A Family of Principal Component Analyses for Dealing with Outliers

J. Eugenio Iglesias[1], Marleen de Bruijne[12], Marco Loog[12], François Lauze[2], and Mads Nielsen[12]

[1] Department of Computer Science, University of Copenhagen, Denmark
[2] Nordic Bioscience A/S, Herlev, Denmark

Abstract. Principal Component Analysis (PCA) has been widely used for dimensionality reduction in shape and appearance modeling. There have been several attempts of making PCA robust against outliers. However, there are cases in which a small subset of samples may appear as outliers and still correspond to plausible data. The example of shapes corresponding to fractures when building a vertebra shape model is addressed in this study. In this case, the modeling of "outliers" is important, and it might be desirable not only not to disregard them, but even to enhance their importance.

A variation on PCA that deals naturally with the importance of outliers is presented in this paper. The technique is utilized for building a shape model of a vertebra, aiming at segmenting the spine out of lateral X-ray images. The results show that the algorithm can implement both an outlier-enhancing and a robust PCA. The former improves the segmentation performance in fractured vertebrae, while the latter does so in the unfractured ones.

1 Introduction

Principal Component Analysis (PCA) is a technique that simplifies data sets by reducing their dimensionality. It is an orthogonal linear transformation that spans a subspace which approximates the data optimally in a least-squares sense (Jolliffe 1986). This is accomplished by maximizing the variance of the transformed coordinates.

If the dimensionality of the data is to be reduced to N, an equivalent formulation of PCA is to find the N-set of orthornormal vectors, grouped in the $\mathbf{P}$ matrix, which minimizes the error made when reconstructing the original data points in the data set. The error is measured in a L_2 norm fashion:

$$C = \sum_{i=1}^{N} \|\mathbf{P}\mathbf{P}^t\mathbf{x}_i - \mathbf{x}_i\|^2 \tag{1}$$

where C is the cost, N is the number of training cases and $\mathbf{x}_i$ are the centered data vectors to approximate.

Least-squares is not robust when outliers are present in the dataset, as they can skew the result from the desired solution, leading to inflated error rates and distortions in statistical estimates (Hampel et al. 1986). Many authors, especially in the neural networks literature, have tried to reduce the impact of outliers on PCA by modifying the cost in

N. Ayache, S. Ourselin, A. Maeder (Eds.): MICCAI 2007, Part II, LNCS 4792, pp. 178–185, 2007.

Equation 1. Xu and Yuille 1991, for example, introduced a binary variable that is zero when a data sample is considered to be an outlier, one otherwise:

$$C_{Xu} = \sum_{i=1}^{N} \left[V_i \|\mathbf{P}\mathbf{P}^t \mathbf{x}_i - \mathbf{x}_i\|^2 + \eta(1 - V_i) \right]$$

where V_i is the set of binary variables. The term $\eta(1 - V_i)$ prevents the optimization from converging to the trivial solution $V_i = 0, \forall i$.

The main disadvantage of this method is that it either completely rejects or includes a sample. Moreover, a single noisy component in a sample vector can make it be discarded completely. Gabriel and Zamir 1979 proposed a similar method in which every single component of each data point is controlled by a coefficient, instead of having one binary weight per vector. Outliers are still considered, but have lower importance. Furthermore, undesired components (known as intra-sample outliers) can be downweighted without discarding the whole sample vector. Several other weighting and penalty terms have more recently been proposed (see for example De la Torre and Black 2003), but the formulation remains essentially the same.

All these approaches aim at reducing the effects of outliers in the model. In this paper, a family of PC analyses, capable of both increasing and decreasing the contribution of outliers in the model, is proposed. The algorithm was tested on a shape model applied to the segmentation of the vertebrae from lateral x-ray images from the spine. In this case, the fractured vertebrae may appear as outliers, but they are the most important cases and should be enhanced rather than disregarded.

