2010 International Conference on Pattern Recognition

Verification under increasing dimensionality

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

***Abstract*—Verification decisions are often based on second order statistics estimated from a set of samples. Ongoing growth of computational resources allows for considering more and more features, increasing the dimensionality of the samples. If the dimensionality is of the same order as the number of samples used in the estimation or even higher, then the accuracy of the estimate decreases significantly. In particular, the eigenvalues of the covariance matrix are estimated with a bias and the estimate of the eigenvectors differ considerably from the real eigenvectors. We show how a classical approach of verification in high dimensions is severely affected by these problems, and we show how bias correction methods can reduce these problems.**

***Keywords*-General Statistical Analysis; high dimensional verification; bias correction;**

I. Introduction

In verification the main objective is to judge a claim that a sample $\bar{x}$ comes from class c (the $^-$ indicates that $\bar{x}$ is a column vector). A common strategy is to accept this claim if the likelihood ratio

$$L\left(c|\bar{x}\right) = \frac{P\left(X = \bar{x}|C = c\right)}{P\left(X = \bar{x}|C \neq c\right)} \tag{1}$$

is above a threshold, where $P\left(X = \bar{x}|C = c\right)$ is the probability of measuring $\bar{x}$ if the class is c. $P\left(X = \bar{x}|C \neq c\right)$ is the probability of measuring $\bar{x}$ if the class is not c, but if the number of classes is high, it can be approximated by $P\left(X = \bar{x}\right)$.

These distributions and their parameters are commonly unknown beforehand. The distributions are often modeled by normal distributions. This distribution model is completely determined by the mean and the second order statistics, which is why second order statistics are very important in verification. There is however a considerable problem with how these statistics are determined.

Second order statistics are described by a covariance matrix $\mathbf{\Sigma}$, which is given by $\mathcal{E}\left\{\left(\bar{x} - \mathcal{E}\left\{\bar{x}\right\}\right)^{\mathrm{T}} \cdot \left(\bar{x} - \mathcal{E}\left\{\bar{x}\right\}\right)\right\}$, where $\mathcal{E}\left\{\right\}$ is the expectation operator. We denote $\mathbf{\Sigma}$ as the population covariance matrix. $\mathbf{\Sigma}$ can be decomposed into $\boldsymbol{E} \cdot \boldsymbol{D} \cdot \boldsymbol{E}^{\mathrm{T}}$, where $\boldsymbol{E}$ is a rotation matrix with each column being an eigenvector of $\mathbf{\Sigma}$ and $\boldsymbol{D}$ is a diagonal matrix, containing the eigenvalues $\bar{\lambda}$ of $\mathbf{\Sigma}$ on the diagonal. These decomposition results can be used to find $\mathbf{\Sigma}^{-1}$, which is needed to evaluate the likelihoods in equation 1.

In practice neither $\mathbf{\Sigma}$ nor its decomposition results are known and the second order statistics have to be estimated from a set of examples, denoted as the training set. Let $\boldsymbol{X}$ be a matrix where each column is a sample from the training set with the mean of that set subtracted. The sample covariance matrix $\hat{\mathbf{\Sigma}}$, which can be used as an estimate of $\mathbf{\Sigma}$, is given by $\frac{1}{N-1}\sum_{k=1}^{N} \boldsymbol{X}^{\mathrm{T}} \cdot \boldsymbol{X}$, where N is the number of samples in the training set. From its decomposition into $\hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{D}} \cdot \hat{\boldsymbol{E}}^{\mathrm{T}}$, we get the sample eigenvalues $\bar{l}$ and sample eigenvectors.

Preferably estimators would give estimates close to the true value, but if the number of samples in the training set is in the same order as the dimensionality of the samples (p), the sample eigenvectors deviate considerably from the population eigenvectors and the sample eigenvalues are significantly biased from the population eigenvalues, so the distribution estimated from the training set is not an optimal estimate for classification.

