Consistent Spectral Methods for Dimensionality Reduction

Malika Kharouf
Charles Delaunay Institut, CNRS UMR 6281
University of Technology of Troyes
Troyes, France
Email: malika.kharouf@utt.fr

Tabea Rebafka
LPSM, UMR 8001
Sorbonne University (Paris 6)
Paris, France
Email: tabea.rebafka@upmc.fr

Nataliya Sokolovska
INSERM UMR S 1166, NutriOmics Team
Sorbonne University (Paris 6)
Paris, France
Email: nataliya.sokolovska@upmc.fr

***Abstract*—This paper addresses the problem of dimension reduction of noisy data, more precisely the challenge to determine the dimension of the subspace where the observed signal lives in. Based on results from random matrix theory, two novel estimators of the signal dimension are proposed in this paper. Consistency of the estimators is proved in the modern asymptotic regime, where the number of parameters grows proportionally with the sample size. Experimental results show that the novel estimators are robust to noise and, moreover, they give highly accurate results in settings where standard methods fail. We apply the novel dimension estimators to several life sciences benchmarks in the context of classification, and illustrate the improvements achieved by the new methods compared to the state-of-the-art approaches.**

I. Introduction

Dimensionality reduction aims at separating signal from noise in order to preserve significant properties of data in a low-dimensional space before analyzing them by further statistical methods. Data representation in a lower dimension is needed in many applications, such as computer vision, natural language processing or bioinformatics, where the number of observed features or parameters has considerably increased mainly due to technical advances. Nevertheless, increasing the number of features does not automatically increase the dimension of the subspace where the signal lives in. Indeed, in speech recognition [11], wireless communications [26], hyperspectral imaging [18], chemometrics [16], medical imaging [4], genomics [22], mathematical finance [17], or wireless communications [14], [3], the signal space dimension is much lower than the number of observed parameters. Thus, a challenge is to determine the low-dimensional signal space in order to project the data onto it, as it is done, for instance, by principal component analysis (PCA) [15]. In this context, a fundamental question is how to determine an optimal minimal dimension of a high-dimensional problem. A major difficulty in real data sets is the presence of noise, making the estimation of the signal space rather involved.

A number of methods to determine an optimal dimension have been proposed (see [6] for an overview). The most prominent method [15] uses the number of principal components that are necessary to explain a given part of the total variance. The scree test [7] which relies on the detection of an elbow in the scree graph, that is the plot of ordered sample eigenvalues, is also widely used in practice. Both methods are rather heuristical. Further methods are the SURE method [27], [28], model-order selection in a Bayesian framework [4], [24] or maximum-likelihood based approaches [25]. Although they achieve reasonable performance, most of them lack theoretical explanations.

Another recent approach introduced by [21] is based on eigengaps, that is the distance between consecutive sample eigenvalues. The novelty of this method (for both white [21] and colored [10] noise) is that a sound mathematical foundation is provided stemming from results in random matrix theory. Notably consistency in the modern asymptotic regime is proved, when both the sample size *and* the number of parameters tend to infinity. This is most relevant for applications where the number of parameters is comparable with or is even larger than the number of observations.

Despite theoretical guarantees, the eigengap method [21] suffers from that it relies on local features of the scree graph. If a single sample eigenvalue is badly estimated, the dimension estimate may be very erroneous. Indeed, in practice high accuracy is only obtained when the signal eigenvalues are well separated.

We propose a method that is robust to the presence of similar or even identical signal eigenvalues, while preserving the strong theoretical properties of the eigengap method. *This is achieved by a more global look on the sample eigenvalues.* It is noteworthy that the consistency of the proposed eigenrange and threshold methods does not depend on strong distributional assumptions like normality as it is the case of the maximum-likelihood approaches and others. Moreover, we address the problem of estimating the variance from the data. *Both the eigenrange and threshold methods perform very competitively on real data.*

The paper is organised as follows. Section II introduces our methods and provides the theoretical foundations. Section III illustrates the performance of the eigenrange and threshold methods in the context of classification on real data. Concluding remarks and perspectives close the paper in section IV.

