Estimation and Clustering of Directional Wave Spectra

Zihao Wu[1], Carolina Euan[2], Rosa M. Crujeiras[3] and Ying Sun[4]

June 16, 2023

Abstract

The directional wave spectrum (DWS) describes the energy of sea waves as a function of frequency and direction. It provides useful information for marine studies and guides the design of maritime structures. One of the challenges in the statistical estimation of DWS is to account for the circular nature of direction. To address this issue, this paper considers the 1-dimensional case of the direction-only DWS (DWSd) and applies the circular regression to smooth the DWSd observations. This paper then improves an existing clustering algorithm by incorporating circular smoothing in the clustering algorithm, automating the determination of the optimal number of clusters, and designing a more appropriate smoothing parameter selection procedure for data with correlated errors. Our simulation studies reveal an improvement in the performance of estimating the underlying DWSd using the circular smoother. Finally, the linear and circular smoothers are compared by clustering two real datasets, one from the Sofar Ocean network and the second from a buoy located at the Red Sea. For the Sofar Ocean data, clustering with the two smoothers results in different number of clusters. For the Red Sea data, a cluster with a peak at the boundary is only identified when the circular smoother is used.

Keywords: wave spectra, clustering, circular regression, data visualization

[1]Department of Statistics and Applied Probability, National University of Singapore, Singapore. E-mail: zihao.wu@u.nus.edu

[2]Department of Mathematics and Statistics, Lancaster Univerisity, Lancaster, LA1, 4YF, UK. E-mail: c.euancampos@lancaster.ac.uk

[3]CITMAga, Universidade de Santiago de Compostela, Spain. E-mail: rosa.crujeiras@usc.es

[4]CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. E-mail: ying.sun@kaust.edu.sa

1 Introduction

The directional wave spectrum (DWS) describes the energy of sea waves as a function of frequency and direction. Real-time DWS can provide information for storm tracing (Hanson & Phillips, 2001) and in-service monitoring systems on ships (Nielsen, 2006). Long-term DWS data, on the other hand, is useful for the determination of dominant wave profiles through cluster analysis (Euan & Sun, 2019). Usages of dominant wave profiles include reducing the computational cost for the estimation of riser fatigue on deepwater floating production systems (Vogel et al., 2016), guiding the design and construction of maritime structures (Boukhanovsky & Guedes Soares, 2009), and helping the exploitation of wave energy (Ribeiro et al., 2020). Furthermore, regional wave climates can be identified by clustering DWS observations, benefiting the studies of coastal and marine processes (Mortlock & Goodwin, 2015).

Despite the great efforts devoted into data collection and derivation of DWS (Benoit, 1993; Boukhanovsky & Guedes Soares, 2009; Hisaki, 1996; Nielsen, 2006; Nielsen & Dam, 2008; Waals et al., 2002; Young, 1994; Yurovskaya et al., 2013), these DWS are still observational and inevitably contain noise. This gives rise to the demand of estimating or smoothing the DWS to better reflect the features of the sea state, for which available solutions are scarce. Note that the derivation of DWS data is named by some authors as "estimation", whereas our aim is to develop a statistical method to estimate the underlying DWS from the derived data via smoothing.

Pascoal & Guedes Soares (2008) presented a parametric method by imposing a parametric constraint to smooth the DWS. Nevertheless, Hinostroza & Guedes Soares (2016) pointed out that parametric modelling fits to the measurements a prescribed shape, which may not be suitable for all scenarios. Therefore, it is of interest to introduce a more flexible non-

parametric method for the estimation of the DWS. Through statistical estimation, our work aims to improve the quality of existing DWS data, thus increase the value of DWS in its various applications.

In real applications, when the DWS does not show interesting features in frequency, it can be reduced to a direction-only DWS (DWSd) by integrating over frequencies. For example, Gorman (2018) considered the derivation of DWSd data with multiple peaks (waves coming from multiple directions) from a wave buoy deployment, and Euan & Sun (2019) further developed a clustering algorithm for DWSd. Although the estimation of the DWSd is a standard regression problem, one challenge in applying statistical tools to DWS is that direction is a circular variable and needs proper treatment beyond classical Euclidean methods. This leads to a regression framework with a circular variable (direction) and linear response (energy).

Di Marzio et al. (2009) introduced a local linear approach to handle kernel regression when dealing with a circular covariate and a linear response. This method has not been explored for estimating DWSd. Hence, the first goal of this paper is to evaluate the suitability of this technique on DWS data by comparing against the ordinary linear kernel smoother (that is, a Nadaraya-Watson smoother over the real line). The second goal of this paper is to develop a clustering algorithm for smoothed DWSd, as most of previous studies have applied cluster analysis on the "raw" DWS data without statistical estimation (Hamilton, 2010; Portilla-Yandún et al., 2015; Ribeiro et al., 2020). Euan & Sun (2019) developed a clustering algorithm for smoothed DWSd using the classical non-parameric regression model with linear covariates. In this work, we will improve the clustering algorithm with a novel procedure to automate the determination of the optimal number of clusters and a more appropriate method for selecting the smoothing parameter of kernel smoothing. With more accurate estimation, better clustering results can be expected. Finally, we will develop

visualization tools to visualize the identified DWSd clusters.

The rest of the paper is organized as follows. Section 2 provides background on DWS and how kernel regression for circular-linear systems can be used in this setting, followed by an introduction to the proposed clustering algorithms for the smoothed DWSd. Section 3 first presents a simulation study where the performance of the circular-linear regression is compared with other alternatives that ignore the circular nature of the direction, and then illustrates how the clustering algorithm works in practice, in different simulated scenarios. Two detailed real data analyses are presented in Section 4. Finally, Section 5 discusses the limitations and possible extensions of the proposed methods.

2 Statistical methodology

This section is devoted to the introduction of the kernel regression approach for the DWSd curves and to describe how these smooth estimates can be used for improving clustering results.

2.1 Circular kernel regression for DWSd

The directional wave spectrum (DWS) denoted by $S(\omega, \theta)$ describes the energy of sea waves as a continuous function of frequency ω and direction θ. In this study, we define the direction-only DWS (DWSd) according to (Euan & Sun, 2019) as

$$g(\theta) = \int S(\omega, \theta) d\omega, \qquad \omega > 0, \theta \in [0, 2\pi) \tag{1}$$

which represents the corresponding total energy in each direction. Suppose that for a collection of directions $\{\theta_l, l = 1, 2, \ldots, n_\theta\}$, we observe the DWSd $\{g^{\text{obs}}(\theta_l), l = 1, 2, \ldots, n_\theta\}$.

These observations are assumed to be generated from the model

$$g^{\text{obs}}(\theta_l) = g(\theta_l) + \epsilon_l, \qquad l = 1, 2, \ldots, n_\theta \tag{2}$$

where $g(\theta)$ is the underlying DWSd and ϵ_l's are random but correlated errors with mean zero and variance σ^2. To estimate $g(\theta)$ while accounting for the circular nature of the covariate (direction), Di Marzio et al. (2009) proposed a local linear approach by fitting, locally, a trigonometric polynomial $\beta_0 + \beta_1 \sin(\cdot - \theta)$. This local fit, for a fixed direction θ, is obtained by solving a least squares problem and the corresponding parameter estimates are given by

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{(\beta_0,\beta_1)}{\operatorname{argmin}} \left\{ \sum_{l=1}^{n_\theta} K_\nu(\theta - \theta_l)[g^{\text{obs}}(\theta_l) - (\beta_0 + \beta_1 \sin(\theta - \theta_l))]^2 \right\} \tag{3}$$

The local linear estimator of $g(\theta)$ for a certain θ is given by $\widehat{g}_\nu(\theta) = \widehat{\beta}_0$. In expression (3), $K_\nu(\cdot - \mu)$ denotes a circular kernel: a circular density which is usually taken as a von Mises model, centered at μ (in the least squares problem, centered at each observation θ_l) and with concentration parameter ν. This parameter controls the smoothness of the estimator, with large concentrations yielding very wiggly estimates of g and with small values of ν giving too smooth curves. As in any other kernel smoothing setting, the selection of the smoothing parameter ν is a crucial task. Note that the behavior of the smoothing parameter in circular regression (a concentration) is opposite to the behavior of the smoothing parameter in linear regression (a bandwidth). In general, we will refer to a smoothing parameter in both settings. If instead of fitting locally a trigonometric polynomial one restricts the least squares problem in (3) to squares differences between the response values and a (locally) constant fit, then a Nadaraya-Watson type estimator is obtained. Despite for real-valued variables, the local linear method usually outperforms Nadaraya-Watson (with most remarkable output differences at the boundaries of the support of the real-valued explanatory variable),

this is not the case for circular covariates, where no boundary effects are found and the two estimators (local linear and local constant) present similar behaviors.

In this work, the kernel regression for circular-linear variables (KRCL) will be evaluated and compared against the ordinary kernel regression smoother (KRL, where L stands for linear) for the estimation of DWSd. The function in the package (Oliveira et al., 2014) and the function in base R are used to perform KRCL and KRL, respectively. Both of them use the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964).

2.2 Clustering algorithms

To investigate whether more accurate estimation leads to better clustering results, this work employs the direction-only Hierarchical Directional Spectra-based (HDSd) clustering algorithm proposed by Euan & Sun (2019). To classify a set of n DWSd observations $\{g_i(\theta), i = 1, 2, \ldots, n\}$, the algorithm runs as follows:

Step 1. Compute the normalized functions $\{g_i^N(\theta), i = 1, 2, \ldots, n\}$, where $g_i^N(\theta) = \dfrac{g_i(\theta)}{\int g_i(\theta) d\theta}$. Define n clusters as $\mathcal{C}_i = \{g_i^N\}$ for $i = 1, 2, \ldots, n$. The representative function of $\mathcal{C}_i$ is g_i^N for $i = 1, 2, \ldots, n$.

Step 2. Compute the dissimilarity between every pair of clusters. Examples of the dissimilarity measure include the total variation distance (TVD) and the squared Euclidean distance (SED), which are defined as

$$\text{TVD}(\mathcal{C}_i, \mathcal{C}_j) = \frac{1}{2} \int \Big| g_i(\theta) - g_j(\theta) \Big| d\theta$$

$$\text{SED}(\mathcal{C}_i, \mathcal{C}_j) = \int \Big[g_i(\theta) - g_j(\theta) \Big]^2 d\theta$$

where g_i and g_j are the representative functions of the clusters $\mathcal{C}_i$ and $\mathcal{C}_j$, respec-

tively.

Step 3. Find the most similar pair $\mathcal{C}_i$ and $\mathcal{C}_j$ according to the dissimilarity measure. Form a new cluster $\mathcal{C}_{\text{new}} = \mathcal{C}_i \cup \mathcal{C}_j$. Remove $\mathcal{C}_i$ and $\mathcal{C}_j$ from the collection of clusters. Record k, the number of remaining clusters and d_k, the dissimilarity between $\mathcal{C}_i$ and $\mathcal{C}_j$ for $k = n-1, n-2, \ldots, 1$. The representative function for $\mathcal{C}_{\text{new}}$ is the mean of all functions in $\mathcal{C}_{\text{new}}$.

