Received: 14 July 2023 | Revised: 15 November 2023 | Accepted: 14 December 2023

DOI: 10.1002/sam.11657

WILEY

RESEARCH ARTICLE

# Nonparametric Bayesian functional clustering with applications to racial disparities in breast cancer

**Wenyu Gao**[1] | **Inyoung Kim**[2] | **Wonil Nam**[3] | **Xiang Ren**[4] | **Wei Zhou**[4] | **Masoud Agah**[3]

[1]Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, North Carolina, USA

[2]Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

[3]Department of Electronic Engineering, Pukyong National University, Busan, Republic of Korea

[4]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

**Correspondence**
Inyoung Kim, Department of Statistics, Virginia Tech., Blacksburg, VA 24061, USA.
Email: inyoungk@vt.edu

**Funding information**
National Science Foundation, Grant/Award Number: 1711699

## Abstract

As we have easier access to massive data sets, functional analyses have gained more interest. However, such data sets often contain large heterogeneities, noises, and dimensionalities. When generalizing the analyses from vectors to functions, classical methods might not work directly. This paper considers noisy information reduction in functional analyses from two perspectives: functional clustering to group similar observations and thus reduce the sample size and functional variable selection to reduce the dimensionality. The complicated data structures and relations can be easily modeled by a Bayesian hierarchical model due to its flexibility. Hence, this paper proposes a nonparametric Bayesian functional clustering and peak point selection method via weighted Dirichlet process mixture (WDPM) modeling that automatically clusters and provides accurate estimations, together with conditional Laplace prior, which is a conjugate variable selection prior. The proposed method is named WDPM-VS for short, and is able to simultaneously perform the following tasks: (1) Automatic cluster without specifying the number of clusters or cluster centers beforehand; (2) Cluster for heterogeneously behaved functions; (3) Select vibrational peak points; and (4) Reduce noisy information from the two perspectives: sample size and dimensionality. The method will greatly outperform its comparison methods in root mean squared errors. Based on this proposed method, we are able to identify biological factors that can explain the breast cancer racial disparities.

**KEYWORDS**

functional clustering, nonparametric Bayesian model, peak point selection, surface-enhanced Raman spectroscopy, WDPM-VS, weighted Dirichlet process mixture

## 1 | INTRODUCTION

Functional analyses have gained more interest as we have easier access to massive data sets. However, such data sets often contain large heterogeneities, noises, and dimensionalities. When generalizing the analyses from vectors to functions, classical methods might not work directly. Thus, noisy information reduction is necessary.

It is typical to consider the noisy information reduction in functional analyses from two perspectives: functional clustering to group similar observations and thus reduce the sample size, as well as functional variable selection to reduce the dimensionality.

It is known that cluster analyses group objects into several clusters such that the objects from the same cluster are more homogeneous than objects from other clusters are. The center of each cluster can represent all the objects within the cluster. As a result, the sample size can be reduced to the number of clusters. Functional cluster analysis is a generalization of point clustering in that the objects are functions. After performing the functional clustering, one only needs to analyze the cluster centers instead of analyzing the original massive heterogeneous functions, which greatly reduces the amount of noise information. Traditional frequentist clustering methods, such as agglomerative hierarchical clustering or k-means, require the number of clusters or positions of cluster centers to be known in advance, which is quite difficult in reality. Furthermore, these methods are not easy to generalize to functional clustering because the definition of similarity measure, such as distance between functions, is not apparent, while traditional cluster analyses highly depend on the similarity metrics. On the contrary, model-based clustering methods, such as finite mixture models, relax the requirements of knowing the number of clusters beforehand by considering it a latent variable [1]. They also allow us to provide statistical inference. However, this requires the prior distribution assumptions. Due to the lack of definition for distribution on functions, as well as the potential of misspecification, a nonparametric model-based clustering method is favored, such as a Dirichlet process mixture (DPM) model [2–5] in Bayesian approach.

This DPM model was proposed by Ferguson [6] and is widely applied [3, 7, 8] due to the property that it automatically clusters. The new observation will be assigned to an existing cluster or a new one, depending on different probabilities. The probabilities of assigning to existing clusters depend on the cluster size, while the probability of assigning to a new cluster depends on a precision parameter. As a result, the larger the number of observations in one cluster, the more probable the new observation will be assigned to this cluster. These properties can be easily interpreted by the Pólya urn representation [2, 4, 5]. As a result, the DPM model does not require the assumption of the number of clusters beforehand. Furthermore, the DPM model requires no distribution family specified as prior, and thus allows for more flexibilities. However, it still requires the assumption of homogeneity. That is, all the observations should share the same prior distribution. Moreover, the clustering results have their own characteristic; that is, the more observations in one cluster, the higher chance that a new observation will be assigned there, but do not take information from data. To further relax the homogeneity assumption and take more advantage of data, a weighted Dirichlet process mixture (WDPM) prior is considered. The WDPM was enlightened by Zellner [9] and applied by Dunson et al. [10]. Instead of assuming all observations share the same prior distribution, as in the DPM, the WDPM allows for multiple candidate prior distributions. Each observation is assigned to one prior distribution with some weight. Thus, the WDPM prior can be seen as a weighted mixture of several DPM priors. The construction of the weight functions can take usage of data information. Dunson et al. [10] proposed a Gaussian-type weight function that depends on the Euclidean distances. Sun et al. [11] further proposed some modifications to those weight functions.

The weighted Dirichlet process (WDP) structure is a special case of the dependent Dirichlet process (DDP) [5, 12, 13], which models the dependency among multiple Dirichlet processes. Müller et al. [14] considered a weighted average between two independent Dirichlet processes, while the WDP allows for multiple processes. There are other well-known methods accounting for different data structures, such as hierarchical structure (HDP) [15] and nested structure (nested DP) [16, 17]. However, our focus is on observations of the same levels. As a result, in this paper, we will only concentrate on the WDP structure. The existing literature studying the WDP mixture (WDPM) models did not focus on functional cluster analyses. Dunson et al. [10] concentrated on density estimations, while Sun et al. [18] applied the WDPM to the error distribution rather than the mean functions. In this paper, we propose a functional clustering method using WDPM. The number of possible candidate priors is also carefully examined. Performances are compared to the traditional DPM priors. Simulation results show that the WDPM will always outperform the DPM priors in terms of root mean squared error (RMSE) on estimated response values. Furthermore, the RMSE will vary by the number of possible candidate priors. More details are explained in Section 3.

Apart from clustering methods, in functional cluster analyses, it is also important to identify peak points on the trends of the cluster functions because they can quantify the unique characteristics among clusters. One possible approach for peak point selection is to apply the variable selection techniques to the potential changing points. The least absolute shrinkage and selection operator (LASSO) [19] is among the most commonly used methods for variable selection, especially for linear models. Thus, we would like to consider Bayesian LASSO, combined with the WDPM prior to perform functional clustering and change point selection simultaneously. Tibshirani [19]

proposed a Bayesian version of LASSO as the maximum a posteriori (MAP) estimate from an independent and identical Laplace (double-exponential) prior. Park and Casella [20] further proposed a conditional Laplace prior that can be extended to a hierarchical conjugate prior, such that the posteriors can be achieved using Gibbs samplers.

In this paper, we want to propose a method that unites both the WDPM and conditional Laplace priors, so that it can perform functional clustering and peak point selection at the same time. To the best of our knowledge, there is no literature studying functional clustering through the WDPM method and perform peak point selection simultaneously. We call our method WDPM-VS, which stands for the combination of WDPM and variable selection. Our comparison method using the DPM prior is similarly called DPM-VS. The performance of our proposed WDPM-VS method is evaluated and compared to DPM-VS method through simulation studies. We will then apply our proposed WDPM-VS method to study breast cancer racial disparities using surface-enhanced Raman spectroscopy (SERS) data, and compare the results with DPM-VS method.

Our main contributions proposing the WDPM-VS method lie in several aspects. To begin with, this method will perform automatic functional clustering and peak point selection at the same time. In this way, we can simultaneously reduce the noise information in functional data analyses from the two perspectives mentioned previously: reduce the sample size and reduce the dimensionality. Thus, this method is beneficial to study SERS data, as this kind of data typically will have intraclass heterogeneities within massive curves. Our proposed method is able to reduce the mass and group the heterogeneities by functional clustering. Concurrently, our method will let the characteristics for each cluster stand out by peak point selection. The selected peak points are useful to identify biological factors explaining the racial disparities. We will elaborate the real data analysis in Section 4. Besides extracting key information from the massive heterogeneity curves with only one model, our proposed method is also shown to outperform the comparison methods, especially with heterogeneity data, through simulation studies.

The above mentioned advantages of our method depend heavily on the Bayesian approach due to various reasons. First, the complicated data structures and relations can be easily modeled by a Bayesian hierarchical model, or developed from a more generic one by changing the prior distributions. That is why we can perform noisy information reduction from the two perspectives through only one model. Second, our proposed method is built from the nonparametric Bayesian method Dirichlet process. This method will cluster automatically without specifying the number of clusters or the cluster centers beforehand. Moreover, there is no distribution assumptions required. Thus, it is easy to generate to functional clustering. The flexibility of Bayesian approach not only provides us a superior method to perform functional clustering, but also makes it possible to perform multiple tasks at the same time. Hence, this paper focuses on the development of Bayesian approaches for functional analyses.

