covariance function, $k((x_1, x_2), (x_1', x_2'))$, which measures the similarity between two distinct pairs of intervals. This structure is what allows the model to capture crucial *interaction effects*-for instance, determining if the predictive power of an interval at $x_1$ is enhanced or diminished when paired with a second interval at $x_2$.

For the simultaneous selection of multiple intervals, a multi-point, filtered-batch Bayesian Optimization strategy is employed. This approach is designed to enhance exploration by proposing several candidate points in each iteration. The process begins after fitting the Gaussian Process surrogate model. First, the domain of previously evaluated points in the 2D search space is partitioned into a set of non-overlapping triangles using Delaunay triangulation. Within each triangle, a local search identifies the point that minimizes the acquisition function (the Lower Credible Bound), yielding a large set of potential candidates—one from each triangular region.

However, not all these candidates are evaluated. A crucial filtering step is applied to ensure that computational effort is focused only on points that are genuinely promising. A performance threshold is set to the minimum RMSEP observed in the previous iteration's samples. A candidate is only retained for evaluation if its predicted performance (i.e., its acquisition function value) is lower than this threshold.

This procedure is visually analogous to the 1D case shown in Fig. 3, where only the local minima (red dots) that dip below the perfor-

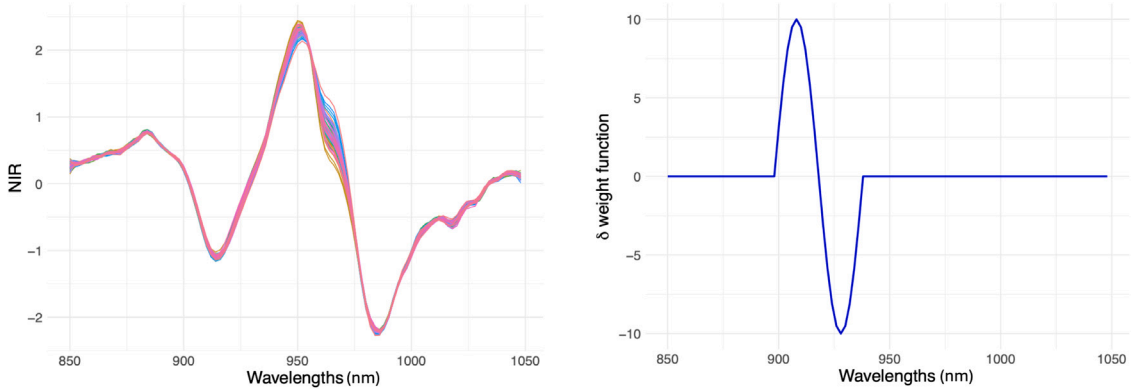


**Fig. 2. Left:** Near-infrared transmission spectra of 48 wheat kernel samples plotted against wavelengths (nm). **Right:** Artificially constructed response variable $y$ showing dependence on wavelengths 900 to 938.

mance benchmark (dashed line) are selected as candidates for the next iteration. All candidates that pass this filter are then evaluated in a single batch, and the resulting data are used to update the surrogate model for the next optimization cycle. This method effectively balances broad exploration across the entire space with a focused exploitation of regions that promise to outperform the current best solution.

## 4. Simulation study

### 4.1. Simulation setup and one–shot experiment

For the simulated experiment, we begin with pre-processed (second derivative followed by SNV [21]) transmission spectra from 48 wheat kernel samples, each measured across 100 wavelengths in 2 nm increments, covering a range of 850–1048 nm (see left panel of Fig. 2), represented by $\mathbf{Z} \in \mathbb{R}^{n \times p}$ (with $n = 48$ samples and $p = 100$ wavelengths). The spectra are real, but the response variable, $\mathbf{w}$, was artificially constructed to reflect specific dependencies on a particular subset of the wavelength domain. To this aim we define the weight function $\delta(x)$, which gives a non-zero weight to the wavelengths between 898 and 938 nm (points 25–45 in the normalized scale 0–100), as follows:

$$\boldsymbol{\delta} = \begin{cases} 10 \times \sin\left(\frac{(x-25)\pi}{10}\right) & \text{for } x \in [25, 45] \\ 0 & \text{for } x \notin [25, 45] \end{cases}$$

Then the artificial response variable is defined as:

$$\mathbf{w} = \delta^T \mathbf{Z} + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, 0.025)$ is Gaussian noise. The resulting $\mathbf{w}$ values have signal components only for wavelengths between 898 and 938 (see right panel of Fig. 2). This design introduces a clear relationship between $\mathbf{w}$ and a localized interval in the spectral data.

To simplify the setting and align with the structure of the synthetic data, we assume that only one predictive interval exists and therefore employ a one-interval selection strategy. The interval width is fixed to 21 variables, corresponding to 42 nm in spectral space, and kept constant throughout the optimization.

This controlled setup allows for a focused evaluation of the method's ability to identify a single informative region. We acknowledge that using a fixed interval width is a simplification for this initial study. While the interval width could be treated as a second hyperparameter, this extension is non-trivial. A variable width can introduce sharp variations or discontinuities into the objective function landscape; for instance, two intervals with nearly identical centres but different widths might include or exclude noisy regions, yielding vastly different prediction errors and therefore challenging the standard GP surrogate to model.

***Results.*** The results of the one shot simulated study demonstrate that the proposed Bayesian optimization framework effectively identifies informative spectral intervals associated with the response variable. As shown in Fig. 3, the method progressively converges towards intervals near 900–938 nm, which aligns with the underlying structure of the artificially generated response given by $\delta$ weight function (Fig. 2-right). Notably, the optimal interval selected corresponds to position 32 out of 100, which maps to a central wavelength of 912 nm—directly overlapping with the true active region, containing 87.5% of the wavelengths (nm).

Fig. 3 also displays the evolution of the Bayesian optimization procedure for the simulated study case. Each plot shows the predicted RMSECV (solid black line) and associated uncertainty bands (shaded area) for each iteration. The blue dots represent the initially evaluated intervals. Red points indicate the local minima of the acquisition function that were selected as candidates for evaluation in the next iteration. In the final iteration, the green star identifies the selected optimal interval centre.

