A Bayesian Approach to Sparse Model Selection in Statistical Shape Models[*]

Ali Gooya[†], Christos Davatzikos[‡], and Alejandro F. Frangi[§]

Abstract. Groupwise registration of point sets is the fundamental step in creating statistical shape models (SSMs). When the number of points on the sets varies across the population, each point set is often regarded as a spatially transformed Gaussian mixture model (GMM) sample, and the registration problem is formulated as the estimation of the underlying GMM from the training samples. Thus, each Gaussian in the mixture specifies a landmark (or model point), which is probabilistically corresponded to a training point. The Gaussian components, transformations, and probabilistic matches are often computed by an expectation-maximization (EM) algorithm. To avoid over- and under-fitting errors, the SSM should be optimized by tuning the required number of components. In this paper, rather than manually setting the number of components before training, we start from a maximal model and prune out the negligible points during the registration by a sparsity criterion. We show that by searching over the continuous space for optimal sparsity level, we can reduce the fitting errors (generalization and specificities), and thereby help the search process for a discrete number of model points. We propose an EM framework, adopting a symmetric Dirichlet distribution as a prior, to enforce sparsity on the mixture weights of Gaussians. The negligible model points are pruned by a quadratic programming technique during EM iterations. The proposed EM framework also iteratively updates the estimates of the rigid registration parameters of the point sets to the mean model. Next, we apply the principal component analysis to the registered and equal-length training point sets and construct the SSMs. This method is evaluated by learning of sparse SSMs from 15 manually segmented caudate nuclei, 24 hippocampal, and 20 prostate data sets. The generalization, specificity, and compactness of the proposed model favorably compare to a traditional EM based model.

Notation.

α	Concentration parameter of Dirichlet distribution
$\boldsymbol{\pi} = [\pi_1, \ldots, \pi_{N_M}]^T$	Vector of Gaussian priors
δ	Thickness of the level set narrow bands

[†]Center of Computational Imaging and Simulation Technologies in Biomedicine, Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield S1 3JD, UK (a.gooya@sheffield.ac.uk).
[‡]Center for Biomedical Image Computing and Analytics, Department of Radiology, University of Pennsylvania, Philadephia, PA 19104 (christos.davatzikos@uphs.upenn.edu).
[§]Center of Computational Imaging and Simulation Technologies in Biomedicine, Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield S1 3JD, UK (a.frangi@sheffield.ac.uk).

$\gamma_j^{(n)}$ Responsibility of jth model point

$\mathbb{T} = \{\mathcal{T}_k\}$ Set of K rigid transformations

$\mathbb{X} = \{\mathcal{X}_k\}$ Set of K training point sets

$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma^2, \ldots, \sigma^2, \sigma'^2)$ Covariance matrix of Gaussians

$\mathbf{m}_j$ Position vector of jth point on $\mathcal{M}$

$\mathcal{T}_k = \{s_k, \mathbf{R}_k, \mathbf{b}_k\}$ Rigid transformation mapping $\mathcal{M}$ to $\mathcal{X}_k$

$\mathcal{X}_k = \{\mathbf{x}_{ki}\}$ kth training point set constructed from kth shape

$\tilde{\mathbf{x}}_{ki} = [\mathbf{x}_{ki}^T, \phi_k(\mathbf{x}_{ki})]^T$ Hybrid vector of ith point on kth shape

$\tilde{\mathbf{x}}_{ki} = [\mathbf{x}_{ki}^T, \phi_k(\mathbf{x}_{ki})]^T]$ Hybrid vector of ith point on kth shape

$\phi(\mathbf{m}_j)$ Signed distance function estimated at $\mathbf{m}_j$

ϕ_k Signed distance function of kth training shape

$\Theta = \{\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$ Set of unknowns

D Dimension of $\mathbf{x}_{ki}$

$E_{kij}^{(n)}$ Posterior of jth Gaussian, given $\tilde{\mathbf{x}}_{ki}$ at nth iteration

K Number of training sample shapes

$N = \sum_{k=1}^{K} N_k$ Total number of points on the shapes

N_k Number of points on $\mathcal{X}_k$

N_M Number of points on the mean model

$Q(\Theta \mid \Theta^{(n)})$ Lower bound on likelihood at nth iteration

z Sparsity parameter (linear in α)

1. Introduction. Statistical shape models (SSMs), originally proposed by Cootes and Taylor in [13], have been historically applied to the automatic segmentation of anatomical structures [37, 36, 41, 55, 57, 11, 47, 39, 50, 28]. By exploiting statistical shape priors, these models increase the robustness of the segmentations against imperfect imaging features such as weak edges and clutter. Constructing an SSM from a training set often involves estimating both the mean shape and the plausible (principal) modes of its variations. For a comprehensive review of these methods, see Heimann and Meinzer [26]. The constructed SSM depends on the shape representation, e.g., a popular approach is binary image representation [49, 15, 31, 2]. In these models, the mean shape is registered to the training sets, usually through diffeomorphic deformation fields as nonfolding and invertible mappings. Next, to identify the modes of variations, principal component analysis (PCA) is directly applied to either the deformation [15] or the velocity fields [31] in the tangent spaces. In a closely related approach, Leventon, Grimson, and Faugeras [37] and Rousson, Paragios, and Deriche [48] represented the shapes by signed distance functions (SDFs). The level set values computed over the entire image domain was concatenated into a vector and subjected to PCA. Although this method conveniently handles topological variations, nonsparse representation frequently requires large computational resources.

More compact forms of shape expression are offered by surface representation. For instance, in one category, the surface is expressed as a vector in a coordinate frame spanned by a set of orthonormal basis functions. Examples include Fourier descriptors [51] and spherical harmonics [6, 33], obtained by solving the Laplace equation on a sphere. The corresponding shape vectors are statistically analyzed by PCA. These methods were originally designed for shapes homeomorphic to a sphere and are not easily extendible to more complicated and nonconvex structures.

Shape representation using point sets has been widely reported in the literature [14, 13, 21, 8, 18, 17] because it is simple to implement. Landmarks are sparsely chosen on the surfaces of the training sets, and concatenated into a vector representing each shape. Evidently, the PCA outcome depends on the order of the points in the concatenation: inconsistent ordering can result in artifactual variation modes. In other words, a pointwise "correspondence" should be specified for every landmark across the training samples. In addition, the actual location of the landmarks affects the quality of PCA. Established ways of tacking these problems are summarized in section 1.1.

Here, we estimate the number of landmarks needed to represent the mean model. This problem can be important when creating an unbiased SSM from unregistered point sets containing differing number of points. Intuitively, including many landmarks on the mean model increases the degree of freedom and results in overfitting in the from of loose registration (and incorrect modes of variations). Conversely, few landmarks deteriorate geometrical details. In fact, the relation between the degree of freedom in a shape model and the under/overfitting problem is also evident in the context of geodesic active contours [10]. In its simplest form, an evolving contour is attracted toward image edges, while its shape is regularized by penalizing its length. This leads into a curve evolution scheme that consists in a curvature weighted term to regularize the shape of the curve [23]. Reducing the weight of curvature often results in an unregularized curve that does not necessarily match the target borders and could overfit to image background, whereas increasing the curvature weight yields overly simplified and underfitted curves. Therefore, motivated by these observations, we hypothesize that there must exist an optimal number of landmarks (or sparsity level as discussed in the rest of the paper) that minimizes generalization and specificity errors.[1]

The optimal number of landmarks should ideally minimize both generalization and specificity errors [18]. In practice, however, the errors may not be minimized by the same number of model points. Thus, rather than obtaining an SSM with a single optimal number of landmarks, we use cross-validation and directly measure the generalization/specificity errors in the left-out samples. We then identify a range of model point numbers giving reasonable error values.

More importantly, rather than searching in the discrete space, we propose a Bayesian framework replacing the discrete number by a continuous hyper-parameter. Starting from a large number of model points, the hyper-parameter controls the sparsity rate of model point pruning during the registration. To achieve this, we formulate groupwise registration as the estimation of a Gaussian mixture model (GMM) from the training points, corresponding each model point to a Gaussian component. In this setting, the sparsity hyper-parameter defines the "concentration" parameter of a symmetric Dirichlet prior imposed on the mixture weights in the GMM. Thus, the problem of selecting the number of model points becomes that of determining the correct sparsity level. We show that this process compacts the SSMs and reduces their generalization and specificity errors, relative to those of established SSMs [30].

In machine learning studies, estimating the number of Gaussians in a GMM is regarded as a particular example of the "model selection" problem [16], in which models with intermediate

---

[1]The first error is related to the distance between an actual and the model predicted shapes. The latter specifies the ability of the model to instantiate correct shapes resembling the training data.

complexity (ratio of free parameters to training data) are known to be optimal in generalization. However, to our knowledge, estimating number of model points as a particular model selection problem has not been fully considered by the medical imaging community. Here, we demonstrate that an intermediate number of model points (sparsity levels) generally improves model efficiencies. To correctly highlight the contributions of this study, we briefly review some of the seminal SSM studies in section 1.1

1.1. Review of related state-of-the-art SSM methods. Manual selection of landmarks and setting their correspondences across training shapes [13] can be a time-consuming and nonreproducible procedure. In fact, landmarks are matched by machine generated approaches, traditionally by the iterative closest point (ICP) [42] algorithm. Other automatic methods improve the matching by including curvatures and local neighborhood features [7, 54, 52]. The robustness of matchings is enhanced by spatial regularization using Markov random fields [43], currents [53], and diffeomorphic mappings [22]. In an alternative approach, Frangi et al. [21] extracted a set of landmarks from a preconstructed atlas and assigned each set an example shape by nonrigid B-spline mappings. However, these methods do not necessarily optimize *groupwise* statistics. Regarding this problem as homologous to the minimum variance principle, Hill and Taylor [27] and Baumberg and Hogg [3] minimized the trace of the covariance matrix by displacing the landmarks in the surface normal direction. In an alternative approach, Kotcheff and Taylor [35] minimized the determinant of the covariance matrix and introduced a free parameter to avoid zero eigenvalues. Inspired by the Occam's razor principle, which favors simple models over complex ones, Davis et al. [18, 17] minimized the length of the binary code specifying the Gaussian distribution of the data. These methods are mathematically elegant but computationally expensive, and because of their parametric surface representation, they are not readily extendible to nonspherical topologies.

Identifying correspondences between landmarks becomes even more challenging when the training sets cannot be exactly matched [44], either because of intrinsic anatomical differences or because the number of points differs among the point clouds. In such cases, *hard-binds* established between the landmarks could introduce dummy modes in the PCA variations. To alleviate this problem, Rangarajan, Chui, and Bookstein [45] proposed the soft-assign procrustes algorithm for matchings two sets of landmarks with different counts. Granger and Pennec [25] proposed EM-ICP, a combination of expectation-maximization (EM) [19] and conventional ICP that enables partial or probabilistic matching between points. Since EM based methods can deal with complex variations, they have been popularly applied in groupwise registrations and SSM constructions.

Chui et al. [12] proposed a forward-backward consistent deformable registration of point sets to the mean model in an EM-like framework. Their algorithm iteratively updates the point correspondences, mean model, and deformations, but it does not consolidate the results by PCA. In a similar but computationally less expensive framework, Hufnagel et al. [30] applied an affine transformation to register the point sets. Based on the EM outcome (i.e., the registered point clouds with partial matches between points), they then performed a heuristic PCA on the "virtually correspondent" points. Recently, Rasoulian Rohling, and Abolmaesumi [46], replaced the affine with a B-spline based registration model, and they applied PCA to the computed deformation fields.

In the above studies, groupwise registration of point sets was reformulated as estimating the underlying probability density function (pdf) from spatially transformed noisy observations of points. The pdf is often represented as a Gaussian mixture model, whose means (estimated by EM) determine the landmark locations. Determining an appropriate number of Gaussian components for a specified number of data points, the so-called model selection problem in machine learning, is ignored in these approaches. However, model selection can avoid under- and overfitting errors and thereby improve the generalization efficiency of the trained pdf. In an SSM context, model complexity can be influenced by the number of points representing the mean shape, which is usually selected before the training phase [25, 12, 30, 46].

The contributions of this study are twofold. First, we propose a Bayesian framework for sparse SSM construction. Within this framework, the number of model points is indirectly determined *during* the training phase. Starting from a maximal number of model points, sparsity is enforced by employing a symmetric Dirichlet prior on the mixture weights. Negligible model points are pruned during each EM iteration by a quadratic programming (QP) technique. The sparsity rate is controlled by a hyper-parameter, namely, the concentration parameter of the Dirichlet prior. As is well known, the weights of the Gaussians in the GMM become sparse when this parameter is less than unity. We optimize the value of this parameter in cross-validation experiments and identify an approximate range that generalizes well globally over various data sets. Second, we simplify the reconstruction of surfaces from points by representing each landmark as a hybrid vector containing the level set and coordinates. Note that our representation is denser than that of Leventon, Grimson, and Faugeras [37], who disregarded the coordinates.