Step 4. Repeat steps 2 and 3 until there is only one cluster left.

Note that the algorithm depends on the distance between each pair of the input DWSd and does not assume any chronological order on the data. This paper proposes two improvements to the algorithm: first, the determination of the optimal number of clusters is automated; second, statistical estimation is incorporated into the process of clustering to obtain a smoothed version of each DWSd in each cluster.

2.2.1 Determination of the Number of Clusters

A scree plot characterizing the clustering result can be obtained by plotting d_k against k, which are recorded in the HDSd algorithm. A sample scree plot is shown in Figure 1(a). From this plot, the *elbow* method is considered to determine k, identifying the optimal number of clusters as a small value of k where the dissimilarity does not present a relevant change.

The elbow is determined according to the idea illustrated in Figure 1(b-d). First, a reference line is formed by connecting the first point $(1, d_1)$ and the n_{ref}-th point $(n_{\text{ref}}, d_{n_{\text{ref}}})$. Denote the height of the reference line at $x = k$ as D_k The optimal number of clusters is then selected as

$$k_{\text{selected}} = \underset{k}{\operatorname{argmax}} \left\{ D_k - d_k \right\}, \quad k = 1, 2, \ldots, n_{\text{ref}} \tag{4}$$

The parameter n_{ref} is introduced because if we just use the rightmost point, the result of k_{selected} will be susceptible to the initial number of clusters. As shown in Figure 1, a bigger n_{ref} may lead to a larger k_{selected} (depending on the shape of the scree plot), and vice versa. In practice, to choose n_{ref} given a new dataset, we suggest starting by choosing n_{ref} = 3-5 times of the number of clusters to expect based on the background knowledge about the data (e.g., theory). If the dataset is noisy (for DWSd, we can check if there are fake peaks), we can decrease n_{ref} by 10-20 percent. It is also recommended to double-check whether the elbow determined is reasonable by plotting the scree plot (these steps will be demonstrated in the Application section).

Hence, a larger number should be chosen for n_{ref} if more clusters are expected to be identified. Conversely, if strong noise is present and/or the clusters are expected to be distinctive, then a smaller n_{ref} is more suitable.

2.2.2 Incorporation of Estimation in Clustering

As pointed out in section 2.1, the smoothing parameter plays a key role in the estimation, so it is crucial to consider an appropriate and data-driven smoothing parameter selection procedure. The leave-one-out cross validation (LCV) ideas have been used both in linear and circular contexts. However, in our setting, the LCV is precluded because the errors in the same observation can be correlated, in which case the LCV smoothing parameter tends to be too small (De Brabanter et al., 2011). Therefore, the following alternative is proposed. First, the HDSd method is applied to the raw observations and the elbow method is used to determine the clusters. The result is called an *initial grouping*. To distinguish between the clusters in the initial grouping and the final clusters, in this paper we use "group" for the former and "cluster" for the latter. Observations in each group are assumed to have the same truth and identically distributed errors. Since the smoothing parameter should only

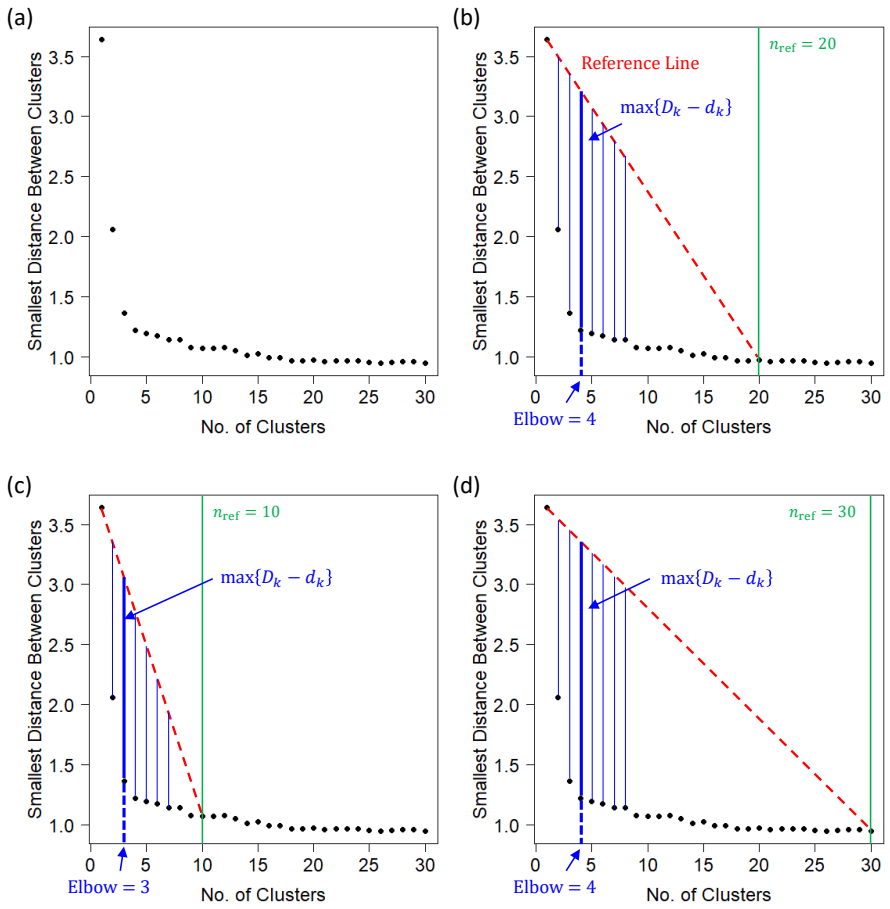

Figure 1. Sample scree plot (a) and the determination of elbow with $n_{\text{ref}} = 20$ (b), 10 (c), and 30 (d). The elbow is determined as the number of clusters that maximizes the distance between the reference line and the smallest distance between clusters.

depend on the shape of the truth and the correlation of the errors, the same value shall be selected for all observations in the same group. The proposed procedure is summarized below:

Step 1. Let $\{g_i^{\text{obs}}(\theta), i = 1, 2, \ldots, n\}$ be DWSd observations from the same group. Compute the mean of the DWSd observations $\bar{g}^{\text{obs}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i^{\text{obs}}(\theta)$.

Step 2. Determine the smoothing parameter as

$$\nu_{\text{selected}} = \underset{\nu}{\text{argmin}} \left\{ \sum_{i=1}^{n} \int \left[\widehat{g}_{i,\nu}(\theta; \nu) - \bar{g}^{\text{obs}}(\theta) \right]^2 d\theta \right\} \tag{5}$$

where $\widehat{g}_{i,\nu}(\theta; \nu)$ denotes the estimated $g_i(\theta)$ with concentration ν.

In this paper, initial groupings and final clusters with size smaller than a certain threshold will be discarded as outlier groups. The threshold will be decided based on the size of the data.

The following steps summarize the full clustering procedure proposed:

Step 1. Apply the HDSd algorithm to the DWSd observations. Use the elbow method to determine the number of clusters and obtain an initial grouping of the observations.

Step 2. Select an appropriate smoothing parameter for each group.

Step 3. For each group, apply the non-parametric estimator with the selected smoothing parameter to obtain a smoothed version of each observation that estimates the true DWSd.

Step 4. Apply the HDSd algorithm to the smoothed observations. Use the elbow method to determine the number of clusters and obtain the final results.

2.2.3 Visualization of the Identified Clusters

In this study, the functional boxplot proposed by Sun & Genton (2011) will be adopted to visualize the identified clusters. Let $\{g_1(\theta), g_2(\theta), \ldots, g_n(\theta)\}$ be a set of DWSd and $\{g_{[1]}(\theta), g_{[2]}(\theta), \ldots, g_{[n]}(\theta)\}$ be the corresponding ordered set according to the decreasing values of the modified band depth (MBD) introduced by López-Pintado & Romo (2009). The MBD measures the centrality of each $g(\theta)$ with respect to the whole set. The curve with the largest depth value ($g_{[1]}(\theta)$) is the median. The upper and lower borders of the box are given by the boundaries of the 50% deepest curves $\min_{r=1,2,\ldots,\lceil n/2 \rceil} g_{[r]}(\theta)$ and $\max_{r=1,2,\ldots,\lceil n/2 \rceil} g_{[r]}(\theta)$. To visualize the distribution of energy over different directions, the functional boxplots will be presented using polar coordinates, where the angle corresponds to the direction. Figure 2 shows an example of functional boxplot and directional functional boxplot produced using the same data.

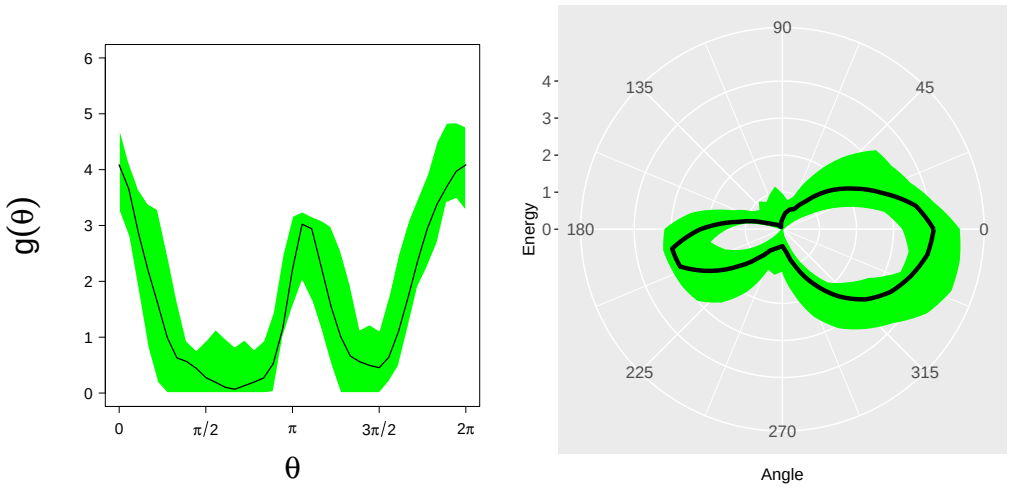

Figure 2. Sample functional boxplot (left) and directional functional boxplot (right) of the total energy $g(\theta)$ in direction θ. The same set of simulated observations are used to generate the two plots. The black curve is the median observation, and the colored area indicates the inter-quartile range. The "energy" scale is equivalent to $g(\theta)$. The numbers correspond to the radius at 90 degrees.

3 Simulation Studies

Two simulation studies are conducted using R (version 4.0.2) (R Core Team, 2020). The first assesses the performance of kernel regression for circular-linear variables (KRCL) in estimating the true DWSd from observations. The second apply KRCL for clustering DWSd observations. Both studies compare KRCL against the ordinary kernel regression smoother (KRL).

