International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Goodness of Prediction for Principal Components of Shape

A Correlation Measure

Ja-Yeon Jeong · Surajit Ray · Qiong Han · Xiaoxiao Liu · Keith E. Muller · Stephen M. Pizer

Received: date / Accepted: date

Abstract The wide variety of statistical shape models available in image analysis and computer vision calls for some criteria and procedure to evaluate them for their effectiveness. In this paper, we introduce a formal correlation measure called goodness of prediction that allows evaluating statistical shape models in terms of their predictive power: the ability to describe an unseen member of the population. Most applications of statistical shape models such as segmentation and classification rely on their predictive power particularly heavily. The correlation measure is designed to analyze statistical shape models that use principal component analysis (PCA) to characterize shape variability. As some geometric shape representations like the m-rep do not form a vector space, the correlation measure initially defined in linear vector space is generalized to a nonlinear manifold by interpreting the measure in terms of geodesic distance in space. Through a systematic procedure for calculating the measure, we analyze the predictive power of statistical shape models as a function of training sample size. Our approach is demonstrated with two different shape representations: the m-rep and the point distribution model. Our experiment results show the usefulness and the benefit of our evaluation method.

Keywords Statistical shape model · Principal component analysis · Goodness of prediction · General linear multivariate model · Correlation · Shape model evaluation

Supported primarily by NCI P01 CA47982, R01 CA095749-01A1 and NIBIB EB000219

Ja-Yeon Jeong · Xiaoxiao Liu · Stephen M. Pizer
Medical Image Display & Analysis Group,
University of North Carolina, Chapel Hill, NC USA

Ja-Yeon Jeong
e-mail: jeong@cs.unc.edu, jayeon.j@gmail.com

Surajit Ray
Department of Mathematics and Statistics,
Boston University, Boston, MA USA

Qiong Han
Center for Visualization & Virtual Environments
Lexington, KY USA

Keith E. Muller
Epidemiology and Health Policy Research,
University of Florida, Gainesville, FL USA

1 Introduction

The main objective of statistical shape models is to provide a probability distribution on the shape that an object can take. Representation and analysis of geometric form is a difficult and challenging problem in image analysis and computer vision. Of special interest are the shape descriptions of 3D objects such as anatomical objects extracted from 3D medical image data or new types of 3D models in computer graphics applications. Many efforts have been made to come up with an effective geometric representation of 2D or 3D objects. Some of the well-known representations in the literature are the followings: active contour models (Malladi et al, 1995a) and its variant geodesic active contour models (Caselles et al, 1995), parametric deformable contour models (Staib and Duncan, 1992), diffeomorphisms from atlases (Joshi, 1997; Christensen et al, 1997), level set models (Malladi et al, 1995b; Tsai et al, 2003), point distribution models (PDM) (Cootes et al, 1995), spherical harmonic models (Brechbühler et al, 1995) and m-reps (Pizer et al, 2003).

Shape is often described in terms of the deformation from a template. Kendall (1977) originally described shape as the geometric information that remains after the variation in location, scaling, and rotational effect is accounted for. That is, two objects have the same shape if one object can be transformed into another object by these translation, scaling, and rotation transformations. So normally some align-

ment methods like Procrustes are applied to sample objects to remove the effects of similarity transformations existing in samples before any statistical estimation is done on training samples. It is assumed in this work that samples are aligned by a sensible alignment method.

Principal component analysis (PCA) has become a very popular method used to analyze shape variability. Cootes et al (1995); Bookstein (1999) were early users of PCA for shape analysis. The usefulness of PCA is two-fold: 1) decomposition of population variables into an efficient reparametrization of the variability observed on the training data and 2) dimension reduction of population variables that allows a focus on the subspace of the original space of population variables. Especially, #2 is a major advantage of PCA in shape analysis because most shape representations presented in the literature have very high dimensional feature spaces due to the complexity of object shape. On the other hand, available training samples are limited due to the cost and time involved in the manual segmentation of images. This kind of data is called high dimension, low sample size (HDLSS) in statistics. The measure we propose in this work applies to statistical shape models that use PCA as their method to describe shape variability.

Given a set of training samples from a population, PCA allows us to extract important directions (features) from these training samples and to use these features to describe new members of the population. The predictability of statistical shape models refers to this power of statistical shape models to predict a new member in the population.

Although there exist in statistics several criteria to judge the appropriateness of any dimension reduction technique, we will mainly concentrate on the criteria of predictability in view of the many practical applications of statistical shape models. In addition, we will touch on the questions of the interpretability and the stability of the extracted directions that are equally important as the predictability: Does the direction have a meaningful interpretation or are they mere mathematical objects?; How do these directions differ from sample to sample, and how many training samples do we need to get a stable estimate of the important directions?

1.1 Motivation, Previous Work, and New Measure

There are several properties that are desired for statistical shape models. First, for geometric representations of an object made from a tuple of spatially sampled primitives there needs to be reasonably good correspondence of the primitives across training cases. Poor correspondence can add noise to training samples that masks real variation of shape of an object, resulting in an unreliable estimation of shape probability distribution. Second, geometric representations need to be efficient so that they can describe shape of an object with minimal number of parameters. For example, to measure the geometric efficiency of representation for curves in 2D, Leonard (2007) introduced the concept of ε-entropy that is the minimum number of ε-balls required to cover the space of a totally bounded metric space and construct an adaptive coding scheme that allows codewords of varying lengths for shape elements in a non-compact space. On the basis of ε-entropy and the adaptive encoding scheme, she theoretically determined conditions in which the medial axis is more efficient than the boundary curves. She shows in her experiments that medial axis holds a tenable position as a shape model in 2D: for all but three out of the 2,322 2D-shapes she analyzed, the medial representation is more efficient. The efficiency of the geometric representations of shape can help to alleviate the HDLSS problem as well as to avoid the over-fitting problem. Third, an estimated shape probability distribution needs to be tight. Fourth, it needs to be unimodal since most statistical data analysis methods employed in shape analysis are based on the assumption of Gaussian distribution of the data. Fifth, a statistical shape model must be able to represent only real instances in the population of the object. Sixth, it must be able to describe a member of the population unseen in a training sample.

Among the few studies in which statistical shape models were evaluated, a key study done by Styner et al (2003) defines three criteria that can assess some of these properties and then compares correspondence of shape models on the basis of the three criteria. The three criteria are defined as follows: compactness as the ability to use a minimal set of parameters, generalization as the ability to describe instances outside of the training set, and specificity as the ability to represent only valid instances of the object in its population. Generalization ability is assessed by doing leave-one-out reconstruction and computing approximation errors of unseen models averaged over the complete set of trials. These measures are defined as functions of the number of shape parameters. Generalization ability and specificity are defined in the *ambient space*, where the models lie physically. In regard to these criteria, Styner et al (2003) examined and compared four methods: a manually initialized subdivision surface method for direct correspondence and three automatic methods - spherical harmonics, minimum description length, and minimum covariance determinant - for model-implied correspondence.

To compute the three criteria, (Styner et al, 2003) proposed to use an approximation error based on a small set of anatomical landmarks that a human expert selects manually on each object. The approximation error is defined as the mean absolute distance (MAD) between the manual landmarks and points corresponding to the same landmarks of the four shape models.

The approach of Styner et al (2003) in the evaluation of the correspondence methods is grounded in an important observation: statistical shape models of good correspondence

are highly likely to have good compactness, good generalization, and good specificity. In fact, these are qualities that a statistical shape model obtains as results of some optimizations to establish correspondence. They do not directly indicate the quality of correspondence of statistical shape models.

While these measures offer legitimate criteria to evaluate correspondence methods of different statistical shape models, they are short on the predictive power of the statistical shape models: the ability to describe unseen members of the population and to describe their frequency of occurrence. This ability of statistical shape models is critical since most applications of statistical shape models heavily rely on their predictive power. One example is model-based segmentation in maximum a posteriori (MAP) framework, which uses a prior distribution of shape of an object to extract the object from a new image. Another example is classification of an object on the basis of its shape, using trained shape prior distributions.

While goodness of fit is of special interest for analytic goals, goodness of prediction is of more importance for goals including generative models. Unfortunately in statistics goodness of prediction of PCA has received far less attention than goodness of fit. Muller (2007) presents a novel method to assess goodness of prediction for principal components. Muller first shows that PCA can be recast as a multivariate regression model by treating the observed variables as responses and principal directions as predictors. Then, goodness of prediction is derived from goodness of fit, a standard statistical measure for second moment accuracy. Finally, he proves that canonical correlations and related measures of association degenerate to constants and that the "univariate approach to repeated measures" test (average squared correlation, generalized variance explained) provides a simple and useful measure of association. He also suggests another measure - squared multiple correlation - to provide more detailed information. Among the several measures he proposes, in this work we adopt the average squared correlation measure to evaluate the predictive power of both linear and nonlinear statistical shape models. The detailed development of this measure is given in sections 2 and 3.

This correlation measure has a clear statistical interpretation in terms of the predictability of statistical shape models. This feature facilitates evaluation and comparison of different approaches to estimate a shape description or different statistical shape models. Furthermore, the average squared correlation is a simple direct measure defined in the shape feature space and is quick and easy to compute. In contrast, the generalization measure in (Styner et al, 2003) is an indirect measure defined in the ambient space where the models lie physically, and it takes a long time to compute.

2 Background

In this section we provide the background necessary to understand goodness of prediction of PCA in a multivariate regression setting. Section 2.1 explains the decomposition of a covariance matrix by PCA taken from (Muller, 2007). Section 2.2 describes the approximation of sample objects given by the mean and major principal coefficients. Section 2.3 gives the basic definitions of the general multivariate linear model and linear regression.

2.1 Decomposition of the Covariance Matrix

Let $Y = Y_1, \ldots, Y_N$ be N sample vectors from a p-variate distribution. Let $\mathbf{Y}$ be a $N \times p$ data matrix with Y_i' as rows, and let Σ be a corresponding $p \times p$ variance covariance matrix.

By PCA (equivalently spectral decomposition), Σ can be written as $\Sigma = \Upsilon \mathrm{D}(\lambda) \Upsilon'$, where $\mathrm{D}(\lambda)$ is a $p \times p$ diagonal matrix of nonnegative eigenvalues $\{\lambda_i\}$ for $i = 1 \ldots p$ and where Υ is a matrix of column eigenvectors of the nonnegative symmetric matrix Σ.

Most of the time $N \ll p$ due to HDLSS situation, i.e., the sample size is much smaller than the dimension of the shape feature. In general, some number of principal directions less than N, say p_a, covers most of the sample's variation, e.g., 80% or 90% of the total variation. Partly because the estimated principal directions explaining the smaller variation of the data are unreliable, we usually take the first p_a eigenvectors to approximate the covariance matrix. For $i > p_a$, the i-th eigenvector estimated from one sample is likely to be different from the i-th eigenvector estimated from another sample, and these later eigenvectors might not appear in the same order. Also, taking only the first p_a eigenvectors reduces the dimension of the original shape feature space considerably, which can be useful in applications of statistical shape models.

