SPECTRAL CLUSTERING FOR HIGH DIMENSIONAL MIXTURE MODELS

Evgeny KUSMENKO, Romain COUILLET

CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

ABSTRACT

In this article we provide new theoretical results concerning the performance of spectral clustering in settings where both the amount and the size of the data are large. We are able to show that kernel graph Laplacian matrices can be approximated by an analytically tractable random matrix model. This approximation is then used in order to characterize the behavior of the eigenvalues of interest as well as the corresponding eigenvectors allowing one in turn to predict the clustering performance and the impact of the kernel choice thereon.

***Index Terms*—** Spectral clustering, kernels, unsupervised learning theory, random matrix theory, Gaussian mixture models

1. INTRODUCTION

The spectral clustering framework has been known to perform well in many unsupervised learning contexts such as speech separation and other complex clustering tasks [1]. Theoretical bounds and guarantees on the performance of spectral clustering have been published, e.g., in [2–4]. However, the exact behavior of the eigenpairs of kernel graph Laplacians used for clustering is not yet sufficiently well understood. It has been shown in [5] for a family of kernels that these challenges can be approached by means of random matrix theory. The aim of this work is to extend these results to the class of kernels depending on the inner product of the features. Our main contributions comprise a tractable random matrix approximation for normalized symmetric graph Laplacians in a setting where the data are drawn from a Gaussian mixture model (GMM) with a given set of parameters, the characterization of the isolated eigenvalues and eigenvectors of this model with regard to clustering capability, and finally the impact of the kernel choice on the clustering performance. Proofs are omitted due to space limitations and will be given in an extended version of this article.

Notation: By $\|\cdot\|$ we denote the Euclidean norm for vectors and the spectral norm for matrices. $\langle\cdot,\cdot\rangle$ stands for the standard inner product. $\mathbb{1}_n$ denotes the n-dimensional column vector full of ones. I_n is the $n\times n$ identity matrix.

This work is funded by ERC MORE EC-120133 and RMT4GRAPH ANR_14_CE28-0006

2. SYSTEM MODEL

Let $x_1,\ldots,x_n\in\mathbb{R}^p$ be n samples drawn from a p-variate K-component GMM. Component a is characterized by its mean vector μ_a and covariance matrix C_a, $a=1,\ldots,K$. Thus, we can express x_i as μ_a+w_i where a is the component observation i belongs to and $w_i\sim\mathcal{N}(0,C_a)$. The set of indices generated by component a is denoted as $\mathcal{C}_a$ with $\bigcup_{a=1}^K\mathcal{C}_a=\{1,\ldots,n\}$ and $\mathcal{C}_a\cap C_b=\emptyset$ for all $a\neq b$. The mixing weights of the GMM are indicated by π_a. We will be concerned with the regime where both the dimensionality and the number of samples are very large which is a common setting in machine learning applications. In asymptotic sense we require $n,p\to\infty$ in such a way that $\frac{p}{n}\to c\in(0,\infty)$.

As it should be inconsequential to clustering, we first subtract the empirical mean from the data. For notational convenience we define the centralized observations and means as $x_i^\circ:=x_i-\frac{1}{n}\sum_{j=1}^n x_j$ and $\mu_a^\circ:=\mu_a-\sum_{i=1}^K\pi_i\mu_i$, respectively. In this study we will focus on the family of kernel functions $k(x_i,x_j)=f\left(\frac{1}{p}\left\langle x_i^\circ,x_j^\circ\right\rangle\right)$ for building a fully connected affinity matrix A with $A_{ij}=A_{ji}=k(x_i,x_j)$ where f may be any three times differentiable function fulfilling $f(0),f'(0)\neq 0$. Clearly, the clustering task is easy if the components are sufficiently distinct. Conversely, if the components tend to coincide, no clustering is possible. Hence, we will concentrate on critical cases given by Assumption 1 where clustering is non-trivial .

