Probabilistic Generative Modelling

Rasmus Larsen and Klaus Baggesen Hilger

Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
{rl, kbh}@imm.dtu.dk, http://www.imm.dtu.dk

Abstract. The contribution of this paper is the adaption of data driven methods for decomposition of tangent shape variability proposed in a probabilistic framework. By Bayesian model selection we compare two generative model representations derived by principal components analysis and by maximum autocorrelation factors analysis.

1 Introduction

For the analysis and interpretation of multivariate observations a standard method has been the application of principal component analysis (PCA) to extract latent variables. Cootes et al. applied PCA to the analysis of tangent space shape coordinates [1]. For various purposes different procedures for PCA using non-Euclidean metrics have been proposed. The maximum autocorrelation factor (MAF) transform proposed by Switzer [2] defines maximum spatial autocorrelation as the optimality criterion for extracting linear combinations of multispectral images. Contrary to this PCA seeks linear combinations that exhibit maximum variance. Because imaged phenomena often exhibit some sort of spatial coherence spatial autocorrelation is often a better optimality criterion than variance. We have previously adapted the MAF transform for analysis of tangent space shape coordinates [3]. In [4] the noise adjusted PCA or the minimum noise fraction (MNF) transformations were used for decomposition of multispectral satellite images. The MNF transform is a PCA in a metric space defined by a noise covariance matrix estimated from the data. For image data the noise process covariance is conveniently estimated using spatial filtering. In [5] the MNF transform is applied to texture modelling in active appearance models [6], and in [7] to multivariate images in extracting a discriminatory representation. Bookstein proposed using bending energy and inverse bending energy as metrics in the tangent space [8]. Using the bending energy puts emphasis on the large scale variation, using its inverse puts emphasis of small scale variation.

2 Methods

In the following two Sections we will describe how to use two methods, maximum autocorrelation factors [2] and minimum noise fractions [4], for decomposing the tangent space coordinates of a set of shapes into a low-dimensional subspace.

The tangent space coordinates are obtained by a generalized Procrustes alignment [9, 10] followed by a projection of the full Procrustes coordinates into the tangent space to the shape space at the full Procrustes mean (e.g. [11]). Let the tangent space coordinates, $\boldsymbol{x}_i = (x_{i11}, \ldots, x_{i1n}, \ldots, x_{id1}, \ldots, x_{idn})^T$, for shapes $i = 1, \ldots, p$ with $j = 1, \ldots, n$ landmarks in $d \in \{2, 3\}$ dimensions be organised in a $p \times dn$ data matrix $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2 \ldots \boldsymbol{x}_p]^T$. Denote the Procrustes (sample) mean shape $\bar{\boldsymbol{x}}$ and let it be centered on (0,0) in 2D and (0,0,0) in 3D, further let the origin of the tangent space coordinate system be the mean shape, then $\boldsymbol{X}$ is doubly centered, i.e. columns as well as rows sum to zero. Additionally, it is assumed that the landmarks are sampled on curves (in 2D) and surfaces (in 3D) that allow for definition of neighbouring landmarks, i.e. in terms of the order along a curve or through a triangulation of landmarks on a surface.

In the following we will consider Q-mode analyses of the matrix $\boldsymbol{X}$. In the case of principal components analysis this is an eigenvalue decomposition of the covariance matrix estimated from observations $\boldsymbol{z}_{lj} = (x_{1lj}, \ldots, x_{plj})^T$, for $j = 1, \ldots, n$, $l = 1, \ldots, d$. These $\boldsymbol{z}_{lj}$ are vectors of a landmark coordinates observed over the set of shapes. The maximum likelihood estimator of the covariance matrix is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{p}\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^T$$

here $\boldsymbol{\Lambda}^2$ is a diagonal matrix containing the eigenvalues of $\hat{\boldsymbol{\Sigma}}$, and $\boldsymbol{V}$ contains the corresponding conjugate eigenvectors. A point distribution model then consists of retaining the $t \leq r$ first principal components. Deviations from the Procrustes mean (in tangent space) can then be modelled by

$$\boldsymbol{x} = \boldsymbol{X}^T\boldsymbol{V}'\boldsymbol{b} \tag{1}$$

where $\boldsymbol{V}'$ is a matrix consisting of the first t columns of $\boldsymbol{V}$, and $\boldsymbol{b}$ defines a set of t parameters of the deformable model.