Let $p_b = p - p_a$. Considering p_a and p_b columns of matrices Υ and $\mathrm{D}(\lambda)$ in two partitions gives $\Upsilon = \begin{bmatrix} \Upsilon_a & \Upsilon_b \end{bmatrix}$, $\lambda = \begin{bmatrix} \lambda_a' & \lambda_b' \end{bmatrix}'$, and

$$\begin{aligned} \Sigma &= \Upsilon \mathrm{D}(\lambda) \Upsilon' \\ &= \begin{bmatrix} \Upsilon_a & \Upsilon_b \end{bmatrix} \begin{bmatrix} \mathrm{D}(\lambda_a) & 0 \\ 0 & \mathrm{D}(\lambda_b) \end{bmatrix} \begin{bmatrix} \Upsilon_a' \\ \Upsilon_b' \end{bmatrix} \\ &= \Upsilon_a \mathrm{D}(\lambda_a) \Upsilon_a' + \Upsilon_b \mathrm{D}(\lambda_b) \Upsilon_b' \\ &= \Phi_a \Phi_a' + \Phi_b \Phi_b' \\ &= \Phi \Phi' , \end{aligned} \tag{1}$$

where $\Phi = \begin{bmatrix} \Phi_a & \Phi_b \end{bmatrix}$, $\Phi_a = \Upsilon_a \mathrm{D}(\lambda_a)^{1/2}$, and $\Phi_b = \Upsilon_b \mathrm{D}(\lambda_b)^{1/2}$. Without loss of generality $\Upsilon' \Upsilon = I_p$ and $\{\lambda_i | \, i = 1 \ldots p\}$ are sorted from largest to smallest. Hence if $\mathrm{rank}(\Sigma) = p_a$, then $\lambda_b = 0$ and $\Sigma = \Upsilon_a \mathrm{D}(\lambda_a) \Upsilon_a' = \Phi_a \Phi_a'$. $\Phi_b \Phi_b' \approx 0$ and $\Sigma_Y \approx \Upsilon_a \mathrm{D}(\lambda_a) \Upsilon_a'$ are assumed when the covariance matrix is approximated with the first $p_a \ll p$ components.

2.2 PCA for Statistical Shape Analysis

In statistical shape analysis, the N rows of $\mathbf{Y}$ correspond to people or images, and the p columns of $\mathbf{Y}$ correspond to features in shape space. With a $p \times 1$ mean shape feature vector μ and an $N \times 1$ column vector $\mathbf{1}_N$ of 1's, the full set of component scores is $(\mathbf{Y}-\mathbf{M})\Upsilon$, where $\mathbf{M} = \mathbf{1}_N\mu'$. Component scores for retaining p_a components are computed by $(\mathbf{Y}-\mathbf{M})\Upsilon_a$. Let $\mathbf{Y}_c$ be the approximating set of component scores $(\mathbf{Y}-\mathbf{M})\Upsilon_a$. Approximating the data with the components gives

$$\begin{aligned}\mathbf{Y} \approx \mathbf{Y}_a &= \mathbf{M} + \mathbf{Y}_c\Upsilon_a' \\ &= \mathbf{M} + (\mathbf{Y}-\mathbf{M})\Upsilon_a\Upsilon_a' ,\end{aligned} \tag{2}$$

with the $N \times p$ matrix $\mathbf{Y}_a$ of rank $p_a \ll p$, while $\Upsilon_a\Upsilon_a'$ is $p \times p$ and of rank $p_a \ll p$.

2.3 General Linear Multivariate Model

The multivariate linear model allows two or more responses to be measured on each independent sampling unit. The definition of the general linear multivariate model given in (Muller and Stewart, 2006) is as follows.

Definition 1 *A general linear multivariate model (GLM)* $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ *with primary parameters* $\mathbf{B}$, Σ *has the following assumptions:*

1. *The rows of the* $N \times p$ *random matrix* $\mathbf{Y}$ *correspond to the independent sampling units, that is, they are mutually independent with* $\mathbf{Y}_i = row_i(\mathbf{Y})$.
2. *The* $N \times q$ *design matrix X has rank*$(\mathbf{X}) = r \le q \le N$ *and is fixed and known without appreciable error, conditional on knowing the sampling units, for data analysis.*
3. *The parameter* $q \times p$ *matrix* $\mathbf{B}$ *is fixed and unknown.*
4. *The mean of* $\mathbf{Y}$ *is* $E(\mathbf{Y}) = \mathbf{XB}$.
5. *The mean of* $\mathbf{E}$ *is* $E(\mathbf{E}) = \mathbf{0}$.
6. *The rows of response matrix* $\mathbf{Y}$ *has finite covariance matrix* Σ_Y*, which is fixed, unknown, and positive definite or positive semidefinite.*

In the multivariate regression model, $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y})$, and
$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{E}(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
which is the maximum likelihood estimator of $\mathbf{B}$. More details about regression analysis and linear models can be found in (Muller and Stewart, 2006; Kleinbaum et al, 1997; Timm, 2002). The theory of multivariate analysis can be found in (Muirhead, 1982; Arnold, 1981).

3 Goodness of Prediction

In most applications of statistical shape models, a mean shape and modes of shape variation (eigenvectors, principal directions) estimated from a set of training models are used to approximate a shape in a new image. Considering how estimated shape statistics are used in applications, the prediction accuracy of estimated shape models can be assessed properly by using estimates from one set to predict shapes in a different set of models. We call a set of training models used for estimating shape statistics a training set and the different set of models a test set.

In this section, we first describe a modified interpretation of the original approach that is proposed in (Muller, 2007) to meet our need. Our focus here is measuring goodness of prediction of the estimated covariance (second moment accuracy), not of the estimated mean (first moment accuracy). We simplify the original approach to that end. Then we present the average squared correlation as the measure of association between the training set and the test set, that is, the goodness of prediction of the estimated covariance. The term "goodness of prediction" here is used in this restricted sense. Full details and proofs of the original approach can be found in (Muller, 2007) and in a series of forthcoming papers.

3.1 PCA as Multivariate Regression

Let $\mathbf{Y}_t$ and $\mathbf{Y}_s$ be the training data and the test data matrices respectively. The subscripts t and s indicate the training and the test set. Our objective is to measure the degree to which the probability distribution estimated from $\mathbf{Y}_t$ describes the probability distribution that appears in $\mathbf{Y}_s$. In this process, the mean estimated from $\mathbf{Y}_t$ is considered to be a true mean. i.e. $E(Y) = \hat{M}_t$. Hat indicates the random estimator of a parameter. We assume that the training mean $\widehat{\mathbf{M}}_t$ is already subtracted from the two data matrices $\mathbf{Y}_t$ and $\mathbf{Y}_s$ in the rest of this subsection.

With $p_a \ll p$ approximating eigenvectors $\widehat{\Upsilon}_{at}$ estimated from a training set $\mathbf{Y}_t$, the component scores $\mathbf{Y}_{cs|t}$ of $\mathbf{Y}_s$ on $\widehat{\Upsilon}_{at}$ are $\mathbf{Y}_s\widehat{\Upsilon}_a$. Then, a multivariate multiple regression model can be formulated by treating the test data matrix $\mathbf{Y}_s$ as responses and the component scores $\mathbf{Y}_{cs|t}$ of the test data as predictors, i.e., $\mathbf{X} = \mathbf{Y}_{cs|t} = \mathbf{Y}_s\widehat{\Upsilon}_{at}$ with

$$\mathbf{Y}_s = \mathbf{XB}_{as|t} + \mathbf{E} . \tag{3}$$

The inverse of $\mathbf{X}'\mathbf{X}$ exists when the rank of the data matrix $\mathbf{Y}_s$ is at least p_a. The least squares estimates of $\mathbf{B}_{as|t}$ and the responses $\mathbf{Y}_s$ are respectively

$$\begin{aligned}&\widehat{\mathbf{B}}_{as|t} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}_s = (\widehat{\Upsilon}_{at}'\mathbf{Y}_s'\mathbf{Y}_s\widehat{\Upsilon}_{at})^{-1}\widehat{\Upsilon}_{at}'\mathbf{Y}_s'\mathbf{Y}_s \quad \text{and} \\ &\widehat{\mathbf{Y}}_s = \mathbf{X}\widehat{\mathbf{B}}_{as|t} = \mathbf{Y}_s\widehat{\Upsilon}_{at}(\widehat{\Upsilon}_{at}'\mathbf{Y}_s'\mathbf{Y}_s\widehat{\Upsilon}_{at})^{-1}\widehat{\Upsilon}_{at}'\mathbf{Y}_s'\mathbf{Y}_s .\end{aligned} \tag{4}$$

When $\mathbf{Y}_t = \mathbf{Y}_s$, that is, when we take the training data as responses (thus omitting subscripts t, s in the derivation be-

low), $\widehat{\mathbf{B}}_{as|t}$ and $\widehat{\mathbf{Y}}_s$ can be simplified as follows:

$$\begin{aligned}\widehat{\mathbf{B}}_{as|t} &= (\widehat{\Upsilon}_a'\mathbf{Y}'\mathbf{Y}\widehat{\Upsilon}_a)^{-1}\widehat{\Upsilon}_a'\mathbf{Y}'\mathbf{Y} \\ &= (\widehat{\Upsilon}_a'(N-1)\widehat{\Sigma}\widehat{\Upsilon}_a)^{-1}\widehat{\Upsilon}_a'\mathbf{Y}'\mathbf{Y} \\ &= (\widehat{\Upsilon}_a'(N-1)\widehat{\Upsilon}D(\widehat{\lambda})\widehat{\Upsilon}\widehat{\Upsilon}_a)^{-1}\widehat{\Upsilon}_a'(N-1)\widehat{\Upsilon}D(\widehat{\lambda})\widehat{\Upsilon} \\ &= \mathrm{D}(\lambda_a)^{-1}\mathrm{D}(\widehat{\lambda}_a)\widehat{\Upsilon}_a' = \widehat{\Upsilon}_a' ,\end{aligned}$$

$$\widehat{\mathbf{Y}}_s = \mathbf{X}\widehat{\mathbf{B}}_{as|t} = \mathbf{Y}\widehat{\Upsilon}_a\widehat{\Upsilon}_a' ,$$

which is the usual approximation of the data as described in (3) for zero mean. Our goal is to measure the association between the estimates $\widehat{\mathbf{Y}}_s$ of test data set in (4) and the test data $\mathbf{Y}_s$ itself.

