Operator norm convergence of spectral clustering on level sets

Bruno Pelletier, Pierre Pudlo

Operator norm convergence of spectral clustering on level sets

Bruno PELLETIER* and Pierre PUDLO†

Abstract

Following Hartigan [1975], a cluster is defined as a connected component of the t-level set of the underlying density, i.e., the set of points for which the density is greater than t. A clustering algorithm which combines a density estimate with spectral clustering techniques is proposed. Our algorithm is composed of two steps. First, a nonparametric density estimate is used to extract the data points for which the estimated density takes a value greater than t. Next, the extracted points are clustered based on the eigenvectors of a graph Laplacian matrix. Under mild assumptions, we prove the almost sure convergence in operator norm of the empirical graph Laplacian operator associated with the algorithm. Furthermore, we give the typical behavior of the representation of the dataset into the feature space, which establishes the strong consistency of our proposed algorithm.

Index Terms: Spectral clustering, graph, unsupervised classification, level sets, connected components.

*Department of Mathematics; IRMAR — UMR CNRS 6625; Université Rennes II; Place du Recteur Henri Le Moal, CS 24307; 35043 Rennes Cedex; France; bruno.pelletier@univ-rennes2.fr

†I3M: Institut de Mathématiques et de Modélisation de Montpellier — UMR CNRS 5149; Université Montpellier II, CC 051; Place Eugène Bataillon; 34095 Montpellier Cedex 5, France; pierre.pudlo@univ-montp2.fr

1 Introduction

The aim of data clustering, or unsupervised classification, is to partition a data set into several homogeneous groups relatively separated one from each other with respect to a certain distance or notion of similarity. There exists an extensive literature on clustering methods, and we refer the reader to Anderberg [1973], Hartigan [1975], McLachlan and Peel [2000], Chapter 10 in Duda et al. [2000], and Chapter 14 in Hastie et al. [2001] for general materials on the subject. In particular, popular clustering algorithms, such as Gaussian mixture models or k-means, have proved useful in a number of applications, yet they suffer from some internal and computational limitations. Indeed, the parametric assumption at the core of mixture models may be too stringent, while the standard k-means algorithm fails at identifying complex shaped, possibly non-convex, clusters.

The class of *spectral clustering* algorithms is presently emerging as a promising alternative, showing improved performance over classical clustering algorithms on several benchmark problems and applications; see e.g., Ng et al. [2002], von Luxburg [2007]. An overview of spectral clustering algorithms may be found in von Luxburg [2007], and connections with kernel methods are exposed in Fillipone et al. [2008]. The spectral clustering algorithm amounts at embedding the data into a feature space by using the eigenvectors of the similarity matrix in such a way that the clusters may be separated using simple rules, e.g. a separation by hyperplanes. The core component of the spectral clustering algorithm is therefore the similarity matrix, or certain normalizations of it, generally called graph Laplacian matrices; see Chung [1997]. Graph Laplacian matrices may be viewed as discrete versions of bounded operators between functional spaces. The study of these operators has started out recently with the works by Belkin et al. [2004], Belkin and Niyogi [2005], Coifman and Lafon [2006], Nadler et al. [2006], Koltchinskii [1998], Giné and Koltchinskii [2006], Hein et al. [2007], among others, and the convergence of the spectral clustering algorithm has been established in von Luxburg et al. [2008].

The standard k-means clustering leads to the optimal quantizer of the underlying distribution; see MacQueen [1967], Pollard [1981], Linder [2002]. However, determining what the limit clustering obtained in von Luxburg et al. [2008] represents for the distribution of the data remains largely an open question. As a

matter of fact, there exists many definitions of a cluster; see e.g., von Luxburg and Ben-David [2005] or García-Escudero et al. [2008]. Perhaps the most intuitive and precise definition of a cluster is the one introduced by Hartigan [1975]. Suppose that the data is drawn from a probability density f on $\mathbb{R}^d$ and let t be a positive number in the range of f. Then a cluster in the sense of Hartigan [1975] is a connected component of the t-level set

$$\mathscr{L}(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}.$$

This definition has several advantages. First, it is geometrically simple. Second, it offers the possibility of filtering out possibly meaningless clusters by keeping only the observations falling in a region of high density, This proves useful, for instance, in the situation where the data exhibits a cluster structure but is contaminated by a uniform background noise, as illustrated in our simulations in Section 4.

In this context, the level t should be considered as a resolution level for the data analysis. Several clustering algorithms have been introduced building upon Hartigan's definition. In Cuevas et al. [2000, 2001], clustering is performed by estimating the connected components of $\mathscr{L}(t)$; see also the work by Azzalini and Torelli [2007]. Hartigan's definition is also used in Biau et al. [2007] to define an estimate of the number of clusters.

In the present paper, the definition of a cluster given by Hartigan [1975] is adopted, and we introduce a spectral clustering algorithm on estimated level sets. More precisely, given a random sample $X_1, \ldots, X_n$ drawn from a density f on $\mathbb{R}^d$, our proposed algorithm is composed of two operations. In the first step, given a positive number t, we extract the observations for which $\hat{f}_n(X_i) \geq t$, where $\hat{f}_n$ is a nonparametric density estimate of f based on the sample $X_1, \ldots, X_n$. In the second step, we perform a spectral clustering of the extracted points. The remaining data points are then left unlabeled.

Our proposal is to study the asymptotic behavior of this algorithm. As mentioned above, strong interest has recently been shown in spectral clustering algorithms, and the major contribution to the proof of the convergence of spectral clustering is certainly due to von Luxburg et al. [2008]. In von Luxburg et al. [2008], the graph Laplacian matrix is associated with some random operator acting on the Banach space of continuous functions. They prove the collectively compact convergence of those operators towards a limit operator. Under mild assumptions,

we strengthen their results by establishing the almost sure convergence in *operator norm*, but in a smaller Banach space (Theorem 3.1). This operator norm convergence is more amenable than the slightly weaker notion of convergence established in von Luxburg et al. [2008]. For instance, it is easy to check that the limit operator, and the graph Laplacian matrices used in the algorithm, are continuous in the scale parameter h.

We also derive the asymptotic representation of the dataset in the feature space in Corollary 3.2. This result implies that the proposed algorithm is strongly consistent and that, asymptotically, observations of $\mathscr{L}(t)$ are assigned to the same cluster if and only if they fall in the same connected component of the level set $\mathscr{L}(t)$.

The paper is organized as follows. In Section 2, we introduce some notations and assumptions, as well as our proposed algorithm. Section 3 contains our main results, namely the convergence in operator norm of the random operators, and the characterization of the dataset embedded into the feature space. We provide a numerical example with a simulated dataset in Section 4. Sections 5 and 6 are devoted to the proofs. At the end of the paper, a technical result on the geometry of level sets is stated in Appendix A, some useful results of functional analysis are summarized in Appendix B, and the theoretical properties of the limit operator are given in Appendix C.

2 Spectral clustering algorithm

2.1 Mathematical setting and assumptions

Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, with common probability measure μ. Suppose that μ admits a density f with respect to the Lebesgue measure on $\mathbb{R}^d$. The *t-level set* of f is denoted by $\mathscr{L}(t)$, i.e.,

$$\mathscr{L}(t) = \big\{x \in \mathbb{R}^d \,:\, f(x) \geq t\big\},$$

for all positive level t, and given $a \leq b$, $\mathscr{L}_a^b$ denotes the set $\{x \in \mathbb{R}^d \,:\, a \leq f(x) \leq b\}$. The differentiation operator with respect to x is denoted by D_x. We assume that f satisfies the following conditions.

Assumption 1. *(i)* f is of class $\mathscr{C}^2$ on $\mathbb{R}^d$; *(ii)* $\|D_x f\| > 0$ on the set $\{x \in \mathbb{R}^d : f(x) = t\}$; *(iii)* f, $D_x f$, and $D_x^2 f$ are uniformly bounded on $\mathbb{R}^d$.

Note that under Assumption 1, $\mathscr{L}(t)$ is compact whenever t belongs to the interior of the range of f. Moreover, $\mathscr{L}(t)$ has a finite number ℓ of connected components $\mathscr{C}_j$, $j = 1, \ldots, \ell$. For ease of notation, the dependence of $\mathscr{C}_j$ on t is omitted. The minimal distance between the connected components of $\mathscr{L}(t)$ is denoted by d_{min}, i.e.,

$$d_{min} = \inf_{i \neq j} \operatorname{dist}\big(\mathscr{C}_i, \mathscr{C}_j\big). \tag{2.1}$$

Let $\widehat{f}_n$ be a consistent density estimate of f based on the random sample $X_1, \ldots, X_n$. The t-level set of $\widehat{f}_n$ is denoted by $\mathscr{L}_n(t)$, i.e.,

$$\mathscr{L}_n(t) = \big\{x \in \mathbb{R}^d : \widehat{f}_n(x) \geq t\big\}.$$

Let $J(n)$ be the set of integers defined by

$$J(n) = \big\{j \in \{1, \ldots, n\} : \widehat{f}_n(X_j) \geq t\big\}.$$

The cardinality of $J(n)$ is denoted by $j(n)$.

Let $k : \mathbb{R}^d \to \mathbb{R}_+$ be a fixed function. The unit ball of $\mathbb{R}^d$ centered at the origin is denoted by B, and the ball centered at $x \in \mathbb{R}^d$ and of radius r is denoted by $x + rB$. We assume throughout that the function k satisfies the following set of conditions.

Assumption 2. *(i)* k is of class $\mathscr{C}^2$ on $\mathbb{R}^d$; *(ii)* the support of k is B; *(iii)* k is uniformly bounded from below on $B/2$ by some positive number; and *(iv)* $k(-x) = k(x)$ for all $x \in \mathbb{R}^d$.

Let h be a positive number. We denote by $k_h : \mathbb{R}^d \to \mathbb{R}_+$ the map defined by $k_h(u) = k(u/h)$.

2.2 Algorithm

The first ingredient of our algorithm is the *similarity matrix* $\mathbf{K}_{n,h}$ whose elements are given by

$$\mathbf{K}_{n,h}(i, j) = k_h(X_j - X_i),$$

and where the integers i and j range over the random set $J(n)$. Hence $\mathbf{K}_{n,h}$ is a random matrix indexed by $J(n) \times J(n)$, whose values depend on the function k_h, and on the observations X_j lying in the estimated level set $\mathscr{L}_n(t)$. Next, we introduce the diagonal *normalization matrix* $\mathbf{D}_{n,h}$ whose diagonal entries are given by

$$\mathbf{D}_{n,h}(i,i) = \sum_{j \in J(n)} \mathbf{K}_{n,h}(i,j), \quad i \in J(n).$$

Note that the diagonal elements of $\mathbf{D}_{n,h}$ are positive.

The spectral clustering algorithm is based on the matrix $\mathbf{Q}_{n,h}$ defined by

$$\mathbf{Q}_{n,h} = \mathbf{D}_{n,h}^{-1} \mathbf{K}_{n,h}.$$

Observe that $\mathbf{Q}_{n,h}$ is a random Markovian transition matrix. Note also that the (random) eigenvalues of $\mathbf{Q}_{n,h}$ are real numbers and that $\mathbf{Q}_{n,h}$ is diagonalizable. Indeed the matrix $\mathbf{Q}_{n,h}$ is conjugate to the symmetric matrix $\mathbf{S}_{n,h} := \mathbf{D}_{n,h}^{-1/2} \mathbf{K}_{n,h} \mathbf{D}_{n,h}^{-1/2}$ since we may write

$$\mathbf{Q}_{n,h} = \mathbf{D}_{n,h}^{-1/2} \mathbf{S}_{n,h} \mathbf{D}_{n,h}^{1/2}.$$

Moreover, the inequality $\|\mathbf{Q}_{n,h}\|_\infty \leq 1$ implies that the spectrum $\sigma(\mathbf{Q}_{n,h})$ is a subset of $[-1;+1]$. Let $1 = \lambda_{n,1} \geq \lambda_{n,2} \geq \ldots \geq \lambda_{n,j(n)} \geq -1$ be the eigenvalues of $\mathbf{Q}_{n,h}$, where in this enumeration, an eigenvalue is repeated as many times as its multiplicity.

