Asymptotic Gaussian Fluctuations of Eigenvectors in Spectral Clustering

Hugo Lebeau, Florent Chatelain, Romain Couillet

Asymptotic Gaussian Fluctuations of Eigenvectors in Spectral Clustering

Hugo Lebeau, Florent Chatelain, Romain Couillet

***Abstract*—The performance of spectral clustering relies on the fluctuations of the entries of the eigenvectors of a similarity matrix, which has been left uncharacterized until now. In this letter, it is shown that the *signal* + *noise* structure of a general spike random matrix model is transferred to the eigenvectors of the corresponding Gram kernel matrix and the fluctuations of their entries are Gaussian in the large-dimensional regime. This CLT-like result was the last missing piece to precisely predict the classification performance of spectral clustering. The proposed proof is very general and relies solely on the rotational invariance of the noise. Numerical experiments on synthetic and real data illustrate the universality of this phenomenon.**

***Index Terms*—Spectral clustering, central limit theorem, kernel matrix, spike eigenvector, Gaussian fluctuations.**

I. Introduction

SPECTRAL clustering is a popular unsupervised classification technique which finds applications in many domains, such as image segmentation [1], text mining [2], and as a general purpose method for data analysis [3], [4], [5]. It relies on the spectrum of a suitably chosen similarity matrix to perform dimensionality reduction before applying a standard clustering algorithm such as K-means. Consider, e.g., the following toy example where n vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ are separated in two clusters $\mathcal{C}^+, \mathcal{C}^-$ centered around $+\boldsymbol{\mu}, -\boldsymbol{\mu}$ respectively, i.e., $\boldsymbol{x}_i = \pm\boldsymbol{\mu} + \boldsymbol{w}_i$ where $\boldsymbol{w}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then, the dominant eigenvector $\hat{\boldsymbol{v}}$ of the Gram kernel matrix $\boldsymbol{K} = \frac{1}{p}[\boldsymbol{x}_i^\top \boldsymbol{x}_j]_{1 \leqslant i,j \leqslant n}$ is an information-theoretically optimal estimator [6] of the vector $\frac{1}{\sqrt{n}}\boldsymbol{j}$ such that $j_i = \pm 1$ if $\boldsymbol{x}_i \in \mathcal{C}^\pm$. In this case, clustering is achieved with the trivial decision rule $\boldsymbol{x}_i \to \mathcal{C}^\pm$ if $\hat{v}_i \gtrless 0$.

The achievable performances of spectral clustering can be theoretically predicted thanks to the study of random matrix models corresponding to similarity matrices. For this purpose, random matrix theory offers powerful tools [7], [8], [9]. In particular, it allows to derive the limiting spectral distribution of the kernel matrix and to predict the position of isolated eigenvalues in *spiked* random matrix models [10], [11], [12]. The latter are of particular importance as, in a wide range of problems, the information of interest can be modeled as a low-rank signal corrupted with noise. In our previous toy example, the data matrix $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \ldots & \boldsymbol{x}_n\end{bmatrix}$ is a rank-one perturbation $\boldsymbol{\mu}\boldsymbol{j}^\top$ of a noise matrix $\boldsymbol{W}$ with i.i.d. $\mathcal{N}(0,1)$ entries. In order to theoretically predict the error rate of spectral clustering for a given signal-to-noise ratio, one must therefore study the behavior of the dominant eigenvectors of the similarity matrix. Tools such as the ones used in [12], [13] allow to express the quality of their alignment with the true underlying signal, i.e., $\frac{1}{\sqrt{n}}|\boldsymbol{j}^\top \hat{\boldsymbol{v}}|$. Although this tells us when an estimation of the signal is possible (depending on the signal-to-noise ratio) and its efficiency, a precise characterization of the fluctuations of the entries of spiked eigenvectors still lacks to rigorously predict the error rate of spectral clustering. Indeed, in our toy example, the expected error rate $\mathbb{P}(\hat{v}_i j_i < 0)$ cannot be expressed unless the law of $\hat{\boldsymbol{v}}$ is known.

Yet, it is often stated that the entries of $\hat{\boldsymbol{v}}$ have Gaussian fluctuations in the large-dimensional regime, so that $\mathbb{P}(\hat{v}_i j_i < 0)$ is a Gaussian integral. In [14], this result is formally stated but no proof is given. Hence, we fill this missing gap with a rigorous proof of this phenomenon for a general spiked random matrix model. Although we stick to a simple *signal* + *noise* model here, the proposed proof is not restricted to Gaussian noise (in fact, the noise only needs to be rotationally invariant) and can easily be adapted to most standard spiked models (such as, notably, the general model considered in [11]). Our result and its proof thus support a wide range of previous works studying the performance of spectral algorithms. The demonstration can be summarized in two simple facts 1) an eigenvector of the kernel matrix can be decomposed into the sum of a deterministic signal part and a random noise part 2) the random part is uniformly distributed on a certain sphere, hence any finite subset of its entries tends to a centered Gaussian vector in the large-dimensional limit.

In this letter, we consider a general *signal* + *noise* random matrix model and briefly recall known results regarding its limiting spectral distribution and the behavior of its dominant eigenvalues and eigenvectors. Then, we show that the entries of the kernel eigenvectors indeed have Gaussian fluctuations in the large-dimensional regime. We present a short, self-contained and general proof which is our main contribution. Finally, we illustrate this result with numerical experiments on synthetic and real data.

