SOS, LOST IN A HIGH DIMENSIONAL SPACE

Anne Hendrikse

De promotiecommissie:

voorzitter en secretaris:	
Prof.dr.ir. A.J. Mouthaan	Universiteit Twente
promotor:	
Prof.dr.ir. C.H. Slump	Universiteit Twente
assistent promotors:	
dr.ir. R.N.J. Veldhuis	Universiteit Twente
dr.ir. L.J. Spreeuwers	Universiteit Twente
referent:	
dr.ir. A.M. Bazen	Uniqkey Biometrics
leden:	
Prof.dr. A. Stein	Universiteit Twente
Prof.dr. H.J. Zwart	Universiteit Twente
Prof.dr. M. Debbah	Supélec, France
Prof.dr. C.A.J. Klaassen	Universiteit van Amsterdam

UNIVERSITY OF TWENTE.

Signals & Systems

Signals & Systems group
P.O. Box 217, 7500 AE
Enschede, the Netherlands

CTIT

CTIT Ph.D. Thesis Series No. 11-223
Centre for Telematics and Information Technology
P.O. Box 217, 7500 AE
Enschede, The Netherlands

Print: Wöhrmann Print Service
Typesetting: LaTeX2e

ISSN 1381-3617 (CTIT Ph.D. Thesis Series No. 11-223)
ISBN 978-90-365-3367-6

SOS, LOST IN A HIGH DIMENSIONAL SPACE

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 1 juni 2012 om 14.45 uur

door

Antonie Johannes Hendrikse
geboren op 2 maart 1983
te Scherpenzeel

Dit proefschrift is goedgekeurd door

De promotor:	Prof.dr.ir. C.H. Slump
De assistent promotors:	dr.ir. R.N.J. Veldhuis
	dr.ir. L.J. Spreeuwers

Samenvatting

Gezichtsherkenningsmethoden gebaseerd op Principle Component Analysis (PCA), bekend als de eigenface methode indien toegepast op gezichtsdata, hebben lange tijd behoord tot de best presterende gezichtsherkenningsmethoden en dienen ook nu nog vaak als referentiemethoden. Eén van de bekendere varianten is de combinatie van eigenfaces met een log likelihood ratio gebaseerde afstandsmaat. De laatste jaren zijn echter geen grote verbeteringen in deze methode gevonden en wordt deze met enige regelmaat overtroffen door andere methoden. Dit ondanks dat bewezen is dat onder bepaalde condities de methode optimaal is.

Eén van de kenmerken van deze methode is dat de prestaties nauwelijks verbeteren met de hogere resolutie foto's die tegenwoordig beschikbaar zijn, terwijl andere algoritmen inmiddels beter presteren op (delen van) deze data. Een verklaring hiervoor ontbreekt echter tot nu toe.

Eén van de effecten die een hogere resolutie heeft, is dat de data een hogere dimensionaliteit krijgen. Het is al langer bekend dat het schatten van tweede orde statistiek, een belangrijk onderdeel van de eigenface methode, in hoog dimensionale data onnauwkeurig kan worden. Dit komt vooral naar voren in de maxima in variantie: deze worden steeds meer bepaald door de random structuren in de specifieke dataset dan door de onderliggende proces parameters. Dit is vooral merkbaar in de schatting van de eigenwaarden: deze zijn significant gebiast.

De bias van een schatter is een niet willekeurige eigenschap en de schatting kan dus worden gecorrigeerd voor deze bias. Een methode ontwikkeld door Karoui biedt op dit moment de beste prestaties, maar helaas is de toepasbaarheid van deze methode op gezichtsdata vrij slecht. Een groot deel van onze studie is daarom gericht op het ontwikkelen van een methode die de bias kan corrigeren uit een eigenwaarde schatting die kan worden gebruikt in de eigenface methode. Hiermee kunnen we tevens onderzoeken of deze verstoring in de tweede orde statistiek schatting kan verklaren waarom de prestaties van de eigenface gebaseerde methoden achterblijven als hoog resolutie beelden worden gebruikt.

De methode die we hiervoor ontwikkeld hebben, is de fixed point eigenvalue correctie. Deze methode is beter toepasbaar op data met de karakteristieken van gezichtsdata dan de Karoui methode, wat onder andere is aangetoond door tests met synthetische data.

De eigenface methode richt zich vooral op het schatten van de statistieken van

de variaties van alle gezichten. Voor herkenning is het echter vooral van belang dat we de structuren vinden in het gezicht die veel variatie vertonen wanneer we verschillende personen vergelijken en maar weinig variatie vertonen als we naar de foto's van één persoon kijken. Dit houdt dus de schatting van de statistieken van 2 distributies in: de variantie van foto's van verschillende personen en de variatie van foto's van dezelfde persoon.

Beide statistieken zijn verstoord door de hoge dimensionaliteit ten opzichte van het beschikbare aantal samples en beide schattingen dienen dus gecorrigeerd te worden. Een naïeve manier om dit te doen is door beide schattingen apart te corrigeren. Echter, doordat correcties apart worden uitgevoerd, kan de verhouding tussen de statistieken van de variatie tussen samples van verschillende personen en de statistieken van de variatie in samples van dezelfde persoon veel groter worden geschat voor sommige structuren dan dat de data ondersteund, wat de herkenning behoorlijk verslechtert. Wij hebben methode ontwikkeld, de eigenwijs correctie, die bij de correctie van de statistieken van de variatie schatting van samples van verschillende personen rekening houdt met de correctie uitgevoerd op de variatie schatting van samples van dezelfde persoon.

Op synthetische data met de karakteristieken van gezichtsdata leveren deze correcties een aanzienlijke verbetering in de herkenningsresultaten op. Als er echter echte gezichtsdata wordt gebruikt, presteert de oude methode om effecten van de hoge dimensies te beperken (PCA dimensie reductie) beter. Het grote verschil tussen de echte gezichtsdata en de synthetische data is echter dat bij de echte data wordt aangenomen dat het impliciete model behorende bij tweede orde statistiek schatting, het vaste positie intensiteitsmodel, een goede beschrijving biedt van de gezichtsdata, terwijl bij synthetische data de data per definitie voldoet aan dit model.

In het laatste deel van het onderzoek hebben wij ons vooral gericht op een mogelijk ander proces dat voor variaties in de gezichtsdata kan zorgen wat slecht te modelleren is met het vaste positie intensiteitsmodel: het positie bronnen model. Dit model blijkt allereerst te kunnen verklaren waarom PCA dimensie reductie beter presteert op gezichtsdata dan de eigenwijs correctie: grote structuren met weinig beweging kunnen nog steeds redelijk gemodelleerd worden met het intensiteitsmodel, maar kleinere structuren met relatief grotere bewegingen zijn slecht te modelleren met het intensiteitsmodel. In hoog resolutie data komen deze bronnen meer voor en waar eigenwijs correctie de invloed van deze bronnen nog eens versterkt, zorgt PCA dimensionaliteitsreductie daarentegen voor laagdoorlaat filtering waardoor de invloed van deze bronnen verminderd wordt.

Daarnaast kan het positie bronnen model ook een aantal karakteristieken van de geschatte karakteristieken van gezichtsdata verklaren, zoals het hoge aantal bronnen nodig voor de modellering van de data en het 1 over f gedrag van de karakteristiek.

Dat het model relevant is voor de beschrijving van gezichtsdata onderbouwen we nog verder door te laten zien dat we met het positie bronnen model een set van foto's met een bewegend oog meer variatie kunnen verklaren dan met de eerste component van het intensiteitsmodel.

Abstract

Face recognition methods based on Principle Component Analysis (PCA), known as the eigen face method, were some of the best performing methods for some time and are still often used as base line in comparisons. One of the best known variants is the combination of eigenfaces with the log likelihood ratio distance measure. The last few years no large improvements have been made and the method is outperformed by other methods regularly, despite the fact that the method has been proven to be optimal under certain conditions.

One of the characteristics of the eigen face method is that its performance hardly increases if nowadays available higher resolution images are used, while other algorithms perform considerably better with (parts of) this data. An explanation for this has been missing until now.

One effect the use of higher resolution images has, is that the data has a higher dimensionality. It has been known for some time that the estimation of Second Order Statistics (SOS), an important part of the eigen face method, becomes increasingly inaccurate with higher dimensionality. This is because the variance maxima are increasingly determined by random structures in the actual data set instead of the data generating process parameters. This is especially noticeable in the eigenvalue estimates: they are biased.

The bias of an estimator is not random; therefore, the estimate can be corrected for this bias. A correction method developed by Karoui has the best performance at the moment, but unfortunately it is very difficult to apply the method to facial image data. A significant part of our study therefore focussed on developing a correction method which can be used to correct the eigenvalue estimates in the eigen face method. With the correction of the bias in estimated eigenvalues from facial data we can also study if the distortion in the SOS estimates is the reason why the performance of the eigen face method is not improved compared to other methods if high resolution images are used.

The method we developed for this purpose is the fixed point bias correction. This method is better suited for data with the characteristics of facial data compared to the Karoui method, which we proved by tests with synthetic data.

The eigenface method focusses mainly on estimating the variations of all the faces. However, for recognition it is more important to find the structures in the face which have a large variation between photos of different persons while they have a low variation between photos of the same person. This involves the estimation of

statistics of two distributions: the variation of photos of different persons and the variation of photos of the same person.

Both of these statistics are distorted by the high dimensionality compared to the number of samples available for the estimation, so both estimates will have to be corrected. A naive approach to do this correction is to correct both distributions independently. However, because the correction is done independently on both distribution estimates, the correction itself can lead to much larger ratios between the statistics of the variations of the samples from different persons over the statistics of the variations of the samples from the same person then is supported by the data, which reduces the verification performance significantly. We developed a method, the eigenwise correction, which takes the corrections on the estimated statistics of the variation of samples from the same person into account in the correction of the estimated statistics of the variation between samples of different persons.

If synthetic data is used, then these corrections provide a significant performance increase in verification results. However, if real facial data is used, then the classic method of PCA dimensionality reduction performs considerably better. The main difference between real facial data and synthetic data is that with real facial data it is assumed that the implicit model corresponding tot the Second Order Statistics estimation, the fixed position intensity model, gives a good description of facial data, while synthetic data adheres to this model by definition.

In the last part of our study we focussed on another process that can lead to variations in facial data which are poorly modeled with the fixed position intensity model: the position sources model. This model can firstly explain why PCA dimensionality reduction gives a higher performance on facial image data than eigenwise correction: large structures with relative little movement can still be reasonable modeled with the intensity model, but at higher resolution, small objects with relatively large movements are present in the data, which are poorly modelled with the intensity model. Eigenwise correction increases the influence of these objects, while PCA dimensionality reduction performs a low pass filtering, reducing the influence of these objects.

Secondly, with the position sources model we can also explain a number of characteristics of the estimated eigenvalues of facial image data, such as the high number of sources required in the modeling of the data and the 1 over f behaviour of the eigenvalue curve.

That the position sources model is relevant for modeling of facial data is further supported by our experiment in which we model a set of images with a moving eye, where we showed that using the position sources model we could explain more variance than the first component of PCA did.

Contents

Chapter 1

Introduction

1.1 Pattern Recognition and face recognition

Face recognition has been an active research area for the past 30 years. Currently over 70 research groups world wide are actively studying the topic [1]. If face recognition is considered as being a particular form of computer vision, then the total effort spent on the subject is even much larger. So is the problem at hand such a difficult problem that it requires this much effort to solve?

An indication of the required effort might be deduced from other areas since the vision problem is not limited to a select group of researchers. In nature, many animals allocate considerable resources either to the processing of visual information or to the frustration of the processing done by other animals. If fact, some animals depend for protection solely on the difficult task of tracking one animal in a group of animals moving crisscross for protection ([2], page 150), despite other difficulties living in a group brings up such as hunger and thirst.

For humans, one significant visual task is face recognition, which is already developing in 6 months old infants, well before they are able to speak [3]. This suggests that face recognition can be learned without supervision. In fact, the ability to detect faces and interpret these faces is so important for humans that Clark claimed that the purpose of female breasts is to enable babies to study their mothers face and its expressions [4] during feeding.

These examples show that these visual tasks may be difficult and require a considerable amount of resources, but still they can be learned and this learning task can be done without supervision. However, little is known what learning algorithms are used in nature, again more specifically, how humans process faces. In this thesis we study the automation of learning and performing these visual tasks. We specifically study the verification problem in which an identity claim is either accepted or rejected based on the comparison of a photo of the claiming person with previous measurements stored of the claimed identity.

1.2 Data driven versus model driven approaches

As noted before, we are not the first ones to study the automation of face recognition. Two approaches were distinguished in the early days of automated face recognition: a feature based approach and a template matching approach [5]. Since the term template is now in use in a different way so it also applies to the feature based approach, we rephrase this distinction to data driven approaches versus model driven approaches. In the model driven approach a very precise face model is used. One example method locates specific facial features in the images and then a vector is constructed by collecting several measures of for example the distances between these features.

In data driven approaches the main idea is to assume a very generic model and determine most of the structure of the data from a training set. In [5] this is done by first aligning face images based on a few detected features and then determine the correlation between the image from the claiming person and compare it to a known image from the claimed identity.

A large difference between the data driven approach and the model driven approach is that the later can never use any additional information available in the images other than already encoded by the model while the former is at least in theory capable of using this information. Of course, the data driven approach does require a training set to find the structure in the data, while the model driven approach does not require any training. The model driven approach usually requires more effort to implement, since all the model specifics have to be implemented.

In practice most algorithms cannot be classified strictly in these two categories; most methods will fall somewhere in between these two extremes. For example the data driven approach in [5] requires several preprocessing steps, based on detection of several facial features. Correlation itself has an implicit data model as we will describe later on.

In the comparison in [5] it was found that the data driven approach outperformed the model driven approach. At the same time the eigenfaces method was introduced [6], a method based on PCA, which has become a well known method and is often used as baseline method in the comparison of recognition methods. This eigenface method is another example of a data driven method.

Since data driven methods determine the structure from the data, it might seem at first sight that the more information provided, the better the results from such methods would be, or at least it should not hurt the results. We will show that this is not the case in biometrics, however.

1.3 Increasing image resolution

One way to increase the information in a training set of images, is to increase the resolution of the images. Beside the low frequency information already available in the lower resolution images, the data driven methods now also have the high frequency content available.

The resolution of the training data has increased significantly in the past, see for example the FRGC2 database. This is due to an increase in camera resolution. For the future it is expected that resolution continues to increase although maybe not as fast as until now due to a change of bottleneck from transistors per square inch to lens properties [7].

However, the eigenface method, introduced in the previous section as an example of a data driven method, does not show an increasing performance with increasing resolution [8] and may even show a drop in performance. So our main research question is:

- Why does providing additional information not always help PCA based methods such as the eigen face method to improve their performance or even damage it and how can we overcome this limitation?

A general answer to this question is given by the fact that the number of samples available in the training set did not keep up with the trend of the increasing resolution, since the number of samples is determined by human effort, and therefore quite costly to increase. Moreover, there are only 7 billion persons or so on the planet, so any training set would be limited to this number. If the dimensionality becomes much higher than the number of training samples, SOS estimators exhibit an overtraining effect: the structure they determine from the data becomes increasingly more based on random variations in the training data instead of the structure of the data generating process.

As it turns out, this overtraining effect can be described in quite some detail, but before we can get to that we first have to go into more detail on how the eigenface method works.

1.4 PCA method as data driven approach

1.4.1 Preprocessing

The implementation of the eigenface method we will use is very similar to the template method of [5]. In the preprocessing stage we assume that the position of the eyes and the mouth are determined. In several databases these coordinates are provided, like for example the FRGC2 database, so we can use those.

With an affine transformation we transform the image such that the position of the eyes and the mouth are at predetermined coordinates to reduce the effect of any movement and rotation of the head and change in distance between camera and face. In the next step pixels in a region of interest are extracted from the face image. The region of interest contains as large a facial area as possible, but neglects facial areas which are highly variable while not containing much reliable identity information, such as the hair area and the borders of the face. The p pixels in the region of interest are concatenated to form one column vector x.

1.4.2 Finding structure in the form of Second Order Statistics

A training set of N images leads to a set of p dimensional samples, which form a cloud of N points in a p dimensional space. As an illustration we show in figure 1.1 a two dimensional synthetic data set. From this point cloud we have to determine some structure of the data.

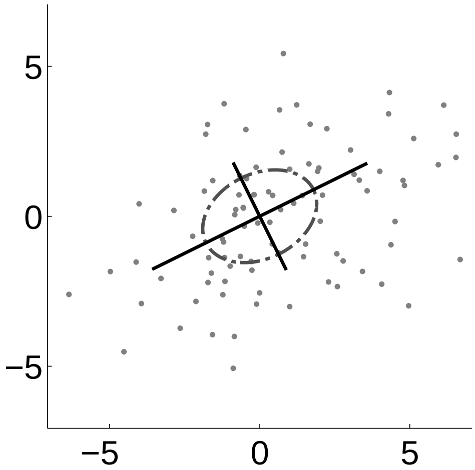


Figure 1.1: Visual interpretation of the eigenvectors and the eigenvalues. The dots represent the samples from some data set. The direction of the dark lines indicate the direction of the eigenvectors, the length of the lines represent the eigenvalues, which are equal to the variance of the data projected on the corresponding eigenvector.

The first step is to model $\boldsymbol{x}$ as a multivariate random variable. The structure of the data is then captured by the distribution of this random variable. A rough description of the structure of the data is by determining the mean and the variations of the data, like indicated by the dash dotted oval.

These two attributes are the first order statistic and Second Order Statistics (SOS) of the data. The first order statistic is known as the mean and is given by:

$$\boldsymbol{\mu} = \mathcal{E}\left\{\boldsymbol{x}\right\} \tag{1.1}$$

The SOS are described by a covariance matrix:

$$\boldsymbol{\Sigma} = \mathcal{E}\left\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\right\} \tag{1.2}$$

With increasing image resolution, the number of pixels of which the image is composed increases and therefore $\boldsymbol{x}$ increases in dimensionality. This might be a disadvantage for several reasons: it causes longer processing times, it requires more storage space and overview of the structure/visual representation of the data is in general more difficult for human understanding if the dimensionality is much larger than 3. Therefore compression of the data is desired. One method of compression is to project the data onto a subspace which still contains most of the variance of the data.

This subspace can be found using PCA. PCA requires the decomposition of the covariance matrix:

$$\boldsymbol{\Sigma} = \boldsymbol{E}\,\boldsymbol{D}\,\boldsymbol{E}^{\mathrm{T}} \tag{1.3}$$

where $\boldsymbol{E}$ is an orthogonal matrix. Each column of $\boldsymbol{E}$ is an eigenvector. $\boldsymbol{D}$ is a diagonal matrix where diagonal element $\boldsymbol{D}_{ii}$ is an eigenvalue corresponding eigenvector $\boldsymbol{E}_{:,i}$, the i^{th} column of $\boldsymbol{E}$. The subspace containing the largest variance of the data is spanned by the eigenvectors corresponding to the largest eigenvalues.

To give a visual interpretation of these eigenvalues and eigenvectors, again consider the 2 dimensional scatter plot in figure 1.1. The dots in this figure represent samples of some data set. The eigenvectors corresponding to the distribution of the data set are represented by the direction of the dark lines. The length of the lines indicate the eigenvalues corresponding to the eigenvectors.

Figure 1.1 shows that the largest eigenvalue represents the largest variance of the data and the corresponding eigenvector gives the direction in which this variance occurs. In cases with more than two dimensions, the second largest eigenvalue gives the largest variance if the first eigenvector is removed from the data and so on.

In mathematics, this process can be described by repetition of the following maximization:

$$\lambda_i \quad = \quad \max \left\{ \frac{\alpha_i^{\mathrm{T}}\,\Sigma\,\alpha_i}{\alpha_i^{\mathrm{T}}\,\alpha_i} \right\} \Bigg|_{|\alpha_i|=1} \tag{1.4}$$

where in each iteration i α_i is a vector in the subspace of the original p dimensional space, orthogonal to the previously found α's, or $\alpha_i^{\mathrm{T}} \cdot \alpha_k = 0 \, \forall \, k = 1 \ldots (i-1)$.

Using SOS implies that a certain model is used for the data generating process. It assumes that images can be described by a mean image to which a weighted sum of a set of base images is added. The information is actually in the weights with which the base images are added. The information sources therefore express themselves in intensity variations at fixed positions in the images. We therefore denote this model by the fixed position intensity sources model. The model is mathematically represented by

$$\boldsymbol{x} = \boldsymbol{B} \cdot \boldsymbol{s} \tag{1.5}$$

Here $\boldsymbol{s}$ is a column vector with every element being one sample from one of the p_s sources. $\boldsymbol{B}$ is a p times p_s matrix where column k determines where and how source k is represented in the image represented by $\boldsymbol{x}$. In other words, if a column of $\boldsymbol{B}$ is reshaped similar to how $\boldsymbol{x}$ has to be reshaped to represent an image, the image resulting from the column of $\boldsymbol{B}$ is a base image and element k of $\boldsymbol{s}$ determines how strongly this base image is represented in image $\boldsymbol{x}$. Under the assumption that the sources are independent and that the columns of $\boldsymbol{B}$ are unitary, $\boldsymbol{E}$ in the decomposition given in equation 1.3 equals $\boldsymbol{B}$. The source signals themselves can be retrieved by $\boldsymbol{s} = \boldsymbol{E}^{\mathrm{T}} \cdot \boldsymbol{x}$.

1.5 Biased eigenvalues

In practice the statistics are unknown, so they have to be estimated from training data. Since in most practical problems the number of samples is limited, variations will occur in the estimates of the variances in all directions. The sample eigenvalues can by determined by finding the directions in which the largest variance in the training set occurs as described for the population eigenvalues in the previous section. Eigenvalue estimation therefore involves maximisation.

If the number of samples is sufficiently large then the largest variances in the training data will occur in a similar direction as the population eigenvectors. However, due to random fluctuations in the data there will be slight variations in the variance estimates, so it is likely that a larger variance occurs just off the population eigenvector. In figure 1.2 we illustrate this by showing the population SOS (solid oval) and two estimates (dash dotted ovals). These variations occur in every direction, so if the dimensionality of the data becomes large, then the effects in each of these directions add up and lead to a large difference in maximum variance. As a result, the largest sample eigenvalue will most likely be larger than the largest population eigenvalue, and this will be so over many experiments. The sample eigenvalues are therefore a biased estimates of the population eigenvalues.

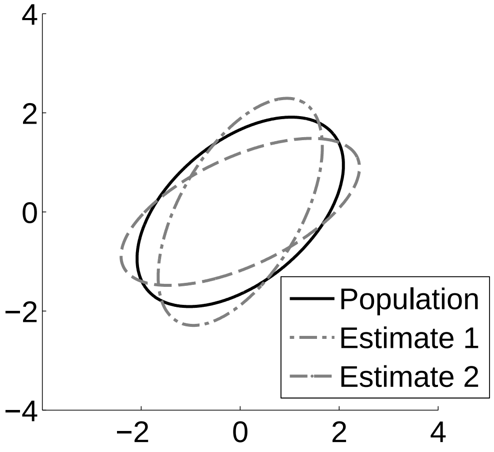


Figure 1.2: Illustration data fluctuations lead to variations in estimates of the SOS. Because of the fluctuations, it is very likely that there is a direction in the neighbourhood of the population eigenvector corresponding to the largest population eigenvalue where the sample estimate is a little larger than the population eigenvalue. Since eigenvalue estimation is a maximization process, the estimate will find this direction and consequently the largest sample eigenvalue will be too large compared to the largest population eigenvalue: the sample eigenvalues are biased.

Since the PCA method depends heavily on the estimation of SOS, the bias in the eigenvalue estimates might be the reason why the PCA method does not profit from an increased resolution.

The bias is a direct result of the variations in the variance estimates. It will therefore be almost negligible if the number of samples N is large compared to the dimensionality p of the data, but the bias has a increasingly distorting effect on the SOS estimate with increasing dimensionality, until the dimensionality becomes larger than the number of samples.

1.5.1 p < N: only a distortion

If the number of dimensions is smaller than the number of samples available for the SOS estimation, then the eigenvalues are biased, but the estimated covariance matrix is invertible. This is for example needed in determining likelihood ratios.

1.5.2 p > N: Singularity problem

If the number of dimensions is larger than the number of samples, at least $p - N$ sample eigenvalues will be zero valued. The estimated covariance matrix becomes singular and is not invertible. As a result, likelihood ratios, which require an inversion of the covariance matrix, cannot be determined.

The fact that $p - N$ sample eigenvalues are necessarily zero also means that due to the bias some information is lost: the process parameters are p variables, the population eigenvalues, while the estimation only has N free parameters, the sample eigenvalues.

1.5.3 Bias correction

Since the bias can have a significant effect on the SOS estimation, it might be the reason why PCA based biometric methods do not benefit from the increased image resolution. Our first derived research question is therefore:

- What (potential) effects does the sample eigenvalue bias have on verification systems and can these effects be reduced?

We try to answer this question by estimating how severe the bias is in the eigenvalues estimated from facial data and try to determine how strongly this influences the verification results. This can be done by bias correction: because the bias is a non random property of the data, it should be possible to remove it or at least reduce its influence.

1.5.4 Classical ad-hoc solutions for bias reduction

To remove the bias from the sample eigenvalues, several ad-hoc solutions already existed.

1.5.4.1 PCA dimensionality reduction

The first ad-hoc solution commonly applied is the PCA dimensionality reduction. This solution is mainly applied to solve one particular effect of the eigenvalue bias, namely the singularity problem. With PCA dimensionality reduction the sample covariance matrix is decomposed into the sample eigenvalues and sample eigenvectors, after which the data is projected on p_{red} sample eigenvectors corresponding to the largest sample eigenvalues. p_{red} is usually chosen a little lower than the number of non zero valued sample eigenvalues, such that the smallest non zero valued sample eigenvalues are removed as well.

This method can be considered a form of bias correction: as described before, the smallest sample eigenvalues are estimated too small, so they should be corrected to a non zero value. In terms of likelihood calculation the removal of the sample eigenvalues has the same effect as assigning them a infinitely large value, so the sample eigenvalues are indeed increased in value. The largest eigenvalues remain unchanged, however.

1.5.4.2 Regularization

Another quick solution to the singularity problem is to add a small value to all the eigenvalues, thereby preserving the order of the eigenvalues, while non remain zero valued. Over time more sophisticated methods have been proposed, which became known as regularisation. In regularisation, a balance is determined between the sample covariance matrix and an identity matrix, which is scaled by the mean of the sample eigenvalues, $\bar{l}$:

$$\bar{\mathbf{\Sigma}}^{c} = (1-\alpha)\,\hat{\mathbf{\Sigma}} + \alpha\bar{l}\boldsymbol{I} \tag{1.6}$$

where α is known as the regularisation constant and has a value between 0 and 1. In general, if the number of samples is large, then covariance matrix estimate will be accurate and α has to be set close to 0. If the number of samples is rather small, the sample estimate is highly inaccurate and α has to be set close to 1. If α is set to 1, no structure is estimated at all. This limit we denote as the regularisation limit.

The regularisation methods differ in how α is set, but in general a distance measure is defined between the population estimate and the sample estimate and the expected value of this distance is minimized. For example, in [9] the distance is determined by Stein's loss criterion, given in equation 1.7.

$$L_{\text{stein}}\left(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}\right) = \operatorname{tr}\left(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\right) - \log\det\left(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\right) - p \tag{1.7}$$

The criterion to be minimized is the risk given by:

$$R\left(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}\right) = \mathcal{E}\left(L_{\text{stein}}\left(\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}\right) | \mathbf{\Sigma}\right) \tag{1.8}$$

Many different regularisation methods have been derived, see for example [10], [11], [12] and [13]. Most of the regularisation algorithms have an α equal to 1 as soon as the training set contains fewer samples than the dimensionality of the samples and are thus equal to the regularisation limit under this condition. An exception is the method presented in [14], which still finds some structure in the sample estimate, even if the dimensionality of the samples is larger than the number of samples available for training.

1.5.5 Bias correction based on bias descriptions

It would be surprising if the ad hoc solutions would provide the best solution to the complicated bias problem, since they are in no way related to any description of the bias. Therefore it is to be expected that better solutions exist. There are already some descriptions available of the bias and some methods have been developed to remove the bias from the eigenvalues.

To answer the first derived research question we therefore also study the following question: are there adjustments of the sample eigenvalues possible such that the likelihood estimates based on the density estimates are more accurate?

We will show that bias correction can indeed provide significant improvements of estimation accuracy and reduces error rates of verification systems if synthetic data is used. However, if real biometric data is used, error rates either do not change significantly or may even go up. If synthetic data is generated with the same SOS as estimated from biometric data, the results do improve. This might indicate that even though we take SOS based systems as an example of data driven method, where most of the structure is estimated from the data and only a minor part is determined by a data model, the few assumptions on which this fixed position intensity sources model is based are still wrong.

1.6 Data modelling error

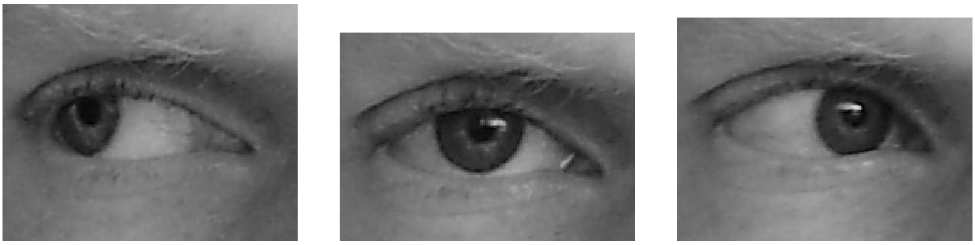
In section 1.4.2 we noted that even though SOS estimation depends largely on training data to determine the structure of data generating process, it still has an implicit model of the data generating process. But since its assumptions are relatively mild, we assumed that although it might not be the most efficient model, it would still be a reasonable approximation of the data generating process and we expected that the bias problem itself is a much larger source of error than any mismatch between the implicit data model of the fixed position intensity sources and the actual data generating process.

But as indicated in the previous section, during our research we found more and more clues that the biased eigenvalues are a minor problem: firstly, we found that bias correction gives significant improvements if applied to synthetic data which adheres to the fixed position intensity sources model, but if bias correction is applied to SOS estimated from facial data, then it does not lead to significant improvements and often it deteriorates the performance. Secondly, the estimated eigenvalue sets of facial data have some remarkable properties, which can hardly be explained by the fixed position intensity sources model combined with bias theory. We will therefore try to find out what the explanation is of the characteristics of the eigenvalue spectra.

To give an explanation of the facial data eigenvalue characteristics, we introduce the concept of position sources. Some of the information in the face is encoded in the position of features instead of their intensities as is assumed by the fixed position intensity sources. So in summary our second research question will be:

- What effect does the presence of position sources in data have on systems based on the fixed position intensity sources model and can it explain the observations made after increasing the image resolution of facial data: the high number of sources estimated, the 1 over f characteristic of the eigenvalue scree plot, the saturation of performance of biometric systems based on SOS estimation and that PCA performs better on real facial data than bias correction?

As an example of such a position source, consider the iris and pupil in the eye. Most of its variations are in its relative position in the eye, depending on the direction in which somebody is looking, as is shown in figure 1.3.



(a) Right

(b) Middle

(c) Left

Figure 1.3: Example of a moving feature in facial images: the iris and pupil

With the introduction of position sources in facial images, we can explain a number of observations we made when estimating SOS: the seemingly infinite number of intensity sources, the 1 over f curve of the eigenvalue plots, and the poor results of bias correction in solving the singularity problem compared to the PCA dimensionality reduction method.

It turns out that position sources can be approximated reasonably well by fixed point intensity sources, if the images are of low resolution, but if the resolution increases, the fixed position intensity source model becomes increasingly inefficient in encoding data containing position sources, and if the model is used in a verification scheme, error rates will go up with increasing resolution. PCA dimensionality reduction effectively performs low pass filtering on the data, therefore making the data fit better to the fixed position intensity model.

It seems that in order to make bias correction useful in biometrics based on facial images, the data should be made more compliant with the intensity sources model either by removing the position sources from the data before doing the SOS estimation or by using different encoding.

1.7 Relations with other disciplines

Before continuing on the more technical details I wish to link the subject at hand to a broader context. In the first section we already noted that the problems with determining structure from image data is not limited to biometrics based on face images, but is a general theme in computer vision. In the following discussion we focus more on the issue of determining structure from data with many parameters and only a few samples. As a result of these many variables or dimensions and only a few training samples, the learned structure becomes more and more based on a random structure in the data rather than on the true parameters of the data generating process.

As we described in section 1.5, the biggest problem in eigenvalue estimation is that it is a maximisation process, so it is affected by random fluctuations. In other fields it is often attempted to predict future events by finding (a combination of) variables with maximum correlation with the quantity to predict. For example, much effort is put into trying to predict stock markets. Many rules of thumb have been introduced, of which the following is a typical example. In the article "De zin en onzin van het januari-effect" [15], the following rule of thumb is studied: if the stock markets indices increased in January, the entire year will be profitable. It is shown that this was not really the case the past 25 years. However, the author continues, if only the first few days are considered, then the prediction of a profitable year is accurate for every year in that data set.

This strategy shows some similarities to the eigenvalue estimation: both methods are maximisation based on covariances. In the stock market prediction, the number of variables is not exactly specified, but the class contains at least 31 variables (any number of days in January) and if group of days is considered as well and maybe part of February is included, the number of variables becomes much larger. Still the training set consists only of 25 samples, so all the biased estimate problems would be present in such a set.

1.7.1 Human decision making

One general rule that seems to follow from the research presented in this thesis is that simply considering all the available variables easily leads to an estimate which is based on a random structure of the data instead of the true structure of the underlying data generating process. Moreover, if the data model is not correct, considering more details will lead to more erroneous estimates. These results seem to support the correctness of some of the strategies humans tend to use when handling information as was discovered in several psychology experiments. For example in [16] it is reported that people often ignore much information in their decision process, even though formal institutions assume that all available information is used.

In sociology it is also remarked that special care has to be taken that the information quality presented in a democracy is of good quality [17], because otherwise the choices of the people will be based on noise variables. This thesis supports the theory that this is not some error of the limited abilities of humans, but it is actually a close to optimal decision making procedure under the given circumstances.

1.7.2 Relation with philosophy

In section 1.5.2 we noted that if the training data consists of fewer samples than the dimensionality of those samples, then the eigenvalue estimation becomes underdetermined: there are multiple population eigenvalue sets leading to the same sample eigenvalue set. In the limit of having many more dimensions than samples, all population eigenvalue sets lead to the same sample eigenvalue set. In philosophy a similar problem is known concerning the choice of theory: given a set of observations, a theory should be chosen from a set of rivalling theories. However, something goes wrong in the choice strategy, since we always have only a limited number of observations available while the number of dimensions, or parameters we try to predict is not fixed, so an infinite number of theories can be constructed.

As an example, consider the fact that up till now the sun has risen every day in my experience. I currently experienced approximately 10,000 sunrises, but does it warrant me that the sun will rise every day? Although my data set seems to support this theory, it also supports the theory that the sun has risen every day until today, but tomorrow the sun will not rise again, or it will rise tomorrow, but after that no more, and so on.

This problem is known as the underdetermination argument in theory choice. The *Strong form of the underdetermination argument* for scientific theories is as follows (page 174 of [18]):

1. For every theory there exists an infinite number of strongly empirically equivalent but incompatible rival theories.

2. If two theories are strongly empirically equivalent then they are *evidentially equivalent*.

3. No evidence can ever support a unique theory more than its strongly empirically equivalent rivals, and theory-choice is therefore radically underdetermined.

It seems therefore that theory choice in philosophy is suffering from the same problems as machine learning: if all variables are considered, then no structure could ever be found from the limited number of examples, without any further prior information.

1.8 Research questions summary

In the remainder of the thesis we will focus on the technical details. In the following chapters we try to answer the question:

- Why does providing additional information not always help PCA based methods such as the eigen face method to improve their performance or even damage it and how can we overcome this limitation?

We study two possible causes in particular: the bias in the sample eigenvalues and a mismatch between the data generating process and the assumed model in SOS estimation. In our study of the bias in the sample eigenvalues, we will particularly focus on the following question:

- What (potential) effects does the sample eigenvalue bias have on verification systems and can these effects be reduced?

Moreover, we are interested whether eigenvalue bias can explain the remarkable observation that adding more information to the system does not improve verification rates and it may actually deteriorate the results.

The results of the study on the eigenvalue bias give several clues that there might be something more fundamentally wrong with the application of SOS estimation in facial data structure discovery. Therefore we study another data model we denote by position sources in chapter 4 and try to answer the derived research question:

- What effect does the presence of position sources in data have on systems based on the fixed position intensity sources model and can it explain the observations made after increasing the image resolution of facial data: the high number of sources estimated, the 1 over f characteristic of the eigenvalue scree plot, the saturation of performance of biometric systems based on SOS estimation and that PCA performs better on real facial data than bias correction?

1.9 Contributions

During our study we developed several new insights and methods. Here we present a summary of these contributions.

In the single distribution SOS estimation we made the following contributions:

- Smooth eigenvalue estimation: the available description of the eigenvalue bias only applies in the limit of p and N both infinitely large. Smooth estimation provides an approach on how to use this bias description in practical cases, where N and p are finite.
- Bootstrap correction. A method to estimate the population eigenvalues given the sample eigenvalues using synthetic data to determine the sample eigenvalues of the current population eigenvalue estimate. Based on these

sample eigenvalues the population eigenvalues estimate is updated. By repeating these steps, an unbiased estimate of the population eigenvalues is obtained.

- Fixed point sample eigenvalue estimation. This is an alternative to determine the sample eigenvalue density if the population eigenvalues are given instead of estimating the sample eigenvalues from synthetic data generated using the given population eigenvalues. One application of this method is given in the next item.
- Fixed point correction. By combining the fixed point sample eigenvalue estimator from the previous item with an iterative update schema similar to the bootstrap method, we developed a new bias correction method.
- Isotonic tree updating. One aspect of both the bootstrap correction and the fixed point correction is that during the update phase the order based on the relative value of the eigenvalue estimates should be preserved. The isotonic tree algorithm is highly parallel programmable and treats all eigenvalues equally in contrast to existing solutions.
- Variance correction. This method can be used to find the actual variance of the data along the sample eigenvectors after successful bias correction.

In the verification setting, where two distributions are of importance, we made three major contributions:

- Within crosstalk on between estimate. We showed that the between class SOS estimates are in fact a mixture of the between class SOS and the within class SOS.
- Limit behaviour of PCA. During our analysis of the effect of the eigenvalue bias on verification systems, we discovered that the classical solution of PCA dimensionality reduction for the singularity problem provides no solution at all if the dimensionality of the data becomes very large. Moreover, in the single distribution case, PCA dimensionality reduction becomes simply a random subspace selector if the dimensionality becomes very large.
- Eigenwise correction. When applying bias correction in the within class distribution estimate and the between class distribution estimate without considering the ratio between the variances of these two distributions, strange effects in the verification can occur: the estimated discriminative capacity of the null space can become arbitrarily large. The eigenwise correction solves this problem.

Finally we made a contribution on the conceptual level:

- Position sources. The large difference we found between experiments with synthetic data and experiments with real facial data made us doubt how accurate the implicit data model used in PCA based methods describes the facial data generation process. We therefore introduced the position sources model. In this model the information is encoded in the position of features rather than in their intensities. We showed that position sources can severely disturb the SOS estimation using the fixed position intensity sources and provide an explanation of several of the characteristics of the SOS estimates of facial image data. We also showed that the position sources model can explain a larger amount of variance of pupil image data than the first principle component.

1.10 Overview

In the following chapters we go more into the details. In Chapters 2 and 3 these details are mainly presented as papers we wrote on these subjects. Both of these chapters start with an introduction to relate the papers with each other and these introductions contain pointers to the most relevant points in these papers. Some of the papers are preceded by a prologue to give some additional pointers on the contents of the papers. At the end of each chapter a conclusion is drawn, based on all the papers presented in the chapter.

In chapter 2 we start with a thorough analysis of the eigenvalue bias in a single distribution. We present two algorithms for bias correction we derived ourselves, the bootstrap method (section 2.2) and the fixed point correction method (section 2.3), and discus how the error in eigenvector estimates should be taken into account. The focus of the experiments is on the correction of single distributions.

In verification we are dealing with two distribution estimates: the within class distribution and the between class distribution. In chapter 3 we go more into the relation between bias correction and verification systems with two estimated distributions. We first study the correction of either the within class estimate or the between class estimate or both. This study shows that particularly the correction of the between class estimate causes problems. In the next paper we explain why this happens and we introduce eigenwise correction which solves this problem.

Based on these analyses we show that eigenvalue bias indeed can have a significant effect in verification systems: bias correction can improve verification results significantly if synthetic data is used which is accurately modeled by the assumed model of SOS estimation. However, if the corrections are applied in experiments with real facial images, verification results go down. Moreover, the scree plot of the estimated sample eigenvalues show several remarkable characteristics. In chapter 4 we introduce the position sources data model and show that this model can both explain the characteristics of the eigenvalue plots and the limited success of bias correction applied to real facial data.

We also show how PCA dimensionality reduction can outperform bias correction if real facial data is used: PCA dimensionality reduction results in low pass filtering of the data, which in turn makes the fixed position intensity sources model fit better to the data (see section 4.3). However, low pass filtering ensures that the additional information available with increased image resolution is not used. We therefore expect that better solutions are possible. In section 4.4 we present a first approach to find a better solution for dealing with the position sources in facial image data.

Chapter 5 concludes the main part of the thesis. In that chapter we review the answers found in the thesis to the research questions. We also take another tour outside pattern recognition field and consider the implications of the limits in SOS estimation in other fields as well.

Chapter 2

Eigenvalue correction

2.1 Introduction

To study the effect of an increasing image resolution on facial verification performance, we formulated one of the research questions in section 1.8 as: "What (potential) effects does the sample eigenvalue bias have on verification systems and can these effects be reduced?" We answer the question in two stages: first we will analyse how the bias affects the SOS estimates of a single distribution in this chapter and we present solutions to improve the SOS estimates of single distributions. In chapter 3 we will deal with the second stage, in which we focus on the fact that verification performance depends on two distributions: the distribution of within class variations and the distribution of between class variations and in particular on the relation between these two distributions.

This chapter contains 3 papers. The first two papers describe two methods for sample eigenvalue bias correction we developed ourselves: section 2.2 is a paper presenting the bootstrap method and section 2.3 contains a paper presenting the fixed point correction method. The bootstrap method is a rather straightforward correction method which does not rely heavily on theoretical analysis of the sample eigenvalues bias. A major disadvantage of the method is that each iteration contains an eigenvalue decomposition, therefore the method significantly increases the required computational time and resources if it is included in eigenvalue estimation.

We removed the eigenvalue decomposition step from the bootstrap algorithm using eigenvalue bias theory. One problem with applying eigenvalue bias theory is that the available theory only holds for the limit case of both the number of samples N and their dimensionality P being infinite. Therefore, a large part of the paper in section 2.3 deals with how the theoretical bias descriptions can be used in practical cases where both N and p are finite. Based on these analyses we developed the fixed point correction.

However, SOS are determined by a combination of eigenvalues and eigenvectors. With eigenvalue correction the bias in the sample eigenvalues can be removed completely, under certain conditions. The eigenvector estimates are still affected by a large p over N correction though. Therefore, the combination of (perfectly) corrected eigenvalues with the estimated eigenvectors still provide a flawed estimate of the SOS. The last paper of this chapter, presented in section 2.4 discusses these problems and presents an additional correction for the eigenvalue estimates to estimate the variances along the estimated eigenvectors. It is demonstrated that the SOS estimates improve by this correction if measured by the Kullback Leibler divergence.

2.2 Bootstrap correction: a simple approach to bias correction[1]

2.2.1 Prologue

The main contribution of the following paper is the introduction of the bootstrap bias correction method. However, it also provides a good introduction to the framework in which the sample eigenvalue bias is usually analysed: instead of considering sets of eigenvalues, eigenvalue distributions are considered, because the bias description is derived for p becoming infinite. This will be explained in section 2.2.3.1, after which the relation between the population eigenvalue distribution and the sample eigenvalue distribution is given in section 2.2.3.2.

The paper also introduces the current state of the art correction method, the Karoui correction (section 2.2.3.3) and it contains the description of the isotonic tree algorithm (section 2.2.3.5), which is our contribution to the field to ensure order preservation during the update of eigenvalue estimates.

The introduction starts with introducing SOS estimation. If the reader is already familiar with the subject (after reading section 1.4), then he/she is recommanded to start at the fifth paragraph.

2.2.2 Introduction

Second order statistics are used extensively in data modeling methods. For example, in Principle Component Analysis (PCA) (see [20]), the second order statistics of high dimensional data are used to find subspaces containing the strongest modes of variation. In Linear Discriminant Analysis (LDA), the ratio of within class and between class variance is used to find the highest discriminating directions (see [21]).

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

[1] Published in [19]: A Bootstrap Approach to Eigenvalue Correction, ICDM '09, 2009

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

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

A problem of high dimensional training sets is that even though $\hat{\mathbf{\Sigma}}$ is an unbiased estimate of $\mathbf{\Sigma}$, $\boldsymbol{l}$ is a biased estimate of $\boldsymbol{\lambda}$. The bias becomes significant if the number of samples is of the same order as the dimensionality of the data. The bias has a negative effect on systems using these estimates as is shown for classification systems in [22, 23] for example.

To reduce this negative effect, the sample eigenvalues could be corrected to remove the bias. In this paper we present a bootstrap approach [24] to eigenvalue correction: our approach iteratively improves eigenvalue estimates without introducing new measurements. Instead, we generate synthetic data in each iteration which we use to update the population eigenvalue estimates.

