Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.151

Distance-weighted discrimination of face images for gender classification

Mónica Benito[a], Eduardo García-Portugués[a]*, J. S. Marron[b] and Daniel Peña[a,c]

Received 15 June 2017; Accepted 22 June 2017

We illustrate the advantages of distance-weighted discrimination for classification and feature extraction in a high-dimension low sample size (HDLSS) situation. The HDLSS context is a gender classification problem of face images in which the dimension of the data is several orders of magnitude larger than the sample size. We compare distance-weighted discrimination with Fisher's linear discriminant, support vector machines and principal component analysis by exploring their classification interpretation through insightful *visuanimations* and by examining the classifiers' discriminant errors. This analysis enables us to make new contributions to the understanding of the drivers of human discrimination between men and women. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: distance-weighted discrimination; feature extraction; Fisher's linear discriminant; gender classification; HDLSS; support vector machines

1 Introduction

Image classification is one of the most important applications of image analysis and, especially during the last two decades, has received significant attention. In particular, automatic identification of human faces from a database of digital images has become increasingly important. Surveys in face recognition can be found in Samal & Iyengar (1992), Valentin et al. (1994) and Zhao et al. (2003), whereas Kawulok et al. (2016) present a compendium of recent research trends in the area. The process of facial recognition has been thoroughly investigated by cognitive psychologists (Bruce, 1988) and questions such as how the brain recognizes a face as male or female have been addressed (Burton et al., 1993), although the mechanisms involved in processing the visual information remain unclear (Wu & Huang, 1990).

In the face classification problem, a certain number of variables are measured in a sample of individuals, and the goal is to build a rule to classify a new face within given groups, for example, males and females. The first approach for face classification, due to Galton (1910), was to use certain characteristics of the population, such as angles and distances between facial landmarks. Later, other approaches based on anatomical considerations were developed for the classification of facial features (Bruce et al., 1992; Farkas et al., 1987; Gizatdinova & Surakka, 2010). However, combining these into a single dataset is highly problematic: different landmark points are used, and distinct measurements are made. A third approach has been based on the whole facial image, instead of upon such descriptive measurements. This approach is facilitated by a suitable database with similar viewpoint, pose and illumination conditions, for effective classification. An example of such a database is shown in Movie 1.

[a]Department of Statistics, Carlos III University of Madrid, 28903 Madrid, Spain
[b]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill 27514 NC, USA
[c]Institute of Financial Big Data, Carlos III University of Madrid, 28903 Madrid, Spain
*Email: edgarcia@est-econ.uc3m.es

(wileyonlinelibrary.com) DOI: 10.1002/sta4.151

Movie 1. Raw and registered images. The registered images were obtained by a generalized Procrustes analysis based on three landmarks: eyes (green) and nose (red). It shows that the registered images are shape comparable and ready for posterior statistical treatment. A blurring has been applied to preserve anonymity. Movie file is available as Supporting Information.

A typical first step in using full images for face recognition is to *rasterize* each image into a vector by, say, column concatenation, and set each entry of the vector as the grey level at a pixel. Let X be the matrix whose columns are these image vectors. Thus, X is $d \times n$, where n is the sample size and d is the dimension of the image vectors, that is, the number of pixels. The task of developing a rule to distinguish men from women is a binary classification problem that is often tackled by *linear classifiers*, that is, classifiers that find a hyperplane that separates the two classes by partitioning the data space. This is the case for all the classifiers considered in this paper. Turk & Pentland (1991) used principal component analysis (PCA) to obtain a reduced rank representation of the faces and used the minimum distance classification (i.e. the first nearest neighbour rule) in that eigenspace, to classify men and women. Belhumeur et al. (1997) applied Fisher's linear discriminant (FLD) function to this problem. They showed that FLD outperformed PCA, perhaps not surprisingly because the latter does not use class information at all. However, as the problem of face classification is a high-dimension low sample size (HDLSS) context, FLD encounters two problems. First, the traditional algorithm cannot be used directly because the estimated within-class scatter matrix is always singular. Second, the high-dimensional image vectors can lead to computational challenges. In order to avoid these difficulties, Yang & Yang (2003) first used PCA for dimensionality reduction and then applied FLD, hence using class information in the second step.

