University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Delaigle, A;Hall, P;Pham, T

Title:
Clustering functional data into groups by using projections

Date:
2019-04-01

Citation:
Delaigle, A., Hall, P. & Pham, T. (2019). Clustering functional data into groups by using projections. Journal of the Royal Statistical Society Series B Statistical Methodology, 81 (2), pp.271-304. https://doi.org/10.1111/rssb.12310.

Persistent Link:
https://hdl.handle.net/11343/285344

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/rssb.12310

This article is protected by copyright. All rights reserved

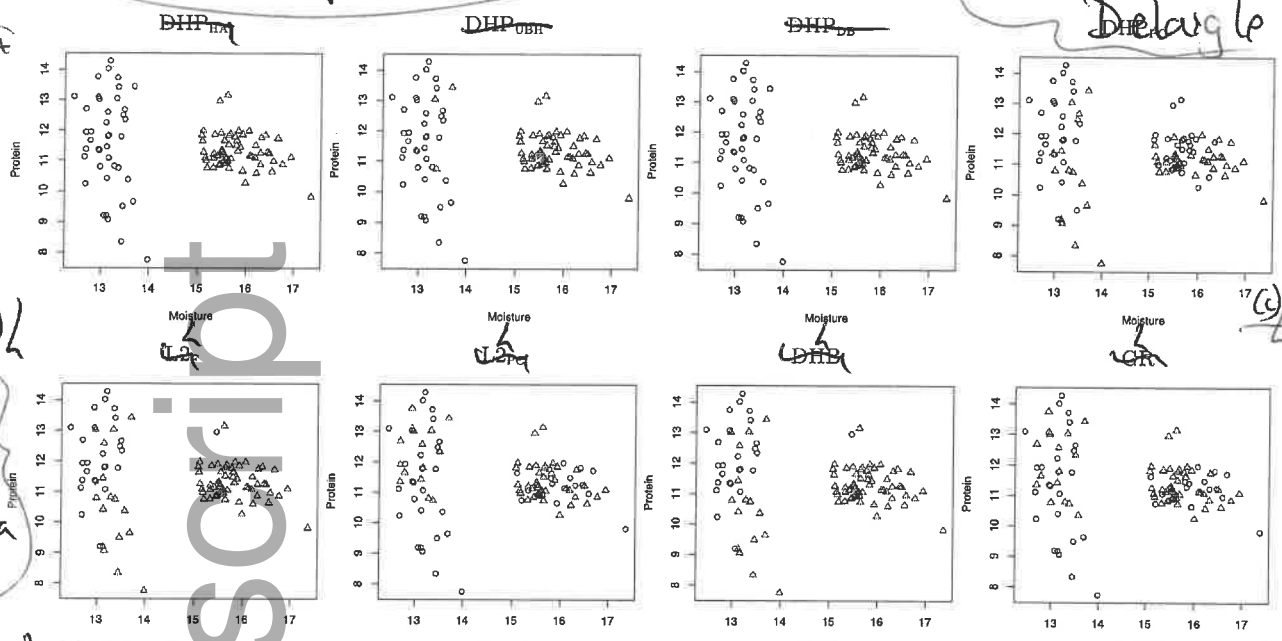


Figure 1: Scatterplots of moisture and protein levels for the wheat data (the blue circles correspond to one cluster and the red triangles to the other) for clusters created using, from left to right and top to bottom, $\mathrm{DHP}_{\mathrm{HA}}$, $\mathrm{DHP}_{\mathrm{UBH}}$, $\mathrm{DHP}_{\mathrm{DB}}$, $\mathrm{DHP}_{\mathrm{PC}}$, $\mathrm{L2}_{\mathrm{F}}$, $\mathrm{L2}_{\mathrm{PC}}$, DHB, GR.

higher than 15% were clustered in another group (no curve had a moisture level in the interval $[14, 15]$). These results are particularly interesting when considering properties of wheat. Specifically, wheat with low moisture level can be stored more safely than wheat with high moisture level, so that moisture level is often used as a quality factor for wheat. In Canada, wheat with moisture level lower than 14.5% can be sold as straight grade seeds; see Jayas et al. (1994), page 347, and `http://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/crop1204`.

The clusters created by $\mathrm{L2}_{\mathrm{F}}$ and DHB correspond roughly to high and low protein levels, whereas the other methods created clusters that do not seem to correspond to specific groups based on neither the protein nor the moisture levels.

For our second example we used the Australian rainfall data described in Delaigle and Hall (2010), and available at `http://rda.ucar.edu/datasets/ds482.1`. The data, depicted in Figure 9 in Appendix B.2, concern 191 rainfall curves $X_i(t)$ collected at Australian weather stations, where $X_i(t)$ represents the average rainfall at time $t \in$

This article is protected by copyright. All rights reserved

landscape - enlarge so that figure and caption just fit within text area
JRSS Ser B
B18356 - Delaigle et al
(a)
(b)
(c)
(d)
Insert (a)-(d) in 8pt Helvetica final size
less #

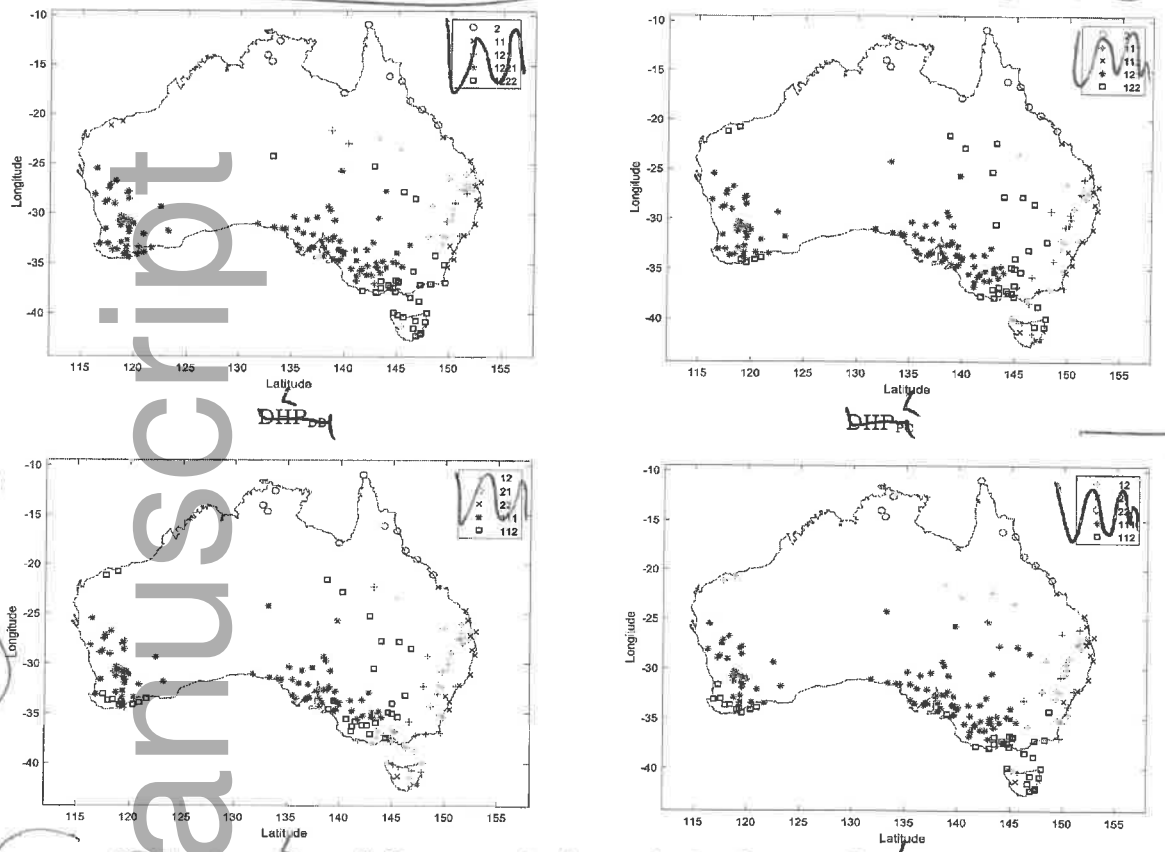


Figure 2: Clusters of rainfall stations in Australia for four methods. From left to right and top to bottom: $\mathrm{DHP}_{\mathrm{HA}}$, $\mathrm{DHP}_{\mathrm{UBH}}$, $\mathrm{DHP}_{\mathrm{DB}}$, $\mathrm{DHP}_{\mathrm{PC}}$. The numbers in the legend are the numbers $k_1, \ldots, k_\ell$ corresponding to the clusters $\chi_{k_1,\ldots,k_\ell}$ created by each method. The four symbols correspond to the four clusters created for $K = 4$.

[1, 365], taken over the years were the ith station has been operating, and smoothed using a local polynomial smoother.

In Delaigle and Hall (2010), these data were classified (manually by the authors) into northern and southern weather stations, depending on their location on the map of Australia. Here our goal is to cluster the data automatically into 2, 3, 4 or 5 groups using each of the eight methods used earlier, and see which clusters are created by each method (the extension of our method to more than two groups, using hierarchical clustering, is discussed in section 5.2).

Author Manuscript

This article is protected by copyright. All rights reserved

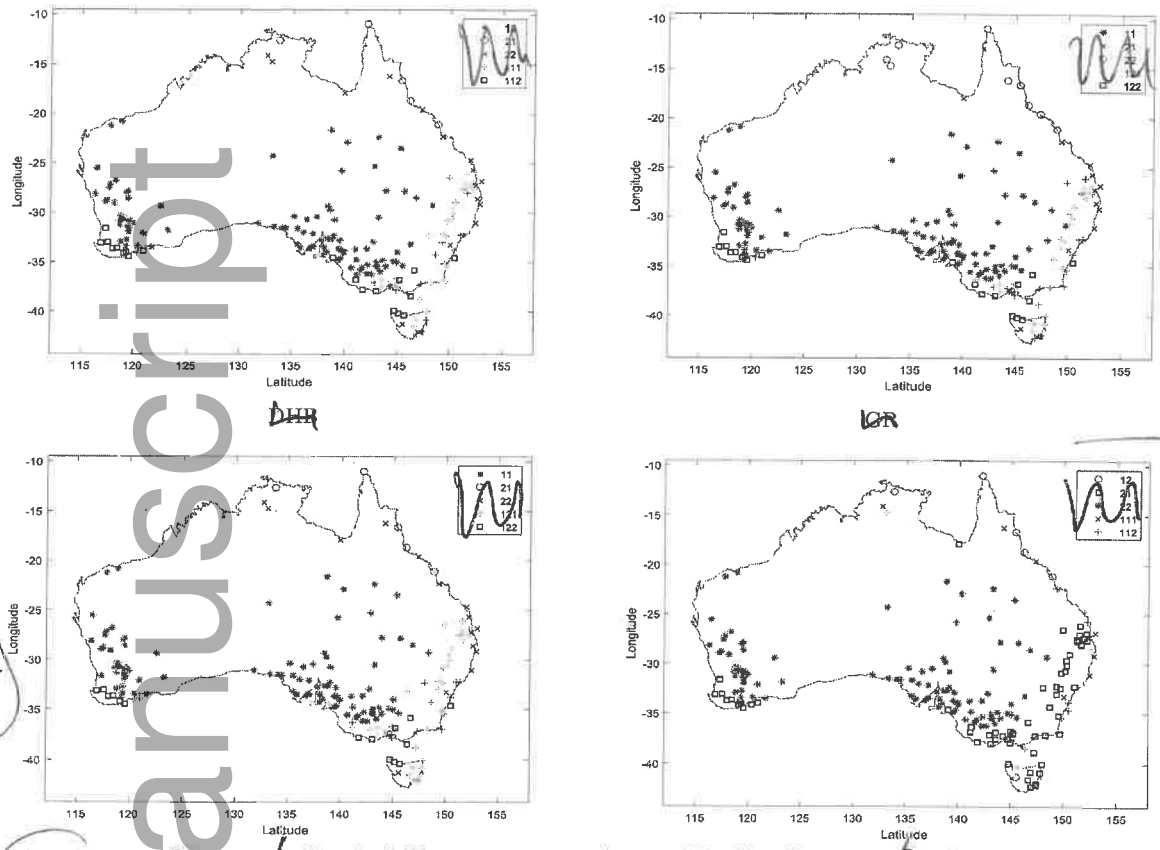


Figure 3: Clusters of rainfall stations in Australia for four methods. From left to right and top to bottom: $L2_F$, $L2_{PC}$, DHB, GR. The numbers in the legend are the numbers $k_1, \ldots, k_\ell$ corresponding to the clusters $\chi_{k_1,\ldots,k_\ell}$ created by each method. The four symbols correspond to the four clusters created for $K = 4$.

For each method, to create $K > 2$ clusters using hierarchical clustering, we proceed iteratively as follows. At the first step, we create 2 clusters. At step k, where we have created $k+1$ clusters, we apply again the clustering method to each of the $k+1$ clusters, which we divide in two clusters. Then, among the resulting $k+1$ possible cluster configurations of $k+2$ clusters, we keep the one that gives the smallest tightness where, for a given configuration of $k+2$ clusters $\mathcal{C}_1, \ldots, \mathcal{C}_{k+2}$, tightness is computed as $\sum_{\ell=1}^{k+2} \sum_{X_i \in \mathcal{C}_\ell} \left\| X_i - \bar{X}_\ell \right\|^2$. For each method, we repeat this procedure until we have reached the maximum number (here 5) of clusters we wish to consider. We denote by $\chi_{k_1,k_2,\ldots,k_\ell}$, with $k_1, \ldots, k_\ell = 1, 2$, a cluster obtained by clustering group χ_{k_1} into

This article is protected by copyright. All rights reserved

enlarge to width 24 picas

JRSS Ser B B18356 — Delaigle et al.

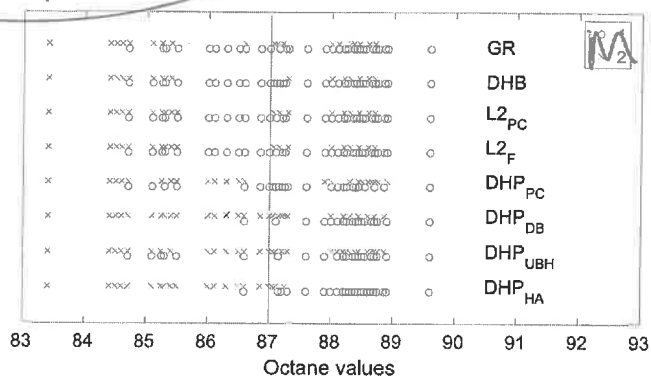


Figure 4: Two clusters of gasoline samples depicted according to their octane value, for eight clustering methods. For each method, the circles (resp., crosses) show the octane value of individuals clustered in group 1 (resp., group 2).

Finally, for our fourth illustration, we used the Berkeley growth data of Tuddenham and Snyder (1954), described in Ramsay and Silverman (2005). As indicated in the introduction, the notion of a perfect cluster is not uniquely defined, and in real examples there may be several insightful ways to cluster the data. In this example, there exists a natural way of clustering the data, which consist of height measurements of $n_1 = 39$ girls (population Π_1) and $n_2 = 54$ boys (population Π_2), taken at 31 time points t between age 1 and 18, which we turned into continuous curves $X_i(t)$, $t \in \mathcal{I} = [1, 18]$, via local polynomial smoothing. We used gender as a benchmark for the "true clusters", to measure the quality of the various clustering methods. However, this is merely for illustration purposes, as unlike for the simulated examples, we do not know what the "ideal clusters" should be nor whether these are even uniquely defined. The curves from each group, as well as the group means, are depicted in Figure 10 in Appendix B.2. The purities and adjusted rand indices obtained for each method are provided in Table 2. Considering boys and girls as the true clusters, our DHP_{HA} and DHP_{UBH} procedures worked particularly well; their performance was not nearly approached by any of the six competing procedures.

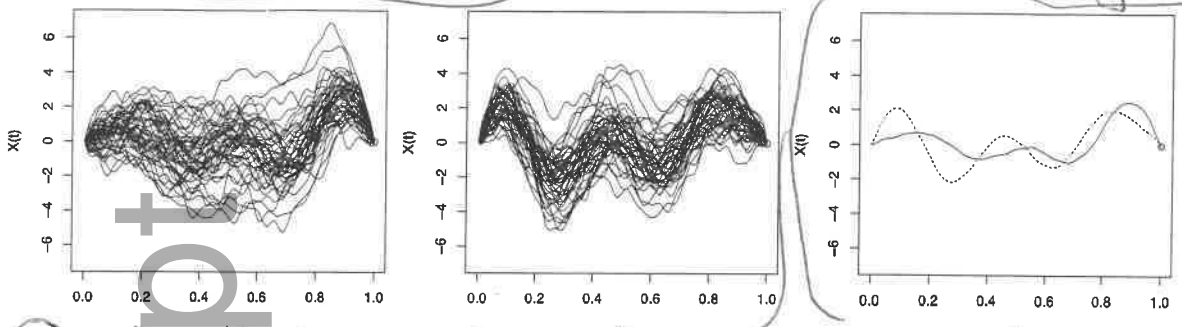
Overall, we conclude from our simulated and real data examples that our DHP_{HA} and DHP_{UBH} methods seem to provide clusters that are able to capture important

Author Manuscript

This article is protected by copyright. All rights reserved

enlarge each part to width 16 picas (both figures)

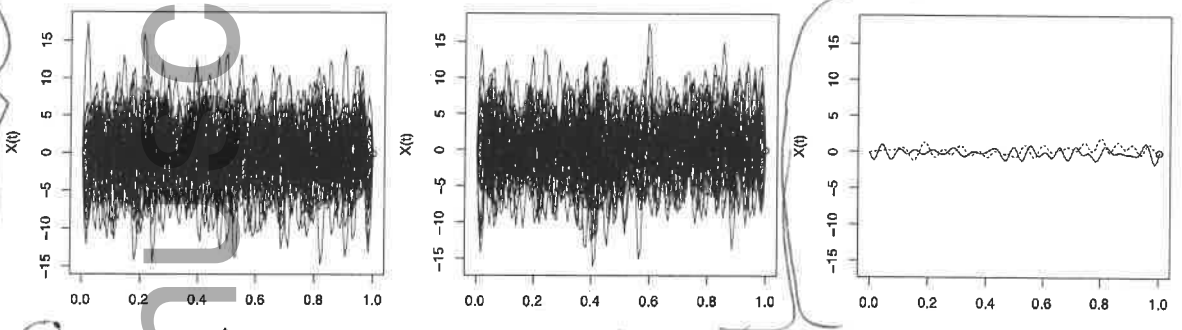
JRSS Se-B B18356 — Delaigle et al



(a) (b) (c)

Figure 5: 50 curves from model (i) coming from population Π_1 (left) or Π_2 (middle), and empirical mean curves from the two groups (right).

Insert (a)–(c) in 8pt Helvetica final size



(a) (b) (c)

Figure 6: 50 curves from model (ii) coming from population Π_1 (left) or Π_2 (middle), and empirical mean curves from the two groups (right).

Figure 10 we display the curves from Π_1, the curves from Π_2, as well as their empirical means, for the Berkeley growth data.

Author Manuscript

This article is protected by copyright. All rights reserved

Author Manuscript

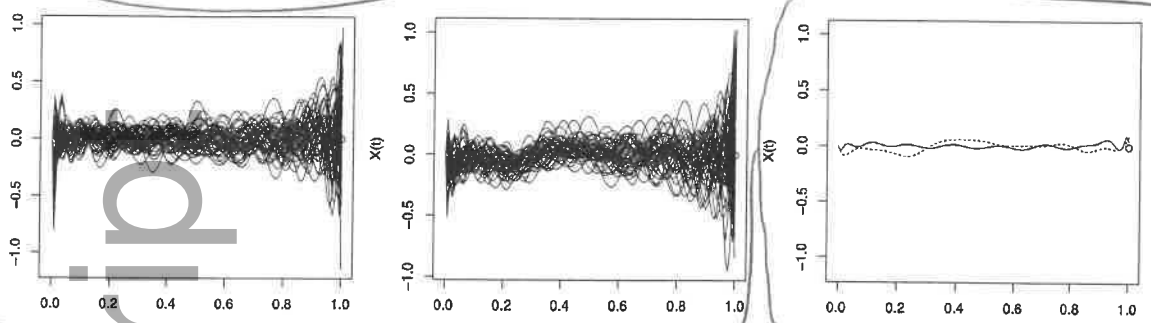


Figure 7: 50 curves from model (iii) coming from population Π_1 (left) or Π_2 (middle), and empirical mean curves from the two groups (right).

