Sample Sufficiency and Number of Modes to Retain in Statistical Shape Modelling

Lin Mei*, Michael Figl, Daniel Rueckert, Ara Darzi, and Philip Edwards*

Dept. of Biosurgery and Surgical Technology Imperial College London, UK
l.mei,eddie.edwards@imperial.ac.uk*

Abstract. Statistical shape modelling is a popular technique in medical imaging, but the issue of sample size sufficiency is not generally considered. Also the number of principal modes retained is often chosen simply to cover a percentage of the total variance. We show that these simple rules are unreliable. We propose a new method that uses bootstrap replication and a *t*-test comparison with noise to decide whether each mode direction has stabilised. We establish mode correspondence by minimising the distance between the space spanned by the replicates and their mean. By retaining only stable modes, our method distinguishes real anatomical variation from modes dominated by random noise. This provides a lower stopping rule when the sample is small and converges as the sample size increases. We use this convergence to determine sample sufficiency. For validation we use synthetic datasets of the left ventricle generated with a known number of structural modes and added noise. Our stopping rule detected the correct number of modes to retain where other methods failed. The methods were also tested on real 2D (22 points) and 3D (500 points) face data, retaining 24 and 70 modes with sample sufficiency being reached at approximately 50 and 150 samples respectively. For a 3D database of the left ventricle (527 points), 319 samples are not sufficient, but at this level we can retain around 55 stable modes. Our method provides a principled foundation for appropriate selection of the number of modes to retain and determination of sample size sufficiency for statistical shape modelling.

1 Introduction

Statistical shape modelling (SSM) is a technique for characterising variation of shape and fitting to unseen shapes. A set of sample shapes is collected and principal component analysis (PCA) is performed to determine the principal modes of shape variation. Since the surfaces are usually extracted from 3D image data the dimensionality of the shape vector will typically be high, perhaps several thousand. The number of samples used in constructing the model varies, but is generally in the range 10-50 [1,2,3,4]. While these training sets are sufficient

* We would like to thank Tyco Healthcare for funding Lin Mei's PhD studentship. We are also grateful to many other members of the Department of Computing and the Department of Biosurgery and Surgical Technology at Imperial College.

D. Metaxas et al. (Eds.): MICCAI 2008, Part I, LNCS 5241, pp. 425–433, 2008.

to prove the principle of a technique, the sample may not be large enough to ensure that the resulting model reflects the true background anatomical variation. Limited literature can be found discussing PCA sample size sufficiency. In the related field of common factor analysis (CFA), early guidelines for a minimum sample size requirement involved either a universal size regardless of the data dimension or a ratio to the number of dimensions. However, there is inconsistency between the suggestions, implying that the minimum size depends on some intrinsic characteristics of the data other than its dimension. MacCallum et al. proposed that for CFA they are communality and overdeterminaton level [5].

Methods which identify the number of modes to retain for a PCA model are called *stopping rules*. A number of methods have been proposed for PCA [6,7,8,9]. The most commonly used rule in SSM is to use a threshold, e.g. 95%, on the cumulative percentage of principal modes' variance [8,10]. However, the choice of threshold is somewhat arbitrary, and we will show in section 5 that the number of modes retained varies with sample size.

Stability measurements for SSM have been proposed to determine the number of modes. Given two shape models trained from different sample sets, Daudin et al [11] used a sum of correlation coefficients between pairs of principal components; Besse et al [12] used a loss function derived from an Euclidean distance between orthogonal projectors; Babalola et al [13] used the Bhattacharya Metric to measure the similarity of PCA models from different sample sets. Resampling techniques such as bootstrapping [11] and jackknifing [12] are used. The distribution of PCA modes across the replicates reflects their distribution in the population, allowing stability analysis to be performed. The selected principal modes span a subspace. Besse et al. proposed a framework for choosing the number of modes based on their spanned-space stability [14]. This method differentiates structural modes and noise-dominated modes when the sample set is large. However, as will be shown in the section 5, this method can only provide a estimation of the number of modes when the sample size is large enough.

In this paper, we establish mode correspondence by minimising the distance between principal spanned spaces. We then apply bootstrapping to estimate the distribution of each eigenmode direction and perform a t-test against pure Gaussian noise to determine the number of modes that should be retained for SSM. This leads to a procedure to test for the sufficiency of the current sample size by convergence of the number of modes retained. These methods are validated on synthetic data generated with a known number of modes, and applied to a real dataset of the left ventricle from MRI and datasets of 2D and 3D faces.