To obtain the simulated DWSd observations we consider a stochastic model for wave systems (Alliot et. al, 2013). The DWS observations are simulated as

$$S^{\text{obs}}(\omega_q, \theta_l) = S(\omega_q, \theta_l) + \epsilon(\omega_q, \theta_l), \qquad q = 1, 2, \ldots, n_\omega,\ l = 1, 2, \ldots, n_\theta \tag{6}$$

where $\epsilon(\omega, \theta) \sim N(0, \sigma^2)$ unconditionally with

$$\text{Cor}\{\epsilon(\omega_i, \theta_i), \epsilon(\omega_j, \theta_j)\} = \exp\left[-\lambda_1 \left|\frac{1}{\omega_i} - \frac{1}{\omega_j}\right| - \lambda_2(1 - \cos(\theta_i - \theta_j))\right] \tag{7}$$

where σ, λ_1 and λ_2 are parameters. Following Alliot et. al (2013) and Ochi (1998), we use the model $S(\omega, \theta) = f(\omega)D(\omega, \theta)$ where $f(\omega)$ is formulated using the Joint North-Sea Wave Project (JONSWAP) spectral family (Hasselmann et al., 1973)

$$f(\omega) \propto \omega^5 \exp(-5\omega_p^4/4\omega^4)\gamma^{\exp(-(\omega-\omega_p)^2/(2\omega_p^2 r^2)} \tag{8}$$

and the directional spreading function given by Longuet-Higgins et al. (1963)

$$D(\omega, \theta) \propto \cos^{2m}\left(\frac{\theta - \theta_0}{2}\right) \tag{9}$$

In (8), $r = 0.07$ if $\omega \leq \omega_p$ and $r = 0.09$ otherwise; $\omega_p = \pi/T_p$ with T_p being the spectral peak period. Both T_p and γ are parameters. In (9), θ_0 is the peak direction, $m = m_{\max}(\omega/\omega_p)^5$ if

$\omega < \omega_p$ and $m = m_{\max}(\omega/\omega_p)^{-5/2}$ otherwise, and $m_{\max}$ is a parameter. This representation is a modelling strategy that is useful for the physical interpretation and simulation of ocean waves.

Each simulated S^{obs} has $n_\omega = 40$ frequency levels and $n_\theta = 40$ directions. Note that $S(\omega, \theta) + \epsilon(\omega, \theta)$ may contain negative values. Since this study focuses on the 1-dimensional situation, the negative values are handled with respect to the corresponding DWSd of the observations $g^{\text{obs}}(\theta)$, i.e.,

$$g^{\text{obs}}(\theta_l) = \max\{0, \int S^{\text{obs}}(\omega, \theta_l) d\omega\}. \tag{10}$$

3.1 Estimation

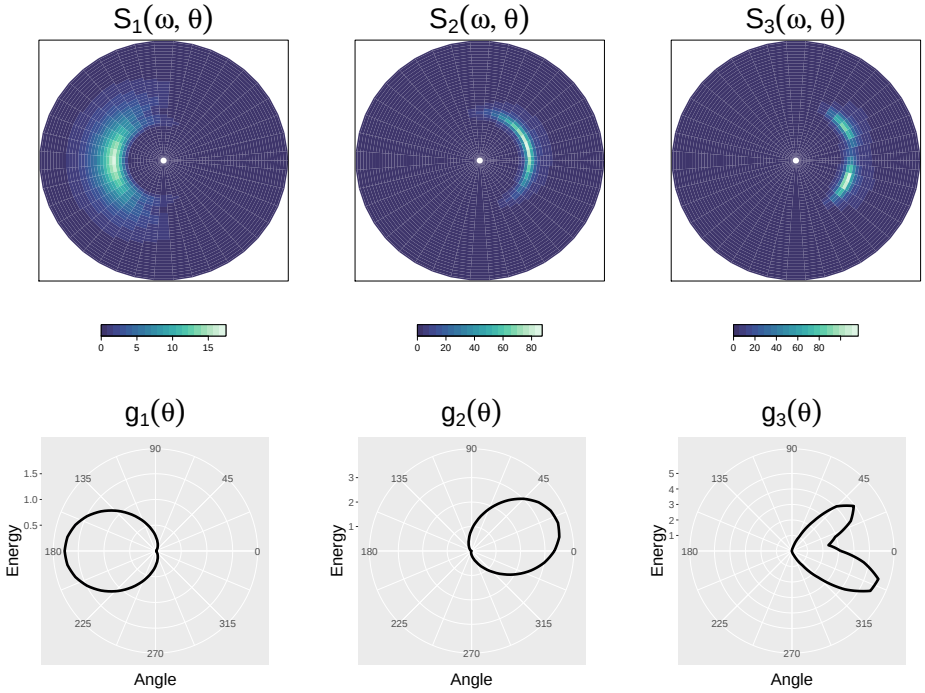

Figure 3. Target DWS and the corresponding DWSd designed for estimation. Each column represents one target. The first row shows the DWS in polar coordinates, where the radius represents the frequency ω, the angle represents the direction θ, and the color scale represents the value of $S(\omega, \theta)$. The second row shows the corresponding DWSd. The "energy" scale is equivalent to $g(\theta)$. The numbers correspond to the radius at 90 degrees.

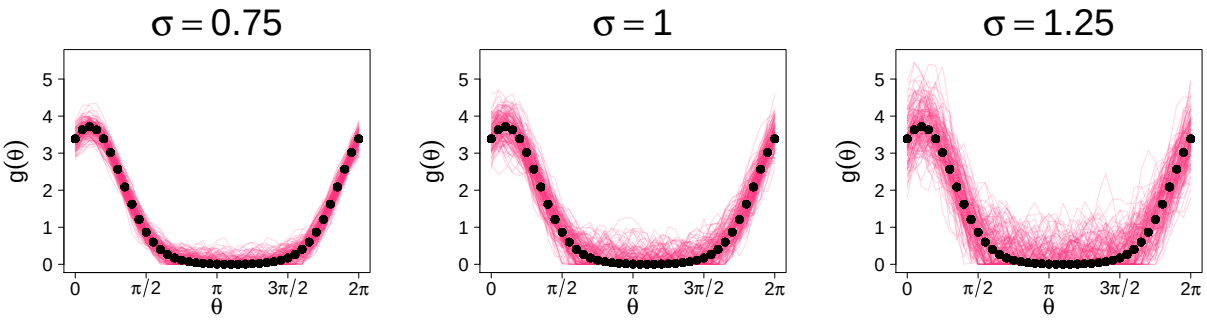

Figure 4. Sample simulated DWSd observations under different choices of σ. Black dots represents the truth ($g_2(\omega)$); pink lines represent the simulated observations.

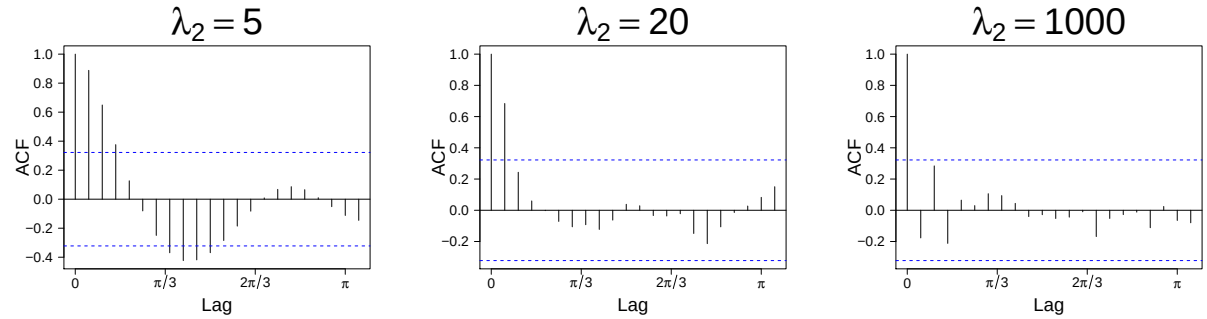

Figure 5. Sample auto-correlation functions under different choices of λ_2. Lag is in terms of θ. An ACF bar longer than the blue dotted line indicates a significant correlation.

Three target DWS S_1, S_2 and S_3 and the corresponding DWSd g_1, g_2 and g_3 are shown in Figure 3. We consider 3 noise levels $\sigma = 0.75, 1, 1.25$ and 3 correlation levels $\lambda_2 = 5, 20, 1000$. For each σ, the extent of noise is illustrated in Figure 4. For each λ_2, the auto-correlation function is plotted in Figure 5. Since the frequency dimension is ignored in DWSd, we fix $\lambda_1 = 3$ throughout the simulation.

We first show the estimations of one randomly generated DWSd observation from g_2 and g_3 by KRCL and KRL in Figure 6. The estimates of KRCL and KRL only differ near the boundary (when θ is close to 0 or 2π). Since KRCL treats the support as circular, it utilizes the information near both boundaries for its estimation. Hence, the KRCL estimates near $\theta = 0$ and $\theta = 2\pi$ are "adjusted" by each other. While this effect may improve the average performance of KRCL estimation, it does not necessarily lead to an estimation closer to the truth in all cases. For example, the sample KRCL estimate for g_2 is only closer to the truth compared to KRL near $\theta = 2\pi$ but not near $\theta = 0$ (Figure 6a).

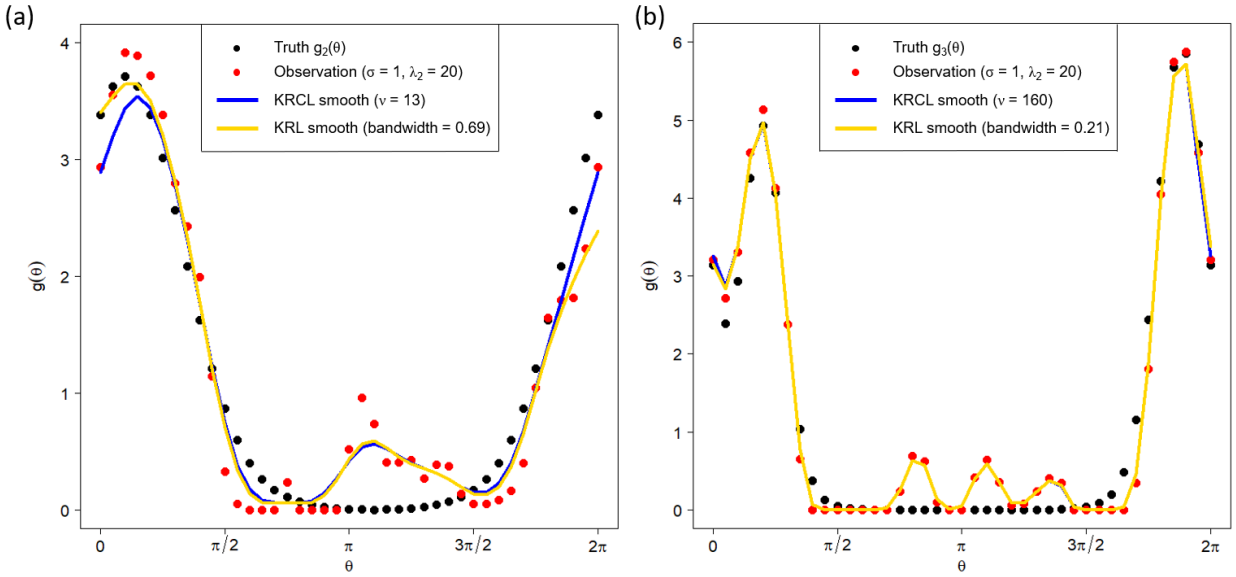

Figure 6. Sample DWSd estimations of g_2 (a) and g_3 (b) using KRCL (blue) and KRL (yellow). The smoothing parameters are selected according to the method described in Section 2.2.2.