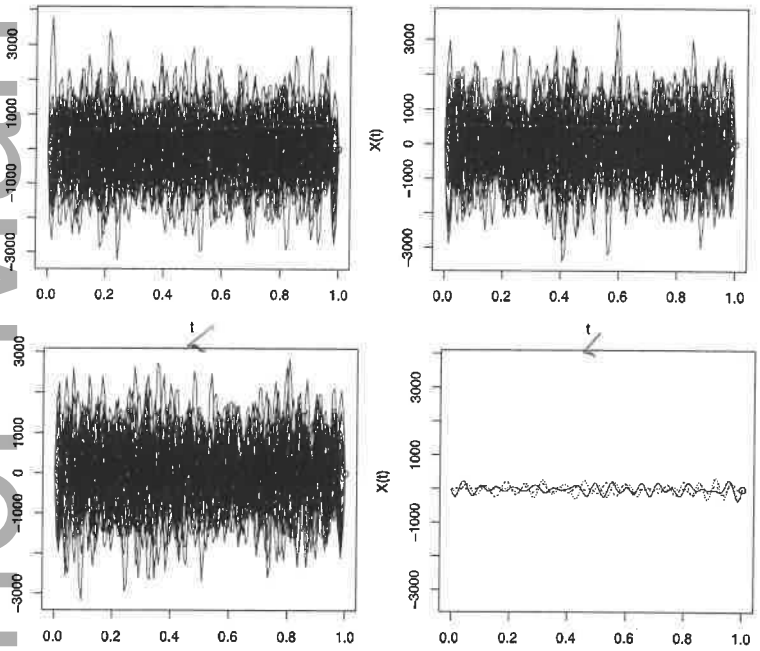


Figure 8: 50 curves from the hierarchical example coming from population Π_1 (top left), Π_2 (top right), Π_3 (bottom left), and empirical mean curves from the three groups (bottom right).

This article is protected by copyright. All rights reserved

enlarge each part to width 16 picas (both figures)

JRSS for B B18356 – Delaigle et al.

Insert (a)-(c) in 8pt Helvetica final size

(a)/

(b)/(c)

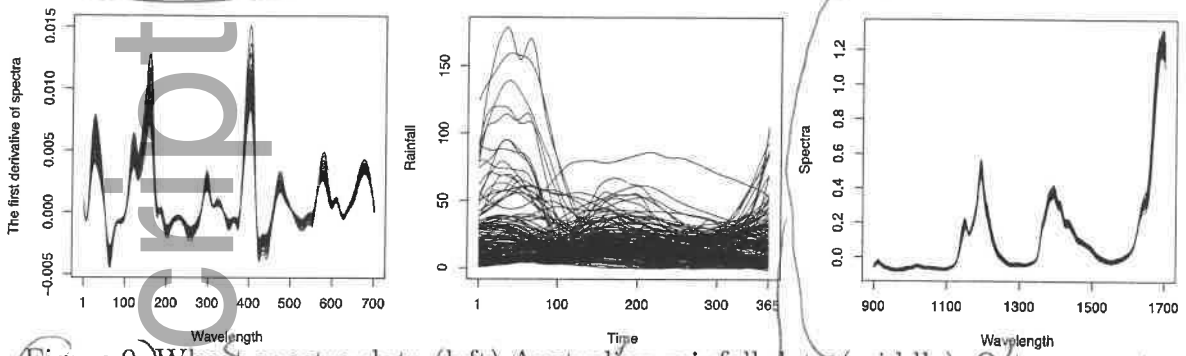


Figure 9: Wheat spectra data (left) Australian rainfall data (middle), Octane spectra curves (right).

Insert (a)-(c) in 8pt Helvetica final size

(a)/

(b)/(c)

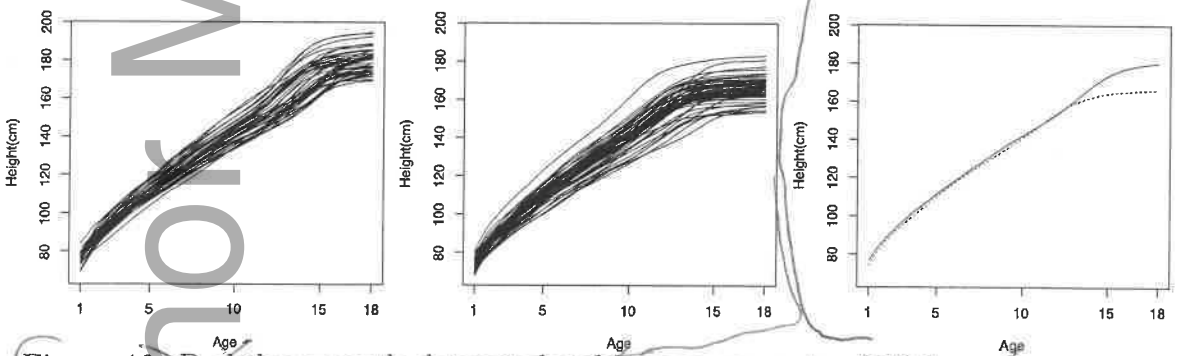


Figure 10: Berkeley growth dataset: height measurements of 39 boys (left) and 54 girls (middle) from age 1 to 18, and mean curves from each group (right).

Author Manuscript

This article is protected by copyright. All rights reserved

J. R. Statist. Soc. B (2019)
81, Part 2, pp. 000–000

JRSS Sec B
B18356 –
Delaigle et al.

Clustering functional data into groups by using projections

Aurore Delaigle, Peter Hall and Tung Pham
School of Mathematics and Statistics and Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers, University of Melbourne, Parkville, VIC 3010, Australia.

Our friend and colleague Peter Hall died in Melbourne, Australia on January 9, 2016, before the revised version of this manuscript was completed. He was an extraordinary researcher and a fantastic man and he will be missed by many.

[Received August 2015. Final revision November 2018]

Abstract. We show that, in the functional data context, by appropriately exploiting the functional nature of the data, it is possible to cluster the observations asymptotically perfectly. We demonstrate that this level of performance can sometimes be achieved by the k-means algorithm as long as the data are projected on a carefully chosen finite dimensional space. In general, the notion of ideal cluster is not clearly defined. We derive our results in the setting where the data come from two populations whose distributions differ at least in terms of means, and where an ideal cluster corresponds to one of these two populations. We propose an iterative algorithm to choose the projection functions in a way that optimises clustering performance, where, to avoid peculiar solutions, we use a weighted least squares criterion. We apply our iterative clustering procedure on simulated and real data, where we show that it works well.

Keywords: asymptotically perfect clustering; functional principal components; Haar basis; k-means; optimal projection

1. Introduction

We consider the two-population functional data clustering problem, where the goal is to cluster, into two groups, curves $X_1, \ldots, X_n$ supported on an interval $\mathcal{I}$, in such a way that data from a same cluster are more similar than data from the other cluster. Recently, in the related functional data classification problem, Delaigle and Hall (2012) have introduced the notion of asymptotically perfect classification. There, they proved that in the two-population functional classification problem, as training sample size increases, it is possible to classify new data correctly with probability

Address for correspondence: Aurore Delaigle,
E-mail: aurored@unimelb.edu

© 2019 Royal Statistical Society

This article is protected by copyright. All rights reserved

1369-7412/19/81000

Author Manuscript

tending to ~~one~~ 1, even in non-pathological cases, as long as the curves are projected on a carefully chosen space of dimension ~~one~~ 1.

We extend the idea of asymptotically perfect performance to the more theoretically challenging functional data clustering problem. Unlike the classification context, we do not have at our disposal a training sample of data from each of the clusters, which, in practice as well, makes the problem more complicated than its classification counterpart. For example, we shall see that, even in cases where we have an analytic formula for a projection that guarantees good performance, we cannot compute it explicitly in practice and need to use instead an iterative procedure. As pointed out by Hennig (2015), there is no unique way to define a perfect cluster. For example, in some applications, it can be reasonable to define ideal clusters directly from the data. Another common approach is based on the assumption that the data come from populations with distinct distributions, and a cluster corresponds to one of the populations; this is the approach that we use in this paper. Specifically, suppose that a clustering algorithm has partitioned the set $\mathcal{X} = \{X_1, \ldots, X_n\}$ of all data into two disjoint parts, $\mathcal{X}_1$ and $\mathcal{X}_2$, and that a detailed analysis of the origins of the data shows that there are just two populations, or subpopulations, Π_1 and Π_2, which differ at least in terms of their means.

We say that the clustering algorithm is asymptotically perfect if, with probability converging to 1 as the total sample size diverges, it asymptotically correctly ascribes the data in any given Π_j to a unique cluster, $\mathcal{X}_j$ say. The latter property means that the proportion of data in Π_j that are ascribed to $\mathcal{X}_j$ by the algorithm converges to 1 as the sample size increases. We establish this property in the case where the two populations differ only in terms of means, where we show that this level of performance is achieved by applying a weighted version of the standard k-means algorithm to a carefully chosen univariate projection of the data; we derive an analytic formula for this projection. Roughly speaking, asymptotically perfect clustering is possible in cases where the difference between the means μ_1 and μ_2 of the two populations is

This article is protected by copyright. All rights reserved

The programs that were used to analyse the data can be obtained from

https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets

large relative to the variability of the data. More specifically, the accumulation over $j = 1, 2, \ldots$ of the square of the projection of $\mu_1 - \mu_2$ on the univariate space generated by the jth eigenfunction of the covariance operator, divided by the corresponding jth eigenvalue, is so large that it is possible to project the data from the two populations on two univariate tight clusters that are infinitely apart from each other. We extend our theoretical results to other distributional settings and to the case of more than two populations that can be clustered hierarchically. In these more complex situations, we might need more than one projection function, and these functions are usually defined only implicitly. We introduce a data-driven way of choosing them.

2 Methodology for clustering

2.1 k-means algorithms for functional data

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote a random sample of random functions, each distributed as the generic function X and supported on the interval $\mathcal{I}$. Recall that the goal of a clustering algorithm is to cluster the set $\mathcal{X}$ in two disjoint parts $\mathcal{X}_1$ and $\mathcal{X}_2$. For $k = 1, 2$, we let X_{ki} denote the ith curve that is classified in the kth group. In this notation, $\mathcal{X}_k = \{X_{k1}, \ldots, X_{kn_k}\}$ where $n_1 + n_2 = n$.

We focus on the k-means algorithm, which is one of the most popular techniques. When the observations are p-vectors V_{ki}, the k-means algorithm chooses the two partitions so as to minimise

$$\sum_{k=1}^{2} \sum_{i=1}^{n_k} \|V_{ki} - \bar{V}_k\|^2, \tag{2.1}$$

where $\bar{V}_k = n_k^{-1} \sum_{i=1}^{n_k} V_{ki}$ and $\|(u_1, \ldots, u_p)\| = (\sum_j u_j^2)^{1/2}$ is the conventional square norm for a p-vector.

In the functional data context, this algorithm can be applied in several ways. A first possibility is to replace the norm of vectors by the L_2 norm of functions or one of their derivatives. Specifically, this version of the algorithm forms the partitions so as to minimise $\sum_{k=1}^{2} \sum_{i=1}^{n_k} \int \{X_{ki}^{(\nu)}(t) - \bar{X}_k^{(\nu)}(t)\}^2 \, dt$, where ν is a positive integer to

This article is protected by copyright. All rights reserved

be chosen by the experimenter, and $\bar{X}_k^{(\nu)} = n_k^{-1} \sum_{i=1}^{n_k} X_{ki}^{(\nu)}$; see for example Tarpey and Kinateder (2003) and Ferraty and Vieu (2006).

This approach is simple to implement, but it is well known that applying k-means to the whole functions is often not the best strategy because these typically contain too much noise. A more commonly used technique consists in applying k-means to the vectors V_{ki} of dimension, say p, obtained by projecting the curves X_{ki} onto a space of finite dimension p. A common way to project a function X onto such a space is to express X as a linear combination of orthonormal functions $\psi_1, \psi_2, \ldots$, that is $X = \sum_{j=1}^{\infty} \alpha_j \psi_j$, where $\alpha_j = X(\psi_j) \equiv \int_{\mathcal{I}} X(t)\psi_j(t)\,dt$, then truncate the infinite sum to p terms (often, but not always, the first p), and take V to be the p-vector of the corresponding coefficients α_j, for example $V = (\alpha_1, \ldots, \alpha_p)$.

Often the ψ_js are taken to be elements of a predetermined basis such as the B-spline basis, a Fourier basis, a wavelet basis or the principal component basis. See for example Abraham et al. (2003), Serban and Wasserman (2005), Auder et al. (2012) and Antoniadis et al. (2013). Sometimes k-means is used only as a preliminary clustering procedure, which is then refined at a second stage. See Chiou and Li (2007), where the preliminary clustering is applied with the ψ_js being eigenfunctions.

A problem when the ψ_js are taken to be elements of a predetermined basis is that they are not chosen in a way that tries to optimise clustering performance, and as a result the coefficients α_j do not necessarily carry the most relevant information for clustering the functions. In a first attempt to overcome this difficulty, Delaigle et al. (2012) suggested projecting each X_i onto a vector $V_i = (X_i(t_1), \ldots, X_i(t_p))$, where the points $t_1, \ldots, t_p$ were selected from the data, so as to try and maximise clustering performance. In Gattone and Rocci (2012) the authors suggested approximating the ψ_js and the X_is by linear combinations of basis functions $\phi_1, \ldots, \phi_q$, with q large, and then selecting the p ψ_js ($p \leq q$) so as to optimise clustering, under a smoothness penalty for the cluster means $\bar{X}_k$. These two approaches are good steps forward in choosing projections so as to optimise clustering performance. In the next section,

This article is protected by copyright. All rights reserved

motivated by theoretical considerations, we introduce a weighted version of the k-means algorithm which guarantees good, indeed sometimes asymptotically perfect, theoretical properties in the functional setting. We shall see in section 4 that it also performs very well in practice.

2.2 A modified k-means algorithm for functional data

Write $\psi_1, \ldots, \psi_p$ for a given, linearly independent sequence of functions defined on $\mathcal{I}$, let $X_i(\psi_j) = \int_{\mathcal{I}} \psi_j X_i$ be the projection of X_i onto the real line in the direction determined by ψ_j, put $\vec{\psi} = (\psi_1, \ldots, \psi_p)$ and let $X_i(\vec{\psi}) = (X_i(\psi_1), \ldots, X_i(\psi_p))$, a p-vector. We propose a weighted version of k-means clustering applied to the p-vectors $X_i(\vec{\psi})$, to determine first the functions $\psi_1, \ldots, \psi_p$ and then the clusters. Here and below the integer p is fixed; see section 4.1 for how to choose it in practice.

The k-means algorithm is usually applied by minimising a measure of the tightness of clusters. In the standard version of k-means for p-vectors used at (2.1), if our clustering algorithm has partitioned $\mathcal{X}$ into two disjoint parts $\mathcal{X}_1$ and $\mathcal{X}_2$ as in section 2.1, then cluster tightness is measured by

$$T_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi}) = \frac{1}{n} \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left\| X_{ki}(\vec{\psi}) - \bar{X}_k(\vec{\psi}) \right\|^2, \tag{2.2}$$

where $X_{ki}(\vec{\psi}) = (\int_{\mathcal{I}} \psi_1 X_{ki}, \ldots, \int_{\mathcal{I}} \psi_p X_{ki})$ and $\bar{X}_k(\vec{\psi}) = n_k^{-1} \sum_{1 \le i \le n_k} X_{ki}(\vec{\psi})$.

The goal of the k-means clustering approach is to find the clustering that produces the minimum $T_2(\vec{\psi})$ of $T_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi})$:

$$T_2(\vec{\psi}) = \min_{\mathcal{X}_1, \mathcal{X}_2} T_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi}). \tag{2.3}$$

Here, the minimum is taken over all sequences of two disjoint, nonempty sets $\mathcal{X}_1$ and $\mathcal{X}_2$ whose union is $\mathcal{X}$. However, the tightness measure at (2.2) can be misleading if it is used to compare different vectors $\vec{\psi}$ of functions. This is because, if X_i is smooth and ψ_j is a highly oscillating function, the value of $\int_{\mathcal{I}} \psi_j X_i$ can be made arbitrarily small simply by increasing the frequency, without violating a standardisation condition such

Author Manuscript

This article is protected by copyright. All rights reserved

as $\|\psi_j\| = 1$. For example, if $\psi_j(t) = c\sin(jt)$ for a constant $c > 0$, and X_i has a bounded derivative on $\mathcal{I}$, then $\int_{\mathcal{I}} \psi_j X_i \to 0$ as j diverges. Therefore $T_2(\mathcal{X}_1, \mathcal{X}_2 \,|\, \vec{\psi})$ can be rendered small simply by choosing the functions ψ_j to have high frequency oscillations, without materially affecting cluster tightness.

This concern suggests that, in the functional data context, a modification of the definition at (2.2) is needed, where the components of each $X_i(\vec{\psi})$ are standardised for scale before being used to calculate $T_2(\mathcal{X}_1, \mathcal{X}_2 \,|\, \vec{\psi})$. Without taking scale into account, T_2 can be small simply because most $X_{ki}(\vec{\psi})$s are small, and not just because they are close to $\bar{X}_k(\vec{\psi})$. With this in mind, we define

$$\bar{X}(\psi_j) = n^{-1} \sum_{i=1}^{n} X_i(\psi_j)\,, \quad \hat{\sigma}(\psi_j)^2 = \frac{1}{n} \sum_{i=1}^{n} \{X_i(\psi_j) - \bar{X}(\psi_j)\}^2\,.$$

As before we partition $\mathcal{X}$ into two disjoint, nonempty sets $\mathcal{X}_1$ and $\mathcal{X}_2$, with $\mathcal{X}_k = \{X_{k1}, \ldots, X_{kn_k}\}$ where $n_1 + n_2 = n$, but we propose measuring the tightness of clusters differently. Specifically, we use the following analogue of $T_2(\mathcal{X}_1, \mathcal{X}_2 \,|\, \vec{\psi})$ to characterise cluster tightness:

$$\widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi}) = \sum_{j=1}^{p} \frac{1}{\hat{\sigma}(\psi_j)^2} \frac{1}{n} \sum_{k=1}^{2} \sum_{i=1}^{n_k} \{X_{ki}(\psi_j) - \bar{X}_k(\psi_j)\}^2\,, \tag{2.4}$$

where $X_{ki}(\psi_j) = \int_{\mathcal{I}} \psi_j X_{ki}$ and $\bar{X}_k(\psi_j) = n_k^{-1} \sum_{1 \le i \le n_k} X_{ki}(\psi_j)$. In section 3 we shall prove that this weighted version of k-means gives particularly good clustering performance, and can even provide asymptotically perfect clustering.

With this version of tightness, the goal of the k-means algorithm is to produce an analogue of $T_2(\vec{\psi})$ at (2.3):

$$\widehat{T}_2(\vec{\psi}) = \min_{\mathcal{X}_1, \mathcal{X}_2} \widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi})\,. \tag{2.5}$$

In order to ensure good clustering performance, we optimise over $\vec{\psi}$ by taking

$$\widehat{\vec{\psi}} = \operatorname*{argmin}_{\vec{\psi} \in Q_n} \widehat{T}_2(\vec{\psi})\,, \tag{2.6}$$

This article is protected by copyright. All rights reserved

where Q_n denotes a class of orthogonal functions which is permitted to become steadily more complex as sample size grows. For example, one approach to approximating members of the class $L_2(\mathcal{I})$ of square integrable functions on $\mathcal{I}$ is to use a complete orthonormal sequence $\chi_1, \chi_2, \ldots$ in the class. Then any $\psi \in L_2(\mathcal{I})$ can be expressed as the L_2 limit as $r \to \infty$ of approximations of the form

$$\psi^{(r)}(x) = \sum_{j=1}^{r} d_j \chi_j(x)\,, \tag{2.7}$$

where $d_1, d_2, \ldots$ is a sequence of real numbers that are elements of a potentially growing set $\mathcal{D}_n$, and are such that $\sum_j d_j^2 < \infty$. We shall discuss the choice of the functions χ_j in section 4.1.

In practice, k-means is usually applied through iterative algorithms, the most popular of which is probably Lloyd's algorithm (Lloyd, 1957, 1982), which is what we used in our numerical work. This algorithm starts by creating an initial partition of the data into two clusters. Then, at each iteration, it partitions the space into a centroidal Voronoi tessellation of $\mathbb{R}^p$, generated by the centres of the clusters. A variety of other approaches are possible. See Telgarsky and Vattani (2010), who argue that an algorithm suggested by Hartigan (1975) is particularly competitive. See also Brusco and Steinley (2007). These algorithms often provide good clustering, although they are sensitive to the initial partition so that they are often applied with several initial random partitions.

3 Theoretical interpretation of clustering methodology

We study theoretical properties of our suggested algorithm. First, in section 3.1, we establish the asymptotic limit of the empirical criterion $\hat{T}_2$, uniformly over the functions ψ_j and the partitions of $\mathbb{R}^p$. Then, in section 3.2, we establish the asymptotically near perfect clustering property in the particular case where the populations differ

Author Manuscript

This article is protected by copyright. All rights reserved

only in terms of means. We extend these results to a more general heteroscedastic setting in section 5.1, and to the case of more than two populations in section 5.2.

