# NGSA: Large scale and high dimensional distribution testing and comparison through random graph

Derwiche Nayef, Elazhari Lina, Guillot Antoine, Haloui Hamza

April 5, 2018

**Abstract**

When working with high-dimensional data, one often faces a number of issues, notably computational complexity, and curse of dimensionality related problems such as distance concentration and polynomial times complexity in either the dimension or the samples size. In the case of two-sample tests and distribution distance estimation, this may render the problem intractable.
Using recent advances in random matrix theories and the analysis of spiked models, we propose a method to compare two samples in linear time in both the dimension and the sample size. When both the sample size and the dimension are growing proportionally,under ideal conditions, the spectrum of the Gram matrix of the sample will converge to a Marczenco-Pastur distribution and will almost surely be in the bulk of the distribution . The gap between the ideal conditions and the real observed samples will induced perturbations in the spectrum. These perturbations take the form of spikes located outside the main bulks, hence the name 'spiked eigenvalues' or "spikes. These spikes "carry the information" and will be used to distinguish samples from different distributions. However, the number of spikes is often low and not sufficient to compute any statistical test. Hence, the Gram matrix is used to generate random graphs, their spectrum converges to a Semi-Circle law and also has spikes when the samples is a perturbation from the ideal case. Sampling enough random graph (and spikes) increases the power of the statistical test and yields great results with large distribution (around $10^6$ points) in large spaces (more than 100 dimensions).

## 1 Motivations

When doing a two sample test on large samples living in high dimensional spaces we face two fundamental problems:

- Computational complexity issues both caused by the large number of samples and the high dimensionality, that both need to be dealt with.
- A form of the curse of high dimensionality: distance notions, and similarity functions as well, loose meaning. For many $K$, distance or similarity measures, for many data distributions $D$, $\forall x, x' \in D \quad K(x, x') \approx \tau$ when the dimension increases. [3]

For instance, one popular method for two samples test is the use of kernel MMD (maximum mean discrepancy) estimators which suffers from both those issues [4] and requires bootstrap to ensure consistency of the test. More recently, linear estimator kernel estimates have been developed, but they are less accurate that the quadratic tests [1]. Still, this method suffers from the curse of dimensionality, our method also propose a way to select an optimal kernel which can discriminate samples in large dimension following recent advances in random matrix theory. [2] (Not done in the current version)

# 2 Methodology

## 2.1 Ideal Model

Let us first review the distribution of the eigenvalues of large var-cov matrices. In fact, these matrices and their endowed graphs have eigenvalues distribution that converges to typical distribution such as these ones :

- the Wigner theorem : Let Let $X_N = (X_{ij}$ ) a symmetric N  N matrix with i.i.d. entries on and above the diagonal with $\mathbb{E}X_{ij} = 0$ and $\mathbb{E}X_{ij}^2 = 1$ and $X_{ij} = X_{ji}$ (for symmetry). Consider the spectrum of Wigner matrix

$$\frac{X_N}{\sqrt{N}}$$

  The histogram of a Wigner matrix converges to the semi-circular distribution

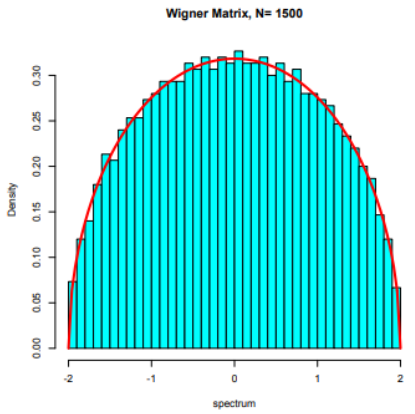



Figure: The semi-circular distribution (in red) with density $x \mapsto \frac{\sqrt{4-x^2}}{2\pi}$

Figure 1: Semi-circle Distribution

- Marcenko-Pastur theorem : With the same hypothesis, let's consider the spectrum of

$$\frac{X_N X_N^*}{n}$$

The histogram of a large covariance matrix converges to Marcenko-Pastur distribution with a given parameter.

