Cognitive Modelling - 02458
Exam Project
Age in Faces

December 2017

Contents

1 Introduction/Motivation

In this report, a study of how humans perceive age will be performed with the intention of determining characteristic facial features that the subjects of our experiment found important when it comes to evaluating age.

2 Method and Theory

2.1 Creation of experimental setup

In order to produce a model to determine age based on facial images, it is necessary to gather some training data. The training data would in the case of this report be a set of human responses, an estimate of age, to a collection of stimuli, in the form of a series of human faces. However, before such an experiment can be set up, it is necessary to develop a nice collection of human faces in various age groups to present as stimulus. In order to get a nice spread of age groups we chose to use the IMDB-WIKI data set [1, 2]. According to the providers, the IMDB-WIKI data set is the largest publicly available data set of face images with gender and age labels for the training of machine learning models (ML). The data set contains 460,723 face images of 20,284 celebrities from IMDB as well as 62,328 images from Wikipedia. For the purpose of this project we chose only to focus on the images from IMDB as these seemed to be of greater quality and had a less skewed distribution of ages as evidenced in Fig. 1.

Having found a large data set with a considerable span in age and gender, the next step was to standardize the images in such a way that they could be compared. This was necessary as to ensure that the same physical features of the human faces would appear in roughly the same pixel patches for each image such that we would have a well-defined average face for both men and women. Another important point was to ensure that we got a relatively even distribution of ages as well as genders. To do this, we loaded the images into MATLAB where-after we wrote a script ensuring that we could manually approve each image before saving it to the final data set to be used for our cognitive experiment. Essentially the script randomly selects one of the pictures from the set with a dimensionality of at least 300 by 300 pixels, and crops it using MATLAB's CascadeObjectDetector. The CascadeObjectDetector is a system object using the Viola-Jones algorithm to detect people's faces, noses, eyes, mouths, or upper body [3]. Thus we simply cropped the input image to the pixels in which a face was recognized and thereafter re-sized it to a 500×500 pixel image. The image was then displayed to the user together with the given labels associated for validation. This was a necessity as some

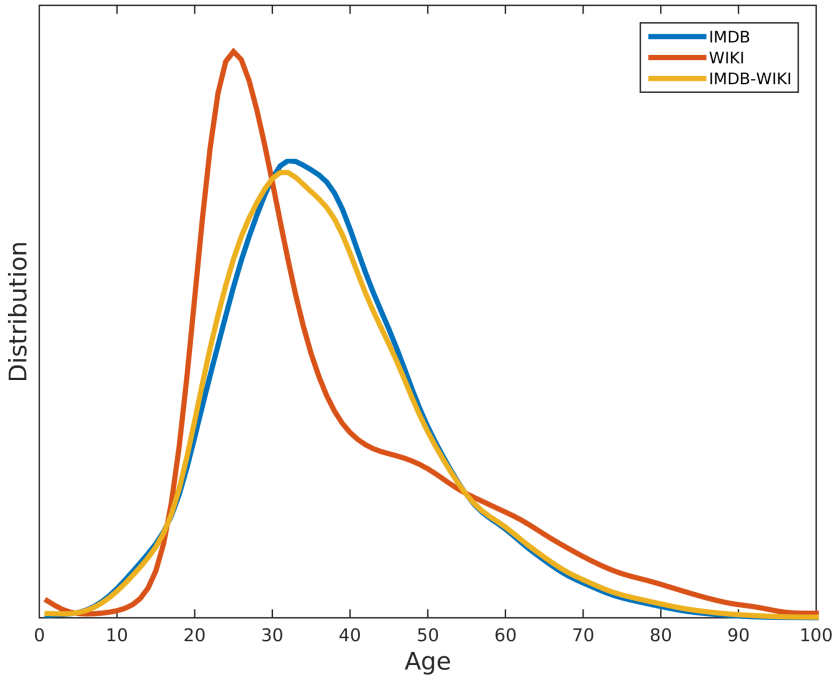

Figure 1: Distribution of ages in the IMDB-WIKI data set

of the images contained more than one person and thus the given label might have been associated with a person different from the one caught by the CascadeObjectDetector. If the image got manually approved, the image would be transformed into grayscale and stored in matrix form to the appropriate predefined age group bucket (a tensor of rank 3) based on the image's age label. The reason for grayscaling the image was not only to reduce the dimensionality of the images, but also to ensure that the conclusions gotten from data processing would be solely based on facial features without the noise generated by colour. A total of 5 age group buckets, linearly splitting an age span from 10 to 80 years, were predefined for both males and females, each of which would ultimately contain 50 unique images.

The result would thus be a data set consisting of 250 images of relatively even age distribution for each gender. The reason for splitting the data set by gender, lies in the inherent difference between male and female faces, and thereby the features important for age. As to ensure that different pictures of the same actor/actress would be excluded from the final data set, the celebrity id (a unique id to distinguish actors/actresses in

the original data set) was stored to a vector and used as a filter for the input images.[1]

Having generated a set consisting of 250 faces, distributed evenly in age, for both males and females, we continued preparing an experiment to test how given subjects perceive age. The experiment was conducted in the following manner. Each subject was asked to look through the 250 images for each gender, typing in the age that first came to mind when viewing the image. After a couple of attempts it was realized that the constructed data set was not completely ideal in the sense that it contained well known actors/actresses, implying that the subjects seemed to have a bias, i.e. an idea of the age of the shown image. This effect was however mitigated to some extent by actively asking the subjects to judge the age solely on the facial features presented, but also naturally by the fact that some of the images were from old movies implying that the knowledge of current age became useless. In total we managed to gather the input of 3 individuals for the male data set and the input of 5 individuals for the female data set.

As to ensure that no outliers would be included in the data for analysis, a filter was implemented checking the difference in the inputted ages for the male and female data sets. If, for a given image, the absolute difference in age was over 20 years, the image was disregarded from the set. This excluded a total of 2 images for the males and 5 images for the females. In Fig. 2 the average male and female face composed of the 248 and 245 images respectively can be seen. This confirms that the images chosen for the experiments indeed are comparable and centered, as we get a relatively clear depiction of a human face.

