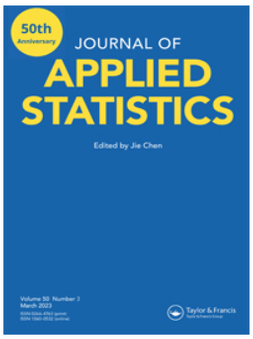


Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Classification of histogram-valued data with support histogram machines

Ilsuk Kang, Cheolwoo Park, Young Joo Yoon, Changyi Park, Soon-Sun Kwon & Hosik Choi

To cite this article: Ilsuk Kang, Cheolwoo Park, Young Joo Yoon, Changyi Park, Soon-Sun Kwon & Hosik Choi (2023) Classification of histogram-valued data with support histogram machines, Journal of Applied Statistics, 50:3, 675-690, DOI: 10.1080/02664763.2021.1947996

To link to this article: https://doi.org/10.1080/02664763.2021.1947996

Published online: 01 Jul 2021.

JOURNAL OF APPLIED STATISTICS
2023, VOL. 50, NO. 3, 675–690
https://doi.org/10.1080/02664763.2021.1947996

Taylor & Francis
Taylor & Francis Group

Classification of histogram-valued data with support histogram machines

Ilsuk Kang[a], Cheolwoo Park[b], Young Joo Yoon[c], Changyi Park[d], Soon-Sun Kwon[e] and Hosik Choi[f]

[a]Department of Statistics, Univ. of Georgia, Athens, GA, USA; [b]Department of Mathematical Sciences, KAIST, Daejeon, The Republic of Korea; [c]Department of Mathematics Education, Korea National Univ. of Education, Cheongju, The Republic of Korea; [d]Department of Statistics, University of Seoul, Seoul, The Republic of Korea; [e]Department of Mathematics, Ajou University, Suwon, The Republic of Korea; [f]Graduate School, Department of Urban Big Data Convergence, University of Seoul, Seoul, The Republic of Korea

ABSTRACT

The current large amounts of data and advanced technologies have produced new types of complex data, such as histogram-valued data. The paper focuses on classification problems when predictors are observed as or aggregated into histograms. Because conventional classification methods take vectors as input, a natural approach converts histograms into vector-valued data using summary values, such as the mean or median. However, this approach forgoes the distributional information available in histograms. To address this issue, we propose a margin-based classifier called support histogram machine (SHM) for histogram-valued data. We adopt the support vector machine framework and the Wasserstein-Kantorovich metric to measure distances between histograms. The proposed optimization problem is solved by a dual approach. We then test the proposed SHM via simulated and real examples and demonstrate its superior performance to summary-value-based methods.

ARTICLE HISTORY
Received 30 September 2020
Accepted 21 June 2021

KEYWORDS
Support vector machines; symbolic data; Wasserstein-Kantorovich metric

2010 MATHEMATICS SUBJECT CLASSIFICATION
62H30

1. Introduction

We live in an era in which we can easily collect and store enormous amounts of data in a short period using modern technology. These data are often summarized into new formats for various reasons. For example, government agencies provide their public data as summary data (e.g. frequency tables and histograms) to ensure that any disclosed information cannot be traced back to an individual. Additionally, by converting single-valued data into histograms or intervals, one can reduce data size and computation time. These summarized data are called symbolic data [3,4].

The literature on symbolic data analysis is growing rapidly, and we review below only those studies directly relevant to our work. For histogram-valued data, there have been

CONTACT Hosik Choi choi.hosik@gmail.com

a few studies on cluster analysis. For instance, Kim and Billard [21] introduced dissimilarity measures and a divisive clustering algorithm for multimodal-valued data, for which a special case is histogram-valued data. Kim and Billard [22] compared several dissimilarity measures for clustering histogram-valued data, while Park *et al.* [27] proposed a hierarchical convex clustering algorithm for histogram-valued data based on the Wasserstein-Kantorovich (WK) distance. Further, Dias and Brito [11] developed a linear regression model for histogram-valued data.

Several classification methods based on support vector machine (SVM, [9]) have been proposed over the last two decades, which are suitable for analyzing symbolic data. For image classification, Chapelle *et al.* [7] divided each dimension in RGB space into 16 bins and embedded the frequencies that were matched to the bin combinations using a quantization step as an input in SVM. In handwriting recognition, Alaei and Roy [1] used histogram-valued data to express the features, instead of using raw vectors, for identifying writing styles. Carrizosa *et al.* [6] introduced an SVM robust to perturbation or uncertainty and showed that the proposed method is applicable to the interval-valued data affected by noise or some type of perturbation. Do and Poulet [13] proposed an approach for applying the Hausdorff distance to the radial basis function kernel for interval-valued data and exploited interactive decision tree algorithms to interpret the SVM results. Angulo *et al.* [2] discussed a different SVM approach, the so-called I-SVM, which combines prior knowledge expressed in an interval form [26] with the original dataset using a linear kernel.

Apart from SVM, various other classification approaches for symbolic data, especially for interval-valued data, have been developed. For example, Duarte Silva and Brito [15] proposed three approaches for applying linear discriminant analysis to interval-valued data, while Duarte Silva and Brito [14] extended this previous research from the distance-based model to parametric models. Ciampi *et al.* [8] proposed the probabilistic decision tree, a tree-growing algorithm, to handle data having uncertainty expressed as interval-valued data. Recently, Dias *et al.* [12] developed a binary classification method for distributional data based on quantile functions.

In this paper, we consider discriminating histogram-valued data for both binary class and multi-class cases. Among the many popular classification methods, we adopt SVM because it has good classification accuracy in general and is flexible enough to be extended to histogram-valued data. Specifically, we develop support histogram machine (SHM), which takes histogram-valued data as an input, and classifies them under the SVM framework. To define the kernel function in the proposed SHM, we measure the distances between histograms using the WK distance, which is known as a suitable metric for histogram-valued data [16,19]. To the best of our knowledge, this is the first study that develops a margin-based classifier for histogram-valued data.

The rest of the paper is organized as follows. We introduce histogram-valued data and the WK distance and describe the proposed SHM method for both binary class and multi-class problems in Section 2. Section 3 presents a simulation study that examines the performance of the proposed method and compares it with those of standard SVM and k-nearest neighborhood (k-NN) methods using vectorized summary values. In Section 4, we apply the proposed method to a real example. Section 5 discusses a further extension of the proposed method to the hinge loss with case-specific parameters.

2. Proposed method

Here, we introduce the notations and definitions used for histogram-valued data and the WK metric in Section 2.1. We propose SHM for binary class problems in Section 2.2 and for multi-class problems in Section 2.3.