The system of eigenvalue estimation and correction is schematically represented in Figure 2.1. In the figure, F represents the entire procedure of data generation and the sample eigenvalue estimation from this data. If the number of samples and the number of dimensions is large enough, F only introduces a bias on the eigenvalues as will be explained later on. Our aim is to obtain the inverse function F^{-1} that compensates for the bias in the sample eigenvalues. A method which applies $\widehat{F^{-1}}$ to sample eigenvalues is called a correction method. We add a superscript ^{c} to symbols representing results of such a correction method.

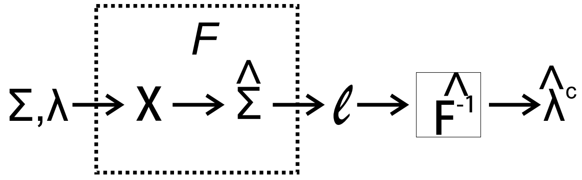


Figure 2.1: Schematic representation of the bias introduction and bias correction in eigenvalue estimation.

The currently available correction methods can roughly be divided in three categories: regularisation methods, corrections based on Steins loss criterion and corrections based on theoretical descriptions of the bias. The regularisation methods are often empirical and lack a strong theoretical foundation (See for example [25]). Correction algorithms based on Stein's loss criterion actually introduce a new bias to reduce the criterion (e.g. [9, 14]).

For the last category bias descriptions are needed. A description of the bias for many data distributions was proved in 1995 [26]. Karoui derived a correction method based on this relation. This method therefore has a strong theoretical basis as opposed to most regularisation algorithms, it reduces the bias as opposed to Stein's loss criterion based methods and it has been shown to reduce bias in real experiments [27]. We therefore consider this method to be one of the state of the art methods so we compare the performance of our bootstrap correction with the performance of this method.

The overview of the remainder of the article is as follows: we start with an analysis of the eigenvalue bias in section 2.2.3.1. In section 2.2.3.2 we introduce the Marčenko Pastur equation which describes this bias. A brief description of the Karoui correction is given in section 2.2.3.3. We then describe the bootstrap correction method (section 2.2.3.4) and compare correction results of the two methods in section 2.2.4 for a number of synthetic data sets. Section 2.2.5 gives a summary and conclusions.

2.2.3 Eigenvalue bias analysis and correction

2.2.3.1 Eigenvalue bias analysis

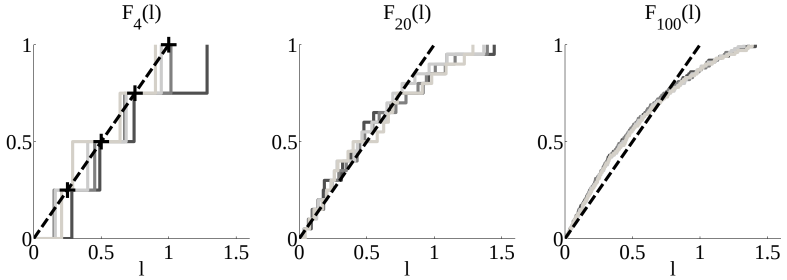
To find the statistics of estimators often Large Sample Analysis (LSA) is performed. In LSA it is assumed that the number of samples grows very large so the statistics of the estimator become a function depending solely on the number of samples, N. In this limit case the sample eigenvalues show no bias. However, for example in biometrics, the number of samples is often in the same order as the number of dimensions p or even lower and LSA cannot be used. Instead in the analysis of the statistics of the sample eigenvalues the following limit may be considered: $N, p \to \infty$ while $\frac{p}{N} \to \gamma$, where γ is some positive constant. Analysis in this limit are denoted as General Statistics Analysis (GSA) [28]. In GSA the sample eigenvalues do have a bias.

Figure 2.2 demonstrates why the limit in GSA is needed. It shows results of three experiments. For each experiment, the population eigenvalues are chosen uniformly between 0 and 1. While γ is kept constant at $\frac{1}{5}$, the number of dimensions is set to 4, 20 and 100 in figures 2.2a, 2.2b and 2.2c respectively. In each figure the population eigenvalue distribution function and the sample eigenvalue distribution functions for 4 repeated experiments are given. Given a set of eigenvalues $\boldsymbol{l}_i, i = 1 \ldots p$, the corresponding distribution function is given by equation 2.1:

$$F_p(l) \quad = \quad \frac{1}{p}\sum_{i=1}^{p} \mathrm{u}\left(l - \boldsymbol{l}_i\right) \tag{2.1}$$

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

Figure 2.2 shows that the empirical sample distribution functions converge with increasing dimensionality. The bias in the estimates is visible because they converge to a different distribution function than the population distribution function. For low dimensional examples, the bias is only a small part of the error in the estimates. The major part of the error is caused by random fluctuations of $F_p(l)$.



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

Figure 2.2: Examples of eigenvalue estimation bias toward the GSA limit. The dashed line indicates the population distribution H_p, the four solid lines are the empirical sample distribution G_p.

2.2.3.2 Marčenko Pastur equation

It turns out that under certain conditions a relation between the sample eigenvalues and the population eigenvalues can be given in the GSA limit. The main proofs and conditions on the input data are given in [29] and [26]. Here we briefly repeat the main points.

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

$$v_p(z) \quad = \quad (1-\gamma)\,\frac{-1}{z} + \gamma \cdot \int \frac{\mathrm{d}G_p(l)}{l-z} \tag{2.2}$$

here $G_p(l)$ is the empirical sample eigenvalue distribution function as given by equation 2.1 and $z \in \mathbb{C}^+$. In the GSA limit, if the population eigenvalue distribution function converges to $H_\infty(\lambda)$, then $G_p(l)$ converges weakly to a $G_\infty(l)$ such that the relation between the corresponding $v_\infty(z)$ and $H_\infty(\lambda)$ is given by equation 2.3.

$$-\frac{1}{v_\infty(z)} \quad = \quad z - \gamma \int \frac{\lambda \cdot \mathrm{d}H_\infty(\lambda)}{1+\lambda\, v_\infty(z)},\ z \in \mathbb{C}^+ \tag{2.3}$$

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

2.2.3.3 Karoui correction method

The Karoui method is based on the assumption that the number of samples and the number of dimensions is high enough such that the conditions given in section 2.2.3.2 are met and equation 2.3 holds. It is also assumed that the density function $h(\lambda)$ exists. The method approximates $h(\lambda)$ by a weighed sum of fixed density functions $p_i(\lambda)$ (see equation 2.4). In our implementation of the method we used a weighed sum of delta pulses and uniform distributions.

$$\hat{h}(\lambda) = \sum a_i \cdot p_i(\lambda),\ \sum a_i = 1,\ a_i \geq 0 \tag{2.4}$$

This approximation is then used in equation 2.3 after substitution of $\mathrm{d}H(\lambda)$ with $h(\lambda)\,\mathrm{d}\lambda$. The empirical sample eigenvalue density function given by equation 2.1 is substituted in equation 2.2 to determine a set of corresponding $v_p(z)$ and z values. The Karoui method then determines the set of a_i values which best satisfies equation 2.3 for this set of z values. For more details, we refer to [27].

2.2.3.4 Bootstrap eigenvalue correction method derivation

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

The general procedure of the method is given schematically in figure 2.3. To start the algorithm an initial estimate of $\boldsymbol{\lambda}$ is needed. We use $\boldsymbol{l}$ as initial estimate. The method then performs a number of iterations, where in each iteration the steps in figure 2.3 are performed starting from the right. In each iteration a synthetic set of white Gaussian distributed data samples $X_{w,n}$ is generated with the same number of samples and the same dimensionality as the original measurement. These samples are scaled so their population eigenvalues are equal to the current estimate of the population eigenvalues $\hat{\boldsymbol{\lambda}}^c_{:,n}$, where n is the current iteration index and subscript $_:$ indicates the entire column vector. From this synthetic data set the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_n$ is estimated. From this matrix the sample eigenvalues $\hat{\boldsymbol{l}}_{:,n}$ are determined. If the cost function $K = \left|\boldsymbol{l} - \hat{\boldsymbol{l}}_{:,n}\right|^2$ is below a threshold, the sample eigenvalues of the real data and the synthetic data are considered to be equal and therefore the current estimate of the population eigenvalues are used as final estimates of the population eigenvalues. If the cost function is not below the threshold, $\hat{\boldsymbol{\lambda}}^c_{:,n}$ is updated via update rule:

$$\hat{\boldsymbol{\lambda}}^c_{:,n+1} = \hat{\boldsymbol{\lambda}}^c_{:,n} - \mu \cdot \frac{\partial K}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}} \tag{2.5}$$

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

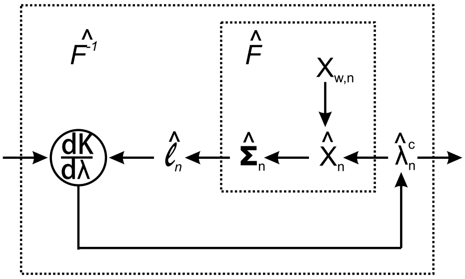


Figure 2.3: Schematic representation of the bootstrap eigenvalue correction

To find the derivative in the right hand side of equation 2.5 we can relate this derivative with the derivative $\frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}}$ via

$$\frac{\partial K}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}} \quad \propto \quad -\left(\boldsymbol{l} - \hat{\boldsymbol{l}}_{:,n}\right)^{\mathrm{T}} \cdot \frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}} \tag{2.6}$$

The derivative $\frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}}$ can be found using the expression for $\hat{\boldsymbol{l}}_{:,n}$ in equation 2.7, where M is the number of synthetic data samples. Note that we use normal distributed synthetic samples although we do not know the real distribution of the data samples. However, according to the Marčenko Pastur equation, in the GSA limit many distributions, including the normal distribution, will have the same relation between sample eigenvalues and population eigenvalues.

$$\begin{aligned} \hat{\boldsymbol{l}}_n \quad &= \quad \operatorname{diag}\left\{ \hat{\mathbf{E}}_n^{\mathrm{T}} \cdot \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}} \cdot \frac{1}{M-1} \cdot \mathbf{X}_{w,n} \cdot \right. \\ & \qquad\qquad \left. \mathbf{X}_{w,n}^{\mathrm{T}} \cdot \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}} \cdot \hat{\mathbf{E}}_n \right\} \end{aligned} \tag{2.7}$$

Here $\hat{\mathbf{E}}_n$ are the eigenvectors of the synthetic data covariance matrix and $\hat{\mathbf{D}}_n^c$ is a diagonal matrix with $\hat{\boldsymbol{\lambda}}^c_{:,n}$ on the diagonal. From equation 2.7 we find the derivative $\frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\boldsymbol{\lambda}}^c_{:,n}}$, given element wise in equation 2.8.

$$\begin{aligned} \frac{\partial \hat{\boldsymbol{l}}_{:,n}}{\partial \hat{\lambda}^c_{j,n}} \quad &= \quad \operatorname{diag}\left\{ 2 \cdot \hat{\mathbf{E}}_n^{\mathrm{T}} \cdot \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}} \cdot \frac{\mathbf{X}_{w,n} \cdot \mathbf{X}_{w,n}^{\mathrm{T}}}{M-1} \cdot \right. \\ & \qquad\qquad \left. \frac{\partial}{\partial \hat{\lambda}^c_{j,n}} \left(\hat{\mathbf{D}}_n^c\right)^{\frac{1}{2}} \cdot \hat{\mathbf{E}}_n \right\} \end{aligned} \tag{2.8}$$

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

2.2.3.5 Order preservation in the bootstrap correction method

When updating the eigenvalues without any constraints, oscillations may occur in which $\hat{\boldsymbol{\lambda}}^c_{i,n}$ switch value with $\hat{\boldsymbol{\lambda}}^c_{i+1,n}$ in each iteration. A order preserving algorithm was presented by Stein in [30]. We used an isotonic tree method, which allows blocks of eigenvalues to have a large change in value simultaneously and the method can also be implemented more efficiently.

The isotonic method first builds a tree: the top layer consists of the current population eigenvalues. Each lower layer is based on the difference and the mean of mean elements of the previous layer above, or $\hat{\boldsymbol{\lambda}}^c_{d,k,l} = \left(\hat{\boldsymbol{\lambda}}^c_{m,2k,l+1} - \hat{\boldsymbol{\lambda}}^c_{m,2k-1,l+1}\right)/2$ and $\hat{\boldsymbol{\lambda}}^c_{m,k,l} = \left(\hat{\boldsymbol{\lambda}}^c_{m,2k-1,l+1} + \hat{\boldsymbol{\lambda}}^c_{m,2k,l+1}\right)/2$ respectively. Here the first subscript of the left hand side symbol indicates whether the element is a mean (m) or a difference (d) element, the second subscript is the index of the element in the layer, and the third index is the layer index. So $\hat{\boldsymbol{\lambda}}^c_{d,k,l}$ is the k^{th} difference term on the l^{th} layer and $\hat{\boldsymbol{\lambda}}^c_{m,k,l}$ is the k^{th} mean term on the l^{th} layer. See the example in table 2.1. To reconstruct the $\hat{\boldsymbol{\lambda}}^c_n$ only the difference terms in all the layers and the final mean term are needed.

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

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

2.2.4 Experiments and discussion

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

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

Table 2.1: Isotonic tree with 6 population eigenvalues.

method and the Karoui method and measured the Levy distance,

$$\begin{aligned} d_L(F,G) &= \inf\{\epsilon \geq 0 : F(x-\epsilon) - \epsilon \leq \\ &\quad G(x) \leq F(x+\epsilon) + \epsilon, \forall\, x\} \end{aligned} \tag{2.9}$$

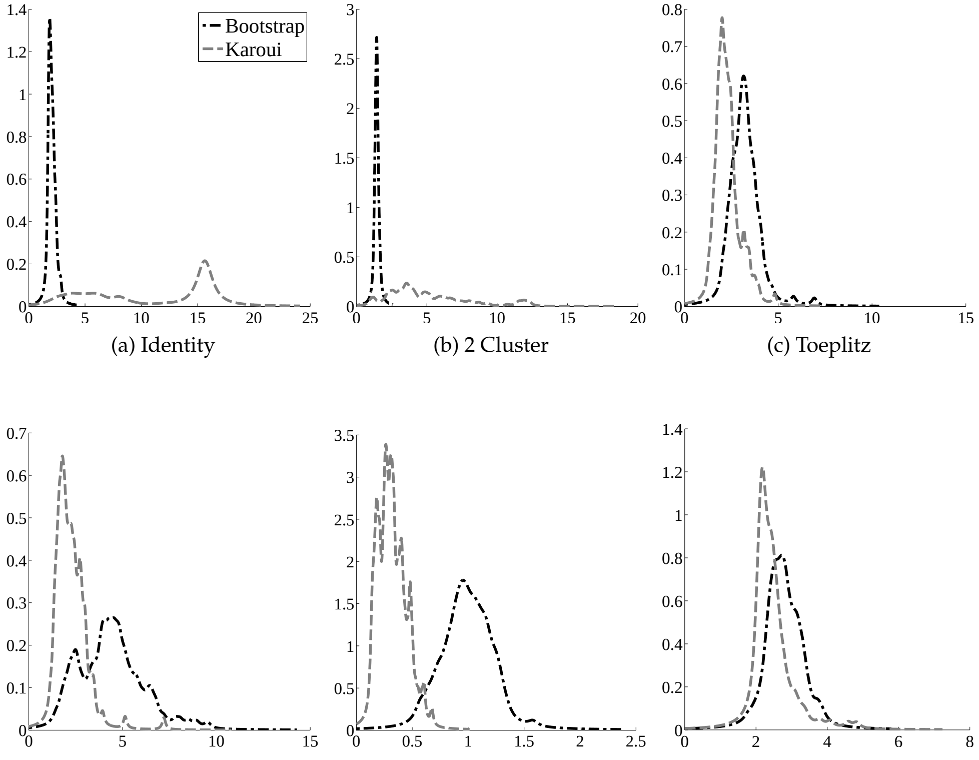
between the corrections and the population eigenvalues. Like Karoui, we calculate a score for a correction result by determining the ratio $d_L(G,H)/d_L(\hat{H}^c,H)$, where $\hat{H}^c$ is the corrected population eigenvalue distribution.

We performed the comparison with 6 different set-ups. For most experiments we drew 501 samples from an 100 dimensional normal distribution. The first three experiments have the same set-up as the experiments in [27]: experiment 1 (identity) has $\lambda_k = 1, \forall k$, experiment 2 (two cluster) has $\lambda_k = 1|k = 1\ldots50$ and $\lambda_k = 2|k = 51\ldots100$ and experiment 3 (Toeplitz) has the eigenvalues of a Toeplitz matrix. In the fourth experiment (Slope) $\lambda_k = 1 + k/100$. In the fifth experiment (one over n) we used $\lambda_k = \frac{1}{k}$, a common model for eigenvalues of facial data [25]. In the sixth experiment (undersampled slope) we draw 201 samples for a 600 dimensional normal distribution with $\lambda_k = 1 + k/600$.

By repeating the experiments collections of scores for the different set-ups are obtained. These score collections are represented by the histograms in figure 2.4. A large score indicates that the corrected population eigenvalue distribution and the population eigenvalue distribution are much more alike than the sample eigenvalue distribution and the population eigenvalue distribution. Therefore which ever method has the most density to the right of the histogram provides the best correction according to the Levy distance measure.

In the identity and two cluster experiments (figures 2.4a and 2.4b) the bootstrap method shows a steady improvement with an average score around 2. The Karoui correction achieves much higher scores, albeit with a huge spread. In the Toeplitz experiment (figure 2.4c) and the slope experiment (figure 2.4d), the bootstrap method has significantly higher scores. In the one over n case (figure 2.4e), the bootstrap method also performs better, however, its scores indicate its estimate has almost the same precision as the sample eigenvalues.

In the undersampled slope case (figure 2.4f) the Bootstrap method still outperforms the Karoui method, however the difference is not as big as in the slope experiment in figure 2.4d. For a more detailed discussion on the results of this experiment an example repetition is shown in figure 2.5, with figure 2.5a containing the population eigenvalue scree plots and figure 2.5b contains the sample eigenvalue scree plots.



(a) Identity

(b) 2 Cluster

(c) Toeplitz

(d) Slope

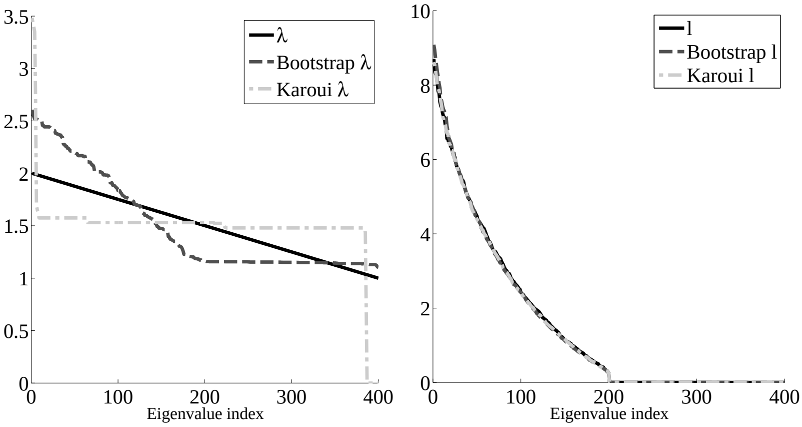
(e) 1/f

(f) Undersampled slope

Figure 2.4: Histogram of scores of both the bootstrap eigenvalue correction and the Karoui eigenvalue correction.

There are large differences between the population eigenvalues. The Karoui correction models the eigenvalues similar to the spiked population model: almost all eigenvalues are equal except for a few considerably larger ones (see [31] and [32]). It also sets a few population eigenvalues to zero. The bootstrap method sets all the population eigenvalues which were estimated as zero by the sample eigenvalue estimator to an equal, non zero value. For the remainder of the eigenvalues a slope is estimated, but with a higher gradient.

Despite the differences in the population eigenvalues, the sample eigenvalue scree plots are almost the same. This shows that the undersampled case is a very difficult problem, which seems to be under determined: multiple population eigenvalue distributions seem to generate the same sample eigenvalue distributions. In practice an equal value for all population eigenvalues in the zero space of the sample eigenvalues is required, since the order of the estimated eigenvectors in the zero space is random. Combining the estimated eigenvectors with different values



(a) Population eigenvalues

(b) 2 Cluster

Figure 2.5: Example repetition with the undersampled slope population eigenvalues.

will introduce arbitrariness. This will lead to arbitrary likelihood estimates for new samples with a component in the sample null space. The bootstrap method corrects all zero sample eigenvalues to almost the same value and seems therefore more applicable to undersampled problems.

We used the Levy distance to be comparable to the tests in [27], but the use of the distance is somewhat arbitrary and causes problems in performance comparisons. The distance is not scale invariant as is demonstrated by the example in figure 2.6. Another problem is that the use of a different criterion may very well result in a different ordering of the methods.

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

2.2.5 Conclusions and recommendations

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

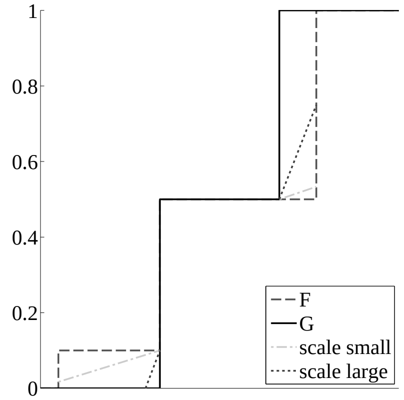


Figure 2.6: Scaling behaviour Levy distance. The dash-dotted line and the dotted line indicate the $F(x-\epsilon) \pm \epsilon$ line in case of small scale and in case of large scale respectively. In the small scale, the infimum is determined by the difference on the left, while in the large scale, the infimum is determined by the difference on the right.

smallest eigenvalues are estimated too small.

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

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

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

To reduce random fluctuations in the gradient estimate, the sample covariance matrix of white samples in equation 2.8 can be constructed using the Marčenko Pastur rule [33].

2.3 Eigenvalue estimation under large but not infinite conditions [2]

2.3.1 Prologue

To reduce the executing time of the bootstrap algorithm the Singular Value Decomposition (SVD) step should be removed. SVD is used to find the relation between population eigenvalues and sample eigenvalues implicitly, so to remove it, another method for determining this relation should be used. This relation is made explicit in the Marčenko Pastur equation, but for the hypothetical situation that both the number of samples (N) and the number of dimensions (p) are infinite. Therefore a major part of the following paper (section 2.2.3.5) focuses on how to make use of the Marčenko Pastur equation in practical problems with both a limited though large N and p.

Based on this analysis we derive a fixed point method to determine the sample eigenvalue distribution if the population eigenvalues are known (section 2.3.3.4). This algorithm is then used in a similar feedback scheme as used in the bootstrap correction to form a new bias correction method: the fixed point correction.

Of course this paper depends heavily on SOS modeling and GSA, so these topics are introduced again. From section 2.3.3.2 new material is covered. However, since this paper depends more heavily on the Marčenko Pastur equation, its introduction is more elaborate and it is therefore advisable to start from theorem 2.3.1.

2.3.2 Introduction

In data modeling, in order to give a meaningful interpretation of input samples, a description of the data generating process is needed. Often little is known about this process beforehand and the description consisting of a model and its parameters has to be derived from a set of examples, called the training set. Since the number of samples is usually limited in this training set, a model is chosen beforehand: the generation of this set is modeled as drawing samples from a random process $P(\boldsymbol{x})$, where the distribution of this random process is approximated by a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

There are two reasons for modeling the distribution with $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Firstly, a normal distribution has the highest entropy for a given variance. Therefore, according to the principle of maximum entropy, the normal distribution is the best choice if no further information about the distribution is available[35]. Secondly, for a multivariate normal distribution only the second order statistics have to be determined. Estimates of higher order statistics in high dimensional data can be highly distorted as shown in[36], but, as we will show, even the estimation of the second order statistics may be severely distorted.

As mentioned before, the parameters of the distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the *population mean* and *population covariance matrix*, are usually unknown and have to be estimated

[2] Smooth Eigenvalue Estimation, EURASIP, 2012 in communication [34]

from the training samples. For the mean the *sample mean*

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{x}_k \tag{2.10}$$

and for the covariance matrix the *sample covariance matrix*

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{k=1}^{N} (\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}) \cdot (\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \tag{2.11}$$

are often used as estimates. Here N is the number of samples in the training set, where each sample is a column vector with p elements, denoted by $\boldsymbol{x}_k$.

It is known that the *sample distribution* $\mathcal{N}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right)$ is not a good estimate of the *population distribution* $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ (See for example our demonstration in[37]), because even though the elements of the sample covariance matrix are unbiased estimates of the elements of the population covariance matrix, the eigenvalues of the sample covariance matrix, the *sample eigenvalues* $L = \{l_k | k = 1 \ldots p\}$, are biased estimates of the eigenvalues of the population covariance matrix, the *population eigenvalues* $\Lambda = \{\lambda_k | k = 1 \ldots p\}$. In [38] it has even been suggested to abandon the estimation of $\hat{\boldsymbol{\Sigma}}$ altogether. In classical LSA sample eigenvalues seem unbiased, because it is assumed that the number of samples is large enough to fully determine the statistics of the sample covariance matrix.

However, many applications evolve in such a way that dimensionality of the sample space increases as fast as or even faster than the number of samples in training sets. For example, in face recognition, the resolution of face images has increased considerably because high resolution devices have become available at modest costs, and the dimensionality of the sample space is related to the image resolution. The number of training samples depends on the number of test subjects available and the effort that can be put in collecting the data. As a result, in one of the databases with the largest number of subjects, the FRGC2 database [[39]] has images of approximately 500 individuals, while the feature vectors can reach a dimensionality of 10000 easily.

If the dimensionality is in the same order or even higher than the number of samples, LSA no longer gives accurate predictions of the statistics of the estimators. In GSA the dimensionality of the samples is also considered and is therefore more applicable than LSA as will be discussed in section 2.3.3.1.

Building on the work of many as is described in [40], using GSA, a relation between the sample eigenvalues and population eigenvalues was determined for a narrow set of sample distributions in[29], the Marčenko Pastur equation. In[26] it was shown that this relation holds for a large set of distributions. Based on this relation a correction of the sample eigenvalues is possible which leads to a more accurate estimate of the population eigenvalues. The basic idea is shown in Figure 2.7.

In Figure 2.7 the first part models how the sample eigenvalues are obtained. In the model, the data generating process generates samples for a training set X by drawing samples from a normal distribution with a set eigenvalues Λ, the *population eigenvalues*. From the training set a sample covariance matrix $\hat{\Sigma}$ is estimated. The decomposition of the matrix results in *sample eigenvalues* $\boldsymbol{L}$. This process can be modeled as a function $B(\boldsymbol{\Lambda}) = \boldsymbol{L}$. Bias correction can then be interpreted as applying an estimate of the inverse of B to the sample eigenvalues, which results in $\hat{\boldsymbol{\Lambda}}^c$, the estimate of the population eigenvalues after correction.

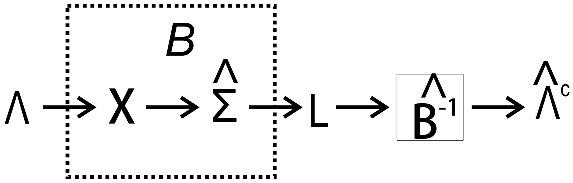


Figure 2.7: Schematic overview of the introduction of the bias in the sample eigenvalues and applying the correction to remove the bias.

One aspect of analysing eigenvalue estimation with GSA is that eigenvalue estimation is considered in the limit that the dimensionality of the samples becomes infinitely large. Therefore, instead of considering the eigenvalue set, an eigenvalue distribution description is used as explained in section 2.3.3.1. The Marčenko Pastur equation in fact does not give a relation between the sample eigenvalues and the population eigenvalues, but between the corresponding distributions in the GSA limit.

Of course, in practice, the dimensionality of the samples and the number of samples are not infinite, and the Marčenko Pastur equation can not be used directly to correct the bias in the sample eigenvalues. However, as we will show in Section 2.3.3.2, by applying a smoothing operation to both the population distribution estimate and the sample eigenvalue distribution estimate the relation between the two smoothed distributions is still accurately described by the Marčenko Pastur relation.

Because the Marčenko Pastur equation does relate the two smoothed distributions, we could develop 2 methods in Section 2.3.3.3, a polynomial method and a fixed point method, which both give a smoothed estimate of the sample eigenvalue density given a set of population eigenvalues. But in practice, bias correction is often desired, which equals to estimating the population eigenvalues corresponding to a set of sample eigenvalues. In Section 2.3.3.4 we derive two methods that can estimate a set of population eigenvalues given a set of sample eigenvalues. The fixed point bias correction method uses the fixed point sample eigenvalue density estimator, which shows that population eigenvalue to sample eigenvalue estimators do have their application.

In Section 2.3.4 we present several experiments. First we illustrate the effectiveness of the two sample eigenvalue density estimation methods: we show that the polynomial method makes good estimates of the sample eigenvalue densities if the number of population eigenvalue clusters is low, but fails if that number increases. We also show that the number of required iterations of the fixed point method increases if we decrease the smoothness of the estimation.

We then compare the fixed point bias correction method with a state-of-the-art-bias correction method by Kaorui ([40]) and a bootstrap bias correction method we presented in [19]. The fixed point method performs well in all experiments, but excels in two real life examples we often encountered in biometrics. In Section 2.3.5 we present conclusions based on these experiments.

2.3.3 Bias of the sample eigenvalues

2.3.3.1 Analysis of the bias in sample eigenvalues

Bias is a statistic of an estimator and in order to find the statistics of estimators, often the classical LSA is performed. In LSA the statistics of an estimator are determined for the limit $N \to \infty$, where N is the number of samples. With LSA, the sample eigenvalues seem to be unbiased. However, in many applications N is of the same order as the dimensionality of the sample space, p and LSA provides inaccurate statistics. In GSA [28], the limit $N, p \to \infty$ while $\frac{p}{N} \to \gamma$ is considered, where γ is some positive constant. Applying GSA to eigenvalue estimation does show a bias in the estimates.

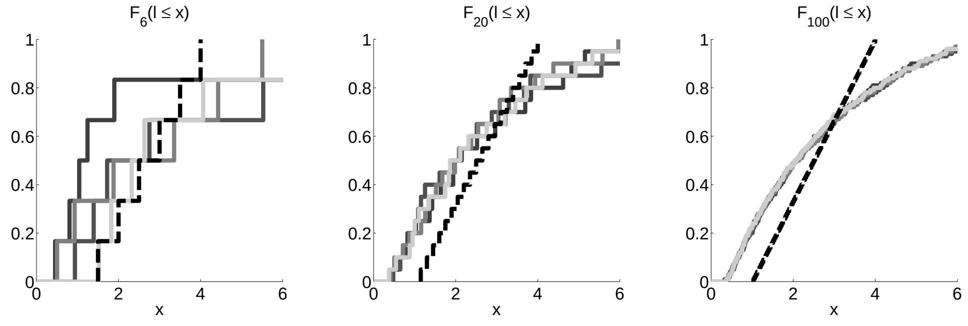
In the following example we demonstrate the situation under consideration in GSA. In the example we measured the sample eigenvalues of synthetic data, with the population eigenvalues uniformly distributed between 1 and 3. To show that the GSA limit "if $p \to \infty$" is relevant we set p to 6, 20 and 100, while keeping $\gamma = \frac{1}{3}$, so $N = 18, 60$ and 300 respectively. From the population eigenvalue sets and the measured sample eigenvalue sets we determined the empirical eigenvalue distribution function, which is given by equation (2.12) for an eigenvalue set $\{x_k | k = 1 \ldots p\}$:

$$F_p(x) = \frac{1}{p} \sum_{k=1}^{p} \mathrm{u}(x - x_k) \tag{2.12}$$

where $\mathrm{u}(x)$ is the unit step function.

In figure 2.8 we show both the empirical population eigenvalue distribution H_p and 4 empirical sample eigenvalue distributions G_p for the different settings of p. If p is low, large variations in the G_p's occur and the bias is only a small component in the estimation error. However, if p increases, the G_p's converge to a fixed distribution, different from H_p. This difference is due to the bias in the eigenvalue estimates.

In [29] an equation was given, which describes the relation between the sample eigenvalue distributions and the population eigenvalue distribution. Originally this relation, from now referred to as the Marčenko Pastur (MP) equation, was proved for a very limited set of data distributions, but based on the work of many others as is described in [40], in [26] it was shown that the same relation holds for a much larger set. The MP equation requires the Stieltjes transform of the empirical sample



(a) 6 dimensions (b) 20 dimensions (c) 100 dimensions

Figure 2.8: Examples of eigenvalue estimation bias toward the GSA limit. The dashed line indicates the population distribution H_p, the four solid lines are the empirical sample distributions G_p.

eigenvalue distribution, which is given by:

$$m_{G_p}(z) = \int \frac{\mathrm{d}G_p(l)}{l-z}, \text{ for } z \in \mathbb{C}^+ \tag{2.13}$$

More specifically, the MP equation uses the Stieltjes transform $v_{G_p}(z)$, which is related to $m_{G_p}(z)$ via

$$v_{G_p}(z) = \left(1 - \frac{p}{n}\right)\frac{-1}{z} + \frac{p}{n} m_{G_p}(z) \tag{2.14}$$

We now quote Theorem 1 from [40], which gives the MP equation and the conditions under which it holds:

Theorem 2.3.1 *Suppose the data matrix X can be written $X = Y\Sigma_p^{\frac{1}{2}}$, where Σ_p is a $p \times p$ positive definite matrix and Y is an $n \times p$ matrix whose entries are independent and identically distributed (real or complex), with $E(Y_{i.j}) = 0$, $E\left(|Y_{i,j}|^2\right) = 1$ and $E\left(|Y_{i,j}|^4\right) < \infty$.*

Call H_p the population spectral distribution, i.e. the distribution that puts mass $1/p$ at each of the eigenvalues of the population covariance matrix, Σ_p. Assume that H_p converges weakly to a limit denoted H_∞. (We write this convergence $H_p \Rightarrow H_\infty$.) Then, when $p, n \to \infty$, and $p/n \to \gamma, \gamma \in (0, \infty)$,

1. *$v_{G_p} \to v_\infty(z)$, a.s, where $v_\infty(z)$ is a deterministic function*

2. *$v_\infty(z)$ satisfies the equation*

$$-\frac{1}{v_\infty(z)} = z - \gamma \int \frac{\lambda \mathrm{d}H_\infty(\lambda)}{1 + \lambda v_\infty(z)}, \forall z \in \mathbb{C}^+ \tag{2.15}$$

3. *The previous equation has one and only one solution which is the Stieltjes transform of a measure.*

Equation (2.15), the MP equation, fully characterizes the sample eigenvalue distribution G_∞ if the population eigenvalue distribution is known. However, the question at hand is to derive a method that reduces the bias in the sample eigenvalues, that is rewrite equation (2.15) in the form $\Lambda = B^{-1}(L)$.

2.3.3.2 Smooth eigenvalue estimation

The Marčenko Pastur equation describes the relation between the sample eigenvalue distribution and the population eigenvalue distribution in the GSA limit, but in practice both N and p are finite. We assume however, that the global characteristics already converge for lower values of N and p and higher values of N and p are only required if very local details have to be considered. To support this assumption, consider the curves in Figure 2.8 again. The empirical distribution function is always staircase shaped as shown in 2.8a, with at most p jumps of at least height $\frac{1}{p}$, so that the curve can only contain local details if p is large enough. For an exact definition of local detail, see appendix 2.3.8.

Based on the assumption of convergence of global characteristics for lower p and N values we show that by varying the imaginary value of z, we can control the influence of local details and global characteristics in equations (2.15) and (2.13). This result we will use later on to derive algorithms which can be used in practical situations.

First we introduce the inverse Stieltjes transform:

$$g(x) = \frac{1}{\pi}\lim_{y\downarrow 0}\Im\left\{m_G(x+\imath y)\right\} \tag{2.16}$$

In order to find the sample eigenvalue density given the population eigenvalue distribution, equation (2.15) has to be solved with $\Im\{z\}\downarrow 0$. However, as long as we use empirical distributions, setting $\Im\{z\} = 0$ will lead to several problems. For example, the Stieltjes transform of the empirical sample eigenvalue distribution will become infinite at the sample eigenvalues and real valued anywhere else. If on the other hand we solve equation (2.15) with $\Im\{z\}$ some fixed positive constant y, we find the empirical sample eigenvalue density convolved with the Cauchy kernel $\frac{1}{\pi}\left[\frac{y}{x^2+y^2}\right]$, as noted in [41].

The factor y therefore seems to have a smoothing effect: local details are filtered out. But this is not limited to the resulting density; setting y non-zero has a similar effect on the Marčenko Pastur equation. Consider the integrands in both the integral of the Stieltjes transform (Equation (2.13)) and the integral in the Marčenko Pastur equation (Equation (2.15)). Both arguments can be rewritten to the form:

$$b(r) = \frac{1}{r-\imath q} \tag{2.17}$$

where r is a real variable and q is a real constant. For example, if we set $r = l - \Re\{z\}$ and $q = \Im\{z\}$ we have the argument of the integral in equation (2.13). The function in equation (2.17) is a generalized circle, a specific kind of Möbius transform [42], which in this case describes a circle in the complex plane with a center $\imath\frac{1}{2q}$ and radius $\frac{1}{2q}$ (see figure 2.9).

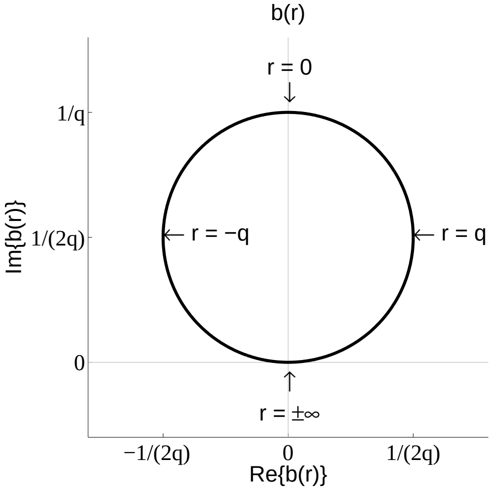


Figure 2.9: The integration arguments used as mapping functions describe a circle in the complex plain with diameter $\frac{1}{q}$. Where each point p is mapped depends on the value of q: if q is small, most points will be mapped in the neighbourhood of $r = \pm\infty$, if q is large more points will be mapped away from $r = \pm\infty$.

The result of the integrals in equations (2.13) and (2.15) is determined by the mapping of the distribution function along this circle. In case q is small, only a small part of the real axis is mapped to other positions than the infinity points. If any probability mass is repositioned, it will still be mapped to the infinity points and so the change will have little effect on the result of the integral, unless the change is in the neighbourhood around $r = 0$. So for small q, the results of the integrals are only sensitive to a small part of the density function.

If on the other hand q is large, much more of the real axis is mapped to other positions than the infinity points. In that case changes of position of density in a large neighbourhood around $r = 0$ have an effect on the result of the integral. In the extreme cases, for $q \downarrow 0$, the integration result is determined by one point on the distribution curve, which gives an explanation of the limit in the inverse Stieltjes transform. If $q \to \infty$, the results are solely determined by the means of the distributions. An exact proof of these claims is given in Appendix 2.3.8.

Because of this mapping the result of the integral is bounded by the circle (see the proof in appendix 2.3.6), which is a more strict property than the well known condition $\Im\{m_G(z)\} \leq 1/\Im\{z\}$ [43]. This limit will be used for choosing a starting point in the fixed point algorithm in section 2.3.3.3 and the limit will be used as an upper bound of the minimal value of $\Im(z)$ for which the fixed point algorithm will converge (appendix 2.3.7).

The observation about the sensitivity of the integral results for variations of the distributions will be used in the following sections where we first derive an algorithm to find a smoothed estimate of the sample eigenvalues if the population eigenvalues are known. We then use this algorithm in a feedback algorithm to find an estimate of the population eigenvalues given that the sample eigenvalues are known.

2.3.3.3 From population eigenvalues to sample eigenvalues

In the previous section we showed that by setting $\Im\{z\}$ low or high we can control how much effect local details of the distributions have on the Stieltjes transform of the empirical sample eigenvalue density and consequently in the MP equation. This means we can approximate the distributions with the empirical distributions if we set $\Im\{z\}$ high enough.

If we substitute the empirical population distribution function $H_p(\lambda) = \frac{1}{p}\sum_{k=1}^{p} u(\lambda - \lambda_m)$ and $\hat{v}_p(z)$ for the population distribution and $v_\infty(z)$ respectively in equation (2.15), we get the following equation:

$$\frac{-1}{\hat{v}_p(z)} = z - \frac{1}{N}\sum_{k=1}^{p}\frac{\lambda_k}{1+\lambda_k\hat{v}_p(z)} \tag{2.18}$$

To find the corresponding sample eigenvalue density $\hat{g}(l)$, $\hat{v}_p(z)$ has to be solved from equation (2.18). We present two solutions: a polynomial method and a fixed point method.

Polynomial method In this section we will derive and analyse a polynomial method, which was already found by Rao and Edelman [44]. Their derivation has a solid embedding in random matrix theory, but therefore it is less focussed on the application we discuss in this article.

We derive the polynomial method by rewriting equation (2.18) to a polynomial expression by multiplying both sides of the equation with $\hat{v}_p(z)\prod_{k=1}^{p}\left(1+\lambda_k\hat{v}_p(z)\right)$. The new expression can be rewritten to an expression of the form $0 = \prod_{k=0}^{p+1} c_k \hat{v}_p^k(z)$, which then can be solved using standard polynomial solving tools.

A problem with this method is that if the number of eigenvalues increases, the order of the polynomial increases and the roots of the higher order polynomial become numerically unstable. As observed in the experiments, the polynomial solution becomes unreliable above 10 eigenvalues. The advantage of the method is that it can solve equation (2.18) for arbitrarily small $\Im\{z\}$ values, even 0.

Fixed point method A second method to solve equation (2.18) is by using a fixed point method. The fixed point method is based on rewriting equation (2.18) to

$$A = z + \frac{1}{N} A \sum_{k=1}^{p} \frac{\lambda_k}{\lambda_k - A} \tag{2.19}$$

where $A = -\frac{1}{\hat{v}_p(z)}$. If we replace the A on the left hand side by A_n and the A's on the right hand side by A_{n-1}, we get an equation similar to the general form of a fixed point method in [45]:

$$A_n = \mathrm{FP}\left(A_{n-1}\right) \tag{2.20}$$

Our hypothesis is that equation (2.19) is indeed a fixed point algorithm where A_n converges to a fixed point if the output of iteration n A_n is repeatedly used as input again in iteration $n+1$. In appendix 2.3.7 we prove convergence for a minimal value of $\Im\{z\}$, but we observed that if we set $\Im\{z\}$ below this value we still get a good approximation, however the number of iterations required increases.

Since the solution should be within the limit circle, as described in section 2.3.3.2, we use the center of that circle as a starting point.

The result of both the polynomial method and the fixed point method is an estimate of the Stieltjes transform of the sample eigenvalue distribution. But in general the sample eigenvalues are required. Using the inverse Stieltjes transform (equation (2.16)) the sample eigenvalue density can be found, but it is convoluted with the Cauchy density, since the chosen values of z have an imaginary value larger than 0. Finding the sample eigenvalue density by deconvolution is hard since the estimate should have zero density for negative values and it should be non negative everywhere. Furthermore, the Cauchy density has an infinite variance and the convolved density is only known on fixed positions ($\Re\{z\}$).

A schematic representation of the developed methods is given in Figure 2.10. On the left, the population eigenvalues Λ are used as input. On these eigenvalues each algorithm applies an estimate of the bias introducing function F, after which the convolution with the Cauchy kernel occurs (S). The result is an estimate of the sample eigenvalue density convolved with the Cauchy kernel $g(\hat{l})$.

Although the methods do not give an estimate of the sample eigenvalues themselves, they can still be useful. One application is to use them to test whether candidate population eigenvalue sets match with the measured sample eigenvalues. First the sample eigenvalue density corresponding to this candidate population eigenvalue set is estimated. If this estimated sample eigenvalue density does not match the empirical density of the measured sample eigenvalues, the candidate population eigenvalue set is probably not a good candidate. One particular candidate could be the measured sample eigenvalues themselves. If they do not match, then the measured sample eigenvalues are probably considerably biased estimates of the original population eigenvalues as well.

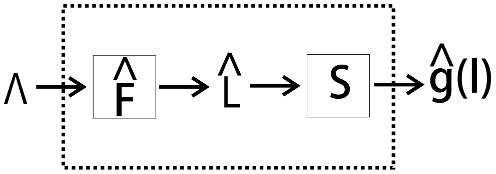


Figure 2.10: Sample eigenvalue density estimation based on population eigenvalues.

2.3.3.4 Sample eigenvalues to population eigenvalues

Although methods for determining the sample eigenvalues if the population eigenvalues are known do have applications (see the example in the previous section), often methods that can determine the population eigenvalues belonging to a set of sample eigenvalues are desired. There already exist several methods designed to perform this action (See [40] and [19]), where the method developed by Karoui can be considered the state of the art at the moment. The method is based on the MP equation as well, but it lacked the explanation of how to deal with finite p and N. Moreover, it also uses estimates the population distribution instead of a set of eigenvalues, making it less suited for a number of practical problems.

Some of these problems are encountered in biometrics, where the distribution of the sample eigenvalues suggests that there are a few significant eigenvalues and the remainder form some bulk. If individual eigenvalues are of importance, than the distribution description used in the Karoui method is less suited, as is also shown in the experimental results presented in section 2.3.4.

We therefore designed two new methods based on the theory and methods presented in the previous sections. Particularly the second method has several advantages over the existing methods. Firstly, it estimates the population eigenvalues directly instead of a density estimate and secondly, as will be shown in the experiments, the method performs better for a number of practical situations.

