Sparse Quantized Spectral Clustering

Zhenyu Liao[1], Romain Couillet[2], Michael W. Mahoney[1]

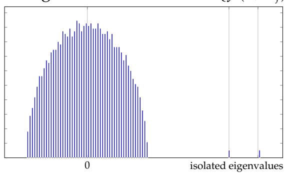
[1]ICSI and Department of Statistics University of California, Berkeley, USA
[2]G-STATS Data Science Chair, GIPSA-lab, University Grenobles-Alpes, France

Introduction

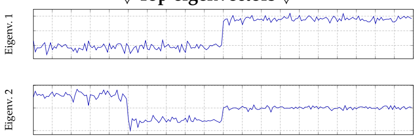
- **Big Data**: number of data n and dimension p both large, thousands or even millions
- Computational challenge: time and/or space complexity $O(n^2)$, unaffordable for low-power devices
- **Idea**: compress machine learning models (e.g., sketching, quantization or binarization), with non-trivial performance-complexity trade-off
- **Objective**: theoretical understanding of performance-complexity trade-off and optimal parameter tuning
- **Example**: unsupervised (kernel) spectral clustering

Reminder on spectral clustering

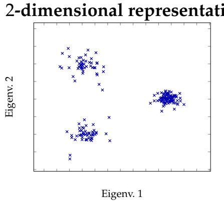
Two-step clustering based on $\mathbf{K} = \{f(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$:



⇓ **Top eigenvectors** ⇓



⇓ $K = 2$**-dimensional representation** ⇓



⇓

EM or k-means clustering.

Computational challenge

- $\mathbf{K} = \{f(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$: pairwise comparison of n data points, require $O(n^2)$ to retrieve top eigenvectors with, e.g., power method
- **Idea**: sparsifying, quantizing, and even binarizing: gain in both time and space!
- **Key object**: eigenspectrum of "compressed" matrix, statistics of top eigenvectors, as a function of **data statistics** and **compression method parameters**!

System model

Assumption 1 (Data: two-class signal-plus-noise mixture). *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be independently (non-necessarily uniformly) drawn from:*

$$\mathcal{C}_1 : \mathbf{x}_i = -\boldsymbol{\mu} + \mathbf{z}_i, \quad \mathcal{C}_2 : +\boldsymbol{\mu} + \mathbf{z}_i \tag{1}$$

*for $\mathbf{z}_i$ having i.i.d. zero-mean, unit-variance, κ-kurtosis, sub-exponential entries. $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] = \mathbf{Z} + \boldsymbol{\mu}\mathbf{v}^\mathsf{T}$ for random $\mathbf{Z} \in \mathbb{R}^{p\times n}$, $\boldsymbol{\mu} \in \mathbb{R}^p$ and **label vector** $\mathbf{v} \in \{\pm 1\}^n$.*

Assumption 2 (High-dimensional asymptotics). *As $n, p \to \infty$, $p/n \to c \in (0, \infty)$ and signal-to-noise ratio (SNR) $\|\boldsymbol{\mu}\|^2 \to \rho \geq 0$.*

Compression as entry-wise nonlinear transformation:

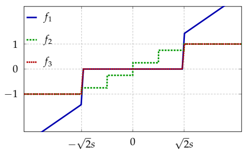
$$\mathbf{K} = \{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}\}_{i,j=1}^n \tag{2}$$

Sparsification: $f_1(t) = t \cdot 1_{|t|>\sqrt{2}s}$
Quantization: $f_2(t) = 2^{2-M}(\lfloor t \cdot 2^{M-2}/\sqrt{2}s \rfloor + 1/2) \cdot 1_{|t|\leq\sqrt{2}s} + \mathrm{sign}(t) \cdot 1_{|t|>\sqrt{2}s}$
Binarization: $f_3(t) = \mathrm{sign}(t) \cdot 1_{|t|>\sqrt{2}s}$

Truncation threshold $s > 0$, number of information bits M.

Key parameters: for each f and $\xi \sim \mathcal{N}(0, 1)$,

$$a_0 = \mathbb{E}[f(\xi)] = 0, \quad \mathbf{a_1} = \mathbb{E}[\xi f(\xi)], \quad \mathbf{a_2} = \mathbb{E}[\xi^2 f(\xi)]/\sqrt{2}, \quad \boldsymbol{\nu} = \mathbb{E}[f^2(\xi)]. \tag{3}$$



f	$\mathbf{a_1}$	ν	$\mathbf{a_2}$
f_1	$\mathrm{erfc}(s) + 2se^{-s^2}/\sqrt{\pi}$	$\mathrm{erfc}(s) + 2se^{-s^2}/\sqrt{\pi}$	**0**
f_2	$\sqrt{\frac{2}{\pi}} \cdot 2^{1-M}(1 + e^{-s^2} + \sum_{k=1}^{2^{M-2}-1} 2e^{-\frac{k^2s^2}{4^{M-2}}})$	$1 - \frac{2^M-1}{4^{M-1}}\mathrm{erf}(s) - \sum_{k=1}^{2^{M-2}-1} \frac{k\,\mathrm{erf}(ks\cdot 2^{2-M})}{2^{2M-5}}$	**0**
f_3	$e^{-s^2}\sqrt{2/\pi}$	$\mathrm{erfc}(s)$	**0**

