Supplementary material: Eigenrange: A Robust Spectral Method for Dimensionality Reduction

Malika Kharouf, Tabea Rebafka, Nataliya Sokolovska

***Abstract*—This document provides additional numerical results in support of the primary article entitled Eigenrange: A robust spectral method for dimensionality reduction. The aim of this document is to illustrate the accuracy of our approach compared to traditional estimators of the dimension of the signal space. First a brief description of several state-of-the-art methods is provided, then the performance of our method is compared to that of the others on simulated data.**

I. Assessment on synthetic data

A. *Description of traditional selection methods*

A *standard method* based on PCA uses the number of principal components to explain a given part, say π, of the total variance in the data [1], [2]. In our simulations we use $\pi = 80\%$ and $\pi = 95\%$.

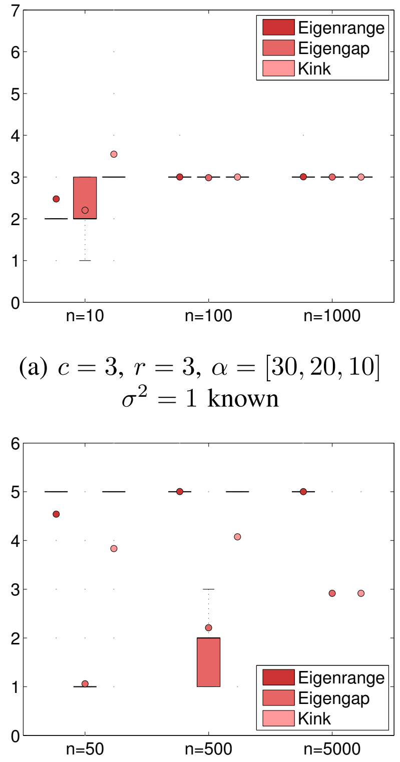
The *kink method* or *scree test* analyses the scree graph, that is the plot of the ordered sample eigenvalues [1]. As the pure noise eigenvalues $\lambda_r, \dots, \lambda_m$ are constant, the scree graph has a (more or less pronounced) kink or elbow at the $(r+1)$-th sample eigenvalue, see Figure **1** in the principal paper.

Another approach relies on so-called eigengaps, that is the distance between consecutive eigenvalues. The *eigengap method*, as formalized in [3], is intended to identify the beginning of the flat part of the scree graph, made up of the pure noise eigenvalues. This method is shown to be consistent when p/n tends to some constant c.

B. *Comparison to the state of the art*

To compare the performance of the different estimators we conducted an intensive simulation study. Here we present detailed results for the two settings of Figure **1** in the principal paper. Recall that in setting (a) there are three times more parameters p than observations n, i.e. $c = 3$. For the signal space dimension we have $r = 3$ and spikes take rather distinct values with $\alpha = [30, 20, 10]$. Here $\sigma^2 = 1$ is supposed to be known. In scenario (b) there are two times more observations than parameters, i.e. $c = 0.5$, which is a somewhat easier setting than the previous one. However, there are several spiked eigenvalues that are very close one to another with $\alpha = [12, 5.3, 5.2, 5.1, 5]$. Moreover, $\sigma^2 = 1.5$ is unknown.

For different values of n, 1000 datasets with Gaussian noise are generated. Figure 1 shows the resulting boxplots for the eigenrange, eigengap and the kink method, while Table I presents the success rates of the different methods, including the standard method (explained variance). The success rate is the proportion of datasets on which the estimator exactly recovers the signal dimension r.



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

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

Fig. 1. Comparison of eigenrange, eigengap and kink method for the estimation of r on 1000 simulated datasets each. The dots in the box plots represent the mean value.

The first observation is that the standard method completely fails for both $\pi = 80\%$ and $\pi = 95\%$, and in general, the signal space dimension is overestimated by far. This is not surprising as the standard method is not conceived to recover the signal space dimension and does not have the property of consistency, but it may be appropriate for different purposes as e.g. visualization. We do not further consider the method based on the explained variance.

Second, the eigenrange method has a very good performance in both settings (σ^2 known/unknown and for different values of c) for any number of observations. Moreover, the accuracy of the eigenrange method increases when n and p increase,

TABLE I
SUCCESS RATES OF EIGENRANGE (ER), EIGENGAP (EG), KINK METHOD, STANDARD METHOD BASED ON THE PART OF EXPLAINED VARIANCE (80% AND 95%) ON 1000 SIMULATED DATASETS EACH.

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

n	ER	EG	Kink	80%	95%
10	0.48	0.42	0.49	0.53	0.00
100	1.00	1.00	1.00	0.00	0.00
1000	1.00	1.00	1.00	0.00	0.00

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

n	ER	EG	Kink	80%	95%
50	0.62	0.00	0.51	0.00	0.00
500	1.00	0.14	0.77	0.00	0.00
5000	1.00	0.48	0.48	0.00	0.00

confirming the consistency of the estimator.

Finally, in setting (a) the eigenrange, eigengap and the kink method all give excellent results and their performance is very similar. Slight differences in accuracy only occur on very small datasets. However, in setting (b) we observe significant differences between the three methods: both the eigengap and the kink method are much less performant than the eigenrange method. The reason for this behavior becomes clear with a look on the scree graph in Figure **1**(b) in the principal paper. Indeed, there are four spiked eigenvalues which are almost identical, so that the scree graph captures a first elbow at the second sample eigenvalue, followed by a relatively flat part until the fifth sample eigenvalue. Hence, in this scenario, it often happens that the kink method relies on this first elbow for the estimation of r, while the eigengap method regularly interprets the flat part (starting at the second eigenvalue) as the beginning of the pure noise eigenvalues. This is our intuition why both methods tend to underestimate the signal space dimension r.

The conclusion of our simulation study is that the accuracy of the eigenrange method is very high and robust in various settings. When all signal eigenvalues take very distinct values, the eigenrange method achieves comparable results to the state of the art. However, it clearly outperforms traditional methods whenever signal eigenvalues take similar values. As in practice generally nothing is known about the values of the signal eigenvalues, we recommend the usage of the eigenrange method to recover the signal dimension r. This is further made clear on some real data examples in Section IV of the principal paper.

REFERENCES

[1] R. Cangelosi and A. Goriely, "Component retention in principal component analysis with application to cDNA microarray data," *Biology Direct*, vol. 2, no. 2, 2007.

[2] J. Harrington and S. Cassidy, *Techniques in Speech Acoustics*. New York: Springer Science and Business Media, New York, 2012.

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