3.2 Measure of Association: Second Moment Accuracy

To measure of association between $\widehat{\mathbf{Y}}_s$ and $\mathbf{Y}_s$, we follow the approach suggested in (Muller, 2007). Let $\mathbf{S}_h$ be the sample covariance matrix of $\hat{\mathbf{Y}}_s$ and $\mathbf{S}_y$ be the sample covariance matrix of $\mathbf{Y}_s$. In the multivariate linear regression model, $\mathbf{S}_h$ is the covariance under the regression, and $\mathrm{tr}(\mathbf{S}_h)$ is the amount of variance explained by the regression [$\mathrm{tr}(\mathbf{X})$ indicates a trace of $\mathbf{X}$]. $\mathbf{S}_y$ is divided into two parts: $\mathbf{S}_y = \mathbf{S}_h + \mathbf{S}_e$. $\mathrm{tr}(\mathbf{S}_y)$ is the amount of total variance, and $\mathrm{tr}(\mathbf{S}_e)$ is the amount of unexplained variance left in $\mathbf{Y}_s$. In other words, $\mathrm{tr}(\mathbf{S}_e)$ represents the amount of variance in $\mathbf{Y}_s$ that remains after accounting for the linear effect of $\mathbf{X}$.

Our goodness of prediction measure is the ratio of variation explained according to the principle of goodness of fit. A univariate approach to repeated measure of goodness of fit $\widehat{\tau}$ (when $t = s$) is given by

$$\widehat{\tau} = \frac{\mathrm{tr}\,(\mathbf{S}_h)}{\mathrm{tr}\,(\mathbf{S}_h + \mathbf{S}_e)} , \tag{5}$$

and is equivalent to the proportion of generalized variance controlled as shown below:

$$\widehat{\tau} = \frac{\mathrm{tr}\left[\widehat{\Upsilon}_a\mathrm{D}(\widehat{\lambda}_a)\widehat{\Upsilon}_a'\right]}{\mathrm{tr}\left[\widehat{\Upsilon}_a\mathrm{D}(\widehat{\lambda}_a)\widehat{\Upsilon}_a' + \widehat{\Upsilon}_b\mathrm{D}(\widehat{\lambda}_b)\widehat{\Upsilon}_b'\right]} = \frac{\sum_{k=1}^{p_a}\widehat{\lambda}_k}{\sum_{k=1}^{p}\widehat{\lambda}_k} .$$

The property of decomposition of the total variation into $\mathbf{S}_e$ and $\mathbf{S}_h$ makes this measure very attractive as it now can be interpreted as the amount of variation explained by the retained directions, whereas $\mathbf{S}_e$ measures the magnitude of the remaining variation.

In general, a goodness of fit test $\widehat{\tau}$ can be calculated as

$$\widehat{\tau} = \frac{\mathrm{tr}\left((\widehat{\mathbf{Y}} - \widehat{\mathbf{M}})'(\widehat{\mathbf{Y}} - \widehat{\mathbf{M}})\right)}{\mathrm{tr}\left((\mathbf{Y} - \widehat{\mathbf{M}})'(\mathbf{Y} - \widehat{\mathbf{M}})\right)} = \frac{\sum_{i=1}^{N}(\widehat{Y}_i - \widehat{\mu})^2}{\sum_{i=1}^{N}(Y_i - \widehat{\mu})^2} . \tag{6}$$

Our goodness of prediction ρ^2, a measure of association, is derived from the goodness of fit by applying $\widehat{\tau}$ to the proposed regression model (3) (when $t \neq s$) and can be written as follows:

$$\widehat{\rho}^2 = \frac{\mathrm{tr}\left((\widehat{\mathbf{Y}}_s - \widehat{\mathbf{M}}_t)'(\widehat{\mathbf{Y}}_s - \widehat{\mathbf{M}}_t)\right)}{\mathrm{tr}\left((\mathbf{Y}_s - \widehat{\mathbf{M}}_t)'(\mathbf{Y}_s - \widehat{\mathbf{M}}_t)\right)} = \frac{\sum_{i=1}^{N}(\widehat{Y}_{si} - \widehat{\mu}_t)^2}{\sum_{i=1}^{N}(Y_{si} - \widehat{\mu}_t)^2} eq : tau(7)$$

where $\widehat{\mu}_t$ is the sample mean estimated from a training set. The numerator and the denominator of $\widehat{\rho}^2$ can be factored into two parts (the derivation is in the Appendix):

$$\begin{aligned}\widehat{\rho}^2 &= \frac{\sum_{i=1}^{N}(\widehat{Y}_{si} - \widehat{\mu}_s)^2 + N(\widehat{\mu}_s - \widehat{\mu}_t)^2}{\sum_{i=1}^{N}(Y_{si} - \widehat{\mu}_s)^2 + N(\widehat{\mu}_s - \widehat{\mu}_t)^2} \\ &= \frac{\mathrm{tr}(\mathbf{S}_h) + N(\widehat{\mu}_s - \widehat{\mu}_t)^2}{\mathrm{tr}(\mathbf{S}_h + \mathbf{S}_e) + N(\widehat{\mu}_s - \widehat{\mu}_t)^2} .\end{aligned} \tag{8}$$

The reason we choose to evaluate the deviation of both $\widehat{Y}_{si}$ and Y_{si} from the mean $\widehat{\mu}_t$ estimated from a training set instead of the mean $\widehat{\mu}_s$ estimated from a test set is to be true to the applications of statistical shape models. In the applications of statistical shape models, $\widehat{\mu}_t$ becomes a template for a new object since there is no way of estimating the mean of objects that are not included in the training set. We can still interpret $\widehat{\rho}^2$ as the amount of variation of a test set explained by the retained principal directions estimated by a training set as long as the mean estimated from a training set is close to the mean estimated from a test set, that is, $\widehat{\mu}_s \approx \widehat{\mu}_t$.

ρ^2 has a value that is between 0 and 1. High values of ρ^2 indicate that the retained modes of shape variation estimated from a training set capture the shape variation of new models well because the amount of total variance explained by the factors in the regression model yields a high value of $\mathrm{tr}(\mathbf{S}_h)$ as a proportion of the fixed total variation. On the other hand, the estimated modes of shape variation that explain less shape variation of new models give lower ρ^2.

The theoretical distribution of ρ^2 can be of special interest for further analysis of the measure of association between $\widehat{\mathbf{Y}}_s$ and $\mathbf{Y}_s$: whether ρ^2 has a unimodal or bimodal distribution; whether its mean or median gives a better summary statistic; whether the distribution is symmetric or skewed. However, we leave this topic as future research.

3.3 Procedure for Iterative Calculation of ρ^2

The goal is to analyze the predictive power of statistical shape models as the size of training sample changes. For each training sample size we calculate quartiles of the distribution of ρ^2, as follows.

INPUT: A sample pool $\mathscr{P}$ of N objects of p geometric features

SETTING: Set the following parameters:

1) The test sample size α

2) The list $\mathscr{L}$ of nt training sample sizes
$\mathscr{L} = \{\beta_i | \beta_1 < \ldots < \beta_{nt}, \beta_{nt} + \alpha \leq N, i = 1, \ldots, nt\}$
3) The number of retained principal directions p_a
4) The number of repetitions R

OUTPUT: ρ^2 values calculated for R times at each training sample size β_i. $R \times nt$ number of ρ^2 values is computed.

PROCEDURE:

For $i = 1, \ldots, nt$	
	For repetitions $= 1, \ldots, R$
Step1	Randomly select two disjoint sets:
	a test set $\mathscr{S}_s$ of size α,
	a training set $\mathscr{S}_t$ of size β_i
Step2	Compute $\widehat{\mu}_t$, eigenvectors $\widehat{\Upsilon}_{at}$ from $\mathscr{S}_t$
Step3	Construct three $\alpha \times p$ matrices:
	a data matrix $\mathbf{Y}_s$ from $\mathscr{S}_s$
	an estimate of the response matrix $\widehat{\mathbf{Y}}_s$ in (4), and
	a mean matrix $\widehat{\mathbf{M}}_t$,
Step4	Compute ρ^2 using (7)

The reason that $\mathscr{S}_s$ and $\mathscr{S}_t$ must be disjoint is to reduce the sampling bias.

We use a box plot to visualize the output ρ^2 values of the procedure. As illustrated in Fig. 4, a box plot has lines at the lower quartile (25th percentile), median (50th percentile), and upper quartile values (75th percentile). Whiskers extend from each end of the box to the adjacent values in the data; by default, the maximum whisker length is 1.5 times of the interquartile range. (The difference between the upper and lower quartiles is called the interquartile range.) Outliers, displayed with a '+' sign, are data with values beyond the ends of the whiskers.

4 Two Statistical Shape Models: PDM & M-reps

In section 5, we will illustrate the use of the procedure just described to evaluate two statistical shape models: PDM and m-reps. In this section we first give brief descriptions of these models. To produce consistency between the PDMs and the m-reps that we analyze, we also describe the means by which a boundary representation (b-rep) PDM is derived from an m-rep.

4.1 PDM

The "point distribution model" is a well-known statistical shape model introduced by Cootes et al (1995). Each object represented by a PDM is captured by a set of boundary points. These boundary points are manually placed in a consistent manner on each of training models and are automatically aligned to minimize the sum of squared distances between corresponding points across the training models. After the alignment, the PDM is obtained by estimating the average positions of the points and major modes of variation by PCA on a training set.

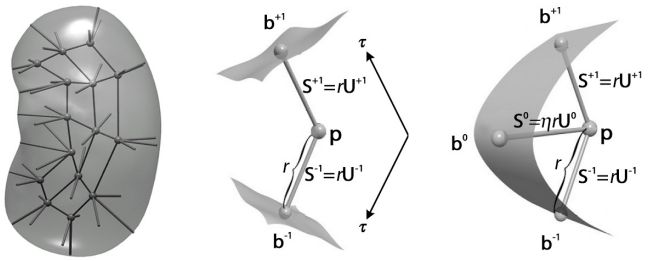


Fig. 1: Left: a single figure m-rep for a kidney and the object boundary implied by it. Middle: an internal atom of two spokes $\mathbf{S}^{+1/-1}$, with τ to parameterize the object interior along the spokes. Right: an end atom with an extra bisector spoke $\mathbf{S}^0$.

4.2 M-rep

An m-rep is an extension of the Blum medial locus (Blum and Nagel, 1978); in the extension the medial locus forms the primitive description. A geometric object of the simplest form is represented by a single continuous medial sheet with boundary. Each sheet corresponds to what we call a figure. On an m-rep, at each point on the sheet two equal length spokes extend to the implied boundary, where the tangent plane of the sheet at the point bisects the spokes. The point with its two spokes is called a *medial atom* (Fig. 1-left). A medial atom is thus a 4-tuple $\{\mathbf{p}, r, \mathbf{U}^{+1/-1}\}$, consisting of the hub position $\mathbf{p} \in \mathbb{R}^3$, the spoke length $r \in \mathbb{R}^+$, and the two spoke directions as two unit vectors $\mathbf{U}^{+1/-1} \in S^2$ (Fig. 1-middle). The two spokes of each atom are $\mathbf{S}^{+1/-1} = r\mathbf{U}^{+1/-1}$.