Assumption 1. *As n and p grow large, the GMM parameters remain bounded as $\frac{1}{p}\operatorname{tr}C_a=\mathcal{O}(1)$, $\frac{1}{p}\operatorname{tr}C_aC_b=\mathcal{O}(1)$, and $\|\mu_a-\mu_b\|=\mathcal{O}(1)$ for all $a,b\in\{1,\ldots,K\}$.*

Spectral clustering does not operate on the affinity matrix A directly but rather on the random graph Laplacian defined as $L:=D-A$ where $D=\operatorname{diag}(A\mathbb{1}_n)$ is the degree matrix of A. However, it is often beneficial to use a modification of this definition as pointed out in [6]. In this work we will concentrate on the spectral properties of the symmetric normalized Laplacian $L_{sym}:=D^{-\frac{1}{2}}LD^{-\frac{1}{2}}=I_n-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. For notational convenience we drop the identity I_n and the minus in the definition thereby obtaining the matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. Clearly, $D^{\frac{1}{2}}\mathbb{1}_n$ is an eigenvector of this matrix and has one as its eigenvalue. Hence, we are able to separate the analysis of this eigenvector from the rest by subtracting its eigenspace

from $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. This leads to a more tractable random matrix model we will focus on, namely,

$$L' := n\frac{f(0)}{f'(0)}\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}} - \frac{D^{\frac{1}{2}}\mathbb{1}_n\mathbb{1}_n^T D^{\frac{1}{2}}}{\mathbb{1}_n^T D\mathbb{1}_n}\right) \quad (1)$$

where we introduced the scaling factors n and $\frac{f(0)}{f'(0)}$ in order to ensure that $\|L'\| = \mathcal{O}(1)$ and to avoid cumbersome notation later on. Note that L' is a symmetric but not necessarily a positive semi-definite matrix.

3. MAIN RESULTS

A norm-consistent tractable approximation of kernel matrix models has been proposed in [7]. We extend this result to the case where the data are generated by a GMM. Since, upon our kernel restriction and Assumption 1, in the large n, p regime, $A_{ij} \to 0$ for all $i \neq j$, the idea is to approximate f by its (sufficiently large order) Taylor series expansion around zero. We will use the definitions

$$M := [\mu_1^\circ, \ldots, \mu_K^\circ], \quad (2)$$

$$J := [j_1, \ldots, j_K], \quad (3)$$

$$P := I_n - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T, \quad (4)$$

$$T := \left\{\frac{1}{p}\operatorname{tr} C_a C_b\right\}_{a,b=1}^K, \quad (5)$$

$$W := [w_1, \ldots, w_n] \quad (6)$$

where $j_a \in \{0,1\}^n$ is the indicator vector of component a with its i-th entry being 1 if $i \in \mathcal{C}_a$ and 0 otherwise.

Theorem 1. *Consider the matrix*

$$\hat{L}' = P\frac{W^TW}{p}P + P\frac{A_{\text{diag}}}{f'(0)}P + PVA_1V^TP \quad (7)$$

where

$$V := \left[\frac{J}{\sqrt{p}}, \frac{1}{\sqrt{p}}PW^TM\right], \quad (8)$$

$$A_1 := \begin{bmatrix} M^TM + \frac{f''(0)}{2f'(0)}T & I_K \\ I_K & 0 \end{bmatrix}, \quad (9)$$

$$A_{\text{diag}} := \operatorname{diag}\Big(J\left\{f\left(\frac{1}{p}\operatorname{tr} C_a\right)\right\}_{a=1}^K \quad (10)$$

$$- f'(0)J\left\{\frac{1}{p}\operatorname{tr} C_a\right\}_{a=1}^K\Big) - f(0)I_n. \quad (11)$$

Then, as $n, p \to \infty$,

$$\left\|L' - \hat{L}'\right\| \to 0 \quad (12)$$

almost surely.

Now using the much more analytically tractable model $\hat{L}'$ we are able to analyze the spectrum of L'. Note that the first two addends in Eq. (7) have high rank scaling with n, p while the third term is a perturbation of rank $2K$. Provided that the components of the GMM are sufficiently different, as per standard random matrix theory results for spiked models, the latter will produce at most $2K$ eigenvalues away from the ensemble of the eigenvalues of the high rank term [8]. Their respective eigenvectors, which align to some extent to J, may then be used for clustering. Clearly, the eigenvalues of $\hat{L}'$ are the values of z solving $\det(\hat{L}' - zI) = 0$. Now, in order to find the isolated eigenvalues we may factor out the high rank part. Then, by applying Sylvester's determinant theorem, the equation may be rewritten as