To implement the spectral clustering algorithm, the data points of the partitioning problem are first embedded into $\mathbb{R}^\ell$ by using the eigenvectors of $\mathbf{Q}_{n,h}$ associated with the ℓ largest eigenvalues, namely $\lambda_{n,1}$, $\lambda_{n,2}$, $\ldots \lambda_{n,\ell}$. More precisely, fix a collection $V_{n,1}$, $V_{n,2}$, $\ldots$, $V_{n,\ell}$ of such eigenvectors with components respectively given by $V_{n,k} = \{V_{n,k,j}\}_{j \in J(n)}$, for $k = 1, \ldots, \ell$. Then the j^{th} data point, for j in $J(n)$, is represented by the vector $\rho_n(X_j)$ of $\mathbb{R}^\ell$ defined by $\rho_n(X_j) := \{V_{n,k,j}\}_{1 \leq k \leq \ell}$. At last, the embedded points are partitioned using a classical clustering method, such as the k-means algorithm for instance.

2.3 Functional operators associated with the matrices of the algorithm

As exposed in the Introduction, some functional operators are associated with the matrices acting on $\mathbb{C}^{J(n)}$ defined in the previous paragraph. The link between

matrices and functional operators is provided by the evaluation map defined in (2.3) below. As a consequence, asymptotic results on the clustering algorithm may be derived by studying first the limit behavior of these operators.

To this aim, let us first introduce some additional notation. For $\mathscr{D}$ a subset of $\mathbb{R}^d$, let $W(\mathscr{D})$ be the Banach space of complex-valued, bounded, and continuously differentiable functions with bounded gradient, endowed with the norm

$$\|g\|_W = \|g\|_\infty + \|D_x g\|_\infty.$$

Consider the non-oriented graph whose vertices are the X_j's for j ranging in $J(n)$. The similarity matrix $\mathbf{K}_{n,h}$ gives random weights to the edges of the graph and the random transition matrix $\mathbf{Q}_{n,h}$ defines a random walk on the vertices of a random graph. Associated with this random walk is the transition operator $Q_{n,h}$: $W\big(\mathscr{L}_n(t)\big) \to W\big(\mathscr{L}_n(t)\big)$ defined for any function g by

$$Q_{n,h}g(x) = \int_{\mathscr{L}_n(t)} q_{n,h}(x,y)g(y)\mathbb{P}_n^t(dy).$$

In this equation, $\mathbb{P}_n^t$ is the discrete random probability measure given by

$$\mathbb{P}_n^t = \frac{1}{j(n)} \sum_{j\in J(n)} \delta_{X_j},$$

and

$$q_{n,h}(x,y) = \frac{k_h(y-x)}{K_{n,h}(x)}, \quad \text{where } K_{n,h}(x) = \int_{\mathscr{L}_n(t)} k_h(y-x)\mathbb{P}_n^t(dy). \tag{2.2}$$

In the definition of $q_{n,h}$, we use the convention that $0/0 = 0$, but this situation does not occur in the proofs of our results.

Given the *evaluation map* $\pi_n : W\big(\mathscr{L}_n(t)\big) \to \mathbb{C}^{J(n)}$ defined by

$$\pi_n(g) = \Big\{g(X_j) \,:\, j \in J(n)\Big\}, \tag{2.3}$$

the matrix $\mathbf{Q}_{n,h}$ and the operator $Q_{n,h}$ are related by $\mathbf{Q}_{n,h} \circ \pi_n = \pi_n \circ Q_{n,h}$. Using this relation, asymptotic properties of the spectral clustering algorithm may be deduced from the limit behavior of the sequence of operators $\{Q_{n,h}\}_n$. The difficulty, though, is that $Q_{n,h}$ acts on $W\big(\mathscr{L}_n(t)\big)$ and $\mathscr{L}_n(t)$ is a random set which

varies with the sample. For this reason, we introduce a sequence of operators $\widehat{Q}_{n,h}$ acting on $W\big(\mathscr{L}(t)\big)$ and constructed from $Q_{n,h}$ as follows.

First of all, recall that under Assumption 1, the gradient of f does not vanish on the set $\{x \in \mathbb{R}^d : f(x) = t\}$. Since f is of class $\mathscr{C}^2$, a continuity argument implies that there exists $\varepsilon_0 > 0$ such that $\mathscr{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$ contains no critical points of f. Under this condition, Lemma A.1 states that $\mathscr{L}(t+\varepsilon)$ is diffeomorphic to $\mathscr{L}(t)$ for every ε such that $|\varepsilon| \leq \varepsilon_0$. In all of the following, it is assumed that ε_0 is small enough so that

$$\varepsilon_0/\alpha(\varepsilon_0) < h/2, \quad \text{where } \alpha(\varepsilon_0) = \inf\big\{\|D_x f(x)\|; x \in \mathscr{L}_{t-\varepsilon_0}^{t}\big\}. \tag{2.4}$$

Let $\{\varepsilon_n\}_n$ be a sequence of positive numbers such that $\varepsilon_n \leq \varepsilon_0$ for each n, and $\varepsilon_n \to 0$ as $n \to \infty$. In Lemma A.1 an explicit diffeomorphism φ_n carrying $\mathscr{L}(t)$ to $\mathscr{L}(t-\varepsilon_n)$ is constructed, i.e.,

$$\varphi_n : \mathscr{L}(t) \xrightarrow{\cong} \mathscr{L}(t-\varepsilon_n). \tag{2.5}$$

The diffeomorphism φ_n induces the linear operator $\Phi_n : W\big(\mathscr{L}(t)\big) \to W\big(\mathscr{L}(t-\varepsilon_n)\big)$ defined by $\Phi_n g = g \circ \varphi_n^{-1}$.

Second, let Ω_n be the probability event defined by

$$\Omega_n = \Big[\|\widehat{f}_n - f\|_\infty \leq \varepsilon_n\Big] \cap \Big[\inf\Big\{\|D_x \widehat{f}_n(x)\|, x \in \mathscr{L}_{t-\varepsilon_0}^{t+\varepsilon_0}\Big\} \geq \frac{1}{2}\|D_x f\|_\infty\Big]. \tag{2.6}$$

Note that on the event Ω_n, the following inclusions hold:

$$\mathscr{L}(t-\varepsilon_n) \subset \mathscr{L}_n(t) \subset \mathscr{L}(t+\varepsilon_n). \tag{2.7}$$

We assume that the indicator function $\mathbf{1}_{\Omega_n}$ tends to 1 almost surely as $n \to \infty$, which is satisfied by common density estimates $\widehat{f}_n$ under mild assumptions. For instance, consider a kernel density estimate with a Gaussian kernel. Then for a density f satisfying the conditions in Assumption 1, we have $\|D_x^{(p)} \widehat{f}_n - D_x^{(p)} f\|_\infty \to 0$ almost surely as $n \to \infty$, for $p = 0$ and $p = 1$ (see e.g., Prakasa Rao [1983]), which implies that $\mathbf{1}_{\Omega_n} \to 1$ almost surely as $n \to \infty$.

We are now in a position to introduce the operator $\widehat{Q}_{n,h} : W\big(\mathscr{L}(t)\big) \to W\big(\mathscr{L}(t)\big)$ defined on the event Ω_n by

$$\widehat{Q}_{n,h} = \Phi_n^{-1} Q_{n,h} \Phi_n, \tag{2.8}$$

and we extend the definition of $\widehat{Q}_{n,h}$ to the whole probability space by setting it to the null operator on the complement Ω_n^c of Ω_n. In other words, on Ω_n^c, the function $\widehat{Q}_{n,h}g$ is identically zero for each $g \in W\big(\mathscr{L}(t)\big)$.

Remark 2.1. Albeit the relevant part of $\widehat{Q}_{n,h}$ is defined on Ω_n for technical reasons, this does not bring any difficulty as long as one is concerned with almost sure convergence. To see this, let $(\Omega, \mathscr{A}, P)$ be the probability space on which the X_i's are defined. Denote by Ω_∞ the event on which $\mathbf{1}_{\Omega_n}$ tends to 1, and recall that $P(\Omega_\infty) = 1$ by assumption. Thus, for every $\omega \in \Omega$, there exists a random integer $n_0(\omega)$ such that, for each $n \geq n_0(\omega)$, ω lies in Ω_n. Besides $n_0(\omega)$ is finite on Ω_∞. Hence in particular, if $\{Z_n\}$ is a sequence of random variables such that $Z_n \mathbf{1}_{\Omega_n}$ converges almost surely to some random variable Z_∞, then $Z_n \to Z_\infty$ almost surely.

3 Main results

Our main result (Theorem 3.1) states that $\widehat{Q}_{n,h}$ converges in operator norm to the limit operator $Q_h : W\big(\mathscr{L}(t)\big) \to W\big(\mathscr{L}(t)\big)$ defined by

$$Q_h g(x) = \int_{\mathscr{L}(t)} q_h(x,y) g(y) \mu^t(dy), \tag{3.1}$$

where μ^t denotes the conditional distribution of X given the event $\big[X \in \mathscr{L}(t)\big]$, and where

$$q_h(x,y) = \frac{k_h(y-x)}{K_h(x)}, \quad \text{with } K_h(x) = \int_{\mathscr{L}(t)} k_h(y-x) \mu^t(dy). \tag{3.2}$$

Theorem 3.1 (Operator Norm Convergence)**.** *Suppose that Assumptions 1 and 2 hold. We have*

$$\big\| \widehat{Q}_{n,h} - Q_h \big\|_W \to 0 \quad \textit{almost surely as } n \to \infty.$$

The proof of Theorem 3.1 is given in Paragraph 5.2. Its main arguments are as follows. First, the three classes of functions defined in Lemma 5.2 are shown to be Glivenko-Cantelli. This, together with additional technical results, leads to uniform convergences of some linear operators (Lemma 5.6).

Theorem 3.1 implies the consistency of our algorithm. We recall that d_{min} given in (2.1) is the minimal distance between the connected components of the level set. The starting point is the fact that, provided that $h < d_{\min}$, the connected components of the level set $\mathscr{L}(t)$ are the recurrent classes of the Markov chain whose transitions are given by Q_h. Indeed, this process cannot jump from one component to the other ones. Hence, Q_h defines the desired clustering via its eigenspace corresponding to the eigenvalue 1.

As stated in Proposition C.2 in the Appendices, the eigenspace of the limit operator Q_h associated with the eigenvalue 1 is spanned by the indicator functions of the connected components of $\mathscr{L}(t)$. Hence the representation of the extracted part of the dataset into the feature space $\mathbb{R}^\ell$ (see the end of Paragraph 2.2) tends to concentrate around ℓ different centroids. Moreover, each of these centroids corresponds to a cluster, i.e., to a connected component of $\mathscr{L}(t)$.

More precisely, using the convergence in operator norm of $\widehat{Q}_{n,h}$ towards Q_h, together with the results of functional analysis given in Appendix B, we obtain the following corollary which describes the asymptotic behavior of our algorithm. Let us denote by $J(\infty)$ the set of integers j such that X_j is in the level set $\mathscr{L}(t)$. For all $j \in J(\infty)$, define $k(j)$ as the integer such that $X_j \in \mathscr{C}_{k(j)}$.

Corollary 3.2. *Suppose that Assumptions 1 and 2 hold, and that h is in* $(0; d_{min})$. *There exists a sequence* $\{\xi_n\}_n$ *of linear transformations of* $\mathbb{R}^\ell$ *such that, for all* $j \in J(\infty)$, $\xi_n \rho_n(X_j)$ *converges almost surely to* $e_{k(j)}$, *where* $e_{k(j)}$ *is the vector of* $\mathbb{R}^\ell$ *whose components are all 0 except the* $k(j)^{\text{th}}$ *component equal to* 1.