2.2 Method used for modelling

Due to the humongous dimensionality of the rank 3 tensors created for image storage as part of the experimental setup, the amount of processing power needed to do real modelling of the data was extensive and hence dimensionality reduction was appropriate, or rather, a necessity. To be more exact, the dimensionality of each rank 3 tensor is $500(\text{height}) \times 500(\text{width}) \times 250(\text{number of gender specific pictures}) = 6.25 \times 10^7$. There exists several methods for dimensionality reduction with the ones touched upon in the course being Principal Component Analysis (PCA) and Independent Component Analysis (ICA). Other methods include auto-encoders (deep artificial neural networks) which can be very efficient, but possess the disadvantage that we would miss the interpretable cognitive properties of our sought model. This is due to the face that auto-encoders make highly non-linear mappings that are hard to track.

[1] That being said, it could have been tremendous to have a data set consisting of 250 pictures of the one and only Samuel L. Jackson (what a **man!**)

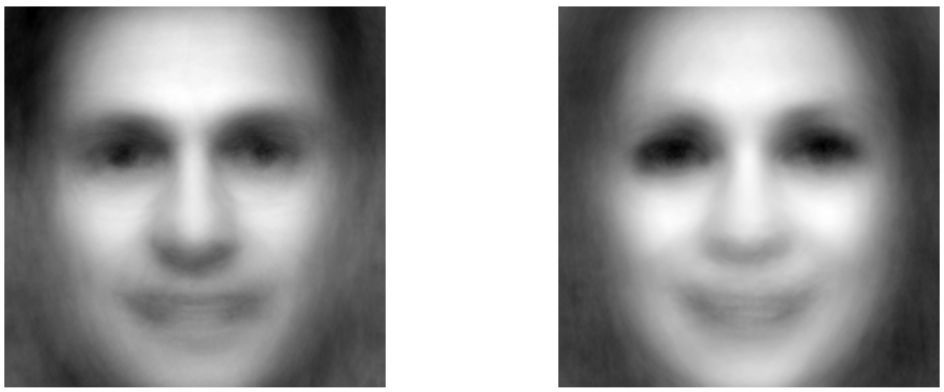

Figure 2: Left: The average face composed of the 248 male images. Right: The average face composed of the 245 female images.

2.2.1 Principal Component Analysis (PCA)

As alluded to in the previous section, a reduction in dimensions will be necessary if any computations and transformations are to be feasible. However, there is another stronger argument for performing a PCA rooted in the modest proposal, that information is represented in an efficient way in the human brain. To elaborate, this rests on the belief that although the human brain receives many signals from its environment, it will not make a one-to-one mapping of these signals to produce a representation of the external environment. Instead, in the essence of being efficient, it will attempt to decorrelate the signals received in the receptors to a representation sufficient for survival [4]. The human brain already consumes 20% of a humans total energy consumption, spending more energy in an attempt to for example render the single hair strand of a tiger would simply not make any sense, as you could use a significantly lower amount of effort and energy in simply identifying the threat of a tiger by the orange and black stripes [2]. In the light of this discussion, one could claim that PCA functions in much the same way. Instead of gathering all the information from the image in question, most of the vital information for cognitive tasks, such as determining the age of a person on the basis of his or her facial features, can be extracted by considering a significantly lower dimensionality.

On a practical level, before performing the PCA, it was necessary to re-scale the images

[2]This is not to say that you should run for your life if you see orange and black stripes as you are strolling through the city of Copenhagen, you should however strongly consider it if you're in an Indian jungle.

from 500×500 pixels to 100×100 pixels to make any further transformations feasible. Although this may sound like a serious reduction of information, we feel it safe to assume that the features important for age-determination would still present themselves in the dimension reduced images. If time allowed for a rerun of the experiment, we would have presented the 100 by 100 pixel images to ensure that no information be lost. We then flattened each of the images to a 100(height) $\times$ 100(width) $= 10,000$ dimensional vector, and placed them as consecutive rows in a N by 10,000 matrix for each gender, where N is the total number of images for the given gender. The matrix was then normalized by subtracting the mean from each of the observations and dividing by the standard deviation pixel-wise. PCA was then evaluated on the resulting normalized matrix. The results can be seen in Fig. 3 where the first 16 eigenfaces have been displayed for both the males and females. An eigenface is in it's essence a principal component that has been restructured to a, in this case, 100 by 100 matrix as to resemble an image. One can therefore think of eigenfaces as being the building blocks of any facial image by the weighted sum of the eigenfaces. For instance a face F could be evaluated as $F = 23\%E_1 + 50\%E_2 - 17\%E_3 + ... + 1\%E_n$, where E_i represents the i'th eigenface, n denotes the number of eigenfaces considered for the reconstruction and the percentages are the weights associated to each eigenface (and can be negative). We will talk more about reconstruction of images later.

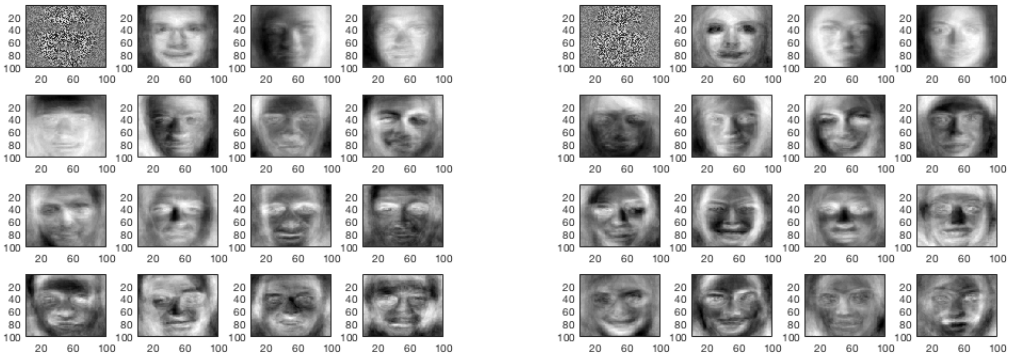

Figure 3: Left: The 16 eigenfaces with highest variance for the males sorted by descending variance. Right: The 16 eigenfaces with highest variance for the females sorted by descending variance.