This paper extends our preliminary work in [24]. Here, we provide the mathematical details and apply our method to different and a larger number of data sets, two of which are nonhomeomorphic to a sphere. Topologically nonspherical data pose special challenges to training SSMs by Davis et al. [17] and Keleman, Szekely, and Gerig [33]. Thus, we compare our method to a closely related approach proposed by Hufnagel et al. [30], which is more flexible compared to [17, 33] and has been used as a reference in an independent recent study [46].

Sparsity has been applied to SSMs by several research groups. For instance, in the sparse shape representation of Zhang et al. [56], any unseen shape is represented by a sparse linear combination of the training samples. Alternatively, Durrleman et al. [20] enforced sparsity on the deformation fields using an L1-type prior, which encodes the shape variations. By contrast, our approach focuses on the number of model points.

The remainder of this paper is organized as follows. Our model is formulated in section 2. Sections 3 and 4 present the results and conclusions, respectively. Detailed mathematical derivations are given in the appendix (section 5).

2. Methods. This section introduces our EM algorithm. This general framework allows us to (1) align given training point clouds and (2) prune out negligible points during the registration. The result is a sparse mean model. We emphasize that in addition to the number of model points, the properties of the principal modes of variations affect the efficacy of the SSM. However, because of its complexity, we do not explicitly optimize the variation modes but rather obtain them by independent PCA after registration. In fact, this problem has been extensively studied in other works, such as [17, 27, 3, 35]. Having achieved these
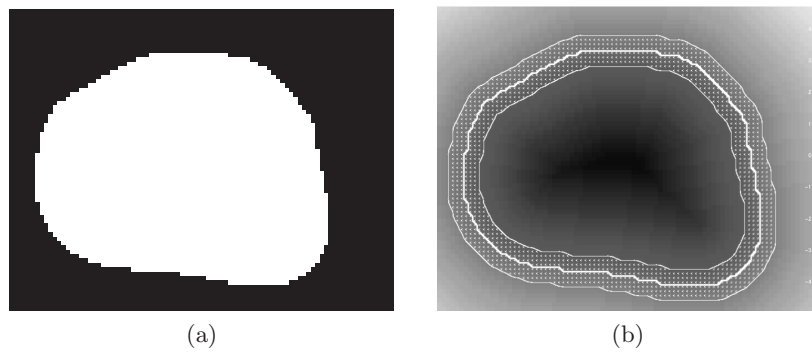
(a)   (b)

Figure 1. *Shape representation using the explained hybrid point sets:* (a) *Binary representation of a shape in the training set.* (b) *Corresponding SDF, with overlaid narrow band (shading) and zero level set contour (thick line). The shape is represented as a set of the points in the narrow band, where each point is defined by concatenating its spatial coordinates and SDF value (i.e., $\tilde{\mathbf{x}} = [\mathbf{x}^T, \phi(\mathbf{x})]^T$). The latter can aid in reconstructing the shape with the correct orientation and topology.*

goals, we construct the SSM in section 2.6 by determining the *virtual* point correspondences [30] and applying the PCA.

2.1. Shape representation. The input data to our algorithm consists of K binary masks, as D dimensional images, from which we constitute our training point sets. To this end, it is customary to sample the points from the surface of the binary masks and thus ignore the orientation of the surfaces (or the polarity of the masks). Without this information, reconstruction of surfaces from points can be handled through various approaches. For instance, Zhao, Osher, and Fedkiw [58] first computed an unsigned distance map from the points by solving the Eikonal equation; then a geodesic active contour was driven toward the point set using advection on the distance map. This step is time-consuming and, during its evolution, contour can be trapped in a local minima. In other related works, the direction of normal vectors to the surface is estimated and inconsistent normal vectors are flipped [29]. Then an SDF is constructed by moving along the normals in both directions of the surface [9]. Without explicit information on the surface normal, its automated extraction from geometry alone is a nontrivial and ambiguous process for complex and closed structures.

Here, we take a different approach and include additional distance features on a surface's narrow band for unambiguous surface reconstruction. As shown in Figure 1, given a binary mask we first construct an SDF, whose zero level set represents the surface of the mask. Next, we collect all the points within a narrow band of thickness δ around the zero level set. Each point is defined as an augmented $D + 1$ dimensional vector and consists of spatial coordinates and the corresponding SDF value. Distance information is conveyed through our EM algorithm to the constructed statistical model and can be used to infer the implied surface from the mean model. To reconstruct a surface, we first interpolate the values of the SDF on a regular grid of voxels and then extract the zero level set (see (3.2)).

2.2. Sparsity and alignment through EM algorithm. Let $\mathbb{X} = \{\mathcal{X}_k\}$, $1 \leq k \leq K$, denote the set of K observed $(D + 1)$ dimensional training point sets defined as $\mathcal{X}_k = \{\tilde{\mathbf{x}}_{ki} \in \mathbb{R}^{D+1} | 1 \leq i \leq N_k\}$. Let ϕ_k denote the SDF from the surface of the kth segmented training shape.

The points specifying $\mathcal{X}_k$ are then defined as $\tilde{\mathbf{x}}_{ki} = [\mathbf{x}_{ki}^T, \phi_k(\mathbf{x}_{ki})]^T$, where $\mathbf{x}_{ki} = [x_{ki}, y_{ki}, z_{ki}]^T$ specify the spatial coordinates of each point. To minimize the number of required points, we consider only those points within a narrow band of thickness δ surrounding the surface. Thus, in the implementation, we have set $D = 3$; however, we retain the general term D in the rest of the paper.

Let $\mathcal{M} = \{\tilde{\mathbf{m}}_j \in \mathbb{R}^{D+1}\}$, $1 \leq j \leq N_M$, be a model point set. Similarly, each model point can be written as $\tilde{\mathbf{m}}_j = [\mathbf{m}_j^T, \phi(\mathbf{m}_j)]^T$, where $\mathbf{m}_j$ denotes the pure spatial coordinates and $\phi(\mathbf{m}_j)$ is the corresponding level set value at that point. In addition, let $\mathbb{T} = \{\mathcal{T}_k\}$ be the set of K rigid transformations defined as $\mathcal{T}_k = \{s_k, \mathbf{R}_k, \mathbf{b}_k\}$, which includes scaling, rotation, and translation parameters. Each $\mathcal{T}_k$ globally transforms the model points $\tilde{\mathbf{m}}_j \in \mathcal{M}$ to the space of $\mathcal{X}_k$ by

$$(2.1) \qquad \mathcal{T}_k \star \tilde{\mathbf{m}}_j \equiv \left[\begin{array}{cc} s_k\mathbf{R}_k & 0 \\ 0 & s_k \end{array} \right] \tilde{\mathbf{m}}_j + \left[\begin{array}{c} \mathbf{b}_k \\ 0 \end{array} \right].$$

Now let $\tilde{\mathbf{x}}_{ki} \in \mathcal{X}_k$ be a noisy observation vector sampled from a Gaussian distribution centered at $\mathcal{T}_k \star \tilde{\mathbf{m}}_j$ such that $p(\tilde{\mathbf{x}}_{ki}|\mathcal{T}_k, \tilde{\mathbf{m}}_j, \boldsymbol{\Sigma}) = \mathcal{N}(\tilde{\mathbf{x}}_{ki}|\mathcal{T}_k \star \tilde{\mathbf{m}}_j, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{diag}$ $(\sigma^2, \dots, \sigma^2, \sigma'^2)$ is a diagonal covariance matrix. In this constraint, the variances of the spatial coordinates are equal but different from that of the level set values, because the spatial coordinates generally have a larger span than the level set values in the narrow bands. Since N_M points exist on $\mathcal{M}$, the conditional pdf of $\tilde{\mathbf{x}}_{ki}$ can be given as a mixture of Gaussians: $p(\tilde{\mathbf{x}}_{ki}|\mathcal{T}_k, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{j=1}^{N_M} \pi_j p(\tilde{\mathbf{x}}_{ki}|\mathcal{T}_k, \tilde{\mathbf{m}}_j, \boldsymbol{\Sigma})$, where $\boldsymbol{\pi} = [\pi_1, \dots, \pi_{N_M}]^T$.

Assuming that all N_k observed points on $\mathcal{X}_k$ are independent, the probability of $\mathcal{X}_k$ can be written as $p(\mathcal{X}_k|\mathcal{T}_k, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{i=1}^{N_k} p(\tilde{\mathbf{x}}_{ki}|\mathcal{T}_k, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$. Finally, to model the pdf of the total observation $\mathbb{X}$, we further assume that the K given points sets are jointly independent and identically distributed such that

$$(2.2) \qquad p(\mathbb{X}|\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{k=1}^{K} p(\mathcal{X}_k|\mathcal{T}_k, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}).$$

Next, we estimate the set of registration parameters and the mixture coefficients through maximum likelihood and Bayesian frameworks, respectively, by maximizing the following probability. The formulation is

$$(2.3) \qquad \begin{aligned} \hat{\mathbb{T}}, \hat{\mathcal{M}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\pi}} &= \underset{\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}}{\operatorname{argmax}} \log[p(\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}|\mathbb{X})] \\ &= \underset{\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}}{\operatorname{argmax}} [\log p(\mathbb{X}|\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) + \log p(\boldsymbol{\pi})]. \end{aligned}$$

Note the following (1) This step generates only the postregistered point sets, which are then subject to an independent PCA to obtain SSM. However, we hypothesize that an optimal registration of the point sets optimizes SSM and reduces the generalization and specificity errors. Thus, regardless of the value of N_M, we optimize the registration parameters. (2) The likelihood term, i.e., $p(\mathbb{X}|\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$ can be infinitely large if every training point corresponds to a model point, i.e., $N_M = N$, [2] which is a controversial result. Therefore, we cannot

---

[2] In that case, the Gaussian components in the mixture will approach infinity and the variance will be zero.

determine an optimal N_M by maximizing the right-hand side in (2.3) w.r.t. N_M. To identify a suitable number of model points, we instead adopt cross-validation schemes, which quantify those generalization and specificity errors versus N_M (or the sparsity level; see section 2.3).

2.3. Prior on the mixture coefficients. In this section, we propose a prior pdf for the set of mixture coefficients $\boldsymbol{\pi}$ defined in the previous section, which enables the sparse estimation of the model points. To enforce sparsity, $\boldsymbol{\pi}$ is drawn from a symmetric Dirichlet distribution function [4], given by

$$(2.4) \qquad p(\boldsymbol{\pi}) = \frac{\Gamma(N_M(N(\alpha-1)+1))}{\Gamma(N(\alpha-1)+1)^{N_M}} \prod_{j=1}^{N_M} \pi_j^{N(\alpha-1)},$$

where N denotes the total number of observed points ($N = \sum_{k=1}^{K} N_k$), Γ is the Gamma function, and α is the *concentration* parameter. Note that the prior given in (2.4) is proper, i.e., the marginalization of the prior over $\boldsymbol{\pi}$ is unity.

When $\alpha = 1$, this distribution reduces to a uniform distribution. For $\alpha > 1$, distribution is maximized when all π_i coefficients equal $1/N_M$. However, when $\alpha < 1$, the power sign becomes negative, and the distribution prefers sparsity on the mixture coefficients, i.e., the π_i's approach zero. Sparsity is enforced by retaining the latter condition. The lower bound on α is determined by $1 - 1/N_M$, as explained in the next subsection.

2.4. Optimization using EM. Since, the direct solution of (2.3) does not have a closed form, we find a tractable solution through the EM algorithm [19]. Given an estimate of the parameters $\Theta = \{\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$ at the current iteration (n), EM maximizes the lower bound on the right-hand side of (2.3), which takes the following form:

$$
\begin{aligned}
Q(\Theta|\Theta^{(n)}) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} \{ & E_{kij}^{(n)}[\log(\pi_j) \\
(2.5) \qquad & + \log(\mathcal{N}(\tilde{\mathbf{x}}_{ki}|\mathcal{T}_k \star \tilde{\mathbf{m}}_j, \boldsymbol{\Sigma}))]\} + \log(p(\boldsymbol{\pi})),
\end{aligned}
$$

where $E_{kij}^{(n)}$ specify the posterior probability that $\tilde{\mathbf{x}}_{ki}$ are sampled from the Gaussian component of mean $\tilde{\mathbf{m}}_j$ in the current iteration n. The EM algorithm proceeds through two iterative steps. In the E-step, the posterior probabilities $E_{kij}^{(n)}$ are updated using the current parameter estimates:

$$(2.6) \qquad E_{kij}^{(n)} = \frac{\pi_j^{(n)} \mathcal{N}(\tilde{\mathbf{x}}_{ki}|\mathcal{T}_k^{(n)} \star \mathbf{m}_j^{(n)}, \boldsymbol{\Sigma}^{(n)})}{\sum_{l=1}^{N_M} \pi_l^{(n)} \mathcal{N}(\tilde{\mathbf{x}}_{ki}|\mathcal{T}_k^{(n)} \star \mathbf{m}_j^{(n)}, \boldsymbol{\Sigma}^{(n)})}.$$

In the M-step, the value of (2.5) is maximized w.r.t. the unknown parameters, giving a set of updated equations. Using the results from the last step, we sequentially update each parameter in the following order: $\mathbb{T}, \mathcal{M}, \boldsymbol{\Sigma}$, and $\boldsymbol{\pi}$. The first set is derived similarly to the procedures outlined in [30] and [40] (see the appendix for the details). This section focuses on the update rule for $\boldsymbol{\pi}$, the main mechanism by which sparsity is enforced.