Corollary 3.2, which is new up to our knowledge, is proved in Section 6. Corollary 3.2 states that the data points embedded in the feature space concentrate on separated centroids. As a consequence, any partitioning algorithm (e.g., *k*-means) applied in the feature space will asymptotically yield the desired clustering. In other words, the clustering algorithm is consistent. Note that if one is only interested in the consistency property, then this result could be obtained through another route. Indeed, it is shown in Biau et al. [2007] that the neighborhood graph with connectivity radius *h* has asymptotically the same number of connected components as the level set. Hence, splitting the graph into its connected components leads to the desired clustering as well. But Corollary 3.2, by giving the asymptotic representation of the data when embedded in the feature space $\mathbb{R}^\ell$, provides additional insight into spectral clustering algorithms. In particular, Corollary 3.2 provides a rationale for the heuristic of Zelnik-Manor and Perona [2004] for automatic selection of the number of groups. Their idea is to quantify the amount

of concentration of the points embedded in the feature space, and to select the number of groups leading to the maximal concentration. Their method compared favorably with the eigengap heuristic considered in von Luxburg [2007].

Naturally, the selection of the number of groups is also linked with the choice of the parameter h. In this direction, let us emphasize that the operators $\widehat{Q}_{n,h}$ and Q_h depend continuously on the scale parameter h. Thus, the spectral properties of both operators will be close to the ones stated in Corollary 3.2, if h is in the neighborhood of the interval $(0; d_{min})$. This follows from the continuity of an isolated set of eigenvalues, as stated in Appendix B. In particular, the sum of the eigenspaces of Q_h associated with the eigenvalues close to 1 is spanned by functions that are close to (in $W(\mathscr{L}(t))$-norm) the indicator functions of the connected components of $\mathscr{L}(t)$. Hence, the representation of the dataset in the feature space $\mathbb{R}^\ell$ still concentrates on some neighborhoods of e_k, $1 \leq k \leq \ell$ and a simple clustering algorithm such as k-means will still give the desired result. To sum up the above, if assumptions 1 and 2 hold, our algorithm is consistent for all h in $(0, h_{max})$ for some $h_{max} > d_{min}$.

Several questions, though, remain largely open. For instance, one might ask if a similar result holds for the classical spectral clustering algorithm, i.e., without the preprocessing step. This case corresponds to taking $t = 0$. One possibility may then be to consider a sequence h_n, with $\lim h_n = 0$ and to the study the limit of the operator Q_{n,h_n}.

4 Simulations

We consider a mixture density on $\mathbb{R}^2$ with four components corresponding to random variables $X_1, \ldots, X_4$ where

(i) $X_1 \sim \mathscr{N}(0, \sigma_1^2 \mathbf{I})$ with $\sigma_1 = 0.2$;

(ii) $X_2 = R_2(\cos\theta_2, \sin\theta_2)$ where $\theta_2 \sim \mathscr{U}([0; 2\pi])$ and $R_2 \sim \mathscr{N}(1, 0.1^2)$;

(iii) $X_3 = R_3(\cos\theta_3, \sin\theta_3)$ where $\theta_3 \sim \mathscr{U}([0; 2\pi])$ and $R_3 \sim \mathscr{N}(2, 0.2^2)$;

(iv) $X_4 \sim \mathscr{U}([-3; 3] \times [-3; 3])$.

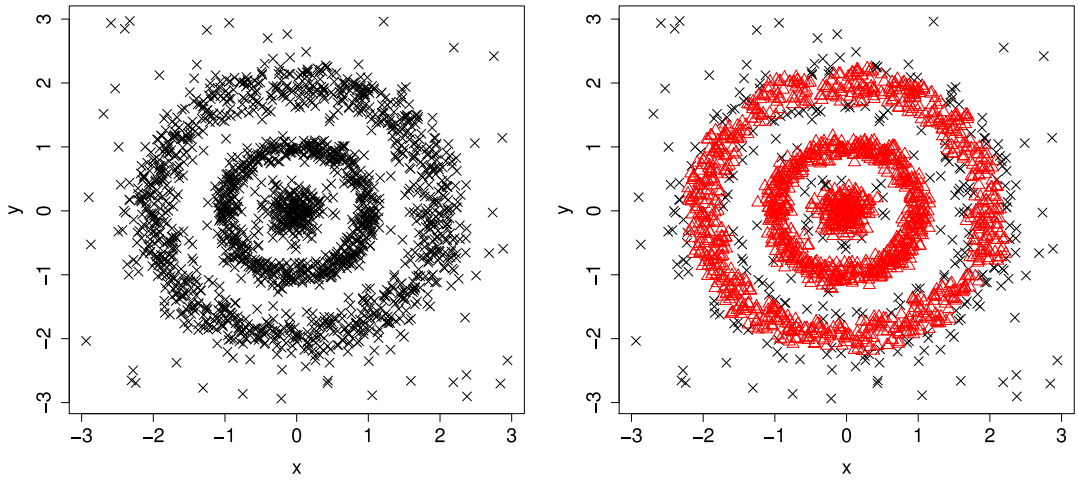

Figure 1: *Left*: simulated points. *Right*: Points belonging to the estimated level set (red triangle) and remaining points (dark cross).

The proportions of the components in the mixture are taken as 10%, 32%, 53% and 5%, respectively. The fourth component (X_4) represents a uniform background noise.

A random sample of size $n = 1,900$ has been simulated according to the mixture. Points are displayed in Figure 1 (left). A nonparametric kernel density estimate, with a Gaussian kernel, has been adjusted to the data. The bandwidth parameter of the density estimate has been selected automatically with cross-validation. A level $t = 0.0444$ has been selected such that 85% of the simulated points are extracted, i.e., 85% of the observations fall in $\mathscr{L}_n(t)$. The extracted and discarded points are displayed in Figure 1 (right). The number of extracted points is equal to $1,615$.

The spectral clustering has been applied to the $1,615$ extracted points, with the similarity function

$$k(x) = \exp(-1/(1-\|x\|)^2)\mathbf{1}\{\|x\| < 1\}.$$

For numerical stability of the algorithm, we considered the eigendecomposition of the symmetric matrix $\mathbf{I} - \mathbf{S}_{n,h}$. Thus, the eigenspace associated with the eigenvalue 1 of the matrix $\mathbf{Q}_{n,h}$ corresponds to the null space of $\mathbf{I} - \mathbf{S}_{n,h}$. The scale

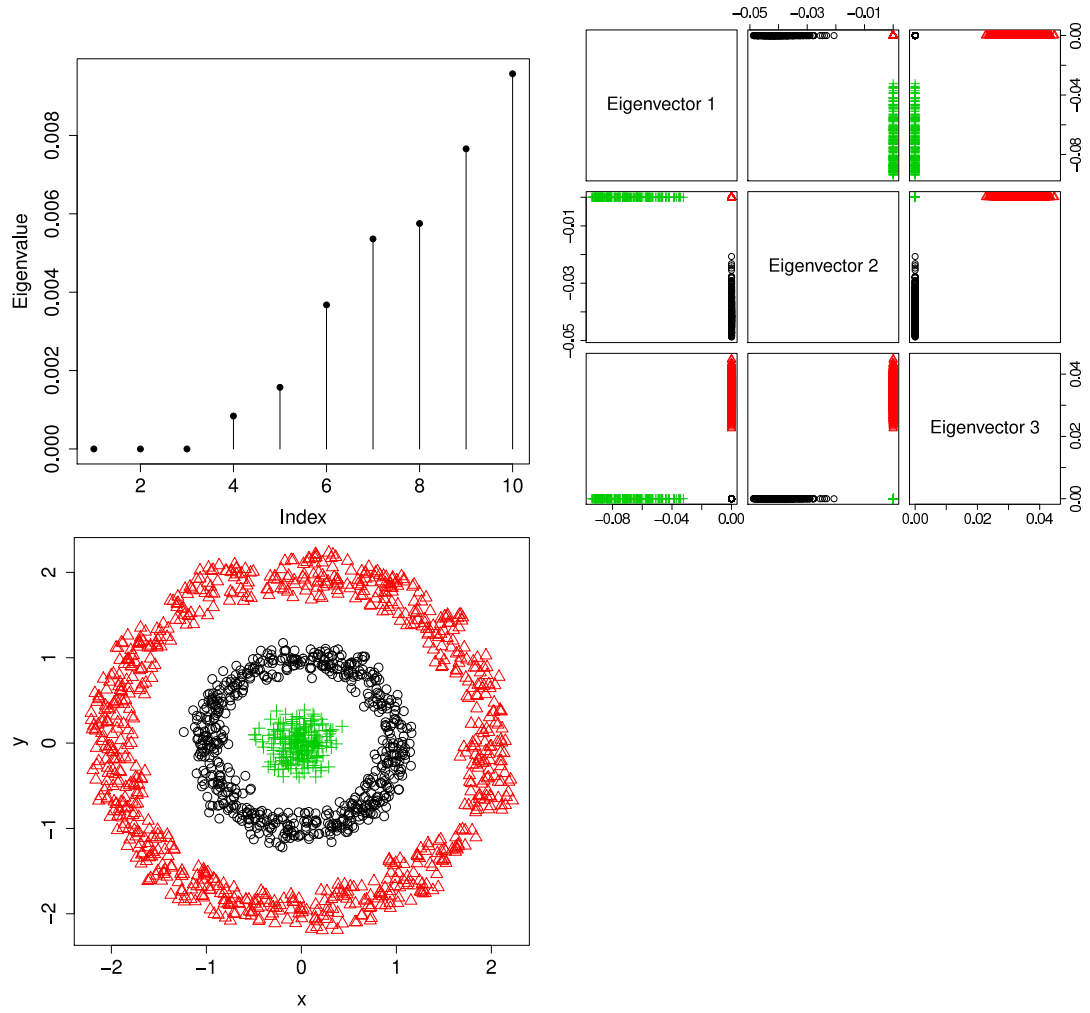

Figure 2: *Top Left*: first 10 eigenvalues, sorted in ascending order. *Top Right*: pairs plots of the first three eigenvectors. It may be seen that the embedded data concentrate around three distinct points in the feature space $\mathbb{R}^3$. *Bottom* Resulting partition obtained by applying a k-means algorithm in the feature space. The color scheme is identical to the representation of the eigenvectors (top-right panel). The three groups are accurately recovered.

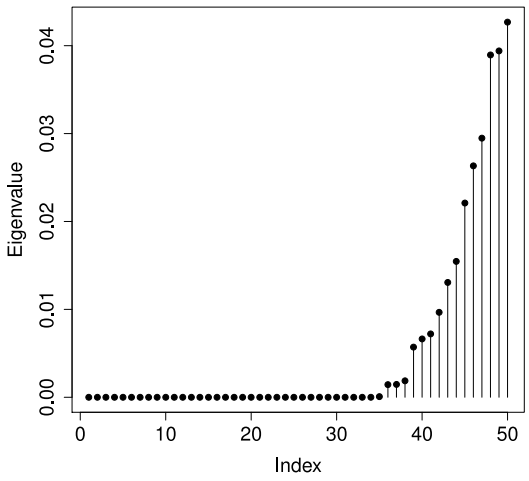

Figure 3: First 50 eigenvalues of the standard spectral clustering algorithm, applied on the initial data set, i.e., without level set pre-processing. A total of 35 eigenvalues are found equal to zero, which leads to 35 inhomogeneous groups, indicating failure of the standard spectral clustering algorithm.

parameter h has be empirically chosen equal to 0.25. The first 10 eigenvalues of $\mathbf{I} - \mathbf{S}_{n,h}$ are represented in Figure 2 (top-left). Three eigenvalues are found equal to zero, indicating three distinct groups. The data is then embedded in $\mathbb{R}^3$ using the three eigenvectors of the null space of $\mathbf{I} - \mathbf{S}_{n,h}$, and the data is partitioned in this space using a k-means clustering algorithm. Pair plots of three eigenvectors of the null space are displayed in Figure 2. It may be observed that the embedded data are concentrated around three distinct points in the feature space. Applying a k-means algorithm in the feature space leads to the partition represented in Figure 2. Note that observations considered as background noise are the discarded points belonging to the complement of $\mathscr{L}_n(t)$. In this example, our algorithm is successful at recovering the three expected groups.

As a comparison, we applied the standard spectral clustering algorithm to the initial data set of size $n = 1,900$. In this case, 35 eigenvalues are found equal to zero (Figure 3). Applying a k-means clustering algorithm in the embedding space $\mathbb{R}^{35}$ leads to 35 inhomogeneous groups (not displayed here), none of which corresponds roughly to the expected groups (the two circular bands and the inner circle). This failure of the standard spectral clustering algorithm is explained by the

presence of the background noise which, when unfiltered, perturbs the formation of distinct groups. While there remains multiple important questions, in particular regarding the choice of the parameter h, these simulations illustrate the added value of combining a spectral clustering algorithm with level-set techniques.

