Joint Clustering and Component Analysis of Correspondenceless Point Sets: Application to Cardiac Statistical Modeling

Ali Gooya[1⋆], Karim Lekadir[2], Xenia Alba[2], Andrew J Swift[3], Jim Wild[3], and Alejandro F Frangi[1]

[1] Centre for Computational Imaging & Simulation Technologies in Biomedicine
The University of Sheffield
a.gooya@sheffield.ac.uk
http//www.cistib.org/
[2] Universitat Pompeu Fabra, Barcelona, Spain
[3] Academic Unit of Radiology, Royal Hallamshire Hospital

Abstract. Construction of Statistical Shape Models (SSMs) from arbitrary point sets is a challenging problem due to significant shape variation and lack of explicit point correspondence across the training data set. In medical imaging, point sets can generally represent different shape classes that span healthy and pathological exemplars. In such cases, the constructed SSM may not generalize well, largely because the probability density function (pdf) of the point sets deviates from the underlying assumption of Gaussian statistics. To this end, we propose a generative model for unsupervised learning of the pdf of point sets as a mixture of distinctive classes. A Variational Bayesian (VB) method is proposed for making joint inferences on the labels of point sets, and the principal modes of variations in each cluster. The method provides a flexible framework to handle point sets with no explicit point-to-point correspondences. We also show that by maximizing the marginalized likelihood of the model, the optimal number of clusters of point sets can be determined. We illustrate this work in the context of understanding the anatomical phenotype of the left and right ventricles in heart. To this end, we use a database containing hearts of healthy subjects, patients with Pulmonary Hypertension (PH), and patients with Hypertrophic Cardiomyopathy (HCM). We demonstrate that our method can outperform traditional PCA in both generalization and specificity measures.

Keywords: Statistical Shape Models, Variational Bayes, Model Selection

1 Introduction

Statistical shape models (SSMs) from point sets, proposed by Cootes *et al.* in [2], are powerful tools in medical imaging to encode the natural variability of anatomical structures. To construct an SSM, traditionally, points are selected on training

⋆ The asterisk shows the corresponding author.

surfaces and point-to-point correspondences are required. By consistently concatenating the points on each training data set, shapes are represented as high-dimensional vectors and assumed to be sampled from a Gaussian distribution; under this hypothesis the major modes of variation are then extracted by PCA. In reality, however, the training data can have a multi-modal distribution and represent various classes of shapes. As a result, no particular class is fully represented by the mean model and the constructed SSM often does not generalize well. To alleviate this problem, Zhang *et al.* [15] proposed sparse non-parametric shape description, and Cootes *et al.* [3] used a Gaussian Mixture Model (GMM) to represent the pdf of the training sets; the shape space is first partitioned and then PCA is applied in each cluster. But, it is likely that clustering of point sets and the estimation of variation modes may mutually benefit from each other. In addition, this approach requires having point-to-point correspondences, and a user-selected a priori number of components, which is difficult.

Establishing point-to-point correspondences across training point sets is a major challenge that undermines the practicality of the SSMs. Manually specifying correspondences over landmarks could be an ambiguous subjective task. 3D automatic techniques based on image registration [5] or minimizing the description length [4] have a varying performance, in particular for complex structures such as the heart. EM-ICP based methods [7, 11] offer more flexibility by computing probabilistic matchings between points and are shown to be robust to the matching errors. Recently, Hufnagel *et al.* [8] proposed a generative model for estimating modes of variation in point sets without resourcing to PCA from point sets with no correspondences. This method, however, still assumes that the distribution is a monomodal Gaussian distribution.

We present a hierarchical clustering scheme to estimate pdfs of unstructured, rigidly aligned, point sets having no point-to-point correspondences. Points at each set are regarded as samples from a low dimensional GMM, whose means are concatenated to form higher dimensional vector. This vector is considered to be a sample drawn from a Mixture of Probabilistic Principal Component Analyzers (PPCA) [13]. The latter is essentially a higher dimensional GMM, where the covariance matrices of its clusters can be decomposed to subspaces of local principal components. An inference algorithm based on variational Bayes (VB) [1] is proposed for unsupervised learning of class labels and variations.

Thanks to this hierarchical structure, the proposed method estimates probabilistic point matchings across the training data sets; and handles mixtures of different shape classes. Another important advantage of the proposed VB approach is that the number of clusters is automatically learned from data. It is noteworthy that, in machine learning, VB has been successfully applied for inferring mixtures of subspace analyzers [6] from training vectors having equal lengths. However, adopting the framework for point sets, as order-less random variables having different cardinalities (point counts), is a challenging problem. In the rest of this paper, we first present our generative model, derive an efficient inference algorithm and finally compare the method to the standard PCA model using cardiac data with different pathologies.

