Likelihood ratio based verification in high dimensional spaces

Anne Hendrikse, Raymond Veldhuis, *Senior Member, IEEE,* and Luuk Spreeuwers

Abstract—The increase of the dimensionality of data sets often lead to problems during estimation, which are denoted as the curse of dimensionality. One of the problems of Second Order Statistics (SOS) estimation in high dimensional data is that the resulting covariance matrices are not full rank, so their inversion, needed for example in verification systems based on the likelihood ratio, is an ill posed problem, known as the singularity problem. A classical solution to this problem is the projection of the data onto a lower dimensional subspace using Principle Component Analysis (PCA) and it is assumed that any further estimation on this dimension reduced data is free from the effects of the high dimensionality.
Using theory on SOS estimation in high dimensional spaces, we show that the solution with PCA is far from optimal in verification systems if the high dimensionality is the sole source of error. For moderate dimensionality it is already outperformed by solutions based on euclidean distances and it breaks down completely if the dimensionality becomes very high. We propose a new method, the fixed point eigenwise correction, which does not have these disadvantages and performs close to optimal.

Index Terms—High dimensional verification, eigenvalue bias correction, variance correction, euclidean distance, Principle Component Analysis, Marčenko Pastur equation, Eigenwise correction, Fixed point eigenvalue correction

1 Introduction

Nowadays during data acquisition more and more variants are measured, resulting in a higher dimensionality of the data, while the number of observations is not increased proportionally. For example, in biometrics based on facial images, the resolution of the images has increased considerably in the last decades, while the number of test subjects has not increased as much. There is also a trend in combining several face representation [1] or even different modalities, such as face images with finger prints. But the large number of variants compared to the number of training samples is not limited to biometrics, it occurs for example in portfolio selection [2], radio signal separation [3] and in gene selection [4] as well. It may seem that the added dimensions only add information, so any system should perform at least as good with the added dimensions as without them. In [5], [6] and [7] it is shown that this is indeed the case if the data structure is known.

However, this theoretical proof seems very often to be contradicted in practice as performance does decrease with increasing dimensionality (known as the curse of dimensionality [8]). In [9] it is argued that the proof is based on knowing the data structure of the data generating process, while in practice only a training set is available to inver this knowledge from. Several factors are known to corrupt this inference: incorrect sampling [9], errors in the measurements [10], [11], modeling errors [12] and errors in the currently used estimators [13], [14].

In this study we focus on the errors in the classically used estimators and show how these errors deteriorate the performance of verification systems based on these estimators and lead to the observed contradiction between the theoretical benefit of adding variants in verification problems and the practically observed curse of dimensionality if this error is the dominant source of error. However, we will also improve these estimators such that added variants will indeed improve the verification performance if they are of similar quality as the already present variants.

It has already been shown that several estimators do not perform well in high dimensional problems. In [15] it is shown that Independent Component Analysis, which uses higher order statistics estimates, is severely affected by high dimensionality of the training data. Second Order Statistics (SOS) estimators can be severely affected by the high dimensionality as well. One effect is that the estimated eigenvalues are biased [13], [16], which results in the covariance matrix becoming singular if the dimensionality p becomes larger than the number of samples (N) used in the estimation. Inversion of these estimated matrices, required e.g. in likelihood estimation, cannot be done.

A classical method to prevent the covariance matrix from becoming singular is to reduce the dimensionality first and then use the covariance matrix estimates in this reduced space. Often Principle Component Analysis (PCA) dimensionality reduction [17] is used for this. In [18] several techniques are discussed which are more suited for dimension reduction for verification. However, most of these techniques are based on

• *A. Hendrikse, R. Veldhuis and L. Spreeuwers are with the Department of EEMCS, University of Twente, Enschede, the Netherlands. E-mail: a.j.hendrikse@ewi.utwente.nl*

Digital Object Indentifier 10.1109/TPAMI.2013.93

SOS estimation and, as we will show, can therefore be severely affected by the high dimensionality as well.

Another approach to the singularity problem is the use of regularisation [2], [18]. Several methods have been suggested, commonly based on cost function minimization of which Stein's loss function [19], [20], [21] is a well known example. However, these approaches do not make use of the theory available on the true underlying issues of the SOS estimation, as is given in [13], [14].

An extreme case of regularisation is to abandon the estimation of covariance matrix completely and use a euclidean distance measure. Since it requires no training, it does not suffer from the singularity problem, however it also lacks the advantage of the structure that can be discovered in the training data (see [18] as well). In [22] it was shown that the euclidean distance outperforms SOS estimates in several statistical tests for even moderately high values of p.

In the following sections we will study the effect of increasing p compared to N available for estimation in verification systems based on SOS, by studying a log likelihood ratio based verification system using likelihood estimates based on SOS estimates. In [23] this problem was studied for the simple case where one of the distributions has all eigenvalues equal. The likelihood ratio is a well known criterion in hypothesis testing [24] and pattern recognition [25] and the criterion is in use in many fields. This system we will discuss in section 2.

The problem of high dimensionality is caused by the added variants to the training data. In section 3 we discuss how the added variants increase the dimensionality and especially how they influence the SOS estimates. Although there are several possibilities to add variants (for example adding variants with less energy), in the remainder of the paper we assume that the added variants are similar to the variants already in the training data, therefore making the bias the dominant factor in the estimate for very large dimensionality. We then present an experiment in section 4 which demonstrates the problems the errors in the SOS estimates cause in a verification system: for even marginally large p the estimates are already outperformed by euclidean distance measures and after a certain p, the added variants start to decrease the systems performance until for very large p the system performs no better than random guessing. Of course, in practice p is never infinitely large, nor are all the variants of comparative energy strength, so a complete breakdown of the system will be rare, but this does point to a problem in the SOS estimators that has to be addressed.

In section 4.2 we therefore present not only observations from the experiment, but also start analysing these observations, using some hypotheses which can be derived from the theory on SOS estimation in high dimensional spaces. The relation between these hypotheses and the theory on SOS estimation is given in section 5.

The theory on SOS estimation in high dimensional spaces can also be used to improve the estimates. In section 6 we present improvements to the individual SOS estimates. However, verification depends on two distribution estimates: one describing the variation between samples from one class and one describing the variation between samples from different classes (a major point in [9]). In the correction of the distribution estimates, the relation between the two should be taken into account. In section 7 we present a method, the eigenwise correction, which can correct the SOS estimates in a verification system.

In section 8 we repeat the experiment of section 2, but using the corrections introduced in sections 6 and 7. The proposed corrections lead to a system converging from the ideal SOS based system for low p to a euclidean distance measure for very large p, while outperforming identity matrix regularisation methods in between. Note that prior information on the eigenvalue distribution can be used to improve the its estimates. An example of such an approach can be found in [26]. In section 9 we draw conclusions.

2 Verification using second order approximations

The purpose of a verification system is to test a claim that a sample $\boldsymbol{x}$ is resulting from a class c. A common approach to this problem is to use the likelihood ratio:

$$R(\boldsymbol{x}, c) = \frac{p(\bar{\boldsymbol{x}} = \boldsymbol{x} | \bar{c} = c)}{p(\bar{\boldsymbol{x}} = \boldsymbol{x} | \bar{c} \neq c)} \tag{1}$$

and only accept a claim if this ratio is above a threshold. Because a threshold is applied, $p(\bar{\boldsymbol{x}} = \boldsymbol{x} | \bar{c} \neq c)$ can be replaced with $p(\bar{\boldsymbol{x}} = \boldsymbol{x})$.

By varying the threshold a trade off can be made between the rate of the genuine claims (x belongs to class c) being rejected (False Rejection Rate (FRR)) and the rate of imposter claims (x belongs to another class) being accepted (False Acceptance Rate (FAR)). According to the Neyman-Pearson lemma [27] this test is optimal for deciding whether $\boldsymbol{x}$ is originating from class c or not for a given FAR [28].

The likelihood ratio approach requires the estimation of the distribution $p(\boldsymbol{x}, c)$ and $p(\boldsymbol{x})$ for which usually a training set is available. One problem is that number of samples (N) in the training set is limited, so determining both the distribution model and its parameters is problematic. A common strategy is to only determine the mean and the SOS of the training set and simply choose the distribution model.

As distribution model usually the Gaussian distribution is used, for three reasons: firstly, it is a well known distribution, which has been in use in statistics and other areas for a long time. Secondly, it is fully determined by the mean and the SOS. Thirdly, since

the Gaussian distribution has the highest entropy for given SOS [29], it is the best approximation according to the maximum entropy principle [30].

In a verification system, the sample model is usually extended as follows: the samples are composed of two parts, a within and a between part, via $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{w}} + \boldsymbol{\mu}_c$. $\boldsymbol{x}_{\mathrm{w}}$ is used to model variations between samples originating from the same class and its distribution is approximated by a normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_{\mathrm{w}})$. The between part $\boldsymbol{\mu}_c$ is used to model variations between samples of different classes and its distribution is approximated by $\mathcal{N}(\boldsymbol{\mu}_{\mathrm{t}}, \boldsymbol{\Sigma}_{\mathrm{b}})$. The distribution of $\boldsymbol{x}$ then becomes $\mathcal{N}(\boldsymbol{\mu}_{\mathrm{t}}, \boldsymbol{\Sigma}_{\mathrm{t}})$ with $\boldsymbol{\Sigma}_{\mathrm{t}} = \boldsymbol{\Sigma}_{\mathrm{w}} + \boldsymbol{\Sigma}_{\mathrm{b}}$. If these distributions are used in equation 1 and we take the logarithm, then the log likelihood ratio becomes:

$$\begin{aligned} L(\boldsymbol{x}, c) = &-(\boldsymbol{x} - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{w}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) \\ &+ (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{t}})^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{t}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{t}}) \end{aligned} \quad (2)$$

aside from a constant and some scaling.

