Factor Analysis for Background Suppression

Kyungim Baek and Bruce A. Draper
Computer Science Department
Colorado State University, Fort Collins, CO, 80523, USA
{baek, draper}@cs.colostate.edu

Abstract

Factor analysis (FA) is a statistical technique similar to principal component analysis (PCA) for explaining the variance in a data set in terms of underlying linear factors. Unlike PCA, however, FA has not been widely exploited for face or object recognition. This paper explains the differences between PCA and FA, and confirms that PCA outperforms FA in a standard face recognition task. However, because FA estimates the unique variance independently for every pixel, we show that the variance estimates from FA can be used to automatically detect and suppress background pixels prior to the application of PCA, and thereby improve the performance of PCA-based object recognition systems.

1. Introduction

Factor analysis (FA [2]) is a popular statistical method for data analysis. Its goal is to explain the correlations among a set of observed variables in terms of a smaller number of relevant and meaningful factors. FA was originally developed in social sciences and psychology, where the major use of FA is to develop objective tests for measurement of qualities such as personality and intelligence [12].

In computer vision, a similar multivariate analysis method called principal component analysis (PCA [8, 14]) has been used extensively. FA and PCA have similar goals – to express data sets in terms of linear combinations of a small set of factors – yet there is a crucial difference between them. FA is concerned with the common variance accounted for by linear factors, and excludes the noise due to unique (i.e. pixel-wise) variance. PCA, on the other hand, finds the basis vectors that optimally explain the total variance; the unique variance is not computed or accounted for separately.

In general, PCA has been used for face and object recognition to the exclusion of FA. The exceptions are systems that fit mixtures of local linear models rather than a single global model, e.g. [5, 7]. These systems exploit an Expectation-Maximization (EM) [4] algorithm for fitting mixtures of factor analyzers to datasets [5, 6]. Recent work by Tipping and Bishop, however, derives a probabilistic model for PCA and extends it to a mixture of local PCA models [13]. As a result, there are now methods to compute mixtures of PCA models as well as mixtures of FA models.

Is there any reason, then, to prefer FA over PCA for face or object recognition? We can find no comparisons in the literature, but data in this paper reflects our experience that PCA outperforms FA on standard face recognition tasks. There are circumstances, however, under which a combination of FA and PCA outperforms either alone. The key observation lies in the analysis by Tipping and Bishop, who show that the difference between FA and PCA lies in the residual error model [13]. PCA assumes that all variances not accounted for by the eigenvectors is drawn from a single zero-mean Gaussian distribution with standard deviation σ. FA, on the other hand, fits a unique standard deviation σ_i to every pixel. This implies that FA may have an advantage over PCA in circumstances where the foreground and background pixels are not separated *a-priori*, since background pixels should receive higher unique variances.

To test this hypothesis, we apply FA to two different versions of the FERET face database [9], with different amounts of background. We find that the resulting unique variance image separates background from foreground pixels quite well. Unfortunately, the linear factors computed by FA do not outperform the PCA basis vectors in a typical face recognition system, but the σ_i values can be used to inversely weight pixels prior to applying PCA. This suppresses background pixels relative to foreground pixels, and improves the performance of PCA.

2. Representing Data as Linear Combination of Components

In pattern recognition, the input patterns are often defined in a large dimensional space where the meaningful features are obscured by noise and complicated dependencies between variables. Therefore, it may be important to reduce the dimensionality of the input data by projecting it onto a smaller and more manageable space in which the relevant features are more explicit. Two useful techniques for dimensionality reduction are factor analysis and principal component analysis.

2.1. Factor Analysis

In FA, a p-dimensional, mean centered observed vector $\mathbf{x} = (x_1, \ldots, x_p)'$ is decomposed into a vector $\mathbf{f} = (f_1, \ldots, f_q)'$ of q latent variables (*factors*), and the vector $\mathbf{u}$ of p independent disturbance variables [2]:

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{u}$$

where $\mathbf{\Lambda}$ is called a *factor loading matrix*, whose elements λ_{ij} determine the importance of factor f_j to x_i. The disturbance term $\mathbf{u}$ accounts for independent noise in each element of $\mathbf{x}$.