In the following sections we will present methods to improve the estimate of the distribution as schematically shown in Figure 1. First we discus in section II-A the statistical framework which we use throughout the rest of the paper. Based on these analysis, the bias in the sample eigenvalues can be described as a function B which takes the population eigenvalues as input and gives the biased sample eigenvalues as result.

Bias correction can then be thought of as applying the inverse of this function to sample eigenvalues which then result in corrected eigenvalues $\bar{\lambda}^{c}$, as shown in Figure 1. In section II-B we present several methods to reduce the bias:

- a method by Karoui, based on the Marčenko Pastur equation, which describes the relation between the population eigenvalues and the sample eigenvalues.
- a bootstrap method, which corrects the eigenvalues in an iterative process, using a bootstrap approach.
- a method by Ledoit and Wolf, based on regularisation.

After the bias correction a variance correction is applied to the eigenvalues to partially compensate the errors in the sample eigenvectors as estimates of the population eigenvectors, as will be explained in Section II-C. This correction is represented by the step leading to $\bar{v}$ in Figure 1.

Our goal is to evaluate the improvement of verification scores with high dimensional data if these corrections are

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.149

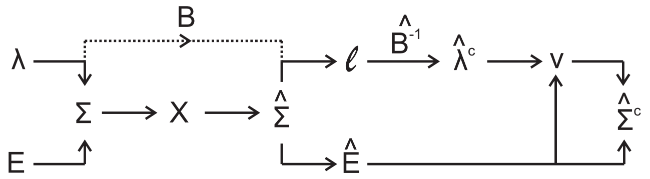


Figure 1. Schematic overview of the estimation and corrections of the second order statistics. Function B represents the introduction of the bias in the sample eigenvalues. Bias correction can be thought of as finding an estimate of the inverse of this function. But since the sample eigenvalues are inaccurate estimates of the population eigenvalues, a second correction of the eigenvalues is needed.

used instead of the classical PCA dimensionality reduction. We therefore present a verification experiment in section III, which is based on the experimentation method of Karoui: we used synthetic data to clearly demonstrate both the effect of the bias and the improvement bias correction gives if the bias is the sole problem. But instead of keeping the ratio between the number of samples and their dimensionality fixed as Karoui did, we keep the number of samples fixed and vary the number of dimensions so we start with much more samples than dimensions and end with much more dimensions than samples. From the results of this experiment we derive conclusions in section IV.

II. Analysis of eigenvalue estimators

We want estimators to be as accurate as possible. One parameter which gives an indication of the accuracy is the bias of an estimator [1]. Due to random fluctuations in the input samples, the estimate of a parameter determined by an estimator will differ. For some estimators, the expected value of the estimator differs from the value of the parameter to be determined. This difference is denoted as the bias of the estimator.

Usually the accuracy of the sample eigenvalues is determined by performing large sample analysis (LSA) under the assumption that fluctuations in the sample eigenvalues are solely dependent on the number of samples used in the estimation. In these analysis the sample eigenvalues seem unbiased. However, if the number of samples used is in the same order as the dimensionality of the problem, the analysis are no longer accurate and different analysis should be done.