2.1. Histogram-valued data and Wassertein-Kantorovich metric

Let $\{\boldsymbol{X}_1, \ldots \boldsymbol{X}_n\}$ be a sample of size n of a p-dimensional histogram-valued variable, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, n$. Also, let $\{y_1, \ldots, y_n\}$ denote the associated class labels with $y_i \in \{-1, +1\}$. Assume that each X_{ij} consists of K_{ij} non-overlapping intervals I_{ijk} with associated relative frequencies p_{ijb} for $b = 1, \ldots, K_{ij}$. That is:

$$X_{ij} = \{(I_{ij1}, p_{ij1}), \ldots, (I_{ijK_{ij}}, p_{ijK_{ij}})\} \tag{1}$$

with $p_{ijb} \geq 0$ and $\sum_{b=1}^{K_{ij}} p_{ijb} = 1$.

To extend SVM to SHM, a metric between histograms needs to be defined. As long as we have the proper distance between histograms, we can induce a kernel function based on the distance and use it for obtaining a classification rule. Gibbs and Su [16] provided a review of 10 different probability metrics. According to Irpino and Verde [19], the WK metric possesses the desirable property that total inertia can be decomposed into within and between groups inertia, which is useful for cluster or classification analysis. Park *et al.* [27] applied the WK metric to convex cluster analysis.

In what follows we introduce the WK metric. Let F_{ij} be the empirical cumulative distribution corresponding to X_{ij} as drawn in Figure 1. Then, the WK distance between histogram-valued observations X_{ij} and X_{lj} for the jth variable is defined as [19]:

$$d_W(X_{ij}, X_{lj}) = \sqrt{\int_0^1 \left(F_{ij}^{-1}(t) - F_{lj}^{-1}(t)\right)^2 \mathrm{d}t}. \tag{2}$$

Instead of calculating the distance as per (2), we use the special case of (2) by assuming the uniformity within the subintervals as in Irpino and Verde [19]. Define the cumulative weight up to the kth subinterval for X_{ij} as

$$w_{ijk} = \begin{cases} 0, & k = 0 \\ \displaystyle\sum_{b=1}^{k} p_{ijb}, & k = 1, \ldots, K_{ij}. \end{cases}$$

Given two univariate histogram-valued observations X_{ij} and X_{lj} for the jth variable, we combine two sets of cumulative weights from X_{ij} and X_{lj} and obtain a common set of cumulative weights $\boldsymbol{w}_j^{(i,l)} = \{w_{j(0)}^{(i,l)}, \ldots, w_{j(m_{ijl})}^{(i,l)}\}$, where $0 \leq w_{j(0)}^{(i,l)} \leq \cdots \leq w_{j(m_{ijl})}^{(i,l)} = 1$. Next, we compute the common relative frequencies $\{\pi_{j1}^{(i,l)}, \ldots, \pi_{jm_{ijl}}^{(i,l)}\}$, $\pi_{jk}^{(i,l)} = w_{j(k)}^{(i,l)} - w_{j(k-1)}^{(i,l)}$ for $k = 1, \ldots, m_{ijl}$. Then, using $\{\pi_{j1}^{(i,l)}, \ldots, \pi_{jm_{ijl}}^{(i,l)}\}$, one can obtain subintervals $\{J_{ij1}, \ldots, J_{ijm_{ijl}}\}$ and $\{J_{lj1}, \ldots, J_{ljm_{ijl}}\}$ that match the quantiles corresponding to $\{\pi_{j1}^{(i,l)}, \ldots, \pi_{jm_{ijl}}^{(i,l)}\}$ under

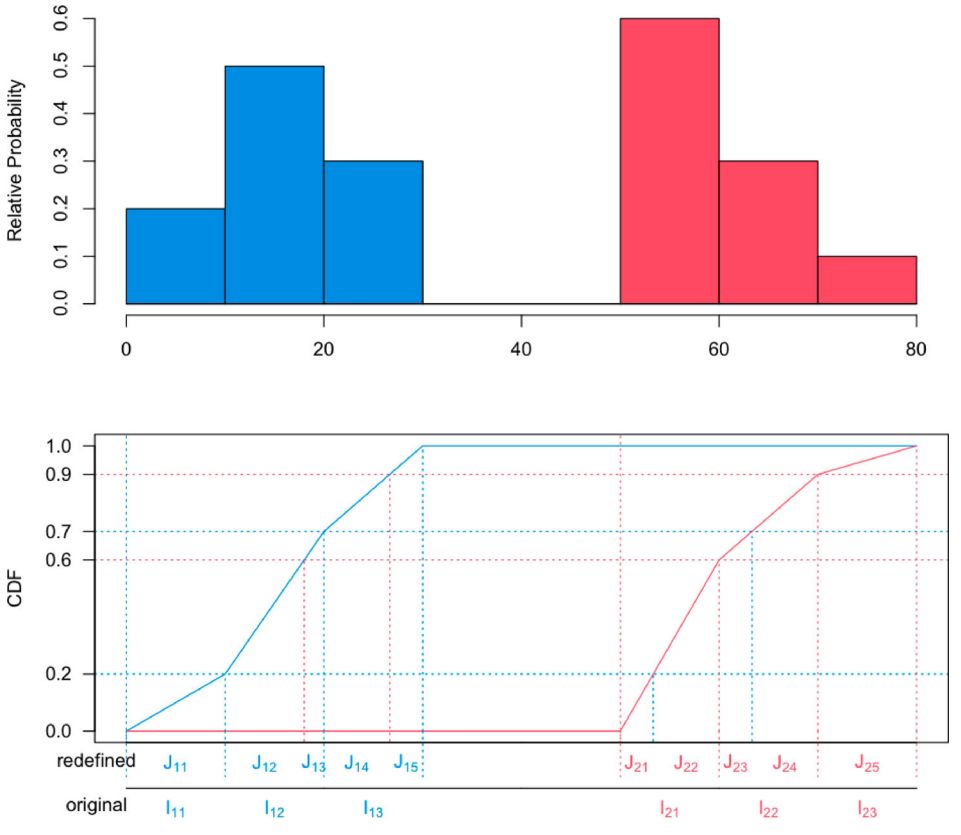


Figure 1. The top panel displays two observed histograms and the bottom shows the corresponding empirical cumulative distribution. The bottom plot illustrates how to obtain the redefined subintervals and the common relative frequency from two histograms.

the uniform distribution assumption within the bins in the histogram of X_{ij} and X_{lj}, respectively. This procedure is illustrated on the bottom panel in Figure 1.