For g_3 (Figure 6b), the difference between the KRCL and KRL estimates near the bound-

ary is less prominent. This is because g_3 has narrow peaks, leading to small bandwidths / large concentrations being selected (so that the two peaks are preserved in the estimates). With a small bandwidth / large concentration, the information used for the estimation at each point is more concentrated. Hence, KRCL utilize less information from the other side of the boundary. As a result, the KRCL estimate becomes more similar to the KRL estimate.

Now we consider replicates. For each combination of $S(\omega, \theta), \sigma$ and λ_2, $M = 100$ observations are generated. After selection of smoothing parameters (according to Section 2.2.2), KRCL and KRL are applied to the observations to obtain an estimate from each observation. The performance is represented by the error, which is measured by the mean of the squared Euclidean distance between the estimates and the truth

$$\frac{1}{M}\sum_{j=1}^{M}\int[\widehat{g}_{i,j,\nu_i}(\theta) - g_i(\theta)]^2 d\theta, \qquad i = 1, 2, 3$$

where ν_i is the smoothing parameter selected for group i and $\widehat{g}_{i,j,\nu_i}(\theta)$ are the estimates from each observation. In order to determine whether KRCL is significantly better than KRL, the Wilcoxon signed rank test is applied to the M pairwise differences in the errors. The alternative hypothesis is that for the same observation, the error of KRCL estimation is less than that of KRL. Since the test is one-sided, a p-value smaller than 0.05 favors KRCL. The results are summarized in Table 1.

For g_1 and g_2, the consistent small p-values indicate that KRCL does give better estimates than KRL. For g_3, however, the performances of the two techniques do not differ significantly. A possible explanation is as follows: non-parametric smoothers tend to underestimate sharp peaks and troughs, because the highest or lowest point is always adjusted by its neighbors. Since KRCL uses also information from the other side of the boundary, its estimates for the two narrow peaks will be more flat compared to KRL. This introduces errors that offset the

advantage of KRCL mentioned before, which is already small due to the small bandwidth / large concentration selected.

Table 1. Median errors between the true $g(\theta)$ and the estimates obtained using KRCL and KRL. The parameters σ and λ_2 indicate the noise level and the within-observation correlation level of the simulated DWSd observations. For each setting, $M = 100$ observations are simulated, and the median error is reported. The Wilcoxon signed rank test is then applied to test against the hypothesis that for each simulated DWSd observation, the KRCL estimate has a higher error than KRL. The p-values are given in brackets.

(a) Median errors of estimating g_1 using KRCL and KRL.

KRCL/KRL (p-value)		λ_2 (correlation)		
		5 (strong)	20 (moderate)	1000 (very weak)
σ	0.75 (weak noise)	0.044/0.044 (0.006)	0.034/0.034 (0.002)	0.015/0.015 (<0.001)
	1 (moderate noise)	0.133/0.134 (0.003)	0.081/0.082 (0.001)	0.044/0.044 (0.006)
	1.25 (strong noise)	0.264/0.264 (0.001)	0.162/0.162 (0.010)	0.095/0.095 (0.004)

(b) Median errors of estimating g_2 using KRCL and KRL.

KRCL/KRL (p-value)		λ_2 (correlation)		
		5 (strong)	20 (moderate)	1000 (very weak)
σ	0.75 (weak noise)	0.172/0.181 (0.003)	0.149/0.160 (<0.001)	0.080/0.088 (<0.001)
	1 (moderate noise)	0.465/0.528 (<0.001)	0.410/0.436 (<0.001)	0.211/0.220 (<0.001)
	1.25 (strong noise)	1.078/1.151 (0.002)	0.792/0.865 (<0.001)	0.421/0.462 (<0.001)

(c) Median errors of estimating g_3 using KRCL and KRL.

KRCL/KRL (p-value)		λ_2 (correlation)		
		5 (strong)	20 (moderate)	1000 (very weak)
σ	0.75 (weak noise)	0.266/0.266 (0.592)	0.213/0.213 (0.508)	0.194/0.183 (0.999)
	1 (moderate noise)	0.680/0.672 (0.657)	0.651/0.635 (0.981)	0.508/0.504 (0.581)
	1.25 (strong noise)	1.375/1.313 (0.323)	1.503/1.475 (0.343)	1.088/1.077 (0.071)

3.2 Clustering

Two sets of the true DWS and the corresponding DWSd are designed for clustering (Figures 7 and 8). For this simulation study, we choose 10 to be the threshold for outlier groups and $n_{\text{ref}} = 20$. We consider 3 noise levels $\sigma = 0.75, 1, 1.25$ and 3 correlation levels $\lambda_2 = 5, 20, 1000$. For each setting, $M = 50$ simulations are conducted. In each iteration, the number of observations simulated from each true DWSd is a random integer between 5 and 50. The

performance is represented by the similarity index between the result and the true clusters, which is given by (Euan & Sun, 2019):

$$\text{sim}(\mathcal{C}, \mathcal{G}) = \frac{1}{B} \sum_{i=1}^{B} \max_{1 \le j \le k} \text{sim}(\mathcal{C}_j, \mathcal{G}_i) \tag{11}$$

where $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_B\}$ are the true clusters (B clusters in total) and $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ is a k-cluster solution and

$$\text{sim}(\mathcal{C}_j, \mathcal{G}_i) = \frac{2|\mathcal{C}_j \cap \mathcal{G}_i|}{|\mathcal{C}_j| + |\mathcal{G}_i|} \tag{12}$$

where $|\mathcal{C}|$ is the number of observations in cluster $\mathcal{C}$.

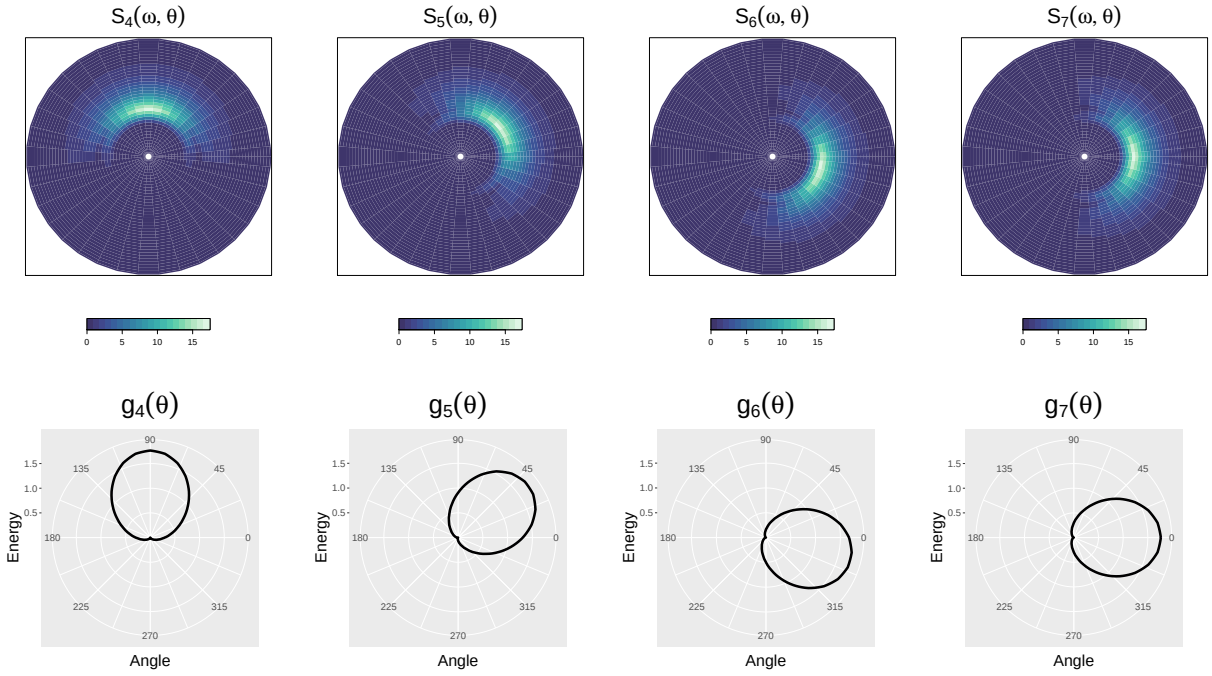

Figure 7. First set of target DWS and the corresponding DWSd designed for clustering. The first row shows the DWS in polar coordinates, where the radius represents the frequency ω, the angle represents the direction θ, and the color scale represents the value of $S(\omega, \theta)$. The second row shows the corresponding DWSd. The "energy" scale is equivalent to $g(\theta)$. The numbers correspond to the radius at 90 degrees.

To determine whether KRCL is more helpful than KRL in clustering, the Wilcoxon signed rank test is applied to the M pairwise differences in the similarity indeces. The alternative

hypothesis is that for the same observation, the similarity index using KRCL is higher than using KRL. Since the test is one-sided, a p-value smaller than 0.05 favors KRCL. As shown by the boxplots of the resulting similarity indeces in Figures 9 and 10, the similarity index is greater than 0.7 most of the time, suggesting effective clustering even when the noise and the correlation are strong. However, the p-value of the Wilcoxon test is not significant for the majority scenarios. This shows that the difference between using KRCL and KRL is small.