Direct density estimation solution The first method is based on the Stieltjes transform of the population eigenvalue distribution. If we are able to determine this transform, then the population eigenvalue density can be found using the inverse transform. The MP equation can be rewritten to give this Stieltjes transform:

$$\begin{aligned} -\frac{(1-\gamma)\,v_\infty(z) + z\,v_\infty^2(z)}{\gamma} &= \int \frac{\mathrm{d}H_\infty(\lambda)}{\lambda - \frac{-1}{v_\infty(z)}} \\ &= m_{H_\infty}\left(\frac{-1}{v_\infty(z)}\right) \end{aligned} \tag{2.21}$$

If we now substitute $\tilde{z}$ for $\frac{-1}{v_\infty(z)}$, we get the follow expression for the Stieltjes transform of the population eigenvalue distribution:

$$m_{H_\infty}(\tilde{z}) = \frac{(1-\gamma)\,\tilde{z} - v^{-1}\left(-\tilde{z}^{-1}\right)}{\tilde{z}^2\gamma} \tag{2.22}$$

where $v^{-1}(c)$ is the function which solves z from $c = v_\infty(z)$.

The reason for choosing $\tilde{z}$ as a parameter and determining the corresponding z instead of choosing z and calculating the corresponding $\tilde{z}$ is twofold: firstly, in section 2.3.3.2 it was noted that if the inverse transform is applied with an argument with a non zero imaginary part, a density is determined which is a convolution of the original density with the Cauchy kernel. The width of the convolution kernel is determined by $\Im\{\tilde{z}\}$. Secondly, the point at which this density is determined is controlled by $\Re\{\tilde{z}\}$. If z is chosen as variable, both parameters are difficult to control.

There are 4 major problems with implementing this method. First of all, the evaluation of equation (2.22) requires an implementation of $v^{-1}(c)$, which is not straightforward. Secondly, the method suffers from numerical instabilities, which are hard to predict in advance. The method also requires deconvolution and combined with the numerical instabilities this can easily lead to large errors. The last problem of the method is that it finds an eigenvalue density description instead of a set of eigenvalues. Because of these problems we do not use the method any further.

Feedback correction In section 2.3.3.3 we derived two algorithms that can estimate a sample eigenvalue density convolved with a Cauchy density corresponding to a set of population eigenvalues. In this section we derive a feedback method which uses the methods from section 2.3.3.3 to correct population eigenvalue estimates.

The global idea (schematically represented in figure 2.11) is as follows: the algorithm starts with an initial estimate of the population eigenvalues ($\hat{\mathbf{\Lambda}}_n^c$ with $n = 1$). The sample eigenvalues corresponding to these population eigenvalues ($\hat{\boldsymbol{L}}_n$) are estimated and compared to the measured sample eigenvalues ($\boldsymbol{L}$). If both sets are not very similar, the estimate of the population eigenvalues is updated and the steps are repeated again.

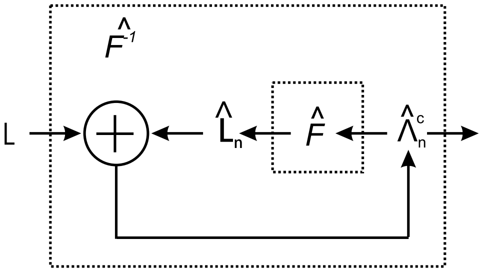


Figure 2.11: Schematic representation of a Feedback correction. In each iteration the sample eigenvalues corresponding to the current estimate of the population eigenvalues are calculated. The estimate of the population eigenvalues is updated by comparing these sample eigenvalues with the measured eigenvalues.

But, as noted in section 2.3.3.3 we do not actually estimate the sample eigenvalues, but the convoluted sample eigenvalue density $\hat{g}_y(l)$. We therefore convolve the empirical distribution of the measured sample eigenvalue set with the Cauchy density, resulting in $g_y(l)$, and compare these two densities.

In order to derive a feedback algorithm which always converges, the best solution is to compare the distribution functions instead of density functions. However,this requires numerical integration of $\hat{g}_y(l)$ which results in a large amplification of the errors in the tails of the density. Besides using densities instead of the distributions, the influence of the tail errors can be reduced even more, by considering the Kullback Leibler divergence as a measure of similarity between $\hat{g}_y(l)$ and $g_y(l)$:

$$d_{KL}\left(\hat{g}_y, g_y\right) = \int \hat{g}_y(l) \log \frac{\hat{g}_y(l)}{g_y(l)} \mathrm{d}l \tag{2.23}$$

where we use the natural logarithm for log.

Due to numerical integration problems, negative parts of the integral can be over estimated, resulting in a negative cost value. We therefore squared the logarithm resulting in the following cost function:

$$K\left(g_y, \hat{g}_y\right) = \int \hat{g}_y(l) \log^2 \frac{\hat{g}_y(l)}{g_y(l)} \mathrm{d}l \tag{2.24}$$

This is still a valid cost function: it is still 0 if and only if $g_y(l) = \hat{g}_y(l)$ and larger than 0 for any mismatch in distributions. Furthermore the focus on the tails is still reduced since $\lim_{a\downarrow 0} a \log^2 \frac{a}{b} = 0$. Note also that we chose $g_y(l)$ as denominator since it is never zero for $y > 0$, because the Cauchy convolution kernel is never zero if $y > 0$.

If the cost function exceeds a preset threshold, the population eigenvalue estimates need to be adjusted. We use gradient descending to find a better estimate. To do the adjustments we need to find an expression for the gradient $\frac{\partial K}{\partial \lambda}\left(g_y, \hat{g}_y\right)$. We first relate the gradient $\frac{\partial K}{\partial \lambda}\left(g_y, \hat{g}_y\right)$ to $\frac{\partial \hat{g}_y(l)}{\partial \lambda}$ via:

$$\frac{\partial K}{\partial \lambda}\left(g_y, \hat{g}_y\right) = \int \log\left(\frac{\hat{g}_y(l)}{g_y(l)}\right)\left(\log\left(\frac{\hat{g}_y(l)}{g_y(l)}\right) + 2\right) \frac{\partial \hat{g}_y(l)}{\partial \lambda} \mathrm{d}l \tag{2.25}$$

Each of the elements of this gradient $\frac{\partial \hat{g}_y(l)}{\partial \lambda}$ can be related to $\frac{\partial}{\partial \lambda_m} A(l + \imath y)$

$$\frac{\partial \hat{g}_y(l)}{\partial \lambda_m} = \frac{1}{\pi \gamma} \Im\left\{A^{-2}(l + \imath y) \frac{\partial}{\partial \lambda_m} A(l + \imath y)\right\} \tag{2.26}$$

where

$$\frac{\partial A}{\partial \lambda_m} = \frac{-\frac{1}{n}\left(\frac{A}{\lambda_m - A}\right)^2}{\left(1 - \gamma - \frac{2}{n}\sum_{k=1}^{p} \frac{A}{\lambda_k - A} - \frac{1}{n}\sum_{k=1}^{p}\left(\frac{A}{\lambda_k - A}\right)^2\right)} \tag{2.27}$$

The feedback correction thus created is represented schematically in figure 2.12. A clear advantage of this algorithm is that the end result is not a density description but a set of population eigenvalues. Another advantage is that the smoothness factor is incorporated without needing to deconvolute the output. A third advantage is that the correction corrects all zero valued sample eigenvalues to the same value, which is a good property as explained in section 2.3.3.5.

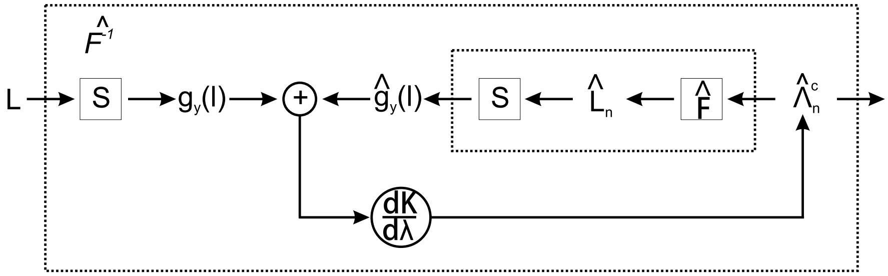


Figure 2.12: Extended schematic representation of our implementation of feedback correction. Instead of comparing sample eigenvalues, smoothed sample eigenvalue densities are compared.

Maintaining order among eigenvalues For most of the methods presented in the previous sections it is necessary that the eigenvalues are sorted in order of value and that they keep this order during updating. If order is not maintained, one of the problems that may occur is oscillation: $\tilde{\lambda}_k$ may switch places with $\tilde{\lambda}_{k+1}$ in one iteration and switch back the next iteration. Other eigenvalue correction methods had the same problem. Therefore Stein presented an algorithm to ensure order preservation during eigenvalue updating [30]. We used an isotonic tree algorithm for this purpose, described in [19], which has several advantages over the algorithm of Stein.

2.3.3.5 Correction of the null space

A problem in eigenvalue correction occurs in underdetermined cases, which are characterised by N being smaller than p. In these case the data matrix has a zero space and $p - N + 1$ sample eigenvalues are necessarily zero, so the correction tries to estimate p population eigenvalues from $N - 1$ non-zero sample eigenvalues.

A related effect of underdetermination is that the sample eigenvectors in the null space form a random orthogonal basis. Without additional information, correction of the zero valued sample eigenvalues with varying values results in randomness in the correction. This suggests that for correction all zero valued sample eigenvalues should be given an equal value.

A more theoretically sound argument for such a correction is based on the maximum entropy theorem (see [35]). The maximum entropy method states that if there are multiple solutions to a problem and all available information has been used to narrow the selection, the best solution is the one with the highest entropy.

The entropy of a multivariate normal distributed random variable is given by

$$\frac{1}{2}\ln\left\{(2\pi \mathrm{e})^{p}\,|\Sigma|\right\} \tag{2.28}$$

This entropy is maximized if the determinant is maximized, which is the product of the eigenvalues. With the constraint that the sum of the eigenvalues remains constant, the maximum of the product is achieved when all eigenvalues are equal. This is thus the maximum entropy solution.

2.3.4 Experimental validation

In the following sections we present 3 experiments: in the first experiment we illustrate some of the characteristics of the population eigenvalue to sample eigenvalue methods. In the second experiment we compare the performance of the fixed point eigenvalue correction method with an implementation we made of a state-of-the-art correction method by Karoui and a bootstrap correction method. In the last experiment we apply the correction method in a verification experiment, with a configuration often encountered in face recognition: a high number of samples with high dimensionality, where the number of samples is smaller than the dimensionality of the samples.

2.3.4.1 Population to sample eigenvalue results

In Section 2.3.3.3 we derived two algorithms to find the sample eigenvalue distribution given a set of population eigenvalues: a polynomial algorithm (section 2.3.3.3) and a fixed point algorithm (section 2.3.3.3). We noted two characteristics of the methods: the polynomial algorithm will have problems if the number of eigenvalue clusters increases and the fixed point method will require more iterations before convergence occurs if the smoothness factor is decreased.

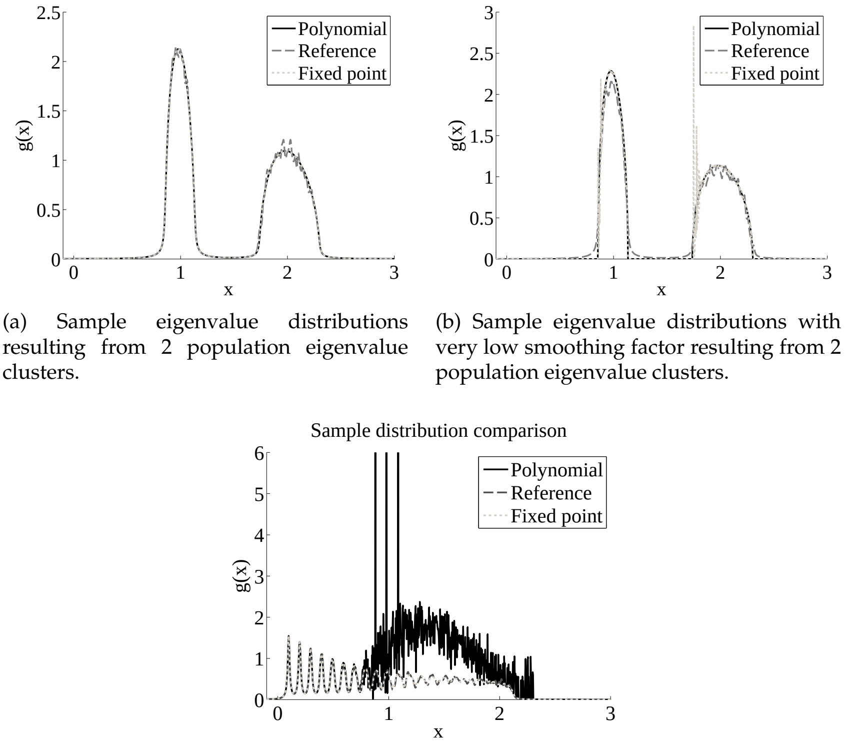
To demonstrate these characteristics we estimated the sample eigenvalue densities in 3 different settings: first we estimate the sample eigenvalue density of belonging to a population eigenvalue set with half of the eigenvalues equal to 1 and the other half equal to 2, with the ratio between the dimensionality of the samples and the number of samples equal to 0.01 and with a smoothness factor y (equation (2.16)) of 0.01 as well. In the second experiment we lower the smoothness factor to 10^{-5}. In the third experiment we set the smoothness factor back to 0.01 but the population eigenvalue set is divided in 20 clusters uniformly distributed between 0.1 and 2.

A reference density is obtained as follows: first a synthetic data set is generated with the same parameters as in the experiments described. Then the sample eigenvalues of synthetic data are calculated. The corresponding empirical density function is then convolved with a Cauchy kernel with the same width as the smoothness factor.

Figure 2.13a shows the estimates of the sample eigenvalue distribution for the first setting. All three estimates are very alike, only the reference estimate shows some local variations due to the use of a limited number of samples.

Figure 2.13b shows that when the smoothness factor is decreased, the fixed point algorithm has not converged on all positions if the number of iterations is kept the same. After increasing the number of iterations, the fixed point algorithm converged on all points again (not shown). Note that the reference distribution is still convolved with a Cauchy kernel of width 0.01 so variations due to local details are kept small.

If the number of eigenvalue clusters is increased, the roots of the polynomial method become unstable and the estimation fails as shown in Figure 2.13c. The fixed point method is still accurate.



(a) Sample eigenvalue distributions resulting from 2 population eigenvalue clusters.

(b) Sample eigenvalue distributions with very low smoothing factor resulting from 2 population eigenvalue clusters.

(c) Sample eigenvalue distributions with 20 population eigenvalue clusters.

Figure 2.13: Sample eigenvalue density predictions based on population eigenvalues.

2.3.4.2 Sample to population eigenvalue results

As noted earlier, the more common problem is how to get from the measured sample eigenvalues to an estimate of the population eigenvalues. Two methods to solve this problem were described in section 2.3.3.4: a direct estimation method and a fixed point feedback loop method.

Direct estimation results Some tests on the direct estimation method (section 2.3.3.4) showed that the method has several implementation flaws. A major flaw is that it results in an estimate of the population eigenvalue density convolved with the Cauchy kernel instead of the population eigenvalues. Because the Cauchy kernel has infinite variance, this poses the problem that the spread in population eigenvalues keeps increasing with an increasing number of eigenvalues. The smaller eigenvalues eventually even end up with values below zero. Because of this flaw, we did no further experiments.

Fixed point correction results The second method is based on using the fixed point algorithm of section 2.3.3.3 in a feedback loop as described in section 2.3.3.4. In [19] we compared an eigenvalue correction method based on bootstrapping with our implementation of the method developed by Karoui. In the next experiment we repeat the comparison but we also include the iterative feedback algorithm.

The experimental set-up is as follows. Synthetic data is generated by drawing N samples from $\mathcal{N}(0, \boldsymbol{D})$, a p-variate normal distribution with zero mean and with diagonal matrix $\boldsymbol{D}$ as covariance matrix. From the data the sample eigenvalues are determined and afterwards these sample eigenvalues are corrected with the three correction methods.

An accuracy score is assigned to each correction result by measuring the Levy distance between the empirical distributions of the sample eigenvalues and the population eigenvalues and dividing this distance by the Levy distance between the empirical distributions of the corrections and the population eigenvalues:

$$\text{score} = \frac{d_L(H, G)}{d_L(H, \hat{H})} \tag{2.29}$$

where the Levy distance d_L between distributions F and G is given by:

$$\begin{aligned} d_L(F, G) &= \inf\{\epsilon \geq 0 : F(x - \epsilon) - \epsilon \leq \\ &\quad G(x) \leq F(x + \epsilon) + \epsilon, \forall x\} \end{aligned} \tag{2.30}$$

After repeating these experiments a number of times, a histogram per correction method can be determined.

We used the Levy distance to make the experiments comparable with the experiments in [40]. But, as we showed in [19], the levy distance has several disadvantages, one being that the distance measure is not scale independent.

We tested 6 different parameter settings. In the first 5 only the distribution of the population eigenvalues vary. We keep the number of dimensions p fixed at 100 and the number of samples N fixed at 500. In the last experiment, we changed N to 201 and p to 600. The settings for the population eigenvalues are given in Table 2.2. Experiments 1, 2 and 4 are repetitions of the experiments done by Karoui. We added experiment 5 because a 100 over f model is a common model for eigenvalues estimated from facial data (see [25], [46] and [47]), even though its limiting distribution is 0 for the GSA limit. Another characteristic of facial data is that it is underdetermined. The performance of the correction methods under such conditions is measured by experiment 6. To compare these performances with the performance if there are more samples than dimensions, experiment 3 is introduced.

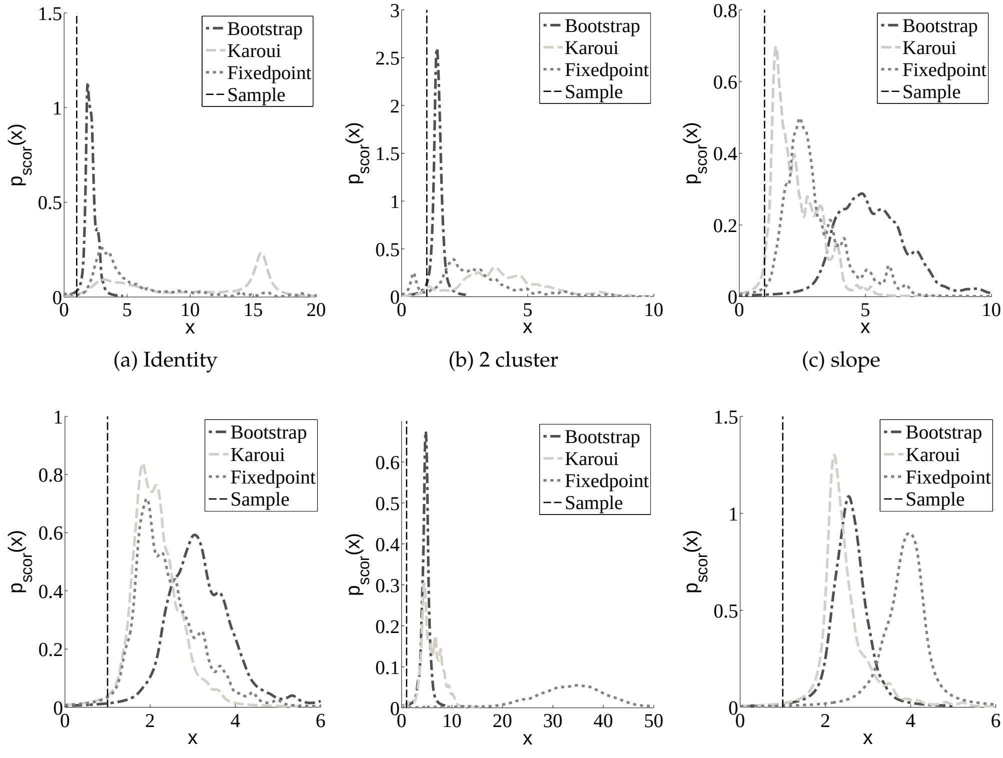
Table 2.2: The population eigenvalues per experiment

Experiment	Λ	Description
1	$\lambda_k = 1$	Identity
2	$\lambda_k = 1 \vert k = 1 \ldots 50$ $\lambda_k = 2 \vert k = 51 \ldots 100$	2 Cluster
3	$\lambda_k = 1 + k/100$	Slope
4	Eigenvalues of Toeplitz matrix	Toeplitz
5	$\lambda_k = 100/k$	**100 over f**
6	$\lambda_k = 1 + k/600$	**underdetermined slope**

Figure 2.14 gives the densities derived from the histograms of the accuracy scores, where the best method is the method that has most of its density on the right. In the Identity experiment the fixed point does a reasonable job although Karoui quite frequently has better scores. In the 2 cluster case, Karoui is only slightly better. In the slope configuration, fixed point outperforms Karoui, but then the bootstrap method has significantly better results. In the Toeplitz case, fixed point is only slightly better than Karoui, but again the bootstrap method outperforms both methods.

So in the experiments set-up by Karoui, the fixed point correction does not excel, but it always performs reasonably. However, in the last two experiments which are based on real life settings, the results are different: in both the 100 over f configuration and the underdetermined slope configuration the fixed point method outperforms both methods clearly.

In Figure 2.15 we show an example repetition of the underdetermined slope experiment. Figure 2.15a gives the scree plots of the population eigenvalue estimates, showing significant differences between estimates, and none of the estimates matches closely with the real population eigenvalues. We estimated the sample eigenvalues belonging to population eigenvalue estimates and show them in Figure 2.15b. Despite the differences in population eigenvalues, the sample eigenvalues seem almost identical. This suggest that the configuration is underdetermined: multiple population eigenvalue sets lead to the same sample eigenvalue set. This hypothesis is further supported by Appendix 2.3.9, which shows that if the dimensionality of the samples continues to increase, in the limit



(a) Identity

(b) 2 cluster

(c) slope

(d) Toeplitz

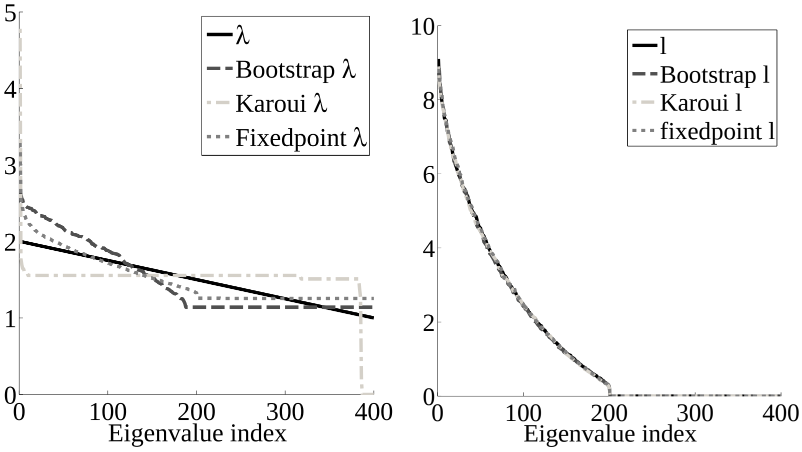
(e) 100 over f

(f) Underdetermined slope

Figure 2.14: Histograms of the Levy distance ratio between the sample eigenvalue distribution and the estimates of the population eigenvalue distributions.

only the mean of the population eigenvalues influences the sample eigenvalue distribution, all other characteristics are lost.

Furthermore, our implementation of the Karoui correction shows that several population eigenvalues are estimated as zero valued. This becomes problematic if the training results are used, for example for likelihood estimates.



(a) Population eigenvalue scree plot

(b) Sample eigenvalue scree plot

Figure 2.15: Example of a repetition with the underdetermined slope configuration.

2.3.4.3 Correction applied in verification experiments

As indicated in the previous section, bias correction can be used to improve likelihood estimates. In biometrics, a common approach to make automated verification decisions (that is, reject or accept the claim that a person has a certain identity based on a comparison of some measured characteristics with a template) is to model both the variations between samples coming from different persons and the variations between samples coming from the same person with normal distributions. The parameters of these distributions are estimated from a set of examples, the training set. For many biometric modalities the number of samples available for training is in the same order as the dimensionality of these samples.

To show that bias reduction can at least in theory improve verification performance, we did a verification experiment with synthetic data, where the parameters of the distributions from which the synthetic samples are drawn, have been set to the estimates obtained from facial image data. The dimensionality p of the facial data samples and the synthetic data is 8762. The training set contained 7047 samples of 400 individuals.

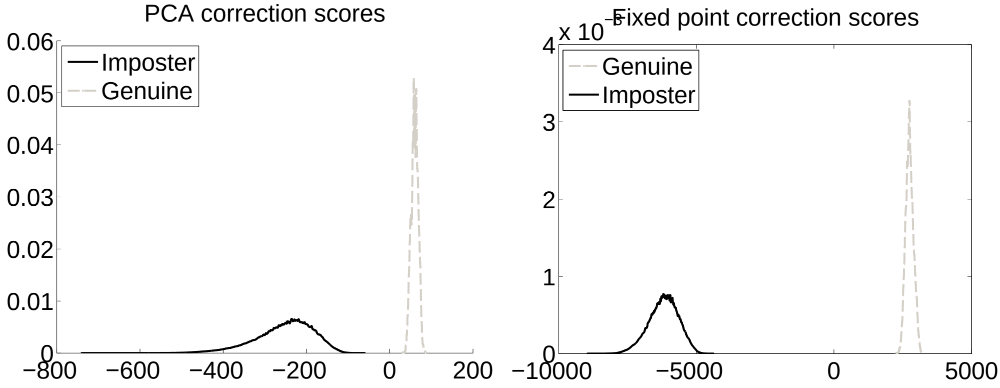
Note that both our implementation of Karoui's method and the bootstrap method can not be used. Karoui can not be used because the system is underdetermined

($N < p$), and as shown in the previous experiments, Karoui results often in zero valued eigenvalues. The evaluation of the likelihood functions requires the inverse of the covariance matrix, which can not be done if some of the eigenvalues are zero.

The bootstrap algorithm requires usually at least 25 iterations to converge, which results in a run time of several days for the values of p and N in the experimental system. This run time is unacceptable in most applications.

In verification there are two kinds of claims: genuine claims, where the claimed identity is indeed the real identity of the person, and impostor claims, where the claimed identity is not the real identity of the person. In our experiment we calculated for each claim a log likelihood ratio score, which is the logarithm of the ratio of the likelihood score that the claim is genuine over the likelihood that the claim is an impostor claim. In Figure 2.16 we show the score histograms achieved with applying classical PCA dimension reduction as bias correction (Figure 2.16a) and with applying the fixed point algorithm (Figure 2.16b) as bias correction.

The results show that the classical PCA reduction method will already result in highly separable likelihood scores (Figure 2.16a), the distance between the two clusters has increased considerably when using the fixed point eigenvalue correction (Figure 2.16b).



(a) Classical PCA dimension reduction

(b) Fixed point population correction

Figure 2.16: Histograms of the log likelihood ratio scores measured in the synthetic facial data verification experiment

We also attempted to do correction in an experiment with real face data. However, we found that correction actually decreases the verification performance. This can be explained with the error in the data model we use as we reported in [47]. The smaller eigenvalues are particularly affected by the modeling error. Since eigenvalue correction will increase these smaller eigenvalues, it can explain why the performance actually decreases.

2.3.5 Conclusion and discussion

We presented a study of estimating population eigenvalues in the case that we have a large, but not infinite, number of samples and a large, but not infinite, number of dimensions. In such problems, the sample eigenvalues are biased, but the Marčenko Pastur (MP) equation only describes the relation between the sample eigenvalue distribution and the sample eigenvalue distribution for the cases that both the number of samples and their dimensionality are infinite, so using the MP equation to remove the bias in practical problems is not straightforward.

To solve this problem, we showed that by setting $\Im\left\{z\right\}$ either small or large in the MP equation we can focus more on local details or global characteristics of the involved eigenvalue distributions, where we assumed that global characteristics converge for much lower p and N values and p and N only have to close to infinite if we are interested at very local characteristics. From these observations we derived methods, one new one to our knowledge, for estimating the sample eigenvalue density for a given set of population eigenvalues. The most important application of these methods is in a feedback algorithm which estimates the population eigenvalues from sample eigenvalues.

In the feedback algorithm, the value of $\Im\left\{z\right\}$ determines how the estimated sample eigenvalue density and the empirical distribution of the measured sample eigenvalues are smoothed before they are compared. Increasing $\Im\left\{z\right\}$ when both p and N reduces the influence of statistical noise in the correction at the price of loosing details of the population eigenvalue density.

We showed that the feedback algorithm particularly outperforms other methods in the underdetermined configurations and configurations where individual eigenvalues are of importance, such as the set described by a 1 over f distribution, which is often encountered in biometrics. In a verification experiment application of the feedback method results in a large increase of the distance between impostor and genuine scores. The difference between the synthetic scores of the classical PCA method and the scores achieved using real data suggests that the bias in the sample eigenvalues is not the only problem in face data though. Eigenvalue correction can actually increase the effect of modeling errors and therefore result in a decreased performance.

2.3.6 Proof result integration along circle stays within circle

Note that the integrals in equations (2.13) and (2.15) can be rewritten to the form:

$$\int \frac{\mathrm{d}F\left(a\left(r\right)\right)}{r - \imath q} \tag{2.31}$$

where r is a real variable, q is a real constant, F is a distribution function and a is a function $\mathbb{R} \rightarrow \mathbb{R}$. We now prove that the result of the integrals stays within the circle described by $\left(r - \imath q\right)^{-1}$, by showing that the norm of the result minus the center of the circle can never exceed the radius of the circle.

$$
\begin{aligned}
&\left|\int \frac{\mathrm{d}F\left(a\left(r\right)\right)}{r-\imath q}-\frac{\imath}{2q}\right| \\
= \quad & \left|\int \frac{1}{2q}\left(\cos\varphi\left(r\right)+\imath\left(\sin\varphi\left(r\right)+1\right)\right)\mathrm{d}F\left(a\left(r\right)\right)-\frac{\imath}{2q}\right| && (2.32) \\
= \quad & \frac{1}{2q}\left|\int \left(\cos\varphi\left(r\right)+\imath\sin\varphi\left(r\right)\right)\mathrm{d}F\left(a\left(r\right)\right)\right| && (2.33) \\
\leq \quad & \frac{1}{2q}\int \left|\cos\varphi\left(r\right)+\imath\sin\varphi\left(r\right)\right|\mathrm{d}F\left(a\left(r\right)\right) && (2.34) \\
= \quad & \frac{1}{2q} && (2.35)
\end{aligned}
$$

where $(r-\imath q)^{-1}=\frac{\imath}{2q}\left(\cos\varphi\left(r\right)+\imath\left(\sin\varphi\left(r\right)+1\right)\right)$.

2.3.7 Proof Fixed Point in Fixed Point solution

In this section we prove that the function in equation (2.19) has a fixed point. According to the Banach fixed point theorem we need to show that $d(A_{n+1}-B_{n+1})\leq q\cdot d(A_n-B_n)$ holds for any two points A and B, where $q<1$ [[45]]. We begin by evaluating the norm of the difference between both points of iteration $n+1$:

$$
\begin{aligned}
\left|A_{n+1}-B_{n+1}\right| &= \left|\gamma\sum_{k=1}^{K}\lambda_k\cdot a_k\left(\frac{A_n}{A_n+\lambda_k}-\frac{B_n}{B_n+\lambda_k}\right)\right| && (2.36) \\
&= \gamma\left|A_n-B_n\right|\left|\sum_{k=1}^{K}\frac{a_k\lambda_k^2}{\left(A_n+\lambda_k\right)\left(B_n+\lambda_k\right)}\right| && (2.37)
\end{aligned}
$$

From this we can derive an expression for the ratio which should be between 0 and 1 according to the theorem:

$$
\begin{aligned}
\frac{\left|A_{n+1}-B_{n+1}\right|}{\left|A_n-B_n\right|} &= \gamma\left|\sum_{k=1}^{K}a_k\frac{\lambda_k}{A_n+\lambda_k}\cdot\frac{\lambda_k}{B_n+\lambda_k}\right| && (2.38) \\
&\leq \gamma\sum_{k=1}^{K}a_k\left|\frac{\lambda_k}{A_n+\lambda_k}\right|\left|\frac{\lambda_k}{B_n+\lambda_k}\right| && (2.39) \\
&\leq \gamma\max\left|\frac{\lambda_k}{A_n+\lambda_k}\right|\max\left|\frac{\lambda_k}{B_n+\lambda_k}\right| && (2.40)
\end{aligned}
$$

Note that the minimum norm of A_n is determined by $\frac{1}{\max|v_\infty(z)|}=\Im\{z\}$, so the minimum norm of A_n is equal to the smoothness factor. Therefore setting the smoothness factor arbitrarily large will result in an arbitrarily low ratio of equation (2.38), guaranteeing convergence after some threshold in the value of the smoothness factor.

A minimum value of the smoothness factor can be derived after which convergence is guaranteed. Assume that $0 < \gamma < 1$. Then if both maximums in equation (2.40) get close to 1, the ratio of equation (2.38) becomes smaller than 1.

We will focus on the first ratio, since the argument on the second ratio is similar. Given $|\lambda + A_n| > \lambda$, the ratio $\left(\max \left|\frac{\lambda}{\lambda + A_n}\right|\right)$ is smaller than 1, therefore convergence is guaranteed. Setting the smoothness factor larger than $2\lambda_{\max}$, where $2\lambda_{\max} = \max \lambda_k$, will result in a minimum norm of A_n of $2\lambda_{\max}$. This results in a ratio lower than 1, guaranteeing convergence of the algorithm. There is even a lower setting for the smoothness factor since A_n attains its minimum norm when it is purely imaginary. In that case, the norm has an upper limit of $\frac{1}{\sqrt{5}}$.

2.3.8 Proof influence of parts of the distribution on the Stieltjes transform

In this section we prove that the Stieltjes transform at z is influenced mostly by changes in the density close to $\Re\{z\}$. Assume we are going to move a part of the density of $G(l)$ around position l_1 with weight β. The part we move we assume to be small enough so $G(l)$ can be written as $G(l) = (1-\beta)\tilde{G}(l) + \beta \mathrm{u}(l - l_1)$. The Stieltjes transform $m_G(z)$ can be written as:

$$\begin{aligned} m_G(z) &= \int \frac{\mathrm{d}G(l)}{l - z} && (2.41)\\ &= \int \frac{1}{l-z} \mathrm{d}\left((1-\beta)\tilde{G}(l) + \beta \mathrm{u}(l - l_1)\right) && (2.42)\\ &= (1-\beta) m_{\tilde{G}}(z) + \beta \frac{1}{l_1 - z} && (2.43)\end{aligned}$$

The derivative of $m_G(z)$ with respect to l_1, normalized for β is then given by:

$$\frac{\mathrm{d}m_G(z)}{\mathrm{d}l_1} = -\frac{1}{(l_1 - z)^2} \qquad (2.44)$$

Its absolute value is maximum when $l_1 = \Re\{z\}$. If we normalize $\left|\frac{\mathrm{d}m_G(z)}{\mathrm{d}l_1}\right|$ with its maximum value, we get:

$$\left(\frac{1}{(\Im\{z\})^2}\right)^{-1} \left|\frac{\mathrm{d}m_G(z)}{\mathrm{d}l_1}\right| = \frac{1}{\left(\frac{l_1 - \Re\{z\}}{\Im\{z\}}\right)^2 + 1} \qquad (2.45)$$

so the norm of the derivative is a function which has a maximum at $\Re\{z\} = l_1$ and its width is proportional to $\Im\{z\}$.

The sample eigenvalue density is solely determined by the imaginary part of the Stieltjes transform (equation (2.16)). The width of the function of the imaginary part of the derivative is also proportional to $\Im\{z\}$, although its extrema are not exactly at $\Re\{z\} = l_1$:

$$\begin{aligned} \Im\left\{\frac{1}{\beta}\frac{\mathrm{d}m_G(z)}{\mathrm{d}l_1}\right\} &= \frac{-2\,\Im\{z\}\,(l_1-\Re\{z\})}{\left((l_1-\Re\{z\})^2+(\Im\{z\})^2\right)^2} && (2.46)\\ &= (\Im\{z\})^{-2}\,\frac{-2\,t}{(t^2+1)^2} && (2.47)\end{aligned}$$

where $t = \frac{l_1-\Re\{z\}}{\Im\{z\}}$.

2.3.9 Underdetermination in high dimensional problems

In the experiments we suggest that if the number of samples N is below the dimensionality of the samples p, the correction of the sample eigenvalues is an underdetermined problem. In the following section we prove that if $\gamma \to \infty$, all characteristics of the population eigenvalue distribution are lost except for its mean, showing that in that limit the sample eigenvalue correction is indeed a severely underdetermined problem. Because $H(\lambda)$ describes the population eigenvalues, $H(\lambda) = 0\,\forall\,\lambda \leq 0$. We also assume the eigenvalues have a supremum $\lambda_{\sup}$, so $H(\lambda) = 0\,\forall\,\lambda > \lambda_{\sup}$.

We start with proving that $O(\|v_\infty(z)\|) = O(\gamma^{-1})$. We first show that $O(\|v_\infty(z)\|) = O(\gamma^a)$, with $a > 0$ leads to a contradiction. Firstly note that $O\left(\left\|\frac{1}{v_\infty(z)}\right\|\right) = O(\gamma^{-a})$. Then note that:

$$\begin{aligned} O\left(\left\|z-\gamma\int\frac{\lambda\mathrm{d}H(\lambda)}{1+\lambda v_\infty(z)}\right\|\right) &= O\left(\left\|z-\gamma\int\frac{\lambda\mathrm{d}H(\lambda)}{\lambda v_\infty(z)}\right\|\right)\\ &= O\left(\left\|z-\gamma\frac{1}{v_\infty(z)}\int\mathrm{d}H(\lambda)\right\|\right)\\ &= O\left(\left\|z-\gamma\frac{1}{v_{\infty(z)}}\right\|\right)\\ &= O\left(\gamma^{1-a}\right)\end{aligned}$$

So the assumption $O(\|v_\infty(z)\|) = O(\gamma^a)$ with $a > 0$ leads to the contradiction that $O\left(\left\|\frac{1}{v_\infty(z)}\right\|\right)$ is both $O(\gamma^{-a})$ and $O\left(\gamma^{1-a}\right)$.

We now assume that $O(\|v_\infty(z)\|) = O\left(\gamma^0\right)$. Note that $O\left(\left\|\frac{1}{v_\infty(z)}\right\|\right) = O\left(\gamma^0\right)$.

$$\begin{aligned} O\left(\left\|z-\gamma\int\frac{\lambda\mathrm{d}H(\lambda)}{1+\lambda v_\infty(z)}\right\|\right) &= \left|O\left(\gamma^0\right)-O(\gamma)\cdot O\left(\gamma^0\right)\right|\\ &= O(\gamma)\end{aligned}$$

So this is again a contradiction: $O\left(\left\|\frac{1}{v_\infty(z)}\right\|\right)$ should be both $O\left(\gamma^0\right)$ and $O\left(\gamma^1\right)$.

Now we assume that $O\left(\|v_{\infty}(z)\|\right) = O\left(\gamma^{a}\right)$ with $a < 0$. Note that $O\left(\left\|\frac{1}{v_{\infty}(z)}\right\|\right) = O\left(\gamma^{-a}\right)$.

$$\begin{aligned} O\left(\left\|z - \gamma\int\frac{\lambda \mathrm{d}H(\lambda)}{1+\lambda v_{\infty}(z)}\right\|\right) &= O\left(\left\|z-\gamma\int\lambda\mathrm{d}H(\lambda)\right\|\right) \\ &= \left|O(1) - O(\gamma)\cdot O(1)\right| \\ &= O(\gamma) \end{aligned}$$

So if we set $a = -1$ both arguments result in $O\left(\left\|\frac{1}{v_{\infty}(z)}\right\|\right) = O(\gamma)$, or $O\left(\|v_{\infty}(z)\|\right) = O\left(\gamma^{-1}\right)$.

Using the fact that $O\left(\|v_{\infty}(z)\|\right) = O\left(\gamma^{-1}\right)$ we can determine the sample eigenvalue distribution if $\gamma \to \infty$.

$$\begin{aligned} \lim_{p\to\infty} -\frac{1}{v(z)} &= \lim_{\gamma\to\infty} z - \gamma\int\frac{\lambda\mathrm{d}H(\lambda)}{1+\lambda v(z)} \\ &= \lim_{\gamma\to\infty} z - \gamma\int\frac{\lambda\mathrm{d}H(\lambda)}{1} \\ &= \lim_{\gamma\to\infty} z - \gamma\bar{\lambda} \end{aligned}$$

So in the limit the Stieltjes transform $v(z)$ converges to

$$v(z) = \frac{1}{\gamma\bar{\lambda} - z}$$

which is the Stieltjes transform of $\mathring{G}(x) = u\left(x - \gamma\bar{\lambda}\right)$, so the sample eigenvalue set will converge to a set of n eigenvalues equal to $\gamma\bar{\lambda}$ and $p-n$ eigenvalues equal to 0, whatever the population eigenvalue distribution, if it has a bounded support. little is known about this process beforehand and the description consisting

2.4 High dimensionality and eigenvector estimation [3]

2.4.1 Prologue

The bias in the sample eigenvalues is not the only error introduced by a high p over N ratio; the sample eigenvectors are affected as well. Ideally the sample eigenvectors are aligned with the population eigenvectors, but with a high p over N ratio, this is typically not the case. This misalignment can not be corrected, however using the inner products between the sample eigenvectors and the population eigenvectors, which can be estimated, an additional correction to the estimated eigenvalues can improve the variance estimate along the sample eigenvectors. We show that under the Kulback Leibler divergence, this improves SOS estimates.

[3] Improved variance estimation along sample eigenvectors, Proceedings of the 30th Symposium on Information Theory in the Benelux, 2009 [37]

The introduction gives again an introduction on GSA, but it also shows how perfect bias correction still does not lead to a perfect estimate of the SOS. This analysis provides a framework on which the additional adjustment of the corrected eigenvalues is based. It is therefore recommended to read the following paper from start.

2.4.2 Introduction

An important aspect in statistic learning is the modeling of data distributions. A common assumption is that the distribution of the data after some preprocessing can be characterised by the second order statistics. The second order statistics of a multidimensional distribution are described by a covariance matrix, which we denote the population covariance matrix. The population covariance matrix is given by $\mathbf{\Sigma} = \mathcal{E}\left(\underline{\tilde{\mathbf{x}}} \cdot \underline{\tilde{\mathbf{x}}}^{\mathrm{T}}\right)$, with $\underline{\tilde{\mathbf{x}}} = \underline{\mathbf{x}} - \mathcal{E}(\underline{\mathbf{x}})$, where $\mathcal{E}()$ is the expectation operator and $\underline{\mathbf{x}}$ is a random variable representing the data generating process.

The covariance matrix can be decomposed such that $\mathbf{\Sigma} = \mathbf{E} \cdot \mathbf{D} \cdot \mathbf{E}^{\mathrm{T}}$, where $\mathbf{E}$ is a rotation matrix. Each column of $\mathbf{E}$ is an eigenvector. Diagonal matrix $\mathbf{D}$ has the eigenvalues of $\mathbf{\Sigma}$ at the diagonal. The i^{th} eigenvector points in the direction with the largest variance after the subspace formed by the first $i-1$ eigenvectors has been removed. The i^{th} eigenvalue gives the variance in this direction. As an example, the black curve in figure 2.17 shows the variance of a two dimensional, zero mean distribution with eigenvalues 2 and 1. The first eigenvector is along the vertical axis and the second eigenvector is along the horizontal axis.

Usually, the population covariance matrix is unknown beforehand and needs to be estimated from a set of examples, the training set. A commonly used estimator is the sample covariance matrix: $\hat{\mathbf{\Sigma}} = \frac{1}{N-1}\mathbf{X} \cdot \mathbf{X}^{\mathrm{T}}$, where each of the N columns of $\mathbf{X}$ consists of a sample from the training set with the mean of the set subtracted. The decomposition results of the sample covariance matrix are denoted by sample eigenvectors and sample eigenvalues. As a result of the limited number of examples, the estimates contain errors. For the sample covariance matrix, these errors have a zero mean: the estimator is unbiased. However, the sample eigenvalues are a biased estimate of the population eigenvalues. We have shown previously that reduction of this bias is possible [48]. In this paper we argue that even if the entire bias in the sample eigenvalues were removed, the combination of corrected sample eigenvalues and uncorrected sample eigenvectors still provides a sub optimal description of the population distribution. We present a study on another solution to correct the sample eigenvalues. The results of the study cannot be used directly in practice because we make extensive use of the population eigenvectors and these are usually unknown in estimation problems. A practical approach is shown in our work in [48]. Because our interest lies in the area of biometrics, some of the experiments use data models used in biometrics.

The remainder of the paper is outlined as follows: in the next section we present some analysis on the sample eigenvalues and eigenvectors, showing that besides the bias in the eigenvalues, the inner product of the sample eigenvectors with the population eigenvalues also shows some sort of bias. This leads to

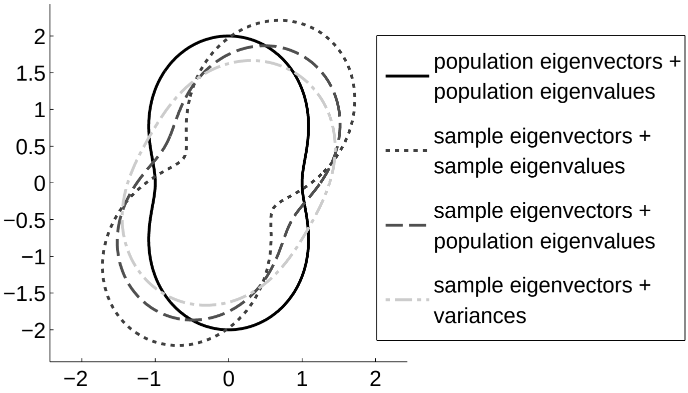


Figure 2.17: Variance curves of approximations of an example distribution.

a new method of replacing the sample eigenvalues: the variance estimate. In section 2.4.4 we compare replacement of the sample eigenvalues with the population eigenvalues and replacement with variances experimentally, using the Kullback Leibler divergence as an evaluation criterion in section 2.4.4.1 and verification scores in section 2.4.4.2. We draw conclusions in section 2.4.5.