2.1 Maximum autocorrelation factors

Let the spatial covariance function of a multivariate stochastic variable, $\boldsymbol{Z}_k$, where k denotes spatial position and Δ a spatial shift, be $\boldsymbol{\Gamma}(\Delta) = \mathrm{Cov}\{\boldsymbol{Z}_k, \boldsymbol{Z}_{k+\Delta}\}$. Evidently $\boldsymbol{\Gamma}^T(\Delta) = \boldsymbol{\Gamma}(-\Delta)$. Then by letting the covariance matrix of $\boldsymbol{Z}_k$ be $\boldsymbol{\Sigma}$ and defining the covariance matrix $\boldsymbol{\Sigma}_\Delta = D\{\boldsymbol{Z}_k - \boldsymbol{Z}_{k+\Delta}\}$, we find $\boldsymbol{\Sigma}_\Delta = 2\boldsymbol{\Sigma} - \boldsymbol{\Gamma}(\Delta) - \boldsymbol{\Gamma}(-\Delta)$ where $\boldsymbol{\Sigma}_\Delta$ is the dispersion of the difference process in lag Δ. We are now able to to compute the covariance between a linear combination of the original variables and the shifted variables

$$\begin{aligned}&\mathrm{Cov}\{\boldsymbol{w}^T\boldsymbol{Z}_k, \boldsymbol{w}^T\boldsymbol{Z}_{k+\Delta}\} = \\ &\boldsymbol{w}^T\boldsymbol{\Gamma}(\Delta)\boldsymbol{w} = \frac{1}{2}\boldsymbol{w}^T(\boldsymbol{\Gamma}(\Delta) + \boldsymbol{\Gamma}(-\Delta))\boldsymbol{w} = \boldsymbol{w}^T(\boldsymbol{\Sigma} - \frac{1}{2}\boldsymbol{\Sigma}_\Delta)\boldsymbol{w}.\end{aligned} \tag{2}$$

Thus the autocorrelation in shift Δ of a linear combination of Z_k is

$$\mathrm{Corr}\{\boldsymbol{w}^T\boldsymbol{Z}_k, \boldsymbol{w}^T\boldsymbol{Z}_{k+\Delta}\} = 1 - \frac{1}{2}\frac{\boldsymbol{w}^T\boldsymbol{\Sigma}_\Delta\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}} = 1 - \frac{1}{2}R(\boldsymbol{w}). \tag{3}$$

In order to maximize this correlation we must minimize the Rayleigh coefficient, $R(\boldsymbol{w})$. This is obtained by choosing as $\boldsymbol{w}$ the conjugate eigenvector corresponding the smallest generalized eigenvalue of $\boldsymbol{\Sigma}_\Delta$ wrt. $\boldsymbol{\Sigma}$. The MAF transform is given by the set of conjugate eigenvectors of $\boldsymbol{\Sigma}_\Delta$ wrt. $\boldsymbol{\Sigma}$, $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m]$, corresponding to the eigenvalues $\kappa_1 \leq \cdots \leq \kappa_m$ [2]. The resulting new uncorrelated variables are ordered so that the first MAF is the linear combination that exhibits maximum autocorrelation. The autocorrelation of the ith component is $1 - \frac{1}{2}\kappa_i$. We assume first and second order stationarity of the data.