3.1 Asymptotic limit of $\widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \,|\, \vec{\psi})$ at (2.4)

In this section we derive the limit of $\widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \,|\, \vec{\psi})$ at (2.4), as $n \to \infty$. This sort of limiting results exist in the multivariate case (see Pollard (1981)), but establishing them in the functional data case requires different arguments because of the intrinsically different nature of the data. Let X denote a generic X_i, put $X(\vec{\psi}) = (X(\psi_1), \ldots, X(\psi_p))$, $x = (x_1, \ldots, x_p)$ and let

$$\sigma(\psi_j)^2 = \mathrm{var}\{X(\psi_j)\}\,. \tag{3.1}$$

Write $F(\cdot \,|\, \vec{\psi})$ for the distribution function of $X(\vec{\psi})$, take $\vec{\mathcal{R}} = (\mathcal{R}_1, \mathcal{R}_2)$ to be a partition of $\mathbb{R}^p$ into two disjoint regions for each of which $\pi(\vec{\psi}, \mathcal{R}_k) \equiv P\{X(\vec{\psi}) \in \mathcal{R}_k\} > 0$, let $\mu_{kj}(\vec{\psi}) = \pi(\vec{\psi}, \mathcal{R}_k)^{-1} \int_{\mathcal{R}_k} x_j \, dF(x \,|\, \vec{\psi})$ and put

$$t_2(\vec{\psi} \,|\, \vec{\mathcal{R}}) = \sum_{j=1}^{p} \frac{1}{\sigma(\psi_j)^2} \sum_{k=1}^{2} \int_{\mathcal{R}_k} \{x_j - \mu_{kj}(\vec{\psi})\}^2 \, dF(x \,|\, \vec{\psi})\,. \tag{3.2}$$

We can interpret $\widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \,|\, \vec{\psi})$, defined at (2.4), as an empirical approximation to $t_2(\vec{\psi} \,|\, \mathcal{R}_1, \mathcal{R}_2)$ when the region $\mathcal{R}_k$ contains $\mathcal{X}_k$, for $k = 1, 2$. More particularly, if we put $\mathcal{X}_k = \mathcal{X} \cap \mathcal{R}_k$ for $k = 1, 2$ then, for fixed $\vec{\psi}$ and $\vec{\mathcal{R}}$, we have $\widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \,|\, \vec{\psi}) \to t_2(\vec{\psi} \,|\, \vec{\mathcal{R}})$ as $n \to \infty$.

In Theorem 1, below, we shall show that this convergence is uniform in $\vec{\psi} \in Q_n$ and $\vec{\mathcal{R}} \in \mathbb{V}_{p,n}$, where Q_n is a set of functions ψ on $\mathcal{I}$ for which

$$\|\psi\| = 1 \text{ for all } \psi \in Q_n\,, \quad \#Q_n \leq a_n\,, \tag{3.3}$$

with $\#$ denoting the number of elements of a set and a_n denoting a sequence of positive numbers diverging to infinity, and, motivated by the fact that the optimal clustering found through minimisation of (2.5) is a centroidal Voronoi tessellation of

This article is protected by copyright. All rights reserved

$\mathbb{R}^p$ (see Hasegawa et al. (1993) and Du et al. (1999)), $\mathbb{V}_{p,n}$ is a set of Voronoi tessellation partitions of $\mathbb{R}^p$ into two cells, satisfying

$$\#\,\mathbb{V}_{p,n} \leqslant b_n\,, \tag{3.4}$$

where b_n denotes a sequence of positive numbers diverging to infinity; see (3.10) below for bounds on a_n and b_n. (Here and throughout the paper, when referring to Voronoi tessellations, these are constructed with the weighted distance as in (2.4).) In particular, $\hat{\vec{\psi}}$ at (2.6) can be viewed as an empirical approximation to a vector of projections into $\mathbb{R}^p$ that minimises the dispersion measure given by t_2.

It is notationally convenient to represent a partition as a sequence $\vec{\mathcal{R}} = (\mathcal{R}_1, \mathcal{R}_2)$ but to establish the theorem if one partition is in $\mathbb{V}_{p,n}$ then we assume without loss of generality that its permutation $(\mathcal{R}_2, \mathcal{R}_1)$ is too. Moreover, we ask that, for a constant $\pi_{\min} > 0$ not depending on n,

$$\inf_{\psi_1,\ldots,\psi_p \in Q_n} \inf_{\vec{\mathcal{R}} \in \mathbb{V}_{p,n}} \inf_{k=1,2} \pi(\vec{\psi}, \mathcal{R}_k) \geqslant \pi_{\min} \tag{3.5}$$

That is, the cells in the Voronoi tessellation do not get too small, in the sense that the probability that $X(\vec{\psi})$ lies in any particular cell is bounded away from zero.

Let $X_1(\psi_j \,|\, \mathcal{R}_k), \ldots, X_n(\psi_j \,|\, \mathcal{R}_k)$ be independent and identically distributed with the distribution of $X(\psi_j)$, conditional on $X(\vec{\psi}) \in \mathcal{R}_k$. Since all clustering steps involve empirical centring and scaling then we may suppose, without loss of generality, that $E(X) = 0$ and $E\|X\|^2 = 1$. For a general random variable R satisfying $E|R| < \infty$, we write $(1 - E)\, R = R - E(R)$. We assume that, for each $\epsilon > 0$,

$$\max_{\ell=1,2} \sup_{\psi_1,\ldots,\psi_p \in Q_n} \max_{1 \leqslant j \leqslant p} \sup_{\vec{\mathcal{R}} \in \mathbb{V}_{p,n}} \sup_{k=1,2} P\left\{ \left| \frac{1}{n} \sum_{i=1}^n (1 - E)\, X_i(\psi_j \,|\, \mathcal{R}_k) \right| > \epsilon\, \sigma(\psi_j)^2 \right\}$$
$$= O\big[\exp\big\{ - C(\epsilon)\, n^c \big\} \big]\,, \tag{3.6}$$

$$\sup_{\psi \in Q_n} P\left\{ \left| \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{I}} \psi\, X_i \right| > \epsilon\, \sigma(\psi)^2 \right\} = O\big[\exp\big\{ - C(\epsilon)\, n^c \big\} \big]\,, \tag{3.7}$$

$$\sup_{\psi \in Q_n} P\left\{ \left| \frac{1}{n} \sum_{i=1}^n (1 - E) \left(\int_{\mathcal{I}} \psi\, X_i \right)^2 \right| > \epsilon\, \sigma(\psi)^4 \right\} = O\big[\exp\big\{ - C(\epsilon)\, n^c \big\} \big]\,, \tag{3.8}$$

Author Manuscript

This article is protected by copyright. All rights reserved

$$\inf_{\psi \in Q_n} \sigma(\psi)^4 \geq C(\epsilon)\, n^{-(1-c)}\,, \tag{3.9}$$

where, here and below, $C(\epsilon)$ denotes a generic positive constant depending on ϵ but not on n, and $0 < c < 1$. Conditions (3.6)–(3.8) assert moderate deviation bounds for the probabilities on the respective left-hand sides. There we ask that the probabilities be exponentially small, but cases where the right-hand sides of (3.6)–(3.8) decrease at a polynomial rather than exponential rate can also be treated using our arguments. See Appendix A.4 for an illustration of these conditions.

Theorem 1. *If (3.5)–(3.9) hold, and the growth rates a_n and b_n are sufficiently low to ensure that, for all $C > 0$,*

$$a_n^p\, b_n \exp\bigl(-C\, n^c\bigr) \to 0 \tag{3.10}$$

as $n \to \infty$, then as n diverges, and for all $\epsilon > 0$,

$$P\Bigl\{ \sup_{\psi_1,\ldots,\psi_p \in Q_n} \sup_{\vec{\mathcal{R}} \in \mathbb{V}_{p,n}} \bigl| \widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi}) - t_2(\vec{\psi} \mid \vec{\mathcal{R}}) \bigr| > \epsilon \Bigr\} \to 0\,. \tag{3.11}$$

If the p-vector $X(\vec{\psi})$ has a well-defined probability density $f(\cdot \mid \vec{\psi})$ that is continuous and positive everywhere in $\mathbb{R}^p$, then the minimum $t_2(\vec{\psi})$ of $t_2(\vec{\psi} \mid \vec{\mathcal{R}})$ over choices of the partition $\vec{\mathcal{R}}$ of $\mathbb{R}^p$ is achieved for a partition $(\mathcal{R}_1^0, \mathcal{R}_2^0)$ that is a centroidal Voronoi tessellation of $\mathbb{R}^p$. (See Schreiber (1998), among others, for discussion of Voronoi tessellations in the context of k-means clustering.) In particular, each $\mathcal{R}_k^0$ is a polygonal prism in $\mathbb{R}^p$; its centroid, defined in the sense of mean with respect to the density $f(\cdot \mid \vec{\psi})$, is the point $\mu_k(\vec{\psi}) = (\mu_{k1}(\vec{\psi}), \ldots, \mu_{kp}(\vec{\psi}))$. (The assumption that $f(\cdot \mid \vec{\psi})$ exists and is continuous and positive can be relaxed, although at the cost of treating issues such as contiguity of the support of $f(\cdot \mid \vec{\psi})$.)

Let $\mathcal{C}_1(\vec{\psi})$ and $\mathcal{C}_2(\vec{\psi})$ denote the convex hulls of $\mathcal{X}_1$ and $\mathcal{X}_2$, where $(\mathcal{X}_1, \mathcal{X}_2)$ is the partition of $\mathcal{X}$ that minimises $\widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi})$. If $\mathcal{R}_1^0$ and $\mathcal{R}_2^0$ are uniquely defined up to permutations then a permutation (depending on the dataset $\mathcal{X}$) of $\mathcal{C}_1(\vec{\psi})$ and $\mathcal{C}_2(\vec{\psi})$ converges in probability, as $n \to \infty$, to $\mathcal{R}_1^0$ and $\mathcal{R}_2^0$, in the following sense: for any

This article is protected by copyright. All rights reserved

fixed, closed p-sphere $\mathcal{S}$ in $\mathbb{R}^p$, there exists a permutation of $\mathcal{C}_1 \cap \mathcal{S}$ and $\mathcal{C}_2 \cap \mathcal{S}$ that converges in probability, as a sequence of random sets, to $\mathcal{R}_1^0 \cap \mathcal{S}$ and $\mathcal{R}_2^0 \cap \mathcal{S}$.

3.2 Asymptotically near perfect clustering when the populations differ only in terms of means

In this section we introduce the notion of asymptotically perfect clustering, where, in the case $p = 1$, minimising the asymptotic version $t_2(\vec{\psi} \,|\, \vec{\mathcal{R}})$ at (3.2) of $\widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \,|\, \vec{\psi})$ at (2.4), with respect to both $\vec{\psi}$ and $\vec{\mathcal{R}}$, can lead to regions $\mathcal{R}_1$ and $\mathcal{R}_2$ each of which contains asymptotically all of the data from just one of the populations, and asymptotically none of the data from the other population. Here asymptotically all and asymptotically none are interpreted in the sense that the proportions of data converge to 1 and 0 respectively, along any sequence of values of $\vec{\psi}$ and $\vec{\mathcal{R}}$ for which $t_2(\vec{\psi} \,|\, \vec{\mathcal{R}})$ converges to its minimum.

We shall simplify our discussion by assuming that the populations differ only in location (see section 5.1 for more general heteroscedastic settings). In this simple case, we can derive explicit results and a simple analytic formula for the unique function ψ to use. We assume that there are two subpopulations Π_1 and Π_2 (see section 5.2 for $K > 2$ populations), and that the data from the kth can be represented as

$$X_{ki} = \nu_k + Z_{ki}\,, \quad 1 \leq i \leq n_k\,, \tag{3.12}$$

where

- The functions Z_{ki} are all distributed as the fixed random function Z, say, with $E(Z) = 0$, the location terms ν_k, for $k = 1, 2$ are fixed, and the populations Π_1 and Π_2 arise in respective proportions ρ_1 and ρ_2, where $\rho_1 + \rho_2 = 1$ and each $\rho_k > 0$. (3.13)

In (3.13) the random function Z is fixed in the sense that its distribution does not depend on k or i. In view of (3.13), a generic X_i can be represented as

$$X = \sum_{k=1}^{2} \nu_k I_k + Z\,, \tag{3.14}$$

Author Manuscript

This article is protected by copyright. All rights reserved

where the zero-one variables I_1 and I_2 are independent of Z, $P(I_k = 1) = 1 - P(I_k = 0) = \rho_k$ and $P(I_1\, I_2 = 0) = 1$.

We shall show that in this setting it is possible to have $t_2(\vec{\psi}\,|\,\vec{\mathcal{R}}) \to 0$, along a sequence of values of $\vec{\psi}$ and $\vec{\mathcal{R}}$. To see how to define explicitly a function ψ that results in asymptotically perfect clustering, write $\kappa(u,v) = \text{cov}\{Z(u), Z(v)\}$ for the covariance function of Z, and let $\kappa(u,v) = \sum_{j=1}^{\infty} \theta_j\, \phi_j(u)\, \phi_j(v)$ denote the spectral decomposition of κ. Here θ_j and ϕ_j are eigenvalues and eigenfunctions, respectively, of the transformation, also denoted by κ, that takes a function ψ to $\kappa(\psi)$ defined by $\kappa(\psi)(u) = \int \kappa(u,v)\, \psi(v)\, dv$. (The dual usage of κ in this setting is common; the context disambiguates notation.) We can write $\nu_k = \sum_{j=1}^{\infty} \nu_{kj}\, \phi_j$ and $Z_{ki} = \sum_{j=1}^{\infty} \theta_j^{1/2}\, Z_{kij}\, \phi_j$, where $\nu_{kj} = \int_{\mathcal{I}} \nu_k\, \phi_j$ and $Z_{kij} = \theta_j^{-1/2} \int_{\mathcal{I}} Z_{ki}\, \phi_j$ is the jth principal component score of Z_{ki}.

Recalling (2.7), define $\psi^{(r)} = \sum_{j=1}^{r} \alpha_j\, \phi_j$. Then it is easy to see that $E\{X_{ki}(\psi^{(r)})\} = \sum_{j=1}^{r} \nu_{kj}\, \alpha_j$ and $\text{var}\{X_{1i}(\psi^{(r)})\} = \text{var}\{X_{2i}(\psi^{(r)})\} = \sum_{j=1}^{r} \theta_j\, \alpha_j^2$, and a calculus of variations argument shows that, over all vectors ψ having the form $\sum_{j \le r} \beta_j\, \phi_j$ for constants β_j, the ratio $|E\{X_{1i}(\psi)\} - E\{X_{2i}(\psi)\}| / [\text{var}\{X_{ki}(\psi)\}]^{1/2}$ is maximised by taking $\alpha_j = \text{const.}\, \theta_j^{-1}\, (\nu_{1j} - \nu_{2j})$, where const. denotes a strictly positive constant. With this definition of α_j, and letting

$$c_r = \sum_{j=1}^{r} \theta_j^{-1}\, (\nu_{1j} - \nu_{2j})^2, \tag{3.15}$$

we also find that

$$\text{const.}^{-1} \bigl| E\{X_{1i}(\psi^{(r)}) - X_{2i}(\psi^{(r)})\} \bigr| = \text{const.}^{-2} \text{var}\{X_{ki}(\psi^{(r)})\} = c_r. \tag{3.16}$$

An important feature of (3.16) is that if c_r diverges to infinity as r increases, then the standard deviation of $X_{ki}(\psi^{(r)})$ becomes negligibly small relative to the difference between the means of the two groups, and so for $k = 1, 2$, data points in the sample $\mathcal{X}$ that come from Π_k are readily identified as a tight cluster around the point $\int_{\mathcal{I}} \psi^{(r)}\, \nu_k$, if we project data in the direction $\psi^{(r)}$. This property remains true if we standardise

This article is protected by copyright. All rights reserved

$\psi^{(r)}$ so that $\|\psi^{(r)}\| = 1$. It is intuitively clear that if c_r, at (3.15), tends to ~~infinity~~ as $r \to \infty$, then, as $r \to \infty$, it is possible to cluster the data perfectly. The next theorem establishes this rigorously.

Theorem 2. *Assume the homoscedastic mixture model defined by (3.12) and (3.13), with κ strictly positive definite and uniformly bounded, and let $\psi^{(r)} = \sum_{j=1}^{r} \theta_j^{-1} (\nu_{1j} - \nu_{2j}) \phi_j$. Then if $\mathbb{E}(Z_{1ij}^4) = \mathbb{E}(Z_{2ij}^4) \equiv \kappa_4 < \infty$ and*

$$\sum_{j=1}^{\infty} (\nu_{1j} - \nu_{2j})^2 \theta_j^{-1} = \infty \tag{3.17}$$

holds, the clustering algorithm is asymptotically perfect, in the sense that, if $\mathcal{R}_1$ and $\mathcal{R}_2$ denote the regions found by minimising (3.2), there ~~exists~~ a permutation $(k(1), k(2))$ of the pair $(1, 2)$ such that, for $\ell = 1, 2$,

$$\lim_{r \to \infty} P\{X(\psi^{(r)}) \in \mathcal{R}_{k(\ell)} \,|\, I_\ell = 1\} = 1\,. \tag{3.18}$$

As noted by Delaigle and Hall (2012) in their functional classification context, asymptotically perfect performance does not usually hold in the multivariate case, except in pathological instances. By contrast, as already highlighted by Delaigle and Hall (2012), condition (3.17) can be satisfied in more standard cases, ~~for example~~ e.g. when, for j large, the eigenvalues θ_j and the difference of the mean coefficients ν_{1j} and ν_{2j} are such that $\theta_j = O\{(\nu_{1j} - \nu_{2j})^2\}$. This makes the clustering problem for functional data unique and particularly interesting.

4 Numerical properties

4.1 Empirical choice of the functions ψ_j

There are many ways to choose the space Q_n at (2.6) in which we seek the functions $\psi_1, \ldots, \psi_p$. Except in the setting of Theorem 2, in general we do not have an explicit expression for the functions ψ_j that should be used for projecting the data, nor for the number p of functions that we should use. In practice, where n is finite we

This article is protected by copyright. All rights reserved

suggest taking each ψ_j to be a linear combination of r basis functions, as in (2.7). Specifically, for $j = 1, \ldots, p$, we take $\psi_j = \sum_{i=1}^{r} d_{j,i}\, \chi_i$, where $\chi_1, \ldots, \chi_r$ are the first r elements of a basis and the $d_{j,i}$s are coefficients to be determined.

In the functional data literature, when choosing bases it is common to use a spline basis, the principal component basis or a Fourier basis. If sample size was infinite, this choice would not matter much. In practice, however, we only have a finite number of observations, which implies that we cannot take r too large. In turn, this implies that, in order to capture the main differences between the two populations, we need to choose the basis functions χ_j carefully. Because its elements are gradually more localised as frequency increases, we suggest using the Haar basis (see Härdle et al. (2012)), or its unbalanced version (Fryzlewicz, 2007), which captures both global and increasingly local trends, and thus offers a good balance between localising and over-localising. It can often capture details without the need for r being too large.

To compute the ψ_js explicitly, we need to choose the $d_{j,i}$s and p in a way that ensures good clustering performance. We suggest choosing p and the $d_{j,i}$s iteratively; in principle we could also choose r iteratively, but choosing both r and the $d_{j,i}$s adaptively is somewhat redundant as we can reduce r artificially simply by setting all the $d_{j,i}$s to zero for appropriate values of j. Therefore, a good alternative and computationally less demanding approach is to fix r to be a large value, and choose only the $d_{j,i}$s adaptively. Motivated by this, in our numerical work we took $r = 16$ (we tried larger values but found this took much more time while not improving the results significantly).

For a fixed value of p, we need find the $d_{j,i}$s that minimise $\hat{T}_2$ at (2.5) by applying Lloyd's algorithm iteratively to a set of $d_{j,i}$s, under the constraint that the ψ_js are orthogonal. This problem is too difficult to be solved in a reasonable amount of time and so we use a greedy algorithm to search for the $d_{j,i}$s. Note that these form a set of p orthnormal vectors in a space of dimension r, that is p vectors located on an

This article is protected by copyright. All rights reserved

r-dimensional unit sphere. Our greedy algorithm iteratively examines a very large collection of p orthogonal vectors on this sphere by starting with an arbitrary set of p orthonormal vectors, and iteratively rotating and reflecting them in two-dimensional subspaces. See Appendix B.1 for details. As in Delaigle et al. (2012), we choose the number p of projections by using a measure of tightness of the clusters, defined by

$$\mathcal{T}_p = \sum_{k=1}^{2} \sum_{X_i \in \mathcal{C}_k(p)} \|X_i - \bar{X}_k\|^2, \tag{4.1}$$

where $\mathcal{C}_k(p)$ denotes the kth cluster of the partition (that is obtained when using p projection functions, $\bar{X}_k = \{\#\mathcal{C}_k(p)\}^{-1} \sum_{i=1}^{n} X_i I\{X_i \in \mathcal{C}_k(p)\}$, $\#\mathcal{C}_k(p)$ denotes the number of observations in cluster $\mathcal{C}_k(p)$ and, on this occasion, $\|\cdot\|$ is the L_2 norm of functions. Specifically, to choose a reasonable value p^* of p, we consider increasing values of p (i.e. $p = 1, 2, \ldots$) until $\mathcal{T}_p$ starts to increase, e.g., or decreases by a too small amount. This can be achieved by choosing $p^* = \inf_p \{p : \mathcal{T}_p - \mathcal{T}_{p+1} \leq \rho \mathcal{T}_1\}$ where, as in Delaigle et al. (2012), ρ denotes a pre-determined small proportion, ~~for example~~ e.g. equal to $\rho = 0.05$, 0.1 or 0.2. As in the classification context of Delaigle et al. (2012), for most methods our numerical investigations indicated that those values of ρ all gave similar results; see section 4.2 for more details about our practical implementation.