II. New estimators of the signal space dimension

In this section we introduce the mathematical framework and present the new eigenrange and threshold methods. We

ISBN 978-90-827970-1-5 © EURASIP 2018

show the consistency of the new estimators when the noise level is known. Then we address the problem of unknown variance, which is most relevant for real applications. First, an estimator of the variance is proposed and is shown to be consistent. Second, an iterative procedure is proposed to deal with unknown noise level.

A. Spiked population model

We consider the additive noise model, where the signal vector $\mathbf{s} \in \mathbb{R}^p$ is corrupted by some additive white noise $\mathbf{e}$:

$$\mathbf{y} = \mathbf{s} + \mathbf{e}\,. \tag{1}$$

The random vectors $\mathbf{s}$ and $\mathbf{e}$ are supposed to be independent and the noise $\mathbf{e}$ has zero mean and covariance $\sigma^2\mathbf{I}_p$, $\sigma^2 > 0$. Denoting the signal's covariance matrix by R_s, the covariance of the observation vector $\mathbf{y}$ verifies $R_y = R_s + \sigma^2\mathbf{I}_p$.

Often the signal $\mathbf{s}$ is a linear combination of a relatively small number of predictors, i.e. $\mathbf{s} = \mathbf{\Sigma x}$ with some $(p \times r)$ matrix $\mathbf{\Sigma}$ and $r < p$. In other words, the signal lives in a proper subspace of $\mathbb{R}^p$ of dimension r. Signal space dimension r is then given by the rank of the signal's covariance matrix R_s.

Denote the non zero eigenvalues of R_s by $\alpha_1 > \cdots > \alpha_r > 0$ and the eigenvalues of R_y by $\lambda_1 \geq \cdots \geq \lambda_p > 0$. In this model, which is also referred to as the *spiked population model*, the eigenvalues verify

$$\lambda_l = \begin{cases} \alpha_l + \sigma^2, & l = 1, \ldots, r \\ \sigma^2, & l > r \end{cases}$$

The first r eigenvalues $\lambda_1, \ldots, \lambda_r$ are called *spikes* and they are larger than the nonspiked ones. As a consequence, the eigenvalues of R_y yield important information on the signal dimension r.

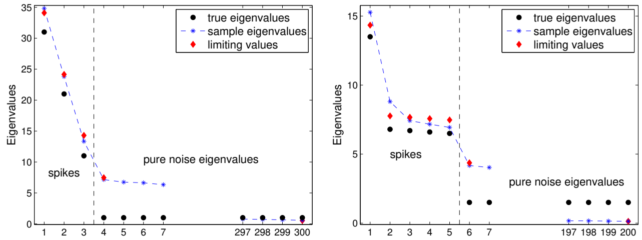
Now let the observations $y_1, \ldots y_n$ be n i.i.d. realizations of $\mathbf{y}$ given by the following spiked model

$$\mathbf{y} = \mathbf{\Sigma^{1/2}x} + \mathbf{e}\,, \tag{2}$$

where, $\mathbf{x} \in \mathbb{R}^p$ is a vector of i.i.d. zero mean and variance 1 entries, and $\mathbf{\Sigma}$ is the theoretical covariance matrix of the observations given by: $\mathbf{\Sigma} = \mathbf{V}\mathrm{diag}\left(\alpha_1, \ldots, \alpha_r, 0, \ldots, 0\right)\mathbf{V}^t$, Then the sample covariance matrix is $\hat{R}_y = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^t$ where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. The associated sample eigenvalues are denoted by $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

B. New estimation methods when σ^2 is known

When many features are observed, the traditional asymptotic results, where the sample size n grows, while the number of features p is fixed, may be rather bad approximations of what happens on finite samples. Generally, results in the modern regime, where both n and p tend to infinity, provide much better approximations of the finite sample situation. Concerning the estimation of the eigenvalues λ_l by the sample ones $\hat{\lambda}_l$, notable works including [2] state consistency in the traditional regime. However, consistency no longer holds when both n and p grow. Recent advances in random matrix theory provide convergence results in the modern asymptotic regime.



