Article:
Gooya, A, Lekadir, K, Castro-Mateos, I et al. (2 more authors) (2018) Mixture of Probabilistic Principal Component Analyzers for Shapes from Point Sets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40 (4). pp. 891-904. ISSN 0162-8828

https://doi.org/10.1109/TPAMI.2017.2700276

Mixture of Probabilistic Principal Component Analyzers for Shapes from Point Sets

Ali Gooya, *Member, IEEE,* Karim Lekadir, Isaac Castro-Mateos, *Student Member, IEEE,* Jose M Pozo, *Member, IEEE,* and Alejandro F Frangi, *Fellow, IEEE*

Abstract—Inferring a probability density function (pdf) for shape from a population of point sets is a challenging problem. The lack of point-to-point correspondences and the non-linearity of the shape spaces undermine the linear models. Methods based on manifolds model the shape variations naturally, however, statistics are often limited to a single geodesic mean and an arbitrary number of variation modes. We relax the manifold assumption and consider a piece-wise linear form, implementing a mixture of distinctive shape classes. The pdf for point sets is defined hierarchically, modeling a mixture of Probabilistic Principal Component Analyzers (PPCA) in higher dimension. A Variational Bayesian approach is designed for unsupervised learning of the posteriors of point set labels, local variation modes, and point correspondences. By maximizing the model evidence, the numbers of clusters, modes of variations, and points on the mean models are automatically selected. Using the predictive distribution, we project a test shape to the spaces spanned by the local PPCA's. The method is applied to point sets from: i) synthetic data, ii) healthy versus pathological heart morphologies, and iii) lumbar vertebrae. The proposed method selects models with expected numbers of clusters and variation modes, achieving lower generalization-specificity errors compared to state-of-the-art.

Index Terms—Generative Modeling, Variational Bayes, Model Selection, Graphical Models, Statistical Shape Models

1 Introduction

Analysis of the natural morphological variability in a given population of shapes has important applications in various fields of sciences, such as archeology [1], [2], biometrics [3], [4], and medical image analysis [5]. Structured shape variability often exists within and across shape classes. Statistical encoding of these features is highly desirable, but depending on the complexity of the shapes, it can be a challenging task. Given a population of training samples, this problem often boils down to estimating the probability density function (pdf) over a shape space, where each sample is represented as a single point. Thus, shape representation becomes a fundamentally important step for their statistical analysis.

A plethora of shape representation methods and their associated spaces exists in the literature [6]. For instance, shapes can be presented as binary masks obtained by warping a "mean" shape. To analyze morphological variability, principal component analysis (PCA) is applied to either the deformation [7], or the velocity fields generating deformation fields [8]. A more compact and natural shape representation can be achieved by continuous or discrete descriptors of the boundary. For example, continuous curves have been used to define shape spaces as infinite dimensional Riemannian manifolds [9], [10], [11], [12]. To study 3D objects, continuous surfaces can be represented by *medial atoms* [13], where PCA is applied in the tangent space spanned by the Riemannian logarithmic mappings at the Karcher mean. In a simpler setting, surfaces can be parametrized with Fourier [14] and spherical harmonics [15], [16], and PCA can be applied to their corresponding vectors of coefficients. Also, invariant shape comparison under re-parametrization has been proposed in [17] as q-maps from $\mathbb{S}^2$ to $\mathbb{R}^3$. However, these methods are largely limited to study shapes that are homeomorphic to a sphere (or closed curves in 2D) and their extensions to more complex structures requires significant theoretical developments.

Boundary description using a discrete set of points is another prominent shape expression approach. Due to their ability to capture variabilities of complex shapes, not necessarily homeomorphic to spheres, point sets have been widely popular. The pioneering work of Kendall *et al.* [18] showed that shapes having N corresponding D dimensional points naturally live on Riemannian manifolds: quotient spaces of pre-shape space of $D(N-1)-1$ dimensional spheres modulo $SO(D)$. The latter denotes the special orthogonal group, which makes the quotient space invariant under rigid transformations. Despite its mathematical elegance, computing a shape pdf in this space becomes very challenging [19]. The non-linearity of the manifold has been approximated by PCA in the tangent spaces, for instance in [20] for a tracking application. In a simpler setting linear Statistical Shape Models (SSMs), proposed in the seminal work of Coote's *et al.* [21], assume that rigidly aligned point sets lie within a Euclidean space. The hypothesized linearity of the model allows for using causal PCA, and has proven to be a pragmatic solution for many shape matching tasks [6], [22]. However, in the presence of large shape variations, the non linearity of the shape space demands more sophisticated analyses (*e.g.* kernel PCA [23], and PCA in the tangent space of Euclidean special group [24]).

Piece-wise linear models can offer sensible solutions for analysis of non-linear data. For morphological variability

• *Authors are with the center of Computational Imaging & Simulation Technologies in Biomedicine, the University of Sheffield, UK. E-mail: a.gooya@sheffield.ac.uk*

analysis, a shape population can be clustered into subgroups each having more localized principal modes of variability. This approach can also be useful from an application point of view for instance in medical imaging; rather than largely deforming a single mean shape, each subgroup can be associated with a particular disorder, gender, age, etc, and the estimated local means can represent more natural average anatomies. To this end, Cootes *et al.* [25] first clustered the point sets by fitting a Gaussian Mixture Model (GMM), and then applied PCA locally. However, this approach requires having point-to-point correspondences, as well as a predetermined number of clusters and modes.

Establishing point-to-point correspondences across training sets is another major challenge of many point set based shape modeling approaches. Landmarks can slide on explicitly parametrized boundaries to minimize the description length of the Gaussian pdf for shapes topologically equivalent to a sphere [26], [27]. Cates *et al.* in [28] optimized positions of dynamic particles on the implicit surfaces to balance the negative entropy of their distribution on each shape with the positive entropy of the ensemble of shapes. This method can be applied to construct statistically compact models from shapes with arbitrary topologies. In [29], Datar *et al.* extended [28] to a shape-age regression model, where the particle positions and regression parameters are recursively optimized. Alternatively, point correspondences can be resolved without having an explicit or implicit boundary description. Chui *et al.* [30] proposed an Expectation Maximization (EM) framework to iteratively refine the correspondences, mean model, and the deformation fields registering the mean to each case in training sets. Hufnagel *et al.* [31] applied affine transformations for registration, and performed a PCA on the heuristically driven "virtually correspondent" points. Similarly, in [32], the authors used rigid transformations to estimate the emerging mean model and its point count by enforcing sparsity, eliminating insignificant model points. A pair-wise deformable point set registration framework, also based on EM, was proposed by Myronenko *et al.* in [33]. Rasoulian *et al.* extended [33] to a group-wise registration scenario in [34]. However, they rely on a post PCA of the deformation fields to derive SSM, and manually select the model parameters. Although the aforementioned methods mitigate the point correspondences, the application of PCA assumes Gaussian pdfs. In the seminal work [35], Vaillant *et al.* proposed shape representation using *currents* defined as discrete set of barycenters of mesh cells and their corresponding surface normal vectors. A Hilbertian inner product directly defined the distance in the space of the currents, avoiding point correspondences problem. In [36], Durrleman *et al.* derived sparse mean and principal variation modes for the currents. Although elegant, only a single mean was considered disallowing decomposition of the shape space into pathological subtypes. Moreover, a proper Gaussian pdf for currents was not fully developed, thus shape probabilities could not be quantified.

In summary, despite some significant contributions, a rigorous development of normalized shape pdfs on non-linear manifolds is still pending, and the existing solutions are complex and computationally expensive. Statistics are often limited to the variations around a single population mean, not necessarily representing any of the shape subpopulations, with an arbitrary number of variation modes. Relaxing the manifold assumption, we propose a full probabilistic framework that captures the non-linearity of the morphological variations through a piecewise linear model by clustering the population into smaller and more homogeneous groups. As a result, the estimated local means are more typical to shape subgroups associated with particular disorders, gender, etc. Our method is based on a fully Bayesian model, which allows a proper statistical determination of all the discrete parameters (number of clusters/modes) from the data.

In this paper, we present a generative model to infer the pdf for unstructured, rigidly aligned point sets having no point-to-point correspondences. The framework is a piecewise linear model for joint clustering of point sets, and estimating the local modes of variations in each cluster. Points at each set are regarded as samples from a low dimensional GMM, whose means are concatenated to form higher dimensional vectors. These vectors are considered as samples from a Mixture of Probabilistic Principal Component Analyzers (PPCA) [37]. The latter is a high dimensional GMM, where the covariance matrices of its clusters are explicitly decomposed to subspaces of local principal as well as random (isotropic) variations. An inference algorithm based on Variational Bayes (VB) [38], [39] is proposed for unsupervised learning of class labels and variations. In summary, the following contributions are made:

- Using mixture of PPCA, a larger class of shape pdfs is modeled, leading to more realistic group means and local variation modes. Unlike [25], variation modes are explicitly modeled, eliminating the post PCA step. Moreover, we handle point sets having no point-to-point correspondences and derive a lemma for shape prediction and projection to the space spanned by the local PPCA's.
- We propose a *full* Bayesian model and provide an explicit tight lower bound on the model evidence given data. By maximizing the later, discrete parameters such as numbers of clusters and variation modes are determined, enabling automatic model selection.
- Ghahramani *et al.* [39] apply VB for inferring mixtures of subspace analyzers from training *vectors* having equal lengths. We extend it to a challenging case where these vectors are latent and infer them given point sets with variable point counts.

This paper is a comprehensive extension to our preliminary conference paper in [40]: i) We have revised the graphical model, treating the precision of the variation modes as random variables for a more consistent model selection performance, ii) An explicit form for the lower bound on the model evidence is derived, and extensive experiments showing a good performance are demonstrated. We also show that models having higher evidence often result in concurrently small generalization and specificity errors, iii) We study an additional large data set containing 100 vertebra models. Being non-homeomorphic to a sphere, these data sets pose special challenges to train SSMs by Davies *et al.* [27] and Kelemen *et al.* [16]. Thus the results are compared to a closely related state-of-the-art method proposed by Rasoulian *et al.* [34], as well as a PCA based approach

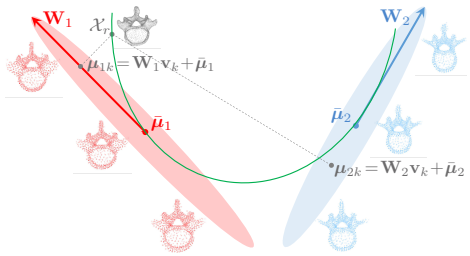


Figure 1. Conceptual representation of the proposed generative model with $J = 2$ PPCA clusters with $L = 1$ principal modes each. Non-linear shape variation (along the green line) is captured in a *piece-wise* form linear around the local means ($\bar{\boldsymbol{\mu}}_j$) using the principal modes ($\mathbf{W}_j$). The projection of a point set $\mathcal{X}_k$ on each space, $\boldsymbol{\mu}_{jk}$, is a linear combination of $\bar{\boldsymbol{\mu}}_j$ (with M model points) and loaded $\mathbf{W}_j$'s ($j = 1, 2$, in this example).

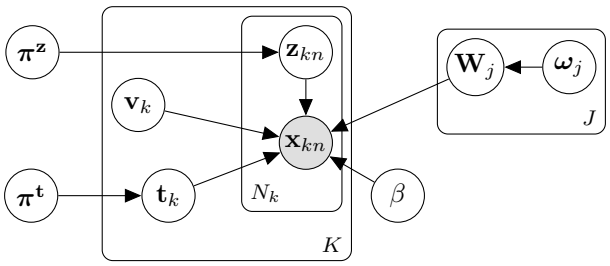


Figure 2. The graphical representation of the proposed model; shaded and hollow circles represent observed and latent variables, respectively, arrows imply the dependencies and plates indicate the number of instances.

proposed by Hufnagel *et al.* [31], iv) A new lemma for shape prediction along with extensive proofs are provided.

In Section 2 our generative model is presented. To derive closed forms for the posteriors, the priors are defined as conjugates to the presumed likelihood distributions. The derivations and the model evidence are given in Appendix A-E, available on-line through supplementary material. In Section 3, we describe our synthetic and real data sets, which are derived from normal and pathological hearts and lumbar vertebra models. The results of model selection and comparison to the state-of-the-art are then provided. Finally, we conclude and discuss the paper in Section 4.