$$\det Q_z^{-1} \cdot \det\left(I_{2K} + A_1V^TPQ_zPV\right) = 0 \quad (13)$$

where Q_z is defined as

$$Q_z := \left(P\frac{W^TW}{p}P + P\frac{A_{\text{diag}}}{f'(0)}P - zI_n\right)^{-1} \quad (14)$$

for z away from the eigenvalues of $P\frac{W^TW}{p}P + P\frac{A_{\text{diag}}}{f'(0)}P$. Using the Gaussian tools for random matrices, a framework introduced in [9, Chapter 2], we derive the following lemma.

Lemma 1. *As $n, p \to \infty$, for all z outside the support of the eigenvalue distribution of $P\frac{W^TW}{p}P + P\frac{A_{\text{diag}}}{f'(0)}P$*

$$Q_z \leftrightarrow \bar{Q}_z := c\operatorname{diag}\left(\{g_a(z)\mathbb{1}_{n_a}\}_{a=1}^K\right) - \left\{\left(\frac{1}{z} + c\frac{g_a(a)g_b(z)}{\sum_{i=1}^K \pi_i g_i(z)}\right)\frac{\mathbb{1}_{n_a}\mathbb{1}_{n_b}^T}{n}\right\}_{a,b=1}^K \quad (15)$$

in the sense that $\frac{1}{n}\operatorname{tr}(D_nQ_z) - \frac{1}{n}\operatorname{tr}\left(D_n\bar{Q}_z\right) \to 0$ and $d_{1,n}^T\left(Q_z - \bar{Q}_z\right)d_{2,n} \to 0$ almost surely for all deterministic Hermitian matrices D_n and deterministic vectors $d_{1,n}, d_{2,n}$ of bounded norms. Thereby, $(g_1, \ldots, g_K)$ is the unique vector of Stieltjes transforms of real measures satisfying the fixed point equations

$$g_a(z) = \left(-zc\left[1 - \frac{1}{z}\frac{A_{\text{diag,a}}}{f'(0)} + \frac{1}{p}\operatorname{tr} C_a\bar{\bar{Q}}_z\right]\right)^{-1} \quad (16)$$

with

$$\bar{\bar{Q}}_z := \left(-z\left[I_p + \sum_{i=1}^K \pi_i g_i(z) C_i\right]\right)^{-1}. \quad (17)$$

For given z, the value of $g_a(z)$ in (16) can be found numerically using a fixed point iteration. Lemma 1 allows us to obtain Theorem 2 by replacing the random matrix Q by its deterministic equivalent $\bar{Q}$.

Theorem 2. *Define*

$$\Gamma_z := \frac{1}{p} J^T P \bar{Q}_z P J, \tag{18}$$

$$G_z := I_K + \left(\frac{f''(0)}{2f'(0)} T - z M^T \bar{\bar{Q}}_z M \right) \Gamma_z \tag{19}$$

and let $\rho \in \mathbb{R}$ be away from the eigenvalue support of $P\frac{W^TW}{p}P + \frac{A_{\text{diag}}}{f'(0)}$ such that G_ρ has a zero eigenvalue of multiplicity m_ρ. Then L' has m_ρ eigenvalues $\lambda_i, \ldots, \lambda_{i+m_\rho-1}$ converging to ρ as $n, p \to \infty$.

This result shows that the analysis of outlying eigenvalues can be done on a completely deterministic $K \times K$ matrix. A general intuition here is that the bigger the distance of these eigenvalues is from the limiting support of $P\frac{W^TW}{p}P + \frac{A_{\text{diag}}}{f'(0)}$ and also from each other, the better clustering results can be expected. However, to make this claim more rigorous we need to study the behavior of the corresponding eigenvectors. Thereby, we omit the analysis of the eigenvector $D^{\frac{1}{2}}\mathbb{1}_n$, the treatment of which is very different from the rest, and will provide it in an extended version of this article.