We denote the centers and radii of the sets of subintervals $\{J_{ij1}, \ldots, J_{ijm}\}$ and $\{J_{lj1}, \ldots, J_{ljm}\}$ by $\{(c_{ij1}, r_{ij1}), \ldots, (c_{ijm}, r_{ijm})\}$ and $\{(c_{lj1}, r_{lj1}), \ldots, (c_{ljm}, r_{ljm})\}$, respectively. Then, the WK distance between two univariate histogram-valued observations X_{ij} and X_{lj} in (2) can be expressed as [19,27]:

$$d_W(X_{ij}, X_{lj}) = \sqrt{\sum_{k=1}^{m_{ijl}} \pi_{jk}^{(i,l)} \left[(c_{ijk} - c_{ljk})^2 + \frac{1}{3}(r_{ijk} - r_{ljk})^2 \right]}.$$

For p-dimensional histogram-valued observations $\boldsymbol{X}_i$ and $\boldsymbol{X}_l$, the WK distance can be written as:

$$d_W(\boldsymbol{X}_i, \boldsymbol{X}_l) = \sqrt{\sum_{j=1}^{p} d_W^2(X_{ij}, X_{lj})}.$$

A visualization of the WK distance between two histograms is demonstrated in Verde and Irpino [28].

The inner product associated with the distance is given as [20]:

$$\langle \boldsymbol{X}_i, \boldsymbol{X}_l \rangle_W = \sum_{j=1}^{p} \sum_{k=1}^{m_{ijl}} \pi_{jk}^{(i,l)} \left(c_{ijk} c_{ljk} + \frac{1}{3} r_{ijk} r_{ljk} \right) \tag{3}$$

where $\langle \cdot, \cdot \rangle_W$ denotes the inner product associated with the distance. In the following subsection, we use the induced inner product (3) as the kernel function in the SHM. A brief derivation of (3) is provided in the Appendix.

2.2. Support histogram machine

The SVM is a large margin classifier, as it searches for an optimal separating hyperplane with maximum separation. It has gained popularity in the statistics and machine learning communities and has been applied to various fields, from biology and engineering to social science. Here, we describe the proposed binary SHM, which is an extension of SVM, to histogram-valued data with two classes. We adopt the kernel function induced by the WK metric described in Section 2.1.

Let $f(\boldsymbol{X}) = \beta_0 + h_K(\boldsymbol{X})$ be a classification rule for the proposed SHM, where β_0 is an intercept term. Consider a feature map ϕ from the data space χ to some feature space $\mathcal{F}$, $\phi : \boldsymbol{X} \in \chi \to \mathcal{F}$. Define an associated kernel function $K(\boldsymbol{X}, \boldsymbol{\theta}) = \langle \phi(\boldsymbol{X}), \phi(\boldsymbol{\theta}) \rangle$. We denote the centers and radii of the intervals associated with $\phi(\boldsymbol{X})$ and $\phi(\boldsymbol{\theta})$ as $\{(c_{j1}^*, r_{jk}^*), \ldots, (c_{jm}^*, r_{jm}^*)\}_{j=1}^p$ and $\{(\theta_{jc_1}, \theta_{jr_1}), \ldots, (\theta_{jc_m}, \theta_{jr_m})\}_{j=1}^p$, respectively, and their common relative frequency $\{\pi_{j1}, \ldots, \pi_{jm}\}_{j=1}^p$. Here, $h_K(\boldsymbol{X})$ belongs to a Reproducing Kernel Hilbert Space (RKHS, $\mathcal{H}_K$), i.e. a space of functions with the reproducing property, $h_K(\boldsymbol{X}) = \langle h_K(\cdot), K(\boldsymbol{X}, \cdot) \rangle_{\mathcal{H}_K}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ is an inner product defined on the corresponding function space, and $K(\cdot, \cdot)$ is the attribute space kernel (see, e.g. [10,29]).

Define

$$K(\boldsymbol{X}, \boldsymbol{\theta}) = \sum_{j=1}^{p} \sum_{k=1}^{m} \pi_{jk} \left(c_{jk}^* \theta_{jc_k} + \frac{1}{3} r_{jk}^* \theta_{jr_k} \right), \tag{4}$$

then the objective function for SHM can be expressed as:

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} [1 - y_i f(\boldsymbol{X}_i)]_+ + \frac{\lambda}{2} \|f\|_{\mathcal{H}_K}^2, \tag{5}$$

where $[1-u]_+ = \max(1-u, 0)$ is a hinge loss function, $\|\cdot\|_{\mathcal{H}_K}^2$ denotes the RKHS norm induced by the WK distance, and λ is a tuning parameter that controls the trade-off between the goodness of data fit measured by the hinge loss and the complexity of classifier f.

Among many different algorithms, we utilize the dual approach developed for SVM in Hastie *et al.* [18]. This algorithm is fast at calculating misclassification error rates over a large amount of candidate tuning parameters that are not fixed in advance. It fits the entire solution path depending on the tuning parameter λ from a large value to zero.

Let the primal objective of the problem (5) be

$$L_P(f) = \sum_{i=1}^{n} [1 - y_i f(\boldsymbol{X}_i)]_+ + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}_K}.$$

It is easy to show that the dual problem can be expressed as follows:

$$L_D(\boldsymbol{\alpha}) = \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{l=1}^{n} h_{il} \alpha_i \alpha_l - \sum_{i=1}^{n} \alpha_i + \text{constant},$$

where α_ls are dual parameters and $h_{il} = y_i y_l K(\boldsymbol{X}_i, \boldsymbol{X}_l)$. Then, the resulting classifier is given as:

$$\hat{f}(\boldsymbol{X}_i) = \hat{\beta}_0 + \frac{1}{\lambda} \sum_{l=1}^{n} \hat{\alpha}_l y_l K(\boldsymbol{X}_i, \boldsymbol{X}_l) \tag{6}$$

where $\hat{\beta}_0$ is an estimate of β_0, and $\hat{\alpha}_l$ is an estimate of α_l. We note that the solution (6) depends only on (3) without directly calculating (4). Our SHM enjoys sparsity in the representation of $f(\boldsymbol{X})$, similar to the standard SVM. As in the standard SVM, we call the samples with nonzero α_i's as support histograms.

The tuning parameter choice is an important issue in solving the problem in (5). One commonly used approach is K-fold cross-validation. Based on Hastie *et al.* [18], a trajectory of α along with the tuning parameter λ (i.e. the entire regularization path) from a sufficiently large to a small number is a piecewise linear function. To take advantage of the piecewise linear property, we select the tuning parameter as follows. Assume that the 10-fold cross-validation method is chosen to select the tuning parameter λ. From 10 training subsets we can fit 10 models $\hat{f}_1, \ldots, \hat{f}_{10}$ and keep them. Then, using the corresponding 10 validation subsets, we can evaluate those models' performance and obtain 10 candidates for the tuning parameter. Each of them shows the performance result (misclassification errors) for each validation subset (e.g. λ_1 is the optimal value from model $\hat{f}_1$). Since the 10 classifiers $\hat{f}_1, \ldots, \hat{f}_{10}$ already include the performance outputs over the entire path of tuning parameter λ, the misclassification errors for the 10 candidate values of the tuning parameter $\lambda_1, \ldots, \lambda_{10}$ can be easily calculated from the 10 classifiers. By averaging the 10 misclassification errors for each tuning parameter candidate from all 10 classifiers, we select the tuning parameter λ with the lowest average misclassification error.