4.2 Numerical results

We compared our clustering algorithm, computed as in section 4.1 with the Haar basis and denoted below by DHP_{HA}, with seven other clustering methods, on both simulated (section 4.2.1) and real (section 4.2.2) data. Our main Matlab code is available ~~at~~ from Courier `http://researchers.ms.unimelb.edu.au/~aurored/Proj_Kmeans.m`; it requires the code available ~~at~~ from Courier `http://researchers.ms.unimelb.edu.au/~aurored/Scaled_Kmeans.m`.

We first considered four commonly used or recently developed approaches: L2_{F}, the standard k-means algorithm applied to the full curves equipped with the functional L_2 distance, as described in the third paragraph of section 2.1; DHB, the

This article is protected by copyright. All rights reserved

clustering method in Delaigle et al. (2012); $L2_{PC}$, the standard k-means algorithm based on (2.1), where $V_{ji} = \int X_i(t)\widehat{\phi}_{M,j}(t)\,dt$, with $\widehat{\phi}_{M,j}$ the jth empirical eigenfunction computed from the data from mixed population (i.e. from the data coming from both Π_1 and Π_2); and GR, the RFRKMn method of Gattone and Rocci (2012) (we used the authors' code, where, as discussed in their section 2.2, we took $p = 1$).

Then, to illustrate the fact that the way we construct the functions ψ_j for our new method (see section 4.1) is important, we considered two alternative ways of computing the ψ_js: DHP_{PC} and DHP_{DB}, where we applied the k-means algorithm based on (2.4) with each ψ_k defined like ψ at (2.7) but instead of the Haar basis as in section 4.1, we took, respectively, $\chi_j = \widehat{\phi}_{M,j}$ and χ_j equal to the jth element of the Daubechies DB2 wavelet basis (see Härdle et al. (2012)). Finally, we also considered our procedure with the Unbalanced Haar basis (DHP_{UBH}) adapted from Fryzlewicz (2007) to our setting, replacing his criterion for choosing a breakpoint of a single curve (see his section 4) by the sum of that criterion over all the data curves.

For all methods except $L2_F$ and GR, we used the tightness-based criterion described below (4.1) to choose the number p of projection functions. For the DHP_{PC} method, since the first few eigenfunctions of the pooled covariance function are often not a very good choice for the basis functions, in order to work reasonably well, the DHP_{PC} method tends to need higher values of p than the other approaches, so that for this method $\rho = 0.05$ works the best (this is what we used in our numerical work for this method). For all other methods, our numerical investigations indicated that $\rho = 0.2$ worked slightly better, which is what we used in our numerical work.

Theorem 2 suggests that, in some cases, our method needs only $p = 1$ function ψ, and that we can chose this function as $\psi = \psi^{(r)} = \sum_{j=1}^{r} \theta_j^{-1}(\nu_{1j} - \nu_{2j})\,\phi_j$. We compared the iterative procedure DHP_{HA} described in section 4.1 with the one where, instead of computing the ψ_js iteratively, we projected the data onto an estimator of the function $\psi^{(r)}$, which we obtained by replacing the θ_js, the ϕ_js and the ν_{kj}s by their empirical estimators computed from the data clustered in two groups at the last

This article is protected by copyright. All rights reserved

step of the above mentioned iterative procedure. Here, we chose r by minimising the tightness $\hat{T}_2$ at (2.5) with respect to r. We obtained almost exactly the same results (not reported here) as with the above iterative procedure, which suggests that the latter is effective.

4.2.1 Simulated data

We applied the eight clustering algorithms to datasets coming from two populations Π_1 and Π_2, where, as in Delaigle and Hall (2012), for $k = 1, 2$ and $i = 1, \ldots, n_k$, the ith observation from the kth population was generated from the model

$$X_{ki}(t) = \sum_{j=1}^{40} \left(\theta_j^{1/2} Z_{kij} + \mu_{jk}\right) \phi_j(t)$$

on a grid of 128 equispaced points t in $\mathcal{I} = [0, 1]$, where the Z_{kij}s were independent standard normal random variables and, for each j, $\phi_j(t) = \sqrt{2}\sin(\pi j t)$. We considered three versions of this model, referred to below as models (i), (ii) and (iii), and which were chosen so as to make the clustering problem reasonably challenging: (i) $\theta_j = j^{-2}$ for $j = 1, \ldots, 40$, $\mu_{jk} = 0$ for $k = 1, 2$ and $j > 6$, $(\mu_{11}, \mu_{21}, \mu_{31}, \mu_{41}, \mu_{51}, \mu_{61}) = (0, -0.30, 0.60, -0.30, 0.60, -0.30)$ and $(\mu_{12}, \mu_{22}, \mu_{32}, \mu_{42}, \mu_{52}, \mu_{62}) = (0, -0.45, 0.45, -0.09, 0.84, 0.60)$; (ii) $\theta_j = \exp[-\{2.1 - (j-1)/20\}^2]$ $\mu_{j1} = 0$ and $\mu_{j2} = 0.2625\,(-1)^{j+1} I\{1 \leq j \leq 3\}$; (iii) $\theta_j = j^{-2}$ $\mu_{j1} = 0$ and $\mu_{j2} = 0.75(-1)^{j+1} I\{1 \leq j \leq 3\}$.

Typical curves from each group for each case, as well as their empirical group means, are depicted in Figures 5 to 7 in Appendix B.2. In both cases, we generated 100 data sets, where each time we considered $n_k = 30$, 50 and 100. To assess the effectiveness of each clustering method, we computed the average purity function and the average adjusted rand index over the 100 simulated data sets, where those two indices are defined as follows in the case of K clusters. For a given data set and for $k = 1, \ldots, K$, let $\mathcal{X}_k^0$ denote the set of observations coming from population Π_k and for each clustering algorithm, let $\mathcal{X}_k$ denote the set of observations clustered in cluster

This article is protected by copyright. All rights reserved

Table 1: 100×purities and 100×adjusted rand indices averaged over 100 samples generated from models (i) to (iii) for several group sizes n_k and clustering methods.

Model	n_k	Methods: DHP$_{HA}$	DHP$_{UBH}$	DHP$_{DB}$	DHP$_{PC}$	L2$_F$	L2$_{PC}$	DHB	GR
		Purities							
(i)	30	74.0	77.0	74.0	82.1	56.3	55.7	56.7	63.7
	50	78.9	78.9	81.6	89.2	55.3	54.9	56.5	64.7
	100	85.0	89.9	85.7	91.5	53.8	53.4	54.6	65.2
(ii)	30	79.8	76.8	59.7	55.5	55.3	55.3	54.9	56.2
	50	82.1	77.4	56.7	54.2	54.8	54.6	54.3	54.9
	100	88.8	87.6	56.9	53.0	53.0	52.7	52.9	53.1
(iii)	30	75.6	78.8	68.9	75.2	69.5	69.1	71.1	74.9
	50	81.6	82.4	68.7	74.5	69.7	69.1	71.0	74.9
	100	84.5	88.5	69.9	84.8	69.7	69.3	70.7	75.6
		Adjusted rand indices							
(i)	30	38.3	43.5	30.5	58.0	0.79	0.35	1.11	7.78
	50	49.4	49.8	45.8	74.8	0.72	0.52	1.65	9.59
	100	66.1	76.7	55.2	80.8	0.40	0.22	0.85	9.55
(ii)	30	39.7	35.0	4.19	0.26	-2e-4	6e-4	-0.10	0.65
	50	46.2	37.6	2.01	0.18	0.39	0.30	0.14	0.46
	100	61.5	58.2	2.33	4e-4	1e-4	-8e-4	-8e-5	9e-4
(iii)	30	32.8	39.5	17.4	31.8	16.2	14.9	20.3	25.8
	50	45.1	46.9	17.6	31.9	16.4	15.1	20.5	25.4
	100	51.6	60.1	18.5	52.0	15.8	15.0	19.0	26.3

k by that algorithm. The purity function (see e.g. Manning et al. (2008)) is defined by

$$\text{purity}(\mathcal{X}_1^0, \ldots, \mathcal{X}_K^0, \mathcal{X}_1, \ldots, \mathcal{X}_K) = \frac{1}{n} \sum_{j=1}^{K} \max_{1 \le k \le K} |\mathcal{X}_k^0 \cap \mathcal{X}_j| .$$

It takes values between 0 and 1; the larger the purity, the more effective the clustering algorithm is. The adjusted rand index is more complicated and we refer to Hubert and Arabie (1985) for a definition. It is often considered to be a better measure of the quality of clusters than purity. Its maximum value is also one when the $\mathcal{X}_k^0$s and the $\mathcal{X}_k$s are in perfect match and the larger the index, the better the match, but it can also take negative values when the partitions are close to being created randomly.

Author Manuscript

This article is protected by copyright. All rights reserved

The results, which are summarised in Table 1, indicate that, overall, our DHP_{HA} method and its unbalanced version DHP_{UBH} worked very well (they were among the best methods for each model), whereas the other methods worked well in some cases but performed very poorly in others. Our DHP_{HA} and DHP_{UBH} approaches improved significantly as the group sizes n_k increased, which reflects the fact that, as the n_k increase, our procedure chooses the projection functions ψ_j so as to optimise clustering performance. The improvement was rather modest or non-existent for the other competing approaches, except for DHP_{PC} and DHP_{DB}, the alternative versions of our method, in models (i) and (iii). The relatively poor performance of DHP_{PC} and DHP_{DB} in model (ii) highlights the importance of the choice of the basis functions χ_j. For example, in this case, since the first few eigenfunctions correspond to θ_j large, they are not a good choice for a basis, since the main differences between the μ_j correspond to θ_j small.

4.2.2 Real data sets

Next we applied the eight clustering methods to four real data sets. The first dataset we considered was described in Kalivas (1997). It consists of a set of moisture and protein levels and near infrared spectra of 100 wheat samples measured at 700 wavelengths. As usual with spectra curves, we took the X_is to be the derivatives of spectra, estimated via spline smoothing as in Ferraty and Vieu (2006). See Figure 9 in Appendix B.2 for a graphical representation of the curves.

We applied the eight clustering methods introduced in section 4.2, and then, for each method, we created a scatterplot of protein versus moisture level (see Figure 1), using different symbols to represent the observations clustered in each group: the blue circles were all clustered in one group, and the red triangles in the other. We can see that the clusters created by our DHP_{HA} procedure (equal to DHP_{DB} and almost equal to DHP_{UBH} in this example) are such that all the data for which moisture level was less than 14% were clustered in a group and those with moisture level

This article is protected by copyright. All rights reserved

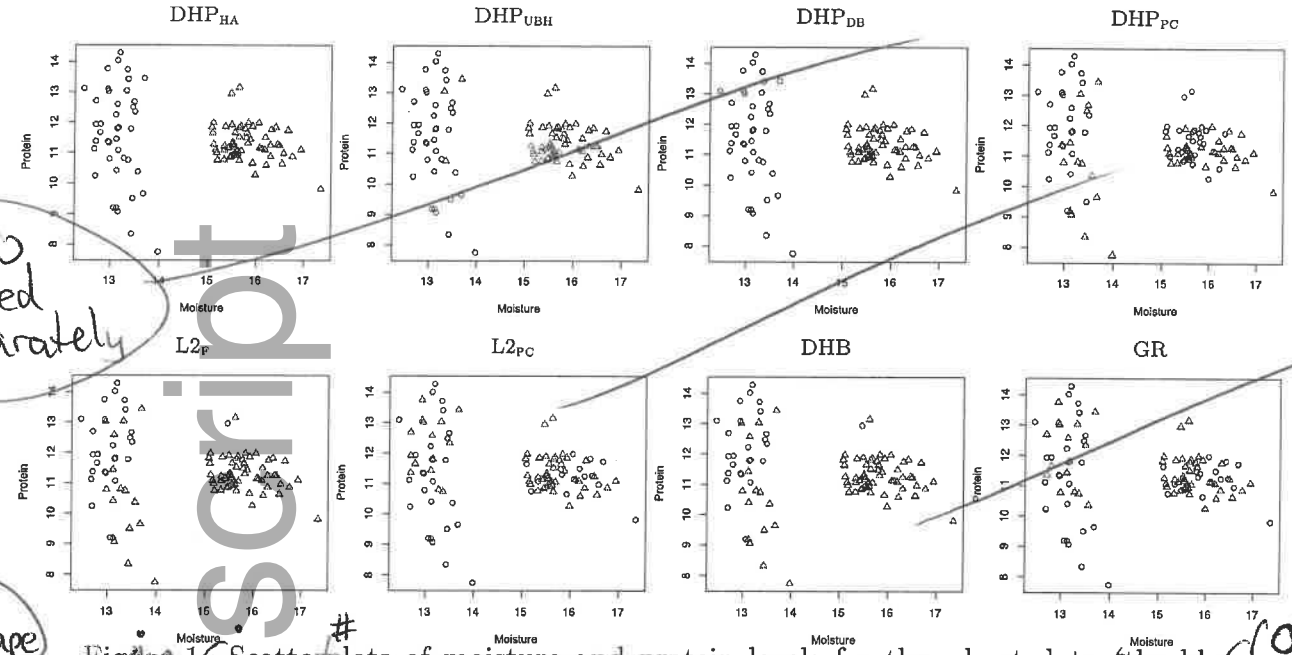


Figure 1: Scatterplots of moisture and protein levels for the wheat data (the blue circles correspond to one cluster and the red triangles to the other) for clusters created using, from left to right and top to bottom, DHP_{HA}, DHP_{UBH}, DHP_{DB}, DHP_{PC}, $L2_F$, $L2_{PC}$, DHB, GR, and (h)

higher than 15% were clustered in another group (no curve had a moisture level in the interval $[14, 15]$). These results are particularly interesting when considering properties of wheat. Specifically, wheat with low moisture level can be stored more safely than wheat with high moisture level, so that moisture level is often used as a quality factor for wheat. In Canada, wheat with moisture level lower than 14.5% can be sold as straight grade seeds; see Jayas et al. (1994), page 347, and `http://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/crop1204`.

The clusters created by $L2_F$ and DHB correspond roughly to high and low protein levels, whereas the other methods created clusters that do not seem to correspond to specific groups based on neither the protein nor the moisture levels.

For our second example we used the Australian rainfall data described in Delaigle and Hall (2010) and available at `http://rda.ucar.edu/datasets/ds482.1`. The data, depicted in Figure 9 in Appendix B.2, concern 191 rainfall curves $X_i(t)$ collected at Australian weather stations, where $X_i(t)$ represents the average rainfall at time $t \in$

This article is protected by copyright. All rights reserved

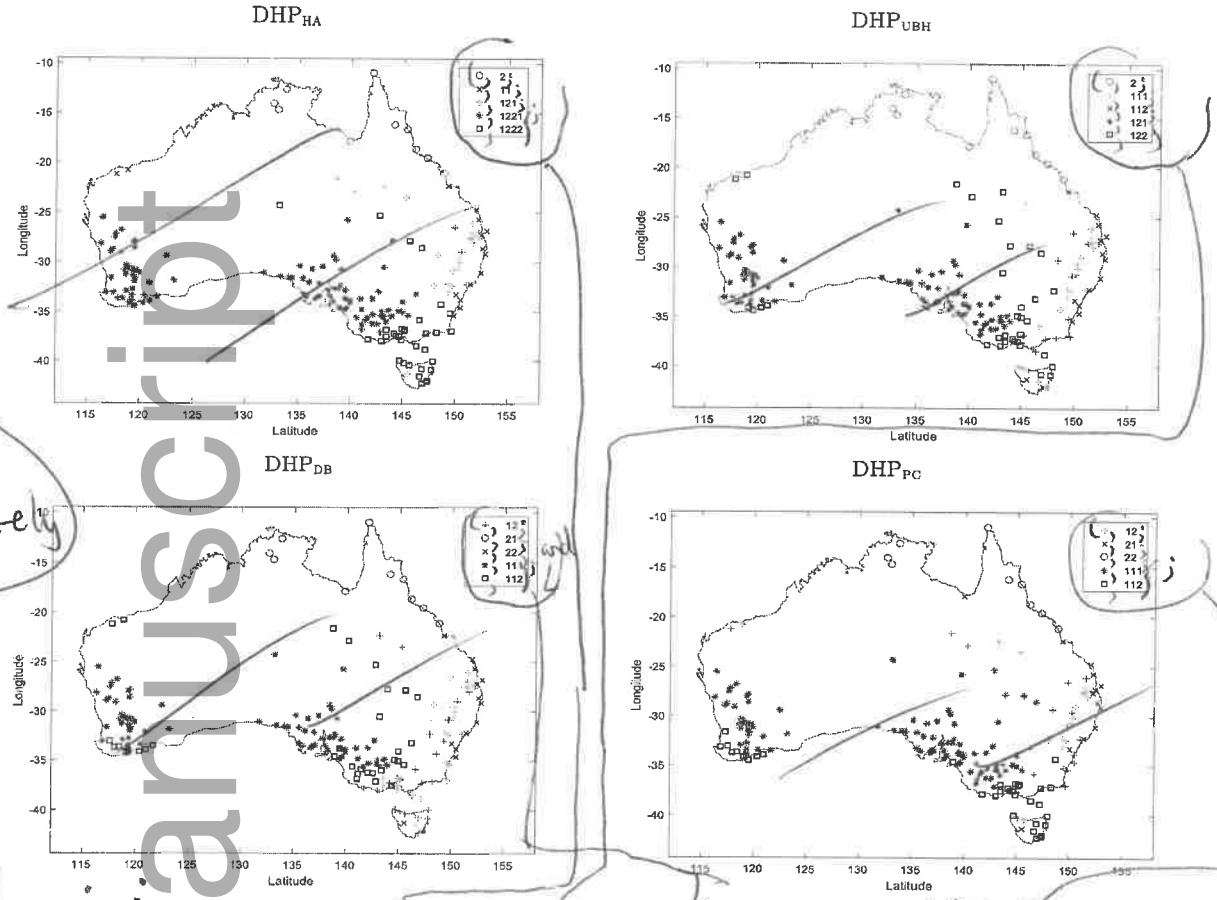


Figure 2. Clusters of rainfall stations in Australia for four methods. From left to right and top to bottom: $\mathrm{DHP}_{\mathrm{HA}}$, $\mathrm{DHP}_{\mathrm{UBH}}$, $\mathrm{DHP}_{\mathrm{DB}}$, $\mathrm{DHP}_{\mathrm{PC}}$. The numbers in the legend are the numbers $k_1, \ldots, k_\ell$ corresponding to the clusters $\chi_{k_1,\ldots,k_\ell}$ created by each method. The four symbols correspond to the four clusters created for $K = 4$.

[1, 365], taken over the years were the ith station has been operating, and smoothed using a local polynomial smoother.

In Delaigle and Hall (2010), these data were classified (manually by the authors) into northern and southern weather stations, depending on their location on the map of Australia. Here our goal is to cluster the data automatically into 2, 3, 4 or 5 groups using each of the eight methods used earlier, and see which clusters are created by each method (the extension of our method to more than two groups, using hierarchical clustering, is discussed in section 5.2).