As evident from Fig. 3, the first eigenface for both males and females looks like random noise, which makes a lot of sense, since we have normalized the data. The eigenface with the second highest variance seems to be a generic male and female face from which most of the gender specific features can be extracted. Another point to make is that, for both

men and women, the next couple of eigenfaces seem to be related to a lighting of the face from the sides and from top and bottom. It is also worth mentioning that some of the less important eigenfaces seem to be concerned with facial symmetry, facial hair and sizes of noses, eyes and mouths. In order to find the appropriate number of eigenfaces to consider for further analysis, we will set a threshold for the variance explained by the eigenfaces and choose the minimum number of eigenfaces required to obtain the threshold.

Figure 4: Reconstruction of a male face using the number of eigenfaces capable of explaining 80%, 90%, 95%, 99% and 100% of the variance respectively.

In Fig. 4, we see the same male face reconstructed using an increasing number of eigenfaces. Specifically they have been created using variance thresholds of 80%, 90%, 95%, 99% and 100% using a total of 27, 64, 110, 195 and 248 respectively (since the given image depicts a person of the male gender). As evident, the difference in the quality of the reconstruction is significantly reduced as the total number of eigenfaces is increased. For the remainder of the report, we will use the 90% variance threshold for the principal components, corresponding to a total of 64 eigenfaces to consider for both blokes and sheilas.

2.3 The Multi-Linear Regression

Ideally, we would like to create an interpretable supervised model which accurately predicts how our subjects perceive age. That would be a model generalizing well which can be identified with a model having a small test error in Fig. 5. To train a supervised learning model such that we can obtain a small error, it is important to consider the complexity of the model one should make. For this purpose, we could fairly easily be able to create an artificial neural network (ANN) which could create some non-linear mappings resulting in a small error. However, it would take the interpretability out of the model since the interpretability of ANNs is very limited. Hence, we decided to create a multi-linear regression model using the Least Absolute Shrinkage and Selection Operator (Lasso) as we would like the model to be interpretable. The Lasso is used for regularization and dimensionality reduction such that only the most important eigenfaces ends up having non-zero weights. To elaborate further on the choice of regularization, i.e.

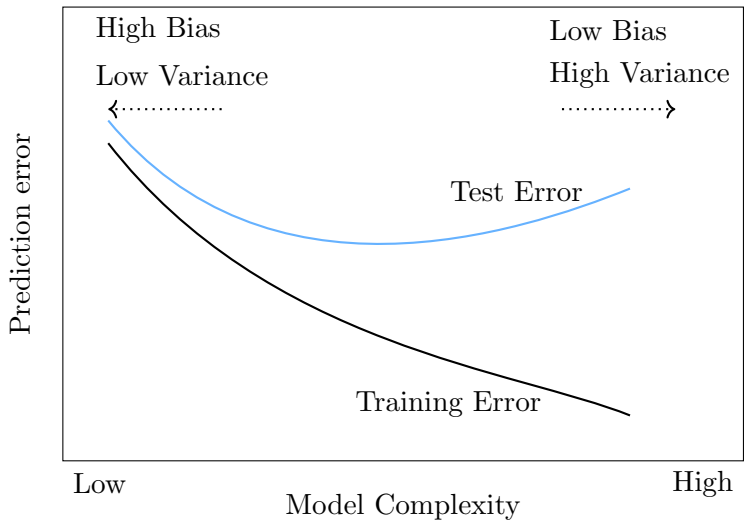

Figure 5: Illustration of the Bias-Variance trade-off in model selection.

the L_1-norm (Lasso), we will first need to introduce the ordinary least squares regression and the objective function that we are trying to minimize.

From the executed experiments done by our subjects, we created two data sets for modelling (one for each gender) with $N_{\text{male}} = 248$ male observations, and $N_{\text{female}} = 245$ female observations. Lets denote an observation by a feature vector $\boldsymbol{x}_i$ and a target y_i which is the perceived age of a given face. The feature vector $\boldsymbol{x}_i$ is a M-dimensional vector consisting of the projection-weights of a given photo onto the PCs found by PCA to achieve a given percentage of the total variance in the original picture for men and women respectively (in this case 90%). The target y_i for each image is taken to be the mean age perceived by our subjects. All data pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$ are then gathered into data sets $(\boldsymbol{X}_{\text{male}}, \boldsymbol{y}_{\text{male}})$ and $(\boldsymbol{X}_{\text{female}}, \boldsymbol{y}_{\text{female}})$ separately.[3]

Having assumed that we can establish a linear relationship between a feature vector $\boldsymbol{x}_i$ and its corresponding target y_i, i.e., we can write a predicted target on the form $y_i = f(\boldsymbol{x}_i, \boldsymbol{w}) = \boldsymbol{x}_i^\top \boldsymbol{w}$ (due to the fact that the data is centered from PCA), where the optimal weights $\boldsymbol{w}^*$ are found by minimizing the objective function in the ordinary least squares regression (OLS) [5],

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 \right\} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{w} \right)^2 \right\} \tag{2.1}$$

[3] With $\boldsymbol{X}_{\text{male}}$ having dimension $N_{\text{male}} \times M$ and $\boldsymbol{y}$ having dimension N_{male}. The only difference to the dimension of the female set being that $N_{\text{male}} \neq N_{\text{female}}$.