2 Methods

2.1 Φ-PCA and α-PCA

In contrast to directly minimizing the squared data reconstruction error as in normal PCA (Equation 1), the presented Φ-PCA algorithm minimizes:

$$C = \sum_{i=1}^{N} \Phi[\|\mathbf{P}\mathbf{P}^t \mathbf{x}_i - \mathbf{x}_i\|^2] \tag{2}$$

where Φ is a twice-differentiable function such that $\Phi(x^2)$ is convex. The fact that Φ is twice-differentiable makes it possible to use Hessian-based methods in the optimization, providing quadratic convergence. The convexity requirement ensures the existance of just one minimum for C.

A simple and at the same time powerful form of the function is $\Phi(x) = x^\alpha$, with $\alpha > 0.5$ in order to accomplish the convexity condition. This special case will be called α-PCA. Large values for α ($\alpha > 1$, in general) will enhance the outliers, as they become more expensive compared to normal cases. In particular, $\alpha = \infty$ would lead to minimizing the L_∞ norm, and hence the maximum reconstruction error over shapes measured in a L_2 norm fashion. On the other hand, smaller values ($0.5 < \alpha < 1$) will have the opposite effect, leading to a more robust PCA. The case $\alpha = 0.5$ minimizes the L_1 norm. Finally $\alpha = 1$ amounts to standard PCA.

The data points $\mathbf{x}_i$ must be centered, which means that their mean must be subtracted from them: $x_i = \mathbf{s}_i - \mu$, where $\mathbf{s}_i$ represents the original, non-zero mean data samples. In the proposed algorithm, the "mean" is no longer the component-wise arithmetic mean of the data points as in the standard PCA, but the vector which minimizes (assuming M dimensions for the data points):

$$C_{\mu^\Phi} = \sum_{i=1}^{N} \Phi[\|\mu^\Phi - \mathbf{s}_i\|^2] = \sum_{i=1}^{N} \Phi\left[\sum_{t=1}^{M} (\mu_t^\Phi - s_{i_t})^2\right] \tag{3}$$

Once the $\mathbf{x}_i$ vectors have been calculated the Φ-PCA, which consists of searching the basis vectors $\mathbf{P}$ that minimize the cost function in Equation 2, can be performed. Numerical methods will be required in both minimizing C and C_μ, as there is no closed-form expression for μ^Φ or $\mathbf{P}$.

An important difference between standard and Φ-PCA is that, in the latter, the principal components have to be recalculated if the desired dimensionality changes. In standard PCA, on the other hand, the first N_1 principal components are common for two analyses with N_1 and N_2 components, assuming that $N_2 > N_1$.

Optimization of the Mean: The expressions for the gradient and the Hessian of the cost C_{μ^Φ} in Equation 3 are quite simple and fast to calculate. Using the component-wise arithmetic mean as initialization, Newton's method converges rapidly to the solution:

$$\mu_{n+1}^\Phi = \mu_n^\Phi - [HC_{\mu^\Phi}(\mu_n^\Phi)]^{-1} \nabla C_{\mu^\Phi}(\mu_n^\Phi),$$

where the gradient ∇C_{μ^Φ} is a column vector consisting of the first-order derivatives:

$$\frac{\partial C_{\mu^\Phi}}{\partial \mu_k^\Phi} = 2 \sum_{i=1}^{N} \Phi' \left[\sum_{l=1}^{M} (\mu_l^\Phi - s_{i_l})^2\right] (\mu_k^\Phi - s_{i_k}),$$

and the Hessian matrix H consists of the second-order derivatives:

$$H_{kk} = \frac{\partial^2 C_{\mu^\Phi}}{\partial \mu_k^{\Phi 2}} = 2 \sum_{i=1}^{N} \left\{ 2\Phi'' \left[\sum_{l=1}^{M} (\mu_l^\Phi - s_{i_l})^2\right] (\mu_k^\Phi - s_{i_k})^2 + \Phi' \left[\sum_{m=1}^{M} (\mu_m^\Phi - s_{i_m})^2\right] \right\}$$

$$H_{ku} = H_{uk} = \frac{\partial^2 C_{\mu^\Phi}}{\partial \mu_k^\Phi \partial \mu_u^\Phi} = 4 \sum_{i=1}^{N} \Phi'' \left[\sum_{l=1}^{M} (\mu_l^\Phi - s_{i_l})^2\right] (\mu_k^\Phi - s_{i_k})(\mu_u^\Phi - s_{i_u})$$