2.4.3 Eigenvector and eigenvalue analysis

2.4.3.1 Eigenvalue analysis

Estimators attempt to obtain parameters from a limited number of samples. As a result, the estimate contains errors. To describe these errors, often Large Sample Analysis (LSA) is performed on the estimator. In LSA the limit behaviour of the error of the estimator is described under the assumption that the number of samples grows to infinity. In LSA, the sample eigenvalues show no bias. However, in biometrics the number of samples (N) is limited and in the same order as the number of dimensions (p) or even lower. Therefore LSA should not be applied.

In General Statistics Analysis (GSA) another limit is considered: $N, p \to \infty$ while $\frac{p}{N} \to \gamma$, where γ is some positive constant. In this limit, the sample eigenvalues do show a bias and a relation between the sample eigenvalues and the population eigenvalues is given for a large class of data distributions by the Marčenko Pastur equation [26]. The bias is such that the largest sample eigenvalue is larger than the largest population eigenvalue and the smallest sample eigenvalue is smaller than the smallest population eigenvalue. In figure 2.17 the dotted line gives the variance estimates if the sample eigenvalues are biased estimates and the sample eigenvectors are off as well.

In the remainder of the article we assume that the bias can be fully compensated for and the remaining fluctuations in the sample eigenvalues are minimal, allowing perfect estimation of the population eigenvalues. However, combining the population eigenvalues with the sample eigenvectors still provides an estimate of the distribution containing significant deviations caused by deviations between sample and population eigenvectors. In the example this is shown by the dashed curve in figure 2.17.

2.4.3.2 General Statistical Analysis of the sample eigenvectors

We expect that the sample eigenvectors are generally off compared to the population eigenvectors as well when studied in GSA. To our knowledge there is no theory available which describes the relation between sample eigenvectors and population eigenvectors similar to the Marčenko Pastur equation, but Anderson did derive an expression for the distribution of the sample eigenvectors if the data is Gaussian [20]. Mestre presented some results on the estimations of subspaces belonging to the eigenvalues in the GSA limit in [49]. In [41] the inner product between the eigenmatrix and a random unitary vector is considered. Our approach differs because we do not consider a random vector, but the population eigenvectors.

To demonstrate that the sample eigenvectors are generally off compared to the population eigenvectors, we did a synthetic eigenvector estimation experiment. In the experiment, synthetic Gaussian data is generated with all eigenvalues chosen uniformly between 0 and 1. The sample covariance matrix is determined, which is decomposed to find the sample eigenvectors. We then determined the component of a sample eigenvector $\hat{\mathbf{E}}_{:,m}$ in the subspace spanned by the K population eigenvectors $\left\{\mathbf{E}_{:,k} | k = 1 \ldots K\right\}$ with the smallest population eigenvalues, which is given by $\sqrt{\sum_{k=1}^{K} \left(\hat{\mathbf{E}}_{:,m}^{\mathrm{T}} \mathbf{E}_{:,k}\right)^2}$.

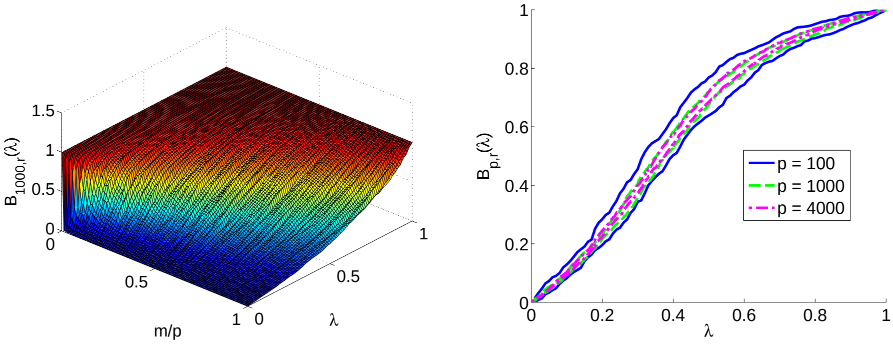
If we repeat this procedure for $K = 0 \ldots p$ we get a function which starts at 0 and increases to 1 for $K = p$. If we consider this component as a function of the largest population eigenvalue instead of the largest index, we can describe the components as a distribution function $B_{p,m}(\lambda)$ given by:

$$B_{p,m}(\lambda) = \sqrt{\sum_{k=1}^{p} \left(\hat{\mathbf{E}}_{:,m}^{\mathrm{T}} \mathbf{E}_{:,k}\right)^2 u(\lambda - \lambda_k)} \tag{2.48}$$

where $u()$ is the step function and λ_k is the k^{th} element of a vector with the population eigenvalues sorted from small to large. We hypothesize that under the GSA limit, $B_{p,m}(\lambda)$ like the eigenvalue distributions converges to a fixed distribution function.

Figure 2.18a shows the distribution functions for all sample eigenvectors. In the experiment $p = 1000$ and $\gamma = \frac{1}{2}$. The sample eigenvalue index is scaled with p so it ranges from 0 to 1. As can be seen from the figure, the smallest sample eigenvectors are confined to the space of the smallest population eigenvectors. The largest sample eigenvectors have a component in almost all population eigenvectors.

Figure 2.18b shows why we hypothesize that $B_{p,m}(\lambda)$ converges in the GSA limit, similar to the eigenvalues. In the figure, we show the mean plus and



(a) Distribution of the component of the sample eigenvectors in the population eigenvector subspace.

(b) Convergence example of $B_{p,\mathbf{r}}(\lambda)$. The curves show the mean plus and minus one standard deviation of $B_{p,\mathbf{r}}(\lambda)$ based on a number of repetitions per configuration.

Figure 2.18: Inner product example of sample eigenvector variation

minus a standard deviation of the middle sample eigenvector for three different configurations: $p = 100$, $p = 1000$ and $p = 4000$, all with $\gamma = \frac{1}{2}$, with the number of repetitions equal to 20, 8 and 4 respectively. The mean curves were all almost indistinguishable, while the standard deviation curves tighten around the mean. It seems that the distribution converges to a fixed distribution in the GSA limit.

Because the sample eigenvector is not in the same direction as the population eigenvector, the corresponding population eigenvalue will in general be an erroneous estimate of the variance along the sample eigenvector. We therefore propose to replace the sample eigenvalues with the variance along the sample eigenvectors ($\mathbf{v}_m$ for sample eigenvector $\hat{\mathbf{E}}_{:,m}$):

$$\mathbf{v}_m \quad = \quad \sum_{k=1}^{p} \left(\left(\hat{\mathbf{E}}_{:,m}^{\mathrm{T}} \cdot \mathbf{E}_{:,k} \right)^2 \lambda_k \right) \tag{2.49}$$

In the example in figure 2.17 this correction leads to the dash-dotted curve. Since this is still not a perfect match with the real distribution, the question is which of the two sample eigenvalue replacements is better: population eigenvalues or variances.

2.4.4 Experimental comparison between population eigenvalue substitution and variance substitution

To compare sample eigenvalue replacement by population eigenvalues with replacement by variances a measure is required which describes how close a replacement distribution is to the population distribution. We chose two measures: the Kullback Leibler divergence and verification rates. With both measures we did an experiment on synthetic data.

2.4.4.1 Comparison based on the Kullback Leibler divergence

The first experiment compares the replacement with population eigenvalues and the replacement with variances by determining the Kullback Leibler divergence[50] between the replacement distributions and the population distribution. For the experiment we generate Gaussian distributed samples. The Kullback Leibler divergence between two Gaussian distributions N_0 and N_1 with covariance matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ respectively and equal mean is given by [51]:

$$d_{\mathrm{KL}}\left(N_0\,|N_1\right) \quad = \quad \frac{1}{2}\left(\log\left(\frac{\det\mathbf{\Sigma}_1}{\det\mathbf{\Sigma}_0}\right) + \mathrm{tr}\left(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0\right) - p\right) \tag{2.50}$$

In the experiment, N_1 was always the population distribution and N_0 was the replacement distribution N_{eig} for replacement with population eigenvalues and N_{var} for replacement with variances.

We used three different population eigenvalue distributions:

- 2 cluster: half of the population eigenvalues have a value of 2 and the other half have a value of 1.
- slope: the population eigenvalues are distributed uniformly between 1 and 2.
- 1/f: the population eigenvalues are set to 1/f, where f is the index of the eigenvalue.

In all configurations we generated 500 samples with a dimensionality of 100. The first two configurations are therefore similar to the experiments by Karoui [27]. The 1/f is an eigenvalue model often used in biometrics. We repeated the experiments 100 times for each configuration.

We determined the divergence improvement between the eigenvalue replacement and the variance replacement by

$$s\left(N_{eig}, N_{var}\right) \quad = \quad \left(d_{KL}\left(N_{eig}|N_1\right) - d_{KL}\left(N_{var}|N_1\right)\right)/d_{KL}\left(N_{eig}|N_1\right) \tag{2.51}$$

Figure 2.19 shows the average and standard deviations of $s\left(N_{eig}, N_{var}\right)$. In both the 2 cluster and the slope configuration the replacement with variances gives a considerable improvement of the density estimation. Even though the improvement in the 1/n configuration is not as large, using variances instead of population eigenvalues is still better.

2.4.4.2 Comparison based on verification experiments

In the second experiment we performed a verification experiment with synthetic data. Repeating the experiment for both the replacement with population eigenvalues and the replacement with variances gives verification rates for both methods. The objective is to determine which method gives the best verification performance.

We designed a verification experiment based on the LDA model, in which all classes are distributed with a normal distribution which only has a different mean

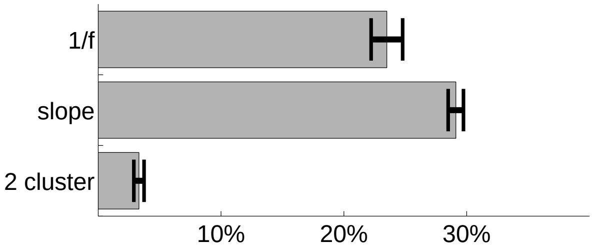


Figure 2.19: Kullback Leibler divergence reduction when replacing sample eigenvalues with variances instead of population eigenvalues. The gray bars indicate the average improvement, while the thick black lines indicate the standard deviation.

per class. We generated a training set with 150 classes with 2 samples per class. The class means ($\boldsymbol{\mu}_c$) have a normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_b$, the samples within class c are distributed $N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_w)$, a normal distribution with mean $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_w$.

For testing we generated a set of probe samples distributed with the same distribution parameters as used for generating the training set, containing 400 classes and 2 samples per class. For enrollment we used the real class means instead of estimating it from a number of samples. All data sets had a dimensionality of 100.

Since LDA and classification is only influenced by the ratio between the within and the between class distributions, one of the distributions can be fixed. We therefore chose to fix $\boldsymbol{\Sigma}_b$ at $\frac{1}{10}\mathbf{I}$, where $\mathbf{I}$ is an identity matrix. For $\boldsymbol{\Sigma}_w$ we can leave the eigenvector matrix also at $\mathbf{I}$ and only have to chose an eigenvalue distribution. We tested four eigenvalue distributions:

- 2 cluster: half of the eigenvalues equal to 0.5 and the other half equal to 1.
- slope distribution: all eigenvalues chosen uniformly between 0.5 and 1.
- 1/f: the f$^{\text{th}}$ eigenvalue is set to 30/f.
- toeplitz: the eigenvalues of the matrix with elements $0.5^{|i-j|}$, where i and j are the row and column indices respectively.

During training, instead of using averages of the class samples, we used the real class means as estimates, so that the estimate of the between class covariance matrix does not contain crosstalk of the within class covariance matrix caused by fluctuations in the class mean estimates. After estimating the within class and between class covariance matrices, we performed four corrections: in the first correction, we kept the sample eigenvalues. In the second correction, we replaced the sample eigenvalues with the population eigenvalues. In the third correction, we replaced the sample eigenvalues with variances. In the last "correction", we used the population eigenvectors and population eigenvalues to get the results in case of correct estimation.

In the verification experiments, we used the log likelihood ratio (equation 2.52) to determine whether a sample originated from the claimed class.

$$L(\mathbf{x}, c) \quad = \quad -(\mathbf{x} - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_w^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) + (\mathbf{x} - \boldsymbol{\mu}_t)^{\mathrm{T}} \boldsymbol{\Sigma}_t^{-1} (\mathbf{x} - \boldsymbol{\mu}_t) \qquad (2.52)$$

If the likelihood ratio for a sample $\mathbf{x}$ and class c combination is above a threshold, the sample is considered to belong to class c. By changing the threshold a trade-off can be made between the ratio of samples incorrectly accepted to a class (False Accept Ratio (FAR)) and the ratio of samples erroneously rejected from a class (False Reject Ratio (FRR)).

In figure 2.20 we show the results of the verification experiments. The curves on the left show Detection Error Trade-off (DET) curves, which plot FRR against the corresponding FAR. In the right column we show the within class eigenvalues for the different corrections. The between class eigenvalues are the same for the eigenvalue correction and the variance correction, namely all $\frac{1}{10}$. In the left column we also show the classification results when the population parameters are used, so the population eigenvalues as well as the population eigenvectors.

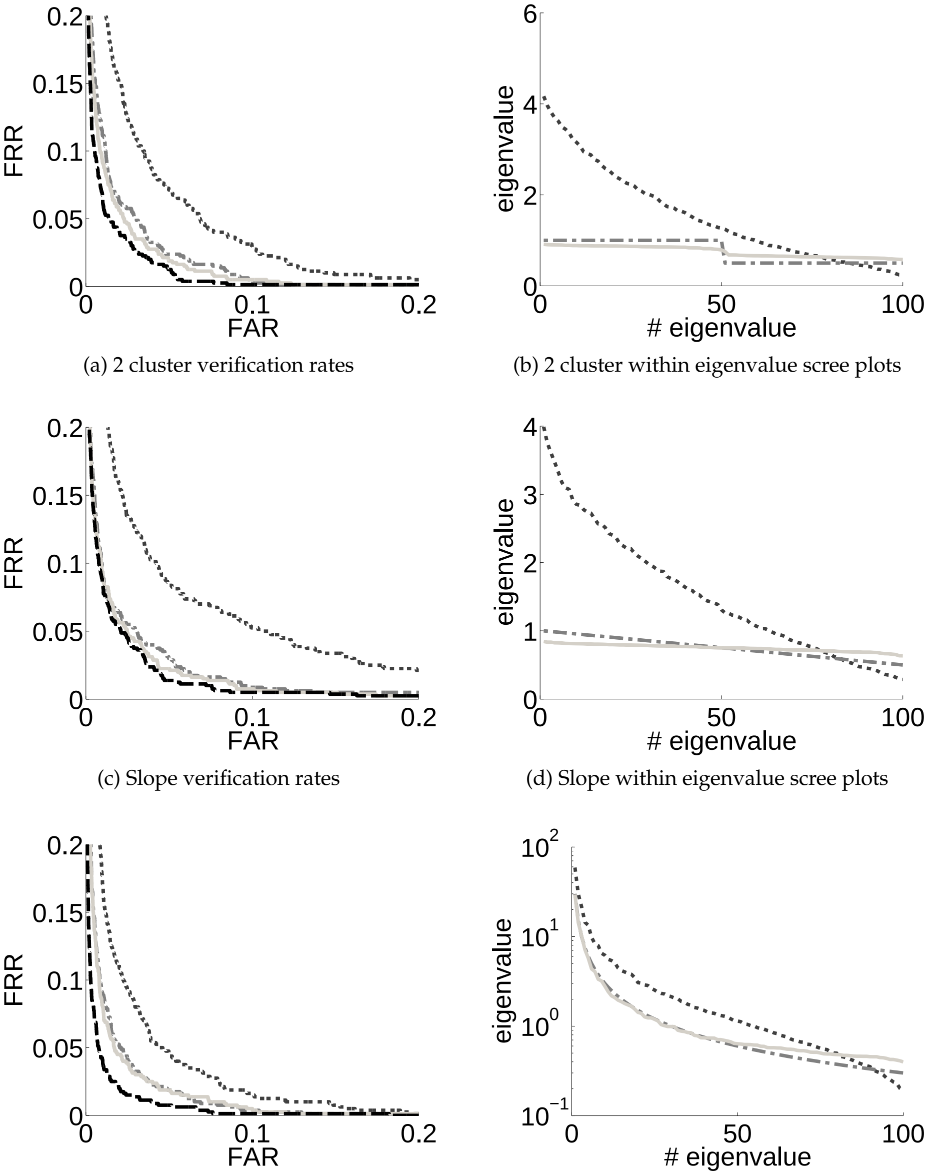
In the DET curves there is a large reduction between no correction and eigenvalue correction in all configurations. The difference between the eigenvalue correction and the variance correction is much smaller, but as the population correction shows, the eigenvalue correction is already much closer to the best estimate than the original sample estimate. The variance correction gives a significant improvement in both the 2 cluster configuration and the slope configuration (also if the experiment is repeated a number of times). The Toeplitz results also seem to have improved with the variance correction, however the difference is too small to be significant.

The scree plots on the right show a general trend we expected: the variance correction reduces the largest eigenvalues while increasing the smallest eigenvalues. With the 1/f configuration however the modification is small. With the Toeplitz, the modification is considerable even though it has only a small effect on the verification rates.

2.4.5 Conclusion & discussion

The combination of population eigenvalues with sample eigenvectors does not lead to an optimal description of the population distribution. We showed that the sample eigenvectors will be off compared to the population eigenvectors, which makes the population eigenvalues a biased variance estimate along the sample eigenvectors. For example, if there is just one population eigenvalue with the maximum value, then none of the sample eigenvectors will be along the corresponding population eigenvector and therefore the maximum variance along a sample eigenvector will be smaller than the maximum population eigenvalue.

We suggested to replace the sample eigenvalues with the variances along the sample eigenvectors. In the experiments, using the variances instead of the population eigenvalues showed as good as and often better estimates of the population distribution. The 2 subset method in [48] already showed a practical implementation of using variances.



(a) 2 cluster verification rates

(b) 2 cluster within eigenvalue scree plots

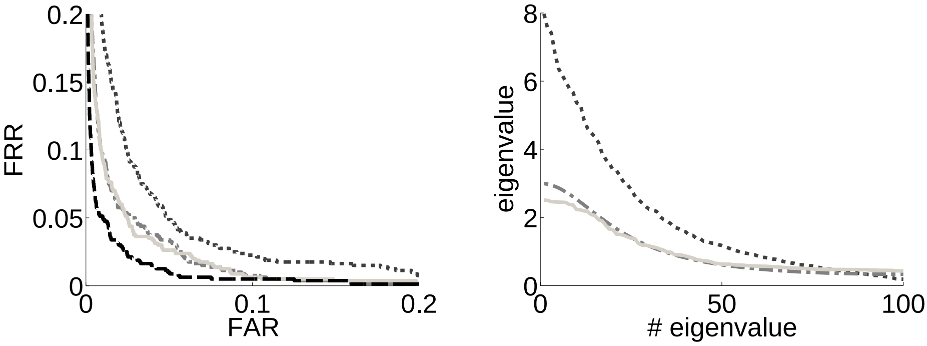
(c) Slope verification rates

(d) Slope within eigenvalue scree plots

(e) 1/f verification rates

(f) 1/f within eigenvalue scree plots

Figure 2.20: Verification experiment results. The curves are the results as follows: dotted = sample eigenvectors with sample eigenvalues, dash dotted = sample eigenvectors with population eigenvalues, solid = sample eigenvectors with variances and dashed = population eigenvectors with population eigenvalues.



(g) Toeplitz verification rates

(h) Toeplitz within eigenvalue scree plots

Figure 2.20: Verification experiment results (cont.)

2.5 Conclusion

In this chapter we improved the sample estimate of a single distribution, by adjusting the estimates of the eigenvalues. This is the first stage of answering the research question: "What (potential) effects does the sample eigenvalue bias have on verification systems and can these effects be reduced?" We first focussed on removing the bias of the sample eigenvalues as estimates of the population eigenvalues and presented two methods for correcting this bias: the bootstrap correction and the fixed point bias correction. The bootstrap method is relatively easy to implement and depends only marginally on theoretical analysis of the bias in the eigenvalues. Its major drawback is that it will always increase the required computer resources by a factor considerably larger than 1 if it is added to an eigenvalue decomposition method.

The fixed point correction depends heavily on the theoretical analysis of the eigenvalue bias, but its requirements on computer resources can be much lighter than the bootstrap correction. Moreover, the method also has a well defined approach to estimation problems with a limited N and p of the training set.

In the next chapter we will focus on the second stage of answering the above mentioned research question and use bias correction in a verification setting. With verification at least two distributions are involved and in particular the relation between these two distributions are of concern. The application of bias correction in verification requires therefore some additional steps.

Chapter 3

Verification and eigenvalue bias

3.1 Introduction

In the previous chapter we discussed how the sample estimate of the SOS of a single distribution could be improved by modifying the sample eigenvalues. The first step was to reduce the bias in the sample eigenvalues, after which in a second step the variances along the sample eigenvectors could be determined. Whether this second step results in better estimates of the SOS depends on the application.

These steps focus on the improving the estimate of a single distribution, getting an as accurate as possible likelihood estimate for a sample from that distribution based on SOS. To answer the research question "What (potential) effect does the sample eigenvalue bias have on verification systems and can these effects be reduced?" we have to study the eigenvalue bias in a verification setting. In verification problems, the likelihood ratio is usually defined by the ratio between the likelihood of a sample based on two different distributions, so the accuracy of a likelihood estimate from a single distribution is not the only concern. We will show that some additional adjustments to the SOS estimates are required in a verification setting.

This chapter contains two papers in which we study the application of bias correction in verification (BCIV). In the first paper (section 3.3) we use several bias correction methods and correct the SOS estimates of the within class variations and the between class variations. This approach we denote with Naive BCIV. The paper leads to some remarkable conclusions:

1. Correction of the within eigenvalues led to significant performance increase while the correction of the between eigenvalues even led to a reduced performance. As stated in the conclusions we suspected at the time that the correction of the within eigenvalues was ok, while the correction of the between eigenvalues led to some problems.
2. After bias correction, still PCA dimensionality reduction (which is itself based on SOS estimation) is required. Bias correction is supposed to correct the SOS

as much as possible for the effects of the limited N. It appears that the current bias correction is not sufficient in solving the problems caused by the limited N.

3. Bias correction with real facial data performs significantly worse compared to bias correction applied to synthetic data.

4. Most correction methods provide little improvement.

Point 1 and partially point 2 will be explained in the second paper of this chapter, presented in section 3.4. In that paper we determine that in verification especially the correction of the between eigenvalues should not be done independently from the within eigenvalue estimates and we introduce a new BCIV method we denote with eigenwise correction which in particular adjusts the correction of the between eigenvalues to take the relation between the within and between eigenvalues into account.

Point number 4 led us to suspect that the existing bias correction methods are not very accurate and hence we tried to improve some of them which led to the methods presented in chapter 2.

Points 2 and 3 led us to question how accurate the model assumed in SOS estimation is. In chapter 4 we will discuss the model assumption and introduce the position sources model which heavily distorts the SOS estimation.

However, before we can start with the main topic of this chapter, a few issues in SOS estimation and verification have to be discussed. The first issue involves a difference in the objective of the verification problem. One view is that with verification the objective is to classify genuine verification attempts from imposter verification attempts. The other view is that the objective is to classify a genuine attempt from the background distribution of all attempts. The first view requires the genuine distribution and a distribution equal to the background distribution, but with the genuine attempts removed, while the second view requires only the genuine attempts and the distribution of all attempts. Section 3.2.1 shows that this mostly a semantic issue which does not affect verification performance.

The second issue is that if a limited number of samples per class is used during training then the SOS estimate of the between class variations contains a crosstalk part of the SOS of the within class variations. This has to be taken into account when determining theoretical optimum estimation results as we will do in the following papers. Moreover, the random fluctuations in the SOS estimates of the within class variations are independent from the random fluctuations in the crosstalk part. Therefore this crosstalk part can sometimes completely determine the verification results. This is discussed in section 3.2.2.

The last issue is about the distribution of the genuine attempts. Due to a limited number of samples used for determining the class centers, the distribution of the genuine attempts is not equal to the within class variations. This again affects the theoretical optimums and verification results in general. This is discussed in detail in section 3.2.2.

3.2 Other issues in Second Order Statistics besides eigenvalue bias [1]

3.2.1 Order equivalence of likelihood ratios

Verification systems are often based on the likelihood ratio. The ratio is based on hypothesis testing. In the case of verification, which hypothesises are tested depends on the view by the designer. In one view it is assumed that hypothesis that an verification attempt is a genuine attempt has to be tested against the hypothesis that a verification attempt is an imposter attempt. In this case the 0 hypothesis is that sample $\boldsymbol{x}$ is originating from class c. Hypothesis 1 is that sample $\boldsymbol{x}$ is not originating from class c. The likelihood of hypothesis 0 is given by $p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}=c\,\right)$. The likelihood of hypothesis 1 is $p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}\neq c\,\right)$. The corresponding likelihood ratio is given by equation 3.1.

$$L_{\mathrm{w}}\left(\boldsymbol{x},c\right) = \frac{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}=c\,\right)}{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}\neq c\,\right)} \tag{3.1}$$

But the density of $p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}\neq c\,\right)$ is usually not available and has be determined from other estimates. One method to determine this density is by:

$$p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}\neq c\,\right) = \frac{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\right)-p\left(\bar{c}=c\right)\cdot p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}=c\,\right)}{p\left(\bar{c}\neq c\right)} \tag{3.2}$$

We will show that this is unnecessary. The density of hypothesis 1 can be replaced with $p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\right)$ which leads to equation 3.3. This will only change the value of the likelihood ratio, the order of the samples will not change: if sample $\boldsymbol{x}_1$ had a lower, equal or higher ratio in equation 3.1, it also has a lower, equal or higher ratio in equation 3.3.

This new ratio is actually the likelihood ratio corresponding to the view that the verification attempt should be tested as either being a genuine attempt or a random attempt.

$$L_{\mathrm{t}}\left(\boldsymbol{x},c\right) = \frac{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}=c\,\right)}{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\right)} \tag{3.3}$$

To prove that the ratios in equation 3.1 and equation 3.1 are order equivalent, we first substitute equation 3.2 in 3.1:

$$\begin{aligned} L_{\mathrm{w}}\left(\boldsymbol{x},c\right) &= \frac{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}=c\,\right)\cdot p\left(\bar{c}\neq c\right)}{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\right)-p\left(\bar{c}=c\right)\cdot p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}=c\,\right)} && (3.4)\\ &= p\left(\bar{c}\neq c\right)\cdot\frac{1}{\frac{p(\bar{\boldsymbol{x}}=\boldsymbol{x})}{p(\bar{\boldsymbol{x}}=\boldsymbol{x}|\bar{c}=c\,)}-p\left(\bar{c}=c\right)} && (3.5)\end{aligned}$$

[1] Notes on Second Order Statistics in Verification, Technical Report, 2011 [52]

We define 3 intermediate likelihood ratios from which the order preservation or inverse ordering easily follows:

$$L'_{\mathrm{w}}\left(\boldsymbol{x},c\right) = \frac{1}{\frac{p(\bar{\boldsymbol{x}}=\boldsymbol{x})}{p(\bar{\boldsymbol{x}}=\boldsymbol{x}|\bar{c}=c\,)} - p\left(\bar{c}=c\right)} \tag{3.6}$$

$$L'_{\mathrm{w}}\left(\boldsymbol{x}_1,c\right) \underset{<}{\overset{>}{=}} L'_{\mathrm{w}}\left(\boldsymbol{x}_2,c\right) \Rightarrow L_{\mathrm{w}}\left(\boldsymbol{x}_1,c\right) \underset{<}{\overset{>}{=}} L_{\mathrm{w}}\left(\boldsymbol{x}_2,c\right) \tag{3.7}$$

$$L''_{\mathrm{w}}\left(\boldsymbol{x},c\right) = \frac{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\right)}{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}=c\,\right)} - p\left(\bar{c}=c\right) \tag{3.8}$$

$$L''_{\mathrm{w}}\left(\boldsymbol{x}_1,c\right) \underset{<}{\overset{>}{=}} L''_{\mathrm{w}}\left(\boldsymbol{x}_2,c\right) \Rightarrow L'_{\mathrm{w}}\left(\boldsymbol{x}_1,c\right) \underset{>}{\overset{<}{=}} L'_{\mathrm{w}}\left(\boldsymbol{x}_2,c\right) \tag{3.9}$$

$$L'''_{\mathrm{w}}\left(\boldsymbol{x},c\right) = \frac{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\right)}{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}\,|\bar{c}=c\,\right)} \tag{3.10}$$

$$L'''_{\mathrm{w}}\left(\boldsymbol{x}_1,c\right) \underset{<}{\overset{>}{=}} L'''_{\mathrm{w}}\left(\boldsymbol{x}_2,c\right) \Rightarrow L''_{\mathrm{w}}\left(\boldsymbol{x}_1,c\right) \underset{<}{\overset{>}{=}} L''_{\mathrm{w}}\left(\boldsymbol{x}_2,c\right) \tag{3.11}$$

To get to L_{t} from L'''_{w} an inversion has to be applied, after which the order sought preservation follows:

$$L_{\mathrm{t}}\left(\boldsymbol{x}_1,c\right) \underset{<}{\overset{>}{=}} L_{\mathrm{t}}\left(\boldsymbol{x}_2,c\right) \Rightarrow L'''_{\mathrm{w}}\left(\boldsymbol{x}_1,c\right) \underset{>}{\overset{<}{=}} L'''_{\mathrm{w}}\left(\boldsymbol{x}_2,c\right) \tag{3.12}$$

$$\Rightarrow L_{\mathrm{w}}\left(\boldsymbol{x}_1,c\right) \underset{<}{\overset{>}{=}} L_{\mathrm{w}}\left(\boldsymbol{x}_2,c\right) \tag{3.13}$$

So only the threshold values change if equation 3.3 is used instead of equation 3.1 in verification: no difference occurs in which genuine and impostor pairs are accepted.

3.2.2 Crosstalk of within covariance matrix on the between covariance matrix estimate

In this section we will demonstrate that the between class covariance matrix is actually an estimate of a mixture of the between class covariance matrix and the within class covariance matrix, so the between class estimate suffers from crosstalk of the within class covariance matrix.

In our data model we assumed that each sample is composed of a within and a between part, via

$$\boldsymbol{x}_k = \boldsymbol{x}_{\mathrm{w},k} + \boldsymbol{\mu}_{l(\boldsymbol{x}_k)} \tag{3.14}$$

The class means are estimated via

$$\begin{aligned}\hat{\boldsymbol{\mu}}_c &= \frac{1}{C}\sum_{l(\boldsymbol{x}_k)=c} \boldsymbol{x}_k = \boldsymbol{\mu}_c + \frac{1}{C}\sum_{l\left(\boldsymbol{x}_{\mathrm{w},k}\right)=c} \boldsymbol{x}_{\mathrm{w},k} && (3.15)\\ &= \boldsymbol{\mu}_c + \boldsymbol{e}_{\mathrm{w},c} && (3.16)\end{aligned}$$

where $\boldsymbol{e}_{\mathrm{w},c}$ represents a non zero remainder of the within part since we use a limit amount of samples to estimate the class mean. As a result, the distribution of $\hat{\boldsymbol{\mu}}_c$ is given by $\mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{t}}, \boldsymbol{\Sigma}_{\mathrm{b}}\right) + \mathcal{N}\left(0, \frac{1}{N_{\mathrm{pc}}}\boldsymbol{\Sigma}_{\mathrm{w}}\right) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{t}}, \boldsymbol{\Sigma}_{\mathrm{b}} + \frac{1}{N_{\mathrm{pc}}}\boldsymbol{\Sigma}_{\mathrm{w}}\right)$. Therefore the expected value of the between class covariance matrix is given by:

$$\mathrm{E}\left\{\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\right\} = \boldsymbol{\Sigma}_{\mathrm{b}} + \frac{1}{N_{\mathrm{pc}}}\boldsymbol{\Sigma}_{\mathrm{w}} \tag{3.17}$$

instead of $\boldsymbol{\Sigma}_{\mathrm{b}}$, which shows that the between class estimate indeed has a crosstalk from the within class estimate.

Without proof we state that $\mathrm{E}\left\{\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}\right\} = \boldsymbol{\Sigma}_{\mathrm{w}}$. Therefore if we estimate $\boldsymbol{\Sigma}_{\mathrm{t}}$ by adding $\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}$ together, we get

$$\mathrm{E}\left\{\hat{\boldsymbol{\Sigma}}_{\mathrm{w}} + \hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\right\} = \left(1 + \frac{1}{N_{pc}}\right)\boldsymbol{\Sigma}_{\mathrm{w}} + \boldsymbol{\Sigma}_{\mathrm{b}} \tag{3.18}$$

Another approach to estimate $\boldsymbol{\Sigma}_{\mathrm{t}}$ is to estimate it directly from the samples without splitting the samples into a within class part and a between class part, as is given by equation 3.19. So does the crosstalk cause an over representation of the within class variations in that approach? After some calculations we can show that this is not the case, although the between class variations are slightly under represented in this estimate, as is shown by the following analysis. First we rewrite the sample estimate of the total to a sum of the within and the between estimate:

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\mathrm{t,se}} &= \frac{1}{N-1}\sum_{k=1}^{N}\Big(\\ &\quad \left(\left(x_k - \hat{\mu}_{l(x_k)}\right) + \left(\hat{\mu}_{l(x_k)} - \hat{\mu}_{\mathrm{t}}\right)\right) \cdot \\ &\quad \left(\left(x_k - \hat{\mu}_{l(x_k)}\right) + \left(\hat{\mu}_{l(x_k)} - \hat{\mu}_{\mathrm{t}}\right)\right)^{\mathrm{T}}\Big) && (3.19)\\ &= \frac{N-C}{N-1}\hat{\boldsymbol{\Sigma}}_{\mathrm{w}} + \frac{N-N_{\mathrm{pc}}}{N-1}\hat{\boldsymbol{\Sigma}}_{\mathrm{b}} && (3.20)\end{aligned}$$

If we now take the expected value of this sum, we get the following:

$$\begin{aligned} & \mathrm{E}\left\{\hat{\boldsymbol{\Sigma}}_{\mathrm{t,se}}\right\} \\ = & \frac{N-C}{N-1}\mathrm{E}\left\{\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}\right\} + \frac{N-N_{\mathrm{pc}}}{N-1}\mathrm{E}\left\{\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}\right\} \quad (3.21) \\ = & \boldsymbol{\Sigma}_{\mathrm{w}} + \left(1 - \frac{N_{\mathrm{pc}}-1}{N-1}\right)\boldsymbol{\Sigma}_{\mathrm{b}} \quad (3.22) \end{aligned}$$

In the estimate the estimate of the within class variation $\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}$ is downscaled with exactly the amount of within class variation present in the between class variation estimate $\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}$. The sample estimate of the total therefore implicitly reduces the influence of the within estimate to compensate for the crosstalk. However, the between is under represented in the estimate.

As a side note, notice that the within class part in $\hat{\boldsymbol{\Sigma}}_{\mathrm{b}}$ is estimated independently from $\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}$. The estimate of $\boldsymbol{\Sigma}_{\mathrm{w}}$ used in $\hat{\boldsymbol{\Sigma}}_{\mathrm{t}}$ is the sum of both estimates and is therefore more accurate than $\hat{\boldsymbol{\Sigma}}_{\mathrm{w}}$.

3.2.3 Effects of limited number of gallery samples

For testing a gallery representing the class centers is composted based on a limited number of gallery samples per class, $N_{\mathrm{pc,gal}}$. Therefore the genuine distribution of the genuine attempts, given by $\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{gal},c}$ is not $\mathcal{N}\left(0, \boldsymbol{\Sigma}_{\mathrm{w}}\right)$, but it is given by:

$$\begin{aligned} & \mathcal{L}\left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{\mathrm{gal},c}\right) \\ = & \mathcal{L}\left(\left(\boldsymbol{\mu}_c + \boldsymbol{x}_{\mathrm{w}}\right) - \left(\boldsymbol{\mu}_c + \boldsymbol{e}_{\mathrm{w,gal},c}\right)\right) \quad (3.23) \\ = & \mathcal{L}\left(\boldsymbol{x}_{\mathrm{w}} + \boldsymbol{e}_{\mathrm{w,gal},c}\right) \quad (3.24) \\ = & \mathcal{N}\left(0, \left(1 + \frac{1}{N_{\mathrm{pc,gal}}}\right)\boldsymbol{\Sigma}_{\mathrm{w}}\right) \quad (3.25) \end{aligned}$$

where $\boldsymbol{e}_{\mathrm{w,gal}}$ represents the error in the gallery class mean estimate, which is the sample mean of the within class parts of the gallery samples belonging to class c.

This new distribution should be used in determining the likelihood of the genuine hypothesis given by $p\left(\bar{x} = \boldsymbol{x} \,|\bar{c} = c\,\right)$, especially for the theoretical limit.

3.3 Naive LDA correction [2]

3.3.1 Prologue

In the following article we attempt to apply bias correction to the distributions of the within class variations and the between class variations individually, without regard

[2] Analysis of eigenvalue correction applied to biometrics, International Conference on Biometrics, 2009[53]

for any relation between the two distributions. The article starts with an introduction to covariance estimation and the decomposition of this matrix, which will probably be already known to the reader. Section 3.3.3.1 starts with explaining the basics of the bias in the sample eigenvalues, which was done before as well. It is therefore recommended to read from section 3.3.3.2, where the bias correction methods used in the remainder of the article are introduced. Most of these algorithms have not been introduced yet.

3.3.2 Introduction

An important aspect of biometrics is data modeling. Modeling the statistics of data by covariance matrices is an example. Two techniques which rely on modeling by covariance matrices are Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

Because the covariance matrix of the data generating process, $\mathbf{\Sigma}$, is usually unknown, it needs to be estimated from a training set. An often used estimate is the sample covariance matrix:

$$\hat{\mathbf{\Sigma}} = \frac{1}{N-1} \mathrm{X} \cdot \mathrm{X}^T \tag{3.26}$$

where the columns of matrix X contain the training samples with the mean subtracted and N is the number of samples in the set.

In the modeling process we are often more interested in functions of the covariance matrix than in the covariance matrix itself. A commonly used function is the decomposition of the covariance matrix in eigenvectors and eigenvalues. The decomposition results we call population eigenvectors and population eigenvalues when derived from $\mathbf{\Sigma}$ and we call them by sample eigenvectors and sample eigenvalues when derived from $\hat{\mathbf{\Sigma}}$. The i^{th} population eigenvalue is denoted by λ_i and the i^{th} sample eigenvalue is denoted by l_i. Though $\hat{\mathbf{\Sigma}}$ is an unbiased estimate of $\mathbf{\Sigma}$ [21], the estimate of λ_i by l_i does have a bias.

In this article, we analyse the effect of this bias with two verification experiments. In the first experiment we use synthetic data so we can compare the verification performance of the system with and without the bias. In both the synthetic data and the real biometric data we compare performance improvement when applying several bias correction algorithms in several configurations.

An analysis of the bias is given in section 3.3.3.1. In section 3.3.3.2 we present a number of algorithms which reduce the bias. In section 3.3.4 we describe the verification system used in the experiments. We indicate where the bias will have its largest effect and where it should be compensated.

In section 3.3.5.1 we present an experiment with synthetic facial data, to determine the effect of the bias when the assumed model is correct. In section 3.3.5.2 we repeat the experiment with real facial data. In section 3.3.6 we present conclusions.

3.3.3 Eigenvalue bias analysis and correction

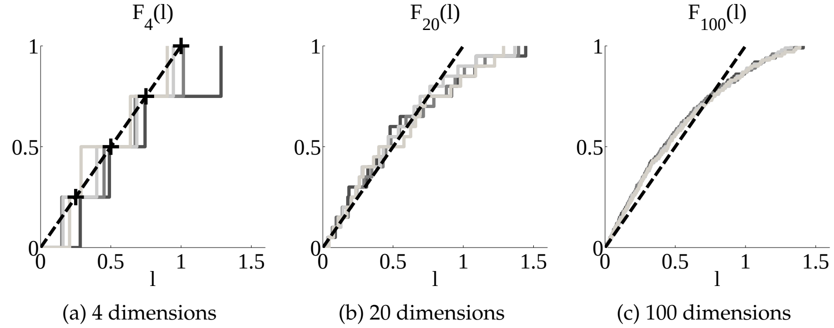
3.3.3.1 Eigenvalue bias analysis

To find the statistics of estimators often Large Sample Analysis (LSA) is performed. The sample eigenvalues show no bias in this limit case where the number of samples is large enough that it solely determines the statistics of the estimator. However, in biometrics, the number of samples is often in the same order as the number of dimensions or even lower. Therefore, in the analysis of the statistics of the sample eigenvalues the following limit may be considered: $N, p \to \infty$ while $\frac{p}{N} \to \gamma$. Here N is the number of samples used, p is the number of dimensions and γ is some positive constant. Analysis in this limit are denoted General Statistics Analysis (GSA) [28]. In GSA the sample eigenvalues do have a bias.

To demonstrate GSA, we estimated sample eigenvalues of synthetic data with population eigenvalues chosen uniformly between 0 and 1. We kept $\gamma = \frac{1}{5}$ while we varied the dimensionality between 4, 20 and 100. In Figure 3.1 we show both the population eigenvalue probability function and the sample eigenvalue probability functions for 4 repetitions, given by

$$F_p(l) = p^{-1} \sum_{i=1}^{p} \mathrm{u}\left(l - l_i\right) \tag{3.27}$$

where $\mathrm{u}\left(l\right)$ is the step function. The empirical probability functions converge with increasing dimensionality, and they converge to a different probability function as the population probability function, due to bias. This example also shows that bias reduction is only possible for a minimum dimensionality, because only then the largest part of the error in l_i as estimate of λ_i is caused by the bias.



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

Figure 3.1: Examples of eigenvalue estimation bias toward the GSA limit. All lines indicate empirical probability functions based on sets of eigenvalues (see equation 3.27). The dashed line indicates the population distribution, the four solid lines are the empirical sample distribution.

3.3.3.2 Eigenvalue bias correction algorithms

The bias is a deterministic error and can therefore be compensated. In this section we present a number of correction algorithms we used in the verification experiments to reduce the bias. The correction algorithms provide new estimates of the population eigenvalues, which are denoted by $\hat{\hat{\lambda}}_i$.

1. The *Muirhead correction* [54] is given by a maximum likelihood estimate of the population eigenvalues:

   $$\hat{\hat{\lambda}}_i = l_i - \frac{1}{n} l_i \sum_{j=1 \ldots i-K, i+K \ldots p}^{p} \frac{l_j}{l_i - l_j} \tag{3.28}$$

   In the original formula K was set to one. However, to prevent strong fluctuations, we set $K = 50$, which is a simplified version of the Stein[30] algorithm.

2. The *Karoui correction* [27] is based on the Marčenko Pastur equation [26] which gives a relation between sample eigenvalues and the population eigenvalues in the limit considered in GSA. The algorithm finds an estimate of the empirical population eigenvalue probability function (Equation 3.27, with l replaced by λ) as a weighed sum of fixed probability functions, in our case a set of delta pulses and bar functions.

3. The *Iterative feedback* algorithm was developed by the authors and is new to our knowledge. To find the population eigenvalues the algorithm starts with an initial guess for the population eigenvalues, $\hat{\hat{\lambda}}_{i,1}$. In the m^{th} iteration of the algorithm, synthetic data is generated with population eigenvalues equal to $\hat{\hat{\lambda}}_{i,m}$. The sample eigenvalues $\hat{l}_{i,m}$ of this synthetic data are determined. $\hat{\hat{\lambda}}_{i,m+1}$ is constructed via $\hat{\hat{\lambda}}_{i,m+1} = \hat{\hat{\lambda}}_{i,m} \cdot \frac{l_i}{\hat{l}_{i,m}}$. These steps are repeated until $\sum_{i=1}^{p} \left(\hat{l}_i - l_i\right)^2$ is below a preset threshold or $m > m_{\max}$.

4. The *Two Subset* correction is a classical technique in statistics to remove bias in estimates, where X is split in two subsets X_1 and X_2. From $(N/2-1)^{-1}\mathrm{X}_1\mathrm{X}_1{}^T$ eigenvectors are estimated, denoted $\hat{\Phi}_1$. The variances in the second set along these estimated eigenvectors are used as $\hat{\hat{\lambda}}_i$'s, so $\hat{\hat{\lambda}}_i = \hat{\Phi}_{1,i}^T \cdot \frac{1}{N_2-1} X_2 X_2^T \cdot \hat{\Phi}_{1,i}$. The $\hat{\hat{\lambda}}_i$'s do not contain the bias of the original estimates. However, since the estimation is performed on half of the original set, the variance of the estimate increases. This might explain why this correction is not commonly used.

3.3.4 Verification system description

3.3.4.1 System setup

In our experiments we test the influence of the bias of eigenvalues in biometric systems, using a well known baseline PCA LDA system in our experiments. In this section we give a brief description of this system. For a more detailed discussion we refer to [55].

The input of the verification system are facial images. On these images some standard preprocessing is done, which results in a data sample $\vec{x}$ for each image. To transform these input vectors to a space where classification is possible, a transformation matrix T is determined in 3 steps based on a training set of example samples. In the first two steps we use PCA to reduce the dimensionality and whiten the data.