This paper is organized as follows: in Section 2, we first explain the functional clustering with the WDPM prior. Then we propose our WDPM-VS prior with nonparametric modeling to perform functional clustering and peak point selection at the same time. Simulation studies are conducted to evaluate the performances of our proposed method in Section 3. In Section 4, we study the breast cancer racial disparities with our WDPM-VS method. Lastly, our concluding remarks are presented in Section 5.

## 2 | NONPARAMETRIC BAYESIAN METHOD

### 2.1 | Functional clustering using WDPM

Consider an unknown relationship:

$$y_{ij} = f_i(\mathbf{X}_{ij}) + \varepsilon_{ij}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, n_i,$$

where $y_{ij}$ is a continuous response variable changing over observation $j$ ($j = 1, \ldots, n_i$) from subject $i$ ($i = 1, \ldots, n$); $n_i$ is the total number of observations for subject $i$, and there are total $n$ subjects; $\mathbf{X}_{ij}$ contains all the $p$ predictor variables at the $j$th observation from the $i$th subject; $p$ represents the total number of covariates; $f_i(\cdot)$ is an unknown function of predictor $\mathbf{X}$s for subject $i$; and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the error term for observation $j$ from subject $i$.

The unknown function $f_i(\cdot)$ can be estimated using a Bayesian approach by considering a prior

$$f_i(\cdot) \mid F \sim F.$$

$F$ does not need to be a known distribution family; rather, it can be an unknown distribution with special characteristics. For example, we could let $F$ have DPM or WDPM priors to perform functional clustering. A DPM prior has two parameters: a precision parameter $\alpha$, which controls the total number of clusters, and a base distribution $F_0$, which determines the characteristics of the model. Thus, we can write

$$f_i \mid F \sim F, F \sim DP(\alpha, F_0),$$

if we let $F$ have the DPM prior. A WDPM prior can be seen as a mixture of $Q$ DPM priors. If we define $Z_i$ as the

indicator of the candidate prior for observation $i$, then

$$Z_i \sim multinomial(1, 2, \cdots, Q; \mathbf{b}_i),$$

where $\mathbf{b}_i = \{b_{i1}, b_{i2}, \cdots, b_{iQ}\}$ are the weight functions of assigning observation $i$ to candidate $q, q = 1, 2, \cdots, Q$. Observations assigned to the same candidate $q$ will share the same DPM prior $DP_q$, that is,

$$f_i \mid (F_1, \cdots, F_Q), Z_i = q \sim F_q, F_q \sim DP_q(\alpha, F_0), \quad q = 1, \cdots, Q.$$

To avoid over-heterogeneity, the Dirichlet processes do not need to be distinguishable. The distinct unknown priors $F_q$ already control the differences. Thus, we consider the prior setting as

$$f_i \mid (F_1, \cdots, F_Q), Z_i = q \sim F_q, F_q \sim DP(\alpha, F_0), q = 1, \cdots, Q.$$

The probabilities for latent variable $Z_i, i = 1, 2, \cdots, n$ are determined by the weight functions $\mathbf{b}_i, i = 1, 2, \cdots, n$. Dunson et al. [10] suggested using

$$b_{iq} = \frac{\gamma_q e^{-\psi \left\| \mathbf{X}_i - \mathbf{X}_q^c \right\|^2}}{\sum_{l=1}^{Q} \gamma_l e^{-\psi \left\| \mathbf{X}_i - \mathbf{X}_l^c \right\|^2}}, q = 1, 2, \cdots, Q,$$

with total number of candidates $Q = n$, the sample size. Here, $\mathbf{X}_q^c$ represents the center value of $\mathbf{X}_i$ s from candidate $q$. However, Sun et al. [11] pointed out that there would be identifiable problems between the hyperparameters $\psi$ and $\gamma$, so they simplified it to an efficient version

$$b_{iq} = \frac{e^{-\psi \left\| \mathbf{X}_i - \mathbf{X}_q^c \right\|^2}}{\sum_{l=1}^{Q} e^{-\psi \left\| \mathbf{X}_i - \mathbf{X}_l^c \right\|^2}}, q = 1, 2, \cdots, Q.$$

To add the variability on different candidates, they also considered a more flexible version

$$b_{iq} = \frac{e^{-\psi_q \left\| \mathbf{X}_i - \mathbf{X}_q^c \right\|^2}}{\sum_{l=1}^{Q} e^{-\psi_l \left\| \mathbf{X}_i - \mathbf{X}_l^c \right\|^2}}, q = 1, 2, \cdots, Q.$$

This weight function might have an overlapping effect with selecting the total number of candidates $Q$. Imagine an extreme case in which $\psi_q$ is extremely large for candidate $q$, then the probability $b_{iq}$ will become negligible. That is equivalent to dropping candidate $q$ and reducing the number of $Q$. As a result, we will focus on the efficient weight function that treats all $\psi_q$ as the same and examine the performance based on the number of $Q$.

## 2.2 | Nonparametric function estimation

To model the unknown function $f_i(\cdot)$, we consider an $s^{th}$-order regression spline. The regression spline is commonly used as a nonparametric approach with fewer parameters, especially compared to wavelet bases. It is also useful for easy interpretations. The locations of bases are determined by the data, so the selection of bases numbers and bases locations can be interpreted by the data as well. In this paper, we use bases selection to perform peak point selection. The unknown function $f_i(X_{ij})$ for the $j^{th}$ observation from the $i^{th}$ subject can be modeled by

$$f_i(\mathbf{X}_{ij}) = \beta_{0i} + \beta_{1i}\mathbf{X}_{ij} + \cdots + \beta_{si}\mathbf{X}_{ij}^s + \sum_{k=1}^{K} \beta_{s+k,i} (\mathbf{X}_{ij} - \boldsymbol{\xi}_k)_+^s,$$

where $K$ is the total number of knots and $\boldsymbol{\xi}_k$ is the position of knot $k$. The function $(\mathbf{X}_{ij} - \boldsymbol{\xi}_k)_+^s$ equals $(\mathbf{X}_{ij} - \boldsymbol{\xi}_k)^s$ if $\mathbf{X}_{ij} - \boldsymbol{\xi}_k \geqslant 0$, and equals 0 otherwise.

If we define

$$\mathbf{X}_{ij}^{(p)} = \left(\mathbf{1} \ \mathbf{X}_{ij} \ \cdots \ \mathbf{X}_{ij}^s\right),$$
$$\boldsymbol{\beta}_i^{(p)} = \left(\beta_{0i} \quad \beta_{1i} \quad \cdots \quad \beta_{si}\right)',$$
$$\mathbf{X}_{ij}^{(np)} = \left\{ (\mathbf{X}_{ij} - \xi_1)_+^s (\mathbf{X}_{ij} - \xi_2)_+^s \cdots (\mathbf{X}_{ij} - \xi_K)_+^s \right\},$$

and

$$\boldsymbol{\beta}_i^{(np)} = \left(\beta_{s+1,i} \quad \beta_{s+2,i} \quad \cdots \quad \beta_{s+K,i}\right)',$$

then the unknown function $f_i(\mathbf{X}_{ij})$ can be written as

$$f_i(\mathbf{X}_{ij}) = \mathbf{X}_{ij}^{(p)} \boldsymbol{\beta}_i^{(p)} + \mathbf{X}_{ij}^{(np)} \boldsymbol{\beta}_i^{(np)}.$$

We call $\mathbf{X}_{ij}^{(p)}$ and $\boldsymbol{\beta}_i^{(p)}$ the parametric component, while $\mathbf{X}_{ij}^{(np)}$ and $\boldsymbol{\beta}_i^{(np)}$ are the nonparametric component. The parametric component relates to polynomial basis functions, which can describe the global behaviors, while the nonparametric component relates to truncated basis functions, which can capture the local behaviors. With this formatting, the nonparametric model can be written as

$$y_{ij} = \mathbf{X}_{ij}^{(p)} \boldsymbol{\beta}_i^{(p)} + \mathbf{X}_{ij}^{(np)} \boldsymbol{\beta}_i^{(np)} + \varepsilon_{ij}, i = 1, 2, \cdots, n,$$
$$j = 1, 2, \cdots, n_i,$$

and the likelihood function is

$$L = \left(2\pi\sigma^2\right)^{-\sum_{i=1}^{n} n_i/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left( y_{ij} - \mathbf{x}_{ij}^{(p)} \boldsymbol{\beta}_i^{(p)} - \mathbf{x}_{ij}^{(np)} \boldsymbol{\beta}_i^{(np)} \right)^2 \right\},$$

where $n_i$ is the number of observations for subject $i$ and there are $n$ total subjects.