Compared to the baseline RMSECV (Root Mean Squared Error of Cross-Validation) from the full-spectrum PLS model (0.47), the RMSECV obtained using the selected interval is substantially lower, 0.352. This highlights that when prior knowledge suggests a sparse informative structure, the one-interval strategy not only improves predictive accuracy but also enhances model interpretability by clearly identifying the relevant spectral region. The convergence behaviour, performance gain, and localized selection underline the potential of this approach in applications where both interpretability and computational efficiency are critical, especially in high-dimensional spectral data.

The horizontal dashed line represents the dynamic performance threshold used for filtering candidates (see Algorithm 1, Step 7). For the first iteration (a), this threshold is the full-spectrum RMSECV, while for subsequent iterations it represents the best RMSECV found in the preceding step.

### 4.2. Monte Carlo evidence

To evaluate the robustness of Algorithm 1 under stochastic conditions and varying uncertainty widths, we conducted a Monte Carlo (MC) experiment. The goal was to determine whether the algorithm consistently detects the correct signal regions, independently of favourable random iterations, and independently of the tightness of the confidence bands (as controlled by the parameter $\kappa$).

Each setting was run for 500 MC iterations, with the following configurations:

(A) **1 True Interval (Width (21 points) 42 nm)**: True interval from 898–938 nm (mapped to [25, 45] on the normalized 100-scale). Detection method also used a single interval with $h = 42$ nm.

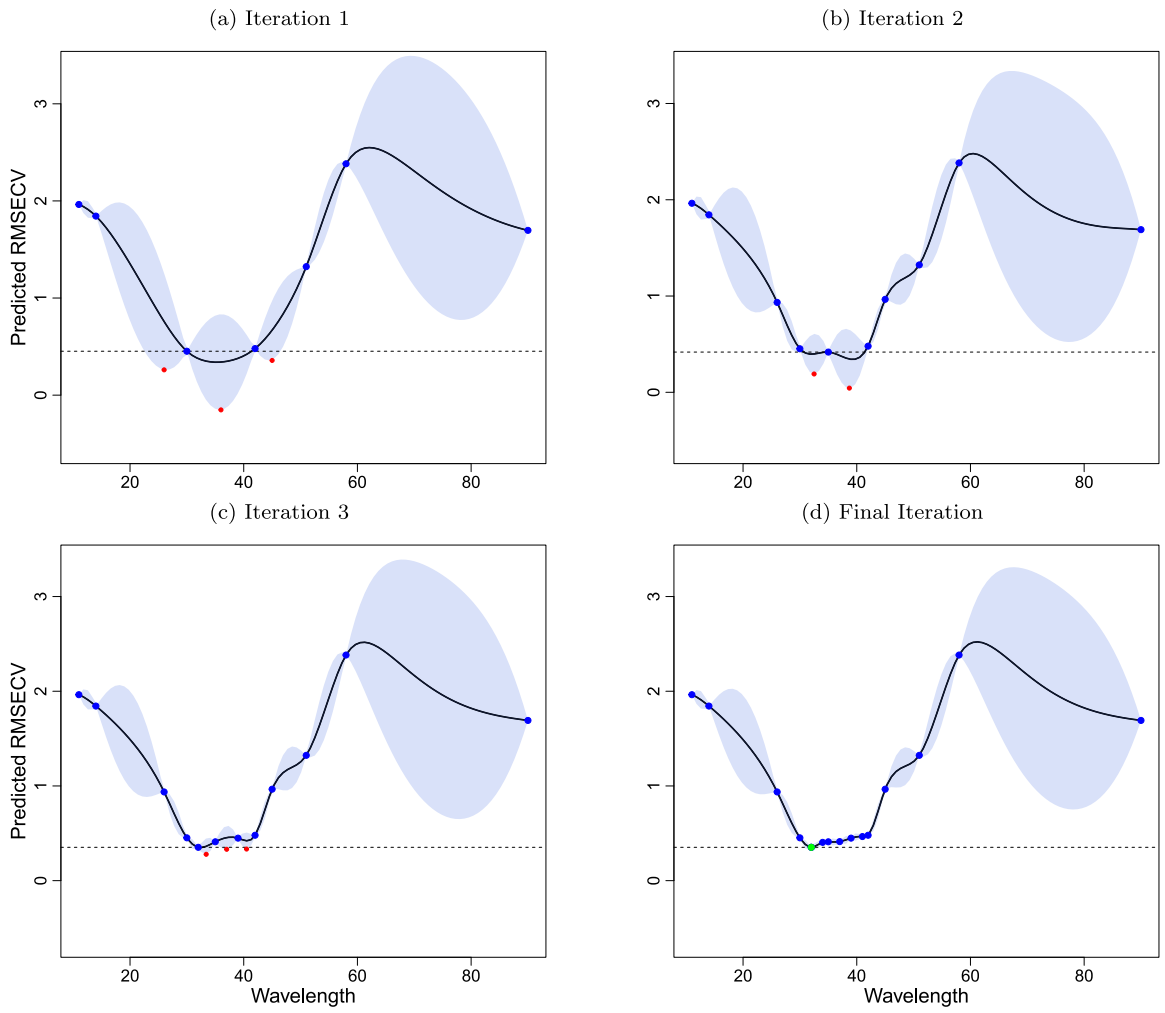


**Fig. 3.** Predicted posterior mean and uncertainty bands of the RMSECV across several iterations. Sampled interval centres are shown as blue dot (•). In subfigures (a-c), candidate intervals for the next evaluation are highlighted in red at the argmin of the lower uncertainty band (•). In subfigure (d), the final selected interval is highlighted in (⋆). The horizontal dashed line represents the dynamic performance threshold used for filtering candidates (see Algorithm 1, Step 7). For the first iteration (a), this threshold is the full-spectrum RMSECV, while for subsequent iterations it represents the best RMSECV found in the preceding step.

(B) **1 True Interval (Width (21 points) 42 nm)**: Same true interval (898–938 nm / [25, 45]). Detection used two intervals with $h = 42$ nm (each).