The parameters of these distributions have to be estimated. The class means $\boldsymbol{\mu}_c$ are estimated by the sample mean of all the samples in the training set belonging to class c, the total mean is estimated by $\hat{\boldsymbol{\mu}}_{\mathrm{t}} = \frac{1}{C}\sum_{c=1}^{C} \hat{\boldsymbol{\mu}}_c$. $\boldsymbol{\Sigma}_{\mathrm{w}}$ and $\boldsymbol{\Sigma}_{\mathrm{b}}$ are estimated by the sample covariance matrices in (3) and (4) respectively, where $l(\boldsymbol{x}_k)$ returns the class label of $\boldsymbol{x}_k$.

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{w}} = \frac{1}{N - C} \sum_{k=1}^{N} \left(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_{l(\boldsymbol{x}_k)}\right) \left(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_{l(\boldsymbol{x}_k)}\right)^{\mathrm{T}} \quad (3)$$

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{b}} = \frac{1}{C - 1} \sum_{c=1}^{C} (\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}}_{\mathrm{t}}) (\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}}_{\mathrm{t}})^{\mathrm{T}} \quad (4)$$

To use these estimates to estimate $L(\boldsymbol{x}, c)$, these estimated covariance matrices have to be inverted. The inverse of a covariance matrix $\boldsymbol{\Sigma}$ is given by $\boldsymbol{E} \cdot \boldsymbol{D}^{-1} \cdot \boldsymbol{E}^{\mathrm{T}}$ where $\boldsymbol{\Sigma} = \boldsymbol{E} \cdot \boldsymbol{D} \cdot \boldsymbol{E}^{\mathrm{T}}$. $\boldsymbol{E}$ is an orthogonal matrix of which each column is an eigenvector of $\boldsymbol{\Sigma}$. $\boldsymbol{D}$ is a diagonal matrix, with the eigenvalues of $\boldsymbol{\Sigma}$ on the diagonal. If a model parameter is decomposed, its results are denoted by population eigenvalues $\boldsymbol{\lambda}$ and population eigenvectors. If an estimate based on training samples is decomposed, its results are denoted by sample eigenvalues $\boldsymbol{l}$ and sample eigenvectors.

To study the effect of the SOS estimation errors on the verification performance, we need to know which parts of the data have the largest influence on the likelihood ratio. This can be determined by determining for which unitary vector $\boldsymbol{w}$ the projection of $\boldsymbol{x}$ on this vector results in the largest variance of $L(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}, c)$. Using the fact that $\mathcal{E}\{L(\boldsymbol{x}, c)\} = 0$ it follows after some calculations that

$$\mathcal{E}\left\{L^2\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}, c\right)\right\} = 4\frac{\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Sigma}_{\mathrm{b}} \cdot \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Sigma}_{\mathrm{t}} \cdot \boldsymbol{w}} \quad (5)$$

This shows that the likelihood ratio is the most sensitive to projections with the largest between class over within class variance ratios. The fraction in equation 5 is the generalized Rayleigh quotient used in Linear Discriminant Analysis (LDA) to find the most discriminating subspace in the data [25], so LDA can be considered as applying PCA dimensionality reduction based on the variance of $L(\boldsymbol{x}, c)$ instead of the variance of the samples themselves. This is very similar to the argument used in [9] to develop the Asymmetric Principle Component technique.

3 Increasing Dimensionality by Adding Variants

A major point in this study is the increase of the dimensionality of the training data by adding variants. However, there are several operations that increase the number of variants. The first option is to add variants with similar characteristics, that is an equal amount of energy and similar discriminative capacity. The second option is to add variants with a very different energy and discriminative capacity. Without loss of generality we assume in the later case that the added variants have considerably less energy and discriminative capacity. A third option is to add variants with much less energy, but which have a higher discriminant power. For the moment we focus on the first two options, which are the extremes.

An example close to adding variants with similar characteristics is the fusion of two biometric modalities (for example face and fingerprint data) or two representations of one modality. An example close to adding variants with lower energy is using images with higher resolution. In natural images usually the most energy is in the lower frequency components [31], [32], which are already present in the low resolution images. The added variants, representing the high frequency components, then have less energy. The discriminative capacity of the added variants is hard to judge without detailed study, but since biometric modalities have reasonable comparable performance, it is reasonable to assume they do not differ too much in discriminative capacity.

In our study it is important how the added variants change the SOS of the data. In the first option, in which the added variants have similar energy and similar discriminative capacity, increasing the dimensionality can be modelled by resampling of the eigenvalue distribution. That is, the number of eigenvalues increases, but their corresponding empirical distribution stays more or less the same, where the empirical distribution $G_p(l)$ for a set of eigenvalues $\boldsymbol{l}$, is given by $\frac{1}{p}\sum_{k=1}^{p} \mathrm{u}(l - l_k)$. Increasing the eigenvalue set shown in figure 1a by this method is demonstrated in figure 1b. This closely matches the General Statistical Analysis (GSA) framework [13], [14].

The second option, in which the added variants have lower energy, can be modeled by extending the tail of curve describing the eigenvalues instead of resampling it as shown in figure 1c. This matches the

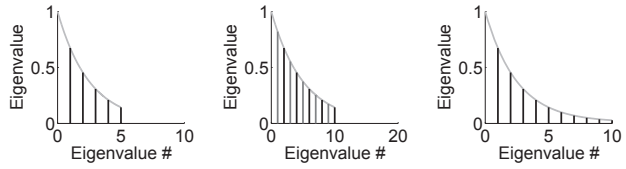


Fig. 1. Demonstration of the two methods to increase the dimensionality of the data. In method one the curve of the eigenvalues is fixed, so increasing the eigenvalue set given in a is extended by the lighter bars in figure b. In method two, the curve is extended to find the new eigenvalues, as is shown in c.

spiked population model presented in [33], [34], [35]. The distorting effect of the increased dimensionality now depends on the fall of of the eigenvalue curve: if the curve falls off rapidly, the larger eigenvalues will not or only marginally become biased with increasing dimensionality [35], just like adding a null space to the data, so effectively p was not increased at all.

As an example, consider that the eigenvalues are distributed exponentially, so $\lambda_k = e^{1-k}$. The total variance of the data is given by $\sum_{k=1}^{p} \lambda_k = \frac{1-e^{-p}}{1-e^{-1}}$, which for $p \to -\infty$ converges to $1 + \frac{1}{e-1}$, so the bulk is small compared to the first few eigenvalues and therefore the bias can only be marginal.

On the other hand, if the curve falls off much slower, then the bulk composed by the smaller eigenvalues will get large enough to also fully affect the estimation of the largest eigenvalues [35] and the results of increased p are similar to the increase of p by interpolation of the eigenvalue curve [36].

An example of this phenomenon is if the eigenvalues are set to $\lambda_k = e^{0.1 \cdot (1-k)} + 0.01$. For $p \to \infty$ the total data variance becomes arbitrarily large. This variance is distributed over the $N-1$ non zero sample eigenvalues, so at least the largest sample eigenvalue becomes arbitrarily large and is therefore a heavily biased estimate of the largest population eigenvalue.

In [9] it is argued that a common curve of the eigenvalues of facial image data is given by 1 over f, where f is the index of the eigenvalue. Such a curve has a slowly decaying tail, and the total variance in such data is given by $\sum_{k=1}^{p} \frac{1}{k}$. For arbitrarily large p this also becomes arbitrarily large and therefore even the largest sample eigenvalue is severely biased for sufficiently large p. To demonstrate that even $l_{\max}$, the largest sample eigenvalue gets biassed for a 1 over f distribution, we estimated $l_{\max}$ from 50 randomly generated data sets with a 1 over f distribution of $\boldsymbol{\lambda}$ with $p = 1000$ and $N = 20$. $l_{\max}$ was 1.17 on average, with a standard deviation of 0.12. However, for moderately N the required p for which the bias significantly affects $l_{\max}$ is much larger than in any real practical situation, so for most practical problems, only the smallest sample eigenvalues will be significantly biased as is the case in [9]. None the less for the limit $p \to \infty$ the estimation adheres to the limits derived in the following sections.

On top of that, as we reported in [12], we have strong clues that such a curve is caused by data generation model errors instead of the true underlying parameters. And since the curve extension option is a mixture between the curve resampling and adding a null space, we examen only curve resampling.

For the discriminative capacity of the added variants we have three options: give the added variants more, less or the same discriminative capacity, where adding features with the same discriminative capacity seems to be the most informing to us. The first option of adding variants with more discriminative capacity seems illogical to us, since if it is known that some features are more discriminative, then they would normally be considered first and the lesser features would be added later on, which corresponds to adding features with smaller discriminative capacity.

However, in practice the discriminative capacity of the variants will be unknown beforehand, so on average the added variants will have approximately the same discriminative capacity. If the added features do have a lower discriminative capacity, then the results presented in the coming sections give an upper limit on verification performance.

4 Second Order Statistics Estimation in High Dimensional Spaces

4.1 A verification experiment with PCA dimensionality reduction

In this section we experimentally demonstrate the weaknesses of the PCA dimensionality reduction in high dimensional verification problems under the curve resampling approach by performing a verification experiment with synthetic data. By using synthetic data we can focus solely on the errors introduced by the sample covariance matrix and not be disturbed by the other known problems, such as outliers in the data [10], [11], the eigenvalue bias introduced by incorrect sampling of some of the involved classes [9] and modeling errors [12]. We use a verification system equal to the system described in the previous sections and vary the dimensionality p of the samples in several iterations between a value much lower than the fixed N and a value considerably larger than N. If $p > N$, at least $p - N$ eigenvalues are zero and the singularity problem occurs as noted in section 1.