In the third step a projection to the most discriminating subspace is determined by modeling each data sample as $\vec{x} = \vec{x}_w + \vec{x}_b$. Variations between samples from the same class are modeled by $\vec{x}_w$, which is distributed as $N(0, \mathbf{\Sigma}_w)$, a multi variate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}_w$. We model the variations between classes by $\vec{x}_b$, which is distributed as $N(\mu_t, \mathbf{\Sigma}_b)$. Since the data is whitened, the most discriminating subspace is the subspace of the largest eigenvalues of $\mathbf{\Sigma}_b$. Therefore the transformation matrix T is given by:

$$\mathrm{T} = \hat{\Phi}_{b,C_2}^T \cdot \hat{\Lambda}_{t,C_1}^{\frac{1}{2}} \cdot \hat{\Phi}_{t,C_1}^T \tag{3.29}$$

where $\hat{\Phi}_{t,C_1}$ are the first C_1 eigenvectors of $\hat{\mathbf{\Sigma}}_t$, the covariance matrix of the training set, and $\hat{\Lambda}_{t,C_1}$ is a diagonal matrix with as diagonal the first C_1 eigenvalues of $\hat{\mathbf{\Sigma}}_t$. $\hat{\Phi}_{b,C_2}$ are the first C_2 eigenvectors of $\hat{\mathbf{\Sigma}}_b$.

After projecting samples in the classification space, we compare sample $\vec{x}$ with class c by calculating a matching score. We accept an identity claim if the score is above a certain threshold. The score is based on the log likelihood:

$$\mathrm{L}(\vec{x}, c) = -\left(\mathrm{T}\cdot\vec{x} - \mu_c\right)^T \cdot \hat{\Sigma}_w^{-1} \cdot \left(\mathrm{T}\cdot\vec{x} - \mu_c\right) + \left(\mathrm{T}\cdot\vec{x} - \mu_t\right)^T \cdot \left(\mathrm{T}\cdot\vec{x} - \mu_t\right) \tag{3.30}$$

3.3.4.2 Modifications for eigenvalue correction

In this verification system, there are two points where eigenvalue correction may improve results: in the whitening step, where the data is scaled based on eigenvalue estimates and in the matching score calculation, where the eigenvalues of the within covariance matrix in the classification space are needed. We perform eigenvalue correction after the dimensionality reduction, but before the whitening step.

At first sight, it seems that the eigenvalues of $\hat{\mathbf{\Sigma}}_t$ need to be corrected. However, under the assumed model, the total covariance matrix $\mathbf{\Sigma}_t$ can be written as $\mathbf{\Sigma}_b + \mathbf{\Sigma}_w$. These matrices are estimated by $(C-1)^{-1}\sum_{c=1}^{C}\mu_c\mu_c^T$ and $(N-C)^{-1}\sum_{i=1}^{N}(\vec{x}_i - \mu_{\ell(\vec{x}_i)})(\vec{x}_i - \mu_{\ell(\vec{x}_i)})^T$ respectively, where C is the number of classes in the training set, μ_c is the mean of the training samples of class c, and $\ell(\vec{x}_i)$ returns the class index of sample $\vec{x}_i$. Because both matrices are estimated with a different number of samples, their eigenvalues have a different bias. We therefore perform the correction in the following manner:

1. Estimate $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_b$.
2. Decompose both covariance matrices in eigenvectors and eigenvalues.
3. Construct new estimates of the covariance matrices using the original eigenvector estimates and the corrected eigenvalues.
4. Sum the two estimates to get a new estimate of $\mathbf{\Sigma}_t$.

The corrected estimate of the covariance matrix is given by $\tilde{\mathbf{\Sigma}}_r = \hat{\Phi}_r \cdot f_{N_r}(\hat{\Lambda}_r) \cdot \hat{\Phi}_r^T$, where r is either w or b and $f_{N_r}(\hat{\Lambda}_r)$ is an eigenvalue correction algorithm.

3.3.5 Experiments

In this section we describe two verification experiments with the system presented in the previous section. In the first experiment we used synthetic facial data while in the second experiment we used real facial data.

3.3.5.1 Synthetic data experiment

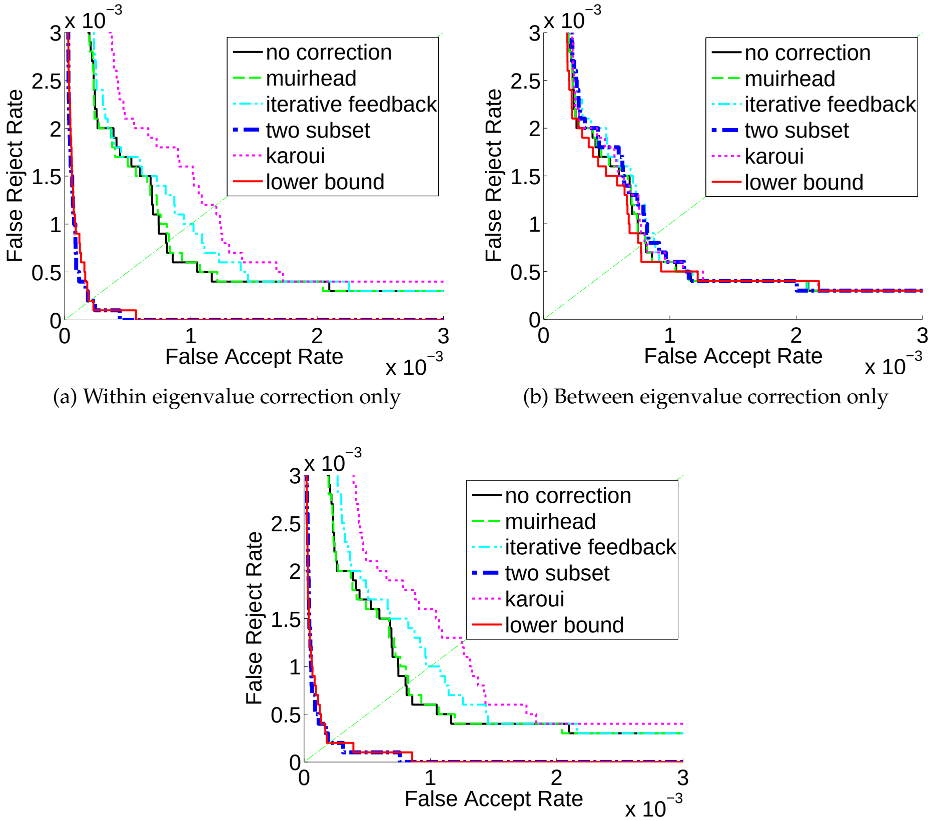
To generate synthetic data close to real facial data, we determined the data structure of a large set of face images in the FRGC database. The data contained 8941 facial images. All facial images were taken under controlled conditions with limited variations in pose and illumination. Also the faces in the facial images had a neutral expression and nobody wore glasses.

We model the facial data with the model in section 3.3.4. For generating synthetic data adhering to this model with parameters close to real facial data, we estimated the within class covariance matrix $\mathbf{\Sigma}_w$ and the between class covariance matrix $\mathbf{\Sigma}_b$ from the FRGC data. Since the eigenvalues of these estimates also contain a bias, we corrected their eigenvalues with the Two Subset correction, knowing from previous experiments that this correction led to better estimates of eigenvalues [48]. We kept μ_t zero.

We generated a small training set of 70 identities, with 4 samples per identity, so the bias should be comparable to small real face data sets. This training set was used to train a verification system. In the dimensionality reduction stage of the training the dimensionality was reduced to 150. In the LDA step, the 60 most discriminating features were retained.

We tested the following corrections: no correction, Muirhead correction, Karoui correction, Iterative Feedback correction, Two Subset correction and a lower bound correction. With the lower bound correction, we use the true covariance matrices of the synthetic data to calculate the actual variances along the estimated eigenvectors and use these values as $\hat{\hat{\lambda}}_i$'s. We assumed this correction would give an indication of the best possible error reduction.

We generated a test set with 1000 identities. For each identity 10 enrollment samples and 10 probe samples were generated. During the experiment 3 configurations were tested: correction of only the within class eigenvalues, correction of only the between class eigenvalues and correction of both the within and the between class eigenvalues. The DET curves of the three configurations are shown in Figure 3.2. In Figure 3.4a we show the relative EER improvement averaged over 5 repetitions.



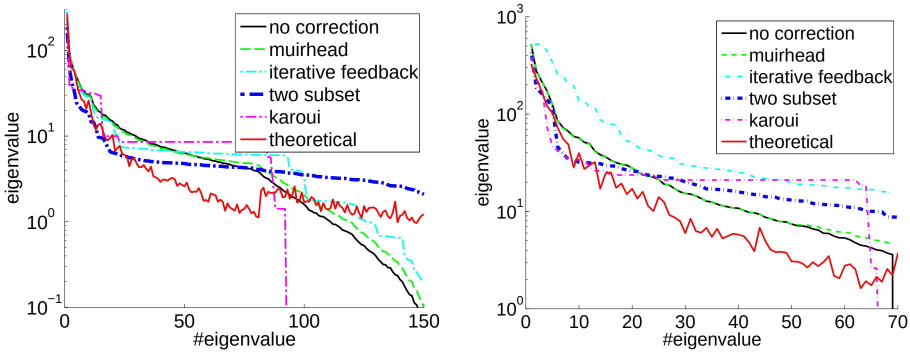
(a) Within eigenvalue correction only

(b) Between eigenvalue correction only

(c) Both within and between class eigenvalue correction

Figure 3.2: DET curves for the synthetic data experiment.

The within class eigenvalues correction configuration shows a large difference between the no correction DET curve and the lower bound correction. Therefore the bias in the within class eigenvalues seems to have a large effect on the error rates. The Two Subset correction achieves on average slightly better results as the lower bound correction, but this is probably due to measurement noise. The performance of the Karoui correction fluctuates when the experiment is repeated. In some repetitions the Karoui correction reduces the error rates by half, but on average it increases the



(a) Within class eigenvalues

(b) Between class eigenvalues

Figure 3.3: Scree plots of the corrected eigenvalues of synthetic data.

error rates as shown in Figure 3.4a.

The between class eigenvalues correction configuration shows hardly any difference between the different correction algorithms. It seems that the bias in the between class eigenvalues have little influence on the verification scores. The curve of both eigenvalue sets corrected shows no significant difference with the within only correction.

In Figure 3.3a and Figure 3.3b we show the corrected within class eigenvalues and between class eigenvalues respectively. The lower bound correction shows considerable fluctuations in the curve. This indicates that the ordering of the sample eigenvectors is wrong.

The lower bound curve is much flatter for the small eigenvalues in the within class correction than the no correction curve. The Two Subset correction also makes the curve much flatter for the smaller eigenvalues, although the eigenvalues are considerably larger than the lower bound correction. Considering the error rates are almost the same, the similarity in flatness seems more important than the actual value of the eigenvalues.

The Karoui correction shows a similar flatness until the 78^{th} eigenvalue. After the 92^{th} eigenvalue, all remaining eigenvalues are set to 0. This seems to have only a small effect on the error rates. This is remarkable since 0 within class variance would indicate very good features, while we know from the lower bound correction that the within class variance is non zero. However, if the between class variance is also zero, the direction will be neglected.

3.3.5.2 FRGC facial data experiment

Eigenvalue correction with synthetic facial data caused a significant reduction of the error rates. In the next experiment we replaced the synthetic facial data with the face data set from the FRGC database. This data set is the set used in the previous experiment to determine the facial data structure.

The data set is split in a training set and a test set. The training set contained 70 randomly chosen identities, with a maximum of 5 samples per identity. The test set contained the remaining 445 identities. At most 5 samples per identity are used for enrolling, at least 1 sample is used as probe per identity.

In the training stage instead of reducing the dimensionality to 150, as described in section 3.3.4, only the null space is removed. After correction of the eigenvalues, the dimensionality is reduced to 150. The correction algorithms described in section 3.3.3.2 are compared.

The experiment is repeated 5 times for the same 3 configurations as in the synthetic data experiment. For each correction algorithm in each configuration we determined the Equal Error Rate (EER). This EER is compared with the no correction EER. The average over 5 repetitions of the relative improvement of EER is shown in figure 3.4b.

The results show that correcting only the between class eigenvalues increases the EER for all correction algorithms. The within correction decreases the EER for most algorithms. Correcting both eigenvalue sets decreases the EER for the iterative feedback algorithm and the Two Subset algorithm. But this decrease in EER is less than the decrease in EER if only the within class eigenvalues are corrected.

Comparing the different correction methods shows that in the within correction and both eigenvalue sets correction the Two Subset correction performs considerably better than the other corrections. The Karoui correction always increases the EER.

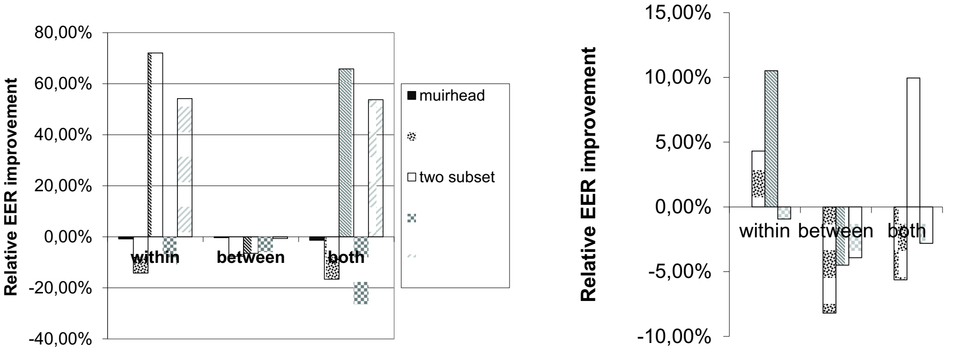
In Figure 3.5 we show the results of the first repetition. The Karoui corrections sets a large set of small eigenvalues to zero. This had remarkably little effect on the error rates. The Two Subset correction on the other hand assigns non zero values to eigenvalues which were originally zero.

Most correction algorithms show a trend: the largest eigenvalues are reduced while the smaller eigenvalues are increased. This effect is the strongest with the Two Subset correction. Since this correction method achieved the lowest error rates, it seems that in face recognition indeed the largest eigenvalues are over estimated while the smallest are under estimated, at least in the within class estimation.

Comparing the results of the real facial data test with the results from the synthetic data shows that the EER's in real data are an order higher than the EER's in synthetic data. This suggests that the model we used is not sufficiently accurate for describing real facial data. However, in both experiments the Two Subset method showed the highest reduction in EER.

3.3.6 Conclusion

We showed that the GSA provides more accurate analysis of the sample eigenvalue estimator than LSA in biometrics: GSA on the estimator predicts that the estimates in biometrics will have a bias, which is observed in synthetic data, especially for the smaller eigenvalues.



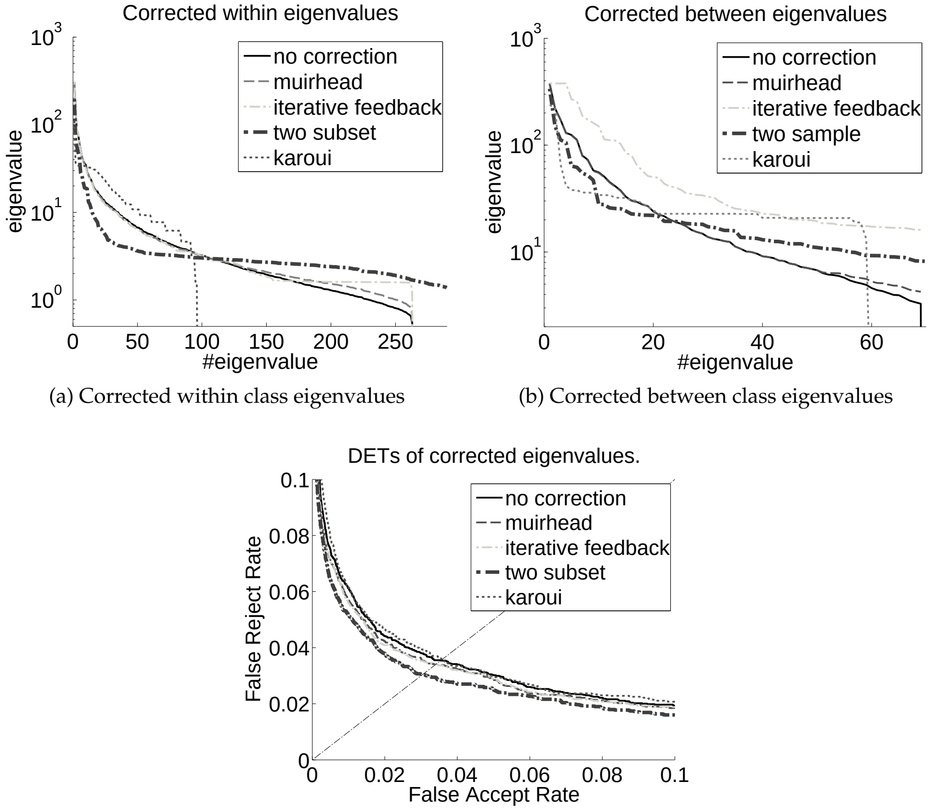
(b) real facial data correction averaged over 5 repetitions.

Figure 3.4: Relative Equal Error Rate improvement for each correction method. There are three configurations: only within class eigenvalues correction, only between class eigenvalues correction and both eigenvalue sets correction.

Correcting only the within class eigenvalues has demonstrated the most effect. This is related to the previous conclusion: the best features are determined by the ratio of between class over within class variance. Therefore the best features probably lie in space spanned by the largest between class eigenvalues and the smallest within class eigenvalues. Since the smaller eigenvalues have more bias, within class correction has the most effect.

The Two Subset correction gave the best improvement of error rates in both the synthetic data experiment and the real facial data experiment. Although the performance of the correction was the same as the synthetic correction, the scree plots did differ. The corrections of the other algorithms did also significantly alter the eigenvalues, but this had little effect on the error rates for most of these corrections. Apparently the actual values of the eigenvalues do not have to be estimated very accurately.

The relative error reduction in the facial data is much lower as in the synthetic data by the Two Subset correction. Also the no correction error rates differ more than an order between the real facial data and the synthetic data. This suggest that the bias in the eigenvalue bias is only a moderate error factor in the real facial data.



(a) Corrected within class eigenvalues

(b) Corrected between class eigenvalues

(c) DET curves of within correction only.

Figure 3.5: Results of the first repetition of real facial data experiment.

3.4 Likelihood ratio based verification in high dimensional spaces [3]

3.4.1 Prologue

In the previous section we tried to apply bias correction rather bluntly in a verification setting. This led to several remarkable observations which require some explanation. In the next paper we perform a thorough analysis of the SOS and the likelihood ratio in a verification setting. With these analysis we show that PCA dimensionality reduction, the classical approach to the problems in verification related to high dimensionality, is far from optimal and even completely fails if the dimensionality is very high. This makes the observation reported in the previous section that PCA dimensionality reduction is required after bias correction even more curious.

The same analysis also shows what goes wrong with the Naive BCIV: if the correction of the between eigenvalues is performed without regard of the within distribution, the null spaces of the data can get an arbitrarily high between class variance over within class variance ratio, suggesting a high discriminative capacity of that subspace which is actually not supported by the data.

Although the first part of the next paper contains several items which have already been covered, it quickly focusses on the bias correction in a verification setting. A quick read through is therefore advised to catch up with section 3.4.4, in which the core problem of PCA dimensionality reduction is demonstrated experimentally. After that section the problems occurring if classical PCA dimensionality reduction is used, are analysed and the main contributions of this paper are presented: a proof is given that classical PCA dimensionality reduction will fail if the dimensionality becomes very large and new methods are given for improved SOS estimation in a verification setting.

3.4.2 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. Intuitively 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.

We will show that this is currently not true for systems based on SOS (such as PCA or LDA based systems) by testing how a verification system based on likelihood estimates using SOS is affected by an increasing dimensionality p of its training data. The likelihood method requires inversion of estimated covariance matrices, but if p is higher than the N available for training, at least $p - N$ of the eigenvalues of these

[3] Likelihood ratio based verification in high dimensional spaces, TPAMI, in communication[56]

matrices will be zero, so inversion is not possible. This is known as the singularity problem and it is one of the problems indicated by "the curse of dimensionality" [57].

A common approach to solve this singularity problem is to use PCA to estimate a subspace in the training data containing most of the variance of the data, while its dimensionality is at least smaller than N. However, PCA dimensionality reduction is not based on a good estimator: after a certain value of p, if p is increased even more, the error rates of the verification system go up instead of at least staying at the same level, and for very high dimensionality, PCA fails completely.

Another approach to the high dimensional verification problem is to use an 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. In [38] it was shown that the euclidean distance outperforms SOS estimates in several statistical tests for even moderately high values of p. We show by experiment that PCA dimensionality reduction is also outperformed by a euclidean distance measure in a verification setting for large p.

There are some theories on how the data dimensionality influences the SOS estimation. One of the theories describes how the eigenvalues estimated from a training set become significantly biased, another theory describes how the eigenvector estimates relate to the eigenvectors of the data generating process. Based on these theories we can derive several hypotheses on the SOS estimation in high dimensional problems and with these hypotheses we can show why PCA dimensionality reduction is outperformed by euclidean distance based methods and why PCA dimensionality reduction breaks down if p is large.

One of the major challenges in these analysis is that the theories are based on the estimation of SOS from a single distribution. But in verification two distributions are of concern: the distribution of the variations of samples originating from the same class and the distribution of samples originating from different classes. As we will show the verification result depends mostly on the estimation of the ratio of the variances of these two distributions. Therefore the theories have to be extended to determine the influence of increasing p on this ratio. Note that in chapter 17 of [58] classification in such a setting is already explored, however only for the case that for one of the distributions all eigenvalues are equal.

The analyses showing the weaknesses of the PCA dimensionality reduction solution can also be used to improve the SOS estimation. These improvements entail: eigenvalue bias correction, variance correction and a new method we denote as eigenwise BCIV. We repeat the experiment in which we compared PCA dimensionality reduction with the euclidean distance based method but now included our improvements. The results show that the improvements indeed outperform the classical PCA dimensionality reduction and converge from the optimum of the sample estimates for low p to the euclidean distance measure for very high p.

In the next section we start with the introduction of verification based on SOS. Section 3.4.4 demonstrates the inadequacy of PCA dimensionality reduction and we explain the results of this experiment. These explanations are based on a few assumptions which are either proved or demonstrated experimentally in section 3.4.5. Section 3.4.6 gives improvements on the SOS estimation of a single distribution based on the analysis in the previous sections. Section 3.4.7 shows how these analyses can be used to improve the verification results. Section 3.4.8 repeats the experiment of section 3.4.4, but now including the improvements. In section 3.4.9 conclusions are presented.

3.4.3 Verification using second order approximations

One of our goals is to determine the performance of PCA dimensionality reduction in a verification experiment. 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:

$$R\left(\boldsymbol{x},c\right)=\frac{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}|\bar{c}=c\right)}{p\left(\bar{\boldsymbol{x}}=\boldsymbol{x}|\bar{c}\neq c\right)} \tag{3.31}$$

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

By varying the threshold a trade off can be made between the rate of the genuine claims (x belongs to class c) being rejected (FRR) and the rate of imposter claims (x belongs to another class) being accepted (FAR). According to the Neyman-Pearson lemma [59] this test is optimal for deciding whether $\boldsymbol{x}$ is originating from class c or not for a given FAR [60].

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

As a 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 [61], it is the best approximation according to the maximum entropy principle [35].

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 $\mathcal{N}\left(0,\boldsymbol{\Sigma}_{\mathrm{w}}\right)$. The between part $\boldsymbol{\mu}_c$ is used to model variations between samples of different classes and its distribution is approximated by $\mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{t}},\boldsymbol{\Sigma}_{\mathrm{b}}\right)$. The distribution of $\boldsymbol{x}$ then becomes $\mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{t}},\boldsymbol{\Sigma}_{\mathrm{t}}\right)$ with $\boldsymbol{\Sigma}_{\mathrm{t}}=\boldsymbol{\Sigma}_{\mathrm{w}}+\boldsymbol{\Sigma}_{\mathrm{b}}$. If these distributions are used in equation 3.31 and we take the logarithm, then the log likelihood ratio becomes:

$$\begin{aligned}L\left(\boldsymbol{x},c\right)=&-\left(\boldsymbol{x}-\boldsymbol{\mu}_c\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{w}}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_c\right)\\&+\left(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{t}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{t}}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{t}}\right)\end{aligned} \tag{3.32}$$

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 $\hat{\boldsymbol{\mu}}_\text{t} = \frac{1}{C}\sum_{c=1}^{C}\hat{\boldsymbol{\mu}}_c$. $\boldsymbol{\Sigma}_\text{w}$ and $\boldsymbol{\Sigma}_\text{b}$ are estimated by the sample covariance matrices in (3.33) and (3.34) respectively, where $l\left(\boldsymbol{x}_k\right)$ returns the class label of $\boldsymbol{x}_k$.

$$\hat{\boldsymbol{\Sigma}}_\text{w} = \frac{1}{N-C}\sum_{k=1}^{N}\left(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_{l(\boldsymbol{x}_k)}\right)\left(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_{l(\boldsymbol{x}_k)}\right)^\text{T} \tag{3.33}$$

$$\hat{\boldsymbol{\Sigma}}_\text{b} = \frac{1}{C-1}\sum_{c=1}^{C}\left(\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}}_\text{t}\right)\left(\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}}_\text{t}\right)^\text{T} \tag{3.34}$$

In many applications functions are applied to the covariance matrices, which commonly require a decomposition of the covariance matrix. In particular we are interested in the evaluation of the probability of samples, which requires the inversion of a covariance matrix. A covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as $\boldsymbol{E}\cdot\boldsymbol{D}\cdot\boldsymbol{E}^\text{T}$, where $\boldsymbol{E}$ 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.

For our analysis in the following sections 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 $L\left(\boldsymbol{w}^\text{T}\boldsymbol{x}, c\right)$. Using the fact that $\mathcal{E}\left\{L\left(\boldsymbol{x}, c\right)\right\} = 0$ it follows after some calculations that

$$\mathcal{E}\left\{L^2\left(\boldsymbol{w}^\text{T}\boldsymbol{x}, c\right)\right\} = 4\frac{\boldsymbol{w}^\text{T}\cdot\boldsymbol{\Sigma}_\text{b}\cdot\boldsymbol{w}}{\boldsymbol{w}^\text{T}\cdot\boldsymbol{\Sigma}_\text{t}\cdot\boldsymbol{w}} \tag{3.35}$$

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 3.35 is the generalized Rayleigh quotient used in LDA to find the most discriminating subspace in the data [21], so LDA can be considered as applying PCA dimensionality reduction based on the variance of $L\left(\boldsymbol{x}, c\right)$ instead of the variance of the samples themselves.

3.4.4 Second order statistics estimation in high dimensional spaces

3.4.4.1 A verification experiment with PCA dimensionality reduction

In this section we will experimentally demonstrate the weaknesses of the PCA dimensionality reduction in high dimensional verification problems, by performing a verification experiment with synthetic data. We used a verification system equal to the system described in the previous sections and we varied the dimensionality p of the samples in several iterations between a value much lower than the fixed number of samples N and a value considerably larger than N. If $p > N$, at least $p - N$ eigenvalues become zero and the singularity problem occurs as noted in section 3.4.2.

To solve the singularity problem, we used the classical PCA dimensionality reduction solution prior to verification and compared the verification performance with two limit cases: a theoretical optimum limit and a regularisation limit. In the theoretical limit we assumed perfect estimation and replaced 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 a method known as regularisation and it turns the probability measures into a euclidean distance such that the log likelihood ratio becomes:

$$\hat{L}_{\mathrm{reglim}}\left(\boldsymbol{x}, c\right) \quad = \quad -\frac{1}{\bar{l}_{\mathrm{w}}}\hat{\boldsymbol{x}}_{\mathrm{w}}^{\mathrm{T}}\hat{\boldsymbol{x}}_{\mathrm{w}} + \frac{1}{\bar{l}_{\mathrm{t}}}\hat{\boldsymbol{x}}_{\mathrm{t,zm}}^{\mathrm{T}}\hat{\boldsymbol{x}}_{\mathrm{t,zm}} \tag{3.36}$$

where $\hat{\boldsymbol{x}}_{\mathrm{w}} = \boldsymbol{x} - \hat{\mu}_c$, $\hat{\boldsymbol{x}}_{\mathrm{t,zm}} = \boldsymbol{x} - \hat{\mu}_{\mathrm{t}}$. $\bar{l}_{\mathrm{w}}$ and $\bar{l}_{\mathrm{t}}$ are the mean of the within sample eigenvalues and the mean of the total sample eigenvalues respectively.

3.4.4.2 Increasing the dimensionality p

To increase p, we have to add features. We chose to do this as follows: first we chose a curve, which we then sampled uniformly to acquire the population eigenvalues. Based on these population eigenvalues we generated the synthetic data. To get higher dimensional data, we repeat the steps, but we maintain the chosen eigenvalue curve; we only sample it with smaller steps.

We have two reasons for taking this approach. First of all, this approach closely matches the usual way to study the eigenvalue bias in SOS estimation: in these studies the eigenvalue estimation is analysed in the limit of $p, N \rightarrow \infty$, which is denoted by GSA, and it is assumed that the population eigenvalues can be described with an empirical distribution which converges for very large p to a fixed distribution. By choosing a fixed curve, we get such behaviour of the population eigenvalues, so in a sense we are performing GSA, but instead of $N \rightarrow \infty$ we keep N at a fixed value.

Secondly, with this choice, the added dimensions have the same structure as the original data, so the added features have the same energy as well as the same discriminative capacity as the original features. The choice of equal discriminative capacity is motivated by the fact that we do not know which features are more informative. If we would know on forehand, it would be illogical to start with the features having a lower discriminative capacity and add the other features later on. If the added features have lower discriminative capacity, the results will be worse compared to what we report here.

The choice to make the energy of the added features equal to the original features is motivated by noting that if (a part of) the original space contains eigenvalues much larger than the added subspace, these eigenvalues are unaffected by the added part (see [33]) and so it is like we added a null space to the data and the added features have no effect.

We expect that with image data, the situation will be somewhere between these two extremes: the added dimensions will have a lower energy, but the largest eigenvalues will be affected by the bias. So the results we report later on will apply in these situations as well, only the breakdown effects will occur for larger values of p.

We performed the experiment for two different choices of the fixed eigenvalue description curve. In both configurations we used 100 classes with 5 samples per class for training (which implies a fixed number of samples for both the within class and the between class estimation) and tested with another 100 classes with 20 samples per class. In the first configuration we chose a 2 cluster like distribution: 10% of the between eigenvalues had a value of 0.5, the remaining 90% had a value of 0.05, but we smoothed the borders between the two clusters so the eigenvalue scree plot follows a tanh curve. The within eigenvalues were chosen such that the total covariance matrix equals the identity matrix.

In the second configuration we chose the within and the between covariance matrices such that the total covariance matrix has eigenvalues exponentially decaying from 1.01 to 0.01. The between eigenvalues are one tenth of the total eigenvalues, making the within eigenvalues 0.9 times the total eigenvalues.

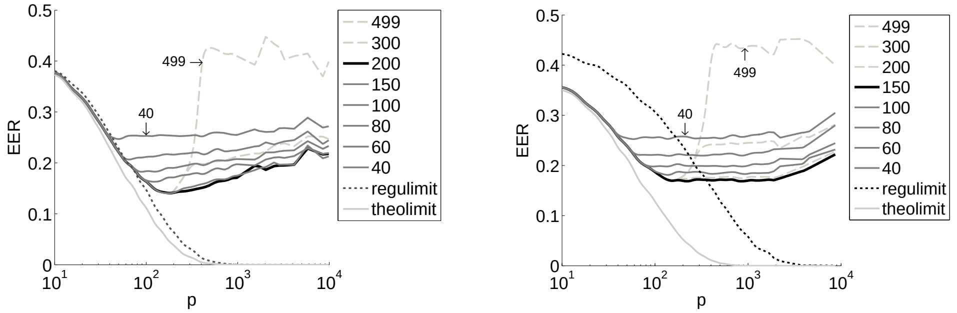
The trade off between FAR and FRR as described in section 3.4.3 can be made based on the Receiver Operating Characteristic (ROC) curve. However, since we want to study the verification performance for many different configurations, we chose to determine only the EER rate, which is the point on the ROC 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.

3.4.4.3 Results

The results of the experiment are shown in Figure 3.6. The figure shows the results of tests in which we fixed the number of components retained after dimensionality reduction. Figure 3.6a shows the EER versus p curves for the 2 cluster configuration, Figure 3.6b shows the curves for the exponential configurations. 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.

Several observations can be made:

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



(a) two cluster, fixed dimensionality reduction

(b) exponential, fixed dimensionality reduction

Figure 3.6: 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.

the theoretical limit and the regularisation limit is small and PCA is already outperformed for $p = 100$.

To explain this, first note that in the regularisation limit, the probability calculations are an euclidean distance calculation in the full dimensional space (equation 3.36). The PCA dimensionality reduction on the other hand will project the data onto a at max N dimensional subspace. Furthermore, as we will prove in section 3.4.5.1, the eigenvalues estimates in this subspace will all be equal valued for large p, so PCA dimensionality reduction turns the probability calculations into an euclidean distance in an at max N dimensional subspace. Since the sample eigenvectors spanning up this subspace are completely random for large p (see section 3.4.5.2), the PCA dimensionality reduction removes information without improving the structure estimate.

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 goes up again. This is highly remarkable, because this implies that projecting high dimensional data on a random bases with a lower dimensionality before performing PCA dimensionality reduction would improve the PCA dimensionality reduction method. Moreover, there seems to be a minimum in EER for the PCA dimensionality reduction method.

We described in the previous observation explanation that the PCA dimensionality reduction turns the probability calculations into euclidean distance calculations in an at max N dimensional subspace with a random basis for very large p. However, if p is in the same order as N, the estimated sample eigenvectors partially depend on the population eigenvectors and the sample eigenvalues also contain some of the structure of the population eigenvalues (see section 3.4.5.2). Therefore the probability calculations in the subspace estimated for smaller p values will be more accurate and hence the error rates will eventually go up with increasing p.

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, we need to focus on the fact that the likelihood ratio depends on the between class variances over the within class variances (equation 3.35). However, 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, as will be shown in section 3.4.5.3. As a result, if PCA dimensionality 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

$$\hat{\mathbf{\Sigma}}_{\mathrm{nzw,w}}^{-1} = \hat{\mathbf{\Sigma}}_{\mathrm{nzw,t}}^{-1} = \frac{p}{(N-C)\,\bar{l}}\mathbf{I} \tag{3.37}$$

and

$$\begin{aligned}\mu_{\mathrm{t}}^{\mathrm{T}}\mu_{\mathrm{t}} + \mu_{\mathrm{w}}^{\mathrm{T}}\mu_{\mathrm{w}} \quad = \quad & \frac{1}{2}\left(\mu_{\mathrm{t}} - \mu_{\mathrm{w}}\right)^{\mathrm{T}}\left(\mu_{\mathrm{t}} + \mu_{\mathrm{w}}\right) + \\ & \frac{1}{2}\left(\mu_{\mathrm{t}} + \mu_{\mathrm{w}}\right)^{\mathrm{T}}\left(\mu_{\mathrm{t}} - \mu_{\mathrm{w}}\right)\end{aligned} \tag{3.38}$$

the log likelihood ratio reduces to

$$\begin{aligned}\hat{L}\left(c, \boldsymbol{x}\right) \quad = \quad & \frac{2p}{(N-C)\,\bar{l}}\left(\hat{\boldsymbol{\mu}}_{\mathrm{nzw},c} - \hat{\boldsymbol{\mu}}_{\mathrm{nzw},t}\right)^{\mathrm{T}} \cdot \\ & \left(\boldsymbol{x}_{\mathrm{nzw}} - \frac{\hat{\boldsymbol{\mu}}_{\mathrm{nzw},c} + \hat{\boldsymbol{\mu}}_{\mathrm{nzw},t}}{2}\right)\end{aligned} \tag{3.39}$$

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

Note that the break down effects occur already for the 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]. If the dimensionality 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 section 3.4.4.3.

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. This effect is however also observable in figure 3.6 if the curves are considered for one fixed p larger than approximately 150. For example consider Figure 3.6b for $p \approx 250$. There is a clear minimum in EER for a dimensionality reduction to $p_{\text{red}} = 150$. It seems therefore 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 suggested that the increase of p had several effects on SOS estimation results. In the next sections we prove some of these effects, while we demonstrate the other effects experimentally.

3.4.5 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. This is mainly due to the fact that the commonly used sample estimators are more and more based on random fluctuations in the samples rather than the actual structure of the data if p increases. One effect is that the sample eigenvalues become biased estimates of the population eigenvalues, as will be described in the next section.

3.4.5.1 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:

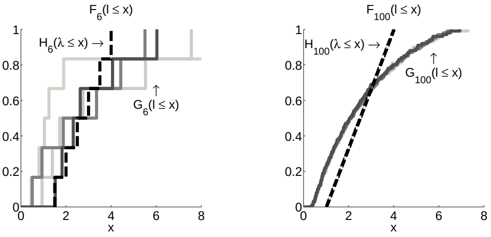
$$\mathcal{E}\left\{\boldsymbol{l}-\boldsymbol{\lambda}\right\} \tag{3.40}$$

The bias of an estimator is typically determined using LSA: that is to evaluate (3.40) under the assumption that $N \rightarrow \infty$. Under this assumption, no bias is observable in the sample eigenvalues. 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 the p is not negligible anymore compared to N. In GSA instead of assuming that $N \rightarrow \infty$ alone, 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. For the sample eigenvalues $\boldsymbol{l}$, the empirical distribution $G_p\left(l\right)$ is given by $\frac{1}{p}\sum_{k=1}^{p} \mathrm{u}\left(l-l_k\right)$.

In figure 3.7 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, once for $p = 6$ and once for $p = 100$ 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 seem to have converged, although not to the population eigenvalue distribution. This is due to the bias in the eigenvalues.



(a) 6 dimensions

(b) 100 dimensions

Figure 3.7: GSA demonstration

In [29] a description known as the 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 [26] it was proven that this relation holds for a much larger set. From this relationship it follows that the bias depends on the ratio $\frac{p}{N}$. The higher this ratio, the more severe the bias. With the trend of ever increasing dimensionality without a matching increase in samples, the bias can therefore be expected to increase.

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 the bias in the sample eigenvalues becomes insignificant and so the sample eigenvalues are almost equal to the population eigenvalues. If $p >> N$ then using the MP equation it can be proven that the sample eigenvalues only depend on the average of the population eigenvalues, $\bar{\lambda}$, and more specific that 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 3.4.4, since this proves that using SOS estimates indeed turns the probability calculations into a euclidean distance in a N dimensional subspace.

Loss of structure in high dimensional problems We now prove that if $p >> N$ then the estimated SOS only depend on $\bar{\lambda}$, the mean of the population eigenvalues. This proof is based on the Marčenko Pastur (MP) equation:

$$-\frac{1}{v(z)} = z - \gamma \int \frac{\lambda \mathrm{d}H(\lambda)}{1+\lambda v(z)} \tag{3.41}$$

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 $G(l)$: $m_G(z) = \int \frac{\mathrm{d}G(l)}{l-z}$. It was proved in [26] that this describes the relation between the sample eigenvalues and the population eigenvalues for a large family of data distributions in the limit of both $p, N \to \infty$. In the remainder of the proof we also assume that $\gamma \to \infty$. Note that these conditions are similar as in [62]. We however also assume that the population eigenvalues have a supremum, so $\frac{\mathrm{d}H}{\mathrm{d}\lambda}(\lambda) = 0 \,\forall\, \lambda \notin [0 \ldots \lambda_{\sup}]$.

First we study the order behaviour of $\|v_\infty(z)\|$ with respect to γ, proving that $\mathcal{O}(\|v_\infty(z)\|) = \gamma^{-1}$. We first show that $\mathcal{O}(\|v_\infty(z)\|) = \gamma^a$, with $a > 0$ leads to a contradiction. Firstly note that this implies that $\mathcal{O}\left(\left\|\frac{1}{v_\infty(z)}\right\|\right) = \gamma^{-a}$. Applying this to the right side of equation 3.41:

$$\begin{aligned}\mathcal{O}\left(\left\|\frac{1}{v_\infty(z)}\right\|\right) = \mathcal{O}\left(\left\|z - \gamma \int \frac{\lambda \mathrm{d}H(\lambda)}{1+\lambda v_\infty(z)}\right\|\right) &\\ = \quad \mathcal{O}\left(\left\|z - \gamma \int \frac{\lambda \mathrm{d}H(\lambda)}{\lambda v_\infty(z)}\right\|\right) &\\ = \quad \mathcal{O}\left(\left\|z - \gamma \frac{1}{v_{\infty(z)}}\right\|\right) = \gamma^{1-a} &\end{aligned} \tag{3.42}$$

this is a contradiction: $\mathcal{O}\left(\left\|\frac{1}{v_\infty(z)}\right\|\right)$ is γ^{-a} and γ^{1-a}.

If we assume that $\mathcal{O}(\|v_\infty(z)\|) = \gamma^0$, then

$$\begin{aligned}\mathcal{O}\left(\left\|\frac{1}{v_\infty(z)}\right\|\right) = \mathcal{O}\left(\left\|z - \gamma \int \frac{\lambda \mathrm{d}H(\lambda)}{1+\lambda v_\infty(z)}\right\|\right) &\\ = \quad \left|\mathcal{O}\left(\gamma^0\right) - \mathcal{O}(\gamma) \cdot \mathcal{O}\left(\gamma^0\right)\right| = \gamma &\end{aligned} \tag{3.43}$$

which is again a contradiction: $\mathcal{O}\left(\left\|\frac{1}{v_\infty(z)}\right\|\right)$ is both γ^0 and γ^1.

If we assume $\mathcal{O}(\|v_\infty(z)\|) = \gamma^a$ with $a < 0$, then

$$\begin{aligned}\mathcal{O}\left(\left\|\frac{1}{v_\infty(z)}\right\|\right) = \mathcal{O}\left(\left\|z - \gamma \int \frac{\lambda \mathrm{d}H(\lambda)}{1+\lambda v_\infty(z)}\right\|\right) &\\ = \quad \mathcal{O}\left(\left\|z - \gamma \int \lambda \mathrm{d}H(\lambda)\right\|\right) &\\ = \quad |\mathcal{O}(1) - \mathcal{O}(\gamma) \cdot \mathcal{O}(1)| = \gamma &\end{aligned} \tag{3.44}$$

So if we set $a = -1$ both arguments result in $\mathcal{O}\left(\left\|v_\infty\left(z\right)\right\|\right) = \gamma^{-1}$. Using this result we can determine the sample eigenvalue distribution if $\gamma \to \infty$:

$$\begin{aligned}
\lim_{p\to\infty} v\left(z\right) &= \lim_{\gamma\to\infty}\left(\gamma\int\frac{\lambda \mathrm{d}H\left(\lambda\right)}{1+\lambda v\left(z\right)}-z\right)^{-1} && (3.45)\\
&= \lim_{\gamma\to\infty}\left(\gamma\int\frac{\lambda \mathrm{d}H\left(\lambda\right)}{1}-z\right)^{-1} && (3.46)\\
&= \frac{1}{\gamma\bar{\lambda}-z} && (3.47)
\end{aligned}$$

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 $H\left(\lambda\right)$, if $H\left(\lambda\right)$ has a bounded support. These findings concur with the findings in [62].

3.4.5.2 Errors in the eigenvectors

Besides a bias in the eigenvalues, errors also occur in the sample eigenvectors: the sample eigenvectors do not align with the population eigenvectors. Since these misalignments are randomly oriented, the average orientation of the sample eigenvectors is still equal to the population eigenvectors, so the misalignment cannot directly be described by a simple bias term like with the sample eigenvalues and the population eigenvalues. In [37] we proposed to study this misalignment problem by looking at the inner product of the sample eigenvectors with the population eigenvectors. We suggested that a relation would exist similar to the relation the MP equation gives between the sample eigenvalues and the population eigenvalues. In [63] attempts to find such a relation are presented.

In these papers it is suggested that the inner product of the sample eigenvalues with the population eigenvalues is influenced by ratio $\frac{p}{N}$ as well. However, since we do not have a description of this relation, we will again focus mainly on the two limit cases considered before in the eigenvalue bias analysis: the case in which $N >> p$ and the case in which $p >> N$.

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 should match at least one population eigenvalue).

In section 3.4.6.2 it will be shown that if $p >> N$ all information about the population eigenvalues is lost, except for its mean value. This means that apparently it makes no difference in the final estimate if the original population eigenvalues are all equal. If all population eigenvalues are equal, then the population eigenvectors form a random basis; any rotation of that basis is a valid set of eigenvectors as well. We therefore make the following hypothesis:

Hypothesis 3.4.1 *Under the same conditions as assumed in section 3.4.5.1 not only most of the structure in the population eigenvalues can not be retrieved from the sample estimates,* *but also all structure in the eigenvectors is lost, making the sample eigenvectors a random basis.*

Although we do not have a formal proof of this hypothesis, we present 3 arguments in favour of it. Firstly, as just described all structure of the population eigenvalues is lost except for their mean during the estimation, 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. The sample eigenvalues in each of these clusters have the same value, so any rotation of the corresponding eigenvectors in the corresponding subspace is a valid set of eigenvectors as well. Therefore, within the two subspaces, hypothesis 3.4.1 is true.

Thirdly, we can demonstrate the hypothesis experimentally. In Figure 3.8 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 over N ratios, which represents the limit of uniform inner product between the sample eigenvector and the population eigenvectors.

