Infinite Warped Multimodal Mixture Models

Sahil Loomba
Wyss Institute for Biologically Inspired Engineering
at Harvard University
Boston, MA 02115
sahil.loomba@wyss.harvard.edu

Abstract

A principled technique of combining data across various modalities for the purpose of machine learning applications, both supervised and unsupervised, remains elusive. In this paper, we propose a nonparametric Bayesian approach called the infinite warped multimodal mixture model to learn representations of data across various observed spaces, by using a shared latent space that maps to observations through a Gaussian Process. We further impose a Dirichlet Process prior on the latent space, which makes for a flexible yet constrained regularization of the latent space, allowing us to uncover patterns in high-dimensional datasets with very few samples. We demonstrate this by applying our method to both synthetic as well as real-world datasets. We further enlist some interesting and far reaching applications of this algorithm, both in machine learning, as well as in cognitive science, for modeling cognitive biases acquired through multiple modalities.

1 Introduction

Probabilistic models have been used for a variety of machine learning tasks, such as clustering of data points, learning representations from observed data, discovering underling patterns in it, and elucidating functional relationships. More often than not in the real world, data corresponding to the same entity can come from various modalities or "views", which adds another flavor to all of the above tasks. This is also true for the way we cognitively process data perceived from our surroundings through different sense organs. While the visual modality is strong because most of our sensory input is through the eyes, it can often play interesting effects on the perception of other modalities, like input through the ears. This suggests that we tend to represent data onto some cognitive latent space that is collectively fed into by various observed modalities.

Taking this idea forward, we present a novel nonparametric Bayesian model which discovers this shared latent space of data representation, while organizing the latent space into a flexible yet constrained structure. Observation spaces can contain high-dimensional data arranged in nonparametric cluster shapes. There are many related approaches in machine learning and probabilistic modeling literature, both for clustering of nonparametric cluster shapes called "manifold learning", and for multimodal learning. But the approach outlined in this paper is most closely related to those of Iwata et al. (2013) and Damianou et al. (2016). The reader is referred to both of these as well as the references therein for a more extensive review of current literature.

In Iwata et al. (2013), the authors introduce a probabilistic generative model of warped mixtures for nonparametric cluster shapes. Instead of assuming an infinite Gaussian mixture model in the observed space, they introduce a latent space mapping to it through a Gaussian Process using an RBF kernel function, and assume an iGMM in the latent space instead. They call this the infinite warped mixture model, or iWMM, which infers the mapping function such that latent coordinates will be well modeled by a mixture of Gaussians. In Damianou et al. (2016), the authors introduce a soft version of the shared latent space model for multiple observation spaces. That is, to do multiview

learning, they assume a common latent space which maps to every observation space through an independent Gaussian Process using an RBF kernel function with separate exponential weights for every latent dimension (called the MDR kernel function). Therefore here, the model infers mapping functions such that the latent space acts as a shared representation across modalities.

Our primary contributions in this paper can be seen as the extension of the above two ideas, armed with the intuition of cognitive generalization on a latent space shared by multiple perceptual streams. We allow for a latent space under a Dirichlet Process prior, which imposes an infinite Gaussian mixture model on this space. For the mapping of this space to each observed space, we use a Gaussian Process prior to flexibly "warp" every Gaussian component of the latent space, allowing us to embed highly nonlinear clusters into lower dimensional manifolds in the latent space. Since eventually we integrate over the mapping functions as well as the latent coordinates, this allows for better propagation of uncertainty through the model. This is again analogous to the way we categorize and generalize over multiple data streams by just recalling the perceived category labels, and not explicitly estimating the "best" shared latent space.

The paper is organized as follows. In section 2 we describe the foundational principles, design and inference over the model, which we refer to as the infinite warped multimodal mixture model (iWMMM). In section 3 we enclose some results and analyses on experiments conducted using both synthetic and real-world datasets. In section 4 we conclude with the applications, limitations and extensions of this very flexible probabilistic model.

2 Infinite Warped Multimodal Mixture Model

The infinite warped multimodal mixture model is a combination of two important standard Bayesian models, namely the Gaussian Process (used here in a multimodal setting), and the infinite Gaussian mixture model. We first describe these two in succinct details, before elucidating on the model below.

