A random matrix analysis of online learning: coping with limited memory resources

Anonymous Authors[1]

Abstract

This article introduces a random matrix framework for the analysis of online learning, a particularly relevant setting for a more sober processing of large amounts of data with limited memory and energy resources. Assuming data $\mathbf{x}_1, \mathbf{x}_2, \dots$ arrives as a continuous flow and a small number L of them can be kept in the learning pipeline, one has only access to the diagonal elements of the Gram kernel matrix: $[\mathbf{K}_L]_{i,j} = \frac{1}{p}\mathbf{x}_i^\top \mathbf{x}_j \mathbf{1}_{|i-j|<L}$. Under a large dimensional data regime, we derive the limiting spectral distribution of the punctured kernel matrix $\mathbf{K}_L$ and study its isolated eigenvalues and eigenvectors, which behave in an unfamiliar way. We detail how these results can be used to perform efficient online kernel spectral clustering and provide theoretical performance guarantees. Our findings are empirically confirmed on image classification tasks. Leveraging on optimality results of spectral methods for clustering, this work offers insights on efficient online clustering techniques for high-dimensional data.

1. Introduction

The ever-increasing amount of data coupled with the need for a more sober use of computational power puts online learning in the spotlight, as a way to deal with numerous and very large data with low memory resources. Be it because the volume of data is too high to be stored or because one is restricted to the sole use of a regular laptop, online learning appears as a handy and frugal way to process information. As data arrives in the learning pipeline, it is processed at a low computational cost before being discarded altogether, thus inducing a limited memory footprint.

Numerous works have proposed various algorithms to cluster data streams in an unsupervised manner (see, e.g., (Ghesmoune et al., 2016) and references therein). Among standard methods are the construction of a graph (Fritzke, 1995) or a tree of clusters (Zhang et al., 1996) which is updated as new data arrives, or else, the formation of clusters using a distance function, as in k-means, (Aggarwal et al., 2003) or a density-based method (Ester et al., 1996). Such algorithms are often adaptations of existing offline algorithms, like OpticsStream (Tasoulis et al., 2007), StreamKM++ (Ackermann et al., 2012), online k-means (Liberty et al., 2015), etc. These techniques operate on the entire feature space and their performance deteriorate as the dimension of the data increases. Therefore, (Aggarwal et al., 2004) proposed to cluster data streams after a projection on a lower-dimensional space. Sketching methods (Keriven et al., 2017; Gribonval et al., 2021) are also convenient to perform large-scale learning on data streams with a limited memory budget; the idea being to summarize the dataset into a single vector computed in one pass over the data.

Adapted from the standard spectral clustering algorithm (von Luxburg, 2007), techniques like incremental spectral clustering (Ning et al., 2010; Dhanjal et al., 2014) have been proposed to handle evolving data. Yet, they become quite memory-demanding when the number of samples grow large. Better suited to streaming applications, the spectral clustering algorithm of (Yoo et al., 2016) constructs a spectral embedding of the stream in one pass by adapting ideas from matrix sketching (Liberty, 2012).

Spectral clustering has indeed remarkably good performances on high-dimensional data as it manages to greatly reduce the dimensionality by keeping just a few leading spectral components. It is therefore computationally less demanding than many other classical clustering algorithms. Moreover, it reaches the optimal phase transition threshold (i.e., it performs better than random guess as soon as theoretically possible) (Onatski et al., 2013) and achieves the optimal clustering error rate in the Gaussian mixture model (Löffler et al., 2020).

It is also of particular interest from a random matrix theory perspective. Following the works of (El Karoui, 2010; Cheng & Singer, 2012) on the spectrum of kernel random matrices, (Couillet & Benaych-Georges, 2016) propose an

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

analysis of kernel spectral clustering with numerous high-dimensional data. Then, (Mai & Couillet, 2017) demonstrate that many standard machine learning algorithms in fact suffer from being ill-used when dealing with such data. Besides, given some data matrix $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix} \in \mathbb{R}^{p\times n}$, (Couillet et al., 2021) show that it is possible to get huge reductions in computational and storage costs with almost no performance loss by puncturing the data, i.e., keeping only a few elements of $\mathbf{X}$ and computing only a few elements of the Gram kernel matrix $\mathbf{K} = \frac{1}{p}\mathbf{X}^\top\mathbf{X}$. In addition, (Liao et al., 2020) demonstrate that, when carefully employed, sparsification and quantization of $\mathbf{K}$ incur negligible performance loss, while providing a great computational gain.

In the light of these numerous benefits of spectral clustering when dealing with high-dimensional data, of the practicality of online learning to handle large data streams with limited memory, and of the promising path shown by random matrix theory towards resource-efficient learning with performance guarantees, the present work introduces an "online spectral learning" algorithm to which we attach a rigorous performance analysis using random matrix theory.

The algorithm goes as follows: supposing that, due to memory limitations, only a small number L of data points can be kept in the pipeline, the computation of the $n \times n$ Gram kernel matrix is limited to the elements which are in a radius L around the diagonal of $\mathbf{K}$. This results in the following punctured kernel matrix model

$$\mathbf{K}_L = \frac{\mathbf{X}^\top\mathbf{X}}{p} \odot \mathbf{T}$$

where $\odot$ denotes the Hadamard product and $\mathbf{T} \in \{0,1\}^{n\times n}$ is a Toeplitz mask: $\mathbf{T}_{i,j} = \mathbf{1}_{|i-j|<L}$. A careful adaption of spectral clustering is then performed on $\mathbf{K}_L$ to retrieve the class information.

In technical terms, the present analysis derives the spectral distribution of $\mathbf{K}_L$ and analyzes the behavior of a few isolated eigenvalues (called *spikes*) which carry information (that is, indicators for the data classes) in their associated eigenvectors. Two new interesting behaviors are observed: unlike classical spectral clustering, due to the Toeplitz filter, the number of informative spikes can potentially grow very large even in the case of binary classification. In addition, the eigenvectors are strongly tainted (in a way "convolved") by the eigenvectors of the Toeplitz mask, which then requires some careful post-processing for classification. Our results particularly shed light on how the learning performance is altered by the dimension of the data and the size of the pipeline, thus providing an analysis of the performance versus cost trade-off of online learning.

In a nutshell, our main contributions may be listed as follows

- we derive the limiting eigenvalue distribution of $\mathbf{K}_L$ as $n, p, L \to +\infty$ for data arising from a Gaussian mixture model: $\mathbf{x}_i \sim \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{I}_p)$;
- for centered data drawn from a two-class mixture $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$, we show that a phase transition phenomenon occurs: depending on the signal power $\|\boldsymbol{\mu}\|$, some eigenvalues of $\mathbf{K}_L$ isolate and their eigenvectors carry information about the classes;
- we propose an algorithm to retrieve information from isolated eigenvectors, thus performing high-dimensional "online spectral clustering";
- simulations of online spectral clustering on Fashion-MNIST and BigGAN-generated images confirm the predicted good behavior of the algorithm and support our theoretical findings.

Proofs and simulations All proofs are deferred to the appendix. Python codes to reproduce simulations are available as supplementary material.

2. Online learning model and problem setting

2.1. General framework

Let $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix} \in \mathbb{R}^{p\times n}$ be a collection of n data samples of dimension p. They are noisy observations of K unknown classes whose centroids are $\begin{bmatrix}\boldsymbol{\mu}_1 & \dots & \boldsymbol{\mu}_K\end{bmatrix} \equiv \mathbf{M} \in \mathbb{R}^{p\times K}$. Also define the $n \times K$ binary matrix $\mathbf{J}$ such that $\mathbf{J}_{i,j} = 1$ if $\mathbf{x}_i$ belongs to class j and 0 otherwise.

We make the following assumptions.

Assumption 2.1. The rows of $\mathbf{J}$ are independent realizations of a multinomial distribution with one trial and K outcomes, i.e., the class of $\mathbf{x}_i$ does not depend on the class of $\{\mathbf{x}_j\}_{j\neq i}$.

Assumption 2.2 (Non-triviality condition)**.** $\mathbf{M}$ is uniformly bounded in spectral norm as $n, p \to +\infty$.

Assumption 2.3. The random matrix $\mathbf{X}$ can be decomposed into a deterministic signal matrix $\mathbf{P} = \mathbf{M}\mathbf{J}^\top$ and a random standard Gaussian noise matrix $\mathbf{Z}$ with independent entries[1]: $\mathbf{X} = \mathbf{P} + \mathbf{Z}$.

Remark 2.4. The non-triviality condition (assumption 2.2) places the work under scenarios of practical relevance, in the sense that the problem is asymptotically (as $n, p, L \to +\infty$) neither too easy nor too hard. The clustering error rate is therefore *not* expected to vanish asymptotically.

In the considered online setting, only the L previously seen data points are kept in memory. Thus, the element

[1]The "interpolation trick" from (Lytova & Pastur, 2009) enables to interpolate the results to non-Gaussian noise, but we keep the Gaussian assumption for simplicity of exposition here.

$\mathbf{K}_{i,j} = \frac{1}{p}\mathbf{x}_i^\top \mathbf{x}_j$ of the Gram kernel matrix can be computed only for $|i-j| < L$. This is represented by the pointwise application of a Toeplitz mask $\mathbf{T} = \left(\mathbf{1}_{|i-j|<L}\right)_{1\leqslant i,j\leqslant n}$ resulting in

$$\mathbf{K}_L = \frac{\mathbf{X}^\top \mathbf{X}}{p} \odot \mathbf{T} \quad \text{with} \quad \mathbf{T} = \begin{bmatrix} 1 & \dots & 1 & & 0 \\ \vdots & \ddots & & \ddots & \\ 1 & & \ddots & & 1 \\ & \ddots & & \ddots & \vdots \\ 0 & & 1 & \dots & 1 \end{bmatrix}.$$

As standard (offline) spectral clustering is "optimal"[2], we argue that spectral clustering on $\mathbf{K}_L$ ought to achieve good performance at least for not too small $(2L-1)/n$ ratios. Our technical goal is thus to first provide a description of the spectral behavior of $\mathbf{K}_L$ as n, p and L are large. To this end, we place ourselves under the regime $n, p, L \to +\infty$ with $p/n \to c \in]0, +\infty[$ and $(2L-1)/n \to \varepsilon \in]0, +\infty[$.

2.2. The circulant approximation

An important trick to derive our main result lies in the fact that the Toeplitz matrix $\mathbf{T}$ can be approximated (Gray, 2006) to some extent by its circulant "version" $\mathbf{C} = \left(\mathbf{1}_{|i-j|<L} + \mathbf{1}_{|i-j|>n-L}\right)_{1\leqslant i,j\leqslant n}$. Denoting $\{\tau_k\}_{0\leqslant k<n}$ and $\{\psi_k\}_{0\leqslant k<n}$ their respective eigenvalues (which depend on n and L), then for *fixed* L and any continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{k=0}^{n-1} |f(\psi_k) - f(\tau_k)| = 0.$$

Remark 2.5. Keep in mind that, in our case, n and L grow together at the same rate. Therefore, approximating $\mathbf{T}$ by $\mathbf{C}$ is only reasonable if ε is sufficiently small.

The core advantage of $\mathbf{C}$ is that, unlike $\mathbf{T}$, its eigendecomposition is well-known:

$$\mathbf{C} = \mathbf{F}\mathbf{\Psi}\mathbf{F}^*$$

where $\mathbf{F}$ is the $n\times n$ Fourier matrix ($\mathbf{F}_{i,j} = \frac{1}{\sqrt{n}} e^{-2\mathrm{i}\pi\frac{ij}{n}}$) and $\mathbf{\Psi} = \operatorname{diag}\left(\psi_k\right)_{0\leqslant k<n}$ is the diagonal matrix of eigenvalues. The latter are a sampling of the Dirichlet kernel:

$$\psi_k = \nu_L\left(\frac{2k\pi}{n}\right) \quad \text{with} \quad \nu_L(x) = \frac{\sin((2L-1)\frac{x}{2})}{\sin(\frac{x}{2})}.$$

In Figure 1 are superimposed to the graph of ν_L the eigenvalues of $\mathbf{C}$ and[3] $\mathbf{T}$. The τ_k's roughly follow the graph of ν_L, as if they were noisy versions of the ψ_k's.

[2] In that it performs better than random guess as soon as theoretically possible (Onatski et al., 2013).

[3] Although there is a natural order for the eigenvalues of $\mathbf{C}$ given by $\psi_k = \nu_L(\frac{2k\pi}{n})$, we use a small trick to get the corresponding order for the eigenvalues of $\mathbf{T}$: after numerically computing them in descending order, we apply the same permutation that maps the eigenvalues of $\mathbf{C}$ in descending order to $(\psi_0, \dots, \psi_{n-1})$. This yields the corresponding $(\tau_0, \dots, \tau_{n-1})$.

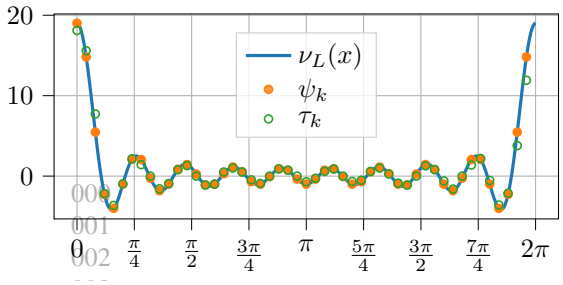


Figure 1. Graph of ν_L on $[0, 2\pi[$ (one period) with a plot of $\psi_k = \nu_L(\frac{2k\pi}{n})$ and τ_k for $0 \leqslant k < n$ (the eigenvalues of $\mathbf{C}$ and $\mathbf{T}$ respectively). **Experimental setting**: $n = 50$, $L = 10$.

3. Main results

Following standard methods in random matrix theory (Couillet & Liao, 2021), the large dimensional spectral behavior of $\mathbf{K}_L$ is accessible through an analysis of the resolvent matrix

$$\mathbf{Q}(z) = (\mathbf{K}_L - z\mathbf{I}_n)^{-1}$$

defined for all $z \in \mathbb{C} \setminus \operatorname{Sp}(\mathbf{K}_L)$, where $\operatorname{Sp}(\mathbf{K}_L)$ denotes the set of eigenvalues of $\mathbf{K}_L$. Notably, the Stieltjes transform of the empirical spectral measure $\mu_n = \frac{1}{n}\sum_{\lambda\in\operatorname{Sp}(\mathbf{K}_L)} \delta_\lambda$ of $\mathbf{K}_L$ (from which the spectral measure itself can be recovered) is the normalized trace of its resolvent:

$$m_n(z) \equiv \int_{\mathbb{R}} \frac{\mu_n(\mathrm{d}t)}{t-z} = \frac{1}{n}\operatorname{tr}\mathbf{Q}(z).$$

The resolvent also encapsulates information about the eigenvectors of $\mathbf{K}_L$: given a closed positively-oriented complex contour Γ circling around an eigenvalue λ of $\mathbf{K}_L$ and leaving all the other eigenvalues outside, $-\frac{1}{2\mathrm{i}\pi}\oint_\Gamma \mathbf{Q}(z)\,\mathrm{d}z = \mathbf{u}\mathbf{u}^*$, where $\mathbf{u}$ is a unit eigenvector associated to λ.[4]

3.1. Large dimensional spectral behavior

Our main theorem provides a deterministic equivalent of the resolvent when the Toeplitz mask $\mathbf{T}$ is approximated by its circulant version $\mathbf{C}$, i.e., $\tilde{\mathbf{Q}}(z) = \left(\tilde{\mathbf{K}}_L - z\mathbf{I}_n\right)^{-1}$ with $\tilde{\mathbf{K}}_L = \frac{\mathbf{X}^\top\mathbf{X}}{p} \odot \mathbf{C}$. Namely, we find a deterministic matrix $\bar{\mathbf{Q}}(z)$ such that, for any sequence of deterministic matrices $\mathbf{A}_n \in \mathbb{R}^{n\times n}$ and vectors $\mathbf{a}_n, \mathbf{b}_n \in \mathbb{R}^n$ of unit norm (spectral norm and Euclidean norm respectively), $\frac{1}{n}\operatorname{tr}\mathbf{A}_n(\tilde{\mathbf{Q}}(z) - \bar{\mathbf{Q}}(z)) \to 0$ and $\mathbf{a}_n^\top(\tilde{\mathbf{Q}}(z) - \bar{\mathbf{Q}}(z))\mathbf{b}_n \to 0$ almost surely as $n, p, L \to +\infty$. This will be simply denoted $\tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\mathbf{Q}}(z)$.

[4] This is only true if λ has multiplicity 1. In the general case, the integral equals the projection matrix on the eigenspace associated to λ.

Theorem 3.1 (Deterministic equivalent of $\tilde{\mathbf{Q}}$). *Let $z \in \mathbb{C} \setminus \limsup_{n,p,L\to+\infty} \mathrm{Sp}(\tilde{\mathbf{K}}_L)$ and define $m(\cdot)$ as the unique Stieltjes transform solution to*

$$1 + zm(z) = \frac{1}{n}\sum_{k=0}^{n-1} \frac{m(z)\psi_k}{1 + \frac{m(z)}{p}\psi_k}. \tag{1}$$

Under assumptions 2.1 – 2.3, if $\left|\frac{2L-1}{p} m(z)\right| < 1$, then

$$\tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\mathbf{Q}}(z) \equiv m(z)\left(\mathbf{I}_n + \frac{\mathbf{P}^\top\mathbf{P}}{p} \odot \mathbf{F}\mathbf{\Lambda}\mathbf{F}^*\right)^{-1}$$

where $\mathbf{\Lambda} = m(z)\mathbf{\Psi}\left(\mathbf{I}_n + \frac{m(z)}{p}\mathbf{\Psi}\right)^{-1}$ is a diagonal matrix, thus $\mathbf{F}\mathbf{\Lambda}\mathbf{F}^*$ is circulant.

