2009 Ninth IEEE International Conference on Data Mining

A bootstrap approach to eigenvalue correction

Anne Hendrikse, Luuk Spreeuwers and Raymond Veldhuis
Signals and Systems group
EEMCS, University of Twente
Enschede, Netherlands
Email: a.j.hendrikse@utwente.nl

***Abstract*—Eigenvalue analysis is an important aspect in many data modeling methods. Unfortunately, the eigenvalues of the sample covariance matrix (sample eigenvalues) are biased estimates of the eigenvalues of the covariance matrix of the data generating process (population eigenvalues). We present a new method based on bootstrapping to reduce the bias in the sample eigenvalues: the eigenvalue estimates are updated in several iterations, where in each iteration synthetic data is generated to determine how to update the population eigenvalue estimates. Comparison of the bootstrap eigenvalue correction with a state of the art correction method by Karoui shows that depending on the type of population eigenvalue distribution, sometimes the Karoui method performs better and sometimes our bootstrap method.**

***Index Terms*—bootstrapping, eigenvalue correction, general statistical analysis, isotonic tree method**

I. Introduction

Second order statistics are used extensively in data modeling methods. For example, in Principle Component Analysis (PCA) (see [1]), the second order statistics of high dimensional data are used to find subspaces containing the strongest modes of variation. In Linear Discriminant Analysis (LDA), the ratio of within class and between class variance is used to find the highest discriminating directions (see [2]).

When applying these methods, it is usually assumed that the data generating process can be modeled with a multivariate probability function $P(\underline{\mathbf{x}})$, where $\underline{\mathbf{x}}$ is a multidimensional random variable. It is then assumed that $P(\underline{\mathbf{x}})$ is reasonably characterised by only the mean and second order statistics. The second order statistics of a multidimensional random variable are described by the covariance matrix, given by $\mathbf{\Sigma} = \mathcal{E}\left(\tilde{\underline{\mathbf{x}}} \cdot \tilde{\underline{\mathbf{x}}}^{\mathrm{T}}\right), \quad \tilde{\underline{\mathbf{x}}} = \underline{\mathbf{x}} - \mathcal{E}(\underline{\mathbf{x}})$, where $\mathcal{E}()$ is the expectancy operator.

The covariance matrix $\mathbf{\Sigma}$ can be decomposed as $\mathbf{\Sigma} = \mathbf{E} \cdot \mathbf{D} \cdot \mathbf{E}^{\mathrm{T}}$. Here, $\mathbf{E} = [\mathbf{e}_1\, \mathbf{e}_2 \ldots \mathbf{e}_\mathbf{p}]$, with $\mathbf{e}_i$ the i^{th} eigenvector, and $\mathbf{D}$ is a diagonal matrix with the eigenvalues on the diagonal. Often the decomposition results are required instead of $\mathbf{\Sigma}$. We denote the eigenvectors and eigenvalues of the decomposition of $\mathbf{\Sigma}$ by population eigenvectors and population eigenvalues respectively.

However, since neither $P(\underline{\mathbf{x}})$ nor $\mathbf{\Sigma}$ are known beforehand, an estimate of $\mathbf{\Sigma}$ has to be obtained from a training set. A commonly used estimator is given by $\hat{\mathbf{\Sigma}} = \frac{1}{N-1}\mathbf{X} \cdot \mathbf{X}^{\mathrm{T}}$, where $\mathbf{X}$ is a matrix in which each column consists of the difference of a training sample and the average of the training samples. N is the number of training samples. The decomposition results of $\hat{\mathbf{\Sigma}}$ are denoted by sample eigenvectors and sample eigenvalues. In a mathematical framework, the column vector consisting of the population eigenvalues is denoted by $\boldsymbol{\lambda}$. The column vector consisting of the sample eigenvalues is denoted by $\boldsymbol{l}$.

A problem of high dimensional training sets is that even though $\hat{\mathbf{\Sigma}}$ is an unbiased estimate of $\mathbf{\Sigma}$, $\boldsymbol{l}$ is a biased estimate of $\boldsymbol{\lambda}$. The bias becomes significant if the number of samples is of the same order as the dimensionality of the data. The bias has a negative effect on systems using these estimates as is shown for classification systems in [3], [4] for example.

To reduce this negative effect, the sample eigenvalues could be corrected to remove the bias. In this paper we present a bootstrap approach [5] to eigenvalue correction: our approach iteratively improves eigenvalue estimates without introducing new measurements. Instead, we generate synthetic data in each iteration which we use to update the population eigenvalue estimates.