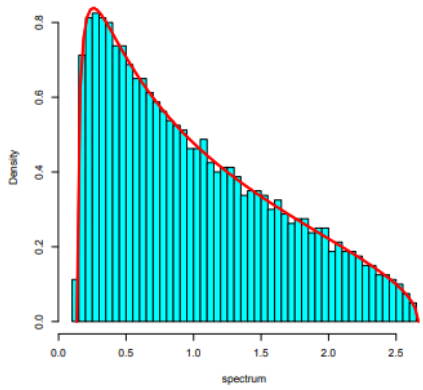



Figure: Marčenko-Pastur's distribution (in red)

Figure 2: Marcenko-Pastur Distribution

So we see that every large matrix with some properties converge to a known distribution. However, if all covariance matrices converge to the same distribution how do we discriminate them?

## 2.2 Gram matrix and kernels

Let x = ($x_1$ , $x_2$..., $x_n$) y = ($y_1$ , $y_2$..., $y_n$) two samples respectively drawn from p and q. We want to test p = q Let K be a symmetric mapping :

$$K : x, y \Rightarrow K(x, y)$$

$$\chi \times \chi \Rightarrow R$$

We denote $K^x = (K(x^i, x^j))$ and $K^y = (K(y^i, y^j))$ the 2 Gram matrices induced by K on x and y. By Aronzjain theorem, there exists

$$\phi : \chi \Rightarrow H$$

where H is a Hilbert space such that K(x,y) $=< \phi(x), \phi(y) >_H$. Then $K^x$ and $K^y$ can be expressed as covariance matrices. We denote

$$K_0^x =< \frac{\phi(x_i) - \phi(\bar{x}_i)}{sd(\phi(x_i))}, \frac{\phi(x_j) - \phi(\bar{x}_j)}{sd(\phi(x_j))} >$$

the centered Gram matrix.

## 2.3 Spike models

The spiked model is a particular case of large covariance matrix defined as a perturbation of rank $k$ of the ideal model:

$$R_N = \mathbf{I_n} + \sum_{k=1}^{l} \theta_l \vec{u_l} u\vec{*}_l$$

The experiments shows that the centered and standardized Gram matrix and the graph generated are rank $k$ perturbation from the ideal case.

## 2.4 Sampling and testing methods

### 2.4.1 The Gram spiked TST

One idea would be to sample many random graphs and compute the distribution of their eigenvalues. In fact we sample a number J of (x) and (y) of size g : $K_1^x ... K_J^x$, $K_1^y ... K_J^y$. If g is large enough, the distribution of the spiked eigenvalues of $K_1^x ... K_J^x$ should be the same and so it is for $K_1^y ... K_J^y$.The comparison of $\bigcup_{i=1}^{J} Sp(K_i^x)$ and $\bigcup_{i=1}^{J} Sp(K_i^y)$ provide a good proxy to compare p and q. Besides, if x and x' are such that $0 \leq K(x, x') \leq 1$ , K(x,x') can be seen as the probability of having an edge between x and x'. Hence, given a Gram matrix $K^x$ with a such a kernel, we can build $G^x = \text{Bernoulli}(K^x(x_i, x_j)$ with $G^x \in S_n(R)$

### 2.4.2 The graph spiked TST

In the matrix $G^x$ all the variables are independent but not iid. The graph matrix is centered and standardized . Hence, one can generate $B^N$ realization of $G_1^x, ..., G_{B^N}^x$, we do hope that $\bigcup_{i=1}^{B^n} Sp^{spike}(G_i^x)$ and $\bigcup_{i=1}^{B^n} Sp^{spike}(G_i^x)$ are good proxy to compare p and q. A U-test is then computed on the spiked eigenvalues and used to compare the distribution.

## 2.5 The algorithm

**Algorithm 1** Gram spiked 2 sample test

**procedure** UTESTGRAM($(x_1, \cdots, x_n)$, $(y_1, \cdots, y_n)$, K,NSAMPLE,SAMPLESIZE,$\alpha$)
    $\lambda_x, \lambda_y \leftarrow [\,], [\,]$
    **for** $i$ in $[0, nSample]$ **do**
        $x', y' \leftarrow \text{sample}(x, \text{sampleSize}), \text{sample}(y, \text{sampleSize})$
        $\lambda_x.\text{append}(\text{Sp}(K_i^{x'}))$
        $\lambda_x.\text{append}(\text{Sp}(K_i^{y'}))$
    $c \leftarrow nSample/dimension$
    $\lambda_x^{Sp}, \lambda_y^{Sp} \leftarrow \lambda_x[\lambda_x > \Gamma_c], \lambda_x[\lambda_x > \Gamma_c]$
    **return** Utest of $(\lambda_x^{Sp}, \lambda_y^{Sp})$ at level $\alpha$