Proof. See appendix B. □

A first observation from theorem 3.1 is that $\bar{\mathbf{Q}}(z)$ is the inverse of a perturbation of the identity which is *not* low rank. This strikingly differs from standard spiked random matrix models (Baik & Silverstein, 2006; Benaych-Georges & Nadakuditi, 2011) where a low-rank perturbation of the identity in the "population" matrix (here $\mathbf{P}$) usually results in the presence of only a few isolated eigenvalues in the "sample" matrix (here $\tilde{\mathbf{K}}_L$). This being said, here, in standard settings, most eigenvalues of $\frac{\mathbf{P}^\top\mathbf{P}}{p} \odot \mathbf{F}\mathbf{\Lambda}\mathbf{F}^*$ are small enough for only a few number of corresponding isolated eigenvalues in the spectrum of $\tilde{\mathbf{K}}_L$ to appear.

Yet, as a result of this full-rank perturbation of the identity property, it may be unclear whether $m(\cdot)$ is the Stieltjes transform of the limiting spectral distribution of $\tilde{\mathbf{K}}_L$ or not (this is important to ensure that isolated eigenvalues are truly informative). This issue is handled along the proof of theorem 3.1 (see proposition B.4) where it is shown that $\frac{1}{n}\operatorname{tr}\tilde{\mathbf{Q}}(z) \to m(z)$ almost surely as $n, p, L \to +\infty$, thereby proving the (almost sure) convergence of the empirical spectral measure of $\tilde{\mathbf{K}}_L$ to a measure μ which is the inverse Stieltjes transform[5] of m.

Remark 3.2 (Link with (Marčenko & Pastur, 1967)). In the particular case $n = 2L - 1$, the mask becomes $\mathbf{C} = \mathbf{1}_n\mathbf{1}_n^\top$ and $\tilde{\mathbf{K}}_L = \mathbf{K}$. And, since $\psi_0 = n$ and $\psi_k = 0$ for $1 \leqslant k < n$, equation 1 becomes

$$zm^2(z) + (cz - c + 1)\, m(z) + c = 0$$

which is the canonical equation defining the Stieltjes transform of the Marčenko-Pastur distribution. The closer ε is to 1, the closer to the Marčenko-Pastur distribution is the limiting spectral distribution of $\tilde{\mathbf{K}}_L$.

[5] If μ has a density $d(x)$ at $x \in \mathbb{R}$, then $d(x) = \lim_{y\downarrow 0} \frac{1}{\pi}\Im m(x + \mathrm{i}y)$.

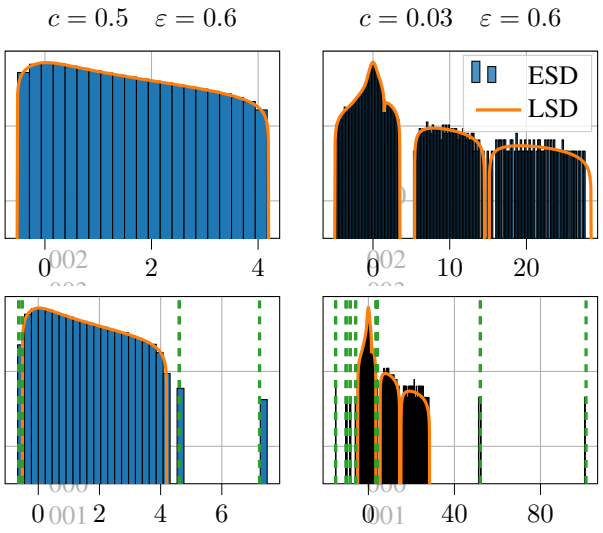


Figure 2. Empirical spectral distribution (ESD) and limiting spectral distribution (LSD) of $\tilde{\mathbf{K}}_L$. **The y-axis is in log scale**. **Top**: noise only, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. **Bottom**: two-class mixture, $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ with $\|\boldsymbol{\mu}\| = 2$. Green dashed lines are the asymptotic positions of the spikes $\bar{\xi}_k$. **Experimental setting**: $n = 2\,500$, $L = 750$ and $p = 1\,250$ (**left**) or $p = 75$ (**right**).

In practice, rather than computing $m(z)$ directly from equation 1, it is easier to solve numerically the following fixed-point equation in η_0

$$\eta_0 = \frac{1}{n}\sum_{k=0}^{n-1} \frac{\psi_k^2}{p\,(1 - z - \eta_0) + \psi_k}$$

and deduce $m(z) = \frac{1}{1-z-\eta_0}$.

Figure 2 displays, in log scale, the empirical spectral distribution of $\tilde{\mathbf{K}}_L$ under two different settings with its limiting spectral distribution computed by inverting the Stieltjes transform given by theorem 3.1. Two kinds of data are presented: noise-only, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, (top row) and a two-class mixture $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ (bottom row). Notice how the shape of the distribution on the left column resembles the Marčenko-Pastur one (yet, some eigenvalues are negative here) while the second distribution has a completely different shape (there even are several bulks) for the same value of ε. This reveals that the parameter c also affects the closeness of the limiting spectral distribution to the Marčenko-Pastur one. Also note that, under the two-class mixture setting, more than one isolated eigenvalue pop out of the limiting support. It now remains to give a close look to their associated eigenvectors to understand how to exploit the latter in a spectral clustering perspective.

3.2. Phase transition and spike behavior

In this section, we thus focus back on our original classification objective. We consider two classes $\mathcal{C}^{\pm}$ whose centroids are $\pm\boldsymbol{\mu}$, i.e.[6], $\mathbf{P} = \boldsymbol{\mu}\mathbf{j}^\top$ with $\mathbf{j}_i = +1$ if $\mathbf{x}_i \in \mathcal{C}^+$ and $\mathbf{j}_i = -1$ if $\mathbf{x}_i \in \mathcal{C}^-$. This corresponds to a two-class mixture with globally centered data.

Because of the rank-one structure, using the relation $\mathbf{M} \odot \mathbf{a}\mathbf{b}^* = [\operatorname{diag} \mathbf{a}]\, \mathbf{M}\, [\operatorname{diag} \mathbf{b}]^*$, the deterministic equivalent of the resolvent has a much simpler expression:

$$\bar{\mathbf{Q}}(z) = m(z)\, [\mathbf{D}_\mathbf{j}\mathbf{F}] \left(\mathbf{I}_n + \frac{\|\boldsymbol{\mu}\|^2}{p} \boldsymbol{\Lambda} \right)^{-1} [\mathbf{D}_\mathbf{j}\mathbf{F}]^*$$

where $\mathbf{D}_\mathbf{j} = \operatorname{diag} \mathbf{j}$ is the diagonal matrix induced by vector $\mathbf{j}$. Now, $\bar{\mathbf{Q}}(z)$ no longer involves a Hadamard product and we already have its eigendecomposition since $\mathbf{I}_n + \frac{\|\boldsymbol{\mu}\|^2}{p}\boldsymbol{\Lambda}$ is diagonal and $\mathbf{D}_\mathbf{j}\mathbf{F}$ is unitary. Note that the columns of $\mathbf{D}_\mathbf{j}\mathbf{F}$ are simply the vectors of the Fourier basis with their sign switched at coordinates i such that $\mathbf{x}_i \in \mathcal{C}^-$.

With a deeper analysis of the resolvent $\bar{\mathbf{Q}}(z)$, the following theorem provides the position of the isolated eigenvalues and the shape of their associated eigenvectors.

Theorem 3.3 (Phase transition, isolated eigenvalues and eigenvector alignments.)**.** *Given an integer $0 \leqslant k < n$, let*

$$\bar{\xi}_k = \frac{\|\boldsymbol{\mu}\|^2 + 1}{p} \psi_k \left(1 + \frac{p}{n} \sum_{l=0}^{n-1} \frac{\psi_l}{\left(\|\boldsymbol{\mu}\|^2 + 1 \right) \psi_k - \psi_l} \right)$$

and

$$\bar{\zeta}_k = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + 1} \left(1 - \frac{p}{n} \sum_{l=0}^{n-1} \left[\frac{\psi_l}{\left(\|\boldsymbol{\mu}\|^2 + 1 \right) \psi_k - \psi_l} \right]^2 \right).$$

The following propositions are equivalent.

1. $\psi_k \neq 0$ *and*[7] $\bar{\xi}_k \notin \operatorname{supp} \mu$.
2. $\bar{\zeta}_k > 0$.
3. $\bar{\xi}_k$ *is a singular point of* $\bar{\mathbf{Q}}(z)$*, i.e., the almost sure asymptotic position of an isolated eigenvalue of* $\tilde{\mathbf{K}}_L$.

Then, in this case, $\mathbf{U}_k = \left[\mathbf{u}_l\right]_{\substack{\psi_k = \psi_l \\ 0 \leqslant l < n}}$ *is an isometric matrix gathering all the eigenvectors of* $\tilde{\mathbf{K}}_L$ *whose associated eigenvalues converge to* $\bar{\xi}_k$ *and*

$$\mathbf{U}_k \mathbf{U}_k^* \leftrightarrow \bar{\zeta}_k\, [\mathbf{D}_\mathbf{j}\mathbf{F}]\, \mathcal{D}_k\, [\mathbf{D}_\mathbf{j}\mathbf{F}]^*$$

where $\mathbf{D}_\mathbf{j} = \operatorname{diag} \mathbf{j}$ *and* $\mathcal{D}_k = \operatorname{diag}\left(\mathbf{1}_{\psi_k = \psi_l} \right)_{0 \leqslant l < n}$.

[6] Consistently with the previous setting, $\mathbf{M} = \begin{bmatrix} +\boldsymbol{\mu} & -\boldsymbol{\mu} \end{bmatrix}$ and $\mathbf{J}_{i,\cdot} = \begin{bmatrix} \mathbf{1}_{\mathbf{x}_i \in \mathcal{C}^+} & \mathbf{1}_{\mathbf{x}_i \in \mathcal{C}^-} \end{bmatrix}$.

[7] Since μ is the limiting spectral distribution of $\tilde{\mathbf{K}}_L$, $\operatorname{supp} \mu = \limsup_{n,p,L \to +\infty} \operatorname{Sp}(\tilde{\mathbf{K}}_L)$.

Proof. See appendix C. □

To better understand this theorem, recall that, in theorem 3.1, we predicted the presence of a few isolated eigenvalues in the spectrum of $\tilde{\mathbf{K}}_L$. Theorem 3.3 details this assertion by specifying the number of spikes ($\# \left\{ \bar{\zeta}_k > 0 \right\}$) and their position $\bar{\xi}_k$. The quantity $\bar{\zeta}_k$ can really be seen as an "indicator of spike" as it tells whether an isolated eigenvalue exists and, if it does, the closer $\bar{\zeta}_k$ is to 1, the better is the "quality" of the information carried in the corresponding eigenvector, i.e., the greater is the signal-to-noise ratio (see Figure 3).

Another difference with classical spiked random matrix models is that each asymptotic spike $\bar{\xi}_k$, which has the same multiplicity as the population spike ψ_k, is rarely simple[8]. However, for finite values of n, p, L, the corresponding eigenvalues of $\tilde{\mathbf{K}}_L$ are not necessarily degenerate (with probability one, they are not), but they have the same limit[9].

One also notices from theorem 3.3 that the number of isolated eigenvalues could potentially grow very large as $\|\boldsymbol{\mu}\|$ increases. Indeed, the value of $\|\boldsymbol{\mu}\|$ at which $\bar{\zeta}_k$ changes sign (i.e, when one or more eigenvalues isolate from the bulk around $\bar{\xi}_k$ during the *phase transition*) is given by

$$1 - \frac{p}{n} \sum_{l=0}^{n-1} \left[\frac{\psi_l}{\left(\|\boldsymbol{\mu}\|^2 + 1 \right) \psi_k - \psi_l} \right]^2 = 0.$$

Therefore, potentially any eigenvalue could leave the bulk, but this is prevented by the non-triviality condition (assumption 2.2): $\|\boldsymbol{\mu}\| = \mathcal{O}_{n,p,L \to +\infty}(1)$. Moreover, since most ψ_k's are small (see Figure 1), the corresponding $\bar{\xi}_k$'s fall into the bulk and there are only a few spikes visible in practice. Yet, it is common to see negative isolated eigenvalues (see Figure 2). Indeed, since ψ_k can be negative, there can be spikes on *both sides* of the spectrum.

When positive, the quantity $\bar{\zeta}_k$ is the asymptotic alignment between the empirical eigenvector $\mathbf{u}_k$ and the corresponding information vector $\mathbf{v}_k = [\mathbf{D}_\mathbf{j}\mathbf{F}]_{\cdot,k} = \mathbf{F}_{\cdot,k} \odot \mathbf{j}$, i.e.,

$$\left| \mathbf{u}_k^* \mathbf{v}_k \right|^2 \xrightarrow[n,p,L \to +\infty]{\text{a.s.}} \bar{\zeta}_k.$$

Thus, $\bar{\zeta}_k$ measures the quality of the empirical eigenvector $\mathbf{u}_k$. Said differently, $\mathbf{u}_k$ is a noisy version of a vector $\mathbf{v}_k = \mathbf{F}_{\cdot,k} \odot \mathbf{j}$ and the noise level is indicated by $0 \leqslant 1 - \bar{\zeta}_k < 1$.

Figure 3 displays the value of $\bar{\zeta}_k^+ = \max(\zeta_k, 0)$ as a function of $\|\boldsymbol{\mu}\|$ for the setting corresponding to the bottom right part of Figure 2. The empirical alignment of the dominant eigenvector $\mathbf{u}_0$ with $\mathbf{v}_0 = \frac{1}{\sqrt{n}}\mathbf{j}$ fits perfectly with the curve

[8] ψ_0, and $\psi_{n/2}$ when n is even, are the only simple eigenvalues of $\mathbf{C}$.

[9] In this case, $\bar{\xi}_k = \bar{\xi}_l$ for all l such that $\psi_k = \psi_l$.

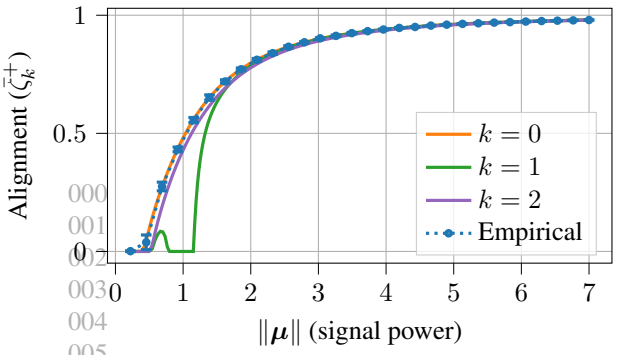


Figure 3. Asymptotic alignment $\bar{\zeta}_k^+$ versus $\|\boldsymbol{\mu}\|$ for three values of k. The empirical alignment is computed as the mean of $|\mathbf{u}_0^* \mathbf{v}_0|^2$ on 10 realizations (error bars indicate the standard deviation). **Experimental setting**: $n = 2\,500$, $p = 75$, $L = 750$.

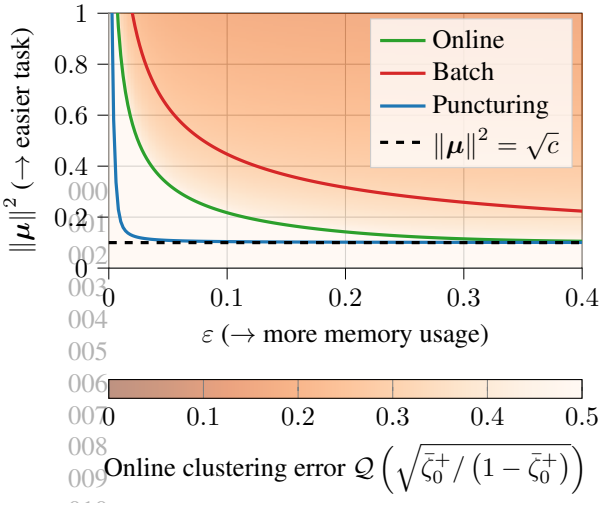


Figure 4. Phase transition position ($\|\boldsymbol{\mu}\|^2$) of the dominant eigenvector of the kernel matrix against the sparsity parameter (ε) with $n/p = 100$. Classification with a given method is only possible above the corresponding curve. The black dashed line is the optimal phase transition (with all the data available). **Green**: online kernel spectral clustering (circulant mask). **Red**: regular kernel spectral clustering with $L = \frac{n\varepsilon+1}{2}$ points. **Blue**: punctured (offline) kernel spectral clustering (Bernoulli mask).

of $\bar{\zeta}_0^+$ predicted by theorem 3.3. Moreover, notice the interesting fact that $\bar{\zeta}_1$ has several phase transitions: as $\|\boldsymbol{\mu}\|$ grows, it appears once, then disappears and appears once again! This is due to the limiting spectral distribution having several bulks under this setting (see Figure 2). The first time this spike appears, it is located between two bulks. It then goes through the rightmost bulk (so it is no longer an isolated eigenvalue thus $\bar{\zeta}_1 \leqslant 0$), and finally goes out on the right edge of the distribution.

This last result may sound awkward and possibly testify of the suboptimality of our approach (when the signal-to-noise ratio increases, the information attached to some eigenvectors vanishes). This conclusion is not so immediate though, as the classification information is still contained within other eigenvectors which, as $\|\boldsymbol{\mu}\|$ increases, *do* carry increasingly clearer information.

3.3. Discussion on the circulant approximation

The approximation of the Toeplitz mask $\mathbf{T}$ by the circulant mask $\mathbf{C}$ used in the previous theorems 3.1 and 3.3 can be seen as a way to remove undesired edge effects, whose size is governed by L.[10] If L is chosen small compared to n, edge effects are expected to be negligible and the previous results can plausibly be extended to the original setting.