One problem now arise, namely, how should we choose Δ. Switzer suggests that we estimate Σ_Δ for a shift in lag 1. Blind source separation by independent components analysis using the Molgedey-Schuster (MS-ICA) algorithm [12] is equivalent to MAF [3]. The purpose of this algorithm is to separate independent signals from linear mixings. MS-ICA does this by exploiting differences in autocorrelation structure between the independent signals. Kolenda et al. [13] use an iterative procedure for identifying the optimal lags based on the sum of pairwise absolute differences between the autocorrelations of the estimated independent components. In this study we use Switzers original suggestion. This is based on the assumption that the noise is separated from the interesting latent variables in terms of autocorrelation already in lag 1. For shape analysis decomposition of the datamatrix $\boldsymbol{X}$ using MAF is carried out in Q-mode. In 2D the difference process covariance matrix $\boldsymbol{\Sigma}_\Delta$ is estimated from the lag 1 difference process of landmarks along the contours of the object. In 3D we estimate the difference process covariance matrix from the differences between the landmark coordinates and a plane fitted to the landmarks in a k^{th}-order neighbourhood.

2.2 Minimum noise fractions

As before we consider a multivariate stochastic variable, $\boldsymbol{Z}_k$. We assume an additive noise structure $\boldsymbol{Z}_k = \boldsymbol{S}_k + \boldsymbol{N}_k$, where $\boldsymbol{S}_k$ and $\boldsymbol{N}_k$ are uncorrelated signal and noise components, with covariance matrices $\boldsymbol{\Sigma}_S$ and $\boldsymbol{\Sigma}_N$, respectively. Thus $\mathrm{Cov}\{\boldsymbol{Z}_k\} = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_S + \boldsymbol{\Sigma}_N$. By defining the signal-to-noise ratio (SNR) as the ratio of the signal variance and the noise variance we find for a linear combination of $\boldsymbol{Z}_k$

$$\mathrm{SNR}_i = \frac{V\{\boldsymbol{w}_i^T \boldsymbol{S}_k\}}{V\{\boldsymbol{w}_i^T \boldsymbol{N}_k\}} = \frac{\boldsymbol{w}_i^T \boldsymbol{\Sigma}_S \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{\Sigma}_N \boldsymbol{w}_i} = \frac{\boldsymbol{w}_i^T \boldsymbol{\Sigma} \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{\Sigma}_N \boldsymbol{w}_i} - 1. \tag{4}$$

So the minimum noise fractions are given by the set of conjugate eigenvectors of $\boldsymbol{\Sigma}$ wrt. $\boldsymbol{\Sigma}_N$, $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m]$, corresponding to the eigenvalues $\kappa_1 \geq \cdots \geq \kappa_m$ [4]. The resulting new variables are ordered so that the first MNF is the linear combination that exhibits maximum SNR. The ith MNF is the linear combination that exhibits the highest SNR subject to it being uncorrelated to the previous MNFs. The SNR of the ith component is $\kappa_i - 1$. If the matrices in Equations (3) and (4) are singular the solution must be found in the affine support of the matrix in the denominator, e.g. by means of a generalized singular value decomposition.

2.3 Evaluation of point distribution models by probabilistic reconstruction

Following Minka [14] we use a probabilistic principal components analysis model for choice of dimensionality. Let a multivariate response $\boldsymbol{X}$ of p dimensions be modelled by a linear combination of a set of basis vectors $\boldsymbol{h}_i$, $i = 1, \ldots, k$ plus noise

$$\boldsymbol{X} = \sum_{i=1}^{k} \boldsymbol{h}_i b_i + \boldsymbol{m} + \boldsymbol{N} = \boldsymbol{H}\boldsymbol{b} + \boldsymbol{m} + \boldsymbol{N} \tag{5}$$