2.3.1 The Least Absolute Shrinkage and Selection Operator (Lasso) and the Ridge regression

As the objective function in the OLS regression does not take into account that the weights might have a prior probability distribution, it does not really match our purpose since we want to find a sparse solution, i.e., a low-dimensional feature vector containing the important eigenfaces for predicting how our subjects perceive facial age. Ideally, the weight distribution would look more like a Laplacian distribution and thus we chose to regularize with Lasso. We also considered the Ridge regression which regularizes with the L_2-norm and assumes that the prior distribution of the weights follow a normal distribution. But applying a Ridge regression would not allow for as sparse solutions as using a Lasso regularization. To elaborate further, looking at Fig. 6 we see the contours which are the weights of the OLS solution. The solution to minimizing the objective functions for the Lasso- and Ridge regressions respectively are seen in the figure as the points where the contours hit the constraint regions, i.e, touches the L_1-norm (Lasso) and L_2-norm (Ridge) regions.[4] Hence this illustrates why weights rarely go all the way to zero for the Ridge regression since there are no edges, whereas the Lasso solution allows for sparse solutions which is also evident from the geometric representation. After having argued why we went for a Lasso regularization it seems appropriate to introduce the objective function for such a regression type,

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - f(\boldsymbol{x}_i, \boldsymbol{w})\right)^2 + \lambda \|\boldsymbol{w}\|_1 \right\} \tag{2.2}$$

where λ is called the regularization strength [6]. From Eq. (2.2) it is evident that if the regularization strength λ is increased, the norm of the weights will have to decrease. Another purpose for regularization, is that it allows for the tuning of model complexity to hit the right amount of bias and variance as shown in Fig. 5. Additionally to using Lasso for model selection and performance evaluation we used Cross-validation (CV), which is a technique which can optimize hyper-parameters (model selection) and give the best possible estimation of the true test error through an idealized quantity called the generalization error [5].[5]

[4] This figure is produced for illustrative purposes and the concept can be extended to higher dimensions.

[5] Further elaboration on this can be found in Appendix A

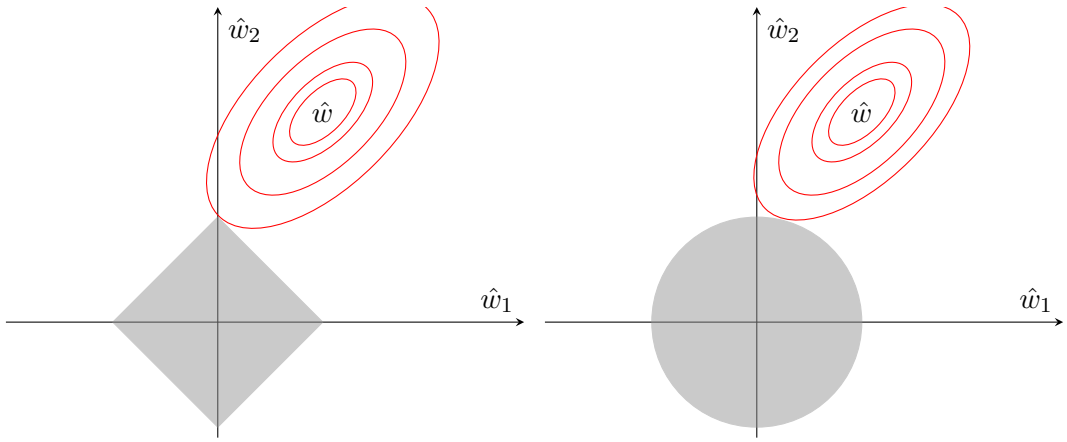

Figure 6: Geometric representation of the solutions to the Lasso (Left pane) and the Ridge (Right pane) respectively.

3 Results

In the following section, the results of the Lasso regression on the male and female data sets will be presented. In order to train the linear model, we need to develop a meaningful target vector for each of the images, based on the experimental input. This was done practically by taking the average of the inputted ages across all the experiments for each of the unique images. Also worth noting, is the fact that the degrees of freedom for the Lasso regression was capped at 40 and that a leave-one-out cross-validation was applied for both genders.

3.1 Male Data Set

We now present the results for the Lasso regression performed on the male data set. In Fig. 7 we see a plot of the weights of the chosen eigenfaces (denoted by the different colours in the plot) as a function of decreasing regularization strength λ. As evident, the lower the strength, the greater the number of eigenfaces and the larger the weights become as discussed in Section 2.

In Fig. 8 we see the evolution of the Mean Square Error (MSE) as a function of decreasing regularization strength λ. As evident, a local minimum has been found (local in the sense that we have constrained the Lasso to only consider up to 40 degrees of freedom) at $\lambda = 1.32$ corresponding to a MSE $= 189.4$ for the optimal model which was validated with leave-one-out cross-validation for the best possible estimate of the generalization error.

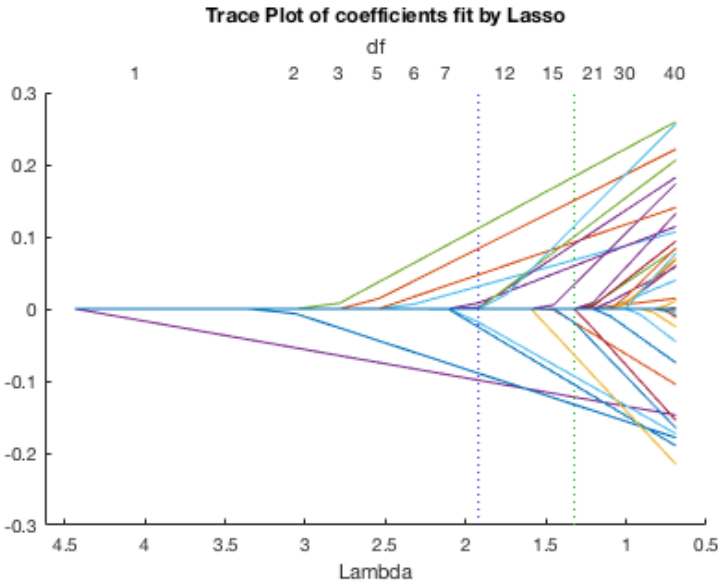

Figure 7: The weights of each eigenface plotted as a function of the λ parameter for the male images.