The system of eigenvalue estimation and correction is schematically represented in Figure 1. In the figure, F represents the entire procedure of data generation and the sample eigenvalue estimation from this data. If the number of samples and the number of dimensions is large enough, F only introduces a bias on the eigenvalues as will be explained later on. Our aim is to obtain the inverse function F^{-1} that compensates for the bias in the sample eigenvalues. A method which applies $\widehat{F^{-1}}$ to sample eigenvalues is called a correction method. We add a superscript ^{c} to symbols representing results of such a correction method.

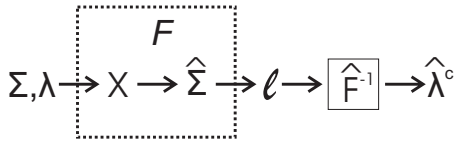


Fig. 1. Schematic representation of the bias introduction and bias correction in eigenvalue estimation.

The currently available correction methods can roughly be divided in three categories: regularisation methods, corrections based on Steins loss criterion and corrections based on theoretical descriptions of the bias. The regularisation methods are often empirical and lack a strong theoretical foundation (See for example [6]). Correction algorithms based on Stein's loss

1550-4786/09 $26.00 © 2009 IEEE
DOI 10.1109/ICDM.2009.111

criterion actually introduce a new bias to reduce the criterion (e.g. [7], [8]).

For the last category bias descriptions are needed. A description of the bias for many data distributions was proved in 1995 [9]. Karoui derived a correction method based on this relation. This method therefore has a strong theoretical basis as opposed to most regularisation algorithms, it reduces the bias as opposed to Stein's loss criterion based methods and it has been shown to reduce bias in real experiments [10]. We therefore consider this method to be one of the state of the art methods so we compare the performance of our bootstrap correction with the performance of this method.

The overview of the remainder of the article is as follows: we start with an analysis of the eigenvalue bias in section II-A. In section II-B we introduce the Marčenko Pastur equation which describes this bias. A brief description of the Karoui correction is given in section II-C. We then describe the bootstrap correction method (section III) and compare correction results of the two methods in section IV for a number of synthetic data sets. Section V gives a summary and conclusions.

II. Eigenvalue bias analysis and correction

A. Eigenvalue bias analysis

To find the statistics of estimators often Large Sample Analysis (LSA) is performed. In LSA it is assumed that the number of samples grows very large so the statistics of the estimator become a function depending solely on the number of samples, N. In this limit case the sample eigenvalues show no bias. However, for example in biometrics, the number of samples is often in the same order as the number of dimensions p or even lower and LSA cannot be used. Instead in the analysis of the statistics of the sample eigenvalues the following limit may be considered: $N, p \to \infty$ while $\frac{p}{N} \to \gamma$, where γ is some positive constant. Analysis in this limit are denoted as General Statistical Analysis (GSA) [11]. In GSA the sample eigenvalues do have a bias.

Figure 2 demonstrates why the limit in GSA is needed. It shows results of three experiments. For each experiment, the population eigenvalues are chosen uniformly between 0 and 1. While γ is kept constant at $\frac{1}{5}$, the number of dimensions is set to 4, 20 and 100 in figures 2a, 2b and 2c respectively. In each figure the population eigenvalue distribution function and the sample eigenvalue distribution functions for 4 repeated experiments are given. Given a set of eigenvalues $\boldsymbol{l}_i, i = 1 \ldots p$, the corresponding distribution function is given by equation 1:

$$F_p(l) \quad = \quad \frac{1}{p} \sum_{i=1}^{p} \mathrm{u}\left(l - \boldsymbol{l}_i\right) \tag{1}$$

here $\mathrm{u}\left(l\right)$ is the step function.

Figure 2 shows that the empirical sample distribution functions converge with increasing dimensionality. The bias in the estimates is visible because they converge to a different distribution function than the population distribution function. For low dimensional examples, the bias is only a small part of the error in the estimates. The major part of the error is caused by random fluctuations of $F_p(l)$.

B. Marčenko Pastur equation

It turns out that under certain conditions a relation between the sample eigenvalues and the population eigenvalues can be given in the GSA limit. The main proofs and conditions on the input data are given in [12] and [9]. Here we briefly repeat the main points.