In this model, it is assumed that the underlying distribution of $\mathbf{f}$ is $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{u}$ follows $\mathcal{N}(\mathbf{0}, \mathbf{\Psi})$, where $\mathbf{\Psi} = diag(\psi_1, \ldots, \psi_p)$. Since it is assumed that the u_i's are uncorrelated, the x_i's are conditionally uncorrelated given $\mathbf{f}$. Also, $cov(\mathbf{x}) = \mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$, and the goal of the FA model is to determine the $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ that best explain the covariance structure of $\mathbf{x}$ [2, 6].

An EM algorithm for a maximum likelihood estimation of FA was proposed in [10] and reviewed in [6] and [11]. Let $\mathbf{X}$ be a data matrix whose rows are n observed sample vectors. Then, the expected log-likelihood for FA is

$$\begin{aligned} \mathcal{L} &= log[\prod_{i=1}^{n}(2\pi)^{-p/2}|\mathbf{\Sigma}|^{-1/2}exp\{-\frac{1}{2}\mathbf{x}_i'\mathbf{\Sigma}^{-1}\mathbf{x}_i\}] \\ &= -\frac{np}{2}log(2\pi) - \frac{n}{2}log|\mathbf{\Sigma}| - \frac{n}{2}trace(\mathbf{C_x\Sigma}^{-1}) \end{aligned}$$

where $\mathbf{C_x} = \mathbf{X}'\mathbf{X}/n$. The EM algorithm maximizes $\mathcal{L}$ by iterating through two steps: In the *E-step*, $E(\mathbf{f}|\mathbf{x}_i)$ and $E(\mathbf{ff}'|\mathbf{x}_i)$, $i = 1, \ldots, n$, are computed using current $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. Then, $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are updated using the newly computed $E(\mathbf{f}|\mathbf{x}_i)$ and $E(\mathbf{ff}'|\mathbf{x}_i)$ in the *M-step* (for the formula, see [6]).

2.2. Principal Component Analysis

PCA is another multivariate analysis method, which has been commonly used in pattern recognition. In PCA, the input vector $\mathbf{x}$ is represented by a linear combination of *principal components*, z_i, defined in the q-dimensional space spanned by the q principal eigenvectors of sample covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{x} = \mathbf{\Upsilon z} \qquad (1)$$

where $\mathbf{\Upsilon}$ is a $p \times q$ orthogonal matrix whose columns are the q principal eigenvectors of $\mathbf{\Sigma}$ and $\mathbf{z} = (z_1, \ldots, z_q)'$, which is computed by $\mathbf{z} = \mathbf{\Upsilon}'\mathbf{x}$. This mapping of $\mathbf{x}$ to a lower dimensional representation $\mathbf{z}$ is optimal in the mean squared error sense. That is, the inverse mapping of $\mathbf{z}$ back into $\mathbf{x}$ by equation (1) has minimum reconstruction error. In fact, if $\mathbf{\Upsilon}$ includes all the eigenvectors with non-zero eigenvalues, the inverse mapping is lossless.

3. Background Suppression by Inverse Variance Weighting

In FA, the variance of a variable is split into two parts:

$$var(x_i) = \sigma_i^2 = \textstyle\sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i$$

The first term, the sum of squared factor loadings across factors, is called the common variance or *communality*, which is the variance accounted for by the factors. The second term, ψ_i is called the *specific* or *unique* variance of x_i, which is independent to that of the other variables.

FA is concerned with the common variance and excludes the unique variance. In this study, however, we are more interested in the latter. The unique variances (i.e. the diagonal elements of $\mathbf{\Psi}$) determine how much of the variability in each input dimension is not attributable to linear factors. Therefore, when the input data is a set of two-dimensional images, it models the independent pixel noise in the dataset. We assume that non-linear variances are attributable to backgrounds since this model is applied to a set of independently collected images rather than a video sequence in which same backgrounds appear repeatedly. Each pixel in the images is inversely weighted by its unique standard deviation as estimated by FA.

4. Experimental Results

In this study, we advocate weighting source pixels by the inverse of their unique standard deviation, in order to suppress background pixels relative to foreground pixels. The recognition algorithm applied after weighting is PCA followed by a nearest-neighbor classifier, as in [9]. Unlike FA, PCA assumes that the noise is spherical and determines the principal components which can account for the total (both common and unique) variance. Therefore, PCA is highly sensitive to variation in pixel noise and the presence of background. To demonstrate the use of FA in suppressing background pixels, we weight pixels prior to PCA and compare the recognition performance to that of PCA performed on an unweighted dataset. A comparison between FA and PCA is also made to confirm that FA alone does not perform as well as PCA or weighted PCA (WPCA).