Support vector machines (SVM), originally developed by Vapnik (1995), are linear classifiers that belong to the class of maximum margin classifiers, that is, the ones that maximize the minimum distance between the two convex hulls of the data points from each class (Hastie et al., 2009). Marron et al. (2007) showed that for HDLSS data SVM has the problem that a large portion of the data are support vectors, that is, data points that lie on the margin boundaries.

Thus, when the observations are projected onto the normal vector of the hyperplane, many of the projections are identical – a property called *data piling* – resulting in reduced generalizability of the classifier. Ahn & Marron (2010) showed the existence of directions, where data from one class project to a single point, and from the other class project to another single point. Among such directions, the one with greatest distance between these points is called the *maximal data piling* (MDP) direction. An important precursor of the MDP direction is the *minimum projected kurtosis direction* of Peña & Prieto (2000, 2001). The formula for MDP is quite similar to that for FLD, with the pooled within-class covariance matrix replaced by the overall covariance matrix. Some algebra reveals that indeed in non-HDLSS situations, the MDP and FLD directions are the same, despite their apparently different algebraic form and quite different behaviour in HDLSS contexts. The data piling tendency of SVM in HDLSS contexts motivated Marron et al. (2007) to develop distance-weighted discrimination (DWD), which was shown (see also Hall et al., 2005) to be better than both SVM and FLD for HDLSS data. See Marron (2015) for an accessible review on DWD, and see Carmichael & Marron (2017) for a deep analysis of the relationship between SVM and MDP.

The main contribution of this paper is twofold. First, we do a head-to-head comparison between DWD and the more commonly used SVM by the use of insightful *visuanimations* (Genton et al., 2015) on the discriminant directions (Section 2.1) and by the examination of the classical discrimination errors on a separate testing dataset (Section 2.3). We complement this comparison with PCA (for the discriminant directions) and FLD (for the discrimination error rate) in order to show the better performance of DWD and SVM. Second, we present in Section 2.2 some novel approaches to deeply understand the canonical differences between male and female faces (which seems automatic by the human perceptual system) that stem from a careful analysis of the DWD classification rule.

2 Experimental results

In order to demonstrate the effectiveness of DWD for facial pattern recognition, we ran a series of experiments and compared our results with those obtained using PCA, FLD and SVM. The experiments were carried out using a training dataset of frontal view face images of former students from Carlos III University of Madrid. The images were recorded with a digital camera, with somewhat similar illumination conditions. The dataset is composed of $n = 108$ face images of size IJ, where $I = 248$ and $J = 186$. The number of men and women are $n_1 = 54$ and $n_2 = 54$, respectively. Rasterization of the images yields vectors in the $\mathbb{R}^d$ space, $d = IJ = 46{,}148$, whose entries are the grey-level values of all the pixels. To eliminate spurious effects due to the location of each face in the image, we registered them using landmarks. Three landmarks, representing the eyes and nose, were automatically selected, and optimal translations and rotations were then computed using a generalized Procrustes analysis, as described in Benito & Peña (2005). Movie 1 shows the results of this registration process. Note that in the raw images, the faces move around quite a bit, while they are much stable in the registered image, and thus are more suitable for addressing the classification challenge.

2.1 Face classification by DWD, PCA and SVM

We first ran an experiment in which the discriminant vectors for DWD, PCA and SVM are computed to separate the training dataset into two groups, men and women. Visualization of these directions gives valuable insights on the discrimination ability of each classifier. Movie 2 gives a graphical summary of the classification performance of DWD (2a), PCA (2b) and SVM (2c) by means of a comparable march along the image projections in the three discrimination directions (green vertical bar, right panels) and the corresponding projected images (left panels). The horizontal span of the right plots was set to 125% of the range of the data for the sake of a better visualization.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.151

The right panel of Movie 2a shows the projections of each data vector onto the DWD direction, where the women are red plus signs, and the men are blue circles. In this and the next plots, the heights of the symbols reflect order in the dataset (for visual separation) and the curves are kernel density estimates, with black showing the overall density and the red and blue showing the corresponding, group-size adjusted, subdensities for the two subpopulations. Women clearly lie to the left and men to the right. Near the middle, the faces appear quite androgynous, depicting a continuum separation of genders. Note that none of the faces in the left panel are members of the dataset, but instead are reconstructions of points in the image space that lie along the DWD direction. DWD was implemented using $C = 100$ as the penalty parameter based on the suggestions in Marron et al. (2007).