The relation requires the following Stieltjes transform $v_p\left(z\right)$ of a distribution function based on the empirical sample eigenvalue distribution function:

$$v_p(z) \quad = \quad \left(1-\gamma\right)\frac{-1}{z} + \gamma \cdot \int \frac{\mathrm{d}G_p\left(l\right)}{l-z} \tag{2}$$

here $G_p(l)$ is the empirical sample eigenvalue distribution function as given by equation 1 and $z \in \mathbb{C}^+$. In the GSA limit, if the population eigenvalue distribution function converges to $H_\infty(\lambda)$, then $G_p(l)$ converges weakly to a $G_\infty(l)$ such that the relation between the corresponding $v_\infty(z)$ and $H_\infty(\lambda)$ is given by equation 3.

$$-\frac{1}{v_\infty(z)} \quad = \quad z - \gamma \int \frac{\lambda \cdot \mathrm{d}H_\infty(\lambda)}{1+\lambda\, v_\infty(z)},\ z \in \mathbb{C}^+ \tag{3}$$

Because of this relation, reduction of the bias in the sample eigenvalues should be possible.

C. Karoui correction method

The Karoui method is based on the assumption that the number of samples and the number of dimensions is high enough such that the conditions given in section II-B are met and equation 3 holds. It is also assumed that the density function $h\left(\lambda\right)$ exists. The method approximates $h(\lambda)$ by a weighed sum of fixed density functions $p_i\left(\lambda\right)$ (see equation 4). In our implementation of the method we used a weighed sum of delta pulses and uniform distributions.

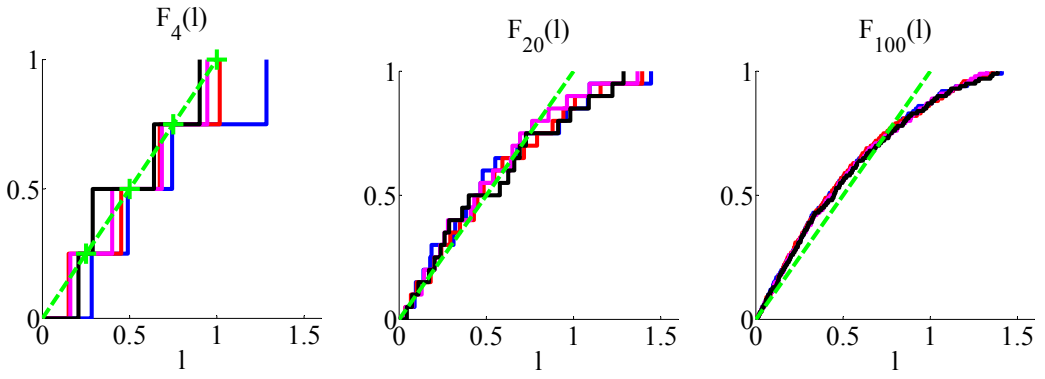
$$\hat{h}(\lambda) = \sum a_i \cdot p_i(\lambda),\ \sum a_i = 1,\ a_i \geq 0 \tag{4}$$

This approximation is then used in equation 3 after substitution of $\mathrm{d}H\left(\lambda\right)$ with $h\left(\lambda\right)\mathrm{d}\lambda$. The empirical sample eigenvalue density function given by equation 1 is substituted in equation 2 to determine a set of corresponding $v_p\left(z\right)$ and z values. The Karoui method then determines the set of a_i values which best satisfies equation 3 for this set of z values. For more details, we refer to [10].

III. Bootstrap eigenvalue correction method derivation

The objective of eigenvalue correction is to find $\boldsymbol{\lambda}$. However, since $\boldsymbol{\lambda}$ is unknown, the objective of the bootstrap correction method we propose is to find a $\hat{\boldsymbol{\lambda}}^c$ such that $\mathrm{F}\left(\hat{\boldsymbol{\lambda}}^c\right) = \mathrm{F}\left(\boldsymbol{\lambda}\right)$.

The general procedure of the method is given schematically in figure 3. To start the algorithm an initial estimate of $\boldsymbol{\lambda}$ is needed. We use $\boldsymbol{l}$ as initial estimate. The method then performs a number of iterations, where in each iteration the steps in



(a) 4 dimensions (b) 20 dimensions (c) 100 dimensions

Fig. 2. Examples of eigenvalue estimation bias toward the GSA limit. The dashed line indicates the population distribution H_p, the four solid lines are the empirical sample distribution G_p.