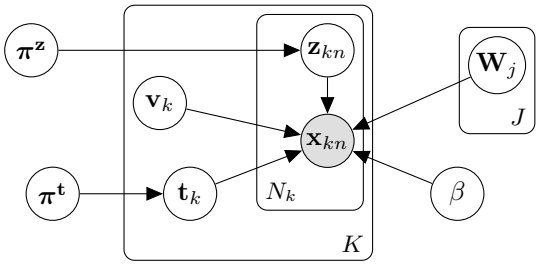


Figure 1. The graphical representation of the proposed model; shaded and hollow circles represent observed and latent variables, respectively, arrows imply the dependencies and plates embrace numbered incidences of events.

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 $\mathbf{W}_j \mathbf{W}_j^T$ [13]. 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.

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, whose mth component equals one ($z_{knm} = 1$) indicates 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 $\beta \mathbf{I}_{D \times D}$. 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 of $\mathbf{x}_{kn}$ is then given by:

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

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):

$$p(\mathbf{W}_j)=\prod_{l=1}^{L}\mathcal{N}(\mathbf{W}_j^{(l)}|\mathbf{0},\alpha_{jl}^{-1}\boldsymbol{I}),\quad p(\mathbf{v}_k)=\mathcal{N}(\mathbf{v}_k|\mathbf{0},\boldsymbol{I}),\quad p(\beta)=\Gamma(\beta|a_0,b_0) \quad (2)$$

The hyper-parameters of the Gamma distribution in the last line are set to $a_0=10^{-3}$ and $b_0=1$ to have a flat prior over β. 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: $p(\boldsymbol{\pi}^{\mathbf{z}})=\mathrm{Dir}(\boldsymbol{\pi}^{\mathbf{z}}|\lambda_0^{\mathbf{z}})$, $p(\boldsymbol{\pi}^{\mathbf{t}})=\mathrm{Dir}(\boldsymbol{\pi}^{\mathbf{t}}|\lambda_0^{\mathbf{t}})$. where we set $\lambda_0^{\mathbf{z}}=\lambda_0^{\mathbf{t}}=10^{-3}$. 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 multi-nomial distributions: $p(\mathbf{z}_{kn}|\boldsymbol{\pi}^{\mathbf{z}})=\prod_{m=1}^{M}(\pi_m^{\mathbf{z}})^{z_{knm}}$, and $p(\mathbf{t}_k|\boldsymbol{\pi}^{\mathbf{t}})=\prod_{j=1}^{J}(\pi_j^{\mathbf{t}})^{t_{kj}}$, where $0\le\pi_m^{\mathbf{z}}$, $0\le\pi_j^{\mathbf{t}}$ are the components m, j of $\boldsymbol{\pi}^{\mathbf{z}}$, $\boldsymbol{\pi}^{\mathbf{t}}$, respectively. We now construct the joint pdf of the sets of all random variables, by assuming (conditional) independence and multiplying the pdfs where needed. Let $\mathbb{X}=\{\mathcal{X}_k\}_{k=1}^{K}$, $\mathbb{Z}=\{\mathcal{Z}_k\}_{k=1}^{K}$, $\mathbb{V}=\{\mathbf{v}_k\}_{k=1}^{K}$, and $\mathbb{T}=\{\mathbf{t}_k\}_{k=1}^{K}$, then the distributions of these variables can be written as:

$$p(\mathbb{W})=\prod_j p(\mathbf{W}_j|\boldsymbol{\alpha}_j),\quad p(\mathbb{Z}|\boldsymbol{\pi}^{\mathbf{z}})=\prod_k p(\mathcal{Z}_k|\boldsymbol{\pi}^{\mathbf{z}})=\prod_k\prod_n p(\mathbf{z}_{kn}),\ p(\mathbb{T}|\boldsymbol{\pi}^{\mathbf{t}})=\prod_k p(\mathbf{t}_k|\boldsymbol{\pi}^{\mathbf{t}})$$

$$p(\mathbb{V})=\prod_k p(\mathbf{V}_k),\quad p(\mathbb{X}|\mathbb{Z},\mathbb{T},\mathbb{W},\mathbb{V},\beta)=\prod_k p(\mathcal{X}_k|\mathcal{Z}_k,\mathbf{t}_k,\beta,\mathbb{W},\mathbf{v}_k), \quad (3)$$

$$p(\mathcal{X}_k|\mathcal{Z}_k,\mathbf{t}_k,\beta,\mathbb{W},\mathbf{v}_k)=\prod_{n=1}^{N_k}p(\mathbf{x}_{kn}|\mathbf{z}_{kn},\mathbf{t}_k,\beta,\mathbb{W},\mathbf{v}_k) \quad (4)$$

Having defined the required distributions through Eqn. (1)-3, the distribution of the *complete* observation is given as

$$p(\mathbb{X},\mathbb{Z},\mathbb{T},\mathbb{W},\mathbb{V},\boldsymbol{\pi}^{\mathbf{t}},\boldsymbol{\pi}^{\mathbf{z}},\beta)=p(\mathbb{X}|\mathbb{Z},\mathbb{T},\mathbb{W},\mathbb{V},\beta)p(\mathbb{Z}|\boldsymbol{\pi}^{\mathbf{z}})p(\mathbb{T}|\boldsymbol{\pi}^{\mathbf{t}})p(\boldsymbol{\pi}^{\mathbf{z}})p(\boldsymbol{\pi}^{\mathbf{t}})p(\mathbb{W})p(\mathbb{V})p(\beta) \quad (5)$$

Figure 1 is a graphical representation for the generative model considered in this paper. Given observations (colored dark gray) as D dimensional points, our problem is to estimate the posterior distributions of all the latent random variables (hollow circles) and hyper-parameters, which include the discrete cluster and the continuous variables (*e.g.* means and modes of variations).