2 Stopping Rule by Stability of Mode Direction

Our stopping rule is based on bootstrap stability analysis on mode directions. This requires correspondence of the PCA modes trained from different replicates.

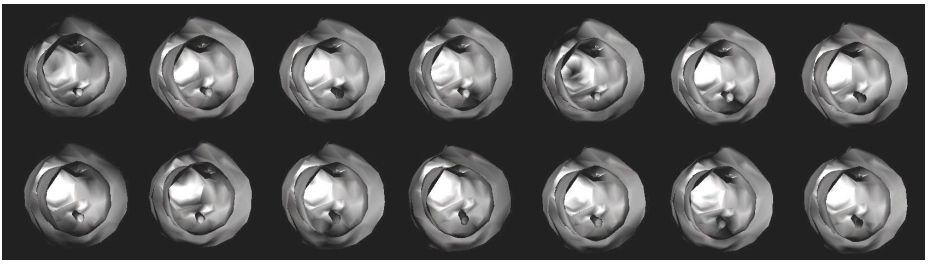

Fig. 1. Comparison of Leading 7 Eigenmodes from two mutually exclusive sets of 50 samples from our 3D heart mesh database. Darker texture implies larger variation.

2.1 Establishing Mode Correspondence

Examining individual modes requires mode correspondence. Normally, this is done by matching those with the same eigenvalue ranks. This method may fail if the variances along two modes are quite similar. As can be seen in figure 1, there can be significant variation of individual mode directions between sample sets. However, the combined modes from different sample sets may still span similar subspaces. Mode alignment can be achieved by minimising the distance between these subspaces.

For the leading PCA modes $\{(\boldsymbol{a}_i, \lambda_i)|\boldsymbol{a}_i| = 1\}$ of an n-dimensional distribution, we define the principal spanned space (PSS) as the subspace $\mathbb{S}^k$ spanned by $\{\boldsymbol{a}_i\}$, where distance measurement used by Besse et al.[12] can be applied:

$$d(\mathbb{A}^k, \mathbb{B}^k) = k - trace(AA^T BB^T) \tag{1}$$

where the columns of A and B are the modes spanning PSS $\mathbb{A}^k$ and $\mathbb{B}^k$.

For two sets of PCA modes, $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, trained from different sample sets of a common distribution, the following rule can be used to establish correspondence. The first mode in $\boldsymbol{a}_i$ corresponds to the mode of a replicate that minimises $d(\mathbb{S}_\mathbb{a}{}^1, \mathbb{S}_\mathbb{b}{}^1)$, and we proceed iteratively. Assume we have already aligned $\mathbb{S}_\mathbb{a}{}^k$, the PSS from the first k modes in $\boldsymbol{a}_i$, to the spanned space $\mathbb{S}_\mathbb{b}{}^k$ from k modes in the replicate $\boldsymbol{b}_i$. The mode in $\boldsymbol{b}_i$ that corresponds to the k+1$^{\text{th}}$ mode in $\boldsymbol{a}_i$ will be the one that minimises $d(\mathbb{S}_\mathbb{a}{}^{k+1}, \mathbb{S}_\mathbb{b}{}^{k+1})$.

2.2 *t*-Test on Mode Stability

There is a risk with tests using the magnitude of the variance that stopping rules will be dominated by the first few modes and fail to identify the correct cut-off point. Also, it is the mode directions that define the basis of a shape model for fitting or synthetic shape generation. Therefore we propose a stopping rule based on the stability of the mode direction only.

Averaged dot-product between corresponding modes and their mean was used as the stability of mode direction [9]. We apply the same principle to the modes

from different bootstrap replicates, but for clarity we use the angles between mode directions. The instability, ξ, of mode $\boldsymbol{a}_i$ is given by:

$$\xi(\boldsymbol{a}_i) = \frac{\sum_{j=1}^{m} arccos(\boldsymbol{a_i}'_j \cdot \widehat{\boldsymbol{\alpha_i}}) \cdot}{m\pi} \tag{2}$$

where $\widehat{\boldsymbol{\alpha_i}}$ is the mean mode vector and m is the number of bootstrap replicates.

Since noise-dominated modes should have higher instability, a threshold on ξ can be used to differentiate them from structural modes. However, the choice for the threshold is arbitrary and is found to be sensitive to the size of replicates. Instead, assuming the distribution of angles between corresponding modes is Gaussian, an one-tailed t-test can be used to establish whether a mode is dominated by noise to a given significance level.