figure 3 are performed starting from the right. In each iteration a synthetic set of white Gaussian distributed data samples $X_{w,n}$ is generated with the same number of samples and the same dimensionality as the original measurement. These samples are scaled so their population eigenvalues are equal to the current estimate of the population eigenvalues $\hat{\boldsymbol{\lambda}}^c_{:,n}$, where n is the current iteration index and subscript $_:$ indicates the entire column vector. From this synthetic data set the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_n$ is estimated. From this matrix the sample eigenvalues $\hat{\boldsymbol{l}}_{:,n}$ are determined. If the cost function $K = \left|\boldsymbol{l} - \hat{\boldsymbol{l}}_{:,n}\right|^2$ is below a threshold, the sample eigenvalues of the real data and the synthetic data are considered to be equal and therefore the current estimate of the population eigenvalues are used as final estimates of the population eigenvalues. If the cost function is not below the threshold, $\hat{\boldsymbol{\lambda}}^c_{:,n}$ is updated via update rule:

$$\hat{\boldsymbol{\lambda}}^c_{:,n+1} \quad = \quad \hat{\boldsymbol{\lambda}}^c_{:,n} - \mu \cdot \frac{\partial K}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}} \tag{5}$$

The parameter μ determines the size of the adjustment steps taken in the method. After this update a new iteration starts and the previously described steps are repeated.

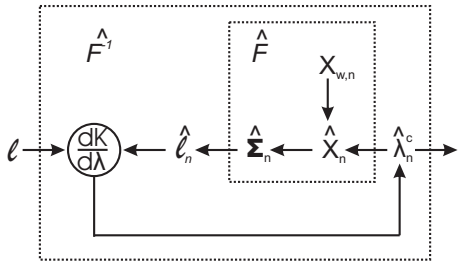


Fig. 3. Schematic representation of the bootstrap eigenvalue correction

To find the derivative in the right hand side of equation 5 we can relate this derivative with the derivative $\frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}}$ via

$$\frac{\partial K}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}} \quad \propto \quad -\left(\boldsymbol{l} - \hat{\boldsymbol{l}}_{:,n}\right)^{\mathrm{T}} \cdot \frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}} \tag{6}$$

The derivative $\frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}}$ can be found using the expression for $\hat{\boldsymbol{l}}_{:,n}$ in equation 7, where M is the number of synthetic data samples. Note that we use normal distributed synthetic samples although we do not know the real distribution of the data samples. However, according to the Marčenko Pastur equation, in the GSA limit many distributions, including the normal distribution, will have the same relation between sample eigenvalues and population eigenvalues.

$$\begin{aligned}\hat{\boldsymbol{l}}_n \quad = \quad & \mathrm{diag}\left\{\hat{\mathbf{E}}_n^{\mathrm{T}} \cdot \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}} \cdot \frac{1}{M-1} \cdot \mathbf{X}_{w,n} \cdot \right. \\ & \left. \mathbf{X}_{w,n}^{\mathrm{T}} \cdot \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}} \cdot \hat{\mathbf{E}}_n \right\}\end{aligned} \tag{7}$$

Here $\hat{\mathbf{E}}_n$ are the eigenvectors of the synthetic data covariance matrix and $\hat{\mathbf{D}}_n^c$ is a diagonal matrix with $\hat{\boldsymbol{\lambda}}^c_{:,n}$ on the diagonal. From equation 7 we find the derivative $\frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}}$, given element wise in equation 8.

$$\begin{aligned}\frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{j,n}} \quad = \quad & \mathrm{diag}\left\{2 \cdot \hat{\mathbf{E}}_n^{\mathrm{T}} \cdot \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}} \cdot \frac{\mathbf{X}_{w,n} \cdot \mathbf{X}_{w,n}^{\mathrm{T}}}{M-1} \cdot \right. \\ & \left. \frac{\partial}{\partial \hat{\boldsymbol{\lambda}}^c_{j,n}} \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}} \cdot \hat{\mathbf{E}}_n \right\}\end{aligned} \tag{8}$$

$\frac{\partial}{\partial \hat{\boldsymbol{\lambda}}^c_{j,n}} \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}}$ is a matrix with zeros except for element $\{j,j\}$, which is $\left(\hat{\boldsymbol{\lambda}}^c_{j,n}\right)^{-\frac{1}{2}}$. Because of this last matrix, the method does not converge if one of the population eigenvalues becomes zero. We solved this problem by using $\boldsymbol{l}$ as initial estimate of $\boldsymbol{\lambda}$, but replacing the zero eigenvalues with the lowest non zero eigenvalue. In our experiments no eigenvalues turned to zero during the iterations. Another solution is to use