2.2 Approximated Inference

If we denote the set of latent variables as $\boldsymbol{\theta}=\{\mathbb{Z},\mathbb{T},\mathbb{W},\mathbb{V},\boldsymbol{\pi}^{\mathbf{t}},\boldsymbol{\pi}^{\mathbf{z}},\beta\}$, direct inference of $p(\boldsymbol{\theta}|\mathbb{X})$ (as our objective) 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 model evidence, *i.e.* $p(\mathbb{X})$ [4], can be decomposed as $p(\mathbb{X}) = \mathcal{L} + \mathrm{KL}(p(\boldsymbol{\theta}|\mathbb{X})||q(\boldsymbol{\theta}))$, where $0 \leq \mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leilber divergence, and

$$\mathcal{L} = \int q(\boldsymbol{\theta}) \ln \frac{p(\mathbb{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \leq p(\mathbb{X}) \tag{6}$$

is a lower bound on $p(\mathbb{X})$. 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. Thus, $q(\boldsymbol{\theta})$ can be computed by maximizing $\mathcal{L}$. We approximate the true posterior as a factorized form, *i.e.* , $q(\boldsymbol{\theta}) = \prod_i q(\theta_i)$, where θ_i refers to any of our latent variables. This factorization leads to the following tractable result: let ε be the variable of interest in $\boldsymbol{\theta}$, and $\xi = \boldsymbol{\theta} - \varepsilon$, then the variational posterior of ε is given by $\ln q(\varepsilon) = \langle \ln p(\mathbb{X}, \boldsymbol{\theta}) \rangle_\xi + \text{const}$, where $p(\mathbb{X}, \boldsymbol{\theta})$ is given in Eqn. (5), $\langle \cdot \rangle_\xi$ denotes the expectation w.r.t. to the product of $q(\cdot)$ of all variable in ξ.

2.3 Update of Posteriors and Hyper-parameters

In this section, we provide equations to update the variational posteriors. Thanks to conjugacy of priors to likelihoods, these derivations are done by inspecting expectations of logarithms and matching posteriors to their corresponding likelihood template forms. Detailed proof of our derivations is skipped for brevity. Starting from $\mathbb{Z}$ variables we have $q(\mathbb{Z}) = \prod_k q(\mathcal{Z}_k) = \prod_{k,m,n} (r_{knm})^{z_{knm}}$ Under this equation, we have $\langle z_{knm} \rangle = r_{knm}$, where the right hand side can be computed using the following relationships:

$$r_{knm} = \frac{\rho_{knm}}{\sum_{m'} \rho_{knm'}}, \ln \rho_{knm} = -\frac{\langle \beta \rangle}{2} \sum_j \langle t_{kj} \rangle \langle |\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2 \rangle + \langle \ln \pi_m^{\mathbf{z}} \rangle \tag{7}$$

The first term can be directly computed using the expectations of $\mathbb{W}$ and $\mathbb{V}$ as follows: $\langle |\mathbf{x}_{kn} - \boldsymbol{\mu}_{jk}^{(m)}|^2 \rangle = |\mathbf{x}_{kn} - \langle \boldsymbol{\mu}_{jk}^{(m)} \rangle|^2 + \mathrm{Tr}[\mathrm{Cov}[\boldsymbol{\mu}_{jk}]^{(m,m)}]$, where the super indexes $(\cdot)$, $(\cdot, \cdot)$ specify the D and $D \times D$ dimensional block numbers of the vector $\langle \boldsymbol{\mu}_{jk} \rangle = \langle \mathbf{W}_j \rangle \langle \mathbf{v}_k \rangle + \bar{\boldsymbol{\mu}}_j$ and the matrix defined by:

$$\mathrm{Cov}[\boldsymbol{\mu}_{jk}] = \langle \mathbf{W}_j \rangle \langle \mathbf{v}_k \mathbf{v}_k^T \rangle \langle \mathbf{W}_j \rangle^T + \sum_l \langle \mathbf{v}_k \mathbf{v}_k^T \rangle_{ll} \mathrm{Cov}[\mathbf{W}_j^{(l)}] \tag{8}$$