Figure 2 shows an example of how the solutions to $\{(\theta_{jc_1}, \theta_{jr_1}), \ldots, (\theta_{jc_m}, \theta_{jr_m})\}$ with $m = 10$ (y axis) vary with the value of $1/\lambda$ (x axis) that controls the complexity of the l_2 norm. We generated the data from Setting 1 described in Section 3.1. It can be seen that the solutions for both center and radius shrink towards zero as λ increases. The green vertical line indicates the optimal λ chosen by 10-fold cross-validation, which is described below. One can also utilize other penalties as in SVM, such as l_1 [32], a combination of l_1 and l_2 [25], and a nonconvex penalty [31].

2.3. An extension to multi-class problem

We extend the proposed SHM from binary to multi-class cases using the one-against-all approach [5]. It is frequently used for multi-class problems because of its relatively simple

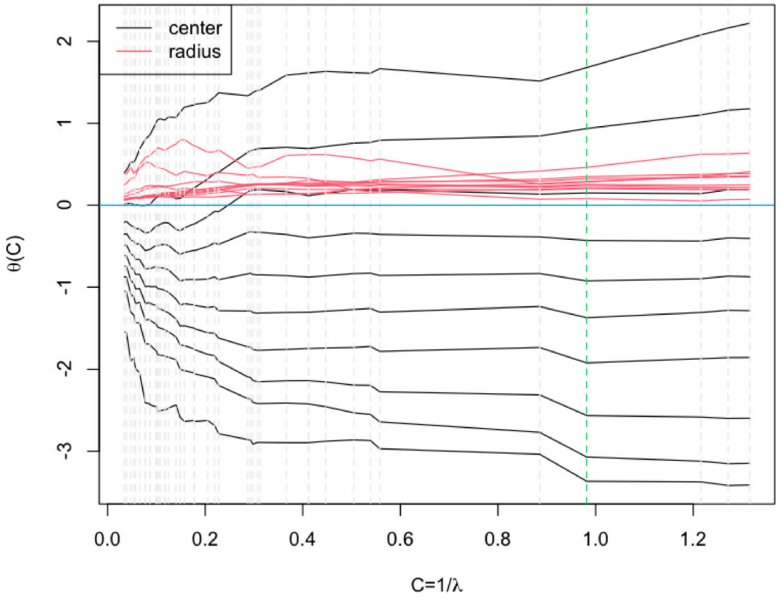


Figure 2. A solution path for SHM. The data are generated from Setting 1 in Section 3.1. The *x* axis represents $1/\lambda$ and the *y* axis the solutions. The black and red solid lines are the solutions for the center and the radius, respectively. The green vertical line indicates the optimal λ chosen by 10-fold cross-validation.

setup of an objective function and computational advantages. Moreover, an extension to histogram-valued data is straightforward by introducing the inner product induced by WK metric to K different classifiers.

Suppose that there are $K > 2$ classes in the data and $y_i \in \{1, \ldots, K\}$ indicates the class to which the ith observation belongs. The one-against-all approach fits K different binary classifiers $\hat{f}_1, \ldots, \hat{f}_K$ separately, and each classifier $\hat{f}_k$ ($k \in \{1, \ldots, K\}$) assigns the observations to the class k versus the rest. This is performed by replacing y_i as a positive class for the observations in class k and as a negative class for those not in class k. The exact path solution can be achieved for all K different classifiers along the entire set of values of the tuning parameter λ. Then, for a single observation X_o in the test data, let $(\hat{f}_1^\lambda(X_o), \ldots, \hat{f}_K^\lambda(X_o))^\top$ be a fitted vector evaluating X_o at a certain value of the tuning parameter λ. Then, a label y_o for X_o is predicted by

$$y_o^\lambda = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \hat{f}_k^\lambda(X_o).$$

3. Simulation study

We perform a Monte Carlo simulation to evaluate the performance of our proposed method. We compare our proposed method to the SVM using vector-valued data (sample means) instead of histograms.

3.1. Binary class cases

There are eight simulation settings for binary classification problems, in which each class follows a different distribution. Additionally, we include the k-NN method using

vector-valued data in the comparison. We generate eight simulation settings as follows. In the first four settings, we consider one-dimensional cases with two classes, having a different distribution for each class. Each class has 100 histogram-valued observations for training, 100 for tuning, and 1,000 for the test set. Each histogram is created from 100 single-valued observations. The single-valued observations of each class are generated from the following distributions:

- Setting 1: $N(0, 9)$, and $N(0.5, 4)$,
- Setting 2 : Gamma$(1, 3)$ and Gamma$(2, 2)$,
- Setting 3 : $0.6N(0, 4) + 0.4N(1, 4)$ and $0.4N(0.5, 9) + 0.6N(1.5, 9)$,
- Setting 4 : $0.7N(0, 4) + 0.3N(1, 9)$ and $0.3N(0, 4) + 0.7N(1, 9)$.

For the next two settings, three-dimensional cases are considered with the same number of histogram-valued observations for training, tuning, and test sets as in the first four settings. We denote $\boldsymbol{\mu}_1 = (0, 0, 0)$ and $\boldsymbol{\mu}_2 = (0.5, 0, 0)$.

- Setting 5 : $MVN(\boldsymbol{\mu}_1, \text{diag}(9, 1, 1))$ and $MVN(\boldsymbol{\mu}_2, \text{diag}(4, 1, 1))$ where $\text{diag}(a, b, c)$ represents a diagonal matrix with diagonal elements a, b and c,
- Setting 6 : $MVt(\boldsymbol{\mu}_1, \Sigma_1)$ and $MVt(\boldsymbol{\mu}_2, \Sigma_2)$ where $\Sigma_1 = \text{diag}(4, 1, 1)$ and $\Sigma_2 = \text{diag}(9, 1, 1)$ with 10 degrees of freedom.

For the remaining two settings, we add noise variables independent of the first three variables in Settings 5 and 6. The number of noise variables is 2, 7 and 17; hence, the total number of variables for classification is $p = 20$, 50, and 100, respectively.

- Setting 7 : Independent $N(0, 1)$ noise variables are added to Setting 5.
- Setting 8 : Independent $N(0, 1)$ noise variables are added to Setting 6.