2 Methods

2.1 Probabilistic Generative Model

Our observation consists of K point sets, denoted as $\mathcal{X}_k = \{\mathbf{x}_{kn}\}_{n=1}^{N_k}, 1 \leq k \leq K$, where $\mathbf{x}_{kn}$ is a D dimensional feature vector corresponding to the nth landmark in the kth point set. The model can be explained as two interacting layers of mixture models. In the first (lower-dimension) layer, $\mathcal{X}_k$ is assumed to be a collection of D-dimensional samples from a GMM with M Gaussian components. Meanwhile, by concatenating the means of the GMM (with a consistent order), a vector representation for $\mathcal{X}_k$ can be derived in $M \cdot D$ dimension. Clustering and linear component analysis for $\mathcal{X}_k$ takes place in this space.

More specifically, we consider a mixture of J probabilistic principal component analyzers (MPPCA). A PPCA is essentially an $M \cdot D$-dimensional Gaussian specified by a mean vector, $\bar{\boldsymbol{\mu}}_j \in \mathcal{R}^{MD}$, $1 \leq j \leq J$, and a covariance matrix having a subspace component in the form of $\mathbf{W}_j\mathbf{W}_j^T$ [37]. Here, $\mathbf{W}_j$ is a $MD \times L$ dimensional matrix, whose column l, *i.e.* $\mathbf{W}_j^{(l)}$, represents one mode of variation for the cluster j. Let $\mathbf{v}_k$ be an L dimensional vector of loading coefficients corresponding to $\mathcal{X}_k$ and let us define: $\boldsymbol{\mu}_{jk} = \mathbf{W}_j\mathbf{v}_k + \bar{\boldsymbol{\mu}}_j$. These vectors can be thought of as variables that bridge the two layers of our model: In the higher dimension, $\boldsymbol{\mu}_{jk}$ is a *re-sampled* representation of $\mathcal{X}_k$ in the space spanned by principal components of the jth cluster; meanwhile, if we partition $\boldsymbol{\mu}_{jk}$ into a series of M subsequent vectors, and denote each as $\boldsymbol{\mu}_{jk}^{(m)}$, we obtain the means of D-dimensional Gaussians of the corresponding GMM. A conceptual representation of the proposed generative model summarizing the outlined descriptions is given in Figure 1. Note that in principle the proposed shape space is only a *piece-wise* linear model: local deformations within each shape class are captured linearly using the corresponding class specific PPCA model; whereas more global deformations are associated with differences across various shape classes (hence captured using multiple PPCA's). In this regard, our proposed model is overall non-linear.

Let $\mathcal{Z}_k = \{\mathbf{z}_{kn}\}_{n=1}^{N_k}$ be a set of N_k, 1-of-M coded latent membership vectors for the points in $\mathcal{X}_k$. Each $\mathbf{z}_{kn} \in \{0,1\}^M$ is a vector of zeros except for its arbitrary mth component, where $z_{knm} = 1$, indicating that $\mathbf{x}_{kn}$ is a sample from the D-dimensional Gaussian m. The precision (inverse of the variance) of Gaussians is globally denoted by $\beta\mathbf{I}_D$. Similarly, let $\mathbf{t}_k \in \{0,1\}^J$ be a latent, 1-of-J coded vector whose component j being one ($t_{kj} = 1$) indicates the membership of the $\mathcal{X}_k$ to cluster j. The conditional pdf for $\mathbf{x}_{kn}$ is then given by:

$$p(\mathbf{x}_{kn}|\mathbf{z}_{kn},\mathbf{t}_k,\beta,\mathbb{W},\mathbf{v}_k) = \prod_{j,m} \mathcal{N}(\mathbf{x}_{kn}|\boldsymbol{\mu}_{jk}^{(m)}, \beta^{-1}\mathbf{I}_D)^{z_{knm}t_{kj}} \quad (1)$$

where $\mathbb{W} = \{\mathbf{W}_j\}_{j=1}^J$ is the set of principal component matrices. To facilitate our derivations, we introduce the following prior distributions over $\mathbf{W}_j$, $\mathbf{v}_k$, and β, which are conjugate to the normal distribution in Eqn. (1):

$$p(\mathbf{W}_j|\boldsymbol{\omega}_j) = \prod_l p(\mathbf{W}_j^{(l)}|\omega_{jl}) = \prod_l \mathcal{N}(\mathbf{W}_j^{(l)}|\mathbf{0}, \omega_{jl}^{-1}\mathbf{I}) \quad (2)$$

$$p(\boldsymbol{\omega}_j) = \prod_l p(\omega_{jl}) = \prod_l \mathrm{Gam}(\omega_{jl}|\varepsilon_0, \eta_0) \quad (3)$$

$$p(\mathbf{v}_k) = \mathcal{N}(\mathbf{v}_k|\mathbf{0}, \mathbf{I}) \quad (4)$$

$$p(\beta) = \mathrm{Gam}(\beta|a_0, b_0). \quad (5)$$

Here, we have assumed the columns of $\mathbf{W}_j$ are statistically independent variables having normal distributions. The precision of the lth distribution is given by the corresponding component of the vector $\boldsymbol{\omega}_j$ and denoted by ω_{jl}. We assume that the latter follows a Gamma distribution as we look for a conjugate form to the Gaussian distribution in Eqn. (2). Conjugacy (of the priors to the likelihood distributions) simplifies our close form derivations of the posteriors. Next, we respectively denote the mixture weights of GMMs and MPPCA by $\boldsymbol{\pi}^{\mathbf{z}}$ and $\boldsymbol{\pi}^{\mathbf{t}}$ vectors, each having a Dirichlet distribution as priors:

$$p(\boldsymbol{\pi}^{\mathbf{z}}) = \mathrm{Dir}(\boldsymbol{\pi}^{\mathbf{z}}|\lambda_0^{\mathbf{z}}), \qquad p(\boldsymbol{\pi}^{\mathbf{t}}) = \mathrm{Dir}(\boldsymbol{\pi}^{\mathbf{t}}|\lambda_0^{\mathbf{t}}). \quad (6)$$

The hyper-parameters are set as $\eta_0 = a_0 = b_0 = 10^{-3}$, $\lambda_0^{\mathbf{z}} = \lambda_0^{\mathbf{t}} = 1.0$ and $\ln \varepsilon_0 = -0.5MD \ln(0.5MD)$ (see Appendix Section E)[1].

The conditional distributions of membership vectors of $\mathbf{z}_{kn}$ (for points) and $\mathbf{t}_k$ (for point sets) given mixing weights are specified by two multinomial distributions:

$$p(\mathbf{z}_{kn}|\boldsymbol{\pi}^{\mathbf{z}}) = \prod_m (\pi_m^{\mathbf{z}})^{z_{knm}}, \qquad p(\mathbf{t}_k|\boldsymbol{\pi}^{\mathbf{t}}) = \prod_j (\pi_j^{\mathbf{t}})^{t_{kj}} \quad (7)$$

where $\pi_m^{\mathbf{z}} \geq 0$, $\pi_j^{\mathbf{t}} \geq 0$ are the components m, j of $\boldsymbol{\pi}^{\mathbf{z}}$, $\boldsymbol{\pi}^{\mathbf{t}}$, respectively. We now construct the joint pdf for the sets of all random variables, by assuming (conditional) independence and multiplying the pdfs where needed. Let $\mathbb{X} = \{\mathcal{X}_k\}_{k=1}^K$, $\mathbb{Z} = \{\mathcal{Z}_k\}_{k=1}^K$, $\mathbb{V} = \{\mathbf{v}_k\}_{k=1}^K$, $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_j\}_{j=1}^J$, and $\mathbb{T} = \{\mathbf{t}_k\}_{k=1}^K$, then the distributions of these variables can be written as:

$$p(\mathbb{W}|\boldsymbol{\Omega}) = \prod_j p(\mathbf{W}_j|\boldsymbol{\omega}_j), \quad p(\boldsymbol{\Omega}) = \prod_j p(\boldsymbol{\omega}_j) \quad (8a)$$

$$p(\mathbb{Z}|\boldsymbol{\pi}^{\mathbf{z}}) = \prod_k p(\mathcal{Z}_k|\boldsymbol{\pi}^{\mathbf{z}}),\ p(\mathcal{Z}_k|\boldsymbol{\pi}^{\mathbf{z}}) = \prod_n p(\mathbf{z}_{kn}|\boldsymbol{\pi}^{\mathbf{z}}) \quad (8b)$$

$$p(\mathbb{T}|\boldsymbol{\pi}^{\mathbf{t}}) = \prod_k p(\mathbf{t}_k|\boldsymbol{\pi}^{\mathbf{t}}), \quad p(\mathbb{V}) = \prod_k p(\mathbf{v}_k) \quad (8c)$$

$$p(\mathbb{X}|\mathbb{Z},\mathbb{T},\mathbb{W},\mathbb{V},\beta) = \prod_k p(\mathcal{X}_k|\mathcal{Z}_k,\mathbf{t}_k,\beta,\mathbb{W},\mathbf{v}_k) \quad (8d)$$

$$p(\mathcal{X}_k|\mathcal{Z}_k,\mathbf{t}_k,\beta,\mathbb{W},\mathbf{v}_k) = \prod_n p(\mathbf{x}_{kn}|\mathbf{z}_{kn},\mathbf{t}_k,\beta,\mathbb{W},\mathbf{v}_k). \quad (8e)$$

Lastly, denoting the set of latent variables as $\boldsymbol{\theta} = \{\mathbb{Z},\mathbb{T},\mathbb{W},\boldsymbol{\Omega},\mathbb{V},\boldsymbol{\pi}^{\mathbf{t}},\boldsymbol{\pi}^{\mathbf{z}},\beta\}$, the distribution of the *complete* observation is modeled as

$$\begin{aligned} p(\mathbb{X},\boldsymbol{\theta}) = {} & p(\mathbb{X}|\mathbb{Z},\mathbb{T},\mathbb{W},\mathbb{V},\beta)p(\mathbb{Z}|\boldsymbol{\pi}^{\mathbf{z}})p(\boldsymbol{\pi}^{\mathbf{z}})p(\mathbb{T}|\boldsymbol{\pi}^{\mathbf{t}})p(\boldsymbol{\pi}^{\mathbf{t}}) \\ & \times p(\mathbb{W}|\boldsymbol{\Omega})p(\boldsymbol{\Omega})p(\mathbb{V})p(\beta). \end{aligned} \quad (9)$$

Figure 2 is a graphical representation for the generative model considered in this paper, which shows the hypothesized dependencies of the variables.

2.2 Approximate Inference

Our objective is to estimate the posterior probabilities of the latent variables, given the observed ones, *i.e.* to infer $p(\boldsymbol{\theta}|\mathbb{X})$. However, this direct inference is analytically intractable thus an approximated distribution, $q(\boldsymbol{\theta})$, is sought. Owing to the dimensionality of the data, we prefer Variational Bayes (VB) over sampling based methods. The VB principle for obtaining $q(\boldsymbol{\theta})$ is explained briefly. The logarithm of the model evidence, *i.e.* $\ln p(\mathbb{X})$ [2], can be decomposed as $\ln p(\mathbb{X}) = \mathcal{L} + \mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbb{X}))$, where $0 \leq \mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leilber divergence, and

$$\mathcal{L} = \int q(\boldsymbol{\theta}) \ln \frac{p(\mathbb{X},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \leq \ln p(\mathbb{X}) \quad (10)$$

is a lower bound on $\ln p(\mathbb{X})$. To obtain $q(\boldsymbol{\theta})$, the KL divergence between the true and the approximated posterior should be minimized. However, this is not feasible because the true posterior is not accessible to us. On the other hand, minimizing KL is equivalent to maximizing $\mathcal{L}$ w.r.t. $q(\boldsymbol{\theta})$ since $p(\mathbb{X})$, as the left side of the relation above, is independent of $q(\boldsymbol{\theta})$. Thus, $q(\boldsymbol{\theta})$ can be computed by maximizing $\mathcal{L}$ as a tight lower bound on $\ln p(\mathbb{X})$.

We approximate the true posterior as a factorized form, *i.e.* , $q(\boldsymbol{\theta}) = \prod_i q(\theta_i)$, where $q(\cdot)$ is the approximated posterior, and θ_i refers to any of our latent variables. This factorization leads to the following tractable result: let θ_i be the variable of interest in $\boldsymbol{\theta}$, then the variational posterior of θ_i can be derived using

$$\ln q(\theta_i) = \langle \ln p(\mathbb{X},\boldsymbol{\theta}) \rangle_{\boldsymbol{\theta}-\theta_i} + \text{o.t.} \quad (11)$$

where $p(\mathbb{X},\boldsymbol{\theta})$ is given in Eqn. (9), $\langle\cdot\rangle_{\boldsymbol{\theta}-\theta_i}$ denotes the expectation w.r.t. to the product of $q(\cdot)$ of all variable in $\boldsymbol{\theta} - \theta_i$, and o.t. refers to terms not depending on θ_i. Notice that these variational posteriors are coupled, thus starting from an initialized status, we iteratively update them until a convergence or a maximum number of iterations is arrived.