Optimization of the Basis: Once the mean has been subtracted from the data points, the cost C in Equation 2 must be minimized. The function has the interesting property that it reaches its global minimum for an orthonormal $\mathbf{P}$ matrix such that $\mathbf{P}^t\mathbf{P} = \mathbf{I}$. This makes it possible not to have to constrain $\mathbf{P}$ to accomplish this condition during the optimization, even if that implies that in general $\mathbf{PP}^t$ will not represent a projection matrix, and hence $\mathbf{PP}^t\mathbf{x}_i - \mathbf{x}_i$ does not express the reconstruction error any longer.

In this minimization problem, only the expression for the gradient is implemented, as the one for the Hessian matrix is too complex and its computation too expensive. Using matrix calculus, all the partial derivatives can be calculated simultaneously:

$$\frac{dC}{d\mathbf{P}} = \frac{d}{d\mathbf{P}} \sum_{i=1}^{N} \Phi \left[\|\mathbf{P}\mathbf{P}^t\mathbf{x}_i - \mathbf{x}_i\|^2 \right] = \sum_{i=1}^{N} \frac{d}{d\mathbf{P}} \Phi \left[(\mathbf{P}\mathbf{P}^t\mathbf{x}_i - \mathbf{x}_i)^t (\mathbf{P}\mathbf{P}^t\mathbf{x}_i - \mathbf{x}_i) \right] = \dots$$

$$\sum_{i=1}^{N} \Phi' [\mathbf{x}_i^t \mathbf{P}\mathbf{P}^t \mathbf{P}\mathbf{P}^t \mathbf{x}_i - \mathbf{x}_i^t \mathbf{x}_i - 2\mathbf{x}_i^t \mathbf{P}\mathbf{P}^t \mathbf{x}_i] \left[-4\mathbf{x}_i \mathbf{x}_i^t \mathbf{P} + 2 \left[\mathbf{x}_i \mathbf{x}_i^t \mathbf{P}\mathbf{P}^t + \mathbf{P}\mathbf{P}^t \mathbf{x}_i \mathbf{x}_i^t \right] \mathbf{P} \right]$$

Once the gradient is known, different standard techniques can be used to update $\mathbf{P}$. In a simple gradient descent scheme, for example:

$$\mathbf{P}_{n+1} = \mathbf{P}_n - k \frac{dC}{d\mathbf{P}}$$

where k is the step size.

Line search can then be used with a normal PCA as initialization in order to quickly find the optimal $\mathbf{P}$. In this algorithm, different step sizes are probed at each iteration, keeping the one that leads to the minimum value of the cost function C. It is important to mention that the orthonormality condition, which would simplify the expressions of the cost and the gradient, cannot be assumed throughout the process, as the $\mathbf{P}$ matrix is being modified unconstrainedly (even though it converges to an orthonormal matrix).

2.2 Shape Models Based on Φ-PCA

In shape models (Cootes et al. 1995), a set of landmarks is defined on a set of previously aligned shapes. One data vector $\mathbf{s}_i$ is built per shape by stacking of the x and y coordinates of the landmarks. Next, the mean is subtracted from them and PCA performed on the resulting $\mathbf{x}_i$ data vectors, aiming at representing the shapes with a lower dimensionality and with a higher specificity than the explicit cartesian coordinates, at the expense of a certain approximation error. The differences between shape models based on standard and Φ-PCA will be described.