(C) **2 True Intervals (Width (11 points) 22 nm)**: Two true intervals, from 878–898 nm and 958–978 nm (mapped to [15, 25] and [55, 65]). Detection method used two intervals with $h = 22$ nm.

In all settings, data were simulated with noise as defined in Section 4.1. The initial sample size was fixed at $S_0 = 10$. Simulations were repeated for two values of $\kappa = \{1.5, 3\}$: corresponding to $1.5\sigma$ and $3\sigma$ uncertainty bands. For each configuration, we report the average Root Mean Squared Error (RMSE), final sample size, number of iterations, and detection accuracy across the 500 repetitions.

The evaluation of interval selection performance is based on four key metrics. The *Overlap over True* is defined as

$$\text{Overlap}_{\text{True}} = \frac{|\hat{I} \cap I|}{|I|},$$

where $I$ denotes the true informative interval and $\hat{I}$ the estimated one. This metric quantifies the proportion of the true region that was successfully recovered. Conversely, the *Overlap over Estimated* is given by

$$\text{Overlap}_{\text{Estimated}} = \frac{|\hat{I} \cap I|}{|\hat{I}|},$$

measuring how accurate or specific the predicted region was. Finally, we report the Centre Error, which measures the absolute difference between the centre of the estimated and true intervals, and the RMSECV to evaluate the predictive performance of the model over the test set. In addition to these, we also report the *final input sample size* and the *terminated step* of the algorithm to assess its computational behaviour and convergence characteristics.

***Results.*** The purpose of this study is to quantitatively assess the algorithm's accuracy in locating the true informative intervals (Table 1) and its computational efficiency (Tables 2) under various conditions. These results confirm that the proposed interval selection algorithm achieves both high predictive accuracy and efficient convergence across all Monte Carlo (MC) scenarios. In Scenario A, where a single true interval of width 21 is detected using a single estimated interval, the algorithm achieves Overlap over True and Estimated above 88% and 83%, respectively, with centre distances below 3.0. Most notably, RMSECV is reduced to 0.3574 with $\kappa = 1.5$, representing a substantial improvement over RMSECV of 0.531 obtained using a full-spectrum PLS model optimized with 7 latent components. This demonstrates that even a single adaptively selected interval can yield more accurate predictions than a full-spectrum model while significantly reducing model complexity.

Scenarios B and C introduce more complex structures that involve narrower or multiple informative intervals. Despite this, the method

**Table 1**
Summary of performance metrics across different Monte Carlo (MC) simulation settings, values between brackets represent standard errors. Each scenario (A, B, and C) corresponds to a specific combination of true interval configuration, interval selection method, and uncertainty width parameter $\kappa$. Scenario A evaluates the case where a single true interval of width 21 (located at 898–938 nm) is detected using a single estimated interval of the same width. Scenario B evaluates the same true interval as A but assumes detection via two estimated intervals of width 11, capturing more localized features within the same region. Scenario C represents the case with two disjoint true intervals: 878–898 nm and 958–978 nm, both of width 11. The detection method here uses two estimated intervals, each of width 11. The reported metrics are defined as follows: Overlap over True is the proportion of the true informative region recovered; Overlap over Estimated measures the specificity of the selected interval; and Centre Error is the absolute distance between the true and estimated interval centres. All reported RMSECV values can be compared against the benchmark RMSECV of 0.531 from a full-spectrum PLS model. Scenario definitions are as follows: (A) A single true interval detected with a single estimated interval. (B) A single true interval detected using two narrower intervals. (C) Two disjoint true intervals detected using two estimated intervals. Values in parentheses are standard errors.

| MC scenario | True interval | $h$ | Intervals selected | $\kappa$ | Overlap true | Overlap estimate | Centre error | RMSECV |
|---|---|---|---|---|---|---|---|---|
| A | 898–938 nm [25, 45] | 21 | 1 | 1.5 | 0.885 (0.030) | 0.843 (0.029) | 2.80 (0.60) | 0.3574 (0.0162) |
| | | | | 3.0 | 0.875 (0.000) | 0.833 (0.000) | 3.00 (0.00) | 0.3521 (0.0000) |
| B | 898–938 nm [25, 45] | 11 | 2 | 1.5 | – | 0.970 (0.080) | 2.66 (1.72) | 0.3508 (0.0035) |
| | | | | 3.0 | – | 1.000 (0.000) | 2.00 (0.00) | 0.3495 (0.0000) |
| C | 878–898 nm [15, 25] | 11 | 2 | 1.5 | 0.39 (0.145) | 0.355 (0.132) | 7.34 (4.61) | 0.3777 (0.0087) |
| | | | | 3.0 | 0.35 (0.000) | 0.318 (0.000) | 7.00 (0.00) | 0.3696 (0.0013) |
| | 958–978 nm [55, 65] | 11 | 2 | 1.5 | 0.768 (0.069) | 0.698 (0.062) | 2.82 (0.69) | 0.3777 (0.0087) |
| | | | | 3.0 | 0.751 (0.010) | 0.683 (0.009) | 2.99 (0.10) | 0.3696 (0.0013) |

maintains strong performance: RMSECVs remain below 0.351 in Scenario B, with overlap over estimated intervals reaching up to 1.000 and centre errors as low as 2.00. In this setting, where two narrow intervals are used to detect a single true interval of width $h = 21$ (42 nm), an Overlap over Estimated of 1.000 indicates that the two selected intervals join up to cover the real wider region. In Scenario C, where two disjoint intervals must be detected, the algorithm achieves Overlap over True exceeding 75% and Overlap over Estimated above 68%, with centre errors under 3.0. Predictive performance also remains high, with RMSECVs of 0.377 and 0.369 for $\kappa = 1.5$ and 3.0, respectively—both outperforming the full-spectrum baseline.

From a computational standpoint, Table 2 shows that the adaptive sampling strategy terminates in a small number of steps (typically 5–8), regardless of scenario complexity. Simpler cases such as Scenario A converge rapidly, requiring only 18–25 final input samples and fewer iterations (mean steps between 4.9 and 6.1). In contrast, more complex cases like Scenario C demand additional evaluations (up to 492 samples), as expected for recovering multiple disjoint regions. Importantly, empirical results show that convergence is consistently achieved regardless of the initial sample configuration or size, highlighting the robustness of the procedure, under both informative and non-informative priors.