2.3 Update of Posteriors

In this section, we provide update equations for the variational posteriors. Due to conjugacy of priors to likelihoods, these derivations are done by inspecting expectations of logarithms and matching posteriors to the corresponding likelihood template forms. To keep our notation uncluttered, we use the following conventions: i) Unless mentioned by an explicit sub-index, $\langle\cdot\rangle$ denotes the expectation w.r.t. the $q(\cdot)$ distributions of all random variables in the angles except for the variable in the query, ii) Sub-indices $(\cdot)_l$ and $(\cdot)_{lr}$ specify element numbers in a vector or matrix, respectively, iii) Parenthetical super-indices $(\cdot)^{(m)}$, $(\cdot)^{(m,n)}$ specify the D and $D \times D$ dimensional block numbers of the MD and $MD \times MD$ vectors and matrices, respectively, iv) A single numbered super-index such as (l) applied to a matrix specifies the lth column in the matrix.

2.3.1 Update of $q(\mathbb{Z})$

Starting from $\mathbb{Z}$ variables, following Eqn. (11) and given Eqn. (1) we have

$$\begin{aligned} \ln q(\mathcal{Z}_k) &= \sum_n \Big[\langle \ln p(\mathbf{x}_{kn}|\mathbf{z}_{kn},\mathbf{t}_k,\beta,\mathbb{W},\mathbf{v}_k) \rangle + \langle \ln p(\mathbf{z}_{kn}|\boldsymbol{\pi}^{\mathbf{z}}) \rangle \Big] \\ &= \sum_{n,m} z_{knm} \Big[\sum_j \langle t_{kj} \rangle \langle \ln \mathcal{N}(\mathbf{x}_{kn}|\boldsymbol{\mu}_{jk}^{(m)}, \beta^{-1}\mathbf{I}_D) \rangle \\ &\quad + \langle \ln \pi_m^{\mathbf{z}} \rangle \Big] + \text{o.t.} \\ &= \sum_{n,m} z_{knm} \ln \rho_{knm} + \text{o.t.} \end{aligned} \quad (12a)$$

$$\ln \rho_{knm} = -\frac{\langle\beta\rangle}{2} \sum_j \langle t_{kj} \rangle \langle |\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2 \rangle + \langle \ln \pi_m^{\mathbf{z}} \rangle. \quad (12b)$$

This result implies that $q(\mathcal{Z}_k) \propto \prod_{m,n} \rho_{knm}^{z_{knm}}$, and given the fact that $\sum_m q(z_{knm}) = 1$, we arrive at the following results

$$q(\mathcal{Z}_k) = \prod_{m,n} (r_{knm})^{z_{knm}}, \quad q(\mathbb{Z}) = \prod_k q(\mathcal{Z}_k) \quad (13)$$

where $r_{knm} = \rho_{knm} / \sum_{m'} \rho_{knm'}$, and $\langle z_{knm} \rangle = r_{knm}$ [38]. Furthermore, by noticing that $\langle \boldsymbol{\mu}_{jk}^{(m)} \rangle = \langle \boldsymbol{\mu}_{jk} \rangle^{(m)}$ and

1. The model evidence is insensitive to these settings due to summation of the hyper-parameters with larger values.

2. More precisely, $p(\mathbb{X})$ is conditioned on parameters with no prior distribution. Hence, it is equivalently referred to as marginal likelihood.

$\text{Cov}[\boldsymbol{\mu}_{jk}^{(m)}] = \text{Cov}[\boldsymbol{\mu}_{jk}]^{(m,m)}$, the first term in Eqn. (12b) can be directly computed using the expectations of $\mathbb{W}$ and $\mathbb{V}$ as

$$\langle|\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2\rangle = |\mathbf{x}_{kn} - \langle\boldsymbol{\mu}_{jk}\rangle^{(m)}|^2 + \text{Tr}\big[\text{Cov}[\boldsymbol{\mu}_{jk}]^{(m,m)}\big] \quad (14)$$

where

$$\langle\boldsymbol{\mu}_{jk}\rangle = \langle\mathbf{W}_j\rangle\langle\mathbf{v}_k\rangle + \bar{\boldsymbol{\mu}}_j, \quad (15a)$$

$$\text{Cov}[\boldsymbol{\mu}_{jk}] = \langle\mathbf{W}_j\rangle\text{Cov}[\mathbf{v}_k]\langle\mathbf{W}_j\rangle^T + \sum_l \langle\mathbf{v}_k\mathbf{v}_k^T\rangle_{ll}\text{Cov}[\mathbf{W}_j^{(l)}] \quad (15b)$$

A proof for (15b) is given in Appendix A (available on line).

2.3.2 Update of $q(\mathbb{T})$

Next, we compute the variational posterior of T variables. Following Eqn. (11) we have

$$\begin{aligned}\ln q(\mathbf{t}_k) &= \sum_j t_{kj}\Big[\sum_{m,n}\langle z_{knm}\rangle\langle\ln\mathcal{N}(\mathbf{x}_{kn}|\boldsymbol{\mu}_{jk}^{(m)}, \beta^{-1}\mathbf{I}_D)\rangle + \langle\ln\pi_j^{\mathbf{t}}\rangle\Big] \\ &= \sum_j t_{kj}\ln\rho'_{kj} + \text{o.t.} \quad (16a)\end{aligned}$$

$$\ln\rho'_{kj} = -\frac{\langle\beta\rangle}{2}\sum_{m,n} r_{knm}\langle|\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2\rangle + \langle\ln\pi_j^{\mathbf{t}}\rangle. \quad (16b)$$

Ignoring all j-independent terms the above equation can be written as

$$\begin{aligned}\ln\rho'_{kj} = &-\frac{\langle\beta\rangle}{2}\sum_m\Big(\sum_n r_{knm}\Big)\langle|\boldsymbol{\mu}_{jk}^{(m)}|^2\rangle + \langle\ln\pi_j^{\mathbf{t}}\rangle \\ &+\langle\beta\rangle\sum_m\langle\boldsymbol{\mu}_{jk}^{(m)}\rangle^T\Big(\sum_n r_{knm}\mathbf{x}_{kn}\Big) + \text{o.t.} \quad (17)\end{aligned}$$

To simplify the rest our notations, we introduce the following auxiliary variables:

$$R_{km} = \sum_n r_{knm} \quad (18a)$$

$$\mathbf{R}_k = \text{Diag}(\underbrace{R_{k1}\cdots R_{k1}}_{D\text{ copies}}, \cdots, \underbrace{R_{kM}\cdots R_{kM}}_{D\text{ copies}}) \quad (18b)$$

$$\bar{\mathbf{x}}_{km} = \sum_n r_{knm}\mathbf{x}_{kn}, \quad \bar{\mathbf{x}}_k = [\bar{\mathbf{x}}_{k1}^T, \cdots, \bar{\mathbf{x}}_{kM}^T]^T. \quad (18c)$$

Plugging Eqn. (18b-18c) back into Eqn. (17), it is easy to see that

$$\ln\rho'_{kj} = \langle\beta\rangle\text{Tr}\big[-\frac{1}{2}\mathbf{R}_k\langle\boldsymbol{\mu}_{jk}\boldsymbol{\mu}_{jk}^T\rangle + \langle\boldsymbol{\mu}_{jk}\rangle\bar{\mathbf{x}}_k^T\big] + \langle\ln\pi_j^{\mathbf{t}}\rangle. \quad (19)$$

Now, comparing Eqn. (16) to Eqn. (12), and following the results obtained in Eqn. (13), we can write

$$q(\mathbf{t}_k) = \prod_j(r'_{kj})^{t_{kj}}, \qquad q(\mathbb{T}) = \prod_k q(\mathbf{t}_k) \quad (20)$$

where $r'_{kj} = \rho'_{kj}/\Sigma_{j'}\rho'_{kj'}$, and $\langle t_{kj}\rangle = r'_{kj}$.

2.3.3 Update of $q(\mathbb{W})$

To obtain the posterior of the principal components, following Eqn. (11), we have

$$\begin{aligned}\ln q(\mathbf{W}_j^{(l)}) = &\sum_{k,n,m} r_{knm}\langle t_{kj}\rangle\langle\ln\mathcal{N}(\mathbf{x}_{kn}|\boldsymbol{\mu}_{jk}^{(m)}, \beta^{-1}\mathbf{I}_D)\rangle_{\mathbf{W}_j^{(\neq l)},\mathbf{v}_k} \\ &+\langle\ln\mathcal{N}(\mathbf{W}_j^{(l)}|\mathbf{0}, \omega_{jl}^{-1}\mathbf{I})\rangle_{\omega_{jl}} + \text{o.t.} \\ = &-\frac{\langle\beta\rangle}{2}\sum_{k,m,n}\langle t_{kj}\rangle r_{knm}\langle|\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2\rangle_{\mathbf{W}_j^{(\neq l)},\mathbf{v}_k} \\ &-\frac{\langle\omega_{jl}\rangle}{2}|\mathbf{W}_j^{(l)}|^2 + \text{o.t.} \quad (21)\end{aligned}$$

Using the auxiliary variables introduced through Eqn. (18b-18c), and in an analogy to the result in Eqn. (19), Eqn. (21) can be written as

$$\begin{aligned}\ln q(\mathbf{W}_j^{(l)}) = &\langle\beta\rangle\sum_k\langle t_{kj}\rangle\text{Tr}\big[\frac{-1}{2}\langle\mathbf{R}_k\boldsymbol{\mu}_{jk}\boldsymbol{\mu}_{jk}^T + \boldsymbol{\mu}_{jk}\bar{\mathbf{x}}_k^T\rangle_{\mathbf{W}_j^{(\neq l)},\mathbf{v}_k}\big] \\ &-\frac{\langle\omega_{jl}\rangle}{2}|\mathbf{W}_j^{(l)}|^2 + \text{o.t.} \quad (22)\end{aligned}$$

In Appendix B we have shown that with further elaboration, the posterior of $\mathbf{W}_j^{(l)}$ is a normal distribution with the following mean and covariance

$$\text{Cov}[\mathbf{W}_j^{(l)}] = \big[\langle\omega_{jl}\rangle\mathbf{I} + \langle\beta\rangle\sum_k\langle t_{kj}\rangle\langle\mathbf{v}_k\mathbf{v}_k^T\rangle_{ll}\mathbf{R}_k\big]^{-1}, \quad (23a)$$

$$\langle\mathbf{W}_j^{(l)}\rangle = \langle\beta\rangle\text{Cov}[\mathbf{W}_j^{(l)}]\sum_k\langle t_{kj}\rangle\mathbf{Q}_{\text{kj}}{}^{(l)}, \quad (23b)$$

$$q(\mathbf{W}_j^{(l)}) = \mathcal{N}\big(\mathbf{W}_j^{(l)}|\langle\mathbf{W}_j^{(l)}\rangle, \text{Cov}[\mathbf{W}_j^{(l)}]\big) \quad (23c)$$

with the auxiliary matrix $\mathbf{Q}_{\text{kj}}$ defined as

$$\begin{aligned}\mathbf{Q}_{\text{kj}} = &\bar{\mathbf{x}}_k\langle\mathbf{v}_k\rangle^T - \mathbf{R}_k\bar{\boldsymbol{\mu}}_j\langle\mathbf{v}_k\rangle^T \\ &-\mathbf{R}_k\langle\mathbf{W}_j\rangle\Big[\langle\mathbf{v}_k\mathbf{v}_k^T\rangle - \text{Diag}(\text{diag}\langle\mathbf{v}_k\mathbf{v}_k^T\rangle)\Big] \quad (24)\end{aligned}$$

where the inner diag operator copies the main diagonal of $\langle\mathbf{v}_k\mathbf{v}_k^T\rangle$ into a vector, and the outer Diag transforms the vector back into a diagonal matrix. Thus, the posteriors for modes of variations are given as $q(\mathbb{W}) = \prod_{j,l} q(\mathbf{W}_j^{(l)})$.