This concludes our argument that the sample eigenvectors form a random basis with respect to the population eigenvectors, which then supports our assumption that for very large p, the PCA dimensionality reduction results in a random subspace.

3.4.5.3 Limits in high dimensional verification

In section 3.4.5.1 we determined the sample eigenvalues for a single distribution for the limit $p \to \infty$ and showed that eigenvalues split into two clusters: a cluster with all eigenvalues equal to $\gamma\bar{\lambda}$, and a cluster with all zero valued eigenvalues. Verification results however depend on the log likelihood ratio, which in turn depends on two distributions: the within class distribution and the between class distribution, and in particular on the relation between these two distributions. So for verification it is more important to determine what in the limit $p \to \infty$ the relation is between the within class distribution and the between class distribution.

To find this relation, we start with the model of section 3.4.3 and determine the limit for the two involved distributions. We assumed a Gaussian distribution for both the within class variations and the between class variations and we assumed that the samples are independent and identically distributed (i.i.d.). A difficulty is that we do not observe these parts separately, we only have the total samples and their class labels. We therefore introduced the estimators $\hat{\mathbf{\Sigma}}_\mathrm{w}$ and $\hat{\mathbf{\Sigma}}_\mathrm{b}$ (equations 3.33 and 3.34 respectively). It is rather straight forward to show that $\hat{\mathbf{\Sigma}}_\mathrm{w}$ only depends on $\boldsymbol{x}_\mathrm{w}$ and its eigenvalues $\boldsymbol{l}_\mathrm{w}$ adhere to the MP equation.

The estimation of $\mathbf{\Sigma}_\mathrm{b}$ is not without problems though: in [52] we show that $\hat{\mathbf{\Sigma}}_\mathrm{b}$ is an estimate of a mixture of $\mathbf{\Sigma}_\mathrm{b}$ and $\mathbf{\Sigma}_\mathrm{w}$ instead of a pure estimate of $\mathbf{\Sigma}_\mathrm{b}$. However, the

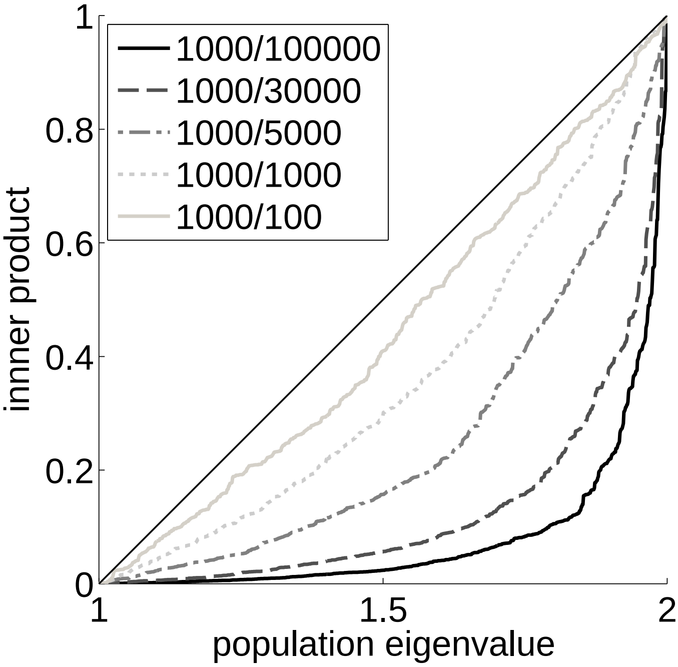


Figure 3.8: Sum of the squared inner products of the first sample eigenvector with all the population eigenvectors for several p/N ratios. The eigenvectors are indexed according to their corresponding eigenvalue.

number of samples for both the crosstalk part and the between part are equal and both sample parts are normally distributed, so the eigenvalues of $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$, $\boldsymbol{l}_{\mathrm{b}}$ also adhere to the MP equation.

Therefore the eigenvalues of both $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$ adhere to the limit described in section 3.4.5.1: for large p, $\boldsymbol{l}_{\mathrm{b}}$ separate into a cluster of $C-1$ eigenvalues with value $\frac{p}{C-1}\bar{\lambda}_{\mathrm{b}}$ and a cluster of $p-C+1$ zero valued eigenvalues. $\boldsymbol{l}_{\mathrm{w}}$ separates into a cluster of $N-C$ eigenvalues with value $\frac{p}{N-C+1}\bar{\lambda}_{\mathrm{w}}$ and a cluster of $p-N+C$ zero valued eigenvalues.

Since we assumed that the samples of our training data are the sum of a within class part and a between class part, the samples themselves are also normally distributed. However, they are not i.i.d.: their distribution depends on the class to which they belong. Therefore the limit distribution of the eigenvalues of $\hat{\mathbf{\Sigma}}_{\mathrm{t}}$, $\boldsymbol{l}_{\mathrm{t}}$, is not described by the MP equation and has to be determined in another way. In the following paragraphs we will 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 3.4.1 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. Since these subspaces are no longer dependent on the population parameters and the samples on which they are based are otherwise independent, both subspaces will be randomly oriented with respect to each other.

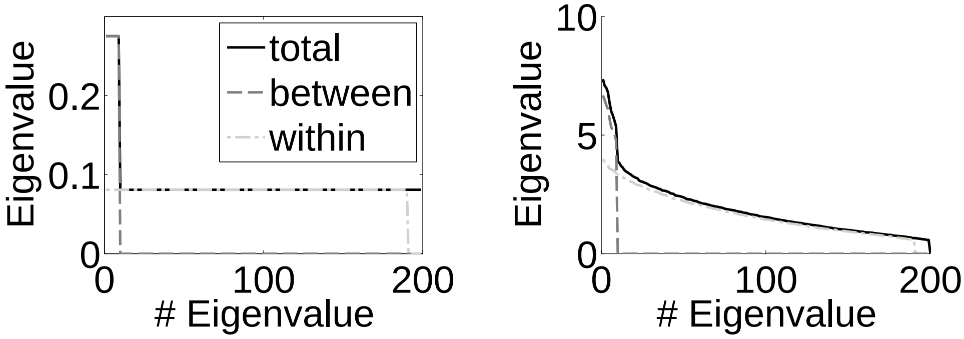
In section 3.4.5.2 it was determined that the limit distribution of the inner product of any of the sample eigenvectors with the population eigenvectors is a ramp function, which we assume is also the limit in probability for any unitary random vector $\boldsymbol{v}$ with an independent basis. The inner product between $\boldsymbol{v}$ and any of the population eigenvectors then becomes a random variable with expectation p^{-1} and a distribution which is scaled by p^{-1} as well.

Therefore 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 $p\to\infty$. Therefore both subspaces will become 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 (or even identification) system: every class is separable, a typical overtraining result. Note as well that this orthogonality was the last point required in the explanations of section 3.4.4.

To demonstrate the limit 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 3.4.4. We generated only 10 classes for the training set, with 20 samples per class. The theoretical limit for this configuration is shown in figure 3.9a. Figure 3.9b shows an estimate for $p = 4000$. Clearly, even for $p = 4000$ the estimate has not converged to the limit yet, but it seems to confirm that the determined limit is correct.



(a) Calculated limit

(b) Estimate for $p = 4000$

Figure 3.9: Second order estimation in a verification scheme for the limit $p \to \infty$.

3.4.6 Improved second order statistics in high dimensional spaces

In the previous sections we described several effects which 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 3.4.4. In section 3.4.4 we also showed, based on the analysis in section 3.4.5, 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 will describe several improvements which will finally result in a system which smoothly changes from theoretical optimal sample estimate if $N >> p$ to the regularisation limit for $p >> N$. These improvements will be experimentally evaluated in section 3.4.7.

3.4.6.1 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 3.10. In figure 3.10 the introduction of the bias 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 estimate of the inverse of $B(\lambda)$ to $\boldsymbol{l}$, resulting in a corrected estimate of the population eigenvalues $\hat{\boldsymbol{\lambda}}^c$.

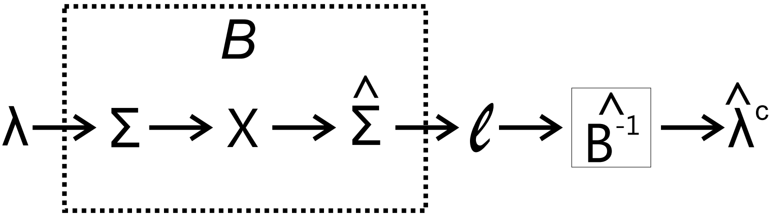


Figure 3.10: Schematic representation of bias correction.

We use 3 bias correction methods in the remainder of this article. The first method is the correction method developed by Karoui in [40]. 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 the estimates of the population eigenvalues still have to be determined.

The second method is the fixed point eigenvalue correction [64]. 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 [14]. It is based on regularisation, but unlike many other regularisation algorithms it is able to handle situations in which $p > N$.

We will compare these algorithms with the classical approach of dealing with the problems related to high dimensionality: the PCA dimensionality reduction. In the experiment of section 3.4.4 we determined that the configuration of a reduction to a fixed dimensionality of 150 components leads to some of the best results for the specific settings of the experiment.

3.4.6.2 Correction limits

In the experiment of section 3.4.8 we will vary p, so all methods are tested for different ratios of $\frac{p}{N}$. In section 3.4.5.1 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 will necessarily contain $N - p + 1$ 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 [35], 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 3.4.5.1, 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 [35], 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 the regularisation limit.

3.4.6.3 Dimensionality reduction by random projection

In the analysis of the SOS estimation for $p >> N$ we determined that the sample eigenvalues are divided into 2 clusters: one cluster of $N - 1$ eigenvalues of value $\frac{p}{N-1}\bar{\lambda}$ and one cluster of zero valued eigenvalues, and the sample eigenvectors are independent from the population eigenvectors. PCA dimensionality reduction therefore turns the probability calculations of equation 3.32 into a euclidean distance measure in a random $N - 1$ dimensional subspace.

This leads to a remarkable possibility for improving classification results: consider a projection of the data on a random basis with a dimensionality somewhat larger than $N - 1$ before applying PCA dimensionality reduction. Note that each of the random basis vectors has on average an equal inner product with all the population eigenvectors, so the high and low discriminative space have an equal portion in this random subspace as in the full space, so we effectively reduced the dimensionality and moved to the left on the p axis in Figure 3.6, aside from some statistical fluctuations, without determining the SOS. If we now estimate the SOS, we have a smaller value of p with still the same N, so the structure we estimate is more based on the population distribution in that subspace.

We will not perform any experiments with this solution, but we use it as an illustration of the weakness of the PCA dimensionality reduction solution for large p over N ratios.

3.4.6.4 Variance correction

In section 3.4.5.2 we noted that besides a bias in the sample eigenvalues there is also a misalignment between the sample eigenvectors and the population eigenvectors. We also noted that the average position of the sample eigenvectors is correct, so we cannot improve upon the estimates of the eigenvectors. However, if we consider the eigenvalues as the estimates of the variances along the eigenvectors, some additional correction can be done on the corrected eigenvalues, which we denote by variance correction.

The basic idea is demonstrated in Figure 3.11. In the figure we show the results of an estimation of the SOS of an artificial estimation problem. The curves in the figure represent the variance in each direction from the origin. The solid dark line gives the population variance curve, so the population eigenvectors are aligned along the vertical axis and the horizontal axis, with corresponding eigenvalues 2 and 1 respectively.

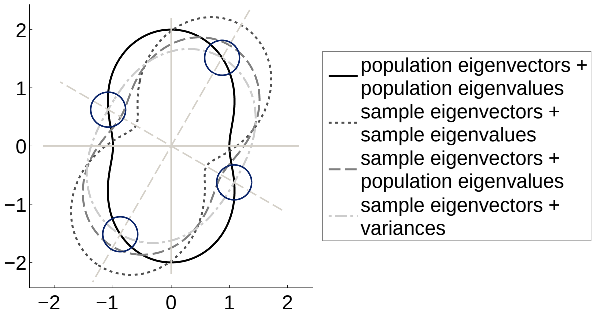


Figure 3.11: Variance correction explanation in a 2D example. The solid straight lines show the population eigenvectors, the dashed straight lines the sample eigenvectors.

A possible sample estimate is given by the dotted curve, so the largest sample eigenvalue is larger than the largest population eigenvalue, and the other sample eigenvalue is smaller than the smallest population eigenvalue, which is typical for a sample estimate. Also, the sample eigenvectors are not aligned with the population eigenvectors.

Now assume that we have a perfect bias correction method, so from the sample eigenvalues we would get a perfect estimate of the population eigenvalues. The resulting variance curve would be a combination of the sample eigenvectors with the population eigenvalues as is represented by the dashed curve. Note that if we consider the eigenvalues a measure of the variances along the sample eigenvectors, then the resulting estimate is a biased estimate: the largest corrected sample eigenvalue is a too large as a variance estimate along its corresponding sample eigenvector while the smallest corrected sample eigenvalue is too small as a variance estimate along its corresponding sample eigenvector.

In section 3.4.5.2 we suggested that the misalignment problem should be studied by examining the inner product of the sample eigenvectors with the population eigenvectors. As it turns out, we do not require the population eigenvectors themselves to determine the square of this inner product. In [37] we explain this and show how we can determine the true variances along the sample eigenvectors given a perfect sample eigenvalue bias correction. If we combine these new variance estimates with the sample eigenvectors we get an estimate represented by the light dot dashed curve in figure 3.11. As the circles show, we now do have a perfect estimate of the variances along the sample eigenvectors as the light solid curve and the dark solid curves cross on the sample eigenvectors, which turned out to give better verification performance in the experiments presented in section 3.4.8.

3.4.7 Bias correction in verification

The verification decision as described in section 3.4.3 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 $\mathbf{\Sigma}_{\mathrm{w}}$ and $\mathbf{\Sigma}_{\mathrm{t}}$, where $\mathbf{\Sigma}_{\mathrm{t}}$ can be split into $\mathbf{\Sigma}_{\mathrm{w}} + \mathbf{\Sigma}_{\mathrm{b}}$.

Estimates of all these matrices are based on a limited amount of samples and so their eigenvalues are biased. The question is how this bias affects the verification process and how we can apply bias correction in verification (BCIV). We have to correct two out of the three sample eigenvalue sets: $\boldsymbol{l}_{\mathrm{w}}$, $\boldsymbol{l}_{\mathrm{b}}$ and $\boldsymbol{l}_{\mathrm{t}}$, which are the sample eigenvalues of $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$, $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{t}}$ respectively. The third set is fixed because there is a fixed relation between $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$, $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{t}}$ as shown in [52].

As described in section 3.4.5.3, only the estimation of $\mathbf{\Sigma}_{\mathrm{w}}$ and $\mathbf{\Sigma}_{\mathrm{b}}$ adhere to the requirements of the MP equation and so the previous analysis and the suggested improvements apply only to their estimates, not the estimate of $\mathbf{\Sigma}_{\mathrm{t}}$. However, if $N \approx 2C$, the effective number of samples for the within estimate and the between estimate 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.

If $\boldsymbol{l}_{\mathrm{b}}$ and $\boldsymbol{l}_{\mathrm{t}}$ are corrected, then the estimate of $\mathbf{\Sigma}_{\mathrm{w}}^{\mathrm{c}}$, needed for the log likelihood ratio (equation 3.32), follows from the subtraction $\hat{\mathbf{\Sigma}}_{\mathrm{t}}^{\mathrm{c}} - \hat{\mathbf{\Sigma}}_{\mathrm{b}}^{\mathrm{c}}$. This subtraction can easily lead to negative eigenvalues, and therefore this correction is not considered any further.

The option to correct $\boldsymbol{l}_{\mathrm{w}}$ and $\boldsymbol{l}_{\mathrm{t}}$ we will denote by improper BCIV, since in general the bias in $\boldsymbol{l}_{\mathrm{t}}$ does not adhere to the MP equation as we just described. This correction option is still considered though since its error is only significant for larger p values or if N significantly differs from $2C$. Based on this considerations and previous tests in [53], we decided in [65] to perform the bias correction on $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{t}}$, using the total number of samples as the amount of samples used for the estimation of $\hat{\mathbf{\Sigma}}_{\mathrm{t}}$.

The last option is to correct $\boldsymbol{l}_{\mathrm{w}}$ and $\boldsymbol{l}_{\mathrm{b}}$. The estimation of $\mathbf{\Sigma}_{\mathrm{b}}$ is not without problems: in [52] we show that $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$ is an estimate of a mixture of $\mathbf{\Sigma}_{\mathrm{b}}$ and $\mathbf{\Sigma}_{\mathrm{w}}$ instead of a pure estimate of $\mathbf{\Sigma}_{\mathrm{b}}$. In combination with eigenvalue bias, this has several effects:

1. Due to the crosstalk, $\mathcal{E}\left\{\hat{\mathbf{\Sigma}}_{\mathrm{w}} + \hat{\mathbf{\Sigma}}_{\mathrm{b}}\right\} = \mathbf{\Sigma}_{\mathrm{t}} + \frac{1}{N_{pc}}\mathbf{\Sigma}_{\mathrm{w}}$, where $N_{\mathrm{pc}} = \frac{N}{C}$. This leads of course to erroneous estimates of $p\left(\boldsymbol{x}\right)$, so $\hat{\mathbf{\Sigma}}_{\mathrm{t}}$ should be estimated by

$\frac{N_{\mathrm{pc}}-1}{N_{\mathrm{pc}}}\hat{\mathbf{\Sigma}}_{\mathrm{w}} + \hat{\mathbf{\Sigma}}_{\mathrm{b}}$.

2. The crosstalk changes the directions for which the likelihood ratio is most sensitive (see the end of section 3.4.3). The crosstalk will change equation 3.35 into

$$4\left(\frac{1}{N_{\mathrm{pc}}}\frac{\boldsymbol{w}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{b,cross}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{w}}\boldsymbol{w}} + \frac{\boldsymbol{w}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{b,nocross}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{w}}\boldsymbol{w}}\right) \tag{3.48}$$

   where $\hat{\mathbf{\Sigma}}_{\mathrm{b,cross}}$ is the estimate of the crosstalk of $\mathbf{\Sigma}_{\mathrm{w}}$ in $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{b,nocross}}$ is the between estimate without the crosstalk. A crosstalk part which is equal to $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$ 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, $\hat{\mathbf{\Sigma}}_{\mathrm{b,cross}}$ is estimated from a different part of the samples than $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$, usually with a different number of samples as well ($N - C$ for $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$ and $C - 1$ for $\hat{\mathbf{\Sigma}}_{\mathrm{b,cross}}$). 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 $\hat{\mathbf{\Sigma}}_{\mathrm{b,cross}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$ solely determine the sensitivity of the log likelihood ratio.

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

The correction of the $\boldsymbol{l}_{\mathrm{b}}$ is however theoretically sound (see section 3.4.5.3). The $\boldsymbol{l}_{\mathrm{w}}$ and $\boldsymbol{l}_{\mathrm{t}}$ correction is no solution for the crosstalk problem, because the difference between $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{t}}$ is solely caused by $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$. So it seems that the $\boldsymbol{l}_{\mathrm{w}}$ and $\boldsymbol{l}_{\mathrm{b}}$ correction is the best option, however, it runs into problems if p is close to or larger than N. This will be explained in the next section.

3.4.7.1 Correction in null spaces and verification

Due to the difference in effective number of samples for $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$ a problem arises if p is considerably larger than 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 3.12), so there is no discriminative distinction between the different orientations. The sample estimates, given by the dark dashed line and the light dash-dotted line, do however 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, which wrongfully suggests that that part of the null space is highly discriminative.

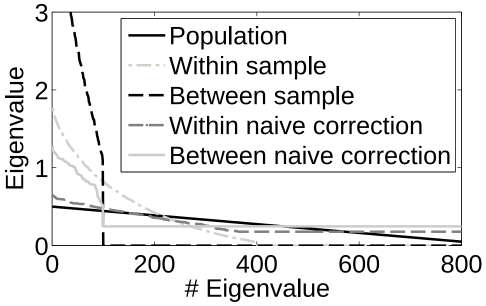


Figure 3.12: Naive BCIV example.

As shown in section 3.4.6.2, bias correction leads to equal values for all eigenvalues in the null space. The null spaces of $\hat{\mathbf{\Sigma}}_\mathrm{w}$ and $\hat{\mathbf{\Sigma}}_\mathrm{b}$ differ both in number of dimensions ($p - (N - C)$ and $p - (C - 1)$ respectively) and orientation. Therefore the average with which the null space of $\hat{\mathbf{\Sigma}}_\mathrm{w}$ is corrected can be much lower than the average of the null space of $\hat{\mathbf{\Sigma}}_\mathrm{b}$, as the example demonstrated. As a result the ratio between the last few eigenvalues can become arbitrarily large and therefore completely determine the verification results (equation 3.35).

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.

3.4.7.2 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 $\hat{\mathbf{\Sigma}}_\mathrm{w}$ and $\hat{\mathbf{\Sigma}}_\mathrm{b}$ are genuine sample covariance matrices in the sense that their eigenvalues follow the general bias relations. The eigenwise BCIV therefore corrects this combination of covariance matrices.

To prevent the problems described in the previous sections the method consists of the following steps, shown schematically in figure 3.13:

1. Estimate $\hat{\mathbf{\Sigma}}_\mathrm{w}$ and $\hat{\mathbf{\Sigma}}_\mathrm{b}$.

2. Decompose $\hat{\mathbf{\Sigma}}_\mathrm{w}$ 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 $\hat{\boldsymbol{\lambda}}^\mathrm{c}_\mathrm{w}$.

4. Use $\hat{\boldsymbol{E}}_\mathrm{w}$ and $\hat{\boldsymbol{\lambda}}^\mathrm{c}_\mathrm{w}$ to determine $\hat{\mathbf{\Sigma}}^\mathrm{ww}_\mathrm{b}$, which is the between covariance matrix if the within data is whitened.

5. Decompose $\hat{\mathbf{\Sigma}}^\mathrm{ww}_\mathrm{b}$ into eigenvectors $\hat{\boldsymbol{E}}^\mathrm{ww}_\mathrm{b}$ and eigenvalues $\boldsymbol{l}^\mathrm{ww}_\mathrm{b}$.

6. Correct $\boldsymbol{l}^\mathrm{ww}_\mathrm{b}$ to get $\hat{\boldsymbol{\lambda}}^\mathrm{c,ww}_\mathrm{b}$.

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 are based on independent parts of the samples, which is the case if the assumed data model described in section 3.4.3 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 encountered when using the naive BCIV? The first three steps are the same as the naive BCIV. In step 4, both matrices can be considered to be scaled and rotated, so argmin$\frac{w^{\mathrm{T}}\cdot\Sigma_{\mathrm{w}}\cdot w}{w^{\mathrm{T}}\cdot\Sigma_{\mathrm{b}}\cdot w}$ remains unchanged, so the order of the basis vectors according to their discriminative ability remains unchanged.

In step 6, bias correction is applied, which is order preserving, so the order of the between eigenvalues stay the same. Since the within 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. As a result, the null space is the least discriminating subspace, and the problem of the naive BCIV is solved.

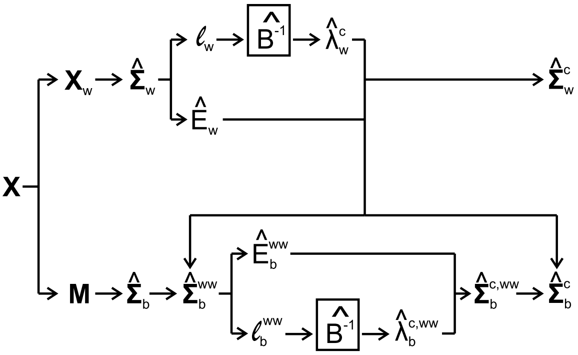


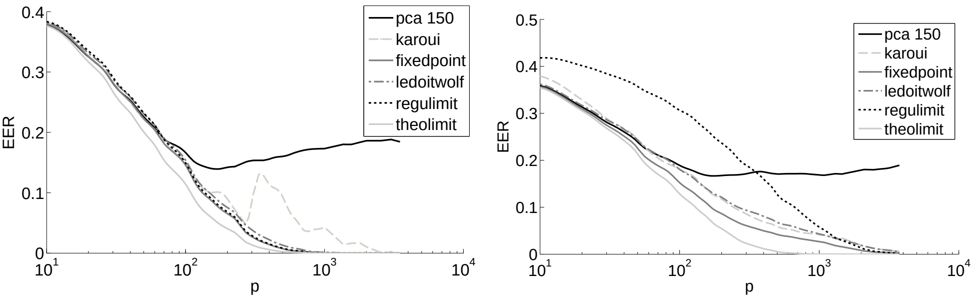
Figure 3.13: Schematic representation of the eigenwise BCIV.

3.4.8 Bias correction in verification approaches comparison

In section section 3.4.4 we presented an experiment which showed that 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, which we will test in this section.

We repeat the experiment of section 3.4.4 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 3.4.6.1: 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 3.14 shows the resulting EER curves for the different methods. As seen before, the PCA method is outperformed by the regularization limit for even modest p values, but it is also outperformed by all 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 like the PCA methods.



(a) All total eigenvalues equal

(b) Exponential total eigenvalue distribution

Figure 3.14: 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 the number of samples 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 3.14a), the almost zero valued eigenvalues still deteriorate the results.

3.4.9 Conclusion and discussion

We studied the effect of increasing the dimensionality p of training data while keeping the number of observations N fixed. 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, because otherwise the estimator could be improved by ignoring the added dimensions.

The classical PCA dimensionality reduction technique does not display this property: in our verification experiments for a certain p, increasing p even more actually increases the EER. A detailed analysis showed that this is because the Second Order Statistics (SOS) estimation gets more and more influenced by random fluctuations in the data instead of the actual SOS of the data generating process.

The analysis of the effects of these random fluctuations show that the sample eigenvalues are biased and the sample eigenvectors are misaligned. Based on these analysis we suggested several improvements for SOS estimation in order to achieve estimators which do show a non increasing EER with increasing p.

We also argued that the optimal estimators should converge from theoretical SOS estimation limit (for low p) to a regularisation limit (for high p). The eigenwise BCIV, based on the fixed point bias correction displays such characteristics, although it does have somewhat larger EER than the regularisation limit for very large p. At lower complexity and resources cost, the Ledoit Wolf correction already showed a large improvement compared to the classical PCA dimensionality reduction method.

These results all require a close fit to the model. Although it was shown in [66] 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. 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 [67, 68]. Therefore, BCIV applied to real data does not improve results as much as shown in our experiments if the data is not correctly modelled.

Furthermore, the variants we added to the data contained a similar discriminative capacity as the original data. If variants are added which have a smaller discriminative capacity than the original set, then the average discriminative capacity decreases. Since all structure is lost for very large p over N ratios, adding these less informing variants may actually increase the error rates of the verification system. The crosstalk of the within class variance on the between class variance estimate combined with the eigenvalue bias inhibits the detection of dimensions with very little discriminative capacity and error rates will go up even further.

So even if the corrections presented here are applied, adding dimensions by for example increasing the resolution of images can still cause error rates to increase.

3.4.10 Epilogue

In this paper we showed how the theory on the effect of high dimensionality on SOS estimation of a single distribution can be extended to verification problems. We showed that solving the singularity problems using the classical PCA dimensionality reduction solution is far from optimal and even fails completely for very large p. We also presented the eigenwise correction, which does perform close to optimal.

However, if eigenwise correction is applied to a verification system with real facial data, the error rates still go up, although not as high as when Naive BCIV is applied. To demonstrate this we performed a verification experiment with facial data from the FRGC database. The results are shown in figure 3.15.

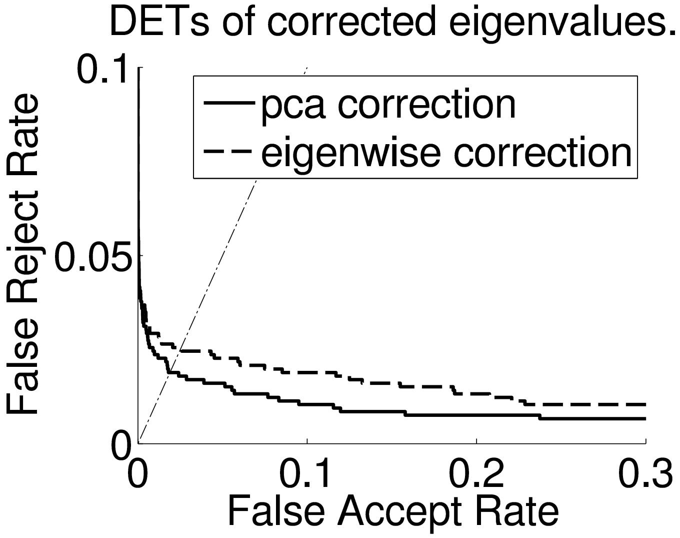


Figure 3.15: DET curves with PCA correction and eigenwise correction applied to real facial data from the FRGC2 database.

The results clearly show that PCA dimensionality reduction outperforms the eigenwise correction if applied to the estimates from real facial data.

3.5 Conclusion

In this chapter we studied the application of bias correction in a verification setting. Straightforward application of bias correction proved to be problematic: correction of the between matrix could lead to an arbitrarily high between over within class variance ratios in the null space of the sample estimates, thereby assigning the null space an arbitrarily high discriminative capacity. With a detailed analysis of SOS estimation in a verification setting, we developed the eigenwise correction, which solved this problem and demonstrated a close to optimal behaviour in synthetic data experiments.

However, in systems using real facial data, the eigenwise correction is still outperformed by the PCA dimensionality reduction, despite the theory showing its weaknesses. This led us to question the assumptions on which SOS estimation is based. In the next chapter we therefore focus on these assumptions and introduce deviations from this model.

Chapter 4

Limitations of intensity sources as data model

> Alice: "If I had a world of my own, everything would be nonsense. Nothing would be what it is because everything would be what it isn't. And contrary-wise; what it is it wouldn't be, and what it wouldn't be, it would. You see?" [69]

4.1 Modeling introduction

Some people tend to see all sorts of things when they look at somebodies face [1] or look into somebodies eyes [2]. However, it has proven to be a challenging task to let the computer make sense of facial image data. A big part of this problem is to detect where and how the information is encoded. That the encoding of information is non trivial is demonstrated in [70] by the outer message, message frame and the inner message model. With that model the author explains that before any receiver can decode any message, it should first be triggered to notice that there is a message (the outer message telling the receiver that this is actually a message), locate it and know how it is encoded (the message frame should indicate this).

This is exactly what is attempted with automated face recognition: the input of the system is a collection of pixels. We want a program that automatically detects that there is a message present (in the form of a face) and automatically decode the identity information from this message. The first step, receiving the outer message by determining if there is a message, is to determine if there is a face in the image and determine its global position. This is already quite developed. However, the problem of how to decode the identity information from the face image is still lacking.

[1] "Everytime I see your face", Live, Birds of Pray

[2] "November rain", Guns N' Roses, Use Your Illusion I

In the previous chapters we found some clues that the implicit model in the classical SOS estimation approach might not be suitable for high resolution images. One of the implicit assumptions in classical SOS estimation is that the information in face images is essentially encoded in the intensity of pixels at fixed positions. In the next chapter we study a different model, the position sources model, which assumes that the information is encoded in the position of features in the face rather than the intensity. We therefore derived a second derived research question, presented in section 1.8: "What effect does the presence of position sources in data have on systems based on the fixed position intensity sources model and can it explain the observations made after increasing the image resolution of facial data: the high number of sources estimated, the 1 over f characteristic of the eigenvalue scree plot, the saturation of performance of biometric systems based on SOS estimation and that PCA performs better on real facial data than bias correction?"

We study this question with one paper, one section with unpublished analysis of position sources in intensity modeled data and one section presenting an unpublished experiment. The first paper, presented in section 4.2, studies the first part of the research question in particular: "What effect does the presence of position sources in data have on systems based on the fixed position intensity sources model?" Although it is not published yet, it is an extension of [47] which has been published. It shows that with a relative simple configuration of position sources already several of the SOS characteristics of facial data can be recreated.

It first demonstrates some of the characteristics of the scree plots obtained if SOS are estimated from facial image data. These scree plots are already used as model for generating synthetic data with SOS similar to facial data in chapters 2 and 3. These characteristics are that the number of estimated sources grow with the number of samples without a clear distinction between signals and noise sources in the scree plot, and that the scree plot can be described by a 1 over f model, where f in this case is the index of the eigenvalue.

After that, the position sources model is introduced and we explain how face information could be encoded in that model. We show that if some of the information is indeed encoded in this manner, this has a significant effect on the estimated SOS based on the intensity sources model and it results in characteristics quite similar to the observed characteristics of facial image data. In a verification experiment it is even shown that the observations of section 3.3 can be explained with this model.

In section 4.3 we give a theoretical analysis of the effect position sources have on the SOS estimates based on the intensity model. In [71] it was already shown that the result of applying PCA to a set of rotated images has some remarkable effects: although there is only one changing source (the angle of rotation) multiple intensity sources are estimated from this data set. Moreover, the eigenvalues do not have the be estimated with an eigenvalue decomposition, they can also be found using Discrete Cosine Transform (DCT).

In section 4.3 we determine the covariance matrix and its decomposition of an data set based on 1 Gaussian feature moving around guided by 2 Gaussian position sources. These analysis show that PCA basically performs a frequency decomposition in that case and PCA dimensionality reduction therefore effectively becomes a low pass filtering operation. This explains why the bias correction introduced in the previous chapter performs so poorly on real facial data while on synthetic data it clearly outperforms PCA dimensionality reduction.

The analysis in sections 4.2 and 4.3 is mainly based on synthetic data. In section 4.4 we therefore report an experiment with real facial data in which the use of position sources gives a more efficient decoding scheme than SOS estimation based on the fixed intensity source model solely.

4.2 Position sources in intensity modeled data [3]

4.2.1 Prologue

We recommend to read the following introduction until the characteristics we try to explain are summarized. The remainder of the introduction of the introduction elaborates on these characteristics and can be skipped if these characteristics are accepted as is. Section 4.2.3.1 largely contains the same introduction on SOS estimation as given before, but with an emphasis on what we mean with intensity sources. We therefore recommend reading this section if the concept of intensity sources remained unclear. Section 4.2.3.2 contains again a general introduction to the bias in the sample eigenvalues and can therefore be skipped. Section 4.2.3.3 introduces the position sources model and we recommend reading the remainder of the article from that point on.

4.2.2 Introduction

Linear image models, like PCA presume that the images can be approximated by a weighed sum of basis images. In face recognition, this assumption implies that facial features have a fixed position in facial images, so the model can be described as the fixed position intensity sources model. We will show that if some of the features do not adhere to this fixed position assumption, they considerably distort the model fitting and consequently deteriorate the performance of biometric systems based on these linear image models.

We introduce position sources to model this position variability of face features. In the position source model we assume that the relevant information is encoded in the position of objects or features rather than in their intensity, so the models description can be denoted by the fixed intensity position sources model. With position sources we show that several of the characteristics of face data can be explained:

1. The high number of intensity sources required in the modeling of facial data.

[3] Position sources as the limiting factor of PCA based face recognition [72], to be published. It is an extension of The effect of position sources on estimated eigenvalues in intensity modeled data, Thirty-first Symposium on Information Theory in the Benelux, 2010 [47]

2. The curve of the eigenvalues can be described for a large part with high accuracy by a 1 over f curve, where f is the eigenvalue index.
3. The effect of using higher resolution images as input on verification performance saturates: after a certain point increasing the image resolution does not improve the biometric system performance and it may even deteriorate the performance.
4. Advanced SOS estimators designed to operate on high dimensional training sets with a relatively low number of samples, like bias correction methods, do not improve systems performance or even deteriorate it.

Especially the third characteristic becomes more and more relevant with the increasing resolution of the images in facial image databases, like for example the FRGC 2 database [39]. Since high resolution images contain more information, it is generally expected that an increase in resolution leads to better performance, but as was shown in [8] experimentally, this is not true for systems based on the linearity assumption.

In [71] it was already shown that model errors can determine the characteristics of the eigenvalue plots: Uenohara and Kanade showed that the eigenvalues of a training set composed of rotated versions of one image can be determined using DCT as well as by doing an eigenvalue decomposition.

Characteristic 4 is based on the common use of PCA dimensionality reduction as a solution to the singularity problems occurring if the number of samples in a training set is lower than the dimensionality of the training samples. This singularity problem is a direct effect of the bias of the eigenvalues estimated from the training set [29, 26, 49]. Several methods exist which can remove the bias from the estimate, and if the assumed model during estimation is accurate, these methods clearly outperform PCA dimensionality reduction as was shown in [65]. However, if bias correction is applied to estimation results based on real facial data, the correction has little effect or even reduces the performance of biometric systems as shown in [48].

To demonstrate that the position source model leads to the characteristics just described, we performed a verification test in which we used synthetic data. The system assumes that the data it receives is generated via the model used in LDA, which uses the model of linear combination of base images. The synthetic data is however generated according to the position sources model.

Before presenting the experiment, we first give a more detailed description of the modeling process in section 4.2.3, where section 4.2.3.1 gives the classical fixed position intensity model. Our position sources model is presented in section 4.2.3.3. In section 4.2.4 the experimental set-up will be discussed, where the verification system used is described in section 4.2.4.1 and the used bias correction method is presented in 4.2.4.2.

Section 4.2.5 describes the experiments and the results, which show that the position sources model is indeed capable of explaining the characteristics of the face data we just described. In section 4.2.6 we draw conclusions and discus the implications of the presence of position sources in data.

4.2.3 Image data modeling

Data modeling is required for two reasons. First of all images usually contain a lot of irrelevant information. Therefore the relevant information has to be extracted from the images, which is often done by performing PCA.

The second reason for data modeling is that we want to determine in a verification setting if a certain face image belongs to a claimed identity. One approach is to compare the likelihood of a sample assuming the claimed identity is correct, $P\left(X|C=c\right)$, with the likelihood of the sample assuming the claimed identity is incorrect, $P\left(X|C\neq c\right)$. To evaluate these likelihoods, the input data has to be modeled. A commonly used model for this problem is the LDA model.

Both PCA and LDA are based on the fixed position intensity model which will be elaborated in the next section.

4.2.3.1 Image data modeling and model training using Second Order Statistics

As stated before, in the fixed position intensity sources model it is assumed that images are a weighted mixture of base images. The information is then encoded in these weights. To determine these weights in new input images, first the base images have to be determined. These can be determined using the following procedure.

First the samples are modeled as a random variable. Its distribution is usually unknown, but it is often assumed that, considering the number of samples available for training, the distribution of the random variable can approximated with a multivariate Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$, which is completely determined by its mean

$$\boldsymbol{\mu}=\mathcal{E}\left\{\boldsymbol{x}\right\} \tag{4.1}$$

and the second order statistics represented by the covariance matrix

$$\boldsymbol{\Sigma}=\mathcal{E}\left\{\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{\mathrm{T}}\right\} \tag{4.2}$$

The covariance matrix can be decomposed into $\boldsymbol{E}\cdot\boldsymbol{D}\cdot\boldsymbol{E}^{\mathrm{T}}$, where $\boldsymbol{E}$ is a rotation matrix and each column is an eigenvector. $\boldsymbol{D}$ is a diagonal matrix with the eigenvalues on the diagonal. The eigenvectors indicate in which direction the largest variances occur in the data and they are the base images as described in the linear model. The eigenvalue indicates the variance in the directions of the eigenvectors.

Since the statistical parameters are usually unknown, they have to be estimated from a set of examples, the training set. The mean can be estimated from the sample mean, $\hat{\boldsymbol{\mu}}=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k$, where $\boldsymbol{x}_k$ is the k^{th} sample of a training set of N samples. $\boldsymbol{\Sigma}$ can be estimated with the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{N-1}\sum_{k=1}^{N}\left(\boldsymbol{x}_k-\boldsymbol{\mu}\right)\cdot\left(\boldsymbol{x}_k-\boldsymbol{\mu}\right)^{\mathrm{T}} \tag{4.3}$$

The parameters of the data generating process will be denoted by the population parameters, while their estimates will be denoted by the sample parameters.

PCA assumes that the structure of the data can be found by determining the direction of largest variation. It is based on the fixed position intensity source model as follows. The samples are constructed by concatenating the pixels of the images after some preprocessing on the images. PCA then uses the model to determine the directions of largest variances, which are given by the eigenvectors corresponding to the M largest eigenvalues.

LDA, like PCA, is also implicitly based on the fixed position intensity model, but instead of finding the direction of largest variances as PCA does, LDA tries to determine the highest discriminating directions. It therefore models the samples $\boldsymbol{x}_k$ as a composition of two parts: $\boldsymbol{x}_k = \boldsymbol{x}_{w,k} + \boldsymbol{\mu}_{c(k)}$, where $\boldsymbol{x}_{w,k}$ differs for each sample and models the variation between samples from the same class and $\boldsymbol{\mu}_{c(k)}$ is the mean of class $c(k)$ and it models variations between samples of different classes.

For both parts the fixed position intensity model is assumed, so $\boldsymbol{x}_w$ and $\boldsymbol{\mu}_c$ are both modeled as random variables with distributions $\mathcal{N}\left(0, \boldsymbol{\Sigma}_w\right)$ and $\mathcal{N}\left(\mu_t, \boldsymbol{\Sigma}_b\right)$ respectively. The distribution of $\boldsymbol{x}$ is then given by $\mathcal{N}\left(\mu_t, \boldsymbol{\Sigma}_t\right)$, where $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_b$.

4.2.3.2 Eigenvalue bias

One effect of increasing the resolution is an increase of the number of pixels, resulting in a higher dimensionality of the samples $\boldsymbol{x}_k$.

In general, the increase in dimensionality is not met with a proportional increase in the amount of samples available for training. If the sample covariance matrix is used, this results in a problem: the sample eigenvalues become significantly biased if the ratio between the dimensionality of the training samples p and the number of training samples N grows [26, 29].

One effect is that if $p > N$, there will be $p - N$ zero valued sample eigenvalues. In [65] it was shown that this bias can be corrected such that if the added dimensions containing a similar ratio between signals and noise, the EER rate in a verification experiment can still go down by adding the data. This would suggest that using high resolution images would indeed improve the performance. However, these results were based on data which is accurately modeled by the intensity model.

4.2.3.3 Position sources

As stated before, both PCA and LDA implicitly model face images by a weighed sum of fixed position intensity sources, the eigenfaces or fisherfaces. This implies that all features in the face are at a fixed position. However, this is not the case for example the pupils, mouth corners and eyebrows. To model the position variability of the face features, we introduce the position sources model. In this model we assume that the information is encoded in the position of features instead of their intensity.

To determine the effect of a position source in intensity modeling, consider the setting in figure 4.1. In the figure, part of an image is shown, with a black pixel of which the position is determined by one position source. Also two pixels are shown (X_1 and X_2). For explanation purposes we assume that the black pixel moves from the left over X_1 and X_2 instead of drawing random samples from the position source

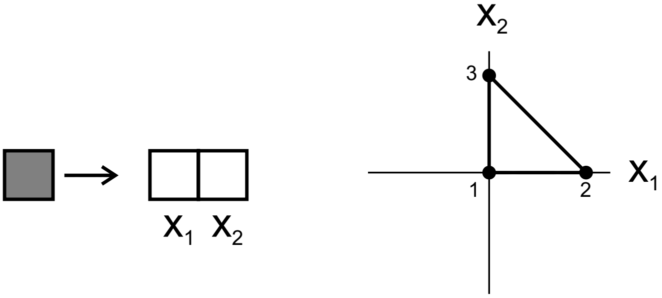


Figure 4.1: Example of 1 position source resulting in the estimation of multiple intensity sources.

distribution. On the right we show a graph on which the intensity of X_1 is plotted against the intensity of X_2 as we move the black square from left to right.

At the start, the black square is on the left and the intensity of both pixels is zero, so we start at position 1. As the black square moves to the right, it starts to overlap pixel X_1 so its intensity increases while X_2 remains unchanged until the black square completely covers X_1. In the graph we then moved from position 1 to position 2.

If the black square moves further, it starts to overlap X_2 while the overlap of X_1 decreases, increasing the intensity of X_2 while decreasing the intensity of X_1, until the black square completely covers X_2 and X_1 is cleared. This is represented in the graph by the line between position 2 and 3. After position 3, the black square moves away from X_2 so the intensity of X_2 decreases as well, until both X_1 and X_2 are no longer covered by the black square and so the curve ends in position 1 again.

The resulting curve on the right shows two directions of variation which requires modeling with two intensity sources. This shows that a single position source can generate multiple intensity source estimates. Moreover, the Gaussian intensity sources density, as suggested by the PCA model will be a highly inaccurate approximation of the curve in figure 4.1.

Note that if the variation of the position of the black square is small, we stay on one line segment of the curve and so only one intensity source is estimated. This suggest that objects with relative large size and low frequency content in their texture may still be accurately modeled with fixed position intensity sources.

4.2.4 Method

4.2.4.1 Verification system

To test the effect of position sources in fixed position intensity modeled data, we will perform a verification experiment. In a verification system the objective is to judge a claim that a sample x originates from class c. If (estimates of) the parameters of the previously described LDA model are known, both the likelihood of x belonging to class c, $P(\boldsymbol{x}|C=c)$, and the likelihood of $\boldsymbol{x}$ not belonging to class c ($P(\boldsymbol{x}|C \neq c)$) can be determined and the logarithm of the ratio between the two likelihoods becomes:

$$\log \frac{P(X|C=c)}{P(X|C \neq c)} = L(\boldsymbol{x}, c) \propto$$
$$(\boldsymbol{x}-\boldsymbol{\mu}_t)^{\mathrm{T}} \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_t) - (\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}_w^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_c) \tag{4.4}$$

where the $P(\boldsymbol{x})$ is used as an approximation of $P(\boldsymbol{x}|C \neq c)$. By setting a threshold above which the claims are accepted, a trade-off can be made between the claims erroneously accepted and rejected.