First, the shapes are aligned with the Procrustes method (Goodall 1991) and their mean calculated. Rotation, translation and scaling are allowed for aligning the shapes. The alignment parameters and the mean are optimized simultaneously, minimizing:

$$C_{align} = \sum_{i=1}^{N} \Phi[\|\mathbf{T}_i(\mathbf{z}_i, \theta_i) - \mu^{\Phi}\|^2] = \sum_{i=1}^{N} \Phi[\|\mathbf{s}_i - \mu^{\Phi}\|^2] = \sum_{i=1}^{N} \Phi[\|\mathbf{x}_i\|^2]$$

where $\mathbf{T}_i(\mathbf{z}_i, \theta_i)$ represents the aligned $\mathbf{s}_i$ shape according to the set of parameters θ_i. The constraint $\mu^t \mu = 1$ prevents the shapes from shrinking towards zero. The iterative algorithm described in Cootes et al. 1995 was used for solving the problem:

1. Normalize the size of the first shape and use it as a first estimate of the mean.
2. Align all the shapes to the current estimate of the mean.
3. Update the estimate of the mean by finding the mean of the aligned shapes.
4. Normalize the size of the new estimate of the mean.
5. Go to step 2 until convergence.

The mean in the third step must be found by numerically minimizing the cost in Equation 3, as already explained. However, as minimizing $\Phi(t^2)$ is equivalent to minimizing $t^2 = \|\mathbf{P}\mathbf{P}^t\mathbf{x} - \mathbf{x}\|^2$, the alignments in the second step can be easily calculated by minimizing the sum of squared distances in the standard way (see Cootes et al. 1995 once more for a simple solution). Another consequence of this property is that the PCA coordinates $\mathbf{b}_i$ of a shape can still be calculated in the same way as in the normal PCA:

$$\mathbf{s}_i \approx \mu^{\Phi} + \mathbf{P}\mathbf{b}_i \tag{4}$$

$$\mathbf{b}_i = \mathbf{P}^t(\mathbf{s}_i - \mu^{\Phi}) \tag{5}$$

3 Experiments

This study is based on a dataset which consists of lateral X-rays from the spine of 141 patients. Vertebrae L1 through L4 were outlined by three different expert radiologists, providing the ground truth of the study. 65 landmarks were extracted for each vertebra using the MDL algorithm described in Thodberg 2003. The same radiologists also provided information regarding the fracture type (wedge, biconcave, crush) and grade (mild, medium, severe) for the vertebrae (see Genant et al. 1993). In addition, they also annotaded the six landmarks used in the standard six-point morphometry (Black et al. 1995, Genant et al. 1993), located on the corners and in the middle point of both vertebra endplates. These points define the anterior, middle and posterior heights, which are used to estimate the fracture grade and type.

Both normal PCA and α-PCA (for different values of α) were applied on the dataset keeping 7 (α-)PCA coordinates, capable of preserving approximately 95% of the total variance in the data in all the cases. For both algorithms the mean and maximum squared reconstruction errors were calculated. The dependence of the error on the number of fractures in the training set was also studied. It should be noted that a higher number of components would achieve better precision and still provide a good trade-off with respect to the specificity of the model, but a smaller amount was kept in this experiment in order to better illustrate the difference between PCA and α-PCA.

Finally, PCA and α-PCA were tested in an active shape model (Cootes et al. 1995) for segmenting the L1-L4 vertebrae in the images. Two shape models were built, one for the six landmarks and the other for the full contour, and the relationship between the (α-)PCA coordinates of both models fitted to a conditional Gaussian distribution. In order to allow for more flexibility in the model, a higher number of principal components was utilized: seven for the six landmarks and eleven for the complete contour, keeping approximately 98% of the total variance in both cases.

The mean of the conditional distribution was used as initialization for the segmentation of the full contour. At each iteration, the gray level information along a profile perpendicular to the contour was used to calculate a desired position for each point at the following looping. The new contour can then be calculated by fitting the model to the new points using Equations 4 and 5. The conditional covariance was used to measure the Mahalanobis distance from the new (α-)PCA coordinates $\mathbf{b}$ to the conditional mean. In case of it being larger than a certain threshold D_{max}, the vector is scaled down $\mathbf{b}' = \mathbf{b}(D_{max}/D(\mathbf{b}))$ to ensure $D(\mathbf{b}) \leq D_{max}$. This way, the solution is constrained to stay close to the six landmarks. The process is repeated until convergence.