We plug (2.4) into (2.5) and keep the relevant terms in $\boldsymbol{\pi}$. We then apply the constraint $\sum_{j=1}^{N_M} \pi_j = 1$ through a Lagrange multiplier λ and maximize the following expression w.r.t. each π_j, $1 \le j \le N_M$:

$$Q(\Theta|\Theta^{(n)}) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} \{[E_{kij}^{(n)} + (\alpha - 1)]\log(\pi_j)\}$$

$$(2.7) \qquad + \lambda \left(\sum_{j=1}^{N_M} \pi_j - 1 \right) + O.T.,$$

where $O.T.$ denotes $\boldsymbol{\pi}$-independent terms. Differentiating (2.7) w.r.t. π_j and eliminating λ after some algebra, we obtain

$$(2.8) \qquad \pi_j^{\star} = \frac{\gamma_j^{(n)} + \alpha - 1}{N_M(\alpha - 1) + 1},$$

where N denotes the total number of the observed points ($N = \sum_{k=1}^{K} N_k$) and $\gamma_j^{(n)} = 1/N \sum_{k=1}^{K} \sum_{i=1}^{N_k} E_{kij}^{(n)}$. This value can be regarded as the *total responsibility* of model point $\mathbf{m}_j$ in explaining the data. When $\alpha = 1$, (2.8) reduces to the classic maximum likelihood based estimation of mixture weights with no sparsity imposed.

Note that (2.8) allows negative mixture weights because no positivity constraint is imposed when deriving (2.7). In fact, it is easily seen that for $1 - 1/N_M \le \alpha$, if $\gamma_j^{(n)} < 1 - \alpha$, then $\pi_j^{\star} < 0$. In other words, if the total responsibility of the model point $\tilde{\mathbf{m}}_{\mathbf{j}}$ falls below $1 - \alpha$, its corresponding mixture weight becomes negative. However, since negative probabilities are invalid, π_j of such model points is set to 0 by the QP scheme explained below. Note that sparsity can increase when α is very close to $1 - 1/N_M$. Hence, we specify in terms of an auxiliary variable as $z \in (0, 1)$, specifically α using $\alpha = (1 - 1/N_M)z + (1 - z)$. In this definition, sparsity is proportional to z.

Having estimated all $\pi_j^{\star}$ values, we update the mixture weights by solving the following convex optimization problem using the generalized sequential minimal optimizer proposed in [32]:

$$\pi_1^{(n)}, \ldots, \pi_{N_M}^{(n)} = \operatorname*{argmin}_{\pi_1, \ldots, \pi_{N_M}} \sum_{j=1}^{N_M} (\pi_j^{\star} - \pi_j)^2$$

$$(2.9) \qquad \text{s.t. } \sum_{j=1}^{N_M} \pi_j = 1 \ \wedge \ 0 \le \pi_j \ \forall j.$$

Equation (2.9) effectively finds the orthonormal projection of $\boldsymbol{\pi}^*$ on the space of eligible probabilities satisfying the above conditions. At the end of the M-step, model points with zero corresponding mixture weight, i.e., $\pi_i^{(n)}$, are identified and pruned out. Next, N_M is updated by counting the remaining *alive* points.

Computationally, we observed this algorithm to be very efficient. Even for a large number of model points ($N_M \simeq 10^4$), the convergence time of QP is typically less than one second.

The method breaks the convex optimization of (2.9) into a series of smaller subproblems by identifying mixture weights that violate the constraints [32].

The model points $\mathcal{M}$ are initialized by a k-means algorithm [4] and all spatial transformations in $\mathbb{T}$ are set to the identity transformation. The initial number of model points N_M is the total number of available points (typically $N - K$ to avoid extreme overfitting) and is later reduced by the imposed sparsity. The algorithm alternately iterates between the E and M-steps until convergence (in our case, this is determined by a small variation in the estimated covariance matrix). Typically, sufficiently good convergence is achieved after 50 iterations of our EM algorithm. Higher numbers of iterations produce no significant change in the results. Thus, the number of iterations was fixed to 50 in all subsequent experiments.

2.5. Brief overview of the EM-ICP method. As discussed earlier, our EM algorithm is motivated by the study of Hufnagel et al. [30]. Besides using an affine transformation model, the Hufnagel et al. study differs from ours in several important ways: (1) we compute the mixture coefficients using (2.8) and (2.9), whereas the Hufnagel et al. approach assumes that all prior probabilities of the mixture components are fixed and equal to $\pi_j = 1/N_M$. Hence, no sparsity is enforced in [30] and N_M must be predetermined. (2) To prevent early trapping in local minima because of sudden drop in σ^2, the variance in [30] was manually reduced by some factor (normally between 0.8 and 0.9). We hypothesize that early drops in σ^2 are prevented by our pruning process, which removes negligible points from the model. Intuitively, we expect that the observed points are sampled from fewer Gaussian components as the pruning proceeds. Thus, the variance of such Gaussians increases to preserve the likelihood of the observed points. (3) The number of EM iterations in [30] was user-defined. While Hufnagel the et al. model requires three free parameters (i.e., the prespecified number of landmarks N_M, a variance reduction factor, and the number of iterations), our model requires a single parameter (z). Therefore, the Hufnagel et al. model is more complex than our model, and its results are more dependent on user choices.

2.6. Construction of SSM. Using the EM algorithm described in the previous section, we first register our training data sets in $\mathbb{X}$ to the mean shape $\mathcal{M}$ by estimating the rigid transformations in $\mathbb{T}$. The resulting mean model is insufficient for predicting unseen samples. To properly construct a generalizable SSM, we should learn the plausible shape variations that exist in the training samples. This subsection explains how the modes of variations are computed from the registered samples. The approach, described in [30], is included here to ensure a self-contained report.

Shape models are popularly constructed by PCA [13]. However, this method requires one-to-one point correspondences between the aligned training data and the mean shape. By contrast, no exact point correspondences are defined in our EM-ICP based method. In fact, point correspondences are identified by the "virtual correspondence" [30], which resolves the problem as follows: for any model point $\tilde{\mathbf{m}}_\mathbf{j}$ a virtually correspondent point, denoted by $\check{\mathbf{m}}_{kj}$, is induced by the training sample $\mathcal{X}_k$ as

$$(2.10) \qquad \check{\mathbf{m}}_{kj} = \sum_{i=1}^{N_M} \frac{E_{kij}}{\sum_l E_{klj}} T_k^{-1} \star \tilde{\mathbf{x}}_{ki},$$

where E_{kij} denotes the posterior probability of $\tilde{\mathbf{x}}_{ki}$ being sampled from the Gaussian component that is specified by the model point $\tilde{\mathbf{m}}_{\mathbf{j}}$. Computing these correspondences for all training samples, we obtain $K+1$ virtually aligned shapes (including the mean shape). Next, each shape is columnwise concatenated to a vector, and PCA is applied to the co-variance matrix of the shape vectors. The result is the average vector $\bar{\mathbf{x}} \in \mathbb{R}^{(D+1)N_M}$ and the matrix of eigenmodes $\mathbf{P} \in \mathbb{R}^{(D+1)N_M \times n}$, where $n \leq K$ is the number of principal components considered in the analysis.[3] Having defined these components, any unseen sample vector $\mathbf{x}$ is then computed as

$$(2.11) \qquad\qquad \mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b},$$

where $\mathbf{b}$ is the vector of coefficients. To eliminate the noise from $\mathbf{x}$, each coefficient is constrained by $|b_i| \leq 3\sqrt{\lambda_i}$, where λ_i is the ith eigenvalue of the co-variance matrix of the training vectors [13].

3. Results. In this section, we first describe the caudate, hippocampus, and prostate data sets and explain the construction of our training point sets from segmented structures. All the data sets are publicly available and downloadable from the repositories of MICCAI Grand Challenges.[4] Since we aimed to evaluate the quality of the statistical models themselves, no image segmentation was performed. In fact, for comparing our SSM with the Hufnagel et al. EM-ICP model, the acquired data sets were segmented by an expert and subjected to cross-validation experiments. Therefore, we used three well-known criteria introduced by Davis et al. [18], i.e., generalization, specificity, and compactness. For brevity, we hereafter refer to our model as SpSSM and to the Hufnagel et al. EM-ICP model as SSM. Finally, we present the principal modes of variations extracted by SpSSM.

3.1. Data specifications. The first data set (caudate) consists of 15 segmented sets acquired from subjects with schizotypal personality disorder [38]. Magnetic resonance (MR) images were acquired by a spoiled gradient-recalled imaging protocol with the parameters, i.e., echo time (TE) = 5 ms, repetition time (TR) = 35 ms, repetition = 1, nutation angle = 45°, field of view = 24 cm, acquisition matrix $256 \times 256 \times 124$, and voxel dimensions = $0.9375 \times 0.9375 \times 1.5$ mm. Each caudate nucleus was manually segmented by an expert (Figure 2(a)).

The second data set contains hippocampus segmentations of both epileptic patients and healthy control individuals. T1-weighted spoiled gradient echo recalled MR images were acquired by two scanners with different acquisition matrices and were subsequently segmented by experts [34]. Twenty-four data sets were randomly selected and the resolutions were unified. The final matrix and voxel sizes were $256 \times 256 \times 124$ and $0.781 \times 0.781 \times 2$ mm, respectively. Sample manual segmentations of the left and right nuclei of a single subject, used as a training set in our experiments, are shown in Figure 2(b).

Our third data set consists of prostate structures that were manually segmented from transversal T2-weighted MR images [1]. Data were acquired from patients with benign (e.g., prostatic hyperplasia) or malignant lesions. Figure 2(c) shows a sample MR prostate image in various slices overlain by an expert segmentation. Since this data set was contributed

---

[3]The number of principle components is usually decided when the sum of the considered n eigenvalues accounts for 98% of the total trace of the covariance matrix.
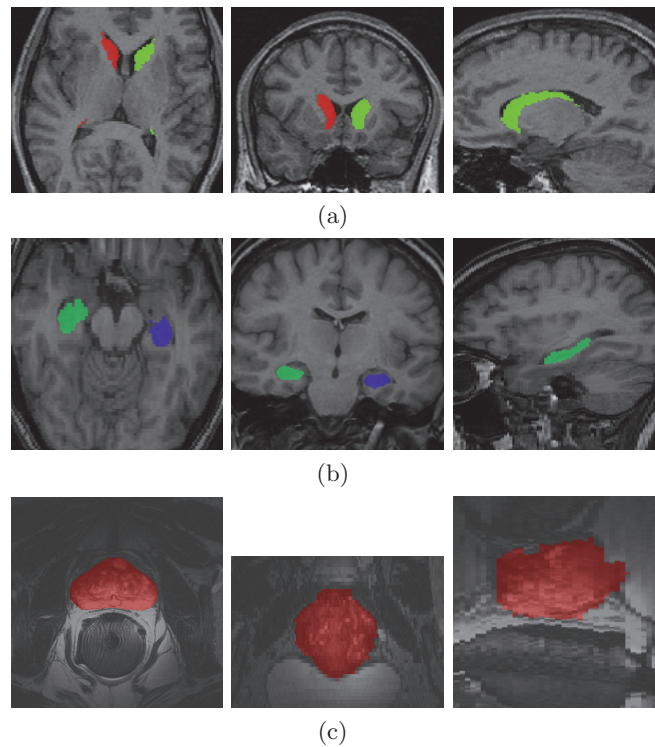
[4]http://www.grand-challenge.org.

(a)

(b)

(c)

Figure 2. *Sample of manually segmented slices of training data:* (a) *caudate nuclei,* (b) *hippocampal nuclei, and* (c) *prostate (shown in axial, coronal, and sagital slices from left to right).*

by multiple centers adopting different scanning protocols, the acquired images with different resolutions. To maintain low computational burden, we chose 20 sets and resampled the expert segmentations to element spacing of $2.0 \times 2.0 \times 2.0$ mm. By implementing this step, we achieve a consistent surface sampling rate across our training sets and consequently preserve the number of points generated for each individual across our populations. Naturally, this steps minimizes the possibility of artifactual bias toward oversampled or undersampled training shapes

3.2. Training point sets. For each of the caudate and hippocampus data-sets, the level sets in the surrounding narrow bands were generated with thickness of $\delta = 1.5$ pixels and, approximately, 10^4 points per set. Next, these point sets were individually reduced to around 300 points by setting $z = 0.95$ and estimating no spatial transformation. Similarly, the level sets of the prostate data sets were generated with a narrow band thickness of 1.5 mm and approximately 3000 points per sample. In the second step, each of these point sets was individually reduced to around 250 points by setting $z = 0.7$. The SSMs are constructed from the resulting point sets in subsequent steps. Several samples of these point sets before and following registration are shown in Figures 3(a) and 3(b), respectively.