2.3.4 Update of $q(\mathbb{V})$

Next, we compute the a variational posterior form for the loading vectors $\mathbb{V}$. By referring to Eqn. (11), for vector $\mathbf{v}_k$ we have

$$\begin{aligned}\ln q(\mathbf{v}_k) = &\sum_{j,n,m}\langle t_{kj}\rangle r_{knm}\langle\ln\mathcal{N}(\mathbf{x}_{kn}|\boldsymbol{\mu}_{jk}^{(m)}, \beta^{-1}\mathbf{I}_D)\rangle_{\mathbf{W}_j,\beta} \\ &+\langle\ln\mathcal{N}(\mathbf{v}_k|\mathbf{0}, \mathbf{I})\rangle + \text{o.t.} \\ = &-\frac{\langle\beta\rangle}{2}\sum_{j,n,m}\langle t_{kj}\rangle r_{knm}\langle|\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2\rangle_{\mathbf{W}_j} - \frac{1}{2}\mathbf{v}_k^T\mathbf{v}_k + \text{o.t.} \\ = &-\frac{\langle\beta\rangle}{2}\sum_j\langle t_{kj}\rangle\text{Tr}\big[\mathbf{R}_k\langle\boldsymbol{\mu}_{jk}\boldsymbol{\mu}_{jk}^T\rangle_{\mathbf{W}_j} - 2\langle\boldsymbol{\mu}_{jk}\rangle_{\mathbf{W}_j}\bar{\mathbf{x}}_k^T\big] \\ &-\frac{1}{2}\mathbf{v}_k^T\mathbf{v}_k + \text{o.t.} \quad (25)\end{aligned}$$

The last identity follows from Eqn. (17) and auxiliary variables introduced in Eqn. (18b-18c). As shown in Appendix C, with further simplification of the right hand side of Eqn. (25), we derive $q(\mathbf{v}_k)$ as the following normal distribution

$$q(\mathbf{v}_k) = \mathcal{N}(\mathbf{v}_k|\langle\mathbf{v}_k\rangle, \text{Cov}[\mathbf{v}_k]), \quad q(\mathbb{V}) = \prod_k q(\mathbf{v}_k) \quad (26a)$$

$$\text{Cov}[\mathbf{v}_k] = \Big[\mathbf{I} + \langle\beta\rangle\sum_j\langle t_{kj}\rangle\langle\mathbf{W}_j^T\mathbf{R}_k\mathbf{W}_j\rangle\Big]^{-1} \quad (26b)$$

$$\langle\mathbf{v}_k\rangle = \langle\beta\rangle\text{Cov}[\mathbf{v}_k]\sum_j\langle t_{kj}\rangle\langle\mathbf{W}_j\rangle^T(\bar{\mathbf{x}}_k - \mathbf{R}_k\bar{\boldsymbol{\mu}}_j) \quad (26c)$$

2.3.5 Update of $q(\beta)$

Similarly, the posterior of the precision variable β can be obtained as follows

$$\begin{aligned}\ln q(\beta) &= \sum_{k,n,m,j} \langle t_{kj}\rangle r_{knm} \langle \ln \mathcal{N}(\mathbf{x}_{kn}|\boldsymbol{\mu}_{jk}^{(m)}, \beta^{-1}\mathbf{I}_D)\rangle \\ &\quad + \ln \text{Gam}(\beta|a_0, b_0) + \text{o.t.} \\ &= \frac{ND}{2}\ln\beta - \frac{\beta}{2}\sum_{k,n,m,j}\langle z_{knm}\rangle\langle t_{kj}\rangle\langle|\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2\rangle \\ &\quad + (a_0 - 1)\ln\beta - b_0\beta + \text{o.t.} \end{aligned} \tag{27}$$

Factoring terms linear in β and $\ln\beta$, it is easy to see that the posterior is a Gamma distribution specified by

$$q(\beta) = \text{Gam}(\beta|a, b), \tag{28a}$$

$$b = b_0 + \frac{1}{2}\sum_{k,n,m,j}\langle z_{knm}\rangle\langle t_{kj}\rangle\langle|\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2\rangle \tag{28b}$$

$$a = a_0 + ND/2. \tag{28c}$$

Notice that to compute the expectation in the right hand side of Eqn. (28b), we use the identity in Eqn. (14). Under these definitions, we have $\langle\beta\rangle = a/b$ and $\langle\ln\beta\rangle = \psi(a) - \ln(b)$, where ψ is the *Digamma* function [38].

2.3.6 Update of $q(\boldsymbol{\pi}^{\mathbf{t}}), q(\boldsymbol{\pi}^{\mathbf{z}})$

Taking the logarithm of Eqn. (9) and the expectation according to Eqn. (11), we have

$$\ln q(\boldsymbol{\pi}^{\mathbf{t}}) = (\lambda_0^{\mathbf{t}} - 1)\sum_j \ln\pi_j^{\mathbf{t}} + \sum_{k,j}\langle t_{kj}\rangle \ln\pi_j^{\mathbf{t}} + \text{o.t.} \tag{29a}$$

$$\ln q(\boldsymbol{\pi}^{\mathbf{z}}) = (\lambda_0^{\mathbf{z}} - 1)\sum_m \ln\pi_m^{\mathbf{z}} + \sum_{k,n,m}\langle z_{knm}\rangle \ln\pi_m^{\mathbf{z}} + \text{o.t.} \tag{29b}$$

Factoring linear forms in $\ln\pi_m^{\mathbf{z}}$ and $\ln\pi_j^{\mathbf{t}}$, it is easy to see that the posteriors of the mixing coefficients are Dirichlet distributions defined by the following identities

$$\lambda_j^{\mathbf{t}} = \lambda_0^{\mathbf{t}} + \sum_k \langle t_{kj}\rangle, \qquad q(\boldsymbol{\pi}^{\mathbf{t}}) = \text{Dir}(\boldsymbol{\pi}^{\mathbf{t}}|\boldsymbol{\lambda}^{\mathbf{t}}) \tag{30a}$$

$$\lambda_m^{\mathbf{z}} = \lambda_0^{\mathbf{z}} + \sum_{k,n}\langle z_{knm}\rangle, \qquad q(\boldsymbol{\pi}^{\mathbf{z}}) = \text{Dir}(\boldsymbol{\pi}^{\mathbf{z}}|\boldsymbol{\lambda}^{\mathbf{z}}) \tag{30b}$$

Using (30a) and (30b), the expectations related to the mixing coefficients are computed as $\langle\pi_m^{\mathbf{z}}\rangle = \lambda_m^{\mathbf{z}}/\sum_{m'}\lambda_{m'}^{\mathbf{z}}$, and $\langle\ln\pi_j^{\mathbf{t}}\rangle = \psi(\lambda_j^{\mathbf{t}}) - \psi(\sum_{j'}\lambda_{j'}^{\mathbf{t}})$.

2.3.7 Update of $q(\boldsymbol{\Omega})$

To compute the posteriors of the precision variables in $\boldsymbol{\Omega}$, we first consider the posterior of ω_{jl}. From Eqn. (9), and following Eqn. (11) we have

$$\begin{aligned}\ln q(\omega_{jl}) &= \langle \ln\mathcal{N}(\mathbf{W}_j^{(l)}|\mathbf{0}, \omega_{jl}^{-1}\mathbf{I})\rangle_{\mathbf{W}_j^{(l)}} + \ln\text{Gam}(\omega_{jl}|\varepsilon_0, \eta_0) \\ &= \frac{MD}{2}\ln\omega_{jl} - \frac{1}{2}\omega_{jl}\langle|\mathbf{W}_j^{(l)}|^2\rangle \\ &\quad + (\varepsilon_0 - 1)\ln\omega_{jl} - \eta_0\omega_{jl} + \text{o.t.} \end{aligned} \tag{31}$$

Therefore $q(\omega_{jl})$ can be written as the following Gamma distribution

$$q(\omega_{jl}) = \text{Gam}(\omega_{jl}|\varepsilon_{jl}, \eta_{jl}), \quad q(\boldsymbol{\Omega}) = \prod_{j,l} q(\omega_{jl}) \tag{32a}$$

$$\eta_{jl} = \eta_0 + \frac{1}{2}\langle|\mathbf{W}_j^{(l)}|^2\rangle, \quad \varepsilon_{jl} = \varepsilon_0 + MD/2. \tag{32b}$$

Furthermore, based on the results obtained in (32a)-(32b), we can compute expectations of $\langle\omega_{jl}\rangle = \varepsilon_{jl}/\eta_{jl}$, and $\langle\ln\omega_{jl}\rangle = \psi(\varepsilon_{jl}) - \ln(\eta_{jl})$.

2.3.8 Update of $\bar{\boldsymbol{\mu}}_j$

Finally, by maximizing the lower bound in Eqn. (10) with w.r.t. $\bar{\boldsymbol{\mu}}_j$, we obtain the following closed form expression

$$\bar{\boldsymbol{\mu}}_j = \Big[\sum_k \langle t_{kj}\rangle \mathbf{R}_k\Big]^{-1}\Big[\sum_k \langle t_{kj}\rangle(\bar{\mathbf{x}}_k - \mathbf{R}_k\langle\mathbf{W}_j\rangle\langle\mathbf{v}_k\rangle)\Big]. \tag{33}$$

2.4 The explicit form for the lower bound

In Appendix E, we have shown that the explicit form of the lower bound on the model evidence can be derived as

$$\begin{aligned}\mathcal{L} = &-a\ln b + \sum_{j,l}\big[-\varepsilon_{jl}\ln\eta_{jl} + \frac{1}{2}\ln|\mathbf{W}_j^{(l)}|\big] \\ &+\ln\Gamma(M\lambda_0^{\mathbf{z}}) - M\ln\Gamma(\lambda_0^{\mathbf{z}}) + \sum_m \ln\Gamma(\lambda_m^{\mathbf{z}}) - \ln\Gamma(\sum_m \lambda_m^{\mathbf{z}}) \\ &+\ln\Gamma(J\lambda_0^{\mathbf{t}}) - J\ln\Gamma(\lambda_0^{\mathbf{t}}) + \sum_j \ln\Gamma(\lambda_j^{\mathbf{t}}) - \ln\Gamma(\sum_j \lambda_j^{\mathbf{t}}) \\ &- \sum_{k,n,m} r_{knm}\ln r_{knm} - \sum_{k,j}\langle t_{jk}\rangle\ln\langle t_{jk}\rangle \\ &+ \sum_k\big[\frac{L}{2} + \frac{1}{2}\big(\ln|\text{Cov}[\mathbf{v}_k]| - \langle|v_k|^2\rangle\big)\big]. \end{aligned} \tag{34}$$

In Section 3.2, we maximize this to select the optimal discrete parameters such as M, L, and J, given data.

2.5 Shape Projection Using Predictive Distribution

For a new test point set $\mathcal{X}_r = \{\mathbf{x}_{rn}\}_{n=1}^{N_r}$, with $K < r$, we can obtain a model projected point set as $\hat{\mathcal{X}}_r = \{\langle\hat{\mathbf{x}}_{rn}\rangle\}_{n=1}^{N_r}$, where

$$\langle\hat{\mathbf{x}}_{rn}\rangle = \int \hat{\mathbf{x}}_{rn} p(\hat{\mathbf{x}}_{rn}|\mathcal{X}_r, \mathbb{X}) d\hat{\mathbf{x}}_{rn}. \tag{35}$$

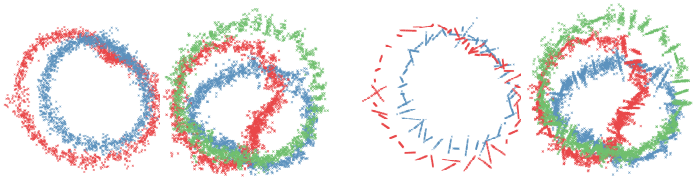
Here, the predictive distribution should be computed by marginalizing the corresponding latent and model variables by

$$\begin{aligned}p(\hat{\mathbf{x}}_{rn}|\mathcal{X}_r,\mathbb{X}) = &\sum_{\mathbf{z}_{rn},\mathbf{t}_r}\int p(\hat{\mathbf{x}}_{rn}|\mathbf{z}_{rn},\mathbf{t}_r,\beta,\mathbb{W},\mathbf{v}_r) \\ &\times p(\mathbf{z}_{rn},\mathbf{t}_r,\beta,\mathbb{W},\boldsymbol{\Omega},\mathbf{v}_r|\mathcal{X}_r,\mathbb{X}) d\mathbb{W} d\mathbf{v}_r d\boldsymbol{\Omega} d\beta. \end{aligned} \tag{36}$$

Because this integral is analytically intractable, we use an approximation for the posterior assuming the factorized form

$$p(\mathbf{z}_{rn},\mathbf{t}_r,\beta,\mathbb{W},\boldsymbol{\Omega},\mathbf{v}_r|\mathcal{X}_r,\mathbb{X}) \approx q(\mathbf{z}_{rn})q(\mathbf{t}_r)q(\mathbf{v}_r)q(\beta)q(\mathbb{W})q(\boldsymbol{\Omega})$$

Thus, having $\mathcal{X}_r$ we iterate over updating $q(\mathbf{z}_{rn})$, $q(\mathbf{t}_r)$ and $q(\mathbf{v}_r)$, and replace $q(\beta)$ and $q(\mathbb{W})$ from the training step, ignoring the influence of $\mathcal{X}_r$ on $\mathbb{W}$ and β. This process isolates the test and training phases. Thus, the generalization errors are directly associated with the quality of the off-line trained models. Under these approximations, we show that a closed form expression for point projection can be obtained using the following lemma



(a) (b)

Figure 3. Clustering and mode estimation of synthetic point sets, color coded by their types. (a) *Left*: overlay of $K = 80$ point sets generated using $\hat{M} = 50$ model points, $\hat{J} = 2$ clusters, and $\hat{L} = 1$ variation mode, *Right*: overlay of $K = 120$ point sets sampled from $\hat{M} = 50$ model points, $\hat{J} = 3$ clusters, and $\hat{L} = 2$ variation modes; (b) Overlay of the estimated corresponding clustering and variation modes. The match of the colors and major structures between (a) and (b), shows a good clustering and estimation of principal variation modes.

