UNIVERSITY OF CALIFORNIA SAN DIEGO

Clustering and Mixture Modeling: Some Methodology and Theory

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

He Jiang

Committee in charge:

Professor Ery Arias-Castro, Chair
Professor Alex Cloninger
Professor Sanjoy Dasgupta
Professor Ronghui Xu
Professor Wenxin Zhou

2022

The Dissertation of He Jiang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to sincerely thank Professor Ery Arias-Castro for being my advisor and providing me with enormous support during my entire graduate school career. This dissertation would be impossible without his professional guidance and kind support.

I would also like to thank Professor Alex Cloninger, Professor Sanjoy Dasgupta, Professor Ronghui Xu, and Professor Wenxin Zhou for being on my committee and providing me with valuable feedback.

Chapter 2, in full, is a version of the paper "K-Means and Gaussian Mixture Modeling with a Separation Constraint", He Jiang and Ery Arias-Castro. It is currently being prepared for submission for publication of the material. The dissertation author is the primary investigator and corresponding author of this material.

Chapter 3, in full, is a version of the paper "Extending the Patra-Sen Approach to Estimating the Background Component in a Two-Component Mixture Model", Ery Arias-Castro and He Jiang. It is currently being prepared for submission for publication of the material. The dissertation author is the primary investigator and corresponding author of this material.

Chapter 4, in full, is a version of the paper "Fitting a Multi-modal Density by Dynamic Programming", He Jiang and Ery Arias-Castro. It is currently being prepared for submission for publication of the material. The dissertation author is the primary investigator and corresponding author of this material.

Chapter 5, in full, is a reprint of the paper "On the Consistency of Metric and Non-Metric K-Medoids", Ery Arias-Castro and He Jiang. It has been accepted for publication in the International Conference on Artificial Intelligence and Statistics 2021. The dissertation author is the primary investigator and corresponding author of this material.

VITA

2017	Bachelor of Science, University of California San Diego
2022	Doctor of Philosophy, University of California San Diego
2015–2022	Teaching Assistant, Department of Mathematics University of California San Diego
2016–2022	Research Assistant, Department of Mathematics University of California San Diego

FIELDS OF STUDY

Major Field: Statistics (Specialization in Clustering and Mixture Modeling)

ABSTRACT OF THE DISSERTATION

Clustering and Mixture Modeling: Some Methodology and Theory

by

He Jiang

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2022

Professor Ery Arias-Castro, Chair

In this dissertation, we propose several methodology in clustering and mixture modeling when the user has some prior knowledge regarding one or more of the clusters or mixture components. We also provide theory in the consistency of K-medoids, and prove its consistency under two versions of ordinal information.

We begin by considering the problem of clustering and mixture modeling in situations where the user has relevant prior knowledge on the center separation between clusters or mixture components. We propose a dynamic programming approach to solving the K-means problem exactly with a separation constraint on the centers in the context of real-valued data, and propose an EM algorithm that incorporates such a constraint in the context of fitting a Gaussian mixture

model.

We then move onto two-component mixture models, where one component plays the role of the background component. We consider situations where the user has prior information on this background component's distribution: symmetric, monotonic, or log-concave. In each setting, we derive estimators for the background component and provide relevant theory. While estimating the log-concave background distribution, we also propose a method to estimate the largest concave minorant.

In addition, we consider the situation where the user has information on the number of modes of a mixture density. We provide a dynamic programming approach to fitting a density function by maximum likelihood when it is constrained to have a given number of modal intervals. When this value is not available, we provide several data-driven ways for selecting it.

For the last project, we establish the consistency of K-medoids in the context of metric spaces. Our general approach applies also to non-metric settings where only an ordering of the dissimilarities is available. We consider two types of ordinal information: one where all quadruple comparisons are available; and one where only triple comparisons are available.

Chapter 1

Introduction

Cluster analysis is widely regarded as one of the most important tasks in unsupervised data analysis (Duran and Odell, 2013; Jain et al., 1999; Kaufman and Rousseeuw, 2009), with broad applications in areas such as statistics, artificial intelligence, biology and economics. It focuses on the task of grouping a set of objects into clusters, such that objects in the same cluster are more similar to each other than to those in different clusters. In many situations, the user might have some prior knowledge on some of the clusters or mixture components, and incorporating such information via constraints could improve the accuracy of the clustering output (Basu et al., 2008). The main part of this dissertation is dedicated to clustering and mixture modeling methods with constraints. We also provide some theory on the consistency of K-medoids.

Chapter 2 is motivated by the desire to create well separated and meaningful clusters or mixture components, and by the dynamic programming approach of (Wang and Song, 2011). Many constraints have been incorporated into variants of K-means, for example on cluster sizes (Banerjee and Ghosh, 2006; Bennett et al., 2000; Nielsen and Nock, 2014), on user-specified pairs of data points that must belong to the same or different clusters (Basu et al., 2004; Davidson and Basu, 2007; Wagstaff et al., 2001). In addition, many constraints have been considered in the context of fitting a mixture model via EM-type algorithms, for example with equality constraints on a subset of parameters (Jamshidian, 2004; Kim and Taylor, 1995; Takai, 2012),

on the variance covariance matrices of Gaussian mixture models (Day, 1969; Hathaway, 1985, 1986; Ingrassia, 2004; Ingrassia and Rocci, 2007, 2011; Rocci et al., 2018), and on penalizing the norms of the mean vectors (Pan and Shen, 2007; Wang and Zhu, 2008). Although many variants of constraints have been proposed, none considers the constraint on separation between the centers of K-means clusters or Gaussian mixture model components, and therefore we introduce this constraint to K-means and Gaussian mixture modeling, and provide algorithms for the two situations respectively.

To do so, we place ourselves in dimension one. We first propose an optimal dynamic programming approach to solving the K-means problem exactly under a center separation constraint, building on (Wang and Song, 2011). In the context of fitting a Gaussian mixture model, we then propose an EM algorithm that incorporates such a constraint. When appropriate prior center separation information is available, our algorithms can improve clustering accuracy significantly.

Chapter 3 is motivated by the wide applications of two-component mixture models and the novel approach of Patra and Sen (2016). Two-component mixture models can be applied in robust statistics to model contamination (Hettmansperger and McKean, 2010; Huber, 1964; Huber and Ronchetti, 2009; Tukey, 1960), and can be applied in multiple testing to model the test statistics under their null hypotheses (Cai and Jin, 2010; Efron, 2007, 2012; Efron et al., 2001; Genovese and Wasserman, 2002; Jin, 2008; Jin and Cai, 2007; Meinshausen and Rice, 2006). Working within the multiple testing framework, Patra and Sen (2016) address the problem of estimating the background proportion with complete information on the background component. Their method focuses on fitting a "largest" known component under the mixture density. Specifically, given a two-component mixture density f representing the density of all the test statistics combined, and letting g_0 denote a completely known density, Patra and Sen (2016) try to estimate the following quantity

$$\theta_0 := \sup\{t : f \geq t g_0\}. \tag{1.1}$$

Inspired by Patra and Sen (2016), we extend their approach to settings where we do not have complete information on the background distribution.

To do so, we consider three widely-encountered situations where the background component's distribution is assumed to be either symmetric, monotone, or log-concave. For each of these settings, we provide an estimator for the background component's proportion and density, and provide numerical implementations. We also provide a confidence interval for the background proportion and a simultaneous confidence band for the background density, and study their consistency results. While implementing the algorithm for log-concave background components, we encountered a situation where we need to compute the largest concave minorant, and provide a numerical implementation based on sequential quadratic programming. We finally show the advantages our method has over existing methods, including requiring much less prior knowledge and being less prone to model misspecification.

Chapter 4 is motivated by the goal to create well-explained densities with a given number of modal intervals, and by the relatively lack of methodology in fitting densities with an arbitrary amount of modes. Not only do modes play an important role in densities (Pearson, 1895), they are also crucial to clustering purposes (Fukunaga and Hostetler, 1975). Although a significant amount of methods for fitting densities with a single mode has been proposed, only a very limited amount of literature considered fitting densities with 2 or more modes, and none implemented a dynamic programming approach. Therefore, we propose a dynamic programming approach to fitting densities with an arbitrary amount of modal intervals, and provide ways for selecting the amount of modal intervals when this information is not given.

To do so, we implement a grid search approach for the knots separating adjacent uni-modal densities. We propose a dynamic programming approach based on this grid, and propose a multi-grid search to reduce computational cost and improve the accuracy of the estimation on knot locations. Based on our approach, we also provide two ways – measure of fit and cross validation – for choosing the number of uni-modal intervals when this information is unknown.

Chapter 5 is motivated by a relatively lack of theoretical treatment on the clustering

method of K-medoids. K-medoids (Kaufman and Rousseeuw, 1987) is a popular alternative to K-means (MacQueen, 1967), where instead of any location in the metric space, cluster centers can only land on the data points. It has the advantage of being more robust than K-means, and also being applicable in non-metric settings, where only a ranking of the dissimilarities between data points are available. Although many variants of K-medoids have been proposed (Kaufman and Rousseeuw, 2009; Park and Jun, 2009; Van der Laan et al., 2003; Wang et al., 2019), the consistency of K-medoids in the more standard setting of clustering points in a metric space has not been previously proved. Also, the use of K-medoids in ordinal settings with only a ranking of the dissimilarities of data points does not seem very widespread. We were only able to find a few literature in fields such as computer vision (Huang et al., 2020; Zadegan et al., 2013; Zhu et al., 2011). Therefore we establish the consistency of K-medoids in the context of metric spaces, and establish its consistency for two types of ordinal information.

To do so, we first prove that K-medoids is asymptotically equivalent to K-means restricted to the support of the underlying distribution under general conditions and a broad range of loss functions. This asymptotic equivalence enables us to apply the work of Pärna (1986) on the consistency of K-means. Our approach also applies to showing consistency in non-metric settings where only an ordering of the dissimilarities is available. We consider two types of ordinal information: quadruple comparisons giving an overall ranking of all pairwise dissimilarities; and triple comparisons giving a ranking relative to each sample point.

Chapter 2

K-Means and Gaussian Mixture Modeling with a Separation Constraint

2.1 Abstract

We consider the problem of clustering with K-means and Gaussian mixture models with a constraint on the separation between the centers in the context of real-valued data. We first propose a dynamic programming approach to solving the K-means problem with a separation constraint on the centers, building on (Wang and Song, 2011). In the context of fitting a Gaussian mixture model, we then propose an EM algorithm that incorporates such a constraint. A separation constraint can help regularize the output of a clustering algorithm, and we provide both simulated and real data examples to illustrate this point.

2.2 Introduction

Cluster analysis is broadly seen as one of the most important tasks in (unsupervised) data analysis (Jain et al., 1999; Kaufman and Rousseeuw, 2009). In some situations, the analyst might have some prior knowledge or relevant information regarding the clusters, and incorporating this information — which may be done via constraints — is thought to improve the accuracy of the clustering output (Basu et al., 2008). In the present paper, we address situations where the analyst has prior knowledge on the separation between the cluster centers. We indeed focus

on methods that rely on centers to define clusters: the K-means criterion and Gaussian mixture models (GMM).

2.2.1 Constrained K-means

A number of constrained variants of K-means have been proposed in the literature, where a significant amount of attention have been placed on the cluster sizes such as imposing a lower bound on the minimum cluster size or aiming for balancing the cluster sizes. For instance, Bennett et al. (2000) addressed the problem of small or even empty clusters by enforcing a size constraint in the cluster assignment step of Lloyd's algorithm; Nielsen and Nock (2014) designed a dynamic K-means algorithm in 1D aimed at optimizing the Bregman divergence that can incorporate a constraint on the minimum cluster size; and Banerjee and Ghosh (2006) proposed an approach to K-means that is able to handle cluster size balance constraints. Constrained K-means has also been widely considered in other areas (Basu et al., 2008). For example, (Basu et al., 2004; Wagstaff et al., 2001) considered forcing some user-specified pairs of observations to belong to the same cluster while forcing others to belong to different clusters. A overview of instance-level constrained clustering, including K-means approaches and related methods, can be found at (Davidson and Basu, 2007). Davidson and Ravi (2005) also incorporated a minimum separation constraint between clusters mandating that any two points assigned to different clusters need to be separated by at least δ, a parameter specified by the analyst. Szkaliczki (2016) considered points arriving sequentially and constraints where clusters have to be of the form $\{x_i, \ldots, x_{i+j}\}$. A recent survey of constrained clustering, including K-means related methods and modern development can be found at Gançarski et al. (2020).

We propose a variant of K-means where a user-specified separation between the cluster centers is enforced which, to the best of our knowledge, is novel. We place ourselves in dimension one, where we are able to solve the resulting constrained optimization problem exactly by building on the dynamic programming approach of Wang and Song (2011).

2.2.2 Constrained Gaussian mixture models

Model based clustering is an important aspect in clustering (Fraley and Raftery, 2002; McLachlan and Peel, 2004), and among parametric finite mixture models, Gaussian mixture models (GMM) are by far the most popular, with their parameters frequently estimated by the EM algorithm or variants (Dempster et al., 1977; McLachlan and Krishnan, 2007).

EM-type algorithms forcing various constraints on the parameters of the mixture model have been considered extensively in the literature. Kim and Taylor (1995), in a more general setting that includes GMM, considered enforcing some equality constraints on a subset of the parameters and provided an EM algorithm based on a projected Newton–Raphson step in the maximization stage. Along the lines, but addressing both equality and inequality constraints, Jamshidian (2004) used a projected gradient step instead, in conjunction with an active set method in order to handle the inequality constraints. See also (Takai, 2012). In the context of linear models in regression, EM-type algorithms were also proposed to handle constraints on the parameters, in particular linear equalities and inequalities (Davis et al., 2012; Shi et al., 2005; Shin et al., 2001; Tan et al., 2007; Tian et al., 2008; Zheng et al., 2012, 2005).

In the context of GMM proper, constraints have been considered for a very long time to reduce the number of parameters, for example, by making all the variances or covariance matrices to be equal. Beyond that, constraints on the variances or eigenvalues of the covariance matrices have been considered extensively, mostly for the purpose of bounding the log likelihood (Day, 1969). In particular, Hathaway (1985) proposed a lower bound on the ratio of any two component standard deviations, which enabled the author to prove the strong consistency of the resulting maximum likelihood estimator. This was followed by (Hathaway, 1986) where a constrained EM algorithm was proposed for maximizing the likelihood with this lower bound constraint on the ratio of any two standard deviations, and also another lower bound constraint on the weights. Generalizing Hathaway's results to higher dimensions, Ingrassia (2004); Ingrassia and Rocci (2007, 2011); Rocci et al. (2018) examined EM algorithms enforcing constraints on

the eigenvalues of the covariance matrices.

In addition to constraints on component variance or eigenvalues, estimation of component means under constraints has also been considered in the literature, where EM-type algorithms have been proposed for that purpose. For example, Pan and Shen (2007) and Wang and Zhu (2008) considered penalizing the L_1 and L_∞ norms of the mean vectors, respectively, to enforce sparsity and/or regularize the model in high dimensions, while Chauveau and Hunter (2013) and Qiao and Li (2015) proposed EM algorithms for linear constraints on the mean parameters.

We consider here, in dimension 1, arbitrary constraints on the separation between adjacent means. We propose an EM algorithm that enforces such constraints, where the M step amounts to solving a quadratic program.

2.2.3 Setting and content

The main focus will be on real-valued data. We do not know how to enforce a separation constraint on the centers in higher dimensions, at least not in such principled fashion. The data points will be assumed ordered without loss of generality, denoted $x_1 \leq \cdots \leq x_N$ and gathered in a vector $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$. Our goal will be to group these data points into K clusters, where K is specified by the user.

The organization of the paper will be as follows. In Section 2.3, we introduce our dynamic programming algorithm for exactly solving K-means in dimension one with a constraint on the minimum separation between the centers. In Section 2.4, we introduce our EM algorithm for (approximately) fitting a Gaussian mixture model by maximum likelihood under the same constraint. It is a true EM algorithm in that the likelihood increases with each iteration. In Section 2.5, we describe some numerical experiments on both simulated data and real data to illustrate the use and accuracy of our methods.

2.3 A dynamic program for separation-constrained K-means

Given data points on the real line, $x_1, \ldots, x_n$, and a desired number of cluster, K, basic K-means is defined as the following optimization problem (MacQueen, 1967)

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} \min_{k=1,\ldots,K} (x_i - \mu_k)^2, \\ \text{over} \quad & \mu_1 < \cdots < \mu_K. \end{aligned}$$

This problem is difficult in general. For one thing, a brute-force approach by grid search would be in dimension K and quickly unfeasible or too costly. The problem is instead most often approached via iterative methods such as Lloyd's algorithm (Lloyd, 1982), which consists in alternating between, for all k, defining Cluster k as the set of points nearest to μ_k and then recomputing μ_k as the barycenter or average of the points forming Cluster k. In dimension 1, however, the K-means problem can be solved by dynamic programming as was established and carried out by Wang and Song (2011).

We are interested in a variant of the K-means problem where the following separation constraint on the centers is enforced: $\mu_{k+1} - \mu_k \geq \delta$ for all $k = 1, \ldots, K-1$. In other words, we consider the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} \min_{k=1,\ldots,K} (x_i - \mu_k)^2, \\ \text{over} \quad & \mu_1 < \cdots < \mu_K \quad \text{satisfying} \quad \mu_{k+1} - \mu_k \geq \delta \text{ for all } k = 1, \ldots, K-1. \end{aligned}$$

The amount of separation is determined by the user-specified parameter $\delta \geq 0$. (Of course, if $\delta = 0$, we recover the basic K-means problem.) Inspired by the work of Wang and Song (2011), we propose a dynamic programming algorithm to solve this constrained variant of K-means. The remaining of this section is dedicated to describing this algorithm, which is encapsulated in

Table 2.1 and Table 2.2.

2.3.1 The DP algorithm of Wang and Song (2011)

Given the sorted data $x = (x_1, \ldots, x_N)$, for $r_1 = 1, \ldots, N$ and $r_2 = r_1, \ldots, N$, we first define the within current cluster center (C) and the within current cluster sum of squares (W):

$$C[r_1, r_2] = \frac{1}{r_2 - r_1 + 1} \sum_{i=r_1}^{r_2} x_i, \tag{2.1}$$

$$W[r_1, r_2] = \frac{1}{r_2 - r_1 + 1} \sum_{i=r_1}^{r_2} (x_i - C[r_1, r_2])^2 \tag{2.2}$$

Each element can be efficiently computed using a simple recursion. For $i = 1, \ldots, N$ and $m = 1, \ldots, K$, define $D[i, m]$ as recording the minimum error sum of squares when grouping the first i data points into m clusters. The main idea of Wang and Song (2011) is to update D dynamically as follows:

$$D[i, m] = \min_{j=m,\ldots,i} D[j-1, m-1] + W[j, i] \ . \tag{2.3}$$

This is carried out for $i = 1, \ldots, N$ and $m = 1, \ldots, K$. (Initialization is $D[0, m] = 0$ for all m and $D[i, 0] = 0$ for all i.)

The method of (Wang and Song, 2011) was designed for solving the basic K-means problem and needs to be modified, in a substantial way, to solve the separation-constrained K-means problem. For illustrative purposes, consider clustering data $x = (-2, 1, 2, 4, 5, 6, 9, 10)$ into $K = 5$ clusters, while keeping each cluster center separated by at least $\delta = 1.75$. Note that without separation the clustering assignment is 12233455 but the centers of Cluster 3 and Cluster 4 are only separated by 1.5, which is less than δ. With separation, the clustering assignment is 12334455. The same example is used to highlight two issues that need to be addressed in order to incorporate the separation constraints:

- The m-th cluster consisting of $x_j, \ldots, x_i$, and the previous cluster, i.e., the $(m-1)$-th cluster

that gives the optimal clustering of $j-1$ data points into $m-1$ clusters, might not be separated by δ in terms of their means. In our example, when grouping $\{x_1,\ldots,x_5\}$ into 3 clusters, the optimal solution is 12233, but when we try to group $\{x_1,\ldots,x_6\}$ into 4 clusters under the separation constraint it becomes impossible to do so as $C[6,6]-C[4,5]<\delta$. (We are not allowing the clusters to be empty.)

- Consider the stage where we have grouped $x_1,\ldots,x_i$ into m clusters, and the last cluster is $x_j,\ldots,x_i$. Then it is not necessarily the case that the first $m-1$ clusters constitute an optimal grouping of $x_1,\ldots,x_{j-1}$, and so the recursion (2.3) is not necessarily valid. In our example, the optimal clustering for grouping $\{x_1,\ldots,x_4\}$ into 3 clusters is 1223, however when grouping $\{x_1,\ldots,x_5\}$ into 4 clusters under the separation constraint the clustering assignment becomes 12334, so that the starting point of cluster 3 has changed from x_4 to x_3.

In brief, (2.3) cannot be used as the updating rule. Although the first issue is easy fix by checking that the separation constraint is satisfied, the second issue is more difficult to deal with and requires us to introduce additional matrices/tensors as described below.

2.3.2 Our DP algorithm with separation constraint

We describe here our algorithm. First initialize:

$$\beta = W[1,N]+1, \qquad D=\beta\, 1_{N\times K}, \qquad I=0_{N\times K}, \qquad B=0_{N\times K}, \qquad U=\beta\, 1_{N\times N\times K}\,, \tag{2.4}$$

where β is a placeholder value chosen to be larger than the maximum possible error sum of squares. Define:

$$\begin{gathered} D[i,1]=\sum_{l=1}^{i}(x_l-C[1,i])^2, \quad I[i,1]=1, \quad B[i,1]=1, \quad U[i,r,1]=0, \\ \text{for } i=1,2,\ldots,N;\ r=i,\ldots,N\,. \end{gathered} \tag{2.5}$$

The D matrix is used to store the optimal error sum of squares under the separation constraint, specifically $D[i,m]$ records the minimal error sum of squares when putting the first i values into m clusters while satisfying separation, if possible (recall we do not allow clusters to be empty). The corresponding entry in the I matrix, $I[i,m]$, is used to record the index of the first data point in the m-th cluster in the optimal clustering of the first i data points into m clusters (the clustering corresponding to $D[i,m]$). $B[i,m]$, on the other hand, is used to store the smallest possible index of the leftmost data point in the m-th cluster over all groupings of the first i data points into m clusters satisfying the separation constraint. Lastly, $U[q,r,m]$ records the optimal error sum of squares for grouping the first $q-1$ data points into $m-1$ clusters, while satisfying the constraint that the $(m-1)$-th cluster and m-th cluster, which consists of data points $x_q,\ldots,x_r$, are separated by δ in means.

In our updates of the above matrices/tensors, we enforce the separation constraint. Suppose that $D[\cdot,m-1], I[\cdot,m-1], B[\cdot,m-1]$ as well as $U[\cdot,\cdot,m-1]$ have been updated, so we are at stage m with the immediate task of updating $D[\cdot,m], I[\cdot,m], B[\cdot,m]$ as well as $U[\cdot,\cdot,m]$. Now consider grouping the first i data points into m clusters. If $B[i-1,m-1]=0$, do nothing (the values stay at the initial ones). Otherwise, for each $j=m,\ldots,i$, we find the first value in the descending sequence $T_j \in \{I[j-1,m-1],\ldots,B[j-1,m-1]\}$ that satisfies $C[j,i]-C[T_j,j-1]\geq\delta$. Note that T_j also depends on i and m, but that is left implicit. The reason for defining T_j as such is because starting from $B[j-1,m-1]$, the closer we are to $I[j-1,m-1]$, the smaller the error sum of squares. Then the main update that guarantees optimality is:

$$D[i,m]=\min_{j=m,\ldots,i}\left(U[T_j,j-1,m-1]+W[T_j,j-1]+W[j,i]\right), \tag{2.6}$$

and

$$U[j,i,m]=U[T_j,j-1,m-1]+W[T_j,j-1]\,. \tag{2.7}$$

Then $I[i,m]$ is updated as the smallest minimizing index in (2.6) and $B[i,m]$ is updated as the smallest j that satisfies $C[j,i]-C[T_j,j-1]\geq\delta$.

Table 2.1. Separation Constrained K-Means.

inputs: C, I, B (all three obtained from Table 2.2), number of clusters K, separation δ

initialize $b = 0_{\{K\}}$, compute $b[K] = I[N, K]$, and set $l = b[K] - 1$
if $K = 1$ **then**
  return b
for $j = I[l, K-1], \ldots, B[l, K-1]$ **do**
  if $C[l+1, N] - C[j, l] \geq \delta$ **then**
    $b[K-1] = j, l = j - 1$
    break
if $K = 2$ **then**
  return b
for $k = (K-2), \ldots, 1$ **do**
  for $j = I[l, k], \ldots, B[l, k]$ **do**
    if $(C[l+1, b[k+2]-1] - C[j, l]) \geq \delta$ **then**
      $b[k] = j, l = j - 1$
      break
return b

Computational complexity

While the algorithm of Wang and Song (2011) has time complexity $O(N^2K)$, our method has computational complexity $O(N^3K)$. While this is true in the worst case, in practice the computational cost appears much lower as moving from $I[j-1, m-1]$ to $B[j-1, m-1]$ and stopping at the first value where the constraint is satisfied considers significantly fewer options than N. The upper bound for space complexity of this algorithm is $O(N^2K)$ — compared to $O(NK)$ for the algorithm of Wang and Song (2011). This is due to the maintaining of the tensor U, but again, this is in the worst case, as in practice U is very sparse (most of its coefficients are equal to the initial value β) and the space complexity is far below the upper bound. We are now ready to use Table 2.2 to find the the optimal error sum of squares under separation and find the corresponding optimal clustering using Table 2.1. We note that the result b of Table 2.1 is a vector indicating the index of each cluster's last datapoint. We also note that in Table 2.2 there is the possibility that the minimum value of p occurs at multiple locations; when this happens, we select the entry with the smallest location index.

Table 2.2. Routine computing matrices C, W, D, I, B.

inputs: data $x_1 \le \cdots \le x_N$, number of clusters K, separation δ

compute C and W as in (2.1) and (2.2)
define β and initialize D, I, B, U as in (2.4) and (2.5)
if $K = 1$ **then**
go to return step
else if $K = 2$ **then**
  for $i = 2, \ldots, N$ **do**
    $p = \beta_{\{i-1\}}$
    for $j = 2, \ldots, i$ **do**
      if $C[j,i] - C[1, j-1] \ge \delta$ **then**
        $p[j-1] = W[j,i] + W[1, j-1]$
        $U[j,i,2] = W[1, j-1]$
    $D[i,2] = \min(p)$
    if $D[i,2] = \beta$ **then**
      $I[i,2] = 0, B[i,2] = 0$
    else
      $I[i,2] = \min(\arg\min(p)) + 1$, $B[i,2] = \min\{a : p[a] < \beta\} + 1$
    go to return step
else
  do everything in $K = 2$ case except go to return step, and:
  for $m = 3, \ldots, K$ **do**
    for $i = m, \ldots, N$ **do**
      if $D[i-1, m-1] < \beta$ **then**
        $p = \beta_{\{i-m+1\}}$
        for $j = m, \ldots, i$ **do**
          if $B[j-1, m-1] \ne 0$ **then**
            if $C[j,i] - C[I[j-1, m-1], j-1] \ge \delta$ **then**
              $p[j-m+1] = D[j-1, m-1] + W[j,i]$
              $U[j,i,m] = D[j-1, m-1]$
            else
              for $t = I[j-1, m-1], \ldots, B[j-1, m-1]$ **do**
                if $C[j,i] - C[t, j-1] \ge \delta$ **then**
                  $p[j-m+1] = \min(p[j-m+1], U[t, j-1, m-1] + W[t, j-1] + W[j,i])$
                  $U[j,i,m] = U[t, j-1, m-1] + W[t, j-1]$
                  break
        $D[i,m] = \min(p)$
        if $D[i,m] = \beta$ **then**
          $I[i,m] = 0, B[i,m] = 0$
        else
          $I[i,m] = \min(\arg\min(p)) + m - 1$, $B[i,m] = \min\{a : p[a] < \beta\} + m - 1$
return D, I, B, C, W, β, U

2.4 An EM algorithm for separation-constrained GMM

A Gaussian mixture model is of the form:

$$\sum_{k=1}^{K} \pi_k f(x; \mu_k, v_k), \tag{2.8}$$

where $f(\cdot; \mu, v)$ is the density of the normal distribution with mean μ and variance v, meaning

$$f(x; \mu, v) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(x-\mu)^2}{2v} \right), \tag{2.9}$$

and $\pi_1, \ldots, \pi_K$ are the mixture weights satisfying

$$\min_k \pi_k > 0, \qquad \sum_k \pi_k = 1. \tag{2.10}$$

For convenience, define $\theta = (\pi, \mu, v)$ where $\pi = (\pi_1, \ldots, \pi_K) \in \mathbb{R}^K$, $\mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K$, and $v = (v_1, \ldots, v_K) \in \mathbb{R}^K$. We then let g_θ denote the mixture (2.8). The working assumption is that the data points, $x_1, \ldots, x_N$, are the realization of a sample of some size N drawn iid from a mixture distribution of the form (2.8), where a draw from that distribution amounts to drawing a cluster index in $\{1, \ldots, K\}$ according to the probability distribution π and then generating a real valued observation according the corresponding normal density — $f(\cdot; \mu_k, v_k)$ if k is the drawn index value. Given the mixture model (2.8), a point x is assigned to the most likely cluster, meaning, to

$$\arg\max_k \pi_k f(x; \mu_k, v_k). \tag{2.11}$$

Thus clustering using a Gaussian mixture model boils down to estimating the parameters of the model. Maximum likelihood estimation[1] is a natural candidate, but difficult to implement,

[1] Interestingly, maximum likelihood is a competitive method even though the likelihood is not bounded. To make sense of what the EM algorithm does, we believe that a useful perspective is to see it as attempting to maximize the likelihood starting from a reasonable initialization, understanding that the E and M steps prevent the algorithm from arriving at pathological values of the parameters.

even in dimension 1, as maximizing the likelihood is not a convex problem and a grid search is too costly given the number of parameters ($3K - 1$). The Expectation-Maximization (EM) algorithm of Dempster et al. (1977) — arguably the most famous approach to fitting a Gaussian mixture model — is an iterative method that attempts to maximize the (log) likelihood:

$$L(\theta, x) = \sum_{i=1}^{N} \log g_\theta(x_i). \tag{2.12}$$

The main idea is based on introducing cluster assignment variables, $z_1, \ldots, z_n$, where $z_i = k$ if x_i was drawn from the k-th component of the mixture. These assignment variables are not observed (they are said to be latent), but are useful nonetheless for thinking about the problem. To that end, let $g_\theta(x, z)$ denote the joint density of (X_i, Z_i). Note that Z_i has distribution π and $X_i \mid Z_i = k$ has the normal distribution with mean μ_k and variance v_k. The algorithm starts with a value of the parameter vector, θ^0, and then alternates between an E step and an M step, until some convergence criterion is satisfied. Suppose we are at the s-th iterate. The E step consists in computing the expectation of the log-likelihood for an arbitrary value of θ with respect to the mixture distribution given by the current parameter estimates:

$$Q(\theta, \theta^s) := \sum_{i=1}^{N} \mathbb{E}_{\theta^s}\left[\log g_\theta(X_i, Z_i) \mid X_i = x_i\right]. \tag{2.13}$$

It turns out that

$$Q(\theta, \theta^s) = \sum_{i=1}^{N} \sum_{k=1}^{K} w_{i,k}^{s+1} \log(\pi_k f(x_i; \mu_k, v_k)), \tag{2.14}$$

where

$$w_{i,k}^{s+1} := \frac{\pi_k^s f(x_i; \mu_k^s, v_k^s)}{\sum_{l=1}^{K} \pi_l^s f(x_i; \mu_l^s, v_l^s)}. \tag{2.15}$$

The M step consists of maximizing the resulting function with respect to θ to yield the updated values of the parameters:

$$\theta^{s+1} = \arg\max_\theta Q(\theta, \theta^s). \tag{2.16}$$

This amounts to the following:

$$\pi_k^{s+1} = \frac{1}{N}\sum_{i=1}^{N} w_{i,k}^{s+1} \tag{2.17}$$

$$\mu_k^{s+1} = \frac{1}{\sum_{i=1}^{N} w_{i,k}^{s+1}}\sum_{i=1}^{N} w_{i,k}^{s+1} x_i \tag{2.18}$$

$$v_k{}^{s+1} = \frac{1}{\sum_{i=1}^{N} w_{i,k}^{s+1}}\sum_{i=1}^{N} w_{i,k}^{s+1}(x_i - \mu_k^{s+1})^2 \tag{2.19}$$

We are interested in a variant of the EM algorithm for maximizing the likelihood under the following separation constraint on the centers: $\delta_{k,1} \le \mu_{k+1} - \mu_k \le \delta_{k,2}$ for all $k = 1, \ldots, K-1$. The separation parameters, $\delta_{k,1}$ and $\delta_{k,2}$, are set by the user. We note that this reduces to separation when we set $\delta_{1,1} = \cdots = \delta_{K-1,1} = \delta$ and $\delta_{1,2} = \cdots = \delta_{K-1,2} = \infty$ (in practice, a very large number). We propose an EM algorithm that incorporates separation2. It is a true EM algorithm in that the likelihood increases with each iteration. The remaining of this section is devoted to introducing the algorithm, which is otherwise compactly given in Table 2.3.