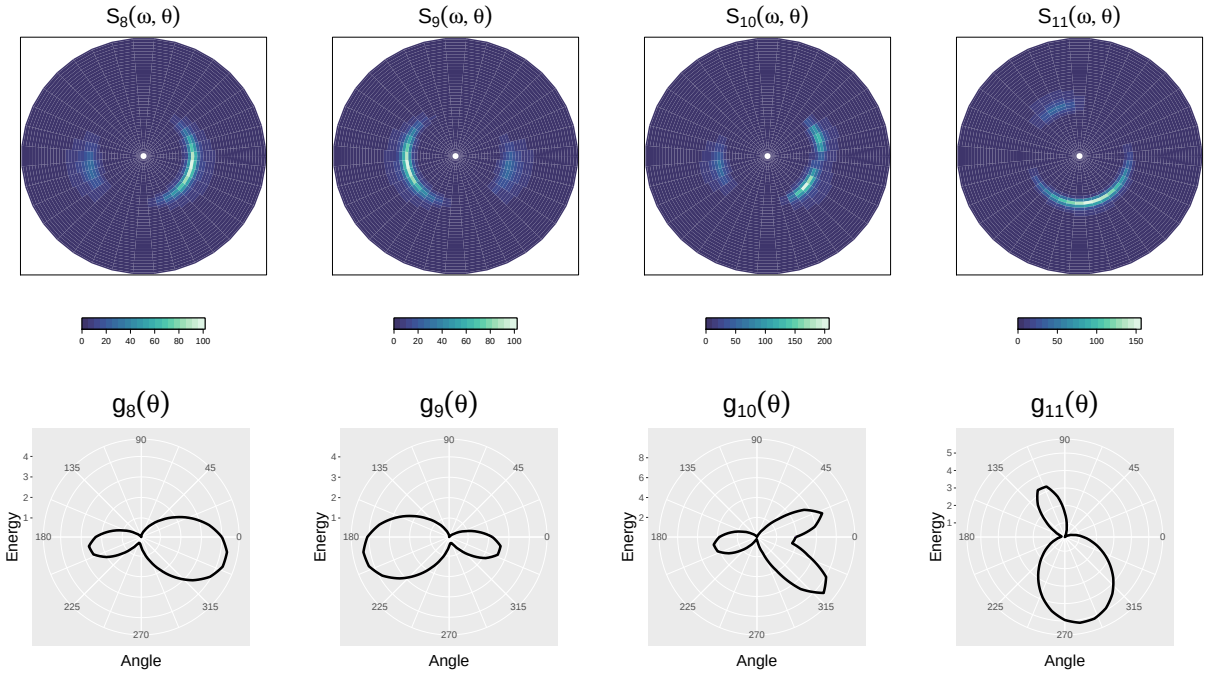

Figure 8. Second set of target DWS and the corresponding DWSd designed for clustering. The first row shows the DWS in polar coordinates, where the radius represents the frequency ω, the angle represents the direction θ, and the color scale represents the value of $S(\omega, \theta)$. The second row shows the corresponding DWSd. The "energy" scale is equivalent to $g(\theta)$. The numbers correspond to the radius at 90 degrees.

For each set, we also present a sample clustering result (Figures 11 and 12). For each DWSd, 50 observations are simulated with $\sigma = 1$ and $\lambda_2 = 20$. The directional functional boxplots depict the estimated truth as well as the uncertainty in the estimation.

In the sample result for the first set (Figure 11), both algorithms with KRCL and KRL

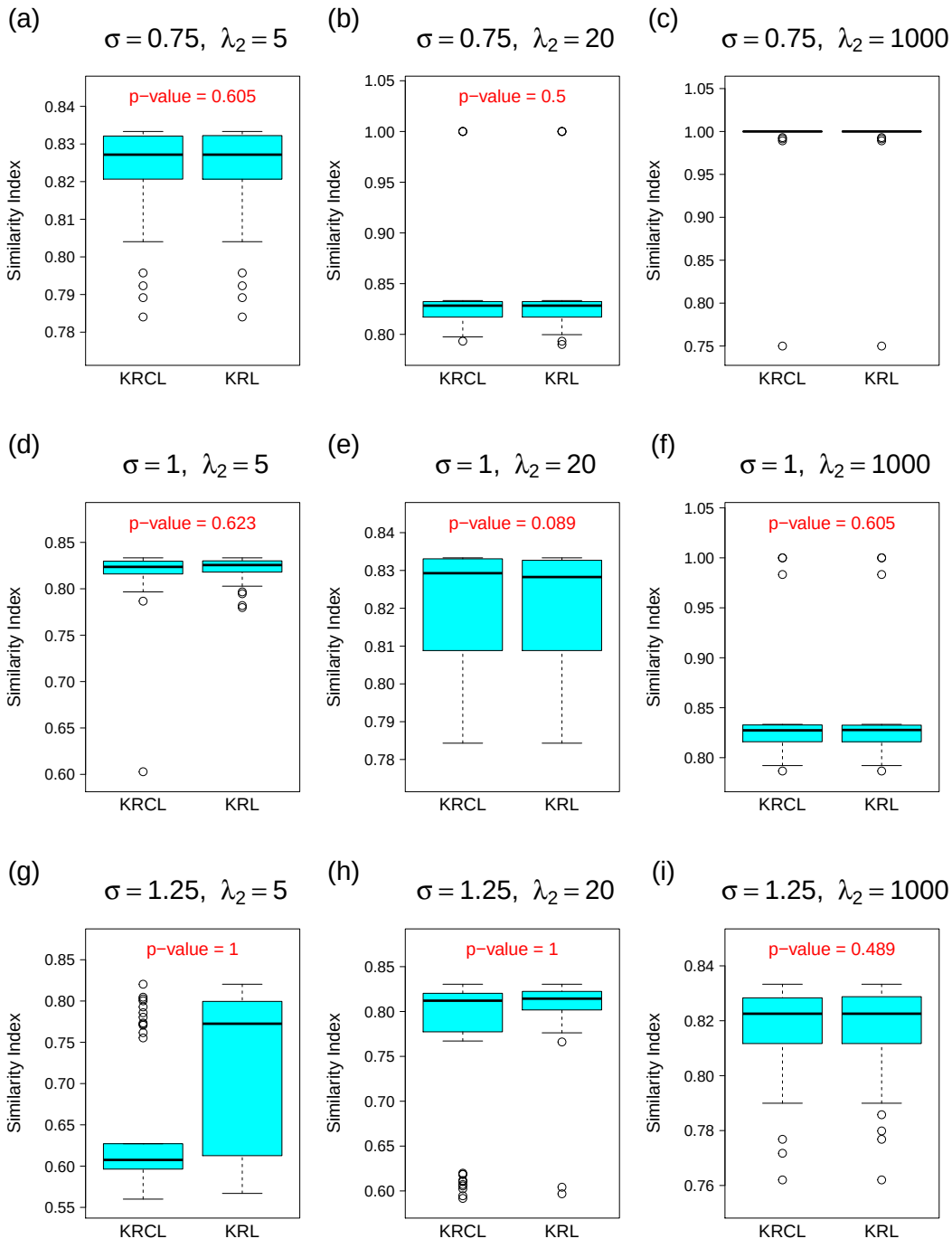

Figure 9. Boxplots of similarity indeces between the true clusters of the first set and the output of the HDSd algorithm when using KRCL / KRL for smoothing. The parameters σ and λ_2 indicate the noise level and the within-observation correlation level of the simulated DWSd observations. For each setting, the Wilcoxon signed rank test is applied to test the hypothesis that the similarity index when using KRL for smoothing is higher than that when using KRCL for smoothing, and the p-value is given in each plot. The p-value for (c) is removed because the results from KRCL and KRL are exactly the same, in which case the Wilcoxon test is not applicable.

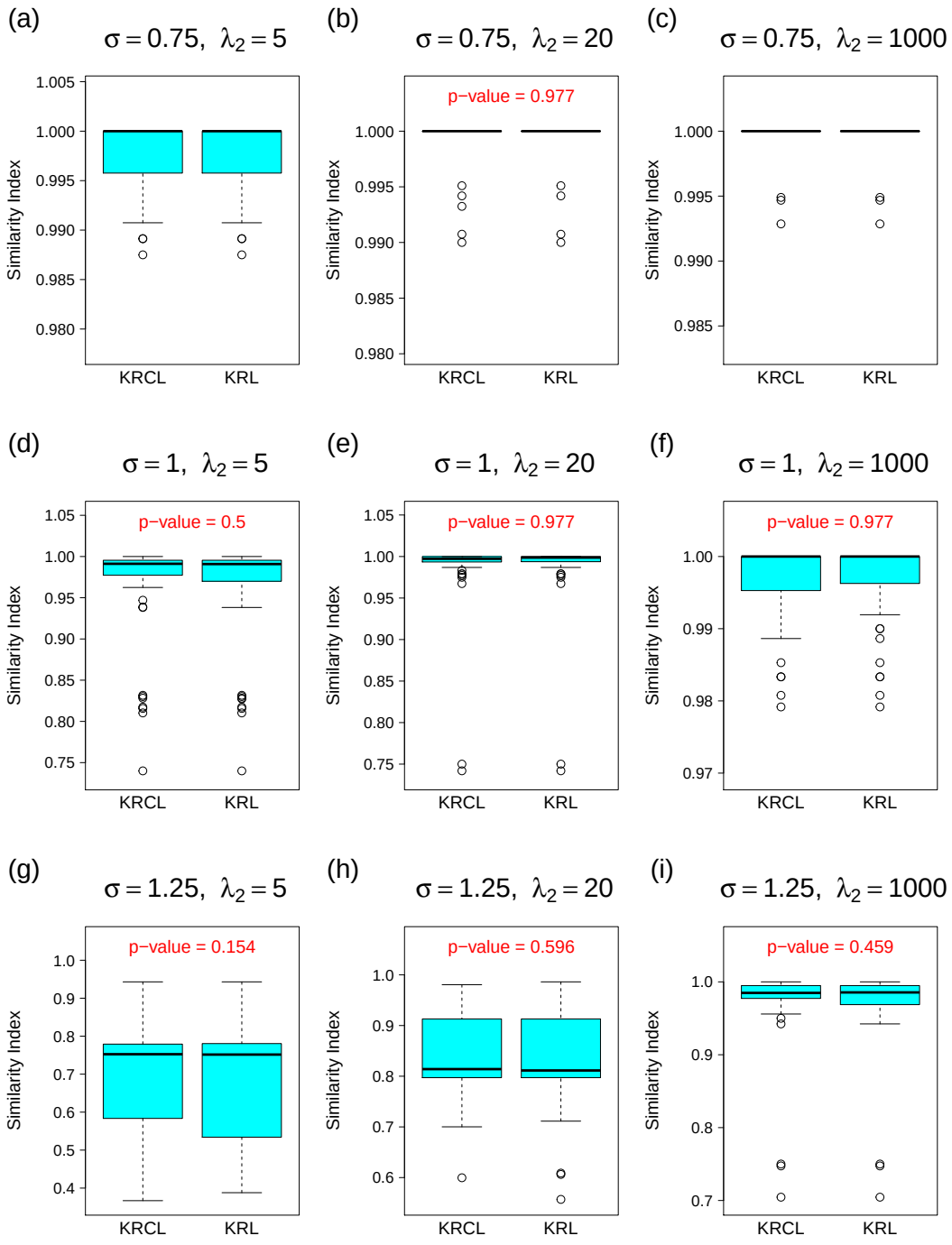

Figure 10. Boxplots of similarity indeces between the true clusters of the second set and the output of the HDSd algorithm when using KRCL / KRL for smoothing. The parameters σ and λ_2 indicate the noise level and the within-observation correlation level of the simulated DWSd observations. For each setting, the Wilcoxon signed rank test is applied to test the hypothesis that the similarity index when using KRL for smoothing is higher than that when using KRCL for smoothing, and the p-value is given in each plot. The p-values for (a) and (c) are removed because the results from KRCL and KRL are exactly the same, in which case the Wilcoxon test is not applicable.