We have developed and implemented three representations of m-reps. The one we have used primarily, called the discrete m-rep, represents each sheet by a discrete grid of atoms, from which we can interpolate a continuous sheet. A discrete m-rep is formed by sampling the medial sheet over a spatially regular lattice to form a mesh of medial atoms. The medial atoms on the edge of the medial sheet correspond to crests of the object boundary. Such an end atom adds a spoke $\mathbf{S}^0$ of length ηr and direction $\mathbf{U}^0$ bisecting $\mathbf{U}^{+1/-1}$, where $\eta \in \mathbb{R}^+$ is a crest sharpness parameter (Fig. 1-right).

4.3 Derivation of B-reps from M-reps

Given an m-rep figure, a subdivision surface method (Thall, 2004) is presently used to generate a smooth object surface. Thall (2004) modified the Catmull–Clark subdivision surface algorithm to interpolate the boundary positions and the normals implied by the spokes.

Deriving a b-rep from an m-rep involves calculating the spoke ends of the m-rep. That is, for the spokes $\mathbf{S}^{+1/-1}$ in each medial atom of the m-rep, its spoke ends are computed as $\mathbf{b}^{+1/-1} = \mathbf{p} + r\mathbf{U}^{+1/-1}$ (Fig. 1-middle). The crest spoke ends of the end medial atoms are computed as $\mathbf{b}^0 = \mathbf{p} + \eta r\mathbf{U}^0$ (Fig. 1-right). Although boundary points other than spoke end points can be sampled from the surface, we decided to choose only the spoke and bisector spoke end surface points from medial atoms so as not to add redundant dimensions to the derived b-rep. Thus, if an m-rep has n interior atoms and m end atoms, the dimension of the corresponding b-rep will be $6\times n+9\times m$ while that of the m-rep is $8\times n+9\times m$. These two representations are not equivalent in the sense that m-reps cannot be constructed from b-reps. B-reps lack necessary nonlinear information of normal directions at the spoke ends to compute the hub positions of the medial atoms. In spite of the inequality, we use the b-rep as the linear representation corresponding to the m-rep.

4.4 Statistics of M-reps

Fletcher et al (2004) realized that because the spoke directions in medial atoms are values on the unit sphere, medial atoms, and thus the tuples of medial atoms that make up discrete m-reps, can be understood to live not on a flat feature space but on a curved feature space known as a "symmetric space" by mathematicians. Fletcher et al (2004) developed a generalized version of PCA called principal geodesic analysis (PGA) for probability density estimation of geometric entities that form a symmetric space. PGA involves computing a Fréchet mean on the actual curved manifold via a gradient descent method, and then doing PCA on the linear space which is tangent at the Fréchet mean. The eigenmodes projected back down onto the curved space make the principal geodesics.

5 Application of ρ^2 on Models in Linear Space

We tested the goodness of prediction measure (7) through the procedure described in section 3.3 on two data sets. One set is made up of synthetic objects, b-reps of simulated ellipsoid m-reps. The other set is made up of real anatomical objects, b-reps of m-reps fitted to right hippocampi. We began with the synthetic ellipsoid data because it allows us to generate as many samples as we want and control the kind of deformations in the samples, thus providing a means of checking properties of the ρ^2 such as the convergence.

The simulation of ellipsoid m-reps and the experiment results are described in sections 5.1 and 5.2. The training of right hippocampus binaries and the experiment results are described in the following sections 5.3 and 5.4.

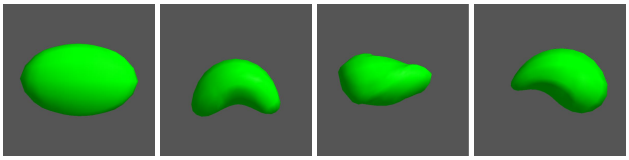


Fig. 2: From left to right, a base ellipsoid m-rep, randomly bent, twisted, and tapered ellipsoid m-reps are shown reflecting the nonlinear variation in the population.

5.1 Simulated Ellipsoid M-reps

We have a simulation program (Han et al, 2007) for generating random ellipsoid deformations as illustrated in Fig. 2. The program applies the composition of random bending, random twisting, and random tapering to a base ellipsoid m-rep $\mathbf{M}_0$ sampled from the Blum medial axis of a standard ellipsoid centered at the origin.

Starting from the base ellipsoid m-rep $\mathbf{M}_0 = \{\mathbf{p}_i, r_i, \mathbf{U}_i^{+1/-1} \mid i = 1,...,N\}$, where $\mathbf{p}_i = (x_i, y_i, z_i)$ and N is the number of medial atoms of $\mathbf{M}_0$, the three deformations are applied to the medial atoms of $\mathbf{M}_0$ in the order of bending, twisting, and tapering.

1. Bending: each atom is translated by $\delta|x_i|^2$ along the $z-$axis, and then rotated around the $y-$axis by the angle between $(1,0,0)$ and the tangent vector $(1,0,2\delta|x_i|)$;
2. Twisting: each atom is rotated around the tangent vector $(1,0,2\delta x_i)$ of a parabola $(x,0,\delta x^2)$ at x_i. The rotation angle is εx_i;
3. Tapering: the radius r_i is scaled by a factor of $e^{\zeta x_i}$, where $|x_i|$ is the distance from the the center of the ellipsoid to each atom along the $x-$axis,;

where $\delta, \varepsilon, \zeta$ are three independent random variables following Gaussian distributions with zero means. Each set of $(\delta_j, \varepsilon_j, \zeta_j)$ sampled independently from Gaussian distributions determines a deformed ellipsoid m-rep $\mathbf{M}_j$ where j is the index to the series of deformed ellipsoid m-reps.

5.2 Experiments on Simulated Ellipsoid B-reps

The base ellipsoid m-rep $\mathbf{M}_0$ consists of a 3×7 grid of medial atoms, where 16 of them are end atoms and 5 of them are internal atoms. The dimension of the m-rep features is $16*9+5*8=184$, and that of its corresponding b-rep features is $16*9+5*6=174$. The radial lengths of the principal axes of the base ellipsoid $\mathbf{M}_0$ are $(0.2625, 0.1575, 0.1181)$ with a ratio of 10:6:4.5. The three parameters $\delta, \varepsilon, \zeta$ were sampled from three independent Gaussian distributions of standard deviations 1.5, 1.047, and 2.12 respectively. We generated 5000 warped ellipsoid m-reps and made warped ellipsoid b-reps from those m-reps.

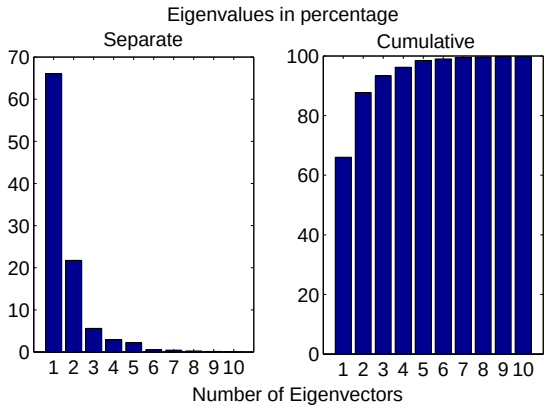


Fig. 3: Two bar graphs of the first 10 eigenvalues in percentage estimated from 5000 simulated warped ellipsoids b-reps. Left: each eigenvalue per mode. Right: cumulative eigenvalues.

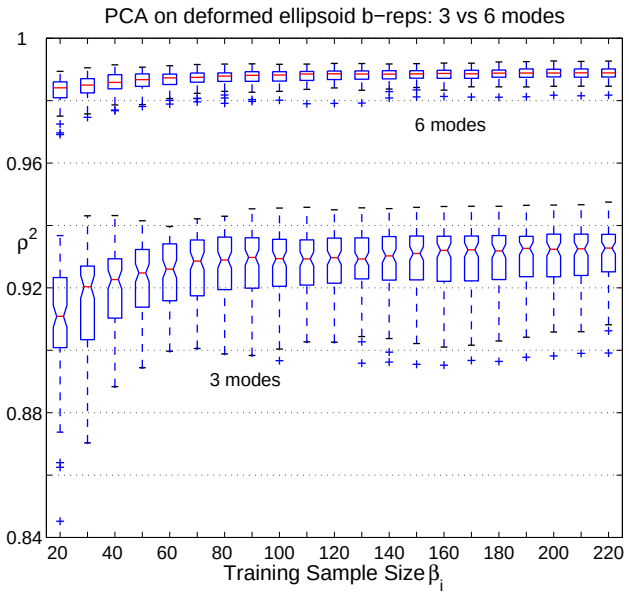


Fig. 4: Box plots of ρ^2 vs. training sample sizes β_i for PCA on warped ellipsoid b-reps with 3 and 6 eigenmodes. 100 independent trials R for each training sample size β_i from 20 to 220 were done using a fixed test sample of size $\alpha = 100$.

Fig. 3 shows the variances of the first ten modes of variation estimated from the 5000 simulated warped ellipsoid b-reps. As there are three independent transformations - bending, twisting, and tapering - applied to the base ellipsoid, PCA on the b-reps of 174 feature dimensions produces three major principal eigenmodes with two trailing eigenmodes. The first three principal eigenmodes explains more than 90% of the total variance.

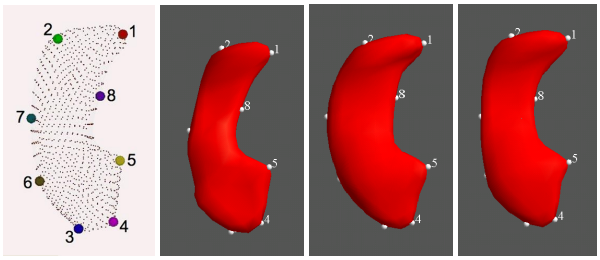


Fig. 5: From left to right, landmark points of hippocampus and three fitted m-reps are shown.

Fig. 4 shows two box plots of ρ^2 where the test set size α is 100. One box plot is for the probability distribution captured by 3 principal modes, and the other is for that captured by 6 principal modes. For each of these box plots the training sample sizes β_i range from 20 to 220 with the increment of 10. A training sample of size β_i and a test sample of size α are randomly drawn 100 times. ρ^2 is computed at each draw.

As expected, the values of ρ^2 are higher at 6 modes than at 3 modes. ρ^2 reaches near convergence at approximately 80 training samples for 3 modes and at about 60 training samples for 6 modes. Since there are only three true deformations in these synthetic data, we can see that ρ^2 converges at a training size much smaller than the feature space dimension. We can also see that ρ^2 values in the box plot (Fig. 4) correspond to the cumulative estimated eigenvalues in Fig. 3. The median ρ^2 starts around 0.91 and converges to around 0.93 for 3 modes, and the median ρ^2 starts a little bit above 0.98 and converges to around 0.99 for 6 modes. Moreover, the range of ρ^2 values (interquartile range) is more spread out at 3 modes than at 6 modes, which indicates that the variation of new cases is captured more stably by 6 modes than by 3 modes.