In the algorithm, the E Step is identical to that in the regular EM algorithm. The M Step is also the same except that the maximization is done under separation2. The constraint is only on the centers, and to isolate that, we adopt an ECM approach (Meng and Rubin, 1993), where we first maximize over the weights, which amounts to the same update (2.17); then maximize over the centers subject to separation2 (see details below); and finally maximize over the variances, which amounts to the same update (2.19).

Proposition 1. *The separation-constrained EM algorithm does not decrease the log likelihood of the model, i.e.,* $L(\theta^{s+1}, x) \ge L(\theta^s, x)$ *for all s.*

Proof. Using Jensen's inequality, we derive:

$$L(\theta,x) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k f(x_i;\mu_k,v_k)\right) \tag{2.20}$$

$$= \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} w_{i,k}^{s+1} \frac{\pi_k f(x_i;\mu_k,v_k)}{w_{i,k}^{s+1}}\right) \tag{2.21}$$

$$\geq \sum_{i=1}^{N}\sum_{k=1}^{K} w_{i,k}^{s+1} \log\left(\frac{\pi_k f(x_i;\mu_k,v_k)}{w_{i,k}^{s+1}}\right) \tag{2.22}$$

$$= Q(\theta,\theta^s) - \sum_{i=1}^{N}\sum_{k=1}^{K} w_{i,k}^{s+1} \log(w_{i,k}^{s+1}). \tag{2.23}$$

Plugging in θ^{s+1} for θ, we obtain:

$$L(\theta^{s+1},x) \geq Q(\theta^{s+1},\theta^s) - \sum_{i=1}^{N}\sum_{k=1}^{K} w_{i,k}^{s+1} \log(w_{i,k}^{s+1}), \tag{2.24}$$

while by definition of the weights in (2.15), we have:

$$L(\theta^s,x) = Q(\theta^s,\theta^s) - \sum_{i=1}^{N}\sum_{k=1}^{K} w_{i,k}^{s+1} \log(w_{i,k}^{s+1}) \tag{2.25}$$

Thus to prove that $L(\theta^{s+1},x) \geq L(\theta^s,x)$, it suffices to prove that $Q(\theta^{s+1},\theta^s) \geq Q(\theta^s,\theta^s)$. As we maximized $Q(\theta = (\pi,\mu,v),\theta^s)$ in every step of the ECM algorithm with other parameters fixed, we have:

$$Q(\theta^{s+1},\theta^s) = Q((\pi^{s+1},\mu^{s+1},v^{s+1}),\theta^s) \tag{2.26}$$

$$\geq Q((\pi^{s+1},\mu^{s+1},v^s),\theta^s) \tag{2.27}$$

$$\geq Q((\pi^{s+1},\mu^s,v^s),\theta^s) \tag{2.28}$$

$$\geq Q((\pi^s,\mu^s,v^s),\theta^s) \tag{2.29}$$

$$= Q(\theta^s,\theta^s). \tag{2.30}$$

□

The maximization over the centers amounts to solving the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{N}\sum_{k=1}^{K}\frac{w_{i,k}^{s+1}}{2v_k^s}(x_i-\mu_k)^2 \\ &\text{over} \quad \mu_1,\ldots,\mu_K \quad \text{such that} \quad \delta_{k,1}\le\mu_{k+1}-\mu_k\le\delta_{k,2} \text{ for all } k=1,2,\ldots,K-1. \end{aligned} \tag{2.31}$$

Define $g_{i,k} = w_{i,k}^{s+1}/2v_k^s$ and

$$G_i = \operatorname{diag}(g_{i,1},\ldots,g_{i,K}), \qquad G=\sum_{i=1}^{n}G_i\ . \tag{2.32}$$

Also define

$$A=\begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0\\ 0 & -1 & 1 & \ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \ldots & -1 & 1\\ 1 & -1 & 0 & \ldots & 0 & 0\\ 0 & 1 & -1 & \ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \ldots & 1 & -1 \end{bmatrix}, \qquad b_0=\begin{bmatrix} \delta_{1,1}\\ \delta_{2,1}\\ \ldots\\ \delta_{K-1,1}\\ -\delta_{1,2}\\ -\delta_{2,2}\\ \ldots\\ -\delta_{K-1,2} \end{bmatrix}, \tag{2.33}$$

where A is a $2K-2$ by K matrix and b_0 is a $2K-2$ vector. With this notation, and let $1^\top$ denote a vector of length K with each entry as 1, the optimization problem (2.31) can be equivalently expressed as follows:

$$\begin{aligned} &\text{minimize} \quad \mu^\top G\mu - 2\sum_{i=1}^{N}x_i 1^\top G_i\mu \\ &\text{subject to} \quad A\mu\ge b_0, \end{aligned} \tag{2.34}$$

which is a quadratic program. In practice, we solve it using the primal dual interior point method of Goldfarb and Idnani (1982, 1983) implemented in the quadprog package[2] in R.

[2] https://cran.r-project.org/web/packages/quadprog/

Table 2.3. Separation Constrained EM Algorithm for GMM.

inputs: data $x_1, \ldots, x_N$, number of clusters K, tolerance value γ

initialization

initialize (π^1, μ^1, v^1) using constrained 1D optimal K-means algorithm as in Section 2.3, if given the same lower bound constraint; otherwise initialize using 1D optimal K-means of (Wang and Song, 2011):

$$\mu_1^1 < \mu_2^1 < \cdots < \mu_K^1, \quad \mu_k^1 = \frac{1}{|J_k|} \sum_{i \in J_k} x_i, \quad \pi_k^1 = \frac{|J_k|}{N}, \quad v_k^1 = \frac{1}{|J_k|} \sum_{i \in J_k} (x_i - \mu_k)^2 \tag{2.35}$$

where J_k denotes the indices that are clustered into the $k-th$ cluster

E step

compute $w_{i,k}$ for $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots K$:

$$w_{i,k}^{s+1} = \frac{\pi_k^s N(x_i | \mu_k^s, v_k^s)}{\sum_{j=1}^K \pi_j^s N(x_i | \mu_j^s, v_j^s)} \tag{2.36}$$

M step

update π_k, $k = 1, 2, \ldots, K$:

$$\pi_k^{s+1} = \frac{1}{N} \sum_{i=1}^N w_{i,k}^{s+1} \tag{2.37}$$

update μ_k, $k = 1, 2, \ldots, K$ by the solution to minimization problem (2.34)
update v_k, $k = 1, 2, \ldots, K$:

$$v_k{}^{s+1} = \frac{1}{\sum_{i=1}^N w_{i,k}^{s+1}} \sum_{i=1}^N w_{i,k}^{s+1} (x_i - \mu_k^{s+1})^2 \tag{2.38}$$

convergence check

if $\max\{\max_k |\pi_k^{s+1} - \pi_k^s|, \max_k |\mu_k^{s+1} - \mu_k^s|, \max_k |v_k^{s+1} - v_k^s|\} \leq \gamma$
  then stop and return $\theta^{s+1} = (\pi^{s+1}, \mu^{s+1}, v^{s+1})$
else
  $s = s + 1$ and go to E step

2.5 Experiments

This section will provide some simulated and real data examples illustrating the application of our algorithms. The constrained algorithms are shown to outperform their unconstrained counterparts in both parameter estimation and in producing a clustering that is more similar to the actual clusters, of course, if the separation constraint reflects the reality of the situation. Our code is publicly available.[3]

2.5.1 Simulations

We start with experiments performed on simulated data using our constrained algorithms. We will first look at the performance of our constrained K-means algorithm over regular K-means as implemented by Wang and Song (2011). We then look at the performance of our constrained EM algorithm for fitting Gaussian mixture models with the regular EM algorithm (Dempster et al., 1977). Finally, we examine the impact of the imposed constraints, including when they are not congruent with the actual situation.

All the simulated data sets were generated from one of the following models:

- Model A: $0.333\,\mathcal{N}(0,1^2) + 0.667\,\mathcal{N}(2,1^2)$
- Model B: $0.45\,\mathcal{N}(0,0.75^2) + 0.1\,\mathcal{N}(2,1.5^2) + 0.45\,\mathcal{N}(4,0.75^2)$
- Model C: $0.2\,\mathcal{N}(0,1^2) + 0.2\,\mathcal{N}(2,1^2) + 0.2\,\mathcal{N}(4,1^2) + 0.2\,\mathcal{N}(6,1^2) + 0.2\,\mathcal{N}(8,1^2)$
- Model D: $0.1\,\mathcal{N}(0,0.25^2) + 0.2\,\mathcal{N}(2,0.75^2) + 0.4\,\mathcal{N}(4,1.25^2)$
  $+ 0.2\,\mathcal{N}(6,0.75^2) + 0.1\,\mathcal{N}(8,0.25^2)$

We first apply constrained and unconstrained K-means on data generated from Model D (Experiment 1) and on data generated from Model B (Experiment 2). The comparison criteria

[3] https://github.com/h8jiang/Center-Separation-Constrained-K-Means-and-EM-Algorithm/

Table 2.4. Errors and Rand Indices of Optimal K-means and Optimal Constrained K-means.

Experiment	**Model**	**Criteria**	K**-means**	**Constrained** K**-means**
Experiment 1	Model D	center error	1.092 (0.276)	0.374 (0.161)
		size error	165.6 (22.9)	119.9 (25.4)
		Rand index	0.786 (0.015)	0.807 (0.014)
Experiment 2	Model B	center error	1.339 (0.393)	0.561 (0.190)
		size error	143.4 (45.9)	58.1 (18.8)
		Rand index	0.834 (0.016)	0.858 (0.014)

are total error on cluster center values and the total difference on the cluster sizes:

$$\sum_{k=1}^{K} |\hat{\mu}_k - \mu_k|, \quad \text{and} \quad \sum_{k=1}^{K} |\hat{n_k} - n_k| \ , \tag{2.39}$$

and also the Rand index. The sample size was set to $N = 500$, the separation to $\delta = 1.95$, and the number of repeats to $R = 1000$. Note that the separation is satisfied in actuality, and this is important. We record the mean value over all R experiments and their standard deviations in Table 2.4. It is quite clear that in these experiments constrained K-means improves significantly over unconstrained K-means in terms of both center estimates and clustering accuracy.

Experiments 3, 4 and 5 will show three examples where applying our constrained EM algorithm with a valid separation constraint results in a large improvement of parameter estimation over the regular EM algorithm. Both algorithms are identically initialized using the constrained K-means algorithm in Section 2.3. We apply these two methods on data generated from Model B in Experiment 3, data generated from Model A in Experiment 4, and data generated from Model C in Experiment 5. The comparison criteria are average error on center values and the average error on all parameters:

$$\frac{1}{K}\sum_{k=1}^{K} |\hat{\mu}_k - \mu_k|, \quad \text{and} \quad \frac{1}{K}\sum_{k=1}^{K} \left(|\hat{\mu}_k - \mu_k| + |\hat{\pi}_k - \pi_k| + |\hat{v}_k - v_k| \right) , \tag{2.40}$$

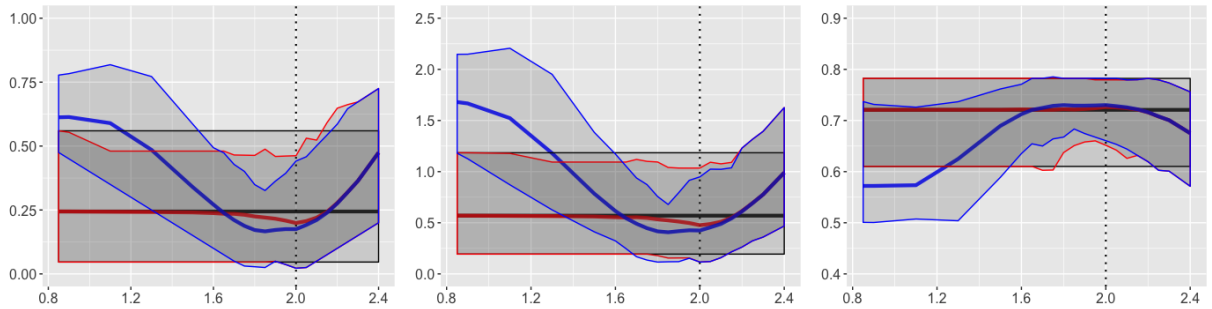
and the Rand index. Here the sample size is $N = 500$, the lower and upper bounds on the

Table 2.5. Errors and Rand Indices of EM Algorithm and Constrained EM Algorithm.

Experiment	Model	Criteria	Regular EM	Constrained EM
Experiment 3	Model B	center error	0.339 (0.205)	0.058 (0.021)
		all parameters error	0.976 (0.296)	0.454 (0.197)
		Rand index	0.893 (0.023)	0.906 (0.013)
Experiment 4	Model A	center error	0.252 (0.155)	0.172 (0.120)
		all parameters error	0.568 (0.320)	0.409 (0.242)
		Rand index	0.715 (0.047)	0.726 (0.040)
Experiment 5	Model C	center error	0.448 (0.231)	0.276 (0.213)
		all parameters error	0.994 (0.415)	0.764 (0.367)
		Rand index	0.810 (0.028)	0.820 (0.030)

separation are $\delta_1 = 1.9$ and $\delta_2 = 2.1$ (same across k), and the number of repeat is $R = 1000$. Note again that the separation constraint is accurate in that it holds in all the models used in this set of experiments. We record the mean value of all R experiments and their standard deviations in Table 2.5. It is clear that in these 3 experiments, the constrained EM algorithm improves over the regular EM algorithm in terms of both parameter estimation and clustering accuracy.

In the last experiments, we look at the impact of the separation constraint on the result of the constrained algorithms (Experiment 6), and also explore how imposing two-sided constraints differs from imposing one-sided constraints (Experiment 7). We focus on the simplest model, Model A, which has only two components. In Experiment 6, we gradually increase δ on a grid of values ranging from 0.85 to 2.4. We generate $R = 100$ data sets from Model A, then for each δ, we fit the model and record all parameters as done in Experiments 3-5. We present the results in 3 plots with 90% confidence bands in Figure 1. We can see that the curves representing regular and one-sided EM tend to overlap on the very left. The one-sided EM starts to become more accurate in parameter estimation compared to regular EM when δ is as low as 0.9, and the improvement increases as we move closer to actual value 2. Once passed 2, the performance starts to deteriorate compared to the regular EM. On the other hand, when the two-sided constraint indeed contains the true separation, 2, the two-sided constrained EM outperforms the one-sided constrained and



(a) error in the estimation of the centers as a function of the separation δ

(b) error in the estimation of all the parameters as a function of the separation δ

(c) Rand index as a function of the separation δ

Figure 2.1. Results of Experiments 6 and 7: black curves represent unconstrained EM; red curves represent constrained EM with a lower bound on separation, δ; blue curves represent two-sided constrained EM, with lower bound on separation being $\delta_1 = \delta$ and upper bound being $\delta_2 = \delta + 0.2$; thick horizontal line is at the mean values of 100 experiments, while the thin horizontal lines give the 90% confidence bands; the actual separation, 2, is indicated by a dotted vertical line. Note that the red and black curves tend to overlap on the left side of plots, and the red and blue curves tend to overlap on the right side of plots.

regular EM. However, quite intuitively, when the two-sided constraint does not include the actual separation inside, the performance of the two-sided constrained EM is significantly worse than the two other algorithms before reaching actual separation value, and the same as the one-sided constrained EM after passing the actual separation value.

2.5.2 Real dataset

In this section we apply our constrained EM Algorithm to the 2010 National Youth Physical Activity and Nutrition Study (NYPANS) dataset provided in (Centers for Disease Control and Prevention, 2010). The dataset consists of measurements on high school students in the United States in 2010. Since the height of students in each grade follows an approximate Gaussian distribution, and there are 4 grades, the height of all the students could form a GMM with $K = 4$. We focus on estimating the parameters of this GMM.

As prior knowledge, we use information from the data table of stature-for-age charts provided by National Center for Health Statistics in (Centers for Disease Control and Prevention,

Table 2.6. Errors and Rand Indices of EM Algorithm and Constrainted EM Algorithm.

Model	Criteria	Regular EM	Constrained EM
NYPANS	center error (times 100)	6.7	0.5
	all parameters error (times 100)	15.6	3.2
	Rand index	0.566	0.596

2001). The two sources are both gathered in the United States, with a time difference of 9 years. From the mean age of students in the NYPANS study in each of the 4 grades, we found the most relevant median height, and in Gaussian distributions the mean height, in these charts, and subtracted the adjacent values. To account for possible slight changes over the years, we add and subtract respectively 0.005, in meters, to each of the differences. The resulting prior knowledge, in units of meters, gives us: $\delta_{1,1} = 0.019023, \delta_{2,1} = 0.00895, \delta_{3,1} = 0.00154$ and $\delta_{1,2} = 0.029023, \delta_{2,2} = 0.01895, \delta_{3,2} = 0.01154$. It is clear from Table 2.6 that the constrained EM, with proper prior knowledge, outperforms regular EM in both parameter estimation and Rand index, and therefore significantly improves on the ability to uncover the underlying components.

2.6 Acknowledgement

This chapter, in full, is a version of the paper "*K*-Means and Gaussian Mixture Modeling with a Separation Constraint", He Jiang and Ery Arias-Castro. It is currently being prepared for submission for publication of the material. The dissertation author is the primary investigator and corresponding author of this material.

Chapter 3

Extending the Patra–Sen Approach to Estimating the Background Component in a Two-Component Mixture Model

3.1 Abstract

Patra and Sen (2016) consider a two-component mixture model, where one component plays the role of background while the other plays the role of signal, and propose to estimate the background component by simply 'maximizing' its weight. While in their work the background component is a completely known distribution, we extend their approach here to three emblematic settings: when the background distribution is symmetric; when it is monotonic; and when it is log-concave. In each setting, we derive estimators for the background component, establish consistency, and provide a confidence band. While the estimation of a background component is straightforward when it is taken to be symmetric or monotonic, when it is log-concave its estimation requires the computation of a largest concave minorant, which we implement using sequential quadratic programming. Compared to existing methods, our method has the advantage of requiring much less prior knowledge on the background component, and is thus less prone to model misspecification. We illustrate this methodology on a number of synthetic and real datasets.

3.2 Introduction

3.2.1 Two component mixture models

Among mixture models, two-component models play a special role. In robust statistics, they are used to model contamination, with the main component representing the inlier distribution, while the remaining component representing the outlier distribution (Hettmansperger and McKean, 2010; Huber, 1964; Huber and Ronchetti, 2009; Tukey, 1960). In that kind of setting, the contamination is a nuisance and the goal is to study how it impacts certain methods for estimation or testing, and also to design alternative methods that behave comparatively better in the presence of contamination.

In multiple testing, the background distribution plays the role of the distribution assumed (in a simplified framework) to be common to all test statistics under their respective null hypotheses, while the remaining component plays the role of the distribution assumed of the test statistics under their respective alternative hypotheses (Efron et al., 2001; Genovese and Wasserman, 2002). In an ideal situation where the p-values can be computed exactly and are uniformly distributed on $[0,1]$ under their respective null hypotheses, the background distribution is the uniform distribution on $[0,1]$. Compared to the contamination perspective, here the situation is in a sense reverse, as we are keenly interested in the component other than the background component. We adopt this multiple testing perspective in the present work.

3.2.2 The Patra–Sen approach

Working within the multiple testing framework, Patra and Sen (2016) posed the problem of estimating the background component as follows. They operated under the assumption that the background distribution is completely known — a natural choice in many practical situations, see for example the first two situations in Section 3.3.3. Given a density f representing the density of all the test statistics combined, and letting g_0 denote a completely known density,

define

$$\theta_0 := \sup\{t : f \geq t g_0\}. \tag{3.1}$$

Note that $\theta_0 \in [0,1]$. Under some mild assumptions on f, the supremum is attained, so that f can be expressed as the following two-component mixture:

$$f = \theta_0 g_0 + (1 - \theta_0) u, \tag{3.2}$$

for some density u. Patra and Sen (2016) aim at estimating θ_0 defined in (3.1) based on a sample from the density f, and implement a slightly modified plug-in approach. Even in this relatively simple setting where the background density — the completely known density g_0 above — is given, information on θ_0 can help improve inference in a multiple testing situation as shown early on by Storey (2002), and even earlier by Benjamini and Hochberg (2000).

3.2.3 Our contribution

We find the Patra–Sen approach elegant, and in the present work extend it to settings where the background distribution (also referred to as the null distribution) — not just the background proportion — is unknown. For an approach that has the potential to be broadly applicable, we consider three emblematic settings where the background distribution is in turn assumed to be symmetric (Section 3.3), monotone (Section 3.4), or log-concave (Section 3.5). Each time, we describe the estimator for the background component (proportion and density) that the Patra-Sen approach leads to, and study its consistency and numerical implementation. We also provide a confidence interval for the background proportion and a simultaneous confidence band for the background density. In addition, in the log-concave setting, we provide a way of computing the largest concave minorant. We address the situation where the background is specified incorrectly, and mention other extensions, including combinations of these settings and in multivariate settings, in Section 3.6.

3.2.4 More related work in multiple testing

The work of Patra and Sen (2016) adds to a larger effort to estimate the proportion and/or the density of the null component in a multiple testing scenario. This effort dates back, at least, to early work on false discovery control (Benjamini and Hochberg, 1995) where (over-)estimating the proportion of null hypotheses is crucial to controlling the FDR and related quantities (Benjamini and Hochberg, 2000; Genovese and Wasserman, 2004; Storey, 2002).

Directly focusing on the estimation of the null proportion, Langaas et al. (2005) consider a setting where the p-values are uniform in $[0,1]$ under their null hypotheses and have a monotone decreasing density under their alternative hypotheses, while Meinshausen and Rice (2006) do not assume anything of the alternative distribution and propose an estimator which is similar in spirit to that of Patra and Sen (2016). Jin and Cai (2007) and Jin (2008) consider a Gaussian mixture model and approach the problem via the characteristic function — a common approach in deconvolution problems. Gaussian mixtures are also considered in (Cai and Jin, 2010; Efron, 2007, 2012), where the Gaussian component corresponding to the null has unknown parameters that need to be estimated.

Some references to estimating the null component that have been or could be applied in the context of multiple testing are given in (Efron, 2012, Ch 5). Otherwise, we are also aware of the very recent work of Roquain and Verzelen (2020), where in addition to studying the 'cost' of having to estimate the parameters of the null distribution when assumed Gaussian, also consider the situation where null distribution belongs to a given location family, and further, propose to estimate the null distribution under an upper bound constraint on the proportion of non-nulls in the mixture model.

Remark 1. Much more broadly, all this connects with the vast literature on Gaussian mixture models (Cohen, 1967; Lindsay and Basak, 1993) and on mixture models in general (Lindsay, 1995; McLachlan and Basford, 1988; McLachlan et al., 2019; McLachlan and Peel, 2004), including two-component models (Bordes et al., 2006; Gadat et al., 2020; Ma and Yao, 2015;

Shen et al., 2018).

3.3 Symmetric background component

We start with what is perhaps the most natural nonparametric class of null distributions: the class of symmetric distributions about the origin. Unlike Roquain and Verzelen (2020), who assume that the null distribution is symmetric around an unknown location that needs to be estimated but is otherwise known, i.e., its 'shape' is known, we assume that the shape is unknown. We do assume that the center of symmetry is known, but this is for simplicity, as an extension to an unknown center of symmetry is straightforward (see our numerical experiments in Section 3.3.2). Mixtures of symmetric distributions are considered in (Hunter et al., 2007), but otherwise, we are not aware of works estimating the null distribution under an assumption of symmetry in the context of multiple testing. For works in multiple testing that assume that the null distribution is symmetric but unknown, but where the goal is either testing the global null hypothesis or controlling the false discovery rate, see (Arias-Castro and Chen, 2017; Arias-Castro and Wang, 2017).

Following the footsteps of Patra and Sen (2016), we make sense of the problem by defining for a density f the following:

$$\pi_0 := \sup\left\{\pi : \exists g \in \mathcal{S} \text{ s.t. } f - \pi g \geq 0 \text{ a.e.}\right\}, \tag{3.3}$$

where $\mathcal{S}$ is the class of even densities (i.e., representing a distribution that is symmetric about the origin). Note that $\pi_0 \in [0,1]$ is well-defined for any density f, with $\pi_0 = 1$ if and only if f itself is symmetric.

Theorem 1. *We have*

$$\pi_0 = \int_{-\infty}^{\infty} h_0(x)dx, \qquad h_0(x) := \min\{f(x), f(-x)\}. \tag{3.4}$$

Moreover, if $\pi_0 > 0$ *the supremum in* (3.3) *is attained by the following density and no other*[1] *:*

$$g_0(x) := \frac{h_0(x)}{\pi_0}. \tag{3.5}$$

Proof. The parameter π_0 can be equivalently defined as

$$\pi_0 = \sup\Big\{ \int h : h \text{ is even and } 0 \le h \le f \text{ a.e.}\Big\}. \tag{3.6}$$

Note that h_0, as defined in the statement, satisfies the above conditions, implying that $\pi_0 \ge \int h_0$. Take h satisfying these same conditions, namely, $h(x) = h(-x)$ and $0 \le h(x) \le f(x)$ for almost all x. Then, for almost any x, $h(x) \le f(x)$ and $h(-x) \le f(-x)$, implying that $h(x) \le f(x) \wedge f(-x) = h_0(x)$. (Here and elsewhere, $a \wedge b$ is another way of denoting $\min(a,b)$.) Hence, $\int h \le \int h_0$ with equality if and only if $h = h_0$ a.e., in particular implying that $\pi_0 \le \int h_0$. We have thus established that $\pi_0 = \int h_0$, and also that $\int h = \pi_0$ if and only if $h = h_0$ a.e.. This not only proves (3.4), but also (3.5), essentially by definition. □

We have thus established that, in the setting of this section, the background component as defined above is given by

$$h_0(x) = \pi_0 g_0(x) = \min\{f(x), f(-x)\}, \tag{3.7}$$

and f can be expressed as a mixture of the background density and another, unspecified, density u, as follows:

$$f = \pi_0 g_0 + (1 - \pi_0)u. \tag{3.8}$$

The procedure is summarized in Table 3.1. An illustration of this decomposition is shown in Figure 3.1. By construction, the density u is such that it has no symmetric background component in that, for almost every x, $u(x) = 0$ or $u(-x) = 0$.

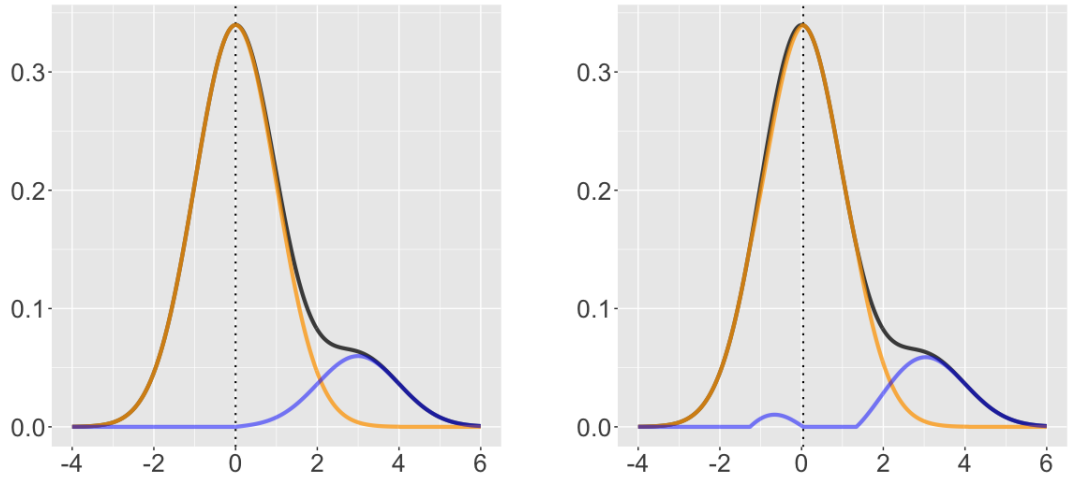
[1] As usual, densities are understood up to sets of zero Lebesgue measure.

Table 3.1. Symmetric background computation.

inputs: density f, given center of symmetry c_0 or candidate center points $\{c_1, c_2, \ldots, c_k\}$

if center of symmetry is not provided **then**
  for $i = 1, \ldots, k$ **do**
  $h_i(x) = \min\{f(x), f(2c_i - x)\}$
  $\pi_i(x) = \int_{-\infty}^{\infty} h_i(x)dx$
  $\beta = \arg\max_i \pi_i(x)$
  $c_0 = c_\beta$
$h_0(x) = \min\{f(x), f(2c_0 - x)\}$
$\pi_0 = \int_{-\infty}^{\infty} h_0(x)dx$
$g_0(x) = h_0(x)/\pi_0$

return c_0, π_0, g_0, h_0



(a) Decomposition of f with center of symmetry specified as 0 (dotted).

(b) Decomposition of f with center of symmetry, 0.04 (dotted), found by maximization.

Figure 3.1. The density f of the Gaussian mixture $0.85\,\mathcal{N}(0,1) + 0.15\,\mathcal{N}(3,1)$, in black, and its decomposition into $\pi_0 g_0$, in orange, and $(1-\pi_0)u$, in blue. We specify the center of symmetry as 0 on the left, and we do not specify the center of symmetry on the right. Notice $\pi_0 = 0.850$ on the left and $\pi_0 = 0.860$ on the right.

3.3.1 Estimation and consistency

When all we have to work with is a sample — $x_1, x_2, \ldots, x_n \in \mathbb{R}$ — we adopt a straightforward plug-in approach: We estimate the density f, obtaining $\hat{f}$, and apply the procedure of Table 3.1, meaning, we compute $\hat{h}_0(x) := \min\{\hat{f}(x), \hat{f}(-x)\}$. If we want estimates for the background density and proportion, we simply return $\hat{\pi}_0 := \int \hat{h}_0$ and $\hat{g}_0 := \hat{h}_0/\hat{\pi}_0$. (By convention, we set $\hat{g}_0$ to the standard normal distribution if $\hat{\pi}_0 = 0$.)