$\hat{\boldsymbol{\lambda}}^c_1$	$\hat{\boldsymbol{\lambda}}^c_2$	$\hat{\boldsymbol{\lambda}}^c_3$	$\hat{\boldsymbol{\lambda}}^c_4$	$\hat{\boldsymbol{\lambda}}^c_5$	$\hat{\boldsymbol{\lambda}}^c_6$
$\hat{\boldsymbol{\lambda}}^c_{m,1,3}$	$\hat{\boldsymbol{\lambda}}^c_{d,1,3}$	$\hat{\boldsymbol{\lambda}}^c_{m,2,3}$	$\hat{\boldsymbol{\lambda}}^c_{d,2,3}$	$\hat{\boldsymbol{\lambda}}^c_{m,3,3}$	$\hat{\boldsymbol{\lambda}}^c_{d,3,3}$
	$\hat{\boldsymbol{\lambda}}^c_{m,1,2}$	$\hat{\boldsymbol{\lambda}}^c_{d,1,2}$		$\hat{\boldsymbol{\lambda}}^c_{m,2,2}$	
		$\hat{\boldsymbol{\lambda}}^c_{m,1,1}$	$\hat{\boldsymbol{\lambda}}^c_{d,1,1}$		

TABLE I
ISOTONIC TREE WITH 6 POPULATION EIGENVALUES.

the gradient with respect to standard deviations. However, the first option proved to give more accurate reconstructions and faster convergence.

A. Order preservation in the bootstrap correction method

When updating the eigenvalues without any constraints, oscillations may occur in which $\hat{\boldsymbol{\lambda}}^c_{i,n}$ switch value with $\hat{\boldsymbol{\lambda}}^c_{i+1,n}$ in each iteration. A order preserving algorithm was presented by Stein in [13]. We used an isotonic tree method, which allows blocks of eigenvalues to have a large change in value simultaneously and the method can also be implemented more efficiently.

The isotonic method first builds a tree: the top layer consists of the current population eigenvalues. Each lower layer is based on the difference and the mean of mean elements of the previous layer above, or $\hat{\boldsymbol{\lambda}}^c_{d,k,l} = \left(\hat{\boldsymbol{\lambda}}^c_{m,2k,l+1} - \hat{\boldsymbol{\lambda}}^c_{m,2k-1,l+1}\right)/2$ and $\hat{\boldsymbol{\lambda}}^c_{m,k,l} = \left(\hat{\boldsymbol{\lambda}}^c_{m,2k-1,l+1} + \hat{\boldsymbol{\lambda}}^c_{m,2k,l+1}\right)/2$ respectively. Here the first subscript of the left hand side symbol indicates whether the element is a mean (m) or a difference (d) element, the second subscript is the index of the element in the layer, and the third index is the layer index. So $\hat{\boldsymbol{\lambda}}^c_{d,k,l}$ is the k^{th} difference term on the l^{th} layer and $\hat{\boldsymbol{\lambda}}^c_{m,k,l}$ is the k^{th} mean term on the l^{th} layer. See the example in table I. To reconstruct the $\hat{\boldsymbol{\lambda}}^c_n$ only the difference terms in all the layers and the final mean term are needed.

The derivative in the update rule can be split into a tree a similar manner. The update is then performed per layer starting from the bottom layer. First the average term is updated via $\hat{\boldsymbol{\lambda}}^c_{m,1,1,2} = \hat{\boldsymbol{\lambda}}^c_{m,1,1} + \mu \mathrm{d}\hat{\boldsymbol{\lambda}}^c_{m,1,1}$. Then the following two steps are repeated for every layer: first, the difference terms are updated via $\hat{\boldsymbol{\lambda}}^c_{d,k,l,2} = \hat{\boldsymbol{\lambda}}^c_{d,k,l} + \mu \mathrm{d}\hat{\boldsymbol{\lambda}}^c_{d,k,l}$. Next the mean terms on layer $l-1$ are updated via $\hat{\boldsymbol{\lambda}}^c_{m,k,l-1,2} = \hat{\boldsymbol{\lambda}}^c_{m,k/2,l,2} \pm \hat{\boldsymbol{\lambda}}^c_{d,k/2,l,2}$, where the decision of doing an addition or a subtraction is based on whether k is odd or even.