5 Proof of the convergence of $\widehat{Q}_{n,h}$ (Theorem 3.1)

5.1 Preliminaries

Let us start with the following simple lemma.

Lemma 5.1. *Let $\{A_n\}_{n\geq 0}$ be a decreasing sequence of Borel sets in $\mathbb{R}^d$, with limit $A_\infty = \cap_{n\geq 0} A_n$. If $\mu(A_\infty) = 0$, then*

$$\mathbb{P}_n A_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \in A_n\} \to 0 \quad \textit{almost surely as } n \to \infty,$$

where $\mathbb{P}_n$ is the empirical measure associated with the random sample $X_1, \ldots, X_n$.

Proof. First, note that $\lim_n \mu(A_n) = \mu(A_\infty)$. Next, fix an integer k. For all $n \geq k$, $A_n \subset A_k$ and so $\mathbb{P}_n A_n \leq \mathbb{P}_n A_k$. But $\lim_n \mathbb{P}_n A_k = \mu(A_k)$ almost surely by the law of large numbers. Consequently $\limsup_n \mathbb{P}_n A_n \leq \mu(A_k)$ almost surely. Letting $k \to \infty$ yields

$$\limsup_n \mathbb{P}_n A_n \leq \mu(A_\infty) = 0,$$

which concludes the proof since $\mathbb{P}_n A_n \geq 0$. □

The operator norm convergence that we expect to prove is a uniform law of large number. The key argument is the fact that the classes of functions of the following lemma are Glivenko-Cantelli. Let g be a function defined on some subset $\mathscr{D}$ of $\mathbb{R}^d$, and let $\mathscr{A}$ be a subset of $\mathscr{D}$. In what follows, for all $x \in \mathbb{R}^d$, the notation $g(x)\mathbf{1}_{\mathscr{A}}(x)$ stands for $g(x)$ if $x \in \mathscr{A}$ and 0 otherwise.

Lemma 5.2. *1. The two collections of functions*

$$\mathscr{F}_1 := \big\{ y \mapsto k_h(y-x)\mathbf{1}_{\mathscr{L}(t)}(y) : x \in \mathscr{L}(t-\varepsilon_0) \big\},$$
$$\mathscr{F}_2 := \big\{ y \mapsto D_x k_h(y-x)\mathbf{1}_{\mathscr{L}(t)}(y) : x \in \mathscr{L}(t-\varepsilon_0) \big\},$$

are Glivenko-Cantelli, where $D_x k_h$ denotes the differential of k_h.

2. Let $r : \mathscr{L}(t) \times \mathbb{R}^d$ be a continuously differentiable function such that
(i) there exists a compact $\mathscr{K} \subset \mathbb{R}^d$ such that $r(x,y) = 0$ for all $(x,y) \in \mathscr{L}(t) \times K^c$;
(ii) r is uniformly bounded on $\mathscr{L}(t) \times \mathbb{R}^d$, i.e. $\|r\|_\infty < \infty$.
Then the collection of functions

$$\mathscr{F}_3 := \left\{ y \mapsto r(x,y) g(y) \mathbf{1}_{\mathscr{L}(t)}(y) \, : \, x \in \mathscr{L}(t), \, \|g\|_{W(\mathscr{L}(t))} \leq 1 \right\}$$

is Glivenko-Cantelli.

Proof. 1. Clearly $\mathscr{F}_1$ has an integrable envelope since k_h is uniformly bounded. Moreover, for each fixed y, the map $x \mapsto k_h(y-x)\mathbf{1}_{\mathscr{L}(t)}(y)$ is continuous, and $\mathscr{L}(t-\varepsilon_0)$ is compact. Hence for each $\delta > 0$, using a finite covering of $\mathscr{L}(t-\varepsilon_0)$, it is easy to construct finitely many L_1 brackets of size at most δ whose union cover $\mathscr{F}_1$; see e.g., Example 19.8 in van der Vaart [1998]. So $\mathscr{F}_1$ is Glivenko-Cantelli. Since k_h is continuously differentiable and with compact support, the same arguments apply to each component of $D_x k_h$, and so $\mathscr{F}_2$ is also a Glivenko-Cantelli class.

2. Set $\mathscr{R} = \{ y \mapsto r(x,y) \, : \, x \in \mathscr{L}(t) \}$. First, since r is continuous on the compact set $\mathscr{L}(t) \times \mathscr{K}$, it is uniformly continuous. So a finite covering of $\mathscr{R}$ of arbitrary size in the supremum norm may be obtained from a finite covering of $\mathscr{L}(t) \times \mathscr{K}$. Hence $\mathscr{R}$ has finite entropy in the supremum norm. Second, set $\mathscr{G} = \{ y \mapsto g(y)\mathbf{1}_{\mathscr{L}(t)}(y) : \|g\|_{W(\mathscr{L}(t))} \leq 1 \}$. Denote by $\mathscr{X}$ the convex hull of $\mathscr{L}(t)$, and consider the collection of functions $\tilde{\mathscr{G}} = \{ \tilde{g} : \mathscr{X} \to \mathbb{R} : \|\tilde{g}\|_{W(\mathscr{X})} \leq 1 \}$. Then $\tilde{\mathscr{G}}$ has finite entropy in the supremum norm; see Kolmogorov and Tikhomirov [1961] and van der Vaart [1994]. Using the surjection $\tilde{\mathscr{G}} \to \mathscr{G}$ carrying $\tilde{g}$ to $\left(\tilde{g}\mathbf{1}_{\mathscr{L}(t)}\right)$, that $\mathscr{G}$ has finite entropy in the supremum norm readily follows. To conclude the proof, since both $\mathscr{R}$ and $\mathscr{G}$ are uniformly bounded, a finite covering of $\mathscr{F}_3$ of arbitrary size δ in the supremum norm may be obtained from finite coverings of $\mathscr{R}$ and $\mathscr{G}$, which yields a finite covering of $\mathscr{F}_3$ by L_1 brackets of size at most 2δ. So $\mathscr{F}_3$ is a Glivenko-Cantelli class. □

We recall that the limit operator Q_h is given by (3.1). The following lemma gives useful bounds on K_h and q_h, both defined in (3.2).

Lemma 5.3. *1. The function K_h is uniformly bounded from below by some positive number on $\mathscr{L}(t-\varepsilon_0)$, i.e.,* $\inf\{K_h(x) : x \in \mathscr{L}(t-\varepsilon_0)\} > 0$*;*

2. The kernel q_h is uniformly bounded, i.e., $\|q_h\|_\infty < \infty$;
3. The differential of q_h with respect to x is uniformly bounded on $\mathscr{L}(t-\varepsilon_0)\times\mathbb{R}^d$, i.e., $\sup\{\|D_x q_h(x,y)\| : (x,y) \in \mathscr{L}(t-\varepsilon_0)\times\mathbb{R}^d\} < \infty$;
4. The Hessian of q_h with respect to x is uniformly bounded on $\mathscr{L}(t-\varepsilon_0)\times\mathbb{R}^d$, i.e., $\sup\{\|D_x^2 q_h(x,y)\| : (x,y) \in \mathscr{L}(t-\varepsilon_0)\times\mathbb{R}^d\} < \infty$.

Proof. First observe that the statements 2, 3 and 4 are immediate consequences of statement 1 together with the fact that the function k_h is of class $\mathscr{C}^2$ with compact support, which implies that $k_h(y-x)$, $D_x k_h(y-x)$, and $D_x^2 k_h(y-x)$ are uniformly bounded.

To prove statement 1, note that K_h is continuous and that $K_h(x) > 0$ for all $x \in \mathscr{L}(t)$. Set

$$\alpha(\varepsilon_0) = \inf\{\|D_x f(x)\|; x \in \mathscr{L}^t_{t-\varepsilon_0}\}.$$

Let $(x,y) \in \mathscr{L}^t_{t-\varepsilon_0} \times \partial\mathscr{L}(t)$. Then

$$\varepsilon_0 \geq f(y) - f(x) \geq \alpha(\varepsilon_0)\|y-x\|.$$

Thus, $\|y-x\| \leq \varepsilon_0/\alpha(\varepsilon_0)$ and so

$$\operatorname{dist}\big(x, \mathscr{L}(t)\big) \leq \frac{\varepsilon_0}{\alpha(\varepsilon_0)}, \quad \text{for all } x \in \mathscr{L}^t_{t-\varepsilon_0}.$$

Recall from (2.4) that $h/2 > \varepsilon_0/\alpha(\varepsilon_0)$. Consequently, for all $x \in \mathscr{L}(t-\varepsilon_0)$, the set $(x+hB/2)\cap\mathscr{L}(t)$ contains a non-empty, open set $U(x)$. Moreover k_h is bounded from below by some positive number on $hB/2$ by Assumption 2. Hence $K_h(x) > 0$ for all x in $\mathscr{L}(t-\varepsilon_0)$ and point 1 follows from the continuity of K_h and the compactness of $\mathscr{L}(t-\varepsilon_0)$. □

In order to prove the convergence of $\widehat{Q}_{n,h}$ to Q_h, we also need to study the uniform convergence of $K_{n,h}$, given in (2.2). Lemma 5.4 controls the difference between $K_{n,h}$ and K_h, while Lemma 5.5 controls the ratio of K_h over $K_{n,h}$.

Lemma 5.4. *As $n \to \infty$, almost surely,*
1. $\displaystyle\sup_{x\in\mathscr{L}(t-\varepsilon_0)} \Big|K_{n,h}(x) - K_h(x)\Big| \to 0$ *and*
2. $\displaystyle\sup_{x\in\mathscr{L}(t-\varepsilon_0)} \Big|D_x K_{n,h}(x) - D_x K_h(x)\Big| \to 0.$

Proof. Let

$$K^{\dagger}_{n,h}(x) := \frac{1}{n\mu(\mathscr{L}(t))}\sum_{i=1}^{n} k_h(X_i - x)\mathbf{1}_{\mathscr{L}_n(t)}(X_i),$$
$$K^{\dagger\dagger}_{n,h}(x) := \frac{1}{n\mu(\mathscr{L}(t))}\sum_{i=1}^{n} k_h(X_i - x)\mathbf{1}_{\mathscr{L}(t)}(X_i).$$

Let us start with the inequality

$$\left|K_{n,h}(x) - K_h(x)\right| \le \left|K_{n,h}(x) - K^{\dagger}_{n,h}(x)\right| + \left|K^{\dagger}_{n,h}(x) - K_h(x)\right|, \tag{5.1}$$

for all $x \in \mathscr{L}(t-\varepsilon_0)$. Using the inequality

$$\left|K_{n,h}(x) - K^{\dagger}_{n,h}(x)\right| \le \left|\frac{n}{j(n)} - \frac{1}{\mu(\mathscr{L}(t))}\right| \|k_h\|_\infty$$

we conclude that the first term in (5.1) tends to 0 uniformly in x over $\mathscr{L}(t-\varepsilon_0)$ with probability one as $n \to \infty$, since $j(n)/n \to \mu\big(\mathscr{L}(t)\big)$ almost surely, and since k_h is bounded on $\mathbb{R}^d$.