Let $\lambda_l, \ldots, \lambda_{l+m_\rho-1}$ be isolated eigenvalues of L' given by Theorem 2, all converging to ρ. Furthermore, let Π_ρ be the projector onto the eigenspace associated with these eigenvalues. Then the $K \times K$ matrix

$$Y := \frac{1}{n} \operatorname{diag}(\pi)^{-1} J^T \Pi_\rho J \operatorname{diag}(\pi)^{-1}, \tag{20}$$

where $\pi = (\pi_1, \ldots, \pi_K)$, carries information on the alignment between Π_ρ and J. In particular, if $m_\rho = 1$ we have $\Pi_\rho = u_l u_l^T$ with u_l being the eigenvector corresponding to eigenvalue λ_l of L' converging to ρ. Up to its sign this eigenvector can be modelled as

$$u_l = \sum_{a=1}^{K} \alpha_a^l \frac{j_a}{\sqrt{n_a}} + \sigma_a^l w_a^l. \tag{21}$$

Thereby, w_a^l is a random vector of norm one orthogonal to j_a, supported on the indices $i \in \mathcal{C}_a$ whereas α_a^l and σ_a^l are the coefficients denoting the component-wise alignment of u_l to j_a and the standard deviation of the eigenvector fluctuation around its component-wise mean $\alpha_a^l \frac{j_a}{\sqrt{n_a}}$, respectively. The diagonal values of Y carry information about the component-wise eigenvector means, namely, $Y_{aa} = |\alpha_a^l|^2$ while for two components a and b the off-diagonal element Y_{ab} provides the sign of the product of these means, i.e, $\operatorname{sgn}(Y_{ab}) = \operatorname{sgn}(\alpha_a^l \alpha_b^l)$. Similarly, the sum of the component-wise variances is given as

$$\sum_{a=1}^{K} (\sigma_a^l)^2 = (1 - \operatorname{tr} Y) m_\rho. \tag{22}$$

In Theorem 3 we show that Y is asymptotically deterministic.

Theorem 3. *As $n, p \to \infty$*

$$\frac{1}{p} J^T \Pi_\rho J = -\Gamma_\rho \sum_{i=1}^{m_\rho} \frac{(V_{r,\rho})_i (V_{l,\rho})_i^T}{(V_{l,\rho})_i^T G'_\rho (V_{r,\rho})_i} + o(1) \tag{23}$$

where $V_{l,\rho}$ and $V_{r,\rho}$ are $K \times m_\rho$ matrices containing left and right eigenvectors of G_ρ corresponding to the eigenvalue zero as their columns, respectively. Furthermore, G'_ρ is the derivative of G_z w.r.t. z evaluated at $z = \rho$.

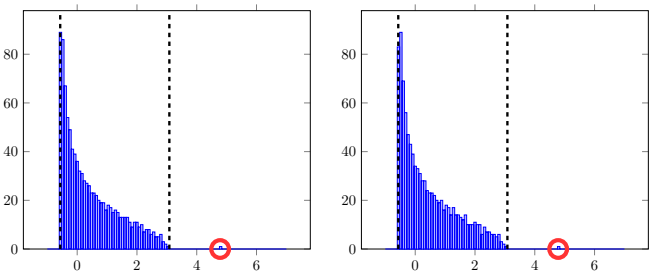
Hence, the statistical parameters of the eigenvectors depend on the same deterministic rank K matrix G_ρ as the positions of their corresponding isolated eigenvalues.

4. SPECIAL CASE

The results given in Section 3 do not allow for an immediate interpretation. Therefore, we now provide a compelling case study making full use of our previously stated findings. Assume that $K = 2$ and let the covariance matrices be $C_1 = I_p + \Delta$, $C_2 = I_p - \Delta$ where Δ is any $p \times p$ matrix satisfying $\operatorname{tr} \Delta = 0$ such that C_1 and C_2 are positive definite and set $\pi_1 = \pi_2 = 0.5$. In this case it can be shown that $g_1(z) = g_2(z)$ have the explicit solution denoted as

$$g_0(z) = c^2 \Big[z^2 - 2 \left(1 + \frac{a_0}{f'(0)} + \frac{1}{c} \right) z + 1 + \frac{1}{c^2} + 2\frac{a_0}{f'(0)} + \left(\frac{a_0}{f'(0)} \right)^2 + \frac{2a_0}{f'(0)\, c} - \frac{2}{c} \Big] \tag{24}$$