To solve the singularity problem, we perform PCA dimensionality reduction solution prior to verification and compare the verification performance with two limit cases: a theoretical optimum limit and a regularisation limit. In the theoretical limit we assume perfect estimation and replace the covariance matrix estimates with the population covariance matrices. In

the second case we do not estimate any second order structure, but we set the sample covariance matrices equal to scaled versions of the identity matrix. This is the limit of regularisation methods and it turns the probability measures into a euclidean distance such that the log likelihood ratio becomes:

$$\hat{L}_{\text{reglim}}(\boldsymbol{x}, c) = -\frac{1}{\bar{l}_{\text{w}}}\hat{\boldsymbol{x}}_{\text{w}}^{\text{T}}\hat{\boldsymbol{x}}_{\text{w}} + \frac{1}{\bar{l}_{\text{t}}}\hat{\boldsymbol{x}}_{\text{t,zm}}^{\text{T}}\hat{\boldsymbol{x}}_{\text{t,zm}} \quad (6)$$

where $\hat{\boldsymbol{x}}_{\text{w}} = \boldsymbol{x} - \hat{\mu}_c$, $\hat{\boldsymbol{x}}_{\text{t,zm}} = \boldsymbol{x} - \hat{\mu}_{\text{t}}$. $\bar{l}_{\text{w}}$ and $\bar{l}_{\text{t}}$ are the mean of the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{\text{w}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{t}}$ respectively.

We perform the experiment for two different choices of the fixed eigenvalue description curve. In both configurations we use 100 classes with 5 samples per class for training (which implies a fixed number of samples for both $\widehat{\boldsymbol{\Sigma}}_{\text{w}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{b}}$) and test with another 100 classes with 20 samples per class. In the first configuration we choose a 2 cluster like distribution: 10% of the between eigenvalues have a value of 0.5, the remaining 90% have a value of 0.05, but we smooth the borders between the two clusters so the eigenvalue scree plot follows a $\tanh$ curve. The within eigenvalues are chosen such that the total covariance matrix equals the identity matrix.

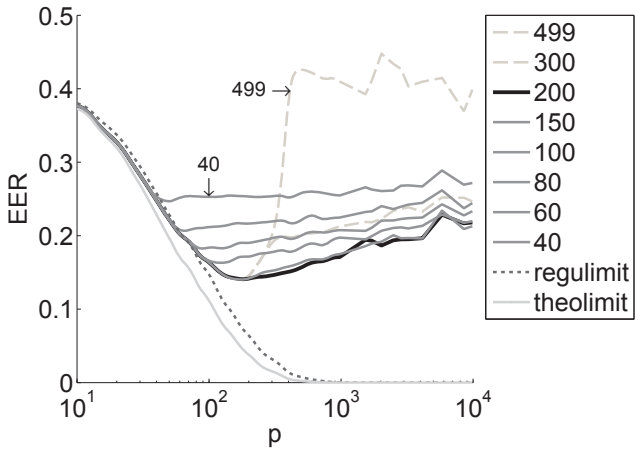
In the second configuration we choose the within and the between covariance matrices such that the total covariance matrix has eigenvalues $\lambda_{\text{t},k} = (1-\alpha)\, e^{-12.5(k/p)} + 0.01\,(1+4\alpha)$. The between eigenvalues are set to $\lambda_{\text{b},k} = 0.1\lambda_{\text{t},k}$, making the within eigenvalues $\lambda_{\text{w},k} = 0.9\lambda_{\text{t},k}$.

The trade off between FAR and FRR as described in section 2 can be made based on the Receiver Operating Characteristic (ROC) curve. However, since we want to study the verification performance for many different configurations, we determine only the Equal Error Rate (EER) rate, which is the point on the ROC curve where FAR equals the FRR. The analysis in the remainder of the paper support the idea that the results can be extended for other points on the ROC curve as well.

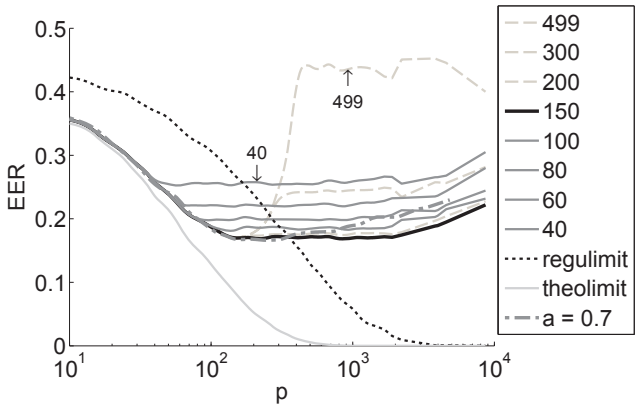
4.2 Results

Figure 2 shows the results of tests in which we fixed the number of components retained after dimensionality reduction. Figure 2a shows the EER versus p curves for the 2 cluster configuration, Figure 2b shows the curves for the exponential configurations ($\alpha = 0$). We also performed tests in which we fixed the total amount of variance retained after the reduction, but the results are lower bound by the best performing reduction to a fixed number of dimensions, so we do not show the results here. The 150 curve for $\alpha = 0.7$ shows that both distributions represent extremes: with increased constant the exponential curves moves towards the two cluster curve.

Several observations can be made:



(a) two cluster, fixed dimension reduction



(b) exponential, fixed dimension reduction

Fig. 2. PCA dimensionality reduction as a solution to the singularity problem compared with a theoretical optimum limit and a regularisation limit. The curves decrease in EER with increasing PCA components retained until a minimum curve of either 150 or 200 components after which the curves become dashed and increase in EER with increasing number of PCA components retained.

4.2.1 Regularisation limit outperforms PCA

PCA dimensionality reduction is already outperformed by the regularisation limit for moderate dimensionality. Even though the exponential configuration is very different from the identity configurations assumed by the regularisation limit, for a dimensionality of around 400 and higher all PCA dimensionality reduction configurations are outperformed by the regularization limit. With the 2 cluster configuration, the difference between the theoretical limit and the regularisation limit is small and PCA is already outperformed for $p = 100$.

To explain this, we hypothesize that in the limit $p \to \infty$ the sample eigenvectors form a random orthogonal basis and the sample eigenvalues cluster into two sets: one set of N non zero equal valued eigenvalues and the remainder are all zero. These hypothesis will be

discussed in sections 5.2 and 5.1.1. This means that for sufficiently large p PCA dimensionality reduction (or any other form of SOS based dimension reduction, such as in [26]) turns the likelihood calculations into a euclidian distance calculation in a random subspace with a dimensionality of at max N, while in the regularisation limit the probability calculations are an euclidean distance calculation in the full dimensional space (equation 6): PCA dimensionality reduction removes information without improving the structure estimate.

Note that this is especially true for the commonly considered "all noise" distribution, where all the eigenvalues are equal [23]. In that case all structure found by PCA is based on random structure and dimension reduction will always under perform compared to the regularisation limit.

4.2.2 EER is not a non decreasing function of p

The EER curves of all PCA dimensionality reduction configurations show a dip: after a certain value of p, the EER increases. This is highly remarkable, because this implies that there is a minimum in EER for the PCA dimensionality reduction method and adding more variants after this point hurts performance.

This can be explained as follows: PCA dimensionality reduction projects the data on a subspace of a dimensionality of at max N. However, only for $p \to \infty$ the estimates are solely based on the random structure in the data: the sample eigenvectors are random and the sample eigenvalues are uniform in this subspace. For $p \approx N$, the estimates still partially depend on the population parameters. Therefore the probability calculations in the subspace estimated for smaller p values are more accurate and hence the error rates eventually go up with increasing p.

4.2.3 The PCA solution still breaks down for large p

In case of retaining (almost) all components with a non zero eigenvalue, the results become highly unstable for larger values of p. Moreover, the EER goes up to almost random guessing.

To explain this, focus on the fact that the likelihood ratio depends on the between class variances over the within class variances (equation 5). We hypothesis that for very large p values, the subspace in which the within class variance estimate is non zero will become orthogonal to the subspace in which the between class variance estimate is non zero (see section 5.3 for proof). As a result, if PCA dimension reduction is applied either the within class covariance matrix is still singular, or the total covariance matrix is identical to the within class covariance matrix. In the latter case, using the fact that

$$\hat{\mathbf{\Sigma}}^{-1}_{\mathrm{nzw,w}} = \hat{\mathbf{\Sigma}}^{-1}_{\mathrm{nzw,t}} = \frac{p}{(N-C)\,\bar{l}}\mathbf{I} \tag{7}$$

and

$$\begin{aligned} \mu_{\mathrm{t}}^{\mathrm{T}}\mu_{\mathrm{t}} + \mu_{\mathrm{w}}^{\mathrm{T}}\mu_{\mathrm{w}} &= \frac{1}{2}\left(\mu_{\mathrm{t}} - \mu_{\mathrm{w}}\right)^{\mathrm{T}}\left(\mu_{\mathrm{t}} + \mu_{\mathrm{w}}\right) + \\ &\quad \frac{1}{2}\left(\mu_{\mathrm{t}} + \mu_{\mathrm{w}}\right)^{\mathrm{T}}\left(\mu_{\mathrm{t}} - \mu_{\mathrm{w}}\right) \end{aligned} \tag{8}$$

the log likelihood ratio reduces to

$$\begin{aligned} \hat{L}\left(c, \boldsymbol{x}\right) &= \frac{2p}{(N-C)\,\bar{l}}\left(\hat{\boldsymbol{\mu}}_{\mathrm{nzw},c} - \hat{\boldsymbol{\mu}}_{\mathrm{nzw},t}\right)^{\mathrm{T}} \cdot \\ &\quad \left(\boldsymbol{x}_{\mathrm{nzw}} - \frac{\hat{\boldsymbol{\mu}}_{\mathrm{nzw},c} + \hat{\boldsymbol{\mu}}_{\mathrm{nzw},t}}{2}\right) \end{aligned} \tag{9}$$

where the subscript nzw means that the corresponding variable is the part in the subspace with non zero within variance. The likelihood ratio is calculated in a subspace of dimensionality 1 of the already random subspace of non zero within class variance, which is just marginally better than random guessing.