5.3 Right Hippocampus M-reps

We ran the procedure on real anatomical object models, 290 right hippocampi b-reps. To have such a large number of samples, we pooled manually segmented binaries of hippocampi from many people regardless of their ages, their mental conditions, or their medications. An m-rep template was fitted into the binaries to extract hippocampus m-rep models through several steps.

The goal of fitting (Merck et al, 2008) is to find a model $\mathbf{M}$ that best describes the target object in a given binary image $\mathbf{I}$. In our m-rep framework, fitting is an optimization process minimizing an objective function $F_{obj}(\mathbf{M}, \mathbf{I})$ that is a weighted sum of two data match functions and of two geometry penalties.

The two data match functions are an image match function F_{img} and a landmark match function F_{lmk}. F_{img} forces

the implied surface of $\mathbf{M}$ match with the boundary voxels of the binary. F_{lmk} enforces explicit correspondences between the landmarks in the binary image either identified by experts or by some programs and the matching surface points of the model.

The two geometry penalties are an irregularity penalty function F_{reg} and an illegality penalty function F_{leg}. F_{reg} penalizes non-uniform spacing of the grid of medial atoms and non-uniform changes in spoke lengths and spoke directions of medial atoms. F_{reg} implicitly contributes to establish correspondence of medial atoms across the training cases. F_{leg} is a penalty unique to a discrete m-rep, making use of a shape operator called $\mathbf{S}_{rad}$ introduced by Damon (2003) for medial geometry. The illegality penalty function tries to prevent local self-intersections or creases from happening in the implied surface of a discrete m-rep.

Given 8 landmarks per binary (Fig. 5-left), we followed steps described as follows to fit hippocampus binaries.

Hippocampi Fitting Steps:

Step1	Fit an initial template model to binaries with high weight on F_{lmk}, F_{reg}, and F_{leg}, resulting in roughly fitted models.
Step2	Discard bad fits resulting from Step1. Train a mean model and a shape space with the remaining fits. Fit the mean model by deforming over the shape space using the same configuration as Step1.
Step3	Refine the fits from Step2 by deforming each medial atom separately with low weight on F_{lmk}, F_{reg}, and zero weight on F_{leg}.

The fitted m-reps are used by PGA to produce the final shape space.

5.4 Experiments on Right Hippocampus B-reps

The hippocampus m-rep consists of a 3×8 grid of medial atoms, where 18 of them are end atoms and 7 of them are internal atoms. The dimension of m-reps is $18*9+6*8=210$, and that of the corresponding b-reps is $18*9+6*6=198$. Three panels on the right side of Fig. 5 show the boundary surfaces of the 3 fitted right hippocampus m-reps.

Fig. 6 shows the variances of the first 40 modes of variation estimated from the 290 right hippocampus b-reps. Unlike the simulated warped ellipsoids, PCA on the right hippocampus produces principal eigenmodes of slowly decreasing variances. It takes more than 30 modes to reach 90% of the total variance.

Fig. 7 shows two box plots of ρ^2 when α =100 and R = 100. One box plot is for the probability distribution captured by 18 principal modes, and the other is for that captured by 36 principal modes. The training sample sizes β_i range from 40 to 200 with the increment of 10.

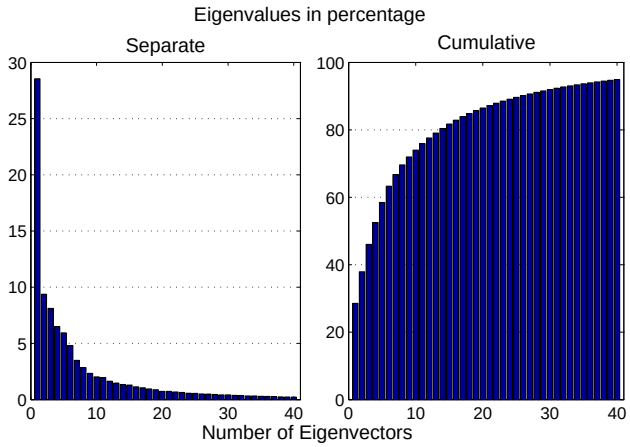


Fig. 6: Two bar graphs of the first 40 eigenvalues in percentage estimated from 290 right hippocampus b-reps. Left: each eigenvalue per mode. Right: cumulative eigenvalues.

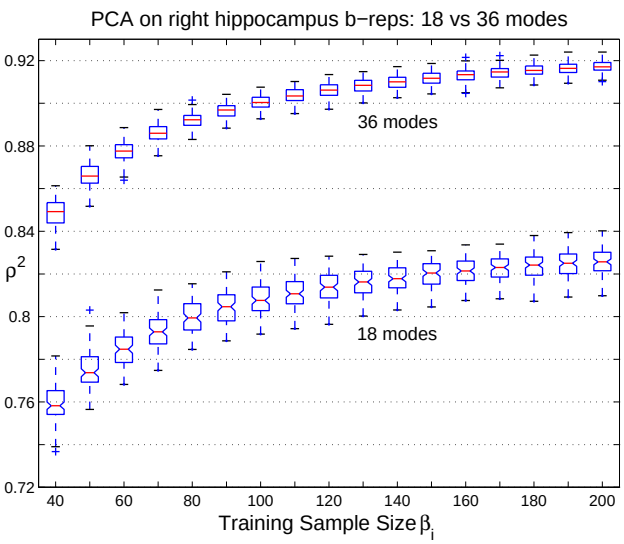


Fig. 7: Box plots of ρ^2 vs. training sample sizes β_i for PCA on right hippocampus b-reps with 18 and 36 eigenmodes. 100 independent trials for each training sample size were done using a fixed test sample of size $\alpha = 90$.

As expected, the values of ρ^2 are higher at 36 modes than at 18 modes. The interquartile range of ρ^2 values is slightly more spread out at 18 modes than at 36 modes as well, which indicates that the estimated shape subspace captures the shape variation in the population more stably at 36 modes than at 18 modes. As for the convergence of ρ^2, it is hard to determine from Fig. 7 the training sample size at which ρ^2 begins to converge within the range of the training sample size tested. However, we can judge the trade-off between the training sample size and the increase of ρ^2 value from Fig. 7. For example, we can see that for 18 modes about

100 samples are enough to estimate the shape variation since ρ^2 value increases very slowly after $\beta_i = 100$.

6 Distance Measures for ρ^2 Evaluation

It is important to assess the validity of our new measure, ρ^2. While it measures the closeness of estimated populations to the real population in the feature space, it is desirable to verify whether the measure's indications in the feature space concur with what happens in the ambient space. To that end, in this section we present other measures based on the standard surface-to-surface distance in the ambient space. With both the correlation measure and the distance measures as function of training sample sizes, the relation between the predictive power measured in the shape feature space and in the ambient space is analyzed.

For a shape model Y, let $B(Y)$ be the set of vertices of its corresponding b-rep in triangle 3D meshes. Let $S(Y)$ be its surface. The mean absolute surface-to-surface distance of two shape models Y_i and Y_j (Aspert et al, 2002) is defined as follows:

$$d_{mad}(Y_i,Y_j) = \frac{1}{N_i+N_j}\Big(\sum_{v\in B(Y_i)} \min_{p'\in S(Y_j)} \|v-p'\| + \sum_{v'\in B(Y_j)} \min_{p\in S(Y_i)} \|v'-p\|\Big),$$

where N_i and N_j indicate the numbers of points in $B(Y_i)$ and $B(Y_j)$ respectively, and $\|*\|$ represents the usual Euclidean norm. The surface-to-surface distances from $S(Y_i)$ to $S(Y_j)$ and from $S(Y_j)$ to $S(Y_i)$ are averaged since they are not equal.

We compute two kinds of surface-to-surface distance measures. One measure is the minimum of all squared mean absolute distances between a model Y in a test set $\mathscr{S}_s$ and all models in a training set $\mathscr{S}_t$:

$$d_m^2(Y,\mathscr{S}_t) = \min_{Y'\in\mathscr{S}_s}\left(d_{mad}^2(Y,Y')\right).$$

Another measure is the squared mean absolute distance between a model Y in a test set $\mathscr{S}_s$ and its projection Y_a (3) on the shape space estimated from a training set $\mathscr{S}_t$:

$$d_p^2(Y,\mathscr{S}_t) = d_{mad}^2(Y,Y_a).$$

We compute these two distance measures following the same procedure for ρ^2 computation described in section 3.3. In applying the procedure, we need a summary statistic for these distance measures to see their change as the training sample size increases since at each iteration d_m^2 and d_p^2 are computed for every model in a test set. The output of the procedure at the end of each iteration is α numbers of d_m^2 and d_p^2 values. So, we use the median as a summary statistic:

$$D_*^2(\mathscr{S}_t,\mathscr{S}_s) = \operatorname*{median}_{Y\in\mathscr{S}_s}\left(d_*^2(Y,\mathscr{S}_t)\right),$$

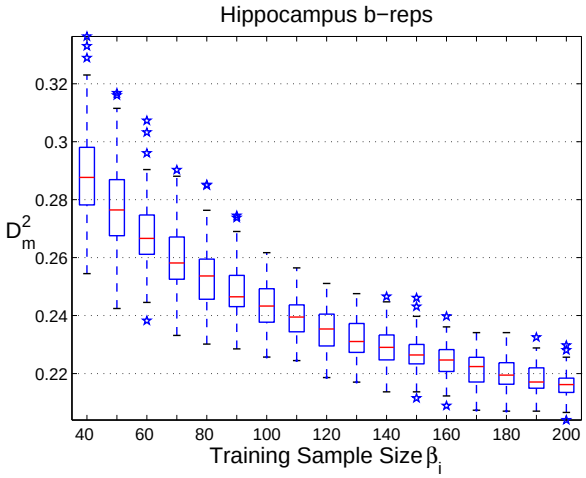


Fig. 8: A box plot of D_m^2 vs. training sample sizes β_i with the same setting as in section 5.4.

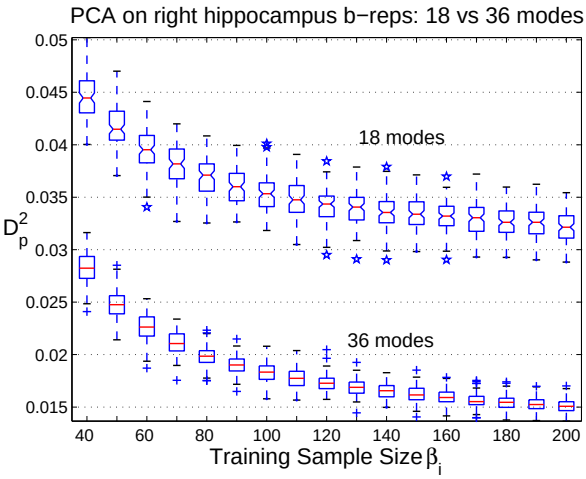


Fig. 9: Two box plots of D_p^2 vs. training sample sizes β_i for PCA on right hippocampus b-reps with 18 and 36 eigenmodes with the same setting as in section 5.4.

where $*$ indicates subscripts m or p for d_m^2 or d_p^2 respectively.