Question to answer

To save X% of computational time and/or space, clustering accuracy drop by Y% (depends on **data SNR**, **dimension**, **sample size**, and **compression parameters**)

Main results

Theorem 1 (Eigenvalue distribution). *As $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, the empirical spectral measure $\omega_\mathbf{K} = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(\mathbf{K})}$ of $\mathbf{K}$ converges to a deterministic limit ω, uniquely defined through its Stieltjes transform $m(z) = \int (t - z)^{-1}\omega(dt)$ solution to*

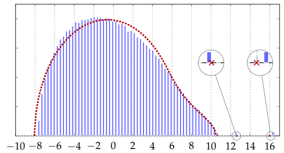
$$z = -\frac{1}{m(z)} - \frac{\nu - a_1{}^2}{c}m(z) - \frac{a_1{}^2 m(z)}{c + a_1 m(z)}. \tag{4}$$

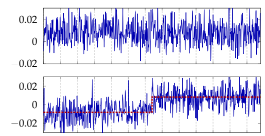
Theorem 2 (Informative spike and a phase transition). *For $a_1 > 0$ and $\mathbf{a_2 = 0}$, define $F(x) = x^4 + 2x^3 + \left(1 - \frac{c\nu}{a_1{}^2}\right)x^2 - 2cx - c$ and $G(x) = \frac{a_1}{c}(1 + x) + \frac{a_1}{x} + \frac{\nu - a_1{}^2}{a_1}\frac{1}{1+x}$ and let γ be the largest real solution to $F(\gamma) = 0$. Then, the largest eigenpair $(\hat{\lambda}, \hat{\mathbf{v}})$ of $\mathbf{K}$ satisfies*

$$\hat{\lambda} \to \lambda = \begin{cases} G(\rho), & \rho > \gamma \\ G(\gamma), & \rho \leq \gamma \end{cases}, \quad \frac{1}{n}|\hat{\mathbf{v}}^\mathsf{T}\mathbf{v}|^2 \to \alpha = \begin{cases} \frac{F(\rho)}{\rho(1+\rho)^3}, & \rho > \gamma \\ 0, & \rho \leq \gamma \end{cases} \tag{5}$$

as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, for SNR $\rho = \lim \|\boldsymbol{\mu}\|^2$.

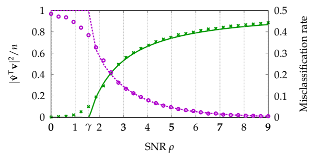
Remark (Spurious non-informative spikes). *If $\mathbf{a_2} \neq \mathbf{0}$, there may be* up to two ***non-informative*** *eigenvalues (with eigenvectors containing only noise) on the* left or right *of the main bulk.*

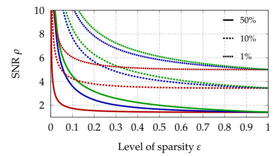




Corollary 1 (Performance of spectral clustering). *Let $a_1 > 0, a_2 = 0$, and $\hat{\mathcal{C}}_i = \mathrm{sign}([\hat{\mathbf{v}}]_i)$ be the estimate of the underlying class $\mathcal{C}_i$ of the datum $\mathbf{x}_i$, with $\hat{\mathbf{v}}^\mathsf{T}\mathbf{v} \geq 0$ for $\hat{\mathbf{v}}$ the top eigenvector of $\mathbf{K}$. As $n, p \to \infty$, the misclassification rate satisfies*

$$\frac{1}{n}\sum_{i=1}^n \delta_{\hat{\mathcal{C}}_i \neq \mathcal{C}_i} \to \frac{1}{2}\mathrm{erfc}(\sqrt{\alpha/(2 - 2\alpha)}).$$





(Left) Eigenvector alignment (**green**) and classif. error (**purple**) versus SNR ρ. **(Right)** Comparison of 1%, 10% and 50% classif. error curves between subsampling (**green**), uniform (**blue**) and selective sparsification f_1 (**red**), as a function of sparsity level ε and SNR ρ.

References

- Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. "Sparse Quantized Spectral Clustering". In: *International Conference on Learning Representations*. 2021.