2.1 Gaussian Process Modeling

A Gaussian Process (GP) is a collection of random variables, any finite number of which have joint Gaussian distributions. A GP is fully parametrized by its mean function $m(x)$ and covariance function $k(x, x')$ (Rasmussen & Williams, 2006). Therefore, GP induces a distribution over functions, and not vectors, as is the case with Gaussian distributions, and is written as $f \sim \mathcal{GP}(m, k)$. Just as random variables are indexed by their position in the vector that is under a Gaussian distribution, function instances are indexed by the input x, as $f(x)$, which is the vale of f that is under a Gaussian process. Hence, if we define $\mu_i = m(x_i)$ and $\Sigma_{ij} = k(x_i, x_j)$, then we can generate a random vector whose coordinates come from $f(x)$ as $\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. More formally, let us consider modeling a function between variables $\mathbf{X} \in \mathbb{R}^{n \times q}$ and $\mathbf{Y} \in \mathbb{R}^{n \times p}$, i.e. $f : \mathbf{X} \to \mathbf{Y}$, where the variables are in the design-matrix form with rows representing a single instance, and columns the dimensions of the feature space (Damianou et al., 2016). Assuming that each dimension of an instance $\mathbf{y}_{i,:}$ is generated by an independent function $\mathbf{f}_{:,j}$ from input $\mathbf{x}_{i,:}$, with some additive Gaussian noise $\epsilon_{ij} \sim \mathcal{N}(0, \beta^{-1})$, we can write:

$$y_{ij} = \mathbf{f}_{:,j}(\mathbf{x}_{\mathbf{i},:}) + \epsilon_{ij} \tag{1}$$

Let $\boldsymbol{\theta}$ represent hyperparameters of the GP prior, that is the paramaeters of the mean and covariance functions. Then we can write all function instances being distributed as:

$$p(\mathbf{f}_{:,j}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{2}$$

where $\mathbf{0}$ refers to the mean vector that is assumed (as is common) to be a zero-vector, and $\mathbf{K}$ refers to the covariance matrix obtained after evaluating the covariance function k, on all instances in $\mathbf{X}$. Furthermore, the assumption of Gaussian noise leads to:

$$p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}, \beta) = \mathcal{N}(\mathbf{f}_{:,j}, \beta^{-1}\mathbf{I}) \tag{3}$$

Then, the marginal likelihood can be written in closed form after integrating out all the mapping functions:

$$p(\mathbf{y}_{:,j}|\mathbf{X}, \boldsymbol{\theta}, \beta) = \int_{\mathbf{f}_{:,j}} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}, \beta) p(\mathbf{f}_{:,j}|\mathbf{X}, \boldsymbol{\theta}) \tag{4}$$

In matrix form, this can be written as:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \beta) &= \prod_{j=1}^{p} \mathcal{N}(\mathbf{0}, \mathbf{K} + \beta^{-1}\mathbf{I}) \\ &= (2\pi)^{-qn/2}|\mathbf{K}|^{-q/2}\text{exp}\left(-\frac{1}{2}\text{tr}(\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{Y})\right) \end{aligned} \tag{5}$$

Maximizing this conditional likelihood for a $\mathbf{X}$ of small dimension q, one can achieve a dimensionality reduction of data from p to q dimensions. This joint optimization over $\mathbf{X}$ and $\boldsymbol{\theta}$ is referred to as the Gaussian latent variable model (GP-LVM) (Lawrence, 2004), which can be shown to be equivalent to a Principle Component Analysis. Under different assumptions of the kernel, i.e. the covariance function k, one can constrain the optimization problem in different ways to obtaining interesting (and fantastical) visualizations. For the purposes of this paper, we consider the popular radial basis function (RBF) kernel (which absorbs the error term) given by:

$$k_{\text{RBF}}(\mathbf{x}_{u,:}, \mathbf{x}_{v,:}) = \alpha \exp\left(-\gamma \sum_{j=1}^{q}(x_{uj} - x_{vj})^2\right) + \delta_{uv}\beta^{-1} \tag{6}$$