We say that $\hat{f} = \hat{f}_n$ is locally uniformly consistent for f if $\mathbb{E}[\operatorname{ess\,sup}_{x \in I} |\hat{f}_n(x) - f(x)|] \to 0$ as $n \to \infty$ for any bounded interval I. (Here and elsewhere, $\operatorname{ess\,sup}_{x \in I} f(x)$ denotes the essential supremum of f over the set I.) We note that this consistency condition is satisfied, for example, when f is continuous and $\hat{f}$ is the kernel density estimator with the Gaussian kernel and bandwidth chosen by cross-validation (Chow et al., 1983).

Theorem 2. *Suppose that $\hat{f}$ is a true density and locally uniformly consistent for f. Then $\hat{h}_0$ is locally uniformly consistent for h_0 and $\hat{\pi}_0$ is consistent, and if $\pi_0 > 0$, then $\hat{g}_0$ is locally uniformly consistent for g_0.*

Proof. All the limits that follows are as the sample size diverges to infinity. We rely on the elementary fact that, for $a_1, a_2, b_1, b_2 \in \mathbb{R}$,

$$\left|\min\{a_1, b_1\} - \min\{a_2, b_2\}\right| \leq \max\{|a_1 - a_2|, |b_1 - b_2|\}, \tag{3.9}$$

to get that, for all x,

$$|\hat{h}_0(x) - h_0(x)| \leq \max\{|\hat{f}(x) - f(x)|, |\hat{f}(-x) - f(-x)|\}, \tag{3.10}$$

implying that $\hat{h}_0$ is locally uniformly consistent for h_0. To be sure, take a bounded interval I,

which we assume to be symmetric without loss of generality. Then

$$\begin{aligned}\operatorname*{ess\,sup}_{x\in I}|\hat{h}_0(x)-h_0(x)| &\le \operatorname*{ess\,sup}_{x\in I}|\hat{f}(x)-f(x)| \vee \operatorname*{ess\,sup}_{x\in I}|\hat{f}(-x)-f(-x)| \\ &= \operatorname*{ess\,sup}_{x\in I}|\hat{f}(x)-f(x)|,\end{aligned}$$

and we then use the fact that $\mathbb{E}[\operatorname{ess\,sup}_I|\hat{f}-f|]\to 0$. (Here and elsewhere, $a\vee b$ is another way of denoting $\max(a,b)$.)

To conclude, it suffices to show that $\hat{\pi}_0$ is consistent for π_0. Fix $\varepsilon>0$ arbitrarily small. There is a bounded interval I such that $\int_I f\ge 1-\varepsilon$. Then, by the fact that $0\le h_0\le f$ a.e.,

$$\int_I h_0 \le \pi_0 = \int h_0 \le \int_I h_0 + \int_{I^c} f \le \int_I h_0 + \varepsilon.$$

(I^c denotes the complement of I, meaning, $I^c=\mathbb{R}\setminus I$.) Similarly, by the fact that $0\le\hat{h}_0\le\hat{f}$ a.e.,

$$\int_I \hat{h}_0 \le \hat{\pi}_0 = \int \hat{h}_0 \le \int_I \hat{h}_0 + \int_{I^c}\hat{f}.$$

From this we gather that

$$\int_I \hat{h}_0 - \int_I h_0 - \varepsilon \le \hat{\pi}_0-\pi_0 \le \int_I \hat{h}_0 - \int_I h_0 + \int_{I^c}\hat{f}.$$

Thus consistency of $\hat{\pi}_0$ follows if we establish that $\limsup\int_{I^c}\hat{f}\le\varepsilon$ and that $\int_I\hat{h}_0-\int_I h_0\to 0$. The former comes from the fact that

$$\begin{aligned}\int_{I^c}\hat{f} &= \int_{I^c}\hat{f}-\int_{I^c}f+\int_{I^c}f \\ &\le \int_I(\hat{f}-f)+\varepsilon \\ &\le |I|\operatorname*{ess\,sup}_I|\hat{f}-f|+\varepsilon\to\varepsilon,\end{aligned}$$

using the fact that f and $\hat{f}$ are densities and that $\int_{I^c} f \le \varepsilon$. ($|I|$ denotes the Lebesgue measure of I, meaning its length when I is an interval.) For the latter, we have

$$\left|\int_I \hat{h}_0 - \int_I h_0\right| \le \int_I |\hat{h}_0 - h_0| \qquad\qquad \le |I| \operatorname*{ess\,sup}_I |\hat{h}_0 - h_0| \to 0,$$

having already established that $\hat{h}_0$ is locally uniformly consistent for h_0. □

Confidence interval and confidence band

Beyond mere pointwise consistency, suppose that we have available a confidence band for f, which can be derived under some conditions on f from a kernel density estimator — see (Chen, 2017) or (Giné and Nickl, 2021, Ch 6.4).

Theorem 3. *Suppose that for some $\alpha \in (0,1)$, we have at our disposal $\hat{f}_l$ and $\hat{f}_u$ such that*

$$\mathbb{P}\left(\hat{f}_l(x) \le f(x) \le \hat{f}_u(x), \textit{ for almost all } x\right) \ge 1 - \alpha. \tag{3.11}$$

Then, with probability at least $1 - \alpha$,

$$\hat{\pi}_l \le \pi_0 \le \hat{\pi}_u, \qquad\qquad \hat{g}_l \le g_0 \le \hat{g}_u \textit{ a.e.}, \tag{3.12}$$

where

$$\hat{\pi}_l := \int \hat{h}_l, \quad \hat{\pi}_u := \int \hat{h}_u, \quad \hat{h}_l(x) := \min\{\hat{f}_l(x), \hat{f}_l(-x)\}, \quad \hat{h}_u(x) := \min\{\hat{f}_u(x), \hat{f}_u(-x)\}.$$

Proof. Let Ω be the event that $\hat{f}_l(x) \le f(x) \le \hat{f}_u(x)$ for almost all x, and note that $\mathbb{P}(\Omega) \ge 1 - \alpha$ by assumption. Assuming that Ω holds, we have, for almost all x,

$$\hat{f}_l(x) \le f(x) \le \hat{f}_u(x), \qquad\qquad \hat{f}_l(-x) \le f(-x) \le \hat{f}_u(-x),$$

and taking the minimum in the corresponding places yields

$$\hat{h}_l(x) \le h_0(x) \le \hat{h}_u(x). \tag{3.13}$$

Everything else follows immediately from this. □

In words, we apply the procedure of Table 3.1 to the lower and upper bounds, $\hat{f}_l$ and $\hat{f}_u$. If the center of symmetry is not provided, then we determine it based on $\hat{f}$, and then use that same center for $\hat{f}_l$ and $\hat{f}_u$.

3.3.2 Numerical experiments

In this subsection we provide simulated examples when the background distribution is taken to be symmetric. We acquire $\hat{f}$ by kernel density estimation using the Gaussian kernel with bandwidth selected by cross-validation (Arlot and Celisse, 2010; Rudemo, 1982; Sheather and Jones, 1991; Silverman, 1986; Stone, 1984), where the cross-validated bandwidth selection is implemented in the kedd package (Guidoum, 2020). The consistency of this density estimator has been proven in (Chow et al., 1983). Although there are many methods that provide confidence bands for kernel density estimator, for example (Bickel and Rosenblatt, 1973; Giné and Nickl, 2010), for consideration of simplicity and intuitiveness, our simultaneous confidence band (in the form of $\hat{f}_l$ and $\hat{f}_u$) used in the experiments is acquired from bootstrapping a debiased estimator of the density as proposed in (Cheng and Chen, 2019). For a comprehensive review on the area of kernel density and confidence bands, we point the reader to the recent survey paper (Chen, 2017) and textbook (Giné and Nickl, 2021, Ch 6.4).

In the experiments below, we carry out method and report the estimated proportion $\hat{\pi}_0$, as well as its 95% confidence interval $(\hat{\pi}_0^{\mathrm{L}}, \hat{\pi}_0^{\mathrm{U}})$. We consider both the situation where the center of symmetry is given, and the situation where it is not. In the latter situation, we also report the background component's estimated center, which is selected among several candidate centers and chosen as the one giving the largest symmetric background proportion.

We are not aware of other methods for estimating the quantity π_0, but we provide a comparison with several well-known methods that estimate similar quantities. To begin with, we consider the method of Patra and Sen (2016), which estimates the quantity θ_0 as defined in (3.1).

We let $\hat{\theta}_0^{\mathrm{PSC}}$, $\hat{\theta}_0^{\mathrm{PSH}}$, $\hat{\theta}_0^{\mathrm{PSB}}$ denote the constant, heuristic, and 95% upper bound estimator on θ_0, respectively. We then consider estimators of the quantity θ, the actual proportion of the given background component f_b in the mixture density

$$f = \theta f_b + (1 - \theta)u, \tag{3.14}$$

where u denotes the unknown component. Note that θ_0 and π_0 may be different from θ. This mixture model may not be identifiable in general, and we discuss this issue down below. We also consider the estimator of (Efron, 2007), denoted $\hat{\theta}^{\mathrm{E}}$, and implemented in package `locfdr` [2]. This method requires the unknown component to be located away from 0, and to be have heavier tails than the background component. In addition, when the p-values are known, Meinshausen and Rice (2006) provide a 95% upper bound on the proportion of the null component, and we include that estimator also, denoted $\hat{\theta}^{\mathrm{MR}}$, and implemented in package `howmany` [3]. Finally, when the distribution is assumed to be a Gaussian mixture, Cai and Jin (2010) provide an estimator, denoted $\hat{\theta}^{\mathrm{CJ}}$, when the unknown component is assumed to have larger standard deviation than the background component. $\hat{\theta}^{\mathrm{CJ}}$ requires the specification of a parameter γ, and following the advice given by the authors, we select $\gamma = 0.2$. Importantly, unlike our method, these other methods assume knowledge of the background distribution. (Note that the methods of Efron (2007) and Cai and Jin (2010) do not necessitate full knowledge of the background distribution, but we provide them with that knowledge in all the simulated datasets.) We summarize the methods used in our experiments in Table 3.2. For experiments in situations where the background component is misspecified, we invite the reader to Section 3.6.1.

[2] https://cran.r-project.org/web/packages/locfdr/index.html
[3] https://cran.r-project.org/web/packages/howmany/howmany.pdf

Table 3.2. Summary of the methods considered in our experiments.

Estimator	Reference	Description	Background information needed
$\hat{\pi}_0, \hat{\pi}_0^{\mathrm{L}}, \hat{\pi}_0^{\mathrm{U}}$	Current paper	Estimator of π_0 with 95% lower and upper confidence bounds.	Requires background distribution to be symmetric (note the requirement becomes monotonic in Section 3.4 or log-concave in Section 3.5).
$\hat{\theta}_0^{\mathrm{PSC}}$, $\hat{\theta}_0^{\mathrm{PSH}}$, $\hat{\theta}_0^{\mathrm{PSB}}$	(Patra and Sen, 2016)	Constant, heuristic, and 95% upper bound estimates of θ_0.	Requires complete knowledge on the background distribution.
$\hat{\theta}_0^{\mathrm{E}}$	(Efron, 2007)	Estimator of θ.	Either requires full knowledge of background distribution or can estimate the background distribution when it has shape similar to a Gaussian distribution centered around 0.
$\hat{\theta}_0^{\mathrm{MR}}$	(Meinshausen and Rice, 2006)	95% upper bound of θ.	Requires complete knowledge on the background distribution.
$\hat{\theta}_0^{\mathrm{CJ}}$	(Cai and Jin, 2010)	Estimator of θ.	Either requires full knowledge of background distribution or can estimate the background distribution when it is Gaussian.

We consider four different situations as listed in Table 3.3. Each situation's corresponding θ, θ_0, and π_0, defined as in (3.3), are also presented, where θ_0 and π_0 are obtained numerically based on knowledge of f. For each model, we generate a sample of size $n = 1000$ and compute all the estimators described above. We repeat this process 1000 times. We transform the data accordingly when applying the comparison methods. The result of our experiment are reported, in terms of the mean values as well as standard deviations, in Table 3.4. It can be seen that in most situations, our estimator achieves comparable if not better performance when estimating π_0 as compared to the other methods for the parameter they are meant to estimate (θ or θ_0). We also note that our method is significantly influenced by the estimation of $\hat{f}$, therefore in situations where $\hat{f}$ deviates from f often, our estimator will likely result in higher error. In addition, it is clear from the experiments that specifying the center of symmetry is unnecessary.

Table 3.3. Simulated situations for the estimation of a symmetric background component, together with the corresponding values of θ, θ_0, π_0 (unspecified center), and π_{00} (given center), obtained numerically (and rounded at 3 decimals).

Model	**Distribution**	θ	θ_0	π_0	π_{00}
S1	$0.85\,\mathcal{N}(0,1)+0.15\,\mathcal{N}(3,1)$	0.850	0.850	0.860	0.850
S2	$0.95\,\mathcal{N}(0,1)+0.05\,\mathcal{N}(3,1)$	0.950	0.950	0.950	0.950
S3	$0.85\,\mathcal{N}(0,1)+0.1\,\mathcal{N}(2.5,0.75)+0.05\,\mathcal{N}(-2.5,0.75)$	0.850	0.851	0.954	0.950
S4	$0.85\,\mathcal{N}(0,1)+0.1\,\mathcal{N}(2.5,0.75)+0.05\,\mathcal{N}(5,0.75)$	0.850	0.850	0.858	0.850

3.3.3 Real data analysis

In this subsection we examine six real datasets where the null component could be reasonably assumed to be symmetric.

We begin with two datasets where we have sufficient information on the background component. The first one is the Prostate dataset (Singh et al., 2002), which contains gene expression levels for $n = 6033$ genes on 102 men, 50 of which are control subjects and 52 are prostate cancer patients. The main objective is to discover the genes that have a different

Table 3.4. A comparison of various methods for estimating a background component in the situations of Table 3.3. For our method, the first and second rows in each situation are for when the center is unspecified, while the third and fourth rows are for when the center is specified to be the origin. Thus $\hat{\pi}_0$ on the first row of each situation is compared with π_0, $\hat{\pi}_0$ on the third row of each situation is compared with π_{00}. Otherwise, the $\hat{\theta}_0^{\mathrm{X}}$ are compared with θ_0, while the $\hat{\theta}^X$ are compared with θ.

Model	Center	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$	**Null**	$\hat{\theta}_0^{\mathrm{PSC}}$	$\hat{\theta}_0^{\mathrm{PSH}}$	$\hat{\theta}_0^{\mathrm{PSB}}$	$\hat{\theta}^{\mathrm{E}}$	$\hat{\theta}^{\mathrm{MR}}$	$\hat{\theta}^{\mathrm{CJ}}$
S1	0.068	0.857	0.571	1	$\mathcal{N}(0,1)$	0.848	0.855	0.893	0.861	0.890	0.893
	(0.052)	(0.021)	(0.059)	(0)		(0.024)	(0.023)	(0.018)	(0.023)	(0.014)	(0.085)
	0	0.835	0.561	1							
	given	(0.022)	(0.058)	(0)							
S2	0.024	0.936	0.631	1	$\mathcal{N}(0,1)$	0.938	0.954	0.985	0.955	0.972	0.963
	(0.044)	(0.017)	(0.066)	(0)		(0.023)	(0.022)	(0.015)	(0.026)	(0.008)	(0.082)
	0	0.925	0.623	1							
	given	(0.019)	(0.067)	(0)							
S3	0.046	0.945	0.597	1	$\mathcal{N}(0,1)$	0.864	0.942	0.937	0.856	0.896	0.707
	(0.050)	(0.019)	(0.063)	(0)		(0.023)	(0.034)	(0.017)	(0.024)	(0.015)	(0.085)
	0	0.930	0.587	1							
	given	(0.020)	(0.064)	(0)							
S4	0.070	0.856	0.574	1	$\mathcal{N}(0,1)$	0.846	0.854	0.891	0.849	0.889	0.713
	(0.056)	(0.021)	(0.061)	(0)		(0.023)	(0.024)	(0.018)	(0.024)	(0.014)	(0.088)
	0	0.833	0.563	1							
	given	(0.022)	(0.060)	(0)							

expression level on the control and prostate patient groups. For each gene, we conduct a two-sided two sample t test on the control subjects and prostate patients, and then transform these t statistics into z values, using

$$z_i = \Phi^{-1}(F_{100}(t_i)),\ i = 1, 2, \ldots, 6033, \tag{3.15}$$

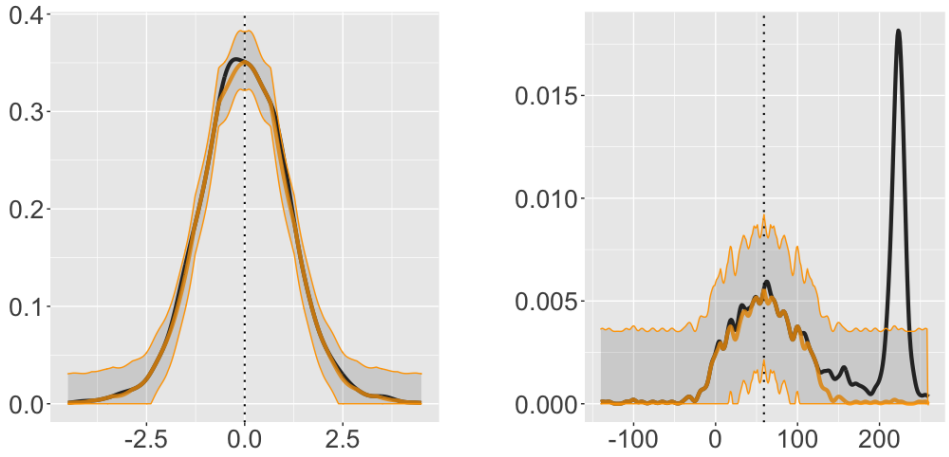
where Φ denotes the cdf of the standard normal distribution, and F_{100} denotes the cdf of the t distribution with 100 degrees of freedom. We work with these $n = 6033$ z values. From (Efron, 2007) the background component here could be reasonably assumed to be $\mathcal{N}(0,1)$. The results of the different proportion estimators compared in Section 3.3.2 are shown in the first row of Table 3.5. The fitted largest symmetric component as well as confidence bands are plotted in Figure 3.2(a).

Next we consider the Carina dataset (Walker et al., 2007), which contains the radial velocities of $n = 1266$ stars in Carina, a dwarf spheroidal galaxy, mixed with those of Milky Way stars in the field of view. As Patra and Sen (2016) stated, the background distribution of the radial velocity, bgstars, can be acquired from (Robin et al., 2003). The various estimators are computed and shown in the second row of Table 3.5. The fitted largest symmetric component as well as confidence bands are plotted in Figure 3.2(b).

Table 3.5. Two real datasets where background component can be reasonably guessed or derived. We compare the same methods for extracting a background symmetric component as in Section 3.3.2. (Note that we work with z values here instead of p-values so our results for the Prostate dataset are slightly different from those reported by Patra and Sen (2016).)

Model	Center	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$	**Null**	$\hat{\theta}_0^{\mathrm{PSC}}$	$\hat{\theta}_0^{\mathrm{PSH}}$	$\hat{\theta}_0^{\mathrm{PSB}}$	$\hat{\theta}^{\mathrm{E}}$	$\hat{\theta}^{\mathrm{MR}}$	$\hat{\theta}^{\mathrm{CJ}}$
Prostate	0	0.977	0.789	1	$\mathcal{N}(0,1)$	0.931	0.941	0.975	0.931	0.956	0.867
Carina	59	0.540	0.071	1	bgstars	0.636	0.645	0.677	0.951	0.664	0.206

Aside from these two real datasets where we know the background distribution, we consider four other real datasets — three microarray datasets and one police dataset — where



(a) Prostate dataset (b) Carina dataset

Figure 3.2. Estimated symmetric component on the Prostate (z values) and Carina (radial velocity) datasets: the black curve represents the fitted density; the center orange curve represents the computed $\hat{h}_0$; the top and bottom orange curves represent the 95% simultaneous confidence bands for h_0; the estimated center of the symmetric component is indicated by a dotted vertical line.

we do not know the null distribution (Efron, 2012). Here, out of the methods considered above, only our method and that of (Efron, 2007) and (Cai and Jin, 2010) are applicable. For the first comparison method we use the MLE estimation as presented in (Efron, 2007, Sec 4), which is usually very close to the result of Central Matching as used in (Efron, 2012). For the second comparison method we use the estimator in (Cai and Jin, 2010, Sec 3.2), although we use $\gamma = 0.1$ as the recommended value $\gamma = 0.2$ leads to significant underestimation of the background proportion. Note that both of these methods are still meant to estimate θ.

The HIV dataset (van't Wout et al., 2003) consists of a study of 4 HIV subjects and 4 control subjects. The measurements of $n = 7680$ gene expression levels were acquired using cDNA microarrays on each subject. We compute t statistics for the two sided t test and then transform them into z values using (3.15), with the degree of freedom being 6 here. We would like to know what proportion of these genes do not show a significant difference in expression levels between HIV and control subjects. The results are summarized in the first row of Table 3.6. The fitted largest symmetric component as well as confidence bands are shown in Figure 3.3(a).

The Leukemia dataset comes from (Golub et al., 1999). There are 72 patients in this study, of which 45 have ALL (Acute Lymphoblastic Leukemia) and 27 have AML (Acute Myeloid Leukemia), with AML being considered more severe. High density oligonucleotide microarrays gave expression levels on $n = 7128$ genes. Following (Efron, 2012, Ch 6.1), the raw expression levels on each microarray, $x_{i,j}$ for gene i on array j, were transformed to a normal score

$$y_{i,j} = \Phi^{-1}\Big(\big(\mathsf{rank}(x_{i,j}) - 0.5\big)/n\Big), \tag{3.16}$$

where $\mathsf{rank}(x_{i,j})$ denotes the rank of $x_{i,j}$ among n raw values of array j. Then t tests were then conducted on ALL and AML patients, and t statistics were transformed to z values according to (3.15), now with 70 degrees of freedom. As before, we would like to know the proportion of genes that do not show a significant difference in expression levels between ALL and AML patients. The results are summarized in the second row of Table 3.6. The fitted largest symmetric component as well as confidence bands are shown in Figure 3.3(b).

The Parkinson dataset comes from (Lesnick et al., 2007). In this dataset, substantia nigra tissue — a brain structure located in the mesencephalon that plays an important role in reward, addiction, and movement — from postmortem the brain of normal and Parkinson disease patients were used for RNA extraction and hybridization, done on Affymetrix microarrays. In this dataset, there are $n = 54277$ nucleotide sequences whose expression levels were measured on 16 Parkinson's disease patients and 9 control patients. We wish to find out the proportion of sequences that do not show significant difference between Parkinson and control patients. The results are summarized in the third row of Table 3.6. The fitted largest symmetric component as well as confidence bands are shown in Figure 3.3(c).

The Police dataset is analyzed in (Ridgeway and MacDonald, 2009). In 2006, based on 500000 pedestrian stops in New York City, each of the city's $n = 2749$ police officers that were regularly involved in pedestrian stops were assigned a z score on the basis of their stop data, in consideration of possible racial bias. For details on computing this z score, we refer the reader

to (Efron, 2012; Ridgeway and MacDonald, 2009). Large positive z values are considered as possible evidence of racial bias. We would like to know the percentage of these police officers that do not exhibit a racial bias in pedestrian traffic stops. The estimated proportion are reported on the last row of Table 3.6. The symmetric component as well as confidence bands are presented in Figure 3.3(d).

Table 3.6. Real datasets where background distribution is unknown and needs to be estimated. We compare the methods for extracting a background symmetric component among those in Section 3.3.2 that apply.

Model	Center	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$	$\hat{\theta}^{\mathrm{E}}$	$\hat{\theta}^{\mathrm{CJ}}$
HIV	-0.62	0.950	0.775	1	0.940	0.926
Leukemia	0.16	0.918	0.639	1	0.911	0.820
Parkinson	-0.18	0.985	0.924	1	0.998	0.993
Police	0.10	0.982	0.767	1	0.985	0.978

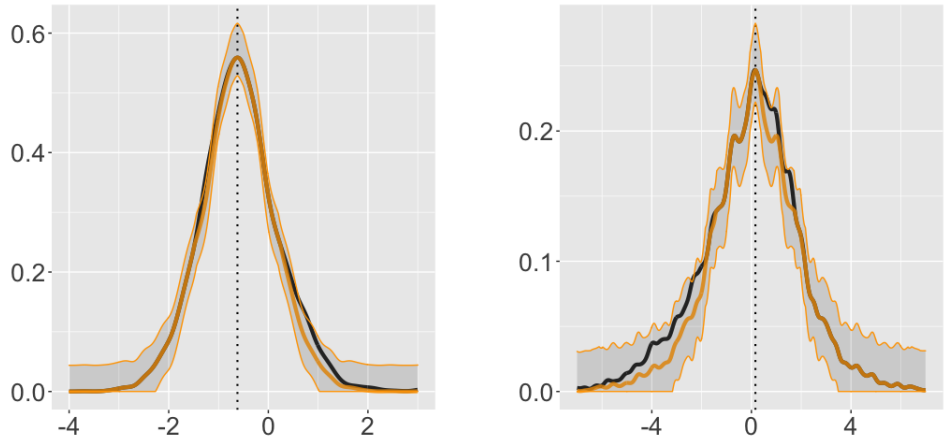
3.4 Monotone background component

In this section, we turn our attention to extracting from a density its monotone background component following the Patra–Sen approach. For this to make sense, we only consider densities supported on $\mathbb{R}_+ = [0, \infty)$. In fact, all the densities we consider in this section will be supported on $\mathbb{R}_+$. For such a density f, we thus define

$$\pi_0 := \sup\left\{\pi : \exists g \in \mathcal{M} \text{ s.t. } f - \pi g \geq 0 \text{ a.e.}\right\}, \tag{3.17}$$

where $\mathcal{M}$ is the class of monotone (necessarily non-increasing) densities on $\mathbb{R}_+$. Note that $\pi_0 \in [0,1]$ is well-defined for any density f, with $\pi_0 = 1$ if and only if f itself is monotone.

Recall that the essential infimum of a measurable set A, denoted $\operatorname{ess\,inf} A$, is defined as the supremum over $t \in \mathbb{R}$ such that $A \cap (-\infty, t)$ has Lebesgue measure zero. Everywhere in this section, we will assume that f is càdlàg, meaning that, at any point, it is continuous from the



(a) HIV dataset

(b) Leukemia dataset

(c) Parkinson dataset

(d) Police dataset

Figure 3.3. Estimated symmetric component on HIV (z values), Leukemia (z values), Parkinson (z values), and Police (z scores) datasets: the black curve represents the fitted density; the center orange curve represents the computed $\hat{h}_0$; the top and bottom orange curves represent the 95% simultaneous confidence bands for h_0; the estimated center of the symmetric component is indicated by a dotted vertical line.

right and admits a limit from the left.

Theorem 4. *Assuming f is càdlàg, we have*

$$\pi_0 = \int_0^\infty h_0(x)dx, \qquad h_0(x) := \operatorname{ess\,inf}\{f(y) : y \le x\}. \tag{3.18}$$

Moreover, if $\pi_0 > 0$ the supremum in (3.17) *is attained by the following density and no other:*

$$g_0(x) := \frac{h_0(x)}{\pi_0}. \tag{3.19}$$

Note that $\pi_0 = 0$ (i.e., f has no monotone background component) if and only if $\operatorname{ess\,inf}\{f(y) : y \le x\} = 0$ for some $x > 0$, or equivalently, if $\{x \in [0,t] : f(x) \le \varepsilon\}$ has positive measure for all $t > 0$ and all $\varepsilon > 0$. If f is càdlàg, this condition reduces to $f(0) = 0$. Also, if f is càdlàg, $h_0(x) = \min\{f(y) : y < x\}$.

Proof. Note that π_0 can be equivalently defined as

$$\pi_0 = \sup\left\{\int h : h \text{ is monotone and } 0 \le h \le f \text{ a.e.}\right\}. \tag{3.20}$$

Note that h_0, as defined in the statement, satisfies the above conditions, implying that $\pi_0 \ge \int h_0$. Take h satisfying these same conditions, namely, h is monotone and $0 \le h(x) \le f(x)$ for almost all x, say, for $x \in \mathbb{R}_+ \setminus A$ where A has Lebesgue measure zero. Take such an x. Then for any $y \le x$ we have $h(x) \le h(y)$, and $h(y) \le f(y)$ if in addition $y \notin A$. Hence,

$$h(x) \le \inf\{f(y) : y < x, y \notin A\} \le \operatorname{ess\,inf}\{f(y) : y < x\} = h_0(x),$$

where the second inequality comes from the fact that A has zero Lebesgue measure and the definition of essential infimum. Hence, $\int h \le \int h_0$ with equality if and only if $h = h_0$ a.e., in particular implying that $\pi_0 \le \int h_0$. We have thus established that $\pi_0 = \int h_0$, and also that $\int h = \pi_0$

if and only if $h = h_0$ a.e.. This not only proves (3.18), but also (3.19). □

We have thus established that, in the setting of this section where f is assumed to be càdlàg, the background component as defined above is given by

$$h_0(x) = \pi_0 g_0(x) = \operatorname{ess\,inf}\{f(y) : y < x\}, \tag{3.21}$$

and f can be expressed as a mixture of the background density and another, unspecified, density u, as follows:

$$f = \pi_0 g_0 + (1 - \pi_0)u. \tag{3.22}$$

The procedure is summarized in Table 3.7. An illustration of this decomposition is shown in Figure 3.4. (In this section, $\mathcal{E}(\sigma)$ denotes the exponential distribution with scale σ and $\mathcal{G}(\kappa, \sigma)$ denotes the Gamma distribution with shape κ and scale σ. Recall that $\mathcal{E}(\sigma) \equiv \mathcal{G}(1, \sigma)$.) By construction, the density u is such that it has no monotone background component in that $\operatorname{ess\,inf}\{u(y) : y < x\} = 0$ for any $x > 0$.

Table 3.7. Monotone background computation.

inputs: density f defined on $[0, \infty)$
$h_0(x) = \operatorname{ess\,inf}\{f(y) : y \le x\}$ $\pi_0 = \int_0^\infty h_0(x)dx$ $g_0(x) = h_0(x)/\pi_0$
return π_0, g_0, h_0

3.4.1 Estimation and consistency

In practice, when all we have is a sample of observations, $x_1, \ldots, x_n \in \mathbb{R}_+$, we first estimate the density, resulting in $\hat{f}$, and then compute the quantities defined in (3.18) and (3.19) with $\hat{f}$ in

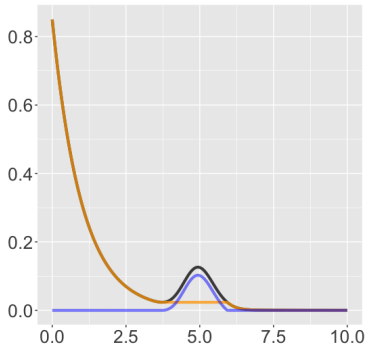


Figure 3.4. The density f of the Gamma mixture $0.85\,\mathcal{E}(1)+0.15\,\mathcal{G}(100,1/20)$, in black, and its decomposition into $\pi_0 g_0$, in orange, and $(1-\pi_0)u$, in blue. Notice that the orange curve has a flat part around the middle and is slightly different from $0.85\,\mathcal{E}(1)$.

place of f. Thus our estimates are

$$\hat{\pi}_0 := \int_0^\infty \hat{h}_0(x)dx, \qquad \hat{h}_0(x) := \operatorname{ess\,inf}\{\hat{f}(y) : y < x\}, \qquad \hat{g}_0(x) := \frac{\hat{h}_0(x)}{\hat{\pi}_0}. \tag{3.23}$$

Theorem 5. *Assume f is càdlàg, and suppose that $\hat{f}$ is a true density, càdlàg, and is locally uniformly consistent for f. Then $\hat{h}_0$ is locally uniformly consistent for h_0 and $\hat{\pi}_0$ is consistent for π_0, and if $\pi_0 > 0$, then $\hat{g}_0$ is locally uniformly consistent for g_0.*

Proof. From the definitions,

$$h_0(x) = \operatorname{ess\,inf}\{f(y) : y < x\}, \qquad \hat{h}_0(x) := \operatorname{ess\,inf}\{\hat{f}(y) : y < x\},$$

so that

$$|h_0(x) - \hat{h}_0(x)| \le \operatorname{ess\,sup}\{|f(y) - \hat{f}(y)| : y < x\},$$

further implying that

$$\operatorname*{ess\,sup}_{x<a} |h_0(x) - \hat{h}_0(x)| \le \operatorname*{ess\,sup}_{y<a} |f(y) - \hat{f}(y)|, \quad \text{for any } a > 0,$$

from which we get that $\hat{h}_0$ is locally uniformly consistent for h_0 whenever $\hat{f}$ is locally uniformly consistent for f.