where $\boldsymbol{N} \in N(\boldsymbol{0}, \boldsymbol{\Sigma_N})$, and $\boldsymbol{b}$ has dimension $k < p$. The vector $\boldsymbol{m}$ defines the mean of $\boldsymbol{X}$, while $\boldsymbol{H}$ and $\boldsymbol{\Sigma}_N$ defines its variance. For PCA the noise variance is spherical, i.e. $\boldsymbol{\Sigma}_N = v\boldsymbol{I}_p$. Furthermore, we assume a spherical Gaussian prior density for $\boldsymbol{b}$, $\boldsymbol{b} \in N(\boldsymbol{0}, \boldsymbol{I}_k)$. For this model the maximum likelihood estimators for the model parameters given observations $\boldsymbol{x}_i$, $i = 1, \ldots, N$ are

$$\hat{\boldsymbol{m}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i, \quad \hat{\boldsymbol{H}} = \boldsymbol{U}_k(\boldsymbol{\Lambda}_k - \hat{v}\boldsymbol{I}_k)^{1/2}\boldsymbol{R}, \quad \hat{v} = \frac{1}{p-k}\sum_{i=k+1}^{p} \lambda_i. \tag{6}$$

Where $\boldsymbol{U}_k$ contains the eigenvectors corresponding to the top k eigenvalues of the maximum likelihood estimate of the dispersion matrix of the observations $\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \hat{\boldsymbol{m}})(\boldsymbol{x}_i - \hat{\boldsymbol{m}})^T$, λ_i is the i^{th} eigenvalue of $\hat{\boldsymbol{\Sigma}}$, the diagonal matrix $\boldsymbol{\Lambda}_k$ contains the corresponding eigenvalues, and $\boldsymbol{R}$ is an arbitrary orthogonal matrix. The likelihood of the data, D, thus becomes

$$p(D|\boldsymbol{H}, \boldsymbol{m}, \boldsymbol{v}) = \\ (2\pi)^{-Np/2}|\boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}} + \boldsymbol{v}\boldsymbol{I}|^{-N/2}\exp(-\frac{1}{2}tr((\boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}} + \boldsymbol{v}\boldsymbol{I})^{-1}N\hat{\boldsymbol{\Sigma}})). \tag{7}$$

Let us instead assume a general unrestricted covariance structure of the noise, which may contain intercorrelated effects. Then it is fairly easily shown that by an initial linear transformation that diagonalises $\boldsymbol{\Sigma}_N$, using the result above, that the maximum likelihood estimate of $\boldsymbol{H}$ consists of the first k minimum noise fraction factors (cf. Eq. (4)). For a given model the loglikelihood (LL) of the data can be estimated. However, with ever increasing model complexity, better reconstruction of the data is obtained, and thus a corresponding increase in LL is observed. The LL estimates must therefore be penalized e.g. by using the Bayesian information criterion (BIC) or Akaikes information criterion (AIC). Given the probability of the data and the degrees of freedom in the model, then BIC and AIC reduce to

$$\text{BIC} = -2\log(p(D|\boldsymbol{H}, \boldsymbol{m}, \boldsymbol{\Sigma_n})) + (k+1)p\log(N), \tag{8}$$

$$\text{AIC} = -2\log(p(D|\boldsymbol{H}, \boldsymbol{m}, \boldsymbol{\Sigma_n})) + 2(k+1)p/N. \tag{9}$$

In order to avoid the bias introduced by estimation parameters and evaluating performance on the same dataset we may apply cross validaion (CV). Let the

variance of the isotropic noise in the model complement subspace λ_e be estimated by the average variance not explained by the model, let $\boldsymbol{\Lambda}$ be a diagonal matrix of the variance $\lambda_i, i = 1, \ldots, k$ in the corresponding orthogonal subspaces of $\boldsymbol{H}$, and let $\lambda_i = \lambda_e, i = k+1, \ldots, p$. Let $\boldsymbol{r}_{mj}$ represent the jth excluded observation projected into the model space given by $\boldsymbol{H}^T(\boldsymbol{x}_{ex,j} - \boldsymbol{m})$, and $\boldsymbol{r}_{ej}$ the residuals in the complement space given by $(\boldsymbol{I} - \boldsymbol{H}\boldsymbol{H}^T)(\boldsymbol{x}_{ex,j} - \boldsymbol{m})$. In this case the loglikelihood of the data on an orthogonal affine model support $\boldsymbol{H}$ is given by