Movie 2b shows the projections of the image vectors onto the PC1 (first principal component) direction. While there is some grouping of men to the right and women to the left, there is also quite a lot of overlap, especially compared with Movie 2a. The inefficiency of PCA at separation of men and women is also clear from the sequence of faces in the left panel, which are less distinctly recognizable as men and women. This is not surprising: PCA does not make use of class information at all, but instead targets only maximal variation in the data. This is apparent from the span of the projections, from -50 to 40 for PCA versus -17 to 12 for DWD. A close look at the sequence of images shows a strong difference in the lower background part for PCA – related with the absence/presence of long hair – which is much more constant in the DWD direction. In fact, this shows that the male–female difference is an important component of the single largest mode of variation in the data.

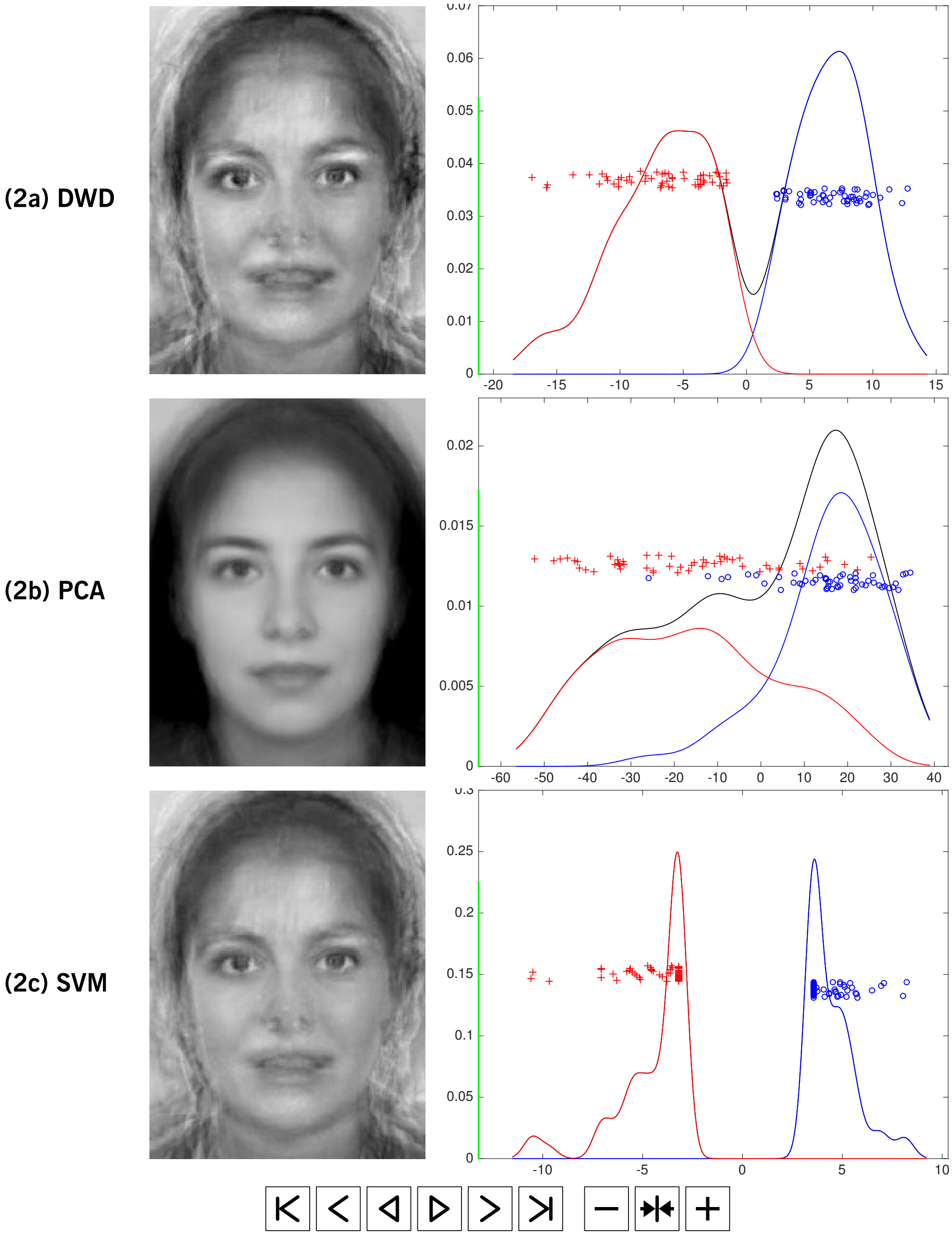
The projections for SVM are shown in Movie 2c. This gives a much better separation of the classes than PCA. However, care is needed here, because it also shows the problem of data piling discussed in the introduction. In particular, many red points are piled at the right end of the red distribution, and blue points are piled at the left end of the blue distribution. Marron et al. (2007) showed that this results in overfitting by the SVM. In contrast, Movie 2a shows a more efficient discrimination of the two groups and no data piling, which is the result of each data point playing a role in finding the discriminant direction in the data. An additional appealing feature of the DWD projected class distributions is their Gaussian shape, a critical property in the good performance of DWD in applications such as microarray batch adjustment (Benito et al., 2004). It is also interesting to compare both methods in terms of the gap between the modes of the subdensities. The gap for SVM in Movie 2c goes, roughly, from -3 to 3, while the gap for DWD is -5 to 7. Bigger gaps are better here, because they indicate real population differences, instead of spurious sampling artifacts driven by overfitting. SVM was fit using $C = 1000$ following Gunn (1998)'s recommendation.