We generate a pure Gaussian noise dataset to compare with the test dataset. All conditions must be the same – the dimensionality, the number of samples in the dataset, the number of replicates, and the number of samples in each replicate. Since we are only interested in mode directions, the level of noise is not important. Let the angle for the first pure noise mode to be $\boldsymbol{\alpha}_1$ and the angle for the i-th mode of the test samples to be $\boldsymbol{a}_i$, The null hypothesis of the t-test is $H_0 : \xi(\boldsymbol{\alpha}_1) > \xi(\boldsymbol{a}_i)$. By rejecting H_0 at a given confidence level, one can safely conclude that the i-th mode is not dominated by noise.

3 Sample Size Sufficiency

Studies on CFA showed that the sample size requirement for a statistical model really depends on certain characteristics of the data that are modelled. For CFA these are communality and overdetermination level [5]. For SSM, such factors could be the compactness and the number of genuine anatomical modes not hidden by noise. With increasing sample size, more PCA modes of the background variance are well covered. Once the training set becomes sufficient, no further modes will be revealed. We propose the following procedure to determine the sample size sufficiency. For a sample set, X, of n samples:

1) Apply PCA on X, to get a set of modes B.
2) Starting with a reasonably small number, n^*, Construct a set of resampled sets $\{X_j^*\}$, in which each set, X_j^* contains n^* samples randomly drawn from X allowing repeats.
3) Apply PCA to $\{X_j^*\}$ to get each set of modes $\{B_j^*\}$ and align them to B using the algorithm described in section 2.1.
4) With modes in $\{B_j^*\}$ aligned, calculate the number of structural modes, k can be tested using our t-Test based stopping rule.
5) Repeat 2-4 with an increased n^*. If k converges before n^* reaches n, we have sufficient samples. Otherwise, further samples are required.

An effective stopping rule for part 4 in this procedure should converge.

4 Real Datasets

As sample data we use a set of 319 surface models of the left ventricle, each with 527 corresponding points. These are derived from 4D CT scans of 29 subjects. Eleven shapes for each subject are chosen at different points in the cardiac cycle. Two other real shape datasets are also used to verify our sample size sufficiency test – 135 samples from the landmarks of the 2D AR face database (22 points) [15] and 150 samples of 3D faces (decimated to 500 points) from University of Notre Dame [16], preprocessed using Papatheodorou's method [17].

5 Validation on Synthetic Data

We have validated previous stopping rules and our method using synthetic data generated using the leading 40 modes of the model built from all the 319 cardiac samples. Gaussian noise with 1mm standard deviation is added to each element of the shape vector. The average noise vector length is 41.3mm, which is significantly larger than variance along the 40th genuine mode which is 9.9mm, stopping rules applied to this dataset should not retain more than 40 modes.

5.1 Validation of Previous Methods

We validated the rule of 95% cumulative variance using synthetic datasets sized from 50 to 200. Compactness plots are shown in the figure 2. With increasing sample size, the number of modes retained by this rule increases beyond 40, where the noise dominates the variance. These noise modes contribute to an increasing proportion of the total variance with increasing sample size, and the number of modes covering 95% of total variance increases accordingly. A similar trend was also found for the real data, strongly suggesting that this rule is unreliable and should not be used.