Lemma 1. *Given the definitions in* (35) *and* (36)*, the projection of point* $\mathbf{x}_{rn}$ *can be computed as*

$$\langle \hat{\mathbf{x}}_{rn} \rangle = \sum_{j,m} \langle t_{jr} \rangle \langle z_{rnm} \rangle \langle \boldsymbol{\mu}_{jr} \rangle^{(m)}. \tag{37}$$

Furthermore, $\hat{\mathbf{x}}_{rn}$ *is placed within the convex hull made by* $\langle \boldsymbol{\mu}_{jr} \rangle^{(m)}$ *points.*

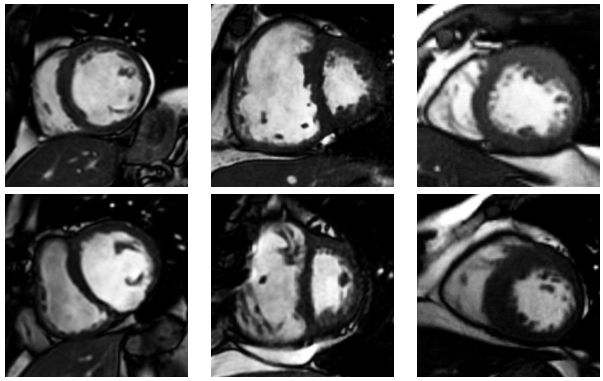
We use (37) to obtain the model predicted point set $\hat{\mathcal{X}}_r$ given $\mathcal{X}_r$. A proof is given in Appendix D (available on line).

2.6 Initialization and Computational Burden

To initialize the clusters of point sets, we adopt ideas from text clustering, where each document (point set) is represented as a *bag of features* (BOF) vector [41], [42]. In order to construct our set of frequent "geometric" words, we consider identifying M locations with dense populations of D dimensional points. To that end, a GMM with M Gaussians is fit to the set of all available points. Next, for any point set such as $\mathcal{X}_k$ an M dimensional BOF vector is constructed by: computing posterior probability of Gaussian component m given a point in $\mathcal{X}_k$, and then summing these posteriors over all points in $\mathcal{X}_k$. Next, the computed vectors are clustered using a k-means algorithm [43], and the labels are used to initialize cluster means and variation modes using the following procedure. For the Gaussian component m in the GMM, a corresponding point from $\mathcal{X}_k$ is identified having the maximum posterior probability in $\mathcal{X}_k$. Iterating over M Gaussian components, all the corresponding points from $\mathcal{X}_k$ are identified and concatenated to form an MD dimensional vector. This procedure is then repeated over K training point sets. Next, by applying PCA at each cluster, we identify the mean $\bar{\boldsymbol{\mu}}_j$, $\mathbf{W}_j$ as the first L components, and $\mathbf{v}_k$ as the projections of the original vectors to these components. Finally, β is initialized as the component wise average L2 difference of the original and the PCA projected vectors. We have observed that for 50 point sets, each having 4000 landmarks, a convergence is achieved by 30 VB iterations in nearly an hour (see Figure 10).

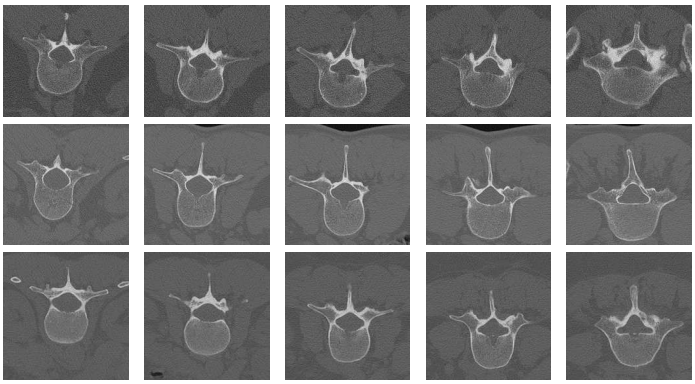
3 Results

We evaluate our method using synthetic and real data sets obtained from cardiac MR and vertebra CT images. The reliability of the lower bound in Eqn. (34) as a criterion to select discrete parameters of the model (*i.e.* , the numbers of: point set clusters J, modes of variations L, and model points M) is demonstrated for all data types.



(a) (b) (c)

Figure 4. Short axis MR images from normal (a), PH (b), and HCM patients (c). Compared to normal hearts: the RV in the PH patients tend to be larger, and LV in the HCM patients appears thicker.



(a) (b) (c) (d) (e)

Figure 5. Short axis CT images from lumbar for three sample patients. Columns (a)-(e) correspond to L1-L5 vertebrae, respectively. Compared to L1-L4 vertebrae, the body of the L5 vertebra in column (e) seem more flattered around the pedicles, forming "bat" like structures.

We also measure generalization and specificity errors, and compare them to the state-of-the-art. Generalization quantifies the error between the actual and the model projected point sets. Specificity is related to the ability of the model to instantiate correct samples resembling the training data. We run cross validations by dividing the point sets into the testing and training subsets. Next, we measure the generalization by quantifying the average distance between the test point sets and their model projected variants. To measure specificity, random point sets are generated, and the average of the minimum distances between each sample and training point sets is computed [27]. Three distance metrics are considered, namely

$$d(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \frac{1}{N_k} \sum_{\mathbf{x}} \min_{\mathbf{y} \in \hat{\mathcal{X}}_k} ||\mathbf{x} - \mathbf{y}||_2$$

where N_k is the number of points in $\mathcal{X}_k$, and $\hat{\mathcal{X}}_k$ denote the model projected point set obtained in (37). Since d is asymmetric, we also compute $d^*(\mathcal{X}_k, \hat{\mathcal{X}}_k) = d(\hat{\mathcal{X}}_k, \mathcal{X}_k)$.

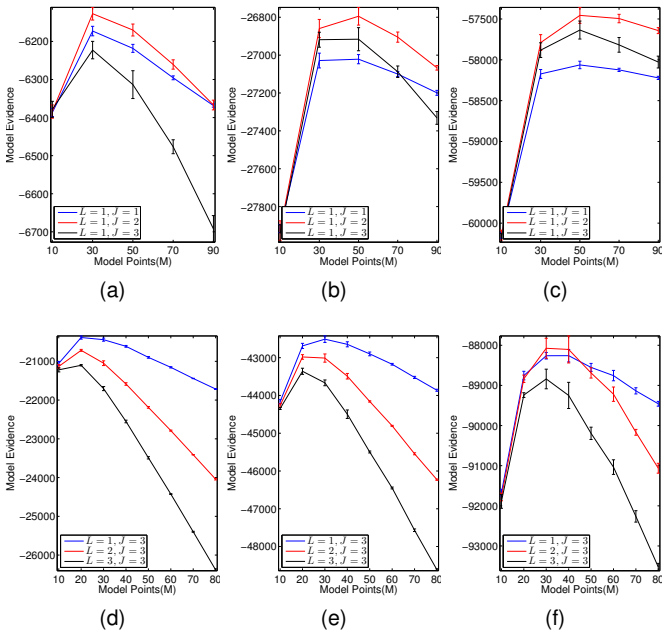


Figure 6. Model evidence versus M for different numbers of point sets generated in Figure 3.(a): (a)$K = 10$, (b)$K = 40$, and (c)$K = 80$ samples from mixture of two clusters each having one mode of variation; (d)$K = 30$, (e)$K = 60$, and (f)$K = 120$ samples from mixture of three clusters with two modes of variations. As the number of training samples (K) increases maximal evidences are attained at $M = \hat{M}$ and correct models are selected.

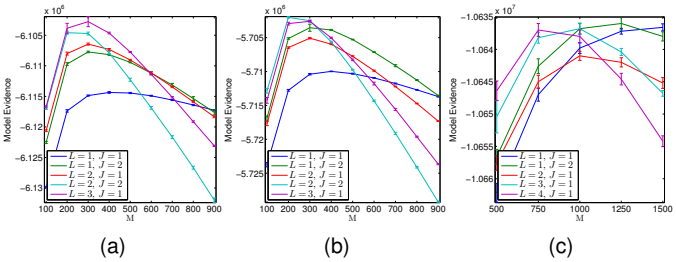


Figure 7. Model evidence versus M for mixtures of: (a)55 Normal-PH, (b)53 Normal-HCM, and (c)100 vertebra point sets.

These quantities measure how two point sets are similar on the average basis but are not suitable to detect difference in the details of $\mathcal{X}_k$ and $\hat{\mathcal{X}}_k$. Therefore, we also measure the Hausdorff distance

$$d_H(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \max\Big(\max_{\mathbf{x}\in\mathcal{X}_k} \min_{\mathbf{y}\in\hat{\mathcal{X}}_k} ||\mathbf{x}-\mathbf{y}||_2, \max_{\mathbf{y}\in\hat{\mathcal{X}}_k} \min_{\mathbf{x}\in\mathcal{X}_k} ||\mathbf{x}-\mathbf{y}||_2\Big).$$

We compare our model to the basic PCA approach proposed in [31], and arguably closest work in the literature proposed by Rasoulian *et al.* in [34]. In an analogy to our method, these methods construct SSMs directly from point sets with no correspondences.

3.1 Description of Point Sets

Synthetic point sets: Given the dependencies of the variables, ancestral sampling [38] was used to draw 2D point sets from our generative model. Setting $M = 50$, we sampled from mixtures of $J = 2$ clusters with $L = 1$ mode of variation, $J = 3$ clusters having $L = 2$ modes, and $J = 3$ clusters with $L = 3$ modes. The cluster means ($\bar{\boldsymbol{\mu}}_j$'s) form radially modulated rings with overlap to make the clustering challenging (see Figure 3.a).

Cardiac point sets: Three groups of cardiac data sets including: 33 normals, 22 subjects with Pulmonary Hypertension (PH), and 20 subjects with Hypertrophic Cardiomyopathy (HCM) were considered. These data sets were acquired using 1.5 MR scanners, resulting in image matrices of $256\times256\times12$ in short axial direction and slice thicknesses of 8-10 mm. To derive cardiac surfaces, the initial shapes were obtained by labeling the MRI slices, then fitting surface meshes to the binary images. Each surface mesh was made using 4000 vertices and registered using [32] to remove scaling, rotation and translation before our analysis.

These subjects differ in their cardiac morphologies. For PH patients, which are associated with pulmonary vascular proliferation [44], complex shape remodeling of both the left and right ventricles occurs. As a result, the RV becomes very dilated, pushing onto the LV, which deforms and loses its roundness [45]. On the other hand, HCM [46] is a condition in which the muscle of the heart shows an excessive thickening, and the most characteristic feature is a hypertrophied LV (asymmetric thickening involving the ventricular septum, see Figure 4).

We ignore the patient labels and cluster two populations made of Normal-PH and Normal-HCM patients, independently. By evaluating the lower bound for each population for different numbers of clusters and modes, and investigate whether the proposed lower bound can correctly identify the underlying number of morphological classes.

Vertebra point sets: The dataset was composed of 20 CT scans from patients suffering from low back pain. The lumbar L1 to L5 vertebrae (shown in Figure 5) in each patient were manually segmented. Then, the binary masks were converted to surface meshes each having 4000 vertices. We then registered these vertices, removing scaling, rotation and translation and used them for the subsequent clustering and variation analysis. The morphologies of lumbar vertebrae are perceived differently, in particular, when L1-L4 vertebra samples in Figure 5.(a)-(d) are compared to L5 samples in (e) having "bat" like patterns. In the next section, we show that the model evidence is maximum when we consider two clusters representing L1-L4 and L5 classes.