Note that the break down effects already occur at smaller values of p for the smaller eigenvalues compared to the larger eigenvalues. This was to be expected from studies showing that the larger eigenvalues are less affected by the bias [33], [34], [35]. If the dimension reduction removes more of the smaller eigenvalues, the breakdown effect occurs for larger p values, which also explains the commonly observed overtraining for fixed p, as noted in the next section.

4.2.4 Overtraining for fixed p

The experiment shows that SOS estimation exhibits overtraining if p becomes large, however in facial biometrics overtraining is often observed for a fixed p: error rates go up if the number of reduction components is chosen too large [9], [37]. This effect can be observed in figure 2 if the curves are considered for a fixed p larger than approximately 150. For example consider Figure 2b for $p \approx 250$. There is a clear minimum in EER for a dimension reduction to $p_{\mathrm{red}} = 150$. It seems that the classical overtraining effect observed in biometrics is related to the overtraining effect of SOS estimation in high dimensional data.

In these explanations we used several hypothesis on the effect increasing p has on SOS estimates. In the next sections we prove some of these hypothesis, while others are demonstrated experimentally.

5 Overtraining in Second Order Statistics estimation

In the previous sections we showed that PCA dimensionality reduction is a far from optimal solution of the singularity problems caused by errors of the sample SOS estimators. The major reason is that these commonly used sample estimators become for increasing p more and more based on random fluctuations in the samples rather than the actual structure of the data. One effect is that the sample eigenvalues become significantly biased estimates of the population eigenvalues as we show next.

5.1 Eigenvalue bias

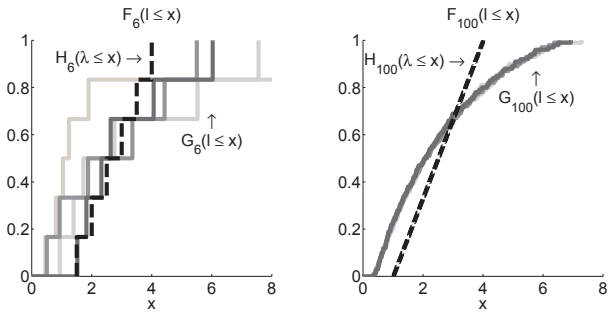
Bias is the expected value of the difference between a parameter value and its estimators expected value. For eigenvalue estimation this equals to:

$$\mathcal{E}\left\{\boldsymbol{l}-\boldsymbol{\lambda}\right\} \tag{10}$$

The bias of an estimator is typically determined using Large Sample Analysis (LSA): that is to evaluate (10) under the assumption that $N \rightarrow \infty$. Under this assumption, the sample eigenvalues seem unbiased. However, in many applications the assumption that N is large enough to solely determine the statistics of the estimate is questionable and therefore the results of LSA may not be a valid approximation.

In practice p is in the same order or larger then N. Therefore, in GSA it is assumed that $N, p \rightarrow \infty$ while $\frac{p}{N} \rightarrow \gamma \in [0, \infty)$. Because the number of eigenvalues depends on p and p becomes very large in these analysis, instead of considering the set of eigenvalues, the corresponding empirical distribution of the eigenvalues is considered.

In figure 3 an example of the GSA limit in eigenvalue estimation is given. In the example synthetic data is generated with the population eigenvalues distributed uniformly between 1 and 3. From this synthetic data the sample eigenvalues are estimated, for $p = 6$ and 100 with $N = 30$ and 500 respectively, keeping $\frac{p}{N} = \frac{1}{5}$. The 2 subfigures show the population eigenvalue distribution $H_p(\lambda)$ (dashed line) and 4 sample eigenvalue distributions $G_p(l)$ (solid lines) per setting. In the 6 dimensional experiment large variations occur between the different sample eigenvalue distributions, but for $p = 100$, the sample eigenvalue distributions have converged, although not to the population eigenvalue distribution. This is due to the bias of the sample eigenvalue estimator.



(a) 6 dimensions (b) 100 dimensions

Fig. 3. GSA demonstration

In [13] a description known as the Marčenko Pastur (MP) equation was given of the relation between the population eigenvalues and the sample eigenvalues in the GSA limit for a limited set of distributions of the samples. In [14] it was proven that this relation holds for a much larger set. The MP equation is given by:

$$-\frac{1}{v(z)} = z - \gamma \int \frac{\lambda \mathrm{d}H(\lambda)}{1+\lambda v(z)} \tag{11}$$

where $v(z) = \gamma \frac{-1}{z} + (1-\gamma) m_G(z)$ and $\mathrm{Im}\{z\} > 0$. $m_G(z)$ is the Stieltjes transform of $G(l)$: $m_G(z) = \int \frac{\mathrm{d}G(l)}{l-z}$. From this relationship it follows that the bias depends on the ratio $\frac{p}{N}$. The higher this ratio, the more severe the bias.

5.1.1 Eigenvalue bias limits

From the MP equation it is in general rather difficult to determine the sample eigenvalues corresponding to a given population eigenvalue set and visa versa. However, two limit cases can be considered in which the MP equation can be used to determine the relation between the sample eigenvalues and the population eigenvalues: the case in which $N >> p$ and the case in which $p >> N$.

If $N >> p$ then LSA is accurate and the bias of $\boldsymbol{l}$ is insignificant. If $p >> N$, using the MP equation it can be proven that $\boldsymbol{l}$ only depends on the average of the $\boldsymbol{\lambda}$, $\bar{\lambda}$, and more specific, the sample eigenvalues split into two clusters: one cluster of N eigenvalues equal to $\gamma \cdot \bar{\lambda}$ and a second cluster of $p-N$ zero valued eigenvalues, as is shown in the next section.

Especially this second limit is used in the explanations of the observations of section 4, since this proves the hypothesis that using SOS estimates indeed turns the probability calculations into a euclidean distance in a N dimensional subspace.

5.1.2 Loss of structure in high dimensional problems

We now prove that if $p >> N$ then the estimated SOS only depend on $\bar{\lambda}$, the population eigenvalue mean.

First note that $\mathcal{O}\left(\left\|\frac{1}{v_\infty(z)}\right\|\right) = \gamma^a$ only leads to no contradiction in $\mathcal{O}\left(\left\|\frac{1}{v_\infty(z)}\right\|\right) = \mathcal{O}\left(\left\|z - \gamma \int \frac{\lambda \mathrm{d}H(\lambda)}{1+\lambda v_\infty(z)}\right\|\right)$, derived from equation (11), if $a = -1$. Using this result we can determine the sample eigenvalue distribution if $\gamma \rightarrow \infty$:

$$\lim_{p\rightarrow\infty} v(z) = \lim_{\gamma\rightarrow\infty} \left(\gamma \int \frac{\lambda \mathrm{d}H(\lambda)}{1+\lambda v(z)} - z\right)^{-1} \tag{12}$$

$$= \frac{1}{\gamma\bar{\lambda} - z} \tag{13}$$

which is the Stieltjes transform of $u\left(l - \gamma\bar{\lambda}\right)$, so the sample eigenvalue set converges to a set of n eigenvalues equal to $\gamma\bar{\lambda}$ and $p-n$ eigenvalues equal to 0, independent of $H(\lambda)$, if $H(\lambda)$ has a bounded support. These findings concur with [38], where the value of the largest eigenvalue is studied under similar conditions, and the results of [33], [34], [35] if the bulk grows large enough.

5.2 Errors in the eigenvectors

Besides a bias in $\boldsymbol{l}$, errors also occur in the sample eigenvector ($\hat{\boldsymbol{\theta}}$): the sample eigenvectors do not align with the population eigenvectors. In [39] and [40] it is shown that this misalignment of $\hat{\boldsymbol{\theta}}$ with population eigenvector ($\boldsymbol{\theta}$) can be studied using the inner product $\boldsymbol{\theta}^{\mathrm{T}} \cdot \hat{\boldsymbol{\theta}}$ and it is argued that a relation exists similar to the relation the MP equation describes between the sample eigenvalues and the population eigenvalues.

The inner products between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ depends on the ratio $\frac{p}{N}$ as well. We focus on the same two limit cases as were considered before in the eigenvalue bias analysis: the case $N >> p$ and the case $p >> N$.

If $N >> p$ the eigenvectors align, so the inner product of a sample eigenvector with a population eigenvector is only non zero if their corresponding eigenvalues are equal (recall that the sample eigenvalues are unbiased, so every sample eigenvalue matches at least one population eigenvalue).

For the case $p >> N$ we hypothesize the following:

Hypothesis 5.1: Under the same conditions as assumed in section 5.1.2 all structure in the eigenvectors is lost, making the sample eigenvectors a random basis.

Although we do not have a complete formal proof of this hypothesis, we present 3 arguments to support it. Firstly, all structure of the population eigenvalues is lost except for their mean during the estimation (see section 5.1.1), so the same sample eigenvalues would be observed if all population eigenvalues are equal. If all population eigenvalues are equal, then any basis is a valid solution for the population eigenvectors and thus also for the sample eigenvalues.

Secondly, the same proof showed that the sample eigenvalues split into two clusters of equal valued eigenvalues. Therefore, within the two subspaces, hypothesis 5.1 is true.

Thirdly, we demonstrate the hypothesis experimentally. In Figure 4 we show the sum of the squared inner products of the sample eigenvector corresponding to the largest sample eigenvalue with the population eigenvectors corresponding to the smallest population eigenvalues. The population eigenvalues are distributed uniformly between 1 and 2. The curves clearly converge to the thin black line for larger p/N, which represents the limit of uniform sample eigenvector population eigenvectors inner product.