Simulations. Python codes to reproduce simulations are available in the following GitHub repository https://github.com/HugoLebeau/asymptotic_fluctuations_spectral_clustering.

II. Model and Main Result

A. Notations

The symbols a, $\boldsymbol{a}$ and $\boldsymbol{A}$ respectively denote a scalar, a vector and a matrix. Their entries are a_i and $A_{i,j}$. The set of positive integers below n is $[n] = \{1, \ldots, n\}$. The

H. Lebeau and R. Couillet are with Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38 000 Grenoble, France, (e-mail: {hugo.lebeau, romain.couillet}@univ-grenoble-alpes.fr).

F. Chatelain is with Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38 000 Grenoble, France (e-mail: florent.chatelain@grenoble-inp.fr).

cardinality of a set $\mathcal{E}$ is $|\mathcal{E}|$. Given an ordered set of indices $\mathcal{I} = (i_1, \dots, i_{|\mathcal{I}|})$, $[\boldsymbol{a}]_{\mathcal{I}}$ is the vector $\begin{bmatrix} a_{i_1} & \dots & a_{i_{|\mathcal{I}|}} \end{bmatrix}^\top$. The norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$. The unit sphere in $\mathbb{R}^n$ is $\mathbb{S}^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$. The $n \times n$ identity matrix is $\boldsymbol{I}_n$. For a real number $x \in \mathbb{R}$, $[x]^+ = \max(0, x)$ and δ_x is the Dirac measure at x. The imaginary unit is denoted i. If the random variable X follows the law $\mathcal{L}$, we write $X \sim \mathcal{L}$. The convergence in distribution of a sequence of random variables $(X_n)_{n \geqslant 0}$ to $\mathcal{L}$ is denoted $X_n \xrightarrow[n \to +\infty]{\mathcal{D}} \mathcal{L}$. Its almost sure convergence to L is denoted $X_n \xrightarrow[n \to +\infty]{\text{a.s.}} L$. The multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is denoted $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The set of eigenvalues of a square matrix $\boldsymbol{A}$ is its spectrum, $\operatorname{Sp} \boldsymbol{A}$. Given two real-valued sequences $(u_n)_{n \geqslant 0}$ and $(v_n)_{n \geqslant 0}$, we write $u_n = \mathcal{O}(v_n)$ if $|u_n / v_n|$ is bounded as $n \to +\infty$ and $u_n \asymp v_n$ if $u_n / v_n \to 1$.

B. Spiked Matrix Model

Consider the following statistical model

$$\boldsymbol{X} = \boldsymbol{P} + \boldsymbol{W} \in \mathbb{R}^{p \times n}, \quad \boldsymbol{P} = \boldsymbol{L} \boldsymbol{V}^\top \tag{1}$$

with $\boldsymbol{L} \in \mathbb{R}^{p \times K}$ and $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_K \end{bmatrix} \in \mathbb{R}^{n \times K}$ such that $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_K$. It models a *low-rank* signal $\boldsymbol{P}$ corrupted by additive Gaussian noise $W_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. In a spectral clustering perspective, K represents the number of classes and $\boldsymbol{P} = \boldsymbol{M} \boldsymbol{J}^\top$ where $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\mu}_1 & \dots & \boldsymbol{\mu}_K \end{bmatrix}$ is a matrix gathering the K cluster means and $J_{i,k} = 1$ if $\boldsymbol{x}_i$ is in the k-th cluster (i.e., $\boldsymbol{x}_i = \boldsymbol{\mu}_k + \boldsymbol{w}_i$) and 0 otherwise. This is congruent with model (1): define the $K \times K$ diagonal matrix $\boldsymbol{D}$ such that $D_{k,k} = n_k$ where n_k is the number of samples belonging to the k-th cluster, then $\boldsymbol{L} = \boldsymbol{M} \boldsymbol{D}^{1/2}$ and $\boldsymbol{V} = \boldsymbol{J} \boldsymbol{D}^{-1/2}$.

Given model (1), we are interested in the reconstruction of $\boldsymbol{V}$ from the dominant eigenvectors of the Gram kernel matrix $\boldsymbol{K} = \frac{1}{p} \boldsymbol{X}^\top \boldsymbol{X}$. We study this problem in the regime where K is fixed and $p, n \to +\infty$ at the same rate, i.e., $0 < c \overset{\text{def}}{=} \limsup p/n < +\infty$. This models the fact that, in practice, the number of samples n is comparable to the number of features p and they are both large. Moreover, we make the following assumptions.

Assumption 1. All classes are of comparable size, i.e., $\liminf n_k / n > 0$ as $p, n \to +\infty$ for all $k \in [K]$.

Assumption 2. $\lim_{p,n \to +\infty} \max_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant k \leqslant K}} \sqrt{n} V_{i,k}^2 = 0$.

Assumption 3. As $n \to +\infty$, the eigenvalues $\ell_1 \geqslant \dots \geqslant \ell_K > 0$ of $\frac{1}{n} \boldsymbol{L}^\top \boldsymbol{L}$ are not degenerate (i.e., have multiplicity one) and the columns of $\boldsymbol{V}$ are ordered accordingly.

