Random Matrix-Improved Kernels for Large Dimensional Spectral Clustering

Hafiz Tiomoko Ali[1], Abla Kammoun[2], Romain Couillet[1,3]

[1]LSS, CentraleSupélec, Université ParisSaclay, France.
[2]KAUST University, Saudi Arabia.
[3]GIPSA-lab, Université Grenoble Alpes, France.

Abstract

Context:

- Kernel spectral clustering in the "large n, large p" regime.
- Gaussian mixture model with "non-trivial" class separability.
- Asymptotic performance analysis and improvement.

Take away:

- Proven suboptimality of "classical" kernels.
- Improved kernel design with optimal separability properties.
- Dramatic performance improvements in simulations.

I – Model and Assumptions

Setting and Basic Assumptions

Data: $x_1, \ldots, x_n \in \mathbb{R}^p$, with $\begin{cases} x_1, \ldots, x_{n_1} & \in \mathcal{C}_1 \\ \ldots & \ldots \\ x_{n-n_k+1}, \ldots, x_n & \in \mathcal{C}_k. \end{cases}$

Class definition:

$$x_i \in \mathcal{C}_a \Leftrightarrow x_i \sim \mathcal{N}(\mu_a, C_a).$$

Growth rates: As $n \to \infty$, k remains fixed, and

$$p/n \to c_0 > 0, \quad n_a/n \to c_a > 0.$$

Assumption (Neyman–Pearson Optimal Separability Rate)

As $p \to \infty$, for all $a, b \in \{1, \ldots, k\}$,

- $\|\mu_a - \mu_b\| = O(1)$
- $\|C_a\|$ *bounded,* $|\operatorname{tr}(C_a - C_b)| = O(\sqrt{p})$, $\operatorname{tr}((C_a - C_b)^2) = O(\sqrt{p})$.
- $\frac{1}{p}\operatorname{tr} C^\circ \to \tau$ *with* $C^\circ \equiv \sum_{a=1}^k \frac{n_a}{n} C_a$ *and* $\tau > 0$.

(Inner Product) Kernel Matrix

Object of interest: With $x_i^\circ = x_i - \frac{1}{n}\sum_{j=1}^n x_j$, define

$$K = P\left\{ f\left(\frac{1}{p}(x_i^\circ)^\mathsf{T} x_j^\circ\right) 1_{i \neq j} \right\} P$$

where $P = I_p - \frac{1}{n} 1_n 1_n^\mathsf{T}$ and f three-times differentiable around 0.

($f(\frac{1}{p}\|x_i - x_j\|^2)$ could be treated similarly)

Objective: Study:

- limiting spectrum of K (eigenvalues + eigenvectors)
- clustering performances.

II – The Importance of $f(t)$ Around $t = 0$

Previous Findings and Motivation

Key Result: As $n, p \to \infty$, and assumption above,

- for all $i \neq j$, irrespective of the class,

$$\frac{1}{p}(x_i^\circ)^\mathsf{T} x_j^\circ \to 0, \quad \frac{1}{p}\|x_i^\circ\|^2 \to \tau, \quad \frac{1}{2p}\|x_i - x_j\|^2 \to \tau.$$

- counter-intuitive curse of dimensionality: **all vectors are far**.
- but allows for Taylor-expansion of K_{ij} around $f(0)$:

$$K_{ij} = f(0) + f'(0)\left[\frac{1}{p}\mu_a^\mathsf{T}\mu_b + \frac{1}{p} w_i^\mathsf{T} w_j + \ldots\right] + \frac{1}{2} f''(0)\left[\frac{1}{p}\operatorname{tr} C_a C_b + \ldots\right] + o(1)$$

(for $x_i = \mu_a + w_i$, $x_j = \mu_b + w_j$)

Consequences:

- [C,Benaych'16] Relaxing assumptions as $\operatorname{tr}((C_a - C_b)^2) = O(p)$:
  - **Spectrum of K is Marcenko–Pastur like**
  - Phase transition phenomenon
  - Special behavior when $f'(0) = 0$!

  $f'(0) = 0$ "kills the noise" **BUT** "kills the means".

- [K,C'17] For $\operatorname{tr}((C_a - C_b)^2) = O(\sqrt{p})$, $\mu_1 = \ldots = \mu_k$ and $f'(0) = 0$:
  - **Spectrum of K is semi-circle like**
  - Phase transition phenomenon
  - But statistical means cannot be used.

A New Kernel Design

For some $\alpha, \beta > 0$, we choose f such that

$$f'(0) = \frac{\alpha}{\sqrt{p}}, \quad \frac{1}{2} f''(0) = \beta.$$

III – Main Results

Theorem (Asymptotic Equivalent for K)

For f such that $f'(0) = \frac{\alpha}{\sqrt{p}}$, $\frac{1}{2}f''(0) = \beta$, as $n, p \to \infty$,

$$\|K - \hat{K}\| \xrightarrow{\text{a.s.}} 0$$

where

$$\sqrt{p}\hat{K} \equiv \alpha P W^\mathsf{T} W P + \beta P \Phi P + U A U^\mathsf{T} - (f(0) + \tau f'(0)) P$$

with

$$W = [w_1, \ldots, w_n], \quad \Phi_{ij} = \sqrt{p}\left[((w_i^\circ)^\mathsf{T} w_j^\circ)^2 - \frac{1}{p^2}\operatorname{tr} C_a C_b\right] 1_{i \neq j}$$

$$U = \left[\frac{J}{\sqrt{p}}, P W^\mathsf{T} M\right], \quad J = [j_1, \ldots, j_k], \quad j_a = (0, \ldots, 1_{n_a}, \ldots, 0)^\mathsf{T}$$

$$A = \begin{bmatrix} \alpha M^\mathsf{T} M + \beta T & \alpha I_k \\ \alpha I_k & 0 \end{bmatrix}, \quad M = [\mu_1^\circ, \ldots, \mu_k^\circ], \quad T = \frac{1}{\sqrt{p}}\{\operatorname{tr} C_a C_b\}.$$

$\Longrightarrow$ Kernel is "Neyman–Pearson optimal", both in means and covariances.