$$\mathrm{LL} = -\frac{1}{2}(n_{ex}(p\log(2\pi) + \sum_{i=1}^{p}\log(\lambda_i)) + \sum_{j=1}^{n_{ex}}(\boldsymbol{r}_{mj}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{r}_{mj} + \lambda_e^{-1}\boldsymbol{r}_{ej}^T\boldsymbol{r}_{ej})) \quad (10)$$

where n_{ex} is the number of observations in the excluded CV set.

3 Results

We demonstrate the properties of the techniques that we propose on a dataset consisting of 2D annotations of the outline of the right and left lung from 115 standard PA chest radiographs. The chest radiographs were randomly selected from a tuberculosis screening program and contained normal as well as abnormal cases. The annotation process was conducted by identification of three anatomical landmarks on each lung outline followed by equidistant distribution of pseudo landmarks along the 3 resulting segments of the outline. In Fig. 1(b) the landmarks used for annotation are shown. Each lung field is annotated independently by two observers [15].

In Fig. 2 results of a five-fold CV study of the loglikelihood is shown. The figure shows the average performance of the generative PCA and MAF models and the one standard deviation bounds for a given model complexity. For the PCA based model the LL analysis attains its maximum at 18 dimensions, whereas the MAF has its maximum at 30 dimensions. Truncation of the models is typically obtained by tracking the lower one standard deviation bound backward, leading to model complexities of 13 and 22 dimensions for respectively the PCA and the MAF analysis. Although the MAF basis indicates a higher rank model, it is important to note that it finds uncorrelated modes of biological variation in a non-Euclidean metric. The modes may thus provide a better separation of signal from noise, and typically the MAF components possess better discriminatory power over the traditional PCs. Fig. 3 shows the most important PCA and MAF components derived from an analysis on all the training data.

4 Conclusion

We have demonstrated data driven methods for PCA and MAF decompositions of tangent space shape variability, and provided a probabilistic framework for selecting the best model and regularization. In our case study PCA performs best in deriving a compact low dimensional model. However, the fact that the MAF

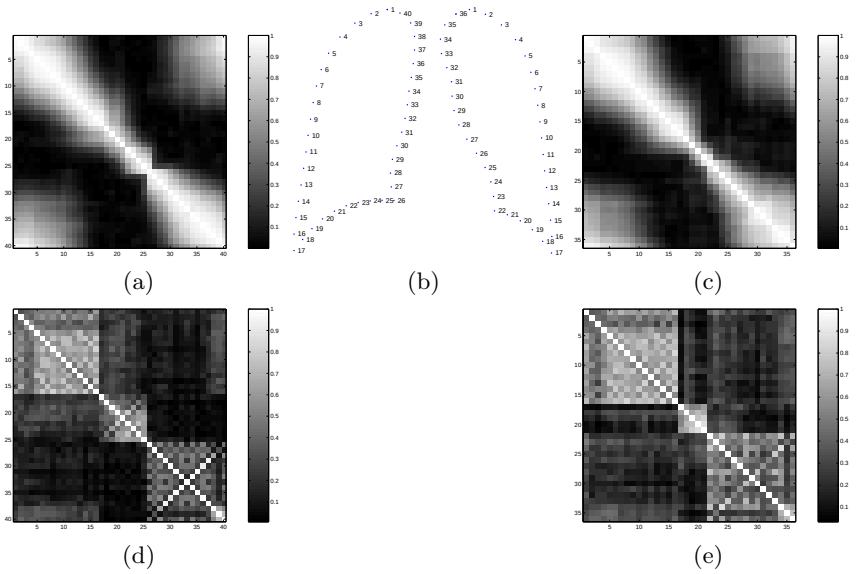

Fig. 1. Landmarks of the left and right lung. Landmark numbers are shown in the middle. The right lung is annotated by 40 landmarks, and the left lung by 36. The anatomical landmarks on the right field are points 1, 17, and 26, on the left field the anatomical landmarks are points 1, 17, and 22. (a),(c) Inter-observer difference canonical correlations between landmarks for the right and left lungs. (d),(e) Inter-neighbour landmark difference canonical correlations between landmark for the right and left lung.