For the remaining of the proof, we can follow in the footsteps of the proof of Theorem 2 based on the fact that, for any $a > 0$,

$$\left| \int_{[0,a]} h_0 - \int_{[0,a]} \hat{h}_0 \right| \leq \int_{[0,a]} |h_0 - \hat{h}_0| \leq a \operatorname*{ess\,sup}_{[0,a]} |h_0 - \hat{h}_0| \to 0,$$

where the limit is in expectation as the sample size increases. □

Confidence interval and confidence band

To go beyond point estimators, we suppose that we have available a confidence band for f and deduce from that a confidence interval for π_0 and a confidence band for g_0.

Theorem 6. *Suppose that we have a confidence band for f as in* (3.11)*. Then* (3.12) *holds with probability at least* $1 - \alpha$*, where*

$$\hat{\pi}_l := \int \hat{h}_l, \quad \hat{\pi}_u := \int \hat{h}_u, \quad \hat{h}_l(x) := \operatorname{ess\,inf}\{\hat{f}_l(y) : y < x\}, \quad \hat{h}_u(x) := \operatorname{ess\,inf}\{\hat{f}_u(y) : y < x\}.$$

The proof is straightforward and thus omitted.

Remark 2. So far, we have assumed that the monotone density is supported on $[0, \infty)$, but in principle we can also consider the starting point as unspecified. If this is the case, similar to what we did for the case of a symmetric component in Table 3.1, we can again consider several candidate locations defining the monotone component's support, and select the one yielding the largest monotone component weight.

3.4.2 Numerical experiments

We are here dealing with densities supported on $[0, \infty)$, and what happens near the origin is completely crucial as is transparent from the definition of $\hat{h}_0$. It is thus important in practice to choose an estimator for f that behaves well in the vicinity of the origin. As it is well known that

kernel density estimators have a substantial bias near the origin, we opted for a different estimator. Many density estimation methods have been proposed to deal with boundary effects, including smoothing splines (Gu, 1993; Gu and Qiu, 1993), local density estimation approaches (Fan, 1993; Hjort and Jones, 1996; Loader, 1996; Park et al., 2002), and local polynomial approximation methods (Cattaneo et al., 2019, 2020). For consideration of simplicity and intuitiveness, we consider kernel density estimation using a reflection about the boundary point (Cline and Hart, 1991; Karunamuni and Alberts, 2005; Schuster, 1985). We acquire $\hat{f}$ from (Schuster, 1985) and, as we did before, we acquire a 95% confidence band $[\hat{f}_l, \hat{f}_u]$ from (Cheng and Chen, 2019). We also note that $\hat{f}$ is consistent for f as shown by Schuster (1985).

We consider two different situations as listed in Table 3.8. Each situation's corresponding θ, θ_0, and π_0, defined as in (3.17), are also presented. We again generate a sample of size $n = 1000$ from each model, and repeat each setting 1000 times. The mean values as well as standard deviations of our method and related methods are reported in Table 3.9.

In the situation M1, our estimator achieves a smaller estimation error for π_0 than all other methods for their corresponding target, either θ or θ_0, even with much fewer information on the background component. In situation M2, our method has a slightly higher error than the error of $\hat{\theta}_0^{\mathrm{PSH}}$, $\hat{\theta}_0^{\mathrm{E}}$, both having complete information on the background component, but lower than that of $\hat{\theta}_0^{\mathrm{PSC}}$ and $\hat{\theta}_0^{\mathrm{CJ}}$.

Table 3.8. Simulated situations for the estimation of a monotone background component, together with the corresponding values of θ, θ_0, π_0, obtained numerically (and rounded at 3 decimals).

Model	**Distribution**	θ	θ_0	π_0
M1	$0.85\ \mathcal{E}(1) + 0.15\ \mathcal{G}(50, 1/10)$	0.850	0.850	0.922
M2	$0.95\ \mathcal{E}(1) + 0.05\ \mathcal{G}(50, 1/10)$	0.950	0.950	0.993

Table 3.9. A comparison of various methods for estimating a monotone background component in the situations of Table 3.8. As always, $\hat{\pi}_0$ is compared with π_0, the $\hat{\theta}_0^{\mathrm{X}}$ are compared with θ_0, while the $\hat{\theta}^X$ are compared with θ.

Model	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$	**Null**	$\hat{\theta}_0^{\mathrm{PSC}}$	$\hat{\theta}_0^{\mathrm{PSH}}$	$\hat{\theta}_0^{\mathrm{PSB}}$	$\hat{\theta}^{\mathrm{E}}$	$\hat{\theta}^{\mathrm{MR}}$	$\hat{\theta}^{\mathrm{CJ}}$
M1	0.920	0.460	1	$\mathcal{E}(1)$	0.843	0.854	0.889	0.841	0.866	0.533
	(0.020)	(0.053)	(0)		(0.022)	(0.022)	(0.018)	(0.028)	(0.013)	(0.085)
M2	0.984	0.519	1	$\mathcal{E}(1)$	0.936	0.953	0.984	0.954	0.964	0.842
	(0.012)	(0.061)	(0)		(0.022)	(0.020)	(0.014)	(0.028)	(0.009)	(0.083)

3.4.3 Real data analysis

In this subsection we consider a real dataset where the background component could be assumed to be monotonic nonincreasing. We look at the Coronavirus dataset (Dong et al., 2021), acquired from the *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE)* at Johns Hopkins University[4]. It is well known that new coronavirus cases are consistently decreasing in the USA currently, and this trend could be seen to begin on Jan 8, 2021 as shown by the New York Times interactive data[5]. For each person infected on or after Jan 8, 2021, we count the number of days between that day and the time they were infected. We are interested in quantifying how monotonic that downward trend in coronavirus infections is. As we do not know the actual background distribution here, and Gaussian distributions are of not particular relevance, we find that none of the other comparison methods in Section 3.3.2 are applicable, and therefore only provide our method's estimate in Table 3.10 and Figure 3.5. Numerically, it can be seen that the background monotonic component accounts for around 96.7% of the new cases arising on or after Jan 8, 2021.

[4] https://github.com/CSSEGISandData/COVID-19

[5] https://www.nytimes.com/interactive/2021/us/covid-cases.html

Table 3.10. Coronavirus dataset where it is of interest to gauge how monotonic the trend is starting in Jan 8, 2021.

Model	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$
Coronavirus	0.967	0.955	0.968

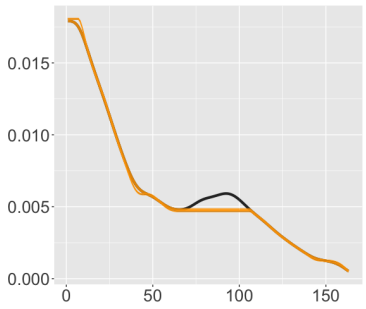


Figure 3.5. Estimated background monotone component on the Coronavirus dataset: the black curve represents the fitted density; the center orange curve represents the computed $\hat{h}_0$; and the top and bottom orange curves represent the 95% simultaneous confidence bands for h_0. (Due to the large amount of data here the confidence bands are very narrow.)

3.5 Log-concave background component

Our last emblematic setting is that of extracting a log-concave background component from a density. Log-concave densities have been widely studied in the literature (Samworth, 2018; Walther, 2009), and include a wide variety of distributions, including all Gaussian densities, all Laplace densities, all exponential densities, and all uniform densities. They have been extensively used in mixture models (Chang and Walther, 2007; Hu et al., 2016; Pu and Arias-Castro, 2020).

Following Patra and Sen (2016), for a density f we define

$$\pi_0 := \sup\left\{\pi : \exists g \in \mathcal{C} \text{ s.t. } f - \pi g \geq 0 \text{ a.e.}\right\}, \tag{3.24}$$

where $\mathcal{C}$ is the class of log-concave densities. Note that $\pi_0 \in [0,1]$, with $\pi_0 = 1$ if and only if f itself is log-concave.

Theorem 7. *π_0 is the value of following optimization problem:*

$$\begin{aligned} &\textit{maximize} \quad \int_{-\infty}^{\infty} h(x)dx \\ &\textit{over} \quad \big\{h : \mathbb{R} \to \mathbb{R}_+,\ \textit{log-concave},\ h \le f\big\}. \end{aligned} \tag{3.25}$$

Indeed, this problem admits a solution, although it may not be unique.

Proof. From the definition in (3.24), it is clear that

$$\pi_0 = \sup\Big\{\int h : h \text{ is log-concave and } 0 \le h \le f \text{ a.e.}\Big\}. \tag{3.26}$$

Thus we only need to show that the problem (3.25) admits a solution — and then show that it may not be unique to complete the proof. Here the arguments are a bit more involved.

Let (h_k) be a solution sequence to the problem (3.25), meaning that $h_k : \mathbb{R} \to \mathbb{R}_+$ is log-concave and satisfies $h_k \le f$, and that $q_k := \int h_k$ converges, as $k \to \infty$, to the value of the optimization problem (3.25), denoted q_* henceforth. Note that $0 \le q_k \le 1$ since $0 \le h_k \le f$ and $\int f = 1$, implying that $0 \le q_* \le 1$. We only need to consider the case where $q_* > 0$, for when $q_* = 0$ the constant function $h \equiv 0$ is a solution. Without loss of generality, we assume that $q_k > 0$ for all k.

Each h_k is log-concave, and from this we know the following: its support is an interval, which we denote $[a_k, b_k]$ with $-\infty \le a_k < b_k \le \infty$; h_k is continuous and strictly positive on (a_k, b_k); and $x \mapsto (\log h_k(y) - \log h_k(x))/(y-x)$ is non-increasing in $x \le y$ and $y \mapsto (\log h_k(y) - \log h_k(x))/(y-x)$ is non-increasing in $y \ge x$. Extracting a subsequence if needed, we assume without loss of generality that $a_k \to a$ and $b_k \to b$ as $k \to \infty$, for some $-\infty \le a \le b \le \infty$.

Define $F(t) = \int_{-\infty}^{t} f(x)dx$ and $H_k(t) = \int_{-\infty}^{t} h_k(x)dx$. We have that H_k is a non-decreasing, and if $s \le t$, $H_k(t) - H_k(s) \le F(t) - F(s)$, because $h_k \le f$. In particular, by Helly's selection theorem (equivalent to Prokhorov's theorem when dealing with distribution functions), we can extract a subsequence that converges pointwise. Without loss of generality, we assume that (H_k) itself

converges, and let H denote its limit. Note that H is constant outside of $[a,b]$. In fact, $H(x) = 0$ when $x < a$, because $x < a_k$ eventually, forcing $H_k(x) = 0$. For $x, x' > b$, we have that $x, x' > b_k$ for large enough k, implying that $H_k(x) = H_k(x')$, yielding $H(x) = H(x')$ by taking the limit $k \to \infty$. Note also that, for all $s \le t$,

$$H(t) - H(s) = \lim_{k\to\infty} (H_k(t) - H_k(s)) \le F(t) - F(s).$$

This implies that H is absolutely continuous with derivative, denoted h, satisfying $h \le f$ a.e..

We claim that h is a solution to (3.25). We already know that $0 \le h \le f$ a.e.. In addition, we also have $\int_{-\infty}^{\infty} h \ge q_*$. To see this, fix $\varepsilon > 0$, and let t be large enough that $\int_t^{\infty} f \le \varepsilon$. Because $h_k \le f$, we have $H_k(t) = q_k - \int_t^{\infty} h_k \ge q_k - \varepsilon$, implying that $H(t) \ge q_* - \varepsilon$ by taking the limit as $k \to \infty$. Hence, $\int_{-\infty}^{\infty} h = H(\infty) \ge H(t) \ge q^* - \varepsilon$, and $\varepsilon > 0$ being otherwise arbitrary, we deduce that $\int_{-\infty}^{\infty} h \ge q_*$. It thus remains to show that h is log-concave.

We establish this claim by proving that, extracting a subsequence if needed, h_k converges to h a.e., and it is enough to do so in an interval. Thus let x_0 and $\Delta > 0$ be such that $[x_0 - 4\Delta, x_0 + 4\Delta] \subset (a,b)$. Note that H is strictly increasing on (a,b), because each H_k is strictly increasing on (a_k,b_k) due to h_k being log-concave. Take any $x_0 - \Delta \le x < y \le x_0 + \Delta$ and let $\delta_k = (\log h_k(y) - \log h_k(x))/(y-x)$. We assume, for example, that $\delta_k \ge 0$, and bound it from above. Let $z_k \in [x_0 - 4\Delta, x_0 - 3\Delta]$ be such that $H_k(x_0 - 3\Delta) - H_k(x_0 - 4\Delta) = h_k(z_k)\Delta$, which exists by the mean-value theorem. Note that $h_k(z_k)\Delta \to \Delta_1 := H(x_0 - 3\Delta) - H(x_0 - 4\Delta)$, so that $h_k(z_k) \ge \Delta_2 := \Delta_1/2\Delta > 0$, eventually. Now, for any z in $[x_0 - 2\Delta, x_0 - \Delta]$, due to $z_k < z < x < y$ and $\log h_k$ being concave,

$$\frac{\log h_k(z) - \log h_k(z_k)}{z - z_k} \ge \frac{\log h_k(y) - \log h_k(x)}{y - x} = \delta_k,$$

which implies

$$h_k(z) \ge h_k(z_k) \exp(\delta_k(z - z_k)) \ge \Delta_2 \exp(\delta_k \Delta).$$

This being true for all such z, we have

$$1 \geq \int_{-\infty}^{\infty} f(z)\mathrm{d}z \geq \int_{-\infty}^{\infty} h_k(z)\mathrm{d}z \geq \int_{x_0-2\Delta}^{x_0-\Delta} h_k(z)\mathrm{d}z \geq \Delta \cdot \Delta_2 \exp(\delta_k \Delta),$$

allowing us to derive $\delta_k \leq M_1 := \Delta^{-1}\log(2/\Delta_1)$. We can deal with the case where $\delta_k < 0$ by symmetry, obtaining that, for all k sufficiently large and all $x_0 - \Delta \leq x < y \leq x_0 + \Delta$,

$$\left|\frac{\log h_k(y) - \log h_k(x)}{y - x}\right| \leq M_1.$$

Let $u_k = h_k(x_0)$. Because h_k is unimodal, either $h_k(x) \leq u_k$ for all $x \leq x_0$ or $h_k(x) \leq u_k$ for all $x \geq x_0$. Extracting a subsequence if needed, and by symmetry, assume that the former is true for all k large enough. Then

$$\begin{aligned}
\Delta_1 &= H(x_0 - 3\Delta) - H(x_0 - 4\Delta) \\
&= \lim_{k\to\infty} \left(H_k(x_0 - 3\Delta) - H_k(x_0 - 4\Delta)\right) \\
&= \lim_{k\to\infty} \int_{x_0-3\Delta}^{x_0-4\Delta} h_k(x)\mathrm{d}x \\
&\leq \liminf_{k\to\infty} \Delta u_k,
\end{aligned}$$

so that $u_k \geq \Delta_2 > 0$, eventually. Assuming so, we have $u_k h_k(x_0) \geq \Delta_1/\Delta \geq \Delta_2$. And we also have $h_k(x_0) \leq f(x_0)$, and together, $|\log h_k(x_0)| \leq M_2 := |\log(\Delta_2)| \vee |\log f(x_0)|$. With the triangle inequality, we thus have, for all $x \in [x_0 - \Delta, x_0 + \Delta]$,

$$|\log h_k(x)| \leq M_1|x - x_0| + |\log h_k(x_0)| \leq M_1\Delta + M_2 =: M_3.$$

The family of functions $(\log h_k)$ (starting at k large enough) is thus uniformly bounded and equicontinuous on $[x_0 - \Delta, x_0 + \Delta]$, so that by the Arzelà–Ascoli theorem, we have that $(\log h_k)$ is precompact for the uniform convergence on that interval. Therefore, the same is true for

(h_k). Let h_∞ be the uniform limit of a subsequence, and note that h_∞ is continuous. h_∞ must also be a weak limit as well, since uniform convergence on a compact interval implies weak convergence on that interval. Therefore $h_\infty = h$ a.e. on that interval, since (h_k) converges weakly to h. Hence, all the uniform limits of (h_k) must coincide with h a.e., and since any such limit must be continuous, it means that they are the same. We conclude that, on the interval under consideration, h is equal a.e. to a continuous function which is the (only) uniform limit of (h_k), and in particular, (h_k) converges pointwise a.e. to h on that interval.

Thus, we have proved that the optimization problem (3.25) has at least one solution. We now show that there may be multiple solutions. This is the case, for instance, when $f = \frac{1}{m} f_1 + \cdots + \frac{1}{m} f_m$ where each f_j is a log-concave density and these densities have support sets that are pairwise disjoint. In that case, any of the components, meaning any f_j, is a solution to (3.25), and these are the only solutions. This comes from the fact that the support of a log-concave distribution is necessarily an interval. □

We have thus established that a density has at least one log-concave background component, and possibly multiple ones, corresponding to the solutions to (3.25). If h is one such solution, then $\pi_0 = \int h$ and we may define the corresponding density as $g = h/\pi_0$ if $\pi_0 > 0$. Then f can be expressed as a mixture of the background density g and another, unspecified, density u, as follows

$$f = \pi_0 g + (1 - \pi_0) u. \tag{3.27}$$

(We do not use the notation h_0 and g_0 here, since these may not be uniquely defined.) The procedure is summarized in Table 3.11. An illustration of this decomposition is shown in Figure 3.6. Note that the density u may have a non-trivial log-concave background component. This is the case, for example, if f is the (nontrivial) mixture of two log-concave densities with disjoint support sets, in which case u is one of these log-concave densities.

Table 3.11. Log-concave background computation. (The input d is used to initialize the function v — discretized as $\mathbf{v}$ below — that is bounded from above by $\log f$. In our experiments, we chose $d = 0.02$.)

---

inputs: equally spaced gridpoints $\mathbf{t} = \{t_1, t_2, \ldots, t_k\}$, density f, a boolean R indicating whether to use the Riemann integral approximation, initialization amount d

---

initialize $\mathbf{w} = (0, 0, \ldots, 0)$ with length k
for $i = 1, \ldots, k$ **do**
  $\mathbf{w}[i] = \log(f(t_i))$
compute $\mathbf{A}$ and $\mathbf{b}$ as in (3.36)
initialize $\mathbf{v} = \mathbf{w} - (d, d, \ldots, d)$
if $R =$ true **then**
  do optimization (3.44) using SQP, and record the optimizer $\mathbf{v}$ and the maximum as π_0
else
  do optimization (3.33) using SQP, and record the optimizer $\mathbf{v}$ and the maximum as π_0

---

return $\mathbf{v}, \pi_0$

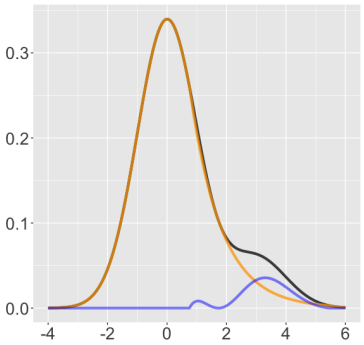


Figure 3.6. The density f of the Gaussian mixture $0.85\,\mathcal{N}(0,1) + 0.15\,\mathcal{N}(3,1)$, in black, and its decomposition into $\pi_0 g$, in orange, and $(1-\pi_0)u$, in blue. The orange curve is acquired by maximizing (3.34), but maximizing the Riemann integral approximation (3.44) gives almost the same result (with the maximum difference between the two curves being 2×10^{-14}). Notice $\pi_0 = 0.931$ as $\pi_0 g$ also takes a non-negligible weight from the smaller component $0.15\,\mathcal{N}(3,1)$.

3.5.1 Estimation and consistency

In practice, based on a sample, we estimate f, resulting in $\hat{f}$, and simply obtain estimates by plug-in as we did before, this time via

$$\begin{aligned} \text{maximize} \quad & \int_{-\infty}^{\infty} h(x)dx \\ \text{over} \quad & \left\{h : \mathbb{R} \to \mathbb{R}_+, \text{ log-concave}, \ h \le \hat{f}\right\}. \end{aligned} \tag{3.28}$$

At least formally, let $\hat{h}$ be a solution to (3.28). We then define

$$\hat{\pi}_0 = \int_{-\infty}^{\infty} \hat{h}(x)dx, \qquad \hat{g}(x) = \frac{\hat{h}(x)}{\hat{\pi}_0}. \tag{3.29}$$

Here, to avoid technicalities, we work under the assumption that the density estimator $\hat{f} = \hat{f}_n$ satisfies

$$\mathbb{E}\left[\operatorname*{ess\,sup}_{|x|\le a} \frac{\max\{\hat{f}_n(x), f(x)\}}{\min\{\hat{f}_n(x), f(x)\}}\right] \to 1, \quad n \to \infty, \quad \forall a > 0. \tag{3.30}$$

This condition is better suited for when the density f approaches 0 only at infinity. Also for the sake of simplicity, we only establish consistency for $\hat{\pi}_0$.

Theorem 8. *When* (3.30) *holds,* $\hat{\pi}_0$ *is a consistent.*

Proof. Fix $\eta > 0$ and $a > 0$, and consider the event, denoted Ω, that

$$\operatorname*{ess\,sup}_{|x|\le a} \frac{\max\{\hat{f}(x), f(x)\}}{\min\{\hat{f}(x), f(x)\}} \le 1 + \eta. \tag{3.31}$$

Because of (3.30), Ω happens with probability tending to 1 as the sample size increases. Let $\varepsilon = 1 - \int_{[-a,a]} f$, which is small when a is large.

Assume that Ω holds. Let h be a solution to (3.25), so that $\pi_0 = \int h$. Then define

$\tilde{h}(x) = (1-\eta)h(x)\mathbb{1}\{|x| \le a\}$ and note that

$$\int \tilde{h} = (1-\eta)\int_{[-a,a]} h \ge (1-\eta)(\pi_0 - \varepsilon),$$

so that $\int \tilde{h}$ is close to π_0 when η is small and a is large. We also note that $\tilde{h}$ is log-concave and satisfies $0 \le \tilde{h} \le (1-\eta)f$ a.e.. Under Ω, we have $f \le (1+\eta)\hat{f}$, and so it is also the case that $\tilde{h} \le (1-\eta)(1+\varepsilon)\hat{f} \le \hat{f}$, assuming as we do that η and ε are small enough. Then $\int \tilde{h} \le \hat{\pi}_0$, by definition of $\hat{\pi}_0$. Gathering everything, we obtain that $(1-\eta)(\pi_0 - \varepsilon) \le \hat{\pi}_0$. By letting $\eta \to 0$ and $a \to \infty$ so that $\varepsilon \to 0$, we have established that $\liminf \hat{\pi}_0 \ge \pi_0$ in probability. (The liminf needs to be understood as the sample size increases.)

The reverse relation, meaning $\limsup \hat{\pi}_0 \le \pi_0$, can be derived similarly starting with a solution $\hat{h}$ to (3.28). □

Confidence interval

Once again, if we have available a confidence band for f, we can deduce from that a confidence interval for π_0.

Theorem 9. *Suppose that we have a confidence band for f as in* (3.11)*. Then*

$$\hat{\pi}_l \le \pi_0 \le \hat{\pi}_u \tag{3.32}$$

holds with probability at least $1-\alpha$, where $\hat{\pi}_l$ and $\hat{\pi}_u$ are the values of the optimization problem (3.25) *with f replaced by $\hat{f}_l$ and $\hat{f}_u$, respectively.*

The proof is straightforward and thus omitted.

3.5.2 Numerical method

Unlike the previous sections, here the computation of our estimator(s) is non-trivial: indeed, after computing $\hat{f}$ with an off-the-shelf procedure, we need to solve the optimization problem (3.28). Thus, in this section, we discuss how to solve this optimization problem.

Although least concave majorants (or equivalently greatest convex minorants) have been considered extensively in the literature, for example in (Franců et al., 2017; Jongbloed, 1998), the problem (3.25) calls for a type of greatest concave minorant, and we were not able to find references in the literature that directly tackle this problem. We do want to mention (Gorokhovik, 2019), where a similar concept is discussed, but the definition is different from ours and no numerical procedure to solve the problem is provided. For lack of structure to exploit, we propose a direct discretization followed by an application of sequential quadratic programming (SQP), first proposed by Wilson (1963). For more details on SQP, we point the reader to (Gill and Wong, 2012) or (Nocedal and Wright, 2006, Ch 18).

Going back to (3.25), where here f plays the role of a generic density on the real line, the main idea is to restrict $v := \log h$ to be a continuous, concave, piecewise linear function. Once discretized, the integral admits a simple closed-form expression and the concavity constraint is transformed into a set of linear inequality constraints.

To setup the discretization, for $k \geq 1$ integer, let $t_{-k,k} < t_{-k+1,k} < \cdots < t_{k-1,k} < t_{k,k}$ be such that $t_{j,k} = -t_{-j,k}$ (symmetry) and $t_{j+1,k} - t_{j,k} = \delta_k$ (equispaced) for all j, with $\delta_k \to 0$ (dense) and $t_{k,k} \to \infty$ as $k \to \infty$ (spanning the real line). Suppose that v is concave with $v \leq \log f$ and to that function associate the triangular sequence $v_{j,k} := v(t_{j,k})$. Then, for each k,

$$\begin{aligned} -v_{j+1,k} + 2v_{j,k} - v_{j-1,k} &= -v(t_{j+1,k}) + 2v(t_{j,k}) - v(t_{j-1,k}) \\ &= 2\Big(v(t_{j,k}) - \tfrac{1}{2}v(t_{j+1,k} + \delta_k) - \tfrac{1}{2}v(t_{j-1,k} + \delta_k)\Big) \geq 0, \quad \text{for all } k, \end{aligned}$$

by the fact that v is concave. In addition, $v_{j,k} = v(t_{j,k}) \leq \log f(t_{j,k}) =: u_{j,k}$ (which are given). Instead of working directly with a generic concave function v, we work with those that are piecewise linear as they are uniquely determined by their values at the grid points if we further restrict them to be $= -\infty$ on $(-\infty, t_{-k,k}) \cup (t_{k,k}, +\infty)$. Effectively, at k, we replace in (3.25) the class $\mathcal{C}$ with the class $\mathcal{C}_k$ of functions h such that $v = \log h$ is log-concave, linear on each interval $[t_{j,k}, t_{j+1,k}]$, $v(x) = -\infty$ for $x < t_{-k,k}$ or $x > t_{k,k}$, and that satisfies $v(t_{j,k}) \leq \log f(t_{j,k})$, for all j.

This leads us to the following optimization problem, which instead of being over a function space is over a Euclidean space:

$$\begin{aligned} &\text{maximize} \quad \Lambda(\mathbf{v}) \\ &\quad \text{over} \quad \mathbf{v} = \left[v_{-k,k}, \ldots, v_{k,k}\right]^\top \quad \text{such that} \quad \mathbf{A}\mathbf{v} \geq \mathbf{b}, \end{aligned} \tag{3.33}$$

where

$$\Lambda(v_{-k}, \ldots, v_k) := \delta_k \sum_{j=-k}^{k-1} \lambda(v_j, v_{j+1}), \tag{3.34}$$

$$\lambda(x,y) := \mathbb{1}\{x \neq y\} \frac{\exp(x) - \exp(y)}{x - y} + \mathbb{1}\{x = y\} \exp(x), \tag{3.35}$$

and where the $2k$ by $k+2$ matrix $\mathbf{A}$ and the $2k$ by 1 vector $\mathbf{b}$ are defined as

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ -1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & -1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -u_1 \\ -u_2 \\ \vdots \\ -u_k \end{bmatrix}. \tag{3.36}$$

As we are not aware of any method that could solve (3.33) exactly, we will use sequential quadratic programming (SQP). In our implementation we use the R package `nloptr` [6] based on an original implementation of Kraft (1988).

Theorem 10. *Assume that f is Lipschitz and locally bounded away from 0. Then the value of discretized optimization problem* (3.33) *converges, as $k \to \infty$, to the value of the original*

[6] https://cran.r-project.org/web/packages/nloptr/index.html

optimization problem (3.25).

The assumptions made on f are really for convenience — to expedite the proof of the result while still including interesting situations — and we do expect that the result holds more broadly. We also note that, in our workflow, the problem (3.33) is solved for $\hat{f}$, and that $\hat{f}$ satisfies these conditions (remember that the sample size is held fixed here) if it is a kernel density estimator based on a smooth kernel function supported on the entire real line like the Gaussian kernel.