To adapt the previous results from $\mathbf{C}$ to $\mathbf{T}$, one only needs to change the eigenvalues and eigenvectors, i.e., replace ψ_k by τ_k, the eigenvalues of $\mathbf{T}$, and replace $\mathbf{F}$ by $\begin{bmatrix}\mathbf{g}_0 & \dots & \mathbf{g}_{n-1}\end{bmatrix} \equiv \mathbf{G}$, an eigenbasis of $\mathbf{T}$.

Very precise predictions on the original model can be made with these simple changes. Comparisons between these and reality are provided in appendix D.

[10] Remove the first and last $L-1$ rows and columns of $\mathbf{C}$ and $\mathbf{T}$ and we are left with the same two Toeplitz matrices.

4. Online spectral clustering of large data

The previous results find direct applications to the online clustering of high-dimensional data.

4.1. Performance vs. cost trade-off in online learning

The phase transition position provided by theorem 3.3 lets us determine under which setting classification is possible or not. Consider the dominant eigenvector $\mathbf{u}_0$. If $\bar{\zeta}_0 \leqslant 0$ then no eigenvalue isolates from the bulk and classification cannot be performed. After the phase transition, $\bar{\zeta}_0 > 0$ and the closer it is to 1, the closer $\mathbf{u}_0$ is to $\mathbf{v}_0 = \frac{1}{\sqrt{n}}\mathbf{j}$. The fluctuations of the entries of $\mathbf{u}_0$ happen to be asymptotically Gaussian and pairwise independent (Kadavankandy & Couillet, 2019) with, for equal-size classes, mean $\pm\sqrt{\bar{\zeta}_0/n}$ and variance $\left(1 - \bar{\zeta}_0\right)/n$. Thus, the asymptotic clustering error is given by $\mathcal{Q}\left(\sqrt{\bar{\zeta}_0^+ / (1 - \bar{\zeta}_0^+)}\right)$, where $\mathcal{Q}$ is the Gaussian tail function: $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-\frac{t^2}{2}}\, \mathrm{d}t$.

Figure 4 shows the phase transition position $\|\boldsymbol{\mu}\|^2$ as a function of $\varepsilon = \frac{2L-1}{n}$, with the asymptotic clustering error of online kernel spectral clustering, when $n/p = 100$. For comparison, the phase transition curves of the following two methods are also represented:

- *Batch clustering*, i.e., standard $L \times L$ kernel spectral clustering with the L data points available in memory.

- *Punctured kernel spectral clustering* (Zarrouk et al., 2020; Couillet et al., 2021), i.e., *offline* clustering performed with a sparsified kernel matrix $\mathbf{K}_\varepsilon = \frac{\mathbf{X}^\top \mathbf{X}}{p} \odot \mathbf{B}$, where $\mathbf{B}_{i,j} = \mathbf{B}_{j,i} \sim \text{Bern}(\varepsilon)$ and $\mathbf{B}_{i,i} = 1$.[11]

As ε grows, the phase transition position of online spectral clustering rapidly reaches the optimal threshold $\|\boldsymbol{\mu}\|^2 = \sqrt{c}$ under which no information can be recovered (regardless of the method used and the data available). Moreover, the clustering error decreases to 0 as $\|\boldsymbol{\mu}\|$ increases (the signal is more powerful). A good compromise between memory usage and performance appears to be $0.1 \lessapprox \varepsilon \lessapprox 0.2$, i.e., $\frac{n}{20} \lessapprox L \lessapprox \frac{n}{10}$ as it keeps L (i.e., the memory usage) small while not impairing much the performance.

Our method performs better (i.e., the phase transition occurs earlier) than the naive setting performing standard clustering on batches of L points available in memory. It is also able to classify the n previous points (and not only the L previous ones) at any time, *although the corresponding data points have left memory*. It is instructive to see that, under the same sparsity level of the kernel matrix, the puncturing method performs better. Yet, this requires the access to n data points to compute $\mathbf{K}_\varepsilon$, which is not possible in online learning.

4.2. Online clustering algorithm

Before diving into the simulations, we detail a clustering algorithm based on our previous results. We here use the banded version of the kernel matrix: $\mathbf{K}_L = \frac{\mathbf{X}^\top \mathbf{X}}{p} \odot \mathbf{T}$ (the circulant mask was only useful for theoretical considerations) and recall the notation of the eigenbasis of $\mathbf{T}$: $\begin{bmatrix}\mathbf{g}_0 & \cdots & \mathbf{g}_{n-1}\end{bmatrix} \equiv \mathbf{G}$.

We consider a data stream of length T (possibly infinite). At each time step, a new vector $\mathbf{x}_t$ arrives and the kernel matrix is updated: $\left[\mathbf{K}_L^{(t)}\right]_{i,j} = \frac{1}{p}\mathbf{x}_{t-n+i}^\top \mathbf{x}_{t-n+j} \mathbf{1}_{|i-j|<L}$.

Remark 4.1 (Choice of n and eigenvector localization). It is important to emphasize that n is *not* the length of the data stream (given by the newly-introduced parameter $T \geqslant n$). As $\mathbf{K}_L$ has size $n \times n$, one can "only" classify the last n points of the stream, even when discarded from the length-L memory (older points are no longer classified though).

The parameter n is left for the user to choose, accounting for L and our previous considerations on the performance (Figure 4) and memory limitations: $\mathcal{O}(Lp + Ln)$ space is needed to store the data *and* the kernel matrix. Moreover, as the graph associated to $\mathbf{K}_L$ becomes sparser ($n \gg L$), its eigenvectors tend to localize (Hata & Nakao, 2017), making classification more challenging.

[11] This of course is not doable with a memory bank of size L but the comparison is interesting as the number of entries in $\mathbf{K}_\varepsilon$ and $\mathbf{K}_L$ is the same.

As per standard kernel spectral clustering, we use the dominant eigenvectors of $\mathbf{K}_L^{(t)}$ to estimate the classes. The last n points of the stream are classified at each time step so each point is classified n times. Then, the final class estimate can be chosen by a majority vote. However, because of the particular shape of the eigenvectors caused by the Toeplitz mask[12] (see Figure 6), standard clustering algorithms such as k-means perform poorly on such spectral embeddings. Therefore, we propose a new way to cluster the data.

Remark 4.2. The eigenvectors of $\mathbf{K}_L^{(t)}$ can be quickly computed at a low cost with a warm start of the power iteration algorithm from the previously computed eigenvectors of $\mathbf{K}_L^{(t-1)}$.

In a binary setting with globally centered data, classification can be performed using only the dominant eigenvector $\mathbf{u}_0^{(t)}$ of $\mathbf{K}_L^{(t)}$. Relying on the alignment of $\mathbf{u}_0^{(t)}$ with $\mathbf{v}_0^{(t)} = \mathbf{g}_0 \odot \mathbf{j}^{(t)}$ (theorem 3.3) and the fact that the coordinates of $\mathbf{g}_0$ have constant sign, the class of $\mathbf{x}_{t-n+i}$ can be estimated from the sign of $\left[\mathbf{u}_0^{(t)}\right]_i$. This online clustering procedure is summarized in Algorithm 1.

Algorithm 1 Online kernel spectral clustering (binary)

Output: class estimators $\left\{\hat{\mathcal{C}}_t^+, \hat{\mathcal{C}}_t^-\right\}_{n \leqslant t \leqslant T}$.
for $t = 1$ to T **do**
  Get a new point $\mathbf{x}_t$ into the pipeline.
  Compute $\mathbf{x}_t^* \mathbf{x}_{t-l}$ for $l = 0$ to $L-1$.
  Update $\mathbf{K}_L^{(t-1)}$ into $\mathbf{K}_L^{(t)}$.
  if $t \geqslant n$ **then**
    $\mathbf{u}_0^{(t)} \leftarrow \text{PowerIteration}(\mathbf{K}_L^{(t)}, \mathbf{u}_0^{(t-1)})$.
    $\hat{\mathcal{C}}_t^\pm \leftarrow \left\{\mathbf{x}_{t-n+i} \mid \left[\mathbf{u}_0^{(t)}\right]_i \gtrless 0\right\}$.
  end if
end for

In appendix E, we also propose a (more complex) online spectral clustering algorithm capable of handling K-class mixtures and test it on Fashion-MNIST images.

Note that this algorithm can easily be adapted to a setting where more than one vector $\mathbf{x}_t$ arrives at each time step (and this quantity does not need to be constant in time).

4.3. Simulations on real-world images

We illustrate our findings with two applications on image classification tasks. We first apply Algorithm 1 on globally centered and scaled VGG-features (Simonyan & Zisserman, 2015) of randomly BigGAN-generated images (Brock et al., 2019) of `tabby` cats and `collie` dogs (see Figure 5). The vectors thus generated have dimension $p = 4\,096$

[12] The dominant eigenvector of $\mathbf{T}$, for example, is not constant, contrary to the first Fourier mode with the circulant mask.

Figure 5. Examples of BigGAN-generated images: `collie` (**top**) and `tabby` (**bottom**).

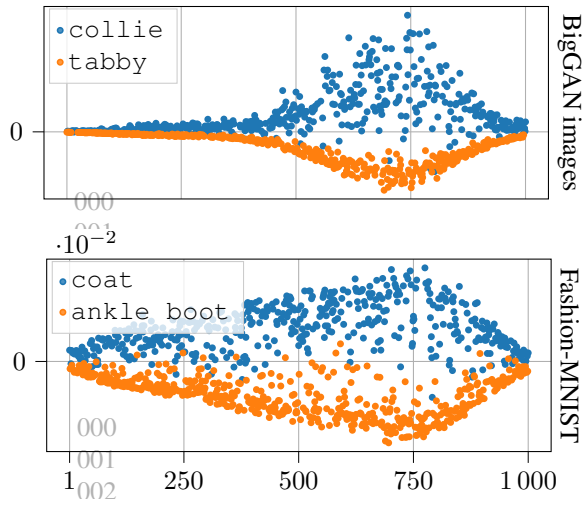


Figure 6. Dominant eigenvector of $\mathbf{K}_L^{(t)}$ with BigGAN-generated images (**top**) and Fashion-MNIST images (**bottom**). **Experimental setting**: $T = 20\,000$, $n = 1\,000$, $p = 4\,096$, $L = 100$ (BigGAN images) and $T = 14\,000$, $n = 1\,000$, $p = 784$, $L = 100$ (Fashion-MNIST).

and simulate a stream of length $T = 20\,000$ with evenly likely cats and dogs. In addition, our algorithm is applied to a stream made of $T = 14\,000$ centered raw-images from the Fashion-MNIST dataset (Xiao et al., 2017). Their dimension is $p = 784$ and we want to discriminate `coat` versus `ankle boot` in an online fashion. In both cases, we choose $n = 1\,000$ and $L = 100$.

Figure 6 shows the shape of the dominant eigenvector $\mathbf{u}_0^{(t)}$ at a given time during the execution of the algorithm. We clearly see a separation between the classes. For both settings, Figure 7 depicts the mean clustering error at $t_0 + \Delta t$ of a data point seen at t_0, as well as the overall classification error obtained after a majority vote (green dashed line), to be compared with the classification error obtained with a $T \times T$ *offline* kernel spectral clustering[13] (black dotted line). The mean classification error remains constant with Δt, thus showing that our algorithm does not lose any discriminative

[13]For which optimality results are known.

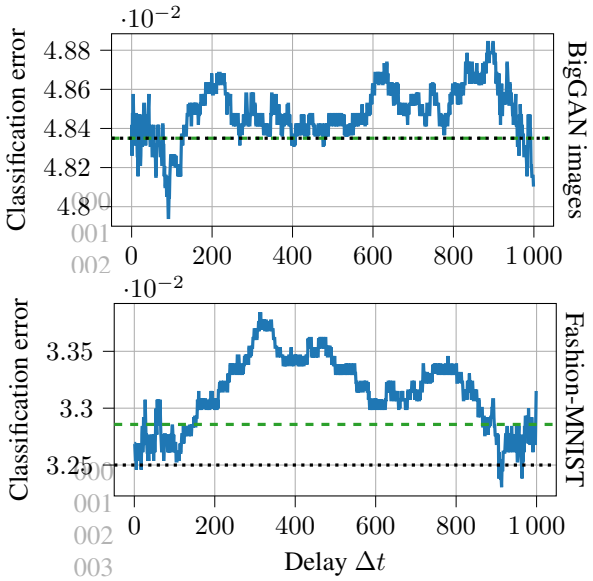


Figure 7. Classification error against delay Δt on BigGAN-generated images (**top**) and Fashion-MNIST images (**bottom**). This is the mean classification error at time $t_0 + \Delta t$ of a point arrived at t_0. The green dashed line indicate the overall classification error when the class is chosen by a majority vote. The black dotted line is the classification error obtained with a $T \times T$ offline kernel spectral clustering. **Experimental setting**: as in Figure 6.

power between t_0 and $t_0 + n - 1$. The classification performances of our algorithm are very close to those of the full spectral clustering but require much less memory resources: $\mathcal{O}(Lp + Ln)$ against $\mathcal{O}(Tp + T^2)$ space for the storage of the data and the kernel matrix.

5. Concluding remarks

Leveraging tools from random matrix theory, the article shows that, under limited memory resources, near-optimal performances on high-dimensional data can be achieved using an online kernel spectral clustering algorithm. By means of a thorough asymptotic analysis, we specify the optimal performances achievable when learning on a data stream, which we exploit to propose a novel efficient clustering algorithm adapted to memory-limited systems.

The article not only introduces a novel algorithm for online classification, but also paves the path towards the question of large-dimensional learning in data streaming with theoretical guarantees. Still, here we miss an information-theoretic result of optimality for the proposed approach (which exists in the unbanded case), a key direction we currently investigate.

References

Ackermann, M. R., Märtens, M., Raupach, C., Swierkot, K., Lammersen, C., and Sohler, C. StreamKM++: A clustering algorithm for data streams. *ACM Journal of Experimental Algorithmics*, 17:2.4:2.1–2.4:2.30, May 2012. ISSN 1084-6654. doi: 10.1145/2133803.2184450. URL https://doi.org/10.1145/2133803.2184450.

Aggarwal, C. C., Yu, P. S., Han, J., and Wang, J. - A Framework for Clustering Evolving Data Streams. In Freytag, J.-C., Lockemann, P., Abiteboul, S., Carey, M., Selinger, P., and Heuer, A. (eds.), *Proceedings 2003 VLDB Conference*, pp. 81–92. Morgan Kaufmann, San Francisco, January 2003. ISBN 978-0-12-722442-8. doi: 10.1016/B978-012722442-8/50016-1. URL https://www.sciencedirect.com/science/article/pii/B9780127224428500161.

Aggarwal, C. C., Han, J., Wang, J., and Yu, P. S. - A Framework for Projected Clustering of High Dimensional Data Streams. In Nascimento, M. A., Özsu, M. T., Kossmann, D., Miller, R. J., Blakeley, J. A., and Schiefer, B. (eds.), *Proceedings 2004 VLDB Conference*, pp. 852–863. Morgan Kaufmann, St Louis, January 2004. ISBN 978-0-12-088469-8. doi: 10.1016/B978-012088469-8.50075-9. URL https://www.sciencedirect.com/science/article/pii/B9780120884698500759.

Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Baik, J. and Silverstein, J. W. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, July 2006. ISSN 0047-259X. doi: 10.1016/j.jmva.2005.08.003. URL https://www.sciencedirect.com/science/article/pii/S0047259X0500134X.

Benaych-Georges, F. and Nadakuditi, R. R. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, May 2011. ISSN 0001-8708. doi: 10.1016/j.aim.2011.02.007. URL https://www.sciencedirect.com/science/article/pii/S0001870811000570.

Brock, A., Donahue, J., and Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv:1809.11096 [cs, stat]*, February 2019. URL http://arxiv.org/abs/1809.11096. arXiv: 1809.11096.

Cheng, X. and Singer, A. The Spectrum of Random Inner-product Kernel Matrices. *arXiv:1202.3155 [math]*, March 2012. URL http://arxiv.org/abs/1202.3155. arXiv: 1202.3155.

Couillet, R. and Benaych-Georges, F. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016. doi: 10.1214/16-EJS1144. URL https://hal.archives-ouvertes.fr/hal-01215343. Publisher: Shaker Heights, OH : Institute of Mathematical Statistics.

Couillet, R. and Liao, Z. *Random Matrix Methods for Machine Learning: When Theory meets Applications*. 2021.

Couillet, R., Chatelain, F., and Le Bihan, N. Two-way kernel matrix puncturing: towards resource-efficient PCA and spectral clustering. *arXiv:2102.12293 [cs, stat]*, May 2021. URL http://arxiv.org/abs/2102.12293. arXiv: 2102.12293.

Dhanjal, C., Gaudel, R., and Clémençon, S. Efficient Eigen-updating for Spectral Graph Clustering. *arXiv:1301.1318 [stat]*, January 2014. URL http://arxiv.org/abs/1301.1318. arXiv: 1301.1318.

El Karoui, N. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, February 2010. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS648. URL https://projecteuclid.org/journals/annals-of-statistics/volume-38/issue-1/The-spectrum-of-kernel-random-matrices/10.1214/08-AOS648.full. Publisher: Institute of Mathematical Statistics.

Ester, M., Kriegel, H., Sander, J., and Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, 1996.

Fritzke, B. A Growing Neural Gas Network Learns Topologies. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995. URL https://papers.nips.cc/paper/1994/hash/d56b9fc4b0f1be8871f5e1c40c0067e7-Abstract.html.

Ghesmoune, M., Lebbah, M., and Azzag, H. State-of-the-art on clustering data streams. *Big Data Analytics*, 1(1):13, December 2016. ISSN 2058-6345. doi: 10.1186/s41044-016-0011-3. URL https://doi.org/10.1186/s41044-016-0011-3.

Gray, R. M. Toeplitz and Circulant Matrices: A Review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, January 2006. ISSN 1567-2190, 1567-2328. doi: 10.1561/0100000006.

URL https://www.nowpublishers.com/article/Details/CIT-006. Publisher: Now Publishers, Inc.