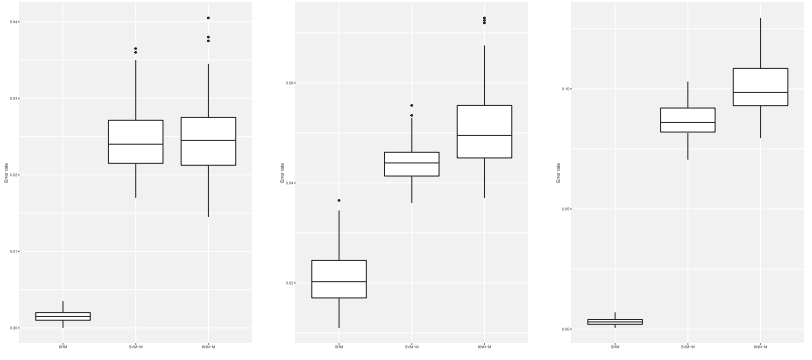
For SVM and SHM, we use the radial basis function $\sigma = 1$ to enhance their flexibility. We choose the tuning parameters in SVM, SHM, and k-NN that minimize misclassification errors in the tuning sets.

The boxplots of the misclassification errors of the three methods for Settings 1–6 with 100 replications are reported in Figure 3. Overall, the proposed SHM outperforms the two other methods with sample means in all settings. The SHM yields substantially lower misclassification errors, indicating that the proposed method using information from histograms is superior to the standard SVM or k-NN using sample means.

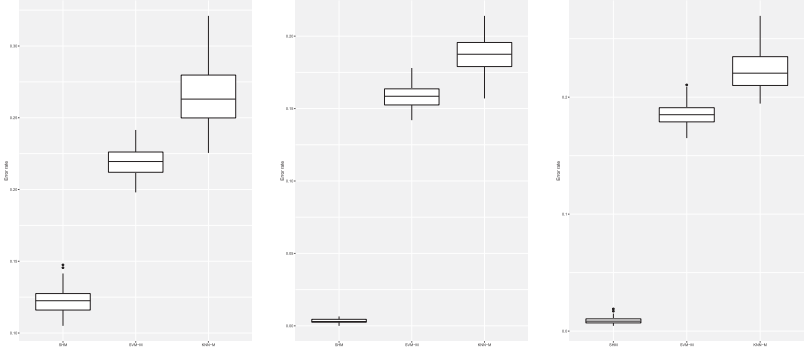
In Settings 7 and 8 (Figure 4), we have added more noise variables to Settings 5 and 6. As the number of noise variables increases, the performances of all methods become slightly worse. However, the pattern of performance remains similar, and the proposed SHM yields the smallest misclassification errors and noticeably outperforms the two methods with sample means.

3.2. Multi-class cases

We perform a simulation study for multi-class cases with the three methods. For SHM and SVM, we take the one-against-all approach. We generate the simulated data in the

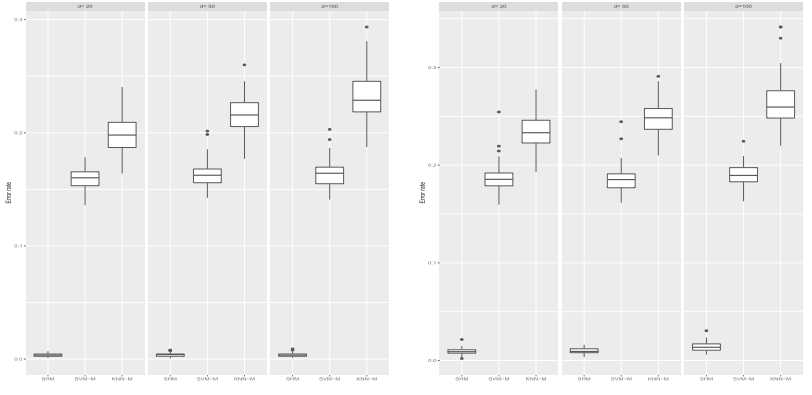


(a) Setting 1 (b) Setting 2 (c) Setting 3



(d) Setting 4 (e) Setting 5 (f) Setting 6

Figure 3. (Binary class cases) Misclassification errors with 100 replications for Settings 1–6. Three methods are compared: SHM, SVM with sample means, and k-NN with sample means. (a) Setting 1. (b) Setting 2. (c) Setting 3. (d) Setting 4. (e) Setting 5 and (f) Setting 6.



(a) Setting 7 (b) Setting 8

Figure 4. (Binary class cases) Misclassification errors with 100 replications for Settings 7 and 8. The total number of variables for classification is $p = 20$, 50, and 100. (a) Setting 7 and (b) Setting 8.

same way as in Section 3.1 with three classes. The first four simulation settings are as follows:

- Setting I: $N(0, 9)$, $N(0.5, 4)$ and $N(-0.4, 4)$,
- Setting II: Gamma$(1, 4)$, Gamma$(2, 2)$ and Gamma$(16, 1/4)$,
- Setting III: $0.6N(0, 4) + 0.4N(1, 4)$, $0.4N(0.5, 9) + 0.6N(1.5, 9)$ and $0.5N(0, 9) + 0.5N(1.5, 4)$,
- Setting IV: $0.7N(0, 4) + 0.3N(1, 9)$, $0.3N(0, 4) + 0.7N(1, 9)$ and $0.5N(0, 4) + 0.5N(1, 9)$.

The next four settings concern multi-dimensional cases. Let $\boldsymbol{\mu}_3 = (1, 0, 0)$.

- Setting V: $MVN(\boldsymbol{\mu}_1, \text{diag}(9, 1, 1))$, $MVN(\boldsymbol{\mu}_2, \text{diag}(4, 1, 1))$ and $MVN(\boldsymbol{\mu}_3, \text{diag}(4, 4, 1))$,
- Setting VI: $MVt(\boldsymbol{\mu}_1, \Sigma_1)$, $MVt(\boldsymbol{\mu}_2, \Sigma_2)$ and $MVt(\boldsymbol{\mu}_3, \Sigma_3)$ where $\Sigma_3 = \text{diag}(4, 4, 1)$ with 10 degrees of freedom,
- Setting VII: Independent $N(0, 1)$ noise variables are added to Setting V,
- Setting VIII: Independent $N(0, 1)$ noise variables are added to Setting VI.

The misclassification error rates of the three methods for Settings I-VI are provided as boxplots in Figure 5 and for Settings VII and VIII in Figure 6. In all eight settings, the proposed SHM shows much lower misclassification error rates than SVM and k-NN with sample means. The proposed SHM can utilize this distributional information in classification, while the other two methods cannot, resulting in a large difference in misclassification error rates.