**Algorithm 2** Graph spiked 2 sample test

**procedure** UTESTGRAPH($(x_1, \cdots, x_n)$, $(y_1, \cdots, y_n)$, K,NSAMPLE,SAMPLESIZE,$\alpha$)
  $\lambda_{\mathrm{x}}, \lambda_{\mathrm{y}} \leftarrow [\,], [\,]$
  **for** $i$ in $[0, nSample]$ **do**
    $\mathrm{x}', \mathrm{y}' \leftarrow \mathrm{sample}(\mathrm{x}, \mathrm{sampleSize}), \mathrm{sample}(\mathrm{y}, \mathrm{sampleSize})$
    Compute $K^{(x')}$ and $K^{(y')}$
    Draw 2 upper triangular unif (0,1) $T_{x'}$ and $T_{y'}$
    $\mathrm{G}^{(\mathrm{x}')} \leftarrow \mathrm{K}^{(\mathrm{x}')}[\mathrm{K}^{(\mathrm{x}')} > \mathrm{T}_{\mathrm{x}'} + \mathrm{T}_{\mathrm{x}'}^{\mathrm{T}}]$
    $\mathrm{G}^{(\mathrm{y}')} \leftarrow \mathrm{K}^{(\mathrm{y}')}[\mathrm{K}^{(\mathrm{y}')} > \mathrm{T}_{\mathrm{y}'} + \mathrm{T}_{\mathrm{y}'}^{\mathrm{T}}]$
    $\lambda_{\mathrm{x}}.\mathrm{append}(\mathrm{Sp}(\mathrm{G}^{(\mathrm{x}')}))$
    $\lambda_{\mathrm{y}}.\mathrm{append}(\mathrm{Sp}(\mathrm{G}^{(\mathrm{y}')}))$
  $\lambda_{\mathrm{x}}^{\mathrm{Sp}}, \lambda_{\mathrm{y}}^{\mathrm{Sp}} \leftarrow \lambda_{\mathrm{x}}[|\lambda_{\mathrm{x}}| > 2], \lambda_{\mathrm{x}}[|\lambda_{\mathrm{x}}| > 2]$
  **return** Utest of $(\lambda_{\mathrm{x}}^{\mathrm{Sp}}, \lambda_{\mathrm{y}}^{\mathrm{Sp}})$ at level $\alpha$

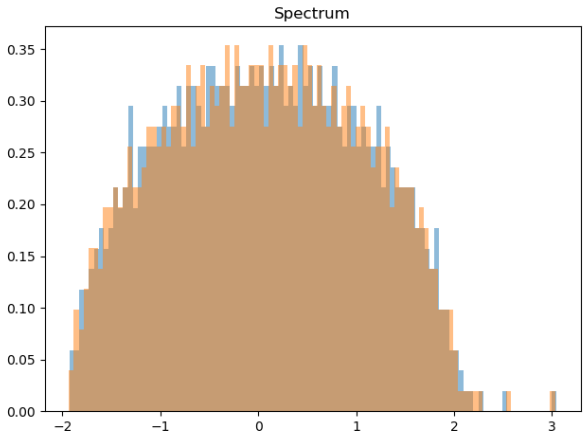
# 3 Experiments

## 3.1 Language Discrimination

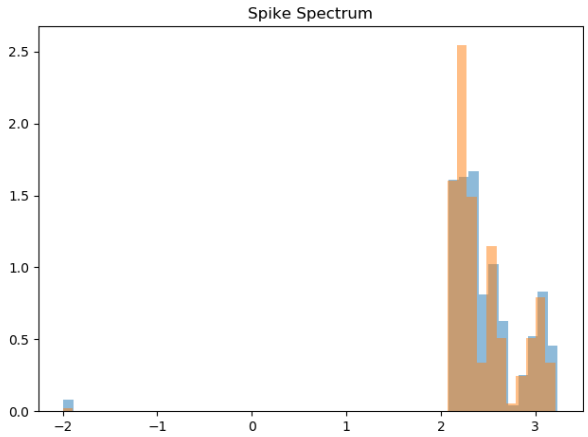
For a first sanity check, we used as data the Word2Vec encoding of English and French from FastText. We hope to find that different languages follow different distributions in sensible representations. The results are in fig (3) In (a) and (b) we sampled matrix of vocabulary similarity (cosine similarity rescaled between 0 and 1 between words in the vector space) to generate random graph and get eigenvalues distribution. In (c) and (d), we took the covariance matrix on the latent dimensions on sample of the vocabulary and generated random graph from it, to get eigenvalues distribution. As expected, most of the eigenvalues fall into the bulk as in the ideal model, they constitute the noise and they overlap for the two languages. The information, the signal, is in the spikes, eg outside of the bulk, and this is where there is far less overlap between the languages. And indeed, if we do a two-sample test on the distribution of the languages, we reject the same distribution hypothesis on the spike but we do not reject this hypothesis if we do it on the whole distribution.