Gribonval, R., Chatalic, A., Keriven, N., Schellekens, V., Jacques, L., and Schniter, P. Sketching Data Sets for Large-Scale Learning: Keeping only what you need. *IEEE Signal Processing Magazine*, 38(5):12–36, September 2021. doi: 10.1109/MSP.2021.3092574. URL https://hal.inria.fr/hal-03350599. Publisher: Institute of Electrical and Electronics Engineers.

Hata, S. and Nakao, H. Localization of Laplacian eigenvectors on random networks. *Scientific Reports*, 7(1):1121, April 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-01010-0. URL https://www.nature.com/articles/s41598-017-01010-0. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Complex networks;Nonlinear phenomena Subject_term_id: complex-networks;nonlinear-phenomena.

Kadavankandy, A. and Couillet, R. Asymptotic Gaussian Fluctuations of Spectral Clustering Eigenvectors. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 694–698, December 2019. doi: 10.1109/CAMSAP45676.2019.9022474.

Keriven, N., Bourrier, A., Gribonval, R., and Pérez, P. Sketching for Large-Scale Learning of Mixture Models. *arXiv:1606.02838 [cs, stat]*, May 2017. URL http://arxiv.org/abs/1606.02838. arXiv: 1606.02838.

Liao, Z., Couillet, R., and Mahoney, M. W. Sparse Quantized Spectral Clustering. *arXiv:2010.01376 [cs, math, stat]*, October 2020. URL http://arxiv.org/abs/2010.01376. arXiv: 2010.01376.

Liberty, E. Simple and Deterministic Matrix Sketching. *arXiv:1206.0594 [cs]*, July 2012. URL http://arxiv.org/abs/1206.0594. arXiv: 1206.0594.

Liberty, E., Sriharsha, R., and Sviridenko, M. An Algorithm for Online K-Means Clustering. In *2016 Proceedings of the Meeting on Algorithm Engineering and Experiments (ALENEX)*, Proceedings, pp. 81–89. Society for Industrial and Applied Mathematics, December 2015. doi: 10.1137/1.9781611974317.7. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974317.7.

Lytova, A. and Pastur, L. Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *The Annals of Probability*, 37(5):1778–1840, 2009. Publisher: Institute of Mathematical Statistics.

Löffler, M., Zhang, A. Y., and Zhou, H. H. Optimality of Spectral Clustering in the Gaussian Mixture Model. *arXiv:1911.00538 [cs, math, stat]*, August 2020. URL http://arxiv.org/abs/1911.00538. arXiv: 1911.00538.

Mai, X. and Couillet, R. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *arXiv:1711.03404 [cs, stat]*, November 2017. URL http://arxiv.org/abs/1711.03404. arXiv: 1711.03404.

Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967. Publisher: IOP Publishing.

Ning, H., Xu, W., Chi, Y., Gong, Y., and Huang, T. S. Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recognition*, 43(1):113–127, January 2010. ISSN 0031-3203. doi: 10.1016/j.patcog.2009.06.001. URL https://www.sciencedirect.com/science/article/pii/S0031320309002209.

Onatski, A., Moreira, M. J., and Hallin, M. Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics*, 41(3):1204–1231, June 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1100. URL https://projecteuclid.org/journals/annals-of-statistics/volume-41/issue-3/Asymptotic-power-of-sphericity-tests-for-high-dim
10.1214/13-AOS1100.full. Publisher: Institute of Mathematical Statistics.

Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, April 2015. URL http://arxiv.org/abs/1409.1556. arXiv: 1409.1556.

Stein, C. M. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151, November 1981. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345632. URL https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-6/Estimation-of-the-Mean-of-a-Multivariate-Normal-D
10.1214/aos/1176345632.full. Publisher: Institute of Mathematical Statistics.

Tasoulis, D. K., Ross, G., and Adams, N. M. Visualising the Cluster Structure of Data Streams. In R. Berthold, M., Shawe-Taylor, J., and Lavrač, N. (eds.), *Advances in Intelligent Data Analysis VII*, Lecture Notes in Computer Science, pp. 81–92, Berlin, Heidelberg, 2007.

Springer. ISBN 978-3-540-74825-0. doi: 10.1007/978-3-540-74825-0_8.

Trench, W. F. Some Spectral Properties of Hermitian Toeplitz Matrices. *SIAM Journal on Matrix Analysis and Applications*, 15(3):938–942, July 1994. ISSN 0895-4798. doi: 10.1137/S0895479892239007. URL https://epubs.siam.org/doi/10.1137/S0895479892239007. Publisher: Society for Industrial and Applied Mathematics.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007. ISSN 1573-1375. doi: 10.1007/s11222-007-9033-z. URL https://doi.org/10.1007/s11222-007-9033-z.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv:1708.07747 [cs, stat]*, September 2017. URL http://arxiv.org/abs/1708.07747. arXiv: 1708.07747.

Yoo, S., Huang, H., and Kasiviswanathan, S. P. Streaming spectral clustering. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 637–648, May 2016. doi: 10.1109/ICDE.2016.7498277.

Zarrouk, T., Couillet, R., Chatelain, F., and Le Bihan, N. Performance-Complexity Trade-Off in Large Dimensional Statistics. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, September 2020. doi: 10.1109/MLSP49062.2020.9231568. ISSN: 1551-2541.

Zhang, T., Ramakrishnan, R., and Livny, M. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, SIGMOD '96, pp. 103–114, New York, NY, USA, June 1996. Association for Computing Machinery. ISBN 978-0-89791-794-0. doi: 10.1145/233269.233324. URL https://doi.org/10.1145/233269.233324.

A. Useful identities

Let $\mathbf{P} + \mathbf{Z} = \mathbf{X} \in \mathbb{R}^{p \times n}$ where $\mathbf{P}$ is a deterministic matrix and $\mathbf{Z}$ is a random matrix with independent entries $\mathbf{Z}_{i,j} \sim \mathcal{N}(0,1)$.

Let $\mathbf{Q} = \left(\frac{\mathbf{X}^\top \mathbf{X}}{p} \odot \mathbf{R} - z \mathbf{I}_n \right)^{-1}$ where $\mathbf{R} \in \mathbb{R}^{n \times n}$ is symmetric with bounded entries and $z \in \mathbb{C} \setminus \mathrm{Sp}\left(\frac{\mathbf{X}^\top \mathbf{X}}{p} \odot \mathbf{R} \right)$.

Proposition A.1.

$$\frac{\partial \mathbf{Q}_{k,l}}{\partial \mathbf{Z}_{i,j}} = -\frac{1}{p} \left(\left[\mathbf{X} \mathbf{D}_{\mathbf{R}_{\cdot,j}} \mathbf{Q} \right]_{i,k} \mathbf{Q}_{j,l} + \mathbf{Q}_{k,j} \left[\mathbf{X} \mathbf{D}_{\mathbf{R}_{j,\cdot}} \mathbf{Q} \right]_{i,l} \right)$$

Proof.

$$\begin{aligned}
\frac{\partial \mathbf{Q}_{k,l}}{\partial \mathbf{Z}_{i,j}} &= \left[\frac{\partial \mathbf{Q}}{\partial \mathbf{Z}_{i,j}} \right]_{k,l} \\
&= -\frac{1}{p} \left[\mathbf{Q} \frac{\partial \left(\mathbf{X}^\top \mathbf{X} \odot \mathbf{R} \right)}{\partial \mathbf{Z}_{i,j}} \mathbf{Q} \right]_{k,l} \\
&= -\frac{1}{p} \sum_{r,s=1}^{n} \mathbf{Q}_{k,r} \mathbf{Q}_{s,l} \frac{\partial}{\partial \mathbf{Z}_{i,j}} \sum_{t=1}^{n} \mathbf{X}_{t,r} \mathbf{X}_{t,s} \mathbf{R}_{r,s} \\
&= -\frac{1}{p} \sum_{r,s,t=1}^{n} \mathbf{Q}_{k,r} \mathbf{Q}_{s,l} \mathbf{R}_{r,s} \left[\frac{\partial \mathbf{X}_{t,r}}{\partial \mathbf{Z}_{i,j}} \mathbf{X}_{t,s} + \mathbf{X}_{t,r} \frac{\partial \mathbf{X}_{t,s}}{\partial \mathbf{Z}_{i,j}} \right] \\
&= -\frac{1}{p} \left(\sum_{s=1}^{n} \mathbf{Q}_{k,j} \mathbf{Q}_{s,l} \mathbf{R}_{j,s} \mathbf{X}_{i,s} + \sum_{r=1}^{n} \mathbf{Q}_{k,r} \mathbf{Q}_{j,l} \mathbf{R}_{r,j} \mathbf{X}_{i,r} \right) \\
\frac{\partial \mathbf{Q}_{k,l}}{\partial \mathbf{Z}_{i,j}} &= -\frac{1}{p} \left(\mathbf{Q}_{k,j} \left[\mathbf{X} \mathbf{D}_{\mathbf{R}_{j,\cdot}} \mathbf{Q} \right]_{i,l} + \mathbf{Q}_{j,l} \left[\mathbf{X} \mathbf{D}_{\mathbf{R}_{\cdot,j}} \mathbf{Q} \right]_{i,k} \right) \qquad \text{since } \mathbf{Q}^\top = \mathbf{Q}.
\end{aligned}$$

□

Lemma A.2 ((Stein, 1981))**.** *Let $Z \sim \mathcal{N}(0,1)$ and $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $\mathbb{E}\left[|f'(Z)| \right] < +\infty$ and $f(z) = \mathrm{o}_{z \to \pm\infty}(e^{z^2})$. Then,*

$$\mathbb{E}\left[Z f(Z) \right] = \mathbb{E}\left[f'(Z) \right].$$

Proof. Using integration by parts,

$$\mathbb{E}\left[Z f(Z) \right] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} z f(z) e^{-\frac{z^2}{2}} \, \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \left[-f(z) e^{-\frac{z^2}{2}} \right]_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f'(z) e^{-\frac{z^2}{2}} \, \mathrm{d}z.$$

In the right-hand side, the first term vanishes since $f(z) = \mathrm{o}_{z \to \pm\infty}(e^{z^2})$ and the second term equals $\mathbb{E}\left[f'(Z) \right]$. □

Proposition A.3. *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a matrix with bounded entries.*

1.
$$\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top \mathbf{Z}}{p} \odot \mathbf{A} \right) \mathbf{Q} \right]_{i,j} = \mathbb{E}\left[\mathbf{A}_{i,i} \mathbf{Q}_{i,j} \right] - \mathbb{E}\left[\frac{1}{p} \mathrm{tr}\left(\mathbf{Q} \left(\frac{(\mathbf{P} + \mathbf{Z})^\top \mathbf{Z}}{p} \odot \left[\mathbf{R}_{\cdot,i} \mathbf{A}_{i,\cdot} \right] \right) \right) \mathbf{Q}_{i,j} \right] + \underset{n,p \to +\infty}{\mathcal{O}}\left(\frac{1}{p} \right)$$

2.
$$\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top \mathbf{P}}{p} \odot \mathbf{A} \right) \mathbf{Q} \right]_{i,j} = -\mathbb{E}\left[\frac{1}{p} \mathrm{tr}\left(\mathbf{Q} \left(\frac{(\mathbf{P} + \mathbf{Z})^\top \mathbf{P}}{p} \odot \left[\mathbf{R}_{\cdot,i} \mathbf{A}_{i,\cdot} \right] \right) \right) \mathbf{Q}_{i,j} \right] + \underset{n,p \to +\infty}{\mathcal{O}}\left(\frac{1}{p} \right)$$

3.
$$\mathbb{E}\left[\left(\frac{\mathbf{P}^\top \mathbf{Z}}{p} \odot \mathbf{A} \right) \mathbf{Q} \right]_{i,j} = -\mathbb{E}\left[\frac{\mathbf{P}^\top (\mathbf{P} + \mathbf{Z})}{p} \mathcal{D}^{(i)}_{\mathbf{A},\mathbf{R}} \mathbf{Q} \right]_{i,j} + \underset{n,p \to +\infty}{\mathcal{O}}\left(\frac{1}{p} \right)$$

where $\mathcal{D}^{(i)}_{\mathbf{A},\mathbf{R}}$ *is a diagonal matrix such that* $\left[\mathcal{D}^{(i)}_{\mathbf{A},\mathbf{R}}\right]_{k,k} = \frac{1}{p}\sum_{l=1}^{n} \mathbf{Q}_{l,l}\mathbf{A}_{i,l}\mathbf{R}_{l,k} = \frac{1}{p}\operatorname{tr}\mathbf{D}_{\mathbf{R}_{\cdot,k}}\mathbf{Q}\mathbf{D}_{\mathbf{A}_{i,\cdot}}$.

Proof. We start with the first equation.

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{A}\right)\mathbf{Q}\right]_{i,j} &= \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\mathbf{Z}_{s,i}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\mathbf{Q}_{r,j}\right] \\
&= \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\frac{\partial\left(\mathbf{Z}_{s,r}\mathbf{Q}_{r,j}\right)}{\partial\mathbf{Z}_{s,i}}\mathbf{A}_{i,r}\right] \qquad \text{using Stein's lemma} \\
&= \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\frac{\partial\mathbf{Z}_{s,r}}{\partial\mathbf{Z}_{s,i}}\mathbf{A}_{i,r}\mathbf{Q}_{r,j} + \mathbf{Z}_{s,r}\mathbf{A}_{i,r}\frac{\partial\mathbf{Q}_{r,j}}{\partial\mathbf{Z}_{s,i}}\right] \\
\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{A}\right)\mathbf{Q}\right]_{i,j} &= \mathbb{E}\left[\mathbf{A}_{i,i}\mathbf{Q}_{i,j}\right] + \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\frac{\partial\mathbf{Q}_{r,j}}{\partial\mathbf{Z}_{s,i}}\right].
\end{aligned}$$

From proposition A.1, we know that

$$\frac{\partial\mathbf{Q}_{r,j}}{\partial\mathbf{Z}_{s,i}} = -\frac{1}{p}\left(\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{\cdot,i}}\mathbf{Q}\right]_{s,r}\mathbf{Q}_{i,j} + \mathbf{Q}_{r,i}\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{i,\cdot}}\mathbf{Q}\right]_{s,j}\right)$$

therefore,

$$\frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\frac{\partial\mathbf{Q}_{r,j}}{\partial\mathbf{Z}_{s,i}}\right] = -\frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\frac{1}{p}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{\cdot,i}}\mathbf{Q}\right]_{s,r}\mathbf{Q}_{i,j}\right] - \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\frac{1}{p}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\mathbf{Q}_{r,i}\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{i,\cdot}}\mathbf{Q}\right]_{s,j}\right].$$

Moreover,

$$\begin{aligned}
\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{\cdot,i}}\mathbf{Q}\right]_{s,r} &= \sum_{r=1}^{n}\sum_{s=1}^{p}\sum_{t=1}^{n}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\mathbf{X}_{s,t}\mathbf{R}_{t,i}\mathbf{Q}_{t,r} \\
&= \sum_{r=1}^{n}\sum_{t=1}^{n}\mathbf{Q}_{t,r}\left[\mathbf{X}^\top\mathbf{Z}\right]_{t,r}\mathbf{R}_{t,i}\mathbf{A}_{i,r} \\
\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{\cdot,i}}\mathbf{Q}\right]_{s,r} &= \operatorname{tr}\left(\mathbf{Q}\left(\mathbf{X}^\top\mathbf{Z}\odot\left[\mathbf{R}_{\cdot,i}\mathbf{A}_{i,\cdot}\right]\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\mathbf{Q}_{r,i}\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{i,\cdot}}\mathbf{Q}\right]_{s,j} &= \sum_{r=1}^{n}\mathbf{Q}_{i,r}\mathbf{A}_{i,r}\left[\mathbf{Z}^\top\mathbf{X}\mathbf{D}_{\mathbf{R}_{i,\cdot}}\mathbf{Q}\right]_{r,j} \\
\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\mathbf{Q}_{r,i}\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{i,\cdot}}\mathbf{Q}\right]_{s,j} &= \left[\mathbf{Q}\mathbf{D}_{\mathbf{A}_{i,\cdot}}\mathbf{Z}^\top\mathbf{X}\mathbf{D}_{\mathbf{R}_{i,\cdot}}\mathbf{Q}\right]_{i,j}.
\end{aligned}$$

So we finally have

$$\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{A}\right)\mathbf{Q}\right]_{i,j} = \mathbb{E}\left[\mathbf{A}_{i,i}\mathbf{Q}_{i,j}\right] - \frac{1}{p}\mathbb{E}\left[\operatorname{tr}\left(\mathbf{Q}\left(\frac{\mathbf{X}^\top\mathbf{Z}}{p}\odot\left[\mathbf{R}_{\cdot,i}\mathbf{A}_{i,\cdot}\right]\right)\right)\mathbf{Q}_{i,j}\right] \underbrace{- \frac{1}{p}\mathbb{E}\left[\mathbf{Q}\mathbf{D}_{\mathbf{A}_{i,\cdot}}\frac{\mathbf{Z}^\top\mathbf{X}}{p}\mathbf{D}_{\mathbf{R}_{i,\cdot}}\mathbf{Q}\right]_{i,j}}_{=\mathcal{O}_{n,p\to+\infty}\left(\frac{1}{p}\right)}.$$

The second equation can be shown in the same way.