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

$$\mathbf{R}_k = \mathrm{Diag}(\underbrace{R_{k1} \cdots R_{k1}}_{D \text{ copies}}, \cdots, \underbrace{R_{kM} \cdots R_{kM}}_{D \text{ copies}}), \bar{\mathbf{x}}_k = [\bar{\mathbf{x}}_{k1}^T, \cdots, \bar{\mathbf{x}}_{kM}^T]^T \tag{9}$$

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

where: $R_{km}=\sum_n r_{knm}$, and $\bar{\mathbf{x}}_{km}=\sum_n r_{knm}\mathbf{x}_{kn}$. Under these definitions, the posteriors of $\mathbb{T}$ is given by $q(\mathbb{T})=\prod_k q(\mathbf{t}_k)=\prod_{k,j}(r'_{kj})^{t_{kj}}$ where, in analogy to Eqn. (7), we have: $\langle t_{kj}\rangle = r'_{kj}$ and

$$r'_{kj}=\frac{\rho'_{kj}}{\Sigma_{j'}\rho'_{kj'}},\quad \ln\rho'_{kj}=\langle\beta\rangle\mathrm{Tr}[-\frac{1}{2}\mathbf{R}_k\langle\boldsymbol{\mu}_{jk}\boldsymbol{\mu}_{jk}^T\rangle+\boldsymbol{\mu}_{jk}\bar{\mathbf{x}}_k^T]+\langle\ln\pi_j^{\mathbf{t}}\rangle \tag{10}$$

The posterior of the principal components is given by

$$q(\mathbb{W})=\prod_{j,l}q(\mathbf{W}_j^{(l)}),\quad q(\mathbf{W}_j^{(l)})=\mathcal{N}(\mathbf{W}_j^{(l)}|\langle\mathbf{W}_j^{(l)}\rangle,\mathrm{Cov}[\mathbf{W}_j^{(l)}]) \tag{11}$$

where the means and covariance matrices are specified as:

$$\begin{aligned}\mathrm{Cov}[\mathbf{W}_j^{(l)}]&=[\alpha_{jl}\boldsymbol{I}+\langle\beta\rangle\sum_k\langle t_{kj}\rangle\langle\mathbf{v}_k\mathbf{v}_k^T\rangle_{ll}\mathbf{R}_k]^{-1}\\ \langle\mathbf{W}_j^{(l)}\rangle&=\langle\beta\rangle\mathrm{Cov}[\mathbf{W}_j^{(l)}](\sum_k\langle t_{kj}\rangle\mathbf{Q}_{\mathrm{kj}}{}^{(l)})\end{aligned} \tag{12}$$

Here, the auxiliary matrix $\mathbf{Q}_{\mathrm{kj}}$ is defined as

$$\mathbf{Q}_{\mathrm{kj}}=(\bar{\mathbf{x}}_k-\mathbf{R}_k\bar{\boldsymbol{\mu}}_j)\langle\mathbf{v}_k\rangle^T-\mathbf{R}_k\langle\mathbf{W}_j\rangle\Big[\langle\mathbf{v}_k\mathbf{v}_k^T\rangle-\mathrm{Diag}(\mathrm{diag}\langle\mathbf{v}_k\mathbf{v}_k^T\rangle)\Big] \tag{13}$$

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. The posterior of $\mathbf{v}_k$ vectors is given by

$$\begin{aligned}&q(\mathbb{V})=\prod_k q(\mathbf{v}_k)=\mathcal{N}(\mathbf{v}_k|\langle\mathbf{v}_k\rangle,\mathrm{Cov}[\mathbf{v}_k]),\ \mathrm{Cov}[\mathbf{v}_k]=\Big[\mathrm{I}+\langle\beta\rangle\sum_j\langle t_{kj}\rangle\langle\mathbf{W}_j^T\mathbf{R}_k\mathbf{W}_j\rangle\Big]^{-1}\\ &\langle\mathbf{W}_j^T\mathbf{R}_k\mathbf{W}_j\rangle=\langle\mathbf{W}_j\rangle^T\mathbf{R}_k\langle\mathbf{W}_j\rangle+\mathrm{Diag}\Big(\mathrm{Tr}[\mathbf{R}_k\mathrm{Cov}[\mathbf{W}_j^{(1)}]],\cdots,\mathrm{Tr}[\mathbf{R}_k\mathrm{Cov}[\mathbf{W}_j^{(L)}]]\Big)\\ &\langle\mathbf{v}_k\rangle=\langle\beta\rangle\mathrm{Cov}[\mathbf{v}_k]\sum_j\langle t_{kj}\rangle\langle\mathbf{W}_j\rangle^T(\bar{\mathbf{x}}_k-\mathbf{R}_k\bar{\boldsymbol{\mu}}_j)\end{aligned} \tag{14}$$

The posterior of the precision β is a Gamma distribution specified by:

$$q(\beta)=\Gamma(\beta|a,b),\ a=a_0+\frac{DN}{2},\ b=b_0+\frac{1}{2}\sum_{k,n,m,j}\langle z_{knm}\rangle\langle t_{kj}\rangle\langle|\mathbf{x}_{kn}-\boldsymbol{\mu}_{jk}^{(m)}|^2\rangle \tag{15}$$