(a) $r = 3$, $\alpha = [30, 20, 10]$, $\sigma^2 = 1$, $c = 3$, $n = 100$, $p = 300$

(b) $r = 5$, $\alpha = [12, 5.3, 5.2, 5.1, 5]$, $\sigma^2 = 1.5$, $c = 0.5$, $n = 400$, $p = 200$

Fig. 1: Illustration of spikes and pure noise eigenvalues, and of the bias of the sample eigenvalues: eigenvalues λ_l (black dots), sample eigenvalues $\hat{\lambda}_l$ (blue stars), theoretical limits of the sample eigenvalues (red diamonds) as $p/n \to c$.

In the pure noise case, where $\mathbf{y} = \mathbf{e}$ and $R_y = \sigma^2\mathbf{I}_p$, the seminal work of Marchenko and Pastur [19] shows that, when $p/n \to c$, all limiting values ϕ_l of the sample eigenvalues $\hat{\lambda}_l$ lie within the interval $[a, b] := [\sigma^2(1 - \sqrt{c})^2, \sigma^2(1 + \sqrt{c})^2]$, which is the support of the so-called Marcenko-Pastur law.

In the additive noise model, the nonspiked sample eigenvalues still tend to lie in the Marchenko-Pastur interval $[a, b]$, while the limits of the spikes are outside [3]. This result holds under the assumption that spikes are sufficiently different from pure noise eigenvalues. More formally, if $\alpha_l > \sigma^2\sqrt{c}$ for $l = 1, \ldots, r$ and $p/n \to c$, then for $l = 1, \ldots, r$

$$\hat{\lambda}_l \longrightarrow \phi_l = \alpha_l + \sigma^2\left(1 + c + \frac{c\sigma^2}{\alpha_l}\right) \quad a.s. \tag{3}$$

Moreover, the first and last pure noise eigenvalues satisfy

$$\begin{aligned} \hat{\lambda}_{r+1} &\longrightarrow b = \sigma^2(1 + \sqrt{c})^2 \quad a.s. \\ \hat{\lambda}_m &\longrightarrow a = \sigma^2(1 - \sqrt{c})^2 \quad a.s. \end{aligned} \tag{4}$$

where $m = \min(n, p)$. It is clear that any procedure relying on sample eigenvalues must take into consideration their bias, that is the difference between the limiting values of the sample eigenvalues $\hat{\lambda}_l$ and the model eigenvalues λ_l. This bias is illustrated in Figure 1.

These asymptotic properties of the sample eigenvalues can be used to derive new estimators of the signal space dimension r. First, as $\phi_l > b$ for $l = 1, \ldots, r$ and $\hat{\lambda}_{r+1} \longrightarrow b$ *a.s.*, the number of sample eigenvalues exceeding b shall be close to r. Taking into account the uncertainty of the sample eigenvalues, it is more appropriate to choose a threshold of the form $b + c_n$, where c_n is a sequence to be chosen. Thus, a first estimator of the signal space dimension is given by

$$\hat{r}^{\text{thresh}} = \#\{l : \hat{\lambda}_l > b + c_n\} = \max\{l : \hat{\lambda}_l > b + c_n\}. \tag{5}$$

We refer to $\hat{r}^{\text{thresh}}$ as the *threshold method.*

Second, the range of the pure noise sample eigenvalues, $\hat{\lambda}_{r+1} - \hat{\lambda}_m$, is about $b - a$, while the distance $\hat{\lambda}_l - \hat{\lambda}_m$ for any $l = 1, \ldots, r$ is significantly larger. From this viewpoint, a

ISBN 978-90-827970-1-5 © EURASIP 2018

natural estimator of the signal dimension r is derived as the number of sample eigenvalues that must be discarded such that the remaining eigenvalues are contained in an interval of approximate length $b-a$. Denote $\delta_l = \hat{\lambda}_{l+1} - \hat{\lambda}_m$ and consider a threshold of the form $b - a + d_n$. Then an estimator of the signal space dimension is defined by

$$\hat{r}^{\text{range}} = \#\{l : \delta_l \geq b - a + d_n\} = \min\{l : \delta_l < b - a + d_n\}. \tag{6}$$