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{P}}{p}\odot\mathbf{A}\right)\mathbf{Q}\right]_{i,j} &= \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\mathbf{Z}_{s,i}\mathbf{P}_{s,r}\mathbf{A}_{i,r}\mathbf{Q}_{r,j}\right] \\
&= \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\mathbf{P}_{s,r}\mathbf{A}_{i,r}\frac{\partial\mathbf{Q}_{r,j}}{\partial\mathbf{Z}_{s,i}}\right] \qquad \text{using Stein's lemma} \\
\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{P}}{p}\odot\mathbf{A}\right)\mathbf{Q}\right]_{i,j} &= -\frac{1}{p}\mathbb{E}\left[\mathrm{tr}\left(\mathbf{Q}\left(\frac{\mathbf{X}^\top\mathbf{P}}{p}\odot[\mathbf{R}_{\cdot,i}\mathbf{A}_{i,\cdot}]\right)\right)\mathbf{Q}_{i,j}\right] \underbrace{-\frac{1}{p}\mathbb{E}\left[\mathbf{Q}\mathbf{D}_{\mathbf{A}_{i,\cdot}}\frac{\mathbf{P}^\top\mathbf{X}}{p}\mathbf{D}_{\mathbf{R}_{i,\cdot}}\mathbf{Q}\right]_{i,j}}_{=\mathcal{O}_{n,p\to+\infty}\left(\frac{1}{p}\right)}.
\end{aligned}$$

We are left to show the third equation.

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\mathbf{P}^\top\mathbf{Z}}{p}\odot\mathbf{A}\right)\mathbf{Q}\right]_{i,j} &= \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\mathbf{P}_{s,i}\mathbf{Z}_{s,r}\mathbf{A}_{i,r}\mathbf{Q}_{r,j}\right] \\
&= \frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\mathbf{P}_{s,i}\mathbf{A}_{i,r}\frac{\partial\mathbf{Q}_{r,j}}{\partial\mathbf{Z}_{s,r}}\right] \qquad \text{using Stein's lemma} \\
&= -\frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\mathbb{E}\left[\frac{1}{p}\mathbf{P}_{s,i}\mathbf{A}_{i,r}\left(\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{\cdot,r}}\mathbf{Q}\right]_{s,r}\mathbf{Q}_{r,j}+\mathbf{Q}_{r,r}\left[\mathbf{X}\mathbf{D}_{\mathbf{R}_{r,\cdot}}\mathbf{Q}\right]_{s,j}\right)\right] \\
&= -\frac{1}{p}\sum_{r=1}^{n}\sum_{s=1}^{p}\sum_{t=1}^{n}\mathbb{E}\left[\frac{1}{p}\mathbf{P}_{s,i}\mathbf{A}_{i,r}\left(\mathbf{X}_{s,t}\mathbf{R}_{t,r}\mathbf{Q}_{t,r}\mathbf{Q}_{r,j}+\mathbf{Q}_{r,r}\mathbf{X}_{s,t}\mathbf{R}_{r,t}\mathbf{Q}_{t,j}\right)\right] \\
&= -\frac{1}{p}\sum_{r=1}^{n}\sum_{t=1}^{n}\mathbb{E}\left[\frac{1}{p}\left[\mathbf{P}^\top\mathbf{X}\right]_{i,t}\mathbf{A}_{i,r}\left(\mathbf{R}_{t,r}\mathbf{Q}_{t,r}\mathbf{Q}_{r,j}+\mathbf{Q}_{r,r}\mathbf{R}_{r,t}\mathbf{Q}_{t,j}\right)\right] \\
&= -\frac{1}{p}\sum_{t=1}^{n}\mathbb{E}\left[\left[\mathbf{P}^\top\mathbf{X}\right]_{i,t}\left(\frac{1}{p}\left[\left(\mathbf{Q}\odot\mathbf{R}\right)\mathbf{D}_{\mathbf{A}_{i,\cdot}}\mathbf{Q}\right]_{t,j}+\left[\mathcal{D}^{(i)}_{\mathbf{A},\mathbf{R}}\right]_{t,t}\mathbf{Q}_{t,j}\right)\right] \\
\mathbb{E}\left[\left(\frac{\mathbf{P}^\top\mathbf{Z}}{p}\odot\mathbf{A}\right)\mathbf{Q}\right]_{i,j} &= -\mathbb{E}\left[\frac{\mathbf{P}^\top\mathbf{X}}{p}\mathcal{D}^{(i)}_{\mathbf{A},\mathbf{R}}\mathbf{Q}\right]_{i,j} \underbrace{-\frac{1}{p}\mathbb{E}\left[\frac{\mathbf{P}^\top\mathbf{X}}{p}\left(\mathbf{Q}\odot\mathbf{R}\right)\mathbf{D}_{\mathbf{A}_{i,\cdot}}\mathbf{Q}\right]_{i,j}}_{=\mathcal{O}_{n,p\to+\infty}\left(\frac{1}{p}\right)}.
\end{aligned}$$

□

B. Proof of theorem 3.1

The study the spectral behavior of $\tilde{\mathbf{K}}_L = \frac{\mathbf{X}^\top\mathbf{X}}{p}\odot\mathbf{C}$ is made through its resolvent

$$\mathbf{Q} = \left(\tilde{\mathbf{K}}_L - z\mathbf{I}_n\right)^{-1}$$

where we have dropped the dependence in z to ease notations.

B.1. Analysis of the model with noise only: $\mathbf{X} = \mathbf{Z}$

In order to find the limiting spectral distribution of $\tilde{\mathbf{K}}_L = \frac{\mathbf{X}^\top\mathbf{X}}{p}\odot\mathbf{C}$, we first consider the simpler model with noise only, i.e.,

$$\mathbf{X} = \mathbf{Z}, \qquad \tilde{\mathbf{K}}_L = \frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C}, \qquad \mathbf{Q} = \left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C} - z\mathbf{I}_n\right)^{-1}.$$

B.1.1. A FIRST EQUIVALENT OF THE RESOLVENT

Let us first consider the following expression of the resolvent

$$\mathbf{Q} = -\frac{1}{z}\mathbf{I}_n + \frac{1}{z}\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p} \odot \mathbf{C}\right)\mathbf{Q}$$

which is a rewriting of $\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{I}_n$.

In order to find a deterministic equivalent of $\mathbf{Q}$, we study its expected value

$$\mathbb{E}\left[\mathbf{Q}_{i,j}\right] = -\frac{1}{z}\delta_{i,j} + \frac{1}{z}\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p} \odot \mathbf{C}\right)\mathbf{Q}\right]_{i,j}.$$

Taking $\mathbf{A} = \mathbf{C}$ in the first equation of proposition A.3, we have

$$z\mathbb{E}\left[\mathbf{Q}_{i,j}\right] = -\delta_{i,j} + \mathbf{C}_{i,i}\mathbb{E}\left[\mathbf{Q}_{i,j}\right] - \mathbb{E}\left[\boldsymbol{\eta}_{i,i}\mathbf{Q}_{i,j}\right] + \underset{n,p,L\to+\infty}{\mathcal{O}}\left(\frac{1}{p}\right)$$

with $\boldsymbol{\eta} \in \mathbb{C}^{n\times n}$ such that

$$\boldsymbol{\eta}_{r,s} = \frac{1}{p}\operatorname{tr}\left(\mathbf{Q}\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p} \odot \left[\mathbf{C}_{\cdot,r}\mathbf{C}_{s,\cdot}\right]\right)\right).$$

Thus, denoting $\mathbf{D}_{\boldsymbol{\eta}} = \boldsymbol{\eta} \odot \mathbf{I}_n$, we have the matrix equivalence $z\mathbf{Q} \leftrightarrow -\mathbf{I}_n + \mathbf{Q} - \mathbf{D}_{\boldsymbol{\eta}}\mathbf{Q}$ from which we deduce that the resolvent is equivalent to a diagonal matrix:

$$\mathbf{Q} \leftrightarrow \left(\mathbf{I}_n - z\mathbf{I}_n - \mathbf{D}_{\boldsymbol{\eta}}\right)^{-1}.$$

B.1.2. ANALYSIS OF THE MATRIX $\boldsymbol{\eta}$

Now taking $\mathbf{A} = \mathbf{C}_{\cdot,r}\mathbf{C}_{s,\cdot}$ in the first equation of proposition A.3, we have

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\eta}_{r,s}\right] &= \frac{1}{p}\sum_{t=1}^n \mathbf{C}_{t,r}\mathbf{C}_{s,t}\mathbb{E}\left[\mathbf{Q}_{t,t}\right] - \frac{1}{p}\sum_{t=1}^n \mathbb{E}\left[\frac{1}{p}\operatorname{tr}\left(\mathbf{Q}\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p} \odot \left[\mathbf{C}_{\cdot,t}\mathbf{C}_{s,\cdot}\right]\right)\right)\mathbf{C}_{t,r}\mathbf{Q}_{t,t}\right] + \underset{n,p,L\to+\infty}{\mathcal{O}}\left(\frac{1}{p}\right)\\ &= \frac{1}{p}\sum_{t=1}^n \mathbf{C}_{t,r}\left(\mathbf{C}_{s,t}\mathbb{E}\left[\mathbf{Q}_{t,t}\right] - \mathbb{E}\left[\boldsymbol{\eta}_{t,s}\mathbf{Q}_{t,t}\right]\right) + \underset{n,p,L\to+\infty}{\mathcal{O}}\left(\frac{1}{p}\right)\end{aligned}$$

and, using the previous matrix equivalent of $\mathbf{Q}$, we can write $\boldsymbol{\eta} \leftrightarrow \bar{\boldsymbol{\eta}}$ where $\bar{\boldsymbol{\eta}}$ is a deterministic matrix such that

$$\bar{\boldsymbol{\eta}}_{r,s} = \frac{1}{p}\sum_{t=1}^n \mathbf{C}_{t,r}\frac{\mathbf{C}_{s,t} - \bar{\boldsymbol{\eta}}_{t,s}}{1 - z - \bar{\boldsymbol{\eta}}_{t,t}}.$$

Therefore, $\bar{\boldsymbol{\eta}}$ has a circulant structure. Indeed, for all $d \in \mathbb{Z}$,

$$\bar{\boldsymbol{\eta}}_{r+d,s+d} = \frac{1}{p}\sum_{t=r-L+1}^{r+L-1}\frac{\mathbf{C}_{s+d,t+d} - \bar{\boldsymbol{\eta}}_{t+d,s+d}}{1 - z - \bar{\boldsymbol{\eta}}_{t+d,t+d}} \qquad d \in \mathbb{Z}.$$

where we write $\bar{\boldsymbol{\eta}}_{i,j}$ for any $i, j \in \mathbb{Z}$ to represent $\bar{\boldsymbol{\eta}}_{(i \mod n),(j \mod n)}$.

B.1.3. FROM $\bar{\boldsymbol{\eta}}$ TO THE LIMITING SPECTRAL DISTRIBUTION

Since $\bar{\boldsymbol{\eta}}$ is circulant, it has a constant diagonal: $\bar{\boldsymbol{\eta}}_{k,k} = \eta_0$. Then, we can recognize a matrix product in the expression of $\bar{\boldsymbol{\eta}}_{r,s}$:

$$\bar{\boldsymbol{\eta}}_{r,s} = \frac{1}{p}\frac{1}{1 - z - \eta_0}\sum_{t=1}^n \mathbf{C}_{t,r}\left(\mathbf{C}_{s,t} - \bar{\boldsymbol{\eta}}_{t,s}\right) = \frac{1}{p}\frac{1}{1 - z - \eta_0}\left[\mathbf{C}\left(\mathbf{C} - \bar{\boldsymbol{\eta}}\right)\right]_{r,s}$$

thus, $\bar{\boldsymbol{\eta}} = \left(p\left(1-z-\eta_0\right)\mathbf{I}_n + \mathbf{C}\right)^{-1}\mathbf{C}^2$ and, since $\eta_0 = \frac{1}{n}\operatorname{tr}\bar{\boldsymbol{\eta}}$,

$$\eta_0 = \frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k^2}{p\left(1-z-\eta_0\right)+\psi_k}.$$

Recalling that $\mathbf{Q} \leftrightarrow \left(\mathbf{I}_n - z\mathbf{I}_n - \mathbf{D}_{\boldsymbol{\eta}}\right)^{-1}$, we can state the following theorem.

Theorem B.1 (Deterministic equivalent of $\mathbf{Q}$ when $\mathbf{X}=\mathbf{Z}$)**.** *Let $z \in \mathbb{C} \setminus \limsup_{n,p,L\to+\infty} \operatorname{Sp}(\tilde{\mathbf{K}}_L)$. Then,*

$$\mathbf{Q} \leftrightarrow m(z)\mathbf{I}_n \qquad \textit{with} \qquad m(z) = \frac{1}{1-z-\eta_0}$$

and η_0 is solution to the fixed-point equation

$$\eta_0 = \frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k^2}{p\left(1-z-\eta_0\right)+\psi_k}.$$

m is the Stieljes transform of the limiting spectral distribution of $\tilde{\mathbf{K}}_L = \frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C}$.

Remark B.2. Notice that $m(z) \neq 0$. Indeed, for a given $z \in \mathbb{C} \setminus \limsup_{n,p,L\to+\infty} \operatorname{Sp}(\tilde{\mathbf{K}}_L)$, $\left(1-z-\eta_0\right)m(z) = 1$ and the fixed-point equation prevent η_0 from going to $\pm\infty$.

Proposition B.3 (Fixed-point equation for $m(z)$)**.** *Under the setting of theorem B.1, $m(z)$ is also solution to a fixed-point equation:*

$$1 + zm(z) = \frac{1}{n}\sum_{k=0}^{n-1}\frac{m(z)\psi_k}{1+\frac{m(z)}{p}\psi_k}.$$

Proof. A rewriting of $\bar{\boldsymbol{\eta}} = \left(p\left(1-z-\eta_0\right)\mathbf{I}_n + \mathbf{C}\right)^{-1}\mathbf{C}^2$ yields another interesting formula:

$$\bar{\boldsymbol{\eta}} = \left(\frac{p}{m(z)}\mathbf{I}_n + \mathbf{C}\right)^{-1}\left(\frac{p}{m(z)}\mathbf{I}_n + \mathbf{C} - \frac{p}{m(z)}\mathbf{I}_n\right)\mathbf{C}$$
$$\bar{\boldsymbol{\eta}} = \mathbf{C} - \left(\mathbf{I}_n + \frac{m(z)}{p}\mathbf{C}\right)^{-1}\mathbf{C}$$

therefore,

$$\eta_0 = \underbrace{\frac{1}{n}\operatorname{tr}\mathbf{C}}_{=1} - \frac{1}{n}\sum_{k=0}^{n-1}\frac{\psi_k}{1+\frac{m(z)}{p}\psi_k}$$

and, since $1-\eta_0 = \frac{1}{m(z)} + z$, we get the result. □

B.2. Analysis of the full model: $\mathbf{X} = \mathbf{P} + \mathbf{Z}$

So far, we have been able to find a deterministic equivalent of the resolvent under the setting where $\mathbf{X}=\mathbf{Z}$, i.e., when the observations are composed of noise only.

Now, we consider the setting where the observations are composed of a signal corrupted with additive noise:

$$\mathbf{X} = \mathbf{P} + \mathbf{Z}, \qquad \tilde{\mathbf{K}}_L = \frac{\mathbf{X}^\top\mathbf{X}}{p}\odot\mathbf{C}, \qquad \mathbf{Q} = \left(\frac{\mathbf{X}^\top\mathbf{X}}{p}\odot\mathbf{C} - z\mathbf{I}_n\right)^{-1}.$$

Let us first prove that the limiting spectral distribution is unchanged.

Proposition B.4.

$$\left|\frac{1}{n}\operatorname{tr}\left(\frac{\left(\mathbf{P}+\mathbf{Z}\right)^\top\left(\mathbf{P}+\mathbf{Z}\right)}{p}\odot\mathbf{C} - z\mathbf{I}_n\right)^{-1} - \frac{1}{n}\operatorname{tr}\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C} - z\mathbf{I}_n\right)^{-1}\right| \xrightarrow[n,p,L\to+\infty]{\text{a.s.}} 0.$$

Proof.

$$\frac{(\mathbf{P}+\mathbf{Z})^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}=\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C}+\mathbf{A}\odot\mathbf{C}$$

with $\mathbf{A}=\frac{\mathbf{Z}^\top\mathbf{P}}{p}+\frac{\mathbf{P}^\top\mathbf{Z}}{p}+\frac{\mathbf{P}^\top\mathbf{P}}{p}$. Notice that, $\frac{\mathbf{Z}^\top\mathbf{P}}{p}$, $\frac{\mathbf{P}^\top\mathbf{Z}}{p}$ and $\frac{\mathbf{P}^\top\mathbf{P}}{p}$ are uniformly bounded in spectral norm (from the non-triviality condition) and their rank is at most K. Thus $\mathbf{A}$ is also uniformly bounded in spectral norm and has rank at most $3K$.

Let $\mathbf{Q}_\mathbf{A}=\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C}+\mathbf{A}\odot\mathbf{C}-z\mathbf{I}_n\right)^{-1}$ and $\mathbf{Q}_0=\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C}-z\mathbf{I}_n\right)^{-1}$. Using singular-value inequalities which can be found in theorems A.12 and A.14 of (Bai & Silverstein, 2010),

$$\begin{aligned}
\left|\frac{1}{n}\operatorname{tr}\mathbf{Q}_\mathbf{A}-\frac{1}{n}\operatorname{tr}\mathbf{Q}_0\right| &= \frac{1}{n}\left|\operatorname{tr}\mathbf{Q}_0\left(\mathbf{A}\odot\mathbf{C}\right)\mathbf{Q}_\mathbf{A}\right| \\
&\leqslant \frac{1}{n}\left\|\mathbf{Q}_0\right\|\left\|\mathbf{Q}_A\right\|\sum_{i=1}^n s_i(\mathbf{A}\odot\mathbf{C}) \qquad \text{from theorems A.12 and A.14} \\
&\leqslant \frac{1}{n}\left\|\mathbf{Q}_0\right\|\left\|\mathbf{Q}_A\right\|\sqrt{n\sum_{i=1}^n s_i^2(\mathbf{A}\odot\mathbf{C})} \qquad \text{from Jensen's inequality} \\
&\leqslant \frac{1}{n}\left\|\mathbf{Q}_0\right\|\left\|\mathbf{Q}_A\right\|\sqrt{n\sum_{i=1}^n s_i^2(\mathbf{A})} \qquad \text{since } \left\|\mathbf{A}\odot\mathbf{C}\right\|_F\leqslant\left\|\mathbf{A}\right\|_F \\
\left|\frac{1}{n}\operatorname{tr}\mathbf{Q}_\mathbf{A}-\frac{1}{n}\operatorname{tr}\mathbf{Q}_0\right| &\leqslant \sqrt{\frac{3K}{n}}\left\|\mathbf{Q}_0\right\|\left\|\mathbf{Q}_A\right\|\left\|\mathbf{A}\right\| = \underset{n,p,L\to+\infty}{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right) \qquad \text{since } \mathbf{A} \text{ has rank at most } 3K.
\end{aligned}$$

□

Since $\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C}-z\mathbf{I}_n\right)^{-1}\leftrightarrow m(z)\mathbf{I}_n$ according to theorem B.1, proposition B.4 justifies that the limiting spectral distribution is unchanged by the presence of signal.