Proof. Let h be a solution to (3.25) so that $\int h = \pi_0$, where π_0 denotes the value of (3.25). Define $v = \log h$ and let $v_{j,k} = v(t_{j,k})$. As we explained above, this makes $\mathbf{v}_k := (v_{-k,k}, \ldots, v_{k,k})$ feasible for (3.33). Therefore $\Lambda(\mathbf{v}_k) \leq \pi_{0,k}$, where $\pi_{0,k}$ denotes the value of (3.33). On the other hand, let $\tilde{v}_k$ denote the piecewise linear approximation to v on the grid, meaning that $\tilde{v}_k(x) = -\infty$ if $x < t_{-k,k}$ or $x > t_{k,k}$, $\tilde{v}_k(t_{j,k}) = v(t_{j,k})$ and $\tilde{v}_k$ linear on $[t_{j,k}, t_{j+1,k}]$ for all j. We then have

$$\Lambda(\mathbf{v}_k) = \int \tilde{h}_k \xrightarrow{k\to\infty} \int h, \tag{3.37}$$

where the convergence is justified, for example, by the fact that $\tilde{h}_k \to h$ pointwise and $0 \leq \tilde{h}_k \leq h$ (because, by concavity of v, $v \geq \tilde{v}_k$), so that the dominated convergence theorem applies. From this we get that

$$\liminf_{k\to\infty} \pi_{0,k} \geq \pi_0. \tag{3.38}$$

In the other direction, let $\mathbf{v}_k$ be a solution of (3.33), so that $\Lambda(\mathbf{v}_k) = \pi_{0,k}$. Let $\tilde{v}_k$ denote the linear interpolation of $\mathbf{v}_k$ on the grid $(t_{-k,k}, \ldots, t_{k,k})$, defined exactly as done above, and let $\tilde{h}_k = \exp(\tilde{v}_k)$. We have that $\tilde{h}_k \leq f$ at the grid points, but not necessarily elsewhere. To work in that direction, fix $a > 0$, taken arbitrarily large in what follows. Because of our assumptions on f, we have that $u := \log f$ is Lipschitz on $[-a, a]$, say with Lipschitz constant L. In particular, with

$\tilde{u}_k$ denoting the linear approximation of u based on the grid, we have

$$|u(x) - \tilde{u}_k(x)| \le L\delta_k =: \eta_k. \tag{3.39}$$

Define $\bar{v}_k(x) = \tilde{v}_k(x) - \eta_k$ if $x \in [-a,a]$ and $\bar{v}_k(x) = -\infty$ otherwise. Note that $\bar{v}_k$ is also concave and piecewise linear, and because $\tilde{h}_k \le f$, $\tilde{v}_k \le \tilde{u}_k$, we have

$$\bar{v}_k(x) = \tilde{v}_k(x) - \eta_k \le \tilde{u}_k(x) - \eta_k \le u(x), \quad \forall x \in [-a,a]. \tag{3.40}$$

In particular, $\bar{h}_k := \exp(\bar{v}_k)$ is feasible for (3.25), implying that $\int \bar{h}_k \le \pi_0$. On the other hand, we have

$$\int \bar{h}_k = \int \exp(\bar{v}_k) = \exp(-\eta_k) \int_{[-a,a]} \exp(\tilde{v}_k) = \exp(-\eta_k) \int_{[-a,a]} \tilde{h}_k, \tag{3.41}$$

and so, because $\tilde{h}_k \le \tilde{f}_k := \exp(\tilde{u}_k)$,

$$\begin{aligned} \pi_{0,k} = \Lambda(\mathbf{v}_k) = \int \tilde{h}_k &= \int_{[-a,a]} \tilde{h}_k + \int_{[-a,a]^c} \tilde{h}_k \\ &= \exp(\eta_k) \int \bar{h}_k + \int_{[-a,a]^c} \tilde{h}_k \\ &\le \exp(\eta_k)\pi_0 + \int_{[-a,a]^c} \tilde{f}_k. \end{aligned}$$

Given $\varepsilon > 0$ arbitrarily small, choose a large enough that $\int_{[-a,a]^c} f \le \varepsilon$. Since

$$\int_{[-a,a]^c} \tilde{f}_k \xrightarrow{k\to\infty} \int_{[-a,a]^c} f, \tag{3.42}$$

for k large enough we have $\int_{[-a,a]^c} \tilde{f}_k \le 2\varepsilon$, implying that $\pi_{0,k} \le \exp(\eta_k)\pi_0 + 2\varepsilon$. Using the fact that $\exp(\eta_k) \to 1$ since $\eta_k = L\delta_k \to 0$, and that $\varepsilon > 0$ is arbitrary, we get that

$$\limsup_{k\to\infty} \pi_{0,k} \le \pi_0, \tag{3.43}$$

concluding the proof. □

We mention that besides the discretization (3.33), we also considered a more straightforward discretization of the integral, effectively replacing Λ in (3.33) with Λ_0 defined as

$$\Lambda_0(v_{-k},\ldots,v_k) := \delta_k\left(\tfrac{1}{2}\exp(v_{-k}) + \tfrac{1}{2}\exp(v_k) + \sum_{j=-k+1}^{k-1}\exp(v_j)\right). \tag{3.44}$$

The outputs returned by these two discretizations, (3.33) and (3.44), were very similar in our numerical experiments.

3.5.3 Numerical experiments

In this subsection we consider experiments where the background component could be assumed to be log-concave. The density fitting process and confidence band acquisition are exactly the same as in Section 3.3.2. We first consider four different situations as listed in Table 3.12. Although the mixture distributions are identical to those in Table 3.3 for the symmetric case, we need to point out that π_0 is different here, as π_0 here corresponds to the largest possible log-concave component as defined in (3.24). We again generate a sample of size $n = 1000$ from each model, and repeat each setting 1000 times.

We note that the output of our algorithm depends heavily on the estimation of the density $\hat{f}$. When the bandwidth selected by cross-validation yields a density with a high frequency of oscillation, the largest log-concave component is very likely to be smaller than the correct value. For an illustrative situation, see Figure 3.7. In the event that such a situation happen, we recommend that the user look at a plot of $\hat{f}$ before applying our procedure, with the possibility of selecting a larger bandwidth. This is what we did, for example, for the Carina and Leukemia datasets in Figure 3.10. In the simulations, to avoid the effect of these issue, we report the median value instead of the mean. These values are reported in Table 3.13.

As can be seen, even with only the assumption of log-concavity, our method is accurate in estimating π_0, with estimation errors ranging from 0.001 to 0.007. Our method frequently

achieves smaller error in estimating π_0 than comparison methods in estimating θ_0 and θ, and does so with less information on the background component. For situations where the background component is specified incorrectly, we invite the reader to Section 3.6.1.

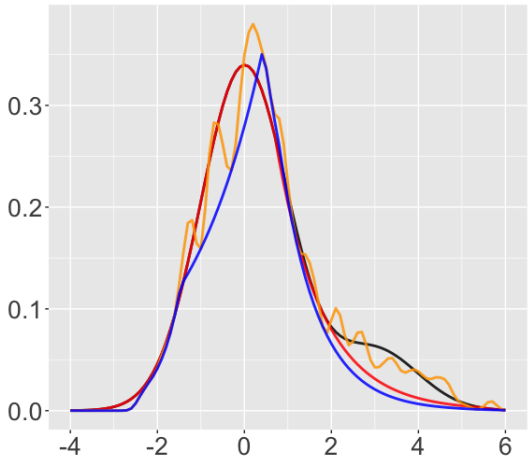


Figure 3.7. An illustrative situation that occurred in the course of our simulations, where a high frequency of oscillation of $\hat{f}$ results in a significantly lower estimation for π_0: a Gaussian mixture density $0.85\,\mathcal{N}(0,1)+0.15\,\mathcal{N}(0,1)$, in black, and corresponding largest log-concave component h, in red, with the fitted density $\hat{f}$, in orange, and fitted largest log-concave component $\hat{h}$, in blue. Here, as before, $\hat{f}$ is obtained by kernel density estimation with bandwidth chosen by cross-validation. The result based on this estimate is $\hat{\pi}_0 = 0.807$, while the actual value is $\pi_0 = 0.931$.

Remark 3. We note that here the SQP algorithm as implemented in the R package nloptr sometimes gives $\hat{h} = 0$ as the largest log-concave component due to some parts of the estimated density $\hat{f}$ being 0. This situation happens occasionally when computing the lower confidence bound, and it is of course incorrect in situations like those in Table 3.12. When this situation occurs, we rerun the algorithm on the largest interval of non-zero $\hat{f}$ values and report the greatest log-concave component acquired on that interval.

3.5.4 Real data analysis

In this subsection we examine real datasets of two component mixtures where the background component could be assumed to be log-concave. We first consider the six real datasets presented in Section 3.3.3, but this time look for a background log-concave component.

Table 3.12. Log-concave simulation situations of Gaussian mixtures, as well as values of θ, θ_0, and π_0, obtained through numerical optimization.

Model	**Distribution**	θ	θ_0	π_0
L1	$0.85\,\mathcal{N}(0,1)+0.15\,\mathcal{N}(3,1)$	0.850	0.850	0.931
L2	$0.95\,\mathcal{N}(0,1)+0.05\,\mathcal{N}(3,1)$	0.950	0.950	0.981
L3	$0.85\,\mathcal{N}(0,1)+0.1\,\mathcal{N}(2.5,0.75)+0.05\,\mathcal{N}(-2.5,0.75)$	0.850	0.851	0.975
L4	$0.85\,\mathcal{N}(0,1)+0.1\,\mathcal{N}(2.5,0.75)+0.05\,\mathcal{N}(5,0.75)$	0.850	0.850	0.946

Table 3.13. A comparison of various methods for estimating a log-concave background component in the situations of Table 3.12. Here we report the median values instead of the mean values (and also report the standard deviations, as before). As always, $\hat{\pi}_0$ should be compared with π_0, $\hat{\theta}_0^X$ should be compared with θ_0, and $\hat{\theta}^X$ should be compared with θ.

Model	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$	**Null**	$\hat{\theta}_0^{\mathrm{PSC}}$	$\hat{\theta}_0^{\mathrm{PSH}}$	$\hat{\theta}_0^{\mathrm{PSB}}$	$\hat{\theta}^{\mathrm{E}}$	$\hat{\theta}^{\mathrm{MR}}$	$\hat{\theta}^{\mathrm{CJ}}$
L1	0.932	0.596	1	$\mathcal{N}(0,1)$	0.850	0.858	0.894	0.861	0.890	0.888
	(0.034)	(0.074)	(0)		(0.024)	(0.023)	(0.018)	(0.023)	(0.014)	(0.085)
L2	0.974	0.662	1	$\mathcal{N}(0,1)$	0.942	0.958	0.988	0.953	0.972	0.964
	(0.028)	(0.080)	(0)		(0.023)	(0.022)	(0.015)	(0.026)	(0.008)	(0.082)
L3	0.969	0.611	1	$\mathcal{N}(0,1)$	0.864	0.948	0.936	0.856	0.896	0.705
	(0.031)	(0.077)	(0)		(0.023)	(0.034)	(0.017)	(0.024)	(0.015)	(0.085)
L4	0.942	0.632	1	$\mathcal{N}(0,1)$	0.848	0.857	0.893	0.849	0.890	0.712
	(0.033)	(0.074)	(0)		(0.023)	(0.024)	(0.018)	(0.024)	(0.014)	(0.088)

When the background is known, the numerical results of the Prostate and Carina datasets can be found in Table 3.14, Figure 3.8(a) and Figure 3.8(b). When the background is unknown, the numerical results of the HIV, Leukemia, Parkinson and Police datasets can be found in Table 3.15, Figure 3.9(a), Figure 3.9(b), Figure 3.9(c) and Figure 3.9(d).

In addition to the above six datasets, we also include here the Old Faithful Geyser dataset (Azzalini and Bowman, 1990). This dataset consists of 272 waiting times, in minutes, between eruptions for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. We attempt to find the largest log-concave component of the waiting times. The results are summarized in the first row of Table 3.15, Figure 3.11(a) and Figure 3.11(b). We want to note from this example that the curve $\hat{h}_u$ acquired from $\hat{f}_u$ leading to the computation of $\hat{\pi}_0^{\mathrm{U}}$, is not the upper confidence bound of h_0, as shown by Figure 3.11(a).

Remark 4. We observe here through these real datasets that compared to symmetric background assumptions in Section 3.3.3, the largest log-concave background component usually has a higher weight than the largest symmetric background component.

Table 3.14. Real datasets where the background log-concave component is known. (Note that we work with z values here instead of p-values so our result in Prostate dataset is slightly different from that in (Patra and Sen, 2016).)

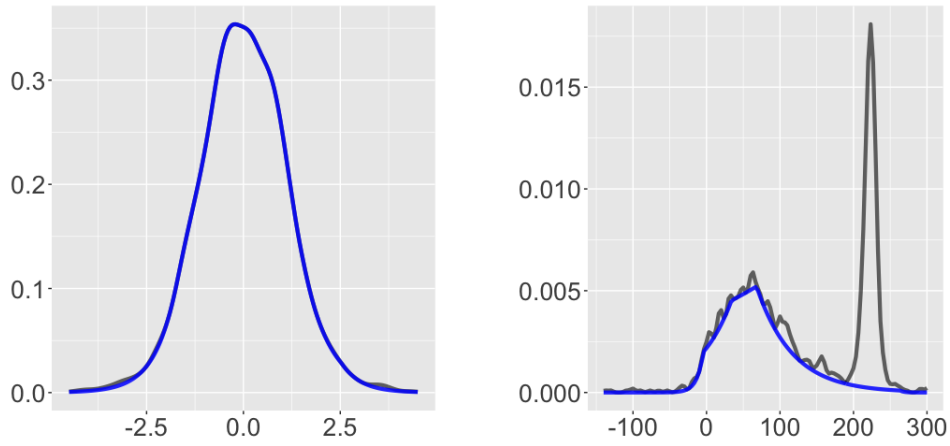
Model	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$	**Null**	$\hat{\theta}_0^{\mathrm{PSC}}$	$\hat{\theta}_0^{\mathrm{PSH}}$	$\hat{\theta}_0^{\mathrm{PSB}}$	$\hat{\theta}^{\mathrm{E}}$	$\hat{\theta}^{\mathrm{MR}}$	$\hat{\theta}^{\mathrm{CJ}}$
Prostate	0.994	0.809	1	$\mathcal{N}(0,1)$	0.931	0.941	0.975	0.931	0.956	0.867
Carina	0.600	0.242	1	bgstars	0.636	0.645	0.677	0.951	0.664	0.206

3.6 Conclusion and discussion

In this paper, we extend the approach of Patra and Sen (2016) to settings where the background component of interest is assumed to belong to three emblematic examples: symmetric, monotone, and log-concave. In each setting, we derive estimators for both the proportion and density for the background component, establish their consistency, and provide confidence

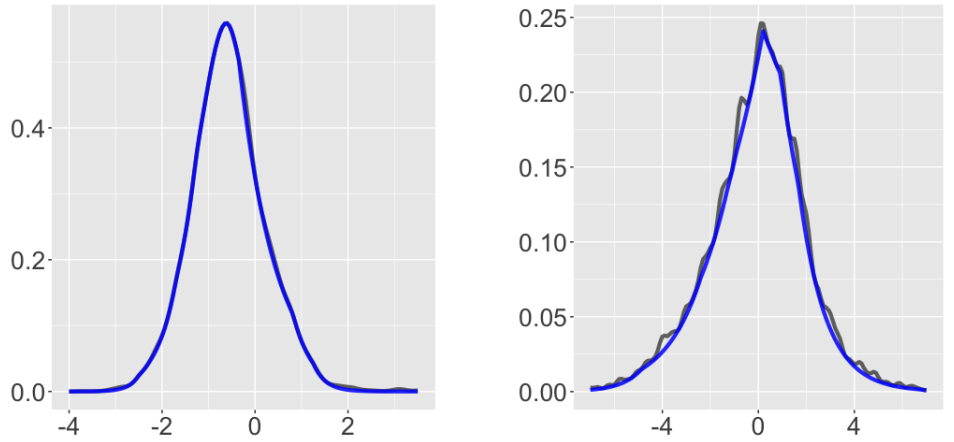
Table 3.15. Real datasets where the background log-concave distribution is unknown. (Note that Efron's method will not run on the Geyser dataset.)

Model	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$	$\hat{\theta}^{\mathrm{E}}$	$\hat{\theta}^{\mathrm{CJ}}$
Geyser	0.693	0.287	1	NA	1
HIV	0.984	0.804	1	0.940	0.926
Leukemia	0.981	0.695	1	0.911	0.820
Parkinson	1	0.605	1	0.998	0.993
Police	0.997	0.765	1	0.985	0.978

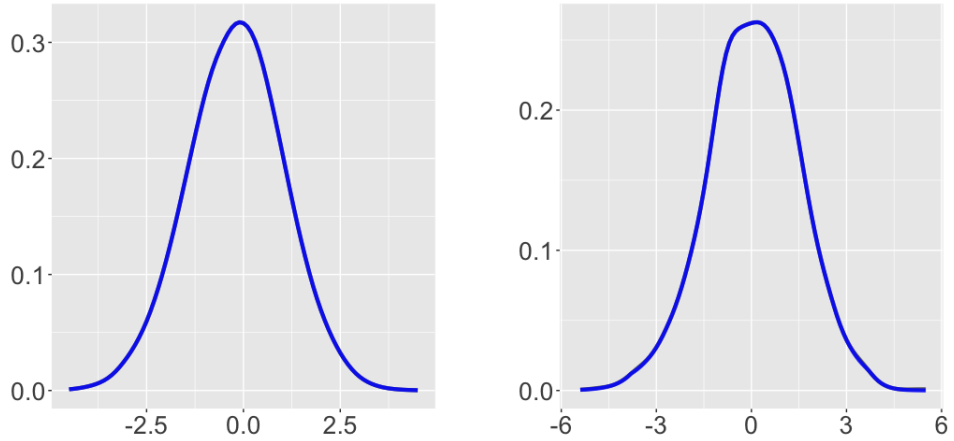


(a) Prostate dataset (b) Carina dataset

Figure 3.8. Estimated background log-concave component on the Prostate (z values) and Carina (radial velocity) datasets: the black curve represents fitted density; the blue curve represents computed $\hat{h}$, one of the largest log-concave components. Due to unusually high frequency of oscillation in $\hat{f}$ in the Carina dataset, we consider increasing the bandwidth from the one acquired by cross-validation, and the corresponding result is shown in Figure 3.10(a).

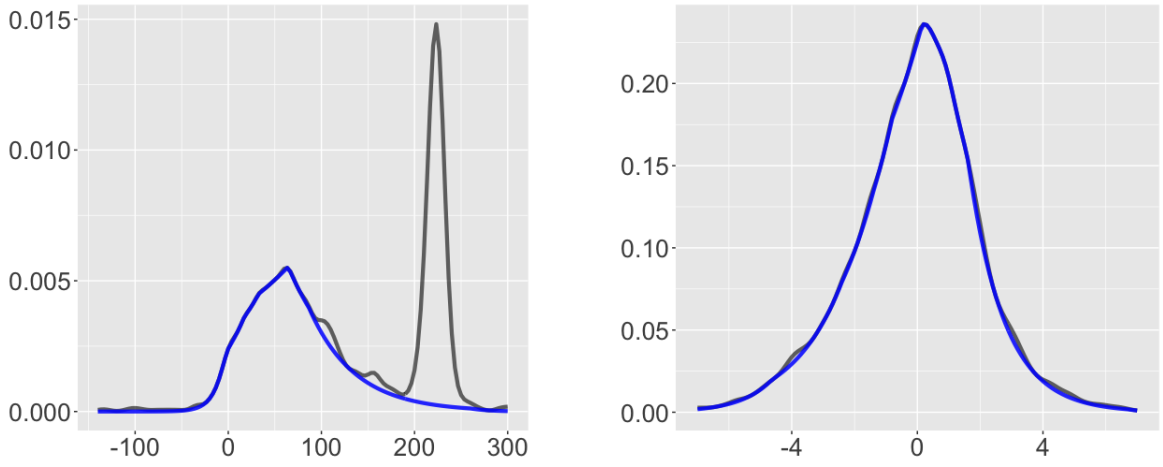


(a) HIV dataset (b) Leukemia dataset



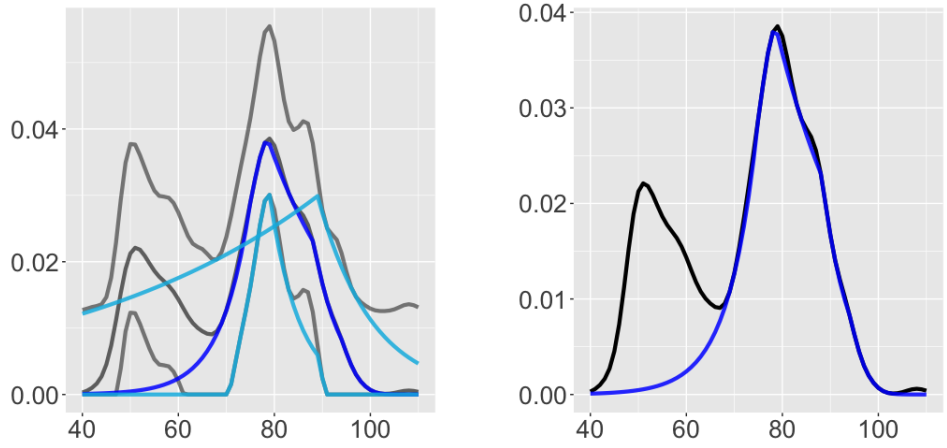
(c) Parkinson dataset (d) Police dataset

Figure 3.9. Estimated background log-concave component on HIV (z values), Leukemia (z values), Parkinson (z values), and Police (z scores) datasets: the black curve represents fitted density; the blue curve represents computed $\hat{h}$, one of the largest log-concave components. Due to the high frequency of oscillation in $\hat{f}$ in the Leukemia dataset, we consider increasing the bandwidth from the one acquired by cross-validation, and the corresponding result is shown in Figure 3.10(b).



(a) Carina dataset (b) Leukemia dataset

Figure 3.10. Carina (radial velocity) and Leukemia (z values) datasets with kernel density, with bandwidth chosen 'by hand' instead of by cross-validation: the black curve represents fitted density with increased bandwidth; the blue curve represents the computed $\hat{h}$. We note that for the Carina dataset, the bandwidth was increased from 3.085 to 6, and $\hat{\pi}_0$ changed from 0.550 to 0.600. For the Leukemia dataset, the bandwidth was been increased from 0.124 to 0.25, and $\hat{\pi}_0$ changed from 0.939 to 0.981.



(a) Geyser dataset with $\hat{h}_l$ and $\hat{h}_u$ (b) Geyser dataset with only $\hat{h}$

Figure 3.11. Estimated background log-concave component on the Geyser (duration) dataset: the black curve represents the fitted density; the blue curve represents the computed $\hat{h}$; the gray curves (left) represent $\hat{f}_l$ and $\hat{f}_u$; the light blue curves (left) represents $\hat{h}_l$ and $\hat{h}_u$. Note from the left plot that $\hat{h}_u$ is only used to compute $\hat{\pi}_0^{\mathrm{U}}$, and is not an upper bound for h.

intervals/bands. However, the important situation of incorrect background distribution and other extensions remain unaddressed, and therefore we discuss them in this section.

3.6.1 Incorrect background specification

As mentioned in Section 3.3.2 and Section 3.5.3, our method requires much less information than comparable methods, and therefore is much less prone to misspecification of the background component. In this subsection, we give an experiment illustrating this point. We consider the mixture model: $0.85\,\mathcal{T}_6 + 0.15\,\mathcal{N}(3,1)$. Instead of the correct null distribution $\mathcal{T}_6$, we take it to be $\mathcal{N}(0,1)$. This situation could happen in situations of multiple testing settings, for example in the HIV dataset in Section 3.3.3. We consider both a symmetric background and log-concave background on this model in Table 3.16, and report the fitted values of our method and comparison methods in Table 3.17. As can be seen, when estimating π_0, our estimator achieves 0.002 error when assuming symmetric background and 0.004 error when assuming log-concave background, less then any other method in comparison that estimates θ_0 or θ. The heuristic estimator of Patra and Sen (2016) has slightly higher error than our method, while the constant estimator of Patra and Sen (2016) and the estimators of Efron (2007) and Cai and Jin (2010) have large errors. The upper confidence bound of Meinshausen and Rice (2006) also becomes incorrect.

Table 3.16. Simulation situations when background specification is incorrect, as well as values of θ, θ_0, π_0, obtained through numerical optimization.

Model	**Background**	**Distribution**	θ	θ_0	π_0
S5	Symmetric	$0.85\,\mathcal{T}_6 + 0.15\,\mathcal{N}(3,1)$	0.850	0.850	0.859
L5	Log-concave	$0.85\,\mathcal{T}_6 + 0.15\,\mathcal{N}(3,1)$	0.850	0.850	0.925

3.6.2 Combinations

Although not discussed in the main part of the paper, some combinations of the shape constraints considered earlier are possible. For example, one could consider extracting a maximal

Table 3.17. Results for the situations of Table 3.16. We provide mean values in S5, and median values in L5.

Model	Center	$\hat{\pi}_0$	$\hat{\pi}_0^{\mathrm{L}}$	$\hat{\pi}_0^{\mathrm{U}}$	**Null**	$\hat{\theta}_0^{\mathrm{PSC}}$	$\hat{\theta}_0^{\mathrm{PSH}}$	$\hat{\theta}_0^{\mathrm{PSB}}$	$\hat{\theta}^{\mathrm{E}}$	$\hat{\theta}^{\mathrm{MR}}$	$\hat{\theta}^{\mathrm{CJ}}$
S5	0.073	0.857	0.541	1	$\mathcal{N}(0,1)$	0.815	0.855	0.877	0.803	0.843	0.816
	(0.057)	(0.021)	(0.057)	(0)		(0.021)	(0.022)	(0.016)	(0.026)	(0.015)	(0.086)
L5		0.921	0.574	1	$\mathcal{N}(0,1)$	0.817	0.856	0.877	0.803	0.843	0.814
		(0.036)	(0.069)	(0)		(0.021)	(0.022)	(0.016)	(0.026)	(0.015)	(0.086)

background that is symmetric *and* log-concave; or one could consider extracting a maximal background that is monotone *and* log-concave. As it turns out, these two combinations are intimately related. Mixtures of symmetric log-concave distributions are considered, for example, in (Pu and Arias-Castro, 2020).

3.6.3 Generalization to higher dimensions

All our examples were on the real line, corresponding to real-valued observations, simply because the work was in large part motivated by multiple testing in which the sample stands for the test statistics. But the approach is more general. Indeed, consider a measurable space, and let $\mathcal{D}$ be a class of probability distributions on that space. Given a probability distribution μ, we can quantify how much there is of $\mathcal{D}$ in μ by defining

$$\pi_0 := \sup\left\{\pi : \exists \nu \in \mathcal{D} \text{ s.t. } \mu \geq \pi\nu\right\}. \tag{3.45}$$

For concreteness, we give a simple example in an arbitrary dimension d by generalizing the setting of a symmetric background component covered in Section 3.3. Although various generalizations are possible, we consider the class — also denoted $\mathcal{S}$ as in (3.3) — of spherically symmetric (i.e., radial) densities with respect to the Lebesgue measure on $\mathbb{R}^d$. It is easy to see

that the background component proportion is given by

$$\pi_0 = \int_{\mathbb{R}^d} h_0(x)dx, \qquad h_0(x) := \min\{f(y) : \|y\| = \|x\|\}, \tag{3.46}$$

and, if $\pi_0 > 0$, the background component density is given by $g_0 := h_0/\pi_0$.

3.7 Acknowledgement

This chapter, in full, is a version of the paper "Extending the Patra-Sen Approach to Estimating the Background Component in a Two-Component Mixture Model", Ery Arias-Castro and He Jiang. It is currently being prepared for submission for publication of the material. The dissertation author is the primary investigator and corresponding author of this material.

We are grateful to Professor Philip Gill for discussions regarding the discretization of the optimization problem (3.25).

Chapter 4

Fitting a Multi-Modal Density by Dynamic Programming

4.1 Abstract

We consider the problem of fitting a probability density function when it is constrained to have a given number of modal intervals. We propose a dynamic programming approach to solving this problem numerically. When this number is not known, we provide several data-driven ways for selecting it. We perform some numerical experiments to illustrate our methodology.

4.2 Introduction

Density estimation is an important task in exploratory data analysis, for example, the histogram is introduced in the most basic courses in statistics, and as such has been the object of some longstanding and intense study in statistics and related fields (Scott, 2015; Sheather, 2004; Silverman, 2018). Rather than assuming that the underlying density comes from a parametric family (e.g., normal), or some smoothness class (e.g., differentiable with some given bound on the derivative), we focus here on the shape constraint that the density has at most a given number of modal intervals. A mode is an important feature and defining the density based on the number of modal intervals is thus intuitive. It also goes hand-in-hand with a modal approach clustering, in particular, as proposed by Fukunaga and Hostetler (1975). In this paper, we propose a dynamic

programming way of fitting a multi-modal density to a numerical (one-dimensional) sample, and discuss data-driven ways of choosing the number of modal intervals when this information is unknown.

4.2.1 Importance of density modes

Modes are important features of densities, and do not fail to catch the eye of the analyst when they manifest themselves in plots of histograms or other estimates of the density, as Pearson (1895) pointed out, "thus the 'mean,' the 'mode,' and the 'median' have all distinct characters important to the statistician". See Figure 4.1 for a classical example. Estimating the modes of the

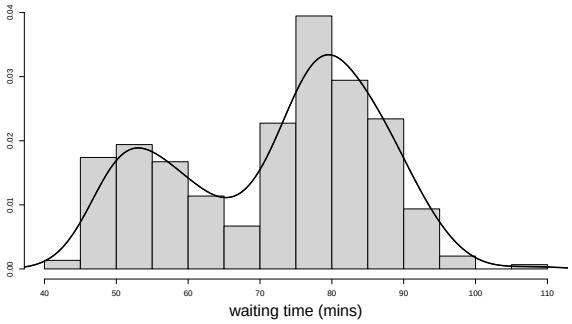


Figure 4.1. A histogram of the waiting times between eruptions of the Old Faithful Geyser, with a an overlay of a kernel density estimate. (This is the geyser dataset in the MASS package in R.)

underlying density has thus been an important problem in statistics which dates back to Parzen (1962), who proposed to first estimate the density by kernel density estimation — pioneered by Rosenblatt (1956) — and then locate the mode of the resulting estimator. The problem of estimating the density modes has been the subject of great, ongoing interest by statisticians since then; see (Chacón, 2020) for a recent review.

In addition to being important features of densities which are likely to attract the attention of the data scientist, modal intervals are also useful when the goal is clustering. Working at the population level as in (Chacón, 2015), a clustering amounts to a partitioning of the population, and Fukunaga and Hostetler (1975) proposed to partition the population according to the basins of attraction of the gradient ascent flow defined by the density. These are called the "descending

manifolds" in Morse theory (Chen et al., 2017), and the ones that pertain to modal intervals provide a partition of the support of the density up to a set of measure zero when the density is regular in a certain sense (specifically, when it is of Morse type). The corresponding methodology is known as the "mean-shift algorithm" (Cheng, 1995) and is now a well-established approach to clustering with its own literature (Carreira-Perpiñán, 2015; Chacón, 2020; Menardi, 2016). This approach to clustering turns out to be very closely related to the level set/cluster tree approach of Hartigan (1975), which itself has generated a good amount of research (Chaudhuri et al., 2014; Rinaldo et al., 2012; Stuetzle and Nugent, 2010). For a connection between these two approaches to clustering, see our recent work (Arias-Castro and Qiao, 2021a,b). The level set approach to clustering is also known to be intimately connected to single-linkage clustering (Hartigan, 1981; Penrose, 1995). We note that, in dimension one — our focus here — modal clustering amounts to partitioning the support of the density according to the intervals defined by the knots (indeed, not the modes or local maxima), and thus is exceedingly simple to execute when that information is available. In practice, of course, inference is based on a sample. Our approach to estimating a multi-modal density returns estimates for the knots separating the modes.

In a closely related field, modes of densities are also of interest in mixture models. For example, the number of modes in a Gaussian mixture model has been the subject of some attention (McLachlan and Rathnayake, 2014). We note, however, that the consideration of modes is not particularly natural in the context of mixtures, for the simple reason that the modes are not directly related to the component centers. When components are too close to each other in a mixture model, it is difficult to interpret the meaning of each component. Fitting a density with a given number of modal intervals can be seen as an alternative offering a clearer interpretation in some settings.

4.2.2 Estimating a uni-modal or multi-modal density

As we just alluded to, a model for multi-modal densities offers a competitive alternative to a mixture model. In this paper, we simply choose as model the entire class of densities with at

most K modes. While a *mode* typically means an isolated local maximum, to better accommodate step density functions — which are important in many respects, including in their connection to maximum likelihood — we here call a density f *K-modal* if there are $\lambda_1 < \cdots < \lambda_{K+1}$ such that f is uni-modal on $[\lambda_k, \lambda_{k+1}]$ for all $k = 1, \ldots, K$, where $\lambda_1 = -\infty$ and $\lambda_{K+1} = \infty$. We call an interval a *modal interval* of f if f is uni-modal on that interval. To be sure, f is *uni-modal* if there is some μ such that f is non-decreasing on $(-\infty, \mu]$ and non-increasing on $[\mu, \infty)$. In that case, μ is a mode of the density f. We also denote $\lambda_2, \lambda_3, \ldots, \lambda_K$ as the knots.