We refer to $\hat{r}^{\text{range}}$ as the *eigenrange method.*

C. Consistency

We show that both the eigenrange estimator $\hat{r}^{\text{range}}$ and the threshold method $\hat{r}^{\text{thresh}}$ are consistent estimators of the signal space dimension r for an appropriate choice of the sequences c_n in (5) and d_n in (6). The proof relies on the rates of convergence of the smallest and the largest pure noise sample eigenvalues when $p/n \to c$. According to [8] and [5],

$$n^{2/3}(\hat{\lambda}_{n,r+1} - b) = O_{\mathbb{P}}(1), \;\; n^{2/3}(\hat{\lambda}_{n,m_n} - a) = O_{\mathbb{P}}(1), \tag{7}$$

where the notation $X_n = O_{\mathbb{P}}(1)$ means that X_n is a stochastically bounded sequence. To ease the understanding of the asymptotics, in this section subscript n is added to all quantities depending on n. An event holds almost surely if it holds with probability tending to 1 as n and p tend to infinity.

Theorem 1. *Let the sequence d_n be such that $d_n \to 0$ and $n^{2/3} d_n \to \infty$ as $n \to \infty$. Suppose that the signal eigenvalues satisfy $\alpha_k > \sigma^2\sqrt{c}$ for $k = 1, \ldots, r$. Then under a moment conditions as in model 2, the eigenrange estimator $\hat{r}^{range}$ defined in* (6) *is a consistent estimator of the dimension of the signal space, that is, as $p/n \to c$,*

$$\hat{r}_n^{range} \longrightarrow r \quad \textit{almost surely.}$$

Proof. Without loss of generality let $\sigma^2 = 1$. Denote $\tilde{d}_n = b - a + d_n$. As $\delta_{n,l} > \delta_{n,l-1}$ for all l, we have

$$\{\hat{r}_n^{\text{range}} = r\} = \{\delta_{n,r} < \tilde{d}_n\} \cap \{\delta_{n,r-1} \geq \tilde{d}_n\},$$

implying that

$$\begin{aligned}\mathbb{P}(\hat{r}_n^{\text{range}} = r) &= 1 - \mathbb{P}\left(\{\delta_{n,r} \geq \tilde{d}_n\} \cup \{\delta_{n,r-1} < \tilde{d}_n\}\right) \\ &\geq 1 - \mathbb{P}(\delta_{n,r} \geq \tilde{d}_n) - \mathbb{P}(\delta_{n,r-1} < \tilde{d}_n).\end{aligned}$$

On the one hand, by (7) and as $n^{2/3} d_n \to \infty$,

$$\begin{aligned}&\mathbb{P}\left(\delta_{n,r} \geq \tilde{d}_n\right) \\ &= \mathbb{P}\left(n^{2/3}(\hat{\lambda}_{n,r+1} - b) - n^{2/3}(\hat{\lambda}_{n,m_n} - a) \geq n^{2/3} d_n\right) \\ &\longrightarrow 0.\end{aligned}$$

On the other hand, it holds that

$$\begin{aligned}\mathbb{P}(\delta_{n,r-1} <\tilde{d}_n) &= \mathbb{P}\left(\hat{\lambda}_{n,r} - \hat{\lambda}_{n,m_n} < d_n + b - a\right) \\ &= \mathbb{P}\left((\hat{\lambda}_{n,r} - \phi_r) - (\hat{\lambda}_{n,m_n} - a) - d_n < b - \phi_r\right) \\ &\longrightarrow 0,\end{aligned}$$

since $b - \phi_r = -\frac{(\alpha_r - \sqrt{c})^2}{\alpha_r} < 0$ and $(\hat{\lambda}_{n,r} - \phi_r) - (\hat{\lambda}_{n,m_n} - a) - d_n \xrightarrow{\mathbb{P}} 0$ by (3), (7) and as $d_n \to 0$. Combining all arguments yields $\mathbb{P}(\hat{r}_n^{\text{range}} = r) \to 1$. This completes the proof. □

The proof showing that

$$\hat{r}_n^{\text{thresh}} \longrightarrow r \quad \textit{almost surely}$$

follows the same lines as the proof of Theorem 1.