4.1. FERET Database

The FERET face recognition database is a set of face images collected by NIST from 1993 to 1997[1]. In this study, we used head-on images only, with 1,196 gallery images, 501 training images, and four different sets of probe images. Using the terminology in [9], the *fb* probe set contains 1,195 images of subjects taken at the same time as the gallery images. The only difference is that the subjects were told to assume a different facial expression than in the gallery images. The *duplicate I* probe set contains 722 images of subjects taken between one minute and 1,031 days after the gallery image was taken. The *duplicate II* probe set is a subset of 234 duplicate I probes, where the probe image is taken at least 18 months after the gallery image. Finally, the *fc* probe set contains 194 images of subjects under significantly different lighting conditions.

The images in this study have been cropped to two different sizes: a standard size of 150×130 pixels (as in [9]), and 200×170 pixels. The individual images are standardized to have zero mean and unit standard deviation. The left column of Figure 1 shows the images of a person from each dataset. Obviously, the larger image includes more background, particularly hair and clothes.

To run FA, the 150×130 images were scaled down to 24×21 pixels, while the 200×170 images were scaled to 25×21 pixels. The matrix Ψ in computed by FA on 1,301 training and gallery images. The diagonal elements form the variance image shown in the right column of Figure 1.

4.2. Recognition Results

As we can see in the variance images in Figure 1, pixels outside the face and around the extreme points of some facial structures – eyebrows, nose and mouth – vary a lot across the dataset, while areas of facial skin have lower unique variance. This is particularly true in the bottom right image, where the background is captured quite well by the unique variances. The standard deviation of each pixel is inversely multiplied to images in the dataset and PCA is performed on the weighted images (WPCA).

The results of comparing WPCA, PCA, and FA are given in Table 1 and Table 2 for the small and large FERET images, respectively. We can see that PCA always outperforms FA. In the comparison, we used 200 factors and the first 201 principal components to make the number of parameters in FA and PCA the same. The factor scores were computed as expected values conditioned on the observation, as in [11].

The weighting process does not help for the smaller images, since they do not include much background to suppress. Most of the unique variance captures internal variations around the facial structures. Statistically[2], there is

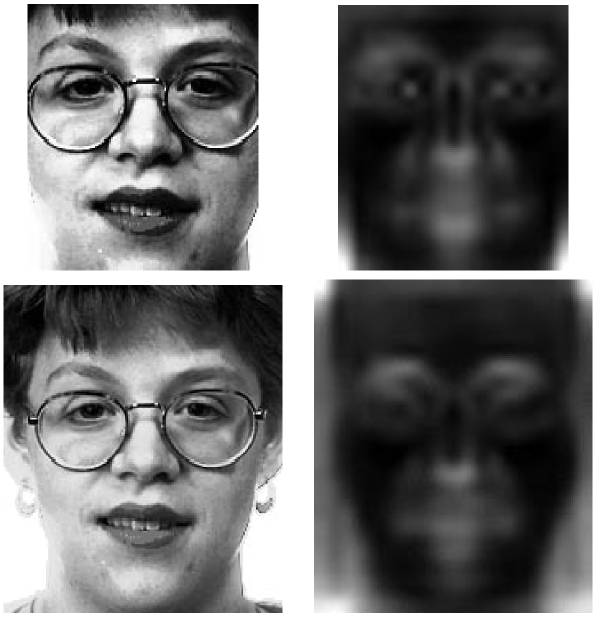

Figure 1. The left column shows an example from the FERET database cropped into two different sizes. On the right, the variance map of the datasets of smaller sized images (top) and larger sized images (bottom) computed by applying FA to combined set of training and gallery images from each dataset.

	PCA	WPCA	FA
fb	1015 (84.93%)	989 (82.76%)	725 (60.67%)
dup I	281 (38.92%)	278 (38.50%)	157 (21.75%)
dup II	36 (15.38%)	36 (15.38%)	15 (6.40%)
fc	63 (32.47%)	67 (34.54%)	9 (4.64%)

Table 1. Performance of PCA, WPCA, and FA on different probe sets. The original image size of the dataset is 150 x 130 pixels.