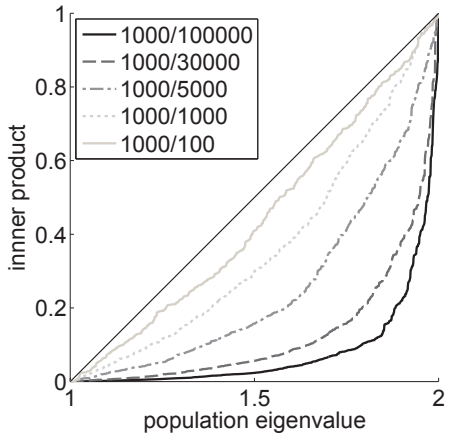


Fig. 4. Sum of the squared inner products of the first sample eigenvector with all the population eigenvectors for several p/N ratios. The eigenvectors are indexed according to their corresponding eigenvalue.

5.3 Limits in high dimensional verification

In section 5.1 we determined $\boldsymbol{l}$ in the limit $p \to \infty$ for a single distribution and showed that $N-1$ eigenvalues are $\gamma\bar{\lambda}$ and the remainder are zero. However, the log likelihood ratio used in verification depends on two distributions: the within class distribution and the between class distribution. So for verification it is more important to determine what the relation is between the within class distribution and the between class distribution in the limit $p \to \infty$.

To find this relation, we first find the limit estimates of the two distributions separately. It is straight forward to show that the within class distribution estimates adhere to the model of section 2, however, as we derived in [41], the between class estimates are based on a mixture of the between class distribution and the within class distribution. But both distributions are sampled equally, so $\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}$ adheres to the model of section 2 as well.

Therefore the eigenvalues of both $\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}$, $\boldsymbol{l}_{\mathrm{w}}$ and $\boldsymbol{l}_{\mathrm{b}}$ respectively, adhere to the limit described in section 5.1: for large p, $\boldsymbol{l}_{\mathrm{b}}$ separate into a cluster of $C-1$ eigenvalues of value $\frac{p}{C-1}\bar{\lambda}_{\mathrm{b}}$ and the rest zero valued. $\boldsymbol{l}_{\mathrm{w}}$ separates into a cluster of $N-C$ eigenvalues of value $\frac{p}{N-C+1}\bar{\lambda}_{\mathrm{w}}$ and the rest zero valued.

In section 2 we derived that the samples x are also normally distributed. However, they are not independent and identically distributed (i.i.d.) and therefore the limits derived in section 2 do not apply to the eigenvalues of $\hat{\boldsymbol{\Sigma}}_{\mathrm{t}}$, $\boldsymbol{l}_{\mathrm{t}}$, so its limit distribution has to be determined in another way. In the following paragraphs we argue that for large p, $\boldsymbol{l}_{\mathrm{t}}$ splits into three clusters: one cluster of the non zero within eigenvalues, one cluster of the non zero between eigenvalues and a cluster of null eigenvalues.

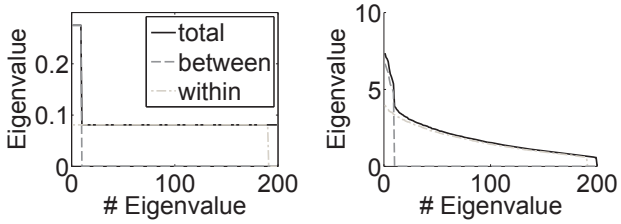
In order to prove this, we have to determine the relation between the subspace in which the within class variance estimate is non zero and the subspace in which the between class variance estimate is non zero. According to hypothesis 5.1 the eigenvectors corresponding to the non zero valued cluster of both the within estimate and the between estimate span a random subspace in the sample space and will be randomly oriented with respect to each other. Therefore, as shown in section 5.2 any basis vector of

one of subspaces will have an inner product with any of the basis vectors of the other subspace proportional to p^{-1}. The sum of the inner products of the $N - C$ basis vectors of the non zero within subspace with the $C - 1$ basis vectors of the non zero between subspace will therefore become proportionally to $\frac{(N-C)\cdot(C-1)}{p}$, which vanishes for $p \to \infty$, so the subspaces are orthogonal in this limit. Since the total covariance estimate is the sum of the within covariance estimate and the between covariance estimate, the decomposition of the total matrix leads to three cluster eigenvalue set as described before.

From these analysis we can also construct a limit for the eigenvectors of the total covariance matrix: they are a combination of the between eigenvectors corresponding to the non zero between eigenvalues with the within eigenvectors corresponding to the non zero within eigenvalues and a random basis spanning the remainder of the space.

The orthogonality also implies a perfect verification system: every class is separable, a typical overtraining result. This orthogonality was also the last point required in the explanations of section 4.

To demonstrate the limit distributions of $\boldsymbol{l}_{\mathrm{w}}$, $\boldsymbol{l}_{\mathrm{b}}$ and $\boldsymbol{l}_{\mathrm{t}}$, we did an estimation experiment with synthetic data. The distribution parameters of the data are the same as in the exponential configuration of the experiment described in section 4. We generated only 10 classes for the training set, with 20 samples per class. The theoretical limit for this configuration is shown in figure 5a. Figure 5b shows an estimate for $p = 4000$. Although for $p = 4000$ the estimate has not yet converged to the limit, it still seems to confirm that the determined limit is correct.



(a) Calculated limit (b) Estimate for $p = 4000$

Fig. 5. Second order estimation in a verification scheme for the limit $p \to \infty$.

6 Improved Second Order Statistics in High Dimensional Spaces

In the previous sections we described several effects an increase of p has on SOS estimation if N is kept at a fixed value. These effects led to some remarkable observations in the experiment described in section 4. In section 4 we also showed, based on the analysis in section 5, how these observations are related to the sample estimators used to estimate the SOS. However, based on these analysis, we can also make improvements on the estimates. In the coming sections we describe several improvements which finally result in a system smoothly changing from theoretical optimal sample estimate if $N >> p$ to the regularisation limit for $p >> N$. These improvements are experimentally evaluated in section 7.

6.1 Bias correction

The eigenvalue bias is a non random distortion of the estimate of population eigenvalues, so it can be removed from this estimate. This is schematically represented in figure 6. In figure 6 the bias introduction is represented as a function $B(\lambda)$ which is applied to the population eigenvalues $\boldsymbol{\lambda}$ and it results in the sample eigenvalues $\boldsymbol{l}$. Bias correction can be thought of as applying an estimated inverse of $B(\lambda)$ to $\boldsymbol{l}$, resulting in corrected population eigenvalue estimates $\hat{\boldsymbol{\lambda}}^c$.

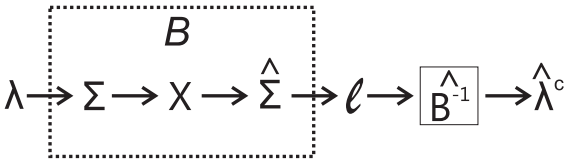


Fig. 6. Schematic representation of bias correction.

We use 3 bias correction methods in the remainder of this article. The first method is the correction method developed by Karoui in [42]. It is based on the MP equation, but instead of estimating the eigenvalues directly, a distribution is estimated which describes the population eigenvalues in the GSA limit. From this distribution estimates of the population eigenvalues themselves still have to be determined.

The second method is the fixed point eigenvalue correction [43]. It is based on a fixed point approach of solving the MP equation which determines the sample eigenvalue distribution of a given population eigenvalue set. By adjusting the population eigenvalue set such that the sample eigenvalue distribution estimated by the fixed point method matches the empirical sample eigenvalue distribution of the data, an estimate of the population eigenvalues of the data can be determined.

The third method was developed by Ledoit and Wolf in [2]. It is based on regularisation, but unlike many other regularisation algorithms it is able to handle situations in which $p > N$.

We compare these algorithms with classical PCA dimensionality reduction. In the experiment of section 4 we determined that the reduction to a fixed dimensionality of 150 components leads to some of the best results for the specific settings of the experiment.

6.2 Correction limits

In the experiment of section 8 we vary p, so all methods are tested for different ratios of $\frac{p}{N}$. In section 5.1.1

we determined that the real bias is hard to determine in advance, and so is the desired function for bias correction. However, the two limit cases of $N >> p$ and $p >> N$ and the case of $p \geq N$ can already be determined in advance.

Many more samples than dimensions, $N >> p$

The sample eigenvalue bias is negligible, so no correction is required. The correction methods should therefore converge to the identity function in this limit, or $\lim_{\gamma\to\infty} \widehat{\boldsymbol{B}^{-1}}(l) = l$. For most methods this is the case except for the Karoui correction, since it estimates distributions instead of sets.

More dimensions than samples, $p \geq N$

The sample estimate necessarily contains $N - p + 1$ zero valued eigenvalues, even if the population eigenvalues are all non zero. Correction of the sample eigenvalues should make these eigenvalues non zero. Moreover, based on the principle of maximum entropy [30], all these zero valued sample eigenvalues should be corrected to an equal, non zero value.

Many more dimensions than samples, $p >> N$

Most of the sample eigenvalues will be zero valued and should be corrected to an equal non zero value. However, as we have shown in section 5.1.1, the sample eigenvalues are only dependent on $\bar{\lambda}$ in this limit, all other characteristics of the population eigenvalues are lost. Therefore, according to the principle of maximum entropy [30], all the sample eigenvalues, zero valued and non zero valued, should be set to $\bar{\lambda}$, which is also the mean of the sample eigenvalues $\bar{l}$. This turns the likelihood calculations into an euclidean distance measure in the full p dimensional space, which is exactly the regularisation limit.