4 Results

4.1 Mean and Maximum Reconstruction Error

Figures 1-a and 1-b show the dependence on α of the sum of squared errors when fitting the model to labelled points. The maximum error decreases with α, as expected, doing it faster for the fractures. The mean error shows how values of α lower than one tend to increase the error in fractures, as they are no longer important in the model, and decrease it in unfractured vertebrae, even if it is not much. It should be noted that unfractured vertebrae are in general quite well modelled already. Values larger than one initially improve the results in fractures, at the expense of making them slightly worse in unfractured vertebrae. Finally, if α increases too much, the model tends to fit merely the most unlikely cases, making the average results worse both for unfractured vertebrae and mild fractures.

4.2 Influence of the Number of Training Fractures

In this experiment, the model was built with all the unfractured vertebrae and different fractions of the total amount of available fractures: from 12.5% to 100% in 12.5% increments. α was set equal to 1.75, providing a good trade-off between the maximum and mean errors in fractured and unfractured vertebrae, according to the results presented in above. Figure 1-c shows that α-PCA is especially useful, clearly outperforming the normal PCA, when the number of fractures in the training set is relatively small.

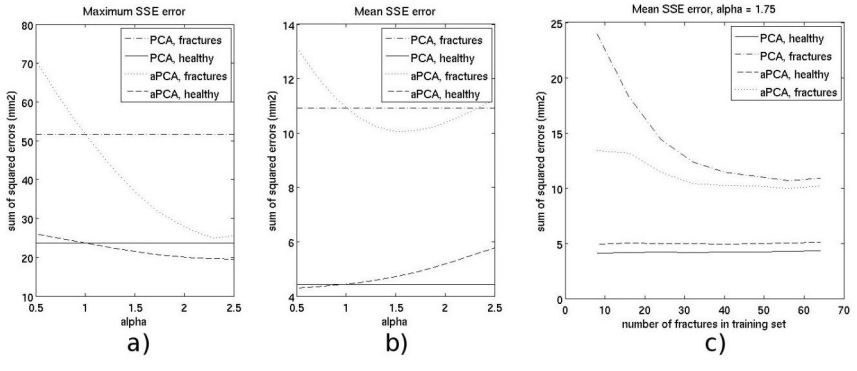

Fig. 1. a) Dependance of the mean sum of squared errors with α. b) Dependence of the maximum sum of squared errors with α. c) Dependence of the mean sum of squared errors with the number of fractures present in the training set for $\alpha = 1.75$.

4.3 Active Shape Model

Vertebrae L1 through L4 were segmented from the available images using a shape model conditioned on the six landmarks annotated by the radiologists, using both standard and α-PCA ($\alpha = 1.75$). The experiments were performed in a leave-one-out fashion: the model used for segmenting a certain image is built upon all the other ones.

Table 1. Mean point-to-line error (in mm) for the different analyses and doublesided p-values for a t-test and a signed rank test with 95% confidence interval

	No. shapes	Error PCA (mm)	Error α-PCA (mm)	p t-test	p signed rank
Unfractured	500	0.44	0.47	$1.79{\cdot}10^{-4}$	$6.05{\cdot}10^{-4}$
Mild fractures	15	0.62	0.56	$1.14{\cdot}10^{-2}$	$1.25{\cdot}10^{-2}$
Medium fractures	38	0.66	0.57	$7.87{\cdot}10^{-3}$	$1.40{\cdot}10^{-2}$
Severe fractures	11	0.97	0.60	$4.78{\cdot}10^{-3}$	$9.77{\cdot}10^{-3}$
All fractures	64	0.70	0.57	$3.10{\cdot}10^{-11}$	$3.53{\cdot}10^{-12}$

The point-to-line errors from the true contour to the output of the algorithm are displayed in Table 1, along with the p-values resulting from a paired, double-sided t-test and a paired, double-sided Wilcoxon signed rank test. The results show that the standard PCA leads to a lower mean error in unfractured vertebrae, but α-PCA provides more uniform results along the different grades of fracture severity, at the expense of a slight increase in the total mean error. Moreover, α-PCA significantly outperforms the standard PCA in fractures, especially in the severe ones. It also has the property of assigning different importance to each case in a continuous manner without requiring fracture information for the training data. If this information was available, it would be possible to build two different models, but then a large number of training fractures would be required. Besides, if two models are fitted, a mistake in the decision about which result to keep could lead to a very bad fit. Regarding the p-values, both tests indicate that the difference in the means between the two setups is significant.