Next, for all $x \in \mathscr{L}(t-\varepsilon_0)$, we have

$$\left|K^{\dagger}_{n,h}(x) - K_h(x)\right| \le \left|K^{\dagger}_{n,h}(x) - K^{\dagger\dagger}_{n,h}(x)\right| + \left|K^{\dagger\dagger}_{n,h}(x) - K_h(x)\right|. \tag{5.2}$$

The first term in (5.2) is bounded by

$$\begin{aligned}\left|K^{\dagger}_{n,h}(x) - K^{\dagger\dagger}_{n,h}(x)\right| &\le \frac{\|k_h\|_\infty}{\mu\big(\mathscr{L}(t)\big)}\frac{1}{n}\left|\sum_{i=1}^{n}\left\{\mathbf{1}_{\mathscr{L}_n(t)}(X_i) - \mathbf{1}_{\mathscr{L}(t)}(X_i)\right\}\right| \\ &= \frac{\|k_h\|_\infty}{\mu\big(\mathscr{L}(t)\big)}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\mathscr{L}_n(t)\Delta\mathscr{L}(t)}(X_i),\end{aligned}$$

where $\mathscr{L}_n(t)\Delta\mathscr{L}(t)$ denotes the symmetric difference between $\mathscr{L}_n(t)$ and $\mathscr{L}(t)$. Recall that, on the event Ω_n, $\mathscr{L}(t-\varepsilon_n) \subset \mathscr{L}_n(t) \subset \mathscr{L}(t-\varepsilon_n)$. Therefore $\mathscr{L}_n(t)\Delta\mathscr{L}(t) \subset \mathscr{L}^{t+\varepsilon_n}_{t-\varepsilon_n}$ on Ω_n, and so

$$0 \le \frac{1}{n}\left|\sum_{i=1}^{n}\left\{\mathbf{1}_{\mathscr{L}_n(t)}(X_i) - \mathbf{1}_{\mathscr{L}(t)}(X_i)\right\}\right|\mathbf{1}_{\Omega_n} \le \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{A_n}(X_i),$$

where $A_n = \mathscr{L}_{t-\varepsilon_n}^{t+\varepsilon_n}$. Hence by Lemma 5.1, and since $\mathbf{1}_{\Omega_n} \to 1$ almost surely as $n \to \infty$, the first term in (5.2) converges to 0 with probability one as $n \to \infty$.

Next, since the collection $\{y \mapsto k_h(y-x)\mathbf{1}_{\mathscr{L}(t)}(y) : x \in \mathscr{L}(t-\varepsilon_0)\}$ is Glivenko-Cantelli by Lemma 5.2, we conclude that

$$\sup_{x \in \mathscr{L}(t-\varepsilon_0)} \left| K_{n,h}^{\dagger\dagger}(x) - K_h(x) \right| \to 0,$$

with probability one as $n \to \infty$. This concludes the proof of the first statement.

The second statement may be proved by developing similar arguments, with k_h replaced by $D_x k_h$, and by noting that the collection of functions $\{y \mapsto D_x k_h(y-x)\mathbf{1}_{\mathscr{L}(t)}(y) : x \in \mathscr{L}(t-\varepsilon_0)\}$ is also Glivenko-Cantelli by Lemma 5.2. □

Lemma 5.5. *As $n \to \infty$, almost surely,*

1. $\sup_{x \in \mathscr{L}(t)} \left| \dfrac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1 \right| \to 0$, *and*

2. $\sup_{x \in \mathscr{L}(t)} \left\| D_x \left[\dfrac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} \right] \right\| \to 0.$

Proof. First of all, K_h is uniformly continuous on $\mathscr{L}(t-\varepsilon_0)$ since K_h is continuous and since $\mathscr{L}(t-\varepsilon_0)$ is compact. Moreover, φ_n converges uniformly to the identity map of $\mathscr{L}(t)$ by Lemma A.1. Hence

$$\sup_{x \in \mathscr{L}(t)} \left| K_h\big(\varphi_n(x)\big) - K_h(x) \right| \to 0 \quad \text{as } n \to \infty,$$

and since $K_{n,h}$ converges uniformly to K_h with probability one as $n \to \infty$ by Lemma 5.4, this proves 1.

We have

$$\begin{aligned} D_x \left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} \right] &= \Big[K_{n,h}\big(\varphi_n(x)\big) \Big]^{-2} D_x \varphi_n(x) \\ &\quad \times \Big[K_{n,h}\big(\varphi_n(x)\big) D_x K_h\big(\varphi_n(x)\big) - K_h\big(\varphi_n(x)\big) D_x K_{n,h}\big(\varphi_n(x)\big) \Big]. \end{aligned}$$

Since $D_x\varphi_n(x)$ converges to the identity matrix I_d uniformly over $x \in \mathscr{L}(t)$ by Lemma A.1, $\|D_x\varphi_n(x)\|$ is bounded uniformly over n and $x \in \mathscr{L}(t)$ by some positive constant C_φ. Furthermore the map $x \mapsto K_{n,h}(x)$ is bounded from below over $\mathscr{L}(t)$ by some positive constant k_{min} independent of x because i) $\inf_{x\in\mathscr{L}(t-\varepsilon_0)} K_h(x) > 0$ by Lemma 5.3, and ii) $\sup_{x\in\mathscr{L}(t-\varepsilon_0)} \left|K_{n,h}(x) - K_h(x)\right| \to 0$ by Lemma 5.4. Hence

$$\left| D_x \left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} \right] \right| \leq \frac{C_\varphi}{k_{min}^2} \left| K_{n,h}(y) D_x K_h(y) - K_h(y) D_x K_{n,h}(y) \right|,$$

where we have set $y = \varphi_n(x)$ which belongs to $\mathscr{L}(t-\varepsilon_n) \subset \mathscr{L}(t-\varepsilon_0)$. At last, Lemma 5.4 gives

$$\sup_{y\in\mathscr{L}(t-\varepsilon_0)} \left| K_{n,h}(y) D_x K_h(y) - K_h(y) D_x K_{n,h}(y) \right| \to 0 \quad \text{almost surely,}$$

as $n \to \infty$ which proves 2. □

We are now almost ready to prove the uniform convergence of empirical operators. The following lemma is a consequence of Lemma 5.2.

Lemma 5.6. *Let $r : \mathscr{L}(t-\varepsilon_0) \times \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function with compact support such that (i) r is uniformly bounded on $\mathscr{L}(t-\varepsilon_0) \times \mathbb{R}^d$, i.e., $\|r\|_\infty < \infty$, and (ii) the differential $D_x r$ with respect to x is uniformly bounded on $\mathscr{L}(t-\varepsilon_0) \times \mathbb{R}^d$, i.e., $\|D_x r\|_\infty := \sup\left\{ \|D_x r(x,y)\| : (x,y) \in \mathscr{L}(t-\varepsilon_0) \times \mathbb{R}^d \right\} < \infty$. Define the linear operators R_n and R on $W\big(\mathscr{L}(t)\big)$ respectively by*

$$R_n g(x) = \int_{\mathscr{L}_n(t)} r\big(\varphi_n(x), y\big) g\big(\varphi_n^{-1}(y)\big) \mathbb{P}_n^t(dy),$$
$$R g(x) = \int_{\mathscr{L}(t)} r(x,y) g(y) \mu^t(dy).$$

Then, as $n \to \infty$,

$$\sup\left\{ \left\| R_n g - R g \right\|_\infty : \|g\|_W \leq 1 \right\} \to 0 \quad \textit{almost surely}.$$

Proof. Set

$$S_n g(x) := \frac{1}{\mu(\mathscr{L}(t))}\frac{1}{n}\sum_{i=1}^{n} r\big(\varphi_n(x),X_i\big)\, g\big(\varphi_n^{-1}(X_i)\big)\mathbf{1}_{\mathscr{L}_n(t)}(X_i),$$
$$T_n g(x) := \frac{1}{\mu\big(\mathscr{L}(t)\big)}\frac{1}{n}\sum_{i=1}^{n} r\big(\varphi_n(x),X_i\big) g(X_i)\mathbf{1}_{\mathscr{L}(t)}(X_i),$$
$$U_n g(x) := \frac{1}{\mu\big(\mathscr{L}(t)\big)}\frac{1}{n}\sum_{i=1}^{n} r\big(x,X_i\big) g(X_i)\mathbf{1}_{\mathscr{L}(t)}(X_i).$$

and consider the inequality

$$\begin{aligned}\big|R_n g(x) - Rg(x)\big| \le \big|R_n g(x) - S_n g(x)\big| + \big|S_n g(x) - T_n g(x)\big| \\ + \big|T_n g(x) - U_n g(x)\big| + \big|U_n g(x) - Rg(x)\big|,\end{aligned} \tag{5.3}$$

for all $x \in \mathscr{L}(t)$ and all $g \in W\big(\mathscr{L}(t)\big)$.

The first term in (5.3) is bounded uniformly by

$$\big|R_n g(x) - S_n g(x)\big| \le \left|\frac{n}{j(n)} - \frac{1}{\mu\big(\mathscr{L}(t)\big)}\right| \|r\|_\infty \|g\|_\infty$$

and since $j(n)/n$ tends to $\mu(\mathscr{L}(t))$ almost surely as $n \to \infty$, we conclude that

$$\sup\Big\{\big\|R_n g - S_n g\big\|_\infty : \|g\|_W \le 1\Big\} \to 0 \quad \text{a.s. as } n \to \infty. \tag{5.4}$$

For the second term in (5.3), we have

$$\begin{aligned}|S_n g(x) - T_n g(x)| &\le \frac{\|r\|_\infty}{\mu\big(\mathscr{L}(t)\big)}\frac{1}{n}\sum_{i=1}^{n}\big|g\big(\varphi_n^{-1}(X_i)\big)\mathbf{1}_{\mathscr{L}_n(t)}(X_i) - g(X_i)\mathbf{1}_{\mathscr{L}(t)}(X_i)\big| \\ &= \frac{\|r\|_\infty}{\mu\big(\mathscr{L}(t)\big)}\frac{1}{n}\sum_{i=1}^{n} g_n(X_i),\end{aligned} \tag{5.5}$$

where g_n is the function defined on the whole space $\mathbb{R}^d$ by

$$g_n(x) = \Big|g\big(\varphi_n^{-1}(x)\big)\mathbf{1}_{\mathscr{L}_n(t)}(x) - g(x)\mathbf{1}_{\mathscr{L}(t)}(x)\Big|.$$

Consider the partition of $\mathbb{R}^d$ given by $\mathbb{R}^d = B_{1,n} \cup B_{2,n} \cup B_{3,n} \cup B_{4,n}$, where

$$\begin{aligned} B_{1,n} &:= \mathscr{L}_n(t) \cap \mathscr{L}(t), \quad B_{2,n} := \mathscr{L}_n(t) \cap \mathscr{L}(t)^c, \\ B_{3,n} &:= \mathscr{L}_n(t)^c \cap \mathscr{L}(t), \quad B_{4,n} := \mathscr{L}_n(t)^c \cap \mathscr{L}(t)^c. \end{aligned}$$

The sum over i in (5.5) may be split into four parts as

$$\frac{1}{n}\sum_{i=1}^{n} g_n(X_i) = I_1(x,g) + I_2(x,g) + I_3(x,g) + I_4(x,g) \tag{5.6}$$

where

$$I_k(x,g) := \frac{1}{n}\sum_{i=1}^{n} g_n(X_i)\mathbf{1}\{X_i \in B_{k,n}\}.$$

First, $I_{4,n}(x,g) = 0$ since g_n is identically 0 on $B_{4,n}$. Second,

$$I_2(x,g) + I_3(x,g) \le \|g\|_\infty \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\mathscr{L}(t)\Delta\mathscr{L}_n(t)}(X_i) \tag{5.7}$$

Applying Lemma 5.1 together with the almost sure convergence of $\mathbf{1}_{\Omega_n}$ to 1, we obtain that

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\mathscr{L}(t)\Delta\mathscr{L}_n(t)}(X_j) \to 0 \quad \text{almost surely.} \tag{5.8}$$

Third,

$$\begin{aligned} I_1(x,g) &\le \sup_{x\in\mathscr{L}(t)} \left| g\big(\varphi_n^{-1}(x)\big) - g(x) \right| \\ &\le \|D_x g\|_\infty \sup_{x\in\mathscr{L}(t)} \|\varphi_n^{-1}(x) - x\| \\ &\le \|D_x g\|_\infty \sup_{x\in\mathscr{L}(t)} \|x - \varphi_n(x)\| \\ &\to 0 \end{aligned} \tag{5.9}$$

as $n \to \infty$ by Lemma A.1. Thus, combining (5.5), (5.6), (5.7), (5.8) and (5.9) leads to

$$\sup\left\{ \big\|S_n g - T_n g\big\|_\infty : \|g\|_W \le 1 \right\} \to 0 \quad \text{a.s. as } n \to \infty. \tag{5.10}$$

For the third term in (5.3), using the inequality

$$\big|r\big(\varphi_n(x), X_i\big) - r\big(x, X_i\big)\big| \le \|D_x r\|_\infty \sup_{x\in\mathscr{L}(t)} \|\varphi_n(x) - x\|$$

we deduce that

$$\big|T_n g(x) - U_n g(x)\big| \le \frac{1}{\mu\big(\mathscr{L}(t)\big)} \|g\|_\infty \|D_x r\|_\infty \sup_{x\in\mathscr{L}(t)} \|\varphi_n(x) - x\|.$$

and so

$$\sup\left\{\left\|T_n g - U_n g\right\|_\infty : \|g\|_W \leq 1\right\} \to 0 \quad \text{a.s. as } n \to \infty, \tag{5.11}$$

by Lemma A.1.