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. Finally, the posteriors of the mixing coefficients are Dirichlet distributions:

$$q(\boldsymbol{\pi}^{\mathbf{t}})=\mathrm{Dir}(\boldsymbol{\pi}^{\mathbf{t}}|\boldsymbol{\lambda}^{\mathbf{t}}),\lambda_j^{\mathbf{t}}=\lambda_0^{\mathbf{t}}+\sum_k\langle t_{kj}\rangle,\ q(\boldsymbol{\pi}^{\mathbf{z}})=\mathrm{Dir}(\boldsymbol{\pi}^{\mathbf{z}}|\boldsymbol{\lambda}^{\mathbf{z}}),\lambda_m^{\mathbf{z}}=\lambda_0^{\mathbf{z}}+\sum_{k,n}\langle z_{knm}\rangle \tag{16}$$

Using Eqn. (16), 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 $\langle\ln\lambda_j^{\mathbf{t}}\rangle=\psi(\lambda_j^{\mathbf{t}})-\psi(\sum_{j'}\lambda_{j'}^{\mathbf{t}})$. Finally, by maximizing Eqn. (6) with regard to $\bar{\boldsymbol{\mu}}_j$ and α_{jl}, we obtain:

$$\bar{\boldsymbol{\mu}}_j=\Big[\sum_k\langle t_{kj}\rangle\mathbf{R}_k\Big]^{-1}\Big[\sum_k\langle t_{kj}\rangle(\bar{\mathbf{x}}_k-\mathbf{R}_k\langle\mathbf{W}_j\rangle\langle\mathbf{v}_k\rangle)\Big], \tag{17}$$

$$\alpha_{jl}=MD/\Big[|\langle\mathbf{W}_j^{(l)}\rangle|^2+\mathrm{Tr}[\mathrm{Cov}[\mathbf{W}_j^{(l)}]]\Big] \tag{18}$$

2.4 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 $\langle\hat{\mathbf{x}}_{rn}\rangle = \int \hat{\mathbf{x}}_{rn} p(\hat{\mathbf{x}}_{rn}|\mathcal{X}_r, \mathbb{X}) d\hat{\mathbf{x}}_{rn}$. Here, the predictive distribution should be computed by marginalizing the corresponding latent and model variables by

$$p(\hat{\mathbf{x}}_{rn}|\mathcal{X}_r, \mathbb{X}) = \sum_{\mathbf{z}_{rn}, \mathbf{t}_r} \int p(\hat{\mathbf{x}}_{rn}|\mathbf{z}_{rn}, \mathbf{t}_r, \beta, \mathbb{W}, \mathbf{v}_r) p(\mathbf{z}_{rn}, \mathbf{t}_r, \beta, \mathbb{W}, \mathbf{v}_r|\mathcal{X}_r, \mathbb{X}) d\mathbb{W} d\mathbf{v}_r d\beta$$

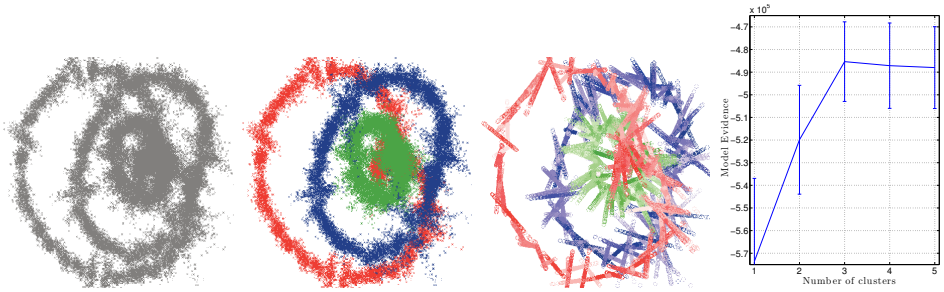
Because this integral is analytically intractable, we use an approximation for the posterior using $p(\mathbf{z}_{rn}, \mathbf{t}_r, \beta, \mathbb{W}, \mathbf{v}_r|\mathcal{X}_r, \mathbb{X}) \approx q(\mathbf{z}_{rn}) q(\mathbf{t}_r) q(\mathbf{v}_r) q(\beta) q(\mathbb{W})$. Thus, having $\mathcal{X}_r$ 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.

2.5 Initialization and Computational Burden

To initialize the model, a GMM with M Gaussians is fit to the set of all points. Next, 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 point set k are identified and concatenated to form an MD dimensional vector. This procedure is then repeated over K training point sets and the obtained vectors are clustered using k-means. 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. In practice, we have observed that for a set of fifty point sets each having 4000 points, sufficient convergence is achieved by 50 VB iterations in nearly an hour.