Given the significance of modes, there is some amount of literature on fitting densities with some constraints on the number of modes, although most of the attention has gone to uni-modal densities, i.e., densities with a single mode, with both methodological and theoretical developments. This effort dates back to the work of Grenander (1956) on the estimation of a monotonic density, and his derivation of the corresponding maximum likelihood estimator, which nowadays bears his name. A monotone density is, of course, uni-modal. Vice versa, a uni-modal density with known mode at μ is non-decreasing on $(-\infty, \mu]$ and non-increasing on $[\mu, \infty)$, so that the maximum likelihood is given by Grenander's estimator applied on each side of μ with the proper monotonicity constraint. When the mode is unknown, the likelihood is simply maximized over the location $\mu \in \mathbb{R}$. Rao (1969) obtained the asymptotic distribution of this estimator pointwise when applied to estimating uni-modal densities with a known mode, and showed this estimator's consistency on intervals not containing the mode. As an extension, Wegman (1970a,b) proposed a maximum likelihood estimator of a uni-modal density when the location of the mode is unknown, and showed that for sufficiently large sample size and on certain regions, this estimator agrees with the maximum likelihood estimator of a uni-modal density with known mode. We note that penalized versions of Grenander's estimator have also been proposed (Sun and Woodroofe, 1996; Woodroofe and Sun, 1993), and their use could be extended to fitting a uni-modal density. Besides maximum likelihood methods, the problem of uni-modal density estimation has also been considered using other approaches, including via the use of splines (Bickel and Fan, 1996; Meyer, 2012; Meyer and Woodroofe, 2004), data

sharpening (Braun and Hall, 2001; Choi and Hall, 1999; Wolters, 2012), and optimal transport (Cumings-Menon, 2018). We also mention Birgé (1987a,b), who derived nonparametric minimax rates for estimating a uni-modal density or decreasing density using well chosen histograms with unequal bin width, and Hengartner and Stark (1995), who computed conservative finite-sample confidence regions for the entire density assuming monotonicity or uni-modality on the density, using linear programming methods.

Given the amount of literature on fitting a uni-modal density, it is natural to consider a fitting multi-modal density with any given number of modal intervals. There is a significant literature on testing for multi-modality, which we review later in the paper in Section 4.4.1. In terms of estimation methodology, the literature is much more scarce. Minnotte and Scott (1993) proposed an exploratory, multi-scale method they named the *mode tree*. A similar method, called *SiZer*, was later proposed by Chaudhuri and Marron (2000, 1999), and later, a rank-based variant was suggested by Dümbgen and Walther (2008). Such methods, although do not offer a way of estimating a multi-modal density, may be seen as offering a way to partition the real-line into (possible) modal intervals. Directly tackling the estimation problem, Wolters and Braun (2018a) considered fitting densities with 1 or 2 modes by adding an additive curve to the regular kernel density. Dasgupta et al. (2021) considered fitting multi-modal densities by first fitting a template density with the given amount of modes, in the spirit of (Cheng et al., 1999), and then transforming this density under a shape-preserving transformation to obtain the final optimal estimate under maximum likelihood.

4.2.3 Contribution and content

We consider fitting a probability density function with a constraint on the number of modal intervals. We introduce a dynamic programming approach to solving this problem numerically which, in principle, may be based on any method for fitting a uni-modal density — in our numerical experiments we use the method recently proposed by Wolters and Braun (2018a). To lighten up the computational burden, we propose a multi-grid search approach.

In addition, we offer data-driven ways for selecting the number of modal intervals when this information is unknown.

The organization of the paper will be as follows. In Section 4.3, we provide our way of fitting a K-modal density via dynamic programming, where K is provided by the user. We also introduce a multi-grid search approach to improving the estimate of the knot values of the density. In Section 4.4, we provide several data-driven options for choosing the number of modal intervals K, when this information is unknown. In Section 4.5, we report on some numerical experiments. We end with a discussion in Section 4.6.

4.3 Dynamic programming method

We have at our disposal a sample of real-valued observations. These data points are assumed ordered (without loss of generality), denoted $x_1 \leq \cdots \leq x_n$, and gathered in a vector $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Given $\mathbf{x}$, our goal is to fit a K-modal density.

4.3.1 Maximum likelihood estimation

It is natural to first consider an estimation by maximum likelihood, which we can formalize as the following optimization problem:

$$\text{maximize} \sum_{i=1}^{n} \log(f(x_i)), \tag{4.1}$$

over f an arbitrary K-modal density.

While when $K = 1$ this leads to the Grenander estimator for fitting a uni-modal density, when $K \geq 2$, this problem is not well-posed due to the fact that the likelihood is not bounded over the class of K-modal densities. In fact, this is already the case when fitting mixtures of two or more components, including Gaussian mixtures, if the variances are left unconstrained. See, e.g., (Day, 1969, Sec 7) or (van der Vaart, 2000, Ex 5.51).

Some solutions to this issue have been suggested. For example, Hathaway (1986)

proposed a constrained EM algorithm when maximizing the likelihood with a lower bound on component weights and an upper bound on the aspect ratio of the component variances, which allows the maximum likelihood estimator to be consistent (Hathaway, 1985).

In our situation, we can think of multiple ways to regularize the optimization problem. One way to do so is to constrain the search for modes to a finite grid of values that avoids all the data points. A further discretization of the problem would involve a search over a finite grid (the same one, or another one) for the $K-1$ knots defining the K candidate modal intervals. For such a $(K-1)$-tuple, the optimization would lead to fitting a uni-modal density on each of the corresponding intervals with a search for the mode done over the mode grid.

4.3.2 Fitting uni-modal densities

Even though the maximum likelihood estimator — at least in its constrained and discretized form — exists, it is known to have an issue, which comes from the Grenander estimator being inconsistent at the mode (Balabdaoui et al., 2011; Bickel and Fan, 1996; Woodroofe and Sun, 1993). Although it is consistent everywhere else and so, practically speaking, this is not particularly important, the resulting fit is not "visually" pleasing, as the value at the mode tends to be much larger than the other values that the density estimator takes, which results in a "spike" at the mode. As our approach can accommodate any method for fitting a uni-modal density, in our implementation we chose another method, that of Wolters and Braun (2018a).

As an example, we generated a sample of size $n = 10,000$ from a Laplace distribution with mean 0 and standard deviation 0.05. We first fitted a uni-modal density with the Grenander estimator (Grenander, 1956) with unknown mode where the mode location was selected to maximize the likelihood. We then fitted a uni-modal density based on adjustedKDE method (Wolters and Braun, 2018a). The result of the fitted densities are shown in Figure 4.2.

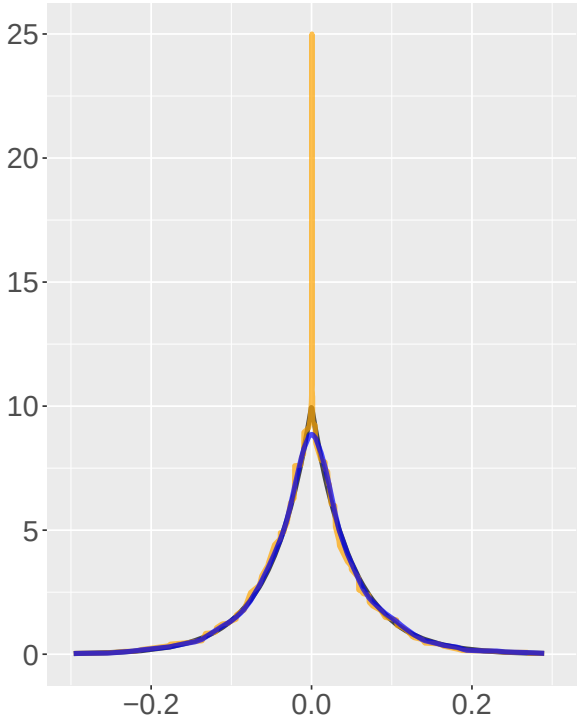


Figure 4.2. The density f of the Laplace distribution with mean 0 and standard deviation 0.05, in black, overlaid with the uni-modal density acquired via maximum likelihood (Grenander, 1956), in orange, and the uni-modal density acquired via adjusted KDE (Wolters and Braun, 2018a), in blue. Although not completely shown in this plot, the MLE estimator takes a value larger than $30,000$ at the mode.

4.3.3 Problem decomposition

We approach the optimization problem (4.1) in the following way, which involves a regularization by discretization and the use of a well-behaved method for fitting a uni-modal density on a given sample. For a candidate density f, let $\lambda = (\lambda_1, \ldots, \lambda_{K+1})$, with $\lambda_2 < \cdots < \lambda_K$, denote the vector of knots defining modal intervals for f, and add $\lambda_1 = -\infty$ and $\lambda_{K+1} = \infty$. Define $\pi_k = \int_{\lambda_k}^{\lambda_{k+1}} f(x)dx$ and assume that $\pi_k > 0$. Let h_k denote the density f conditional on $[\lambda_k, \lambda_{k+1}]$, i.e.,

$$h_k(x) = \frac{f(x)}{\pi_k} \mathbb{1}\{\lambda_k \leq x < \lambda_{k+1}\},$$

and note that it is uni-modal by construction. We will refer to $\pi_k h_k$ as the "rescaled" version of h_k.

Realizing that the set of knots $(\lambda_2, \ldots, \lambda_K)$, the set of weights $(\pi_1, \ldots, \pi_K)$, and the set of uni-modal densities $h_1, \ldots, h_K$, can "move" independently of each other, we may re-express the maximization problem (4.1) as follows:

$$\text{maximize} \sum_{k=1}^{K} \sum_{i=1}^{n} \log\big(\pi_k h_k(x_i)\big) \mathbb{1}\{\lambda_k \leq x_i < \lambda_{k+1}\}, \tag{4.2}$$

over $\lambda_2 < \cdots < \lambda_K$ chosen among grid points $g_2, \ldots, g_M$; over $\pi_1, \ldots, \pi_K > 0$ such that $\sum_{k=1}^{K} \pi_k = 1$; and over $h_1, \ldots, h_K$ uni-modal densities with h_k supported on $[\lambda_k, \lambda_{k+1}]$. Again, the fitting of these uni-modal densities will be eventually done by a well-behaved method, as otherwise the likelihood is not bounded and the problem above is not well-posed.

Given λ, for each k, let h_k^λ denote the uni-modal density fitted by our method of choice based on the sample falling in the interval $[\lambda_k, \lambda_{k+1})$. Note that h_k^λ is supported on $[\lambda_k, \lambda_{k+1}]$. Having computed these, the maximizing set of weights maximizes

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \log\big(\pi_k h_k^\lambda(x_i)\big) \mathbb{1}\{\lambda_k \leq x_i < \lambda_{k+1}\} = \sum_{k=1}^{K} n_k^\lambda \log \pi_k + \sum_{k=1}^{K} \sum_{i=1}^{n} \log\big(h_k^\lambda(x_i)\big) \mathbb{1}\{\lambda_k \leq x_i < \lambda_{k+1}\}, \tag{4.3}$$

where n_k^λ denote the amount of data points in $[\lambda_k, \lambda_{k+1})$. Since the second term on the right-hand side does not depend on the weights, the solution is the one maximizing the first term, which is $\pi_1^\lambda, \ldots, \pi_K^\lambda$, where $\pi_k^\lambda = n_k^\lambda / n$. The corresponding estimate for the density is simply

$$f^\lambda(x) = \sum_{k=1}^{K} \pi_k^\lambda h_k^\lambda(x) \mathbb{1}\{\lambda_k \le x < \lambda_{k+1}\}.$$

It thus remain to optimize with respect to the knot set λ. The search will be over a finite grid $g_2 < \cdots < g_M$, to which we add $g_1 = -\infty$ and $g_{M+1} = \infty$. We are able to implement this search in an efficient way.

4.3.4 Dynamic programming approach

As described in (Bradley et al., 1977, Chap 11), dynamic programming is an optimization method that turns a large and complex problem into a sequence of smaller and simpler problems, where its essential characteristic is a multi-stage optimization procedure. In our situation, we turn a maximization problem over K-modal densities into a sequence of maximizations over uni-modal densities. This is possible because of the fact that in (4.2) the maximization over the uni-modal densities $h_1, \ldots, h_K$ can be done with respect to each h_k independently, without coordination. We provide an example to illustrate our method in action in Figure 4.3.

To set up the dynamic programming recursion, we define two matrices. We first let $\mathbf{S}$ be an M-by-M upper-triangular matrix, where $\mathbf{S}[i, j]$, $i = 1, \ldots, M$, and $j = i, \ldots, M$, denote the log-likelihood value of the (rescaled) uni-modal density fitted on $[g_i, g_{j+1}]$. In formula,

$$\mathbf{S}[i, j] = \sum_{l=1}^{n} \log\left(h_{ij}(x_l)\right) \mathbb{1}\{g_i \le x_l < g_{j+1}\} + n_{ij} \log\left(\frac{n_{ij}}{n}\right), \tag{4.4}$$

where h_{ij} denote the uni-modal density estimated on $[g_i, g_{j+1}]$, and n_{ij} denote the number of data points in that interval.

In addition to $\mathbf{S}$, let $\mathbf{D}$ be an M-by-K lower-triangular matrix, where $\mathbf{D}[m, k]$, $k = 1, \ldots, K$,

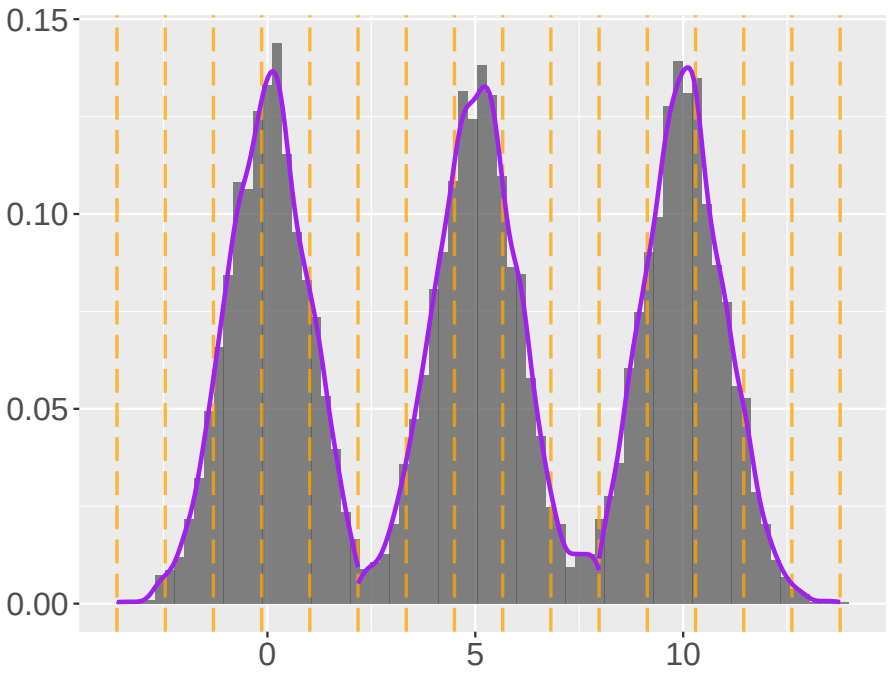


Figure 4.3. An illustrative example on our methodology. It is based on a sample of size $n = 10,000$ generated from the mixture model $\frac{1}{3}\mathcal{N}(0,1) + \frac{1}{3}\mathcal{N}(5,1) + \frac{1}{3}\mathcal{N}(10,1)$. The number of modes for this mixture is $K = 3$. In dark grey we plot a histogram of the sample. We consider the default $M = 5K - 1 = 14$ grid points as candidate locations for the knots $\boldsymbol{\lambda} = (\lambda_2, \lambda_3)$. We plot these grid points, $\mathbf{g} = (g_2, \ldots, g_{15})$, in vertical dotted orange lines, together with the extremes, in the left and right dotted orange lines. The m-th interval of interest, is located between the m-th and $(m+1)$-th dotted orange lines, where $m = 1, 2, \ldots, 15$. The purple line is a drawing of the density found by our dynamic programming method with knowledge of the correct number of modes.

$m = k, \ldots, M$, denote the maximum log-likelihood of fitting k (rescaled) uni-modal densities on $[g_1, g_{m+1}]$. By definition, only entries $\mathbf{D}[m,k]$, $m \geq k$ carry meaning as it would be impossible to fit a k-modal density on $[g_1, g_{m+1}]$ if $m < k$ as we assume that the density has at most one modal interval in any interval of the form $[g_k, g_{k+1}]$.

We initialize $\mathbf{D}$ as follows

$$\mathbf{D}[m,1] = \mathbf{S}[1,m], \; m = 1, 2, \ldots, M. \tag{4.5}$$

From there, the dynamic programming recursion takes the form

$$\mathbf{D}[m,k] = \max_{k-1 \le i \le m-1} \mathbf{D}[i,k-1] + \mathbf{S}[i+1,m], \tag{4.6}$$

for $k = 2,\ldots,K$ and $m = k,\ldots,M$.

To find the actual values of λ, we also create an M-by-K matrix $\mathbf{I}$. In the setting with the highest log-likelihood, $\mathbf{I}[m,k]$ indicates the index of the first interval of the k-th uni-modal density, when fitting k uni-modal densities on the first m intervals. By definition, the initialization is,

$$\mathbf{I}[m,1] = 1. \tag{4.7}$$

Then $\mathbf{I}[m,k]$ is updated as the maximizing index i in (4.6).

The algorithm for finding the optimal log-likelihood, as well as the optimal knot values of the K-modal density, is summarized in Table 4.1, while the algorithm for finding $\mathbf{S}$, $\mathbf{D}$, and $\mathbf{I}$ is now summarized in Table 4.2.

Table 4.1. Dynamic programming approach for fitting a density with K modes, which returns the maximum likelihood $\mathbf{D}[M,K]$ and the optimal knots $\lambda^\star$

inputs: $\mathbf{D}$ and $\mathbf{I}$ (from Table 4.2), number of modes K, grid $\mathbf{g}$ of size M

initialize $\mathbf{b} = \mathbf{0}_{K+1}$ and $\mathbf{b}[K+1] = M+1$
initialize $l = M$
for $k = K,\ldots,1$ **do**
  $\mathbf{b}[k] = \mathbf{I}[l,k]$
  $l = \mathbf{I}[l,k] - 1$
$\lambda^\star = \mathbf{g}[\mathbf{b}]$
return $\lambda^\star$, $\mathbf{D}[M,K]$

4.3.5 Multi-grid search

We note that an initial grid search over a large number of grid points is time consuming, so we first use a coarse grid, and then refine locally around each pair of knots using a finer grid.

Table 4.2. Rountine for computing **S**, **D** and **I**

inputs: data **x**, number of modes K, grid **g** of size M

compute **S** as in (4.4)
initialize **D** as in (4.5) and **I** as in (4.7)
if $K = 1$ **then**
go to return step
else if $K \geq 2$ **then**
  for $k = 2, \ldots, K$ **do**
    for $m = k, \ldots, M$ **do**
      initialize $\mathbf{p} = \mathbf{0}_{m-k+1}$
      for $i = k-1, \ldots, m-1$ **do**
        $\mathbf{p}[i-k+2] = \mathbf{D}[i, k-1] + \mathbf{S}[i+1, m]$
        $\mathbf{D}[m, k] = \max(\mathbf{p})$
        $\mathbf{I}[m, k] = \arg\max(\mathbf{p}) + k - 1$
return **S**, **D**, **I**

Suppose we have found $\lambda^\star = \{\lambda_2^\star, \ldots, \lambda_K^\star\}$ as optimal solution for the knots on a grid containing M intervals. We then consider, for each k, a local grid of size L spanning the interval $[\lambda_k^* - r, \lambda_k^* + r]$; this local grid is denoted as a vector θ_k. We then optimize the log-likelihood over all the combinations of $K-1$ values[1], one from each θ_k. The parameters L and r can be set by the user, although the parameter r is preferably chosen so that $2r < \delta^* = \min_k(\lambda_{k+1}^* - \lambda_k^*)$.

4.4 Selecting the number of modes

Our dynamic programming approach in the previous section is based on knowing the value of K. In this section, we introduce two data-driven ways for selecting K when it is unknown.

4.4.1 Literature on testing multi-modality

There is a significant amount of literature on testing for multi-modality. For instance, Silverman (1981) introduced a way of using kernel density estimates to investigate multi-modality, where the smoothing parameter is chosen automatically. This method was followed up

[1]The left and right end points can be taken as the minimum and maximum values of the data points, respectively.

by Hall and York (2001); Mammen et al. (1992), who considered its calibration and asymptotic characteristics. Hartigan and Hartigan (1985) proposed the dip test, which measures multi-modality in a sample by the maximum difference between the empirical distribution function and a chosen uni-modal distribution. Later, Müller and Sawitzki (1991) considered the excess mass functional, which measures excessive empirical mass in comparison with multiples of uniform distributions. From the perspective of hierarchical clustering, Hartigan and Mohanty (1992) designed a test for uni-modality based on the runt size in single linkage hierarchical clustering, defined as the number of points in a cluster's smallest subcluster. For an early survey see, e.g., (Fischer et al., 1994). Testing of the significance of a mode was considered by Chacón and Duong (2013); Cheng and Hall (1999); Duong et al. (2008); Genovese et al. (2016); Godtliebsen et al. (2002); Minnotte (1997); Rufibach and Walther (2010), among others. Hengartner and Stark (1995) developed a lower confidence bound for the number of modes. Multi-scale methods aiming at partitioning the real-line into modal intervals, such as those of Minnotte and Scott (1993), Chaudhuri and Marron (2000, 1999), and Dümbgen and Walther (2008), already mentioned in the introduction, are based on testing the significance of modes.

We note that closely related to the problem of selecting the number of modes is the problem of selecting the number of clusters (Rousseeuw, 1987; Wang, 2010) or choosing the number of components in a mixture model (Celisse, 2014; McLachlan, 1987; McLachlan and Rathnayake, 2014).

4.4.2 Our approaches

We offer two data-driven options here for choosing the number of modes K, using our dynamic programming approach in Section 4.3.4.

Measure of fit

Let $\Delta(F, G)$ be a measure of fit for comparing the distribution functions F and G. An example is the one used for the Kolmogorov–Smirnov test which corresponds to the supnorm,

i.e., $\Delta(F,G) = \sup_x |F(x) - G(x)|$, which is particularly easy to intuit. Let $\hat{F}$ denote the empirical distribution of $\mathbf{x}$ and $\hat{F}_K$ denote the distribution function under the maximum likelihood criterion with K modes. In the spirit of Hartigan and Hartigan (1985), the general idea is to 'look' at how $\Delta(\hat{F}_K, \hat{F})$ varies with K. For the particular choice of the supnorm, one could set a threshold τ, for example, $\tau = 0.01$, and select $K^* = \min\{K : \Delta(\hat{F}_K, \hat{F}) \leq \tau\}$.

Cross validation

We can also adopt a cross validation procedure. For a comprehensive review on applying cross validation procedures to model selection problems, we invite the reader to (Arlot and Celisse, 2010). For every candidate K for the number of modes, we can acquire a cross-validation log-likelihood. Specifically, if we are doing 5-fold cross validation, we would fit a K-modal density on the datapoints in the training folds, and then evaluate the log-likelihood of this model on the datapoints in the validation fold. We take an average over the folds at the end. We note that the log-likelihood could also be replaced by the L_2 norm, i.e., the integrated mean squared error. Due to computational reasons, this method is suited if the largest candidate for K is not too large, for example, $K = 5$.

Remark 1. *We note that since the selection of K involves potentially a large amount of overhead computations, one might want to use a coarser computation of the density, which can then be refined when the choice of K is settled.*

4.5 Numerical experiments

In this section, we provide some numerical experiments to illustrate our dynamic programming method and its application to determining the number of modes.

4.5.1 Application to real dataset

We first apply our method to the Old Faithful Geyser dataset (Azzalini and Bowman, 1990) in the MASS package of R. This dataset contains 272 waiting times, in minutes, indicating

the duration between eruptions for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. For this dataset, we consider fitting a density with the widely used $K = 2$ modes, for example, in (Pu and Arias-Castro, 2020). As initialization, we select $M = 5K$. We also use the multi-grid search in Section 4.3.5, with $L = 15$ and $r = \delta^*(1/2 - 1/2L)$.

In addition to our method, we provide the estimates of the regular kernel density estimator(KDE) with bandwidth chosen by the "rule of thumb" as in (Silverman, 1986, page 28). We also provide the density constrained to have two modes (Wolters and Braun, 2018a), acquired from adding an additive curve to the regular kernel density, implemented in the R package scdensity (Wolters and Braun, 2018b).

The histogram of the waiting times, and the 3 resulting densities, are shown in Figure 4.4. The regular KDE achieves a negative log-likelihood of 1160.584; the method of Wolters and Braun (2018a) achieves a negative log-likelihood of 1167.776; and our method achieves a negative log-likelihood of 944.366. We also note that in this dataset, our method gives a density that fits the data better both when the density of the histogram is low and when it is high.

4.5.2 Selecting the number of modes

We next consider selecting the number of modes when it is not known. To do so, we generate random mixture models, and check the performance of our method on these random mixtures.

We start by considering random Gaussian mixture models. We first uniformly sample the number of components from $1, 2, \ldots, 5$. Once given the number of components, we randomly generate each components' mean values uniformly from $[0, 10]$, and standard deviation values exponentially with parameter 1. We note that for each of these randomly generated Gaussian mixture models, the actual number of modes is not always the same as the number of components, and therefore we also compute the number of modes in each model by the amount of locations where the derivative changes from increasing to decreasing. For each randomly generated model, we generate a sample of size $n = 10,000$. We consider here possible candidates of K ranging

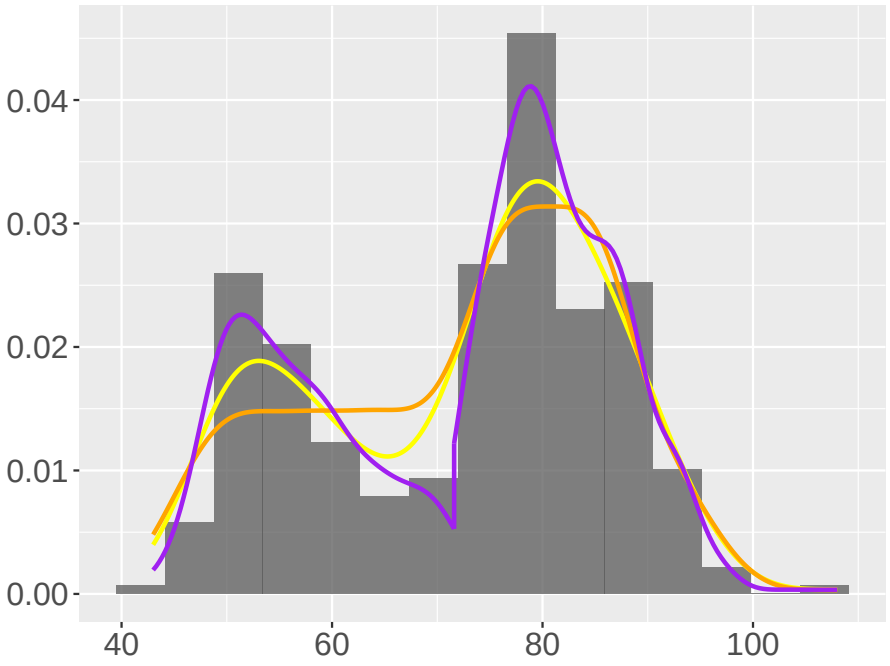


Figure 4.4. Histogram of waiting times, in minutes, of the Geyser dataset, overlaid with density estimators acquired by regular KDE (in yellow), by additive curve (in orange), and by dynamic programming (in purple). We note that our density acquired from dynamic programming (in purple) is interpolated between the area where the two modal densities meet, i.e. around 71.6.

Table 4.3. Amount of times (out of 100) for choosing the actual number of modes by measure of fit, among data from randomly generated mixture models. We select the smallest K that has a KS distance less than γ. We present 0.01 and 0.05 as the standard γ values, and also present the γ value resulting in the largest amount of correct selections.

Random Mixture Distribution	γ	**Amount of Times for Selecting the Actual K (out of 100)**
Gaussian	0.01	76
	0.05	59
	0.0162	78
Laplace	0.01	66
	0.05	50
	0.0104	68

from $1, 2, \ldots, 5$. (This is mainly for computational considerations, as our experiments necessarily involve replicates. In practice, the range of K can be much larger.) For each possible value of K, we initialize $M = 5K$. We also use the multi-grid search in Section 4.3.5 with $L = 5$ and $r = \delta^*(1/2 - 1/2L)$. We generate $B = 100$ random Gaussian mixtures.

We then consider random Laplace mixture models. We use the same sampling method as the Gaussian case to generate the number of components, the mean values, and the standard deviations of each component. We also compute the actual number of modes like in the Gaussian case. Similarly, we generate $B = 100$ random Laplace mixtures. We use these data from random mixtures in the following two subsections to determine the number of modal intervals.

Measure of fit

In this subsection, we determine the number of uni-modal intervals based on a measure of fit as in Section 4.4.2. Specifically, we use the supnorm, as detailed in that same subsection. The result on the number of times the number of uni-modal intervals has been correctly identified is summarized in Table 4.3.

Table 4.4. Amount of times (out of 100) for choosing the actual number of modes by cross validation, among data from randomly generated mixture models. We select the smallest K under each given decreasing percentage of improvement. We present 1% as the default percentage, and also present the percentage value resulting in the largest amount of correct selections.

Random Mixture Distribution	**Percentage (in %)**	**Amount of Times for Selecting the Actual K (out of 100)**
Gaussian	1	66
	0.08	75
Laplace	1	53
	0.17	62

Cross validation

In this subsection we conduct simulations to determine the number of modes based on cross validation as in Section 4.4.2. Here we use 5 folds, and set $L = 3$. At any given value K, if increasing the number of modes from K to $K+1$ does not result in an improvement of the log-likelihood by more than 1%, we select the K as the number of modes; otherwise we move on to comparing $K+1$ and $K+2$; and so on. The result on the number of times the correct number of modes is selected is summarized in Table 4.4.

4.6 Discussion

In this paper, we considered a dynamic programming approach to fitting a probability density function when it is constrained to have a given number of modal intervals. Based on this approach, we then gave two data driven ways for selecting the number of modal intervals when it is not given.

It is in principle possible to develop a similar dynamic programming approach to fitting a density whose shape is constrained in a certain way over a number of intervals defining a partition of the real line. For example, we can consider fitting a density which is log-concave (Walther, 2009) in each of K intervals defining a partition. Importantly, though, this dynamic programming approach seems limited to real-valued data.

4.7 Acknowledgement

This chapter, in full, is a version of the paper “Fitting a Multi-modal Density by Dynamic Programming”, He Jiang and Ery Arias-Castro. It is currently being prepared for submission for publication of the material. The dissertation author is the primary investigator and corresponding author of this material.

Chapter 5

On the Consistency of Metric and Non-Metric K-medoids

5.1 Abstract

We establish the consistency of K-medoids in the context of metric spaces. We start by proving that K-medoids is asymptotically equivalent to K-means restricted to the support of the underlying distribution under general conditions, including a wide selection of loss functions. This asymptotic equivalence, in turn, enables us to apply the work of Pärna (1986) on the consistency of K-means. This general approach applies also to non-metric settings where only an ordering of the dissimilarities is available. We consider two types of ordinal information: one where all quadruple comparisons are available; and one where only triple comparisons are available. We provide some numerical experiments to illustrate our theory.