Even under high uncertainty settings (i.e., when $\kappa = 3$), the algorithm demonstrates stable behaviour across all simulation scenarios. The selected intervals remain consistently aligned with the true informative regions, and both the overlap metrics and centre localization errors remain within tight bounds. This robustness under increased variance in the confidence bands highlights the method's ability to efficiently balance exploration and exploitation, avoiding overfitting to noise while still converging to relevant spectral features. Notably, although larger $\kappa$ values tend to induce broader uncertainty estimates, the algorithm compensates adaptively by requiring only a modest increase in the number of sampled points, maintaining convergence within a practical number of iterations.

Overall, the algorithm outperforms full-spectrum modelling in both performance and parsimony, making it especially suited for high-dimensional spectroscopic applications, where identifying and focusing on informative regions may be important.

## 5. Real data applications

In this paper we used two sets of NIR datasets, which we detail below, to validate our proposed method.

**Table 2**
Computational performance and convergence metrics for the Monte Carlo simulation. Final input sample size is the total number of unique interval centres sampled and evaluated by the algorithm before termination. Terminated step indicates the number of Bayesian optimization iterations required for convergence. All values are reported as the mean (and standard deviation in parentheses) over 500 MC repetitions.

| Scenario | $\kappa$ | RMSECV | Final input sample size | Terminated step |
|---|---|---|---|---|
| A | 1.5 | 0.3574 (0.0162) | 18.83 (3.88) | 4.94 (1.69) |
| | 3.0 | 0.3521 (0.0000) | 24.74 (5.09) | 5.20 (1.10) |
| B | 1.5 | 0.3508 (0.0035) | 126.85 (26.12) | 6.99 (1.22) |
| | 3.0 | 0.3495 (0.0000) | 305.83 (55.38) | 8.20 (1.12) |
| C | 1.5 | 0.3777 (0.0087) | 234.28 (51.06) | 6.88 (1.09) |
| | 3.0 | 0.3696 (0.0013) | 491.64 (91.16) | 8.23 (0.89) |

***Corn datasets.*** Four Near-Infrared (NIR) datasets for corn samples were obtained from http://www.eigenvector.com/data/Corn/index.html. Each dataset includes 80 corn samples measured using three different NIR spectrometers of the same type, with each raw spectrum comprising 700 wavelength points at 2 nm intervals, covering the range from 1100 to 2498 nm. Some limited trials with these data suggest that pretreatment does not substantially improve the results when using the full spectrum, so none was applied here. The properties of interest in these datasets are oil, protein, and starch content. Following the Kennard-Stone algorithm [22], each dataset was divided into a training set (60 samples) and an separate test set (20 samples).

***Diesel fuel datasets.*** Six NIR datasets for diesel fuels were obtained from http://www.eigenvector.com/data/SWRI/index.html. Each raw spectrum consists of 401 wavelength points taken at 2 nm intervals within the range of 750 to 1550 nm. The target properties for analysis include the boiling point at 50% recovery (boiling point), density, cetane number (CN), freezing temperature (Freeze), total aromatics (aromatics), and viscosity. In this data set the train-test distribution was not uniform and the partitions were 85/28 (boiling point), 85/28 (CN), 88/28 (Freeze), 88/30 (aromatics) and 87/29 (viscosity).

***Results.*** The performance of our proposed Bayesian optimization approach for interval selection was evaluated across multiple datasets

**Table 3**
Comparison of interval selection performance across different sigma-based uncertainty thresholds ($2\sigma$ and $3\sigma$) and interval widths (10% and 20%). Results are reported for the *Corn* (m5) dataset, considering different response variables. Metrics include RMSECV (train), RMSEP (test), number of PLS components, and selected intervals centres. The units for all RMSE values are the same as the units of the corresponding reference property. The number of PLS components was optimized for each model using 10-fold cross-validation on the training set. The row "Sig. Difference" indicates the models (Full PLS = 1, BiPLS = 2, FiPLS = 3) over which the proposed method showed statistically significant improvements.

| Dataset | Response variable | Metrics | $2\sigma$ | | $3\sigma$ | |
|---|---|---|---|---|---|---|
| | | | 10% | 20% | 10% | 20% |
| Corn (m5) | Oil | RMSECV (Train) | 0.014 | 0.015 | 0.014 | 0.015 |
| | | RMSEP (Test) | 0.018 | 0.018 | 0.018 | 0.018 |
| | | Num. Comp. | 12 | 16 | 12 | 16 |
| | | Selected intervals | (283, 599) | (323, 534) | (283, 599) | (323, 534) |
| | | Sig. Difference | 1 | 1,2,3 | 1 | 1,2,3 |
| | Starch | RMSECV (Train) | 0.072 | 0.063 | 0.072 | 0.063 |
| | | RMSEP (Test) | 0.119 | 0.062 | 0.119 | 0.062 |
| | | Num. Comp. | 15 | 15 | 15 | 15 |
| | | Selected intervals | (185, 350) | (371, 467) | (185, 350) | (371, 467) |
| | | Sig. Difference | 2,3 | 1,2,3 | 2,3 | 1,2,3 |
| | Protein | RMSECV (Train) | 0.013 | 0.029 | 0.013 | 0.029 |
| | | RMSEP (Test) | 0.010 | 0.022 | 0.010 | 0.022 |
| | | Num. Comp. | 16 | 17 | 16 | 17 |
| | | Selected intervals | (340, 514) | (297, 479) | (340, 514) | (297, 479) |
| | | Sig. Difference | 1,2,3 | 1,2,3 | 1,2,3 | 1,2,3 |

and response variables. Results are presented in terms of RMSEP (test), RMSECV (train), number of latent components, and selected intervals, under varying uncertainty thresholds and training set proportions. To assess whether the observed improvements over benchmark methods (Full PLS, BiPLS, FiPLS-see Tables B.1 and B.2) were statistically significant, we applied the test proposed [23], which evaluates differences in predictive ability based on RMSEP. Statistically significant improvements are indicated in the "Sig. Difference" rows of the tables.