Assumption 3 is only to simplify the presentation of the results so it is not necessary, but often verified in practice. However, Assumption 2 states that $\boldsymbol{V}$ must be *delocalized*, i.e., not sparse. It is naturally verified for spectral clustering as a result of Assumption 1 since $\boldsymbol{V} = \boldsymbol{J} \boldsymbol{D}^{-1/2}$, but the results presented below concern the statistical model (1), which is more general and also encompasses PCA for example [13].

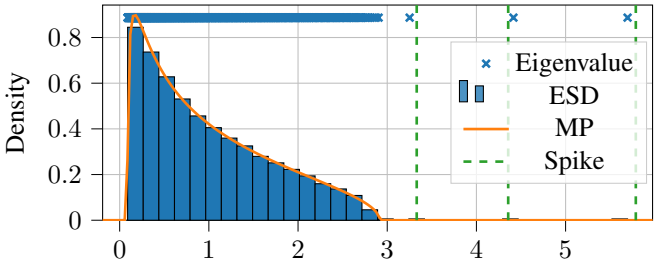

Fig. 1. Empirical Spectral Distribution (ESD) of $\boldsymbol{K} = \frac{1}{p} \boldsymbol{X}^\top \boldsymbol{X}$ and Marčenko-Pastur Distribution (MP). The green dashed lines are the positions ξ_k of isolated eigenvalues predicted by Theorem 1. **Experimental setting**: $n = 1000$, $p = 2000$, $K = 3$, $(n_1, n_2, n_3) = (333, 334, 333)$, $(\|\boldsymbol{\mu}_1\|, \|\boldsymbol{\mu}_2\|, \|\boldsymbol{\mu}_3\|) = (3, 4, 5)$.

C. Eigenvalue Distribution and Spiked Eigenvalues

We briefly recall known results on model (1) in order to set the ground for our main result in Theorem 2.

Firstly, the empirical spectral distribution of $\boldsymbol{K}$, that is $\frac{1}{n} \sum_{\lambda \in \operatorname{Sp} \boldsymbol{K}} \delta_\lambda$, converges weakly almost surely to the Marčenko-Pastur distribution $\mu_{\text{MP}} = [1-c]^+ \delta_0 + \nu$ where ν has density supported on $[E_-, E_+]$ with $E_\pm = (1 \pm \sqrt{c^{-1}})^2$ [15], [7], [8], [9]. In other words, as $p, n \to +\infty$, the histogram of eigenvalues of $\boldsymbol{K}$ approaches μ_{MP}, as depicted in Figure 1.

Due to the low-rank perturbation $\boldsymbol{P}$, the K dominant eigenvalues of $\boldsymbol{K}$ may isolate themselves from the bulk characterized by the Marčenko-Pastur distribution if their corresponding signal-to-noise ratios (the eigenvalues of $\frac{1}{n} \boldsymbol{L}^\top \boldsymbol{L}$) are large enough — they are then called *spikes*. Their behavior is specified in the following theorem, which is a particular case of Theorem 2 in [13].

Theorem 1 (Spikes)**.** *Let $(\lambda_k, \hat{\boldsymbol{v}}_k)_{k \in [K]}$ denote the dominant eigenvalue-eigenvector pairs of $\boldsymbol{K}$ such that $\lambda_1 \geqslant \dots \geqslant \lambda_K$. Then, for all $k \in [K]$,*

$$\lambda_k \xrightarrow[p,n \to +\infty]{\text{a.s.}} \xi_k \overset{\text{def}}{=} \begin{cases} \frac{(\ell_k + c)(\ell_k + 1)}{\ell_k c} & \text{if } \ell_k > \sqrt{c} \\ E_+ & \text{otherwise} \end{cases},$$

$$\left| \boldsymbol{v}_k^\top \hat{\boldsymbol{v}}_k \right|^2 \xrightarrow[p,n \to +\infty]{\text{a.s.}} \zeta_k \overset{\text{def}}{=} \begin{cases} 1 - \frac{\ell_k + c}{\ell_k (\ell_k + 1)} & \text{if } \ell_k > \sqrt{c} \\ 0 & \text{otherwise} \end{cases}.$$

This result states that λ_k leaves the bulk if, and only if, $\ell_k > \sqrt{c}$ (this is a well-known phase transition phenomenon [16]) and further gives its almost sure asymptotic position ξ_k. This is illustrated in Figure 1. Moreover, Theorem 1 indicates the almost sure asymptotic alignment ζ_k of the corresponding eigenvector $\hat{\boldsymbol{v}}_k$ with the underlying signal $\boldsymbol{v}_k$. This is depicted in Figure 2 (top row).

It should be noted that the previous results are not restricted to Gaussian noise: up to a control on the moments of the distribution, they can be generalized thanks to an "interpolation trick" [17, Corollary 3.1]. In addition, a similar spectral behavior is observed with non-i.i.d. noise following the realistic assumption that it is *concentrated* [18], [19].