We conducted an extensive simulation study to calibrate the sequences c_n and d_n, and concluded that the best choice for both r^{thresh} and r^{range} is given by

$$c_n = d_n = \sigma^2 n^{1/20 - 2/3}. \tag{8}$$

D. Consistent estimator of the variance σ^2

In practice the noise level σ^2 is generally unknown, though required in (8) for the estimators. As the smallest non null sample eigenvalue $\hat{\lambda}_m$ converges to $a = \sigma^2(1 - \sqrt{c})^2$, a consistent estimator of the unknown variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{\hat{\lambda}_m}{(1 - \sqrt{c})^2}. \tag{9}$$

Theorem 2. *If $c \neq 1$, the estimator $\hat{\sigma}^2$ of σ defined in* (9) *is consistent, i.e.*

$$\hat{\sigma}^2 \longrightarrow \sigma^2, \qquad n/p \to c, n \to \infty.$$

Proof. By (4). □

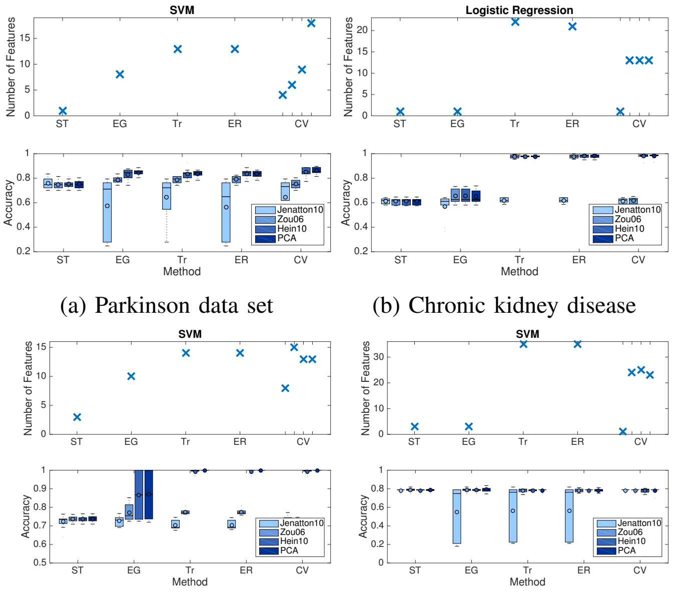
In the case of unknown σ^2, both the eigenrange and the threshold estimator can be used with $\hat{\sigma}^2$ instead of σ in (8).

E. An iterative procedure when σ^2 is unknown

An alternative approach to deal with the unknown σ^2 case consists in an iterative procedure (similar to that in [21]) to estimate both r and σ^2. More precisely, we alternate the estimation of σ^2 (using the current value r^{curr} of r) and the estimation of r by the eigenrange or the threshold method (with σ^2 replaced by its current estimate $\hat{\sigma}^2_{\text{curr}}$) until convergence. To estimate σ^2 we use that the mean of the pure noise sample eigenvalues $\hat{\lambda}_l$ converges to σ^2. The whole procedure is given in Algorithm 1.

Algorithm 1 Iterative procedure when σ^2 is unknown

```
1: Initialize $r^{\text{curr}} = 0$.
2: Estimate $\sigma^2$ by $\hat{\sigma}^2_{\text{curr}} = \frac{1}{p - r^{\text{curr}}} \sum_{l=r^{\text{curr}}+1}^{p} \hat{\lambda}_l$.
3: Compute the eigenrange estimate $\hat{r}^{\text{eigen}}$ (threshold estimate
   $\hat{r}^{\text{thresh}}$) where $\sigma^2$ is replaced with $\hat{\sigma}^2_{\text{curr}}$ in (8).
4: If $r^{\text{curr}} < \hat{r}$, update $r^{\text{curr}} = \hat{r}$ and return to 2, otherwise
   $r^{\text{curr}}$ is the final estimate of $r$.
```