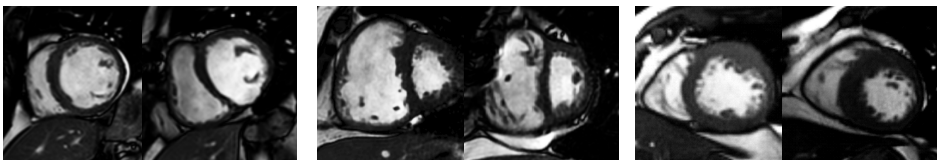
3 Results

We evaluate our method on both synthetic and real data sets of cardiac MRI as follows. The reliability of the lower bound as a criterion to select the number of clusters of point sets is evaluated in both data types. We also measure generalization and specificity errors, and compare them to the standard PCA based SSM on the real data sets. Generalization ability is 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 randomly divide the available point sets into the testing and training subsets and trained the model using latter. Next, we measure the generalization and specificity using the testing, and model generated point sets as explained in [4]. To measure the distances between point sets, we considered: $d(\mathcal{X}_k, \hat{\mathcal{X}}_k) = 1/N_k \sum_{\mathbf{x}} \min_{\mathbf{y} \in \hat{\mathcal{X}}_k} ||\mathbf{x} - \mathbf{y}||_2$. Here, N_k is the number of points in $\mathcal{X}_k$, and $\hat{\mathcal{X}}_k$ denote the model projected point set. Furthermore, since d is asymmetric, we also compute $\hat{d}(\mathcal{X}_k, \hat{\mathcal{X}}_k) = d(\hat{\mathcal{X}}_k, \mathcal{X}_k)$ and report both.



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

Figure 2. Clustering and mode estimation of synthetic point sets. (a) overlay of 750 point sets, (b) corresponding color separated ground true clusters, (c) estimated labels (colors), GMM centroids showing local modes of variations, (d) lower bound $\mathcal{L}$ on the model evidence versus number of clusters, indicating $J = 3$ as the optimal number.



(a) (b) (c)

Figure 3. Short axis MR images from normal (a), PH (b), and HCM patients (c).

3.1 Synthetic Dataset

We investigate the problem of selecting a proper number of clusters, J, from the data by generating synthetic point sets using ancestral sampling [1]. We expect the model evidence, *i.e.* $p(\mathbb{X}|J)$, to reach a maximum for the proper number of clusters used to generate data, due to the marginalization of the latent variables. Three distinctive 2D point set patterns are generated as the cluster means. To help visualization, by setting $L = 1$, a single mode of variation per cluster is considered and sampled from $p(\mathbf{W}_j)$. Next, a set of 250 point sets for each cluster is generated by sampling from $p(\mathbf{v}_k)$ and by applying the corresponding variations at each point (local Gaussians in Figure 2), by adding the measurement noise (by the precision of $\beta = 1$). The number of points in each point set is around 100, but this is variable over the population and thus no correspondence is assumed. For $1 \leq J \leq 5$, we repeated 15 rounds of experiments with variable patterns (one shown in Figure 2), and recorded the mean and one standard deviation on $\mathcal{L}$ (Figure 2(d)). As shown, for $J = 3$ a maximum for the model evidence is correctly found, and the color match between (b)-(c) indicates correct clustering of point sets. In addition, the linear patterns of GMM means, *i.e.* $\boldsymbol{\mu}_{jk}^{(m)}$, in (c) match the local structures of the points in (d), showing that variation modes are estimated reasonably.

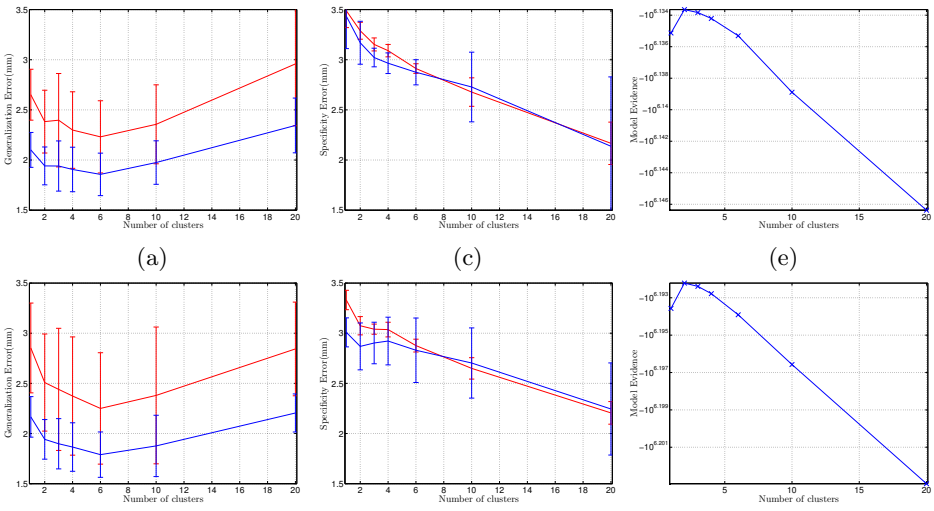


Figure 4. Quantitative results for models trained by clustering normal-HCM (top row) and normals-PH (bottom row) cases versus number of clusters, (a-b) Generalization errors as distances between the model projected and original test sets (red), and vice versa (blue), (c-d) Specificity errors as distances between model generated and training point sets (red), and vice versa (blue). (e-f) show the model evidecene.