6.3 Variance correction

If eigenvalue bias correction is applied, then very accurate estimates of the population eigenvalues can be obtained. However, the SOS are described by the combination of eigenvalues and eigenvectors: the eigenvalues give the maxima of the variances of the data and the eigenvectors describe in which directions they occur. In other words, the corrected eigenvalues give an estimate of the variance along the population eigenvectors. Since the sample eigenvectors differ from the population eigenvectors, the corrected sample eigenvalues do not give an accurate estimate of the variance of the population distribution along the sample eigenvectors. With an additional correction, the variance correction [39], these variances can be determined. This idea is further studied in [40].

7 Bias correction in verification

The verification decision as described in section 2 is based on a comparison between the variances of samples from the same class with variances of samples between different classes. These variances are captured in the estimates of $\boldsymbol{\Sigma}_\mathrm{w}$ and $\boldsymbol{\Sigma}_\mathrm{t}$, where $\boldsymbol{\Sigma}_\mathrm{t}$ can be split into $\boldsymbol{\Sigma}_\mathrm{w} + \boldsymbol{\Sigma}_\mathrm{b}$.

Estimates of all these matrices are based on a limited amount of samples and so their eigenvalues are biased. In this section we show how the corrections presented in the previous sections can be applied to these estimates, which we denote by bias correction in verification (BCIV), improving the verification results. Only two out of the three sample eigenvalue sets $\boldsymbol{l}_\mathrm{w}$, $\boldsymbol{l}_\mathrm{b}$ and $\boldsymbol{l}_\mathrm{t}$ have to be corrected, where $\boldsymbol{l}_\mathrm{w}$, $\boldsymbol{l}_\mathrm{b}$ and $\boldsymbol{l}_\mathrm{t}$ are the sample eigenvalues of $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$, $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{t}$ respectively. The third set is fixed because there is a fixed relation between $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$, $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{t}$ as shown in [9], [41].

Let the superscript ^{c} indicate a corrected estimate. If $\boldsymbol{l}_\mathrm{b}$ and $\boldsymbol{l}_\mathrm{t}$ are corrected, then $\hat{\boldsymbol{\Sigma}}^\mathrm{c}_\mathrm{w}$, needed for the log likelihood ratio (equation 2), follows from the subtraction $\hat{\boldsymbol{\Sigma}}^\mathrm{c}_\mathrm{t} - \hat{\boldsymbol{\Sigma}}^\mathrm{c}_\mathrm{b}$. This subtraction easily leads to negative eigenvalues, and therefore this correction option is not considered any further.

The option to correct $\boldsymbol{l}_\mathrm{w}$ and $\boldsymbol{l}_\mathrm{t}$ we denote by improper BCIV, since in general the bias in $\boldsymbol{l}_\mathrm{t}$ does not adhere to the MP equation as shown in section 5.3. This correction option is still considered though, since its error is only significant for larger p values or if N significantly differs from $2C$, because if $N \approx 2C$, the effective number of samples for estimating $\boldsymbol{\Sigma}_\mathrm{w}$ and $\boldsymbol{\Sigma}_\mathrm{b}$ are almost the same, and if the distributions of $\boldsymbol{\lambda}_\mathrm{w}$ and $\boldsymbol{\lambda}_\mathrm{b}$ are also quite similar, then the distribution of the total sample eigenvalues can be approximated by the MP equation. Based on these considerations and previous tests in [44], we decided in [45] to perform the bias correction on $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{t}$, using the total number of samples as the amount of samples used for the estimation of $\hat{\boldsymbol{\Sigma}}_\mathrm{t}$.

The last option is to correct $\boldsymbol{l}_\mathrm{w}$ and $\boldsymbol{l}_\mathrm{b}$. The estimation of $\boldsymbol{\Sigma}_\mathrm{b}$ is not without problems: in [41] we show that $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ is an estimate of a mixture of $\boldsymbol{\Sigma}_\mathrm{b}$ and $\boldsymbol{\Sigma}_\mathrm{w}$ instead of a pure estimate of $\boldsymbol{\Sigma}_\mathrm{b}$. In combination with eigenvalue bias, this has several effects:

1) Due to the crosstalk, $\mathcal{E}\left\{\hat{\boldsymbol{\Sigma}}_\mathrm{w} + \hat{\boldsymbol{\Sigma}}_\mathrm{b}\right\} = \boldsymbol{\Sigma}_\mathrm{t} + \frac{1}{N_{pc}}\boldsymbol{\Sigma}_\mathrm{w}$, where $N_\mathrm{pc} = \frac{N}{C}$. This leads to erroneous estimates of $p(\boldsymbol{x})$, so $\hat{\boldsymbol{\Sigma}}_\mathrm{t}$ should be estimated by $\frac{N_\mathrm{pc}-1}{N_\mathrm{pc}}\hat{\boldsymbol{\Sigma}}_\mathrm{w} + \hat{\boldsymbol{\Sigma}}_\mathrm{b}$.
2) The crosstalk changes the directions for which the likelihood ratio is most sensitive (see the end of section 2). The crosstalk will change equation 5 into

$$4\left(\frac{1}{N_\mathrm{pc}}\frac{\boldsymbol{w}^\mathrm{T}\hat{\boldsymbol{\Sigma}}_\mathrm{b,cross}\boldsymbol{w}}{\boldsymbol{w}^\mathrm{T}\hat{\boldsymbol{\Sigma}}_\mathrm{w}\boldsymbol{w}} + \frac{\boldsymbol{w}^\mathrm{T}\hat{\boldsymbol{\Sigma}}_\mathrm{b,nocross}\boldsymbol{w}}{\boldsymbol{w}^\mathrm{T}\hat{\boldsymbol{\Sigma}}_\mathrm{w}\boldsymbol{w}}\right) \quad (14)$$

where $\hat{\boldsymbol{\Sigma}}_\mathrm{b,cross}$ is the estimate of the crosstalk of

$\boldsymbol{\Sigma}_\mathrm{w}$ in $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{b,nocross}}$ is the between estimate without the crosstalk. A crosstalk part which is equal to $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ will give an equal increase of the variance of L in all directions, so it does not change the order of the directions according to the sensitivity of L. However, $\hat{\boldsymbol{\Sigma}}_{\mathrm{b,cross}}$ is estimated from a different part of the samples than $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$, usually with a different number of samples as well ($N-C$ for $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ and $C-1$ for $\hat{\boldsymbol{\Sigma}}_{\mathrm{b,cross}}$). Therefore both the bias of the eigenvalues and the sample eigenvectors will differ for the two estimates. In fact, if the data is really noisy (large within class variances), then the differences between $\hat{\boldsymbol{\Sigma}}_{\mathrm{b,cross}}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ solely determine the sensitivity of the log likelihood ratio.

3) Again due to the difference between $\hat{\boldsymbol{\Sigma}}_{\mathrm{b,cross}}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$, simply subtracting a fraction of $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ from $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ cannot remove the crosstalk. This would have been possible if no bias was present.

The correction of the $\boldsymbol{l}_\mathrm{b}$ is however theoretically sound (see section 5.3). Note that focussing on $\boldsymbol{l}_\mathrm{w}$ and $\boldsymbol{l}_\mathrm{t}$ correction is no solution for the crosstalk problem, because the difference between $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{t}$ is solely caused by $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$. So it seems that the $\boldsymbol{l}_\mathrm{w}$ and $\boldsymbol{l}_\mathrm{b}$ correction is the best option for BCIV. However, it runs into problems if p is close to or larger than N as will be explained in the next section.

7.1 Correction in null spaces and verification

Due to the difference in effective number of samples for estimating $\boldsymbol{\Sigma}_\mathrm{w}$ and $\boldsymbol{\Sigma}_\mathrm{b}$ a problem arises if $p >>$N. We illustrate this problem with a synthetic data experiment where we attempt to estimate SOS from a training set with $p = 800$. The samples originate from 100 classes and for each class 5 samples were generated. Both the within class and the between class eigenvalues are uniformly distributed between 0.5 and 0.05 (solid line in figure 7), so there is no discriminative distinction between the different orientations. However, the sample estimates, given by the dark dashed line and the light dash-dotted line, do show a large discriminative difference.

Bias correction of the two separate distributions reduces this somewhat, as shown by the correction curves (the lighter dashed line and the light solid line), but in the null space something odd can be observed: the bias correction causes a between over within ratio which is considerably larger than 1, wrongfully suggesting that that part of the null space is highly discriminative. In [46] a very similar problem was observed in a recognition experiment in which the eigenvalues of an estimate related to $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ are regularised. They found that the entire null space has too large a weight in the verification decision.

To explain this, note that bias correction leads to equal valued eigenvalues in the null space (section 6.2). The null spaces of $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ differ both in number of dimensions ($p-(N-C)$ and $p-(C-1)$ respectively) and orientation. Therefore the value with which the null space of $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ is corrected can be much lower than the value of the null space of $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$, as the example demonstrates, resulting in a arbitrarily large ratio between the last few eigenvalues which then fully determines verification results (equation 5).

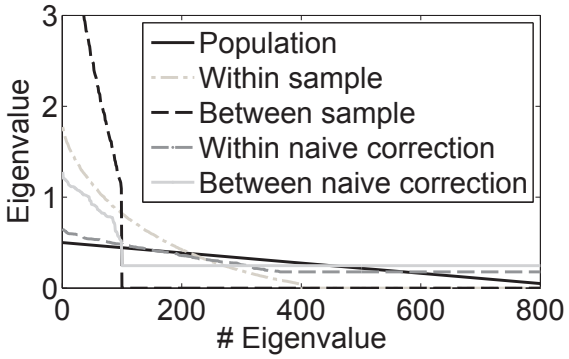


Fig. 7. Naive BCIV example.

This problem arises because if $\boldsymbol{l}_\mathrm{w}$ and $\boldsymbol{l}_\mathrm{b}$ are corrected as just described, the ratio of between over within variance is not taken into account. We therefore denote this correction by naive BCIV.