analysis expands into a higher rank representation is not necessarily undesirable. In fact the higher dimensional MAF model attains comparable capability in generalizing to the data. In effect, it provides a more detailed image of the signal present in the data in probing for uncorrelated biological modes of variation.

5 Acknowledgements

The work was supported by the Danish Technical Research Council under grant number 26-01-0198 which is hereby gratefully acknowledged. The authors thank Dr. Bram van Ginneken for use of the lung annotation data set.

References

1. T. F. Cootes, G. J. Taylor, D. H. Cooper, and J. Graham, "Training models of shape from sets of examples," in *British Machine Vision Conference: Selected Papers 1992*, (Berlin), Springer-Verlag, 1992.
2. P. Switzer, "Min/max autocorrelation factors for multivariate spatial imagery," in *Computer Science and Statistics* (L. Billard, ed.), pp. 13–16, Elsevier Science Publishers B.V. (North Holland), 1985.

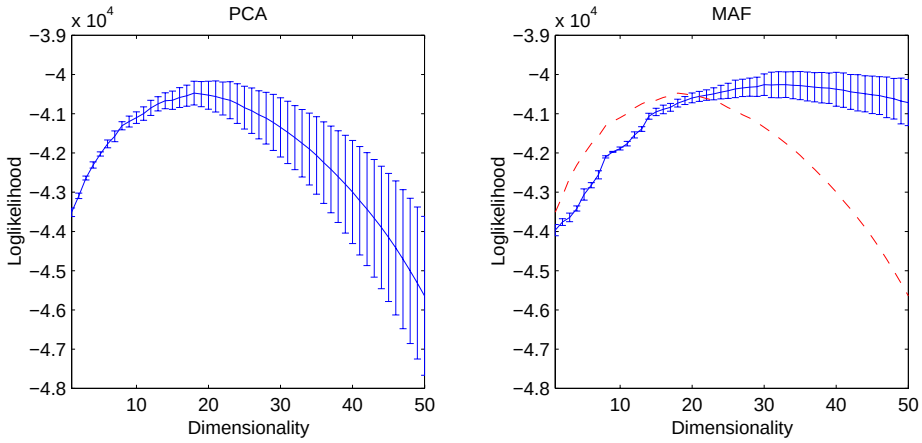

Fig. 2. Results of a five fold cross validation analysis of the 2D lung data. The loglikelihood is shown as a function of increasing model complexity. The mean is shown with one standard deviations error-bars. In the left plot a PC based model is applied, and the right MAF based model. For comparison the dashed curve in the right plot shows the average performance of the PCA basis.

3. R. Larsen, H. Eiriksson, and M. B. Stegmann, "Q-MAF shape decomposition," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2001, 4th International Conference, Utrecht, The Netherlands*, vol. 2208 of *Lecture Notes in Computer Science*, pp. 837–844, Springer, 2001.
4. A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, pp. 65–74, Jan. 1988.
5. K. B. Hilger, M. B. Stegmann, and R. Larsen, "A noise robust statistical texture model," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2002, 5th International Conference, Tokyo, Japan*, 2002. 8 pp. (submitted).
6. T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proceedings of the European Conf. On Computer Vision*, pp. 484–498, Springer, 1998.
7. K. B. Hilger, A. A. Nielsen, and R. Larsen, "A scheme for initial exploratory data analysis of multivariate image data," in *Proceedings of the Scandinavian Image Analysis Conference, SCIA'01, Bergen, Norway, 11–14 June 2001*, 2001. pp. 717–724.
8. F. L. Bookstein, *Morphometric tools for landmark data*. Cambridge University Press, 1991. 435 pp.
9. J. C. Gower, "Generalized Procrustes analysis," *Psychometrika*, vol. 40, pp. 33–50, 1975.
10. C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society, Series B*, vol. 53, no. 2, pp. 285–339, 1991.
11. I. L. Dryden and K. Mardia, *Statistical Shape Analysis*. Chichester: John Wiley & Sons, 1998. xx + 347 pp.