D. Fluctuations of Spiked Eigenvectors Entries

The convergence of $|\boldsymbol{v}_k^\top \hat{\boldsymbol{v}}_k|^2$ to ζ_k stated in Theorem 1 is an important result which justifies the use of the dominant

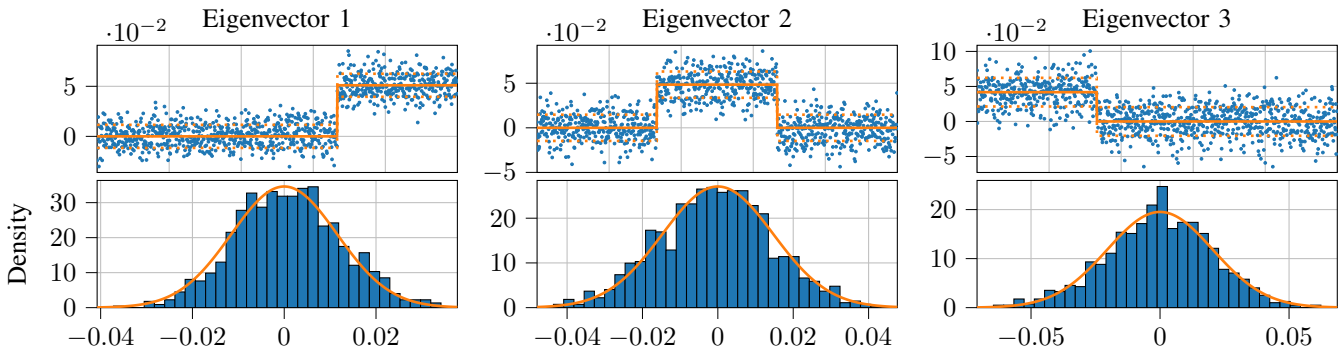

Fig. 2. Dominant eigenvectors of $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$. **Top**: Coordinates of $\hat{\boldsymbol{v}}_k$ (blue) and the underlying signal $\sqrt{\zeta_k}\boldsymbol{v}_k$ (orange) with ζ_k given in Theorem 1. The dotted orange lines are the $\pm 1\sigma$-error curves deduced from Theorem 2. **Bottom**: Histogram of the entries of $\hat{\boldsymbol{v}}_k - \sqrt{\zeta_k}\boldsymbol{v}_k$ (blue) and probability density function of $\mathcal{N}(0, \frac{1-\zeta_k}{n})$ (orange). **Experimental setting**: like in Figure 1.

eigenvectors of $\boldsymbol{K}$ as estimators of the underlying signal $\boldsymbol{V}$. Yet, it is not enough to characterize its reconstruction performance. Indeed, the fluctuations of the entries of $\hat{\boldsymbol{v}}_k$ must be known to fully characterize *how* it is aligned with $\boldsymbol{v}_k$.

Consider, e.g., the multi-class spectral clustering problem with $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{J}^\top$. Here, $[\boldsymbol{v}_k]_i = J_{i,k}/\sqrt{n_k}$. Hence, $\boldsymbol{x}_i$ is classified in the k-th class if $[\hat{\boldsymbol{v}}_k]_i > [\hat{\boldsymbol{v}}_{k'}]_i$ for all $k' \neq k$. The reconstruction performance thus depends on the probability of correct classification $\mathbb{P}([\hat{\boldsymbol{v}}_k]_i > [\hat{\boldsymbol{v}}_{k'}]_i \mid J_{i,k} = 1)$. In the theorem below, we show that the entries of $\hat{\boldsymbol{v}}_k$ asymptotically have Gaussian fluctuations around those of $\boldsymbol{v}_k$ with variance $(1-\zeta_k)/n$, as illustrated in the bottom row of Figure 2.

Theorem 2. *For all finite ordered set of indices $\mathcal{I} = (i_1, \ldots, i_{|\mathcal{I}|}) \subset [n]$ and $k \in [K]$,*

$$\frac{\sqrt{n}\left[\hat{\boldsymbol{v}}_k - \sqrt{\zeta_k}\boldsymbol{v}_k\right]_{\mathcal{I}}}{\sqrt{1-\zeta_k}} \xrightarrow[p,n\to+\infty]{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{|\mathcal{I}|})$$

with $\hat{\boldsymbol{v}}_k$ such that $\boldsymbol{v}_k^\top \hat{\boldsymbol{v}}_k \geqslant 0$ (otherwise, consider $-\hat{\boldsymbol{v}}_k$).

This result invokes the quantity ζ_k introduced in Theorem 1, which quantifies the alignment of $\hat{\boldsymbol{v}}_k$ with $\boldsymbol{v}_k$. Theorem 2 specifies that $[\hat{\boldsymbol{v}}_k]_{\mathcal{I}}$ behaves like $\mathcal{N}(\sqrt{\zeta_k}[\boldsymbol{v}_k]_{\mathcal{I}}, \frac{1-\zeta_k}{n}\boldsymbol{I}_{|\mathcal{I}|})$ in the large-dimensional regime. That is, the more $\hat{\boldsymbol{v}}_k$ is aligned with $\boldsymbol{v}_k$ (i.e., the closer ζ_k is to 1), the more it concentrates around $\sqrt{\zeta_k}\boldsymbol{v}_k$, since the variance is $(1-\zeta_k)/n$. Furthermore, the entries of $[\hat{\boldsymbol{v}}_k]_{\mathcal{I}}$ are *asymptotically independent* for any *finite* ordered set of indices $\mathcal{I}$. In the multi-class spectral clustering problem considered above, since $\hat{\boldsymbol{v}}_k$ and $\hat{\boldsymbol{v}}_{k'}$ are asymptotically independent if $k' \neq k$ [20, Theorem 4], Theorem 2 yields

$$\mathbb{P}([\hat{\boldsymbol{v}}_k]_i > [\hat{\boldsymbol{v}}_{k'}]_i \mid J_{i,k} = 1) \asymp \Phi\left(\sqrt{\frac{n}{n_k}\frac{\zeta_k}{2-(\zeta_k+\zeta_{k'})}}\right)$$

where $\Phi : x \mapsto \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\mathrm{d}t$ is the Gaussian cumulative distribution function.