2.2 Infinite Gaussian Mixture Model

A k-component Gaussian mixture model is a standard generative model for data, wherein the generative process involves first selecting one of the k Gaussian components c according to some probability π_c, and then generating the data point according to the selected Gaussian's parametrization $\boldsymbol{\mu}_c$ and $\boldsymbol{R}_c$, where $\boldsymbol{R}$ refers to the precision matrix (inverse of the covariance matrix). That is, the probability of data $\mathbf{X}$ can be written as:

$$p(\mathbf{x}_{i,:}|\{\pi_c, \boldsymbol{\mu}_c, \mathbf{R}_c\}_{c=1}^{k}) = \sum_{c=1}^{k} \pi_c \mathcal{N}(\boldsymbol{\mu}_c, \mathbf{R}_c^{-1}) \tag{7}$$

Let the indicator variables z_i suggest which component/class $\mathbf{x}_{i,:}$ belongs to. Now, the mixing proportions $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_k\}$ are given a symmetric Dirichlet prior distribution, with the concentration parameter α/k, i.e.:

$$p(\boldsymbol{\pi}|\alpha) \sim \text{Dirichlet}(\alpha/k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^{k} \pi_j^{\alpha/k-1} \tag{8}$$

where mixing proportions sum to one. Given $\boldsymbol{\pi}$, the prior for occupation numbers n_j is multinomial, and the distribution of indicator variables becomes:

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{j=1}^{k} \pi_j^{n_j} \tag{9}$$

Combining equations 8 and 9, we can integrate over the mixing proportions as follows:

$$\begin{aligned} p(\mathbf{z}|\alpha) &= \int_{\boldsymbol{\pi}} p(\boldsymbol{z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha) \\ &= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{j=1}^{k} \frac{\Gamma(n_j + \alpha/k)}{\Gamma(\alpha/k)} \end{aligned} \tag{10}$$

Since the indicator variables are discrete, to be able to do Gibbs sampling for inference over the variable space, we can convert equation 10 to conditional prior for an indicator of interest, given every other indicator variable:

$$p(z_i = j|\mathbf{z}_{-i}, \alpha) = \frac{n_{-i,j} + \alpha/k}{n - 1 + \alpha} \tag{11}$$

where a negative index denotes all indexes other than the one mentioned. Now, say we do not have a fixed number of clusters. This would refer to a k-component GMM where $k \to \infty$, known as

the infinite Gaussian mixture model or, iGMM (Rasmussen, 1999). Note that marginalizing over the "infinite" mixing proportions allows us to take a limit of equation 11, and work with the "finite" number of indicator variables, writing their conditional priors as:

$$\begin{aligned} p(z_i = j|\mathbf{z}_{-i}, \alpha) &= \frac{n_{-i,j}}{n-1+\alpha} \text{ , where } n_{-i,j} > 0 \\ p(z_i \neq z_{i'} \forall i \neq i'|\mathbf{z}_{-i}, \alpha) &= \frac{\alpha}{n-1+\alpha} \text{ , for all other (unoccupied) components combined} \end{aligned} \tag{12}$$

This type of a prior on the mixture proportions is called a Dirichlet Process (DP) prior, with concentration paramater α.

2.3 Model Representation

Let us elaborate on extending the GP-LVM model into a multimodal setting, referred to as the shared GP-LVM (Damianou et al., 2016). Say we have κ sets of observations, called modalities or views, corresponding to the same entity, given by $\mathcal{Y} = \{Y^k\}_{k=1}^{\kappa}$. Let us assume that all views are generated from the same latent space $\mathbf{X}$, by κ independent GP mappings, each with their set of kernel hyperparameters $\mathbf{\Theta} = \{\boldsymbol{\theta}_{k=1}^{\kappa}\}$ and noise parameters $\boldsymbol{\beta} = \{\beta^k\}_{k=1}^{\kappa}$. (That is, given the latent space, observation spaces are independent of each other.) For this shared GP-LVM we can rewrite equation 5 as:

$$\begin{aligned} p(\mathcal{Y}|\mathbf{X}, \mathbf{\Theta}, \boldsymbol{\beta}) &= \prod_{k=1}^{\kappa} \prod_{j=1}^{p} \mathcal{N}(\mathbf{0}, \mathbf{K}^k + (\beta^k)^{-1}\mathbf{I}) \\ &= (2\pi)^{-qn\kappa/2} \prod_{k=1}^{\kappa} \left(|\mathbf{K}^k|^{-q/2} \right) \exp\left(-\frac{1}{2} \sum_{k=1}^{\kappa} \text{tr}\left((\mathbf{Y}^k)^T (\mathbf{K}^k)^{-1} \mathbf{Y}^k \right) \right) \end{aligned} \tag{13}$$