For the Corn data set consider one interval method, with interval width of 10% (70 nm) and 20% (140 nm), with an uncertainty threshold of $2\sigma$ and $3\sigma$. Across all four response variables-Oil, Starch, and Protein—our method demonstrates consistent and statistically significant improvements over baseline methods. For Oil, the RMSEP values are consistently lower than Full PLS (0.018 vs. 0.066), and significance tests confirm better performance compared to all baselines (1, 2, 3) under all settings. In the case of Starch, our method shows gains especially under the 20% setting (RMSEP 0.062 vs. 0.119 for Full PLS), with significance against all three baselines. For Protein, test RMSE drops from 0.086 (Full PLS) to 0.013, with consistent statistical superiority (1, 2, 3) across all settings-showing that our method successfully isolates informative wavelength intervals (see Table 3 for a full summary).

For the Diesel data set consider one interval method, with interval width of 5% (20 nm), 10% (40 nm) and 20% (80 nm), with an uncertainty threshold of $2\sigma$ and $3\sigma$. Our method achieves competitive or superior performance in most cases. In Boiling Point, RMSEP drops to 2.249 under the 20% $2\sigma$ setting, outperforming Full PLS (4.151) and matching BiPLS and FiPLS in most cases. Significant gains are observed in rows 2 and 3, particularly where intervals such as (138, 246) and (198, 324) are selected. For Cetane Number, the method holds RMSEP near 2.16 while using only 3 latent components—with significant differences against baselines (1, 3), indicating effective dimension reduction. In Freeze, the performance is particularly strong: RMSEP falls to as low as 1.892, and significance holds across all three baselines for all configurations. Density shows extremely low errors (RMSEP = 0.00075), matching or outperforming all competitors and again achieving significance over all methods. For Total Aromatics, our method reduces test error while maintaining low complexity, significantly outperforming Full PLS and BiPLS in several configurations. Finally, in Viscosity, our method consistently beats Full PLS (0.066 vs. 0.102 RMSEP), with statistical differences over 2 and 3-confirming its robustness even in more challenging regression targets (the full results are presented in Table 4).

## 6. Discussion

In this paper, we set out to address limitations in interval selection for Partial Least Squares analysis of spectral data, aiming to improve flexibility and precision in identifying significant spectral regions. The primary objective was to develop a method that leverages Bayesian optimization within the Interval Partial Least Squares framework, enabling a more adaptive interval selection process. Our proposed approach combines random sampling and GP regression, introducing a probabilistic surrogate model that efficiently guides interval selection.

The novelty of this work lies in the application of Bayesian optimization to the interval selection process in chemometric analysis. Traditional ad-hoc interval selection methods, often lack the adaptability needed to capture the nuanced variations present in spectral data. Our Bayesian optimization-based approach, by contrast, is simple yet highly flexible, adapting to the data and allowing a more extensive search across potential intervals. This flexibility is particularly advantageous in high-dimensional settings, where exhaustive search approaches are computationally prohibitive.

A key feature of our algorithm is its ability to quantify uncertainty across all possible subintervals without requiring full exploration. The GP model serves as a surrogate, mapping sampled intervals to prediction accuracy measures and providing both mean and uncertainty estimates of the RMSE for untested intervals. This approach makes efficient use of available data, allowing us to pinpoint promising spectral regions even when only a subset of intervals has been evaluated.

To further assess the stability and robustness of our method, we performed a comprehensive Monte Carlo experiment under varying signal complexities and initial sampling conditions. The results demonstrated that the algorithm consistently converges to the correct informative regions with high accuracy, regardless of the true interval structure or the nature of the initial samples—whether informative or not. Across all scenarios, the method outperformed full-spectrum PLS in terms of predictive accuracy. These findings confirm that Bayesian optimization not only guides interval selection effectively, but also provides flexibility with respect to prior knowledge, making the approach broadly applicable in practical settings. These results confirm that Bayesian optimization not only guides interval selection effectively but also provides flexibility in terms of the initial knowledge input.

Moreover, our methodology accommodates the selection of multiple intervals, either simultaneously or sequentially, being the latter more computationally efficient since one can condition on a previously selected interval. This characteristic broadens the applicability

**Table 4**
Comparison of interval selection performance across different sigma-based uncertainty thresholds ($2\sigma$ and $3\sigma$) and interval widths (10% and 20%). Results are reported for the *Diesel* dataset, considering different response variables. Metrics include RMSECV (train), RMSEP (test), number of PLS components, and selected intervals. The units for all RMSE values are the same as the units of the corresponding reference property. The number of PLS components was optimized for each model using 10-fold cross-validation on the training set. The row "Sig. Difference" indicates the models (Full PLS = 1, BiPLS = 2, FiPLS = 3) for which the proposed method showed statistically significant improvements.