5.2 Introduction

Cluster analysis is widely regarded as one of the most important tasks in unsupervised data analysis (Jain et al., 1999; Kaufman and Rousseeuw, 2009). In this paper, we consider several center based clustering methods. Specifically, we show the asymptotic equivalence of K-means and K-medoids, and use this equivalence to prove the consistency of K-medoids in metric and non-metric (i.e., ordinal) settings.

5.2.1 K-means and K-medoids

The problem of K-means can be traced back to the 1960's to early work of MacQueen (1967). As the problem is computationally difficult in higher dimensions or when the number of clusters is large, it is instead most often approached via iterative methods such as Lloyd's algorithm (Lloyd, 1982). Leaving these computational challenges aside, assuming the problem is solved exactly, the consistency of K-means as a method has been thoroughly addressed in the literature. Early in this line of work, Pollard (1981) established the consistency of K-means in Euclidean spaces. Pärna (1986) extended the result to separable metric spaces, while Pärna (1988, 1990, 1992) examined the particular situation of Hilbert and Banach spaces, where the existence of an optimal solution had been considered by Herrndorf (1983) and Cuesta and Matrán (1988).

The problem of K-medoids dates back to the 1980's to work of Kaufman and Rousseeuw (1987), who in the process proposed the Partition Around Medoids (PAM) iterative algorithm. Van der Laan et al. (2003) discovered that the original PAM has problem with recognizing rather small clusters, and defined a new version of PAM based on maximizing average silhouette, as defined by Kaufman and Rousseeuw (1990). Later Park and Jun (2009) proposed a computationally simpler version of PAM akin to Lloyd's algorithm for K-means. See (Kaufman and Rousseeuw, 2009, Ch 2). In a setting where the goal is the clustering of data sequences, Wang et al. (2019) established an exponential consistency result for K-medoids itself (when solved exactly). To the best of our knowledge, however, the consistency of K-medoids in the more standard setting of clustering points in a metric space has not been previously established.

We establish the consistency of K-medoids by first showing that K-medoids is asymptotically equivalent to K-means restricted to the support of the underlying distribution, and then leveraging the work of Pärna (1986) on the consistency of K-means in metric spaces.

5.2.2 Ordinal K-medoids

Beyond the more standard setting where the distances are available to us, we also consider ordinal settings where only an ordering of the distances is available. Even when the dissimilarities are available, turning them into ranks, and thus only working with the underlying ordinal information, can be attractive in situations where the numerical value of the dissimilarities has little meaning besides providing an ordering. This is the case, for example, in psychological experiments where human subjects are tasked with rating some items in order of preference. Working with ranks also has the advantage of added robustness to outliers.

Statisticians and other data scientists have dealt with ordinal information for decades. Without going too far afield into rank-based inference (Hájek and Sidák, 1967) or ranking models (Bradley and Terry, 1952), there is non-metric scaling, aka ordinal embedding, which is the problem of embedding a set of items based on an ordering of their pairwise dissimilarities, with pioneering work in the 1960's by Shepard (1962a,b) and Kruskal (1964). The consistency of ordinal embedding — by which we mean any solution to the problem assuming one exists — was already considered by Shepard (1966), and more thoroughly addressed only recently by Kleindessner and Luxburg (2014) and Arias-Castro (2017).

Even closer to our situation, in the area of clustering, we know that hierarchical clustering with either single or complete linkage (or the less popular median linkage) only use the ordinal information, as can be seen from the fact that the output grouping remains the same if the dissimilarities are transformed by the application of a monotonically increasing function. The well-known clustering method DBSCAN of Ester et al. (1996) can be seen as a robust variant of single linkage, in its nearest-neighbor formulation, only relies on ordinal information as well. On the other hand, hierarchical clustering with either average linkage or Ward's criterion does not have that property. The use of K-medoids in ordinal settings does not seem nearly as widespread. In fact, we could only find a few references where the idea was proposed, scattered across various fields such as computer vision (Zhu et al., 2011) and data mining (Zadegan et al., 2013). In the

context of an application to the clustering of pictures of human faces, Zhu et al. (2011) proposed a rank order distance (ROD) based on a sum of individual ranks, acquired from triple comparisons, and then applied single linkage hierarchical clustering with this distance. They argued that this distance was more appropriate for their particular application than the more standard L_1 distance. In a followup work, Huang et al. (2020) proposed a kernel variant of ROD. With the intention of making the clustering result less sensitive to initialization and potential outliers, Zadegan et al. (2013) proposed the concept of hostility index based on a sum of ranks obtained from triple comparisons. Aside from these, Achtert et al. (2006) proposed a dissimilarity based on the distance to the ℓ-th nearest neighbor, which can therefore be implemented based solely on ordinal information.

Besides putting ordinal K-medoids in the context of ordinal data, as we just did, we establish its consistency for two types of ordinal information: quadruple comparisons giving an overall ranking of all pairwise dissimilarities; and triple comparisons giving a ranking relative to each sample point.

5.2.3 Setting and content

We consider the problem of clustering some data points in a metric space into k clusters, where k is given. The metric space is denoted $(\mathcal{X}, \mathsf{d})$ and assumed to be a locally compact Polish space. The sample is denoted $x_1, \ldots, x_n$ and assumed to have been drawn from a Borel probability measure Q assumed to have bounded support[1] containing at least k points. We will let Q_n denote the empirical distribution, namely, $Q_n(B) := \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \in B\}}$ for any set $B \subset \mathcal{X}$. For two sets $A, B \subset \mathcal{X}$, define

$$\mathsf{H}(A|B) := \sup_{a \in A} \inf_{b \in B} \mathsf{d}(a,b), \tag{5.1}$$

so that the Hausdorff distance between A and B is $\max\{\mathsf{H}(A|B), \mathsf{H}(B|A)\}$.

The organization of the paper will be as follows. In Section 5.3, we prove the asymptotic

[1] This is for convenience. See (Pärna, 1986).

equivalence of K-means and K-medoids, and deduce from that the consistency of K-medoids in the metric setting. In Section 5.4, we consider two ordinal settings, based on quadruple and triple comparisons respectively, and establish the consistency of K-medoids in each case using the equivalence result from Section 5.3. We provide numerical experiments along the way to illustrate our theoretical results. Our work is greatly inspired by that of Pärna (1986), and we will refer to his work often.

Remark 2. *We want to mention that all our results apply when $\mathcal{X}$ is a finite dimensional Banach space and Q has a density with respect to the Lebesgue measure which is bounded and has compact support.*

5.3 Consistency of K-medoids

For a k-tuple $A \subset \mathcal{X}$, consider the risk

$$L(A, Q) = \int_{\mathcal{X}} \min_{a \in A} \phi(\mathsf{d}(x, a)) dQ(x), \tag{5.2}$$

where $\phi : [0, \infty) \to [0, \infty)$ is a loss function assumed to be non-decreasing, continuous, and such that $\phi(d) = 0$ if and only if $d = 0$ — all these assumptions being rather standard. By K-means we mean the result of the following optimization problem:

$$\text{minimize } L(A, Q_n) \quad \text{over } A \subset \mathcal{X},\ |A| = k. \tag{5.3}$$

And by K-medoids we mean the same optimization problem but restricted to k-tuples made of sample points:

$$\text{minimize } L(A, Q_n) \quad \text{over } A \subset \mathcal{X}_n,\ |A| = k, \tag{5.4}$$

where $\mathcal{X}_n := \{x_1, \ldots, x_n\}$. Note that

$$L(A, Q_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{a \in A} \phi(\mathsf{d}(x_i, a)). \tag{5.5}$$

It is well-known that, as formulated, (5.3) and (5.4) can behave quite differently. Take for example the case of the real line with Q the uniform distribution on $[-2,-1]\cup[1,2]$. When $k=1$, in the large-sample limit, the origin is the unique solution to K-means problem, while -1 and 1 are the solutions to the K-medoids problem. Instead, we consider the following restricted form of K-means:

$$\text{minimize } L(A,Q_n) \quad \text{over } A\subset \text{supp}(Q),\ |A|=k, \tag{5.6}$$

where $\text{supp}(Q)$ denotes the support of Q. The analyst cannot consider this problem when the support of Q is unknown, which is typically the case. But this optimization problem is only used as a device to analyze the asymptotic behavior of K-medoids.

Theorem 1. *In the present context, K-medoids* (5.4) *is asymptotically equivalent to K-means* (5.6)*, which in turn is asymptotically equivalent to population version of the same problem, namely*

$$\textit{minimize } L(A,Q) \quad \textit{over } A\subset \text{supp}(Q),\ |A|=k. \tag{5.7}$$

We conclude that, if A_n^* is a solution to (5.4)*, then in probability,*

$$L(A_n^*,Q)\xrightarrow{n\to\infty}\min_{|A|=k}L(A,Q). \tag{5.8}$$

Remark 3. *As discussed in (Cuesta and Matrán, 1988; Pärna, 1990, 1992), a K-means problem may not have a solution. In our situation, however, we are assuming that the space is a locally compact Polish space, and a solution can be shown to exist by a simple compactness argument together with our assumptions on ϕ (and the fact that the distance function is always continuous in any metric space it equips). This applies to* (5.3)*,* (5.6) *and* (5.7)*.*

Proof. Since everything happens within the support of Q, we may assume without loss of generality that Q is supported on the entire space, meaning that $\text{supp}(Q)=\mathcal{X}$. And since we assume $\text{supp}(Q)$ to be bounded, we are effectively assuming that $\mathcal{X}$ is bounded, and therefore compact since it is assumed to be locally compact.

The asymptotic equivalence of (5.6) and (5.7) is the consistency result of Pärna (1986). It can be deduced easily from the arguments we present below, which themselves are by-and-large adapted from (Pärna, 1986). So all we are left to do is prove that (5.4) is asymptotically equivalent to (5.6). To be sure, by this we mean that, if A_n^* is a solution to the former and A_n a solution to the latter, then

$$\left|L(A_n^*, Q_n) - L(A_n, Q_n)\right| \xrightarrow{n\to\infty} 0, \tag{5.9}$$

in probability. Because by definition $L(A_n^*, Q_n) \geq L(A_n, Q_n)$, all we need to show is that

$$\limsup_{n\to\infty} L(A_n^*, Q_n) - L(A_n, Q_n) \leq 0. \tag{5.10}$$

The remaining of the proof consists of three steps. We first show in Lemma 2 below that $L(A, Q_n) \to L(A, Q)$ as $n \to \infty$, uniformly over A. We then show in Lemma 4 further down that $A \mapsto L(A, Q)$ is uniformly continuous. The last step consists in using these results in conjunction with the 'squeeze theorem'.

By the uniform convergence established in Lemma 2, we have

$$\lim_{n\to\infty} |L(A_n^*, Q_n) - L(A_n^*, Q)| = 0, \tag{5.11}$$

as well as

$$\lim_{n\to\infty} |L(A_n, Q_n) - L(A_n, Q)| = 0. \tag{5.12}$$

Therefore, all we need to show is that

$$\limsup_{n\to\infty} L(A_n^*, Q) - L(A_n, Q) \leq 0. \tag{5.13}$$

For every point in A_n find the closest sample point, and gather all these in B_n^*.

Note that by Lemma 1,

$$h_n := \mathsf{H}(A_n|B_n) = \max_{a\in A_n}\min_{b\in B_n^*}\mathsf{d}(a,b) \xrightarrow{n\to\infty} 0, \tag{5.14}$$

in probability. Hence, by Lemma 4, we have

$$\limsup_{n\to\infty} L(B_n^*,Q) - L(A_n,Q) \le \limsup_{n\to\infty}\omega(h_n) = 0. \tag{5.15}$$

With the fact that $L(A_n^*,Q_n) \le L(B_n^*,Q_n)$ by definition of A_n^*, together with the uniform convergence also giving

$$\lim_{n\to\infty} |L(B_n^*,Q_n) - L(B_n^*,Q)| = 0, \tag{5.16}$$

we thus conclude that (5.13) holds. □

Lemma 1. *Assuming that $\mathcal{X}$ is compact and that* $\mathrm{supp}(Q) = \mathcal{X}$*, in probability,*

$$\mathsf{H}(\mathcal{X}|\mathcal{X}_n) = \sup_{x\in\mathcal{X}}\min_{i\in[n]}\mathsf{d}(x,x_i) \to 0, \quad n\to\infty. \tag{5.17}$$

Proof. The arguments are standard and follow from the definition of $\mathrm{supp}(Q)$. Indeed, $\mathrm{supp}(Q)$ is the complement of the largest open set D in $\mathcal{X}$ such that $Q(D) = 0$. Since $\mathrm{supp}(Q) = \mathcal{X}$ by assumption, it must be that $Q(B(x,r)) > 0$ for all $x \in \mathcal{X}$ and all $r > 0$, where $B(x,r)$ is defined as the closed ball centered at x with radius r. Fix $r > 0$. Because $\mathcal{X}$ is compact there is $y_1,\ldots,y_m \in \mathcal{X}$ such that $\mathcal{X} = \bigcup_j B(y_j,r)$. By the triangle inequality,

$$\begin{aligned}
&\mathsf{H}(\mathcal{X}|\mathcal{X}_n) \ge 2r \\
&\Leftrightarrow\ \exists x : \min_i \mathsf{d}(x,x_i) \ge 2r \\
&\Rightarrow\ \exists j : \min_i \mathsf{d}(y_j,x_i) \ge r,
\end{aligned}$$

and by the union bound, this implies that

$$\begin{aligned}
&\mathbb{P}(\mathsf{H}(\mathcal{X}|\mathcal{X}_n) \geq 2r) \\
&\leq \sum_j \mathbb{P}(\min_i \mathsf{d}(y_j, x_i) \geq r) \\
&= \sum_j (1 - Q(B(y_j, r))^n \\
&\leq m(1 - \min_j Q(B(y_j, r))^n \\
&\to 0, \quad n \to \infty.
\end{aligned}$$

Since $r > 0$ is arbitrary, the claim is established. □

5.3.1 Uniform convergence lemma

Lemma 2. *Assuming that $\mathcal{X}$ is compact, we have, in probability,*

$$\limsup_{n\to\infty} \sup_{|A|\leq k} |L(A, Q_n) - L(A, Q)| = 0. \tag{5.18}$$

The rest of this subsection is devoted to proving this lemma. It is enough to prove the variant where $|A| \leq k$ is replaced by $|A| = k$. The proof is very similar to the proof of (Pärna, 1986, Lem 1), with some differences. We provide a full proof for the sake of completeness.

Note that, like $L(A, Q)$, $L(A, Q_n)$ can be expressed as an integral:

$$L(A, Q_n) = \int_{\mathcal{X}} \min_{a\in A} \phi(\mathsf{d}(x, a)) dQ_n(x). \tag{5.19}$$

To each finite set A, we associate the following function

$$f_A(x) = \min_{a\in A} \phi(\mathsf{d}(x, a)). \tag{5.20}$$

Define the following class of functions

$$\mathcal{F} = \{f_A : A \subset \mathcal{X}, |A| = k\}. \tag{5.21}$$

Lemma 3 (Th 3.2 of (Rao, 1962))**.** *Let $\mathcal{F}$ be a family of continuous functions on a separable metric space $\mathcal{X}$ which is equicontinuous and admits a continuous envelope (there is g continuous such that $|f(x)| \leq g(x)$ for all $f \in \mathcal{F}$). In this context, suppose that (μ_n) is a sequence of measures on $\mathcal{X}$ converging weakly to μ, another measure on $\mathcal{X}$ with $\int g d\mu_n \to \int g d\mu < \infty$. Then we have:*

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} \left| \int f d\mu_n - \int f d\mu \right| = 0. \tag{5.22}$$

We apply this result with $\mu_n = Q_n$ and $\mu = Q$, for which the weak convergence is satisfied with probability 1 (Varadarajan, 1958). The existence of an envelope function g satisfying the requirements for the function class of interest, $\mathcal{F}$ above, is here immediate since for any A,

$$0 \leq f_A(x) \leq \phi(\text{diam}(\mathcal{X})) < \infty, \tag{5.23}$$

so that we may take $g \equiv \phi(\text{diam}(\mathcal{X}))$. It only remains to show $\mathcal{F}$ is equicontinuous. This amounts to showing that, for any $y_0 \in \mathcal{X}$ and any $\varepsilon > 0$, there exists a $\delta > 0$, such that $|f_A(y_0) - f_A(y)| < \varepsilon$ for any k-tuple A and any $y \in B(y_0, \varepsilon)$.

For the given y_0 and y, we denote $a(y_0)$ and $a(y)$ closest points in A to them so that

$$\min_{a\in A} \mathsf{d}(y_0, a) - \min_{a\in A} \mathsf{d}(y, a) = \mathsf{d}(y_0, a(y_0)) - \mathsf{d}(y, a(y)).$$

By definition and the triangle inequality,

$$\begin{aligned}
&\mathsf{d}(y_0, a(y_0)) - \mathsf{d}(y, a(y)) \\
&\leq \mathsf{d}(y_0, a(y)) - \mathsf{d}(y, a(y)) \\
&\leq \mathsf{d}(y_0, y),
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\mathsf{d}(y_0, a(y_0)) - \mathsf{d}(y, a(y)) \\
&\geq \mathsf{d}(y_0, a(y_0)) - \mathsf{d}(y, a(y_0)) \\
&\geq -\mathsf{d}(y_0, y).
\end{aligned}$$

We thus deduce that

$$\left| \min_{a \in A} \mathsf{d}(y_0, a) - \min_{a \in A} \mathsf{d}(y, a) \right| \leq \mathsf{d}(y_0, y). \tag{5.24}$$

Since ϕ is assumed to be continuous, it is uniformly continuous on $[0, \mathrm{diam}(\mathcal{X})]$. Let ω denote its modulus of continuity on that interval so that

$$\left|\phi(d) - \phi(d')\right| \leq \omega(|d - d'|), \quad \forall d, d' \in [0, \mathrm{diam}(\mathcal{X})].$$

We then have

$$\begin{aligned}
&|f_A(y_0) - f_A(y)| && (5.25)\\
&= \left| \min_{a \in A} \phi(\mathsf{d}(y_0, a)) - \min_{a \in A} \phi(\mathsf{d}(y, a)) \right| && (5.26)\\
&= \left| \phi\Big(\min_{a \in A} \mathsf{d}(y_0, a) \Big) - \phi\Big(\min_{a \in A} \mathsf{d}(y, a) \Big) \right| && (5.27)\\
&\leq \omega(\mathsf{d}(y_0, y)), && (5.28)
\end{aligned}$$

using the monotonicity of ϕ along the way. We have proved that $\mathcal{F}$ is indeed equicontinuous. Therefore the proof of Lemma 2 is complete.

5.3.2 Uniform continuity lemma

Lemma 4. *For any two sets $A, B \subset \mathcal{X}$, we have*

$$L(B,Q) \leq L(A,Q) + \omega(\mathsf{H}(A|B)), \tag{5.29}$$

where ω is the modulus of continuity of ϕ on $[0, \mathrm{diam}(\mathcal{X})]$.

The rest of this subsection is devoted to proving this lemma. Fix two sets $A, B \subset \mathcal{X}$, and let $h := \mathsf{H}(A|B)$. For any $a \in A$, define b_a as the closest point in B to a. Notice that by definition:

$$\mathsf{d}(a, b_a) \leq h, \tag{5.30}$$

and thus with the triangle inequality, for any point x we have:

$$\mathsf{d}(x,a) \geq \mathsf{d}(x,b_a) - \mathsf{d}(a,b_a) \geq \mathsf{d}(x,b_a) - h. \tag{5.31}$$

Taking minimums we get:

$$\min_{a \in A} \mathsf{d}(x,a) \geq \min_{a \in A} \mathsf{d}(x,b_a) - h \geq \min_{b \in B} \mathsf{d}(x,b) - h. \tag{5.32}$$

Using the fact that ϕ is non-decreasing, we then have:

$$\min_{a \in A} \phi(\mathsf{d}(x,a)) - \min_{b \in B} \phi(\mathsf{d}(x,b)) \tag{5.33}$$

$$= \phi\Big(\min_{a \in A} \mathsf{d}(x,a)\Big) - \phi\Big(\min_{b \in B} \mathsf{d}(x,b)\Big) \tag{5.34}$$

$$\geq -\omega(h). \tag{5.35}$$

Therefore, by integrating with respect to Q, we obtain:

$$\begin{aligned}
&L(A,Q) - L(B,Q) \\
&= \int \min_{a\in A} \phi(\mathsf{d}(x,a)) dQ(x) - \int \min_{b\in B} \phi(\mathsf{d}(x,b)) dQ(x) \\
&= \int \left[\min_{a\in A} \phi(\mathsf{d}(x,a)) - \min_{b\in B} \phi(\mathsf{d}(x,b)) \right] dQ(x) \\
&\geq -\omega(h).
\end{aligned}$$

5.3.3 Simulations

We report on a simple experiment illustrating the asymptotic equivalence established in Theorem 1. To keep a balance between the necessity to probe an asymptotic result (n large enough) and computational feasibility (n not too large), we choose to work with a sample of size $n = 2000$. We generate data from two equally weighted Gaussian distributions in R^2, centered at $(-0.5, 0)$ and $(0.5, 0)$, each with covariance $0.05 \times I_2$. Each setting is repeated 50 times. The result of this experiments is summarized in Table 5.1. As can be seen from this experiment, although varying according to different metrics and loss functions, the performance of K-means and K-medoids are indeed very similar.

5.4 Consistency of ordinal K-medoids

In this section we consider the problem of clustering with only an ordering of the dissimilarities. We consider two such orderings, one based on quadruple comparisons and another based on triple comparisons. We apply the results from Section 5.3 to show that, in both cases, K-medoids is consistent.

5.4.1 Quadruple comparisons

First we consider a situation in which all quadruple comparisons of the form 'Is $\mathsf{d}(x_i, x_j)$ larger or smaller than $\mathsf{d}(x_l, x_m)$?' are available. Equivalently, this is a situation in which a

Table 5.1. Mean values and standard deviations of the Average Center Error (error) and the Adjusted Rand Index (ARI) of K-means and K-medoids for various metrics and loss functions.

		K-means	K-medoids
L_1	error [$\times 10^{-2}$]	1.2 (0.4)	1.8 (0.7)
	ARI	0.780 (0.017)	0.778 (0.016)
$\sqrt{L_2}$	error [$\times 10^{-2}$]	8.9 (1.7)	11.7 (2.5)
	ARI	0.784 (0.020)	0.782 (0.022)
L_2	error [$\times 10^{-3}$]	9.4 (3.2)	12.3 (4.3)
	ARI	0.789 (0.017)	0.789 (0.017)
${L_2}^2$	error [$\times 10^{-4}$]	1.1 (0.9)	2.3 (1.7)
	ARI	0.785 (0.016)	0.785 (0.016)
L_∞	error [$\times 10^{-3}$]	8.9 (3.4)	12.0 (4.2)
	ARI	0.785 (0.021)	0.783 (0.020)

complete ordering of the pairwise dissimilarities is available.

For $i \in [n]$, let $R_i(a)$ denote the rank of $\mathsf{d}(x_i, a)$ among $\{\mathsf{d}(x_l, x_m) : l < m\}$, and for a k-tuple A, define

$$S_{\text{rank}}(A, Q_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{a \in A} \frac{R_i(a)}{\frac{n(n-1)}{2}}. \tag{5.36}$$

By ordinal K-medoids we mean the following optimization problem:

$$\text{minimize } S_{\text{rank}}(A, Q_n) \quad \text{over } A \subset \mathcal{X}_n,\ |A| = k. \tag{5.37}$$

This problem can be posed with the available information, and thus in principle can be solved. The equivalent restricted variant of ordinal K-means corresponds the following optimization problem:

$$\text{minimize } S_{\text{rank}}(A, Q_n) \quad \text{over } A \subset \text{supp}(Q),\ |A| = k. \tag{5.38}$$

As before, the latter is used as a bridge to show that the former is asymptotically equivalent to

the population version of this K-means problem, which is given by

$$\text{minimize } S(A,Q) \quad \text{over } A \subset \text{supp}(Q),\ |A| = k, \tag{5.39}$$

where

$$S(A,Q) := \int \min_{a\in A} G(\mathsf{d}(x,a))dQ(x), \tag{5.40}$$

with

$$G(t) := \mathbb{P}(\mathsf{d}(X,X') \le t), \tag{5.41}$$

X, X' being independent with distribution Q.

Here is the missing link between S_{rank} and S.

Lemma 5. *The following holds in probability:*

$$\limsup_{n\to\infty} \sup_{|A|\le k} |S_{\text{rank}}(A,Q_n) - S(A,Q_n)| = 0. \tag{5.42}$$

Proof. Let $\hat{G}_n$ denote the empirical distribution function of all the pairwise distances between sample points, meaning,

$$\hat{G}_n(t) := \frac{2}{n(n-1)} \sum_{l<m} 1_{\{\mathsf{d}(x_l,x_m)\le t\}}.$$

By the law of large numbers for U-statistics, in probability, $\hat{G}_n(t) \to G(t)$ as $n \to \infty$ for every fixed t. The Glivenko–Cantelli lemma does not quite apply as the pairwise distances are not an iid sample, but the two ingredients are there (van der Vaart, 2000): pointwise convergence as just stated, and the fact that $\hat{G}_N$ and G are both distribution functions in that they both are non-decreasing from 0 to 1 on $[0,\infty)$. Hence,

$$\varepsilon_n := \sup_t |\hat{G}_n(t) - G(t)| \xrightarrow{n\to\infty} 0,$$

in probability. We then have:

$$\begin{aligned} R_i(a) &= \sum_{l<m} 1_{\{\mathsf{d}(x_l,x_m)\le \mathsf{d}(x_i,a)\}} \\ &= \frac{n(n-1)}{2}\hat{G}_n(\mathsf{d}(x_i,a)) \\ &= \frac{n(n-1)}{2}G(\mathsf{d}(x_i,a)) \pm \frac{n(n-1)}{2}\varepsilon_n, \end{aligned}$$

giving

$$\begin{aligned} S_{\text{rank}}(A,Q_n) &= \frac{1}{n}\sum_{i=1}^{n}\min_{a\in A} G(\mathsf{d}(x_i,a)) \pm \varepsilon_n \\ &= S(A,Q_n) \pm \varepsilon_n, \end{aligned}$$

for any finite A, which establishes the result. □

Establishing the consistency of ordinal K-medoids is now a straightforward consequence of Theorem 1. We need to assume that G defined above is continuous, which is the case in the canonical situation of Remark 2.

Theorem 2. *In the present context, if A_n^* is a solution to ordinal K-medoids in the form of* (5.37)*, then in probability,*

$$S(A_n^*,Q) \xrightarrow{n\to\infty} \min_{|A|=k} S(A,Q). \tag{5.43}$$

Proof. By Lemma 5, we have that ordinal K-medoids (5.37) is asymptotically equivalent to the following problem:

$$\text{minimize } S(A,Q_n) \quad \text{over } A\subset \mathcal{X}_n,\ |A|=k. \tag{5.44}$$

But S is exactly as L in Section 5.3, with G replacing ϕ there, and since G satisfies the same properties assumed of ϕ, Theorem 1 applies to yield the claim. □

5.4.2 Triple comparisons

We turn to a situation in which only triple comparisons of the form 'Is $\mathsf{d}(x_i, x_j)$ larger or smaller than $\mathsf{d}(x_i, x_l)$?' are available. We do assume that all of these comparisons are on hand. Equivalently, this is a situation in which an ordering of the pairwise dissimilarities involving a particular point are available.

Hence, we work here with the ranks (re)defined as follows. For $i \in [n]$, let $R_i(a)$ denote the rank of $\mathsf{d}(x_i, a)$ among $\{\mathsf{d}(x_i, x_j) : j \neq i\}$, and for a k-tuple A, define

$$S_{\text{rank}}(A, Q_n) = \frac{1}{n} \sum_{i=1}^{n} \min_{a \in A} \frac{R_i(a)}{n-1}. \tag{5.45}$$

Ordinal K-medoids and (restricted) ordinal K-means are otherwise defined as before. The population equivalent to ordinal K-means is now given by

$$\text{minimize } S(A, Q) \quad \text{over } A \subset \text{supp}(Q),\ |A| = k, \tag{5.46}$$

where

$$S(A, Q) := \int \min_{a \in A} G^x(\mathsf{d}(x, a)) dQ(x), \tag{5.47}$$

with

$$G^x(t) := \mathbb{P}(\mathsf{d}(x, X') \leq t), \tag{5.48}$$

X' having distribution Q.

Lemma 6. *The following holds in probability:*

$$\limsup_{n \to \infty} \sup_{|A| \leq k} |S_{\text{rank}}(A, Q_n) - S(A, Q_n)| = 0. \tag{5.49}$$

Proof. Define

$$\hat{G}_{n,i}(t) := \frac{1}{n-1} \sum_{j \neq i} 1_{\{\mathsf{d}(x_i,x_j) \leq t\}}.$$

By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, for each i and any $\varepsilon > 0$, we have:

$$\mathbb{P}\Big(\sup_t |\hat{G}_{n,i}(t) - G^{x_i}(t)| > \varepsilon \Big) \leq 2\exp(-2(n-1)\varepsilon^2).$$

With this, and the union bound, we obtain:

$$\varepsilon_n := \max_i \sup_t |\hat{G}_{n,i}(t) - G^{x_i}(t)| \xrightarrow{n\to\infty} 0,$$

in probability. We then have:

$$\begin{aligned} R_i(a) &= \sum_{j \neq i} 1_{\{\mathsf{d}(x_i,x_j) \leq \mathsf{d}(x_i,a)\}} \\ &= (n-1)\hat{G}_{n,i}(\mathsf{d}(x_i,a)) \\ &= (n-1)G^{x_i}(\mathsf{d}(x_i,a)) \pm (n-1)\varepsilon_n, \end{aligned}$$

giving

$$\begin{aligned} S_{\text{rank}}(A,Q_n) &= \frac{1}{n} \sum_{i=1}^{n} \min_{a \in A} G^{x_i}(\mathsf{d}(x_i,a)) \pm \varepsilon_n \\ &= S(A,Q_n) \pm \varepsilon_n, \end{aligned}$$

for any finite A, which establishes the result. □

The following is our consistency result for K-medoids based on triple comparisons. It is not an immediate consequence of Theorem 1, but the proof arguments are parallel. We need to make additional assumption that $(x,t) \mapsto Q(B(x,t))$ is continuous on $\mathcal{X} \times (0,\infty)$. This is the case in the canonical situation of Remark 2.

Theorem 3. *In the present context, if A_n^* is a solution to ordinal K-medoids based on triple comparisons, then in probability,*

$$S(A_n^*, Q) \xrightarrow{n\to\infty} \min_{|A|=k} S(A, Q), \tag{5.50}$$

now with S defined as in (5.47).

Proof. By Lemma 6, we have that ordinal K-medoids is asymptotically equivalent to the following problem:

$$\text{minimize } S(A, Q_n) \quad \text{over } A \subset \mathcal{X}_n,\ |A| = k. \tag{5.51}$$

But unlike the situation in Theorem 2, now S is *not* exactly as L in Section 5.3, complicating matters a little bit. Nevertheless, the proof arguments are parallel to those underlying Theorem 1. As we did there, we need to establish uniform convergence and uniform continuity. As before, we may assume without loss of generality that $\mathcal{X}$ is compact and that $\operatorname{supp}(Q) = \mathcal{X}$. In that case, $(x,t) \mapsto Q(B(x,t))$ is uniformly continuous, and we let Ω denote its modulus of continuity so that

$$\left|Q(B(x,s)) - Q(B(y,t))\right| \le \Omega(\mathsf{d}(x,y), |s-t|),$$

for all $x, y \in \mathcal{X}$ and all $s, t > 0$.