Finally, note that although the SVM faces in the left panel are recognizable as female in the left and male in the right, the DWD images in Movie 2a are sharper and give an overall impression of a better male–female separation. In particular, in DWD faces the outline of the faces is more distinct, the forehead seems more rounded for women and squared for men, the eyebrows are hairier in men, the women's eyes are more open and women smile a little more. All of these ideas suggest that when a new face image is classified the performance of DWD will be superior to the SVM classification. This will in fact be seen in Section 2.3.

2.2 Face recognition insights from DWD

Figure 1 gives insights into face recognition through a new image aimed to highlight the facial aspects that drive gender classification. The entries of the DWD vector (i.e. loadings) are coded with colours, white for 0, darker shades of red for stronger negative (associated with women) and darker shades of blue for stronger positive (associated with men). These contrasts on the DWD vector show that the more discriminant facial aspects are the eyebrows, eyes, nose, lips, chin and the outline of the head (in which the presence of short/long hair plays a key role).

We focus next on these facial aspects by repeating the DWD analysis for the same dataset, but this time restricting the input vectors to the most relevant features. In particular, we use only pixels that lie within rectangles chosen as


(2a) DWD

(2b) PCA

(2c) SVM

Movie 2. Projection of the data image vectors onto the distance-weighted discrimination (DWD), principal component analysis (PCA) and support vector machines (SVM) discriminant directions, showing the continuous separation between women (red plusses) and men (blue circles). Green line indicates a march along the projections in the discriminant direction, with corresponding reconstructed images on the left. DWD shows a good separation of women and men, going through androgyny. PCA has some discrimination capability, but not as strong as DWD. SVM shows good male–female separation although some facial features are more clear in the DWD plot, and SVM overfits in terms of data piling. Movie files and a static comparison figure are available as Supporting Information.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.151

in Benito & Peña (2005), which correspond to the eyebrows, eyes, nose, lips and chin. Figure 2 shows the result of projecting the full image vectors into each of the DWD directions that were independently obtained for each of these face regions. The first direction is computed just using the eyebrows; the second one is obtained by only considering the eyes; and the third, fourth and fifth are those corresponding to the nose, lips and chin, respectively.