	PCA	WPCA	FA
fb	1046 (87.53%)	1087 (90.96%)	843 (70.54%)
dup I	285 (39.47%)	308 (42.66%)	186 (25.76%)
dup II	30 (12.82%)	40 (17.09%)	21 (8.97%)
fc	108 (55.67%)	93 (47.94%)	37 (19.07%)

Table 2. Performance of PCA, WPCA, and FA on different probe sets. The original image size of the dataset is 200 x 170 pixels.

[1] http://www.itl.nist.gov/iad/humanid/feret/

[2] To test statistical significance, we apply McNemar's pairwise-

no significant difference in performance between the two methods, except on the fb probe set where PCA outperforms WPCA. This makes sense, since the fb images were taken immediately after the gallery images. The backgrounds (e.g. hair, clothes) are therefore the same for each individual in the gallery and fb data sets, so the background represents useful information rather than noise. In all instances where the background is different between the probe and gallery, WPCA performs as well as PCA.

However, with larger images that include more background, WPCA outperforms PCA. For the fb, dup I, and dup II probe sets, WPCA performs significantly better than PCA with over 99% confidence. Compared to the results on the set of smaller images shown in Table 1, this shows that the proposed weighting process, using the unique variances captured by FA, helps suppress the background. The only exception is the fc probe set.

It can be argued that the proposed method seems to have an inherent contradiction when it is applied to face recognition as described in this section. FA captures high variance in pixels around facial structures (Figure 1), which are often considered prominent features in face recognition, and the proposed process seems to suppress influences of those features. The results in Table 1, however, show that suppressing these pixels does not drop the performance; there is no significant difference between PCA and WPCA. It may be that movable features are less predictable than fixed features that are not suppressed, or it may be because face recognition is more sensitive to holistic or configural properties than localized features [3]. Indeed, localized features may be more important for analysing facial expressions [1] than reconizing individual identities.

We cannot explain the anomalous result on the fc probe set. The fc probe set is the smallest probe set, and the lighting for the fc images is from a different direction than in the other three probe sets. The fc images are also darker than the other images, so there is less dynamic range in the source pixels. We do not yet understand why either of these factors should differentially effect WPCA vs PCA.

5. Conclusion

A method for automatically suppressing backgrounds prior to face or object recognition is presented. This technique has the property that it improves recognition rates when background is present in the images, but doesn't significantly reduce performance when no or little background is present. The WPCA technique is also well grounded in terms of classical statistics.

Although general and well-grounded, the applications of this technique to face recognition are limited. The shape of a human face is well known, and background can be eliminated through a simple mask. In other object recognition tasks, however, objects may be registered automatically. Under these circumstances, a technique for automatically detecting and suppressing background pixels may be critical.

References

[1] M. Bartlett, G. Donato, J. Movellan, J. Hager, P. Ekman, and T. Sejnowski. Image representations for facial expression coding. In S. Solla, T. Leen, and K. Mueller, editors, *Advances in Neural Information Processing Systems*, pages 886–892. MIT Press, Cambridge, MA, 2000.

[2] A. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications*. John Wiley & Sons, Inc., New York, 1994.

[3] I. Biederman and P. Kalocsai. Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society: Biological Sciences*, 352:1203–1219, 1997.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–39, 1977.

[5] B. Frey, A. Colmenarez, and T. Huang. Mixtures of local linear subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

[6] Z. Ghahramani and G. Hinton. The em algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, 1997.

[7] G. Hinton, M. Revow, and P. Dayan. Recognizing handwritten ditits using mixtures of linear models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 1015–1022. MIT Press, 1995.

[8] M. Kirby and L. Sirovich. Applications of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[9] H. Moon and J. Phillips. Analysis of pca-based face recognition algorithms. In K. Boyer and J. Phillips, editors, *Empirical Evaluation Techniques in Computer Vision*. IEEE Computer Society Press, Los Alamitos, CA, 1998.

[10] D. Rubin and D. Thayer. Em algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[11] L. Saul and M. Rahim. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 8(2):115–125, 1999.

[12] B. Tabachnick and L. Fidell. *Using Multivariate Statistics*. Allyn & Bacon, Inc., Boston, 2000.

[13] M. Tipping and C. Bishop. Mixtures of principal component analyzers. *Neural Computation*, 11(2), 1999.

[14] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

difference test.