3.3. Quantitative model performance measures. Our SSM was compared to the Hufnagel et al. model [30] in an exhaustive series of cross-validation experiments. The generalization, specificity, and compactness measures were quantified for each model [18, 17].
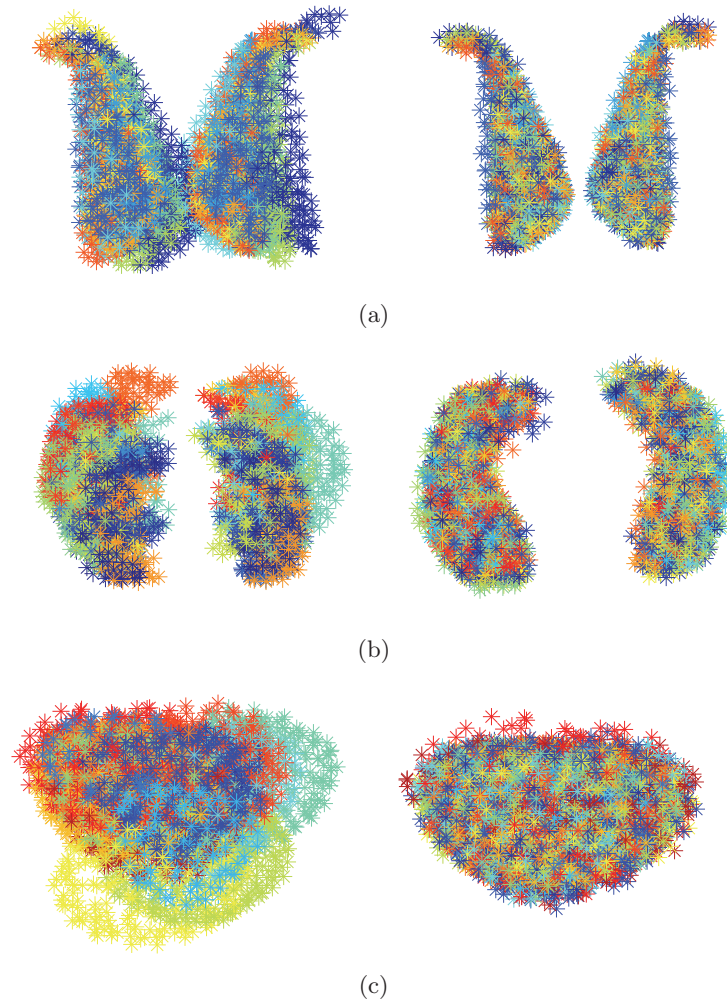
Figure 3. *Registration of training point clouds by our method:* (a) *caudate nuclei,* (b) *hippocampal nuclei, and* (c) *prostate samples (left and right panels display the original and the registered clouds, respectively.)*

The generalization performance indicates the ability of the trained model to predict unseen test samples [18, 4]. The generalization in each fold of our cross-validation experiments was evaluated in four steps: (1) Coregister the training set to the mean model $\mathcal{M}$ using the EM algorithm. (2) Apply PCA to the coregistered training sets by (2.10) and (2.11) and compute the matrix of eigen modes P and the mean of the registered samples $\bar{X}$. (3) Align each of the point clouds in the left-out test set with the trained model $\bar{X}$ by reapplying the EM algorithm but without updating $\mathcal{M}$. Set X_j as the test set, which is now registered to $\bar{X}$. (4) Project X_j on the P space by (2.10) and (2.11) and thereby obtain X. The spatial normalization effects on the training sets imposed by scaling, rotation and translation operations are visible in the sample registrations shown in Figure 3. From these registered point clouds, we compute the PCA space and project the left-out test point sets on that space.

Next, we quantify the difference between X_j (the reference point set) and $\hat{X}_j$ (the model-predicted point set). The smaller the difference, the better the generalization. Following [30],

our first measure is the Hausdorff distance, defined for two given point sets as follows:

$$(3.1) \qquad d_H(X_j, \hat{X}_j) = \max\{\max_{\tilde{\mathbf{x}} \in X_j} \min_{\tilde{\mathbf{y}} \in \hat{X}_j} d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \max_{\tilde{\mathbf{y}} \in X_j} \min_{\tilde{\mathbf{x}} \in \hat{X}_j} d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})\},$$

where $d(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = ||\tilde{\mathbf{x}} - \tilde{\mathbf{y}}||_2$ denotes the Euclidean distance between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. Thus, our first generalization measure is defined as $G = \frac{1}{K} \sum_{j=1}^{K} d_H(X_j, \hat{X}_j)$

The Hausdorff distance properly quantifies the differences between details and between sharp geometrical features, such as the elongated tails of caudate nuclei. The overall similarity of the shapes represented by X and X_j requires an alternative measure, here defined as the average distance of the surfaces reconstructed from these point sets. To reconstruct a surface, the level set values on the three-dimensional grid of voxels are first interpolated from the scattered points by

$$(3.2) \qquad \hat{\phi}(\mathbf{v}_n) = \sum_m \phi(\mathbf{x}_m) k(||\mathbf{v}_n - \mathbf{x}_m||),$$

where $\hat{\phi}$, ϕ are the interpolated and available level set values, $k(\cdot)$ is a radial basis kernel function, and $\mathbf{v}_n$, $\mathbf{x}_n$ represent the positional vectors of the query voxel n and scattered point m, respectively. This interpolation is implemented using a fast and efficient library [5]. Gaussian kernels were used and the bandwidth of the basis functions was set to 6 mm and 30 mm for the caudate/hippocampus and prostate data sets, respectively. These values were retained throughout the experiments. Next, the zero level set is extracted from the three-dimensional volume and a surface mesh is reconstructed [9].

The specificity determines the ability of the model to produce a valid output. The specificity was computed by randomly sampling the SSMs within the range of valid parameters, $[-3\sqrt{\lambda_k}, 3\sqrt{\lambda_k}]$ and measuring $S = \frac{1}{R} \sum_{r=1}^{R} \min_{k \in \{1 \dots K\}} d_H(X_r, \mathcal{X}_k)$. Here, R is the number of randomly generated data sets, X_r is the random shape, and $\mathcal{X}_k$ is the kth training sample. The minimalism of the parameter set is quantified by the compactness measure, that is, the cumulative sum of c eigenvalues of the covariance matrix of the aligned training samples, normalized over the trace of the covariance matrix: $C(c) = \sum_{k=1}^{c} \lambda_k / \sum_{k=1}^{K} \lambda_k$

3.4. Experiments. In this section, we compare our sparse SSM with the Hufnagel et al. model [30]. The parameters space of both models were subjected to cross-validation experiments, and the generalization, specificity, and compactness were evaluated at each point. The generalization was quantified using left-out test samples, whereas the specificity was determined from training sets. In SpSSM, the sparsity parameter z was varied from 0.01 to 0.99 in 0.1 increments. However, the dimensions of the parameters in SSM were prohibitively large because the N_M, number of EM iterations, and variance reduction factor (section 2.5) were freely selectable. Thus, N_M was restricted to 100–3700 (approximate number of total available points $N = \sum_k N_k$) in increments of 400, while the other parameters were assigned typical values.

3.4.1. Generalization. We performed threefold cross-validations on our caudate data. In each fold, 10 point sets were used for training and 5 point sets were left as test samples. We

Table 1

Comparison of the minimum Hausdorff and surface distances (in mm) using SSM [30] and our sparse model (SpSSM). Measurements in bold are statistically different. The associated p-values are listed in the following rows.

	Caudate	Hippocampus	Prostate
Hausdorff distances			
SSM	5.4±0.81	6.0±1.16	21.46±2.42
SpSSM	**4.67±0.66**	**5.47±0.89**	**13.54±2.99**
p-value	2.13×10^{-8}	8.2×10^{-4}	≈ 0
Average surface distances			
SSM	1.42±0.14	1.24±0.18	1.43±0.16
SpSSM	1.39±0.13	**1.14± 0.18**	**1.18±0.09**
p-value	0.1875	8.88×10^{-4}	≈ 0

found that for any preselected z or N_M, the EM algorithm converged to slightly different registration results (different local minima) depending on the initialization achieved by k-means. To minimize such randomness in the created shape models, we report the results of five experiments.

Figure 4 shows the generalization results. Evidently, the Hausdorff and surface distances obtained by SSM depend on the number of EM iterations (specified by different colors in Figure 4(a) and 4(b)). In particular, the smallest Hausdorff distance is achieved after 15 EM iterations; implementing more iterations does not improve the results. The distance improves as the number of model points increases to 500, but no absolute distance improvement is observed thereafter. Comparing the results of SSM and SpSSM (Figure 4(c) and (d)), we observe that our model reduces the errors for $0.2 \leq z \leq 0.6$. The Hausdorff distance obtained by the optimal SpSSM (at $z = 0.4$) 4.67 ± 0.66 mm is significantly less than the minimum Hausdorff distance obtained by SSM (5.40 ± 0.81 mm) after 15 EM iterations with $N_M = 2500$. Table 1 summarizes the results of statistical paired t-tests of the difference analysis and the corresponding p-values. The minimum average surface distances of the caudate data are not significantly different between the models, suggesting that our SpSSM better captures the details (such as corners and tails), whose contributions to surface distance are attenuated by averaging, but which significantly contribute to the Hausdorff distances.

Next, we quantified the generalization by performing fourfold cross-validations on hippocampus data sets. In each fold, 18 sets were used for training and 6 sets were left as test samples. To minimize randomness, which could be introduced by different k-means initializations, we report the results of three experiments. Here, the variance reduction factor is set to 0.8 and the results are shown in Figure 5. Again, the distances computed for SSM depend on the number of EM iterations, but more iterations do not improve the generalization efficiency (Figure 5(a) and (b)). The Hausdorff and surface distances are minimized at 6.0 ± 1.16 mm ($N_M = 500$) and 1.24 ± 0.18 mm ($N_M = 1300$), respectively. The results of SpSSM are shown in Figure 5(c) and (d). The best distances of SpSSM are smaller than those of SSM when $0.1 \leq z$. In SpSSM, the Hausdorff and surface distances are minimized at 5.47 ± 0.89 mm ($z = 0.7$) and 1.14 ± 0.18 mm, respectively. Both these errors are significantly less than their SSM counterparts (p-values are shown in Table 1).

Finally, we quantified the generalization on the prostate data sets by conducting 5 rounds of cross-validations. In each fold, 15 point sets were used for training, and 5 sets were left
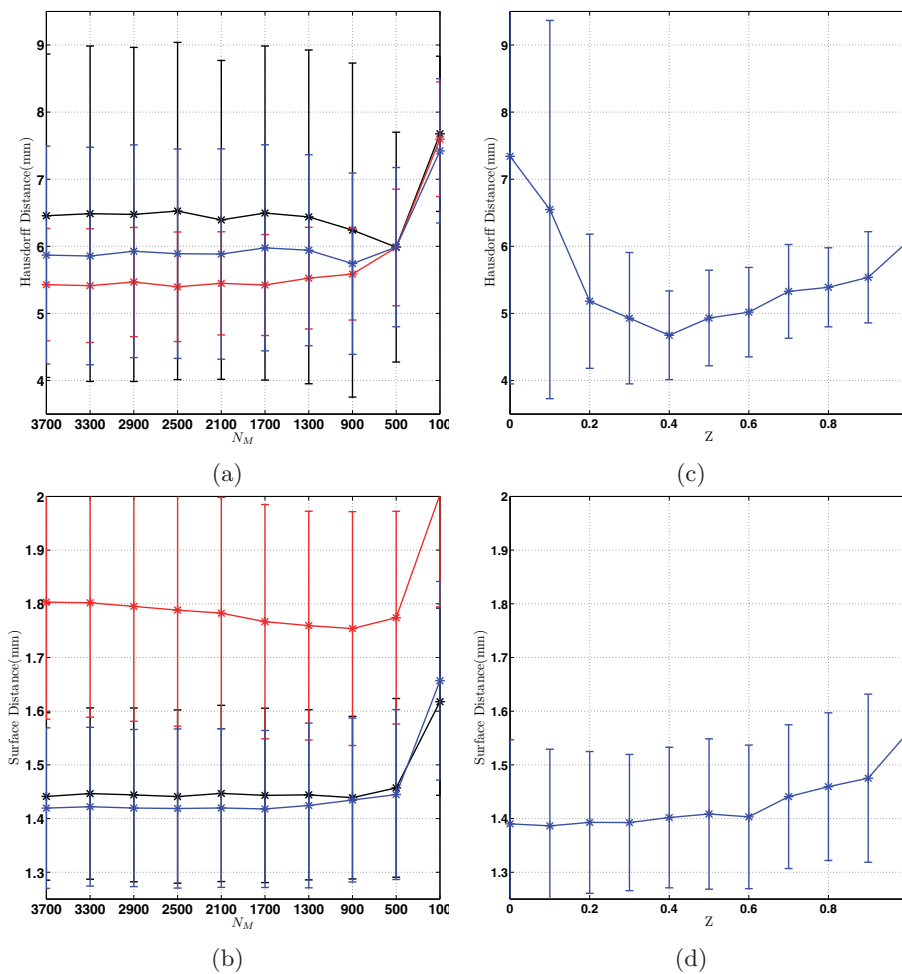
Figure 4. *Generalization errors obtained for* 15 *caudate nucleus data sets (see section* 3.4*). Shown are the averages of* 5 *rounds of threefold cross validation experiments:* (a) *Hausdorff and* (b) *surface distances using the classic EM-ICP method* [30] *with different numbers of EM iterations (black:* 25*; blue:* 20*; red:* 15*; and variance reduction factor:* 0.85*). For comparison, the Hausdorff and surface distances obtained using our SpSSM are shown in* (c) *and* (d)*, respectively. Our model reduces the Hausdorff and average surface distances (for* $0.2 \leq z \leq 0.6$*).*

for testing. Figure 6(a) and (b) shows how the SSM outcomes depend on the number of EM iterations in this data set. Comparing these to the SpSSM results in Figure 6(c) and (d), we observe that the distances are reduced in SpSSM over the entire range of z range, indicating better generalization by this method. The minimal Hausdorff and surface distances are significantly smaller in the optimal SpSSM than in the optimal SSM (Table 1). The differences between the model results are larger in the prostate data set than in the caudate and hippocampal data sets, probably because of larger variability in prostate data sets.