Additionally, as discussed in Damianou et al. (2016), the non-convexity of the above objective function requires a rather neat initialization of the latent space for this approach to reach a good solution in a reasonable amount of exploration. While for a unimodal approach, such an initialization can be performed using spectral methods or other non-linear dimensionality reduction techniques on the modality, no such approach exists in a multimodal setting. To resolve this issue (at least to some extent), the authors suggest doing a "soft" separation of the latent space for every corresponding modality, instead of a "hard" separation into shared and private latent spaces for every view. This can be done by tweaking the kernel function, by using a different weight for every latent dimension. This is referred to as manifold relevance determination (MRD) (Damianou, 2015). We can rewrite the kernel equation in 6 as:

$$k_{\text{MRD}}(\mathbf{x}_{u,:}, \mathbf{x}_{v,:}) = \alpha \exp\left(-\sum_{j=1}^{q} \gamma_j (x_{uj} - x_{vj})^2 \right) + \delta_{uv}\beta^{-1} \tag{14}$$

In this equation, γ_j encodes the relative relevance of latent dimension q in determining the covariance between uth and vth data point. Note that in the multimodal setting, we have a different MRD kernel for every view.

Additionally, we must put priors over the parameters of every Gaussian component, to allow for full Bayesian learning. We place a shared Gaussian prior on the means and a shared Wishart prior on the precision matrices of all components, since these are the conjugate priors to a Gaussian distribution with unknown mean and variance, enabling us to do exact inference by marginalizing over the component parameters (Iwata et al., 2013).

$$p(\boldsymbol{\mu}_c, \mathbf{R}_c|\mathbf{u}, r, \mathbf{S}, \nu) = \mathcal{N}\left(\mathbf{u}, (r\mathbf{R}_c)^{-1}\right) \mathcal{W}(\mathbf{S}^{-1}, \nu) \tag{15}$$

Moreover, we put vague hyperpriors on all of these iGMM priors (namely $\alpha, \mathbf{u}, r, \mathbf{S}, \nu$), and on the kernel hyperparameters of the shared GP-LVM (namely $\mathbf{\Theta}, \boldsymbol{\beta}$). We are now prepared to describe the representation of the infinite warped multimodal mixture model, which is essentially an iGMM stacked on top of a soft-shared GP-LVM, as shown in figure 1. (Note that we are using η as the concentration parameter, to avoid notational ambiguity with the kernel hyperparameter α.) It can be described using the following generative process:

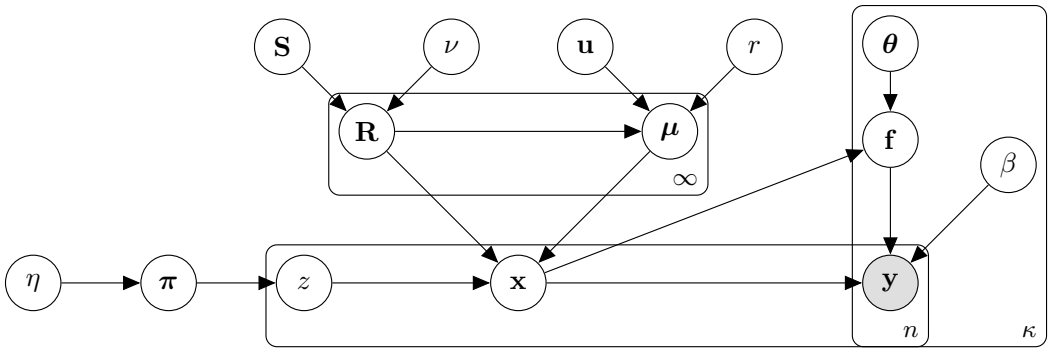


Figure 1: A graphical model representation of the infinite warped multimodal mixture model, where the shaded and unshaded nodes indicate observed and latent variables respectively, and plates indicate repetition. Note that hyperpriors have been excluded for clarity.