For the males a total of 16 eigenfaces were chosen with the optimal Lasso as evident in the left pane of Fig. 9. Out of these 16 eigenfaces, 7 had negative weights associated with them whilst 9 had positive weights associated with them. The sign of the weights determine whether a positive projection onto one of the eigenfaces will correspond to an increase in the model's prediction of an age or not. The eigenfaces with negative weights can thus be thought of as young and vice versa.

Taking a closer look at the 16 chosen eigenfaces depicted in Fig. 9, it becomes evident that the first couple of eigenfaces have "younger" characteristics in the sense that they have narrow noses, smoother skin and less noise surrounding the eyes. On the other hand, the last couple of eigenfaces seem to have wider noses and significantly more facial hair, which seems to indicate that facial hair makes you look older. It is also evident that the later eigenfaces have more glasses also indicating that the older a given man is the more likely he is to problems with sight.

Since we now have an understanding of how to interpret the chosen eigenfaces, we can proceed and try to look at the inverse problem, i.e., try to predict a face given an age. By Section 2, we have assumed a linear relationship between the age and our eigenfaces, i.e.,

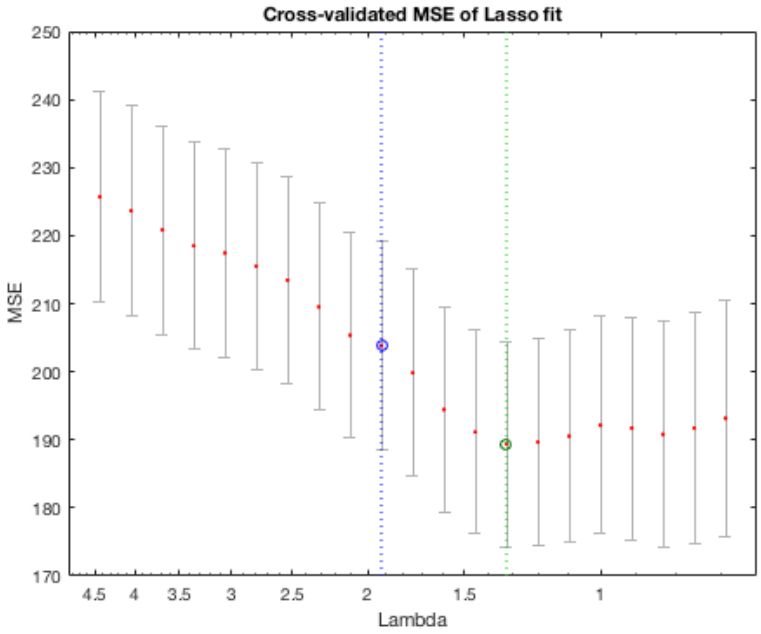

Figure 8: The Mean Squared Error (MSE) plotted as a function of the λ parameter for the male images.

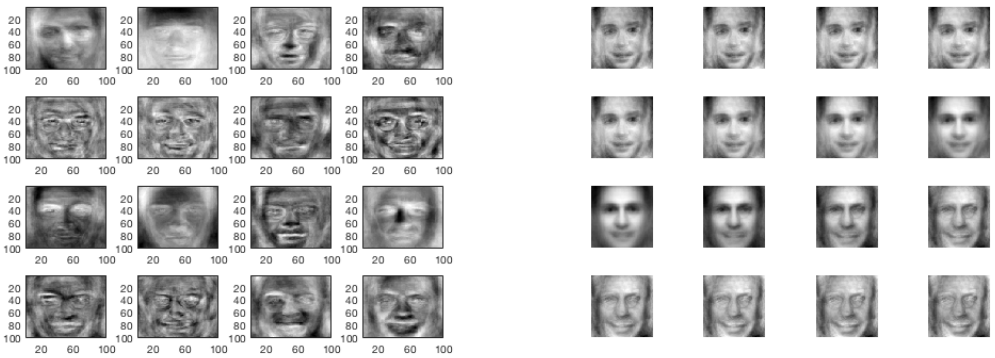

Figure 9: Left: The chosen eigenfaces from the Lasso regression ordered from the eigenface with the largest negative weight to the eigenface with the largest positive weight. Right: A linearly varying age based on the linear extension of the Lasso model for males, from the mean age minus 30 years to the mean age plus 30 years.

$y_i = \boldsymbol{x}_i^\top \boldsymbol{w}$. Since the feature vector $\boldsymbol{x}_i$ is just a straight line in a M-dimensional vector space and we know that only 16 eigenfaces contribute to the estimation of perceived age, we can write an observation $\boldsymbol{x}$ as a constant α times the contributing weights and

their orthogonal complement (all of the non-contributing eigenfaces). Mathematically speaking we can write the linear relationship between targets and inputs as,

$$\begin{aligned} y_i &= \boldsymbol{x}_i^\top \boldsymbol{w} \\ &= \left(\alpha_i \boldsymbol{w}^\top + \boldsymbol{w}_\perp^\top\right)\boldsymbol{w} \\ &= \alpha_i \boldsymbol{w}^\top \boldsymbol{w} \\ &= \alpha_i \|\boldsymbol{w}\|^2 \end{aligned} \tag{3.1}$$

where $\alpha_i = \dfrac{y_i}{\|w\|^2}$ and hence we see that the age perceived is proportional to α. By applying the thoughts behind Eq. (3.1), we can express a feature vector $\boldsymbol{x}_i = \alpha_i \boldsymbol{w}$ and thereby be able to reconstruct an image of arbitrary age. This has been done on the average male face (Fig. 2) for the average age (corresponding to the intercept of the Lasso regression) minus 30 years to the average age plus 30 years, the results of which can be seen in the right pane of Fig. 9. Looking at the figure, similarities to the discoveries found from analysis of the plots in the left pane of Fig. 9 are evident. First and foremost, the eyes on the young male is clearly more well-defined. Likewise, another similarity is the appearance of glasses as the male is tweaked towards the elderly man in the bottom-right image. The third comment on the contributing eigenfaces holds for this evolution as well as the nose of the male clearly gets bigger as "time passes" (which is also a known fact for all human beings).