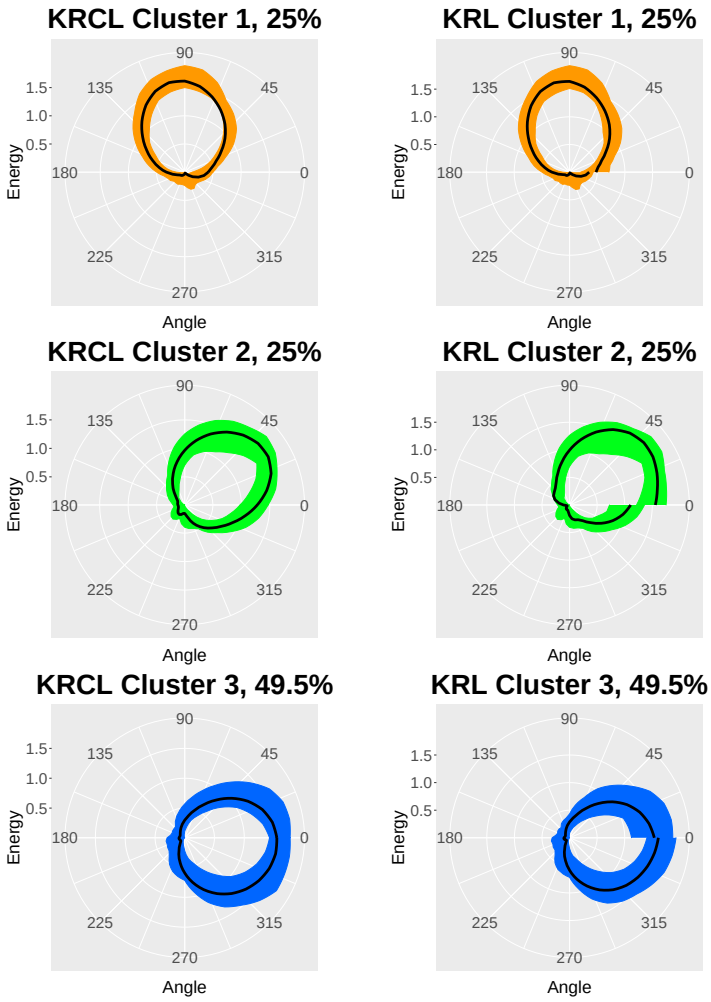

Figure 11. Directional functional boxplots of the clusters obtained by the HDSd algorithm with KRCL / KRL smoothing on a set of 200 observations simulated based on the first set of true DWSd (Figure 7). 50 observations are simulated from each profile. The percentage in the title of each diagram indicates the percentage of observations that belong to this cluster. Similar clusters obtained with KRCL and KRL smoothing are given the same color. The percentages do not sum up to 100% due to removal of outliers. The "energy" scale is equivalent to $g(\theta)$. The numbers correspond to the radius at 90 degrees.

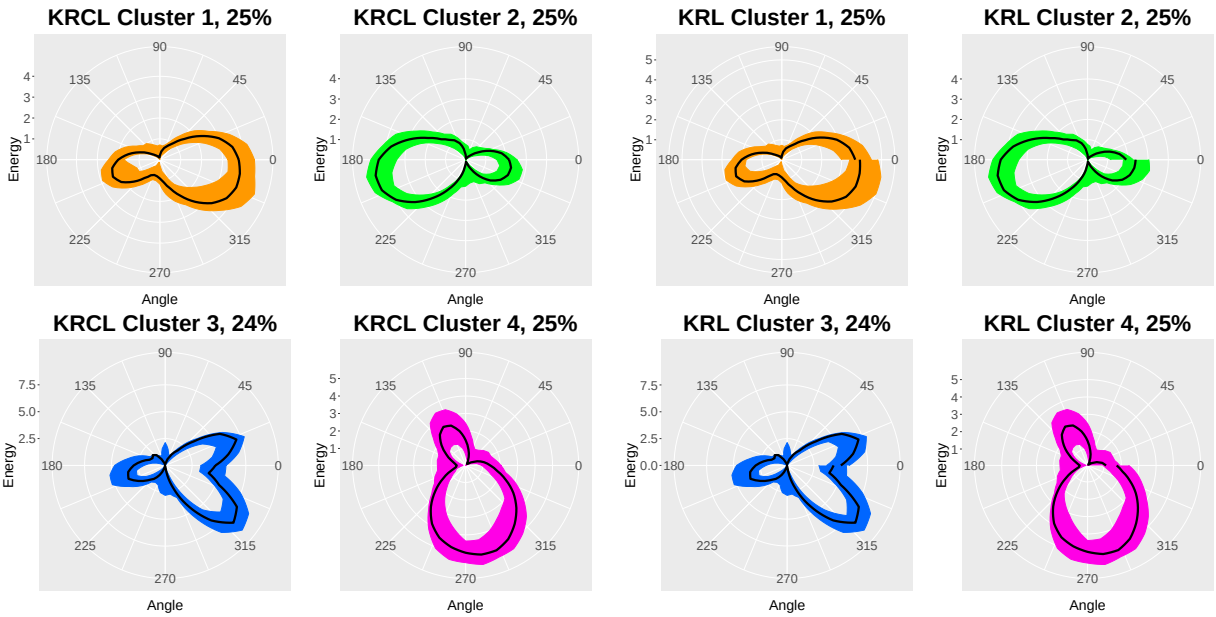

Figure 12. Directional functional boxplots of the clusters obtained by the HDSd algorithm with KRCL / KRL smoothing on a set of 200 observations simulated based on the second set of true DWSd (Figure 8). 50 observations are simulated from each profile. The percentage in the title of each diagram indicates the percentage of observations that belong to this cluster. Similar clusters obtained with KRCL and KRL smoothing are given the same color. The percentages do not sum up to 100% due to removal of outliers. The "energy" scale is equivalent to $g(\theta)$. The numbers correspond to the radius at 90 degrees.

fail to separate g_6 and g_7. This is expected as the directions of the peaks only differ by 15 degrees, which is small compared to the width of the peaks. Both KRCL and KRL removes one observation from cluster 3 (corresponding to g_6 and g_7) as an outlier, hence the percentages do not sum up to 100%. In contrast, the sample result for the second set identifies all clusters correctly, as the peaks have very different directions. Again, two observations are treated as outliers by both KRCL and KRL and removed from cluster 3 (corresponding to g_{10}).

4 Application

In this section, we apply the proposed clustering method to two DWS datasets, one from the Sofar Ocean network[1], and the other collected from a buoy located at the Red Sea (Farrar et al., 2009). The corresponding DWSd are calculated and used for this investigation.

4.1 Sofar Ocean Data

The data from Sofar Ocean consists of the predicted DWS of each hour from 7 Dec to 14 Dec 2021 at 34.5°N, 20°E. Since this is a small set with just 168 observations, we choose a smaller number of 5 to be the threshold for outlier groups. We expect to have 3 to 4 clusters within one week of observations, so we choose $n_{\text{ref}} = 10$.

The directional functional boxplots of the identified clusters are displayed in Figure 13. It can be observed that the first and the last two clusters are consistently identified by both methods. However, while KRL separates clusters 2 and 3, KRCL merges them into a single cluster (cluster 2). The scree plots show that the elbows are chosen appropriately: increasing the number of clusters does not lead to a significant change in the smallest dissimilarity. The selected bandwidths/concentrations are given in Table 2.

[1] https://docs.sofarocean.com/wave-spectra

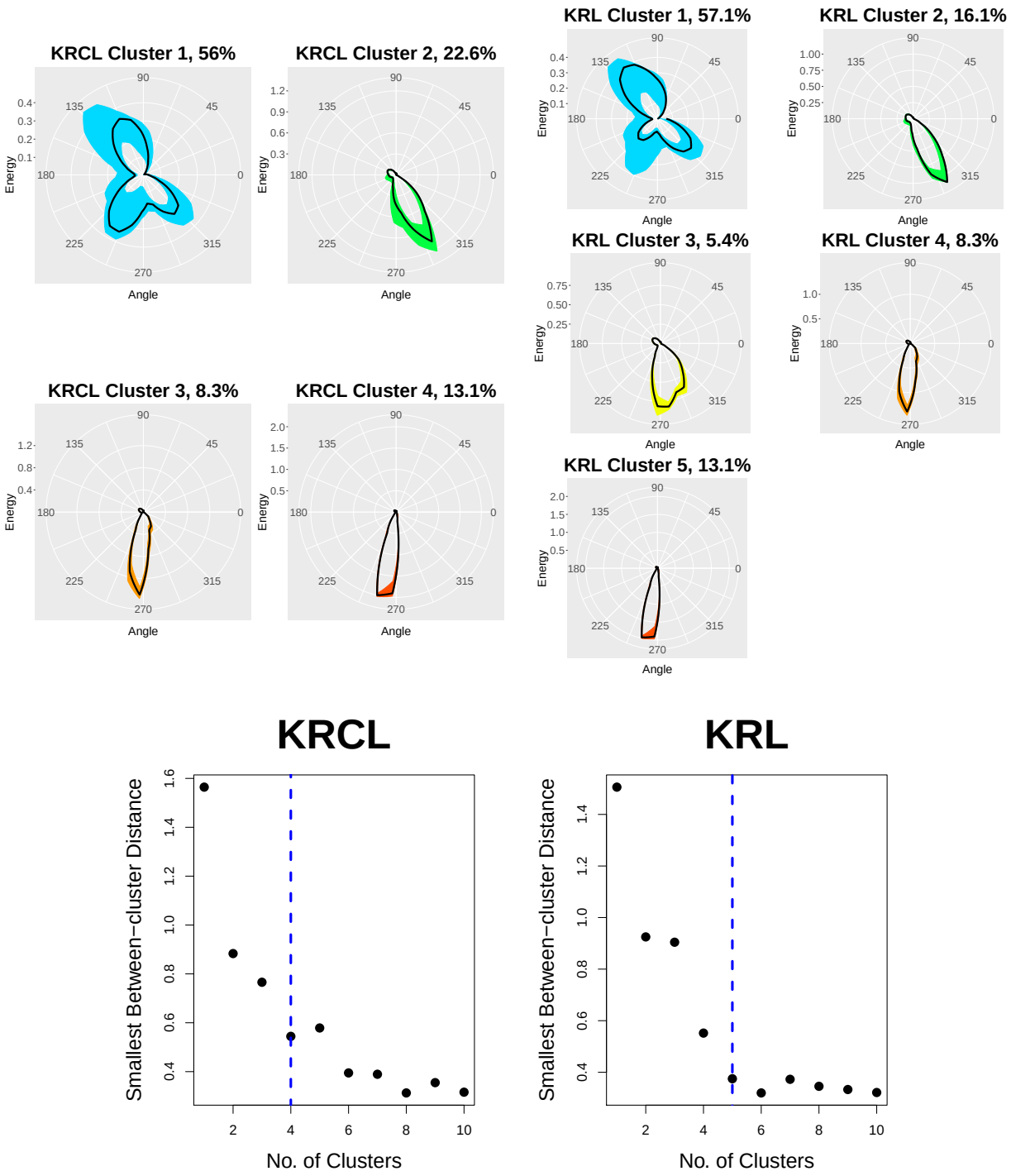

Figure 13. Directional functional boxplots (upper) and scree plots (lower) obtained by the HDSd algorithm with KRCL / KRL smoothing on the Sofar Ocean data. The percentage in the title of each boxplot indicates the percentage of observations that belong to this cluster. Similar clusters obtained with KRCL and KRL smoothing are given the same color. The "energy" scale is equivalent to $g(\theta)$. The numbers correspond to the radius at 90 degrees. The dashed blue lines in the scree plots indicate the elbows determined.