As previously, we consider a rewriting of $\mathbf{Q}^{-1}\mathbf{Q}=\mathbf{I}_n$,

$$\mathbf{Q}=-\frac{1}{z}\mathbf{I}_n+\frac{1}{z}\left(\frac{(\mathbf{P}+\mathbf{Z})^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}\right)\mathbf{Q}$$

and we study the expected value of $\mathbf{Q}_{i,j}$,

$$\mathbb{E}\left[\mathbf{Q}_{i,j}\right]=-\frac{1}{z}\delta_{i,j}+\frac{1}{z}\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot\mathbf{C}\right)\mathbf{Q}\right]_{i,j}+\frac{1}{z}\mathbb{E}\left[\left(\frac{\mathbf{Z}^\top\mathbf{P}}{p}\odot\mathbf{C}\right)\mathbf{Q}\right]_{i,j}+\frac{1}{z}\mathbb{E}\left[\left(\frac{\mathbf{P}^\top\mathbf{Z}}{p}\odot\mathbf{C}\right)\mathbf{Q}\right]_{i,j}+\frac{1}{z}\mathbb{E}\left[\left(\frac{\mathbf{P}^\top\mathbf{P}}{p}\odot\mathbf{C}\right)\mathbf{Q}\right]_{i,j}.$$

$\mathbf{P}^\top\mathbf{P}$ is deterministic so there is no work to do on the last term of the sum. Expanding the other terms yields (see proposition A.3)

$$z\mathbb{E}\left[\mathbf{Q}_{i,j}\right]=-\delta_{i,j}+\mathbb{E}\left[\mathbf{C}_{i,i}\mathbf{Q}_{i,j}\right]-\mathbb{E}\left[\boldsymbol{\kappa}_{i,i}\mathbf{Q}_{i,j}\right]-\mathbb{E}\left[\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\mathcal{D}^{(i)}_{\mathbf{C},\mathbf{C}}\mathbf{Q}\right]_{i,j}+\mathbb{E}\left[\left(\frac{\mathbf{P}^\top\mathbf{P}}{p}\odot\mathbf{C}\right)\mathbf{Q}\right]_{i,j}+\underset{n,p,L\to+\infty}{\mathcal{O}}\left(\frac{1}{p}\right)$$

with $\boldsymbol{\kappa}\in\mathbb{C}^{n\times n}$ such that

$$\boldsymbol{\kappa}_{r,s}=\frac{1}{p}\operatorname{tr}\left(\mathbf{Q}\left(\frac{(\mathbf{P}+\mathbf{Z})^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\left[\mathbf{C}_{\cdot,r}\mathbf{C}_{s,\cdot}\right]\right)\right)$$

and $\mathcal{D}^{(i)}_{\mathbf{C},\mathbf{C}}$ is a diagonal matrix such that $\left[\mathcal{D}^{(i)}_{\mathbf{C},\mathbf{C}}\right]_{k,k}=\frac{1}{p}\sum_{l=1}^n\mathbf{Q}_{l,l}\mathbf{C}_{i,l}\mathbf{C}_{l,k}$.

Proposition B.5.

$$\boldsymbol{\kappa}\leftrightarrow\bar{\boldsymbol{\eta}}.$$

Proof. Similarly to the proof of proposition B.4, we consider a matrix $\mathbf{A}$ uniformly bounded in spectral norm whose rank is at most K, representing $\frac{\mathbf{Z}^\top\mathbf{P}}{p}$, $\frac{\mathbf{P}^\top\mathbf{Z}}{p}$ or $\frac{\mathbf{P}^\top\mathbf{P}}{p}$ and we make use of singular-value inequalities.

$$\begin{aligned}
\frac{1}{p}\left|\mathrm{tr}\left(\mathbf{Q}\left(\mathbf{A}\odot[\mathbf{C}_{\cdot,r}\mathbf{C}_{s,\cdot}]\right)\right)\right| &= \frac{1}{p}\left|\mathrm{tr}\left(\mathbf{Q}\mathbf{D}_{\mathbf{C}_{\cdot,r}}\mathbf{A}\mathbf{D}_{\mathbf{C}_{s,\cdot}}\right)\right| \\
&\leqslant \frac{1}{p}\sum_{i=1}^n s_i(\mathbf{Q}\mathbf{D}_{\mathbf{C}_{\cdot,r}}\mathbf{A}\mathbf{D}_{\mathbf{C}_{s,\cdot}}) \\
&\leqslant \frac{1}{p}\sum_{i=1}^n s_i(\mathbf{Q})s_i(\mathbf{D}_{\mathbf{C}_{\cdot,r}}\mathbf{A}\mathbf{D}_{\mathbf{C}_{s,\cdot}}) \\
&\leqslant \frac{1}{p}\|\mathbf{Q}\|\sum_{i=1}^n s_i(\mathbf{D}_{\mathbf{C}_{\cdot,r}}\mathbf{A}\mathbf{D}_{\mathbf{C}_{s,\cdot}}) \\
&\leqslant \frac{K}{p}\|\mathbf{Q}\|\left\|\mathbf{D}_{\mathbf{C}_{\cdot,r}}\mathbf{A}\mathbf{D}_{\mathbf{C}_{s,\cdot}}\right\| \qquad \text{since } \mathbf{A} \text{ has rank at most } K \\
\frac{1}{p}\left|\mathrm{tr}\left(\mathbf{Q}\left(\mathbf{A}\odot[\mathbf{C}_{\cdot,r}\mathbf{C}_{s,\cdot}]\right)\right)\right| &\leqslant \frac{K}{p}\|\mathbf{Q}\|\|\mathbf{A}\| = \underset{n,p,L\to+\infty}{\mathcal{O}}\left(\frac{1}{p}\right) \qquad \text{since } \left\|\mathbf{D}_{\mathbf{C}_{\cdot,r}}\right\| = \left\|\mathbf{D}_{\mathbf{C}_{s,\cdot}}\right\| = 1.
\end{aligned}$$

Hence,

$$\boldsymbol{\kappa}_{r,s} = \underbrace{\frac{1}{p}\mathrm{tr}\left(\mathbf{Q}\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{p}\odot[\mathbf{C}_{\cdot,r}\mathbf{C}_{s,\cdot}]\right)\right)}_{=\boldsymbol{\eta}_{r,s}} + \underset{n,p,L\to+\infty}{\mathcal{O}}\left(\frac{1}{p}\right)$$

□

So far, we have

$$z\mathbf{Q} \leftrightarrow -\mathbf{I}_n + \mathbf{Q} - \eta_0\mathbf{Q} + \left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}\right)\mathbf{Q} \qquad \text{i.e.} \qquad \mathbf{Q} \leftrightarrow \left(\frac{1}{m(z)}\mathbf{I}_n + \frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}\right)^{-1}.$$

The analysis of $\left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}\right)\mathbf{Q}$ is summarized in the following proposition.

Proposition B.6.

$$\left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}\right)\mathbf{Q} \leftrightarrow \left(\frac{\mathbf{P}^\top\mathbf{P}}{p}\odot\mathbf{F}\boldsymbol{\Psi}\left(\mathbf{I}_n+\frac{m(z)}{p}\boldsymbol{\Psi}\right)^{-1}\mathbf{F}^*\right)\mathbf{Q}.$$

Proof. From assumption 2.1, all diagonal entries of $\mathbf{Q}$ are statistically equivalent. Thus, we can have a simple matrix equivalent of $\mathcal{D}^{(i)}_{\mathbf{C}^t,\mathbf{C}}$ for all integer $t \geqslant 1$:

$$\left[\mathcal{D}^{(i)}_{\mathbf{C}^t,\mathbf{C}}\right]_{k,k} = \frac{1}{p}\sum_{l=1}^n \mathbf{Q}_{l,l}\left[\mathbf{C}^t\right]_{i,l}\mathbf{C}_{l,k} \leftrightarrow \frac{1}{p}\frac{\mathrm{tr}\,\mathbf{Q}}{n}\sum_{l=1}^n\left[\mathbf{C}^t\right]_{i,l}\mathbf{C}_{l,k} \leftrightarrow \frac{m(z)}{p}\left[\mathbf{C}^{t+1}\right]_{i,k}$$

where the last equivalence is justified by proposition B.4.

Now, we can notice the following recurrence relation

$$\begin{aligned}
\left[\left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}^t\right)\mathbf{Q}\right]_{i,j} &\leftrightarrow -\left[\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\mathcal{D}^{(i)}_{\mathbf{C}^t,\mathbf{C}}\mathbf{Q}\right]_{i,j} + \left[\left(\frac{\mathbf{P}^\top\mathbf{P}}{p}\odot\mathbf{C}^t\right)\mathbf{Q}\right]_{i,j} \\
&\leftrightarrow -\frac{m(z)}{p}\sum_{k=1}^n\left[\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\right]_{i,k}\left[\mathbf{C}^{t+1}\right]_{i,k}\mathbf{Q}_{k,j} + \left[\left(\frac{\mathbf{P}^\top\mathbf{P}}{p}\odot\mathbf{C}^t\right)\mathbf{Q}\right]_{i,j} \\
\left[\left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}^t\right)\mathbf{Q}\right]_{i,j} &\leftrightarrow -\frac{m(z)}{p}\left[\left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\odot\mathbf{C}^{t+1}\right)\mathbf{Q}\right]_{i,j} + \left[\left(\frac{\mathbf{P}^\top\mathbf{P}}{p}\odot\mathbf{C}^t\right)\mathbf{Q}\right]_{i,j}.
\end{aligned}$$

In particular, for all integer $T \geqslant 1$,

$$\left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p} \odot \mathbf{C}\right)\mathbf{Q} \leftrightarrow \left(-\frac{m(z)}{p}\right)^T \left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p} \odot \mathbf{C}^{T+1}\right)\mathbf{Q} + \sum_{t=0}^{T-1}\left(-\frac{m(z)}{p}\right)^t \left(\frac{\mathbf{P}^\top\mathbf{P}}{p} \odot \mathbf{C}^{t+1}\right)\mathbf{Q}.$$

We know that $\|\mathbf{C}\| = (2L-1)$ and, using the fact that the spectral norm of a pointwise product (as well as a regular matrix product) can be bounded by the product of the spectral norms (see theorem A.19 of (Bai & Silverstein, 2010)), we have

$$\left\|\left(-\frac{m(z)}{p}\right)^T \left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p} \odot \mathbf{C}^{T+1}\right)\mathbf{Q}\right\| \leqslant \left|\frac{m(z)}{p}\right|^T \left\|\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p}\right\| |2L-1|^{T+1} \|\mathbf{Q}\|.$$

Thus, if $\left|\frac{2L-1}{p}m(z)\right| < 1$,

$$\left(\frac{\mathbf{P}^\top(\mathbf{P}+\mathbf{Z})}{p} \odot \mathbf{C}\right)\mathbf{Q} \leftrightarrow \left(\frac{\mathbf{P}^\top\mathbf{P}}{p} \odot \left[\sum_{t=0}^{+\infty}\left(-\frac{m(z)}{p}\right)^t \mathbf{C}^{t+1}\right]\right)\mathbf{Q}.$$

And, since $\mathbf{C} = \mathbf{F}\mathbf{\Psi}\mathbf{F}^*$,

$$\sum_{t=0}^{+\infty}\left(-\frac{m(z)}{p}\right)^t \mathbf{C}^{t+1} = \mathbf{F}\mathbf{\Psi}\left(\mathbf{I}_n + \frac{m(z)}{p}\mathbf{\Psi}\right)^{-1}\mathbf{F}^*$$

which completes the proof. □

We can now state the following theorem.

Theorem B.7 (Deterministic equivalent of $\mathbf{Q}$ when $\mathbf{X} = \mathbf{P}+\mathbf{Z}$)**.** *Let $z \in \mathbb{C} \setminus \limsup_{n,p,L\to+\infty} \mathrm{Sp}(\tilde{\mathbf{K}}_L)$. If $\left|\frac{2L-1}{p}m(z)\right| < 1$, then*

$$\mathbf{Q} \leftrightarrow m(z)\left(\mathbf{I}_n + \frac{\mathbf{P}^\top\mathbf{P}}{p} \odot \mathbf{F}\mathbf{\Lambda}\mathbf{F}^*\right)^{-1} \qquad \textit{with} \qquad \mathbf{\Lambda} = m(z)\mathbf{\Psi}\left(\mathbf{I}_n + \frac{m(z)}{p}\mathbf{\Psi}\right)^{-1}.$$

Remark B.8. This is coherent with the result of theorem B.1 when $\mathbf{P} = \mathbf{0}$.

Remark B.9. From proposition B.3, we see that $1 + zm(z) = \frac{1}{n}\operatorname{tr}\mathbf{\Lambda}$.

C. Proof of theorem 3.3

In this section, we use the following notation:

$$\mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L) = \limsup_{n,p,L\to+\infty} \mathrm{Sp}(\tilde{\mathbf{K}}_L)$$

C.1. Spikes

Here, $\mathbf{P} = \boldsymbol{\mu}\mathbf{j}^\top$. Let us state a much more tractable expression of the deterministic equivalent of the resolvent.

Theorem C.1 (Deterministic equivalent of $\mathbf{Q}$ when $\mathbf{P} = \boldsymbol{\mu}\mathbf{j}^\top$)**.** *Let $z \in \mathbb{C} \setminus \mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)$. If $\left|\frac{2L-1}{p}m(z)\right| < 1$, then*

$$\mathbf{Q} \leftrightarrow \bar{\mathbf{Q}} = m(z)\left[\mathbf{D}_\mathbf{j}\mathbf{F}\right]\left(\mathbf{I}_n + \frac{\|\boldsymbol{\mu}\|^2}{p}\mathbf{\Lambda}\right)^{-1}\left[\mathbf{D}_\mathbf{j}\mathbf{F}\right]^*$$

where $\mathbf{D}_\mathbf{j}$ is the diagonal matrix induced by vector $\mathbf{j}$.

Proof. From theorem B.7,

$$\begin{aligned}
\mathbf{Q} &\leftrightarrow m(z)\left(\mathbf{I}_n + \frac{\|\boldsymbol{\mu}\|^2}{p}\mathbf{j}\mathbf{j}^\top \odot \mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^*\right)^{-1} \\
&\leftrightarrow m(z)\left(\mathbf{I}_n + \frac{\|\boldsymbol{\mu}\|^2}{p}\mathbf{D}_\mathbf{j}\mathbf{F}\boldsymbol{\Lambda}\mathbf{F}^*\mathbf{D}_\mathbf{j}\right)^{-1} \\
\mathbf{Q} &\leftrightarrow m(z)\left[\mathbf{D}_\mathbf{j}\mathbf{F}\right]\left(\mathbf{I}_n + \frac{\|\boldsymbol{\mu}\|^2}{p}\boldsymbol{\Lambda}\right)^{-1}\left[\mathbf{D}_\mathbf{j}\mathbf{F}\right]^* \qquad \text{since } \mathbf{D}_\mathbf{j}\mathbf{F} \text{ is a unitary matrix.}
\end{aligned}$$

□

The sought-after *spikes* which encapsulate the information about our data are the singular points of the resolvent. Therefore, their asympotical position is given by the solution in z to

$$1 + \frac{\|\boldsymbol{\mu}\|^2}{p}\frac{m(z)\psi_k}{1+\frac{m(z)}{p}\psi_k} = 0 \qquad 0 \leqslant k < n.$$

Since $\tilde{\mathbf{K}}_L$ is symmetric, all solutions are real. Moreover, there cannot be any spike inside $\mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)$ (the eigenvalue must be isolated). Therefore, we are only interested in solutions outside $\mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)$.

If $\psi_k = 0$, there is no solution, whereas if $\psi_k \neq 0$,

$$1 + \frac{\|\boldsymbol{\mu}\|^2}{p}\frac{m(z)\psi_k}{1+\frac{m(z)}{p}\psi_k} = 0 \iff m(z) = \frac{-1}{\frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k}$$

and, supposing $z \in \mathbb{C} \setminus \mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)$, we have, from proposition B.3,

$$z = \frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k + \frac{1}{n}\sum_{l=0}^{n-1}\frac{\psi_l}{1-\frac{\psi_l}{\left(\|\boldsymbol{\mu}\|^2+1\right)\psi_k}}.$$

Proposition C.2 (Singular points of $\bar{\mathbf{Q}}$)**.** *Let*

$$\bar{\xi}_k = \frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k\left(1 + \frac{p}{n}\sum_{l=0}^{n-1}\frac{\psi_l}{\left(\|\boldsymbol{\mu}\|^2+1\right)\psi_k - \psi_l}\right) \qquad 0 \leqslant k < n.$$

The set of singular points of $\bar{\mathbf{Q}}$ *is* $\left\{\bar{\xi}_k \mid 0 \leqslant k < n, \psi_k \neq 0\right\} \cap \left(\mathbb{C} \setminus \mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)\right)$.