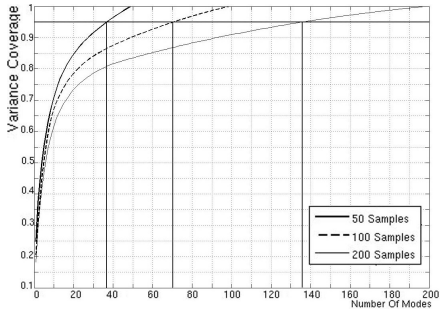

Fig. 2. 95% thresholded compactness plots for synthetic datasets

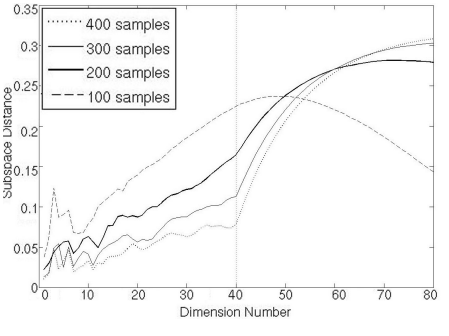

Fig. 3. Instability of PSS for synthetic datasets

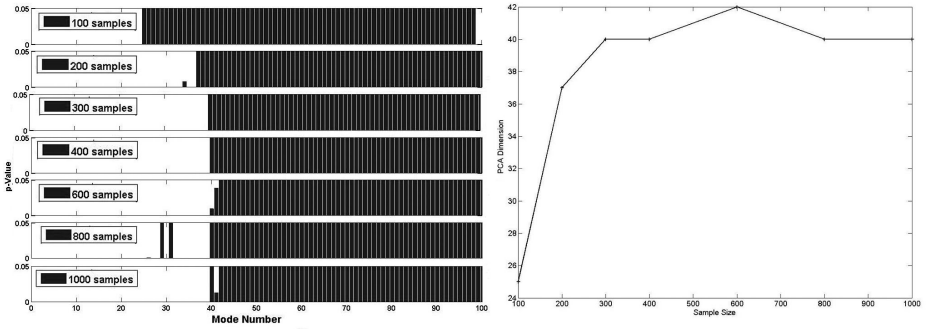

Fig. 4. *t*-Test Based stopping rule on synthetic cardiac datasets of different size

The method of Besse et al [14] was validated with synthetic datasets sized from 100 to 400. A plot of instability, measured as the distance between subspaces spanned by different replicates, is shown in figure 3. Although this method provides a visible indication of the correct number of modes to retain when the sample size is sufficiently large, it cannot identify the lower number of modes that should be retained when the sample size is insufficient.

5.2 Validation of *t*-Test Comparison with Noise

Our method was validated with synthetic datasets sized from 100 to 1000. Figure 4 shows the bar graphs of the p-Value (up to our 0.05 confidence level) from *t*-Tests for each mode trained from different sample sizes. The number of modes to retain versus the sample size is also shown. Our stopping rule does not have the tendency of going beyond 40 under large sample sizes. It also identifies a lower number of stable modes to retain for smaller sample sizes. It appears a sample size of around 300 is sufficient.

6 Results on Real Datasets

Figure 5 shows the results of sample sufficiency tests applied to the real cardiac dataset. The plot on the left side is the p-value for different replicate sizes. It shows that the number of modes determined by our stopping rule with 0.05 confidence level does not converge before the replicate size reaches the total sample size. This suggests that 319 samples are not enough. However, if an SSM is built from these samples, the number of modes to retain should be around 55.

Figure 6 shows sample size sufficiency tests on the real face datasets. For the 2D dataset, the plot obviously converges at 24 modes with 50 samples. With the 3D faces, the graph appears close to convergence at around 70 modes for the 150 samples. These results suggest both face datasets are sufficient.

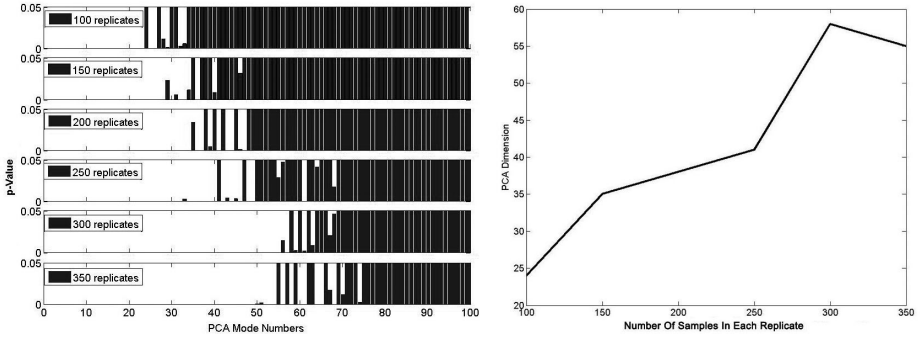

Fig. 5. Real heart dataset sufficiency test

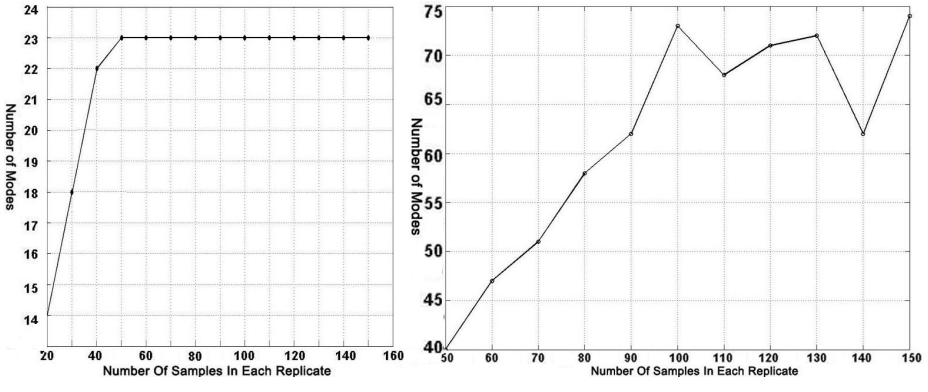

Fig. 6. Result of real face datasets sufficiency test. Left: 2D faces; Right: 3D faces.