## 2.3 | Vibrational peak selection

With the introduction of parameter $\boldsymbol{\beta}$ from the regression spline, we can project the priors for $f_i$ to the priors for $\boldsymbol{\beta}_i = \left\{ \left(\boldsymbol{\beta}_i^{(p)}\right)' \left(\boldsymbol{\beta}_i^{(np)}\right)' \right\}'$. To distinguish the prior for functions from prior for vectors, we write the priors as $\boldsymbol{\beta}_i \mid G_0 \sim G_0$ for the parametric Bayesian setting; $\boldsymbol{\beta}_i \mid G \sim G, G \sim DP(\alpha, G_0)$ for the DPM setting; and $\boldsymbol{\beta}_i \mid \left(G_1, \cdots, G_Q\right), Z_i = q \sim G_q, G_q \sim DP(\alpha, G_0), q = 1, \cdots, Q,$ for the WDPM setting.

By projection to priors on $\boldsymbol{\beta}$, selection on the parametric portion $\boldsymbol{\beta}^{(p)}$ s helps determine the shape of the unknown function $f$, while selection on the nonparametric portion $\boldsymbol{\beta}^{(np)}$ s indicates the important changing points. Both are useful to explain the unknown behavior $f$. Thus, we consider bases selection on both parametric and nonparametric components as peak point selection, which is equivalent to variable selection on all parameters $\boldsymbol{\beta}$. To perform the variable selection, we can set $G_0$ to a conditional Laplace distribution for conjugacy [20] with probability density

$$p\left(\boldsymbol{\beta}_i|\sigma^2\right) = \prod_{d=0}^{s+K} \frac{\lambda_i}{2\sqrt{\sigma^2}} e^{-\lambda_i|\beta_{i,d}|/\sqrt{\sigma^2}},$$

where $\sigma^2$ is the common variance for the model errors. It is essential to condition on the $\sigma^2$ because it assures a unimodal posterior. With this conditional Laplace prior, we can derive an equivalent set of hierarchical conjugate priors by introducing a set of parameters $\left\{\tau_{i,0}^2, \tau_{i,1}^2, \tau_{i,2}^2, \cdots, \tau_{i,s+K}^2\right\}$ with the same dimension as $\boldsymbol{\beta}_i$:

$$\begin{aligned} &\boldsymbol{\beta}_i \mid \sigma^2, \left\{\tau_{i,0}^2, \tau_{i,1}^2, \tau_{i,2}^2, \cdots, \tau_{i,s+K}^2\right\} \\ &\quad \sim N\left\{\mathbf{0}, \sigma^{2*} diag\left(\tau_{i,0}^2, \tau_{i,1}^2, \tau_{i,2}^2, \cdots, \tau_{i,s+K}^2\right)\right\}, \\ &\tau_{i,d}^2, d = 0,1,2, \cdots, s \sim Exponential\left\{\frac{\lambda_i^{2,(p)}}{2}\right\}, \\ &\tau_{i,d}^2, d = s+1, \cdots, s+K \sim Exponential\left\{\frac{\lambda_i^{2,(np)}}{2}\right\}. \end{aligned} \quad (1)$$

The hyperparameter $\lambda_i$ is set to be different for the parametric and nonparametric components, and labeled $\lambda_i^{(p)}$ and $\lambda_i^{(np)}$. This is because the nonparametric component for $\boldsymbol{\beta}_i$ controls the local behaviors and needs a smoothing parameter. Thus, $\left\{\tau_{i,s+1}^2, \cdots, \tau_{i,s+K}^2\right\}$, corresponding to the nonparametric parameters, need to separate out from $\left\{\tau_{i,0}^2, \tau_{i,1}^2, \cdots, \tau_{i,s}^2\right\}$, corresponding to the parametric component. $\lambda_i^2/2$ is the rate parameter for the exponential distribution. Additionally, if we integrate out the parameters $\boldsymbol{\tau}_i$, we will achieve the original conditional Laplace prior.

If we set $G_0$ according to (1), we will obtain the parametric Bayesian prior. We are setting different priors for each observation $i$; thus, the model is fitted separately over observations and will lose the generality of the whole data set. However, examining the overall relationship is our top priority, though it eliminates individual differentiations. Therefore, we set all $\boldsymbol{\beta}_i$ s as the same and only fit a united relationship for the whole data set. The parametric Bayesian prior cannot perform functional clustering, so that we will focus on the DPM and WDPM priors. Yet, the nonparametric Bayesian priors are also comparable to the parametric Bayesian priors if we set the base distribution $G_0$ as (1).

The hyperparameter $\lambda_i$, either parametric or nonparametric, is also considered to vary by observation. This is the penalty parameter in frequentist LASSO, and it might behave differently for observations from different groups. According to Park and Casella [20], the squared value $\lambda_i^2$ can have a conjugate prior as

$$\lambda_i^2 \sim Gamma(r, \delta),$$

where $r$ is the shape parameter and $\delta$ is the rate parameter. The common variance $\sigma^2$ has its conjugacy with an inverse gamma prior (shape parameter $a$ and rate parameter $b$):

$$\sigma^2 \sim IG(a, b).$$

As a result, our proposed WDPM-VS method has the prior as

$$\begin{aligned} \beta_i &= \left(\left(\beta_i^{(p)}\right)' \quad \left(\beta_i^{(np)}\right)'\right)', \\ \beta_i \mid \left(G_1, \cdots, G_Q\right), Z_i &= q \sim G_q, q = 1, \cdots, Q, \\ Z_i \mid \mathbf{b}_i &\sim multinomial\left[1, \cdots, Q; \mathbf{b}_i\right], \\ G_q &\sim DP(\alpha, G_0), G_0 \equiv \text{conditional Laplace}\left(\sigma^2, \lambda_i\right), \end{aligned}$$

which is equivalent to

$$\begin{aligned} \beta_i \mid \sigma^2, \left\{\tau_{i,0}^2, \tau_{i,1}^2, \cdots, \tau_{i,s+K}^2\right\} &\sim N\left\{\mathbf{0}, \sigma^2 * diag\left(\tau_{i,0}^2, \tau_{i,1}^2, \tau_{i,2}^2, \cdots, \tau_{i,s+K}^2\right)\right\}, \\ \tau_{i,d}^2, d = 0, 1, 2, \cdots, s &\sim Exponential\left\{\frac{\lambda_i^{2,(p)}}{2}\right\}, \\ \tau_{i,d}^2, d = s+1, \cdots, s+K &\sim Exponential\left\{\frac{\lambda_i^{2,(np)}}{2}\right\}, \\ \sigma^2 &\sim IG(a, b), \\ \lambda_i^2 &\sim Gamma(r, \delta). \end{aligned} \quad (2)$$

Due to conjugacy and computing efficiency, we use grid search to select the hyperparameters $\alpha$ (precision parameter from Dirichlet process) and $\psi$ (hyperparameter in the weight function $\mathbf{b}$). The other priors from (2) are all

conjugate, so a Gibbs sampling algorithm can be applied to solve for the posteriors. Detailed procedures are summarized in Appendix A.

## 3 | SIMULATION

We examined the performances of our proposed method through simulation studies in three different scenarios. In the first two simulation settings, we investigated the performance of functional clustering using WDPM when three different functions are generated. In the thrid simulation setting, we investigated the performance of functional clustering and peak point selection together from WDPM-VS method. The number of possible candidate priors $Q$ is also studied through the simulations.

Our proposed WDPM and WDPM-VS method are compared to the parametric Bayesian prior, DPM, and DPM-VS prior with $G_0$ set as Equations (1). Root mean squared errors (RMSEs) calculated between estimated ($\widehat{y}_{ij}$) and true ($y_{ij}$) response values, defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{\left(y_{ij}-\widehat{y}_{ij}\right)^2\right\}}{\sum_{i=1}^{n} n_i}},$$

are used as comparison criteria, where $\widehat{y}_{ij}$ are calculated via different methods. The number of possible candidates $Q$ is also examined through the simulations.

### 3.1 | Simulation settings

Consider nonparametric model:

$$y_{ij} = f_i\left(x_{ij}\right) + \varepsilon_{ij}, j = 1, 2, \cdots, n_i, i = 1, 2, \cdots, n,$$

where $x_{ij} \sim Unif(-2, 2)$, $\varepsilon_{ij} \sim N(0,0.01)$ and $n = 50$. That is, for each subject $i$, there are $n_i$ locations to form the function. The nonparametric function $f_i(x)$ s are considered differently in two settings.

#### 3.1.1 | Setting 1

Consider $n_i = 20$. The mean function $f_i(x)$ s are generated from three groups, containing both parametric and nonparametric functions:

$$f_i(x) = \begin{cases} \cos(\pi x) & \text{if } i = 1, 2, \cdots, 17, \\ x^2 & \text{if } i = 18, 19, \cdots, 34, \\ x & \text{if } i = 35, 36, \cdots, 50. \end{cases}$$

All three groups have nearly the same sizes, so this is a balanced scenario. Performances on both linear and nonlinear models are examined.