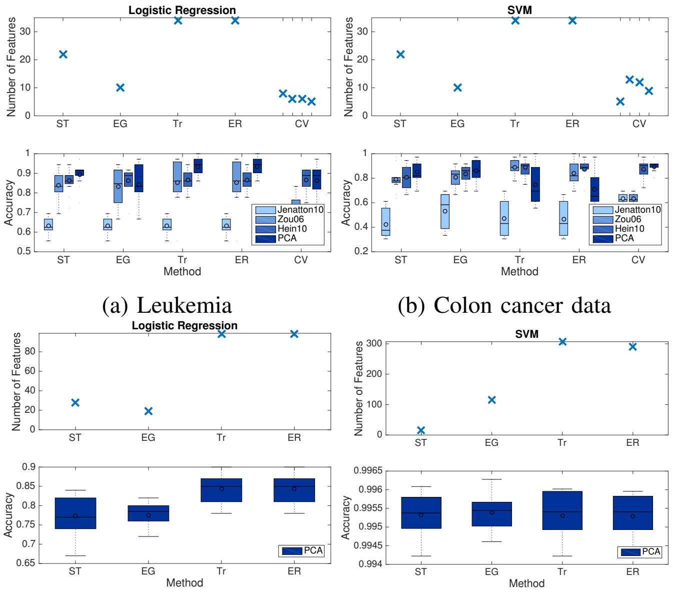
(a) Parkinson data set

(b) Chronic kidney disease

(c) Diabetic retinopathy

(d) SPECTF heart data

Fig. 2: Estimated ranks and performance for data sets with a small number of observations and features.



(a) Leukemia

(b) Colon cancer data

(c) Arcene data

(d) p53 mutants set

Fig. 3: Estimated ranks and performance for scenarios where the number of parameters is much bigger than the number of observations (above), and where the number of both observations and features is getting big.

III. Application to Classification

To illustrate the importance of accurate estimation of the signal space dimension, we show how the eigenrange and threshold methods improve results in the context of classification tasks.

To increase computational efficiency and also to reduce noise in data, several versions of PCA with sparse loadings have been proposed, namely the structured sparse PCA [13] to which we refer in the following as *Jenatton10*, sparse PCA that we call *Zou06* method [29], and the inverse power method applied to sparse PCA called *Hein10* [12]. The data used in our experiments are benchmarks from life sciences (most of them from the UCI Machine Learning repository[1]).

- Parkinson data set is composed of 197 observations and 22 features containing biomedical voice measurements. The task consists in to predict whether a patient is healthy or ill.
- Chronic kidney disease data set classifies 400 patients into ill and healthy based on 24 medical attributes such as blood pressure, bacteria, hemoglobin, etc.
- Diabetic retinopathy Debrecen cohort contains information extracted from images whether an image contains signs of diabetic retinopathy or not. The number of instances is 1151, and the number of features is quite limited, namely, 19.
- SPECTF heart data set includes 267 observations and 44 features corresponding to SPECT images for patients with and without cardiac problems.
- Molecular classification of leukemia data set [9] contains gene expressions of 72 patients and 3562 genes.
- Colon cancer data set of [1] consists of 62 patients and 2000 gene expressions of colon adenocarcinoma tissues.
- The aim of p53 mutants data set is to models mutant p53 transcriptional activity (active or not) from biophysical simulations data. There are 31159 instances and 5408 continuous features.
- Arcene task is dedicated to distinguish cancer versus normal patterns from mass-spectrometric data. The number of observations is equal to 200, and the number of continuous features equals 10000.

While choosing the data sets, we tried to consider three realistic scenarios. The first scenario is a setting where the number of observations and original dimensions are relatively small. The gene expression data reflect a scenario where the number of observations is much smaller than the number of features, and the p53 mutations data sets can be seen as a case where both the numbers of observations and of features are big.

We find optimal dimensions of new reduced data using the scree test (ST), eigengap (EG), eigenrange (ER), and thresholding (Tr) methods. We run the logistic regression and SVM on the reduced data to make predictions. We perform 10-fold cross validation, and boxplot the obtained accuracies. Note that the estimated ranks and projections are learned from training data. We also test a heuristic which is computationally expensive and that finds an optimal number of components by cross validation (CV).

Figure 2 illustrates the performance and the estimated ranks for Parkinson, Chronic Kidney, Diabetic Retinopathy,

[1] http://archive.ics.uci.edu/ml/datasets.html

and SPECTF Heart data sets. Figure 3 shows our results on Leukemia, Colon Cancer, p53 Mutant, and Arcene tasks. The logistic regression and the SVM reached quite a similar performance on the reduced data, and for each data set, we show the results for a method which achieved a better accuracy.