We prove Theorem 2 in Section III below. The proof hinges on the rotational invariance of the noise (Lemma 1). In fact, it does not need the entries of $\boldsymbol{W}$ to be distributed according to the Gaussian law, but only that its distribution be invariant under isometries. This makes it a very general proof, which can easily be adapted to most standard spiked models as those discussed, e.g., in [9, §2.5.4].

III. Proof of Main Result

Consider the tangent-normal decomposition

$$\hat{\boldsymbol{v}}_k = \sum_{\kappa=1}^{K} \tau_\kappa \boldsymbol{v}_\kappa + \sqrt{1-\|\boldsymbol{\tau}\|^2}\,\hat{\boldsymbol{v}}_k^\sharp \tag{2}$$

where $\hat{\boldsymbol{v}}_k^\sharp = (\boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\top)\frac{\hat{\boldsymbol{v}}_k}{\sqrt{1-\|\boldsymbol{\tau}\|^2}}$ is a unit-norm vector orthogonal to the span of $\boldsymbol{V}$ and $\boldsymbol{\tau} = \begin{bmatrix}\tau_1 & \ldots & \tau_K\end{bmatrix}^\top$ with $\tau_\kappa = \boldsymbol{v}_\kappa^\top \hat{\boldsymbol{v}}_k$ measuring the cosine between $\boldsymbol{v}_\kappa$ and $\hat{\boldsymbol{v}}_k$. Let $\boldsymbol{O}$ be an $n \times n$ orthogonal matrix such that $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$ — i.e., a rotational symmetry about the span of $\boldsymbol{V}$ — and $\tilde{\boldsymbol{K}} \overset{\text{def}}{=} \boldsymbol{O}\boldsymbol{K}\boldsymbol{O}^\top$. Then, since $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{V}^\top + \boldsymbol{W}$,

$$\tilde{\boldsymbol{K}} = \frac{1}{p}\Big([\boldsymbol{O}\boldsymbol{V}]\boldsymbol{L}^\top\boldsymbol{L}[\boldsymbol{O}\boldsymbol{V}]^\top + [\boldsymbol{O}\boldsymbol{V}]\boldsymbol{L}^\top[\boldsymbol{W}\boldsymbol{O}^\top] + [\boldsymbol{W}\boldsymbol{O}^\top]^\top\boldsymbol{L}[\boldsymbol{O}\boldsymbol{V}]^\top + [\boldsymbol{W}\boldsymbol{O}^\top]^\top[\boldsymbol{W}\boldsymbol{O}^\top]\Big).$$

Lemma 1. *$\boldsymbol{W}$ and $\boldsymbol{W}\boldsymbol{O}^\top$ are identically distributed.*

Proof. The distribution of $[\boldsymbol{W}\boldsymbol{O}^\top]_{i,j} = \sum_{k=1}^{n} W_{i,k}O_{j,k}$ is $\mathcal{N}(0,1)$ and $\mathrm{Cov}([\boldsymbol{W}\boldsymbol{O}^\top]_{i,j}, [\boldsymbol{W}\boldsymbol{O}^\top]_{i',j'})$ is 1 if $(i,j) = (i',j')$ and 0 otherwise. Hence $[\boldsymbol{W}\boldsymbol{O}^\top]_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. □

According to the previous lemma, $\tilde{\boldsymbol{K}}$ follows the same model as $\boldsymbol{K}$ since $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$. Therefore, its k-th dominant eigenvector can likewise be decomposed as

$$\tilde{\boldsymbol{v}}_k = \sum_{\kappa=1}^{K} \tilde{\tau}_\kappa \boldsymbol{v}_\kappa + \sqrt{1-\|\tilde{\boldsymbol{\tau}}\|^2}\,\tilde{\boldsymbol{v}}_k^\sharp$$

with $\tilde{\tau}_\kappa = \boldsymbol{v}_\kappa^\top\tilde{\boldsymbol{v}}_k$ and $\tilde{\boldsymbol{v}}_k^\sharp = (\boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\top)\frac{\tilde{\boldsymbol{v}}_k}{\sqrt{1-\|\tilde{\boldsymbol{\tau}}\|^2}}$ *identically distributed to* $\hat{\boldsymbol{v}}_k^\sharp$. Yet, $\tilde{\boldsymbol{v}}_k = \boldsymbol{O}\hat{\boldsymbol{v}}_k$. Thus, $\hat{\boldsymbol{v}}_k^\sharp$ and $\boldsymbol{O}\hat{\boldsymbol{v}}_k^\sharp$ are identically distributed for all $n \times n$ orthogonal matrix $\boldsymbol{O}$ such that $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$. Consequently, denoting η the probability distribution of $\hat{\boldsymbol{v}}_k^\sharp$ and $\boldsymbol{V}^\perp = \{\boldsymbol{w} \in \mathbb{R}^n \mid \boldsymbol{V}^\top\boldsymbol{w} = \boldsymbol{0}\}$, then, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{n-1} \cap \boldsymbol{V}^\perp$, we have $\mathrm{d}\eta(\boldsymbol{x}) = \mathrm{d}\eta(\boldsymbol{O}\boldsymbol{x}) = \mathrm{d}\eta(\boldsymbol{y})$ with $\boldsymbol{O} = \boldsymbol{I}_n - \frac{(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^\top}{1-\boldsymbol{x}^\top\boldsymbol{y}}$ satisfying $\boldsymbol{O}\boldsymbol{V} = \boldsymbol{V}$. This shows that $\hat{\boldsymbol{v}}_k^\sharp$ is uniformly distributed on $\mathbb{S}^{n-1} \cap \boldsymbol{V}^\perp$.