#### 3.1.2 | Setting 2

Consider $n_i = 20$. The mean function $f_i(x)$ s are written as

$$f_i(x) = \beta_{i1}x + \beta_{i2}x^2 + \beta_{i3}(x-\xi_1)_+^2 + \beta_{i4}(x-\xi_2)_+^2 + \beta_{i5}(x-\xi_3)_+^2,$$

where the parameters $\boldsymbol{\beta}_i$ follow from the WDP distribution. An unbalanced scenario is considered. We set two candidate priors so that the first quarter of observations is generated from the same Dirichlet process, while the rest are generated from another Dirichlet process. That is,

$$\boldsymbol{\beta}_i = \begin{cases} DP_1\left(\alpha_1, G_{0,1}\right) & \text{if } i = 1, 2, \cdots, 12, \\ DP_2\left(\alpha_2, G_{0,2}\right) & \text{if } i = 13, 14, \cdots, 50. \end{cases}$$

We considered that $DP_1$ and $DP_2$ share the same parameters to avoid over-heterogeneity, that is, $\alpha_1 = \alpha_2 = 2$, and

$$G_{0,1} = G_{0,2} = N\left\{\mathbf{0}, \begin{pmatrix} \sigma_p^2 = 25 & 0 & 0 & 0 & 0 \\ 0 & \sigma_p^2 = 25 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{np}^2 = 5 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{np}^2 = 5 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{np}^2 = 5 \end{pmatrix}\right\}.$$

In this setting, we examine nonparametric mean functions with heterogeneous data.

#### 3.1.3 | Setting 3

Consider $n_i = 100$. We generate the unknown function $f_i$ equally from three groups, containing both parametric and nonparametric functions, where the nonparametric function contains unequal changing points:

$$f_i(x) = \begin{cases} \frac{6}{10}\beta_{30,17}\left\{\frac{x-\min(x)}{\max(x)-\min(x)}\right\} + \frac{4}{10}\beta_{3,11}\left\{\frac{x-\min(x)}{\max(x)-\min(x)}\right\} & \text{if } i = 1, 2, \cdots, 17, \\ x^2 & \text{if } i = 18, 19, \cdots, 34, \\ 0.5x & \text{if } i = 35, 36, \cdots, 50, \end{cases}$$

where $\beta_{a,b}(x)$ is the density function of a beta distribution with parameters $a$ and $b$. The first function is modified from Wang and Wahba [21] and Montoya et al. [22]. This is a function with unequal changing points.

## 3.2 | Simulation results

### 3.2.1 | Setting 1

Setting 1 generates data from a balanced case and fits the model using WDPM method. Although the data are homogeneous, the WDPM prior still outperforms the DPM prior. Results are summarized in Figure 1, and Table A.1.1, with comparisons among various methods and the raw data.

Figure 1D demonstrates results from the WDPM prior with 50 candidates. The number 50 is selected based on the RMSEs from a number of candidates from 1 to 50. We found that the WDPM fitting can capture the characteristics of the raw data very well. It greatly outperforms the parametric Bayesian approach, illustrated in Figure 1B, and the DPM fitting, shown in Figure 1C. Although the data are homogeneous, the DPM prior still cannot separate the groups well. It requires a large precision parameter value to further separate the subjects, while the WDPM can control the separation through different candidates. The plot for run time (in hours) is also summarized in Figure A.1.1.

### 3.2.2 | Setting 2

We examine an unbalanced scenario for the WDPM method. Data possess more heterogeneities, while WDPM still captures the data structure well and beats the DPM prior and parametric Bayesian prior. Compared to Setting 1, our proposed method with the WDPM prior beats the comparison methods more in terms of the RMSEs as shown in Figure 2. This result supports that, when data are more heterogeneous, our proposed method performed much better than traditional approaches. Detailed results from setting 2 are shown in Table A.1.2. The plot for run time (in hours) is also summarized in Figure A.1.2.

### 3.2.3 | Setting 3

The performance of the WDPM-VS method is compared to the parametric Bayesian prior and DPM-VS prior with different candidate number $Q$ through 50 simulations. As suggested by Ruppert et al. [23], we selected total number

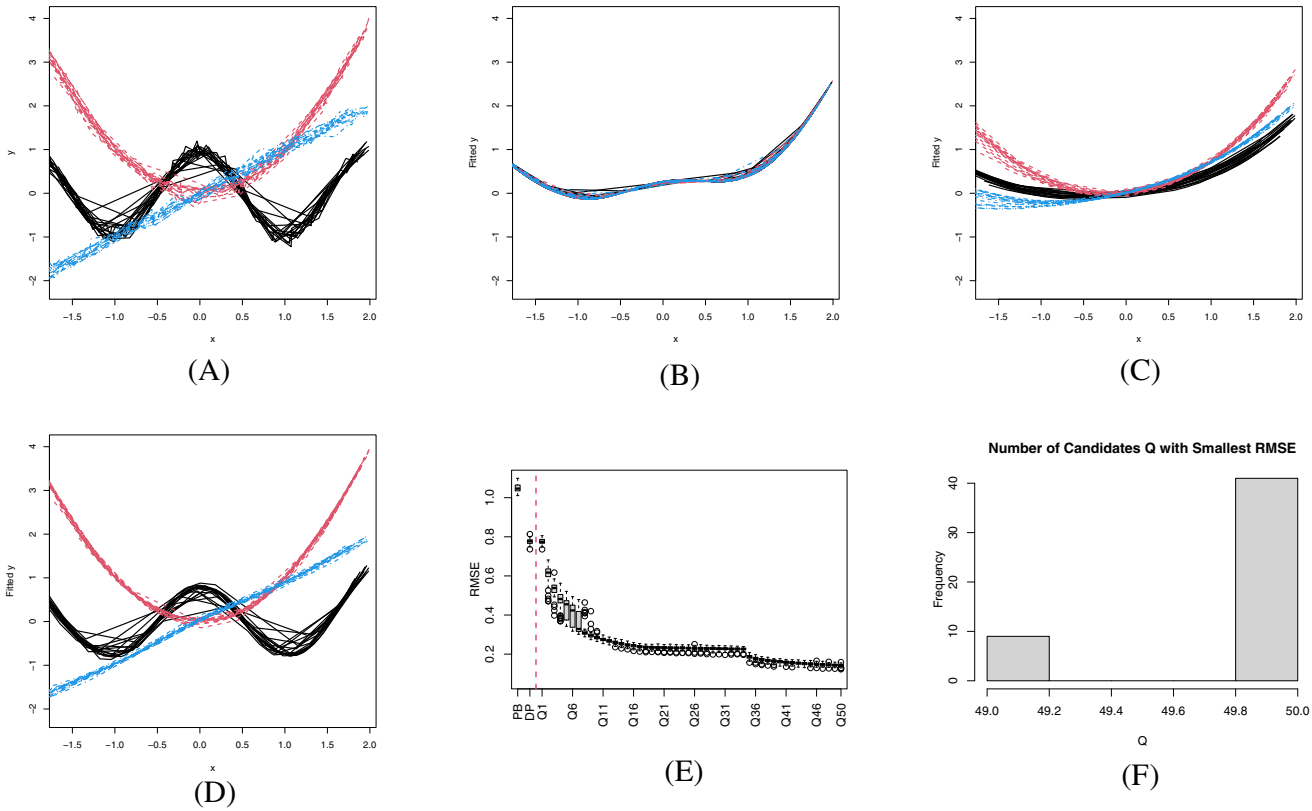



**FIGURE 1** Results from Simulation Setting 1: WDPM Model with Balanced Data. Figures Show the Raw Data (A) and the Fitted Functional Estimates Using Parametric Bayesian Approach (B) prior (C), and WDPM prior with 50 Candidate Priors (D). (E) Shows the Boxplot of RMSEs Through 50 Simulations From Different Methods: Parametric Bayesian Prior (PB), DPM Prior (DP) and WDPM Prior with Candidate Q (Q + Number). Boxes to the Right of the Red Dashed Line Are Results From WDPM. (F) Examines the Best Q Selected Based on RMSEs Through 50 Simulations.