Theorem (Limiting Eigenvalue Distribution)

As $n, p \to \infty$,

$$\nu_n \equiv \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(K)} \xrightarrow{\mathcal{L}} \nu$$

with ν given by its Stieltjes transform $m(z) = \int \frac{\nu(d\lambda)}{\lambda - z}$, unique solution of

$$\frac{1}{m(z)} = -z + \frac{\alpha}{p}\operatorname{tr} C^\circ\left(I_p + \frac{\alpha m(z)}{c_0} C^\circ\right)^{-1} - \frac{2\beta^2}{c_0} m(z)\left(\frac{1}{p}\operatorname{tr}(C^\circ)^2\right)^2.$$

Mixed Marcenko–Pastur & Wigner Spectrum

- $PW^\mathsf{T}WP$: Marcenko–Pastur like spectrum
- $P\Phi P$: semi-circle (Wigner) like spectrum
- UAU^T: produces spikes under phase transition!

Here for $f(x) = \frac{1}{2}\beta\left(x + \frac{1}{\sqrt{p}}\frac{\alpha}{\beta}\right)^2$,

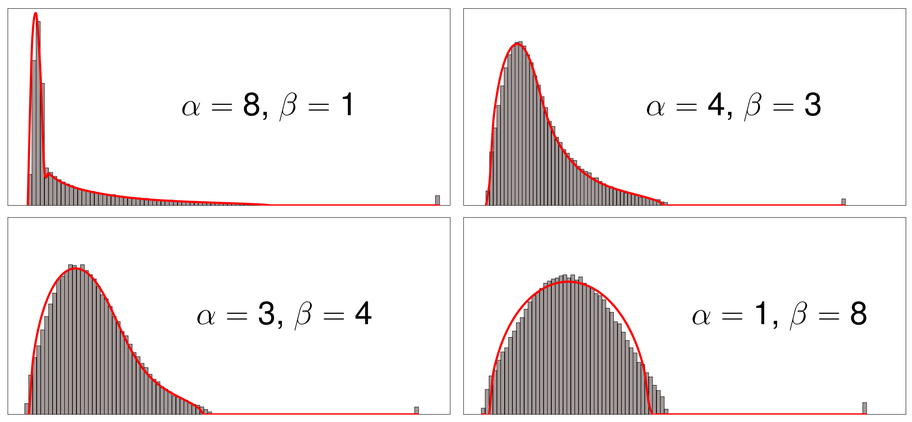

Figure: Eigenvalues of K versus limiting law, $p = 2048$, $n = 4096$, $k = 2$, $n_1 = n_2$, $\mu_i = 3\delta_i$.

IV – Application

Performance of Kernel Spectral Clustering

DATASETS	$\|\mu_1 - \mu_2\|^2$	$\frac{1}{\sqrt{p}}\operatorname{tr}(C_1 - C_2)^2$	RATIO
MINIST (DIGITS 1, 7)	613	1990	3.2
MINIST (DIGITS 3, 6)	441	1119	2.5
MINIST (DIGITS 3, 8)	212	652	9.0
EEG (SETS A, E)	2.4	109	45.4

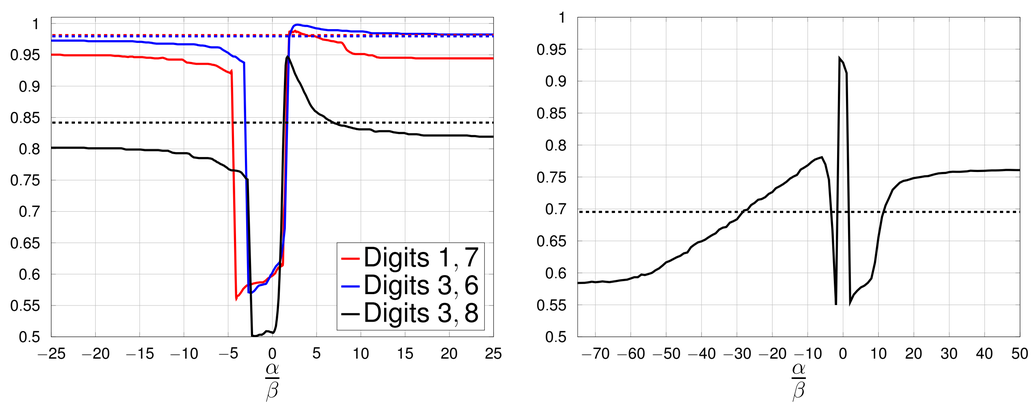

Figure: Spectral clustering accuracy of MNIST and EEG, versus Gaussian kernel (dashed).

Open Questions

- Online estimation of optimal (α, β)
- Similar behavior for kernel $f(\frac{1}{\sqrt{p}}(x_i^\circ)^\mathsf{T} x_j^\circ)$ but difficult to analyze.

Email: hafiz.tiomokoali@centralesupelec.fr, abla.kammoun@kaust.edu.sa, romain.couillet@centralesupelec.fr