Then, $\hat{\boldsymbol{v}}_k^\sharp$ can be written as $\boldsymbol{U}\boldsymbol{u}$ where $\boldsymbol{u}$ is uniformly distributed on $\mathbb{S}^{n-1-K} \subset \mathbb{R}^{n-K}$ and $\boldsymbol{U} \in \mathbb{R}^{n\times(n-K)}$ is such that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{n-K}$ and $\boldsymbol{U}^\top\boldsymbol{V} = \boldsymbol{0}$ (the columns of $\boldsymbol{U}$ form

an orthonormal basis of $\boldsymbol{V}^\perp$ in $\mathbb{R}^n$). We use the following theorem to identify the asymptotic distribution of $\sqrt{n}[\hat{\boldsymbol{v}}_k^\sharp]_\mathcal{I}$.

Theorem 3 ([21], [22])**.** *The characteristic function of a vector $\boldsymbol{w}$ uniformly distributed on $\mathbb{S}^{n-1}$ is given by $\varphi_{\boldsymbol{w}}(\boldsymbol{t}) \overset{def}{=} \mathbb{E}[e^{\mathrm{i}\boldsymbol{t}^\top \boldsymbol{w}}] = \Omega_n(\|\boldsymbol{t}\|)$ where Ω_n is such that $r \mapsto \Omega_n(r\sqrt{n})$ converges uniformly in $r \geqslant 0$ to $r \mapsto e^{-\frac{r^2}{2}}$ as $n \to +\infty$.*

Let $\boldsymbol{t} \in \mathbb{R}^n$ be such that $t_i = 0$ if $i \notin \mathcal{I}$. The characteristic function of $\sqrt{n}[\hat{\boldsymbol{v}}_k^\sharp]_\mathcal{I}$ is

$$\begin{aligned}\varphi_{\sqrt{n}[\hat{\boldsymbol{v}}_k^\sharp]_\mathcal{I}}(t_{i_1}, \ldots, t_{i_{|\mathcal{I}|}}) &= \mathbb{E}\left[e^{\mathrm{i}\sqrt{n}\boldsymbol{t}^\top \boldsymbol{U}\boldsymbol{u}}\right] \\ &= \Omega_{n-K}(\sqrt{n}\left\|\boldsymbol{U}^\top \boldsymbol{t}\right\|)\end{aligned}$$

and $\|\boldsymbol{U}^\top \boldsymbol{t}\| = \sqrt{\|\boldsymbol{t}\|^2 - \|\boldsymbol{V}^\top \boldsymbol{t}\|^2} = \|\boldsymbol{t}\| + \mathcal{O}(\|\boldsymbol{V}^\top \boldsymbol{t}\|^2)$. According to Assumption 2, $\sqrt{n}\|\boldsymbol{V}^\top \boldsymbol{t}\|^2 \to 0$ as $p, n \to +\infty$, thus $\Omega_{n-K}(\sqrt{n}\|\boldsymbol{U}^\top \boldsymbol{t}\|) = \Omega_{n-K}(\sqrt{n-K}\|\boldsymbol{t}\| + \epsilon_n)$ with $\epsilon_n \to 0$ as $p, n \to +\infty$ and

$$\begin{aligned}\left|\Omega_{n-K}(\sqrt{n-K}\,\|\boldsymbol{t}\| + \epsilon_n) - e^{-\frac{1}{2}\|\boldsymbol{t}\|^2}\right| \leqslant \\ \left|\Omega_{n-K}(\sqrt{n-K}\,\|\boldsymbol{t}\| + \epsilon_n) - e^{-\frac{1}{2}\left[\|\boldsymbol{t}\| + \epsilon_n/\sqrt{n-K}\right]^2}\right| \\ + \left|e^{-\frac{1}{2}\left[\|\boldsymbol{t}\| + \epsilon_n/\sqrt{n-K}\right]^2} - e^{-\frac{1}{2}\|\boldsymbol{t}\|^2}\right|.\end{aligned}$$

As $p, n \to +\infty$, the first term vanishes from the uniform convergence given in Theorem 3 and the second term vanishes by continuity. Therefore, $\varphi_{\sqrt{n}[\hat{\boldsymbol{v}}_k^\sharp]_\mathcal{I}}(t_{i_1}, \ldots, t_{i_{|\mathcal{I}|}}) \to e^{-\frac{\|\boldsymbol{t}\|^2}{2}}$ and, by Lévy's continuity theorem [23], we can conclude that $\sqrt{n}[\hat{\boldsymbol{v}}_k^\sharp]_\mathcal{I} \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{|\mathcal{I}|})$ as $p, n \to +\infty$. And, finally, given decomposition (2), Theorem 1 and the independence of $\hat{\boldsymbol{v}}_k$ and $\hat{\boldsymbol{v}}_{k'}$ if $k' \neq k$ [20, Theorem 4],

$$\hat{\boldsymbol{v}}_k = \sqrt{\zeta_k}\boldsymbol{v}_k + \sqrt{1-\zeta_k}\hat{\boldsymbol{v}}_k^\sharp + \boldsymbol{\varepsilon} \quad \text{with} \quad \|\boldsymbol{\varepsilon}\| \xrightarrow[p,n\to+\infty]{\text{a.s.}} 0.$$

This concludes the proof of Theorem 2.