1. Draw mixture weights $\boldsymbol{\pi} \sim \mathcal{DP}(\eta)$
2. For each component $c = 1, \ldots, \infty$
   (a) Draw precision matrix $\mathbf{R}_c \sim \mathcal{W}(\boldsymbol{S}^{-1}, \nu)$
   (b) Draw mean $\boldsymbol{\mu}_c \sim \mathcal{N}(\boldsymbol{u}, (r\mathbf{R}_c)^{-1})$
3. For each entity $i = 1, \ldots, n$
   (a) Draw latent assignment $z_i \sim \text{Multinomial}(\boldsymbol{\pi})$
   (b) Draw latent coordinates $\mathbf{x}_{i,:} \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \mathbf{R}_{z_i}^{-1})$
4. For each view $k = 1, \ldots, \kappa$
   (a) Compute kernel K^k
   (b) For each observed dimension $j = 1, \ldots, p^k$
      i. Draw function $\mathbf{f}_{:,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^k)$
      ii. For each observation $i = 1, \ldots, n$
         A. Draw feature $y_{ij} \sim \mathcal{N}(\mathbf{f}_{:,j}(\mathbf{x}_{i,:}), (\beta^k)^{-1})$

2.4 Inference

The posterior over $\mathbf{X}$ and $\mathbf{z}$ can be obtained by using Markov chain Monte Carlo (MCMC). Since the component assignments are discrete, and we can analytically integrate out Gaussian parameters, and use collapsed Gibbs sampling for finding the assignment posterior $p(\mathbf{z}|\mathbf{X}, \mathbf{u}, r, \mathbf{S}, \nu, \eta)$. (The exact expressions of posteriors for the Gaussian parameters and the component assignments are lengthy and not enclosed here. For more, the reader is referred to Iwata et al. (2013).) For the continuous latent coordinates, we can use Hybrid Monte Carlo (HMC) (Duane et al., 1987) to sample from the posterior $p(\mathbf{X}|\mathbf{z}, \mathcal{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{u}, r, \mathbf{S}, \nu)$. HMC is used to both optimize on $\mathbf{X}$, as well as on the kernel hyperparameters $\boldsymbol{\Theta}, \boldsymbol{\beta}$. This requires computing gradients of the log of the unnormalized posterior $\log p(\mathcal{Y}|\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\beta}) + \log p(\mathbf{X}|\mathbf{z}, \mathbf{u}, r, \mathbf{S}, \nu)$, with respect to both $\mathbf{X}$ and the hyperparameters. For $\mathbf{X}$, the second term can be simply written as:

$$\frac{\partial \log p(\mathbf{X}|\mathbf{z}, \mathbf{u}, r, \mathbf{S}, \nu)}{\partial \mathbf{x}_{i,:}} = -\nu_{z_i} \mathbf{S}_{z_i}^{-1} (\mathbf{x}_{i,:} - \mathbf{u}_{z_i}) \tag{16}$$

For the first term, since every observation modality is independent given the latent space, we first decompose the probability of every view separately, and then compute for a view $\mathbf{Y}$ using the chain rule (and equation 5 and 14) as (the indexing on view has been omitted for clarity):

$$\begin{aligned} \frac{\partial \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \beta)}{\partial \mathbf{K}} &= -\frac{1}{2} p \mathbf{K}^{-1} + \frac{1}{2} \mathbf{K}^{-1} \mathbf{Y}\mathbf{Y}^T \mathbf{K}^{-1} \\ \frac{\partial k(\mathbf{x}_{u,:}, \mathbf{x}_{v,:})}{\partial \mathbf{x}_{u,:}} &= -2\alpha \exp\left(-\sum_{j=1}^{q} \gamma_j (x_{uj} - x_{vj})^2 \right) (\mathbf{x}_{u,:} - \mathbf{x}_{v,:}) \end{aligned} \tag{17}$$

For hyperparameters, the second term will differentiate to 0. For the first term, we again use the chain rule, with the second term of the chain product as:

$$\begin{aligned}
\frac{\partial k(\mathbf{x}_{u,:}, \mathbf{x}_{v,:})}{\partial \alpha} &= \exp\left(-\sum_{j=1}^{q} \gamma_j (x_{uj} - x_{vj})^2\right) \\
\frac{\partial k(\mathbf{x}_{u,:}, \mathbf{x}_{v,:})}{\partial \beta} &= 1 \text{ if } u = v \text{ else } 0 \\
\frac{\partial k(\mathbf{x}_{u,:}, \mathbf{x}_{v,:})}{\partial \gamma_j} &= -\alpha \exp\left(-\sum_{j=1}^{q} \gamma_j (x_{uj} - x_{vj})^2\right)(x_{uj} - x_{vj})
\end{aligned} \tag{18}$$