3.4.2. Specificity. The specificities of both models were compared by the Hausdorff distances, which are more discriminative than the surface distances, as seen in the previous
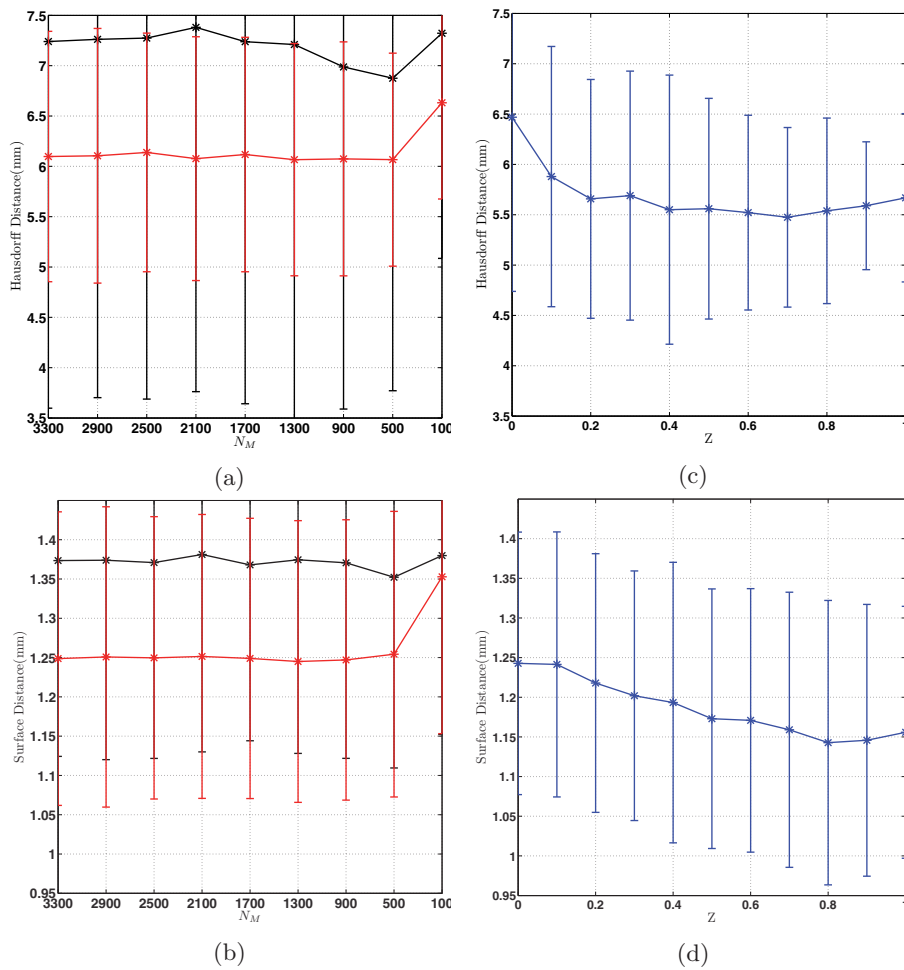
Figure 5. *Generalization errors in 24 hippocampal data sets (section 3.4). Shown are the averages of 3 rounds of fourfold cross validation experiments: Hausdorff (a) and surface (b) distances using the classic EM-ICP method [30] with different numbers of EM iterations (black: 25; red: 20; variance reduction factor: 0.8). For comparison, the Hausdorff and surface distances obtained by our sparse SSM are in (c) and (d), respectively. Our model reduces Hausdorff distance while preserving the quality of the distances (for $0.1 \leq z$). The errors in (c) and (d) are slightly increased when the sparsity is over imposed.*

subsection. Furthermore, for SSM, we consider only the EM iteration numbers and variance reduction factors yielding the best generalization outcome.

Figure 7(a)–(c) shows the specificity results for SSM. Comparing these graphs to the corresponding specificities for SpSSM in (d)–(f), we confirm that our proposed model improves specificity of the caudate data sets for $0.1 \leq z \leq 0.6$ and of the hippocampal data set for $0.01 \leq z \leq 0.9$. For the prostate data set, improvement is observed across the entire range of z. Furthermore, analogous to the generalization errors in Figure 4, 5, and 6, the specificity errors are minimized at intermediate sparsity values for all three data sets. However, the values that optimize the specificity are not necessarily those that minimize the generalization. Therefore, an optimal SpSSM model that minimizes both specificity and generalization errors is not easily determined.
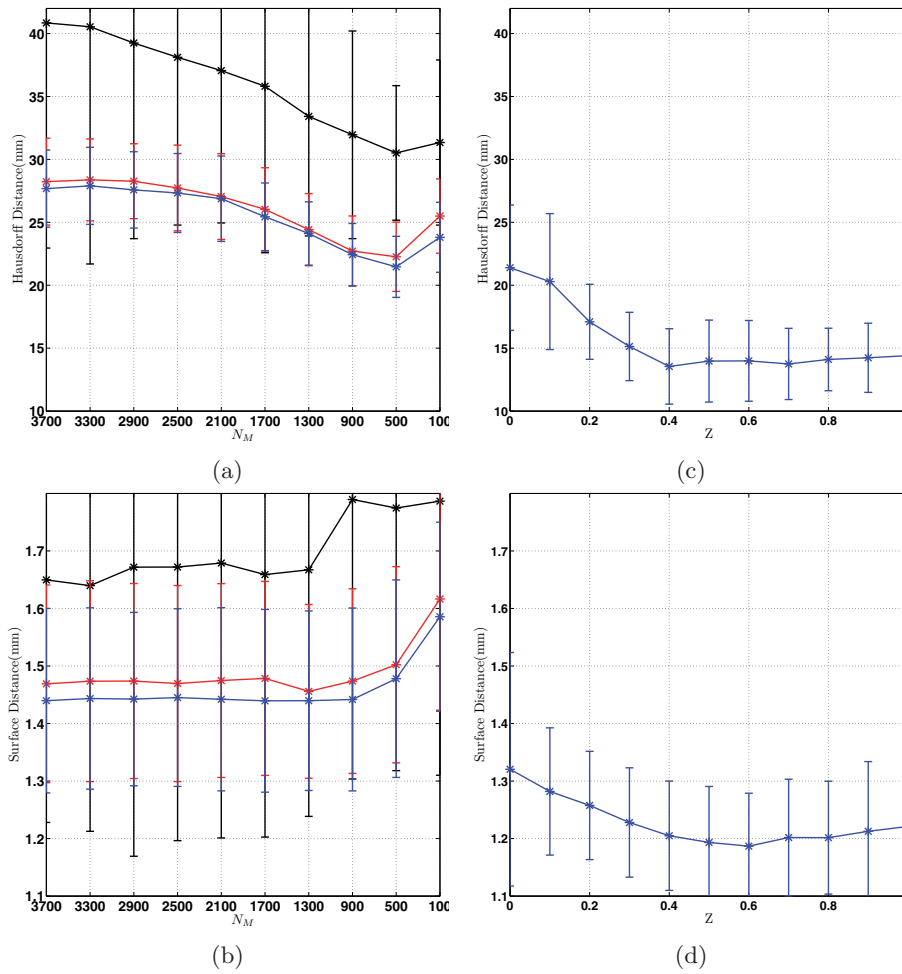
Figure 6. *Generalization errors obtained for* 20 *prostate data sets (section* 3.4*). Shown are the averages of* 5 *rounds of fourfold cross validation experiments: Hausdorff* (a) *and surface* (b) *distances obtained by the EMICP method in* [27] *with different numbers of EM iterations (black:* 30*; blue:* 25*; red:* 20*; and variance reduction factor:* 0.85*). For comparison, the Hausdorff and surface distances obtained by our sparse SSM are shown in* (c) *and* (d)*, respectively. In general, our model reduces the Hausdorff and surface distances.*

3.4.3. Compactness. Figure 8 shows the compactnesses results of SSM and SpSSM. As the sparsity parameter z increases (as N_M decreases), SpSSM (SSM) becomes more compact. However, in both cases, models with good generalization and specificity generally have intermediate compactness. Although the compactness overlaps the three data sets, the proposed sparse models with $0.1 \leq z$ (for caudate) and $0.4 \leq z$ (for hippocampus and prostate) are at least as compact as SSM with adequate point counts (strong generalization performance requires $100 < N_M$). Therefore, we conclude that (optimal) SpSSMs are generally more compact than SSMs.

3.4.4. Optimality interval for sparsity parameter. Figure 9 summarizes the results of our quantitative comparisons. In this figure, for each data type we have highlighted the ranges of z, where SpSSM outperforms SSM in various evaluation metrics. By considering the intersection of these ranges shared across various data types, we can identify a range of
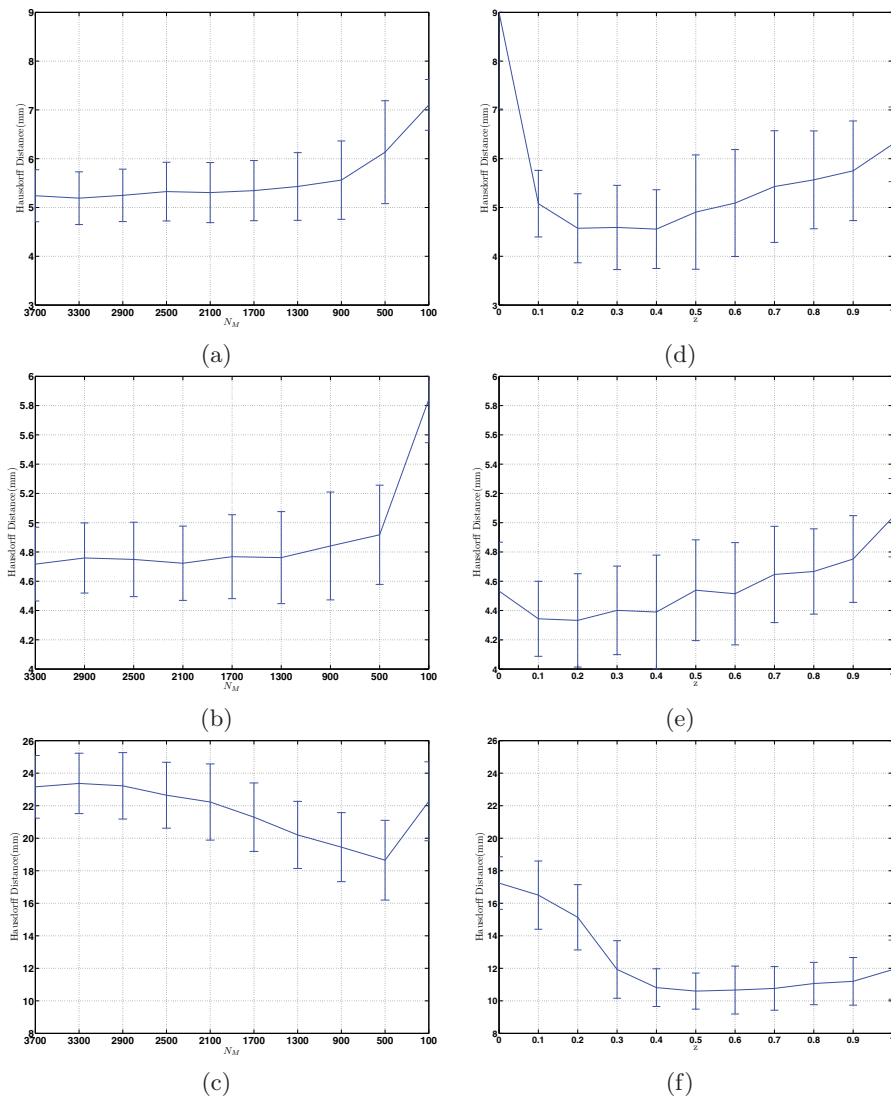
Figure 7. (a)–(c) *Specificity errors in the caudate, hippocampus, and prostate data sets using the EM-ICP method in* [30]. *The corresponding errors obtained by SpSSM are shown in* (d)–(f). *The proposed method improves the specificities in the caudate and hippocampus data sets for* $0.1 \leq z \leq 0.6$ *and* $0.01 \leq z \leq 0.9$, *respectively. In the prostate data set (bottom row), the specificities are improved across the entire range* z.

optimality of z, in which the proposed SpSSM is *globally* superior to SSM (regardless of the data type, N_M, and evaluation criterion). As seen in Figure 9, this range is lower bounded at $z = 0.4$ by the compactness of SpSSM on the prostate and upper bounded at $z = 0.6$ by the specificity on the hipposcampus data sets. Thus, we hypothesize that a *safe* value for z could be selected in the range of $0.4 \leq z \leq 0.6$ if no cross-validation experiments are considered. The interval naturally favors models with intermediate complexities, where sparsity is in neither of extremities.