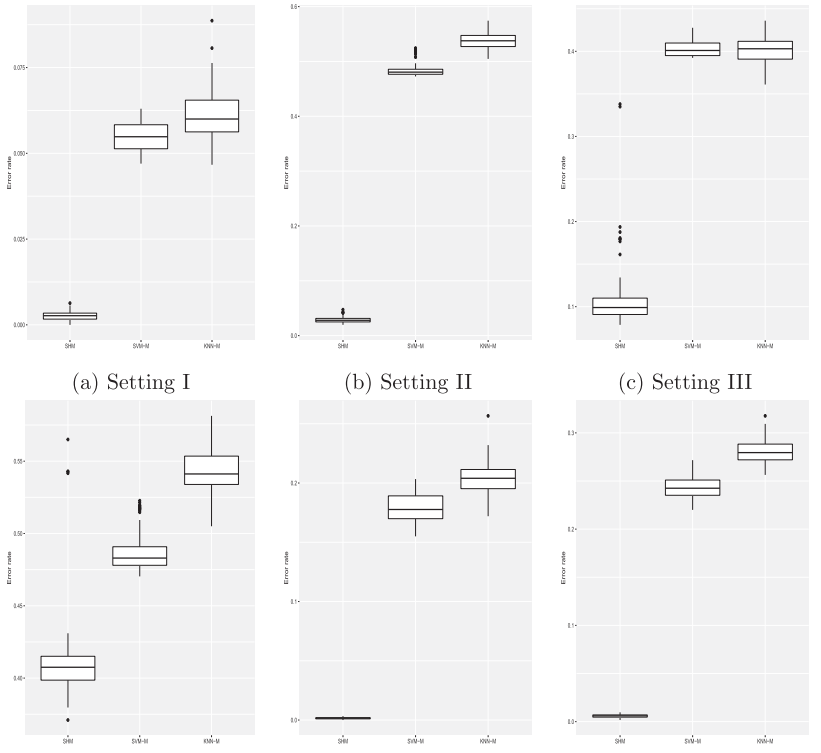
In summary, the simulation demonstrates that the proposed SHM outperforms the methods with sample means. This finding suggests that using distributional information is more beneficial than using single-valued summary values for the classification of histogram-valued data.

4. Real data analysis

We demonstrate the performance of the proposed SHM method and compare it with those of the SVM and k-NN with summary values using the Fashion-MNIST dataset. We apply the 10-fold cross-valid approach to choose the tuning parameters for the three methods.

The Fashion-MNIST dataset (e.g. [30]) consists of pictures of clothes and shoes, captured from Zalando's article. This dataset is available for download from Keras library. Each picture contains 786 pixels (28×28) in grayscale. There are 60,000 pictures in the training set and 10,000 in the test set. The 10 labels are evenly available for both the training and test sets, meaning there are 6000 pictures in the training and 1000 in test one for each label. The 786 pixels within each picture are vectorized and saved to create histogram-valued data.

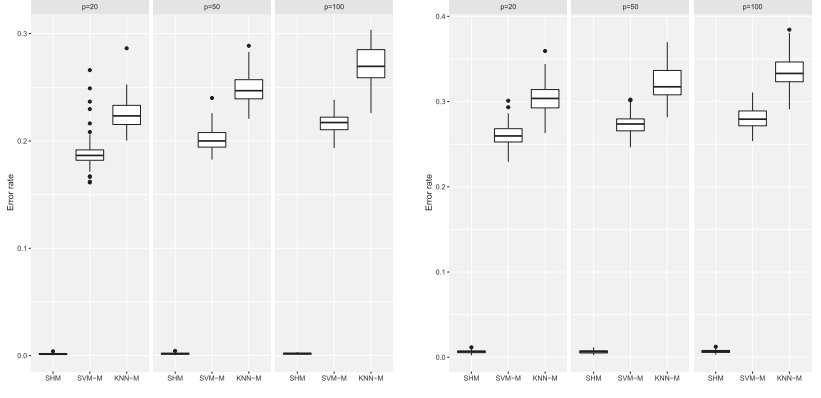
First, we focus on binary classification and choose two out of the 10 categories in the data: Pullover and Sandal (see Figure 7(a,b)). Figure 7(c–f) show two histogram examples for each class. Both exhibit right skewed shapes, but the histograms of the Sandal class are more extremely skewed, which characterizes the different shapes of distribution for each class. To ensure that the histograms from the two labels contrast in terms of shape, the histogram equalization technique [17] is utilized, which is popular in digital image processing.



(a) Setting I (b) Setting II (c) Setting III

(d) Setting IV (e) Setting V (f) Setting VI

Figure 5. (Multi-class cases) Misclassification errors with 100 replications for Settings I–VI. Three methods are compared: SHM, SVM with sample means, and k-NN with sample means to classify three classes. (a) Setting I. (b) Setting II. (c) Setting III. (d) Setting IV. (e) Setting V and (f) Setting VI.



(a) Setting VII (b) Setting VIII

Figure 6. (Multi-class cases) Misclassification errors with 100 replications for Settings VII and VIII. The total number of variables for classification is $p = 20$, 50, and 100. (a) Setting VII and (b) Setting VIII.

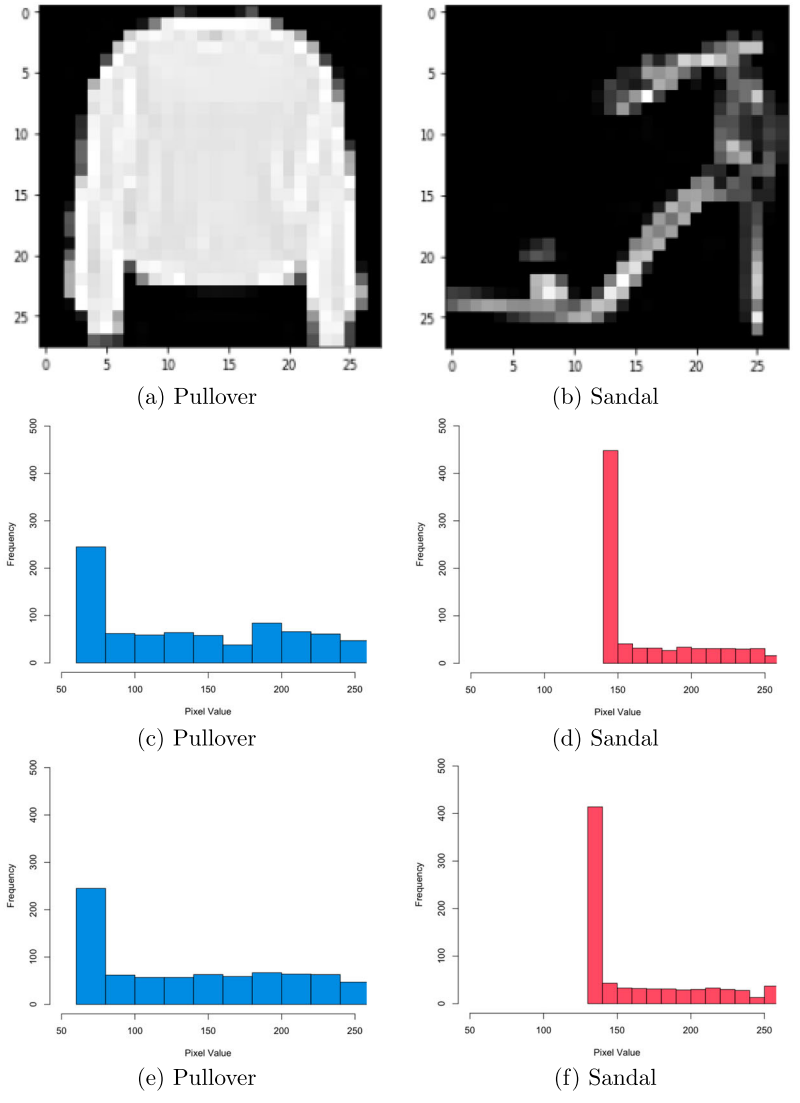


Figure 7. Example of (a) Pullover and (b) Sandal label. (c)–(f) display the selected histograms from two different labels.