At last, for the fourth term in (5.3), since the function r satisfies the conditions of the second statement in Lemma 5.2, we conclude by Lemma 5.2 that

$$\sup\left\{\left\|U_n g - R g\right\|_\infty : \|g\|_W \leq 1\right\} \to 0 \quad \text{a.s. as } n \to \infty. \tag{5.12}$$

Finally, reporting (5.4), (5.10) and (5.11) in (5.3) yields the desired result. □

5.2 Proof of Theorem 3.1

We will prove that, as $n \to \infty$, almost surely,

$$\sup\left\{\left\|\widehat{Q}_{n,h} g - Q_h g\right\|_\infty : \|g\|_W \leq 1\right\} \to 0 \tag{5.13}$$

and

$$\sup\left\{\left\|D_x\left[\widehat{Q}_{n,h} g\right] - D_x\left[Q_h g\right]\right\|_\infty : \|g\|_W \leq 1\right\} \to 0 \tag{5.14}$$

To this aim, we introduce the operator $\widetilde{Q}_{n,h}$ acting on $W(\mathscr{L}(t))$ as

$$\widetilde{Q}_{n,h} g(x) = \int_{\mathscr{L}_n(t)} q_h(\varphi_n(x), y) g\big(\varphi_n^{-1}(y)\big) \mathbb{P}_n^t(dy).$$

Proof of (5.13) For all $g \in W\big(\mathscr{L}(t)\big)$, we have

$$\left\|\widehat{Q}_{n,h} g - Q_h g\right\|_\infty \leq \left\|\widehat{Q}_{n,h} g - \widetilde{Q}_{n,h} g\right\|_\infty + \left\|\widetilde{Q}_{n,h} g - Q_h g\right\|_\infty. \tag{5.15}$$

First, by Lemma 5.3, the function $r = q_h$ satisfies the condition in Lemma 5.6, so that

$$\sup\left\{\|\widetilde{Q}_{n,h} g - Q_h g\|_\infty : \|g\|_W \leq 1\right\} \to 0 \tag{5.16}$$

with probability one as $n \to \infty$.

Next, since $\|q_h\|_\infty < \infty$ by Lemma 5.3, there exists a finite constant C_h such that,

$$\|\widetilde{Q}_{n,h}g\|_\infty \le C_h \quad \text{for all } n \text{ and all } g \text{ with } \|g\|_W \le 1. \tag{5.17}$$

By definition of $q_{n,h}$, for all x,y in the level set $\mathscr{L}(t)$, we have

$$q_{n,h}(x,y) = \frac{K_h(x)}{K_{n,h}(x)} q_h(x,y). \tag{5.18}$$

So

$$\begin{aligned}\left|\widehat{Q}_{n,h}g(x) - \widetilde{Q}_{n,h}g(x)\right| &= \left|\frac{K_n\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1\right| \left|\widetilde{Q}_{n,h}g(x)\right| \\ &\le C_h \sup_{x\in\mathscr{L}(t)} \left|\frac{K_n\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1\right|,\end{aligned}$$

where C_h is as in (5.17). Applying Lemma 5.5 yields

$$\sup\left\{\|\widehat{Q}_{n,h}g - \widetilde{Q}_{n,h}g\|_\infty : \|g\|_W \le 1\right\} \to 0 \tag{5.19}$$

with probability one as $n \to \infty$. Reporting (5.16) and (5.19) in (5.15) proves (5.13).

Proof of (5.14) We have

$$\begin{aligned}&\left\|D_x\left[\widehat{Q}_{n,h}g\right] - D_x\left[Q_h g\right]\right\|_\infty \\ &\qquad \le \left\|D_x\left[\widehat{Q}_{n,h}g\right] - D_x\left[\widetilde{Q}_h g\right]\right\|_\infty + \left\|D_x\left[\widetilde{Q}_{n,h}g\right] - D_x\left[Q_h g\right]\right\|_\infty. \end{aligned}\tag{5.20}$$

The second term in (5.20) is bounded by

$$\left\|D_x\left[\widetilde{Q}_{n,h}g\right] - D_x\left[Q_h g\right]\right\|_\infty \le \|D_x\varphi_n\|_\infty \|R_n g - Rg\|_\infty,$$

where

$$\begin{aligned}R_n g(x) &:= \int_{\mathscr{L}_n(t)} (D_x q_h)(\varphi_n(x),y) g\big(\varphi_n^{-1}(y)\big)\mathbb{P}_n^t(dy) \quad \text{and} \\ Rg(x) &:= \int_{\mathscr{L}(t)} (D_x q_h)(\varphi_n(x),y) g\big(\varphi_n^{-1}(y)\big)\mu^t(dy).\end{aligned}$$

By lemma A.1, $x \mapsto D_x\varphi_n(x)$ converges to the identity matrix I_d of $\mathbb{R}^d$, uniformly in x over $\mathscr{L}(t)$. So $\|D_x\varphi_n(x)\|$ is bounded by some finite constant C_φ uniformly over n and $x \in \mathscr{L}(t)$ and

$$\left\| D_x\left[\widetilde{Q}_{n,h}g\right] - D_x\left[Q_h g\right] \right\|_\infty \leq C_\varphi \left\| R_n g - Rg \right\|_\infty .$$

By Lemma 5.3, the map $r : (x,y) \mapsto D_x q_h(x,y)$ satisfies the conditions in Lemma 5.6. Thus, $\|R_n g - Rg\|_\infty$ converges to 0 almost surely, uniformly over g in the unit ball of $W(\mathscr{L}(t))$, and we deduce that

$$\sup\left\{ \left\| D_x\left[\widetilde{Q}_{n,h}g\right] - D_x\left[Q_h g\right] \right\|_\infty : \|g\|_W \leq 1 \right\} \to 0 \quad \text{a.s. as } n \to \infty. \tag{5.21}$$

For the first term in (5.20), observe first that there exists a constant C'_h such that, for all n and all g in the unit ball of $W\left(\mathscr{L}(t)\right)$,

$$\left\| R_{n,h} g \right\|_\infty \leq C'_h, \quad \text{for all } n \text{ and all } g \text{ with } \|g\|_W \leq 1, \tag{5.22}$$

by Lemma 5.3.

On the one hand, we have

$$\begin{aligned} D_x\left[q_{n,h}(\varphi_n(x),y)\right] = {} & \frac{K_h\left(\varphi_n(x)\right)}{K_{n,h}\left(\varphi_n(x)\right)} D_x\varphi_n(x)(D_x q_h)\left(\varphi_n(x),y\right) \\ & + D_x\left[\frac{K_h\left(\varphi_n(x)\right)}{K_{n,h}\left(\varphi_n(x)\right)}\right] q_h\left(\varphi_n(x),y\right). \end{aligned}$$

Hence,

$$D_x\left[\widehat{Q}_{n,h}g(x)\right] = \frac{K_h\left(\varphi_n(x)\right)}{K_{n,h}\left(\varphi_n(x)\right)} D_x\varphi_n(x) R_n g(x) + D_x\left[\frac{K_h\left(\varphi_n(x)\right)}{K_{n,h}\left(\varphi_n(x)\right)}\right] \widetilde{Q}_{n,h}g(x).$$

On the other hand, since $D_x\left[q_h\left(\varphi_n(x),y\right)\right] = D_x\varphi_n(x)(D_x q_h)\left(\varphi_n(x),y\right)$,

$$D_x\left[\widetilde{Q}_{n,h}g(x)\right] = D_x\varphi_n(x) R_n g(x).$$

Thus,

$$D_x\Big[\widehat{Q}_{n,h}g(x)\Big] - D_x\Big[\widetilde{Q}_h g(x)\Big] = D_x\left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)}\right]\widetilde{Q}_{n,h}g(x) + \left(\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1\right) D_x\varphi_n(x) R_n g(x).$$

Using the inequalities (5.17) and (5.22), we obtain

$$\left\|D_x\Big[\widehat{Q}_{n,h}g\Big] - D_x\Big[\widetilde{Q}_h g\Big]\right\|_\infty \le C_h \sup_{x\in\mathscr{L}(t)}\left|D_x\left[\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)}\right]\right| + C_h' C_\varphi \sup_{x\in\mathscr{L}(t)}\left|\frac{K_h\big(\varphi_n(x)\big)}{K_{n,h}\big(\varphi_n(x)\big)} - 1\right|.$$

and by applying Lemma 5.5, we deduce that

$$\sup\left\{\left\|D_x\Big[\widehat{Q}_{n,h}g\Big] - D_x\Big[\widetilde{Q}_h g\Big]\right\|_\infty : \|g\|_W \le 1\right\} \to 0 \quad \text{a.s. as } n\to\infty. \tag{5.23}$$

Reporting (5.21) and (5.23) in (5.20) proves (5.14). □

6 Proof of Corollary 3.2

Let us start with the following proposition, which relates the spectrum of the functional operator $\widehat{Q}_{n,h}$ with the one of the matrix $\mathbf{Q}_{n,h}$.

Proposition 6.1. *On Ω_n, we have $\pi_n\Phi_n\widehat{Q}_{n,h} = \mathbf{Q}_{n,h}\pi_n\Phi_n$ and the spectrum of the functional operator $\widehat{Q}_{n,h}$ is $\sigma(\widehat{Q}_{n,h}) = \{0\}\cup\sigma(\mathbf{Q}_{n,h})$.*

Proof. Recall that the evaluation map π_n defined in (2.3) is such that $\mathbf{Q}_{n,h}\pi_n = \pi_n Q_{n,h}$, and that, on Ω_n, $\widehat{Q}_{n,h} = \Phi_n Q_{n,h}\Phi_n^{-1}$. Moreover, since $\widehat{Q}_{n,h}$ and $Q_{n,h}$ are conjugate, their spectra are equal. Thus, there remains to show that $\sigma(Q_{n,h}) = \{0\}\cup\sigma(\mathbf{Q}_{n,h})$.

Remark that $Q_{n,h}$ is a finite rank operator, and that its range is spanned by the maps $x \mapsto q_{n,h}(x, X_j)$, for $j \in J(n)$. Thus its spectrum is composed of 0 and its eigenvalues. By the relation $\mathbf{Q}_{n,h}\pi_n = \pi_n Q_{n,h}$, it immediately follows that if g is an eigenfunction of $Q_{n,h}$ with eigenvalue λ, then $V = \pi_n(g)$ is an eigenvector of $\mathbf{Q}_{n,h}$ with eigenvalue λ. Conversely, if $\{V_j\}_j$ is an eigenvector of $\mathbf{Q}_{n,h}$, then with some easy algebra, it may be verified that the function g defined by

$$g(x) := \sum_{j \in J(n)} V_j \, q_{n,h}(x, X_j)$$

is an eigenfunction of $Q_{n,h}$ with the same eigenvalue. □

The spectrum of Q_h may be decomposed as $\sigma(Q_h) = \sigma_1(Q_h) \cup \sigma_2(Q_h)$, where $\sigma_1(Q_h) = \{1\}$ and where $\sigma_2(Q_h) = \sigma(Q_h) \setminus \{1\}$. Since 1 is an isolated eigenvalue, there exists η_0 in the open interval $(0;1)$ such that $\sigma(Q_h) \cap \{z \in \mathbb{C} : |z-1| \leq \eta_0\}$ is reduced to the singleton $\{1\}$. Moreover, 1 is an eigenvalue of Q_h of multiplicity ℓ, by proposition C.2. Hence by Theorem B.1, $W\big(\mathscr{L}(t)\big)$ decomposes into $W\big(\mathscr{L}(t)\big) = M_1 \oplus M_2$ where $\dim(M_1) = \ell$.