3.2 Model Selection

In this section, our objective is to evaluate the lower bound for different settings of model parameters (J, L, M), and verify that the maximum $\mathcal{L}$ (model evidence) is attained at the correct values of those parameters. To that end, we use the synthetic point sets generated with known ground truth parameters ($\hat{J}$, $\hat{L}$, and $\hat{M}$). By varying the number of training point sets (K), we show that for adequately large K, the correct model (with maximum $\mathcal{L}$ at $\hat{J}$, $\hat{L}$, and $\hat{M}$) is selected. Furthermore, to remove the bias made by initialization, we fit the model 5 times and report the means values and standard deviations.

3.2.1 *Model evidence versus variable M*

Figure 6 shows the variation of the $\mathcal{L}$ versus M for the synthetic point sets shown in Figure 3.a (generated from

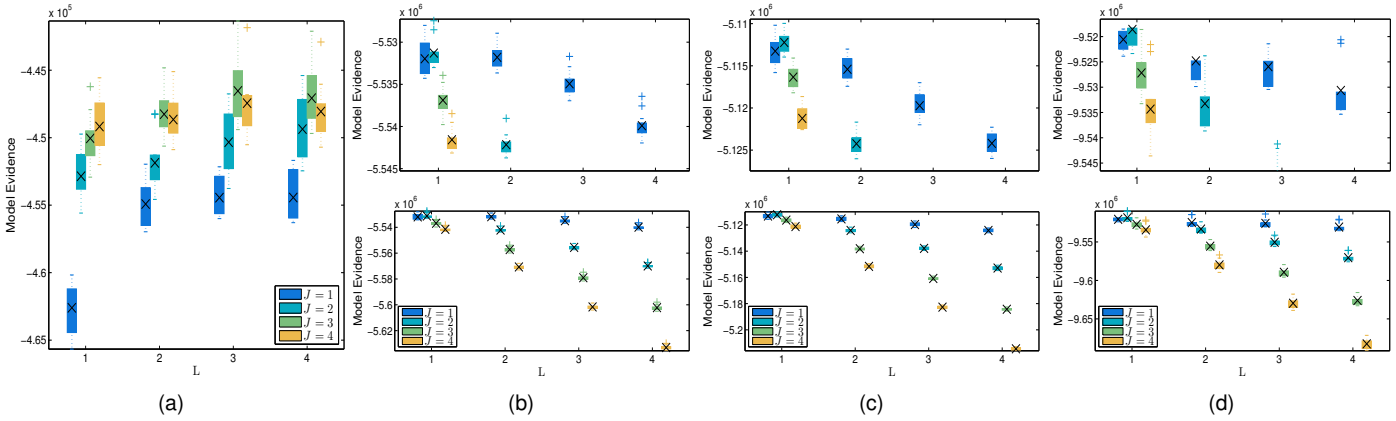


Figure 8. Model evidence versus L and J parameters, evaluated through ten fold cross-validations. At each fold, a model was fit to the training data, thus obtaining 10 different $\mathcal{L}$ values for various mixtures of: (a) Synthetic samples obtained with ground truth parameters of $\hat{L} = 3$ and $\hat{J} = 3$, (b) Normal-PH, (c) Normal-HCM, and (d) vertebra point sets. The upper charts in (b)-(d) correspond to the lower counterparts in a finer scale. Comparing the means (denoted by crosses), it can be noticed that maximal model evidences ($\mathcal{L}$) happen in the correct underlying model in (a), and clinically plausible models in (b)-(d), indicating presence of two clusters.

mixtures having ground truth parameters of $\hat{M} = 50$, $\hat{J} = 2$, $\hat{L} = 1$, and $\hat{M} = 50$, $\hat{J} = 3$, $\hat{L} = 2$). For each setting, the number of observed point sets K varied in an increasing order through panels (a)-(c), and (d)-(f), respectively. In both cases, when K is small (*i.e.* (a) and (d)), the model evidence is maximal for overly simple models with M, J and L values smaller than expected ground truth parameters. However, as K increases, the maximals evidences are correctly attained at the corresponding $\hat{M}$, $\hat{L}$, and $\hat{J}$ (*i.e.* (c) and (f)) used to generate the data; suggesting that the maximum $\mathcal{L}$ can identify the correct model with sufficient training samples. Note that due to marginalization over latent variables (*i.e.* , $p(\mathbb{X}) = \int p(\mathbb{X}, \boldsymbol{\theta}) d\boldsymbol{\theta}$) in full Bayesian models, the lower bound on $p(\mathbb{X})$ is penalized for large models [38] having redundant number of parameters/model points.

A similar set of experiments were conducted using three mixtures of real data sets made of 55 Normal-PH, 53 Normal-HCM, and 100 vertebra point sets. Fixing L and J values, we vary M and show the results in Figure 7. Notice that in these cases the "correct" models are not known *a priori*, therefore we only justify the selected models based on clinical or physiological interpretations. Considering clustering of Normal-PH cases in (a), we notice the maximum $\mathcal{L}$ is achieved in $M = 300$, $L = 3$, and $J = 1$, suggesting the presence of only one cluster. This clinically controversial result, however, can be explained using our analysis of synthetic point sets. In fact, referring to Figure 6, we already observed that having a sufficient number of training point sets is crucial to discover the correct underlying model. Hence, we believe that having more cases of PH and normal cases will lead into selection of models having more clusters and larger M values.

The evaluation of $\mathcal{L}$ for Normal-HCM mixture in Figure 7.(b) shows that the model specified by $M = 200$, $L = 2$, and $J = 2$ has the largest evidence, which is the expected number of clusters due to presence of two types of heart models in the mixture. For vertebra data sets, Figure 7.(c) reveals that the model having $M = 1250$, $L = 1$ and $J = 2$ clusters is optimal. The two clusters correspond to L5 and L1-L4 vertebrae, which is expected due to large morphological discrepancy between these groups (see Figure 5).

Notice that M is significantly larger for optimal models trained using vertebrae samples compared to cardiac data sets, which can be due to the larger number of available training samples in the former case. Indeed, for synthetic data, we saw that with increasing number of training samples, models with larger M values show higher evidence. In the rest of our analysis with cardiac point sets, however, we use models having $M = 800$ points, $J = 2$ classes and $L = 1$ mode of variation. This allows us to represent the general structure of the heart using an adequate numbers of model points.

3.2.2 *Model evidence versus variable L and J*

Next, we investigate the suitability of $\mathcal{L}$ to select the correct number of clusters and modes, under fixed M values (obtained from previous section). Because our objective is also to link the model evidence to specificity and generalization errors in the next section, we perform 10 folds cross-validations, obtaining 10 $\mathcal{L}$ values for each L and J settings. Also, for the rest of experiments using synthetic samples, we generate a new set of 150 point sets from a model having $\hat{J} = 3$ clusters and $\hat{L} = 3$ modes of variations. As shown in Figure 8.(a), the maximal evidence is correctly found for this setting. Also, evaluations using mixtures of Normal-PH, Normal-HCM, and Vertebrae cases in Figure 8.(b)-(d) reveals the maximum of $\mathcal{L}$ for $L = 1$ mode and $J = 2$ clusters, which seem to be plausible models.

3.2.3 *Sensitivity to hyper-parameters*

As stated before, due to conjugacy the variational posteriors have the same form as the prior distributions. However, more importantly, their parameters are updated to secondary values different from the hyper-parameters used in the priors. For instance, through (30a) and (30b), $\boldsymbol{\lambda}^{\mathbf{t}}$ and $\boldsymbol{\lambda}^{\mathbf{z}}$ (controlling sparsity of the Dirichlet distributions) are effectively determined by data related terms as long as $\lambda_0^{\mathbf{t}} \ll \sum_k \langle t_{kj} \rangle$ and $\lambda_0^{\mathbf{z}} \ll \sum_{k,n} \langle z_{knm} \rangle$. Under these conditions, exact setting of the hyper-parameters is not critical.

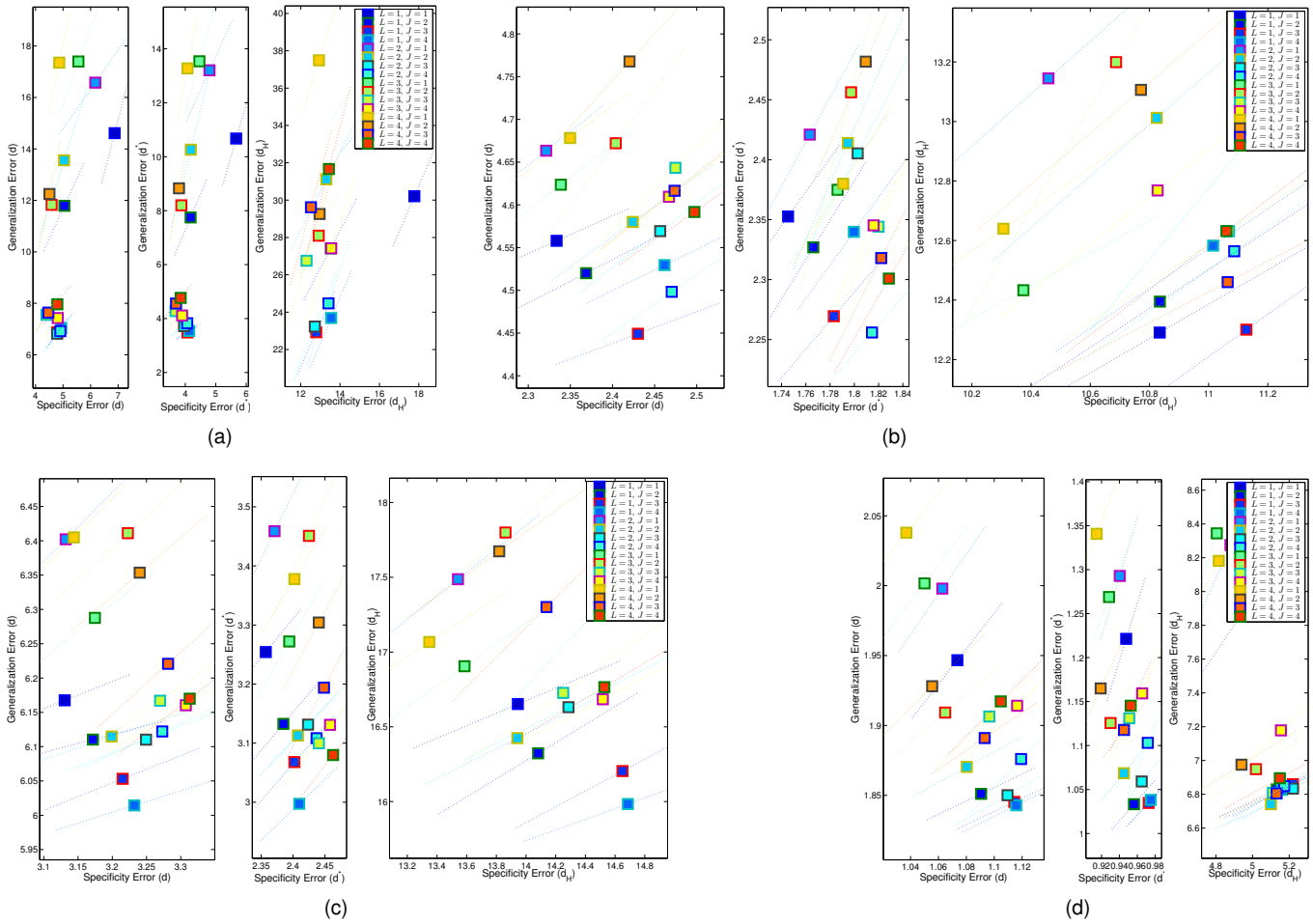


Figure 9. Generalization and Specificity errors (in $[mm]$) of models trained through 10-fold cross-validations using point sets from mixtures of: (a)Synthetic samples from a model having the ground truth parameters of $\hat{L}, \hat{J} = 3$, (b)Normal-PH, (c)Normal-HCM, and (d)vertebrae. In each panel, d, d^*, and d_H distances are quantified from left to right. For each model, markers and dotted lines indicate the corresponding average and rough variability of the errors, respectively. The models having maximal evidences (in Figure 8) are placed competitively close to the lower-left corner in each graph, indicating concurrent small generalization and specificity errors.

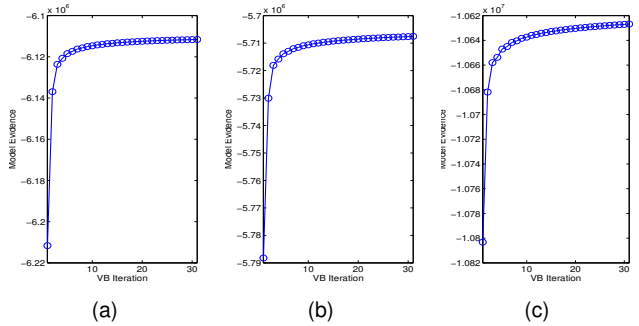


Figure 10. Model evidence versus VB iterations showing convergence and maximizing the lower bound for mixtures of: (a) Normal-PH, (b)Noraml-HCM, and (c)vertebra point sets.