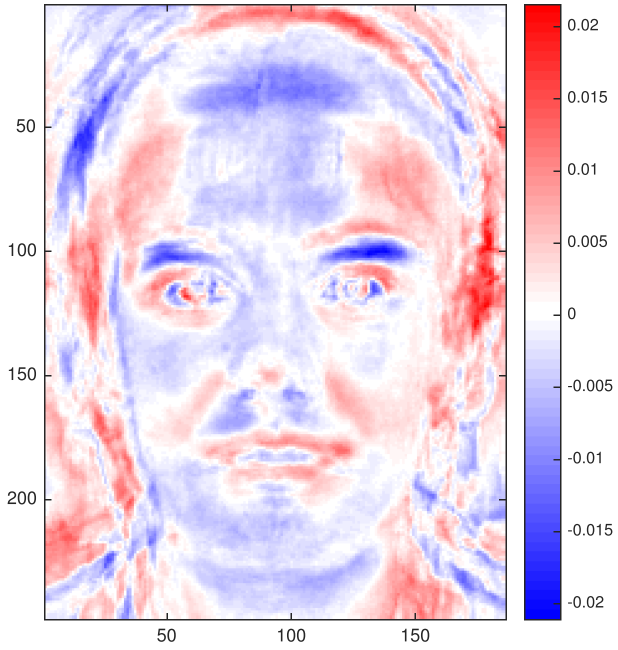


Figure 1. Highlighting of discriminant pixels in the gender classification problem using distance-weighted discrimination (DWD) loadings. White codes no difference and shade of red (blue) shows magnitude of negative (positive) entry in the DWD vector, associated with females' (males') features. It shows that head shape, together with pixels in the eyebrows, eyes, nose, lips and chin are important to the discrimination.

The univariate distributions shown in the diagonal plots of Figure 2 indicate how well each region classifies men versus women using only pixels in that region. Note that both the eyebrows' and eyes' pixels alone produce a complete separation of the subpopulations. For the other regions, we observe some small overlap. The off-diagonal plots are the corresponding pairwise scatterplots: in the second column of the first row, the two-dimensional projection is of the full image vectors on the DWD directions found by using the eyebrows and eyes, respectively. Note that there is even better discrimination in the two dimensional plot than in either one dimensional version. Also insightful are the black lines indicating the two DWD directions, showing that they are nearly orthogonal, that is, that each component brings essentially independent information to the discrimination method. Thus, using both together improves generalization ability of the combined classifier. Similar orthogonality holds for the other pairs of directions as well, suggesting that the different facial features tend to work independently in assisting with the overall good performance of DWD. Although none of the nose, lips and chin give perfect separation, they all provide useful information to the overall discrimination. This is consistent with the much larger gap between the combined DWD subpopulations in Movie 2a, relative to the gaps shown in Figure 2. Overall, we now have several new DWD directions each of which discriminate some part of the population well, but not everyone. In particular, some women fall into that category because of their smile, others because of their eyebrows.