## 3.2 Language Clustering

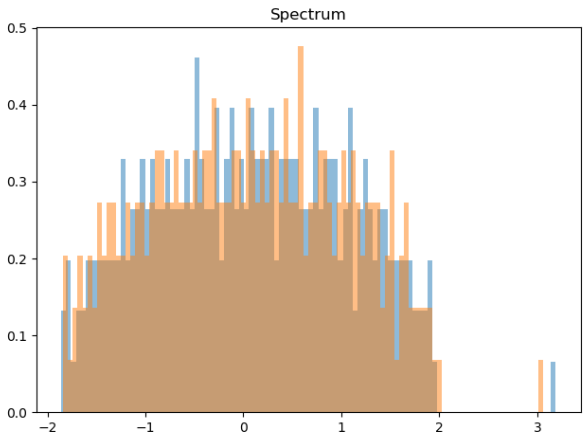
As a second experiment, we wanted to see if the use of spiked values allowed us to recover a meaningful notion of distance. We computed again the eigenvalues distribution using the similarity matrix in the latent dimensions of several languages Word2Vec encodings. We then computed the pairwise Wasserstein distance between the eigenvalues distributions. When doing it with the whole distribution, we find that all the distance have close values and it is difficult to discriminate meaningfully, as often with distance in high-dimensional spaces. However, if compute the distance with spiked eigenvalues distribution, we have a much more relevant and discriminative distance measure. We show in fig (4) this pairwise distance matrix, with clusters. We see that English, German en French, constitutes a distinct bock, there is also another block with Romanian and Albanian, while Chinese is very far from all other languages.



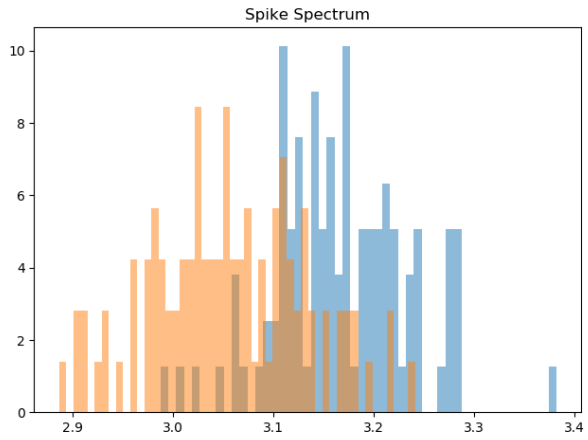

Whole Distribution




Spiked Values




Whole Distribution




Spiked Values

Figure 3: Histogram of eigenvalues distribution for English (orange) and French (blue)

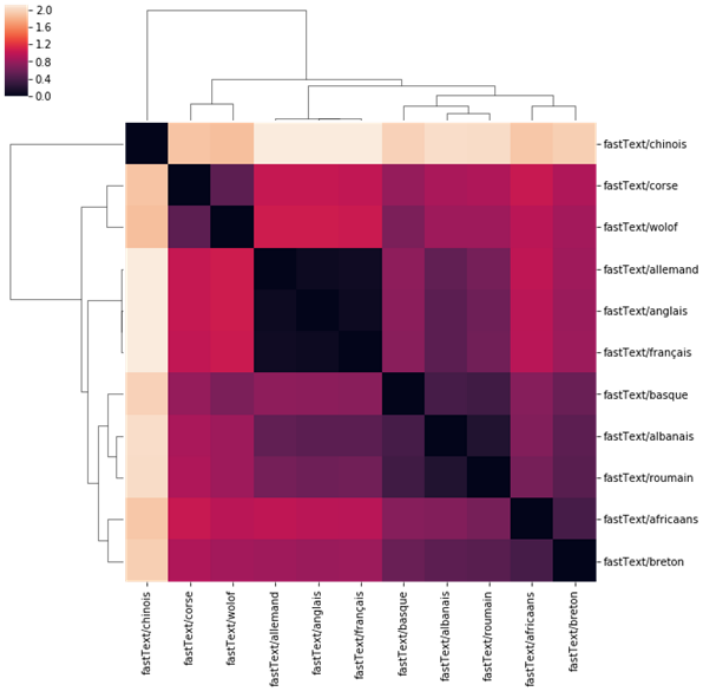



Figure 4: Wasserstein distance between spiked eigenvalues distribution of different languages sampled random graph on latent dimension similarity)