We have observed that the proposed eigenrange and threshold methods tend to find a bigger rank than the scree test and the eigengap which are quite stringent. The accuracies of the proposed methods either achieve the state-of-the-art performance, or outperform it. We have noticed, that the standard PCA reaches in our experiments an optimal performance, and, taking into consideration that the computational complexity of the considered sparse PCA methods is quite high, we decided to run only the standard PCA on two biggest data sets. On the Arcene data (Figure 3, c)) we clearly see that the eigenrange and threshold methods find a rank which leads to a better performance than the state-of-the-art methods.

IV. Conclusion

Dimensionality reduction is a challenge, especially in applications where data are noisy. We proposed two highly accurate estimators of the signal dimension, based on global statistics involving sample eigenvalues, that outperform such state-of-the-art methods as scree test and eigengap that focus on local features of the scree graph. Based on recent advances of random matrix theory, we formulated our main theoretical result summarized in Theorem 1. The simulated and real data experiments confirm our theoretical findings and, moreover, show the robustness in situations where the state-of-the-art methods fail, that is when spiked eigenvalues are close to each other. The novel eigenrange and threshold methods are of great interest for applications where assumptions on the spiked eigenvalues or on the distribution of the observations are difficult to assert. A further advantage of the proposed method is that it is straightforward to implement, and it is not computationally expensive.

Our method is not directly connected to classification, since the proposed dimensionality reduction procedure is completely unsupervised. However, we have tested the algorithm to reduce data before applying a prediction method, and we observed a clearly beneficial effect. It is promising to integrate the eigenrange and threshold approaches into other methods that rely on the eigenstructure of a data matrix. In particular, we are currently investigating its extension to the problem of spectral clustering.

Acknowledgment

This work was supported by PEPS 2016 and 2017, CNRS France.

References

[1] A. Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:67456750, 1999.

[2] T. W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, 34:122–148, 1963.

[3] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97:1382–1408, 2006.

[4] C. F. Beckmann and S. M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging*, 23(2):137–152, 2004.

[5] F. Benaych-Georges, A. Guionnet, and M. Maïda. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electronic Journal of Probability*, 16, 2011.

[6] R. Cangelosi and A. Goriely. Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*, 2(2), 2007.

[7] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1:245276, 1966.

[8] R. Couillet and W. Hachem. Fluctuations of spiked random matrix models and failure diagnosis in sensor networks. *IEEE Transactions on Information Theory*, 59(1):509–525, 2013.

[9] T.R. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[10] A. Halimi, P. Honeine, M. Kharouf, C. Richard, and J.-Y. Tourneret. Estimating the Intrinsic Dimension of Hyperspectral Images Using a Noise-Whitened Eigengap Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):3811–3821, 2016.

[11] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 02 1995.

[12] M. Hein and T. Buehler. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. In *NIPS*, 2010.

[13] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *AISTATS*, 2010.

[14] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[15] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[16] S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19 – 32, 2008.

[17] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(3):391–397, 2000.

[18] D. Landgrebe. Hyperspectral image data analysis as a high dimensional signal processing problem. *Special Issue of the IEEE Signal Processing Magazine*, 19:17–28, 2002.

[19] V. A. Marcenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sbornik*, 1:457–486, 1967.

[20] A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, November 2010.

[21] D. Passemier and J. F. Yao. On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices : Theory and Applications*, 1, 2012.

[22] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, August 2006.

[23] B. Rosenow. Determining the optimal dimensionality of multivariate volatility models with tools from random matrix theory. *Journal of Economic Dynamics and Control*, 32(1):279–302, January 2008.

[24] A.-K. Seghouane and A. Cichocki. Bayesian estimation of the number of principal components. *Signal Processing*, 87(3):562–568, 2007.

[25] P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

[26] E. Telatar. Capacity of multi-antenna Gaussian channels. *Eur. Trans. Telecomm. ETT*, 10(6):585–596, November 1999.

[27] M. O. Ulfarsson and V. Solo. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. Signal Processing*, 56(12):5804–5816, 2008.

[28] M. O. Ulfarsson and V. Solo. Selecting the number of principal components with SURE. *IEEE Signal Process. Lett.*, 22(2):239–243, 2015.

[29] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.