Finally, two radiographs which have been segmented with standard and α-PCA ($\alpha = 1.75$) are displayed along with the contour provided by the radiologists in figure 2. They both correspond to severe fractures. α-PCA provides a better approximation of the real shape, especially around the points in which it changes its direction rapidly.

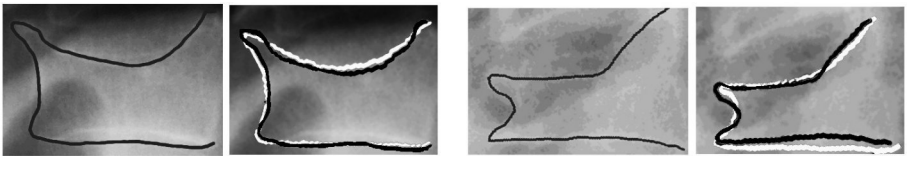

Fig. 2. Segmentation examples. For each pair, the image on the left corresponds to the ground truth and the image on the right to the shape model-based segmentation, both for standard PCA (white) and α-PCA (black).

5 Discussion and Conclusion

A family of modified PC analyses has been presented in this paper. The family deals with outliers in the data set in an optimal way according to a predefined function, whose shape determines whether the importance of outliers increases or decreases compared with normal PCA. The family $\Phi(x) = x^{\alpha}$ is proposed, but others could be used. Compared to other methods in the literature, the one presented here has the ability of enhancing or disregarding outliers with just one compact and simple formulation.

In most applications it is desirable to reduce the influence of abnormal cases on the principal components. However, if PCA is utilized in a segmentation method, it is essential to be able to adapt to such cases, whose correct processing might be even more important than that of the normal ones. In this paper, α-PCA was tested in the creation of a vertebra shape model, giving a higher importance to abnormal cases without requiring prior knowledge on which of the shapes are fractured or present other abnormalities, such as osteophytes. The segmentation accuracy was improved in such cases.

It should finally be noted that the conditional model used in the segmentation algorithm assumes a Gaussian distribution for the PCA coordinates of the shapes. Using α-PCA instead of standard PCA makes the distribution resemble less a Gaussian, indirectly affecting the segmentation results. This fact may also affect further statistics analysis if the PCA coordinates are for example used to estimate the fracture grade.

References

Jolliffe, I.: Principal component analysis. Springer, New York (1986)

Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W.: Robust statistics:the approach based on influence functions. Wiley, New York (1986)

Xu, L., Yuille, A.: Robust principal component analysis by self organizing rules based on statistical physics approach. IEEE Trans. Neural Networks 6, 71–86 (1991)

Gabriel, K., Zamir, S.: Lower rank approximation of matrices by least squares with any choice of weights. Technometrics 21(21), 489–498 (1979)

De la Torre, F., Black, M.J.: A framework for robust subspace learning. International Journal of Computer Vision, 117–142 (2003)

Goodall, C.R.: Procrustes methods in the statistical analysis of shape. supervised methods: a comparative study on a public database. J. R. Stat. Soc. Medical Image Analysis 53, 285–339 (1991)

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models - their training and application. Comput Vis. Image Underst. 61(1), 38–59 (1995)

Thodberg, H.H.: Minimum Description Length Shape and Appearance Models. In: Proceedings of Information Processing in Medical Imaging, Springer, Heidelberg (2003)

Black, D.M., Palermo, L., Nevitt, M.C., Genant, H.K., Epstein, R., San Valentín, R., Cummings, S.R.: Comparison of methods for defining prevalent vertebral deformities: the study of osteoporotic Fractures. J. Bone Miner. Res. 10, 890–902 (1995)

Genant, H.K., Wu, C.Y., van Kuijk, C., Nevitt, M.C.: Vertebral Fracture Assesment Using a Semiquantitative Technique. J. Bone Miner. Res. 8(9), 1137–1148 (1993)