This article is protected by copyright. All rights reserved

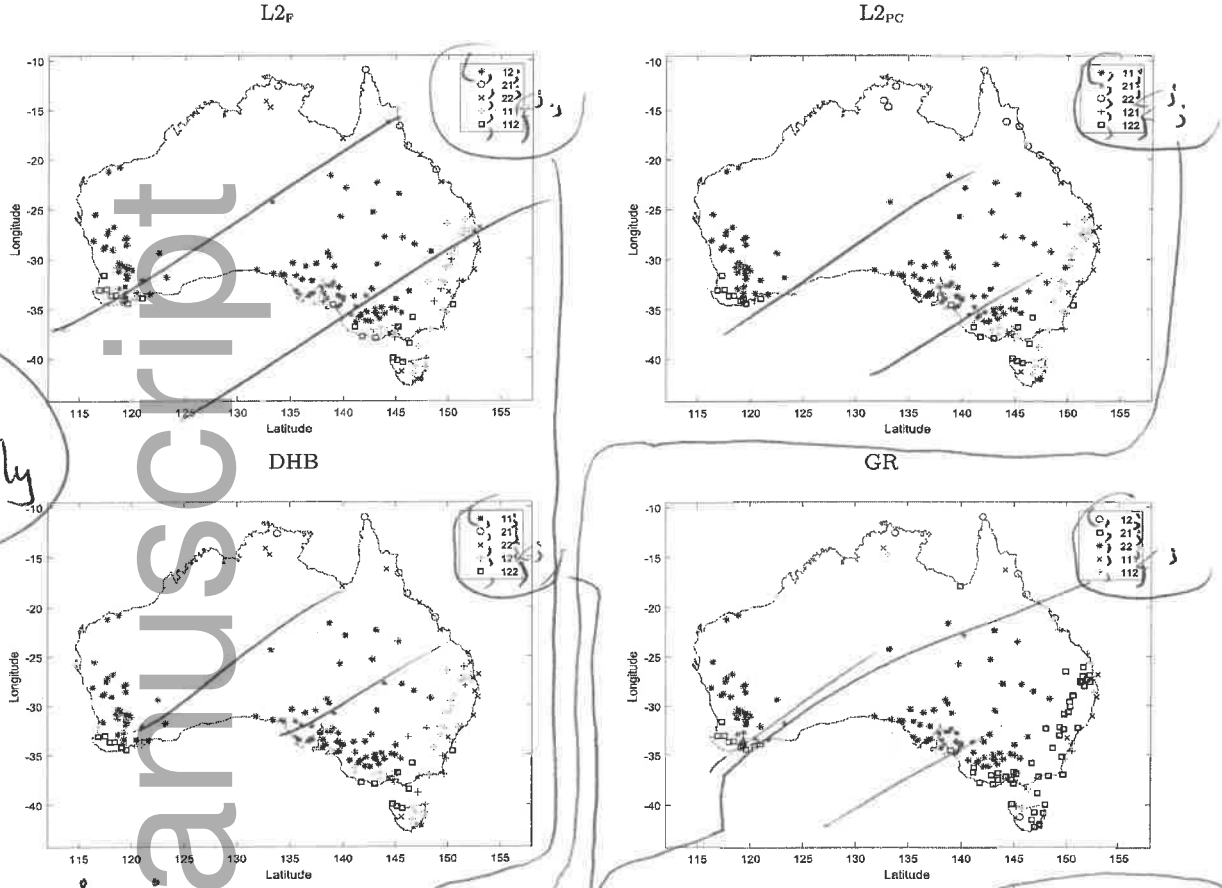


Figure 3: Clusters of rainfall stations in Australia for four methods. From left to right and top to bottom: $L2_F$; $L2_{PC}$; DHB; GR. The numbers in the legend are the numbers $k_1, \ldots, k_\ell$ corresponding to the clusters $\chi_{k_1,\ldots,k_\ell}$ created by each method. The four symbols correspond to the four clusters created for $K = 4$.

For each method, to create $K > 2$ clusters using hierarchical clustering, we proceed iteratively as follows. At the first step, we create 2 clusters. At step k, where we have created $k+1$ clusters, we apply again the clustering method to each of the $k+1$ clusters, which we divide in two clusters. Then, among the resulting $k+1$ possible cluster configurations of $k+2$ clusters, we keep the one that gives the smallest tightness where, for a given configuration of $k+2$ clusters $C_1, \ldots, C_{k+2}$, tightness is computed as $\sum_{\ell=1}^{k+2} \sum_{X_i \in C_\ell} \|X_i - \bar{X}_\ell\|^2$. For each method, we repeat this procedure until we have reached the maximum number (here 5) of clusters we wish to consider. We denote by $\chi_{k_1,k_2,\ldots,k_\ell}$ with $k_1, \ldots, k_\ell = 1, 2$, a cluster obtained by clustering group χ_{k_1} into

This article is protected by copyright. All rights reserved

two groups, $\chi_{k_1,1}$ and $\chi_{k_1,2}$, followed by clustering group χ_{k_1,k_2} into two groups,..., followed by clustering group $\chi_{k_1,k_2,\ldots,k_{\ell-1}}$ into two groups.

Figures 2 and 3 depict the clusters obtained in this way for all eight methods, displayed on a map of Australia. The five clusters are displayed in different colours; we also display the four clusters created for $K = 4$ by using a different symbol for each of those four clusters. We also display the values of $k_1, \ldots, k_\ell$ for each cluster, so that the clusters obtained for $K = 2$ and $K = 3$ can be deduced from these figures. For example, for our DHP_{HA} method, for $K = 5$ we obtained the clusters χ_2, $\chi_{1,1}$, $\chi_{1,2,1}$, $\chi_{1,2,2,1}$ and $\chi_{1,2,2,2}$. The clusters for $K = 4$ are χ_2, $\chi_{1,1}$, $\chi_{1,2,1}$ and $\chi_{1,2,2} = \chi_{1,2,2,1} \cup \chi_{1,2,2,2}$, those for $K = 3$ are χ_2, $\chi_{1,1}$ and $\chi_{1,2} = \chi_{1,2,1} \cup \chi_{1,2,2}$, and those for $K = 2$ are $\chi_1 = \chi_{1,2} \cup \chi_{1,1}$ and χ_2.

It can be seen that, overall, our method DHP_{HA}, and, a bit less clearly, DHP_{UBH}, DHP_{DB} and DHP_{PC}, clustered the stations into groups that are geographically separated. The other clustering methods produced clusters which seem less easy to interpret on a map of Australia, with stations from various geographical locations spread over several clusters, although there might be some other logical interpretation of the clusters created by those methods.

Our third example comes from Kalivas (1997). The dataset consists of spectra of 60 gasoline samples (see Figure 9 in Appendix B.2) and their octane values. We applied all eight clustering methods to these data and the clusters created by our method correspond, with a purity of 91%, to the gasoline samples that have octane level below 87 (group 1) and octane level greater or equal to 87 (group 2). These results are quite interesting because in the standard octane rating from the US government, gasoline with octane level below 87 is considered of not good enough quality for vehicle use; see `https://www.fueleconomy.gov/feg/octane.shtml`. The other methods created clusters which do not seem to be very connected to the octane level. See Figure 4, where we depict the clusters created by each method according to the octane level.

This article is protected by copyright. All rights reserved

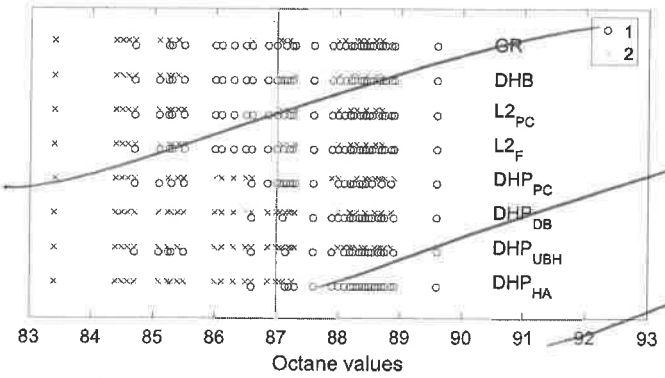


Figure 4: Two clusters of gasoline samples depicted according to their octane value, for eight clustering methods. For each method, the circles (resp., crosses) show the octane value of individuals clustered in group 1 (resp., group 2).

Finally, for our fourth illustration, we used the Berkeley growth data of Tuddenham and Snyder (1954), described in Ramsay and Silverman (2005). As indicated in the introduction, the notion of a perfect cluster is not uniquely defined, and in real examples there may be several insightful ways to cluster the data. In this example, there exists a natural way of clustering the data, which consist of height measurements of $n_1 = 39$ girls (population Π_1) and $n_2 = 54$ boys (population Π_2), taken at 31 time points t between age 1 and 18, which we turned into continuous curves $X_i(t)$, $t \in \mathcal{I} = [1, 18]$, via local polynomial smoothing. We used gender as a benchmark for the "true clusters", to measure the quality of the various clustering methods. However, this is merely for illustration purposes, as unlike for the simulated examples, we do not know what the "ideal clusters" should be nor whether these are even uniquely defined. The curves from each group, as well as the group means, are depicted in Figure 10 in Appendix B.2. The purities and adjusted rand indices obtained for each method are provided in Table 2. Considering boys and girls as the true clusters, our DHP_{HA} and DHP_{UBH} procedures worked particularly well; their performance was not nearly approached by any of the six competing procedures.

Overall, we conclude from our simulated and real data examples that our DHP_{HA} and DHP_{UBH} methods seem to provide clusters that are able to capture important

This article is protected by copyright. All rights reserved

Table 2: 100×purities and 100×adjusted rand indices (Arand) for eight clustering methods applied to the Berkeley data

	Methods							
	$\mathrm{DHP}_{\mathrm{HA}}$	$\mathrm{DHP}_{\mathrm{UBH}}$	$\mathrm{DHP}_{\mathrm{DB}}$	$\mathrm{DHP}_{\mathrm{PC}}$	$\mathrm{L2}_{\mathrm{F}}$	$\mathrm{L2}_{\mathrm{PC}}$	DHB	GR
Purities	90.3	92.5	74.2	86.0	64.5	64.5	60.2	64.5
Arand	64.7	71.9	22.5	51.4	7.42	7.42	3.07	7.42

differences between groups of individuals. In most of the cases we considered, the clusters they created were easier to interpret than those created by other methods.

5 Extensions

5.1 Heteroscedastic case

Asymptotically perfect clustering is possible in broader settings than those discussed earlier. However, in more general cases it is usually not possible to find an analytic formula for the functions ψ_j that maximise the ratio between the differences of the projected means and the standard deviations. This makes the results more implicit and less elegant than those in Theorem 2, but the lack of analytic formulae is irrelevant in practice since we choose the ψ_js, and the number of them, in a data-driven way.

In this section, we extend our main result to the case where the data come from two populations Π_1 and Π_2, and the data from the kth population can be written as

$$X_{ki} = \nu_k + Z_{ki}\,, \tag{5.1}$$

where the covariance function of Z_{ki} is equal to κ_k, where, κ_1 and κ_2 may differ, and, for a given k, the Z_{ki}s are identically distributed with mean zero. Let κ denote covariance function of the mixed centered population, and let $\kappa(u,v) = \sum_{j=1}^{\infty} \theta_j \phi_j(u)\phi_j(v)$ and $\kappa_k(u,v) = \sum_{j=1}^{\infty} \theta_{kj} \phi_{kj}(u)\phi_{kj}(v)$, $k = 1, 2$, denote the spectral decompositions of κ and κ_k. For $k = 1, 2$, write $\nu_k = \sum_{j=1}^{\infty} \nu_{kj}\phi_j$ and $Z_{ki} = \sum_{j=1}^{\infty} \theta_{kj}^{1/2} Z_{kij}\,\phi_{kj}$, where $\nu_{kj} = \int_{\mathcal{I}} \nu_k\,\phi_j$ and $Z_{kij} = \theta_{kj}^{-1/2} \int_{\mathcal{I}} Z_{ki}\,\phi_{kj}$.

For any integer $r > 0$, let $\psi^{(r)} = \sum_{j=1}^{r} \alpha_j \phi_j$. For $k = 1, 2$, we have $\nu_k(\psi^{(r)}) \equiv \mathbb{E}\{X_{ki}(\psi^{(r)})\} = \sum_{j=1}^{r} \nu_{kj}\alpha_j$ and $\sigma_k^2(\psi^{(r)}) \equiv \mathrm{Var}\{X_{ki}(\psi^{(r)})\} = \sum_{j=1}^{\infty} \theta_{kj} \left(\sum_{\ell=1}^{r} \alpha_\ell \delta_{k\ell j}\right)^2$,

This article is protected by copyright. All rights reserved

where we used the notation $\delta_{k\ell j} = \int_{\mathcal{I}} \phi_{kj}\phi_{\ell}$. The following result shows conditions on the α_j's under which asymptotically perfect clustering is possible.

Theorem 3. *Assume the heteroscedastic mixture model defined by* (5.1) *and* (3.13), *and let* $\psi^{(r)} = \sum_{j=1}^{r} \alpha_j \phi_j$. *Then if* $\sup_{k,j=1,2,\ldots} \mathbb{E}(Z_{kij}^4) < \infty$ *and*

$$\lim_{r\to\infty} |\nu_1(\psi^{(r)}) - \nu_2(\psi^{(r)})| / \max\left\{\sigma_1(\psi^{(r)}), \sigma_2(\psi^{(r)})\right\} = \infty\,, \tag{5.2}$$

$$\sup_{r\in\mathbb{N}} \max\left\{\frac{\sigma_1(\psi^{(r)})}{\sigma_2(\psi^{(r)})}, \frac{\sigma_2(\psi^{(r)})}{\sigma_1(\psi^{(r)})}\right\} < C < \infty \tag{5.3}$$

hold, the clustering algorithm is asymptotically perfect, in the sense that, if $\mathcal{R}_1$ *and* $\mathcal{R}_2$ *denote the regions found by minimizing* (3.2), *there exists a permutation* $(k(1), k(2))$ *of the pair* $(1, 2)$ *such that, for* $\ell = 1, 2$,

$$\lim_{r\to\infty} P\{X(\psi^{(r)}) \in \mathcal{R}_{k(\ell)} \mid I_\ell = 1\} = 1\,. \tag{5.4}$$

Condition (5.2) is similar to condition (3.17); it requires that the ratio between the mean difference and the standard deviation diverges to infinity. Condition (5.3) is a technical assumption. It assumes that no population is much more spread around its mean than the other.

5.2 More than two clusters

All the results we have presented so far assumed that the data could be clustered into $K = 2$ groups. In particular, note from Theorem 2 that the theoretical function $\psi^{(r)}$ used to project the data is a scaled squared distance between the means of the two populations, and it is not clear how to extend this concept of binary comparison to more than two populations. One exception is the case where we can reasonably assume that the data follow a binary hierarchical structure, that is, where we can apply hierarchical clustering. In that case, we can partition the data into $K > 2$ groups by sequentially applying our binary clustering procedure. As usual with hierarchical clustering, some caution is required when applying this technique as the weaknesses

Author Manuscript

This article is protected by copyright. All rights reserved

of that clustering approach carry over from the multivariate case to the functional case. See our discussion in section 6.

There, once the sample $\mathcal{X}$ has been split in two clusters $\mathcal{X}_1$ and $\mathcal{X}_2$ using the methodology introduced in the previous sections, our clustering procedure can be applied to the clusters $\mathcal{X}_k$, $k = 1, 2$, which are then partitioned into two clusters $\mathcal{X}_{k,1}$ and $\mathcal{X}_{k,2}$. This process can be iterated several times, i.e., applying again the same procedure, each cluster $\mathcal{X}_{k,j}$, $k, j = 1, 2$, can be further split into two clusters $\mathcal{X}_{k,j,1}$ and $\mathcal{X}_{k,j,2}$, and those new clusters can themselves be split in two clusters, etc. At each step, when a cluster is split in two, only the data from that cluster are used when applying our procedure.

As noted by Hastie et al. (2009), often the clusters created at each splitting level are shown, in the hope that some of those clusters can help the experimenter uncover interesting properties of their data. In the multivariate setting, there have been some attempts at developing procedures which can automatically detect which subclusters are "true clusters", but most of them are quite informal and require subjective choices of tuning parameters which do not work universally well. See Everitt et al. (2011) for a discussion on these and related issues. The difficulty of choosing which clusters are "true clusters" is also closely related to the fact that the notion of "true cluster" is not clearly determined in real applications, as noted in the ~~abstract~~ and ~~the introduction.~~

Since the theoretical properties of clusters obtained by repeatedly dividing clusters are similar, in this section we derive properties obtained only at the first splitting level, where the clusters $\mathcal{X}_1$ and $\mathcal{X}_2$ are each divided in two clusters. ~~That is,~~ we consider the case where the data come from four different populations Π_k, $k = 1, \ldots, 4$. For $k = 1, \ldots, 4$, we assume that the data from Π_k can be represented as at (3.12), where, to simplify our discussion, we assume that the Z_{ki}s have the same covariance; the heteroscedastic case can be treated as in section 5.1. Specifically we assume that

> For $k = 1, \ldots, 4$, the functions Z_{ki} are all distributed as the random function Z with $E(Z) = 0$, the location terms ν_k are fixed and the populations Π_k arise in respective proportions ρ_k, where $\sum_{k=1}^{4} \rho_k = 1$ and each $\rho_k > 0$. (5.5)

This article is protected by copyright. All rights reserved

Our goal in this section is to show that under appropriate regularity conditions, if hierarchical clustering is performed, at each split into two clusters, it is possible to find a direction such that, if the data are projected along that direction then at the first stage of the clustering algorithm, the clusters $\mathcal{X}_1$ and $\mathcal{X}_2$ asymptotically correspond, respectively, to $\Pi_1 \cup \Pi_2$ and $\Pi_3 \cup \Pi_4$; at the second stage of the clustering algorithm, $\mathcal{X}_1$ and $\mathcal{X}_2$ are split in two clusters $\mathcal{X}_{1,1}$ and $\mathcal{X}_{1,2}$, and $\mathcal{X}_{2,1}$ and $\mathcal{X}_{2,2}$, which asymptotically correspond, respectively, to Π_1 and Π_2, and Π_3 and Π_4.

The spectral decomposition of the covariance function κ of Z can be expressed as $\kappa(u,v) = \sum_{j=1}^{\infty} \theta_j \phi_j(u) \phi_j(v)$. Next, let $\nu_{12} = (\rho_1 \nu_1 + \rho_2 \nu_2)/(\rho_1 + \rho_2)$ and $\nu_{34} = (\rho_3 \nu_3 + \rho_4 \nu_4)/(\rho_3 + \rho_4)$, respectively denote the mean of the pooled Π_1 and Π_2, and Π_3 and Π_4. For $k = 1, \ldots, 4$ write $\nu_k = \sum_{j=1}^{\infty} \nu_{k,j} \phi_j$ and $Z_{ki} = \sum_{j=1}^{\infty} \theta_j^{1/2} Z_{kij} \phi_j$, where $\nu_{k,j} = \int_{\mathcal{I}} \nu_k \phi_j$ and $Z_{kij} = \theta_j^{-1/2} \int_{\mathcal{I}} Z_{ki} \phi_j$, and for $(k,\ell) = (1,2)$ and $(3,4)$, write $\nu_{k\ell} = \sum_{j=1}^{\infty} \nu_{k\ell,j} \phi_j$ where $\nu_{k\ell,j} = \int_{\mathcal{I}} \nu_{k\ell} \phi_j$ and let $\sigma^2_{[k,r]} = \sigma^2_{[\ell,r]} = \sum_{j=1}^{r} (\nu_{k,j} - \nu_{\ell,j})^2 \theta_j^{-1}$. Finally, for $(k,\ell) = (1,2)$ and $(3,4)$, let $\mu^{(r)}_{[k\ell]} = \sum_{j=1}^{r} \theta_j^{-1} (\nu_{k,j} - \nu_{\ell,j})(\nu_{12,j} - \nu_{34,j})$ and $\sigma^2_{[r]} = \sum_{j=1}^{r} (\nu_{12,j} - \nu_{34,j})^2 \theta_j^{-1}$.

The following theorem can be proved in the same way as Theorem 2. It shows one scenario for $K = 4$ where asymptotically perfect clustering is possible, as long as we recompute the function $\psi^{(r)}$ each time a cluster is split in two subclusters. The conditions can be satisfied in cases similar to the homoscedastic case where $K = 2$. See section 6 for a discussion about more general cases and weaknesses of the hierarchical approach.

Theorem 4. *Assume the model defined by (3.12) and (5.5) and that* $\kappa_4 := \mathbb{E}(Z_{k1}^4) = \ldots = \mathbb{E}(Z_{k4}^4) < \infty$.

(i) Let $\psi^{(r)} = \sum_{j=1}^{r} \theta_j^{-1} (\nu_{12,j} - \nu_{34,j}) \phi_j$. *If* $\sum_{j=1}^{\infty} (\nu_{12,j} - \nu_{34,j})^2 \theta_j^{-1} = \infty$ *holds and* $|\mu^{(r)}_{[12]}| + |\mu^{(r)}_{[34]}| < \min_{i=1,\ldots,4} \sqrt{(\rho_i/2)} \times C \sigma^2_{[r]}$ *for all* r, *where* C *is a constant such that* $C\{3 + \min_{i=1,\ldots,4} \sqrt{(\rho_i/2)}\} \leq 1$, *and if* $\mathcal{R}_1$ *and* $\mathcal{R}_2$ *denote the regions found by minimising (3.2), there exists a permutation* $(k(1), k(2))$ *of the pair* $(1,2)$ *such that, for* $\ell = 1, 3$, $\lim_{r \to \infty} P\{X(\psi^{(r)}) \in \mathcal{R}_{k(\ell)} \mid I_\ell = 1 \text{ or } I_{\ell+1} = 1\} = 1$.