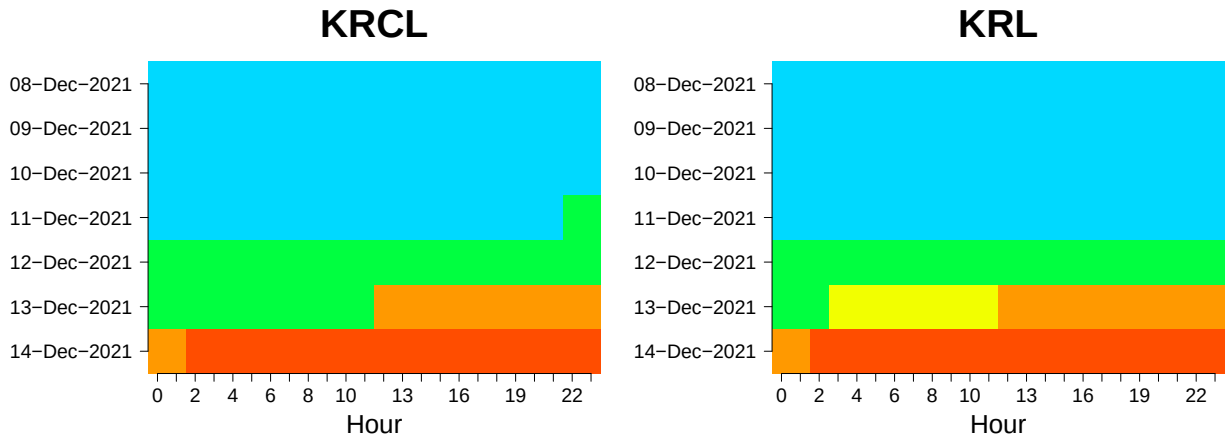

Figure 14. Color-coded calendar plot of the clusters in the Sofar Ocean data obtained by the HDSd algorithm with KRCL / KRL smoothing. Each color represents a cluster. Similar clusters obtained with KRCL and KRL smoothing are given the same color.

Figure 14 plots the identified cluster of each hour, with each color representing one cluster. For most hours, the two methods appear to be consistent. This can be attributed to the fact that no peaks in this dataset appear at the boundary, so the advantages of KRCL have little effect.

Table 2. Selected concentration parameter for KRCL and bandwidth for KRL for clustering Sofar data. Note that the groups are initial groups, not final clusters (see Section 2.2.2).

Group	ν for KRCL	Bandwidth for KRL
1	16.5	0.610
2	444	0.276
3	261	0.247
4	545	0.156

4.2 Red Sea Data

This data consists of the observed DWS on the Red Sea near Thuwal (22.2°N, 38.5°E) from January to March 2010. The buoy was installed in 2009 and was the first structure that measures the wind and wave on the Red Sea (Farrar et al., 2009). The DWS, $S(\omega, \theta)$, where $\omega \in (0, 0.5]$ and $\theta \in [0, 2\pi)$, are derived from buoy data recorded at a frequency of 2 Hz

during the first 17 minutes of each hour. The Red Sea data contains DWSd from more directions, including the boundary (in the direction of the east), which will be a challenge for both methods. Since this dataset is larger, we choose 10 to be the threshold for outlier groups. We choose $n_{\text{ref}} = 30$ as we expect to have 5 to 8 clusters.

Table 3. Selected concentration parameter for KRCL and bandwidth for KRL for clustering Red Sea buoy data. Note that the groups are initial groups, not final clusters (see Section 2.2.2).

Group	ν for KRCL	Bandwidth for KRL
1	691	1.14
2	0.733	10.0
3	719	1.21
4	13.4	0.683
5	12.6	0.721
6	12.8	0.729

The directional functional boxplots of the identified clusters are displayed in Figure 15. The calendar plots are shown in Figure 16. The selected bandwidths/concentrations are given in Table 3. Compared to the Sofar Ocean dataset, more differences appear between the results of KRCL and KRL.

First, the DWSd identified by KRL are broken at the boundary, as shown in Figure 15, KRL clusters 1 to 3. This is expected, as KRL does not model direction as a circular variable, which means there is little relationship between the energy at the boundaries $\theta = 0$ and $\theta = 2\pi$. On the other hand, KRCL treats $\theta = 0$ and $\theta = 2\pi$ as connected and leads to continuous estimates at the boundaries.

Next, we compare each cluster identified by KRCL and KRL. Notably, cluster 3 of KRCL is not identified by KRL. Further inspection (clues can also be found in Figure 16, the calendar plot) suggests that in the results of KRL, this cluster is absorbed into cluster 2, which seems to be a mix of DWSd in different directions.

Moreover, clusters 4 and 5 of KRL match exactly to cluster 5 of KRCL. This observation

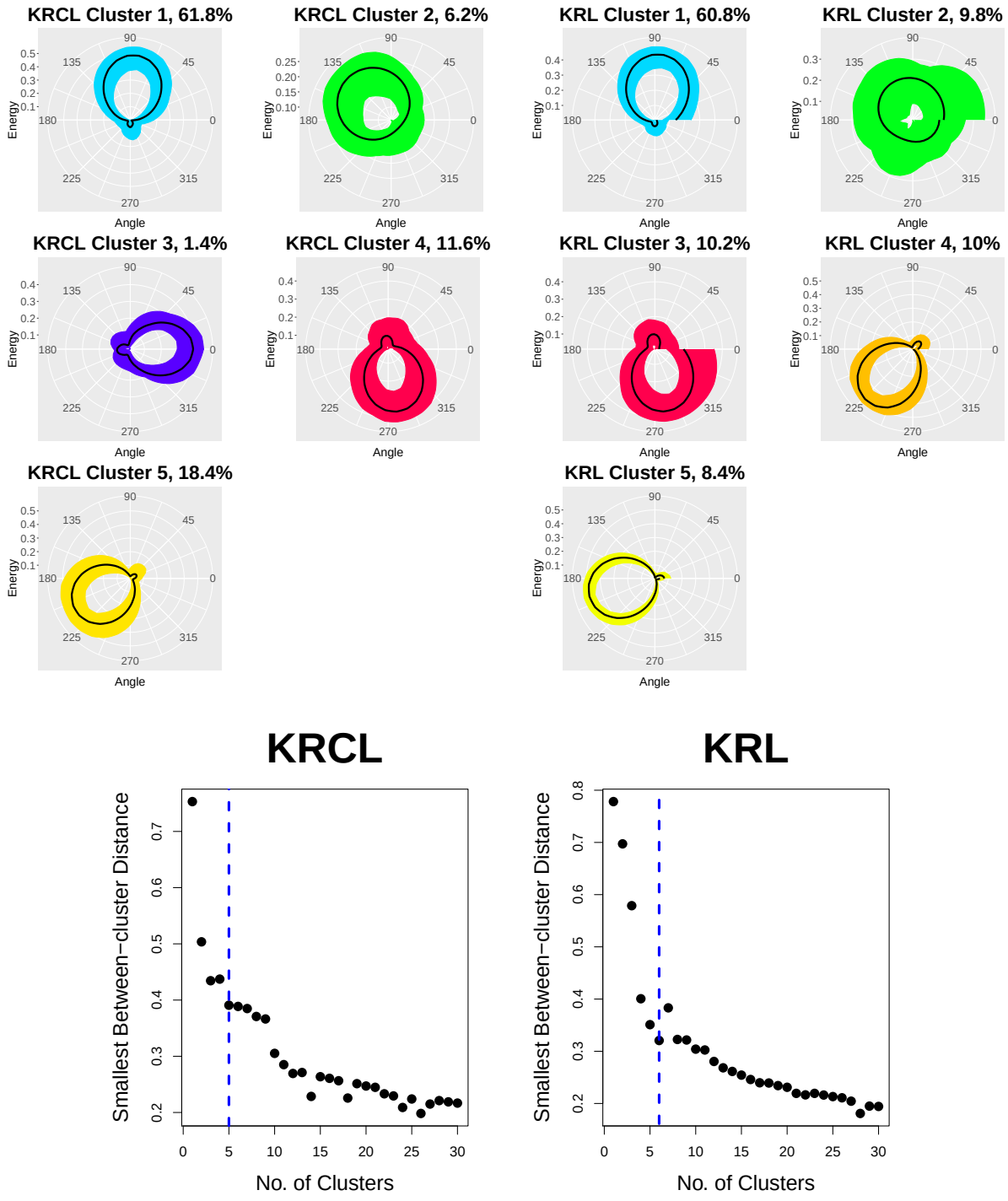

Figure 15. Directional functional boxplots (upper) and scree plots (lower) obtained by the HDSd algorithm with KRCL / KRL smoothing on the Red Sea buoy data. For KRL, one of the clusters determined is treated as an outlier due to a size smaller than 10. The percentage in the title of each boxplot indicates the percentage of observations that belong to this cluster. Similar clusters obtained using KRCL and KRL smoothing are given the same color. The "energy" scale is equivalent to $g(\theta)$. The numbers correspond to the radius at 90 degrees. The dashed blue lines in the scree plots indicate the elbows determined.