Thus, by iteratively alternating between Gibbs sampling of $\mathbf{z}$ and HMC of $\mathbf{X}$, we obtain samples from the posterior $p(\mathbf{X}, \mathbf{z}|\mathcal{Y}, \Theta, \mathbf{u}, r, \mathbf{S}, \nu, \eta)$.

2.5 Feature Importance

$$f_i^y = T_{ii} = \sum_j \sum_k Y_{ji} Y_{ki} K_{jk}^{-1}$$
$$f_i^x = \gamma_i$$

3 Experimental Results

The primary motivation of this paper is to create a nonparametric Bayesian model that can find nonlinearly separable clusters in high-dimensional multimodal data, even in the limit of few data points. That is, we have presented a manifold learning algorithm for data with few samples in multiple high dimensional feature spaces, which can capture certain high-level features of interest. In this multimodal setting, the notion of "supervision" gets dissolved, since one modality is essentially supervising the other. Thus, one could even imagine defining a modality called "labels", which could somehow encode the class labels of every data point, and thus fully supervised classification can be achieved using this method. To test whether this multimodal supervision works, we run experiments on both a toy and a real-world dataset.

3.1 Toy Dataset

The infinite warped mixture modal defined in Iwata et al. (2013) shows considerable success in clustering non-Gaussian shaped datasets, such as the 2-curve, 3-semi, 2-circle and pinwheel datasets. Another arguably more challenging benchmark for manifold clustering is the double Swiss roll. We use a 2D version of the double Swiss roll for our experiments, with as few as 200 points, 100 in each of the two manifolds. First, we use the singleview version of our algorithm, which is equivalent to iWMM using the MRD kernel function. For all experiments in this paper, MCMC is run for 10,000 iterations. When number of latent dimensions $q = 2$, looking at a posterior sample reveals that the algorithm uncovers 3 clusters, which correspond to the two arms of the inner arc of the double Swiss roll, and its combined outer arm (see figure 2). That is, it is unable to separate the arms of the outer arc. On more careful observation, we see that where the data density is high, in the inner arc, the clustering is very good, while where it is lower, in the outer arc, it merges the two clusters into one Gaussian. This could be an evidence of the Bayes' Occam's razor at work. Next, in the multiview version where labels are the other view, encoded by a Boolean $0/1$, we get 3 clusters again, however this time the observed space is rather oddly segmented (see figure 3). After looking at the latent space, we see that the algorithm in some sense does separate out the two arms of the roll. While the first dimension seems to correspond to some notion of "radius of the arc", the second dimension seems to concord with the "angular position in the arc". However, the arcs themselves have been cropped up. But, although the clustering is not appropriate, the labels view has managed to supervise more structure onto the latent space, and hence onto the other observed space. This is further exemplified when we look at a 2D projected posterior sample of the latent space with $q = 3$ in this supervised setting, as in figure 4. Another key point to consider is the variation in the kernel hyperparameter γ_j, since using the MRD kernel allows us to weight every latent dimension differently for different