This article is protected by copyright. All rights reserved

Table 3: 100×purities and 100×adjusted rand indices averaged over 100 samples generated from the hierarchical example, with group sizes $n_k = 30$, 50 and 100 and for eight clustering methods

n_k	DHP_{HA}	DHP_{UBH}	DHP_{DB}	DHP_{PC}	L2_{F}	L2_{PC}	DHB	GR
Purities								
30	77.7	81.2	71.6	60.8	43.0	42.4	45.5	42.4
50	83.1	84.9	74.8	65.7	42.0	42.4	45.0	41.5
100	92.5	90.0	75.5	72.4	40.9	41.2	44.1	40.6
Adjusted rand indices								
30	56.0	63.3	42.9	29.5	0.02	0.02	0.04	0.02
50	66.0	70.4	48.5	38.9	0.02	0.02	0.04	0.01
100	82.9	79.9	50.7	50.8	0.02	0.02	0.04	0.01

(a) *For* $(k,k') = (1,2)$ *and* $(k,k') = (3,4)$, *let* $\psi_{kk'}^{(r)} = \sum_{j=1}^{r} \theta_j^{-1}(\nu_{k,j} - \nu_{k',j})\phi_j$. *If* $\sum_{j=1}^{\infty}(\nu_{k,j} - \nu_{k',j})^2\theta_j^{-1} = \infty$ *holds, and if* $\mathcal{R}_{11}$ *and* $\mathcal{R}_{12}$ *(resp.* $\mathcal{R}_{21}$ *and* $\mathcal{R}_{22}$*) denote the regions found by minimising (3.2) using only the data clustered in* $\mathcal{R}_1$ *(resp.* $\mathcal{R}_2$*), there exists a permutation* $(k(1), k(2))$ *of the pair* $(1,2)$ *such that, for* $\ell = 1,2$, $\lim_{r\to\infty} P\{X(\psi_{12}^{(r)}) \in \mathcal{R}_{1k(\ell)} \mid I_\ell = 1\} = 1$, *and, for* $\ell = 3,4$, $\lim_{r\to\infty} P\{X(\psi_{34}^{(r)}) \in \mathcal{R}_{2k(\ell)} \mid I_\ell = 1\} = 1$.

Note that in Theorem 4, for simplicity, we assume that the Z_{ki}s have the same covariance function for all four populations. This assumption could be eased as in Theorem 3 as long as conditions similar to the one at (5.3) are assumed to ensure that one of the projected populations is not much more widespread than the other. Likewise, Theorem 4 deals only with a scenario where the populations are well clustered when projected using a single function but in many other cases, several projection functions would be needed to achieve good clustering performance.

To illustrate our procedure in this setting, we generated data from three populations. For $k = 1,2,3$ and $i = 1,\ldots,n_k$, the ith observation from the kth population was generated as $X_{ki}(t) = \sum_{j=1}^{40}(\theta_j^{1/2}Z_{kij} + \mu_{jk})\phi_j(t)$ for 128 equispaced points t in

This article is protected by copyright. All rights reserved

$\mathcal{I} = [0, 1]$; the Z_{kij}s were independent standard normal random variables and we took $\phi_j(t) = \sqrt{2}\sin(\pi j t)$, $\theta_j = j^{-3}$, $(\mu_{11}, \mu_{21}, \mu_{31}, \mu_{41}, \mu_{51}, \mu_{61}) = (0, 0.2, 0.1, -0.2, 0.08, 0.18)$, $(\mu_{12}, \mu_{22}, \mu_{32}, \mu_{42}, \mu_{52}, \mu_{62}) = (0.2, 0.1, -0.1, 0.2, -0.04, 0.16)$, $(\mu_{13}, \mu_{23}, \mu_{33}, \mu_{43}, \mu_{53}, \mu_{63}) = (-0.22, -0.31, 0.4, -0.08 - 0.02, -0.25)$ and $\mu_{jk} = 0$ for $j > 6$.

Then we applied hierarchical clustering with the eight methods introduced in section 4.2.1. Each time, we first created two clusters and then applied again binary clustering to both clusters. That is, we split cluster 1 in two clusters, thereby obtaining three clusters. Then, instead, we split cluster 2 in two clusters, thereby obtaining another configuration of three clusters. To choose which of the two configurations of three clusters was the most likely, we chose the one that minimised tightness $\sum_{\ell=1}^{3}\sum_{X_i \in C_\ell} \|X_i - \bar{X}_\ell\|^2$, where, for a given configuration, C_ℓ denotes the ℓth cluster. The purities and adjusted rand indices are presented in Table 3 for each method. These numbers indicate that on average our DHP_{HA} and DHP_{UBH} methods outperformed the other methods.

6 Discussion: more than two clusters

As discussed at the beginning of section 5.2, near perfect clustering performance is only possible when the data can be reasonably clustered by hierarchical clustering. In particular, as in the standard multivariate case, hierarchical clustering has difficulty handling groups of very different sizes or very large groups, which can be incorrectly split into parts that are merged with other clusters.

Theorem 4 considers one of the scenarios where our method is ensured to give good clustering performance in the $K = 4$ population context. Note that $\mu^{(r)}_{[k\ell]}$ is the difference between the means of populations Π_k and Π_ℓ projected via the function $\psi^{(r)}$, and $\sigma^2_{[r]}$ is equal to the difference between the means of the pooled Π_1 and Π_2 and the pooled Π_3 and Π_4 projected via the function $\psi^{(r)}$. Therefore, in Theorem 4, the condition $|\mu^{(r)}_{[12]}| + |\mu^{(r)}_{[34]}| < \min_{i=1,\ldots,4} \sqrt{(p_i/2)} \times C\sigma^2_{[r]}$ ensures that, after projection via $\psi^{(r)}$, the difference between the pooled means of Π_1 and Π_2, and of Π_3 and Π_4,

This article is protected by copyright. All rights reserved

is sufficiently large compared to the difference between the means of Π_1 and Π_2, and to that between the means of Π_3 and Π_4. Moreover, the condition $\sum_{j=1}^{\infty}(\nu_{12,j} - \nu_{34,j})^2\theta_j^{-1} = \infty$ means that together, populations Π_1 and Π_2 form a tight cluster around their pooled mean, which is infinitely apart from the cluster created by Π_3 and Π_4. This prevents situations where a projected population would overlap with other populations, in which case near perfect clustering would not be possible.

Other scenarios or values of K can be handled similarly, each time requiring that at any binary splitting step of the procedure, when projected via the function $\psi^{(r)}$ (whose formula depends on the scenario considered) the populations involved are separated into two clearly distinct clusters, each of which contains one or more entire populations. This also implies that several configurations of K clusters need to be investigated and compared ~~In order~~ to decide which ~~one~~ is the most suitable. Indeed, in cases where, at a given binary splitting step, the two new clusters contain an unequal number of populations, one of the two clusters will need to be split subsequently more often than the other. If, at any stage of the procedure, there is one population that does not belong very distinctly to one of the clusters only, then there are risks that this population will be split between two different clusters, a situation which is irreversible at later stages of the iterative procedure. Also, even if the entire populations can be aggregated into well distinct clusters, if the sample is such that the group sizes are very different, then the algorithm may fail.

Acknowledgements

Research supported by grants and fellowships from the Australian Research Council (DP170102434, FT130100098 and FL110100003). The Australian weather data we used in the paper were assembled by the Australian Bureau of Meteorology. They are available from the Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory `http://rda.ucar.edu/datasets/ds482.1`. Bob Dattore is acknowledged for providing the data. The wheat and the octane data are available in Shang and Hyndman's (2014) `fds` R package. The Berkeley growth data are available in Ramsay et al.'s (2014) R `fda` package. We thank the editor, the associate editor and two reviewers for their helpful comments which helped significantly improve the paper.

This article is protected by copyright. All rights reserved

References

Abraham, C., Cornillon, P.A., Matzner-Løber, E. and Molinari, N. (2003) Unsupervised curve clustering using B-splines. *Scand. J. Stat.* **30**, 581–595.

Antoniadis, A., Brossat, X., Cugliari, J. and Poggi, J. M. (2013) Clustering functional data using wavelets. *Int. J. Wavelets Multi.* **11**, 1350003.

Auder, B. and Fischer, A. (2012) Projection-based curve clustering. *J. Stat. Comput. Sim.* **82**, 1145–1168.

Brusco, M. J. and Steinley, D. (2007) A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. *Psychometrika* **72**, 583–600.

Chiou, J.M. and Li, P.L. (2007) Functional clustering and identifying substructures of longitudinal data. *J. Roy. Statist. Soc. Ser. B* **69**, 679–699.

Delaigle, A. and Hall, P. (2010) Defining probability density for a distribution of random functions. *Ann. Stat.* **38**, 1171–1193.

Delaigle, A. and Hall, P. (2012) Achieving near perfect classification for functional data. *J. Roy. Statist. Soc. Ser. B* **74**, 267–286.

Delaigle, A., Hall, P. and Bathia, N. (2012) Componentwise classification and clustering of functional data. *Biometrika* **99**, 299–313.

Du, Q., Faber, V. and Gunzburger, M. (1999) Centroidal Voronoi tessellations: applications and algorithms. *SIAM review* **41**, 637–676.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*, 5th ed. John Wiley & Sons, Ltd., New York.

Ferraty, F. and Vieu, P. (2006) *Nonparametric functional data analysis.* Springer Series in Statistics. Springer, New York.

Fryzlewicz, P. (2007) Unbalanced Haar technique for nonparametric function estimation. *J. Am. Stat. Assoc.* **102**, 1318–1327.

Gattone, S.A. and Rocci, R. (2012) Clustering curves on a reduced subspace. *J. Comput. Graph. Stat.* **21**, 361–379.

Hartigan, J.A. (1975) *Clustering algorithms.* Wiley, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction*, 2nd Ed. Springer Verlag.

Hasegawa, S., Imai, H., Inaba, M., Katoh, N. and Nakano, J. (1993) Efficient algorithms for variance-based k-clustering. in *Proceedings of the First Pacific Conference on Computer Graphics and Applications*, World Scientific, River Edge, NJ, pp. 75–89.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (2012) *Wavelets, approximation, and statistical applications.* Lecture Notes in Statistics, 129, Springer.

Hennig, C. (2015) What are the true clusters? *Pattern Recognition Letters* **64**, 53–62.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classif.* 2, 193–218.

This article is protected by copyright. All rights reserved

Jayas, D.S., White, N.D. and Muir, W.E. (1994) Stored-grain ecosystems. CRC Press.

Kalivas, J.H. (1997) Two data sets of near infrared spectra. *Chemometr. Intell. Lab.* **37**, 255–259.

Lloyd, S.P. (1957) Least square quantization in PCM. *Technical Report* Bell Telephone Laboratories.

Lloyd, S.P. (1982) Least squares quantization in PCM. *IEEE T. Inform. Theory* **28**, 129–137.

Manning, C.D., Raghavan, P. and Schütze, H., (2008) *Introduction to Information Retrieval.* Cambridge University Press.

Pollard, D. (1981) Strong consistency of k-means clustering. *Ann. Stat.* **9**, 135–140.

Ramsay, J.O. and Silverman, B.W. (2005) *Functional data analysis*, second edn. Springer, New York.

Ramsay, J.O. Wickham, H., Graves, S. and Hooker, G. (2014) fda: Functional data analysis. R package version 2.4.3. `http://CRAN.R-project.org/package=fda`.

Schreiber, T. (1998) A Voronoi diagram based adaptive k-means-type clustering algorithm for multidimensional weighted data. In *Computational Geometry Methods, Algorithms, and Applications* (Eds H. Bieri and H. Noltemeier), pp. 265–275. Springer, New York.

Serban, N. and Wasserman, L. (2005) CATS: clustering after transformation and smoothing. *J. Am. Stat. Assoc.* **100**, 990–999.

Shang, H.L. and Hyndman, R.J. (2013) fds: Functional data sets. R package version 1.7. `http://CRAN.R-project.org/package=fds`.

Tarpey, T. and Kinateder, K.J. (2003) Clustering functional data. *J. Classif.* **20**, 93–114.

Telgarski, M. and Vattani, A. (2010) Hartigan's method: k-means clustering without Voronoi. *J. Mach. Learn. Res.* **9**, 820–827.

Tuddenham, R.D. and Snyder, M.M. (1954) Physical growth of California boys and girls from birth to age 18. *University of California Publications in Child Development* **1**, 183–364.

A Technical arguments

A.1 Proof of Theorem 1

Let

$$S_{kj}(\vec{\psi}) = \frac{1}{n_k} \sum_{i=1}^{n_k} (1 - E) \left\{ X_{ki}(\psi_j) - \mu_{kj}(\vec{\psi}) \right\}^2,$$

This article is protected by copyright. All rights reserved

and observe that

$$\frac{1}{n_k}\sum_{i=1}^{n_k}\{X_{ki}(\psi_j)-\bar{X}_k(\psi_j)\}^2$$
$$=\frac{1}{n_k}\sum_{i=1}^{n_k}\{X_{ki}(\psi_j)-\mu_{kj}(\vec{\psi})\}^2-\{\bar{X}_k(\psi_j)-\mu_{kj}(\vec{\psi})\}^2.$$

Therefore, by (2.4),

$$\widehat{T}_2(\mathcal{X}_1,\mathcal{X}_2\mid\vec{\psi})=\sum_{j=1}^{p}\frac{1}{\hat{\sigma}(\psi_j)^2}\,\frac{1}{n}\sum_{k=1}^{2}n_k\Big[S_{kj}(\vec{\psi})+E\{X_{k1}(\psi_j)-\mu_{kj}(\vec{\psi})\}^2$$
$$-\{\bar{X}_k(\psi_j)-\mu_{kj}(\vec{\psi})\}^2\Big].$$

Note that, by (3.2),

$$t_2(\vec{\psi}\mid\vec{\mathcal{R}})=\sum_{j=1}^{p}\frac{1}{\sigma(\psi_j)^2}\sum_{k=1}^{2}\pi(\vec{\psi},\mathcal{R}_k)\,E\{X_{k1}(\psi_j)-\mu_{kj}(\vec{\psi})\}^2,$$

where $X_{k1}(\psi_j)$ denotes a random variable with the distribution of $X(\psi_j)$ conditional on $X(\vec{\psi})\in\mathcal{R}_k$. Hence,

$$\Big|\widehat{T}_2(\mathcal{X}_1,\mathcal{X}_2\mid\vec{\psi})-t_2(\vec{\psi}\mid\vec{\mathcal{R}})\Big|$$
$$\le\sum_{j=1}^{p}\frac{1}{\hat{\sigma}(\psi_j)^2}\,\frac{1}{n}\sum_{k=1}^{2}n_k\Big[\big|S_{kj}(\vec{\psi})\big|+\{\bar{X}_k(\psi_j)-\mu_{kj}(\vec{\psi})\}^2\Big]$$
$$+\sum_{j=1}^{p}\Big|\frac{1}{\hat{\sigma}(\psi_j)^2}-\frac{1}{\sigma(\psi_j)^2}\Big|\sum_{k=1}^{2}\frac{n_k}{n}\,E\{X_{k1}(\psi_j)-\mu_{kj}(\vec{\psi})\}^2$$
$$+\sum_{j=1}^{p}\frac{1}{\sigma(\psi_j)^2}\sum_{k=1}^{2}\Big|\frac{n_k}{n}-\pi(\vec{\psi},\mathcal{R}_k)\Big|\,E\{X_{k1}(\psi_j)-\mu_{kj}(\vec{\psi})\}^2. \qquad (A.1)$$

Since $E\|X\|^2\le 1$ then, in view of (3.5),

$$E\{X_{k1}(\psi_j)-\mu_{kj}(\vec{\psi})\}^2\le E\{X_{k1}(\psi_j)\}^2$$
$$=\frac{1}{\pi(\vec{\psi},\mathcal{R}_k)}\,E\Big[X(\psi_j)^2\,I\{X(\vec{\psi})\in\mathcal{R}_k\}\Big]$$
$$\le\frac{1}{\pi(\vec{\psi},\mathcal{R}_k)}\,E\Big[\|X\|^2\|\psi_j\|^2\,I\{X(\vec{\psi})\in\mathcal{R}_k\}\Big]$$

This article is protected by copyright. All rights reserved

$$\leq \frac{1}{\pi(\vec{\psi}, \mathcal{R}_k)} E\|X\|^2 \leq \pi_{\min}^{-1} .$$

This property and (A.1) imply that

$$\begin{aligned} &\left| \widehat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi}) - t_2(\vec{\psi} \mid \vec{\mathcal{R}}) \right| \\ &\quad \leq \sum_{j=1}^{p} \left\{ \left| \frac{1}{\hat{\sigma}(\psi_j)^2} - \frac{1}{\sigma(\psi_j)^2} \right| + \frac{1}{\sigma(\psi_j)^2} \right\} \\ &\qquad \times \frac{1}{n} \sum_{k=1}^{2} n_k \left[\left| S_{kj}(\vec{\psi}) \right| + \left\{ \bar{X}_k(\psi_j) - \mu_{kj}(\vec{\psi}) \right\}^2 \right] \\ &\qquad + \sum_{j=1}^{p} \left| \frac{1}{\hat{\sigma}(\psi_j)^2} - \frac{1}{\sigma(\psi_j)^2} \right| \pi_{\min}^{-1} \\ &\qquad + \sum_{j=1}^{p} \frac{\pi_{\min}^{-1}}{\sigma(\psi_j)^2} \sum_{k=1}^{2} \left| \frac{n_k}{n} - \pi(\vec{\psi}, \mathcal{R}_k) \right| . \end{aligned} \tag{A.2}$$

The fourth series in (A.2) is relatively easy to bound, as follows. First, note that for the definition of $\mathcal{R}_k$ given below (3.2),

> the random integers n_1 and n_2, where n_k denotes the number of functions X_i, for $1 \leq i \leq n$, that satisfy $X_i(\vec{\psi}) \in \mathcal{R}_k$, jointly have a multinomial distribution with parameters n and π_1 and π_2, where $\pi_k = \pi(\vec{\psi}, \mathcal{R}_k)$. (A.3)

Result (A.3), assumption (3.9) and large deviation properties of the binomial distribution (Hoeffding's inequality) imply that, for all $\epsilon > 0$, there exist constants $C_1 = C_1(\epsilon) > 0$ and $C_2 = C_2(\epsilon) > 0$ such that

$$\begin{aligned} &\max_{k=1,2} P\left\{ \left| n^{-1} n_k - \pi(\vec{\psi}, \mathcal{R}_k) \right| > \sigma(\psi_j)^2 \, \epsilon \right\} \\ &\qquad \leq 2 \exp\left\{ - C_1 \, n \, \sigma(\psi_j)^4 \right\} \leq 2 \exp\left(- C_2 \, n^c \right) , \end{aligned}$$

uniformly in $\psi_1, \ldots, \psi_p \in Q_n$ and $\vec{\mathcal{R}} \in \mathbb{V}_{p,n}$. Therefore,

$$P\left\{ \sum_{j=1}^{p} \frac{\pi_{\min}^{-1}}{\sigma(\psi_j)^2} \sum_{k=1}^{2} \left| \frac{n_k}{n} - \pi(\vec{\psi}, \mathcal{R}_k) \right| > \epsilon \right\} = O\left[\exp\left\{ - C(\epsilon) \, n^c \right\} \right] , \tag{A.4}$$

uniformly in the same sense, where we interpret $C(\epsilon)$ as a generic constant.

This article is protected by copyright. All rights reserved

Next we bound the first factor in the third series on the right-hand side of (A.2). Note that, since $E\|X\|^2 = 1$ and $\|\psi\| = 1$ for $\psi \in Q_n$, we have $\sigma(\psi)^2 \le 1$. Therefore, if $\psi \in Q_n$ and $\epsilon = 4\,\delta^2$, where $0 < \delta \le \frac{1}{2}$, then

$$
\begin{aligned}
P\{|\hat\sigma(\psi)^{-2} - \sigma(\psi)^{-2}| > \epsilon\} &= P\Big\{|\hat\sigma(\psi)^2 - \sigma(\psi)^2| > \epsilon\,\hat\sigma(\psi)^2\,\sigma(\psi)^2\Big\} \\
&\le P\Big[|\hat\sigma(\psi)^2 - \sigma(\psi)^2| > \epsilon\,\{\sigma(\psi)^4 - \sigma(\psi)^2\,|\hat\sigma(\psi)^2 - \sigma(\psi)^2|\}\Big] \\
&= P\Big\{|\hat\sigma(\psi)^2 - \sigma(\psi)^2| > \frac{\epsilon\,\sigma(\psi)^4}{1+\epsilon\,\sigma(\psi)^2}\Big\} \le P\big\{|\hat\sigma(\psi)^2 - \sigma(\psi)^2| > \tfrac{1}{2}\,\epsilon\,\sigma(\psi)^4\big\} \\
&\le P\left\{\left|\frac{1}{n}\sum_{i=1}^n (1-E)\left(\int_{\mathcal{I}} \psi\,X_i\right)^2\right| > \delta^2\,\sigma(\psi)^4\right\} \\
&\qquad + P\left\{\left|\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{I}} \psi\,X_i\right| > \delta\,\sigma(\psi)^2\right\} \\
&= O\big[\exp\{-C(\epsilon)\,n^c\}\big], \qquad \text{(A.5)}
\end{aligned}
$$

where the identity holds uniformly in $\psi \in Q_n$ and $\vec{\mathcal{R}} \in \mathbb{V}_{p,n}$ and follows from (3.7) and (3.8).