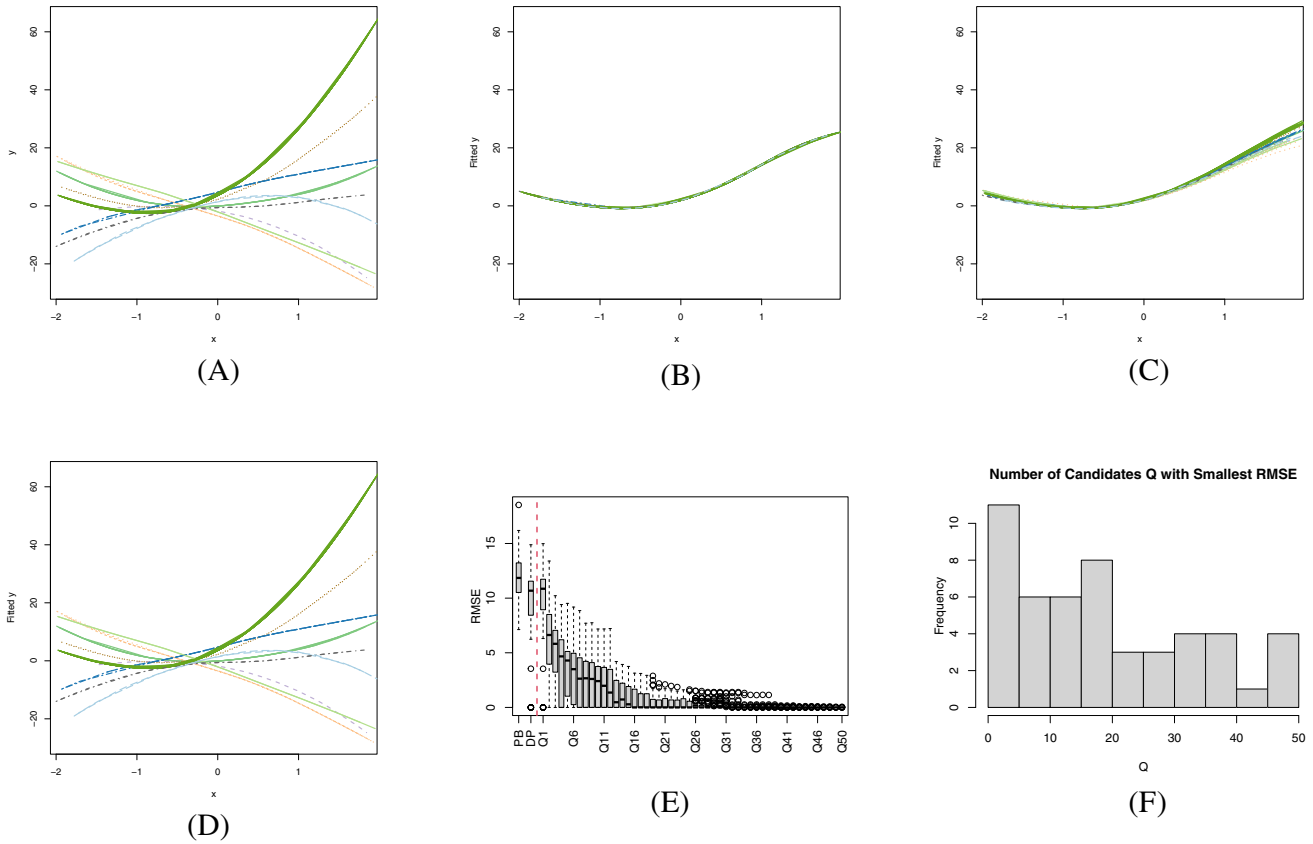



**FIGURE 2** Results from Simulation Setting 2: WDPM Model with Unbalanced Data. Figures Show the Raw Data (A) and the Fitted Functional Estimates Using Parametric Bayesian Approach (B) prior (C), and WDPM prior with 2 Candidate Priors (D). (E) Shows the Boxplot of RMSEs Through 50 Simulations From Different Methods: Parametric Bayesian Prior (PB), DPM Prior (DP) and WDPM Prior with Candidate Q (Q + Number). Boxes to the Right of the Red Dashed Line Are Results From WDPM. (F) Examines the Best Q Selected Based on RMSEs Through 50 Simulations.

of knots $K$ as

$$K = \min\left(\frac{\text{number of unique } x_{ij}}{4}, 35\right),$$

and each knot location $\xi_k$ as

$$\xi_k = \left(\frac{k+1}{K+2}\right)\text{th sample quantile of the unique } x_{ij}.$$

Results are summarized in Figure 3, Table A.1.3, with comparisons among various methods and the raw data. Figure 3D demonstrates results from the WDPM prior with 39 candidates. The number 39 is selected based on the RMSEs from a number of candidates from 1 to 50. We found that the WDPM fitting has comparable results to the truth (Figure 3A), and it greatly outperforms the parametric Bayesian (Figure 3B) and DPM-VS (Figure 3C) priors. The plot for run time (in hours) is also summarized in Figure 4.

## 3.3 | Investigations on number of potential candidates

The number of possible candidate priors $Q$ plays an essential role with the WDPM-VS prior. We investigated the performance of $Q$ by RMSEs and selected the $Q$ with the smallest RMSE as our final WDPM-VS model. Comparison results among candidate numbers 1 to 50 through boxplots and histograms from the 50 simulations are shown in Figure 3E,F, as well as Tables A.1.1–A.1.3.

The DPM-VS prior has a similar RMSE to the WDPM-VS prior when $Q = 1$. This is because WDPM is actually a mixture of DPM priors. When $Q = 1$, there is only one DPM prior, so it is equivalent to having the traditional DPM prior. As $Q$ increases, the RMSEs and their variances decrease in general. They usually have smaller values than the parametric Bayesian prior. However, when $Q$ is extremely large (over 42), the RMSE values and their variances grow rapidly and can be even larger than that of the parametric prior. Thus, we prepared a zoomed boxplot

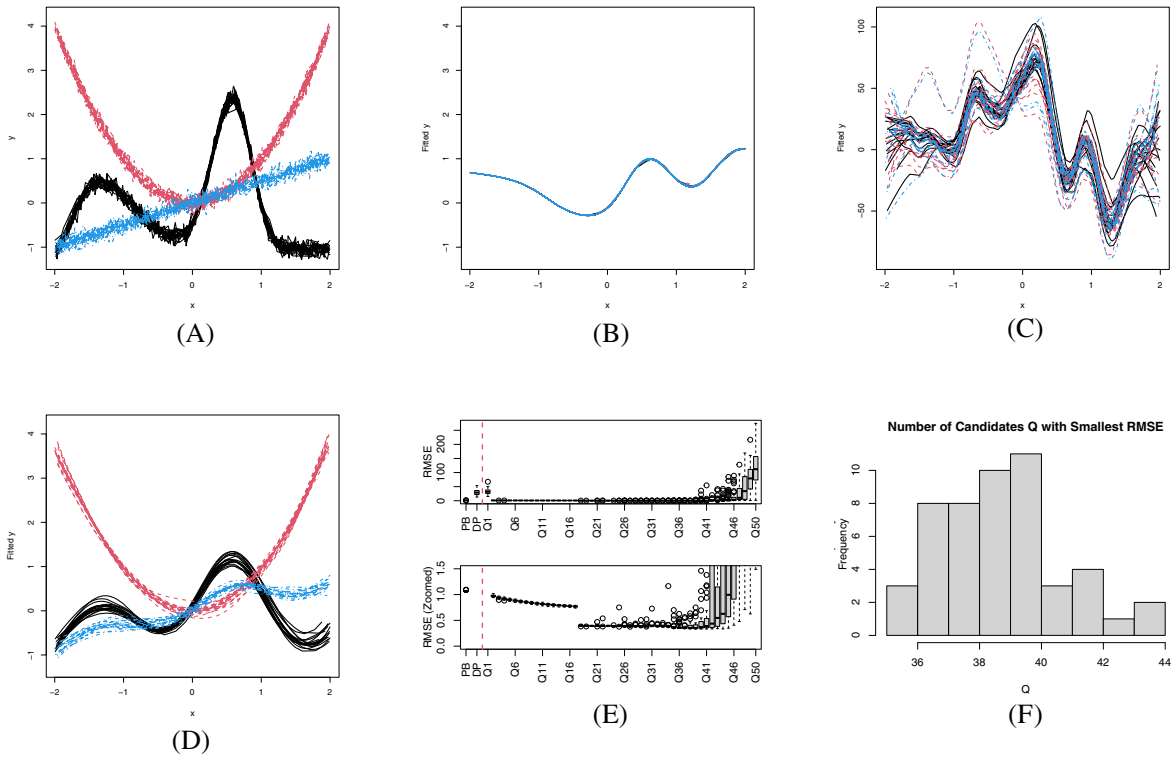



**FIGURE 3** Results from Simulation Setting 3: WDPM-VS Model with Balanced Data. Figures Show the Raw Data (A) and the Fitted Functional Estimates Using Parametric Bayesian Approach (B), DPM-VS prior (C), and WDPM-VS prior with 39 Candidates (D). (E) Shows the Boxplot of RMSEs Through 50 Simulations From Different Methods: Parametric Bayesian Prior (PB), DPM-VS Prior (DP) And WDPM-VS Prior with Candidate Q (Q + Number). Boxes to the Right of the Red Dashed Line Are Results From WDPM. The Plot Above is the Original Boxplot while the Plot Below Zoomed to RMSEs Lower Than 1.5. (F) Examines the Best Q Selected Based on RMSEs Through 50 Simulations.

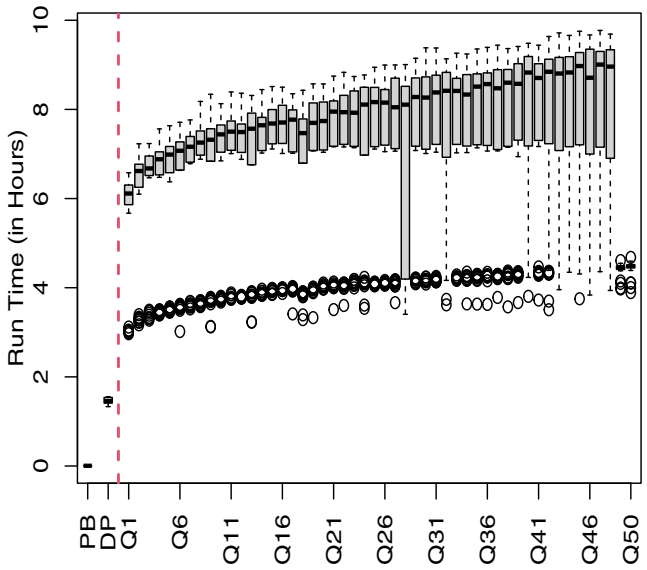



**FIGURE 4** Timing Results from Simulation Setting 3: WDPM-VS Model with Balanced Data. Boxplot of Computation Time in Hours Through 50 Simulations From Different Methods: Parametric Bayesian Prior (PB), DPM Prior (DP) And WDPM Prior with Candidate Q (Q + Number). Boxes to the Right of the Red Dashed Line Are Results From WDPM.

of the RMSEs, ranging from 0 to 1.5 (Figure 3E), to compare the majority models with lower RMSE values. For the majority of the $Q$ values, the WDPM-VS model has smaller RMSE values with larger $Q$, which are smaller than those from the DPM-VS and parametric priors.