For k even, it may happen that $\hat{\boldsymbol{\lambda}}^c_{m,k+1,l-1,2} > \hat{\boldsymbol{\lambda}}^c_{m,k,l-1,2}$. In that case the mean terms k and $k+1$ are updated via $\hat{\boldsymbol{\lambda}}^c_{m,k+1,l-1,2} = \hat{\boldsymbol{\lambda}}^c_{m,k,l-1,2} = \hat{\boldsymbol{\lambda}}^c_{m,k/2,l,2} + \hat{\boldsymbol{\lambda}}^c_{d,k/2,l,2} \cdot \frac{\hat{\boldsymbol{\lambda}}^c_{d,k/2,l,2} + \hat{\boldsymbol{\lambda}}^c_{d,1+k/2,l,2}}{\hat{\boldsymbol{\lambda}}^c_{m,1+k/2,l,2} - \hat{\boldsymbol{\lambda}}^c_{m,k/2,l,2}}$. The update of the population eigenvalues is completed by taking $\hat{\lambda}^c_{n+1,k} = \hat{\lambda}^c_{m,k,l_{max},2}$. In case of an odd number of mean elements on a layer, the last mean element is just copied from the lower level.

IV. EXPERIMENTS AND DISCUSSION

To evaluate the performance of the bootstrap correction we generated a number of synthetic data sets with a number of chosen population eigenvalues set beforehand. We corrected the sample eigenvalues measured from these sets with the bootstrap method and the Karoui method and measured the Levy distance,

$$d_L(F,G) = \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon, \forall x\} \quad (9)$$

between the corrections and the population eigenvalues. Like Karoui, we calculate a score for a correction result by determining the ratio $d_L(G,H)/d_L(\hat{H}^c,H)$, where $\hat{H}^c$ is the corrected population eigenvalue distribution.

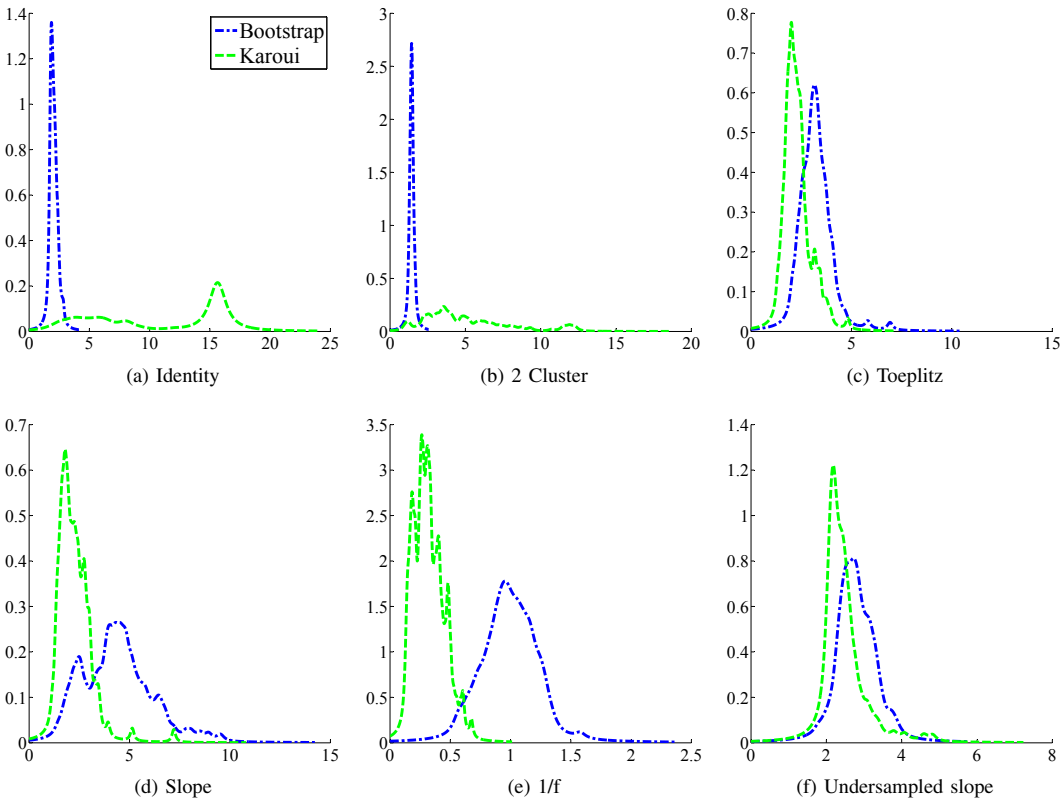
We performed the comparison with 6 different set-ups. For most experiments we drew 501 samples from an 100 dimensional normal distribution. The first three experiments have the same set-up as the experiments in [10]: experiment 1 (identity) has $\lambda_k = 1, \forall k$, experiment 2 (two cluster) has $\lambda_k = 1 | k = 1 \ldots 50$ and $\lambda_k = 2 | k = 51 \ldots 100$ and experiment 3 (Toeplitz) has the eigenvalues of a Toeplitz matrix. In the fourth experiment (Slope) $\lambda_k = 1 + k/100$. In the fifth experiment (one over n) we used $\lambda_k = \frac{1}{k}$, a common model for eigenvalues of facial data [6]. In the sixth experiment (undersampled slope) we draw 201 samples for a 600 dimensional normal distribution with $\lambda_k = 1 + k/600$.