Another point to consider is the fact that the older/younger the image the less relative change there is in consecutive images which seems to indicate that the Lasso model has been constrained to such a degree that it gains better generalization error by guessing on ages close to the mean. In other words, we only see significant changes in the images for ages close to the mean age of approximately 43.7 years.

3.2 Female Data Set

We now present the results for the Lasso regression performed on the female data set. Similarly to the male Lasso, Fig. 10 represents a plot of the weights of the chosen eigenfaces as a function of decreasing regularization strength λ. Again, we see that the lower the strength, the greater the number of eigenfaces and the larger the weights become.

In Fig. 11 we see the evolution of the MSE as a function of decreasing regularization strength λ. As evident, a local minimum has been found at $\lambda = 0.75$ corresponding to MSE=169.3 for the optimal model.

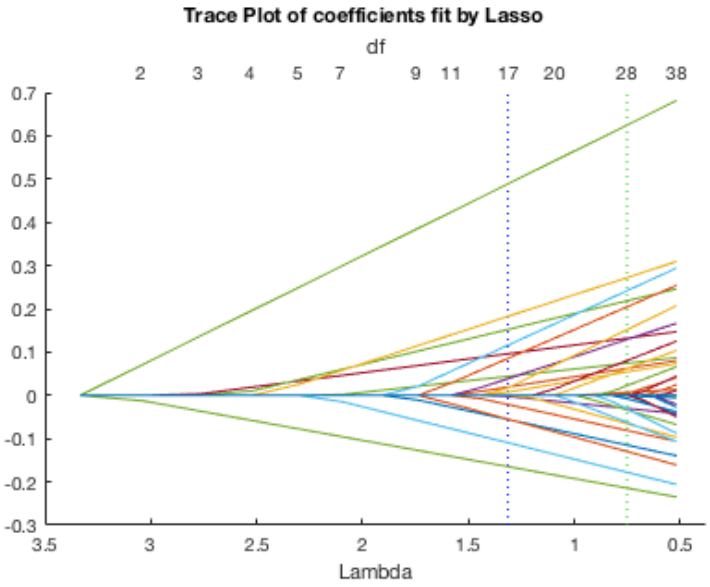

Figure 10: The weights of each eigenface plotted as a function of the λ parameter for the female images.

For the females a total of 28 eigenfaces were chosen, in which the first 12 have negative weights. In the left pane of Fig. 12, the first 8 eigenfaces are the ones with the most negative weights and the final 8 eigenfaces are the ones with the highest positive weights. As for the males, we can see that the eigenfaces in the top left corner of the image patch seem to possess younger facial features such as small nose, smooth skin and large eyes. On the contrary we see that the eigenfaces in the lower right corner are more grainy which could correspond to wrinkles, also evident are the smaller eyes and larger noses. In the right pane of Fig. 12 we have displayed images of ages linearly varying from the Lasso intercept of 43.2 years minus 30 years to the intercept plus 30 years based on the average female face (Fig. 2). As for the males we see that the images in the extreme points of high/low age have similarities to the eigenfaces chosen by the Lasso regression. Also evident is the same trend with the variance being low and centered about the mean image, which as for the males, points towards a too harsh constraint on the part of the Lasso regression.

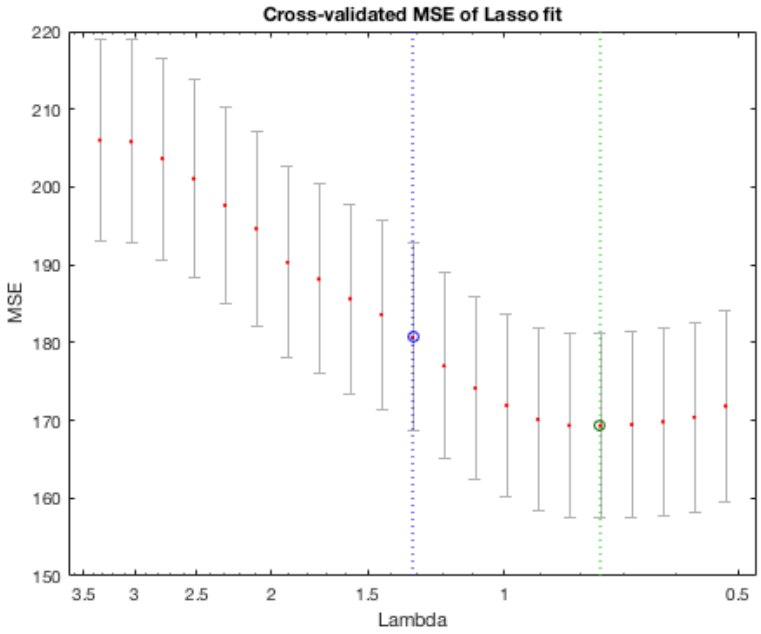

Figure 11: The Mean Squared Error (MSE) plotted as a function of the λ parameter for the female images.

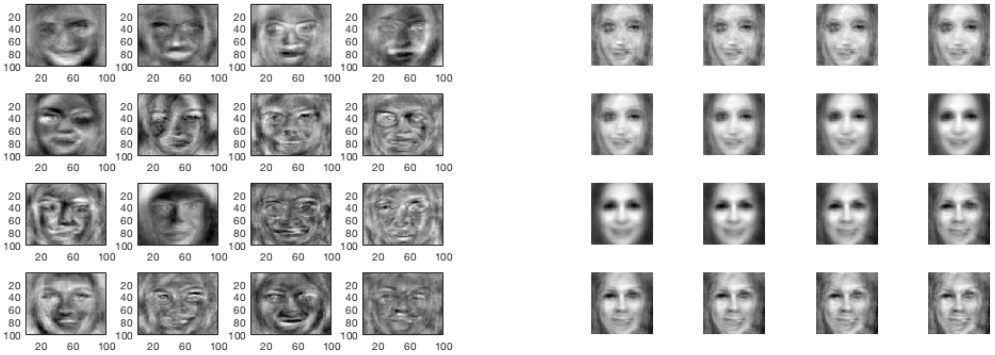

Figure 12: Left: The 16 most important chosen eigenfaces from the Lasso regression ordered from the eigenface with the largest negative weight to the eigenface with the largest positive weight. Right: A linearly varying age based on the linear extension of the Lasso model for males, from the mean age minus 30 years to the mean age plus 30 years.