3.4.5. Sensitivity to initialization. As explained, the means of the Gaussian components in our EM algorithm are initialized using k-means algorithm. We repeat k-means clustering
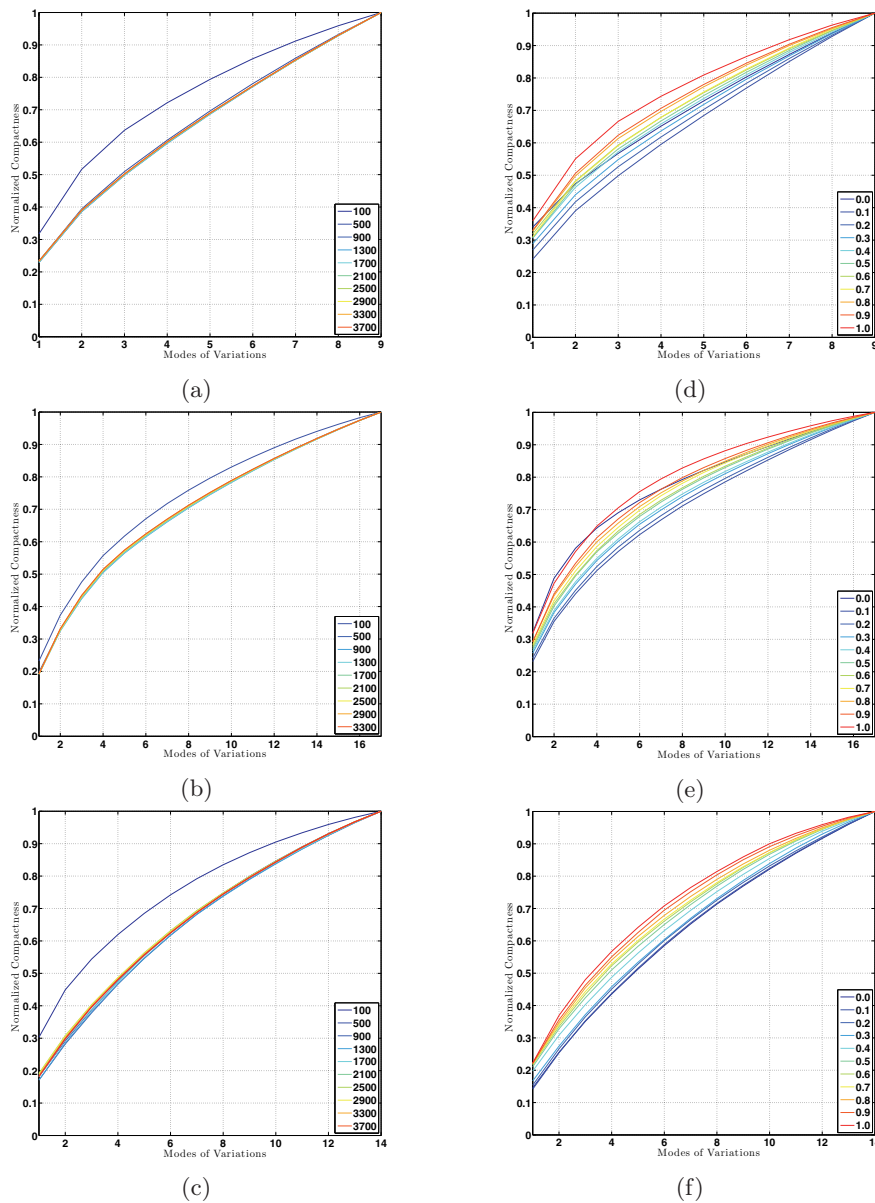
Figure 8. *Compactness, defined as the scree plots of the covariance matrices of the coaligned training samples: (a)–(c) Measured by applying SSM [30] to the caudate, hippocampus, and prostate data sets; (d)–(e) Corresponding values measured using SpSSM. The SpSSM tends to become more compact than SSM when $0.1 \leq z$ and $0.3 \leq z$ are chosen for caudate/hippocampus and prostate data sets respectively. (SSMs with $N_M = 100$ are excluded, due to their poor generalization and specificity.)*

of all data points, and then identify the best initialization by picking a k-means outcome that has a minimal clustering cost. In order to investigate the sensitivity of the final results to initialization, we ran cross-validations measuring generalization of SpSSM under different numbers of k-means runs. The rationale is that a robust model construction should be
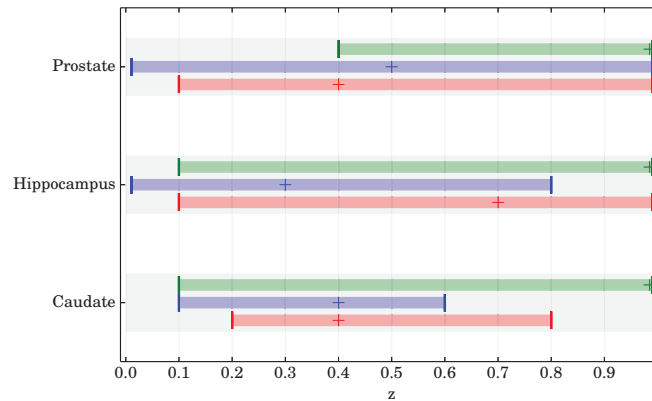
Figure 9. *Sparsity ranges where SpSSM outperforms SSM with regard to generalization (red), specificity (blue), and compactness (green) critera. As seen for $0.4 \leq z \leq 0.6$, SpSSM is globally superior to SSM, regardless of the data type, the evaluation metric, and N_M (and other parameters used to construct SSMs). Markers denote the optimal sparsity values.*

Table 2
The influence of initialization on the generalization errors (in mm) obtained by SpSSM, measured as Hausdorff distances for different number of k-means initializing rounds (measured at $z = 0.5$).

	Caudate	Hippocampus	Prostate
K-means rounds			
1	4.98 ± 0.82	5.54 ± 1.05	13.99 ± 3.4
10^1	4.92 ± 0.64	5.52 ± 1.12	13.98 ± 3.23
10^2	5.01 ± 0.79	5.51 ± 1.0	13.77 ± 3.33
10^3	5.03 ± 0.78	5.53 ± 1.13	14.06 ± 3.2

minimally dependent on k-means outcomes. Thus, the measures should be stable across different numbers of initialization rounds. For these experiments, we fixed $z = 0.5$ and varied k-means rounds from 1 to 10^3. Then we measured the Hausdorff distance of PCA projected and original point sets as before. The results, shown in Table 2, are not statistically different across various numbers of k-means initializations, suggesting that our sparse model construction is well robust to initialization.

3.4.6. Nonlinear relationship between sparsity and z. Figure 10 shows the relationships between z and the number of points at convergence, and between z and sparsity levels in the optimal SpSSM, determined by cross-validation. The converged sparsity is most sensitive at $z = 0.01$ and falls almost exponentially as z increases. For instance, as z varies from 0.4 to 0.8, the sparsity of the estimated prostate model changes by less than 5%. This varying sensitivity trend may guide the selection of z in the absence of cross-validation experiments and measurements of the Hausdorff and surface distances.

3.4.7. Qualitative descriptions. Figure 11 shows the mean models estimated by SpSSM for different values of z. Minimal ($z = 0.01$) leads to anatomically incorrect models, with several artificial structures (left panels of Figure 11). For example, unrealistic features appear in the tails of the estimated caudate in Figure 11(a) and on the surface of the estimated prostate in Figure 11(g). These artifacts are introduced by overfitting, which occurs when
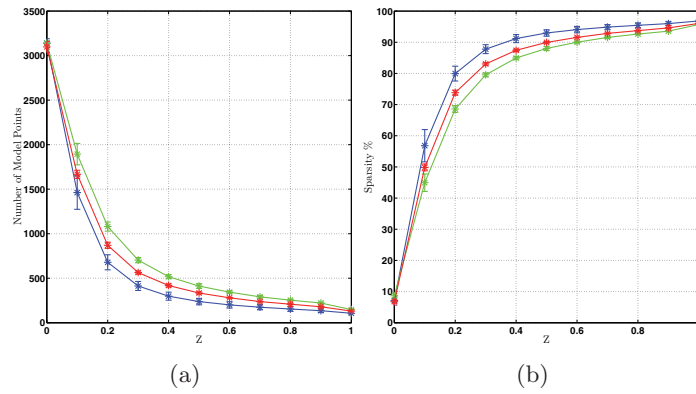
(a) (b)

Figure 10. *Number of model points* (a) *and sparsity levels* (b) *versus z obtained in cross-validation experiments, prostate (blue), hippocampus (red), and caudate data sets (green).*
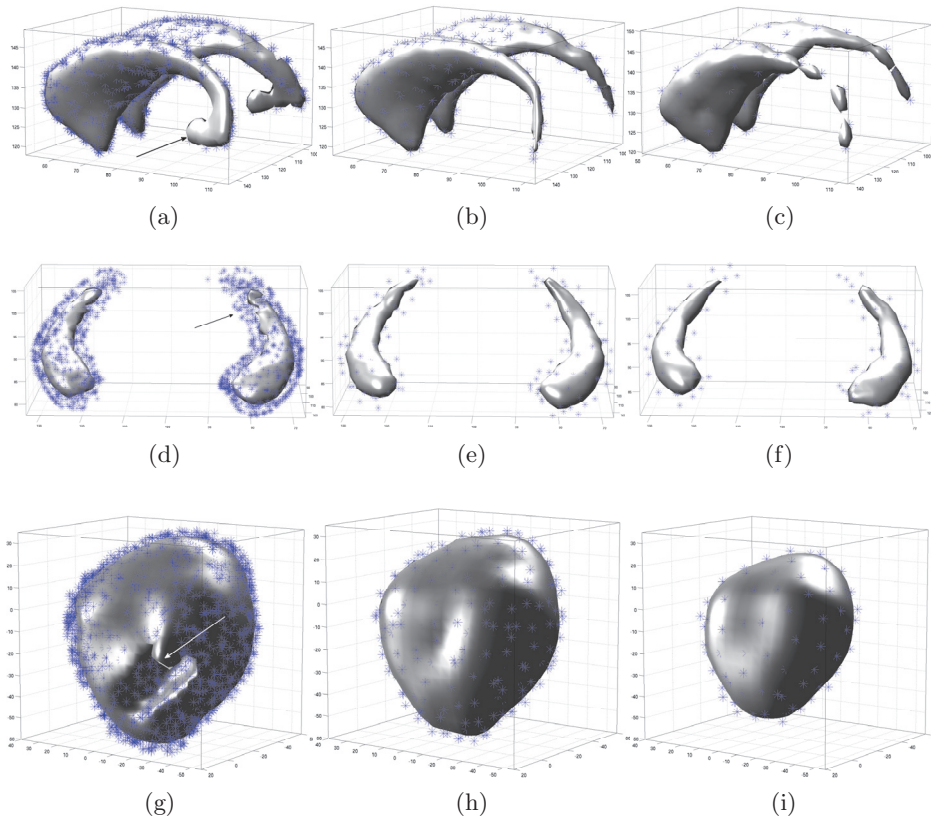

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 11. *Estimated mean shapes from segmented training samples* (a)–(c) *10 caudate nuclei at* $z = 0.01$ *(3194 points),* $z = 0.4$ *(511 points), and* $z = 0.99$ *(145 points);* (d)–(f) *18 hippocampal nuclei at* $z = 0.01$ *(3149 points),* $z = 0.7$ *(243 points), and* $z = 0.99$ *(139 points);* (g)–(i) *15 of prostate at* $z = 0.01$ *(3061 points),* $z = 0.4$ *(361 points), and* $z = 0.99$ *(122 points). Arrows in the left column show the artificial structures generated by overfitting when* $N \approx N_M$ *(section* 3.4*).*
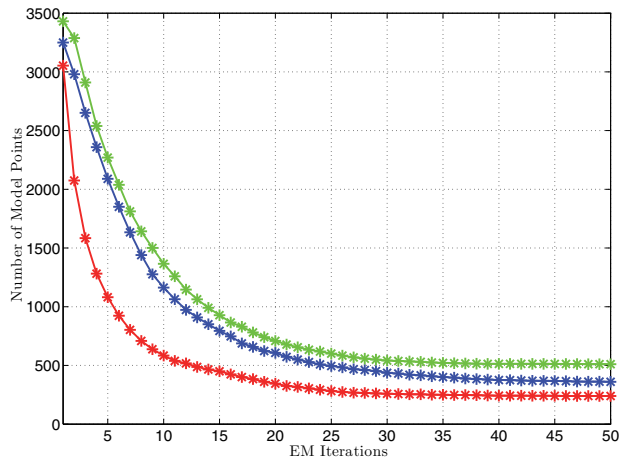
Figure 12. *Convergence behavior of the number model points versus EM iterations: prostate (blue), hippocampus (red), and caudate data sets (green).*

the number of unknown/free parameters approaches the total number of known data. In our case, the numbers of free parameters approximate $N_M \times D$, i.e., the number of points in the mean model multiplied by the dimension of each point. The total number of known data is the number of points in training sets multiplied by $(D+1)$. Conversely, overimposing the sparsity (for example, setting $z = 0.99$) leaves insufficient model points for capturing the anatomical details. The latter, known as an underfitting issue, is illustrated in the right column of Figure 11, i.e., the caudates are disjointed in Figure 11(c), and the hippocampus geometry is coarsened in Figure 11(f). In contrast, in the central panels (Figure 11(b), (e), and (h)), SpSSMs are implemented using z values that minimize the Hausdorff generalization errors. Evidently, the these SpSSMs preserve both the overall geometry and their details, without introducing artifacts.