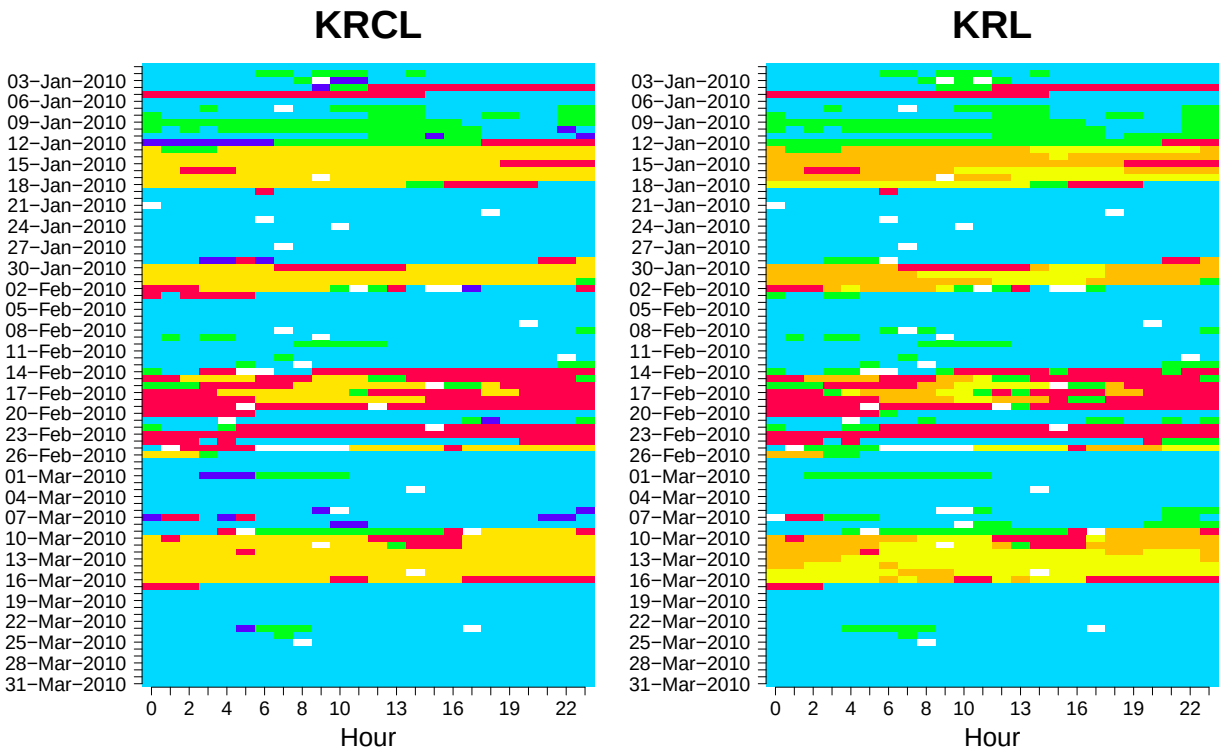

Figure 16. Color-coded calendar plot of the clusters in the Red Sea buoy data obtained by the HDSd algorithm with KRCL / KRL smoothing. Each color represents a cluster. Similar clusters obtained with KRCL and KRL smoothing are given the same color. White cells represent missing data.

is similar to the case for the Sofar Ocean data, where cluster 2 and 3 of KRL is merged into cluster 2 of KRCL. A possible explanation for the separations in KRL's results is that the clusters' profiles at the boundary are modelled inaccurately, which leads to exaggerated within-group differences.

Finally, the scree plots show that the smallest dissimilarity continues to decrease beyond the selected number of clusters. Based on our choice of $n_{\text{ref}} = 30$, our procedure (as described in Section 2.2.1) decides that the decrease does not have a relevant meaning, because the clusters are not so distinct. A larger n_{ref} should be chosen if those not-so-distinct clusters are expected. In the supplementary code, different choices of n_{ref} are explored for both the Sofar and Red Sea data, and the scree plots are obtained.

5 Conclusion

In this paper, we applied the kernel smoother for circular-linear variables (circular regression) proposed by Di Marzio et al. (2009) for the estimation of direction-only directional wave spectrum (DWSd) data. The circular regression is compared against the ordinary linear kernel smoother in the estimation of circular-linear data. The circular regression takes into account the circular nature of the support (boundaries are connected), while the ordinary kernel smoother does not. A procedure for smoothing parameter selection is developed for DWSd data with correlated errors in each observation. The results from the simulation study suggest that the estimates from circular regression have significantly lower errors than ordinary kernel regression.

Our study is then extended to the clustering of DWSd data. We expected that the accuracy of clustering would be improved with a better estimation technique, as it helps to remove the noise in the observations. To utilize the estimation techniques for noise removal

during the clustering process, a clustering workflow is developed based on the Hierarchical Directional Spectra-based (HDSd) algorithm proposed by Euan & Sun (2019). For fair comparisons, the determination of the optimal number of clusters based on the scree plot is automated.

Finally, we applied our proposed clustering procedure to two sets of real data – the Sofar Ocean data (168 predicted DWSd) and the Red Sea buoy data (2137 DWSd observations collected at the Red Sea). Differences are observed in the clustering results of both datasets. For the Sofar Ocean data, clustering with circular regression results in fewer clusters (4 clusters) compared to the ordinary kernel smoother (5 clusters), which is due to the combination of two similar clusters determined by the kernel smoother. For the Red Sea buoy data, the ordinary kernel smoother misses a cluster at the boundary (that is, close to 0 and 2π) and splits a cluster (southwest) into two compared to the circular regression. In addition, the profiles of the clusters smoothed by the ordinary kernel smoother are discontinuous at the boundary. This suggests that the circular regression may be a more appropriate technique for such circular-linear data.

As a first attempt, this study only applies the circular regression to 1-dimensional DWSd data. An extension to the 2-dimensional case is desirable where the circular regression methodology needs developing. Considering both ω and θ would require to perform nonparametric regression on a cylinder. In this case, the nonparametric regression method would involve cylindrical kernels, but a simple way to approach this would be to consider product kernels, as suggested by García-Portugués et al. (2013). Another possible extension is to consider a different model of the noise, such as the multiplicative noise model. Further research is required to evaluate the effectiveness of the method on other types of real data. For the selection of smoothing parameter and determination of the optimal number of clusters, the procedures proposed in this paper are among many other possibilities, which also worth

investigating.

Acknowledgement

We would like to thank the two reviewers and the associate editor for providing valuable comments, questions and suggestions to make the revised version more concise and clearer. This research was supported by King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR) under Award No: OSR-2019-CRG7-3800. Work by Rosa M. Crujeiras was supported by Project MTM2016-76969-P from the AEI cofunded by the European Regional Development Fund (ERDF), the Competitive Reference Groups 2017-2020 (ED431C 2017/38) from the Xunta de Galicia through the ERDF.

References

Ailliot, P, Maisondieu, C., Monbet, V. (2013). Dynamical partitioning of directional ocean wave spectra. *Probabilistic Engineering Mechanics* (33), 95-102.

Benoit, M. (1993). Practical comparative performance survey of methods used for estimating directional wave spectra from heave-pitch-roll data. *In Coastal Engineering 1992* (pp. 62-75).

Boukhanovsky, A. V., & Guedes Soares, C. (2009). Modelling of multipeaked directional wave spectra. *Applied Ocean Research, 31*(2), 132-141.

De Brabanter, K., De Brabanter, J., Suykens, J. A., & De Moor, B. 2011. Kernel Regression in the Presence of Correlated Errors. *Journal of Machine Learning Research, 12*(6).

Di Marzio, M., Panzera, A., & Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statistics & Probability Letters, 79*(19), 2066-2075.

Euan, C. & Sun, Y. 2019. Directional Spectra Based Clustering for Visualizing Patterns of Ocean Waves and Winds. *Journal of Computational and Graphical Statistics, 28*(3), 659-670.

Farrar, J. T., Lentz, S., Churchill, J., Bouchard, P., Smith, J., Kemp, J., ... & Hosom, D. (2009). King Abdullah University of Science and Technology (KAUST) mooring deployment cruise and fieldwork report. *Technical report, Woods Hole Oceanographic Institution, WHOI-KAUST-CTR-2009*, 2.

García-Portugués, E., Crujeiras, R.M. & González-Manteiga, W. (2013) Kernel density estimation for directional–linear data, *Journal of Multivariate Analysis, 121*, 152-175.

Gorman, R. M. (2018). Estimation of directional spectra from wave buoys for model validation. *Procedia Iutam, 26*, 81-91.

Hamilton, L. J. (2010). Characterising spectral sea wave conditions with statistical clustering of actual spectra. *Applied Ocean Research, 32*(3), 332-342.

Hanson, J. L., & Phillips, O. M. (2001). Automated analysis of ocean surface directional wave spectra. *Journal of atmospheric and oceanic technology, 18*(2), 277-293.

Hasselmann, K., Barnett, T. P., Bouws, E., Carlson, H., Cartwright, D. E., Enke, K., ... & Walden, H. (1973). Measurements of wind-wave growth and swell decay during the Joint North Sea Wave Project (JONSWAP). *Ergaenzungsheft zur Deutschen Hydrographischen Zeitschrift, Reihe A*.

Hinostroza, M. A., & Guedes Soares, C. (2016). Nonparametric estimation of directional wave spectra using two hyperparameters. *Maritime Technology and Engineering, 3*, 287-293.

Hisaki, Y. (1996). Nonlinear inversion of the integral equation to estimate ocean wave spectra from HF radar. *Radio science, 31*(1), 25-39.

Longuet-Higgins, M. S., Cartwright, D. E., and Smith, N. D. (1963). Observations of the Directional Spectrum of Sea Waves Using the Motions of a Floating Buoy. *Ocean Wave Spectra*, Proceedings of a Conference, Easton, Maryland, Prentice-Hall: National Academy of Sciences, 111–136.

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American statistical Association, 104*(486), 718-734.

Mortlock, T. R., & Goodwin, I. D. (2015). Directional wave climate and power variability along the Southeast Australian shelf. *Continental Shelf Research, 98*, 36-53.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications, 9*(1), 141-142.

Nielsen, U. D. (2006). Estimations of on-site directional wave spectra from measured ship responses. *Marine Structures, 19*(1), 33-69.

Nielsen, U. D. (2008). Introducing two hyperparameters in Bayesian estimation of wave spectra. *Probabilistic Engineering Mechanics, 23*(1), 84-94.

Ochi, M. (1998). Ocean Waves: The Stochastic Approach (Cambridge Ocean Technology Series). *Cambridge: Cambridge University Press*.

Oliveira, M., Crujeiras, R. M., & Rodríguez Casal, A. (2014). NPCirc: An R package for nonparametric circular methods. *Journal of Statistical Software*.

Pascoal, R., & Guedes Soares, C. (2008). Non-parametric wave spectral estimation using vessel motions. *Applied Ocean Research, 30*(1), 46-53.

Portilla-Yandún, J., Cavaleri, L., & Van Vledder, G. P. (2015). Wave spectra partitioning and long term statistical distribution. *Ocean Modelling, 96*, 148-160.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

Ribeiro, P. J. C., Henriques, J. C. C., Campuzano, F. J., Gato, L. M. C., & Falcão, A. F. O. (2020). A new directional wave spectra characterization for offshore renewable energy applications. *Energy, 213*, 118828.

Sun, Y., & Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics, 20*(2), 316-334.

Vogel, M., Hanson, J., Fan, S., Forristall, G. Z., Li, Y., Fratantonio, R., & Jonathan, P. (2016, May). Efficient environmental and structural response analysis by clustering of directional wave spectra. In *Offshore Technology Conference*. OnePetro.

Waals, O. J., Aalbers, A. B., & Pinkster, J. A. (2002, January). Maximum likelihood method as a means to estimate the directional wave spectrum and the mean wave drift force on a dynamically positioned vessel. *In International Conference on Offshore Mechanics and Arctic Engineering* (Vol. 36142, pp. 605-613).

Watson, G. S. (1964). Smooth regression analysis. Sankhyā: *The Indian Journal of Statistics, Series A*, 359-372.

Young, I. R. (1994). On the measurement of directional wave spectra. *Applied Ocean Research, 16*(5), 283-294.

Yurovskaya, M. V., Dulov, V. A., Chapron, B., & Kudryavtsev, V. N. (2013). Directional short wind wave spectra derived from the sea surface photography. *Journal of Geophysical Research: Oceans, 118*(9), 4380-4394.