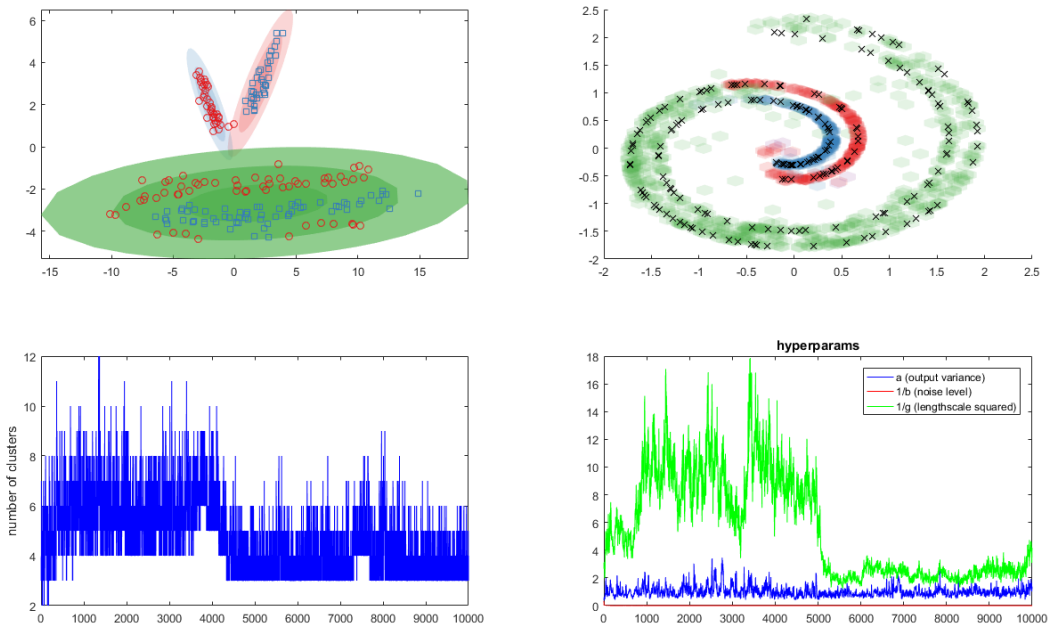


Figure 2: A posterior sample for the double Swiss roll dataset in the singleview setting, when latent space is of 2 dimensions
1L: Latent Space, 1R: Observed Space, 2L: Variation in number of clusters as MCMC progresses, 2R: Variation in kernel hyperparameters as MCMC progresses (averaged over all latent dimensions)

views, which wouldn't have been possible with a regular RBF kernel. As seen in figure 5, the 3rd latent dimension collapses entirely for both views, 2nd dimension is a private latent space for the data view, and 1st dimension is a shared latent space for both data and label views. The fact that the labels only make use of the first latent dimension is reflected in the way the labels view supervises latent space structure, as in figure 4.

3.2 Real-world Dataset

The ultimate test of manifold learning lies not in carefully synthesized datasets, but in the real-world. We therefore use the extended Yale face database B (Georghiades et al., 2001; Lee at al., 2005), which is a collection of different individuals under different lighting conditions. We use a version with cropped grayscale images of size 168x192, which we downsample to 32x32. We make use of faces of two individuals, with illumination angles varying both in the azimuth and polar directions. We divide the faces into two views where every entity is associated with (1) a positive azimuth angle $+\theta^\circ$ corresponding to right-illumination-view, and (2) a negative azimuth angle $-\theta^\circ$ corresponding to left-illumination-view. This gives us a total of 58 entities, 29 per person, with an observation in each view. When the iWMMM is run with $q = 2$, we obtain 3 very neat Gaussian clusters, 1 (roughly) belonging to one subject, and the other 2 (roughly) belonging to the other (see figure 6). On plotting the faces on the latent space, we notice that within each cluster, the x-axis seems representative of the deviation of the azimuth angle of illumination from 0°, and the y-axis seems to indicate the polar angle of illumination. All of these are very high-level features of the dataset, which are easily recovered by iWMMM. If we push q up to 8, we again obtain 3 clusters, 1 belonging to each of the subjects, and 1 common to both. Variation along the diagonal seems indicative of the polar angle, while the two individuals are maintained in roughly two separate manifolds along this diagonal. On the other hand, for the singleview setting when $q = 2$, although we obtain 2 Gaussian clusters which are a true representation of the number of low dimensional manifolds, they aren't very distinctly separated even in the 2D latent space. Clearly, by organizing our dataset into two different views (left/right illumination), we add more information to the structure of the observed space, and hence of the latent space.