D_p^2 is more interesting to us than D_m^2. D_m^2 indicates how close the test models are to training models and depends only on the choice and the size of $\mathscr{S}_t$. On the other hand, D_p^2 indicates the ability of the estimated shape space to approximate a new model in its ambient space. D_p^2 depends not only on the choice and the size of $\mathscr{S}_t$ but also on the number of retained principal directions from $\mathscr{S}_t$.

6.1 Application of D_*^2 to Right Hippocampus B-reps

We tested the two distances measure on the right hippocampus b-reps. Figs. 8 and 9 show box plots for D_m^2 and D_p^2 respectively. Since D_m^2 is independent of the number of re-

tained eigenmodes, Fig. 8 has only one box plot of D_m^2. As the training sample size increases, we can see that values and the interquartile range of D_m^2 decreases. Fig. 9 clearly indicates that with more principal modes and larger training samples we get more accurate approximations of new instances in the population, which is consistent with what we have observed in the ρ^2 plot (Fig. 7). The tests with D_*^2 distance measures provide empirical evidence that our proposed goodness prediction measure defined in the shape feature space reflects the shape variation appearing in the ambient space.

7 Goodness of Prediction ρ_d^2 for Curved Manifolds

7.1 Two Possible Extensions of ρ^2

Our goodness of prediction measure (7) (equivalently (8)) does not directly apply to models such as m-reps that live in a nonlinear curved manifold. However, as already indicated in the equation (7), the numerator and the denominator of (7) can be interpreted via distances from an estimated training mean $\widehat{\mu}_t$. The following two equations simply show rewriting of (7) and (8) in terms of a more general metric function d:

1. from (7)

$$\widehat{\rho}_d^2 = \frac{\sum_{i=1}^N d^2(\widehat{Y}_{si}, \widehat{\mu}_t)}{\sum_{i=1}^N d^2(Y_{si}, \widehat{\mu}_t)}\,, \tag{9}$$

2. from (8)

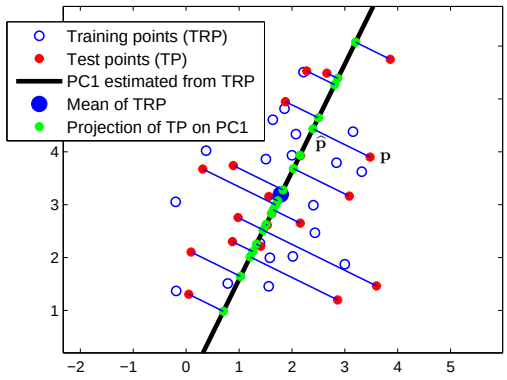
$$\widehat{\rho}_d^2 = \frac{\sum_{i=1}^N d^2(\widehat{Y}_{si}, \widehat{\mu}_s) + N d^2(\widehat{\mu}_s, \widehat{\mu}_t)}{\sum_{i=1}^N d^2(Y_{si}, \widehat{\mu}_s) + N d^2(\widehat{\mu}_s, \widehat{\mu}_t)}\,. \tag{10}$$

These two expressions for $\widehat{\rho}_d^2$ show a natural extension of the distance decomposition of the total variance $\mathbf{S}_y$ into $\mathbf{S}_h$ and $\mathbf{S}_e$. However, the equality of these two expressions, which holds for linear feature spaces, does not strictly hold for nonlinear spaces. Also, for nonlinear spaces the numerator and the denominator in equation (9) cannot be interpreted as $\mathrm{tr}(\mathbf{S}_h) + N(\widehat{\mu}_s - \widehat{\mu}_t)^2$ and $\mathrm{tr}(\mathbf{S}_h + \mathbf{S}_e) + N(\widehat{\mu}_s - \widehat{\mu}_t)^2$ respectively, as they can for linear spaces (equations (7) and (8)).

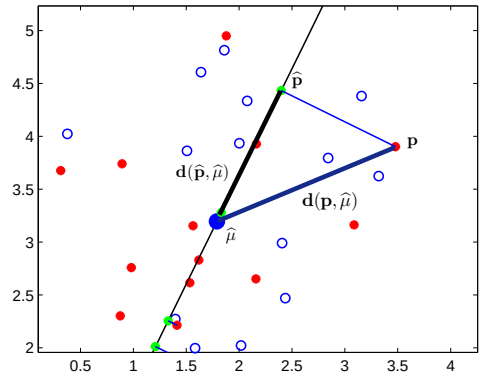
7.2 Riemannian Symmetric Spaces and Tangent Space at a Point of a Manifold

The nonlinear shape models that we deal with in this paper are elements of curved differential manifolds called "Riemannian symmetric spaces" with an associated *Riemannian metric* $d(\cdot,\cdot)$. The main property of differential manifolds is that locally they behave like Euclidean space. Thus, for every point p in a given differential manifold $M \in \mathbb{R}^d$, a linear subspace that best approximates a neighborhood of p in M can be associated. The linear subspace is called a *tangent space at the point p* and is denoted by T_pM.

Two key functions that map points between T_pM and M are the *Riemannian exponential map* and the *Riemannian log map*. The Riemannian exponential map at $p \in M$ denoted by $Exp_p : T_pM \to M$ is a diffeomorphism in a neighborhood $U \subset T_pM$ mapping a point $x \in U$ to a point $Exp_p(x) \in M$ along the geodesic from p to $Exp_p(x)$. A geodesic in a manifold M is the shortest smooth curve segment between two points in M. In the Euclidean space a straight line is the geodesic path between two points. Its inverse mapping



(a)



(b)

Fig. 10: (a) Two sets of 20 random vectors chosen from a multivariate normal distribution. Hollow blue dots indicate training points from which the mean and the first principal direction are estimated. Red dots indicate test points projected onto the first principal direction going through the estimated mean. Green dots on the first principal directors are projections of test points (red dots), that is, approximations of test points with the training mean and the first principal component. (b) Zoomed plot of (a). This plot shows the distances in the numerator and denominator of the correlation measure formula for one test point $\mathbf{P}$.

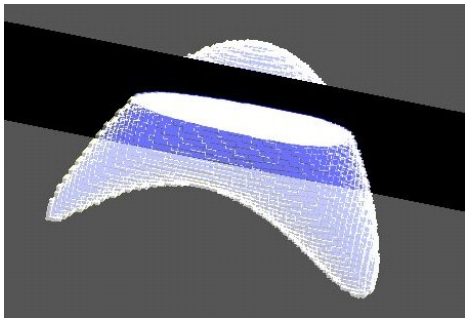


Fig. 11: A $(128 \times 128 \times 128)$ bent and tapered binary ellipsoid

is called the Riemannian log map and denoted by $Log_p : Exp_p(U) \rightarrow T_pM$.

It is useful to select a metric on T_pM such that distances to p on the Riemannian manifold are equal to those on T_pM. That is, distances from a point x on the tangent plane to p denoted as $||x||$ are equal to the geodesic distance $d(Exp_p(x), p)$ on the manifold.

7.3 ρ_d^2 for Nonlinear Shape Models

(9) and (10) provide the two possible extensions of ρ^2 to a curved manifold suggested in section 7.1. Recall that (9) and (10) are not equivalent in the curved manifold. We choose to use (9) as the goodness of prediction for nonlinear shape models in the curved manifold because it has a nice interpretation in the tangent space: the geodesic distance $d^2(Y, \widehat{\mu}_t)$ for $Y \in M$ is equal to $||Log_{\widehat{\mu}_t}(Y)||$ in $T_{\widehat{\mu}_t}M$. Also, the decomposition of the total variance $\mathbf{S}_y$ holds in the tangent space $T_{\widehat{\mu}_t}M$.

Fig. 10 shows the graphical view of the equation (9). One set of points indicates a training set, and the other set of points indicates a test set. The line is the first principal direction going through a mean estimated from the training set. In the subspace (line in this example) spanned by the first principal direction, points in a test set are approximated by their projections on the first principal direction. The denominator in equation (9) is the sum of the distances from the training mean to each point in a test set, and the numerator is the sum of the distances from the training mean to projections of points in a test set onto the subspace.

8 Application of ρ_d^2 on Models in Nonlinear Space

8.1 Deformed Binary Ellipsoids

A synthetic test population was created from an ellipsoid deformed from the original ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1$ by a

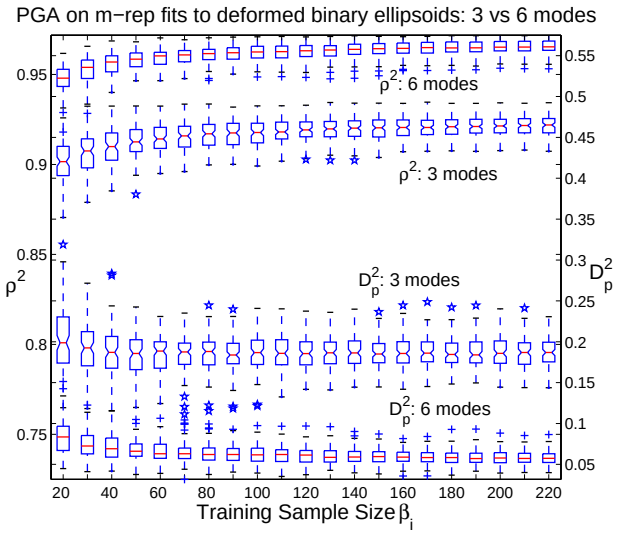


Fig. 12: Two box plots of ρ_d^2 vs. training sample sizes β_i for PGA on m-reps fits to deformed binary ellipsoids with 3 and 6 eigenmodes (left y-axis). 100 independent trials for each training sample size were done using a fixed test sample of size $\alpha = 100$. Two box plots of D_p^2 vs. training sample sizes β_i with the same setting (rigth y-axis).

set of 1000 diffeomorphisms of the form (Han et al, 2007)

$$\Psi_{\delta,\varepsilon,\zeta}(x,y,z) \equiv \begin{bmatrix} x \\ e^{\zeta x}(y\cos(\varepsilon x) + z\sin(\varepsilon x)) \\ e^{\zeta x}(y\cos(\varepsilon x) + z\sin(\varepsilon x) + \delta x^2 \end{bmatrix},$$

where $\delta, \varepsilon,$ and ζ are parameters to control bending, twisting, and tapering respectively. δ, ε, and ζ follow $N(0,(1.5)^2)$, $N(0,(1.047)^2)$, and $N(0,(2.12)^2)$ respectively, where $N(\mu,\sigma^2)$ is the normal distribution with mean μ and standard deviation σ. In this experiment, we again used a standard ellipsoid with axis lengths of $(0.2625, 0.1575, 0.1181)$ centered at the origin, i.e., $a = 0.2625$, $b = 0.1575$, $c = 0.1181$. These parameters were sampled 1000 times from the three normal distributions, and the resulting deformations were applied to the standard ellipsoid. The results were 1000 $(128 \times 128 \times 128)$ binary images of warped ellipsoids. Fig. 11 shows a case of deformed ellipsoid binaries.