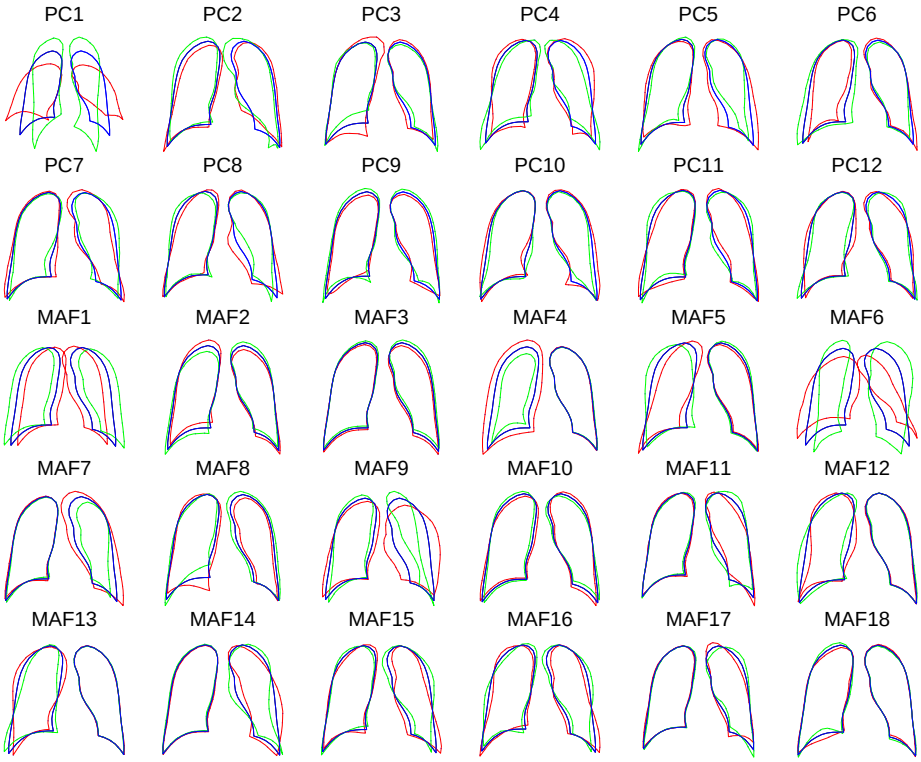

Fig. 3. The most important modes of variation derived from the PCA and the MAF analysis. The blue curve is the mean shape, and the green and red curves are ± 5 standard deviations as observed in the training set.

12. L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.
13. T. Kolenda, L. K. Hansen, and J. Larsen, "Signal detection using ica: Application to chat room topic spotting," in *Proceedings of the 3rd International Conference on Independent Components Analysis and Blind Signal Separation (ICA'2001), San Diego, USA, December 9-13* (Lee, Jung, Makeig, and Sejnowski, eds.), 2001.
14. T. P. Minka, "Automatic choice of dimensionality of PCA," in *Advances in Neural Information Processing Systems 13* (T. K. Leen, T. G. Dietterich, and V. Tresp, eds.), pp. 598–604, MIT Press, 2001.
15. B. van Ginneken, *Computer-Aided Diagnosis in Chest Radiographs.* PhD thesis, Image Sciences Institute, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands, 2001. 184 pp.