### 3.3 Document Classification

Finally, to have a more tangible indication of how meaningful the distance on spiked eigenvalues is, we wanted to try to use it to do unsupervised classification. We selected a few documents from the NewsGroup20 dataset, in two categories. We computed the pairwise distance as before, and we observe two main block that are significantly homogeneous in their categories 5.

It is a very interesting result because it is totally unsupervised. We took a sensible and reasonable metrics, the Wasserstein distance on the eigenvalues distribution on the spikes from a similarity matrix on the latent dimensions interactions, on a space which is is linked to the words co-occurrences, but this distance was in no way built for a specific task. Also one should bear in mind that the categories of the documents in the dataset are quite noisy, subjectively and not clearly defined, which makes it difficult for off-the-shelf unsupervised classification.

## 4 Conclusions

Our approach has shown conclusive results on text analysis and a novel use of random graphs for general two sample test and distribution comparison. Future works should focus on two points:

- Selection of an appropriate kernel to discriminate the two samples usage [2].

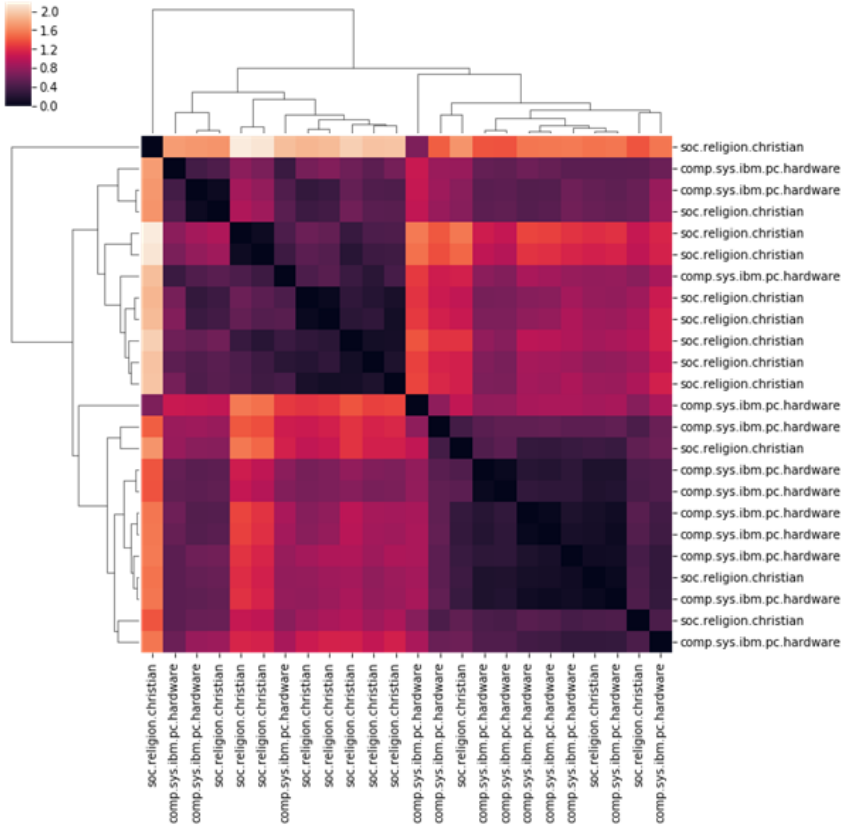



Figure 5: Wasserstein distance between spiked eigenvalues distribution of different documents sampled random graph on latent dimension similarity)

- Comparison of the method performance with MMD based method.

# References


[1] Optimal kernel choice for large-scale two-sample tests.

[2] Random matrix-improved kernels for large dimensional spectral clustering.

[3] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[4] Raschg Scholkopf Gretton, Bogwartd and Smola. A kernel two-sample test. In *Journal of Machine Learning Research*, 2012.