In Figure 12, the number of model points (N_M) is plotted as a function of the EM iteration number for SpSSM with minimal Hausdorf generalization errors. The N_M is stabilized within 50 iterations regardless of the data sets. On an Intel Core i3-2100 CPU @ 3.10 GHz $\times$ 4 system, running Linux, the entire process stabilizes within 5 minutes. In particular, the runtime is not significantly affected by the QP component incorporated within our EM algorithm.

3.5. Visualization of principal modes of variations. To demonstrate that the proposed model appropriately captures shape variations, the optimal mean models and their variations along the first and second principal components are shown in Figure 13. These models are estimated by the above described cross-validation experiments. To enhance the visualization, the mean model is rendered in red, and the first and second modes of variations are overlain in blue and green, respectively. In the caudate data sets, variation along the first component mainly causes tails to shorten and elongate (Figure 13(a)), whereas variation along the second component causes small oscillations in both nuclei (Figure 13(b)). Similarly, in the hippocampus data sets, the first variation mode mainly corresponds to deformations at the distal end of the structures in Figure 13(c)), whereas the second mode causes horizontal displacement of
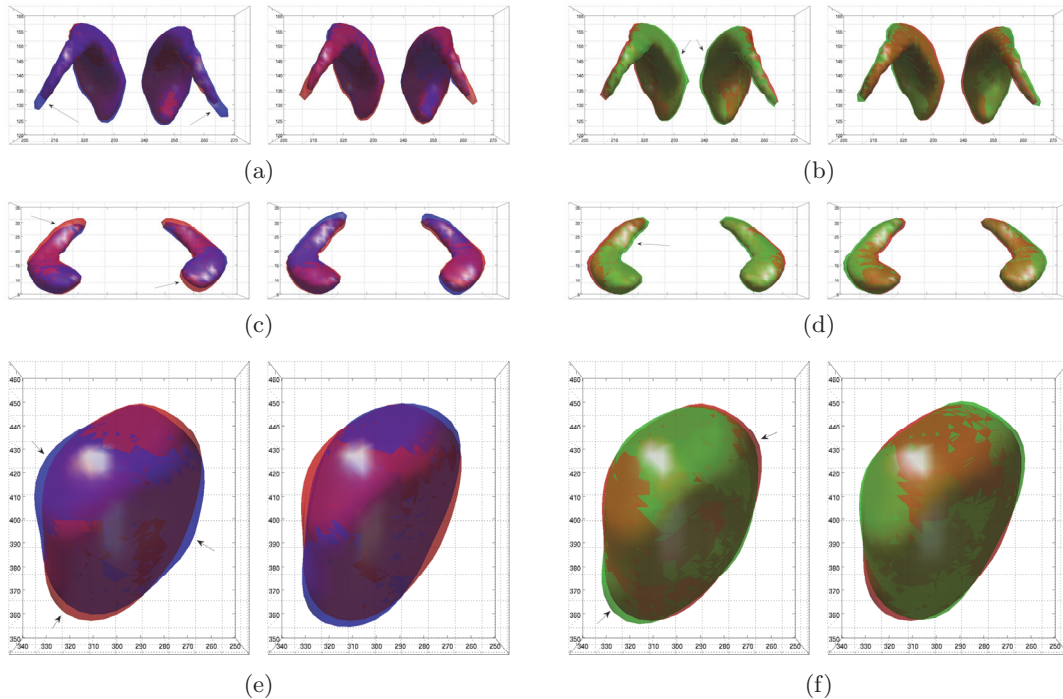
Figure 13. *First and second principal modes of variation in the* (a)–(b) *caudate* (c)–(d), *hippocampi, and* (e)–(f) *prostate populations. In all figures, the mean shape* ($\bar{X}$) *is rendered in red, and other colors specify the positive and negative variations along with the principal modes* $\bar{X} \pm 3\sqrt{\lambda_{1,2}}P_{1,2}$. *Arrows point to the locations of significant variations.*

both nuclei (Figure 13(d)). Finally, the first variation modes of the prostate data sets induces contraction and extensions along the major axes of the structure (Figure 13(e)), while the second mode causes smaller, unidirectional variations.

4. Conclusion. The fundamental hypothesis of this study is that the SSM of an object can be rendered by an optimal number of points. The number of model points (N_M) should be sufficiently small to avoid overfitting yet large enough to capture the structural details. We determined (a range of) N_M by an indirect approach based on a Bayesian estimation framework, which imposes sparsity on the number of mixture components using a symmetric Dirichlet distribution. The proposed method assumes the maximal number of available points and removes points with smallest posterior probabilities in each EM iteration, until no negligible model points are identified. The rate of the pruning process is approximately specified by z, a sparsity controlling hyper-parameter.

We successfully applied our method to learn the variations within segmented caudate, hippocampus, and prostate data sets. The first two of these data sets are nonhomeomorphic to a sphere, rendering them unsuitable for training by the methods proposed in [17] and [33]. Therefore, we compared our sparse model to the Hufnagel et al. model [30] and demonstrated that our model improved the generalization, specificity, and compactness measures. The method proposed by Hufnagel et al. [29] is a generic framework to handle points with any

dimensionality. We used the same point sets (with level sets) to implement and compare both models. Therefore, we argue the improved performance of the sparse models proposed in this paper owes to model construction rather than to the shape representation.

The benefit of exchanging a discrete, N_M, by a continuous, z, parameter can be explained as follows: (1) We have shown that cross-validation over z is advantageous to that over N_M, because if we compare the best possible metrics that could be achieved through each cross-validation scheme (i.e., the minima of generalization/specificity mean errors, and the maxima of compactness through Figures 4–8), we see improvements in favor of cross-validation over z. We argue this improvement responds to the weight estimation for Gaussian components in the mixture. (2) Although the evaluation metrics are optimized differently over various shapes, by intersecting suitable z intervals across all experiments in Figure 9, we identified $0.4 \leq z \leq 0.6$ as a range that might generalize well over different data sets.

In the current framework, z is determined though exhaustive cross-validation experiments, which are computationally demanding and time-consuming. In future study, we will estimate this hyper-parameter by introducing a prior over z, following a Bayesian approach.

5. Appendix. Here, we derive update equations for the parameters $\mathcal{T}, \mathcal{M}$, and Σ, following the principles outlined in [30] and [40]. However, since we newly represent the points as hybrid vectors and adopt a rigid transformation model, our derivations considerably differ from those outlined in [30] and [40].

5.1. Update of spatial transformation parameters $\mathbb{T}$. For a specific value of k, the transformation parameters are updated by (2.5). In the following section, we retain only the $\mathcal{T}_k$ dependent terms and denote other nonrelevant terms as $O.T.$ We have

$$
Q(\Theta|\Theta^{(n)}) = -\frac{1}{2} \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} \big[\|\mathbf{x}_{ki} - s_k \mathbf{R}_k \mathbf{m}_j - \mathbf{b}_k\|^2 / \sigma^2
$$

$$
(5.1) \qquad\qquad + (\phi_k(\mathbf{x}_{ki}) - s_k \phi(\mathbf{m}_j))^2 / \sigma'^2 \big] + O.T.,
$$

which must be maximized for $\mathbf{b}_k, \mathbf{R}_k$, and s_k. Assigning the gradient of w.r.t. $\mathbf{b}_k$ to zero, the update equation is obtained as

$$
(5.2) \qquad\qquad \mathbf{b}_k = \bar{\mathbf{x}}_k - s_k \mathbf{R}_k \bar{\mathbf{m}}_k,
$$

where the barycenters are defined as

$$
S_k = \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)},
$$

$$
\bar{\mathbf{x}}_k = 1/S_k \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} \mathbf{x}_{ki},
$$

$$
(5.3) \qquad\qquad \bar{\mathbf{m}}_k = 1/S_k \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} \mathbf{m}_j.
$$

The update of the rotation matrices under the following constraints requires elaborations:

$$(5.4) \qquad \mathbf{R}_k^T \mathbf{R}_k = \mathbf{I} \ \wedge \ \det(\mathbf{R}_k) = 1.$$

We start with the following lemma [40].

Lemma 1. Let $\mathbf{R}_{D \times D}$ be an unknown rotation matrix and $\mathbf{A}_{D \times D}$ be a known real matrix. Let $\mathbf{USSV}^T$ be a singular value decomposition (SVD) of $\mathbf{A}$, where $\mathbf{UU}^T = \mathbf{VV}^T = \mathbf{I}$ and $\mathbf{SS}$ is a diagonal matrix.

Then the optimal rotation matrix $\mathbf{R}$ that maximizes $\mathrm{tr}(\mathbf{A}^T \mathbf{R})$ is $\mathbf{R} = \mathbf{UCV}^T$, where $\mathbf{C} = \mathrm{diag}(1, 1, \ldots, 1, \det(\mathbf{UV}^T))$.

This lemma implies that we must reformulate (5.1) in terms of the matrix traces. Thereby, given that $\| \cdot \|^2 = (\cdot)^T(\cdot)$ and $\mathbf{R}_k^T \mathbf{R}_k = \mathbf{I}$, we insert (5.2) into (5.1) to obtain

$$(5.5) \qquad Q(\Theta | \Theta^{(n)}) \propto \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T \mathbf{R}_k (\mathbf{m}_j - \bar{\mathbf{m}}_k).$$

Now, using the circular property of the trace operator, we can rewrite (5.5) as

$$Q(\Theta | \Theta^{(n)}) \propto \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} \mathrm{tr}[(\mathbf{m}_j - \bar{\mathbf{m}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T \mathbf{R}_k],$$

which is to be maximized w.r.t. $\mathbf{R}_k$. Therefore, the matrix $\mathbf{A}$ in Lemma 1 is given by

$$(5.6) \qquad \mathbf{A}_k = \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{m}_j - \bar{\mathbf{m}}_k)^T.$$

Hence, we identify $\mathbf{U}, \mathbf{V}$, and $\mathbf{C}$ matrices by SVD decomposition of $\mathbf{A}_k$ and then set $\mathbf{R}_k = \mathbf{UCV}^T$.

To determine the scaling factor s_k, we take the partial derivative of (5.1) w.r.t. s_k and equate it to zero, thereby obtaining

$$(5.7) \qquad s_k = \frac{\mathrm{tr}(\mathbf{A}_k^T \mathbf{R}_k)/\sigma^2 + \sum_{i,j} E_{kij}^{(n)} \phi_k(\mathbf{x}_{ki}) \phi(\mathbf{m}_j)/\sigma'^2}{\mathrm{tr}(\mathbf{D}_k)/\sigma^2 + \sum_{i,j} E_{kij}^{(n)} \phi^2(\mathbf{m}_j)/\sigma'^2},$$

where the matrix $\mathbf{D}_k$ is defined as

$$(5.8) \qquad \mathbf{D}_k = \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} (\mathbf{m}_j - \bar{\mathbf{m}}_k)(\mathbf{m}_j - \bar{\mathbf{m}}_k)^T.$$

Narrow band level sets satisfy $\|\phi(\mathbf{m}_j)\| \ll \|\mathbf{m}_j - \bar{\mathbf{m}}_k\|$ and $\|\phi(\mathbf{x}_{ki})\| \ll \|\mathbf{x}_{ki} - \bar{\mathbf{x}}_k\|$. Therefore, the scales can be simply approximated as $s_k \approx \mathrm{tr}(\mathbf{A}_k^T \mathbf{R}_k)/\mathrm{tr}(\mathbf{D}_k)$.

5.2. Update of model points $\mathcal{M}$. To update the model points $\tilde{\mathbf{m}}_j = [\mathbf{m}_j^T, \phi(\mathbf{m}_j)]^T$, we compute the gradients of (2.5) w.r.t. to $\mathbf{m}_j$ and $\phi(\mathbf{m}_j)$ independently and equate them to zero. Noting that $\mathbf{R}_k^T \mathbf{R}_k = \mathbf{I}$, we obtain the following results:

$$(5.9) \qquad \mathbf{m}_j = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N_k} E_{kij}^{(n)} \mathbf{R}_k^T (\mathbf{x}_{ki} - \mathbf{b}_k)}{\sum_{k=1}^{K} \sum_{i=1}^{N_k} E_{kij}^{(n)} s_k},$$

$$(5.10) \qquad \phi(\mathbf{m}_j) = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N_k} E_{kij}^{(n)} \phi_k(\mathbf{x}_{ki})}{\sum_{k=1}^{K} \sum_{i=1}^{N_k} E_{kij}^{(n)} s_k}.$$

5.3. Update of the covariance matrix Σ. Next, we derive the update equations of the variances by solving $\partial Q(\Theta|\Theta^{(n)})/\partial \sigma = 0$ and $\partial Q(\Theta|\Theta^{(n)})/\partial \sigma' = 0$. The final results are given as below:

$$(5.11) \qquad \sigma^2 = \frac{1}{ND} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} \|\mathbf{x}_{ki} - s_k \mathbf{R}_k \mathbf{m}_j - \mathbf{b}_k\|^2$$

$$(5.12) \qquad \sigma'^2 = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_M} E_{kij}^{(n)} (\phi_k(\mathbf{x}_{ki}) - s_k \phi(\mathbf{m}_j))^2,$$

where N denotes the total number of observed points ($N = \sum_{k=1}^{K} N_k$).