where $a_0 = f(1) - f'(0) - f(0)$. It can be concluded



(a) Histogram of L' (b) Histogram of $\hat{L}'$

Fig. 1. Comparison of the histograms of L' and $\hat{L}'$

from (24) that the eigenvalues of $P\frac{W^TW}{p}P + P\frac{A_{\text{diag}}}{f'(0)}P$ are distributed according to a shifted version of the Marchenko-Pastur law with support $[S_-, S_+] \cup \delta_{c<1}\{0\}$ where

$$S_\pm = \frac{a_0}{f'(0)} + \left(1 \pm \sqrt{\frac{1}{c}} \right)^2. \tag{25}$$

Focusing on isolated eigenvalues greater than S_+ we identify that their existence as per Theorem 2 is equivalent to

$$0 \geq g_0(\rho) \geq -\frac{1}{1 + \sqrt{c}}, \tag{26}$$

where

$$g_0(\rho) \in \left\{ -\frac{1}{2}s \pm \frac{1}{2}\sqrt{s^2 - \frac{4}{\frac{f''(0)}{2f'}\frac{1}{p}\operatorname{tr}\Delta^2}} \right\} \tag{27}$$

with

$$s := 1 + \frac{1 + \frac{1}{4}\left\|\mu_1 - \mu_2\right\|^2}{\frac{f''(0)}{2f'}\frac{1}{p}\operatorname{tr}\Delta^2}. \tag{28}$$

We find that there is at maximum one such eigenvalue λ_1. For the matrix Y of its corresponding eigenvector u_1 we obtain the explicit solution

$$Y = \frac{2cg_0(\rho)\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}}{g_0'(\rho)\left(2\frac{f''(0)}{f'(0)}\frac{1}{p}\operatorname{tr}\Delta^2 + \frac{4}{(1+g_0(\rho))^2}\left\|\mu_1\right\|^2\right)}. \tag{29}$$

Due to the symmetry of this model we are able to infer that the component-wise mean magnitudes and variances of this eigenvector do not depend on the component. Therefore, we obtain from (22) that

$$\left(\sigma_1^1\right)^2 = \left(\sigma_2^1\right)^2 = \frac{1}{2}\left(1 - \operatorname{tr} Y\right). \tag{30}$$

Assuming that the fluctuations have a Gaussian distribution, we are able to express the probability of clustering error, i.e., the probability that a sample gets associated with the wrong component, based on this eigenvector only as

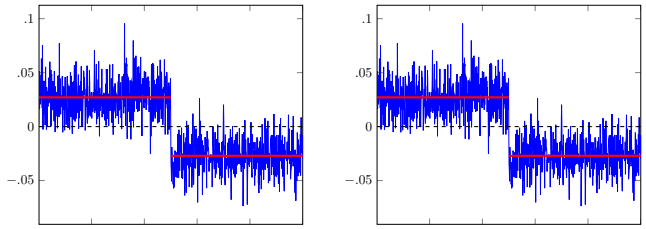
$$P_{\text{err}} = \Phi\left(-\sqrt{\frac{2Y_{11}}{1 - 2Y_{11}}}\right) \tag{31}$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

To demonstrate the accuracy of our results we instantiate the specialized model with the parameters $p = 1200$, $n = 1000$, and $c = 1.2$. We set $\Delta = \operatorname{diag}\left(0.3I_{600}, -0.3I_{600}\right)$ and $\mu_1 = 4\sqrt{\frac{1}{p}}\left\{\delta_{i\leq 600}\right\}_{i=1}^{p}$, $\mu_2 = 4\sqrt{\frac{1}{p}}\left\{\delta_{601\leq i}\right\}_{i=1}^{p}$ and we choose f to be the (indefinite) Sigmoid kernel [10], i.e.,