Observe also that

$$
\begin{aligned}
&P\{|S_{kj}(\vec\psi)| > 3\pi_{\min}^{-1}\epsilon\} \\
&\quad = P\left[\left|\frac{1}{n_k}\sum_{i=1}^{n_k}(1-E)\left\{X_{ki}(\psi_j)^2 - 2\,\mu_{kj}(\vec\psi)\,X_{ki}(\psi_j) + \mu_{kj}(\vec\psi)^2\right\}\right| > \frac{3\epsilon}{\pi_{\min}}\right] \\
&\quad \le P\left\{\left|\frac{1}{n_k}\sum_{i=1}^{n_k}(1-E)\,X_{ki}(\psi_j)^2\right| > \frac{\epsilon}{\pi_{\min}}\right\} \\
&\qquad + P\left\{\left|\frac{\mu_{kj}(\vec\psi)}{n_k}\sum_{i=1}^{n_k}(1-E)\,X_{ki}(\psi_j)\right| > \frac{\epsilon}{\pi_{\min}}\right\}. \qquad \text{(A.6)}
\end{aligned}
$$

To bound the second probability on the far right-hand side of (A.6) we observe that

$$
\begin{aligned}
|\mu_{kj}(\vec\psi)| &= \frac{1}{\pi(\vec\psi,\mathcal{R}_k)}\Big|E\big[X(\psi_j)\,I\{X(\vec\psi)\in\mathcal{R}_k\}\big]\Big| \\
&\le \frac{1}{\pi(\vec\psi,\mathcal{R}_k)}E\big[|X(\psi_j)|\,I\{X(\vec\psi)\in\mathcal{R}_k\}\big]
\end{aligned}
$$

Author Manuscript

This article is protected by copyright. All rights reserved

$$\leq \frac{1}{\pi(\vec{\psi}, \mathcal{R}_k)} \Big\{ E\Big[X(\psi_j)^2 \, I\{X(\vec{\psi}) \in \mathcal{R}_k\} \Big] \Big\}^{1/2}$$
$$\leq \frac{1}{\pi(\vec{\psi}, \mathcal{R}_k)} \Big\{ E\Big[\|X\|^2 \|\psi_j\|^2 \, I\{X(\vec{\psi}) \in \mathcal{R}_k\} \Big] \Big\}^{1/2}$$
$$\leq \frac{1}{\pi(\vec{\psi}, \mathcal{R}_k)} \big(E\|X\|^2\big)^{1/2} \leq \pi_{\min}^{-1} ,$$

where we used (3.5), and the fact that $\|\psi_j\| = 1$ for $\psi_j \in Q_n$, as prescribed by (3.3), and $E\|X\|^2 = 1$. Therefore, for any $\epsilon > 0$,

$$P\left\{ \left| \frac{\mu_{kj}(\vec{\psi})}{n_k} \sum_{i=1}^{n_k} (1 - E)\, X_{ki}(\psi_j) \right| > \frac{\epsilon}{\pi_{\min}} \right\} \leq P\left\{ \left| \frac{1}{n_k} \sum_{i=1}^{n_k} (1 - E)\, X_{ki}(\psi_j) \right| > \epsilon \right\}, \tag{A.7}$$

and more simply,

$$P\Big\{ \big| \bar{X}_k(\psi_j) - \mu_{kj}(\vec{\psi}) \big|^2 > \epsilon^2 \Big\} = P\left\{ \left| \frac{1}{n_k} \sum_{i=1}^{n_k} (1 - E)\, X_{ki}(\psi_j) \right| > \epsilon \right\}. \tag{A.8}$$

Replacing ϵ by $\epsilon\, \sigma(\psi_j)^2$ in (A.6) and (A.7), and by $\epsilon\, \sigma(\psi_j)$ in (A.8), we deduce from (3.6) and (A.6)–(A.8) that

$$P\Big[\big| S_{kj}(\vec{\psi}) \big| + \big\{ \bar{X}_k(\psi_j) - \mu_{kj}(\vec{\psi}) \big\}^2 > 3\pi_{\min}^{-1} \epsilon \sigma(\psi_j)^2 + \epsilon^2 \sigma(\psi_j)^2 \Big]$$
$$\leq \sum_{\ell=1}^{2} (3 - \ell)\, P\left\{ \left| \frac{1}{n_k} \sum_{i=1}^{n_k} (1 - E)\, X_{ki}(\psi_j)^\ell \right| > \epsilon\, \sigma(\psi_j)^2 \right\}$$
$$= O\big[\exp\{ - C(\epsilon)\, n^c \} \big] , \tag{A.9}$$

uniformly in $\psi_1, \ldots, \psi_p \in Q_n$ and $\vec{\mathcal{R}} \in \mathbb{V}_{p,n}$. (Note that $\sigma(\psi_j) \leq 1$.)

Recall from (A.3) that n_k, a random variable, has the binomial $\mathrm{Bi}\{n, \pi(\vec{\psi}, \mathcal{R}_k)\}$ distribution. From this property, using (3.5) and Hoeffding's inequality, it can be shown that

$$\max_{k=1,2} P\big(|n_k - \pi_k\, n| > \tfrac{1}{2}\, \pi_k\, n \big) \leq \exp(-C_3\, n) ,$$

where $C_3 > 0$.

Combining this property (A.2), (A.4), (A.5) and (A.9) we deduce that, for each $\epsilon > 0$,

$$P\Big\{ \Big| \hat{T}_2(\mathcal{X}_1, \mathcal{X}_2 \mid \vec{\psi}) - t_2(\vec{\psi} \mid \vec{\mathcal{R}}) \Big| > \epsilon \Big\} = O\big[\exp\{ - C(\epsilon)\, n^c \} \big] , \tag{A.10}$$

This article is protected by copyright. All rights reserved

uniformly in $\psi_1, \ldots, \psi_p \in Q_n$ and $\vec{\mathcal{R}} \in \mathbb{V}_{p,n}$. Result (3.11) follows from (3.3), (3.5), (3.6) and (A.10).

A.2 Proof of Theorem 2

For each j, let $\tilde{\nu}_j = \nu_{1j} - \nu_{2j}$, and without loss of generality, assume that $\nu_2 = 0$. To simplify the notation, we do not explicitly indicate the dependence of ψ on r, so that we let $\psi = \psi^{(r)}$. Let $\mu_2 = \mathbb{E}\{X_{2i}(\psi)\} = 0$ and $\mu_1 = \mathbb{E}\{X_{1i}(\psi)\}$. We deduce from the calculations preceding the theorem that $\mu_1 = \sum_{j=1}^{r} \theta_j^{-1} \nu_{1j}^2 = \sum_{j=1}^{r} \theta_j^{-1} \tilde{\nu}_j^2$ and $\mathrm{var}\{X_{2i}(\psi)\} = \mathrm{var}\{X_{1i}(\psi)\} = \sum_{j=1}^{r} \theta_j^{-1} \tilde{\nu}_j^2 = \mu_1$.

Letting f denote the density function of $X_{2i}(\psi)$, the density function of $X_{1i}(\psi)$ is equal to $f(x - \mu_1)$ and the density function of $X(\psi)$ can be written as $g(x) = \rho_1 f(x - \mu_1) + \rho_2 f(x)$. When $p = 1$ with $\psi = \psi^{(r)}$, it can be proved that finding the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ minimising (3.2) is equivalent to finding

$$(\bar{x}_1, \bar{x}_2) = \arg\min_{t_1, t_2} \int \min_{i=1,2} \|x - t_i\|^2 g(x)\, dx, \tag{A.11}$$

and then taking $\mathcal{R}_1$ and $\mathcal{R}_2$ to be determined by $(-\infty, (\bar{x}_1 + \bar{x}_2)/2]$ and $((\bar{x}_1 + \bar{x}_2)/2, \infty)$. Without loss of generality, $\bar{x}_1 < \bar{x}_2$, so that $\mathcal{R}_1 = (-\infty, (\bar{x}_1 + \bar{x}_2)/2]$ and $\mathcal{R}_2 = ((\bar{x}_1 + \bar{x}_2)/2, \infty)$. For $i = 1, 2$, let $\mathbb{E}_i$ and $\mathbb{P}_i$ denote the expectation and the probability conditionally on being in the ith population, respectively.

We start by deriving two results which will be useful in the sequel. First, we derive an upper bound for the minimum value of the objective function on the right-hand side of (A.11), as follows:

$$\begin{aligned}
\int \min_{i=1,2} \|x - \bar{x}_i\|^2 g(x)\, dx &\le \int \min_{i=1,2} \|x - \mu_i\|^2 g(x)\, dx \\
&= \int \min_{i=1,2} \|x - \mu_i\|^2 \rho_1 f(x - \mu_1)\, dx + \int \min_{i=1,2} \|x - \mu_i\|^2 \rho_2 f(x)\, dx \\
&\le \int \|x - \mu_1\|^2 \rho_1 f(x - \mu_1)\, dx + \int \|x - \mu_2\|^2 \rho_2 f(x)\, dx \\
&= \rho_1 \int \|x - \mu_1\|^2 f(x - \mu_1)\, dx + \rho_2 \int \|x\|^2 f(x)\, dx
\end{aligned}$$

This article is protected by copyright. All rights reserved

$$= \int \|x\|^2 f(x)\, dx$$
$$= \mu_1. \tag{A.12}$$

Second, recalling that $\mathbb{E}(Z_{2ij}^4) = \kappa_4$, we note that

$$\mathbb{E}\Big\{X_{2i}(\psi)^4\Big\} = \mathbb{E}\Big(\sum_{j=1}^r \frac{\tilde{\nu}_j}{\theta_j^{1/2}} Z_{2ij}\Big)^4 = \sum_{j=1}^r \frac{\tilde{\nu}_j^4}{\theta_j^2}\mathbb{E}\Big(Z_{2ij}^4\Big) + 3\sum_{j\neq k=1}^r \frac{\tilde{\nu}_j^2\tilde{\nu}_k^2}{\theta_j\theta_k}\mathbb{E}\Big(Z_{2ij}^2\Big)\mathbb{E}\Big(Z_{2ik}^2\Big)$$
$$= \sum_{j=1}^r \frac{\tilde{\nu}_j^4}{\theta_j^2}\kappa_4 + 3\sum_{j\neq k=1}^r \frac{\tilde{\nu}_j^2\tilde{\nu}_k^2}{\theta_j\theta_k} \leq 3\kappa_4\Big(\sum_{j=1}^r \frac{\tilde{\nu}_j^4}{\theta_j^2} + \sum_{j\neq k=1}^r \frac{\tilde{\nu}_j^2\tilde{\nu}_k^2}{\theta_j\theta_k}\Big)$$
$$= 3\kappa_4\mu_1^2.$$

In the next paragraphs, we construct balls A_1 and A_2 which are such that all the individuals in those balls are correctly clustered as long as r is large enough. Then, we prove that as $r \to \infty$, the probability that these two balls contain the entire population tends to 1. Together, these two results prove the theorem. Note that since we project the data on a space of dimension $p = 1$, the balls are in fact intervals, but we use the terminology "balls" because it makes it simpler to refer to the center and the radius of the balls.

For $j = 1, 2$, we let A_j denote the ball of center μ_j with radius $c_1\mu_1$, where c_1 is a positive constant that will be determined later. Since $\mu_2 = 0$, we have, for $j = 1, 2$,

$$\int_{A_j} \|x - \mu_j\|^2 f(x - \mu_j)\, dx = \int_{A_2} \|x\|^2 f(x)\, dx = \mathbb{E}_2\Big\{\|X_2(\psi) - \mu_2\|^2 1_{A_2}\Big\}$$
$$= \mu_1 - \mathbb{E}_2\Big\{\|X_2(\psi) - \mu_2\|^2 1_{A_2^c}\Big\}$$
$$\geq \mu_1 - \Big\{\mathbb{E}_2\|X_2(\psi) - \mu_2\|^4\Big\}^{1/2}\Big\{\mathbb{E}_2 1_{A_2^c}\Big\}^{1/2}$$
$$\geq \mu_1 - 2\sqrt{\kappa_4}\mu_1 \Big\{\mathbb{P}_2\Big\{X_2(\psi) \in A_2^c\Big\}\Big\}^{1/2} \tag{A.13}$$
$$\geq \mu_1 - 2\sqrt{\kappa_4}\mu_1 \Big\{\frac{\mathbb{E}_2\|X_2(\psi) - \mu_2\|^2}{c_1^2\mu_1^2}\Big\}^{1/2}$$
$$\geq \mu_1 - \frac{2\sqrt{\kappa_4}}{c_1}\mu_1^{1/2}. \tag{A.14}$$

Let $k \geq 2$ be a positive integer and for $j = 1, 2$, let $A_{j,k}$ be the ball of center μ_j with radius $kc_1\mu_1$. Assume that there exists $j \in \{1, 2\}$ such that neither $\bar{x}_1$ nor $\bar{x}_2$

Author Manuscript

This article is protected by copyright. All rights reserved

belong to the ball $A_{j,k}$. For $x \in A_j$, $\|x - \mu_j\| \leq c_1\mu_1$ and for any $i \in \{1,2\}$,

$$\|x - \bar{x}_i\| \geq \|\bar{x}_i - \mu_j\| - \|x - \mu_j\| \geq kc_1\mu_1 - c_1\mu_1 = (k-1)c_1\mu_1 \geq (k-1)\|x - \mu_j\| .$$

Therefore, using (A.19)

$$\begin{aligned}\int \min_{i=1,2} \|x - \bar{x}_i\|^2 g(x)\, dx &\geq \int_{A_j} \min_{i=1,2} \|x - \bar{x}_i\|^2 g(x)\, dx \geq \int_{A_j} (k-1)^2 \|x - \mu_j\|^2 g(x)\, dx \\ &\geq \int_{A_j} (k-1)^2 \|x - \mu_j\|^2 \rho_j f(x - \mu_j)\, dx \\ &\geq (k-1)^2 \rho_j \left(\mu_1 - \frac{2\sqrt{\kappa_4}}{c_1}\mu_1^{1/2}\right).\end{aligned}$$

Choose k such that $(k-1)^2 \min_{i=1,2} \rho_i > 2$ and $c_1 = (5k)^{-1}$. Then the balls $A_{j,k}$ are disjoint and when r is sufficiently large so that $\sqrt{\mu_1} > 20k\sqrt{\kappa_4}$,

$$\begin{aligned}\int \min_{i=1,2} \|x - \bar{x}_i\|^2 g(x)\, dx &\geq (k-1)^2 \rho_j \left(\mu_1 - \frac{2\sqrt{\kappa_4}}{c_1}\mu_1^{1/2}\right) \geq 2\left(\mu_1 - 10k\sqrt{\kappa_4}\mu_1^{1/2}\right) \\ &> 2\left(\mu_1 - \frac{1}{2}\mu_1^{1/2}\mu_1^{1/2}\right) > \mu_1 ,\end{aligned}$$

which contradicts (A.12). Thus, each ball $A_{j,k}$ must contain one $\bar{x}_i$ and since the balls are disjoint, they must contain exactly one $\bar{x}_i$.

Without loss of generality, assume that $\bar{x}_i \in A_{i,k}$ for $i = 1, 2$. Then $\|\bar{x}_i - \mu_i\| \leq kc_1\mu_1$. Let B_i be the ball centered at $\bar{x}_i$ with radius $(k+1)c_1\mu_1$. Then B_i contains A_i. Now, for any $x \in B_1$, we have

$$\begin{aligned}\|x - \bar{x}_2\| &\geq \|x - \mu_2\| - \|\mu_2 - \bar{x}_2\| \geq \|x - \mu_2\| - kc_1\mu_1 \geq \|\mu_1 - \mu_2\| - \|x - \mu_1\| - kc_1\mu_1 \\ &\geq \|\mu_1 - \mu_2\| - \|x - \bar{x}_1\| - \|\bar{x}_1 - \mu_1\| - kc_1\mu_1 \geq \mu_1 - (3k+1)c_1\mu_1 \\ &\geq (k+1)c_1\mu_1 \geq \|x - \bar{x}_1\| .\end{aligned}$$

Thus, if $x \in B_1$, then x belongs to $\mathcal{R}_1$. The same argument applies to the case $x \in B_2$. Hence, for $i = 1, 2$, B_i belongs to the $\mathcal{R}_i$. Since B_i contains A_i, all the points in A_i are correctly clustered. Thus, all the points of population Π_i which are wrongly clustered, must lie in the set A_i^c. Hence, the error rate of clustering, which is equal to

$$\rho_2 P\{X(\psi^{(r)}) \in \mathcal{R}_1 \,|\, I_2 = 1\} + \rho_1 P\{X(\psi^{(r)}) \in \mathcal{R}_2 \,|\, I_1 = 1\}$$

Author Manuscript

This article is protected by copyright. All rights reserved

is bounded by

$$\sum_{i=1}^{2} \mathbb{P}_i(A_i^c) \leq 2\frac{\mathbb{E}_2\|X_2(\psi) - \mu_2\|^2}{c_1^2\mu_1^2} = \frac{2}{c_1^2\mu_1},$$

which tends to zero as $r \to \infty$.

A.3 Proof of Theorem 3

For each j, let $\tilde{\nu}_j = \nu_{1j} - \nu_{2j}$ and without loss of generality, assume that $\nu_2 = 0$. To simplify the notation, we do not explicitly indicate the dependence of ψ on r, so that we let $\psi = \psi^{(r)}$. Let $\mu_2 = \mathbb{E}\{X_{2i}(\psi)\} = 0$ and $\mu_1 = \mathbb{E}\{X_{1i}(\psi)\}$. Recall from the calculations before the theorem that $\mu_1 = \sum_{j=1}^{r} \alpha_j\nu_{1j} = \sum_{j=1}^{r} \alpha_j\tilde{\nu}_j$ and $\sigma_k^2(\psi) = \mathrm{Var}\{X_{ki}(\psi)\} = \sum_{j=1}^{\infty} \theta_{kj}\left(\sum_{i=1}^{r} \alpha_i\delta_{kij}\right)^2$. For simplicity, write $\sigma_k^2(\psi)$ as σ_k^2.

Letting f_2 denote the density function of $X_{2i}(\psi)$, and $f_1(x - \mu_1)$ be the density function of $X_{1i}(\psi)$, the density function of $X(\psi)$ can be written as $g(x) = \rho_1 f_1(x - \mu_1) + \rho_2 f_2(x)$. As in the homoscedastic case, finding the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ minimising (3.2) is equivalent to finding

$$(\bar{x}_1, \bar{x}_2) = \arg\min_{t_1,t_2} \int \min_{i=1,2} \|x - t_i\|^2 g(x)\, dx, \tag{A.15}$$

and then taking $\mathcal{R}_1$ and $\mathcal{R}_2$ to be determined by $(-\infty, (\bar{x}_1 + \bar{x}_2)/2]$ and $((\bar{x}_1 + \bar{x}_2)/2, \infty)$. Without loss of generality, $\bar{x}_1 < \bar{x}_2$, so that $\mathcal{R}_1 = (-\infty, (\bar{x}_1 + \bar{x}_2)/2]$ and $\mathcal{R}_2 = ((\bar{x}_1 + \bar{x}_2)/2, \infty)$. For $i = 1, 2$, let $\mathbb{E}_i$ and $\mathbb{P}_i$ denote the expectation and the probability conditional on being in the ith population respectively.

Similarly to the homoscedastic case, we have

$$\begin{aligned}\int \min_{i=1,2} \|x - \bar{x}_i\|^2 g(x)\, dx &\leq \rho_1 \int \|x - \mu_1\|^2 f_1(x - \mu_1)\, dx + \rho_2 \int \|x\|^2 f_2(x)\, dx \\ &= \rho_1\sigma_1^2 + \rho_2\sigma_2^2. \end{aligned} \tag{A.16}$$

Second, recalling that $\mathbb{E}(Z_{kij}^4) \leq \kappa_4$ and letting $\rho_{kj} = \theta_{kj}^{1/2}\sum_{i=1}^{r} \alpha_i\delta_{kij}$, we have, for $\ell = 1, 2$,

$$\mathbb{E}\left[\{X_{\ell i}(\psi) - \mu_\ell(\psi)\}^4\right] = \mathbb{E}\left[\left\{\sum_{j=1}^{\infty} Z_{\ell ij}\theta_{\ell j}^{1/2}\left(\sum_{i=1}^{r} \alpha_i\delta_{\ell ij}\right)\right\}^4\right] = \mathbb{E}\left(\sum_{j=1}^{\infty} Z_{\ell ij}\rho_{\ell j}\right)^4$$

This article is protected by copyright. All rights reserved

$$= \sum_{j=1}^{\infty} \rho_{\ell j}^4 \mathbb{E}\big(Z_{\ell ij}^4\big) + 3 \sum_{j \neq k=1}^{\infty} \rho_{\ell j}^2 \rho_{\ell k}^2 \mathbb{E}\big(Z_{\ell ij}^2\big) \mathbb{E}\big(Z_{\ell ik}^2\big)$$

$$\leq 3\kappa_4 \Big(\sum_{j=1}^{\infty} \rho_{\ell j}^4 + \sum_{j \neq k=1}^{\infty} \rho_{\ell j}^2 \rho_{\ell k}^2 \Big)$$

$$= 3\kappa_4 \Big(\sum_{j=1}^{\infty} \rho_{\ell j}^2 \Big)^2 = 3\kappa_4 \sigma_\ell^4 .$$

In the next paragraphs, we construct balls A_1 and A_2 which are such that all the individuals in those balls are correctly clustered as long as r is large enough. Then, we prove that, as $r \to \infty$, the probability that these two balls contain the entire projected population tends to 1. Together, these two results prove the theorem.