4 Discussion

As seen in Section 3, the Lasso solution tends to minimize the error by truncating a lot of weights at 0 hence making it a sparse solution. Considering the fact that it actually prefers solutions where it solely predicts ages close to the mean age of the data sets, it is evident that the model has low variance and high bias. Actually the solution is too sparse since it does not allow for predictions in the outer regions of the age-spectrum of the data. Hence, even though we can collect some valuable knowledge from the Lasso regression it would be nice to create a model which allows for more variance in its predictions. To alleviate the pain of too little variance, we discussed if sequential feature selection (i.e. forward selection) could do the trick. In forward selection, one starts by creating all possible linear models with only 1 eigenface. Then the eigenface which alone gives the lowest value to the chosen dissimilarity measure will be added to the final feature list. Afterwards new linear models are trained with the first chosen eigenface *and* a new eigenface where all possibilities are addressed again, choosing the next feature, i.e., the one that minimizes the dissimilarity measure the most. This procedure is continued until the chosen dissimilarity measure cannot be reduced by adding another feature or reaching a preset feature limit. During the final stages of our project work, we had a chance to construct such a model. This should allow for weights with greater absolute value compared to those of the Lasso regression.

4.1 Sequential Feature Selection - Forward Selection

4.1.1 Results of the Forwards selection model

The dissimilarity measure chosen for the forward selection model was a MSE based on the results of a linear regression of the training data compared to the actual results of the test data. The test and training data was split using a 10-fold cross-validation. Running the forward selection model with the before-described dissimilarity measure, under the constraint that the maximum number of features to include is 16 (corresponding to the number of chosen eigenfaces for the men by the Lasso regression), we get a MSE of 137.0 for both male and female. This is of course a measure of the training error which is inherently smaller than that of the generalization error, which makes it hard to compare one-to-one. However, since the weights in general have a greater absolute value compared to that of the Lasso regression, one can claim that they allow for more variance and thus better predictions.

In Fig. 13 and Fig. 14 we see the results of the sequential feature selection model for the males and females respectively. The left pane in both images shows the 16 eigenfaces

chosen by the forward selection whereas the right pane shows the linear variation in age. Interestingly many of the features commented upon in Section 3 still hold (some of the eigenfaces are actually identical). Examples include the facial hair and glasses for the men, and the grainy images for the women.

Another interesting point arises from the fact that the variance in the linearly varying age for both the males and females seem to be higher in comparison to the Lasso results. This goes hand in hand with the statement that the sum of the weights is higher.

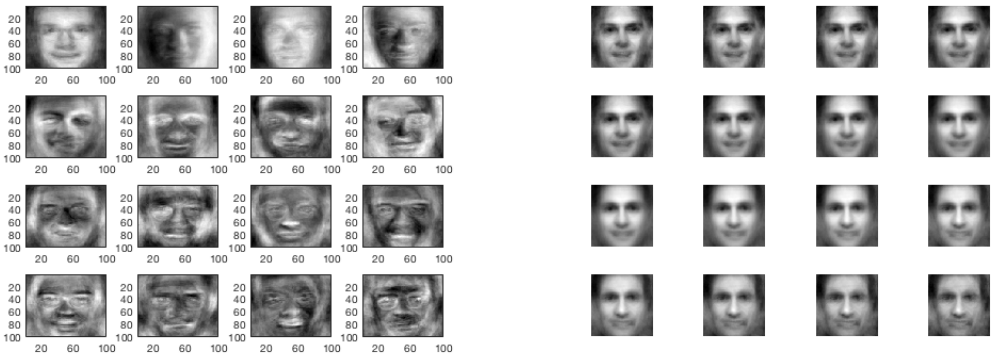

Figure 13: Left: The chosen eigenfaces from the Forward Feature Selection ordered from the eigenface with the largest negative weight to the eigenface with the largest positive weight. Right: A linearly varying age based on the linear extension of the Forward Feature Selection linear model for males, from the mean age minus 30 years to the mean age plus 30 years.

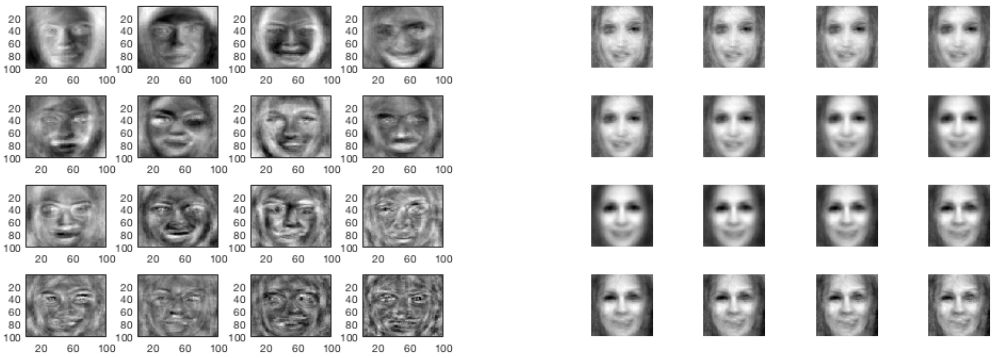

Figure 14: Left: The chosen eigenfaces from the Forward Feature Selection ordered from the eigenface with the largest negative weight to the eigenface with the largest positive weight. Right: A linearly varying age based on the linear extension of the Forward Feature Selection linear model for females, from the mean age minus 30 years to the mean age plus 30 years.

4.2 Future prospects within the scope of this project

Since the project phase of this course was *"only"* 2.5 ECTS points of time, we have not had the time to go into detail with as many aspects of the project as we would have liked. Hence there are various parts of the project which could have benefited from additional work. Some of which will be discussed in this subsection.