IV. Numerical Experiments

To illustrate this result, we conduct a first experiment on synthetic data following model (1) with $K = 3$ classes of equal size and $(\|\boldsymbol{\mu}_1\|, \|\boldsymbol{\mu}_2\|, \|\boldsymbol{\mu}_3\|) = (3, 4, 5)$. The $\boldsymbol{x}_i$'s are ordered by class. The spectral distribution of $\boldsymbol{K}$ is plotted in Figure 1 and Figure 2 shows the dominant eigenvectors with the histograms of residuals $\hat{\boldsymbol{v}}_k - \sqrt{\zeta_k}\boldsymbol{v}_k$. We observe a very good fit of the latter to the probability density function of $\mathcal{N}(0, \frac{1-\zeta_k}{n})$ — the $\hat{\boldsymbol{v}}_k$'s exactly correspond to a deterministic signal $\sqrt{\zeta_k}\boldsymbol{v}_k$ corrupted by additive centered Gaussian noise. The *signal* + *noise* structure of model (1) has been transferred to the spectral estimator of $\boldsymbol{V}$.

Then, we conduct a second experiment on the Fashion-MNIST dataset [24] consisting of 28×28 images of clothes separated in 10 classes of size 7 000 each. We select two classes k_1, k_2 and perform binary spectral clustering using the dominant eigenvector of $\boldsymbol{K} = \frac{1}{p}\boldsymbol{X}^\top \boldsymbol{X}$ where the columns of $\boldsymbol{X}$ are the 784 pixels of the images from classes k_1 and k_2. The dimension of $\boldsymbol{X}$ is thus $784 \times 14\,000$. Here, we assume a similar model as our toy example in the introduction: $\boldsymbol{X} = \boldsymbol{\mu}\boldsymbol{j}^\top + \boldsymbol{W}$ where $j_i = \pm 1$ depending on the class of the i-th image. Thus, according to Theorem 2,

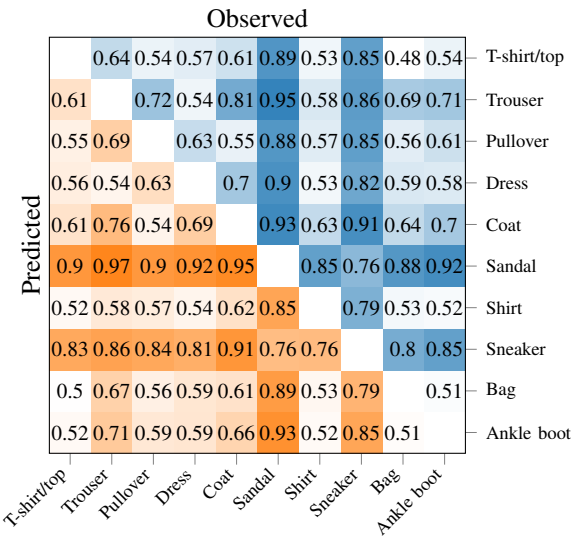

Fig. 3. Observed (upper right, blue) and predicted (lower left, orange) classification accuracies of binary spectral clustering on the Fashion-MNIST dataset [24].

the i-th entry of the dominant eigenvector $\hat{\boldsymbol{v}}$ asymptotically follows $\mathcal{N}(\sqrt{\zeta}\frac{j_i}{\sqrt{n}}, \frac{1-\zeta}{n})$ and, given $\boldsymbol{j}$, ζ can be estimated as $(\sum_{i=1}^n \frac{j_i}{\sqrt{n}}\hat{v}_i)^2$. We can then compare the observed accuracy $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\hat{v}_i j_i > 0}$ to the one expected from Theorem 2, $\mathbb{P}(\hat{v}_i j_i > 0) \asymp \Phi(\sqrt{\frac{\zeta}{1-\zeta}})$. The results are presented in Figure 3 for each pair of classes k_1, k_2.

We find a very good agreement between the observed and predicted accuracies, regardless of whether the problem is easy (e.g., Trouser vs Sandal) or hard (e.g., Bag vs Ankle Boot). This observation confirms the general scope of Theorem 2: starting from real data $\boldsymbol{X}$ which is clearly *not* Gaussian, the normal distribution naturally emerges in the fluctuations of the entries of the large-dimensional eigenvector $\hat{\boldsymbol{v}}$.