For $j = 1, 2$, we let A_ℓ denote the ball of center μ_ℓ with radius $c_1\mu_1$, where c_1 is a positive constant that will be determined later. We have, for $\ell = 1, 2$,

$$\int_{A_\ell} \|x - \mu_\ell\|^2 f_\ell(x - \mu_\ell)\, dx = \mathbb{E}_\ell \Big\{ \|X_\ell(\psi) - \mu_\ell\|^2 1_{A_\ell} \Big\}$$

$$= \sigma_\ell^2 - \mathbb{E}_\ell \Big\{ \|X_\ell(\psi) - \mu_\ell\|^2 1_{A_\ell^c} \Big\}$$

$$\geq \sigma_\ell^2 - \Big\{ \mathbb{E}_\ell \|X_\ell(\psi) - \mu_\ell\|^4 \Big\}^{1/2} \Big(\mathbb{E}_\ell 1_{A_\ell^c} \Big)^{1/2}$$

$$\geq \sigma_\ell^2 - 2\sqrt{\kappa_4}\sigma_\ell^2 \times \mathbb{P}_\ell \big\{ X_\ell(\psi) \in A_\ell^c \big\}^{1/2} \tag{A.17}$$

$$\geq \sigma_\ell^2 - 2\sqrt{\kappa_4}\sigma_\ell^2 \times \Big\{ \frac{\mathbb{E}_\ell \|X_\ell(\psi) - \mu_\ell\|^2}{c_1^2 \mu_1^2} \Big\}^{1/2}$$

$$\geq \sigma_\ell^2 - \frac{2\sqrt{\kappa_4}}{c_1} \times \frac{\sigma_\ell^3}{\mu_1} \tag{A.18}$$

$$= \sigma_\ell^2 \Big(1 - \frac{2\sqrt{\kappa_4}\sigma_\ell}{c_1 \mu_1} \Big) . \tag{A.19}$$

Let $k \geq 2$ be a positive integer and for $\ell = 1, 2$, let $A_{\ell,k}$ be the ball of center μ_ℓ with radius $kc_1\mu_1$. Assume that there exists $\ell \in \{1, 2\}$ such that neither $\bar{x}_1$ nor $\bar{x}_2$ belong to the ball $A_{\ell,k}$. For $x \in A_\ell$, $\|x - \mu_\ell\| \leq c_1\mu_1$ and for any $i \in \{1, 2\}$, as in the homoscedastic case we have $\|x - \bar{x}_i\| \geq (k-1)\|x - \mu_\ell\|$. Therefore, as in the homoscedastic case and using (A.19)

$$\int \min_{i=1,2} \|x - \bar{x}_i\|^2 g(x)\, dx \geq \int_{A_\ell} (k-1)^2 \|x - \mu_\ell\|^2 \rho_\ell f_\ell(x - \mu_\ell)\, dx$$

Author Manuscript

This article is protected by copyright. All rights reserved

$$\geq (k-1)^2 \rho_\ell \sigma_\ell^2 \left\{ 1 - \frac{2\sqrt{\kappa_4}\sigma_\ell}{c_1 \mu_1} \right\}.$$

Choose k such that $(k-1)^2 \min_{i=1,2} \rho_i > 2C^2$ and $c_1 = (5k)^{-1}$. Then the balls $A_{j,k}$ are disjoint and when r is sufficiently large so that $\mu_1 > 20k\sqrt{\kappa_4}\sigma_\ell$,

$$\int \min_{i=1,2} \|x - \bar{x}_i\|^2 g(x)\, dx \geq (k-1)^2 \rho_\ell \sigma_\ell^2 \left\{ 1 - \frac{2\sqrt{\kappa_4}\sigma_\ell}{c_1 \mu_1} \right\} \geq 2C^2 \sigma_\ell^2 \left\{ 1 - \frac{10k\sqrt{\kappa_4}\sigma_\ell}{\mu_1} \right\}$$
$$> C^2 \sigma_\ell^2 > \max\left\{\sigma_1^2, \sigma_2^2\right\} \geq \rho_1 \sigma_1^2 + \rho_2 \sigma_2^2,$$

which contradicts (A.16). Thus, each ball $A_{j,k}$ must contain one $\bar{x}_i$ and since the balls are disjoint, they must contain exactly one $\bar{x}_i$.

The rest of the proof is similar to the homoscedastic case, leading to the error rate of clustering, which is bounded by

$$\sum_{i=1}^{2} \mathbb{P}_i\left(A_i^c\right) \leq \sum_{i=1}^{2} \frac{\mathbb{E}_i \|X_i(\psi) - \mu_i\|^2}{c_1^2 \mu_i^2} = \frac{\sigma_1^2 + \sigma_2^2}{c_1^2 \mu_1^2},$$

which tends to zero as $r \to \infty$.

A.4 Illustration of the conditions used in Theorem 1

To illustrate condition (3.10), let Q_n denote the set of functions $\psi \in L_2(\mathcal{I})$ having the form $\psi^{(r)}$ at (2.7) where each $d_j \in \mathcal{D}_n$ and $\mathcal{D}_n$ contains no more than n^{C_1} elements, for some $C_1 > 0$, and including zero. Then we can take a_n, at (3.3), to be given by $a_n = n^{C_1 r} = \exp\{C_1 r \log(n)\}$. Unless a function $\psi \in L_2(\mathcal{I})$ has a particularly slowly converging expansion in terms of the basis functions χ_j, the construction in terms of a polynomial grid given in this paragraph ensures that ψ is approximated, in an L_2 sense, at a polynomial rate by functions in Q_n. This can often be extended to approximations in the supremum metric, if the suprema of the absolute values of the functions χ_j and their first derivative diverge at no faster than a polynomial rate in j.

Let $\mathbb{V}_{p,n}$ be a set of $b_n = n^{C_2}$ scaled centroidal Voronoi tessellations, where $C_2 > 0$ is a constant. Then, the left-hand side of (3.10) is bounded above by

$$\exp\{p\, C_1 r \log(n)\} n^{C_2} \exp\left(-C n^c\right),$$

This article is protected by copyright. All rights reserved

which converges to zero if $r = o\{n^c / \log(n)\}$. The latter condition is therefore sufficient to ensure (3.10). Condition (3.5) can be ensured by discarding tessellations for which at least one region $\mathcal{R}_k$ fails to contain at least a given fixed proportion of points in the dataset.

To illustrate conditions (3.6)–(3.9), take for example X to be a Gaussian process, or a mixture of Gaussian processes, such that

$$\inf_{\psi \in Q_n} \sigma(\psi)^2 \geq n^{-c_1},$$

where $0 < c_1 < \frac{1}{8}$. Then (3.6)–(3.9) hold with $c = \frac{1}{2} - 4c_1$. (Equation (3.8) is the determining factor here.)

B Additional numerical results

B.1 Details for computing the $d_{j,i}$s used in section 4.1

Recall from section 4.1 that, for $j = 1, \ldots, p$, we compute the jth projection function ψ_j as $\psi_j = \sum_{i=1}^{r} d_{j,i}\, \chi_i$, where χ_i denotes the ith element of the Haar basis. Recall that, for each $j \neq j'$, we want the functions ψ_j and $\psi_{j'}$ to be orthonormal. Therefore, we need to choose the coefficients $d_{j,i}$ and $d_{j',i'}$ so that

$$\sum_{i,i'=1}^{r} d_{j,i} d_{j',i'} \int_{\mathcal{I}} \chi_i(t) \chi_{i'}(t)\, dt = 0. \tag{B.1}$$

Let $\mathbf{d}_j = (d_{j,1}, \ldots, d_{j,r})^\top$. Since the χ_is are orthonormal, (B.1) is equivalent to $\mathbf{d}_j^\top \mathbf{d}_{j'} = \mathbb{1}\{j = j'\}$. In other words, the $\mathbf{d}_j$s need to be orthonormal vectors in $\mathbb{R}^r$. Searching for all possible combinations of orthonormal vectors in $\mathbb{R}^r$ would be too computationally intensive, and so instead we suggest using a greedy algorithm, which iteratively updates the components of the vectors $\mathbf{d}_j$ two by two, through consecutive rotations and reflections while maintaining orthonormality. The following notations will be useful. Recall that $\mathbf{d}_j \in \mathbb{R}^r$, and note that each of the r dimensions corresponds to an axis. For $k = 0, \ldots, 90$, let $\alpha_k = 2\pi k/180$ and, for two of the r axis,

This article is protected by copyright. All rights reserved

say i and i', let $\mathcal{R}_{\alpha_k,i,i'}$ denote the rotation of angle α_k on the plane determined by the ith and i'th axes. Finally, let $\mathcal{R}^{i,i'}$ denote the reflection about the ith axis in the plane determined by the ith and i'th axes. To compute the $\mathbf{d}_j$s:

Step 1: Start with p arbitrarily chosen orthonormal vectors $\mathbf{d}_j = (d_{j,1}, \ldots, d_{j,r})^\top$, $j = 1, \ldots, p$.

Step 2: For $i = 1, \ldots, r-1$

for $i' = i+1, \ldots, r$

For $j = 1, \ldots, p$ and $k = 1, \ldots, 90$, let $(\delta_{j,1}^{k,0}, \delta_{j,2}^{k,0})$ be the pair obtained by applying to $(d_{j,i}, d_{j,i'})$ the rotation $\mathcal{R}_{\alpha_k,i,i'}$, and let $(\delta_{j,1}^{k,1}, \delta_{j,2}^{k,1})$ be the pair obtained by applying $\mathcal{R}^{i,i'}$ to $(\delta_{j,1}^{k,0}, \delta_{j,2}^{k,0})$. Let

$$\Delta = \{(\delta_{j,1}^{k,t}, \delta_{j,2}^{k,t})_{j=1}^p, 1 \leq k \leq 90, t = 0, 1\}$$

denote the set of all possible p-pairs created by these rotations and reflections. Update the value of $(d_{j,i}, d_{j,i'})_{j=1}^p$ by taking

$$(d_{j,i}, d_{j,i'})_{j=1}^p = \text{argmin}_{(a_{j0},a_{j1})_{j=1}^p \in \Delta} \hat{T}_2(\vec{\psi}_a),$$

where $\vec{\psi}_a = (\psi_{a,1}, \ldots, \psi_{a,p})^\top$, and, for $j = 1, \ldots, p$, $\psi_{a,j} = \sum_{i=1}^r d_{j,i}^a \chi_i$ with $(d_{j,i}^a, d_{j,i'}^a) = (a_{j0}, a_{j1})$ and, for $\ell \notin \{i, i'\}$, $d_{j,\ell}^a = d_{j,\ell}$.

Step 3: Repeat step 2 once.

B.2 Graphs of simulated and real data

In Figures 5 to 7, we depict a random sample of 50 curves from Π_1 and of 50 curves from Π_2, as well as their empirical means, where the curves are drawn from respectively models (i) to (iii) introduced in section 4.2.1. Figure 8 shows a random sample of 50 curves from Π_1, 50 curves from Π_2 and 50 curves from Π_3, as well as their empirical means, where the curves are drawn from the hierarchical example introduced in section 5.2.

In Figure 9, we depict all the curves from the wheat, rainfall and octane data. In

This article is protected by copyright. All rights reserved

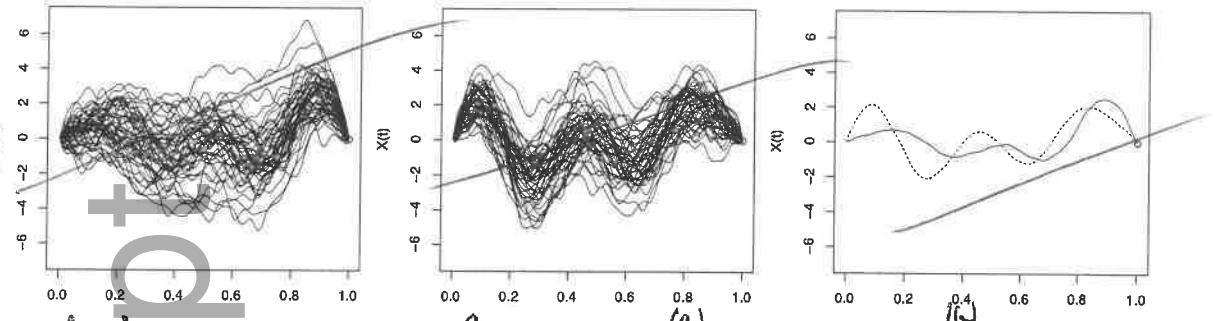


Figure 5. 50 curves from model (i) coming from population Π_1 (left) or Π_2 (middle), and empirical mean curves from the two groups (right).

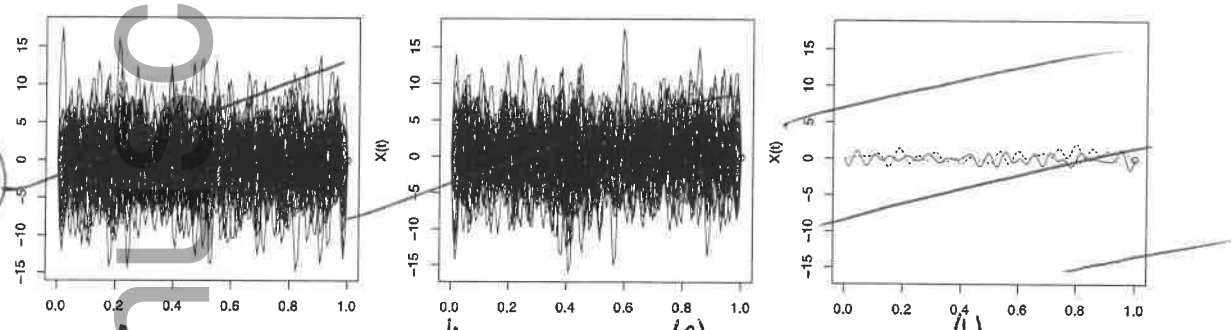


Figure 6. 50 curves from model (ii) coming from population Π_1 (left) or Π_2 (middle), and empirical mean curves from the two groups (right).

In Figure 10 we display the curves from Π_1 and the curves from Π_2, as well as their empirical means, for the Berkeley growth data.

This article is protected by copyright. All rights reserved

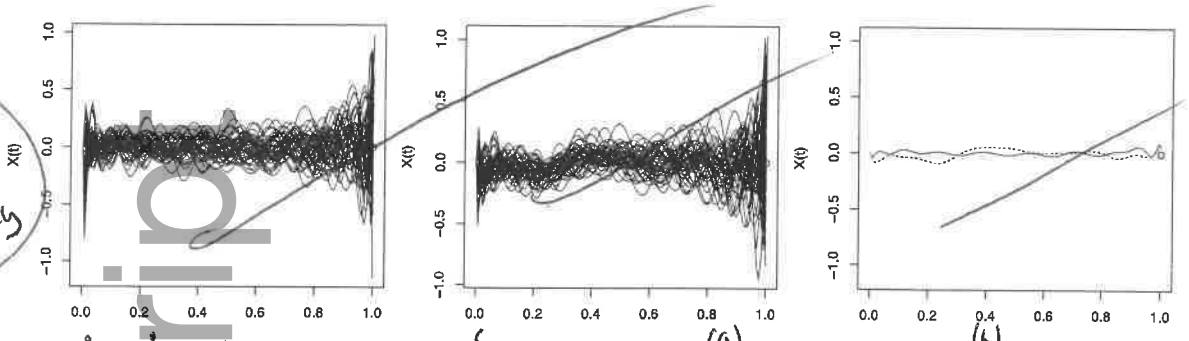


Figure 7: 50 curves from model (iii) coming from population Π_1 (left) or Π_2 (middle), and empirical mean curves from the two groups (right).

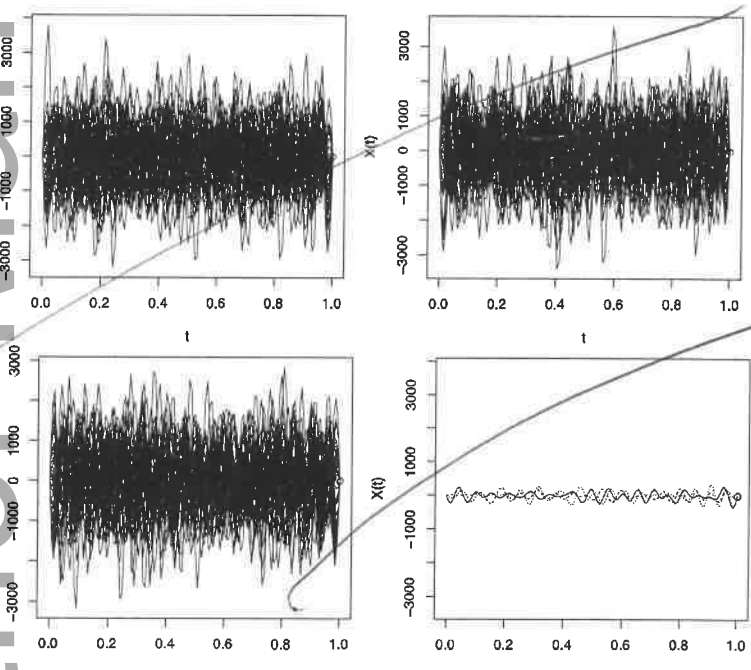


Figure 8: 50 curves from the hierarchical example coming from population Π_1 (top left), Π_2 (top right), Π_3 (bottom left), and empirical mean curves from the three groups (bottom right).

This article is protected by copyright. All rights reserved

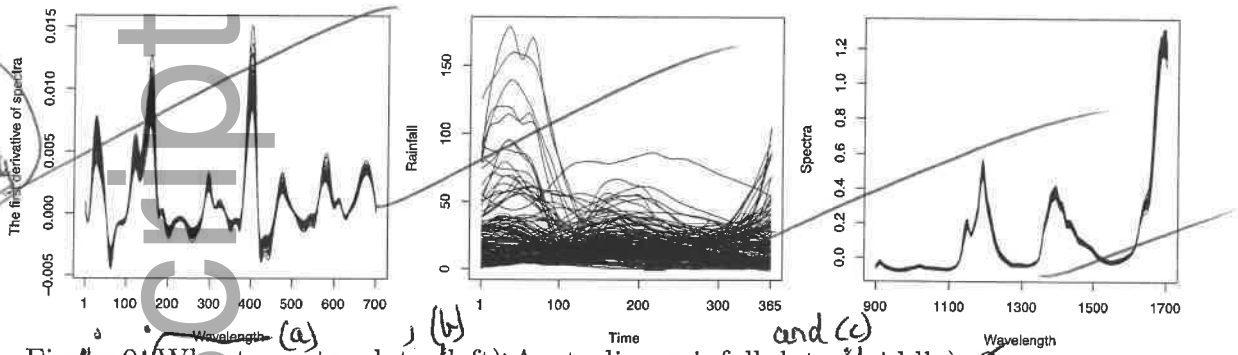


Figure 9: Wheat spectra data (left), Australian rainfall data (middle), Octane spectra curves (right).

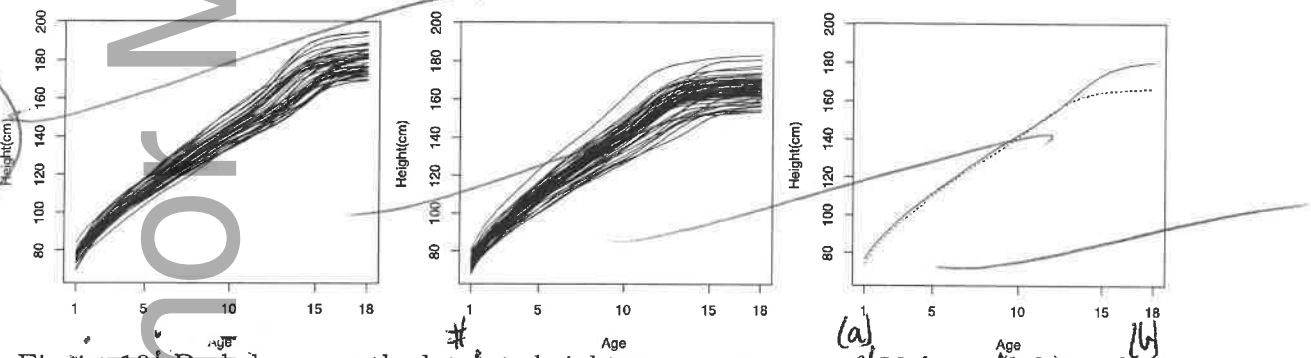


Figure 10: Berkeley growth dataset: height measurements of 39 boys (left) and 54 girls (middle) from age 1 to 18, and mean curves from each group (right).

Author Manuscript

This article is protected by copyright. All rights reserved