Histogram equalization is an image enhancement technique that improves image contrast. We also try two different dimensions: $p = 1$ by using the entire image (one histogram per image) and $p = 2$ by splitting each image in half (two histograms per image).

For SVM and SHM, the radial basis function with $\sigma = 1$ is used for $p = 1$ and $\sigma = 100$ for $p = 2$. The misclassification errors of the three methods (SHM, SVM with sample medians, and k-NN with sample medians) for the Fashion-MNIST dataset are displayed in Table 1. As most histograms do not have a bell shape or are symmetric, the sample median is used as a summary value for each image and applied to SVM and k-NN. For k-NN with $p = 1$, there is a tie issue for this dataset; thus, we add a small noise generated from $N(0, 0.001^2)$ to each observation to address this issue. In Table 1, the evaluation of

Table 1. Misclassification errors for the binary case of the three methods for the Fashion-MNIST dataset.

	SHM	SVM	k-NN
$p = 1$	0.013	0.144	0.066 (0.003)
$p = 2$	0.006	0.164	0.020

Note: For k-NN with $p = 1$, we add a small noise generated from $N(0, 0.001^2)$ to each observation. The evaluation of k-NN is summarized with mean misclassification error rate and its standard deviation (between parentheses) over 1000 replications.

Table 2. Misclassification errors for the 3-class and 4-class cases of the three methods (i.e. SHM, SVM with sample medians, and k-NN with sample medians) for the Fashion-MNIST dataset.

	SHM	SVM	k-NN
3-class	0.0959 (0.0029)	0.3041 (0.008)	0.279 (0.006)
4-class	0.1833 (0.0023)	0.4791 (0.017)	0.450 (0.007)

Note: The mean misclassification error rate and its standard deviation (between parentheses) over the replications are reported.

k-NN is summarized with the mean misclassification error rate and its standard deviation (between parentheses) over 1,000 replications. From Table 1, the proposed SHM yields the lowest misclassification errors for both p=1 and 2. Both SHM and k-NN show improved performance from p=1 to p=2, whereas SVM produces a higher misclassification error for $p = 2$.

Next, we consider the 3-class and 4-class cases with p=1 by adding the Trouser/Pant (for 3-class) and Ankle Boot categories (for 4-class). To expedite the computation, we match the main class's sample sizes and the remaining $K-1$ classes (where $K = 3$ or 4) in the one-against-all approach by subsampling. We repeat this subsampling 50 times for the 3-class case and 25 times for the 4-class case. For k-NN, we add a small noise generated from $N(0, 0.001^2)$ to each observation and repeat the process 1000 times, as in the binary case. In Table 2, the evaluations of all three methods are summarized with the mean misclassification error rate and its standard deviation (between parentheses) over the replications, as mentioned previously.

5. Discussion

We propose the SHM methodology for histogram-valued data with the WK metric. Instead of using the SVM's inner product, we adopted the inner product induced by the WK metric. Our numerical study shows that the proposed method is superior to the standard SVM or k-NN with sample summary values. This result implies that using histograms instead of single-valued summarized values would provide more accurate classification results.

Still, there is room for improvement for the proposed method when handling histogram-valued data with mislabeling. In such a case, a more robust approach would be necessary; introducing case-specific parameters can be an alternative method. Several studies have adopted case-specific parameters in regression to identify potential outliers [23,24]. In future research, we can incorporate case-specific parameters in our objective function to account for potential mislabels and make the classifier more robust.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research of Young Joo Yoon was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03028121). The research of Changyi Park was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2015R1D1A1A01059984). The research of Hosik Choi was supported by the Basic Science Research Program through the NRF funded by the Ministry of Education (2017R1D1A1B05028565).

References

[1] A. Alaei and P.P. Roy, *A new method for writer identification based on histogram symbolic representation*, 14th International Conference on Frontiers in Handwriting Recognition, Heraklion, 2014, pp. 216–221.
[2] C. Angulo, D. Anguita, L.G. Abril, and J.A. Ortega, *Support Vector Machines for Interval Discriminant Analysis*, Neurocomput. 71 (2008), pp. 1220–1229.
[3] L. Billard and E. Diday, *From the statistics of data to the statistics of knowledge: Symbolic data analysis*, J. Am. Stat. Assoc. 98 (2003), pp. 470–487.
[4] L. Billard and E. Diday, *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester, 2007.
[5] L. Bottou, C. Cortes, J.S. Denker, H. Druncker, I. Guyon, L. Jackel, Y. LeCun, U.A. Muller, E. Sackinger, P. Simard, and V. Vapnik, *Comparison of classifier methods: a case study in handwritten digit recognition*, Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 – Conference C: Signal Processing (Cat. No.94CH3440–5), 1994, pp. 77–82 vol.2. https://doi.org/10.1109/ICPR.1994.576879.
[6] E. Carrizosa, J. Gordillo, and F. Plastria, *Classification problems with imprecise data through separating hyperplanes*, Tech. Rep. MOSI/33, MOSI Department, Vrije Universiteit Brussel, 2007.
[7] O. Chapelle, P. Haffner, and V. Vapnik, *Support vector machines for histogram-based image classification*, IEEE Trans. Neural Netw. 10 (1999), pp. 1055–1064.
[8] A. Ciampi, E. Diday, J. Lebbe, E. Périnel, and R. Vignes, *Growing a tree classifier with imprecise data*, Pattern Recognit. Lett. 21 (2000), pp. 787–803.
[9] C. Cortes and V. Vapnik, *Support-vector networks*, Mach. Learn. 20 (1995), pp. 273–297.
[10] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge, 2000.
[11] S. Dias and P. Brito, *Linear regression model with histogram-valued variables*, Stat. Anal. Data. Min. 8 (2015), pp. 75–113.
[12] S. Dias, P. Brito, and P. Amaral, *Discriminant analysis of distributional data via fractional programming*, Eur. J. Oper. Res. Available at https://doi.org/10.1016/j.ejor.2021.01.025.
[13] T. Do and F. Poulet, *Kernel–based Algorithms and Visualization for Interval Data Mining*, *Sixth IEEE International Conference on Data Mining – Workshops (ICDMW'06)*, 2006, pp. 295–299. https://doi.org/10.1109/ICDMW.2006.103
[14] A.P. Duarte Silva and P. Brito, *Discriminant analysis of interval data: An assessment of parametric and distance-based approaches*, J. Classif. 32 (2006), pp. 516–541.
[15] A.P. Duarte Silva and P. Brito, *Linear discriminant analysis for interval data*, Comput. Stat. 21 (2006), pp. 289–308.
[16] A.L. Gibbs and F.E. Su, *On choosing and bounding probability metrics*, Int. Stat. Rev. 70 (2002), pp. 419–435.
[17] R.C. Gonzales and R.E. Woods, *Digital Image Processing*, Prentice Hall, New Jersey, 2002.