7.2 Eigenwise bias correction in verification

In the previous sections we demonstrated that both the naive BCIV and the improper BCIV have disadvantages. We developed a third method, the eigenwise BCIV, which does not suffer from the problems of either the improper BCIV or the naive BCIV. As said before, only $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$ are genuine sample covariance matrices in the sense that these estimates adhere to the conditions of the MP equation. So eigenwise BCIV corrects these.

To prevent the problems of naive BCIV, the method consists of the following steps, shown schematically in figure 8:

1) Estimate $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ and $\hat{\boldsymbol{\Sigma}}_\mathrm{b}$.
2) Decompose $\hat{\boldsymbol{\Sigma}}_\mathrm{w}$ into its eigenvectors $\hat{\boldsymbol{E}}_\mathrm{w}$ and eigenvalues $\boldsymbol{l}_\mathrm{w}$.
3) Correct $\boldsymbol{l}_\mathrm{w}$ to get a less biased estimate $\hat{\boldsymbol{\lambda}}_\mathrm{w}^\mathrm{c}$ using the correction methods from section 6.
4) Use $\hat{\boldsymbol{E}}_\mathrm{w}$ and $\hat{\boldsymbol{\lambda}}_\mathrm{w}^\mathrm{c}$ to determine $\hat{\boldsymbol{\Sigma}}_\mathrm{b}^\mathrm{ww}$, which is the between covariance matrix if the within data is whitened.
5) Decompose $\hat{\boldsymbol{\Sigma}}_\mathrm{b}^\mathrm{ww}$ into eigenvectors $\hat{\boldsymbol{E}}_\mathrm{b}^\mathrm{ww}$ and eigenvalues $\boldsymbol{l}_\mathrm{b}^\mathrm{ww}$.
6) Correct $\boldsymbol{l}_\mathrm{b}^\mathrm{ww}$ to get $\hat{\boldsymbol{\lambda}}_\mathrm{b}^\mathrm{c,ww}$, again using the methods from section 6.
7) Combine $\hat{\boldsymbol{E}}_\mathrm{b}^\mathrm{ww}$ and $\hat{\boldsymbol{\lambda}}_\mathrm{b}^\mathrm{c,ww}$ to get a new estimate of the between matrix in the within whitened space $\hat{\boldsymbol{\Sigma}}_\mathrm{b}^\mathrm{c,ww}$ without bias.
8) From $\hat{\boldsymbol{\Sigma}}_\mathrm{b}^\mathrm{c,ww}$ determine the corrected between matrix in the original input space $\hat{\boldsymbol{\Sigma}}_\mathrm{b}^\mathrm{c}$.

Two assumptions form the basis of this algorithm. Firstly, the within covariance estimate and the between covariance estimate $\hat{\Sigma}_\mathrm{w}$ and $\hat{\Sigma}_\mathrm{b}$ should be based on independent parts of the samples, which is the case if the assumed data model described in section 2 is correct. Secondly, scaling of the data before bias correction is allowed as long as the scaling parameters are independent from the bias generating process.

So does this algorithm solve the problems of naive BCIV? The first three steps have no effect on the discriminative ordering of the null space. In step 4, both matrices can be considered to be scaled and rotated, so $\operatorname{argmin}\frac{\boldsymbol{w}^\mathrm{T}\cdot\boldsymbol{\Sigma}_\mathrm{w}\cdot\boldsymbol{w}}{\boldsymbol{w}^\mathrm{T}\cdot\boldsymbol{\Sigma}_\mathrm{b}\cdot\boldsymbol{w}}$ remains unchanged and therefore the order of the basis vectors according to their discriminative ability remains unchanged.

In step 6, bias correction is applied, which is order preserving: the order of the between eigenvalues stay the same. Since the within class matrix is white, the ordering of the eigenvectors according to their discriminating capacity remains the same as well. The transform back to the original space in step 8 has no effect on the discriminating ratio, similar to step 4. So, the null space is still the least discriminating subspace, and the problems of the naive BCIV are solved.

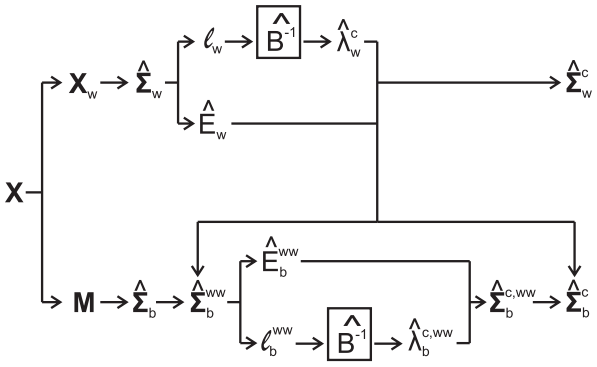


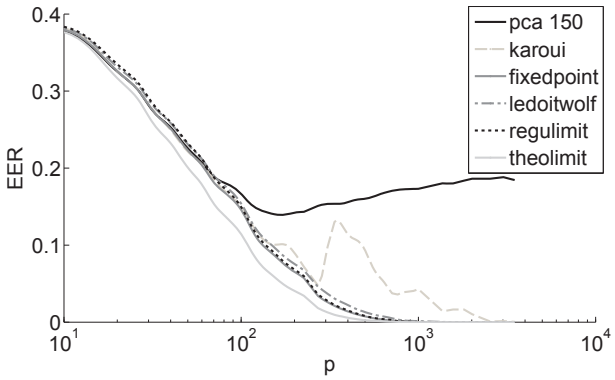
Fig. 8. Eigenwise BCIV schematically.

8 Bias correction in verification approaches comparison

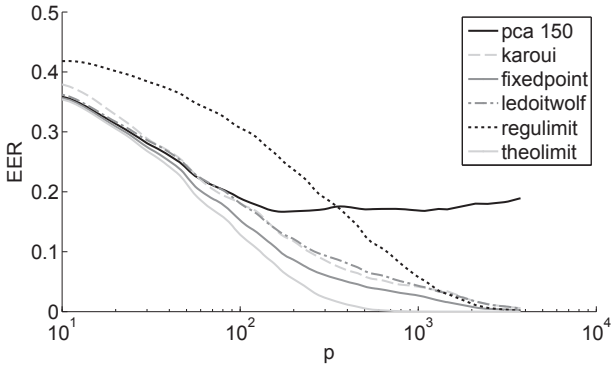
In section 4 we presented an experiment which showed that if eigenvalue bias is the dominant factor in SOS estimation, PCA dimensionality reduction is outperformed by regularization limit for even modest values of p. In the sections afterwards we presented explanations of these results based on theoretical and experimental considerations and we suggested some improvements of the SOS estimation process. We now repeat the experiment and include the improvements.

We repeat the experiment of section 4 but we only use the PCA method with dimensionality reduction to a fixed number of 150 components and compare it not only to the theoretical limit and the regularisation limit, but also with eigenwise BCIV based on the 3 bias correction methods presented in section 6.1: the karoui correction, the Ledoit Wolf correction and the fixed point correction. For the fixed point correction, the smoothing factor s has to be set. Since there is no pre-described method on how to set s, we determined a relation for s with p experimentally, which turned out to be $s \approx \min(0.3, 0.006 \cdot p)$.

Figure 9 shows the resulting EER curves for the different methods. Again, the PCA method is outperformed by the regularization limit for even modest p values, but it is also outperformed by the other methods. Moreover, the eigenwise corrections are not limited to the apparent minimum of the PCA limit, nor do their EER curves show a dip.



(a) All total eigenvalues equal



(b) Exponential total eigenvalue distribution

Fig. 9. EER versus p experiments with different eigenvalue distributions.

The correction methods do seem to match the ideal behaviour of a transition from close to the theoretical limit for small p to close to the regularisation limit for large p. However, they end up slightly above the regularisation limit, so improvement is still possible.

The eigenwise correction based on the fixed point method outperforms the other methods for almost all values of p, but it is also the most complex one. The Ledoit Wolf method already gives a large improvement over the classical PCA dimensionality reduction, although its complexity and processing time is limited. Note that it outperforms the regularisation limit for values of p between 100 and 500 in the exponential total eigenvalues configuration. This means it would

outperform many of the other regularisation methods as well, since they are equal to the regularisation limit if p is larger than N used in the estimation.

Furthermore, it should be noted that the Karoui correction still leaves some of the eigenvalues equal to zero for p above 100. We solved this problem by setting these values equal to the smallest non zero valued eigenvalue estimate, but as can be seen in the uniform total configuration (figure 9a), the almost zero valued eigenvalues still deteriorate the results.

9 Conclusion and Discussion

We studied the effect of increasing the dimensionality p of training data while keeping the number of observations N fixed under conditions that the sample eigenvalue bias becomes the dominant problem in the estimation results. From a good estimator it is expected that adding dimensions results in an estimate which is at least as good as the estimate based on the original data if the added dimensions are of the same quality, otherwise the estimator could simply be improved by ignoring the added dimensions.

The classical PCA dimensionality reduction technique does not display this property if the added space is equally informing. Using GSA we showed that this is because if the eigenvalues do not fall off quick enough, the SOS estimation gets more and more influenced by random fluctuations in the data instead of the actual SOS of the data generating process, resulting in biased sample eigenvalues and misaligned sample eigenvectors. Based on these analysis we suggested several improvements for SOS estimation.

We also argued that estimators should converge from the population SOS ($p << N$) to a regularisation limit ($p >> N$). Eigenwise BCIV has such characteristics, although it is slightly worse than the regularisation limit for $p >> N$. At lower complexity and resources cost, the Ledoit Wolf correction already showed a large improvement compared to the classical PCA dimensionality reduction method.

If the eigenvalues do fall off fast enough though, after a certain p the added dimensions have a similar effect as adding a null space, and PCA dimensionality reduction can efficiently solve singularity issues.