| Dataset | Response variable | Metrics | $2\sigma$ | | | $3\sigma$ | | |
|---|---|---|---|---|---|---|---|---|
| | | | 5% | 10% | 20% | 5% | 10% | 20% |
| Diesel | Boiling Point | RMSECV (Train) | 3.858 | 3.766 | 3.393 | 3.956 | 3.766 | 3.300 |
| | | RMSEP (Test) | 4.151 | 2.456 | 3.018 | 3.656 | 2.456 | 3.197 |
| | | Num. Comp. | 12 | 8 | 12 | 11 | 8 | 13 |
| | | Selected intervals | (158, 376) | (138, 246) | (198, 324) | (158, 368) | (138, 246) | (198, 321) |
| | | Sig. Difference | | 2,3 | | | 2,3 | |
| | Cetane Number | RMSECV (Train) | 1.968 | 1.974 | 1.956 | 1.968 | 1.974 | 1.963 |
| | | RMSEP (Test) | 2.178 | 2.169 | 2.161 | 2.178 | 2.169 | 2.169 |
| | | Num. Comp. | 3 | 3 | 3 | 3 | 3 | 3 |
| | | Selected intervals | (135, 383) | (125, 373) | (105, 353) | (135, 383) | (125, 373) | (105, 352) |
| | | Sig. Difference | 1,3 | 1,2 | 1 | 1,3 | 1,2 | 1 |
| | Freeze | RMSECV (Train) | 2.244 | 2.182 | 2.118 | 2.244 | 0.118 | 2.179 |
| | | RMSEP (Test) | 1.987 | 2.023 | 2.079 | 1.987 | 0.072 | 1.892 |
| | | Num. Comp. | 6 | 9 | 12 | 6 | 8 | 12 |
| | | Selected intervals | (162, 289) | (269, 329) | (104, 253) | (162, 289) | (134, 252) | (74, 250) |
| | | Sig. Difference | 1,2,3 | 1,2,3 | 1,2,3 | 1,2,3 | 1,2,3 | 1,2,3 |
| | Density | RMSECV (Train) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | | RMSEP (Test) | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | | Num. Comp. | 10 | 14 | 12 | 10 | 15 | 12 |
| | | Selected intervals | (159, 247) | (141, 256) | (215, 324) | (159, 247) | (142, 237) | (215, 324) |
| | | Sig. Difference | | | | | | |
| | Total Aromatics | RMSECV (Train) | 0.517 | 0.481 | 0.501 | 0.517 | 0.481 | 0.501 |
| | | RMSEP (Test) | 0.495 | 0.479 | 0.538 | 0.495 | 0.479 | 0.538 |
| | | Num. Comp. | 13 | 10 | 10 | 13 | 10 | 10 |
| | | Selected intervals | (138, 391) | (129, 377) | (109, 355) | (138, 391) | (129, 377) | (109, 355) |
| | | Sig. Difference | 1,3 | 1,3 | 2 | 1,2 | 1,3 | 2 |
| | Viscosity | RMSECV (Train) | 0.121 | 0.118 | 0.119 | 0.123 | 0.126 | 0.119 |
| | | RMSEP (Test) | 0.102 | 0.072 | 0.081 | 0.121 | 0.066 | 0.081 |
| | | Num. Comp. | 14 | 8 | 14 | 14 | 7 | 14 |
| | | Selected intervals | (155, 228) | (134, 252) | (62, 156) | (152, 228) | (137, 247) | (62, 156) |
| | | Sig. Difference | | | 2,3 | | | 2,3 |

**Table B.1**
Performance comparison between Full PLS, BiPLS, and FiPLS on the Corn (m5) dataset across different and interval widths (10% and 20%) and response variables and wavelength retention levels. Metrics reported include RMSE on train/test sets, number of latent variables, and percentage of retained wavelengths.

| Dataset | Response variable | Metrics | Full PLS | BiPLS | | FiPLS | |
|---|---|---|---|---|---|---|---|
| | | | | 10% | 20% | 10% | 20% |
| Corn (m5) | Oil | RMSE Train | 0.07342 | 0.01392 | 0.06629 | 0.01673 | 0.06070 |
| | | RMSE Test | 0.06627 | 0.01403 | 0.06980 | 0.01315 | 0.06980 |
| | | Num. Comp. | 7 | 14 | 8 | 12 | 9 |
| | | Retained Wavelength (%) | 100 | 10 | 40 | 20.3 | 60.4 |
| | Starch | RMSE Train | 0.824 | 0.119 | 0.837 | 0.095 | 0.861 |
| | | RMSE Test | 0.817 | 0.146 | 0.817 | 0.086 | 0.817 |
| | | Num. Comp. | 1 | 13 | 1 | 15 | 1 |
| | | Retained Wavelength (%) | 100 | 20.3 | 20.1 | 30.4 | 80.6 |
| | Protein | RMSE Train | 0.1277 | 0.019 | 0.077 | 0.016 | 0.056 |
| | | RMSE Test | 0.0863 | 0.013 | 0.079 | 0.013 | 0.034 |
| | | Num. Comp. | 11.0000 | 15 | 15 | 15 | 15 |
| | | Retained Wavelength (%) | 100.0 | 10.1 | 20.1 | 20.3 | 40.3 |

of our approach to cases where multiple spectral regions contribute to predictive accuracy, providing a more comprehensive view of the data.

Our Bayesian optimization-based interval selection method demonstrated consistent and statistically significant improvements over Full PLS, BiPLS, and FiPLS across multiple datasets and response variables. Using the RMSEP-based significance test described in Fearn [23], we confirmed that the selected intervals led to better predictive performance in most scenarios. On the Corn dataset, the method showed notably lower test errors – especially for Protein and Starch – while maintaining low model complexity. In the Diesel dataset, our approach performed competitively across all targets, with strong results for Boiling Point, Freeze, and Density. These findings highlight the method's ability to identify informative spectral regions and enhance prediction accuracy across diverse chemometric applications.

While our results demonstrate the effectiveness of the proposed framework, it is important to acknowledge its current limitations and avenues for future research. A key simplification in this study was the use of a fixed, pre-defined interval width. As noted, this may not be optimal when informative spectral bands differ in size. A major advantage of our BO approach is the potential handling of the interval width as a parameter in the surrogate model. However, this is a non-trivial extension. It would introduce a second dimension to the optimization and could create a more complex objective function where small changes in width lead to large changes in performance. Therefore, our focus on a fixed-width approach serves as a crucial first step in

**Table B.2**
Performance comparison between Full PLS, BiPLS, and FiPLS on the Diesel dataset across different and interval widths (10% and 20%) and response variables and wavelength retention levels. Metrics reported include RMSE on train/test sets, number of latent variables, and percentage of retained wavelengths.