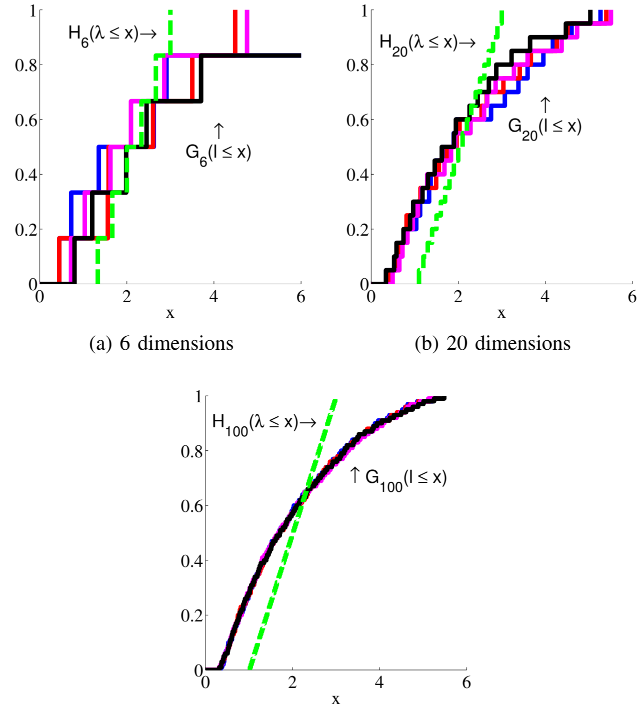
A. General statistical analysis

General Statistical Analysis (GSA, [2]) may be more applicable in high dimensional data. In GSA the following limit is considered: $N, p \to \infty$ under the condition that $\frac{p}{N} \to \gamma$, where γ is some positive constant. In GSA, Marčenko and Pastur proved a relation between the empirical distribution function belonging to the set of population eigenvalues and the set of sample eigenvalues ([3]) which Silverstein showed to hold for a large set of data distributions ([4]). An empirical distribution function $F_p\left(x\right)$ of a set of eigenvalues $x_k, k = 1 \ldots p$ is given by:

$$F_p\left(x\right) = \frac{1}{p}\sum_{k=1}^{p} \mathrm{u}\left(x - x_k\right) \qquad (2)$$

where $u(x)$ is the Heavyside step function.

As an example of these analysis we did an eigenvalue estimation experiment with synthetic data. The population eigenvalues were chosen uniform between 1 and 3. We varied the dimensionality p between 6, 20 and 100, while keeping γ fixed on $\frac{1}{5}$. We estimated the sample eigenvalues and show their empirical distribution function $G_p\left(l \leq x\right)$ in figures 2a, 2b and 2c, with 4 repetitions of the estimation (the four solid lines). It is clear that with increasing p, the 4 estimates converge, but to a distribution differing from the population distribution $H_p\left(\lambda \leq x\right)$ (the dashed line).



(a) 6 dimensions

(b) 20 dimensions

(c) 100 dimensions

Figure 2. Example of convergence of the empirical sample eigenvalue distribution to a fixed distribution function in the GSA limit.

B. Bias correction

Since a bias in an estimator is a non random property, it may be possible to remove it from the estimate. Several methods have been introduced to correct the eigenvalues for this bias. We use 3 different bias correction methods: a method proposed by Karoui ([5]), a method based on bootstrapping ([6]) and a method proposed by Ledoit and Wolf ([7]).

The Karoui method is based on the Marčenko Pastur equation ([5]). In [6] we argued that the Karoui method is the current state of the art of the correction methods available, but we presented a correction method based on bootstrapping competitive in performance with the Karoui method. It particularly out performed Karoui if the number of samples was smaller than the number of dimensions. This is of interest since we study the classification performance under a growing number of dimensions.

A large group of eigenvalue correction methods are not designed to reduce the bias in the eigenvalues, but reduce a criterion known as Stein's loss by shrinking the sample eigenvalues toward the mean of the sample eigenvalues, commonly known as regularisation. However, many of those methods require that the eigenvalues are distinct or non zero or both (see [8], [9] or [10] for example), which is not the case if $p > N$. We used the method proposed by Ledoit and Wolf (LW, [7]), which does not have these requirements.

We compare the eigenvalue corrections with a classical approach in biometrics, where zero valued sample eigenvalues are removed by first doing a dimension reduction to a fixed number of dimensions by applying PCA ([11]).

C. Variance estimation improvement

However, the bias in the eigenvalues is not the only error in the estimate of the distributions. As indicated before, the sample eigenvectors will also differ from the population eigenvectors. To our knowledge, no method exists that improves the density estimate based on a description relating the population eigenvectors and the sample eigenvalues similar to how the Marčenko Pastur equation relates the population eigenvalues and the sample eigenvalues, although [12] is an example of advancements in that direction.