The deformed binary ellipsoids are different from the deformed ellipsoid m-reps described in section 5.1. The difference lies in whether the deformations are applied to medial atoms of the base ellipsoid m-rep or to the ambient space of the base ellipsoid.

The fitting of m-reps to the binary ellipsoids follows the same steps taken for fitting the right hippocampus (section 5.3). Six landmarks are used for the fitting: two end points of three ellipsoid axes.

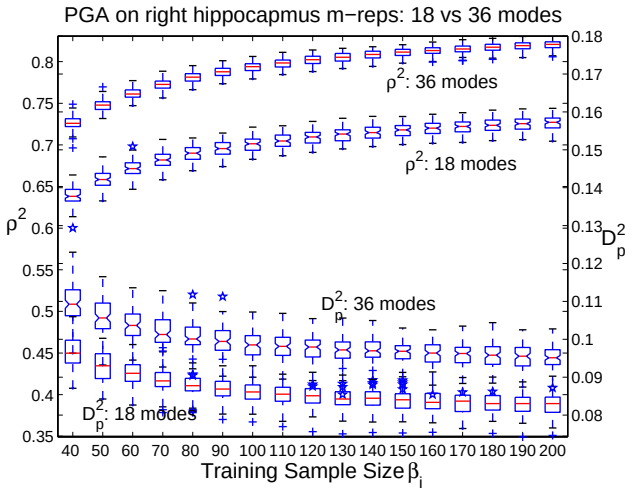


Fig. 13: Two box plots of ρ_d^2 vs. training sample sizes β_i for PGA on m-reps fits to deformed binary ellipsoids with 18 and 36 eigenmodes (left y-axis). 100 independent trials for each training sample size were done using a fixed test sample of size $\alpha = 90$. Two box plots of D_p^2 vs. training sample sizes β_i with the same setting (right y-axis).

8.2 Experiment on M-rep Fits to Deformed Binary Ellipsoids

The fitted ellipsoid m-reps consist of a 3×7 grid of medial atoms. The settings of the procedure here are the same as the settings for the simulated warped ellipsoid b-reps.

The results here are consistent with those for the simulated warped ellipsoid b-reps. As shown in Fig. 12, the values of ρ_d^2 are higher at 6 modes than at 3 modes, and the values of D_p^2 are smaller at at 6 modes than at 3 modes. Both measures begin to converge at approximately 60 training samples. The convergence starts at the training size much smaller than the feature space dimension since there are only 3 true deformations in population of this synthetic data. The interquartile range of ρ^2 and D_p^2 values is more spread out at 3 modes than at 6 modes, as well.

8.3 Experiment on Right Hippocampus M-rep

ρ_d^2 was tested through the procedure (section 3.3) on the 290 right hippocampus m-reps (section 5.3). Fig. 13 shows the results of the procedure as two box plots of ρ_d^2 and D_p^2 when α =90 and R =100. One box plot is for the probability distribution captured by 18 principal modes, and the other is for that captured by 36 principal modes. The training sample sizes β_i range from 40 to 200 with the increment of 10.

The results of ρ_e^2 on the right hippocampus m-reps are very similar to the results of ρ^2 on the corresponding b-reps. The values of ρ_d^2 are higher at 36 modes than at 18 modes and

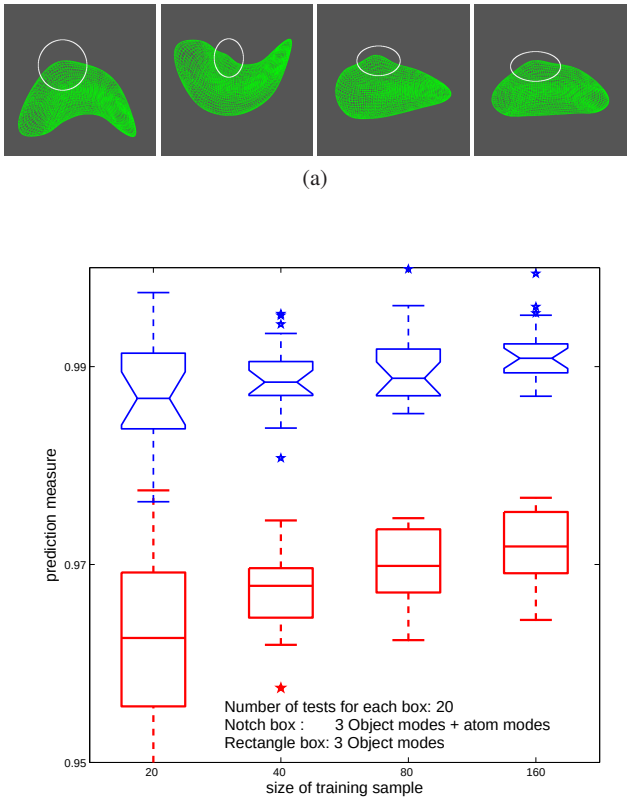


Fig. 14: (a) Four simulated ellipsoids with local deformation (bump - circle). (b) Two box plots of ρ_d^2 vs. training sample sizes β_i for multiscale PGA on 3 object modes followed by 4 atom modes.

the values of D_p^2 are smaller at at 36 modes than at 18 modes. It is still difficult to determine when ρ_d^2 and D_p^2 converges within the range of the training sample size tested.

8.4 Evaluation of a Coarse-to-fine Shape Prior

Our goodness of prediction measure ρ_d^2 has also proven to be effective in the evaluation of the multiscale shape priors. Liu et al (2008) proposed a coarse-to-fine shape prior for the probabilistic segmentation to enable local refinement in the m-rep framework, which aims to capture the small level of detail components of the object shape variation that PCA-based approximations are likely to miss out, as pointed out in (Nain et al, 2005) and (Davatzikos et al, 2003). The approach developed in (Liu et al, 2008) is to decompose the variation space into two scale levels: object-scale and atom-scale prior. As the object-scale prior describes the shape changes of the object as a whole, the atom-scale prior captures the small level of detail components of the shape changes that are left from the object-scale shape space. The atom-

scale prior is constructed using the residual space, that is, the remainders of the actual shape space from the shape subspace described by the object-scale prior.

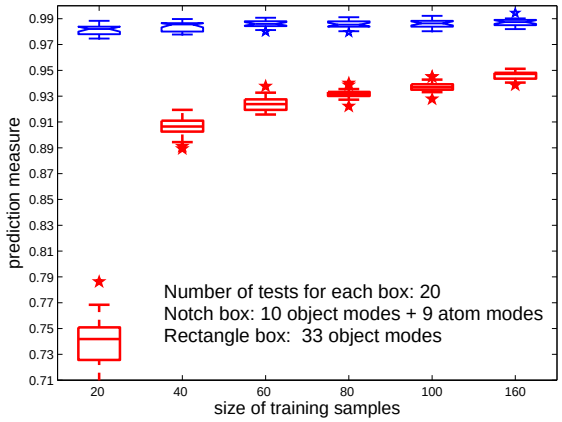
To show the robustness of the estimated coarse-to-fine shape prior, Liu et al (2008) adopted our procedure and viewed ρ_d^2 against training sample sizes. $\widehat{Y}_{si}$ in (9) becomes the multiscale projection that refines the object-scale approximation of Y_{si} by adding each atom-scale approximation in the residual space.

The two scale shape priors were first tested on a synthetic data set of 1000 warped ellipsoids m-reps. These model ellipsoids shown in Fig. 14a were produced by applying a relatively small amount of local perturbation on the hub position of one selected atom on top of the three global deformations described in section 5.1.

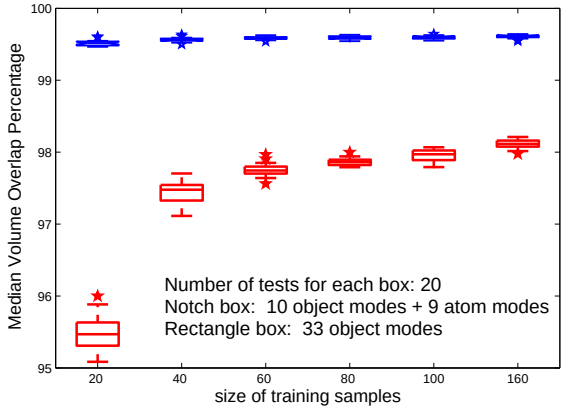
Fig. 14b shows the two box plots of ρ_d^2 computed using the object-scale shape prior vs the coarse-to-fine shape prior. One box plot shows the ρ_d^2 values for the shape prior captured by three object-scale principal modes, and the other shows those for the multiscale shape prior captured by three object-scale principal modes followed by a few eigenmodes of the selected atom. ρ_d^2 cleary indicates the improvement of the predictability when the coarse-to-fine shape prior is used.

Liu et al (2008) then examined the advantage of the coarse-to-fine shape prior on simulated real anatomical structures. The simulation took 51 eigenmodes estimated from the 290 well-fitted right hippocampus m-reps described in section 5.3 and did Gaussian random sampling on the 51 eigenmodes. The 51 eigenmodes explain 95% of the total variation observed in the 290 fitted right hippocampus m-reps. With this data, the object-scale shape prior (the 33 object eigenmodes) was compared with the coarse-to-fine shape prior (the 10 object eigenmodes followed by the 9 atom eigenmodes) using the ρ_d^2 and the volume overlap measure. The Dice Similarity Coefficient (Crum et al, 2006) on the volumes of the test models and the corresponding object-scale and atom-scale approximations was used as the volume overlap measure. Fig. 15a shows the comparison.

We can clearly see not only the benefit of the coarse-to-fine shape prior over the object shape prior by ρ_d^2 but also the consistency between the volume overlap measure and the ρ_d^2 measure. In addition to the distance measures we introduced in section 6 to show the validity of our ρ_d^2 measures in ambient space, this consistency demonstrated between the the volume overlap measure and the ρ_d^2 measure confirms that our goodness of prediction measure ρ_d^2 defined in the feature space does indeed reflect what happens in the ambient space.



(a)



(b)

Fig. 15: (a) Two box plots of ρ_d^2 vs. training sample sizes β_i for multiscale PGA on simulated hippocampus m-reps with 10 object modes followed by 9 atom modes. (b) Corresponding box plots of volume overlap between the test models and their projected models.

9 Conclusion & Discussion

Our work has been motivated by the need to have a quantitative measure to evaluate the predictive power of a statistical shape model. We proposed a novel statistical correlation measure ρ^2 called the goodness of prediction. It is designed to judge the predictive power of a PCA-based statistical shape model to analyze the major shape variation observed in the training sample. The measure is formally derived by interpreting PCA in terms of the multivariate linear regression model, and it is interpreted as a ratio of the variation of a new data explained by the retained principal directions estimated from a training data.