We do not claim that dimension reduction has no use in practice, because these results are obtained with data closely fitting the model. Although it was shown in [36] that even deviations that can be modeled as an random matrix addition to the covariance estimate can still be analysed with random matrix theory, in [47] we showed that other deviations from the model can distort the estimation considerably and can easily explain the 1 over f characteristic [26] of facial image data. Bias correction actually increases the error rates in those cases. Real data will most likely contain these kinds of deviations. Moreover, real data sets also contain measurement errors, which require robust estimation techniques as presented in for example [11], [48]. Also incorrect sampling of some of the involved classes can lead to a eigenvalue bias as well, as was shown in [9]. Therefore, BCIV is not as effective if applied to real data and dimensionality reduction methods can lead to performance improvements [26]. Furthermore, prior information on the eigenvalue distribution can be used for specific estimation problems, like is done in [26] for 1 over f like distribution.

However, if the data model is correct, then the analyses give a lower limit to the errors: we assumed that the added variants contained similar discriminative capacity as the original data. If variants are added which have a smaller discriminative capacity, then the average discriminative capacity decreases and error rates reduction is not guaranteed. So even if eigenwise BCIV is applied, adding variants by e.g. increased image resolution might still increase error rates.

References

[1] X. Tan and B. Triggs, "Fusing gabor and lbp feature sets for kernel-based face recognition," in *Proc. of the 3rd int. conf. AMFG*, ser. AMFG'07, 2007, pp. 235–249.

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

[3] X. Mestre, "Estimating the eigenvalues and associated subspaces of correlation matrices from a small number of observations," in *Proc. of the 2nd Int. Symp. on Communications, Control and Signal Processing*, Marrakech, 2006.

[4] R. D. Uriarte and S. A. de Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, pp. 3+, 2006.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*, 2nd ed. Wiley-Interscience, Nov. 2001.

[6] G. V. Trunk, "A problem of dimensionality: A simple example," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 1, pp. 306–307, 1979.

[7] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, 2000.

[8] R. E. Bellman, *Adaptive control processes - A guided tour*. Princeton, New Jersey, U.S.A.: Princeton University Press, 1961.

[9] X. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 931–937,, 2009.

[10] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: The approach based on influence functions*, ser. Probability and mathematical statistics. John Wiley and Sons, 1986.

[11] M. Hubert, P. J. Rousseeuw, and K. van den Branden, "ROBPCA: a new approach to robust principal component analysis," *Technometrics*, vol. 47, pp. 64–79, 2005.

[12] A. J. Hendrikse, R. N. J. Veldhuis, and L. J. Spreeuwers, "The effect of position sources on estimated eigenvalues in intensity modeled data," in *31th Symp. on Inform. Theory in the Benelux*, 2010, pp. 105–112.

[13] 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.

[14] 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.

[15] J. Särelä and R. Vigário, "Overlearning in marginal distribution-based ica: analysis and solutions," *The Journal of Machine Learning Research*, vol. 4, no. 7-8, pp. 1447–1469, 2004.

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

[17] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 19, no. 7, pp. 711–720, 1997.
[18] X. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Proc. Mag.*, vol. 28, no. 2, pp. 16–26, 2011.
[19] 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.
[20] 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.
[21] S. Srivatava, "Distribution-based bayesian minimum expected risk for discriminant analysis," *IEEE International Symposium on Information Theory 2006*, pp. 2294–2298, July 2006.
[22] Z. D. Bai and H. Saranadasa, "Effect of high dimension: By an example of a two sample problem," in *Statistica Sinica*, no. 6. National Sun Yat-sen University, 1996, pp. 311–329.
[23] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.
[24] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, California, USA: Duxbury, Thomson Learning, 2002.
[25] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., 1990.
[26] 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.
[27] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Phil. Trans. R. Soc. London*, vol. 231, no. 694-706, pp. 289–337, 1933.
[28] M. Soltane, N. Doghmane, and N. Guersi, "Face and speech based multi-modal biometric authentication," *International Journal of Advanced Science and Technology*, vol. 21, August 2010.
[29] D. Middleton, *An Introduction to Statistical Communication Theory*. McGraw-Hill, 1960.
[30] E. Jaynes, "On the rationale of maximum entropy methods," in *Proceedings of the IEEE, Special issue on spectral estimation*, vol. 70, 1982, pp. 939–952.
[31] D. Tolhurst, Y. Tadmor, and T. Chao, "The amplitude spectra of natural images," *Ophthalmic and Physiological Optics*, 1992.
[32] S. A. R. P. Millane and W. H. Hsiao, "Scaling and power spectra of natural images," in *Proceedings of Image and Vision Computing New Zealand 2003*, 2003.
[33] 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.
[34] D. Paul, "Asymptotics of the leading sample eigenvalues for a spiked covariance model," Dep. of Statistics, Stanford University, Tech. Rep., 2004.
[35] I. M. Johnstone, "On the distribution of the largest principle component," Dep. of Statistics, Stanford, Tech. Rep., 2000.
[36] G. Pan, "Strong convergence of the empirical distribution of eigenvalues of sample covariance matrices with a perturbation matrix," *J. Multivar. Anal.*, vol. 101, pp. 1330–1338, July 2010.
[37] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "Eigenvalue correction results in face recognition," in *29th Symp. on Inform. Theory in the Benelux*, 2008, pp. 27–35.
[38] B. B. Chen and G. M. Pan, "Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero," *Accepted, Bernoulli*, 2011.
[39] A. Hendrikse, R. Veldhuis, and L. Spreeuwers, "Improved variance estimation along sample eigenvectors," in *Proc. of the 30th Symp. on Information Theory in the Benelux*, 2009, pp. 25–32.
[40] O. Ledoit and S. Péché, "Eigenvectors of some large sample covariance matrices ensembles," Institute for Empirical Research in Economics, Tech. Rep. iewwp407, Mar. 2009.
[41] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "Notes on second order statistics in verification," University of Twente, Enschede, the Netherlands, Tech. Rep., 2011.
[42] N. El Karoui, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *Annals of Statistics*, vol. 36, no. 6, pp. 2757–2790, 2008.
[43] A. J. Hendrikse, R. N. J. Veldhuis, and L. J. Spreeuwers, "Smooth eigenvalue estimation," University of Twente, Enschede, the Netherlands, Tech. Rep., 2011.
[44] A. J. Hendrikse, R. N. J. Veldhuis, L. J. Spreeuwers, and A. M. Bazen, "Analysis of eigenvalue correction applied to biometrics," in *Advances in Biometrics, Alghero, Italy*, ser. Lecture Notes in Computer Science, vol. 5558/2009. Berlin / Heidelberg: Springer Verlag, June 2009, pp. 189–198.
[45] A. Hendrikse, R. Veldhuis, and L. Spreeuwers, "Verification under increasing dimensionality," *Pattern Recognition, International Conference on*, pp. 589–592, 2010.
[46] X. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition," *Electronics Letters*, vol. 42, no. 19, pp. 1089–1090, 2006.
[47] A. J. Hendrikse, R. N. J. Veldhuis, and L. J. Spreeuwers, "The effect of position sources on estimated eigenvalues in intensity modeled data," in *Thirty-first Symposium on Information Theory in the Benelux*, 2010, pp. 105–112.
[48] S. Serneels and T. Verdonck, "Principal component analysis for data containing outliers and missing elements," *Comput. Stat. Data Anal.*, vol. 52, no. 3, pp. 1712–1727, 2008.

Anne Hendrikse Anne Hendrikse studied Electrical Engineering at the University of Twente, The Netherlands from 2001 until 2007. The subject of his MSc thesis is FRICA: Face Recognition with Independent Component Analysis. He is currently pursuing a PhD degree at the SAS group of the Faculty for Electrical Engineering, Mathematics and Informatics at the University of Twente. His thesis subject is second order statistics estimation in face recognition.

Raymond Veldhuis Raymond Veldhuis graduated from The University of Twente, The Netherlands in 1981. From 1982 until 1992 he worked as a researcher at Philips Research Laboratories in Eindhoven in various areas of digital signal processing. In 1988 he received the PhD degree from Nijmegen University on a thesis entitled Adaptive Restoration of Lost Samples in Discrete-Time Signals and Digital Images. From 1992 until 2001 he worked at the IPO (Institute of Perception Research) Eindhoven. Raymond Veldhuis is now an associate professor at The University of Twente, working in the fields of biometrics, pattern recognition and signal processing and leading a research theme in the area of biometrics.

Luuk Spreeuwers Luuk Spreeuwers studied Electrical Engineering at the University of Twente, Netherlands from 1982-1988. The subject of his MSc thesis was Neural Networks in adaptive Control. In 1992 he obtained his PhD from the University of Twente. The title of his PhD-thesis is: Image Filtering with Neural Networks. Applications and Performance Evaluation. Subsequently Luuk Spreeuwers worked as a postdoc at the International Institute for Aerospace and Earth Sciences (ITC) in Enschede, Netherlands, the University of Twente in a SION project on 3-D image analysis of aerial image sequences and in Budapest at the Hungarian Academy of Sciences on a 3-D textures ERCIM project. From 1997-1999 he worked for an Amarican-Hungarian company: Mindmaker, Ltd. in Budapest as senior researcher in the area of image processing. From 1999-2005 Luuk Spreeuwers worked on 3-D modeling and segmentation of the left ventricle of the human heart in MRI at the Image Sciences Institute of the University Medical Center in Utrecht, the Netherlands. Luuk Spreeuwers is a member of the Dutch Society for Pattern Recognition and Image Processing (NVPHBV).

Currently he investigates face recognition in surveillance applications at the Systems and Signals group of the Faculty for Electrical Engineering, Mathematics and Informatics at the University of Twente. His main interests are model-based image processing.