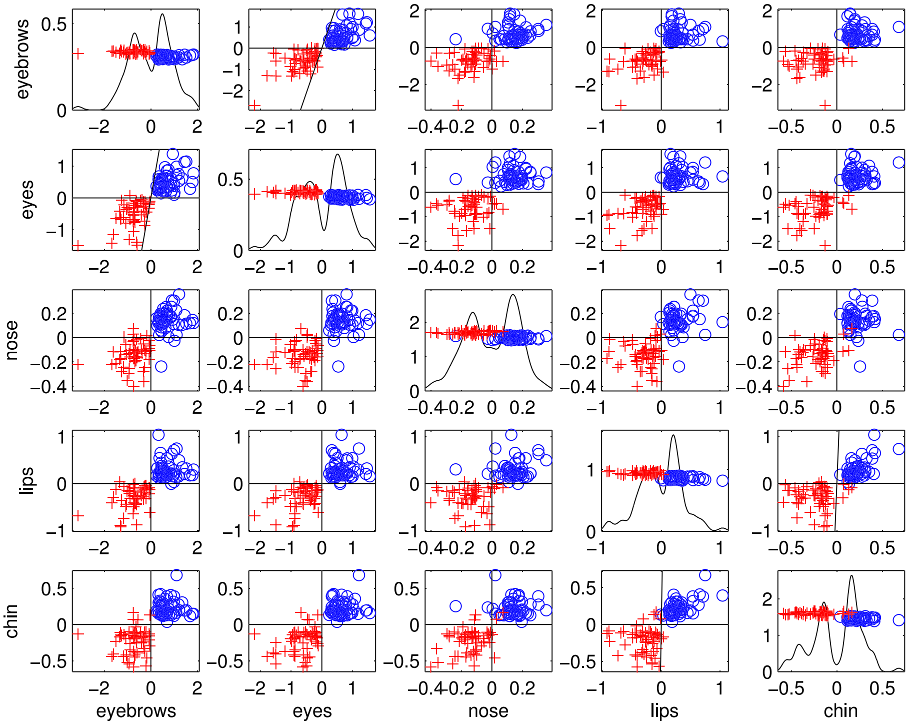


Figure 2. Projection of the full image vectors onto the distance-weighted discrimination (DWD) directions defined by the most important regions. The plots along the diagonal show the one-dimensional projection onto each direction. The off-diagonal plots are two-dimensional projections by pairs. The symbols and colours are as in Movie 2, and black lines represent the two DWD directions. It shows that eyebrows and eyes give the most male–female discrimination, and others contribute weaker but nearly independent additional information.

Table I. Classification error on the testing dataset for distance-weighted discrimination (DWD), support vector machines (SVM) and Fisher's linear discriminant (FLD).

Classification matrix	DWD	SVM	FLD
Men classified as men	58	57	56
Women classified as women	40	38	15
Men classified as women	0	1	2
Women classified as men	2	4	27
Error rate	2%	5%	29%

Note: It shows superior performance of DWD, which is consistent with each of the aforementioned results.

2.3 Classification errors for DWD, SVM and FLD

The last experiment evaluates the performance of DWD, SVM and FLD (with Moore–Penrose pseudo-inverse for the covariance matrix) in the classification of male and female faces. To that aim, we classify an independent testing dataset with the aforementioned classifiers, after training them with the above studied $n = 108$ images. The testing dataset consists of 100 images, 58 men and 42 women, which were obtained under similar illumination conditions and viewpoint as the training data. Similarly as for the training dataset, there was a preprocessing step for registering the images using a generalized Procrustes analysis.

Table I shows the classification error rates (defined as number misclassified, divided by the total) in the classification of the testing dataset when the full image vectors are considered. The results are quite consistent with what we have shown in the aforementioned sections: SVM is much better than FLD, and DWD is substantially better than SVM.

3 Conclusions

We have illustrated the benefits of DWD over SVM, FLD and PCA in a challenging HDLSS discrimination problem: the classification of male and female faces from a dataset of facial images. When performing classification, SVM exhibited signs of overfitting in terms of data piling, FLD was significantly worse than SVM and DWD showed the best performance, both in terms of classification error and in terms of a wider, Gaussian-shaped and interpretable separation of subpopulations. PCA showed some discrimination ability, but it was clearly overcome by any of the aforementioned classifiers.

The classification by DWD automatically identified key features for male–female discrimination that contributed nearly orthogonal DWD directions to the classifier: eyebrows, eyes, nose, lips, chin and the outline of the head. Indeed, each of the regions of pixels associated with eyebrows and eyes delivered perfect DWD classification in the training dataset. Careful interpretation of the DWD separating vector showed its success in capturing well-known physical differences between male and female faces and also revealed other characteristic features of male and female facial images that were less immediate, such as the differences of smiles (women tend to smile more) and eyes (women's eyes are more open).

Acknowledgements

J. S. Marron acknowledges support from the National Science Foundation under grant no. 633074, E. García-Portugués from project MTM2016-76969-P, and D. Peña from project ECO2015-66593-P, both projects funded by the Spanish Ministry of Economy, Industry and Competitiveness, and ERDF.

References

Ahn, J & Marron, JS (2010), 'The maximal data piling direction for discrimination', *Biometrika*, **97**(1), 254–259.

Belhumeur, PN, Hespanha, JP & Kriegman, DJ (1997), 'Eigenfaces vs. Fisherfaces: recognition using class specific linear projection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(7), 711–720.