C.2. Alignments and phase transition

Let us denote $\{(\xi_k, \mathbf{u}_k)\}_{0\leqslant k<n}$ the pairs eigenvalue-eigenvector of $\tilde{\mathbf{K}}_L$. From the definition of the resolvent, we know that

$$\mathbf{Q} = \sum_{l=0}^{n-1}\frac{\mathbf{u}_l\mathbf{u}_l^*}{\xi_l - z}.$$

Therefore, with Cauchy's integral formula and a positively oriented closed contour Γ_k circling around ξ_k and leaving the other eigenvalues outside, we can have access to the quantity

$$\sum_{\substack{0\leqslant l\leqslant n-1\\ \xi_l=\xi_k}} \mathbf{u}_l\mathbf{u}_l^* = -\frac{1}{2\mathrm{i}\pi}\oint_{\Gamma_k}\mathbf{Q}(z)\,\mathrm{d}z$$

which is simply $\mathbf{u}_k\mathbf{u}_k^*$ when the associated eigenvalue has multiplicity one. Then, we can calculate the alignment of any vector $\mathbf{v} \in \mathbb{C}^n$ with the eigenspace associated to ξ_k:

$$\sum_{\substack{0 \leqslant l \leqslant n-1 \\ \xi_l = \xi_k}} |\mathbf{v}^* \mathbf{u}_l|^2 = -\frac{1}{2\mathrm{i}\pi} \oint_{\Gamma_k} \mathbf{v}^* \mathbf{Q}(z) \mathbf{v} \, \mathrm{d}z.$$

Using the deterministic equivalent of $\mathbf{Q}$, we have the following result.

Proposition C.3 (Spike alignments)**.** *For $0 \leqslant k < n$ such that $\bar{\xi}_k$ is a singular point of $\bar{\mathbf{Q}}$, let Γ_k be a positively oriented closed contour circling around $\bar{\xi}_k$ and leaving all the $\bar{\xi}_l \neq \bar{\xi}_k$ outside.*

$$-\frac{1}{2\mathrm{i}\pi} \oint_{\Gamma_k} \bar{\mathbf{Q}}(z) \, \mathrm{d}z = \bar{\zeta}_k \left[\mathbf{D_j F}\right] \mathcal{D}_k \left[\mathbf{D_j F}\right]^*$$

where

$$\bar{\zeta}_k = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + 1} \left(1 - \frac{p}{n} \sum_{l=0}^{n-1} \left[\frac{\psi_l}{\left(\|\boldsymbol{\mu}\|^2 + 1\right) \psi_k - \psi_l} \right]^2 \right) \qquad \textit{and} \qquad \mathcal{D}_k = \operatorname{diag}\left(\mathbf{1}_{\psi_k = \psi_l} \right)_{0 \leqslant l < n}.$$

Proof. By residue calculus,

$$-\frac{1}{2\mathrm{i}\pi} \oint_{\Gamma_k} \bar{\mathbf{Q}}(z) \, \mathrm{d}z = -\left[\mathbf{D_j F}\right] \left[\lim_{z \to \bar{\xi}_k} \left(z - \bar{\xi}_k\right) m(z) \left(\mathbf{I}_n + \frac{\|\boldsymbol{\mu}\|^2}{p} \boldsymbol{\Lambda}(z) \right)^{-1} \right] \left[\mathbf{D_j F}\right]^*.$$

Let $0 \leqslant l < n$. If $\psi_l \neq \psi_k$, then

$$\lim_{z \to \bar{\xi}_k} \frac{\left(z - \bar{\xi}_k\right) m(z)}{1 + \frac{\|\boldsymbol{\mu}\|^2}{p} \frac{m(z)\psi_l}{1 + \frac{m(z)}{p}\psi_l}} = 0$$

whereas if $\psi_l = \psi_k$, L'Hôpital's rule yields

$$\begin{aligned} \lim_{z \to \bar{\xi}_k} \frac{\left(z - \bar{\xi}_k\right) m(z)}{1 + \frac{\|\boldsymbol{\mu}\|^2}{p} \frac{m(z)\psi_l}{1 + \frac{m(z)}{p}\psi_l}} &= \frac{m(\bar{\xi}_k)}{\frac{\mathrm{d}}{\mathrm{d}z} \left[1 + \frac{\|\boldsymbol{\mu}\|^2}{p} \frac{m(z)\psi_l}{1 + \frac{m(z)}{p}\psi_l} \right]_{z = \bar{\xi}_k}} \\ &= \frac{m(\bar{\xi}_k) \left(1 + \frac{m(\bar{\xi}_k)}{p}\psi_l\right)^2}{\frac{\|\boldsymbol{\mu}\|^2}{p} m'(\bar{\xi}_k)\psi_l}. \end{aligned}$$

Recalling that $m(\bar{\xi}_k) = \frac{-1}{\frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k}$, we have $1 + \frac{m(\bar{\xi}_k)}{p}\psi_k = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+1}$. Hence,

$$\begin{aligned} \lim_{z \to \bar{\xi}_k} \frac{\left(z - \bar{\xi}_k\right) m(z)}{1 + \frac{\|\boldsymbol{\mu}\|^2}{p} \frac{m(z)\psi_l}{1 + \frac{m(z)}{p}\psi_l}} &= \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + 1} \frac{1}{\frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k} \frac{m(\bar{\xi}_k)}{m'(\bar{\xi}_k)} \\ &= -\frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + 1} \frac{m^2(\bar{\xi}_k)}{m'(\bar{\xi}_k)}. \end{aligned}$$

Let us calculate an expression of $\frac{m^2(\bar{\xi}_k)}{m'(\bar{\xi}_k)}$. Differentiating in z the fixed-point equation of proposition B.3 yields

$$m(z) + z m'(z) = \frac{1}{n} \sum_{r=0}^{n-1} \frac{m'(z)\psi_r}{\left(1 + m(z)\psi_r/p\right)^2}$$

thus,

$$\frac{m^2(\bar{\xi}_k)}{m'(\bar{\xi}_k)} = -\bar{\xi}_k m(\bar{\xi}_k) + \frac{1}{n}\sum_{r=0}^{n-1} \frac{m(\bar{\xi}_k)\psi_r}{\left(1+m(\bar{\xi}_k)\psi_r/p\right)^2}$$

$$= 1 - \frac{1}{n}\sum_{r=0}^{n-1} \frac{m(\bar{\xi}_k)\psi_r}{1+m(\bar{\xi}_k)\psi_r/p} + \frac{1}{n}\sum_{r=0}^{n-1} \frac{m(\bar{\xi}_k)\psi_r}{\left(1+m(\bar{\xi}_k)\psi_r/p\right)^2} \qquad \text{from proposition B.3}$$

$$\frac{m^2(\bar{\xi}_k)}{m'(\bar{\xi}_k)} = 1 - \frac{p}{n}\sum_{r=0}^{n-1} \left[\frac{m(\bar{\xi}_k)\psi_r/p}{1+m(\bar{\xi}_k)\psi_r/p}\right]^2$$

and we just need to remember that $m(\bar{\xi}_k) = \frac{-1}{\frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k}$ to get the result. □

We can now state the following proposition which defines the phase transition position as the value of $\|\boldsymbol{\mu}\|$ at which $\bar{\zeta}_k$ changes sign.

Proposition C.4 (Phase transition)**.** *For* $0 \leqslant k < n$*,*

$$\bar{\xi}_k \textit{ is a singular point of } \bar{\mathbf{Q}} \iff \bar{\zeta}_k > 0.$$

Proof. Let us consider a singular point $\bar{\xi}_k$ of $\bar{\mathbf{Q}}$.

As a Stieljes transform, m is increasing on all connected components of $\mathbb{R} \setminus \mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)$ and the restriction of its functional inverse $z(\cdot)$ to the real line, here denoted $x(\cdot)$, is also growing on every connected component of $m(\mathbb{R} \setminus \mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L))$. Then, as $\bar{\xi}_k$ is outside $\mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)$, it implies $x'\left(\frac{-1}{\frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k}\right) > 0$.

We have

$$x(m) = -\frac{1}{m} + \frac{1}{n}\sum_{l=0}^{n-1} \frac{\psi_l}{1+m\psi_l/p}$$

$$x'(m) = \frac{1}{m^2} - \frac{p}{n}\sum_{l=0}^{n-1} \left[\frac{\psi_l/p}{1+m\psi_l/p}\right]^2$$

thus

$$x'\left(\frac{-1}{\frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k}\right) > 0 \iff 1 - \frac{p}{n}\sum_{l=0}^{n-1} \left[\frac{\psi_l}{\left(\|\boldsymbol{\mu}\|^2+1\right)\psi_k - \psi_l}\right]^2 > 0.$$

Therefore, if $\bar{\xi}_k$ is a singular point of $\bar{\mathbf{Q}}$, then $\bar{\zeta}_k > 0$.

Conversely, if $\bar{\xi}_k$ is not a singular point of $\bar{\mathbf{Q}}$, then either $\psi_k = 0$ or $\bar{\xi}_k \in \mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)$. If $\psi_k = 0$, we immediately see that $\bar{\zeta}_k = \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2+1}(1-p) < 0$.

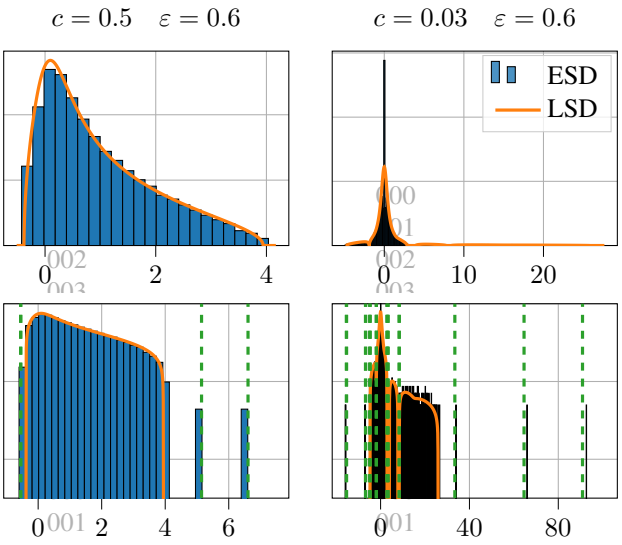
On the other hand, if $\bar{\xi}_k \in \mathrm{Sp}_\infty(\tilde{\mathbf{K}}_L)$ and $\psi_k \neq 0$ then $x'\left(\frac{-1}{\frac{\|\boldsymbol{\mu}\|^2+1}{p}\psi_k}\right) \leqslant 0$ (otherwise $\bar{\xi}_k$ would be a spike) and $\bar{\zeta}_k \leqslant 0$. □

D. Predictions with a Toeplitz mask

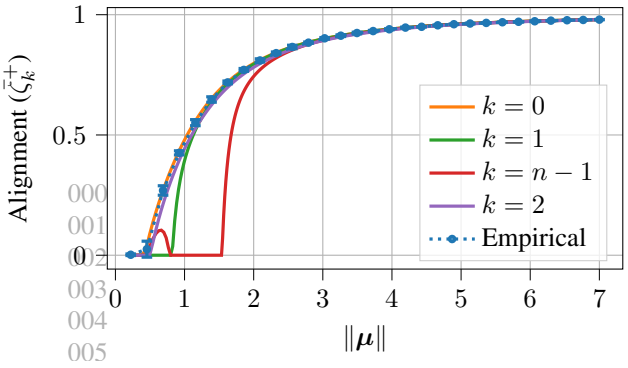
Figures 8a and 8b compare simulations with a Toeplitz mask and the predictions of theorems 3.1 and 3.3 with the ψ_k's replaced by the τ_k's and $\mathbf{F}$ replaced by $\mathbf{G}$.

Apart from extra mass around 0 in the second setting ($c = 0.03$ and $\varepsilon = 0.6$), the shape of the limiting spectral distribution is very well predicted, as well as the position of the isolated eigenvalues. Empirical alignments $|\mathbf{u}_0^* \mathbf{v}_0|^2$ are also fit well the predicted curve.

Note that, contrary to the circulant mask, the eigenvalues of $\mathbf{T}$ are mostly simple (see theorem 5 of (Trench, 1994)). Thus, we also represent $\bar{\zeta}_{n-1}^+$ is Figure 8b, which was confounded with $\bar{\zeta}_1^+$ in Figure 3 ($\psi_1 = \psi_{n-1}$ but $\tau_1 \neq \tau_{n-1}$).



(a) Empirical spectral distribution (ESD) and limiting spectral distribution (LSD) of $\mathbf{K}_L$. **Top**: noise only, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. **Bottom**: two-class mixture, $\mathbf{x}_i \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ with $\|\boldsymbol{\mu}\| = 2$. **The y-axis is in log scale**. Green dashed lines are the asymptotic positions of the spikes $\bar{\xi}_k$. **Experimental setting**: $n = 2500$, $L = 750$ and $p = 1250$ (**left**) or $p = 75$ (**right**).



(b) Asymptotic alignment $\bar{\zeta}_k^+$ between versus $\|\boldsymbol{\mu}\|$ for three values of k. The empirical alignment is computed as the mean of $|\mathbf{u}_0^* \mathbf{v}_0|^2$ on 10 realizations (error bars indicate the standard deviation). **Experimental setting**: $n = 2500$, $p = 75$, $L = 750$ ($c = 0.03$, $\varepsilon = 0.6$).

Figure 8. Predictions of theorems 3.1 and 3.3 adapted for a Toeplitz mask.

E. K-classes online kernel spectral clustering algorithm

E.1. General presentation and simulations

We use a set of spike eigenvectors $\left\{\mathbf{u}_k^{(t)}\right\}_{k\in\mathcal{K}}$ (with a set if indices $\mathcal{K}$) to estimate the $|\mathcal{K}|$-dimensional "trend" of each class. That is, denoting $\mathcal{C}[t]$ the class of $\mathbf{x}_t$, we consider the following model

$$\left[\mathbf{u}_k^{(t)}\right]_i = \left[\mathbf{h}_{k,\mathcal{C}[t-n+i]}^{(t)} + \boldsymbol{\epsilon}_k^{(t)}\right]_i$$

where, for $k \in \mathcal{K}$, $\mathbf{h}_{k,\mathcal{C}}^{(t)} \in \mathbb{R}^n$ is the "trend" of class $\mathcal{C}$ and $\boldsymbol{\epsilon}_k^{(t)}$ is a centered noise vector. A deeper analysis of the deterministic equivalent of theorem 3.1 is needed to properly understand the behavior of the vectors $\mathbf{h}_{k,\mathcal{C}}^{(t)}$. From our general understanding so far, it is expected that they are linear combinations of a few dominant eigenvectors of $\mathbf{T}$. Using this approach, we are able to estimate the trends from $\left\{\mathbf{u}_k^{(t)}\right\}_{k\in\mathcal{K}}$ (see the left part of Figure 9). Each point is then associated to the class whose curve is the nearest. The details of this algorithm are given in the following subsection.

This algorithm is tested on a stream made of $T = 21\,000$ centered raw-images from the Fashion-MNIST dataset (Xiao et al., 2017). Their dimension is $p = 784$ and we want to discriminate between `trouser`, `coat` and `ankle boot` images in an online fashion. We choose $n = 1\,000$ and $L = 100$ and we use the 5 dominant eigenvectors of $\mathbf{K}_L^{(t)}$ for the estimation.

In figure 9 are displayed the shape of the dominant eigenvector $\mathbf{u}_0^{(t)}$ at a given time during the execution of the algorithm with the estimated trends of each class[14] (left) and the mean clustering error at $t_0 + \Delta t$ of a data point seen at t_0 with the overall classification error obtained after a majority vote (right). The classification error curve is U-shaped: classes are better estimated around $t_0 + \frac{n}{2}$ than t_0 or $t_0 + n - 1$. This can be understood by the slightly-localized shape of $\mathbf{u}_0^{(t)}$ (Figure 6, bottom) — it is easier to discriminate between the trends in the middle of the eigenvector than on its edges. Nevertheless, the majority vote counteract this weakness and the overall classification error touches the bottom of the U-shape.

Remark E.1. In a binary setting, Algorithm 1 does not suffer this limitation as class estimates are directly given by the sign

[14] This is only the first dimension of a 5-dimensional trend.

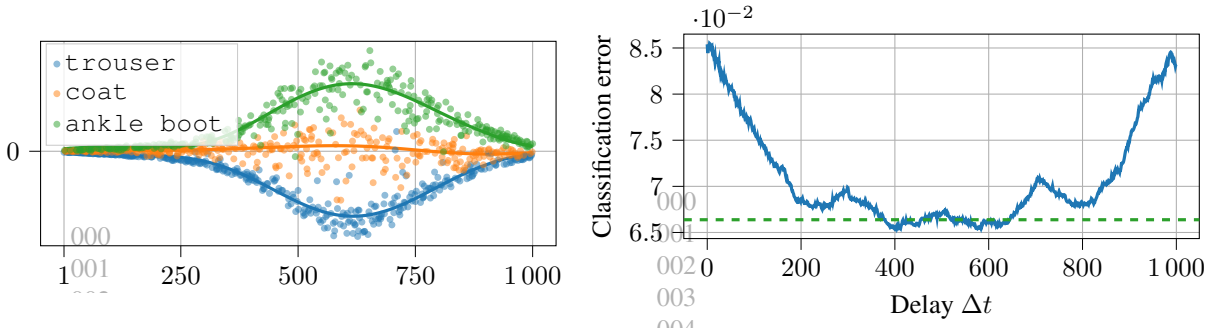


Figure 9. Clustering on Fashion-MNIST images (`trouser` vs. `coat` vs. `ankle boot`). **Left**: dominant eigenvector of $\mathbf{K}_L^{(t)}$. Solid curves are the estimated trend of each class$\mathbf{h}_{k,\mathcal{C}}^{(t)}$. **Right**: Classification error against delay Δt. This is the mean classification error at time $t_0 + \Delta t$ of a point arrived at t_0. The green dashed line indicate the overall classification error when the class is chosen by a majority vote. **Experimental setting**: $T = 21\,000$, $n = 1\,000$, $p = 784$, $L = 100$.

of the coordinates of $\mathbf{u}_0$ (no trend needs to be estimated).