For the uniform convergence, the proof of Lemma 2 proceeds as before until the very

end where instead

$$|f_A(y_0) - f_A(y)| \tag{5.52}$$

$$= \left| \min_{a \in A} G^{y_0}(\mathsf{d}(y_0, a)) - \min_{a \in A} G^{y}(\mathsf{d}(y, a)) \right| \tag{5.53}$$

$$= |G^{y_0}(\mathsf{d}(y_0, A)) - G^{y}(\mathsf{d}(y, A))| \tag{5.54}$$

$$\leq \Omega(\mathsf{d}(y_0, y), |\mathsf{d}(y_0, A) - \mathsf{d}(y, A)|) \tag{5.55}$$

$$\leq \Omega(\mathsf{d}(y_0, y), \mathsf{d}(y_0, y)) \tag{5.56}$$

$$\to 0, \quad \text{when } \mathsf{d}(y_0, y) \to 0. \tag{5.57}$$

For the uniform continuity, the proof of Lemma 4 proceeds as before except that

$$\min_{a \in A} G^{x}(\mathsf{d}(x, a)) - \min_{b \in B} G^{x}(\mathsf{d}(x, b)) \tag{5.58}$$

$$= G^{x}(\mathsf{d}(x, A)) - G^{x}(\mathsf{d}(x, B)) \tag{5.59}$$

$$\geq -\Omega(0, h), \tag{5.60}$$

with $h := \mathsf{H}(A|B)$ as in that proof, so that the statement of that lemma continues to hold but with $\omega(t) := \Omega(0, t)$. □

5.4.3 Simulations

We again report on a numerical experiment showcasing the results derived in this section in the context of ordinal clustering. We chose to work with a sample of size $n = 750$. We generate data from three equally weighted Gaussian distributions in two dimensions, centered at $(-0.5, 0)$, $(0.5, 0)$ and $(0, \sqrt{3}/2)$, each with covariance $0.05 \times I_2$. Each setting is repeated 50 times. The result of our experiment is summarized in Table 5.2. As can be seen from this experiment, K-medoids based on ordinal information performs nearly as well as K-medoids based on the full dissimilarity information.

Table 5.2. Mean values and standard deviations of the Average Center Error (error) and the Adjusted Rand Index (ARI) for various metrics and loss functions for K-medoids based on triple-comparisons (TC), quadruple-comparisons (QC), and the actual distances (KM).

		TC	**QC**	**KM**
L_1	error [$\times 10^{-2}$]	4.4	3.6	3.7
		(1.3)	(1.3)	(1.2)
	ARI	0.924	0.924	0.925
		(0.014)	(0.014)	(0.014)
$\sqrt{L_2}$	error [$\times 10^{-1}$]	1.8	1.6	1.7
		(0.3)	(0.2)	(0.3)
	ARI	0.928	0.929	0.929
		(0.016)	(0.015)	(0.016)
L_2	error [$\times 10^{-2}$]	3.2	2.7	2.7
		(0.8)	(0.9)	(0.9)
	ARI	0.934	0.933	0.933
		(0.016)	(0.016)	(0.016)
${L_2}^2$	error [$\times 10^{-3}$]	1.4	0.9	0.8
		(0.7)	(0.5)	(0.5)
	ARI	0.931	0.931	0.930
		(0.020)	(0.019)	(0.020)
L_∞	error [$\times 10^{-2}$]	3.0	2.6	2.7
		(1.1)	(1.0)	(1.0)
	ARI	0.918	0.917	0.918
		(0.017)	(0.017)	(0.017)

5.5 Discussion

In this paper, we have shown the asymptotic equivalence of K-means and K-medoids, and used this equivalence to prove the consistency of K-medoids in metric and non-metric situations.

5.5.1 Consistency of the solution

Our consistency results are on the value of the optimization problem defining K-medoids in the various settings we considered. Specifically, we showed in each case that $T(A_n^*, Q) \to_{n\to\infty} \min_A T(A, Q)$, in probability, where T is an appropriate criterion (either L or one of the two variants of S) and A_n^* is the solution to K-medoids. What about the behavior of the solution A_n^* itself?

Here the situation is completely generic: if the solution to the population problem, namely $A_{\text{opt}} := \arg\min_A T(A, Q)$, is unique, then $A_n^* \to_{n\to\infty} A_{\text{opt}}$, again in probability. This is simply due to the fact that in our setting we can reduce the situation to when $\mathcal{X}$ is compact, and in all cases we considered $A \mapsto T(A, Q)$ is continuous.

5.5.2 Clustering after embedding?

It might be possible to establish the consistency of ordinal K-medoids building on the consistency of ordinal embedding. This route appears unnecessarily sophisticated, however, in particular in light of a more straightforward approach that we built on the work of Pärna (1986). And from a computational standpoint, performing K-medoids in the ordinal setting has essentially the same complexity as in the regular (i.e., metric) setting, while methods for ordinal embedding tend to be much more demanding in computational resources.

5.5.3 A 'bad' variant of K-medoids

In the setting where triple comparisons are available, instead of defining the ranks as we did, we could have worked with the following definition. For $i \in [n]$, let $R_i(a)$ denote the rank of $\mathsf{d}(x_i, a)$ among $\{\mathsf{d}(x_j, a) : j \in [n]\}$. Although the resulting method can be analyzed in very

much the same way, it turns out to not be useful for the purpose of clustering. This is due to the fact that the corresponding optimization problem accepts a large range of solutions. To see this, consider the case $k = 1$. With the corresponding definition of S_{rank}, we have that

$$S_{\text{rank}}(a, Q_n) = \frac{1 + 2 + \cdots + n}{n(n-1)}, \tag{5.61}$$

for all $a \in \{x_1, ..., x_n\}$. And the problem persists for other values of k. For another example, consider clustering points distributed uniformly between $[-1, 1]$ into $k = 2$ clusters. It is clear that the correct population centers for K-means here are $\{-1/2, 1/2\}$. However, it can be seen that for any $1/2 \leq c \leq 1$, $A = \{-c, c\}$ also achieves the optimal population risk.

5.6 Acknowledgement

This chapter, in full, is a reprint of the paper "On the Consistency of Metric and Non-Metric K-Medoids", Ery Arias-Castro and He Jiang. It has been accepted for publication in the International Conference on Artificial Intelligence and Statistics 2021. The dissertation author is the primary investigator and corresponding author of this material.

We are very grateful to Professor Kalev Pärna for sharing with us his papers, which we could not otherwise access.

Bibliography

Achtert, E., C. Böhm, and P. Kröger (2006). Deli-clu: boosting robustness, completeness, usability, and efficiency of hierarchical clustering by a closest pair ranking. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 119–128. Springer.

Arias-Castro, E. (2017). Some theory for ordinal embedding. *Bernoulli 23*(3), 1663–1693.

Arias-Castro, E. and S. Chen (2017). Distribution-free multiple testing. *Electronic Journal of Statistics 11*(1), 1983–2001.

Arias-Castro, E. and W. Qiao (2021a). An asymptotic equivalence between the mean-shift algorithm and the cluster tree. *arXiv preprint arXiv:2111.10298*.

Arias-Castro, E. and W. Qiao (2021b). Moving up the cluster tree with the gradient flow. *arXiv preprint arXiv:2109.08362*.

Arias-Castro, E. and M. Wang (2017). Distribution-free tests for sparse heterogeneous mixtures. *TEST 26*(1), 71–94.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys 4*, 40–79.

Azzalini, A. and A. W. Bowman (1990). A look at some data on the Old Faithful Geyser. *Journal of the Royal Statistical Society: Series C 39*(3), 357–365.

Balabdaoui, F., H. Jankowski, M. Pavlides, A. Seregin, and J. Wellner (2011). On the Grenander estimator at zero. *Statistica Sinica*, 873–899.

Banerjee, A. and J. Ghosh (2006). Scalable clustering algorithms with balancing constraints. *Data Mining and Knowledge Discovery 13*(3), 365–395.

Basu, S., A. Banerjee, and R. J. Mooney (2004). Active semi-supervision for pairwise constrained clustering. In *SIAM International Conference on Data Mining*, pp. 333–344.

Basu, S., I. Davidson, and K. Wagstaff (2008). *Constrained Clustering: Advances in Algorithms,*

Theory, and Applications. CRC Press.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B 57*(1), 289–300.

Benjamini, Y. and Y. Hochberg (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics 25*(1), 60–83.

Bennett, K., P. Bradley, and A. Demiriz (2000). Constrained K-means clustering. Technical Report MSR-TR-2000-65, Microsoft Research.

Bickel, P. J. and J. Fan (1996). Some problems on the estimation of unimodal densities. *Statistica Sinica*, 23–45.

Bickel, P. J. and M. Rosenblatt (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics 1*(6), 1071–1095.

Birgé, L. (1987a). Estimating a density under order restrictions: Nonasymptotic minimax risk. *The Annals of Statistics*, 995–1012.

Birgé, L. (1987b). On the risk of histograms for estimating decreasing densities. *The Annals of Statistics*, 1013–1022.

Bordes, L., S. Mottelet, and P. Vandekerkhove (2006). Semiparametric estimation of a two-component mixture model. *The Annals of Statistics 34*(3), 1204–1232.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*(3/4), 324–345.

Bradley, S. P., A. C. Hax, and T. L. Magnanti (1977). *Applied Mathematical Programming*. Addison-Wesley.

Braun, W. J. and P. Hall (2001). Data sharpening for nonparametric inference subject to constraints. *Journal of Computational and Graphical Statistics 10*(4), 786–806.

Cai, T. T. and J. Jin (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics 38*(1), 100–145.

Carreira-Perpiñán, M. A. (2015). A review of Mean-Shift algorithms for clustering. *arXiv preprint arXiv:1503.00687*.

Cattaneo, M. D., M. Jansson, and X. Ma (2019). lpdensity: Local polynomial density estimation and inference. *arXiv preprint arXiv:1906.06529*.

Cattaneo, M. D., M. Jansson, and X. Ma (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association 115*(531), 1449–1455.

Celisse, A. (2014). Optimal cross-validation in density estimation with the L2-loss. *The Annals of Statistics 42*(5), 1879–1910.

Centers for Disease Control and Prevention (2001). Data table of stature-for-age charts. https://www.cdc.gov/growthcharts/html_charts/statage.htm#males.

Centers for Disease Control and Prevention (2010). National youth physical activity and nutrition study (NYPANS). https://www.cdc.gov/healthyyouth/data/yrbs/nypans.htm.

Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science 30*(4), 518–532.

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review 88*(1), 122–141.

Chacón, J. E. and T. Duong (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics 7*, 499–532.

Chang, G. T. and G. Walther (2007). Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis 51*(12), 6242–6251.

Chaudhuri, K., S. Dasgupta, S. Kpotufe, and U. Von Luxburg (2014). Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory 60*(12), 7900–7912.

Chaudhuri, P. and J. Marron (2000). Scale space view of curve estimation. *The Annals of Statistics 28*(2), 408–428.

Chaudhuri, P. and J. S. Marron (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association 94*(447), 807–823.

Chauveau, D. and D. Hunter (2013). ECM and MM algorithms for Normal mixtures with constrained parameters. https://hal.archives-ouvertes.fr/hal-00625285.

Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology 1*(1), 161–187.

Chen, Y.-C., C. R. Genovese, and L. Wasserman (2017). Statistical inference using the Morse-

Smale complex. *Electronic Journal of Statistics 11*(1), 1390–1433.

Cheng, G. and Y.-C. Chen (2019). Nonparametric inference via bootstrapping the debiased estimator. *Electronic Journal of Statistics 13*(1), 2194–2256.

Cheng, M.-Y., T. Gasser, and P. Hall (1999). Nonparametric density estimation under unimodality and monotonicity constraints. *Journal of Computational and Graphical Statistics 8*(1), 1–21.

Cheng, M.-Y. and P. Hall (1999). Mode testing in difficult cases. *The Annals of Statistics 27*(4), 1294–1315.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17*(8), 790–799.

Choi, E. and P. Hall (1999). Miscellanea. data sharpening as a prelude to density estimation. *Biometrika 86*(4), 941–947.

Chow, Y.-S., S. Geman, and L.-D. Wu (1983). Consistent cross-validated density estimation. *The Annals of Statistics 11*(1), 25–38.

Cline, D. B. H. and J. D. Hart (1991). Kernel estimation of densities with discontinuities or discontinuous derivatives. *Statistics 22*(1), 69–84.

Cohen, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics 9*(1), 15–28.

Cuesta, J. and C. Matrán (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probability Theory and Related Fields 78*(4), 523–534.

Cumings-Menon, R. (2018). Shape-constrained density estimation via optimal transport. *arXiv preprint arXiv:1710.09069*.

Dasgupta, S., D. Pati, I. H. Jermyn, and A. Srivastava (2021). Modality-constrained density estimation via deformable templates. *Technometrics*, 1–12.

Davidson, I. and S. Basu (2007). A survey of clustering with instance level constraints. *ACM Transactions on Knowledge Discovery from Data 1*(1-41), 2–42.

Davidson, I. and S. Ravi (2005). Clustering with constraints: feasibility issues and the K-means algorithm. In *SIAM International Conference on Data Mining*, pp. 138–149.

Davis, K. A., C. G. Park, and S. K. Sinha (2012). Testing for generalized linear mixed models with cluster correlated data under linear inequality constraints. *Canadian Journal of Statistics 40*(2),

243–258.

Day, N. E. (1969). Estimating the components of a mixture of Normal distributions. *Biometrika 56*(3), 463–474.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Dong, E., H. Du, and L. Gardner (2021). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet: Infectious Diseases 20*(5), 533–534.

Dümbgen, L. and G. Walther (2008). Multiscale inference about a density. *The Annals of Statistics 36*(4), 1758–1785.

Duong, T., A. Cowling, I. Koch, and M. P. Wand (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis 52*(9), 4225–4242.

Duran, B. S. and P. L. Odell (2013). *Cluster analysis: a survey*, Volume 100. Springer Science & Business Media.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics 35*(4), 1351–1377.

Efron, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association 96*(456), 1151–1160.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Volume 96, pp. 226–231.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics 21*(1), 196–216.

Fischer, N., E. Mammen, and J. S. Marron (1994). Testing for multimodality. *Computational Statistics & Data Analysis 18*(5), 499–512.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Franců, M., R. Kerman, and G. Sinnamon (2017). A new algorithm for approximating the least concave majorant. *Czechoslovak Mathematical Journal 67*(4), 1071–1093.

Fukunaga, K. and L. Hostetler (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory 21*(1), 32–40.

Gadat, S., J. Kahn, C. Marteau, and C. Maugis-Rabusseau (2020). Parameter recovery in two-component contamination mixtures: The L_2 strategy. *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques 56*(2), 1391–1418.

Gançarski, P., B. Crémilleux, G. Forestier, and T. Lampert (2020). Constrained clustering: current and new trends. In *A Guided Tour of Artificial Intelligence Research*, pp. 447–484. Springer.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B 64*(3), 499–517.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *The Annals of Statistics 32*(3), 1035–1061.

Genovese, C. R., M. Perone-Pacifico, I. Verdinelli, and L. Wasserman (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B 78*(1), 99–126.

Gill, P. E. and E. Wong (2012). Sequential quadratic programming methods. In J. Lee and S. Leyffer (Eds.), *Mixed Integer Nonlinear Programming*, pp. 147–224. Springer.

Giné, E. and R. Nickl (2010). Confidence bands in density estimation. *The Annals of Statistics 38*(2), 1122–1170.

Giné, E. and R. Nickl (2021). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Godtliebsen, F., J. Marron, and P. Chaudhuri (2002). Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics 11*(1), 1–21.

Goldfarb, D. and A. Idnani (1982). Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical Analysis*, pp. 226–239. Springer.

Goldfarb, D. and A. Idnani (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming 27*(1), 1–33.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*(5439), 531–537.

Gorokhovik, V. V. (2019). Minimal convex majorants of functions and Demyanov–Rubinov exhaustive super(sub)differentials. *Optimization 68*(10), 1933–1961.

Grenander, U. (1956). On the theory of mortality measurement: part ii. *Scandinavian Actuarial Journal 1956*(2), 125–153.

Gu, C. (1993). Smoothing spline density estimation: a dimensionless automatic algorithm. *Journal of the American Statistical Association 88*(422), 495–504.

Gu, C. and C. Qiu (1993). Smoothing spline density estimation: theory. *The Annals of Statistics 21*(1), 217–234.

Guidoum, A. C. (2020). Kernel estimator and bandwidth selection for density and its derivatives: The kedd package. *arXiv preprint arXiv:2012.06102*.

Hájek, J. and Z. Sidák (1967). Theory of rank tests.

Hall, P. and M. York (2001). On the calibration of silverman's test for multimodality. *Statistica Sinica*, 515–536.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc.

Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association 76*(374), 388–394.

Hartigan, J. A. and P. M. Hartigan (1985). The dip test of unimodality. *The Annals of Statistics*, 70–84.

Hartigan, J. A. and S. Mohanty (1992). The runt test for multimodality. *Journal of Classification 9*(1), 63–70.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for Normal mixture distributions. *The Annals of Statistics 13*(2), 795–800.

Hathaway, R. J. (1986). A constrained EM algorithm for univariate Normal mixtures. *Journal of Statistical Computation and Simulation 23*(3), 211–230.

Hengartner, N. W. and P. B. Stark (1995). Finite-sample confidence envelopes for shape-restricted densities. *The Annals of Statistics*, 525–550.

Herrndorf, N. (1983). Approximation of vector-valued random variables by constants. *Journal of Approximation Theory 37*(2), 175–181.

Hettmansperger, T. P. and J. W. McKean (2010). *Robust Nonparametric Statistical Methods*. CRC Press.

Hjort, N. L. and M. C. Jones (1996). Locally parametric nonparametric density estimation. *The*

Annals of Statistics 24(4), 1619–1647.

Hu, H., Y. Wu, and W. Yao (2016). Maximum likelihood estimation of the mixture of log-concave densities. *Computational Statistics & Data Analysis 101*, 137–147.

Huang, T., S. Wang, and W. Zhu (2020). An adaptive kernelized rank-order distance for clustering non-spherical data with high noise. *International Journal of Machine Learning and Cybernetics*, 1–13.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics 35*(1), 73 – 101.

Huber, P. J. and E. M. Ronchetti (2009). *Robust Statistics*. Wiley.

Hunter, D. R., S. Wang, and T. P. Hettmansperger (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics 35*(1), 224–251.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate Normal mixture models. *Statistical Methods and Applications 13*(2), 151–166.

Ingrassia, S. and R. Rocci (2007). Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis 51*(11), 5339–5351.

Ingrassia, S. and R. Rocci (2011). Degeneracy of the EM algorithm for the MLE of multivariate Gaussian mixtures and dynamic constraints. *Computational Statistics & Data Analysis 55*(4), 1715–1725.

Jain, A. K., M. N. Murty, and P. J. Flynn (1999). Data clustering: a review. *ACM Computing Surveys (CSUR) 31*(3), 264–323.

Jamshidian, M. (2004). On algorithms for restricted maximum likelihood estimation. *Computational Statistics & Data Analysis 45*(2), 137–157.

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society: Series B 70*(3), 461–493.

Jin, J. and T. T. Cai (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association 102*(478), 495–506.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics 7*(3), 310–321.

Karunamuni, R. J. and T. Alberts (2005). A generalized reflection method of boundary correction in kernel density estimation. *Canadian Journal of Statistics 33*(4), 497–509.

Kaufman, L. and P. Rousseeuw (1987). Clustering by means of medoids. In *Statistical Data Analysis Based on the L_1 Norm Conference, Neuchatel, 1987*, pp. 405–416.

Kaufman, L. and P. Rousseeuw (1990). Finding groups in data: An introduction to cluster analysis. *Hoboken NJ John Wiley & Sons Inc 725*.

Kaufman, L. and P. Rousseeuw (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*, Volume 344. John Wiley & Sons.

Kim, D. K. and J. M. Taylor (1995). The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *Journal of the American Statistical Association 90*(430), 708–716.

Kleindessner, M. and U. Luxburg (2014). Uniqueness of ordinal embedding. In *Conference on Learning Theory*, pp. 40–67.

Kraft, D. (1988). A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, Oberpfaffenhofen: Institut für Dynamik der Flugsysteme.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*(1), 1–27.

Langaas, M., B. H. Lindqvist, and E. Ferkingstad (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society: Series B 67*(4), 555–572.

Lesnick, T. G., S. Papapetropoulos, D. C. Mash, J. Ffrench-Mullen, L. Shehadeh, M. De Andrade, J. R. Henley, W. A. Rocca, J. E. Ahlskog, and D. M. Maraganore (2007). A genomic pathway approach to a complex disease: Axon guidance and Parkinson disease. *Public Library of Science Genetics 3*(6), e98.

Lindsay, B. G. (1995). *Mixture models: theory, geometry and applications*, Volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics.

Lindsay, B. G. and P. Basak (1993). Multivariate normal mixtures: a fast consistent method of moments. *Journal of the American Statistical Association 88*(422), 468–476.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory 28*(2), 129–137.

Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics 24*(4), 1602–1618.

Ma, Y. and W. Yao (2015). Flexible estimation of a semiparametric two-component mixture model with one parametric component. *Electronic Journal of Statistics 9*(1), 444–474.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

Mammen, E., J. S. Marron, and N. I. Fisher (1992). Some asymptotics for multimodality tests based on kernel density estimates. *Probability Theory and Related Fields 91*(1), 115–132.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 36*(3), 318–324.

McLachlan, G. J. and K. E. Basford (1988). *Mixture models: Inference and Applications to Clustering*. Marcel Dekker, Inc.

McLachlan, G. J. and T. Krishnan (2007). *The EM Algorithm and Extensions*, Volume 382. John Wiley & Sons.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual Review of Statistics and its Application 6*, 355–378.

McLachlan, G. J. and D. Peel (2004). *Finite Mixture Models*. John Wiley & Sons.

McLachlan, G. J. and S. Rathnayake (2014). On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4*(5), 341–355.

Meinshausen, N. and J. Rice (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics 34*(1), 373–393.

Menardi, G. (2016). A review on modal clustering. *International Statistical Review 84*(3), 413–433.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika 80*(2), 267–278.

Meyer, M. C. (2012). Nonparametric estimation of a smooth density with shape restrictions. *Statistica Sinica*, 681–701.

Meyer, M. C. and M. Woodroofe (2004). Consistent maximum likelihood estimation of a unimodal density using shape restrictions. *Canadian Journal of Statistics 32*(1), 85–100.

Minnotte, M. C. (1997). Nonparametric testing of the existence of modes. *The Annals of Statistics 25*(4), 1646–1660.

Minnotte, M. C. and D. W. Scott (1993). The mode tree: A tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics 2*(1), 51–68.

Müller, D. W. and G. Sawitzki (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association 86*(415), 738–746.

Nielsen, F. and R. Nock (2014). Optimal interval clustering: Application to Bregman clustering and statistical mixture learning. *IEEE Signal Processing Letters 21*(10), 1289–1292.

Nocedal, J. and S. J. Wright (2006). *Numerical Optimization*. Springer.

Pan, W. and X. Shen (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research 8*, 1145–1164.

Park, B., W. Kim, and M. Jones (2002). On local likelihood density estimation. *Annals of Statistics 30*(5), 1480–1495.

Park, H.-S. and C.-H. Jun (2009). A simple and fast algorithm for K-medoids clustering. *Expert systems with applications 36*(2), 3336–3341.

Pärna, K. (1986). Strong consistency of K-means clustering criterion in separable metric spaces. *Tartu Riikl. Ul. Toimetised 733*, 86–96.

Pärna, K. (1988). On the stability of K-means clustering in metric spaces. *Tartu Riikl. Ul. Toimetised 798*, 19–36.

Pärna, K. (1990). On the existence and weak convergence of K-centres in Banach spaces. *Tartu Ulikooli Toimetised 893*, 17–287.

Pärna, K. (1992). Clustering in metric spaces: some existence and continuity results for K-centers. In *Analyzing and Modeling Data and Knowledge*, pp. 85–91. Springer.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*(3), 1065–1076.

Patra, R. K. and B. Sen (2016). Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B 78*(4), 869–893.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A 186*, 343–414.

Penrose, M. D. (1995). Single linkage clustering and continuum percolation. *Journal of Multivariate Analysis 53*(1), 94–109.

Pollard, D. (1981). Strong consistency of K-means clustering. *The Annals of Statistics*, 135–140.

Pu, X. and E. Arias-Castro (2020). An EM algorithm for fitting a mixture model with symmetric log-concave densities. *Communications in Statistics-Theory and Methods 49*(1), 78–87.

Qiao, M. and J. Li (2015). Gaussian mixture models with component means constrained in pre-selected subspaces. arXiv preprint arXiv:1508.06388.

Rao, B. P. (1969). Estimation of a unimodal density. *Sankhyā: The Indian Journal of Statistics, Series A*, 23–36.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, 659–680.

Ridgeway, G. and J. M. MacDonald (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association 104*(486), 661–668.

Rinaldo, A., A. Singh, R. Nugent, and L. Wasserman (2012). Stability of density-based clustering. *Journal of Machine Learning Research 13*, 905.

Robin, A. C., C. Reylé, S. Derriere, and S. Picaud (2003). A synthetic view on structure and evolution of the Milky Way. *Astronomy & Astrophysics 409*(2), 523–540.

Rocci, R., S. A. Gattone, and R. Di Mari (2018). A data driven equivariant approach to constrained Gaussian mixture modeling. *Advances in Data Analysis and Classification 12*(2), 235–260.

Roquain, E. and N. Verzelen (2020). False discovery rate control with unknown null distribution: is it possible to mimic the oracle? *arXiv preprint arXiv:1912.03109*.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics 27*(3), 832 – 837.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics 9*(2), 65–78.

Rufibach, K. and G. Walther (2010). The block criterion for multiscale inference about a density,

with applications to other multiscale problems. *Journal of Computational and Graphical Statistics 19*(1), 175–190.

Samworth, R. J. (2018). Recent progress in log-concave density estimation. *Statistical Science 33*(4), 493–509.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and Methods 14*(5), 1123–1136.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Sheather, S. J. (2004). Density estimation. *Statistical Science*, 588–597.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B 53*(3), 683–690.

Shen, Z., M. Levine, and Z. Shang (2018). An MM algorithm for estimation of a two component semiparametric density mixture with a known component. *Electronic Journal of Statistics 12*(1), 1181–1209.

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika 27*(2), 125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika 27*(3), 219–246.

Shepard, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology 3*(2), 287–315.

Shi, N.-Z., S.-R. Zheng, and J. Guo (2005). The restricted EM algorithm under inequality restrictions on the parameters. *Journal of Multivariate Analysis 92*(1), 53–76.

Shin, D. W., C. G. Park, and T. Park (2001). Testing for one-sided group effects in repeated measures study. *Computational Statistics & Data Analysis 37*(2), 233–247.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological) 43*(1), 97–99.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. CRC Press.

Silverman, B. W. (2018). *Density Estimation for Statistics and Data Analysis*. Routledge.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw,

A. V. D'Amico, and J. P. Richie (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell 1*(2), 203–209.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics 12*(4), 1285–1297.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B 64*(3), 479–498.

Stuetzle, W. and R. Nugent (2010). A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics 19*(2), 397–418.

Sun, J. and M. Woodroofe (1996). Adaptive smoothing for a penalized npmle of a non-increasing density. *Journal of statistical planning and inference 52*(2), 143–159.

Szkaliczki, T. (2016). clustering.sc.dp: Optimal clustering with sequential constraint by using dynamic programming. *R Journal 8*(1), 318–327.

Takai, K. (2012). Constrained EM algorithm with projection method. *Computational Statistics 27*(4), 701–714.

Tan, M., G.-L. Tian, H.-B. Fang, and K. W. Ng (2007). A fast EM algorithm for quadratic optimization subject to convex constraints. *Statistica Sinica 17*(3), 945–964.

Tian, G.-L., K. W. Ng, and M. Tan (2008). EM-type algorithms for computing restricted MLEs in multivariate Normal distributions and multivariate t-distributions. *Computational Statistics & Data Analysis 52*(10), 4768–4778.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, and H. Mann (Eds.), *Contributions to Probability and Statistics*, pp. 448–485. Stanford University Press.

Van der Laan, M., K. Pollard, and J. Bryan (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation 73*(8), 575–584.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van't Wout, A. B., G. K. Lehrman, S. A. Mikheeva, G. C. O'Keeffe, M. G. Katze, R. E. Bumgarner, G. K. Geiss, and J. I. Mullins (2003). Cellular gene expression upon human immunodeficiency virus Type 1 infection of $CD4^+$-T-cell lines. *Journal of Virology 77*(2), 1392–1402.

Varadarajan, V. S. (1958). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960) 19*(1/2), 23–26.

Wagstaff, K., C. Cardie, S. Rogers, and S. Schrödl (2001). Constrained K-means clustering with background knowledge. In *International Conference on Machine Learning*, Volume 1, pp. 577–584.

Walker, M. G., M. Mateo, E. W. Olszewski, O. Y. Gnedin, X. Wang, B. Sen, and M. Woodroofe (2007). Velocity dispersion profiles of seven dwarf spheroidal galaxies. *The Astrophysical Journal Letters 667*(1), L53.

Walther, G. (2009). Inference and modeling with log-concave distributions. *Statistical Science 24*(3), 319–327.

Wang, H. and M. Song (2011). Ckmeans.1d.dp: Optimal K-means clustering in one dimension by dynamic programming. *The R Journal 3*(2), 29.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika 97*(4), 893–904.

Wang, S. and J. Zhu (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics 64*(2), 440–448.

Wang, T., Q. Li, D. J. Bucci, Y. Liang, B. Chen, and P. K. Varshney (2019). K-medoids clustering of data sequences with composite distributions. *IEEE Transactions on Signal Processing 67*(8), 2093–2106.

Wegman, E. J. (1970a). Maximum likelihood estimation of a unimodal density function. *The Annals of Mathematical Statistics 41*(2), 457–471.

Wegman, E. J. (1970b). Maximum likelihood estimation of a unimodal density, ii. *The Annals of Mathematical Statistics 41*(6), 2169–2174.

Wilson, R. B. (1963). *A Simplicial Algorithm for Concave Programming*. Ph. D. thesis, Harvard University, Graduate School of Business Administration.

Wolters, M. A. (2012). A greedy algorithm for unimodal kernel density estimation by data sharpening. *Journal of Statistical Software 47*(i06).

Wolters, M. A. and W. J. Braun (2018a). Enforcing shape constraints on a probability density estimate using an additive adjustment curve. *Communications in Statistics-Simulation and Computation 47*(3), 672–691.

Wolters, M. A. and W. J. Braun (2018b). A practical implementation of weighted kernel density estimation for handling shape constraints. *Stat 7*(1), e202.

Woodroofe, M. and J. Sun (1993). A penalized maximum likelihood estimate of $f(0+)$ when f

is non-increasing. *Statistica Sinica*, 501–515.

Zadegan, S. M. R., M. Mirzaie, and F. Sadoughi (2013). Ranked K-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-Based Systems 39*, 133–143.

Zheng, S., J. Guo, N.-Z. Shi, and G.-L. Tian (2012). Likelihood-based approaches for multivariate linear models under inequality constraints for incomplete data. *Journal of Statistical Planning and Inference 142*(11), 2926–2942.

Zheng, S., N. Shi, and J. Guo (2005). The restricted EM algorithm under linear inequalities in a linear model with missing data. *Science in China Series A: Mathematics 48*(6), 819–828.

Zhu, C., F. Wen, and J. Sun (2011). A rank-order distance based clustering algorithm for face tagging. In *CVPR 2011*, pp. 481–488. IEEE.