By repeating the experiments collections of scores for the different set-ups are obtained. These score collections are represented by the histograms in figure 4. A large score indicates that the corrected population eigenvalue distribution and the population eigenvalue distribution are much more alike than the sample eigenvalue distribution and the population eigenvalue distribution. Therefore which ever method has the most density to the right of the histogram provides the best correction according to the Levy distance measure.

In the identity and two cluster experiments (figures 4a and 4b) the bootstrap method shows a steady improvement with an average score around 2. The Karoui correction achieves much higher scores, albeit with a huge spread. In the Toeplitz experiment (figure 4c) and the slope experiment (figure 4d), the bootstrap method has significantly higher scores. In the one over n case (figure 4e), the bootstrap method also performs better, however, its scores indicate its estimate has almost the same precision as the sample eigenvalues.

In the undersampled slope case (figure 4f) the Bootstrap method still outperforms the Karoui method, however the difference is not as big as in the slope experiment in figure 4d. For a more detailed discussion on the results of this experiment an example repetition is shown in figure 5, with figure 5a containing the population eigenvalue scree plots and figure 5b contains the sample eigenvalue scree plots.

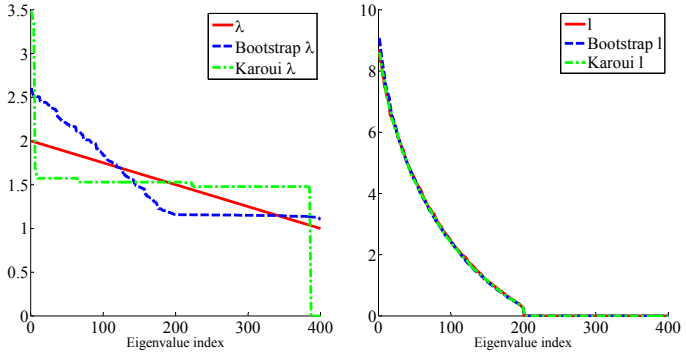
There are large differences between the population eigenvalues. The Karoui correction models the eigenvalues similar to the spiked population model: almost all eigenvalues are equal except for a few considerably larger ones (see [14] and [15]). It also sets a few population eigenvalues to zero. The bootstrap method sets all the population eigenvalues which were estimated as zero by the sample eigenvalue estimator to



(a) Identity (b) 2 Cluster (c) Toeplitz

(d) Slope (e) 1/f (f) Undersampled slope

Fig. 4. Histogram of scores of both the bootstrap eigenvalue correction and the Karoui eigenvalue correction.



(a) Population eigenvalues (b) 2 Cluster

Fig. 5. Example repetition with the undersampled slope population eigenvalues.

an equal, non zero value. For the remainder of the eigenvalues a slope is estimated, but with a higher gradient.

Despite the differences in the population eigenvalues, the sample eigenvalue scree plots are almost the same. This shows that the undersampled case is a very difficult problem, which seems to be under determined: multiple population eigenvalue distributions seem to generate the same sample eigenvalue distributions. In practice an equal value for all population eigenvalues in the zero space of the sample eigenvalues is required, since the order of the estimated eigenvectors in the zero space is random. Combining the estimated eigenvectors with different values will introduce arbitrariness. This will lead to arbitrary likelihood estimates for new samples with a component in the sample null space. The bootstrap method corrects all zero sample eigenvalues to almost the same value and seems therefore more applicable to undersampled problems.

We used the Levy distance to be comparable to the tests in [10], but the use of the distance is somewhat arbitrary and causes problems in performance comparisons. The distance

is not scale invariant as is demonstrated by the example in figure 6. Another problem is that the use of a different criterion may very well result in a different ordering of the methods.