The major shift of our perspective in evaluating a statistical shape model is that we are more concerned about evaluating how well an estimated shape model fits the new

data rather than how well an estimated shape model fits the parameters of the population probability distribution. It is a reasonable stance to take considering that the ability of an estimated shape probability distribution to describe a new object is the major concern in most applications of statistical shape models.

The novelty of the measurement lies in its being computed using two disjoint sets of samples. One set of samples is used for the estimation of a shape space by PCA, and the other set of samples is used for the evaluation of the estimated shape space, which exactly reflects the situation happening in the applications of statistical shape models.

Moreover, we proposed a procedure to compute the goodness of prediction against the training sample sizes, which allows inferring the training sample size that ensures capturing certain amount of shape variation present in objects unseen from training samples. The procedure was experimentally evaluated on synthetic warped ellipsoid b-reps and real anatomical right hippocampus b-reps. The results were visualized using box plots that show the median and the interquartile range of the ρ^2 values of a large number of independent trials.

We tested a slight variation of the proposed procedure although we did not report the results in this paper. We further aligned each training set to tighten the distribution before doing PCA (Step2 in section 3.3) since we were concerned the use of pre-aligned data might bring a bias in our estimation of ρ^2. This additional alignment on the right hippocampus data did not bring any noticeable difference in ρ^2 values. Thus we did not include this additional alignment step in the procedure, concluding that the pre-alignment on the pooled data is sufficient to remove any non-shape related transformations in the models.

We extended ρ^2 for linear shape representations to ρ_d^2 for nonlinear shape representations that form Riemannian symmetric spaces. ρ_d^2 carries the same statistical meaning as ρ^2 because the geodesic distance from a training mean in a Riemannian space is equivalent to the Euclidean distance in a corresponding tangent space at the training mean. ρ_d^2 was also tested on the synthetic ellipsoid m-reps and real anatomical right hippocampus m-reps. The results of ρ_d^2 on these m-rep data are consistent with those of ρ^2 on the corresponding b-rep data.

The ρ_d^2 measure was also empirically verified by two surface distance measures and a volume overlap measure to prove that ρ_d^2 really reflects what happens in the ambient space where the model lies.

Our experiments showed the usefulness and the versatility of the procedure. It yields an appropriate training sample size for some retained number of eigenvalues and judging the trade-off between training sample size and the amount of variation explained by the retained number of eigenvalues. However, the major drawback of the proposed procedure is that it needs many data samples to see the convergence of median ρ_d^2 values and of the interquartile range of ρ_d^2 values with respect to the training sample size . Another issue with the procedure is the number of independent trials R. In Bayesian statistics, R must be several thousand to have a statistical significance, which is impractical in many applications.

In spite of these disadvantages, the measure itself is easy and fast to compute, and its statistical interpretation is simple to understand. It is also quite flexible to apply for any statistical shape model. We already showed in other work that applying the measure to evaluate multiscale shape priors is straightforward. We hope this measure will be found to be useful in evaluating other statistical shape models and will motivate further exploration in the evaluation of statistical shape models.

Appendix

We can break the numerator of (7) into two parts as follows:

$$\begin{aligned}(\widehat{Y}_{si}-\widehat{\mu}_t)^2 = \quad & (\widehat{Y}_{si}-\widehat{\mu}_s)^2+(\widehat{\mu}_s-\widehat{\mu}_t)^2 \\ & +2(\widehat{Y}_{si}-\widehat{\mu}_s)(\widehat{\mu}_s-\widehat{\mu}_t)\ ,\end{aligned} \tag{11}$$

$$\begin{aligned}\sum_{i=1}^{N}(\widehat{Y}_{si}-\widehat{\mu}_t)^2 &= \sum_{i=1}^{N}(\widehat{Y}_{si}-\widehat{\mu}_s)^2+\sum_{i=1}^{N}(\widehat{\mu}_s-\widehat{\mu}_t)^2 \\ &= \mathrm{tr}(\mathbf{S}_h)+N(\widehat{\mu}_s-\widehat{\mu}_t)^2\ ,\end{aligned}$$

where $\widehat{\mu}_s$ is the sample mean estimated from a test set. The cross product term of (11) disappears after summation since $\sum_{i=1}^{N}(\widehat{Y}_{si}-\widehat{\mu}_s)=\sum_{i=1}^{N}\widehat{Y}_{si}-N\widehat{\mu}_s=0$ and $\widehat{\mu}_s-\widehat{\mu}_t$ is constant. Similarly, its denominator is decomposed into two parts as follows:

$$\begin{aligned}\sum_{i=1}^{N}(Y_{si}-\widehat{\mu}_t)^2 &= \sum_{i=1}^{N}(Y_{si}-\widehat{\mu}_s)^2+\sum_{i=1}^{N}(\widehat{\mu}_s-\widehat{\mu}_t)^2 \\ &= \mathrm{tr}(\mathbf{S}_y)+N(\widehat{\mu}_s-\widehat{\mu}_t)^2 \\ &= \mathrm{tr}(\mathbf{S}_h+\mathbf{S}_e)+N(\widehat{\mu}_s-\widehat{\mu}_t)^2\ .\end{aligned}$$

Acknowledgements The authors would like to thank J. Stephen Marron for lending us his insight and having scientific discussions with us during the development of this work. The authors also thank Martin Styner and Christine Xu for providing the hippocampus data and the MeshValmet program to compute surface distance and especially all members in MIDAG for their constant help to data preparation.

References

Arnold SF (1981) The Theory of Linear Models and Multivariate Analysis. Wiley, New York

Aspert N, Santa-Cruz D, Ebrahimi T (2002) Mesh: Measuring errors between surfaces using the hausdorff distance. In: IEEE International Conference in Multimedia and Expo, vol 1, pp 705–708

Blum H, Nagel R (1978) Shape description using weighted symmetric axis features. Pattern Recognition 10:167–180

Bookstein FL (1999) Morphometric Tools for Landmark Data: Geometry and Biology. Cambridge University Press

Brechbühler C, Gerig G, Kübler O (1995) Parametrization of closed surfaces for 3-d shape description. Comput Vis Image Underst 61(2):154–170, DOI http://dx.doi.org/10.1006/cviu.1995.1013

Caselles V, Kimmel R, Sapiro G (1995) Geodesic active contours. IEEE Computer Society Press, Los Alamitos, CA, pp 694–699

Christensen GE, Joshi SC, Miller MI (1997) Volumetric transformation of brain anatomy. IEEE Trans Med Imaging 16(6):864–877

Cootes T, Taylor C, Cooper D, Graham J (1995) Active shape models – their training and application. Computer Vision and Image Understanding 61(1):38–59

Crum WR, Camara O, Hill DLG (2006) Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Transactions on Medical Imaging 25(11):1451–1461

Damon J (2003) Smoothness and geometry of boundaries associated to skeletal structures i: Sufficient conditions for smoothness. Annales de Institut Fourier 53(6):1941–1985

Davatzikos C, Xiaodong T, Shen D (2003) Hierarchical active shape models, using the wavelet transform. IEEE Transactions on Medical Imaging 22(3):414–423

Fletcher PT, Lu C, Pizer SM, Joshi S (2004) Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Transactions on Medical Imaging 23(8):995–1005

Han Q, Merck D, Levy J, Villarruel C, Chaney E, Pizer SM (2007) Geometrically proper models in statistical training. In: Proc. of Information Processing in Medical Imaging, (Nico Karssemeijer and Boudewijn Lelieveldt, eds.), vol 4584, pp 751–762

Joshi S (1997) Large deformation diffeomorphisms and gaussian random fields for statistical characterization of brain submanifolds. PhD thesis, Dept. of Electrical Engineering, Sever Institute of Technology, Washington Univ.

Kendall DG (1977) The diffusion of shape. Advances in Applied Probability 9(3):428–430, URL http://www.jstor.org/stable/1426091

Kleinbaum DG, Kupper LL, Muller KE, Nizam A (1997) Applied Regression Analysis and Multivariable Methods. Duxbury Press

Leonard K (2007) Efficient shape modeling: -entropy, adaptive coding, and boundary curves -vs- blums medial axis. International Journal of Computer Vision 74(2):183–199, DOI 10.1007/s11263-006-0010-3, URL http://dx.doi.org/10.1007/s11263-006-0010-3

Liu X, Jeong JY, Levy JH, Saboo R, Chaney EL, Pizer SM (2008) A large-to-fine-scale shape prior for probabilistic segmentations using a deformable m-rep. In: IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA)

Malladi R, Sethian JA, Vemuri BC (1995a) Shape modeling with front propagation: A level set approach. IEEE Transactions on Pattern Analysis and Machine Intelligence (2):158–175

Malladi R, Sethian JA, Vemuri BC (1995b) Shape modeling with front propagation: A level set approach. IEEE Trans Pattern Anal Mach Intell 17(2):158–175

Merck D, Tracton G, Saboo R, Levy J, Chaney E, Pizer S, Joshi S (2008) Training models of anatomic shape variability. Medical Physics 35(7)

Muirhead RJ (1982) Aspects of Multivariate Statistical Theory. Wiley

Muller KE (2007) Goodness of prediction for principal components, including high dimension, low sample size, work in progress

Muller KE, Stewart PW (2006) Linear Model Theory: Univariate, Multivariate, and Mixed Models. Wiley-Interscience

Nain D, Haker S, Bobick A, Tannenbaum A (2005) Multiscale 3d shape analysis using spherical wavelets. In: Medical Image Computing and Computer Assisted-Intervention

Pizer S, Fletcher T, Fridman Y, Fritsch D, Gash A, Glotzer J, Joshi S, Thall A, Tracton G, Yushkevich P, Chaney E (2003) Deformable m-reps for 3d medical image segmentation. International Journal of Computer Vision - Special UNC-MIDAG issue, (O Faugeras, K Ikeuchi, and J Ponce, eds) 55(2):85–106

Staib LH, Duncan JS (1992) Boundary finding with parametrically deformable models. IEEE Transactions on Pattern Analysis and Machine Intelligence (11):1061–1075

Styner MA, Rajamani KT, Nolte LP, Zsemlye G, Szekely G, Taylor CJ, Davies RH (2003) Evaluation of 3d correspondence methods for model building. In: Information Processing in Medical Imaging (IPMI)

Thall A (2004) Deformable solid modeling via medial sampling and displacement subdivision. PhD thesis, Dept. of Computer Science, University of North Carolina

Timm NH (2002) Applied Multivariate Analysis. Springer

Tsai A, Yezzi A, Wells W, Tempany C, Tucker D, Fan A, Grimson E, Willsky A (2003) A shape-based approach to the segmentation of medical imagery using level sets. IEEE Transactions on Medical Imaging 22(2):137–153