Benito, M, Parker, J, Du, Q, Wu, J, Xiang, D, Perou, CM & Marron, JS (2004), 'Adjustment of systematic microarray data biases', *Bioinformatics*, **20**(1), 105–114.

Benito, M & Peña, D (2005), 'A fast approach for dimensionality reduction with image data', *Pattern Recognition*, **38**(12), 2400–2408.

Bruce, V (1988), *Recognising Faces*, Essays in cognitive psychology, *Lawrence Erlbaum Associates, Inc.*, Hillsdale, NJ.

Bruce, V, Burton, AM & Craw, I (1992), 'Modelling face recognition', *Philosophical Transactions of the Royal Society, Series B*, **335**(1273), 121–127.

Burton, AM, Bruce, V & Dench, N (1993), 'What's the difference between men and women? Evidence from facial measurement', *Perception*, **22**(2), 153–176.

Carmichael, I & Marron, J (2017), 'Geometric insights into support vector machine behavior using the KKT conditions', *arXiv:1704.00767*.

Farkas, LG, Munro, IR & Kolar, J (1987), Relationships of Profile Segment Inclinations in the Faces of North American Caucasians. In Farkas, LG & Munro, IR (eds.), Anthropometric Facial Proportions in Medicine, Charles C. Thomas, Springfield, IL, pp. 67–78.

Galton, F (1910), 'Numeralized profiles for classification and recognition', *Nature*, **83**, 127–130.

Genton, MG, Castruccio, S, Crippa, P, Dutta, S, Huser, R, Sun, Y & Vettori, S (2015), 'Visuanimation in statistics', *Stat*, **4**(1), 81–96.

Gizatdinova, Y & Surakka, V (2010), 'Automatic edge-based localization of facial features from images with complex facial expressions', *Pattern Recognition Letters*, **31**(15), 2436–2446.

Gunn, SR (1998), 'Support vector machines for classification and regression', *Image Speech and Intelligent Systems Research Group, University of Southampton*, **14**, 85–86.

Hall, P, Marron, JS & Neeman, A (2005), 'Geometric representation of high dimension, low sample size data', *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **67**(3), 427–444.

Hastie, T, Tibshirani, R & Friedman, J (2009), *The Elements of Statistical Learning*, Second edition Springer Series in Statistics, *Springer*, New York.

Kawulok, M, Celebi, ME & Smolka, B (2016), *Advances in Face Detection and Facial Image Analysis*, *Springer Publishing Company, Inc.*

Marron, JS (2015), 'Distance-weighted discrimination', *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**(2), 109–114.

Marron, JS, Todd, MJ & Ahn, J (2007), 'Distance-weighted discrimination', *Journal of the American Statistical Association*, **102**(480), 1267–1271.

Peña, D & Prieto, FJ (2000), 'The kurtosis coefficient and the linear discriminant function', *Statistics & Probability Letters*, **49**(3), 257–261.

Peña, D & Prieto, FJ (2001), 'Cluster identification using projections', *Journal of the American Statistical Association*, **96**(456), 1433–1445.

Samal, A & Iyengar, PA (1992), 'Automatic recognition and analysis of human faces and facial expressions: a survey', *Pattern Recognition*, **25**(1), 65–77.

Turk, M & Pentland, A (1991), 'Eigenfaces for recognition', *Journal of Cognitive Neuroscience*, **3**(1), 71–86.

Valentin, D, Abdi, H, O'Toole, AJ & Cottrell, GW (1994), 'Connectionist models of face processing: a survey', *Pattern Recognition*, **27**(9), 1209–1230.

Vapnik, VN (1995), *The Nature of Statistical Learning Theory*, *Springer-Verlag*, New York.

Wu, CJ & Huang, JS (1990), 'Human face profile recognition by computer', *Pattern Recognition*, **23**(3), 255–259.

Yang, J & Yang, J (2003), 'Why can LDA be performed in PCA transformed space?', *Pattern Recognition*, **36**(2), 563–566.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.151

The ISI's Journal for the Rapid
Dissemination of Statistics Research

Zhao, W, Chellappa, R, Phillips, PJ & Rosenfeld, A (2003), 'Face recognition: a literature survey', *ACM Computing Surveys*, **35**(4), 399–458.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.