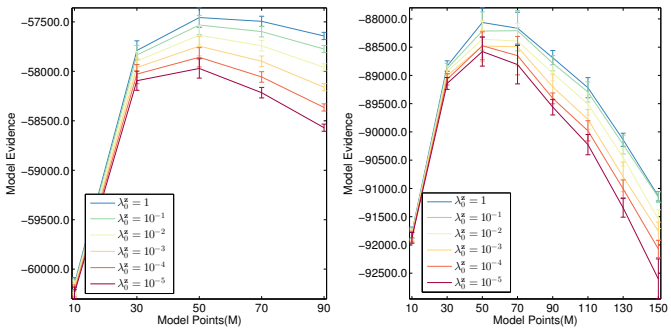


Figure 11. Model evidence for synthetic point sets with $L=1, J=2$ (a), and $L=2, J=3$ (b), showing robust maximums at $M = \hat{M} = 50$ for different $\lambda_0^{\mathbf{z}}$ values.

A similar update mechanism for other parameters can be seen in (28b)-(28c), and (32b). Consequently, the behavior of the model evidence in (34), as a function of these secondary parameters, remains relatively invariant to initial settings of the hyper-parameters. This is shown in Figure 11, where the model evidence for the synthetic point sets in Figure 3 is plotted by varying the number of Gaussian components (M) and $\lambda_0^{\mathbf{z}}$. As seen, reducing the latter penalizes models with large M's more heavily (favoring more sparse models). However, across this range of $\lambda_0^{\mathbf{z}}$, the maximal evidences are correctly found in $M = 50$ used to generate data sets.

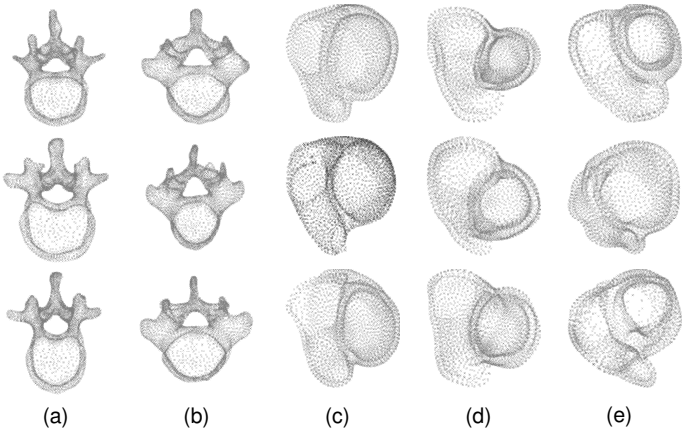


Figure 12. Sample clustered point sets as: L1-L4(a) and L5(b) vertebrae, Normal(c), PH(d), and HCM(e) hearts.

3.2.4 Generalization/specificity versus model evidence

We noticed that, given enough training data, the lower bound in Eqn. (34) can select a correct or clinically plausible model. A natural question that arises here is: how does the model evidence relate to more tangible quality measures such as specificity and generalization errors? To answer the question, we quantify these errors for the range of models trained through our cross-validations in Figure 8. For every model and mixture type considered, we measure the errors in terms of d/d^*, and d_H distances introduced earlier.

Figure 9 shows how model evidence relates to specificity and generalization errors for the mixtures of: Synthetic ($\hat{L}, \hat{J} = 3$), Normal-PH, Normal-HCM, and Vertebrae in panels (a)-(d), respectively. It is interesting to notice that models with largest evidence values (in Figure 8) correspond to those placed generally close to the lower-left corners of generalization-specificity planes in Figure 9, indicating *concurrently* small errors of both types. For instance, the models specified by $J = 2$ and $L = 1$, showing largest evidence in Figure 8.(d), is the closest to the lower left corners in the left (d) and middle (d^*) panels in Figure 9.(d). These observation suggest that models with higher evidence generally avoid large errors in both benchmarks.

3.2.5 Validation of maximization and clustering errors

Figure 10.(a)-(c) show how the proposed lower bound on the model evidence is maximized by iterating through the update equations of the posteriors for the Normal-PH, Normal-HCM and Vertebra data sets, respectively. As seen, a good convergence is usually achieved within 30 VB iterations, experimentally validating our derivations. Finally, we noticed that in clustering mixture of cardiac point sets, 2 out of 22 PH cases, and 2 out of 20 HCM cases were clustered as normal data sets. Moreover, none of the L5 vertebrae were clustered in L1-L4 group and vice versa. Sample clustered point sets are shown in Figure 12.

3.2.6 Comparison to State-of-The-Art

Having our models selected, we now compare them to state-of-the-art in terms of generalization and specificity errors. We consider the method proposed by Rasoulian *et al.* [34] because the hypotheses taken by this method resemble our assumptions: it constructs SSMs directly from a group of point sets, and handles complex topologies and lack of point-to-point correspondences. We additionally compare our model to the PCA based approach proposed by Hufnagel *et al.* [31] at equal number of model points (M).

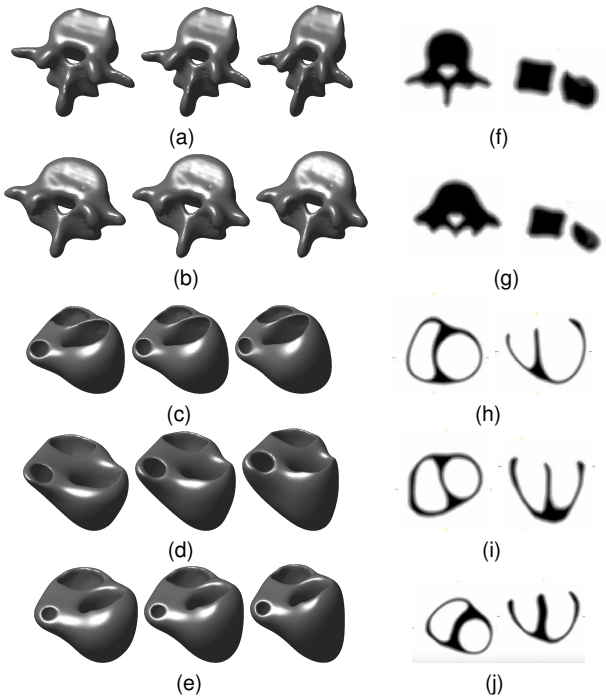


Figure 13. Means and variation modes for L1-L4(a), L5(b), Normal(c), PH(d), and HCM(e) point sets with mean models in the middle and variations in opposite directions at two sides, (f-j) axial and coronal cross sections of the mean models for each population.

For real point sets, we see that the models with highest evidence have $L = 1$ mode and $J = 2$ clusters (see Figure 8), therefore, we chose these models for further comparison. Furthermore, we set $M = 800$ for the mixtures of cardiac point sets, and $M = 1250$ for the vertebrae when constructing PCA models using the reference methods. To determine the number of required PCA modes, we select the minimum number of modes covering 95% of the trace of the covariance matrix. The results summarized in Table 1 show that in the majority of the average distances, our approach outperforms the methods proposed in [34] and [31]. This can be due to the model averaging mechanism that exists in (37), *i.e.* weighting the prediction according to clustering ($\langle t_{kj} \rangle$), and soft-correspondence variables ($\langle z_{knm} \rangle$).

3.3 Visualization, and qualitative results

To visualize the means and variation modes for the real point sets, we use our implementation of the proposed method in [47] to reconstruct surfaces from the computed mean point sets. We first construct an unsigned distance map from the points by fitting 2D planes to local point subsets, then computing distances from these planes. Next, a geodesic active contour is driven towards the point set using advection on the distance map.

The 3D representations, cross sections of the means, and principal variation modes are visualized in Figure 13.

Table 1
Generalization and Specificity errors (in mm) for the methods proposed in [34], [31], and the selected models with $L = 1$, $J = 2$ (significant differences are in bold (p-value < 0.001)).

		Generalization			Specificity		
		Norm-PH	Norm-HCM	Vertebrae	Norm-PH	Norm-HCM	Vertebrae
Method proposed in [34]	d	**3.89±0.23**	5.96±0.25	**1.60±0.17**	3.68±0.02	4.96±0.04	1.47±0.03
	d^*	3.17±0.47	4.48±0.49	1.72±0.27	2.82±0.06	3.95±0.10	1.50±0.07
	d_H	20.75±3.11	27.60±4.70	12.17±1.92	13.75±0.93	18.56±1.21	7.82±0.45
Method proposed in [31]	d	6.95±1.03	6.01±0.22	1.82±0.26	5.18±0.10	5.21±0.07	1.73±0.05
	d^*	8.72±0.95	3.58±0.23	1.1±0.21	4.21±0.14	3.22±0.11	1.10±0.06
	d_H	26.2±2.25	18.60±3.09	7.44±1.72	15.57±1.12	14.31±0.79	6.29±0.27
Our method ($J = 2, L = 1$)	d	4.52±0.18	6.11±0.11	1.85±0.07	**2.36±0.35**	**3.17±0.42**	**1.09±0.15**
	d^*	**2.32±0.28**	**3.13±0.28**	**1.03±0.11**	**1.76±0.17**	**2.38±0.22**	**0.95±0.09**
	d_H	**12.39±1.80**	**16.32±2.02**	**6.82±0.80**	**10.83±2.75**	14.08±3.35	**5.13±1.55**

As seen in panels (a) and (b), the mean of L5 cluster is significantly wider in centrum and shorter in the transverse process, compared to the mean of L1-L4 cluster. Moreover, the variation around the latter generally involves an expansion of the vertebra body in the lateral direction and changes over length and rotation of the transverse processes. Considering the fact that we have normalized the scaling during the registration, this mode of variation matches the second principal mode extracted by Rasoulian *et al.* [34].

Furthermore, in the normal heart, shown in Figure 13.(c) and (h), LV is significantly larger than RV, and when compared to PH ((d) and (f)) and HCM ((e) and (j)), it is more spherical. On the other hand, in the PH heart, the RV is evidently dilated and the LV loses its roundness. Finally, significant thickening of the septum and shrinkage of LV are noticeable in the HCM heart. These morphological variations have been reported for both pathologies [45], [46].

4 Conclusion

In this paper, we proposed a generative model to compute a pdf for point sets with no point-to-point correspondences. The pdf is formulated in a hierarchical fashion; in D-dimension, the points in each point set are assumed to be samples from a mixture of Gaussians. Similar to [30], we establish soft point-to-point correspondences across the Gaussian centroids, rather than the observed points. This enables us to effectively transform the point sets to consistent high dimensional vector representations, made by concatenating the spatial coordinates of the corresponding Gaussian centroids of each point set. The key aspect of the framework, however, is that these high-dimensional vectors are assumed to be samples from mixtures of principal component analyzers [37], extending it to handle point sets.

It is important to notice that information flows in both directions in the hierarchy. In fact, estimating the means and modes of variations of clusters in higher dimension constraints the Gaussian centroids in the lower dimension. We designed a variational Bayesian (VB) method to infer the approximate posteriors of unknown variables given data.

Using VB, we were able to compute a tight lower bound on the model evidence. We showed that by maximizing the lower bound, we could select models having correct numbers of model points M, modes of variations L, and clusters J, provided that a sufficient number of training data exists. To this end, we relied on mixtures of synthetic point sets sampled from a ground truth model (*e.g.* see Figure 8). We also observed that with inadequate data, simple models (with smaller M, L, or J values) have larger evidence (*e.g.* see Figure 6). This is because when computing model evidence in a full Bayesian setting, the model complexity is penalized due to marginalization over hidden variables. In the presence of inadequate samples, this penalization dominates the model evidence, undermining the fitness (likelihood) term, and favoring simpler models. We also investigated the model selection problem using real point sets representing mixtures of healthy and diseased hearts and lumbar vertebrae. Although, in this cases the true models were not available, the selected models having two clusters were reasonable due to either clinical interpretation or our perception of the morphologies of the structures.

We project a given point set to the space of the proposed mixture of PPCAs using the trained predictive distribution, by providing (37). Using this relationship, we measured the specificity and generalization of various models and established a link between these errors and the model evidence (Figure 9). It is interesting to see that models with maximum evidence are located competitively close to lower left corners in generalization-specificity planes for each of the considered distance types. This observation suggests that the models selected by our VB approach generally show small errors of both types. We also compared our framework to the arguably closest statistical shape modeling approach proposed by Rasoulian *et al.* [34]. The results in Table 1 indicate that the proposed model outperforms Rasoulian *et al.* and achieves better generalization and specificity errors. In addition, our method clusters the point sets into various groups indicating their pathological conditions or intrinsic morphological properties.