Split the spectrum of $\widehat{Q}_{n,h}$ as $\sigma\big(\widehat{Q}_{n,h}\big) = \sigma_1\big(\widehat{Q}_{n,h}\big) \cup \sigma_2\big(\widehat{Q}_{n,h}\big)$, where

$$\sigma_1\big(\widehat{Q}_{n,h}\big) = \sigma\big(\widehat{Q}_{n,h}\big) \cap \big\{z \in \mathbb{C} : |z-1| < \eta_0\big\}.$$

By Theorem B.1, this decomposition of the spectrum of $\widehat{Q}_{n,h}$ yields a decomposition of $W\big(\mathscr{L}(t)\big)$ as $W\big(\mathscr{L}(t)\big) = M_{n,1} \oplus M_{n,2}$, where $M_{n,1}$ and $M_{n,2}$ are stable subspaces under $\widehat{Q}_{n,h}$. Statements 4 and 6 of Theorem B.2, together with Proposition 6.1, gives the following convergences.

Proposition 6.2. *The first ℓ eigenvalues $\lambda_{n,1}, \lambda_{n,2}, \ldots, \lambda_{n,\ell}$ of $\mathbf{Q}_{n,h}$ converge to 1 almost surely as $n \to \infty$ and there exists $\eta_0 > 0$ such that, for all $j > \ell$, $\lambda_{n,j}$ belongs to $\{z : |z-1| \geq \eta_0\}$ for n large enough, with probability one.*

In addition to the convergence of the eigenvalues of $\mathbf{Q}_{n,h}$, the convergence of eigenspaces also holds. More precisely, let Π be the projector on M_1 along M_2 and Π_n the projector on $M_{n,1}$ along $M_{n,2}$. Statements 2, 3, 5 and 6 of Theorem B.2 leads to

Proposition 6.3. *Π_n converges to Π in operator norm almost surely and the dimension of $M_{n,1}$ is ℓ for all large enough n.*

Denote by $E_{n,1}$ the subspace of $\mathbb{R}^{J(n)}$ spanned by the eigenvectors of $\mathbf{Q}_{n,h}$ corresponding to the eigenvalues $\lambda_{n,1}, \ldots \lambda_{n,\ell}$. If n is large enough, we have the following isomorphisms of vector spaces:

$$\Pi_n : M_1 \xrightarrow{\cong} M_{n,1} \quad \text{and} \quad \pi_n \Phi_n : M_{n,1} \xrightarrow{\cong} E_{n,1}, \tag{6.1}$$

where, strictly speaking, the isomorphisms are defined by the restriction of Π_n and $\pi_n \Phi_n$ to M_1 and $M_{n,1}$, respectively.

The functions $g_{n,k} := \Pi_n \mathbf{1}_{\mathscr{C}_k}$, $k = 1, \ldots, \ell$ are in $M_{n,1}$ and converges to $\mathbf{1}_{\mathscr{C}_k}$ in W-norm. Then, the vectors $\vartheta_{n,k} = \pi_n(g_{n,k} \circ \varphi_n^{-1})$ are in $E_{n,1}$ and, as $n \to \infty$,

$$\vartheta_{n,k,j} = \Pi_n(\mathbf{1}_{\mathscr{C}_k}) \circ \varphi_n^{-1}(X_j) \to \mathbf{1}_{\mathscr{C}_k}(X_j) = \begin{cases} 1 & \text{if } k = k(j), \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

Since $V_{n,1}, \ldots, V_{n,\ell}$ form a basis of $E_{n,1}$, there exists a matrix ξ_n of dimension $\ell \times \ell$ such that

$$\vartheta_{n,k} = \sum_{i=1}^{\ell} \xi_{n,k,i} V_{n,i}.$$

Hence the j^{th} component of $\vartheta_{n,k}$, for all $j \in J(n)$, may be expressed as

$$\vartheta_{n,k,j} = \sum_{i=1}^{\ell} \xi_{n,k,i} V_{n,i,j}.$$

Since $\rho_n(X_j)$ is the vector of $\mathbb{R}^\ell$ with components $\{V_{n,i,j}\}_i$, the vector $\vartheta_{n,\bullet,j} = \{\vartheta_{n,k,j}\}_k$ of $\mathbb{R}^\ell$ is related to $\rho_n(X_j)$ by the linear transformation ξ_n, i.e.,

$$\vartheta_{n,\bullet,j} = \xi_n \rho_n(X_j).$$

The convergence of $\vartheta_{n,\bullet,j}$ to $e_{k(j)}$ then follows from (6.2) and Corollary 3.2 is proved.

Acknowledgments. This work was supported by the French National Research Agency (ANR) under grant ANR-09-BLAN-0051-01.

References

M. Anderberg. *Cluster Analysis for Applications*. Academic Press, New-York, 1973.

A. Azzalini and N. Torelli. Clustering via nonparametric estimation. *Stat. Comput.*, 17:71–80, 2007.

M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *Learning theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 486–500. Springer, Berlin, 2005.

M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 624–638. Springer, Berlin, 2004.

G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM Probab. Stat.*, 11:272–280, 2007.

F. R. K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 1997.

R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21: 5–30, 2006.

A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *Canadian Journal of Statistics*, 28:367–382, 2000.

A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, 36:441–459, 2001.

R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley Interscience, New-York, 2000.

M. Fillipone, F. Camastra, F. Masulli, and S. Rovetta. A survey of kernel and spectral methods for clustering. *Pattern Recognition*, 41(1):176–190, 2008.

L. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *Ann. Statis.*, 36(3):1324–1345, 2008.

E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006.

J. Hartigan. *Clustering Algorithms*. Wiley, New-York, 1975.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New-York, 2001.

M. Hein, J.-Y. Audibert, and U. von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8:1325–1368, 2007.

J. Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer-Verlag, Berlin, 1995.

T. Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17:277–364, 1961.

V. I. Koltchinskii. Asymptotics of spectral projections of some random matrices approximating integral operators. In *High dimensional probability (Oberwolfach, 1996)*, volume 43 of *Progr. Probab.*, pages 191–227. Birkhäuser, Basel, 1998.

T. Linder. Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer, Vienna, 2002.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkely Symp. Math. Statist. Prob.*, volume 1, pages 281–297, 1967.

G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New-York, 2000.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag, London, 1993.

J. W. Milnor. *Morse theory*. Annals of Mathematics Studies, No. 51. Princeton University Press, Princeton, N.J., 1963.

B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Difusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.*, 21(1):113–127, 2006.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 849–856. MIT Press, 2002.

D. Pollard. Consistency of k-means clustering. *Ann. Statis.*, 9(1):135–140, 1981.

B. L. S. Prakasa Rao. *Nonparametric functional estimation*. Probability and Mathematical Statistics. Academic Press Inc., New York, 1983.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

A. W. van der Vaart. Bracketing smooth functions. *Stochastic Process. Appl.*, 52 (1):93–105, 1994.

U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.

U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*, 2005.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Ann. Statis.*, 36(2):555–586, 2008.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Eighteenth Annual Conference on Neural Information Processing Systems (NIPS)*, 2004.

A Geometry of level sets

The proof of the following result is adapted from Theorem 3.1 in Milnor [1963] p.12 and Theorem 5.2.1 in Jost [1995] p.176.

Lemma A.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function of class $\mathscr{C}^2$. Let $t \in \mathbb{R}$ and suppose that there exists $\varepsilon_0 > 0$ such that $f^{-1}\big([t-\varepsilon_0; t+\varepsilon_0]\big)$ is non empty, compact and contains no critical point of f. Let $\{\varepsilon_n\}_n$ be a sequence of positive numbers such that $\varepsilon_n < \varepsilon_0$ for all n, and $\varepsilon_n \to 0$ as $n \to \infty$. Then there exists a sequence of diffeomorphisms $\varphi_n : \mathscr{L}(t) \to \mathscr{L}(t-\varepsilon_n)$ carrying $\mathscr{L}(t)$ to $\mathscr{L}(t-\varepsilon_n)$ such that:*

1. $\sup_{x \in \mathscr{L}(t)} \|\varphi_n(x) - x\| \to 0$ *and*

2. $\sup_{x\in\mathscr{L}(t)} \|D_x\varphi_n(x) - I_d\| \to 0,$

as $n \to \infty$, where $D_x\varphi_n$ denotes the differential of φ_n and where I_d is the identity matrix on $\mathbb{R}^d$.

Proof. Recall first that a one-parameter group of diffeomorphisms $\{\varphi_u\}_{u\in\mathbb{R}}$ of $\mathbb{R}^d$ gives rise to a vector field V defined by

$$V_x g = \lim_{u\to 0} \frac{g\big(\varphi_u(x)\big) - g(x)}{u}, \quad x \in \mathbb{R}^d,$$

for all smooth function $g : \mathbb{R}^d \to \mathbb{R}$. Conversely, a smooth vector field which vanishes outside of a compact set generates a unique one-parameter group of diffeomorphisms of $\mathbb{R}^d$; see Lemma 2.4 in Milnor [1963] p. 10 and Theorem 1.6.2 in Jost [1995] p. 42.

Denote the set $\{x \in \mathbb{R}^d : a \le f(x) \le b\}$ by $\mathscr{L}_a^b$, for $a \le b$. Let $\eta : \mathbb{R}^d \to \mathbb{R}$ be the non-negative differentiable function with compact support defined by

$$\eta(x) = \begin{cases} 1/\|D_x f(x)\|^2 & \text{if } x \in \mathscr{L}_{t-\varepsilon_0}^t, \\ (t+\varepsilon_0 - f(x))/\|D_x f(x)\|^2 & \text{if } x \in \mathscr{L}_t^{t+\varepsilon_0}, \\ 0 & \text{otherwise.} \end{cases}$$

Then the vector field V defined by $V_x = \eta(x) D_x f(x)$ has compact support $\mathscr{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$, so that V generates a one-parameter group of diffeomorphisms

$$\varphi_u : \mathbb{R}^d \to \mathbb{R}^d, \quad u \in \mathbb{R}.$$

We have

$$D_u\big[f\big(\varphi_u(x)\big)\big] = \langle V, D_x f\rangle_{\varphi_u(x)} \ge 0,$$

since η is non-negative. Furthermore,

$$\langle V, D_x f\rangle_{\varphi_u(x)} = 1, \quad \text{if } \varphi_u(x) \in \mathscr{L}_{t-\varepsilon_0}^t$$

Consequently the map $u \mapsto f\big(\varphi_u(x)\big)$ has constant derivative 1 as long as $\varphi_u(x)$ lies in $\mathscr{L}_{t-\varepsilon_0}^t$. This proves the existence of the diffeomorphism $\varphi_n := \varphi_{-\varepsilon_n}$ which carries $\mathscr{L}(t)$ to $\mathscr{L}(t-\varepsilon_n)$.

Note that the map $u \in \mathbb{R} \mapsto \varphi_u(x)$ is the integral curve of V with initial condition x. Without loss of generality, suppose that $\varepsilon_n \le 1$. For all x in $\mathscr{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$, we have

$$\|\varphi_n(x) - x\| \le \int_{-\varepsilon_n}^0 \big\|D_u\big(\varphi_u(x)\big)\big\|\, du \le \varepsilon_n/\beta(\varepsilon_n) \le \varepsilon_n/\beta(\varepsilon_0)$$

where we have set

$$\beta(\varepsilon) := \inf\big\{\|D_x f(x)\| : x \in \mathscr{L}_{t-\varepsilon}^{t+\varepsilon}\big\} > 0.$$

This proves the statement 1, since $\varphi_n(x) - x$ is identically 0 on $\mathscr{L}(t+\varepsilon_0)$.