4.2.4.2 Eigenvalue bias test and correction

In the experiment the second order characteristics are estimated from high dimensional data. As described in section 4.2.3.2, this can lead to a severe bias in the estimated eigenvalues. To test whether the bias might have a significant effect at all, we follow a procedure shown schematically in figure 4.2a. The first part of the figure shows how the population eigenvalues $\boldsymbol{\lambda}$ and sample eigenvalues $\boldsymbol{l}$ are related: the sample eigenvalues are estimated from a data set $\boldsymbol{X}$ which is generated with the population eigenvalues as input parameters. Note that for the bias check, the involved eigenvectors are irrelevant, since the sample eigenvalue bias is independent of the population eigenvectors.

To determine if the sample eigenvalues are significantly biased these steps are repeated using the sample eigenvalues $\boldsymbol{l}$ as input parameters and generate a control data set $\boldsymbol{X}_{\mathrm{ctr}}$. If the eigenvalues $\boldsymbol{l}_{\mathrm{ctr}}$ estimated from this control set are close to the sample eigenvalues, then the sample eigenvalues might be reasonably unbiased.

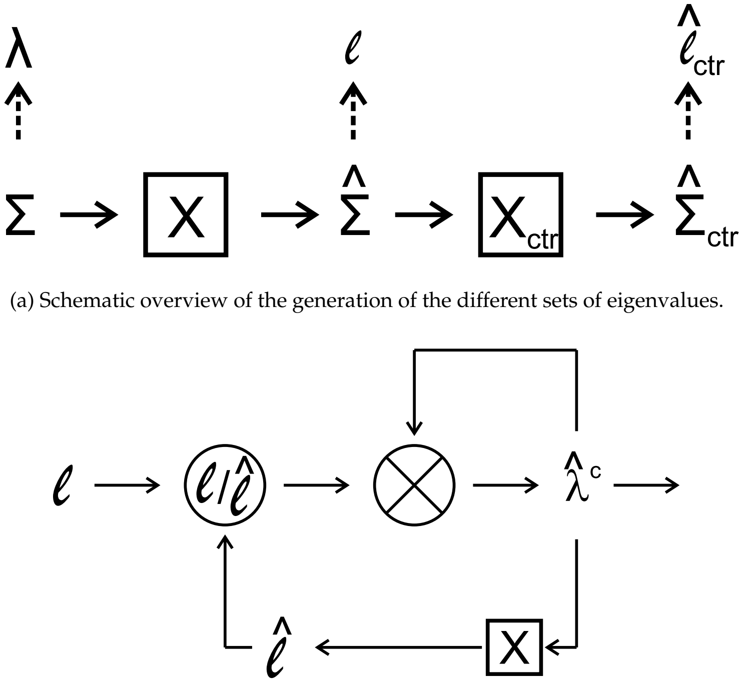
If the bias is significant, it should be corrected for which several correction methods exist. Due to the specific configuration of the experiment, we used a method very similar to the method presented in [48].

A schematic representation of the method is given in figure 4.2b. The sample eigenvalues $\boldsymbol{l}$ are used as an initial estimate of the population eigenvalues $\hat{\boldsymbol{\lambda}}^c$ and synthetic data are generated with these eigenvalues as population eigenvalues. From these data an estimate of the sample eigenvalues $\hat{\boldsymbol{l}}$ is determined which is then compared to $\boldsymbol{l}$ by determining the ratio between the two. Based on this ratio $\hat{\boldsymbol{\lambda}}^c$ is updated. If the ratio $\boldsymbol{l}/\hat{\boldsymbol{l}}$ is close enough to 1, the repetitions are stopped.

In order to prevent divisions by zero, a similar approach as proposed in [25] is followed: all the zero valued eigenvalues are set to the smallest non zero eigenvalue. We modified this approach to include a small set of the smallest non zero eigenvalues as well and set the values to the smallest eigenvalue not included in this set.

4.2.5 Experiments

We want to show that the position sources model can explain the characteristics observed if facial data is used in systems based on the fixed position intensity sources model. Therefore we perform two experiments with synthetic data, in which the



(a) Schematic overview of the generation of the different sets of eigenvalues.

(b) Schematic representation of the bias correction method

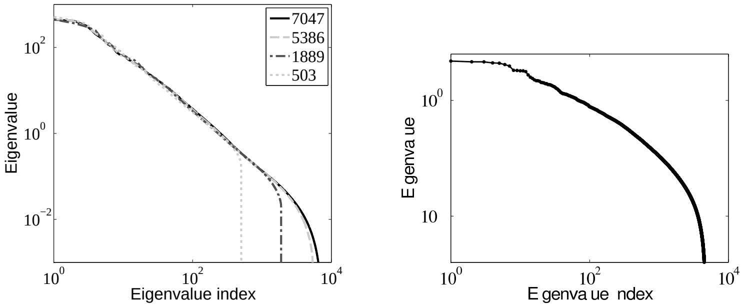
Figure 4.2: Eigenvalue estimation schematics

data is generated as described by the position sources model. In the first experiment we demonstrate characteristics 1 (the high number of estimated sources) and 2 (the eigenvalue curve can be described by a 1 over f function). The second experiment is a verification experiment and it demonstrates that characteristics 3 and 4 can also be explained by the modeling of data of position sources with fixed position intensity sources.

4.2.5.1 Experiment 1: Scree plot characteristics

For the first experiment we used facial images from the FRGC2 database [39]. We used images with frontal faces, no expressions, no glasses and a neutral expression. After preprocessing we estimated the covariance matrix of the set of samples. The scree plots showing the sample eigenvalues of this covariance matrix for different numbers of input classes are shown in figure 4.3a. As can be seen, adding samples to the training set increases the number of estimated sources. Moreover the largest part of the curves can be accurately modeled by a 1 over f model (the nearly straight line in the log-log plot), where f is the eigenvector number.

To demonstrate that position source can explain these characteristics, we generated synthetic images of 80 by 80 pixels. Each image contained 4 squares of



(a) Scree plots of the FGRC2 Facial Data for different numbers of samples.

(b) Scree plots of synthetic data

Figure 4.3: Scree plot comparison between facial data and synthetic data with position sources.

20 by 20 pixels. The 4 mean positions of these 4 squares were uniformly distributed along the diagonal from the top left corner to the lower right corner of the images. The position of each object deviated from its mean for the generated images with a variance of 6^2 in both horizontal and vertical directions.

We generated a training set of 5000 images and determined the covariance matrix of this training set. The resulting scree plot is shown in figure 4.3b. Although the plot is not a perfect match with the FRGC2 data scree plot, it shares several of its characteristics: it starts almost flat, but after the first few eigenvalues it follows a 1 over f curve rather accurately for a large part, although the fit of the FRGC 2 scree plot with the 1 over f curve is tighter.

So these experiments show that the position sources model is capable of explaining the first two characteristics of the results obtained if modeling facial image data with the fixed position intensity sources model. To demonstrate that the position sources model can also explain the third and fourth characteristics of face data in fixed position intensity sources modeled systems we did a verification experiment with synthetic data.

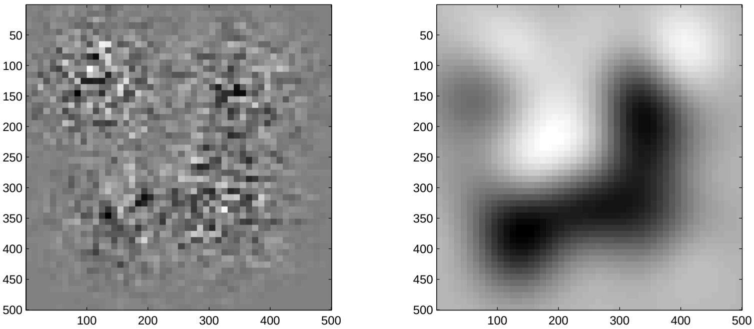
4.2.5.2 Experiment 2: Verification characteristics

In the experiment we use data containing position sources as input for a verification system which models the input data with the fixed position intensity sources model. The synthetic images of 50 by 50 pixels contained 4 objects, where each object was constructed by placing a Gaussian window over an image where the pixel intensities were determined by drawing from a Gaussian distribution with zero mean and a standard deviation of 1. In this way the objects have a texture with a spectrum covering all frequencies.

To position the objects in the images contained in the data sets used for training and testing, 8 sources were generated, which control the horizontal and the vertical position of the 4 objects. The average of the sources, $\boldsymbol{\mu}_t$, is set so that each object is placed in a quadrant starting in the top left corner and proceeding clockwise. The sources are generated according to the LDA model. For the training set we generated $C = 400$ classes. The mean of each class, $\boldsymbol{\mu}_c, c \in [1, 400]$, was determined by drawing samples from a normal distribution $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_b)$. The between covariance matrix $\boldsymbol{\Sigma}_b$ has a random orthogonal matrix as eigenvectors and eigenvalues distributed uniformly between 0.1 and 3.075.

For each class in the training set we generated 5 samples per class by drawing samples from $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_w)$. The within class covariance matrix $\boldsymbol{\Sigma}_w$ has the same eigenvectors as $\boldsymbol{\Sigma}_b$ and eigenvalues distributed uniformly between 1.9250 and 4.9, such that the total covariance matrix, $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_b$ equals $5 \cdot \boldsymbol{I}_8$, where $\boldsymbol{I}_8$ is a 8 x 8 identity matrix. For the test set we generated 100 classes, with 10 samples per class for enrollment and 5 samples per class as probes.

One sample of the training set is given in figure 4.4a, where the 4 objects can be distinguished. Note that images are generated by simply adding the contributions of the objects, so no occlusion due to object overlap occurs.

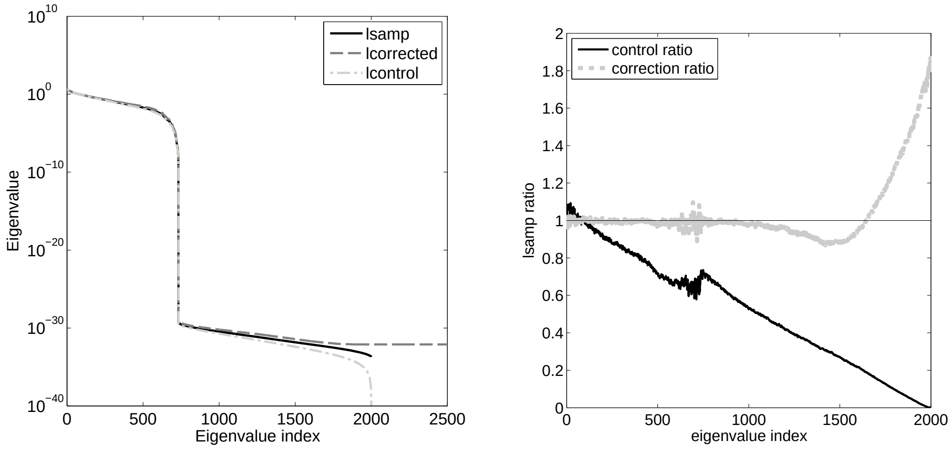


(a) Sample with features containing all frequencies

(b) Sample with features after low pass filtering

Figure 4.4: Synthetic image examples. Each sample contains 4 objects with positions varying between samples.

Figure 4.5a shows the scree plot of the eigenvalues of the estimated within class covariance matrix (the solid line, the other lines will be described in section 4.2.5.2). It shows two clusters of eigenvalues: one cluster significantly larger than zero (all eigenvalues above 10^{-4} and a second cluster of eigenvalues below 10^{-30}. A large section of the first part can be described by an exponential decaying curve. Note that we have 2000 samples for training and from this set we also estimated 400 class means, so we could only have 1600 non zero sample eigenvalues. The scree plot contains 2000 non zero eigenvalues, but the second cluster has very small valued eigenvalues so they seem to be the result of numerical problems. Still, the first cluster consists of over 700 eigenvalues, suggesting an equal amount of signal sources, while only 8 sources were used to generate the data.



(a) Eigenvalue scree plot

(b) Eigenvalue ratios

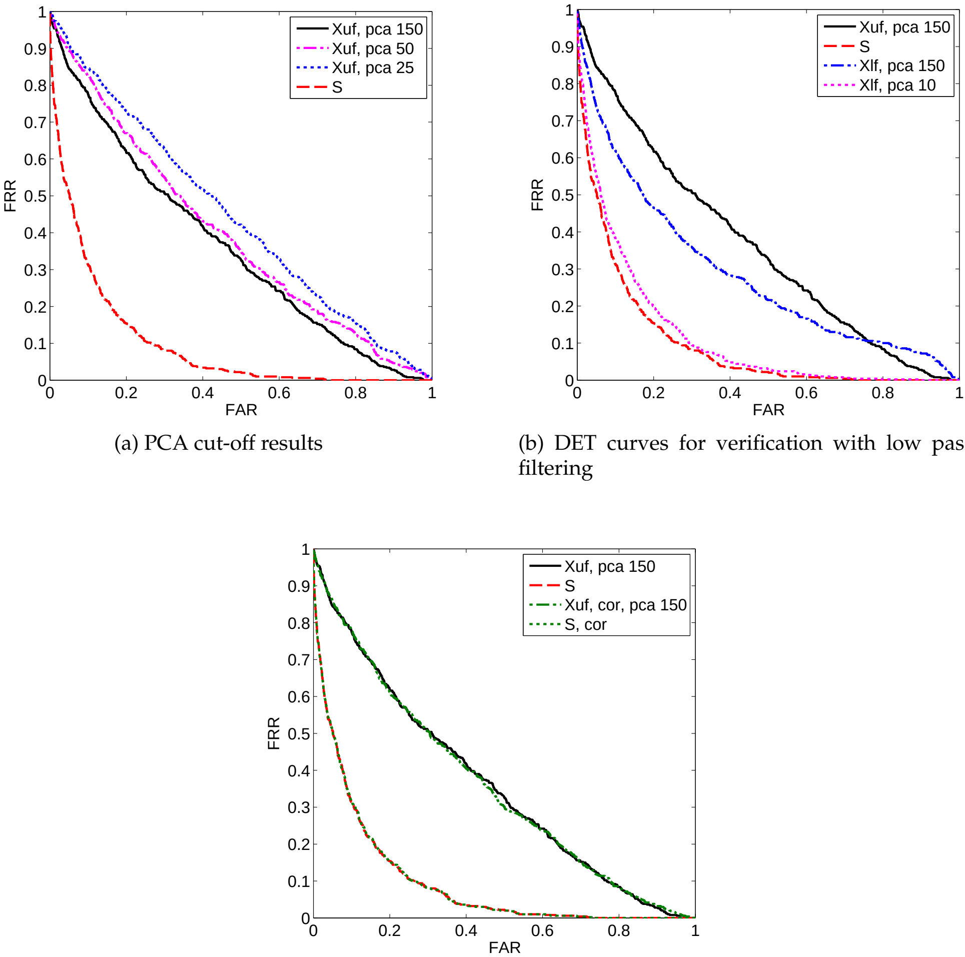
Figure 4.5: Within sample eigenvalue scree plots.

After training the system with the training set, we tested the performance with the test set. First we determined the class means of the gallery set, using the mean of the 10 samples per class. Next every sample in the probe set is compared with all the gallery classes and the log likelihood ratio of the probe belonging to the gallery class is determined. By varying the threshold on the log likelihood ratio above which a claim is accepted, we can control how many probe samples are falsely rejected from their class (FRR) versus how many probe samples are falsely accepted at some other class than their own (FAR).

Figure 4.6a shows the DET curves for several system configurations. As preprocessing step we applied a PCA dimensionality reduction. The best result was obtained if 150 components were retained (the darkest solid line). Increasing the number of components first has little effect on the curve, until after retaining 200 components. Retaining more components starts to increase the error rates, until at 500 components, the results are almost equal to random guessing.

Note that we used only 8 sources to generate the data. However, if we decrease the number of components below 150, the error rates go up again as shown by the second and third curves (the dash-dotted and dotted curves 'Xuf, pca 50' and 'Xuf, pca 25'), where we retained 50 and 25 components respectively.

The lowest curve (dashed curve in the lower right corner) shows the DET curve if we use the original 8 signals as input. Clearly there is a large difference in performance of the system if the images are used and if the original signals are used, which is solely caused by the error in the assumed model.



(a) PCA cut-off results

(b) DET curves for verification with low pass filtering

(c) DET curves for verification with bias correction

Figure 4.6: Detection Error Rates of synthetic verification experiments. X means image data, uf means unfiltered, lf means low pass filtered, pca # means dimensionality reduction with PCA to # components, S is the real position sources signal, and cor means bias correction applied.

To demonstrate characteristics 3 and 4, we made two additions two this experiment, as described in the next two sections.

low pass filtering In the analysis at the end of section 4.2.3.3 we hypothesized that large objects with low frequency textures and low variance in position may still be accurately modeled with the fixed position intensity sources model. The objects in the previous experiment had high frequency content in their textures. To test the low frequency hypothesis, we repeat the experiment, but we apply a low pass filter to the images before using them as input.

The filter was of the same size as the images (50 by 50). The sigma of the filter was 5. In figure 4.4b shows the same sample as figure 4.4a but after low pass filtering.

After this we applied again PCA dimensionality reduction and performed the verification experiment as described before. This results in a detection error curve (dash-dotted curve in figure 4.6b, 'Xlf, pca 150') which already outperforms the unfiltered curve, except for the part where FAR > 0.8. However, if we reduce the number of components to 10, we get a curve very close to the curve resulting from using the original 8 position signals (dotted curve, 'Xlf, pca 10').

Low pass filtering seems therefore an appropriate method for making the data fit the fixed position intensity sources model. However, the filtering operation removes all the additional information available if using high resolution images instead of the low resolution images. This seems therefore to explain the behaviour described in characteristic 3, which describes that there is a minimum in error rates if the resolution of face data is increased, after which the error rates even go up slightly if even higher resolution images are used.

bias correction Characteristic 4 describes that methods like bias correction, although theoretically leading to better estimates of the sample distributions, effect little result in practical verification experiments. We also applied bias correction in our verification experiment.

Figure 4.5a shows besides the within class sample eigenvalues also the results of the test for the severity of bias in the sample eigenvalues as described in section 4.2.4.2. The dot-dashed line shows the control eigenvalues of the synthetic 4 objects data. The differences between the eigenvalues are difficult to see, since the variations in eigenvalues required plotting the curves with a logarithmic scale on the vertical axis. Therefore we show in figure 4.5b the ratio between the control eigenvalues and the sample eigenvalues (the thin solid line). The control eigenvalues do have some bias.

Using the correction method described in section 4.2.4.2, we corrected the eigenvalues so that the ratio between the measured sample eigenvalues and the sample eigenvalues belonging to the corrected eigenvalue set is almost one for most of the larger eigenvalues as shown by the thick line.

To test how severe the effect of the bias in the eigenvalues is, we performed eigenvalue correction right after the estimation of the within class covariance matrix and the total covariance matrix. First we did the correction using the unfiltered data and after the correction we applied a dimensionality reduction to 150 components. The resulting curve (dash-dotted curve almost on top of the unfiltered 150 PCA solid line in figure 4.6c) is very close to the curve without applying the correction, so the bias seems to have only a minor effect on the error rates in the verification experiment, compared to the error introduced by the mismodeling.

After that we also applied the bias correction to the original 8 dimensional position signal. The resulting curve (the dotted curve also in the lower right corner of figure 4.6c) is almost the same as using the not corrected signal. This is however not surprising, since we used close to 2000 samples for estimating the covariance matrix of this 8 dimensional signal.

4.2.6 Conclusion

In face recognition the well known methods PCA and LDA model the data implicitly with a fixed position intensity model. We introduced the position sources model and showed that the fixed position intensity model fails if the data contains position sources. Using the position sources model we were able to recreate the characteristics observed if facial data is used in a system which models the data with the fixed position intensity sources model:

1. The estimation results in a large number of estimated intensity sources, while the synthetic data are generated from only a few sources.
2. The curve of the estimated eigenvalues fit a 1 over f curve for a very large part.
3. Increased resolution does not result in better verification performance but actually reduces it.
4. Bias correction has very little effect, in spite of its theoretical higher performance than PCA dimensionality reduction.

These observations suggest that position sources set a bound on the resolution of face images that can be accurately modeled by the fixed position intensity sources model. Therefore, in order to make use of the additional information high resolution images provide, first a more accurate model should be designed for facial images, which takes position changes of features into account. One solution is to detect these feature and remove them from the data before processing this data with a system based on the fixed position intensity source model.

Low pass filtering of the input data thus provides a solution to the position sources problem. However, it also removes any high resolution information images.

4.3 PCA and frequency decomposition

4.3.1 Introduction

In our analysis of SOS estimation, in particular solutions for the singularity problem, we determined that the classical sample estimate is unusable and has to be fixed one way or another. A classical solution for this problem is the application of PCA dimensionality reduction. As we showed in the previous chapters, PCA is a far from optimal solution: even a simple euclidean distance measure outperformed PCA dimensionality reduction for moderate dimensionality of the training samples.

Based on these results it was expected that if real data is used, applying correction methods would lead to large improvements in biometric verification systems. However, quite the opposite seems to be the case: PCA dimensionality reduction outperforms most methods with ease, and is certainly not overwhelmingly outperformed by any of the methods. So how is this possible?

As noted before, the analysis showing the weakness of PCA in high dimensional problems is performed on synthetic data, which necessarily adheres to the assumed data model, since the data is generated using that model. However, if the sample eigenvalues of face data is studied, a very typical character can be observed: the eigenvalues are accurately modelled by a 1 over f curve.

In the previous sections we showed that the position sources model is capable of explaining the observed characteristics of the sample eigenvalues estimated from facial data. It however did not explain the success of PCA with facial data, despite of its serious flaws in high dimensional estimation problems. In the coming section we will take a more theoretical approach to the analysis of the effect of position sources in intensity modelled data.

In particular, we will derive an equation describing the elements of a covariance matrix from a set of images, where the images are constructed from moving a single object around. The object itself is a Gaussian blob: its intensity depends on a Gaussian exponential formula.

After we find this expression for the covariance matrix, we use numerical methods to calculate the eigenvalue decomposition of this matrix. As it turns out, the results of this decomposition have to be considered in polar coordinates. It then turns out that the eigenvalue decomposition is in fact a form of frequency decomposition, with the largest eigenvalues corresponding to an eigenvector which resembles a low pass frequency filter.

This then does provide a possible explanation of the success of PCA dimensionality reduction. By retaining only the larger components, high frequency components are filtered from the data and as we already showed in the previous sections, low pass filtering of data containing position sources makes the data more fitted to an intensity sources model, therefore leading to more accurate estimates of the data generating signals.

4.3.2 Derivation of the covariance matrix elements with arbitrary matrices

Suppose we have a Gaussian feature, with its position determined by a 2D Gaussian variable. With this model the intensity at position a: $\boldsymbol{p}_a$ will be:

$$i_a(\boldsymbol{x}) = \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{p}_a)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}(\boldsymbol{x}-\boldsymbol{p}_a)\right] \tag{4.5}$$

where $\boldsymbol{x}$ is a random variable controlling the feature position, with a distribution given by:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{4.6}$$

To determine the covariance matrix of images created according to this model, we need to determine the cross covariance of the intensities of position a with position b. This is given by:

$$\begin{aligned}
\operatorname{cov}(i_a, i_b) &= \mathcal{E}\{(i_a - \mathcal{E}\{i_a\})\cdot(i_b - \mathcal{E}\{i_b\})\} && (4.7)\\
&= \mathcal{E}\{i_a i_b - i_a\mathcal{E}\{i_b\} - i_b\mathcal{E}\{i_a\} + \mathcal{E}\{i_a\}\mathcal{E}\{i_b\}\} && (4.8)\\
&= \mathcal{E}\{i_a i_b\} + \mathcal{E}\{i_a\}\mathcal{E}\{i_b\} && (4.9)
\end{aligned}$$

This requires the calculation of several expectations. We start with the simple expectation of the intensity at position a.

$$\begin{aligned}
\mathcal{E}\{i_a\} &= \int i_a(\boldsymbol{x})\; p(\boldsymbol{x})\; \mathrm{d}\boldsymbol{x} && (4.10)\\
&= \int \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{p}_a)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}(\boldsymbol{x}-\boldsymbol{p}_a)\right]\cdot && (4.11)\\
&\quad \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\mathrm{d}\boldsymbol{x} && (4.12)\\
&= \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left[-\frac{1}{2}\cdot\right.\\
&\quad \left.\Big((\boldsymbol{x}-\boldsymbol{p}_a)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}(\boldsymbol{x}-\boldsymbol{p}_a) + (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big)\right]\mathrm{d}\boldsymbol{x} && (4.13)
\end{aligned}$$

The exponential in the previous equation can be rewritten to a standard quadratic form using the following hypothesis:

Hypothesis 4.3.1 *The product of Gaussian exponentials,* A *and* B *result in another Gaussian exponential with a scaling factor, so:*

$$\begin{aligned}
\exp\Big((\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{A}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{A}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{A}})\Big)\cdot\exp\Big((\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{A}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{A}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{A}})\Big)&\\
= C\cdot\exp\Big((\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{R}})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{R}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{R}})\Big)& && (4.14)
\end{aligned}$$

Moreover, the parameters of the replacement are given by:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathrm{R}}^{-1} &= \boldsymbol{\Sigma}_{\mathrm{A}}^{-1} + \boldsymbol{\Sigma}_{\mathrm{B}}^{-1} && (4.15)\\
\boldsymbol{\mu}_{\mathrm{R}} &= \boldsymbol{\Sigma}_{\mathrm{R}}\cdot\Big(\boldsymbol{\Sigma}_{\mathrm{A}}^{-1}\boldsymbol{\mu}_{\mathrm{A}} + \boldsymbol{\Sigma}_{\mathrm{B}}^{-1}\boldsymbol{\mu}_{\mathrm{B}}\Big) && (4.16)\\
&= \Big(\mathbf{I} + \boldsymbol{\Sigma}_{\mathrm{A}}\boldsymbol{\Sigma}_{\mathrm{B}}^{-1}\Big)^{-1}\boldsymbol{\mu}_{\mathrm{A}} + \Big(\mathbf{I} + \boldsymbol{\Sigma}_{\mathrm{B}}\boldsymbol{\Sigma}_{\mathrm{A}}^{-1}\Big)^{-1}\boldsymbol{\mu}_{\mathrm{B}} && (4.17)\\
\tilde{C} &= \boldsymbol{\mu}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{A}}^{-1}\boldsymbol{\mu}_{\mathrm{A}} + \boldsymbol{\mu}_{\mathrm{B}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{B}}^{-1}\boldsymbol{\mu}_{\mathrm{B}} - \boldsymbol{\mu}_{\mathrm{R}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{R}}^{-1}\boldsymbol{\mu}_{\mathrm{R}} && (4.18)
\end{aligned}$$

where $C = \exp\left(-\frac{1}{2}\tilde{C}\right)$.

The hypothesis is proven if we can show that

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{A})^\mathrm{T}\,\boldsymbol{\Sigma}_\mathrm{A}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{A})+(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{B})^\mathrm{T}\,\boldsymbol{\Sigma}_\mathrm{B}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{B})&\\ =(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{R})^\mathrm{T}\,\boldsymbol{\Sigma}_\mathrm{R}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{R})+\tilde{C}&\end{aligned} \tag{4.19}$$

Starting from the left hand side of the equation

$$\begin{aligned}&(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{A})^\mathrm{T}\,\boldsymbol{\Sigma}_\mathrm{A}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{A})+(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{B})^\mathrm{T}\,\boldsymbol{\Sigma}_\mathrm{B}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{B})\\
=\;& \boldsymbol{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{x}-\boldsymbol{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{\mu}_\mathrm{A}+\boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{\mu}_\mathrm{A}+\\
& \boldsymbol{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{x}-\boldsymbol{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{\mu}_\mathrm{B}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{\mu}_\mathrm{B} && (4.20)\\
=\;& \boldsymbol{x}^\mathrm{T}\left(\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)\boldsymbol{x}-\left(\boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)\boldsymbol{x}-\\
& \boldsymbol{x}\left(\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)+\boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{\mu}_\mathrm{A}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{\mu}_\mathrm{B} && (4.21)\\
=\;& \boldsymbol{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{R}^{-1}\boldsymbol{x}-\boldsymbol{\mu}_\mathrm{R}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{R}^{-1}\boldsymbol{x}-\boldsymbol{x}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{R}^{-1}\boldsymbol{\mu}_\mathrm{R}+\boldsymbol{\mu}_\mathrm{R}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{R}^{-1}\boldsymbol{\mu}_\mathrm{R}+\tilde{C} && (4.22)\end{aligned}$$

, if

$$\boldsymbol{\Sigma}_\mathrm{R}^{-1} = \boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\Sigma}_\mathrm{B}^{-1} \tag{4.23}$$

and

$$\begin{aligned}\boldsymbol{\mu}_\mathrm{R}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{R}^{-1} &= \boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1} && (4.24)\\
\boldsymbol{\mu}_\mathrm{R}^\mathrm{T} &= \left(\boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)\cdot\boldsymbol{\Sigma}_\mathrm{R} && (4.25)\\
&= \left(\boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)\left(\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)^{-1} && (4.26)\\
&= \boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\left(\left(\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)\cdot\boldsymbol{\Sigma}_\mathrm{A}\right)^{-1}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\left(\left(\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)\cdot\boldsymbol{\Sigma}_\mathrm{B}\right)^{-1} && (4.27)\\
&= \boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\left(\mathbf{I}+\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{\Sigma}_\mathrm{A}\right)^{-1}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\left(\mathbf{I}+\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{\Sigma}_\mathrm{B}\right)^{-1} && (4.28)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\mu}_\mathrm{R}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{R}^{-1}\boldsymbol{\mu}_\mathrm{R}+C &= \boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{\mu}_\mathrm{A}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{\mu}_\mathrm{B} && (4.29)\\
C &= \boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{\mu}_\mathrm{A}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{\mu}_\mathrm{B}-\boldsymbol{\mu}_\mathrm{R}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{R}^{-1}\boldsymbol{\mu}_\mathrm{R} && (4.30)\\
&= \boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}\boldsymbol{\mu}_\mathrm{A}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\boldsymbol{\mu}_\mathrm{B}-\\
&\quad\left(\boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)\boldsymbol{\Sigma}_\mathrm{R}\left(\boldsymbol{\mu}_\mathrm{A}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{A}^{-1}+\boldsymbol{\mu}_\mathrm{B}^\mathrm{T}\boldsymbol{\Sigma}_\mathrm{B}^{-1}\right)^\mathrm{T} && (4.31)\end{aligned}$$

Using hypothesis 4.3.1 we can rewrite equation 4.13 and solve it:

$$
\begin{aligned}
\mathcal{E}\left\{i_a\right\} &= \frac{\mathrm{C_a}}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \\
&\quad \int \exp\left[-\frac{1}{2}\cdot\Big(\left(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{R,a}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{R,a}}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{R,a}}\right)\Big)\right]\mathrm{d}\boldsymbol{x} && (4.32)\\
&= \frac{\left|\boldsymbol{\Sigma}_{\mathrm{R,a}}\right|^{1/2}}{\left|\boldsymbol{\Sigma}\right|^{1/2}}\mathrm{C_a}\int\frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_{\mathrm{R,a}}\right|^{1/2}} \\
&\quad \exp\left[-\frac{1}{2}\cdot\Big(\left(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{R,a}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{R,a}}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{R,a}}\right)\Big)\right]\mathrm{d}\boldsymbol{x} && (4.33)\\
&= \frac{\mathrm{C_a}}{\left|\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}+\mathbf{I}\right|^{1/2}} && (4.34)
\end{aligned}
$$

with

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\mathrm{R,a}}^{-1} &= \boldsymbol{\Sigma}_{\mathrm{F}}^{-1}+\boldsymbol{\Sigma}^{-1} && (4.35)\\
\boldsymbol{\mu}_{\mathrm{R,a}} &= \left(\mathbf{I}+\boldsymbol{\Sigma}_{\mathrm{F}}\boldsymbol{\Sigma}^{-1}\right)^{-1}\boldsymbol{p}_a+\left(\mathbf{I}+\boldsymbol{\Sigma}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\right)^{-1}\boldsymbol{\mu} && (4.36)\\
\tilde{\mathrm{C}}_{\mathrm{a}} &= \boldsymbol{p}_a^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{p}_a+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{R,a}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{R,a}}^{-1}\boldsymbol{\mu}_{\mathrm{R,a}} && (4.37)
\end{aligned}
$$

The expectation of the intensity at position b adheres to the same relation, but with the parameters of position b substituted.

The last unknown term in the covariance calculation of equation 4.9 is $\mathcal{E}\left\{i_a i_b\right\}$:

$$
\begin{aligned}
\mathcal{E}\left\{i_a i_b\right\} &= \int i_a\left(\boldsymbol{x}\right)\cdot i_b\left(\boldsymbol{x}\right)\cdot p\left(\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x} && (4.38)\\
&= \int\exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{p}_a\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\left(\boldsymbol{x}-\boldsymbol{p}_a\right)\right]\cdot && (4.39)\\
&\quad \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{p}_b\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\left(\boldsymbol{x}-\boldsymbol{p}_b\right)\right]\cdot && (4.40)\\
&\quad \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right]\mathrm{d}\boldsymbol{x} && (4.41)
\end{aligned}
$$

Using hypothesis 4.3.1 we take the first 2 Gaussian exponentials together and get

$$
\begin{aligned}
\mathcal{E}\left\{i_a i_b\right\} &= \exp\left[-\frac{1}{4}\left(\boldsymbol{p}_a-\boldsymbol{p}_b\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\left(\boldsymbol{p}_a-\boldsymbol{p}_b\right)\right]\cdot && (4.42)\\
&\quad \int\exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\bar{\boldsymbol{p}}\right)^{\mathrm{T}}\left(\frac{1}{2}\boldsymbol{\Sigma}_{\mathrm{F}}\right)^{-1}\left(\boldsymbol{x}-\bar{\boldsymbol{p}}\right)\right]\cdot && (4.43)\\
&\quad \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right]\mathrm{d}\boldsymbol{x} && (4.44)
\end{aligned}
$$

where $\bar{\boldsymbol{p}} = \frac{1}{2}\left(\boldsymbol{p}_a + \boldsymbol{p}_b\right)$. We apply hypothesis 4.3.1 again to reduce the number of Gaussian exponentials in the equation to 1:

$$\begin{aligned}
\mathcal{E}\left\{i_a i_b\right\} &= \exp\left[-\frac{1}{2}\tilde{C}_{\mathrm{ab}}\right] \frac{\left|\boldsymbol{\Sigma}_{\mathrm{R,ab}}\right|^{1/2}}{\left|\boldsymbol{\Sigma}\right|^{1/2}} \int \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_{\mathrm{R,ab}}\right|^{1/2}} \\
&\quad \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{R,ab}}\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{R,ab}}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{R,ab}}\right)\right] \mathrm{d}\boldsymbol{x} &(4.45)\\
&= \exp\left[-\frac{1}{2}\tilde{C}_{\mathrm{ab}}\right] \frac{\left|\boldsymbol{\Sigma}_{\mathrm{R,ab}}\right|^{1/2}}{\left|\boldsymbol{\Sigma}\right|^{1/2}} &(4.46)\\
&= \exp\left[-\frac{1}{2}\tilde{C}_{\mathrm{ab}}\right] \frac{1}{\left|2\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}+\mathbf{I}\right|^{1/2}} &(4.47)
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathrm{R,ab}}^{-1} &= 2\boldsymbol{\Sigma}_{\mathrm{F}}^{-1} + \boldsymbol{\Sigma}^{-1} &(4.48)\\
\boldsymbol{\mu}_{\mathrm{R,ab}} &= \left(\mathbf{I}+\frac{1}{2}\boldsymbol{\Sigma}_{\mathrm{F}}\boldsymbol{\Sigma}^{-1}\right)^{-1}\bar{\boldsymbol{p}} + \left(\mathbf{I}+2\boldsymbol{\Sigma}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\right)^{-1}\boldsymbol{\mu} &(4.49)\\
\tilde{C}_{\mathrm{ab}} &= \boldsymbol{p}_a^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{p}_a + \boldsymbol{p}_b^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{p}_b + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{R,ab}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{R,ab}}^{-1}\boldsymbol{\mu}_{\mathrm{R,ab}} &(4.50)
\end{aligned}$$

These results can now be substituted in equation 4.9:

$$\operatorname{cov}\left(i_a, i_b\right) = \frac{C_{\mathrm{ab}}}{\left|2\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}+\mathbf{I}\right|^{1/2}} + \frac{C_{\mathrm{a}}\cdot C_{\mathrm{b}}}{\left|\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}+\mathbf{I}\right|^{1/2}} \tag{4.51}$$

4.3.3 Determining the eigenvectors

In the previous section we derived an expression for the elements of the covariance matrix. In this section we try to find its decomposition. The expression in equation 4.51 describes a continues case, so a first step is to find make a discretisation by choosing an image resolution.

In the remainder of the section we assume that the Gaussian feature is uniform in all directions, so $\boldsymbol{\Sigma}_{\mathrm{F}} = \sigma_{\mathrm{F}}^2\mathrm{I}$. Furthermore we place the mean position of the feature in the center of the images. In our early analysis we assumed that the images are square, but it turned out that this caused artifacts in the decomposition, so we then assumed the use of only a circular area in the images and left the other pixels blank.

We used matlab to do the decomposition of the covariance matrix. In figure 4.7 the eigenvectors are shown. Apparently there is some randomness in the decomposition: eigenvectors # 2 and # 3 (figures 4.7b and 4.7c respectively) are identical up to a rotation of 90 degrees. However, the rotation of both images, represent another valid set, so this is some arbitrary factor. This set of two eigenvectors which have an identical image up to some rotation with a random initial rotation is very frequent. The independence of the eigenvectors on the initial rotation could have been predicted, since equation 4.51 is in fact rotation independent for the chosen parameters.

Because of the apparent circular properties of the images of the eigenvectors, the images are best studied in polar coordinates. We can then describe the eigenvectors by their behaviour along the radial axis and their behaviour along the rotation axis. The eigenvectors can be divided into two classes: a set of eigenvectors which are uniform on the rotation axis, for example eigenvectors # 1 and # 28 as shown in Figures 4.7a and 4.7e, and eigenvectors which have a sinusoidal behaviour along the rotation axis, for example eigenvectors # 2, # 3, # 4 and # 29 shown in figures 4.7b, 4.7c, 4.7d and 4.7f respectively.

The two classes of eigenvectors are ordered in a specific manner. The first eigenvector is a rotation uniform eigenvector, which removes the average on each circle around the centre. Then sets of two rotation sinusoidal eigenvectors occur with the first set having the base frequency, the second set having 2 times the base frequency and so on until another rotation uniform eigenvector occurs again, which has a higher frequency content on the radial axis compared to the previous rotation uniform eigenvector. After each rotation uniform eigenvectors, the rotation sinusoidal eigenvectors have are more focussed on the outside of circle in the image, so they have a radial component which is almost zero for small radius values. We did not find a clear relation for when exactly a switch from the rotation sinusoidal eigenvectors to a rotation uniform eigenvector occurs.

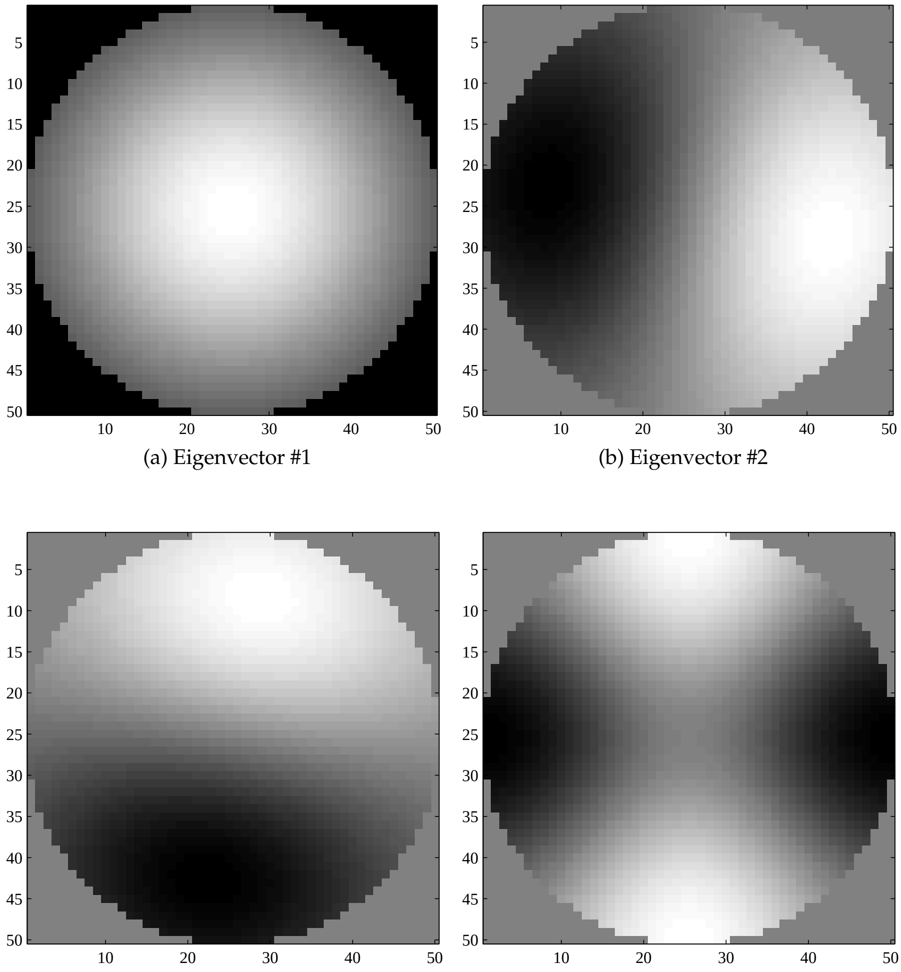
In figure 4.8 we show a representation of eigenvector 4 in polar coordinates. In figure 4.8a a 2 dimensional representation of the eigenvector in polar coordinates is given, which clearly show the periodicity. In figure 4.8b we show the polar representation for one fixed radius component and matched it with a sine function, which is apparently very accurate.

The radius plot for a fixed angle, given in figure 4.8c, has mainly a low frequency component, although we did not find a matching function. For higher components more local minima and maxima occur in the curve.

Figure 4.9c shows the scree plot of the eigenvalues. As determined in the previous sections, the feature, with its position determined by just two sources, causes the estimation of much more intensity sources. The first part of the scree plot seems to follow an exponential decay, although the curve flattens out for eigenvalues larger than 50. Note as well that the two neighbouring rotation sinusoidal eigenvectors have equal corresponding eigenvalues. The flattening of the curve might be explained by numerical problems: figures 4.9a and 4.9b show the eigenvectors # 63 and # 64 respectively, but where there is a clear structure in # 63, no structure can be observed in # 64.

4.3.4 Conclusion

Our study of the effect of a moving Gaussian feature in intensity modelled data showed that the eigenvectors of the covariance matrix form a set of frequency basis vectors, so PCA seems to perform some kind of frequency decomposition.



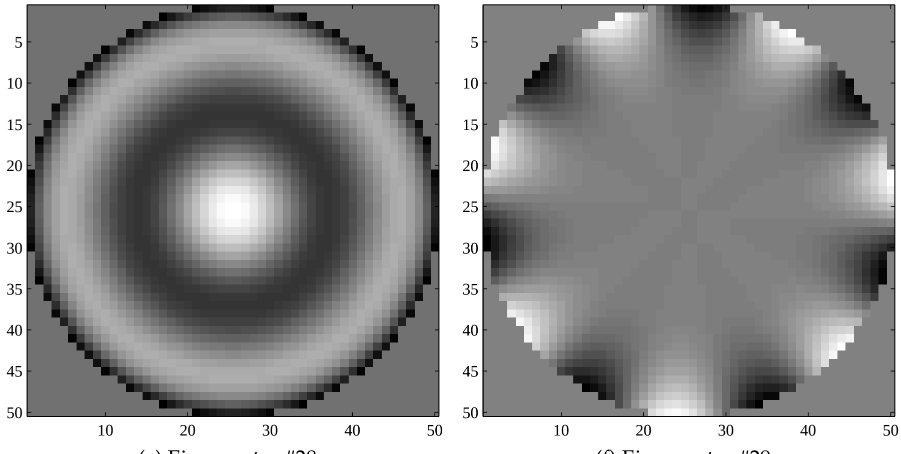
(a) Eigenvector #1

(b) Eigenvector #2

(c) Eigenvector #3

(d) Eigenvector #4

Figure 4.7: Eigenvectors of 2D Gaussian moving feature.



(e) Eigenvector #28

(f) Eigenvector #29

Figure 4.7: Eigenvectors of 2D Gaussian moving feature (cont.)