V. Conclusion

After recalling known results on spectral clustering under a general *signal* + *noise* random matrix model, we have shown that the entries of spiked eigenvectors have Gaussian fluctuations in the large-dimensional regime. This formalizes and clearly states a result which is often implicitly assumed in many problems, without ever being actually proven. The proposed proof relies solely on the rotational invariance of the noise. It is thus very general and can easily be extended to most standard spike models. Numerical experiments have demonstrated the universality of this phenomenon: the Gaussian behavior of the entries of spike eigenvectors can even be observed on real unprocessed data. This allows to accurately predict the classification performance of spectral clustering. An interesting problem for future work is to understand how these results extend to more exotic spike models such as [25].

REFERENCES

[1] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, Aug. 2000, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: https://ieeexplore.ieee.org/document/868688

[2] C. Brew and S. S. im Walde, "Spectral clustering for german verbs," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002, pp. 117–124. [Online]. Available: https://aclanthology.org/W02-1016.pdf

[3] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, vol. 14. MIT Press, 2002. [Online]. Available: https://papers.nips.cc/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html

[4] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007. [Online]. Available: https://doi.org/10.1007/s11222-007-9033-z

[5] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in *Proceedings 2001 IEEE International Conference on Data Mining*, Nov. 2001, pp. 107–114. [Online]. Available: https://ieeexplore.ieee.org/document/989507

[6] A. Onatski, M. J. Moreira, and M. Hallin, "Asymptotic power of sphericity tests for high-dimensional data," *The Annals of Statistics*, vol. 41, no. 3, pp. 1204–1231, Jun. 2013, publisher: Institute of Mathematical Statistics. [Online]. Available: https://projecteuclid.org/journals/annals-of-statistics/volume-41/issue-3/Asymptotic-power-of-sphericity-tests-for-high-dimensional-data/10.1214/13-AOS1100.full

[7] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*. Springer, 2010, vol. 20.

[8] L. A. Pastur and M. Shcherbina, *Eigenvalue Distribution of Large Random Matrices*, ser. Mathematical Surveys and Monographs. American Mathematical Society, 2011, no. 171.

[9] R. Couillet and Z. Liao, *Random Matrix Methods for Machine Learning*. Cambridge: Cambridge University Press, 2022. [Online]. Available: https://www.cambridge.org/core/books/random-matrix-methods-for-machine-learning/6B681EB69E58B5F888EDB689C160C682

[10] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X0500134X

[11] F. Benaych-Georges and R. R. Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Advances in Mathematics*, vol. 227, no. 1, pp. 494–521, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001870811000570

[12] R. Couillet and F. Benaych-Georges, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016, publisher: Shaker Heights, OH : Institute of Mathematical Statistics. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01215343

[13] R. Couillet, F. Chatelain, and N. L. Bihan, "Two-way kernel matrix puncturing: towards resource-efficient PCA and spectral clustering," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 2156–2165, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v139/couillet21a.html

[14] A. Kadavankandy and R. Couillet, "Asymptotic Gaussian Fluctuations of Spectral Clustering Eigenvectors," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec. 2019, pp. 694–698.

[15] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, p. 457, 1967, publisher: IOP Publishing.

[16] J. Baik, G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *The Annals of Probability*, vol. 33, no. 5, pp. 1643–1697, Sep. 2005, publisher: Institute of Mathematical Statistics. [Online]. Available: https://projecteuclid.org/journals/annals-of-probability/volume-33/issue-5/Phase-transition-of-the-largest-eigenvalue-for-nonnull-complex-sample/10.1214/009117905000000233.full

[17] A. Lytova and L. Pastur, "Central limit theorem for linear eigenvalue statistics of random matrices with independent entries," *The Annals of Probability*, vol. 37, no. 5, pp. 1778–1840, 2009, publisher: Institute of Mathematical Statistics.

[18] N. El Karoui, "Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond," *The Annals of Applied Probability*, vol. 19, no. 6, pp. 2362–2405, Dec. 2009, publisher: Institute of Mathematical Statistics. [Online]. Available: https://projecteuclid.org/journals/annals-of-applied-probability/volume-19/issue-6/Concentration-of-measure-and-spectra-of-random-matrices--Applications/10.1214/08-AAP548.full

[19] C. Louart and R. Couillet, "Concentration of Measure and Large Random Matrices with an application to Sample Covariance Matrices," Jan. 2021, arXiv:1805.08295 [math]. [Online]. Available: http://arxiv.org/abs/1805.08295

[20] R. Couillet and W. Hachem, "Fluctuations of Spiked Random Matrix Models and Failure Diagnosis in Sensor Networks," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 509–525, Jan. 2013, conference Name: IEEE Transactions on Information Theory. [Online]. Available: https://ieeexplore.ieee.org/document/6320635

[21] I. J. Schoenberg, "Metric Spaces and Completely Monotone Functions," *Annals of Mathematics*, vol. 39, no. 4, pp. 811–841, 1938, publisher: Annals of Mathematics. [Online]. Available: https://www.jstor.org/stable/1968466

[22] A. G. M. Steerneman and F. van Perlo-ten Kleij, "Spherical distributions: Schoenberg (1938) revisited," *Expositiones Mathematicae*, vol. 23, no. 3, pp. 281–287, Sep. 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0723086905000034

[23] P. Billingsley, *Probability and Measure*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2012.

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

[25] H. Lebeau, R. Couillet, and F. Chatelain, "A Random Matrix Analysis of Data Stream Clustering: Coping With Limited Memory Resources," in *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022, pp. 12 253–12 281, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v162/lebeau22a.html