In [13] we suggested an empirical method for finding such a relation and we will use this to adjust the sample eigenvalues again. The general idea is to adjust the sample eigenvalues to the real variances along the sample eigenvectors instead of the population eigenvalues. This is because the population eigenvalues give the variances along the population eigenvectors, not the variances along the sample eigenvectors. Note that the LW correction requires no variance correction, since it is not based on eigenvalue bias correction.

III. Experiments

After both the bias correction and the variance correction of the sample eigenvalues, recombination of these corrected values and the sample eigenvectors leads to a new density estimate. We studied the effect of this improved density estimate in a verification scheme.

In our experiment we use synthetic data, which is generated according to the model used in Linear Discriminant Analysis (LDA, [11]): each sample of class c is generated by drawing samples from a distribution $N\left(\bar{\mu}_c, \mathbf{\Sigma}_w\right)$, a normal distribution with mean $\bar{\mu}_c$ and covariance $\mathbf{\Sigma}_w$. The mean of each class is generated by drawing from a distribution $N\left(0, \mathbf{\Sigma}_b\right)$. We chose the matrices such that $\mathbf{\Sigma}_t = \mathbf{\Sigma}_w + \mathbf{\Sigma}_b$ is the identity matrix. This means that the only other parameters of significance are the eigenvalues of $\mathbf{\Sigma}_b$, for which we use two configurations: in one configuration we set the between eigenvalues uniformly between 0.02 and 0.2. In the other configuration we distribute them exponentially between 0.02 and 0.2.

To judge the claim that the class is c given measurement $\bar{x}$, we need to estimate both $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_t$ from the training data. In our experiment we generate a training set containing samples from 100 classes where each class has 5 samples. We estimate $\mathbf{\Sigma}_t$ by the sample covariance matrix of the total training set. To find an estimate of $\mathbf{\Sigma}_w$, we first estimate the sample covariance matrix for each set of samples belonging to one class and then take the average of these sample covariance matrices. Both matrices are corrected as described in the previous sections.

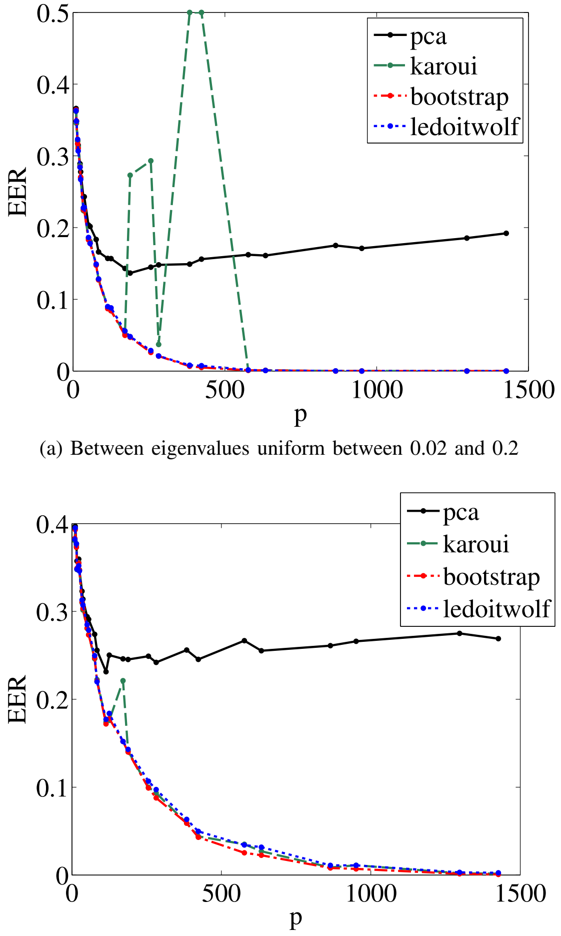
To measure the verification performance of the corrected density estimates, we generated a test set of 100 classes with 20 samples per class. In our verification scheme, the claim that sample $\bar{x}$ belongs to class c is accepted if the likelihood from equation 1 is above a preset threshold. By varying this threshold, the probability of rejecting a true claim (False Rejection Rate, FRR) can be either decreased or increased while the probability of accepting a false claim (False Accept Rate, FAR) is either increased or decreased. By choosing the right threshold, both error rates can be made equal, resulting in an Equal Error Rate (EER).