7 Discussion

We propose a stopping rule for determining the number of modes to include for SSM based on mode direction stability. For a synthetic cardiac dataset generated with 40 real structural modes plus added noise our method converges correctly where conventional methods did not. We provide mode correspondence by minimising the distance between the principal spanned spaces rather than by the rank of their variances. We apply our stopping rule in a procedure we introduce to determine PCA sample size sufficiency. Results on real data suggest 319 samples are not sufficient for SSM of left ventricle (527 points) where both cardiac deformation and population variance are combined, but around 55 modes can be retained. However, 150 samples is sufficient for the 3D face meshes (500 points), where around 70 modes are retained. There is no trivial relationship between dimension of the shape vector, number of true anatomical modes and the required sample size. A more sophisticated method such as ours should be used instead.

It is hoped that our techniques will be adopted by those researchers working in SSM. Currently the number of samples used in most published studies is unlikely to be sufficient in the sense described in this paper. We hope that medical imaging researchers will gather more data and combine their sample sets in the aim of sufficient sample size for producing a standard, validated SSM for each organ. We aim to begin this process by building a significant database for two anatomical regions, the thorax and the lower abdomen. In time we intend to make the data and resulting models freely available to other research groups.

We provide what we believe to be the first principled test for sample sufficiency and determination of the number of modes to retain for SSM. Our method is also applicable to other applications of PCA and related fields.

References

1. Lee, S.L., Horkaew, P., Caspersz, W., Darzi, A., Yang, G.Z.: Assessment of shape variation of the levator ani with optimal scan planning and statistical shape modeling. Journal Of Computer Assisted Tomography 29, 154–162 (2005)
2. Rueckert, D., Frangi, A.F., Schnabel, J.A.: Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. IEEE Transactions On Medical Imaging 22, 1014–1025 (2003)
3. Lotjonen, J., Kivisto, S., Koikkalainen, J., Smutek, D., Lauerma, K.: Statistical shape model of atria, ventricles and epicardium from shortand long-axis MR images. Medical Image Analysis 8, 371–386 (2004)
4. Heimann, T., Wolf, I., Meinzer, H.P.: Active shape models for a fully automated 3d segmentation of the liver - an evaluation on clinical data. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 41–48. Springer, Heidelberg (2006)
5. MacCallum, R., Widaman, K., Zhang, S., Hong, S.: Sample size in factor analysis. Psychological Methods 4, 84–99 (1999)
6. Osborne, J., Costello, A.: Sample size and subject to item ratio in principal components analysis. Practical Assessment, Research and Evaluation 9(11) (2004)
7. Jackson, D.: Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74(8), 2204–2214 (1993)
8. Jolliffe, I.: Principal Component Analysis, 2nd edn. Springer, Heidelberg (2002)
9. Sinha, A., Buchanan, B.: Assessing the stability of principal components using regression. Psychometrika 60(3), 355–369 (2006)
10. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Training models of shape from sets of examples. In: Proc. British Machine Vision Conference, pp. 266–275. Springer, Berlin (1992)
11. Daudin, J., Duby, C., Trecourt, P.: Stability of principal component analysis studied by the bootstrap method. Statistics 19, 341–358 (1988)
12. Besse, P.: PCA stability and choice of dimensionality. Statistics& Probability 13, 405–410 (1992)
13. Babalola, K., Cootes, T., Patenaude, B., Rao, A., Jenkinson, M.: Comparing the similarity of statistical shape models using the bhattacharya metric. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 142–150. Springer, Heidelberg (2006)

14. Besse, P., de Falguerolles, A.: Application of resampling methods to the choice of dimension in PCA. Computer Intensive Methods in Statistics. In: Hardle, W., Simar, L. (eds.), pp. 167–176. Physica-Verlag, Heidelberg (1993)
15. Cootes, T.: The AR face database 22 point markup, http://www.isbe.man.ac.uk/~bim/data/tarfd_markup/tarfd_markup.html
16. University of Notre Dame Computer Vision Research Laboratory: Biometrics database distribution, http://www.nd.edu/~cvrl/UNDBiometricsDatabase.html
17. Papatheodorou, T.: 3D Face Recognition Using Rigid and Non-Rigid Surface Registration. PhD thesis, VIP Group, Department of Computing, Imperial College, London University (2006)