Table 1. Generalization and Specificity errors (in mm) for PCA and the proposed method at $J = 2, 6$. Significant differences between results of PCA and our method are indicated in bold (p-value < 0.001). Double lines separate $d|\hat{d}$ distances computed from the models trained by clustering PH or HCM patients versus normals.

	Generalization				Specificity			
	PH		HCM		PH		HCM	
PCA	2.9±0.4	2.9±0.4	3.1±0.4	3.2±0.5	3.2±0.3	3.2±0.3	3.5±0.4	3.5 ± 0.6
$J = 2$	**2.5±0.5**	**1.9±0.2**	**2.4±0.3**	**1.9±0.2**	3.0±0.1	**2.8±0.2**	3.3±0.1	3.2±0.2
$J = 6$	**2.2±0.5**	**1.7±0.2**	**2.2±0.3**	**1.8±0.2**	**2.8±0.1**	**2.8±0.3**	**2.9±0.1**	**2.8±0.1**

3.2 Cardiac Datasets

We apply the proposed approach to the analysis of cardiac data sets, which are known to display significant variability and geometrical complexity. We consider three groups of individuals, 36 normal cases, 20 subjects with Pulmonary Hypertension (PH), and 20 subjects with Hypertrophic Cardiomyopathy (HCM). The data acquisition of these data sets were done using a balanced Steady State Free Precession protocol under various brands of 1.5 MR scanners, resulting in image matrices of 256×256 in short axial direction and slice thicknesses of $8 - 10$mm.

These subjects differ in the properties of the cardiac shapes. For PH patients, which are associated with pulmonary vascular proliferation [12], complex shape

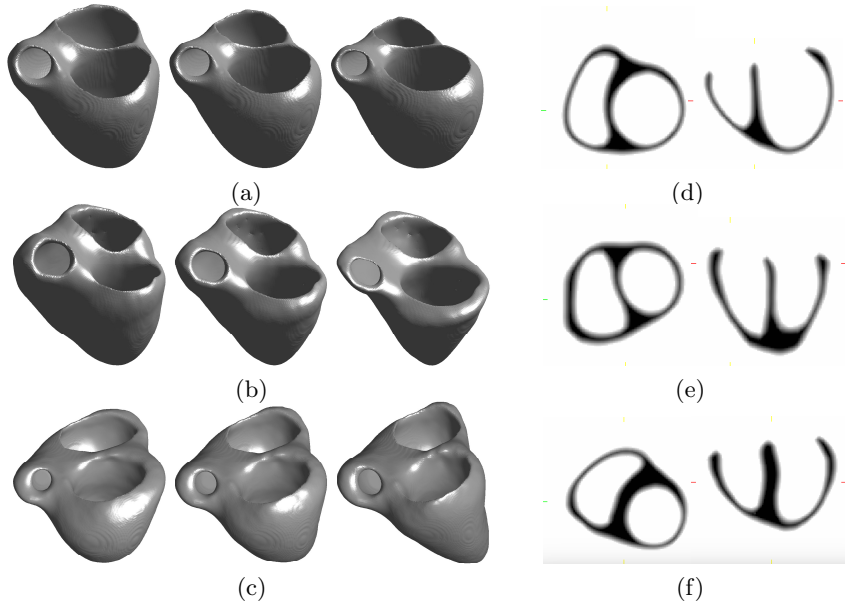


Figure 5. Means and variation modes for normal (a), PH (b), and HCM (c) cases, with mean models in the middle and variations in opposite directions at two sides, (d-f) axial and coronal cross sections of the mean models for each population.

remodeling of both the left and right ventricles occurs (see Fig. 3). As a result, the RV becomes very dilated, pushing onto the LV, which deforms and loses its roundness [14]. On the other hand, HCM [9] 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 prominently involving the ventricular septum) without abnormal enlargement of the ventricular cavities.

To derive cardiac surfaces, the initial shape was obtained by labeling the MRI slices, thus obtaining binary masks, then a volume mesh was generated form the binary images, finally, we extracted a surface mesh and considered vertices of the mesh as a point set. Next, we registered the point sets removing scaling, rotation the translation effects. Because we want to compare our method to standard PCA, point correspondences between training sets was needed and established using the projection method proposed in [10]. However, to implement our proposed method we ignore this information.

We cluster mixtures of normal-PH and normal-HCM cases in two sets of experiments. We randomly pick 33 (20 normals and 13 pathological) cases, ignore the labels and perform clustering. We then sample random point sets from the trained generative model and compute specificity. To quantify generalization on the test point sets, we use Eqn. (19) to compute model projected point sets for the test set. The number of clusters, *i.e.* J, is varied from 1 to maximum 20.