4.2.1 More appropriate dissimilarity measure for Lasso

At the moment, we use the MSE as dissimilarity measure for choosing the right model in the Lasso regression model. However, other dissimilarity measures should be considered and an example of such a dissimilarity measure could be to count the number observations a given model fails to predict within a given range of a target.

4.2.2 A professional data-set

Since the data used for our modelling is taken from an open database where the images are mainly taken by paparazzi or screen-shots from movies, the lighting of the images is not the same, photos are taken from different angles and the facial expressions vary a lot. Hence it would have been beneficial to create a professional data set made for such an experiment. It would have been beneficial to use a data set which did not contain familiar faces in the form of actors and actresses as this inevitably introduces some bias to how our subjects perceived age.

5 Conclusion

From our cognitive experiment in which we tried to model how humans perceive age by looking at facial characteristics, we found some of the general characteristics that matter when humans (in this case our subjects) perceive age. Examining the results from our male experiment by looking at the right pane of Fig. 9, we found that some of the facial characteristics which our subjects tended to perceive as characteristics of young males include smaller and more narrow noses, smaller probability of wearing glasses as well less noise around the mouth. On the contrary, elderly males tend to have more noise around the mouth which might be because our subjects associate beards and wrinkles as a characteristic of more mature males. Furthermore, as human noses continue to grow throughout a lifetime it makes sense that we see a correlation between bigger noses and perceived age by our subjects. Another clear trend was the fact that our subjects

perceived the natural reduction of sight throughout a lifetime to be correlated with maturity and an increase in age, which can be seen by the appearance of glasses in the right pane of Fig. 9. The same trends goes for the linear evolution of varying age of females in the right pane of Fig. 12 where our subjects again perceived a correlation between loss/reduction of sight and age, as well as with nose enlargement and age, and furthermore, found make-up and wrinkles to be correlated to maturity and age which is seen on Fig. 12 as noise and darker pixels around the eyes. However, the noise around the mouth typically won't resemble beard for women but rather wrinkles.
By assuming that we could establish a linear relationship between age and input in Section 2, we restricted the model complexity a lot. Additionally by adding regularization (Lasso) as we sought a low-dimensional model, we restricted the model complexity further introducing quite a lot of bias to our model and thus reduced the variance as seen in Fig. 5. As we found that the constructed models for each gender with Lasso chose to predict around the mean, we decided to try and create a low-dimensional model with a different approach, i.e., by sequential feature selection. With the models made by sequential feature selection it was found that the models were quite similar, choosing some of the same eigenfaces to be non-zero and additionally found the same characteristics as of the first models to be important. Examples include the comments on facial hair and glasses, as well as the grainy images for women.
Moreover, we found that even though we got some interesting insights, it was hard to predict the perceived facial ages by our subjects accurately with a linear model and hence increasing the variance (model complexity) would have been appropriate if we were to make further progress towards making a model being able to accurately predict how humans perceive age based on facial characteristics.

References

[1] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[2] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, July 2016.

[3] Michael J. Jones and Paul Viola. Dex: Deep expectation of apparent age from a single image. *International Journal of Computer Vision*, 2004.

[4] Tobias Andersen. Natural world statistics. University Lecture, 2017.

[5] Morten Mørup, Mikkel N. Schmidt, and Tue Herlau. *Introduction to Machine Learning and Data Mining*, volume 7. DTU, 2017.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

A A short note on Cross-validation

Cross-validation (CV) is a beautiful technique which can alleviate the pain of tuning hyperparameters when finding the optimal model and estimate the generalization error. CV was used for every model created in the code phase of this project. Therefore the procedure is outlined in pseudo-code in Algorithm 1[6].

Algorithm 1 K-Fold Cross-Validation for Model selection and $\hat{E}^{\text{gen}}_{\mathcal{M}}$ estimation

Require: K_1 folds in outer loop for estimation of the generalization error
Require: K_2 folds in inner loop for model selection
Require: S models to cross-validate: $\mathcal{M}_1, ..., \mathcal{M}_S$
Ensure: $\hat{E}^{\text{gen}}_{\mathcal{M}*}$
  for $i = 1, ..., K_1$ **do**
    Outer CV loop. The data set, $\mathcal{D}$ is split into K_1 folds
    The i'th split of $\mathcal{D}$ is $\mathcal{D}^{\text{par}}_i, \mathcal{D}^{\text{test}}_i$
    for $j = 1, ..., K_2$ **do**
      Inner CV loop doing K_2 splits for model selection testing S models
      The j'th split of $\mathcal{D}^{\text{par}}_i$ is $\mathcal{D}^{\text{train}}_j, \mathcal{D}^{\text{val}}_j$
      for $s = 1, ..., S$ **do**
        Train $\mathcal{M}_s$ on $\mathcal{D}^{\text{train}}_j$
        Let $E^{\text{val}}_{\mathcal{M}_{s,j}}$ be the *validation error* of the model $\mathcal{M}_s$ when it is *tested* on $\mathcal{D}^{\text{val}}_j$
      end for
    end for
    For each s compute $\hat{E}^{\text{gen}}_s = \sum_{j=1}^{K_2} \frac{|\mathcal{D}^{\text{val}}_j|}{\mathcal{D}^{\text{par}}_i} E^{\text{val}}_{\mathcal{M}_s,j}$
    Select the optimal model $\mathcal{M}^* = \mathcal{M}_{s^*}$ where $s^* = \text{argmin}_s \hat{E}^{\text{gen}}_s$
    Train $\mathcal{M}^*$ on $\mathcal{D}^{\text{par}}_i$
    Let E^{test}_i be the *test error* of the model $\mathcal{M}^*$ when it is tested on $\mathcal{D}^{\text{test}}_i$
  end for
  Compute the estimate of the generalization error: $\hat{E}^{\text{gen}} = \sum_i^{K_1} \frac{|\mathcal{D}^{\text{test}}_i|}{N} E^{\text{test}}_i$

[6]The idea of outlining the Algorithm in pseudo-code is due to [5]