Although the RMSEs generally decrease with the increase of the number $Q$, the decreasing rate keeps reducing. This result implies that although a larger $Q$ value is better, a relatively large $Q$ may already be optimal. When taking account of computing time shown in Figure 3 as well as Figures A.1.1–A.1.2, a relatively large $Q$ will be more optimal than the largest $Q$.

## 4 | BREAST CANCER RACIAL DISPARITIES USING SURFACE-ENHANCED RAMAN SPECTROSCOPY

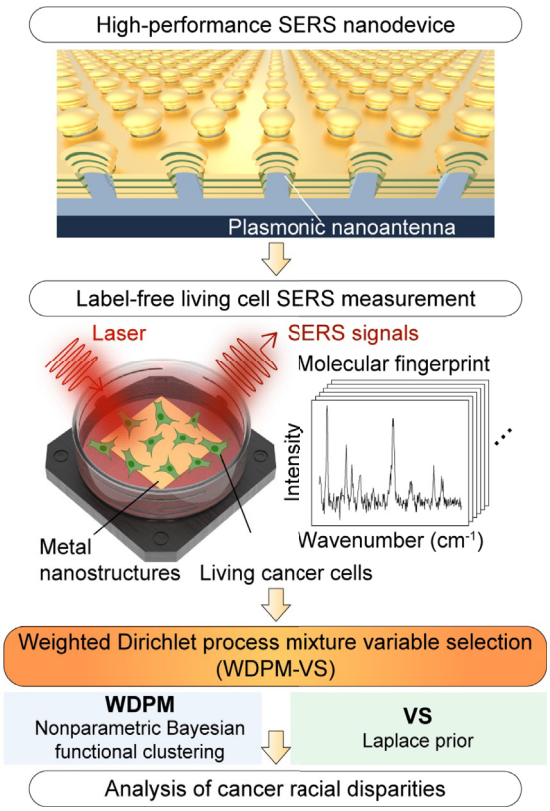
We applied our proposed method to a real data application, where we examined the WDPM-VS approach through the surface-enhanced Raman spectroscopy (SERS) data to study breast cancer racial disparities. Results show that our

proposed method outperforms the comparison DPM-VS method. Through the analysis results using our proposed model, we are able to identify the important biological factors affecting racial disparities.

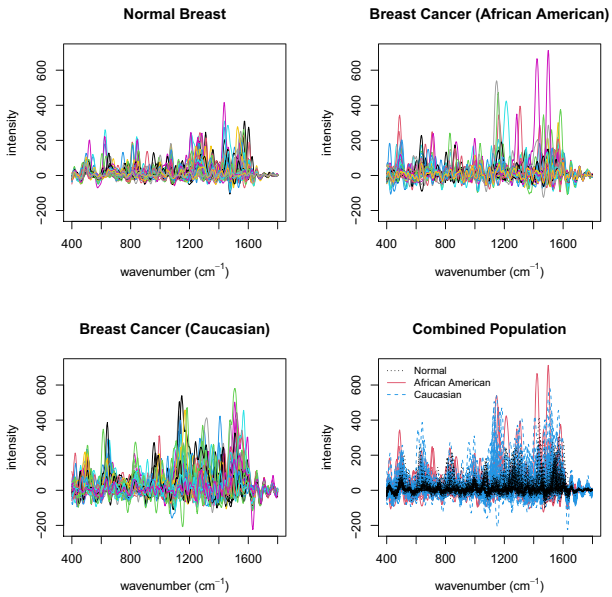
## 4.1 | Motivation and data description

Breast cancer is the second most common cancer and the leading cause of cancer death in American women [24]. The National Institutes of Health (NIH) shows that different incidence and mortality rates for breast cancer exist among various racial populations. For example, Caucasian women are more likely to develop breast cancer than African American women are [25]. Recent literature has shown that biological factors might heavily affect racial disparities [26]. Studies on the biological factors thus have been rapidly developed with advanced nanotechnologies, such as SERS [27, 28]. The label-free SERS approach can directly measure the biomolecular fingerprint information from living cells in a real-time and non-destructive manner, so that sample preparation and chemical modifications can be significantly omitted [28–30]. A new cost-effective SERS device which consists of 3D arrays of nanolaminate nanoantennas with high sensitivity and good uniformity has been developed by the Nano-enabled Photonics-Electronics Devices and Systems (NePEDS) lab from Virginia Tech [31]. Extracellular SERS signals from living breast cells were measured by this practical high-performance SERS device with an intimate analyses on classification between cancer and normal groups have already been studied by the lab. However, classification among racial groups has been challenging. The flowchart of how to collect label-free living cell spectroscopic data from this SERS device with the proposed statistical analysis is shown in Figure 5A. The measured SERS spectra via molecular profiling can be shown in curves of Raman signal intensities versus wavenumbers as shown in Figure 5B.

The wavenumber, determined by the energy of a molecular vibrational mode, has a one-to-one match and can be used as the biomolecular fingerprint information




(A)




(B)

FIGURE 5 Nanolamined SERS device and signal functions: (A) Using high-performance nanolaminated SERS device, collect molecule profiling of label-free living breast cancer and normal cells, and conduct WDPM-VS for functional clustering among racial disparities groups and simultaneously select vibrational molecular fingerprint associated to cancer racial disparities; (B) Signal Intensities Versus Raman Wavenumbers by Racial Populations: Breast Normal Women, Breast Cancer Patients for African American Women and Breast Cancer Patients for Caucasian Women, Along with Combined Populations.

to explain the cause of racial disparities. Hence, we want to identify the wavenumbers that have different behaviors in the intensities for racial groups. However, the large intraclass variations due to cellular and additional cancerous heterogeneity add difficulties to compare across racial groups, because the noise information might produce overlaps across groups. Therefore, it is desirable to reduce the amount of noise information and make each group distinguishable. The noises exist in two directions: a large number of heterogeneously behaved signal curves, as well as the massive peak points on the curves. Therefore, we apply our proposed WDPM-VS method to reduce the massive noisy information within each racial group.

Three different racial groups were considered: women without breast cancer, Caucasian women with breast cancer, and African American women with breast cancer. The Raman signals were all measured with wavenumbers from 400 to 1800. The sample sizes for the three groups were not the same. We measured 85 cells for Caucasian women, 78 for African American women, and 95 for women without breast cancer. Figure 5B shows the original data by each racial group, as well as a combination of the three groups. It is clear that the three racial groups behave differently. However, it is difficult to identify the differences from the overlapping plot.

## 4.2 | Analyses and results

Figure 6 shows the fitted results using both DPM-VS and WDPM-VS priors. The DPM-VS prior has only one cluster for each racial group, while the WDPM-VS results retain the variations in the data. It keeps the intraclass heterogeneities, and thus is preferred. The RMSEs of WDPM-VS are also lower than those of DPM-VS for all racial groups.

From the fitted plot, we can observe that the normal group is quite stable, while the patient groups have more variations. In general, the African American group has a univariate structure. However, a few clusters show large variations around wavenumbers 1100–1500 to separate it from the normal group. On the other hand, the Caucasian group has large heterogeneities. We further investigated the parameter estimations. For easy comparison, we scaled each estimated parameter relative to the largest absolute value from the same observation. If the absolute value of the scaled parameter was larger than 0.5, we considered it important and selected it. Using the DPM-VS method, we only selected one parameter corresponding to the linear basis. Using the WDPM-VS approach, we also only selected the linear basis for the normal group and the African American group, but there are more selections for the Caucasian group. Among the 85 observations, 79 selected the linear basis and 20 selected the knot at 1351–1531. If we considered all peaks selected with at least 10 observations, we found important peak