| Dataset | Response variable | Metrics | Full PLS | BiPLS | | | FiPLS | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | 5% | 10% | 20% | 5% | 10% | 20% |
| Diesel | Boiling Point | RMSE Train | 3.599 | 3.626 | 3.735 | 3.579 | 3.176 | 3.552 | 3.707 |
| | | RMSE Test | 2.439 | 2.302 | 2.503 | 2.573 | 2.390 | 2.494 | 2.249 |
| | | Num. Comp. | 12 | 8 | 11 | 9 | 15 | 10 | 13 |
| | | Retained Wavelength (%) | 100 | 57.6 | 61.3 | 60.6 | 20.9 | 81.8 | 80.8 |
| | Cetane Number | RMSE Train | 2.097 | 1.931 | 2.135 | 2.005 | 1.984 | 1.905 | 2.019 |
| | | RMSE Test | 2.189 | 2.163 | 2.174 | 2.136 | 2.181 | 2.142 | 2.137 |
| | | Num. Comp. | 2 | 5 | 3 | 6 | 4 | 6 | 6 |
| | | Retained Wavelength (%) | 100 | 31.4 | 51.1 | 60.6 | 20.9 | 40.9 | 60.6 |
| | Freeze | RMSE Train | 2.463 | 2.423 | 2.342 | 2.414 | 2.180 | 2.285 | 2.416 |
| | | RMSE Test | 2.238 | 2.223 | 2.234 | 2.143 | 2.122 | 2.260 | 2.562 |
| | | Num. Comp. | 12 | 10 | 9 | 9 | 10 | 8 | 5 |
| | | Retained Wavelength (%) | 100 | 36.7 | 40.9 | 60.6 | 20.9 | 51.1 | 40.4 |
| | Density | RMSE Train | 0.00109 | 0.00088 | 0.00103 | 0.00106 | 0.00099 | 0.00104 | 0.00113 |
| | | RMSE Test | 0.00075 | 0.00060 | 0.00061 | 0.00057 | 0.00066 | 0.00077 | 0.00075 |
| | | Num. Comp. | 15 | 15 | 15 | 15 | 13 | 15 | 15 |
| | | Retained Wavelength (%) | 100 | 31.4 | 81.8 | 40.4 | 47.1 | 71.6 | 101.0 |
| | total Aromatics | RMSE Train | 0.603 | 0.531 | 0.507 | 0.563 | 0.514 | 0.539 | 0.515 |
| | | RMSE Test | 0.538 | 0.530 | 0.479 | 0.549 | 0.491 | 0.518 | 0.498 |
| | | Num. Comp. | 12 | 13 | 15 | 13 | 13 | 13 | 15 |
| | | Retained Wavelength (%) | 100 | 41.9 | 20.4 | 40.4 | 26.2 | 61.3 | 40.4 |
| | Viscosity | RMSE Train | 0.126 | 0.122 | 0.125 | 0.204 | 0.116 | 0.120 | 0.238 |
| | | RMSE Test | 0.071 | 0.059 | 0.062 | 0.163 | 0.063 | 0.065 | 0.157 |
| | | Num. Comp. | 11 | 12 | 12 | 7 | 11 | 13 | 3 |
| | | Retained Wavelength (%) | 100 | 89.0 | 71.6 | 40.4 | 78.6 | 61.3 | 60.6 |

establishing the viability of BO for this problem. Future work will concentrate on this simultaneous optimization of interval position and width to enhance the method's adaptability.

A key advantage of the proposed method is its computational strategy. Rather than fitting models for every possible interval or combination, Bayesian optimization leverages a probabilistic model to prioritize regions with high expected gains. This reduces the number of PLS evaluations required and eliminates the need for manual tuning of thresholds or stopping criteria, making the method both efficient and fully automated.

It is also important to discuss the computational cost in more detail. The primary efficiency gain of our framework is in sample efficiency—that is, minimizing the number of expensive function evaluations (fitting and validating a PLS model), rather than raw CPU time. An exhaustive grid search might evaluate hundreds of intervals, whereas our method intelligently selects only the most promising candidates. Our Monte Carlo study provides a quantitative illustration of this efficiency. As shown in Table 2, the algorithm consistently converges in a small number of iterations, typically between 5 and 8 steps. For a single-interval search, this required the evaluation of only 20–25 PLS models on average to find the optimal region. This stands in stark contrast to grid-based methods, whose cost scales linearly for a single interval but grows combinatorially when searching for multiple intervals, quickly becoming computationally prohibitive. This highlights the practical value and scalability of the Bayesian Optimization approach.

## CRediT authorship contribution statement

**Nicolás Hernández:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Yoonsun Choi:** Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tom Fearn:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

## Ethics approval

Authors have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

## Consent for publication

Authors give consent for publication.

## Declaration of competing interest

None of the authors has a conflict of interest.

## Acknowledgements

We thank the reviewers and associate editor for all the valuable comments and suggestions, which helped us to improve the quality of the manuscript.

## Appendix A. Bayesian optimization - iPLS

See Algorithm 1

## Appendix B. Full PLS and interval-PLS models

(See Tables B.1 and B.2).

## Data availability

Data is available on Eigenvector.

**Algorithm 1** Bayesian optimization for Interval Selection in PLS

**Require:** Spectral data $\mathbf{Z} \in \mathbb{R}^{n \times p}$; Response vector $\mathbf{w} \in \mathbb{R}^n$; Interval width $h$; Exploration parameter $\kappa$; Initial sample size $S_0$; Max iterations $J_{\max}$.

**Initialization**

**Step 0:** Train a benchmark PLS model using the full spectrum and compute the baseline RMSECV, denoted $y_F$

**Step 1:** Randomly sample $S_0$ initial interval centres $\mathbf{x} = \{x_1, \ldots, x_{S_0}\}$ from the spectral domain, ensuring boundary coverage.

**Step 2:** For each initial centre $x_i$, train a PLS model on the interval $[x_i - h/2, x_i + h/2]$ and compute its corresponding $y$ =RMSECV.

**Step 3:** Form the dataset $\mathcal{D}_0 = \{(x_i, y_i)\}_{i=1}^{S_0}$

**Bayesian optimization Loop**

Set iteration counter $j = 0$.

Repeat the following steps:

**Step 4** (Fit Surrogate): Fit a Gaussian Process model to the current dataset $\mathcal{D}_j$)

**Step 5** (Acquisition Function): Use the fitted GP to compute the acquisition function over the entire domain, $\alpha(x; \mathcal{D}_j) = \kappa\sigma(x) - \mu(x)$.