For an improved shape analysis several research directions can be considered: i) In the current implementation, the points were simply defined using their three dimensional coordinate values. Consequently, the point-to-point correspondences between shapes are only established based on the spatial co-ordinate values. A more accurate correspondence can be established if *shape context* proposed by Belongie *et al.* [48], or shortest path description proposed by Ling *et al.* [49] is utilized to make the points more distinctive. ii) Modeling shapes as GMMs with isotropic

covariances (parametrized by β^{-1}), the density of mesh vertices is assumed to be uniform across the shape, suggesting that the proposed model is best suitable for point sets with homogeneous point distribution. A locally variable form of the covariance matrices can be implemented in the cost of increased model complexity, demanding more data. iii) Designed to process unstructured point clouds, the current framework ignores the connectivity of vertices from surface meshes. Connectivities can be encoded using an enriched/higher dimensional point description based on heat kernel signatures [50], [51], which are invariant to isometric transformations. However, this evidently comes at the expense of increased demand for computational power.

In summary, the proposed model provides a significant versatility for statistical shape analysis, eliminating the need for specifying the point-to-point correspondences, number of model points, clusters, or modes of variation. Being fully Bayesian, the method favors statistically compact models [38], where only one or two variation modes are suggested per each subspace. This is in line with [28], where likewise our framework, the model complexity is penalized and statistically "thin" models with minimal number of modes are obtained. Furthermore, although the method infers point set labels in an unsupervised fashion, it can easily handle supervised and semi-supervised training scenarios simply by not updating the observed labels.

Although, we did not consider a natural shape modeling on manifolds, overall, the model is non-linear and captures the shape variations using a mixture of smaller linear (PPCA) subspaces. In this regard, the framework presents a good compromise between accuracy and pragmatic usability. Being a general model to analyze multi-dimensional point sets, we are currently investigating to further the applications of the proposed framework. The code has been efficiently implemented in Matlab, and is available for download (see http://www.cistib.org).

Acknowledgement This project was funded by the Marie Skodowska-Curie Individual Fellowship (Contract Agreement 625745), granted to A. Gooya.

REFERENCES

[1] F. Spoor, P. Gunz, S. Neubauer, S. Stelzer, N. Scott, A. Kwekason, and M. Dean, "Reconstructed homo habilis type OH 7 suggests deep-rooted species diversity in early homo," *Nature*, no. 519, pp. 83–86, 2015.

[2] P. Lestrel, C. Wolfe, and A. Bodt, "Mandibular shape analysis in fossil hominins: Fourier descriptors in norma lateralis," *Journal of Comparative Human Biology*, vol. 64, no. 4, pp. 247–272, 2013.

[3] P. Yan and K. W. Bowyer, "Biometric recognition using 3d ear shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1297–1308, 2007.

[4] N. Duta, "A survey of biometric technology based on hand shape," *Pattern Recognition*, vol. 42, no. 11, pp. 2797–2806, 2009.

[5] T. Heimann and H. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Medical Image Analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[6] L. Younes, "Spaces and manifolds of shapes in computer vision: An overview," *Image and Vision Computing*, vol. 30, no. 6, pp. 389–397, 2012.

[7] T. F. Cootes, C. J. Twinging, K. O. Babalola, and C. J. Taylor, "Diffeomorphic statistical shape models," *Image Vision Computing*, vol. 26, no. 3, pp. 326–332, 2008.

[8] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," *Neuroimage*, vol. 23, pp. S151–S160, 2004.

[9] L. Younes, "Computable elastic distances between shapes," *SIAM Journal on Applied Mathematics*, vol. 58, no. 2, pp. 565–586, 1998.

[10] E. Klassen, A. Srivastava, W. Mio, and S. H. Joshi, "Analysis of planar shapes using geodesic paths on shape spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 372–383, 2004.

[11] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu, "Statistical shape analysis: Clustering, learning, and testing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 590–602, 2005.

[12] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1415–1428, 2011.

[13] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 995–1005, 2004.

[14] L. H. Staib and J. S. Duncan, "Boundary finding with parametrically deformable models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 11, pp. 1061–1075, 1992.

[15] C. Brechbuhler, G. Gerig, and O. Kohler, "Parametrization of closed surfaces for 3D shape description," *Computer Vision Image Understandings*, vol. 61, no. 2, pp. 154–170, 1995.

[16] A. Kelemen, G. Szekely, and G. Gerig, "Elastic model-based segmentation of 3D neurological data sets," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 828–839, 1999.

[17] S. Kurtek, E. Klassen, Z. Ding, S. W. Jacobson, J. L. Jacobson, M. J. Avison, and A. Srivastava, "Parameterization-invariant shape comparisons of anatomical surfaces," *IEEE Trans. Med. Imag.*, vol. 30, no. 3, pp. 849–858, 2011.

[18] D. G. Kendall, "A survey of the statistical theory of shape," *Statistical Science*, pp. 87–99, 1989.

[19] H.-L. Le, "Explicit formulae for polygonally generated shape-densities in the basic tile," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 101, no. 02. Cambridge Univ Press, 1987, pp. 313–321.

[20] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1896–1909, 2005.

[21] T. F. Cootes and C. J. Taylor, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 10, pp. 38–59, 1995.

[22] C. Lindner, P. Bromiley, M. C. Ionita, and T. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1862–1874, 2015.

[23] D. Cremers, T. Kohlberger, and C. Schnörr, "Shape statistics in kernel space for variational image segmentation," *Pattern Recognition*, vol. 36, no. 9, pp. 1929–1943, 2003.

[24] M. S. Hefny, T. Okada, M. Hori, Y. Sato, and R. E. Ellis, "A liver atlas using the special euclidean group," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Springer, 2015, pp. 238–245.

[25] T. F. Cootes and C. J. Taylor, "A mixture model for representing shape variation," *Image and Vision Computing*, vol. 17, no. 8, pp. 567–573, 1999.

[26] R. H. Davies, C. J. Twinning, T. F. Cootes, C. J. Waterton, and C. J. Taylor, "A minimum description length approach to statistical shape modelling," *IEEE Trans. Med. Imag.*, vol. 21, no. 5, pp. 525–537, 2002.

[27] R. H. Davies, C. J. Twinning, T. F. Cootes, and C. J. Taylor, "Building 3-D statistical shape models by direct optimization," *IEEE Trans. Med. Imag.*, vol. 29, no. 4, pp. 961–982, 2010.

[28] J. Cates, P. T. Fletcher, M. Styner, M. Shenton, and R. Whitaker, "Shape modeling and analysis with entropy-based particle systems," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2007, pp. 333–345.

[29] M. Datar, J. Cates, P. T. Fletcher, S. Gouttard, G. Gerig, and R. Whitaker, "Particle based shape regression of open surfaces with applications to developmental neuroimaging," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2009, pp. 167–174.

[30] H. Chui, A. Rangarajan, J. Zhang, and C. M. Leonard, "Unsupervised learning of an atlas from unlabeled point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 160–172, 2004.

[31] H. Hufnagel, X. Pennec, J. Ehrhardt, N. Ayache, and H. Handels, "Generation of a statistical shape model with probabilistic point correspondences and EM-ICP," *International Journal for Computer Assisted Radiology and Surgery*, vol. 5, pp. 265–273, 2008.

[32] A. Gooya, C. Davatzikos, and A. F. Frangi, "A bayesian approach to sparse model selection in statistical shape models," *SIAM Journal on Imaging Sciences*, vol. 8, no. 2, pp. 858–887, 2015.

[33] A. Myronenko and S. Xubo, "Point set registration: Coherent point drift," *IEEE Tran. on Pat. Anal. and Mach. Intel.*, vol. 32, no. 12, pp. 2262–2275, 2010.

[34] A. Rasoulian, R. Rohling, and P. Abolmaesumi, "Group-wise registration of point sets for statistical shape models," *IEEE Trans. Med. Imag.*, vol. 31, no. 11, pp. 2025–2033, 2012.

[35] M. Vaillant and J. Glaunès, "Surface matching via currents," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2005, pp. 381–392.

[36] S. Durrleman, X. Pennec, A. Trouvé, and N. Ayache, "Statistical models of sets of curves and surfaces based on currents," *Medical image analysis*, vol. 13, no. 5, pp. 793–808, 2009.

[37] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Comput.*, vol. 11, no. 2, pp. 443–482, 1999.

[38] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[39] Z. Ghahramani and M. J. Beal, "Variational inference for bayesian mixtures of factor analysers," in *Adv in Neur Infor Proc Sys 12*. MIT Press, 2000, pp. 449–455.

[40] A. Gooya, K. Lekadir, X. Alba, A. J. Swift, J. Wild, and A. F. Frangi, "Joint clustering and component analysis of correspondenceless point sets: Application to cardiac statistical modeling," in *Informaiton Processing in Medical Imaging*, 2015, vol. 9123, pp. 858–887.

[41] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 436–442.

[42] A. M. Bronstein, M. M. Bronstein, L. J. Guibas, and M. Ovsjanikon, "Shape Google: Geometric words and expressions for invariant shape retrieval," *ACM Transactions on Graphics*, vol. 30, no. 1, pp. 1295–1312, 2008.

[43] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[44] A. J. Swift and et. al., "Diagnostic accuracy of cardiovascular magnetic resonance imaging of right ventricle morphology and function in assessment of suspected pulmonary hypertension results from the ASPIRE registry," *J Card Mag Res*, vol. 14, no. 40, 2012.

[45] N. Voelkel, R. Quaife, L. Leinwand, R. Barst, and et. al., "Right ventricular function and failure: a report," *Circ.*, vol. 114, pp. 1883–91, 2006.

[46] B. Maron and et. al., "Hypertrophic cardiomyopathy. Interrelations of clinical manifestations, pathophysiology, and therapy," *N Engl J Med.*, vol. 316, pp. 780–844, 1987.

[47] 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, July 2001, pp. 194–202.

[48] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[49] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, Feb. 2007.

[50] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," in *Computer graphics forum*, vol. 28, no. 5. Wiley Online Library, 2009, pp. 1383–1392.

[51] M. M. Bronstein and I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1704–1711.

Ali Gooya obtained a MSc in Electrical Engineering from Tehran University and a PhD in Information Science from the University of Tokyo (Monbusho scholarship) where he was awarded a post-doctoral fellowship by Japan Society of Promotion of Science (2008). He then moved to the University of Pennsylvania, Philadelphia, USA, and worked on tumour image segmentation/registration till 2011. Subsequently, he served as an Assistant Professor in Tarbiat Modares University, Tehran. In 2014, he was awarded an IIF Marie-Curie Fellowship for statistical modeling of morphology and function from population in University of Sheffield, where he recently works as a Lecturer in Medical image Computing in the department of EEE. His research interest includes probabilistic machine learning, variational Bayesian inference, and graphical models.

Karim Lekadir studied mathematics and engineering in France, before receiving his PhD in computer science in 2009 from Imperial College London. He is a Ramon y Cajal senior researcher at the Department of ICT at Universitat Pompeu Fabra, Barcelona. His research interests include medical image analysis, computational modeling, and biomedical data analytics.

Isaac Castro-Mateos obtained his PhD in biomedical image computing from the University of Sheffield in 2016, working in the context of the European Project MySpine. He is currently a research associate in the perinatal department at King's College of London. His research interests include medical imaging processing, statistical modeling and pattern recognition.

Jose Maria Pozo is a Research Fellow at the University of Sheffield, where he leads the musculoskeletal team in the CISTIB and contributes to several National and European projects, such as MySpine and MD-Paedigree. Jose's research interests include the development of geometrical and image computing methods for the extraction of quantitative information and patient-specific models from 3D or 2D image modalities (CT, MRI, DXA).

Alejandro F Frangi graduated with an BEng/MEng in Telecommunications Engineering from the Technical University of Catalonia (Barcelona) in 1996. In 1997 he obtained a grant from the Dutch Ministry of Economic Affairs to pursue his PhD at the Image Sciences Institute (www.isi.uu.nl) of the University Medical Center Utrecht on model-based cardiovascular image analysis. He is currently Professor of Biomedical Image Computing (h-index $>$ 47) at the University of Sheffield (USFD), leading the Center for Computational Imaging & Simulation Technologies in Biomedicine. Prof Frangi has been principal investigator or scientific coordinator of over 25 national and European projects. He is Associate Editor of IEEE TMI, MedIA, SIAM Journal Imaging Sciences. His main research interests are in medical image computing, medical imaging and image-based computational physiology.