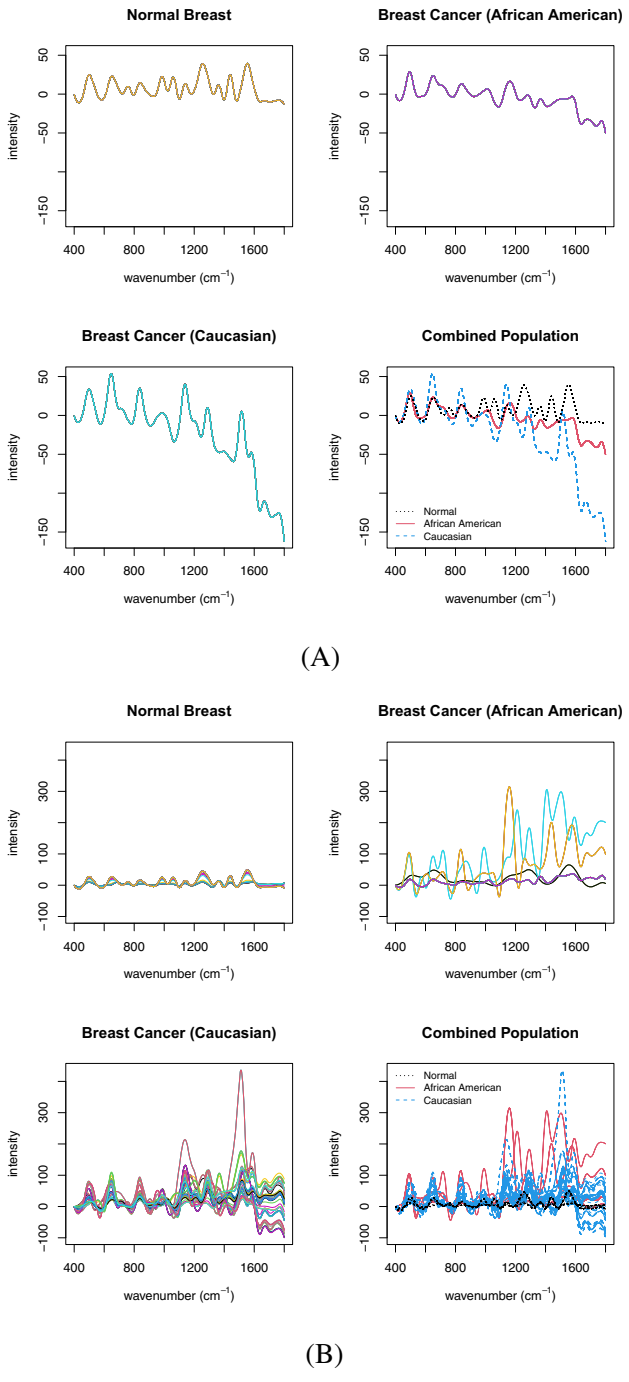



**FIGURE 6** Results from Real Data Application. Clustering Results by DPM-VS and WDPM-VS Priors with Nonparametric Function for Raman Signal Intensities Versus Raman Shifts/Wavenumbers by Racial Populations: Women Without Breast Cancer, Breast Cancer Patients for African American Women and Breast Cancer Patients for Caucasian Women, Along with Combined Populations.

**TABLE 1** Breast cancer racical disparities: wavenumber at selected peaks, possible correlated bioattributions, and their references.

| Selected Wavenumber (cm−1) | Possible corresponding attributions | Reference |
|---|---|---|
| 685 | Ring breathing mode of guanine | Chan et al. [32] |
| 1291 | Cytosine | Ruiz-Choca et al. [33] |
| 1316 | Collagen and lipid, Amide III protein | Stone et al. [34], Stone et al. [38] |
| 1422 | Deoxyribose | Ruiz-Choca et al. [33] |
| 1569 | Tryptophan, guanine | Stone et al. [34], Lau et al. [35] |
| 1596 | Phenylalanine, Amide I protein | Chan et al. [32], Sigurdsson et al. [37] Dukor [36] |
| 1620 | Tryptophan, Amide I protein | Chan et al. [32], Sigurdsson et al. [37] Dukor [36] |

locations around wavenumbers 600–700, 1270–1470, and 1550–1750, which contain vibrational molecular fingerprints associated to cancer racial disparities.

The single and/or combination of peaks act as a fingerprint to identify a racial group. The peaks can be correlated to possible biochemical attributes, such as the follows: (1) Ring breath mode of guanien [32] at peak 685, (2) Cytosine [33] at peak 1291, (3) Collagen and lipid, Amide III protein ([34],0) at peak 1316, (4) Deoxyribose [33] at peak 1422, (5) Dexoxyribose [34, 35] at peak 1569, (6) Phenylalanine, Amide I protein [32, 36, 37] at peak 1596, and (7)Tryptophan, Amide I protein [32, 36, 37] at peak 1620.

Hence, some of these wavenumbers have already been identified to have corresponding attributions, which are listed in Table 1. These attributions can be helpful to explain the biological factors separating the Caucasian group.

## 5 | DISCUSSION

In this paper, we proposed the WDPM-VS method that performs functional clustering and peak point selection at the same time. Our proposed method can simultaneously perform the following tasks: (1) Automatic cluster without specifying the number of clusters or cluster centers beforehand; (2) Cluster for heterogeneously behaved functions; (3) Select vibrational peak points; and (4) Reduce noisy information from the two perspectives: sample size and dimensionality. Based on simulations, our method outperforms comparison methods in root mean squared errors (RMSE). Our method also beats comparisons in real data application in that it reduces the noise information from the two perspectives mentioned above, while it maintains data structure when identifying critical signals. We have examined the performance of our proposed method through simulations and real data applications. More advanced theoretical justifications are beyond the scope of this paper, but can be conducted for future work. These will help explain the rational behind the results in more general settings.

Estimation results of the WDPM-VS prior depend highly on the total number of candidates. In general, the RMSE will decrease with an increase in the number of candidates, within a reasonable range. However, the decreasing rate will drop after a certain number. Further investigations are necessary to find the general relationship between the optimal number of candidates and the total number of subjects, considering computation efficiency, through various simulation studies and theoretical proofs. Meanwhile, further studies can work on developing efficient algorithms to improve the computation efficiency.

We investigated the breast cancer racial disparities to compare different racial groups. Due to the large intra-class heterogeneities, we reduced the noise by clustering and peak point selections. Results show that people without breast cancer have more stable Raman signal curves, while breast cancer patients have more variations. Intensities for Caucasian women show important changes at wavenumbers 600–700, 1270–1470, and 1550–1750. Some of these wavenumbers have already been identified to have corresponding attributions, which can be used to explain the biological factors separating the Caucasian group. We note that these finds need to be further validated biologically.

## ACKNOWLEDGMENTS

This study was supported in part by the National Science Foundation grant number 1711699.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