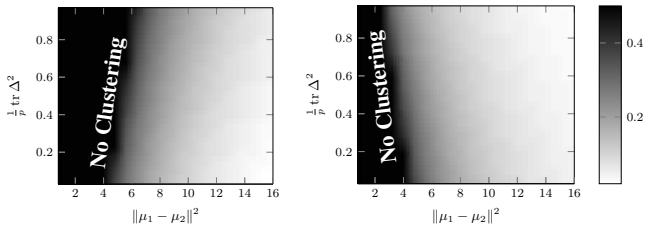
$$f\left(\frac{\langle x_i, x_j\rangle}{p}\right) = \tanh\left(\alpha\frac{\langle x_i, x_j\rangle}{p} + \theta\right) \tag{32}$$

with $\alpha = 1.2$ and $\theta = 1$. To ensure visual interpretability of the results, the samples are generated in component-wise order (which does not affect the spectrum). Figure 1 shows the eigenvalue histograms of the matrices L' and its approximation $\hat{L}'$ with the predicted bounds $S_{\pm}$ depicted by vertical dashed lines. Note that in both Fig. 1 (a) and (b) there is exactly one isolated eigenvalue $\lambda_1 \approx 4.86$. The corresponding eigenvectors and their theoretical means are depicted in Fig. 2. Approximately 5.0% of the samples cannot be clustered correctly. The simulation is consistent with our theoretical prediction yielding $\rho \approx 4.80$ and $P_{\text{err}} \approx 0.049$. Figure 3



(a) Eigenvector of L' (b) Eigenvector of $\hat{L}'$

Fig. 2. Eigenvectors corresponding to the largest eigenvalue and their theoretical means (thick red lines)



(a) Sigmoid parameter $\theta = 1$ (b) Sigmoid parameter $\theta = -1$

Fig. 3. P_{err} as a function of $\frac{1}{p}\operatorname{tr}\Delta^2$ and $\|\mu_1 - \mu_2\|^2$

demonstrates the behavior of the theoretical error probability P_{err} when only u_1 is used for clustering as a functional of the differences of the component means and the parameter Δ. $P_{\text{err}} = 0.5$ means that clustering is completely random whereas P_{err} close to zero suggests a good component separability. Note that the impact of the covariance matrix and the means on the clustering quality can be drastically influenced by the sign of $\frac{f''(0)}{f'(0)}$ which we change in Fig. 3 (a) in (b) by choosing different values for the Sigmoid parameter θ.

5. CONCLUSIONS

The spectral analysis conducted in this work enables important insights into the eigenpair behavior of random graph Laplacians. We characterized the existence and positions of isolated eigenvalues and provided a probabilistic model of the corresponding eigenvectors essential to spectral clustering. Explicit interpretable results as well as an error probability function were given for a specialized model and the impact of the GMM parameters on the possibility to cluster was shown to be influenced by the signs of the kernel derivatives. Simulations illustrated the high accuracy of our theoretical findings in finite size settings. Since the used tools are not restricted to GMMs it is likely that our results will generalize to other generative and even non-parametric models. We believe that our work can serve as a strong basis for a better understanding and the enhancement of the final step of spectral clustering often performed using the naive K-means algorithm.

6. REFERENCES

[1] Francis R. Bach and Michael I. Jordan, "Learning spectral clustering, with application to speech separation," *Journal of Machine Learning Research*, vol. 7, pp. 1963–2001, Dec. 2006.

[2] Ravi Kannan, Santosh Vempala, and Adrian Vetta, "On clusterings: Good, bad and spectral," *J. ACM*, vol. 51, no. 3, pp. 497–515, May 2004.

[3] Santosh Vempala and Grant Wang, "A spectral algorithm for learning mixture models," *Journal of Computer and System Sciences*, vol. 68, no. 4, pp. 841–860, 2004.

[4] Ravindran Kannan and Santosh Vempala, "Spectral algorithms," *Foundations and Trends in Theoretical Computer Science*, vol. 4, no. 34, pp. 157–288, 2008.

[5] Romain Couillet and Florent Benaych-Georges, "Understanding big data spectral clustering," *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Cancun, Mexico*, 2015.

[6] Ulrike von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

[7] Noureddine E. Karoui, "The spectrum of kernel random matrices," *Annals of Statistics*, vol. 38, no. 1, pp. 1–50, 2010.

[8] Florent Benaych-Georges and Raj Rao Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[9] Leonid Pastur and Mariya Shcherbina, *Eigenvalue Distribution of Large Random Matrices*, American Mathematical Society Providence, RI, 2011.

[10] Hsuan-Tien Lin and Chih-Jen Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," *submitted to Neural Computation*, pp. 1–32, 2003.