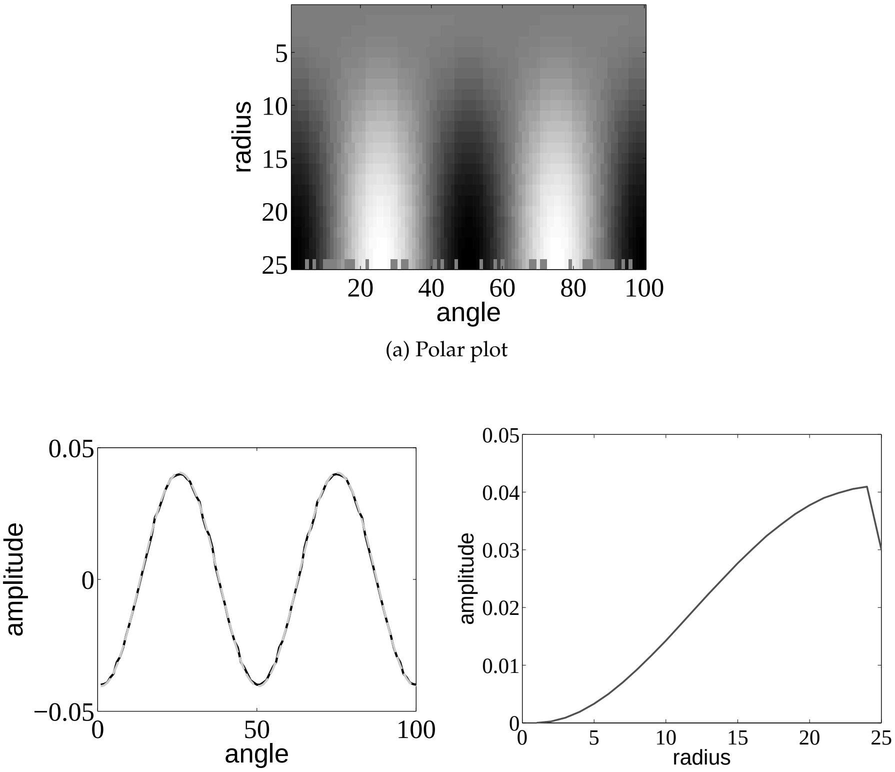
Projecting the data on only the lower PCA components therefore resembles low pass filtering, which is, as we showed before, one approach to allow the intensity sources model to estimate the original signals of data with position sources. This gives an explanation why PCA dimensionality reduction performs better in face data than bias correction as a solution to the singularity problem: although it is a poor solution to the singularity problem, it does also solve errors in the assumptions of how the data is generated.

4.4 Example of position source encoding in face recognition: Eye template

The previous sections demonstrated that the presence of position sources have a strong disturbing effect on the SOS estimation based on the intensity sources model. We made several attempts to deal with position sources in facial data. In the coming sections we report our thoughts on the subject and show some preliminary results. The basic methodology we chose is to try and remove the position information before estimating the SOS based on the position sources model.

4.4.1 Introduction

To demonstrate that position sources can be used successfully in facial data, we have chosen to limit ourselves to the eye region, but the results can easily be extended to images containing the entire face. In the coming sections we will model the iris using

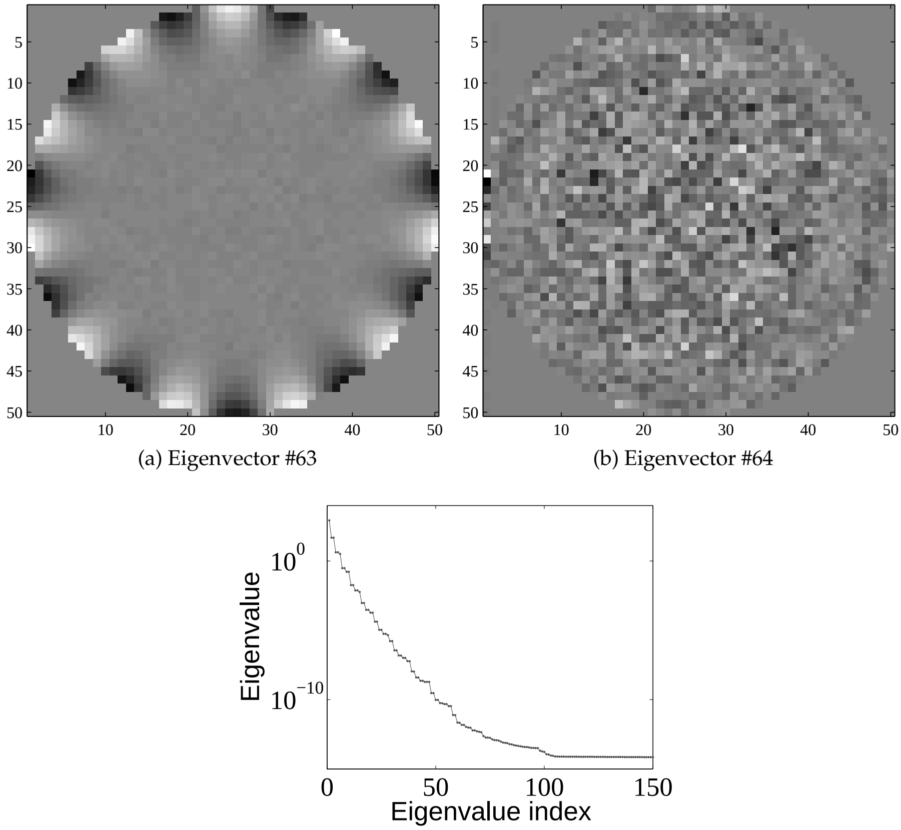


(a) Polar plot

(b) Angle plot

(c) Radius plot

Figure 4.8: Different representations of eigenvector #4



(a) Eigenvector #63

(b) Eigenvector #64

(c) Scree plot

Figure 4.9: The end of structure

one position source and show that this provides a more efficient coding scheme than the classical fixed position intensity source model.

There is a big advantage for choosing the eye as an area to limit ourselves to: there are already a large number of publications on modeling the eyes and extracting features from it. Examples are iris detection and recognition [73, 74, 75, 76], eye tracking applications [77] and even face recognition [78, 79, 80, 81, 82, 83, 84, 85, 86, 5, 87, 88].

Although more complex encoding models have already been developed, we used a very simple model to demonstrate the effect. The procedure we will follow is as follows: we start with frames of a video containing an eye. In each frame we first detect the corners of the eye. Based on these corners a Region of Interest (ROI) is extracted from the frame, after which two training schemes are used to find the structure in the data: one scheme based on the fixed position intensity sources model and the other based on a combination of the fixed position intensity sources model and the position sources model. We compared the training results from both schemes to determine that the second scheme is able to explain more variance with a position source than the strongest intensity source found in the first scheme.

As input data a video is used showing a part of the face containing one eye which tracks a finger from left to right and visa versa for a couple of times. Two examples are shown in figure 4.10a and 4.10b. The position of the eye corners and the iris were set manually.

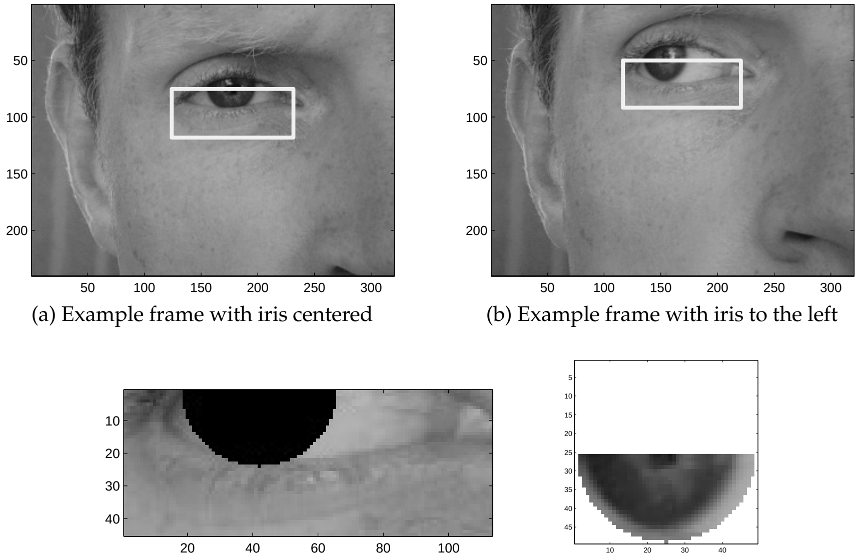
The first training scheme is based on classical PCA: the pixels in the ROI of each frame are concatenated into a column vector and PCA is applied to the set of these column vectors. PCA is designed to explain as much of the variance of the data with as few components as possible. We therefore determined how much variance the first PCA component explains (which is the value of the first eigenvalue).

The second training scheme starts by locating the position of the iris in each frame. The data in the frame is then split into two sets: an iris set and a non iris set. In the non iris set, the pixels in the iris area are blacked out and not taken into account any further. In later stages this means we have data with missing values. In the iris set, the pixels in the iris area are only considered.

In the second scheme we modeled the iris rather crudely with a fixed radius circle. Figures 4.10c and 4.10d demonstrate how in the second training scheme the sample shown in figure 4.10b is split into a sample for the non iris set and the iris set respectively.

Since the iris was tracking a finger from left to right, its movement should essentially be modeled with one position source. However there is also vertical movent, as shown in figure 4.11b by the dots which represent iris positions in the different frames. This can be explained by partly by the viewing angle of the camera. We therefore fitted a curve and used the corresponding function to connect the horizontal position of the iris with the vertical position of the iris, so the iris model is controlled by one position source.

The objective of PCA is to find a representation of the data explaining the most of the variance of the data with as few components as possible. We will show that modeling the iris with one position source explains more variance than the strongest



(a) Example frame with iris centered

(b) Example frame with iris to the left

(c) Eye with iris area left out

(d) Iris area

Figure 4.10: Examples of the moving iris modeling. The white rectangles in the top images indicate the ROI. The bottom two images show the processing results of the second training scheme in which the iris and the non iris parts are separated.

principle component. To find the variance explained by the position source we first split the original variance σ^2 in a iris term and a non iris term using a mask function $M(q,k)$ which equals 0 if pixel q in frame k is part of the iris and is otherwise 1:

$$\begin{aligned}
\sigma^2 &= \sum_{q=1}^{p} \frac{1}{N-1} \sum_{k=1}^{N} \left(X(q,k) - \mu(q)\right)^2 && (4.52)\\
&= \sum_{q=1}^{p} \frac{1}{N-1} \sum_{k=1}^{N} M(q,k) \cdot \left(X(q,k) - \mu(q)\right)^2 + \\
& \quad \sum_{q=1}^{p} \frac{1}{N-1} \sum_{k=1}^{N} \left(1 - M(q,k)\right) \cdot \left(X(q,k) - \mu(q)\right)^2 && (4.53)
\end{aligned}$$

where the first part represents the non iris part and the second part represents the iris part.

Encoding with the iris position source gives each set their own means, $\boldsymbol{\mu}'$ and $\boldsymbol{\mu}''$, and alters indexing of the pixels in the iris area. Therefore the new variance with position source encoding is given by:

$$\begin{aligned}
\sigma'^2 &= \sum_{q=1}^{p} \frac{1}{N-1} \sum_{k=1}^{N} M(q,k) \cdot \left(X(q,k) - \mu'(q)\right)^2 + \\
& \quad \sum_{q'=1}^{p'} \frac{1}{N-1} \sum_{k=1}^{N} \left(1 - M'(q',k)\right) \cdot \left(X'(q',k) - \mu''(q')\right)^2 && (4.54)
\end{aligned}$$

where p' is the number of pixels in the iris area, M' is the iris mask projected in the iris space, and X' are the iris samples extracted from the frame. The variance explained by the position source is then given by $\sigma^2 - \sigma'^2$.

4.4.2 Results

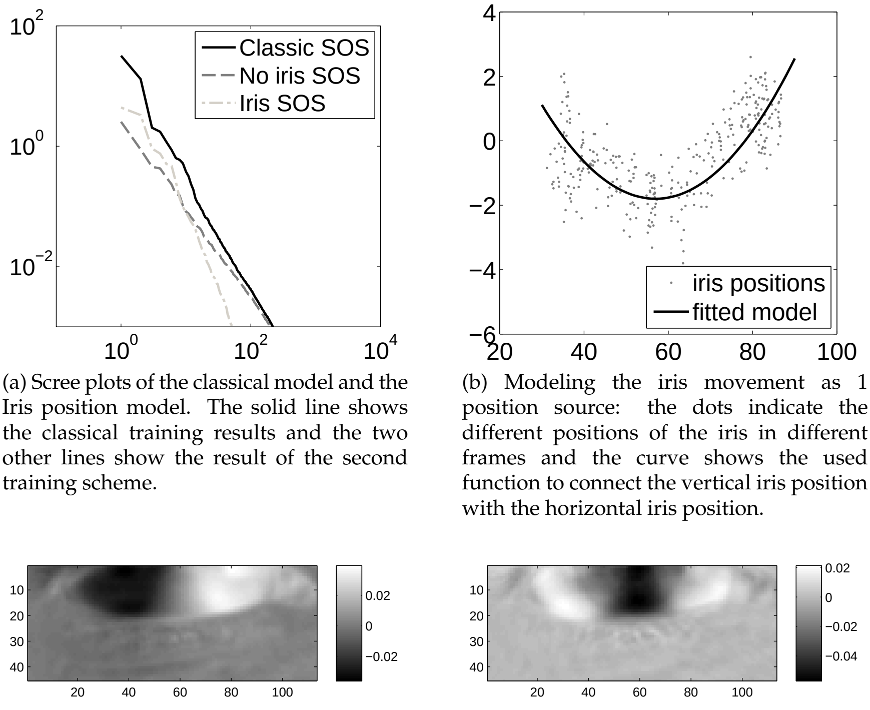
After training we compared the variance explained by the first component of the classical PCA approach with the total reduction in variance by incorporating position sources. Those results are presented in table 4.1, which clearly shows that the position sources are explaining more of the variance of the data, which supports the hypothesis that position sources are not efficiently encoded with intensity sources.

Table 4.1: Variance explained by the different training methods

	PCA component 1	Position source
explained variance	31.9	37.6

The scree plots of the eigenvalues determined during training are shown in figure 4.11b. The solid line represents the eigenvalues of the classical PCA training. It clearly shows a double log relationship, although it is more close to k^{-2}, instead of the k^{-1} behaviour previously observed in facial data.

The no iris set shows a similar behaviour, so this might indicate that the remaining data still contains position sources. In the iris set, the eigenvalues drop off towards zero more quickly so it seems that data can be represented by the fixed position intensity sources model quite efficiently.



(a) Scree plots of the classical model and the Iris position model. The solid line shows the classical training results and the two other lines show the result of the second training scheme.

(b) Modeling the iris movement as 1 position source: the dots indicate the different positions of the iris in different frames and the curve shows the used function to connect the vertical iris position with the horizontal iris position.

(c) Classical model eigenvector 1

(d) Classical model eigenvector 2

Figure 4.11: Training results with classical model and the position model

Inspection of the eigenvectors of the classical training approach further supports our analysis in section 4.3. Figures 4.11c and 4.11d show eigenvectors 1 and 2 respectively, which clearly show a frequency decomposition: eigenvector 1 filters the base frequency of the movement from left to right, while eigenvector 2 filters the first harmonic.

4.4.3 Conclusion and suggestions

The previous experiment showed on the least that the classical approach based on PCA is not the most efficient encoding scheme of moving eye data, which it should be if the fixed position intensity sources model was accurate. Further more, the results also match the results the analysis of section 4.3: the classical approach performs a frequency decomposition of the movement.

This all suggests that indeed the position sources should be taken care of in facial data. The method we used here is very crude and should be improved. For the extension the following points should be taken into account:

- We want to take the current model as starting point and gradually extent it instead of starting all over again.
- One extension could be to handle position sources and reduce if not completely remove their presence in the data before passing the data on to a system based on SOS estimation assuming intensity sources, similar to what we did in the previous experiment.
- In this experiment we choose the model for the iris on forehand and as a result only 2 position sources could be estimated (which were then brought back to only one position source). Therefore the model only works for modeling the iris instead of any arbitrary position source in face images. If we make the model more general and we assume that every pixel value could be the result of two position sources, similar to the intensity sources model which allows every pixel to be controlled by its own intensity source, the number of parameters increases by a factor of 3. Therefore, such a system is inherently underdetermined. Additional constraints should be posed on such a system if any structure is to be estimated at all.

A possible next step is to use a more complex eye model which is already available in literature. One example is the model present in [89]. However with that model the number of parameters has increased to around 20, depending on the exact implementation. It remains to be seen if all these apparent position sources can explain more variance then using the intensity sources model as was the case with the iris position source.

4.5 Conclusion

The second derived research question posed in section 1.8 is: "What effect does the presence of position sources in data have on systems based on the fixed position intensity sources model and can it explain the observations made after increasing the image resolution of facial data: the high number of sources estimated, the 1 over f characteristic of the eigenvalue scree plot, the saturation of performance of biometric systems based on SOS estimation and that PCA performs better on real facial data than bias correction?" We studied the effect of position sources with both synthetic data and real facial data and derived from these experiments the following answer to this question.

By assuming that face data is not accurately modelled by the fixed position intensity model, we can explain several characteristics observed in the SOS estimated from face data:

- The seemingly high number of sources involved in the face data generation process without a clear distinction between the relevance of the different sources.
- The 1 over f characteristic of the eigenvalue spectra.

With position sources we are also able to explain why bias reduction techniques sort little effect on biometric system performance and why it is outperformed by classic PCA dimensionality reduction: firstly, data generated with the position sources model can be made more fitting to the fixed position intensity sources model by applying a low pass filter to the data. Secondly, we showed that PCA analysis of a moving source results in a frequency decomposition of the data. Therefore, applying a PCA dimensionality reduction seems to resemble a low pass filtering operation. This provides an explanation of the success of PCA dimensionality reduction applied to facial data even though in theory, if the data adheres to the fixed position intensity sources model, the PCA dimensionality reduction is far from optimal and will even break down completely for very high dimensionality.

Chapter 5

Conclusion

In this thesis we sought to explain the counter intuitive observation that adding new information by increasing the number of variables observed does not always increase the accuracy of biometric systems and can even be harmful. We focussed on the application of face verification based on the log likelihood ratio using SOS estimation and formulated our main research question as follows:

- Why does providing additional information not always help PCA based methods such as the eigen face method to improve their performance or even damage it and how can we overcome this limitation?

We studied two major options which can provide an answer to this question: the bias in the estimation of the eigenvalues and a deviation of the model used in this estimation. We therefore formulated two derived research questions.

Chapters 2 and 3 dealt with the derived research question:

- What (potential) effects does the sample eigenvalue bias have on verification systems and can these effects be reduced?

In chapter 2 we showed that the SOS estimated from high dimensional data contain some serious flaws: the sample eigenvalues are biased estimates of the population eigenvalues and the sample eigenvectors do not align with the population eigenvectors.

The bias in the sample eigenvalues is described by the Marčenko Pastur equation, but this law only holds for the GSA limit, where both the number of samples N and their dimensionality p grow infinitely large. To apply this equation to practical problems, we introduced smooth eigenvalue estimation, where by setting a smoothness factor we could use the equation in practical problems with finite N and p values.

We then showed the bias in the sample eigenvalues can be reduced significantly by bias correction methods. Chapter 2 presents two of such methods we developed: the bootstrap correction and the fixed point correction of which the later relies heavily on the smooth eigenvalue estimation. The major advantage of this method over the existing method by Karoui is that the fixed point bias correction corrects the eigenvalues directly instead of estimating a distribution and the method works for underdetermined cases, where all zero valued eigenvalues are corrected to an equal non zero eigenvalue. This last part is needed in verification using the log likelihood ratio, since an inversion of the covariance matrix is required.

Although the bias in the sample eigenvalues can (theoretically) be removed completely, still errors remain: the variance estimates along the sample eigenvectors are still wrong, since the corrected eigenvalues give the variances along the population eigenvectors and these sets of eigenvectors are not aligned. It is not surprising that this misalignment error can not be fixed because otherwise we could obtain a perfect estimate from just a few samples, which seems very unlikely. However, the total amount of misalignment can be estimated and a second correction can be made to find the variances along the sample eigenvectors. In experiments with synthetic data this variance correction reduced the Kullback Leibler divergence between the population distribution and the corrected sample distribution significantly.

In chapter 3 we studied the effect of the bias in a verification setting. The system we studied was a verification system based on log likelihood estimation. The log likelihood estimation requires the inverse of the estimated covariance matrices, but because of the eigenvalue bias, these matrices become singular when p becomes larger than N. The classical solution to this problem is to perform PCA dimensionality reduction before estimating the likelihoods. We showed that this solution is far from optimal if the dimensionality gets even higher and is out performed by the simple euclidean distance for high dimensional problems.

We then studied the application of the improved SOS estimation methods in a verification setting. It turned out that applying the correction naively to the two involved distributions, the within class distribution and the between class distribution, leads to arbitrary between over within class variance ratios, which results in almost random verification behaviour. We introduced the eigenwise correction method to solve this problem. The method takes the correction done on the within class estimate into account in the correction of the between class estimate.

With synthetic data we showed that the eigenwise correction combined with the fixed point correction gives better performance on high dimensional verification systems than the classical PCA dimensionality reduction. The combination using the fixed point bias correction also outperformed the combination using the bias correction developed by Karoui. However, with real facial data, PCA dimensionality reduction outperforms eigenwise correction. This lead to the derived research question:

- What effect does the presence of position sources in data have on systems based on the fixed position intensity sources model and can it explain the observations made after increasing the image resolution of facial data: the high number of sources estimated, the 1 over f characteristic of the eigenvalue scree plot, the saturation of performance of biometric systems based on SOS estimation and that PCA performs better on real facial data than bias correction?

Chapter 4 presents arguments for the hypothesis that facial data is not accurately modelled by the intensity sources model implicitly assumed by SOS estimation:

- Some features in the face are better described by their position than their intensity. For example pupils show much movement while the intensity fluctuations are much smaller.
- Several characteristics of the estimated SOS of facial data are quite easily explained by the effect position sources have if modeled with the intensity sources model:
  - The high number of intensity sources required in the modeling of facial data. Only a few position sources can lead to the estimation of a high number of intensity sources.
  - The curve of the eigenvalues can be described for a large part with high accuracy by a 1 over f curve, where f is the eigenvalue index. Using only a few position sources, an eigenvalue curve already closely resembling such a curve can be generated.
  - The effect of using higher resolution images as input on verification performance saturates: after a certain point increasing the image resolution does not improve the biometric system performance and it may even deteriorate the performance.
  - Advanced SOS estimators designed to operate on high dimensional training sets with a relatively low number of samples, like bias correction methods, do not improve systems performance or even deteriorate it.

The property that the scree plot of sample eigenvalues estimated from facial data can be described by a 1 over f curve is problematic for bias correction: if it is an accurate estimate then even though the first few eigenvalues are significantly larger than the remainder of the eigenvalues individually, the total energy in the bulk will become infinite, so the largest eigenvalues will be affected strongly by the bias. In fact, in the GSA limit, the corresponding distribution limit of the population eigenvalues would be a step function with the step occurring at $\lambda = 0$, $H(\lambda) = \mathrm{u}(\lambda)$.

The last two characteristics are the result of the effect described in section 4.3. The saturation effect is explained by the fact that large features with rather small movement compared to their size can still be modeled accurately with the intensity model, while small features with large movements are poorly captured with the intensity sources model. Higher resolution images will contain more and more features of the second kind, so the intensity sources model becomes less fit to describe the data with increase in resolution up to a point that this effect dominates the effect of the increased amount information by the higher resolution.

The effect that PCA seemingly outperforms the advanced correction methods in the reduction of the bias can be explained as follows. PCA dimensionality besides removing the zero valued eigenvalues also performs a low pass filtering operation on the data, resulting in data which is more accurately modeled by the intensity model. This low pass filtering operation is the result of the following: if data containing position sources is modeled with the fixed position sources model and the SOS are estimated, then the estimation results can be described as a frequency decomposition where the basis spanned by the eigenvectors corresponding to the largest eigenvalues contain the low frequency content of the data and the smaller the eigenvalues get, the higher the frequency content of the data their corresponding eigenvectors represent. Since PCA dimensionality reduction removes the smallest eigenvalues, it retains only the low frequency content.

5.1 Future work

To improve verification systems and let them take advantage of the high frequency content available nowadays, the next step would be to develop a method capable of handling position sources as well as intensity sources, something similar to what PCA does for just the intensity sources. We reported some steps in that direction, however the combination of the intensity sources model with the position sources model is inherently underdetermined as explained in section 4.4.3. Therefore additional constraints should be introduced. We chose a fixed parameterized model. The choice of a parameterized eye model makes the modeling approach less general applicable though.

A more general approach could be to take a tree structured approach: first analyse the low frequency content of the images and detect if some variations are better described by movement or by intensity changes. Then analyse the slightly higher frequency content, taking the previously detected local position changes into account, and so on. This local frequency decomposition has many similarities with the gabor wavelet [90], in which a frequency decomposition is made local by applying a Gaussian kernel. This new approach differs by allowing position changes between the different frequency parts. However, a big question with this approach is again: at which point does the estimated structure become more dependent on random structures in the data rather than the process parameters of the data generating process?

5.2 Relations with other fields

> "Free will. It's like butterfly wings: Once touched, they never get off the ground." [91]

In section 1.7 we already noted that the subject studied in this thesis is closely related to topics in other fields. After the conclusions we drew in the previous sections, I want to explain how I think these findings may aid in the discussion in other fields as well. Since these topics are not the main focus of my studies, any of the following remarks are given without proof.

5.2.1 Free will

One of these topics I want to address is the discussion on whether ours is a deterministic world or a world governed by chance. Often the underlying topic in this discussion is that if the world is fully deterministic, then there can be no free will: in a deterministic world, if all the inputs of a person are known, his/her behaviour is completely predictable.

This issue has already been marked as unsolvable with the introduction of quantum mechanics and the Heisenberg uncertainty principle, which points out that it is impossible for us mortals to determine the complete state of the universe with absolute certainty [92]. However, at macroscopic level, the world might still be accurately described as deterministic so the discussion is still valid.

The limited number of samples issue may provide some valuable elements to this discussion. Assuming that human behaviour is mostly determined by the inter connections of the neurons in the brain, then the amount of parameters involved is related to the number of synapses in the brain. The number of neurons in the brain is estimated to be in the order of 10^{11}, which are interconnected by 10^{14} synapses [93]. Even though the time of observing someone's behaviour is difficult to translate into an exact number of training samples, it seems that the observation time is only sufficient to construct a very general model, even for an observer which is present all the time (the subject itself).

Therefore, even if the world is completely deterministic, it is not an accurate model to base behaviour towards other people on, since we cannot estimate its parameters accurately. In a sufficiently complex world, the free will model may therefore provide a more accurate model to base behaviour on compared to a predeterministic model.

5.2.2 Rational choice theory

A closely related subject is the rational choice theory, often used in economics. Although there are many different versions of the theory, the basic assumption is that people will always act to get the best products at the lowest price, or in other words, in ones own interests. This model is often perceived to be in conflict with altruistic behaviour, let alone predict it and it is argued that norm driven behaviour is counter efficient as is noted in [94]: "En daarmee stuit ik op de mythe dat efficiëntie en gelijkheid noodzakelijkerwijze trade-offs zijn." (With that I encounter the myth that efficiency and equality are necessarily trade-offs).

However, the model used in rational choice theory describes the decision process as being a maximisation process on the cost benefit analysis of the available options. If many parameters are involved in the decision process, then that optimisation can suffer from exactly the same problem as eigenvalue estimation. Therefore options may seem highly profitable at low cost, while in fact they are no better then any other option or even worse.

Under the assumption that the decision has to consider many variables (a sufficiently complex environment) membership of a group can be helpful in several ways: first of all the experience of other members effective increase the number of samples available in the estimation process, for example by sharing experience or by customs. Secondly, group membership may reduce the complexity of the problem, for example by providing safety and again customs. So the problem of cost estimation in a complex world might also be a part of bridging the gap between rational decision theory and social and altruistic behaviour observed in the wild.

References

[1] M. Grgic and K. Delac, "Research groups in face recognition." http://www.face-rec.org/research-groups/, 2011.

[2] K. Lorenz, *Agressie bij dier en mens*. Amsterdam, Nederland: Uitgeverij Ploegsma bv, 1984. original title: Das sogenannte Böse. Zur Naturgeschichte der Agression.

[3] C. A. Nelson, "The development and neural bases of face recognition," *Infant and Child Development*, vol. 10, pp. 3–18, March - June 2001.

[4] M. E. Clark, *In search of human nature*. Routledge, London; New York, 2002.

[5] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042–1052, 1993.

[6] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[7] N. Myhrvold, "Not so fast in dismissing moore's law." http://www.luminous-landscape.com/essays/not-so-fast.shtml, July 2009.

[8] B. J. Boom, G. M. Beumer, L. J. Spreeuwers, and R. N. J. Veldhuis, "The effect of image resolution on the performance of a face recognition system," in *Proceedings of the Ninth International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore, Malaysia*, pp. 409–414, December 2006.

[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," *Pattern Recognition Letters*, vol. 16, no. 3, pp. 307–312, 1995.

[11] S. Srivatava, "Distribution-based bayesian minimum expected risk for discriminant analysis," *IEEE International Symposium on Information Theory 2006*, pp. 2294–2298, July 2006.

[12] 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.), pp. 411–429, Elsevier Science Publishers B.V., 1985.

[13] M. J. Daniels and R. E. Kass, "Shrinkage estimators for covariance matrices," *Biometrics*, vol. 57, no. 4, pp. 1173–1184, 2001.

[14] 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, pp. 603–621, December 2003.

[15] F. Huibers, "De zin en onzin van het januari-effect." `http://www.nuzakelijk.nl/column-fred-huibers/2409358/zin-en-onzin-van-januari-effect.html`, December 2010.

[16] K. Falkenberg, "Disclosed to death," *Forbes Magazine*, June 2010.

[17] A. I. Goldman, *Knowledge in a Social World*. Oxford University Press, February 1999.

[18] J. Ladyman, *Understanding philosophy of science / James Ladyman*. Routledge, London ; New York, 2002.

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

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

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

[22] 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*, pp. 27–35, 2008.

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

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

[25] 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.

[26] 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.

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

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

[29] 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.

[30] C. Stein, "Lectures on the theory of estimation of many parameters," *Journal of Mathematical Sciences*, vol. 34, pp. 1371–1403, July 1986.

[31] 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.

[32] D. Paul, "Asymptotics of the leading sample eigenvalues for a spiked covariance model," tech. rep., Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305, 2004.

[33] I. M. Johnstone, "On the distribution of the largest principle component," tech. rep., Dep. of Statistics, Stanford University, 2000.

[34] A. J. Hendrikse, R. N. J. Veldhuis, and L. J. Spreeuwers, "Smooth eigenvalue estimation," *Submitted to EURASIP*, 2012.

[35] E. Jaynes, "On the rationale of maximum entropy methods," in *Proceedings of the IEEE, Special issue on spectral estimation*, vol. 70, pp. 939–952, 1982.

[36] 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.

[37] 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*, pp. 25–32, 2009.

[38] Z. D. Bai and H. Saranadasa, "Effect of high dimension: By an example of a two sample problem," in *Statistica Sinica*, no. 6, pp. 311–329, National Sun Yat-sen University, 1996.

[39] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, (Washington, DC, USA), pp. 947–954, IEEE Computer Society, 2005.

[40] 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.

[41] Z. D. Bai, “Methodologies in spectral analysis of large dimensional random matrices, a review,” in *Statistica Sinica*, no. 9, pp. 611–677, National University of Singapore, 1999.

[42] E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis*. Pearson Education, 3 ed., 2003.

[43] S. Tsai, “Characterization of stieltjes transforms,” 2000.

[44] N. R. Rao and A. Edelman, “The polynomial method for random matrices,” *Foundations of Computational Mathematics*, vol. 8, pp. 649–702, November 2008.

[45] V. I. Istrăţescu, *Fixed point theory*, vol. 7 of *Mathematics and its Applications*. Dordrecht: D. Reidel Publishing Co., 1981.

[46] B. Moghaddam and A. Pentland, “Probabilistic visual learning for object representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 696–710, 1997.

[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*, pp. 105–112, 2010.

[48] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, “Eigenvalue correction results in face recognition,” in *Twenty-ninth Symposium on Information Theory in the Benelux*, pp. 27–35, 2008.

[49] X. Mestre, “Estimating the eigenvalues and associated subspaces of correlation matrices from a small number of observations,” in *Proceedings of the Second International Symposium on Communications, Control and Signal Processing*, (Marrakech (Morocco)), 2006.

[50] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Annals of Mathematical Statistics*, vol. 22, pp. 49–86, 1951.

[51] M. Tumminello, F. Lillo, and R. N. Mantegna, “Kullback-leibler distance as a measure of the information filtered from multivariate data,” *Physical Review E*, vol. 76, p. 031123, 2007.

[52] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, “Notes on second order statistics in verification,” tech. rep., University of Twente, Enschede, the Netherlands, 2011.

[53] 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*, vol. 5558/2009 of *Lecture Notes in Computer Science*, (Berlin / Heidelberg), pp. 189–198, Springer Verlag, June 2009.

[54] R. J. Muirhead, *Aspects of multivariate statistical theory*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, inc, 1982.

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

[56] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "Likelihood ratio based verification in high dimensional spaces," *accepted for second review after major revision in Transactions on Pattern Analysis and Machine Intelligence*, 2011.

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

[58] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[59] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[60] 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.

[61] D. Middleton, *An Introduction to Statistical Communication Theory*. McGraw-Hill, 1960.

[62] 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.

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

[64] A. J. Hendrikse, R. N. J. Veldhuis, and L. J. Spreeuwers, "Smooth eigenvalue estimation," tech. rep., University of Twente, Enschede, the Netherlands, 2011.

[65] A. Hendrikse, R. Veldhuis, and L. Spreeuwers, "Verification under increasing dimensionality," *Pattern Recognition, International Conference on*, pp. 589–592, 2010.

[66] 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.

[67] 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.

[68] 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.

[69] L. Carroll, *Alice's adventures in wonderland*. Alexander Macmillan, 1865.

[70] D. R. Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid*. New York, NY, USA: Basic Books, Inc., 1979.

[71] M. Uenohara and T. Kanade, "Optimal approximation of uniformly rotated images: Relationship between karhunen–loeve expansion and discrete cosine transform," *IEEE Transactions on Image Processing*, vol. 7, pp. 116–119, 1998.

[72] A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "Face recognition based on second order statistics and high resolution images," in *to be published*, 2011.

[73] T.-H. Min and R.-H. Park, "Eyelid and eyelash detection method in the normalized iris image using the parabolic hough model and otsu's thresholding method," *Pattern Recogn. Lett.*, vol. 30, no. 12, pp. 1138–1143, 2009.

[74] M. Adam, F. Rossant, F. Amiel, B. Mikovicova, and T. Ea, "Reliable eyelid localization for iris recognition," in *ACIVS '08: Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*, (Berlin, Heidelberg), pp. 1062–1070, Springer-Verlag, 2008.

[75] A. Basit, M. Javed, and M. Anjum, "Eyelid detection in localized iris images," in *Emerging Technologies, 2006. ICET '06. International Conference on*, pp. 157 –159, 13-14 2006.

[76] Y. K. Jang, B. J. Kang, and K. R. Park, "A study on eyelid localization considering image focus for iris recognition," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1698–1704, 2008.

[77] P. R. Tabrizi and R. A. Zoroofi, "Open/closed eye analysis for drowsiness dectection," in *Image Processing Theory, Tools and Applications*, pp. 1–7, Nov 2008.

[78] K.-M. Lam and H. Yan, "An analytic-to-holistic approach for face recognition based on a single frontal view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 673–686, 1998.

[79] Q. Chen, W.-k. Cham, and K.-k. Lee, "Extracting eyebrow contour and chin contour for face recognition," *Pattern Recognition*, vol. 40, no. 8, pp. 2292–2300, 2007.

[80] F. Huang and J. Su, “Face contour detection using geometric active contours,” in *Proceedings of the 4th World Congress on Intelligent Control and Automation*, vol. 3, pp. 2090–2093, 2002.

[81] V. Vezhnevets and A. Degtiareva, “Robust and accurate eye contour extraction,” in *Proc. Graphicon-2003*, pp. 81–84, 2003.

[82] C. Grecos and M. Yang, “An eye detector based on cues and heuristics with a good accuracy/complexity trade-off,” *Adaptive Hardware and Systems, NASA/ESA Conference on*, vol. 0, pp. 492–497, 2008.

[83] B. d. Brito Leite, E. T. Pereira, H. M. Gomes, L. R. Veloso, C. E. d. Nascimento Santos, and J. M. d. Carvalho, “A learning-based eye detector coupled with eye candidate filtering and pca features,” in *SIBGRAPI '07: Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing*, (Washington, DC, USA), pp. 187–194, IEEE Computer Society, 2007.

[84] Z. Zheng, J. Yang, and L. Yang, “A robust method for eye features extraction on color image,” *Pattern Recognition Letters*, vol. 26, no. 14, pp. 2252–2261, 2005.

[85] E. Ardizzone, M. Cascia, and M. Morana, “Probabilistic corner detection for facial feature extraction,” in *ICIAP '09: Proceedings of the 15th International Conference on Image Analysis and Processing*, (Berlin, Heidelberg), pp. 461–470, Springer-Verlag, 2009.

[86] K.-M. Lam and H. Yan, “Locating and extracting the eye in human face images,” *Pattern Recognition*, vol. 29, no. 5, pp. 771–779, 1996.

[87] M. H. Khosravi and R. Safabakhsh, “Human eye sclera detection and tracking using a modified time-adaptive self-organizing map,” *Pattern Recogn.*, vol. 41, no. 8, pp. 2571–2593, 2008.

[88] D. J. Williams and M. Shah, “A fast algorithm for active contours and curvature estimation,” *CVGIP: Image Underst.*, vol. 55, no. 1, pp. 14–26, 1992.

[89] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, “Feature extraction from faces using deformable templates,” *International Journal of Computer Vision*, vol. 8, pp. 99–111, August 1992.

[90] O. Pichler, A. Teuner, and B. J. Hosticka, “A comparison of texture feature extraction using adaptive gabor filtering, pyramidal and tree structured wavelet transforms,” *Pattern Recognition*, vol. 29, no. 5, pp. 733 – 742, 1996.

[91] J. Milton, “The devil's advocate.” [DVD], 1997.

[92] S. Hawking, *A brief history of time: from the big bang to black holes*.

[93] R. Williams and K. Herrup, “The control of neuron number,” *Annual Review of Neuroscience*, vol. 11, pp. 423–453, 1988.

[94] I. van Staveren, "Wat is de meest dringende maatschappelijke kwestie van dit moment?." `http://www.groene.nl/2011/wetenschappers/irene-van-staveren`, 2011.

[95] A. J. Hendrikse, R. N. J. Veldhuis, and L. J. Spreeuwers, "Component ordering in independent component analysis based on data power," in *Proceedings of the 28th Symposium on Information Theory in the Benelux, Enschede, The Netherlands* (R. N. J. Veldhuis and H. S. Cronie, eds.), (Eindhoven), pp. 211–218, Werkgemeenschap voor Informatie- en Communicatietechniek, June 2007.

[96] R. N. J. Veldhuis, A. M. Bazen, W. Booij, and A. J. Hendrikse, "Hand-geometry recognition based on contour parameters," in *SPIE Biometric Technology for Human Identification II, Orlando, FL, USA* (A. K. Jain and N. K. Ratha, eds.), (Washington), pp. 344–353, SPIE – The Int. Society for Optical Engineering, March 2005.

[97] R. N. J. Veldhuis, A. M. Bazen, W. D. T. Booij, and A. J. Hendrikse, "A comparison of hand-geometry recognition methods based on low- and high-level features," in *15th Annual Workshop on Circuits Systems and Signal Processing (ProRISC), Veldhoven, The Netherlands*, (Utrecht, The Netherlands), pp. 326–330, Technology Foundation STW, November 2004.

List of publications

2011

- A. J. Hendrikse, R. N. J. Veldhuis, and L. J. Spreeuwers, "Smooth eigenvalue estimation," *Submitted to EURASIP*, 2012
- A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "Face recognition based on second order statistics and high resolution images," in *to be published*, 2011
- A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "Likelihood ratio based verification in high dimensional spaces," *accepted for second review after major revision in Transactions on Pattern Analysis and Machine Intelligence*, 2011

2010

- 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*, pp. 105–112, 2010
- A. Hendrikse, R. Veldhuis, and L. Spreeuwers, "Verification under increasing dimensionality," *Pattern Recognition, International Conference on*, pp. 589–592, 2010

2009

- 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*, pp. 25–32, 2009
- A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "A bootstrap approach to eigenvalue correction," in *ICDM '09*, pp. 818–823, IEEE Computer Society Press, December 2009
- 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*, vol. 5558/2009 of *Lecture Notes in Computer Science*, (Berlin / Heidelberg), pp. 189–198, Springer Verlag, June 2009

2008

- A. J. Hendrikse, L. J. Spreeuwers, and R. N. J. Veldhuis, "Eigenvalue correction results in face recognition," in *Twenty-ninth Symposium on Information Theory in the Benelux*, pp. 27–35, 2008

2007

- A. J. Hendrikse, R. N. J. Veldhuis, and L. J. Spreeuwers, "Component ordering in independent component analysis based on data power," in *Proceedings of the 28th Symposium on Information Theory in the Benelux, Enschede, The Netherlands* (R. N. J. Veldhuis and H. S. Cronie, eds.), (Eindhoven), pp. 211–218, Werkgemeenschap voor Informatie- en Communicatietechniek, June 2007

2005

- R. N. J. Veldhuis, A. M. Bazen, W. Booij, and A. J. Hendrikse, "Hand-geometry recognition based on contour parameters," in *SPIE Biometric Technology for Human Identification II, Orlando, FL, USA* (A. K. Jain and N. K. Ratha, eds.), (Washington), pp. 344–353, SPIE – The Int. Society for Optical Engineering, March 2005

2004

- R. N. J. Veldhuis, A. M. Bazen, W. D. T. Booij, and A. J. Hendrikse, "A comparison of hand-geometry recognition methods based on low- and high-level features," in *15th Annual Workshop on Circuits Systems and Signal Processing (ProRISC), Veldhoven, The Netherlands*, (Utrecht, The Netherlands), pp. 326–330, Technology Foundation STW, November 2004

Acronyms

SAS Signals and Systems

ROI Region of Interest

PCA Principle Component Analysis

LDA Linear Discriminant Analysis

FAR False Accept Ratio

FRR False Reject Ratio

EER Equal Error Rate

DET Detection Error Trade-off

ROC Receiver Operating Characteristic

DCT Discrete Cosine Transform

SVD Singular Value Decomposition

LSA Large Sample Analysis

GSA General Statistics Analysis

SOS Second Order Statistics

BCIV bias correction in verification

i.i.d. independent and identically distributed

N the number of samples

p the number of dimensions

MP Marčenko Pastur

MDL Minimum Description Length

Nawoord

Tijdens het laatste deel van mijn afstuderen bij de vakgroep signalen en systemen vertelde Raymond mij over het eigenwaarde probleem, wat al langer bekend was binnen de patroonherkenning. Hij had samen met Asker een eerste oplossing ontwikkeld maar daar mocht wel eens wat uitgebreider naar gekeken worden en dus was Raymond op zoek naar iemand die eens wat dieper op dat onderwerp in kon gaan. Met mijn interesse in de wiskundige kant van patroonherkenning was dit een goede match.

Omdat het onderwerp tweede orde statistiek schatting in hoog dimensionale data toch wat abstract is, zijn de bijdragen van mensen om mij heen wat moeilijk te onderscheiden. Desondanks hebben zij grote invloed gehad op de totstandkoming van dit proefschrift, waar ik hier voor wil bedanken. Het meest direct zichtbaar is het aandeel van mijn directe begeleiders Raymond en Luuk (zeker als je de tekst vergelijkt met de eerste versies). Het onderwerp van mijn onderzoek verschilt behoorlijk van de onderwerpen waar zij zich doorgaans mee bezig houden, desalniettemin heb ik veel nuttige feedback en discussies gehad. Daarnaast weet ik niet zeker of Raymond alles wat ik te vertellen had helemaal verwachtte, des te groter is mijn dank en respect voor de vrijheid die ik gekregen heb om mij te verdiepen in de tweede orde statistiek schatting in hoog dimensionale data. Daarnaast wil ik ook mijn promotor Kees bedanken voor de geboden mogelijkheden.

Het onderzoek met hoog dimensionale data stelt vrij hoge eisen aan de computer hardware. Geert-jan heeft een heel aardig rekenmonster voor mij weten te regelen. Samen met Henny heeft hij mij ook bij de nodige praktische problemen vooruit geholpen. Bij meer organisatorische problemen heb ik veel hulp gekregen van Anneke en Sandra.

Ik heb tijdens mijn promotie ook veel gehad aan mijn mede promovendi, onder andere tips over conferenties en software (vooral over dat laatste heb ik nog wel eens een discussie had met Almar). Met Bas heb ik zelfs nog een weekje na de ICB op Sardinië rondgereden. Om niet volledig te verdwalen in hoog dimensionale problemen, waren er aan het begin van mijn promotie de vrijdag middag borrels, die soms nog wel eens wat langer duurde. Hieruit volgden de wintersport vakanties met Rene, Bas, Dirk-Jan, Gerbert, Jelle en anderen, waar ik ook nu elk jaar weer naar uitkijk.

Voor mij is het erg belangrijk om het zitten achter de computer af te wisselen met beweging, waar de campus in Twente voldoende mogelijkheden voor biedt. Een van mijn favoriete sporten is volleybal. Aangezien volleybal pas leuk wordt met meer mensen, wil ik de teamgenoten van bedrijfssport, Vasata, Harambee, Nuvo'68 en VC Natlab bedanken. Ik hoop velen van jullie nog tegen te komen, onder andere bij toernooien. Over sport gesproken, dan moet ik de groep jaargenoten bedanken dat ze nog elk jaar met mij een weekend willen mountainbiken (behalve Vinnie dan).

De afronding van mijn promotie werk heb ik in Eindhoven gedaan. Bij GN Resound heb ik de tijd gekregen om de puntjes op de i te zetten en alvast aan een nieuw onderwerp in signaal bewerking te werken, waarvoor mijn dank.

Waar ik gekomen ben zou nooit gelukt zijn zonder de steun van mijn familie. Allereerst Dineke, Petra, Corine, Kees en Hans, ik heb het erg getroffen met jullie. Daar zijn in de afgelopen jaren Pieter en Bart bij gekomen. Pappa en mamma, bedankt voor de ondersteuning die zo nu en dan hard nodig was.

Anne Hendrikse
Eindhoven, the Netherlands
April 2012