**Step 6** (Find Candidates): Identify all local minima of the acquisition function $\alpha(x; \mathcal{D}_j)$. Denote this set of candidate centres as $\{x_{j,k}\}_{k=1}^{K_j}$

**Step 7** (Filter Candidates): For each candidate $x_{j,k}$, retain it only if its Lower Credible Bound satisfies the following condition:

$$\mu(x_{j,k}) - \kappa\sigma(x_{j,k}) < \begin{cases} y_F, & \text{if } j = 1 \\ \min\limits_{k'} y_{j-1,k'}, & \text{if } j > 1 \end{cases}$$

Let the set of filtered candidates be $\{\tilde{x}_{j,k}\}_{k=1}^{\tilde{K}_j}$

**Step 8** (Check Convergence): If the set of filtered candidates is empty ($\tilde{K}_j = 0$), or a pre-defined maximum number of iterations is reached, terminate the loop.

**Step 9** (Evaluate & Update): For each filtered candidate $\tilde{x}_{j,k}$, train a new PLS model and compute its RMSECV, $\tilde{y}_{j,k}$

**Step 10**: Augment the dataset: $\mathcal{D}_{j+1} = \mathcal{D}_j \cup \left\{(\tilde{x}_{j,k}, \tilde{y}_{j,k})\right\}_{k=1}^{\tilde{K}_j}$

**Step 11:** Increment iteration counter: $j \leftarrow j + 1$.

## References


[1] Yong-Huan Yun, Hong-Dong Li, Bai-Chuan Deng, Dong-Sheng Cao, An overview of variable selection methods in multivariate analysis of near-infrared spectra, TRAC Trends Anal. Chem. 113 (2019) 102–115.

[2] L. Nørgaard, A. Saudland, J. Wagner, J. Pram Nielsen, L. Munck, S. Balling Engelsen, Interval partial least-squares regression (i pls): A comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (3) (2000) 413–419.

[3] Isabelle Guyon, Andre Elisseeff, An introduction to variable and feature selection.

[4] Girish Chandrashekar, Ferat Sahin, A survey on feature selection methods, Comput. Electr. Eng. (ISSN: 0045-7906) 40 (1) (2014) 16–28, http://dx.doi.org/10.1016/j.compeleceng.2013.11.024, URL https://www.sciencedirect.com/science/article/pii/S0045790613003066.

[5] J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, Microchem. J. (ISSN: 0026-265X) 47 (1) (1993) 60–66, http://dx.doi.org/10.1006/mchj.1993.1012, URL https://www.sciencedirect.com/science/article/pii/S0026265X8371012X.

[6] F. Guillaume Blanchet, Pierre Legendre, Daniel Borcard, Forward selection of explanatory variables, Ecology (ISSN: 1939-9170) 89 (9) (2008) 2623–2632, http://dx.doi.org/10.1890/07-0986.1, URL https://onlinelibrary.wiley.com/doi/abs/10.1890/07-0986.1, _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/07-0986.1.

[7] Riccardo Leardi, Application of genetic algorithm-PLS for feature selection in spectral data sets, J. Chemom. 14 (2000) 643–655, http://dx.doi.org/10.1002/1099-128X(200009/12)14:5/6<643::AID-CEM621>3.0.CO;2-E, John Wiley and Sons Ltd, Issue: 5-6 Journal Abbreviation: J. Chemometr.

[8] Riccardo Leardi, Genetic algorithms in chemometrics and chemistry: a review, J. Chemom. (ISSN: 1099-128X) 15 (7) (2001) 559–569, http://dx.doi.org/10.1002/cem.651, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.651, _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.651.

[9] Sergey Kucheryavskiy, mdatools–r package for chemometrics, Chemometr. Intell. Lab. Syst. 198 (2020) 103937.

[10] Svante Wold, Michael Sjöström, Lennart Eriksson, Pls-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109–130.

[11] Xiaobo Zou, Jiewen Zhao, Yanxiao Li, Selection of the efficient wavelength regions in ft-nir spectroscopy for determination of ssc of 'Fuji'apple based on bipls and fipls models, Vib. Spectrosc. 44 (2) (2007) 220–227.

[12] Riccardo Leardi, Lars Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom.: A J. Chemom. Soc. 18 (11) (2004) 486–497.

[13] Lars Munck, J. Pram Nielsen, Birthe Møller, Susanne Jacobsen, Ib Søndergaard, S.B. Engelsen, L. Nørgaard, R. Bro, Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics, Anal. Chim. Acta 446 (1–2) (2001) 169–184.

[14] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, Nando De Freitas, Taking the human out of the loop: A review of bayesian optimization, Proc. IEEE 104 (1) (2015) 148–175.

[15] Roman Garnett, Bayesian Optimization, Cambridge University Press, 2023.

[16] Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, Si-Hao Deng, Hyperparameter optimization for machine learning models based on bayesian optimization, J. Electron. Sci. Technol. 17 (1) (2019) 26–40.

[17] Marc C. Kennedy, Anthony O'Hagan, Predicting the output from a complex computer code when fast approximations are available, Biometrika 87 (1) (2000) 1–13.

[18] Jasper Snoek, Hugo Larochelle, Ryan P. Adams, Practical Bayesian Optimization of Machine Learning algorithms, in: Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., 2012, URL https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

[19] Michael L. Stein, Interpolation of spatial data, in: Springer Series in Statistics, Springer, New York, NY, 1999, http://dx.doi.org/10.1007/978-1-4612-1494-6, ISBN: 978-1-4612-7166-6 978-1-4612-1494-6, URL http://link.springer.com/10.1007/978-1-4612-1494-6.

[20] Aurore Delaigle, Peter Hall, et al., Methodology and theory for partial least squares applied to functional data, Ann. Stat. 40 (1) (2012) 322–352.

[21] R.J. Barnes, Mewa Singh Dhanoa, Susan J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (5) (1989) 772–777.

[22] Ronald W. Kennard, Larry A. Stone, Computer aided design of experiments, Technometrics 11 (1) (1969) 137–148.

[23] Tom Fearn, Testing differences in predictive ability: A tutorial, J. Chemom. (2024) e3549.