Here, the overall classification error is 6.638% while a standard $T \times T$ *offline* kernel spectral clustering has only a 3.662% error rate.

E.2. Details of the algorithm

We consider a set $\mathcal{K}$ of indices of spikes and the following model for $\mathbf{u}_k^{(t)}$, $k \in \mathcal{K}$,

$$\left[\mathbf{u}_k^{(t)}\right]_i = \left[\mathbf{h}_{k,\mathcal{C}[t-n+i]}^{(t)} + \boldsymbol{\epsilon}_k^{(t)}\right]_i \qquad 1 \leqslant i \leqslant n$$

where $\mathbf{h}_{k,\mathcal{C}}^{(t)} \in \mathbb{R}^n$ is the trend of class $\mathcal{C}$ and $\boldsymbol{\epsilon}_k^{(t)}$ is a centered noise vector.

Our goal is to estimate the trend $\mathbf{h}_{k,\mathcal{C}}^{(t)}$ from the eigenvectors $\left\{\mathbf{u}_k^{(t)}\right\}_{k\in\mathcal{K}}$. Since we assume they are linear combinations of a few dominant eigenvectors of $\mathbf{T}$, we define a set of indices $\mathcal{K}_*$ specifying the eigenvectors $\{\mathbf{g}_k\}_{k\in\mathcal{K}_*}$ which we expect the $\mathbf{h}_{k,\mathcal{C}}^{(t)}$'s being linear combinations of.

We denote $\hat{\mathcal{C}}^{(t)}[s]$ the class of $\mathbf{x}_{t-n+s}$ estimated at time t.

In order to compute an estimation $\left\{\hat{\mathcal{C}}^{(t)}[i]\right\}_{1\leqslant i\leqslant n}$ of the classes at a given time t, we propose a two-step algorithm. Firstly, we compute a rough estimation $\left\{\hat{\mathcal{C}}_0^{(t)}[i]\right\}_{1\leqslant i\leqslant n}$ of the classes by following the K paths with an exponential smoothing in the coordinates of the eigenvectors $\left\{\mathbf{u}_k^{(t)}\right\}_{k\in\mathcal{K}}$, this is called the *pre-classification* step. Then, we refine this estimation with projections on $\operatorname{span}\{\mathbf{g}_k\}_{k\in\mathcal{K}_*}$, this is the *classification* step.

In the following, we drop the time dependency when it is not needed to ease notations.

E.2.1. Pre-classification step

Given the number of classes K and the eigenvectors $\{\mathbf{u}_k\}_{k\in\mathcal{K}}$, we consider the set of n points in $\mathbb{R}^{|\mathcal{K}|}$ defined by the coordinates of each eigenvector: $[\mathbf{u}_\mathcal{K}]_i = ([\mathbf{u}_k]_i)_{k\in\mathcal{K}}$ for $1 \leqslant i \leqslant n$. As i goes from 1 to n, these points draw K paths. The goal is to guess which path (and therefore which class) each point belong to.

Let us suppose we have already estimated $\hat{\mathcal{C}}_0[1], \ldots, \hat{\mathcal{C}}_0[i-1]$ and the first $i-1$ coordinates of the vectors $\left\{\tilde{\mathbf{h}}_k\right\}_{k\in\mathcal{K}}$ such that $\left[\tilde{\mathbf{h}}_k\right]_j$ is an estimation of $\left[\mathbf{h}_{k,\hat{\mathcal{C}}_0[j]}\right]_j$ (initialization is discussed later). As for $\{\mathbf{u}_k\}_{k\in\mathcal{K}}$, we see $\left\{\tilde{\mathbf{h}}_k\right\}_{k\in\mathcal{K}}$ as a set of n points in $\mathbb{R}^{|\mathcal{K}|}$, which have to be estimated. The estimation of the i-th point $\left[\tilde{\mathbf{h}}_\mathcal{K}\right]_i$ is induced by the class estimate $\hat{\mathcal{C}}_0[i]$ —

the corresponding path is updated with an exponential smoothing:

$$\left[\tilde{\mathbf{h}}_{\mathcal{K}}\right]_i = \mathcal{E}_\alpha(i, \mathbf{u}_{\mathcal{K}}, \tilde{\mathbf{h}}_{\mathcal{K}}, \hat{\mathcal{C}}_0[i]) \equiv \frac{\alpha \left[\mathbf{u}_{\mathcal{K}}\right]_i + M \left[\tilde{\mathbf{h}}_{\mathcal{K}}\right]_{I[\hat{\mathcal{C}}_0[i],i]}}{\alpha + M}$$
$$\text{where} \qquad M = \frac{1-\alpha}{i - I[\hat{\mathcal{C}}_0[i], i]} \left[1 + \frac{1-\alpha}{\alpha}\left(1 - (1-\alpha)^{i - I[\hat{\mathcal{C}}_0[i],i] - 1}\right)\right],$$

$I[\hat{\mathcal{C}}_0[i], i] = \max\left\{1 \leqslant j \leqslant i-1 \mid \hat{\mathcal{C}}_0[j] = \hat{\mathcal{C}}_0[i]\right\}$ is the index of the last seen point in $\hat{\mathcal{C}}_0[i]$ and $\alpha \in [0,1]$ is the smoothing parameter. The reasons for such a formula are detailed in appendix F.

However, $\hat{\mathcal{C}}_0[i]$ is chosen as the class which minimizes the growth of the corresponding path:

$$\hat{\mathcal{C}}_0[i] = \underset{\hat{\mathcal{C}} \in \{\hat{\mathcal{C}}_1, \dots, \hat{\mathcal{C}}_K\}}{\arg\min} \frac{\left\| \mathcal{E}_\alpha(i, \mathbf{u}_{\mathcal{K}}, \tilde{\mathbf{h}}_{\mathcal{K}}, \hat{\mathcal{C}}) - \left[\tilde{\mathbf{h}}_{\mathcal{K}}\right]_{I[\hat{\mathcal{C}}, i]} \right\|}{i - I[\hat{\mathcal{C}}, i]}.$$

Indeed, by doing so, we minimize the Lipschitz constant of the estimated trend and ensure some regularity.

From the regularity of the true trend, $\mathbf{h}_{k,\mathcal{C}}$ is almost flat on its very first coordinates. Therefore, we can initialize the values $\hat{\mathcal{C}}_0[i]$ for $1 \leqslant i \leqslant H$ with a standard clustering algorithm applied to $\{[\mathbf{u}_{\mathcal{K}}]_i\}_{1 \leqslant i \leqslant H}$. H is a parameter which should be taken as small as possible to stay in a domain where the trends are almost flat while still having a few representatives of each class. The computation of $\left\{\left[\tilde{\mathbf{h}}_{\mathcal{K}}\right]_i\right\}_{1 \leqslant i \leqslant H}$ follows from the class estimates, as presented above.

We found that a hierarchical clustering algorithm and $H \approx 10K$ worked well for the initialization. As for the smoothing parameter, a good value is $\alpha \approx 0.15$.

The pre-classification step is summarized in Algorithm 2.

Algorithm 2 Pre-classification

Input: K, $\{\mathbf{u}_k\}_{k \in \mathcal{K}}$.
Parameters: H, α.
Output: $\left\{\hat{\mathcal{C}}_0[i]\right\}_{1 \leqslant i \leqslant n}$.
Set $\hat{\mathcal{C}}_0[i]$ for $i = 1$ to H with agglomerative clustering.
for $i = 1$ to H **do**
  $\left[\tilde{\mathbf{h}}_{\mathcal{K}}\right]_i \leftarrow \mathcal{E}_\alpha(i, \mathbf{u}_{\mathcal{K}}, \tilde{\mathbf{h}}_{\mathcal{K}}, \hat{\mathcal{C}}_0[i])$
end for
for $i = H+1$ to n **do**
  $\hat{\mathcal{C}}_0[i] \leftarrow \arg\min_{\hat{\mathcal{C}} \in \{\hat{\mathcal{C}}_1, \dots, \hat{\mathcal{C}}_K\}} \frac{\left\| \mathcal{E}_\alpha(i, \mathbf{u}_{\mathcal{K}}, \tilde{\mathbf{h}}_{\mathcal{K}}, \hat{\mathcal{C}}) - \left[\tilde{\mathbf{h}}_{\mathcal{K}}\right]_{I[\hat{\mathcal{C}}, i]} \right\|}{i - I[\hat{\mathcal{C}}, i]}$
  $\left[\tilde{\mathbf{h}}_{\mathcal{K}}\right]_i \leftarrow \mathcal{E}_\alpha(i, \mathbf{u}_{\mathcal{K}}, \tilde{\mathbf{h}}_{\mathcal{K}}, \hat{\mathcal{C}}_0[i])$
end for

E.2.2. Classification step

The class estimates obtained after the pre-classification step usually are not very satisfying but still are a good basis to estimate $\mathbf{h}_{k,\mathcal{C}}$ with regressions.

In the second step of the algorithm, we are given a set $\{\mathbf{g}_k\}_{k \in \mathcal{K}_*}$ of eigenvectors of $\mathbf{T}$. It is supposed that the trends $\{\mathbf{h}_{k,\mathcal{C}}\}_{k \in \mathcal{K}}$ are mixtures of these eigenvectors.

From class estimates $\left\{\hat{\mathcal{C}}[i]\right\}_{1 \leqslant i \leqslant n}$, we can compute an estimation $\hat{\mathbf{h}}_{\mathcal{K},\hat{\mathcal{C}}}$ of the trend of each estimated class $\hat{\mathcal{C}}$ with a linear regression

$$\hat{\mathbf{h}}_{k,\hat{\mathcal{C}}} = \mathbf{v}_{\mathcal{K}_*} \boldsymbol{\beta}_{k,\hat{\mathcal{C}}} \qquad \text{where} \qquad \boldsymbol{\beta}_{k,\hat{\mathcal{C}}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{K}_*|}}{\arg\min} \left\| [\mathbf{u}_k]_{\hat{\mathcal{C}}} - [\mathbf{v}_{\mathcal{K}_*}]_{\hat{\mathcal{C}}} \boldsymbol{\beta} \right\|^2$$

where we use the notation $[\cdot]_{\hat{\mathcal{C}}}$ to represent the restriction to $\hat{\mathcal{C}}$.

Then, new class estimates can be computed by associating each point to the class whose trend is the closest:

$$\hat{\mathcal{C}}[i] = \underset{\hat{\mathcal{C}} \in \left\{\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_K\right\}}{\arg\min} \left\| [\mathbf{u}_{\mathcal{K}}]_i - \left[\hat{\mathbf{h}}_{\mathcal{K},\hat{\mathcal{C}}}\right]_i \right\|.$$

We repeat this process until convergence of the class estimates. The classification step is summarized in Algorithm 3.

Algorithm 3 Classification

Input: K, $\left\{\hat{\mathcal{C}}_0[i]\right\}_{1 \leqslant i \leqslant n}$, $\{\mathbf{u}_k\}_{k \in \mathcal{K}}$, $\{\mathbf{v}_k\}_{k \in \mathcal{K}_*}$.
Output: $\left\{\hat{\mathcal{C}}[i]\right\}_{1 \leqslant i \leqslant n}$.
for $i = 1$ to n **do**
  $\hat{\mathcal{C}}[i] \leftarrow \hat{\mathcal{C}}_0[i]$
end for
repeat
  for $\hat{\mathcal{C}} \in \left\{\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_K\right\}$ **do**
    $\hat{\mathbf{h}}_{\mathcal{K},\hat{\mathcal{C}}} \leftarrow \mathbf{v}_{\mathcal{K}_*} \left([\mathbf{v}_{\mathcal{K}_*}]_{\hat{\mathcal{C}}}^\top [\mathbf{v}_{\mathcal{K}_*}]_{\hat{\mathcal{C}}} \right)^{-1} [\mathbf{v}_{\mathcal{K}_*}]_{\hat{\mathcal{C}}}^\top [\mathbf{u}_{\mathcal{K}}]_{\hat{\mathcal{C}}}$
  end for
  for $i = 1$ to n **do**
    $\hat{\mathcal{C}}[i] \leftarrow \arg\min_{\hat{\mathcal{C}} \in \left\{\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_K\right\}} \left\| [\mathbf{u}_{\mathcal{K}}]_i - \left[\hat{\mathbf{h}}_{\mathcal{K}}^{\hat{\mathcal{C}}}\right]_i \right\|$
  end for
until convergence

E.2.3. Final algorithm

Algorithm 4 Online Kernel Spectral Clustering

Input: K, $\mathcal{K}$, $\{\mathbf{g}_k\}_{k \in \mathcal{K}_*}$.
Parameters: H, α.
Output: $\left\{\hat{\mathcal{C}}_t[s]\right\}_{\substack{1 \leqslant s \leqslant n \\ n \leqslant t \leqslant T}}$.
for $t = 1$ to T **do**
  Get a new point $\mathbf{x}_t$ into the pipeline.
  Compute $\mathbf{x}_t^* \mathbf{x}_{t-l}$ for $l = 0$ to $L-1$.
  Update $\mathbf{K}_L^{(t-1)}$ into $\mathbf{K}_L^{(t)}$.
  $\mathbf{u}_{\mathcal{K}}^{(t)} \leftarrow$ PowerIteration$(\mathbf{K}_l^{(t)}, \mathbf{u}_{\mathcal{K}}^{(t-1)})$.
  if $1 \leqslant t \leqslant n$ **then**
    Do an iteration as in Algorithm 2.
  end if
  if $t \geqslant n$ **then**
    Compute $\left\{\hat{\mathcal{C}}_t[s]\right\}_{1 \leqslant i \leqslant n}$ according to Algorithm 3 with $\left\{\hat{\mathcal{C}}_{t-1}[s]\right\}_{1 \leqslant i \leqslant n-1}$.
  end if
end for

In an online fashion, pre-classification can be performed as a warm-up during the first n time steps. Then, as $t \geqslant n$, only the classification step is needed: the classes $\left\{\hat{\mathcal{C}}_{t-1}[s]\right\}_{1 \leqslant s \leqslant n}$ estimated at $t-1$ (or during pre-classification if $t = n$) serve as a good basis to estimate the classes at time t (both $\hat{\mathcal{C}}_{t-1}[s]$ and $\hat{\mathcal{C}}_t[s+1]$ are estimates of the class of $\mathbf{x}_{t-s}$). Moreover, the few interesting eigenvectors $\mathbf{u}_{\mathcal{K}}^{(t)}$ of $\mathbf{K}_L^{(t)}$ can be quickly computed with a power iteration algorithm starting at $\mathbf{u}_{\mathcal{K}}^{(t-1)}$ (they do not differ much from one time step to another). The final algorithm is presented in Algorithm 4.

F. Exponential smoothing with missing data

Let $(\mathbf{y}_t)_{t\in\mathbb{N}}$ be a time series. Assume we want to compute its trend $(\mathbf{s}_t)_{t\in\mathbb{N}}$. A common technique to do so is to perform an exponential smoothing:

$$\mathbf{s}_0 = \mathbf{y}_0 \qquad \text{and} \qquad \mathbf{s}_{t+1} = \alpha\mathbf{y}_{t+1} + (1-\alpha)\,\mathbf{s}_t \quad \forall t \in \mathbb{N}$$

where $\alpha \in [0,1]$ is the smoothing parameter. It acts as a low-pass filter which removes high-frequency noise.

Let us now assume that we do not have access to $(\mathbf{y}_t)_{t\in\mathbb{N}}$ at each time step and we want to compute $\mathbf{s}_{t+h}$ ($h \geqslant 1$) with $\mathbf{y}_{t+h}$ and $\mathbf{s}_t$ only. Expanding the recurrence relation, we have

$$\mathbf{s}_{t+h} = \alpha\mathbf{y}_{t+h} + \alpha\sum_{k=1}^{h-1}(1-\alpha)^k\,\mathbf{y}_{t+h-k} + (1-\alpha)^h\,\mathbf{s}_t.$$

We propose to replace the unknown values $\mathbf{y}_{t+h-k}$ for $1 \leqslant k \leqslant h-1$ by the linear interpolation of the trend:

$$\begin{aligned}\mathbf{s}_{t+h} &= \alpha\mathbf{y}_{t+h} + \alpha\sum_{k=1}^{h-1}(1-\alpha)^k\left[\frac{k}{h}\mathbf{s}_t + \frac{h-k}{h}\mathbf{s}_{t+h}\right] + (1-\alpha)^h\,\mathbf{s}_t \\ &= \alpha\mathbf{y}_{t+h} + \frac{\alpha}{h}\left(\mathbf{s}_t\sum_{k=1}^{h-1}k\,(1-\alpha)^k + \mathbf{s}_{t+h}\sum_{k=1}^{h-1}k\,(1-\alpha)^{h-k}\right) + (1-\alpha)^h\,\mathbf{s}_t.\end{aligned}$$

Using the following formulae,

$$\sum_{k=1}^{h-1}k\,(1-\alpha)^k = \frac{1-\alpha}{\alpha}\left(1-h\,(1-\alpha)^{h-1}\right) + \left(\frac{1-\alpha}{\alpha}\right)^2\left(1-(1-\alpha)^{h-1}\right)$$

$$\text{and} \qquad \sum_{k=1}^{h-1}k\,(1-\alpha)^{h-k} = \frac{1-\alpha}{\alpha}\,(h-1) - \left(\frac{1-\alpha}{\alpha}\right)^2\left(1-(1-\alpha)^{h-1}\right)$$

we have

$$\left(\alpha + \frac{1-\alpha}{h}\left[1 + \frac{1-\alpha}{\alpha}\left(1-(1-\alpha)^{h-1}\right)\right]\right)\mathbf{s}_{t+h} = \alpha\mathbf{y}_{t+h} + \frac{1-\alpha}{h}\left[1 + \frac{1-\alpha}{\alpha}\left(1-(1-\alpha)^{h-1}\right)\right]\mathbf{s}_t.$$