REFERENCES

[1] *MICCAI Grand Challenge: Prostate MR Image Segmentation* 2012, http://promise12.grand-challenge.org/.

[2] B. Awants and J. C. Gee, *Geodesic estimation for large deformation anatomical shape averaging and interpolation*, NeuroImage, 23 (2004), pp. S139–S150.

[3] A. Baumberg and D. Hogg, *An adaptive eigenshape model*, in Proceedings of BMVC, 1995, pp. 87–96.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2009.

[5] S. Bochkanov and V. Bystritsky, *ALGLIB*, http://www.alglib.net.

[6] C. Brechbuhler, G. Gerig, and O. Kohler, *Parametrization of closed surfaces for 3D shape description*, Computer Vision Image Understandings, 61 (1995), pp. 154–170.

[7] A. D. Brett, A. Hill, and C. J. Taylor, *A method of automatic landmark generation for automated 3D PDM construction*, Image Vis. Comput., 18 (2000), pp. 739–748.

[8] C. Butakoff and A. F. Frangi, *A framework for weighted fusion of multiple statistical models of shape and appearance*, IEEE Trans. Pattern Anal. Mach. Intel., 28 (2006), pp. 1847–1857.

[9] J. C. Carr, R. K. Beatson, J. B. Cherrie, T. J. Mitchell, W. R. Fright, B. C. McCallum, and T. R. Evans, *Reconstruction and representation of 3d objects with radial basis functions*, in Proceedings of SIGGRAPH 01, ACM, Springer, New York, 2001, pp. 67–76.

[10] V. Caselles, R. Kimmel, and G. Sapiro, *Geodesic active contours*, Int. J. Comput. Vis., 22 (1997), pp. 61–97.

[11] X. Chen, J. Graham, C. Hutchinson, and L. Muir, *Automatic generation of statistical pose and shape models for articulated joints*, IEEE Trans. Medical Imaging, 33 (2014), pp. 372–383.

[12] H. Chui, A. Rangarajan, J. Zhang, and C. M. Leonard, *Unsupervised learning of an atlas from unlabeled point sets*, IEEE Trans. Pattern Anal. Mach. Intel., 26 (2004), pp. 160–172.

[13] T. F. Cootes and C. J. Taylor, *Active shape models-their training and application*, Comput. Vis. Image Understanding, 61 (1995), pp. 38–59.

[14] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, *Training models of shape from sets of examples*, in Proceedings of the British Machine Vision Conference, Springer-Verlag, New York, 1992, pp. 9–18.

[15] T. F. Cootes, C. J. Twinging, K. O. Babalola, and C. J. Taylor, *Diffeomorphic statistical shape models*, Image Vis. Comput., 26 (2008), pp. 326–332.

[16] A. Corduneanu and C. M. Bishop, *Variational bayesian model selection for mixture distributions*, in Proceedings of the Eighth International Conference on Artificial Intelligence and Statistics, Morgan Kaufmann, San Francisco, 2001, pp. 27–34.

[17] R. H. Davis, C. J. Twinning, T. F. Cootes, and C. J. Taylor, *Building 3-D statistical shape models by direct optimization*, IEEE Trans. Medical Imaging, 29 (2010), pp. 961–982.

[18] R. H. Davis, C. J. Twinning, T. F. Cootes, C. J. Waterton, and C. J. Taylor, *A minimum description length approach to statistical shape modelling*, IEEE Trans. Medical Imaging, 21 (2002), pp. 525–537.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc., 39 (1977), pp. 1–38.

[20] S. Durrleman, M. Prastawa, J. R. Korenberg, S. Joshi, A. Touvx, and G. Gerig, *Topology preserving atlas construction from shape data without correspondence using sparse parameters*, in Medical Image Computing and Computer Assisted Intervention, Lecture Notes in Comput. Sci. 15, 2012, pp. 223–230.

[21] A. F. Frangi, D. Rueckert, J. A. Schnabel, and W. J. Niessen, *Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling*, IEEE Trans. Medical Imaging, 21 (2002), pp. 1151–1165.

[22] L. Garcin, A. Rangarajan, and L. Younes, *Non rigid registration of shapes via diffeomorphic point matching and clustering*, in Proceedings of the International Conference on Image Processing, IEEE, 2004, pp. 3299–3302.

[23] A. Gooya, H. Liao, K. Matsumiya, K. Masamune, Y. Masutani, and T. Dohi, *A variational method for geometric regularization of vascular segmentation in medical images*, IEEE Trans. Image Processing, 17 (2008), pp. 1295–1312.

[24] A. Gooya, E. Mousavi, C. Davatzikos, and H. Liao, *A Bayesian approach for construction of sparse statistical shape models using Dirichlet distribution*, in Proceedings of the 5th International Workshop on Medical Imaging and Augmented Reality (MIAR), 2013, pp. 144–152.

[25] S. Granger and X. Pennec, *Multi-scale EM-ICP: A fast and robust approach for surface registration*, in Proceedings of the European Conference on Computer Vision (ECCV 2002), Springer, New York, 2002, pp. 418–432.

[26] T. Heimann and H. Meinzer, *Statistical shape models for 3D medical image segmentation: A review*, Medical Image Anal., 13 (2009), pp. 543–563.

[27] A. Hill and C. J. Taylor, *Automatic landmark generation for point distribution models*, in Proceedings of the 5th BMVC, 1994, pp. 429–438.

[28] C. Hoogendoorn, N. Duchateau, D. Sanchez-Quintana, T. Whitmarsh, F. M. Sukno, M. De Craene, K. Lekadir, and A. F. Frangi, *A high-resolution atlas and statistical model of the human heart from multislice CT*, IEEE Trans. Medical Imaging, 32 (2013), pp. 28–44.

[29] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, *Surface reconstruction from unorganized points*, SIGGRAPH Comput. Graph., 26 (1992), pp. 71–78.

[30] H. Hufnagel, X. Pennec, J. Ehrhardt, N. Ayache, and H. Handels, *Generation of a statistical shape model with probabilistic point correspondences and EM-ICP*, Int. J. Comput. Assisted Radiology Surgery, 5 (2008), pp. 265–273.

[31] S. Joshi, B. Davis, M. Jomier, and G. Gerig, *Unbiased diffeomorphic atlas construction for computational anatomy*, NeuroImage, 23 (2004), pp. S151–S160.

[32] S. S. Keerthi and E. G. Gilbert, *Convergence of a generalized SMO algorithm for SVM classifier design*, Machine Learning, 46 (2002), pp. 351–360.

[33] A. Kelemen, G. Szekely, and G. Gerig, *Elastic model-based segmentation of 3D neurological data sets*, IEEE Trans. Medical Imaging, 18 (1999), pp. 828–839.

[34] K. J. Khorosh, K. V. Elisevich, S. Patel, and H. Soltanian-Zadeh, *Dataset of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques*, Neuroinformatics, 9 (2011), pp. 335–346.

[35] A. C. W. Kotcheff and C. J. Taylor, *Automatic construction of eigenshape models by direct optimisation*, Medical Image Anal., 2 (1998), pp. 303–314.

[36] H. Lamecker, T. Lange, and M. Seebass, *A statistical shape model for the liver*, in Medical Image Computing and Computer Assisted Intervention, Lecture Notes in Comput. Sci. 2489, Springer, Berlin, 2002, pp. 421–427.

[37] M. E. Leventon, W. E. L. Grimson, and D. O. Faugeras, *Statistical shape influence in geodesic active contours*, in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 1316–1323.

[38] J. Levitt, R. McCarley, C. Dickey, M. Volgmaier, M. Niznikeiwicz, M. L. Seidman, Y. Hirayasu, A. Ciszewski, R. Kikinis, F. Jolesz, and M. Shenton, *MRI study of caudate nucleus volume and its cognitive correlates in neuroleptic-naive patients with schizotypal personality disorder*, Amer. J. Psychiatry, 159 (2002), pp. 1190–1197.

[39] C. Lindner, S. Taigarajah, J. M. Wikinson, G. A. Wallis, and T. F. Cootes, *Fully automatic segmentation of the proximal femur using random forest regression voting*, IEEE Trans. Medical Imaging, 32 (2013), pp. 1462–1472.

[40] A. Myronenko and S. Xubo, *Point set registration: Coherent point drift*, IEEE Trans. Pattern Anal. Mach. Intel., 32 (2010), pp. 2262–2275.

[41] S. Ordas, E. Oubel, R. Leta, F. Carreras, and A. F. Frangi, *A statistical shape model of the heart and its applications to model-based segmentation*, in Proceedings of SPIE, Medical Imaging: Physiology, Function, and Structure from Medical Images, 2007.

[42] J. B. Paul and N. D. McKay, *A method for registration of 3-D shapes*, IEEE Trans. Pattern Anal. Mach. Intel., 14 (1992), pp. 239–256.

[43] R. Paulsen and K. Hilger, *Shape modelling using Markov random field restoration of point correspondences*, in Inform. Process. Med. Imag., 2003, pp. 1–12.

[44] M. Pereanez, K. Lekadir, C. Butakoff, C. Hoogendoorna, and A. F. Frangi, *A framework for the merging of pre-existing and correspondenceless 3D statistical shape models*, Medical Image Anal., 18 (2014), pp. 1044–58.

[45] A. Rangarajan, H. Chui, and F. L. Bookstein, *The soft-assign procrustes matching algorithm*, in Information Processing in Medical Imaging, Springer, New York, 1997, pp. 29–42.

[46] A. Rasoulian, R. Rohling, and P. Abolmaesumi, *Group-wise registration of point sets for statistical shape models*, IEEE Trans. Medical Imaging, 31 (2012), pp. 2025–2033.

[47] A. Rasoulian, R. Rohling, and P. Abolmaesumi, *Lumbar spine segmentation using a statistical multi-vertebrae anatomical shap+pose model*, IEEE Trans. Medical Imaging, 32 (2014), pp. 1890–1900.

[48] M. Rousson, N. Paragios, and R. Deriche, *Implicit active shape models for 3D segmentation in MR imaging*, in Proceedings of Medical Image Computing and Computer Assisted Intervention, St. Mallo, France, 2004, pp. 209–216.

[49] D. Rueckert and A. F. Frangi, *Automatic construction of 3D statistical deformation models of the brain using non-rigid registration*, IEEE Trans. Medical Imaging, 22 (2003), pp. 1014–1025.

[50] D. Shen, Y. Zhan, and C. Davatzikos, *Segmentation of prostate boundaries from ultrasound images using statistical shape model*, IEEE Trans. Medical Imaging, 22 (2003), pp. 539–551.

[51] L. H. Staib and J. S. Duncan, *Boundry finding with parametrically deformable models*, IEEE Trans. Pattern Anal. Mach. Intel., 14 (1992), pp. 1061–1075.

[52] G. Subsol, J. Thirion, and N. Ayache, *A scheme for automatically building three-dimensional morphometric anatomical atlases: Application to a skull atlas*, Medical Image Anal., 2 (1998), pp. 37–60.

[53] M. Vaillant and J. Glaunes, *Surface matching via currents*, in Information Processing in Medical Imaging, Lecture Notes in Comput. Sci. 3565, 2005, Springer, New York, pp. 381–392.

[54] Y. Wang, B. S. Peterson, and L. H. Staib, *Shape-based 3D surface correspondence using geodesics and local geometry*, in Proceedings of IEEE CVPR, 2000, pp. 644–651.

[55] T. Wu, W. Liao, and N. Dai, *Three-dimensional statistical model for gingival contour reconstruction*, IEEE Trans. Biomed. Eng., 59 (2012), pp. 1086–1093.

[56] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. N. Metaxas, and X. S. Zhou, *Deformable segmentation via sparse shape representation*, in Medical Image Computing and Computer Assisted Intervention, Lecture Notes in Comput. Sci., 6892, Springer, New York, 2011, pp. 451–458.

[57] X. Zhang, T. Jie, K. Deng, Y. Wu, and X. Li, *Automatic liver segmentation using a statistical shape model with optimal surface detection*, IEEE Trans. Biomed. Eng., 57 (2010), pp. 828–839.

[58] H. K. Zhao, S. Osher, and R. Fedkiw, *Fast surface reconstruction using the level set method*, in Proceedings of the IEEE Workshop on Variational and Level Set Methods in Computer Vision, Vancouver, BC, Canada, 2001, pp. 194–202.