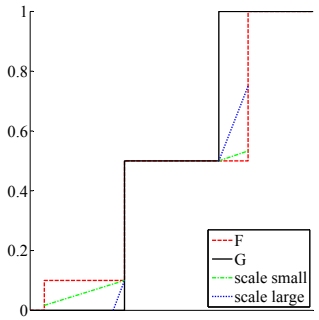


Fig. 6. Scaling behaviour Levy distance. The dash-dotted line and the dotted line indicate the $F(x-\epsilon)\pm\epsilon$ line in case of small scale and in case of large scale respectively. In the small scale, the infimum is determined by the difference on the left, while in the large scale, the infimum is determined by the difference on the right.

It appears that both methods are still biased: the Karoui method seems to be biased toward clusters, what could be explained by the approximation of the population density by the weighed sum of fixed distribution functions. The bootstrap method still over estimates the largest eigenvalues and under estimates the smallest eigenvalues, probably due to inaccuracies in the gradient estimate.

V. Conclusions and recommendations

We introduced the bootstrap correction method, which corrects the bias in sample eigenvalues, and we did a performance comparison between the bootstrap method and the Karoui method. Both methods do improve the estimate of the eigenvalues for most cases, except for the undersampled case. The Karoui correction performs better on clustered densities while the bootstrap method performs better on smooth distributions of the population eigenvalues. Both methods still leave a bias in the eigenvalues after correction: the Karoui correction introduces more clusters, while the bootstrap method still estimates the largest eigenvalues too large, while the smallest eigenvalues are estimated too small.

The comparison of the methods using the Levy distances is still somewhat arbitrarily: the order may change when the scale of the distributions is changed and the ordering by the Levy distance may very well differ from the order found using other measures. In the undersampled case the use of the Levy distance becomes even more problematic: different population eigenvalue distributions seem to lead to the same sample distributions, while the Levy distance between the population distributions can be considerable.

In the undersampled case there is a clear advantage for the bootstrap method: the smallest population eigenvalues which were estimated to be zero with the sample estimator are assigned an equal value. Because the original sample eigenvalues were all zero, the corresponding sample eigenvectors are a random basis in the zero space. Replacing the zero valued sample eigenvalues by non zero corrected eigenvalues introduces some arbitrariness, unless the non zero corrected eigenvalues all have the same value.

In the tests we performed we did not focus on the convergence rate of the two methods with respect to the GSA limit. As we described previously, the Karoui correction relies heavily on the Marčenko Pastur relation, while the bootstrap method only uses this relation to justify the use of Gaussian distributed synthetic data samples. The Karoui correction may therefore rely more on the GSA limit and have reduced performance for lower dimensional problems.

To reduce random fluctuations in the gradient estimate, the sample covariance matrix of white samples in equation 8 can be constructed using the Marčenko Pastur rule [16].

References

[1] T. W. Anderson, *An introduction to multivariate statistical analysis*, 2nd ed., ser. Wiley series in probability and mathematical statistics. John Wiley & Sons, 1984.

[2] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[3] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "Eigenvalue correction results in face recognition," in *29th Symp. on Information Theory in the Benelux*, 2008, pp. 27–35.

[4] P. Xu, G. N. Brock, and R. S. Parrish, "Modified linear discriminant analysis approaches for classification of high-dimensional microarray data," *Computational Statistics & Data Analysis*, vol. 53, no. 5, pp. 1674 – 1687, 2009.

[5] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications*, ser. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, January 1997.

[6] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 383–394, 2008.

[7] D. K. Dey and C. Srinivasan, "Estimation of a covariance matrix under stein's loss," *The Annals of Statistics*, vol. 13, no. 4, pp. 1581–1591, 1985.

[8] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *J. of Empirical Finance*, vol. 10, no. 5, pp. 603–621, 2003.

[9] J. W. Silverstein, "Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices," *J. Multivar. Anal.*, vol. 55, no. 2, pp. 331–339, 1995.

[10] N. El Karoui, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *ArXiv Mathematics e-prints*, september 2006.

[11] V. Girko, *Theory of Random Determinants*. Kluwer, 1990.

[12] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR - Sbornik*, vol. 1, no. 4, pp. 457–483, 1967.

[13] C. Stein, "Lectures on the theory of estimation of many parameters," *J. of Mathematical Sciences*, vol. 34, no. 1, pp. 1371–1403, 1986.

[14] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, 2006.

[15] D. Paul, "Asymptotics of the leading sample eigenvalues for a spiked covariance model," Dep. of Statistics, Stanford University, Tech. Rep., 2004.

[16] I. M. Johnstone, "On the distribution of the largest principle component," Dep. of Statistics, Stanford University, Tech. Rep., 2000.