At each number, we compare both measures to the PCA model, taking 13-15 number of modes to cover 95% of the full trace of the covariance matrix.

Figure 4 shows the quantitative results obtained as described above. It can be seen that the generalization distances, *i.e.* , d (blue) and $\hat{d}$ (red), are minimum at $J = 6$. Compared to PCA, both generalization and specificity improve (indicated in Table 1). Also, note that as the number of clusters increases and the variations within each cluster are eliminated, specificity error improves. To understand this behavior, consider an extreme case, where every training point set becomes the cluster of it own. In that case, all the intra-cluster variations are eliminated and the trained model becomes strictly specific to the training data. It is noteworthy that the model evidence is maximized at $J = 2$ (see Figure 4(c)), which is the expected number of clusters at each experiment. The discrepancy between this number of clusters and $J = 6$ where the generalization error is minimum could be due to the approximations that are made in computing the predictive distribution. Nevertheless, as shown in Table 1, at both $J = 2, 6$ results are significantly improved over PCA model.

Next, we evaluate the clustering efficiency by comparing the estimated labels of the point sets to ground truths. For these experiments, by setting $L = 2$, $J = 2$, we independently apply the method to cluster all the available normal-PH and normal-HCM cases. We observe that at both experiments only 2 out of 53 cases are clustered incorrectly as normals cases. The first mode of variation both in 3D and in longitudinal cross sections are visualized in Figure 5. It can be seen that, in the normal heart (see (a) and (d)), LV is significantly larger than RV, and when compared to PH and HCM, it is more spherical. On the other hand, in the PH heart ((b) and (e)), the RV is evidently dilated and the LV loses it roundness. Finally, significant thickening of the septum and shrinkage of LV are noticeable in the HCM heart ((c) and (f)). These morphological variations in the normal heart have been reported for both pathologies [9, 14].

4 Conclusion

We proposed a unified framework for joint clustering and component analysis of point sets. We modeled the pdf of point sets as a mixture of principal component analyzers, where the labels of the point sets and variations of clusters are derived through a Variational Bayesian framework. The method is flexible to handle point sets with no point-to-point correspondences. We showed that the method can identify the number of clusters automatically, and outperform traditional PCA based SSMs in generalization and specificity. Application of the proposed framework to heart data sets shows successful clustering of normal and pathological cases, as well as extraction of their intra-cluster variations.

Acknowledgement This project was funded by the Marie Skodowska-Curie Individual Fellowship (Contract Agreement 625745).

References

1. Bishop, C.M.: Pattern recognition and machine learning. Springer (2009)
2. Cootes, T.F., Taylor, C.J.: Active shape models-their training and application. Computer Vision and Image Understanding 61(10), 38–59 (1995)
3. Cootes, T.F., Taylor, C.J.: A mixture model for representing shape variation. Imag. and Visi. Comp. 17(8), 567–574 (1999)
4. Davis, R.H., Twinning, C.J., Cootes, T.F., Taylor, C.J.: Building 3D statistical shape models by direct optimization. IEEE Tran on Med Imag 29(4), 961–82 (2010)
5. Frangi, A.F., Rueckert, D., Schnabel, J.A., Niessen, W.J.: Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling. IEEE Transactions on Medical Imaging 21(9), 1151–1165 (2002)
6. Ghahramani, Z., Beal, M.J.: Variational inference for bayesian mixtures of factor analysers. In: In Adv in Neur Infor Proc Sys 12. pp. 449–455. MIT Press (2000)
7. Granger, S., Pennec, X.: Multi-scale EM-ICP: A fast and robust approach for surface registration. In: Euro Conf on Comp Vis. pp. 418–432. Sprin (2002)
8. Hufnagel, H.: A Probabilistic Framework for Point-Based Shape Modeling in Medical Image Analysis. Springer (2011)
9. Maron, B., et. al.: Hypertrophic cardiomyopathy. Interrelations of clinical manifestations, pathophysiology, and therapy. N Engl J Med. 316, 780–844 (1987)
10. Pereanez, M., Lekadir, K., Butakoff, C., Hoogendoorna, C., Frangi, A.F.: A framework for the merging of pre-existing and correspondenceless 3D statistical shape models. Medical Image Analysis 18(7), 1044–58 (2014)
11. Rasoulian, A., Rohling, R., Abolmaesumi, P.: Group-wise registration of point sets for statistical shape models. IEEE Tran on Med Imag 31(11), 2025–2033 (2012)
12. Swift, A., 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 14(40) (2012)
13. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. Neural Comput. 11(2), 443–482 (1999)
14. Voelkel, N., Quaife, R., Leinwand, L., Barst, R., et. al.: Right ventricular function and failure: a report. Circ. 114, 1883–91 (2006)
15. Zhang, S., Zhan, Y., Dewan, M., Metaxas, D.N., Zhou, X.S.: Towards robust and effective shape modeling: Sparse shape composition. Med Imag Anal 16, 265–277 (2012)