For the statement 2, observe that $\varphi_u(x)$ satisfies the relation

$$\varphi_u(x) - x = \int_0^u D_v\big(\varphi_v(x)\big)dv = \int_0^u V\big(\varphi_v(x))\big)dv.$$

Differentiating with respect to x yields

$$D_x\varphi_u(x) - I_d = \int_0^u D_x\varphi_v(x) \circ D_x V\big(\varphi_v(x)\big)dv.$$

Since f is of class $\mathscr{C}^2$, the two terms inside the integral are uniformly bounded over $\mathscr{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$, so that there exists a constant $C > 0$ such that

$$\|D_x\varphi_n - I\|_x \leq C\varepsilon_n,$$

for all x in $\mathscr{L}_{t-\varepsilon_0}^{t+\varepsilon_0}$. Since $\|D_x\varphi_n - I\|_x$ is identically zero on $\mathscr{L}(t+\varepsilon_0)$, this proves the statement 2. □

B Continuity of an isolated finite set of eigenvalues

In brief, the spectrum $\sigma(T)$ of a bounded linear operator T on a Banach space is upper semi-continuous in T, but not lower semi-continuous; see Kato [1995]IV§3.1 and IV§3.2. However, an isolated finite set of eigenvalues of T is continuous in T, as stated in Theorem B.2 below.

Let T be a bounded operator on the $\mathbb{C}$-Banach space E with spectrum $\sigma(T)$. Let $\sigma_1(T)$ be a finite set of eigenvalues of T. Set $\sigma_2(T) = \sigma(T) \setminus \sigma_1(T)$ and suppose that $\sigma_1(T)$ is separated from $\sigma_2(T)$ by a rectifiable, simple, and closed curve Γ. Assume that a neighborhood of $\sigma_1(T)$ is enclosed in the interior of Γ. Then we have the following theorem; see Kato [1995], III.§6.4 and III.§6.5.

Theorem B.1 (Separation of the spectrum)**.** *The Banach space E decomposes into a pair of supplementary subspaces as $E = M_1 \oplus M_2$ such that T maps M_j into M_j ($j = 1,2$) and the spectrum of the operator induced by T on M_j is $\sigma_j(T)$ ($j = 1,2$). If additionally the total multiplicity m of $\sigma_1(T)$ is finite, then* $\dim(M_1) = m$.

Moreover, the following theorem states that a finite system of eigenvalues of T, as well as the decomposition of E of Theorem B.1, depends continuously of T, see Kato [1995], IV.§3.5. Let $\{T_n\}_n$ be a sequence of operators which converges to T in norm. Denote by $\sigma_1(T_n)$ the part of the spectrum of T_n enclosed in the interior of the closed curve Γ, and by $\sigma_2(T_n)$ the remainder of the spectrum of T_n.

Theorem B.2 (Continuous approximation of the spectral decomposition)**.** *There exists a finite integer n_0 such that the following holds true.*
1. Both $\sigma_1(T_n)$ and $\sigma_2(T_n)$ are nonempty for all $n \geq n_0$ provided this is true for T.
2. For each $n \geq 0$, the Banach space E decomposes into two subspaces as $E = M_{n,1} \oplus M_{n,2}$ in the manner of Theorem B.1, i.e. T_n maps $M_{n,j}$ into itself and the spectrum of T_n on $M_{n,j}$ is $\sigma_j(T_n)$.
3. For all $n \geq n_0$, $M_{n,j}$ is isomorphic to M_j.

4. If $\sigma_1(T)$ is a singleton $\{\lambda\}$, then every sequence $\{\lambda_n\}_n$ with $\lambda_n \in \sigma_1(T_n)$ for all $n \geq n_0$ converges to λ.
5. If Π is the projector on M_1 along M_2 and Π_n the projector on $M_{n,1}$ along $M_{n,2}$, then Π_n converges in norm to Π.
6. If the total multiplicity m of $\sigma_1(T)$ is finite, then, for all $n \geq n_0$, the total multiplicity of $\sigma_1(T_n)$ is also m and $\dim(M_{n,1}) = m$.

C Markov chains and limit operator

For the reader not familiar with Markov chains on a general state space, we begin by summarizing the relevant part of the theory.

C.1 Background materials on Markov chains

Let $\{\xi_i\}_{i\geq 0}$ be a Markov chain with state space $\mathscr{S} \subset \mathbb{R}^d$ and transition kernel $q(x,dy)$. We write P_x for the probability measure when the initial state is x and E_x for the expectation with respect to P_x. The Markov chain is called *(strongly) Feller* if the map

$$x \in \mathscr{S} \mapsto Qg(x) := \int_{\mathscr{S}} q(x,dy)g(y) = \mathbb{E}_x f(\xi_1)$$

is continuous for every bounded, measurable function g on $\mathscr{S}$; see Meyn and Tweedie [1993], p. 132. This condition ensures that the chain behaves nicely with the topology of the state space $\mathscr{S}$. The notion of irreducibility expresses the idea that, from an arbitrary initial point, each subset of the state space may be reached by the Markov chain with a positive probability. A Feller chain is said *open set irreducible* if, for every points x,y in $\mathscr{S}$, and every $\eta > 0$,

$$\sum_{n\geq 1} q^n(x, y+\eta B) > 0,$$

where $q^n(x,dy)$ stands for the n-step transition kernel; see Meyn and Tweedie [1993], p. 135. Even if open set irreducible, a Markov chain may exhibit a periodic behavior, i.e., there may exist a partition $\mathscr{S} = \mathscr{S}_0 \cup \mathscr{S}_1 \cup \ldots \cup \mathscr{S}_N$ of the state space such that, for every initial state $x \in \mathscr{S}_0$,

$$P_x[\xi_1 \in \mathscr{S}_1, \xi_2 \in \mathscr{S}_2, \ldots, \xi_N \in \mathscr{S}_N, \xi_{N+1} \in \mathscr{S}_0, \ldots] = 1.$$

Such a behavior does not occur if the Feller chain is *topologically aperiodic*, i.e., if for each initial state x, each $\eta > 0$, there exists n_0 such that $q^n(x, x+\eta B) > 0$ for every $n \geq n_0$; see Meyn and Tweedie [1993], p. 479.

Next we come to ergodic properties of the Markov chain. A Borel set A of $\mathscr{S}$ is called *Harris recurrent* if the chain visits A infinitely often with probability 1 when started at any point x of A, i.e.,

$$P_x\left(\sum_{i=0}^{\infty} \mathbf{1}_A(\xi_i) = \infty\right) = 1$$

for all $x \in A$. The chain is then said to be *Harris recurrent* if every Borel set A with positive Lebesgue measure is Harris recurrent; see Meyn and Tweedie [1993], p. 204. At least two types of behavior, called evanescence and non-evanescence, may occur. The event $[\xi_n \to \infty]$ denotes the fact that the sample path visits each compact set only finitely many often, and the Markov chain is called *non-evanescent* if $P_x(\xi_n \to \infty) = 0$ for each initial state $x \in \mathscr{S}$. Specifically, a Feller chain is Harris recurrent if and only if it is non-evanescent; see Meyn and Tweedie [1993], Theorem 9.2.2, p. 212.

The ergodic properties exposed above describe the long time behavior of the chain. A measure ν on the state space is said *invariant* if

$$\nu(A) = \int_{\mathscr{S}} q(x,A)\nu(dx)$$

for every Borel set A in $\mathscr{S}$. If the chain is Feller, open set irreducible, topologically aperiodic and Harris recurrent, it admits a unique (up to constant multiples) invariant measure ν; see Meyn and Tweedie [1993], Theorem 10.0.1 p. 235. In this case, either $\nu(\mathscr{S}) < \infty$ and the chain is called *positive*, or $\nu(\mathscr{S}) = \infty$ and the chain is called *null*. The following important result provides one with the limit of the distribution of ξ_n when $n \to \infty$, whatever the initial state is. Assuming that the chain is Feller, open set irreducible, topologically aperiodic and positive Harris recurrent, the sequence of distribution $\{q^n(x,dy)\}_{n\geq 1}$ converges in total variation to $\nu(dy)$, the unique invariant probability distribution; see Theorem 13.3.1 of Meyn and Tweedie [1993], p. 326. That is to say, for every x in $\mathscr{S}$,

$$\sup_g \left\{ \left| \int_{\mathscr{S}} g(y)q^n(x,dy) - \int_{\mathscr{S}} g(y)\nu(dy) \right| \right\} \to 0 \quad \text{as } n \to \infty,$$

where the supremum is taken over all continuous functions g from $\mathscr{S}$ to $\mathbb{R}$ with $\|g\|_\infty \leq 1$.

C.2 Limit properties of Q_h

With the definitions and results from the previous paragraph, we may now study the properties of the limit clustering induced by the operator Q_h. The transition kernel $q_h(x,dy) := q_h(x,y)\mu^t(dy)$ defines a Markov chain with state space $\mathscr{L}(t)$. Recall that $\mathscr{L}(t)$ has ℓ connected components $\mathscr{C}_1, \ldots, \mathscr{C}_\ell$ and that under Assumption 3, h is strictly lower than d_{min}, the minimal distance between the connected components.

Proposition C.1. *1. The chain is Feller and topologically aperiodic.*
2. When started at a point x in some connected component of the state space, the chain evolves within this connected component only.
3. When the state space is reduced to some connected component of $\mathscr{L}(t)$, the chain is open set irreducible and positive Harris recurrent.

Proof. 1. Since the similarity function k_h is continuous, with compact support hB, the map

$$x \mapsto Q_h g(x) = \int_{\mathscr{L}(t)} q_h(x,dy)g(y)$$

is continuous for every bounded, measurable function g. Moreover, k_h is bounded from below on $(h/2)B$ by Assumption 2. Thus, for each $x \in \mathscr{L}(t)$, $n \geq 1$ and $\eta > 0$, $q_h^n(x, x+\eta B) > 0$. Hence, the chain is Feller and topologically aperiodic.

2. Without loss of generality, assume that $x \in \mathscr{C}_1$. Let y be a point of $\mathscr{L}(t)$ which does not belong to $\mathscr{C}_1$. Then $\|y-x\| \geq d_{min} > h$ so that $q_h(x,y) = 0$. Whence,

$$P_x(\xi_1 \in \mathscr{C}_1) = q_h(x, \mathscr{C}_1) = \int_{\mathscr{C}_1} q_h(x,y)\mu^t(dy) = \int_{\mathscr{L}(t)} q_h(x,y)\mu^t(dy) = 1.$$

3. Assume that the state space is reduced to $\mathscr{C}_1$. Fix $x, y \in \mathscr{C}_1$ and $\eta > 0$. Since $\mathscr{C}_1$ is connected, there exists a finite sequence $x_0, x_1, \ldots x_N$ of points in $\mathscr{C}_1$ such that $x_0 = x$, $x_N = y$, and $\|x_i - x_{i+1}\| \leq h/2$ for each i. Therefore

$$q_h^N(x, y+\eta B) \geq P_x(\xi_i \in x_i + \eta B \text{ for all } i \leq N) > 0$$

which proves that the chain is topologically aperiodic.

Since $\mathscr{C}_1$ is compact, the chain is non-evanescent, and so it is Harris recurrent. Recall that $k(x) = k(-x)$ from Assumption 2. Therefore $k_h(y-x) = k_h(x-y)$ which yields

$$K_h(x)q_h(x,dy)\mu^t(dx) = K_h(y)q_h(y,dx)\mu^t(dy).$$

By integrating the previous relation with respect to x over $\mathscr{C}_1$, one may verify that $K_h(x)\mu^t(dx)$ is an invariant measure. At last $\int_{\mathscr{C}_1} K_h(x)\mu^t(dx) < \infty$, which proves that the chain is positive. □

Proposition C.2. *If g is continuous and $Q_h g = g$, then g is constant on the connected components of $\mathscr{L}(t)$.*

Proof. We will prove that g is constant over $\mathscr{C}_1$. Proposition C.1 provides one with a unique invariant measure $\nu_1(dy)$ when the state space is reduced to $\mathscr{C}_1$. Fix x in $\mathscr{C}_1$. Since $g = Q_h g$, $g = Q_h^n g$ for every $n \geq 1$. Moreover by Proposition C.1, the chain is open set irreducible, topologically aperiodic, and positive Harris recurrent on $\mathscr{C}_1$. Thus, $q_h^n(x,dy)$ converges in total variation norm to $\nu_1(dy)$. Specifically,

$$Q_h^n g(x) \longrightarrow \int_{\mathscr{C}_1} g(y)\nu_1(dy) \quad \text{as } n \to \infty.$$

Hence, for every x in $\mathscr{C}_1$,

$$g(x) = \int_{\mathscr{C}_1} g(y)\nu_1(dy),$$

and since the last integral does not depend on x, it follows that g is a constant function on $\mathscr{C}_1$. □