We let the dimensionality grow from 10 dimensions (much smaller than the $N = 500$ training samples) to 1300 (considerably larger than N). The results are shown in Figure 3.

The PCA correction reduction projects the data along the first 150 sample eigenvectors of the training set (no reduction occurs if $p < 150$). In the results it shows that although this approach prevents the singularity problems, so a likelihood estimate can be obtained at all, the final EER rates start even to increase slightly if p gets larger than 150.

With both bias correction methods and LW correction the EER decreases with increasing p, even after p becomes larger than N. The similarity of the curves of the bias corrections may be caused by using the same variance correction method for both bias corrections.

In the results of the Karoui correction several outliers occur. This can be explained as follows. The correction method actually tries to estimate the population eigenvalue distribution function instead of the set of population eigenvalues themselves. Since the bias reduces the smallest eigenvalues, all the population eigenvalues should be larger than the smallest sample eigenvalue, so Karoui allows no density for λ smaller than the smallest sample eigenvalue. But as the dimensionality of the training samples is increased to the



(a) Between eigenvalues uniform between 0.02 and 0.2

(b) Between eigenvalues exponential between 0.02 and 0.2

Figure 3. Equal Error Rates (EER) under increasing dimensionality.

same order as the number of samples, the smallest sample eigenvalue decreases until it becomes zero when the number of training samples is smaller than their dimensionality.

So when the number of dimensions approaches the number of samples, Karoui sometimes estimates population density for λ close or equal to zero. This results in small to zero valued population eigenvalues, which will than solely determine the within class probability estimate. The EER will therefore be based on a small subspace of the sample spaces, causing a huge increase in EER.

IV. Conclusions and discussion

We investigated the effect of increasing the number of dimensions p in a verification setting, which is similar to adding more features. If p increases to the same order as the number of training samples, the bias in the sample eigenvalues and the difference between population eigenvectors and sample eigenvectors leads to serious errors in verification systems.

The classical approach of performing PCA dimension reduction leads to an undesirable solution where EER increases if the number of features is increased. Bias correction combined with variance correction in general lead to an improved estimate of the involved distributions, but Karoui every now and then fails on the correction of the smallest eigenvalues, resulting in huge mistakes in likelihood estimates. On the other hand the Bootstrap approach requires significant computational resources when the involved data sets become large. The LW correction gives slightly worse EER rates, but is still a substantial improvement compared to the PCA solution.

References

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

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

[3] 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.

[4] 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.

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

[6] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, “A bootstrap approach to eigenvalue correction,” in *ICDM ’09*. IEEE Computer Society Press, December 2009, pp. 818–823.

[7] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[8] S. P. Lin and M. D. Perlman, “A monte carlo comparison of four estimators of a covariance matrix,” in *Multivariate Analysis - VI*, P. Krishnaiah, Ed. Elsevier Science Publishers B.V., 1985, pp. 411–429.

[9] 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.

[10] T. Takeshita and J. ichiro Toriwaki, “Experimental study of performance of pattern classifiers and the size of design samples,” *Patt. Recog. Lett.*, vol. 16, no. 3, pp. 307–312, 1995.

[11] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. on Pat. Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[12] O. Ledoit and S. Péché, “Eigenvectors of some large sample covariance matrices ensembles,” Institute for Empirical Research in Economics, Tech. Rep. iewwp407, Mar. 2009.

[13] A. Hendrikse, R. Veldhuis, and L. Spreeuwers, “Improved variance estimation along sample eigenvectors,” in *Proceedings of the 30th Symposium on Information Theory in the Benelux*, 2009, pp. 25–32.