*Inyoung Kim* https://orcid.org/0000-0002-5975-4582


## REFERENCES


1. F. Chamroukhi and H. D. Nguyen, *Model-based clustering and classification of functional data*, Wiley Interdiscip. Rev. 9 (2019), no. 4, e1298.
2. D. Blackwell and J. B. MacQueen, *Ferguson distributions via Pólya urn schemes*, Ann. Stat. 1 (1973), 353–355.
3. T. S. Ferguson, "*Bayesian density estimation by mixtures of Normal distributions*," *Recent advances in statistics: Papers in honor of Herman Chernoff on his sixtieth birthday*, Academic Press, Cambridge, MA, 1983, pp. 287–302.
4. S. N. MacEachern and P. Müller, *Estimating mixtures of Dirichlet process models*, J. Comput. Graph. Stat. 7 (1998), 223–238.
5. S. N. MacEachern, *Dependent dirichlet processes*. Technical Report, Department of Statistics, Ohio State University, Columbus Ohio, 2000.
6. T. S. Ferguson, *A Bayesian analysis of some nonparametric problems*, Ann. Stat. 1 (1973), 209–230.
7. S. Basu and S. Chib, *Marginal likelihood and Bayes factors for Dirichlet process mixture models*, J. Am. Stat. Assoc. 98 (2003), no. 461, 224–235.
8. L. A. Hannah, D. M. Blei, and W. B. Powell, *Dirichlet process mixtures of generalized linear models*, J. Mach. Learn. Res. 12 (2011), 1923–1953.
9. A. Zellner, "*On assessing prior distributions and Bayesian regression analysis with g-prior distributions*," *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti 233–243*, North-Holland, Amsterdam, 1986.
10. D. B. Dunson, N. Pillai, and J. Park, *Bayesian density regression*, J. Roy. Stat. Soc. Ser. B 69 (2007), no. 2, 163–183.
11. P. Sun, I. Kim, and K. Lee, *Dual-semiparametric regression using weighted Dirichlet process mixture*, Comput. Stat. Data Anal. 117 (2018), 162–181.
12. S. N. MacEachern, "*Dependent nonparametric processes*," *ASA proceedings of the section on Bayesian statistical science Amer*, Statist. Assoc, Alexandria, VA, 1999.
13. F. A. Quintana, P. Müller, A. Jara, and S. N. MacEachern, *The dependent Dirichlet process and related models*, Stat. Sci. 37 (2022), no. 1, 24–41.
14. P. Müller, F. Quintana, and G. Rosner, *A method for combining inference across related nonparametric Bayesian models*, J. Roy. Stat. Soc. Ser. B 66 (2004), no. 3, 735–749.
15. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, *Hierarchical dirichlet processes*, J. Am. Stat. Assoc. 101 (2006), no. 476, 1566–1581.
16. A. Rodriguez, D. B. Dunson, and A. E. Gelfand, *The nested Dirichlet process*, J. Am. Stat. Assoc. 103 (2008), no. 483, 1131–1154.
17. A. Rodriguez and D. B. Dunson, *Functional clustering in nested designs: Modeling variability in reproductive epidemiology studies*, Anna. Appl. Stat. 8 (2014), no. 3, 1416–1442.
18. P. Sun, I. Kim, and K. Lee, *Flexible weighted dirichlet process mixture modelling and evaluation to address the problem of forecasting return distribution*, J. Nonparamet. Stat. 32 (2020), no. 4, 989–1014.
19. R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Methodol. 58 (1996), no. 1, 267–288.
20. T. Park and G. Casella, *The bayesian lasso*, J. Am. Stat. Assoc. 103 (2008), no. 482, 681–686.
21. Y. Wang and G. Wahba, *Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals*, J. Stat. Comput. Simul. 51 (1995), no. 2–4, 263–279.
22. E. L. Montoya, N. Ulloa, and V. Miller, *A simulation study comparing knot selection methods with equally spaced knots in a penalized regression spline*, Int. J. of Stat. Probab. 3 (2014), no. 3, 96.
23. D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric regression (No. 12)*, Cambridge University Press, New York, 2003.
24. American Cancer Society, *Cancer facts and figures 2019*, American Cancer Society, Atlanta, GA, 2019.
25. R. T. Chlebowski, Z. Chen, G. L. Anderson, T. Rohan, A. Aragaki, D. Lane, N. C. Dolan, E. D. Paskett, A. McTiernan, F. A. Hubbell, L. L. Adams-Campbell, and R. Prentice, *Ethnicity and breast cancer: Factors influencing differences in incidence and outcome*, J. Natl. Cancer Inst. 97 (2005), no. 6, 439–448.
26. T. Keenan, B. Moy, E. A. Mroz, K. Ross, A. Niemierko, J. W. Rocco, S. Isakoff, L. W. Ellisen, and A. Bardia, *Comparison of the genomic landscape between primary breast cancer in African American versus white women and the association of racial differences with tumor recurrence*, J. Clin. Oncol. 33 (2015), no. 31, 3621–3627.
27. J. Kneipp, H. Kneipp, and K. Kneipp, *SERS—A single-molecule and nanoscale tool for bioanalytics*, Chem. Soc. Rev. 37 (2008), no. 5, 1052–1060.
28. C. Zong, M. Xu, L. J. Xu, T. Wei, X. Ma, X. S. Zheng, R. Hu, and B. Ren, *Surface-enhanced Raman spectroscopy for bioanalysis: Reliability and challenges*, Chem. Rev. 118 (2018), no. 10, 4946–4980.
29. G. Kuku, M. Altunbek, and M. Culha, *Surface-enhanced raman scattering for label-free living single cell analysis*, Anal. Chem. 89 (2017), no. 21, 11160–11166.
30. X. S. Zheng, I. J. Jahn, K. Weber, D. Cialla-May, and J. Popp, *Label-free SERS in biological and biomedical applications: Recent progress, current challenges and opportunities*, Spectrochim. Acta A Mol. Biomol. Spectrosc. 197 (2018), 56–77.
31. W. Nam, X. Ren, S. A. S. Tali, P. Ghassemi, I. Kim, M. Agah, and W. Zhou, *Refractive-index-insensitive nanolaminated SERS substrates for label-free raman profiling and classification of living cancer cells*, Nano Lett. 19 (2019), no. 10, 7273–7281.
32. J. Chan, D. Tayloer, T. Zwerdling, S. Lane, and K. Ihara, *Micro-raman spectroscopy detects individual neoplastic and normal hematopoietic cells*, Biophys. J. 90 (2006), 648–656.
33. A. Ruiz-Chica, M. A. Medina, F. Sanchez-Jimenez, and F. J. Ramirez, *Characterization by Raman spectroscopy of conformational changes on guanine-cytosine and adenine-thymine oligonucleotides induced by aminooxy analogues of sermidine*, J. Raman Spectrosc. 35 (2004), 93–100.
34. N. Stone, C. Kendall, N. Shepherd, P. Crow, and H. Barr, *Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers*, J. Raman Spectrosc. 33 (2002), 564–573.
35. D. Lau, Z. Huang, H. Lui, K. Berean, M. Morrison, and H. Zeng, *Raman spectroscopy for optical diagnosis in normal and cancerous tissue of the nasopharynx—preliminary findings*, Lasers Surg. Med. 32 (2020), 210–214.

36. R. K. Dukor, "*Vibrational spectroscopy in the detection of cancer*," *Handbook Virational Spectroscopy*, Hobojen, 2006, pp. 3335–3361.
37. S. Sigurdsson, P. A. Philipsen, L. Hansen, and J. Larsen, *Detection of skin cancer by classification of raman spectra*, IEEE Transact. Biomed. Eng. 51 (2004), 1784–1793.
38. N. Stone, C. Kendall, J. Smith, P. Crow, and H. Barr, *Raman spectroscopy for identification if epithelial caners*, J Faraday Dicuss. 126 (2004), 141–157.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** W. Gao, I. Kim, W. Nam, X. Ren, W. Zhou, and M. Agah, *Nonparametric Bayesian functional clustering with applications to racial disparities in breast cancer*, Stat. Anal. Data Min.: ASA Data Sci. J. **17** (2024), e11657. https://doi.org/10.1002/sam.11657

## APPENDIX A

### A.1 MCMC SAMPLING PROCEDURE

Our proposed WDPM-VS method can obtain its posterior distributions through Gibbs sampling due to conjugacy. Assume we finally arrive at $L$ distinct clusters, with $L^{(i)}$ as the number of distinct clusters without observation $i$. For each cluster $l$, we have a unique set of values $\{\boldsymbol{\beta}_l^*, \boldsymbol{\tau}_l^{2,*}, \lambda_l^{2,*}\}$, which comes from candidate $q$, and is labeled $C_l = q$. Each $\{\boldsymbol{\beta}_l^*, \boldsymbol{\tau}_l^{2,*}, \lambda_l^{2,*}\}$ can be matched to multiple observations $\{\boldsymbol{\beta}_i, \boldsymbol{\tau}_i^2, \lambda_i^2\}$ with $L_i = l$, where $L_i$ is the cluster label for observation $i$. Further define $n_g^-$ to be the number of observations without observation $i$ from group $g$, where $g$ can be a cluster or a candidate. Denote $\mathbf{D}_\tau = diag\left(\tau_{i,0}^2, \tau_{i,1}^2, \tau_{i,2}^2, \cdots, \tau_{i,s+K}^2\right)$. The posterior distributions for WDPM-VS proceed in the following steps:

Step H0: Select initial values for $\mathbf{L}$, $\mathbf{C}$, $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, $\boldsymbol{\lambda^2}$, and $\sigma^2$;

Step H1: Sample label $L_i, i = 1, 2, \cdots, n$ from

$$l, l = 1, \cdots, L^{(i)}$$

with probability proportional to

$$\frac{b_{i,C_l}^* n_l^{-*} N\left(\mathbf{X}_i \boldsymbol{\beta}_l, \sigma^2\right)}{\alpha + n_{C_l}^-},$$

and from

$$L^{(i)} + 1$$

with probability proportional to

$$\alpha^* N\left(0, \sigma^2 I_p + \mathbf{X}_i \sigma^2 \mathbf{D}_\tau \mathbf{X}_{i'}\right) * \sum_{q=1}^{Q} \frac{b_{iq}}{\alpha + n_q^-};$$

Step H2: Sample unique $\boldsymbol{\beta}_l^*, l = 1, 2, \cdots, L$ from

$$N\left\{\left(\sum_{i:L_i=l} \mathbf{X}_i' \mathbf{X}_i + \mathbf{D}_\tau^{-1}\right)^{-1} \sum_{i:L_i=l} \mathbf{X}_i' y_i, \sigma^{2*}\left(\sum_{i:L_i=l} \mathbf{X}_i' \mathbf{X}_i + \mathbf{D}_\tau^{-1}\right)^{-1}\right\};$$

Step H3: Sample unique $1/\boldsymbol{\tau}_{d,l}^{2,*}, d = 0,1,2, \cdots, s+K, l = 1, 2, \cdots, L$ from

$$Inverse - Gaussian\left(\sqrt{\frac{\lambda_l^{2*}\sigma^2}{\boldsymbol{\beta}_{d,l}^2}}, \lambda_l^2\right),$$

where $1/\boldsymbol{\tau}_{d,l}^{2,*}$ and $\lambda_l^2$ can be either parametric or nonparametric;

Step H4: Sample unique $\lambda_l^2, l = 1, 2, \cdots, L$ from

$$Gamma\left(p + r, \frac{\sum_{d=0}^{s+K} \tau_{d,l}^2}{2} + \delta\right);$$

Step H5: Sample candidate label $C_l, l = 1, 2, \cdots, L$ from

$$q, q = 1, \cdots, Q$$

with probability proportional to

$$\frac{\prod_{i:L_i=l} b_{iq}}{\sum_{m=1}^{Q} \prod_{i:L_i=l} b_{im}};$$

Step H6: Sample common variance $\sigma^2$ from

$$IG\left\{a + \frac{\left(\sum_{i=1}^{n} n_i + L^*(s+K+1)\right)}{2}, b + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2} + \frac{\boldsymbol{\beta}' \mathbf{D}_\tau \boldsymbol{\beta}}{2}\right\}.$$