[18] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, *The entire regularization path for the support vector machine*, J. Mach. Learn. Res. 5 (2004), pp. 1391–1415.
[19] A. Irpino and R. Verde, *A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data*, in *Data Science and Claissfication*, V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, eds., Springer, Berlin, 2006, pp. 185–192.
[20] A. Irpino and R. Verde, *Basic statistics for distributional symbolic variables: A new metric-based approach*, Adv. Data. Anal. Classif. 9 (2015), pp. 143–175.
[21] J. Kim and L. Billard, *Dissimility measures and divisive clustering for symbolic mulimodal-valued data*, Comput. Statist. Data Anal. 56 (2012), pp. 2795–2808.
[22] J. Kim and L. Billard, *Dissimility measures for histogram-valued observations*, Comm. Statist. Theory Methods 42 (2013), pp. 283–303.
[23] Y. Lee, S.N. MacEachern, and Y. Jung, *Regularization of case-specific parameters for robustness and efficiency*, Stat. Sci. 27 (2012), pp. 350–372.
[24] S. Lee, H. Shin, and S.H. Lee, *Label-noise resistant logistic regression for functional data classification with an application to Alzhemimer disease study*, Biometrics 72 (2016), pp. 1325–1335.
[25] W. Li, J. Zhu, and H. Zou, *The doubly regularized support vector machine*, Stat. Sin. 16 (2006), pp. 589–615.
[26] O.L. Mangasarian, J.W. Shavlik, and E.W. Wild, *Knowledge-based kernel approximation*, J. Mach. Learn. Res. 5 (2011), pp. 1127–1141.
[27] C. Park, H. Choi, C. Delcher, Y. Wang, and Y.-J. Yoon, *Convex clustering analysis for histogram-valued data*, Biometrics 75 (2019), pp. 603–612.
[28] R. Verde and A. Irpino, *Histogram data analysis based on Wasserstein distance* [Conference presentation]. Theory and Application of High-dimensional Complex and Symbolic Data Analysis in Economics and Management Science (2011, October 27–29), Beijing, China. Available at http://www.modulad.fr/sda11/HCSDA11-Verde.pdf.
[29] G. Wahba, *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.
[30] H. Xiao, K. Rasul, and R. Vollgraf, *Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms*, preprint (2017). Available at arXiv:1708.07747.
[31] H. Zhang, J. Ahn, X. Lin, and C. Park, *Gene selection using support vector machines with nonconvex penalty*, Bioinformatics 22 (2006), pp. 88–95.
[32] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, *1-norm support vector machines*. In Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS'03), MIT Press, Cambridge, MA, 2003, pp. 49–56.

Appendix. Derivation of (3)

For the redefined histograms $X_{ij} = \{(J_{ij1}, \pi_{j1}^{(i,l)}), \ldots, (J_{ijm_{ijl}}, \pi_{jm_{ijl}}^{(i,l)})\}$ and $X_{lj} = \{(J_{lj1}, \pi_{j1}^{(i,l)}), \ldots, (J_{ljm_{ijl}}, \pi_{jm_{ijl}}^{(i,l)})\}$, the inverse of the empirical cumulative function with the uniformity assumption within the subintervals is defined as [19]

$$F_{ij}^{-1}(t) = \underline{x}_{ijk} + \frac{t - w_{j(k-1)}^{(i,l)}}{\pi_{jk}^{(i,l)}} |J_{ijk}|,$$

where $\underline{x}_{ijk}$ is the lower bound of J_{ijk} and $|J_{ijk}|$ is the length of the interval J_{ijk}. Hence, the subinterval J_{ijk} is calculated as $J_{ijk} = (F_{ij}^{-1}(w_{j(k-1)}^{(i,l)}), F_{ij}^{-1}(w_{j(k)}^{(i,l)}))$, where $w_{j(k)}^{(i,l)} = \sum_{u=1}^{k} \pi_{ju}^{(i,l)}$ and $w_{j(0)}^{(i,l)} = 0$. The function $F_{lj}^{-1}(t)$ can be similarly defined. Then, the inner product based on $d_W(\cdot, \cdot)$ is defined as

follows [20]:

$$
\begin{aligned}
&\langle \boldsymbol{X}_i, \boldsymbol{X}_l \rangle_W \\
&\quad = \sum_{j=1}^{p} \sum_{k=1}^{m_{ijl}} \int_{w_{j(k-1)}^{(i,l)}}^{w_{j(k)}^{(i,l)}} \left(F_{ij}^{-1}(t) F_{lj}^{-1}(t) \right) \mathrm{d}t \\
&\quad = \sum_{j=1}^{p} \sum_{k=1}^{m_{ijl}} \int_{w_{j(k-1)}^{(i,l)}}^{w_{j(k)}^{(i,l)}} \left(c_{ijk} - r_{ijk} + 2 r_{ijk} \frac{t - w_{j(k-1)}^{(i,l)}}{\pi_{jk}^{(i,l)}} \right) \left(c_{ljk} - r_{ljk} + 2 r_{ljk} \frac{t - w_{j(k-1)}^{(i,l)}}{\pi_{jk}^{(i,l)}} \right) \mathrm{d}t \\
&\quad = \sum_{j=1}^{p} \sum_{k=1}^{m_{ijl}} \pi_{jk}^{(i,l)} \int_0^1 \left(c_{ijk} + r_{ijk}(2t-1) \right) \left(c_{ljk} + r_{ljk}(2t-1) \right) \mathrm{d}t \\
&\quad = \sum_{j=1}^{p} \sum_{k=1}^{m_{ijl}} \pi_{jk}^{(i,l)} \left(c_{ijk} c_{ljk} + \frac{1}{3} r_{ijk} r_{lik} \right).
\end{aligned}
$$

See Irpino and Verde [19,20] for more details.