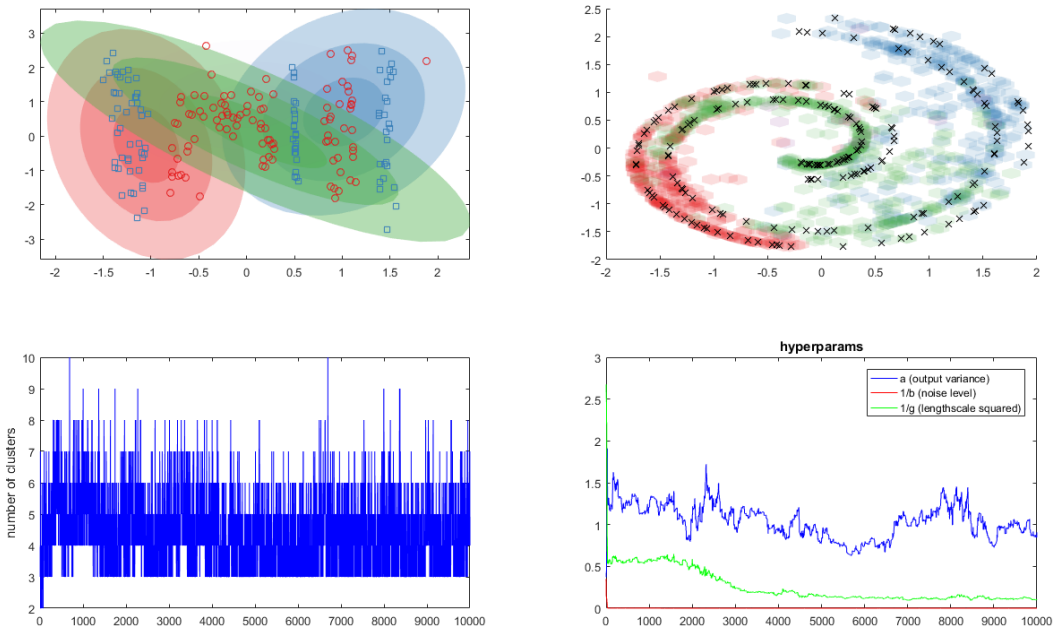


Figure 3: A posterior sample for the double Swiss roll dataset in the multiview setting, when latent space is of 2 dimensions
1L: Latent Space, 1R: Observed Space, 2L: Variation in number of clusters as MCMC progresses, 2R: Variation in kernel hyperparameters as MCMC progresses (averaged over all latent dimensions and views)

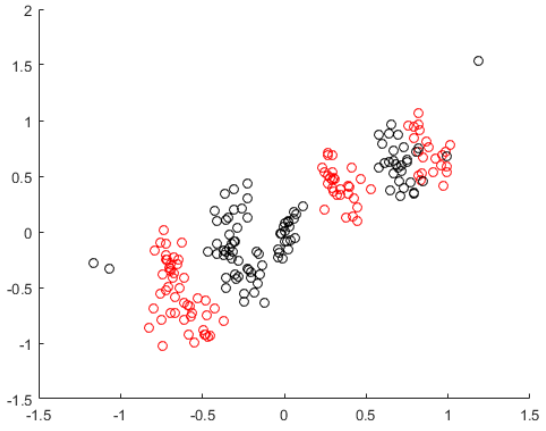
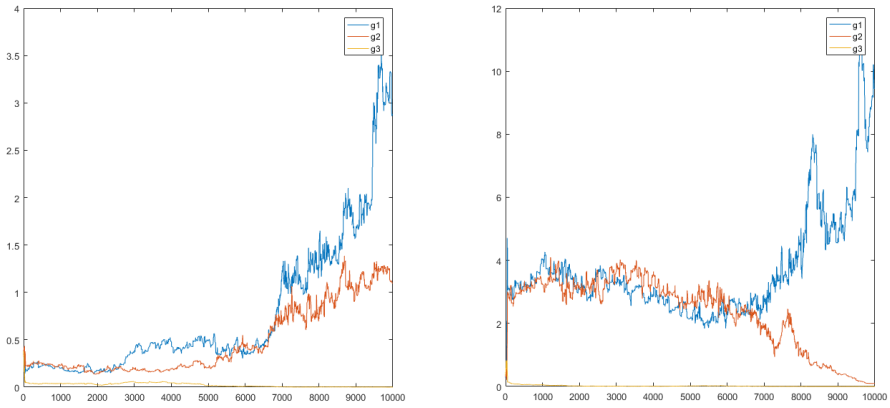


Figure 4: A posterior sample for the double Swiss roll dataset in the multiview setting, (when latent space is of 3 dimensions,) plotted using random projection onto 2D.



(a) View 1: Data

(b) View 2: Labels

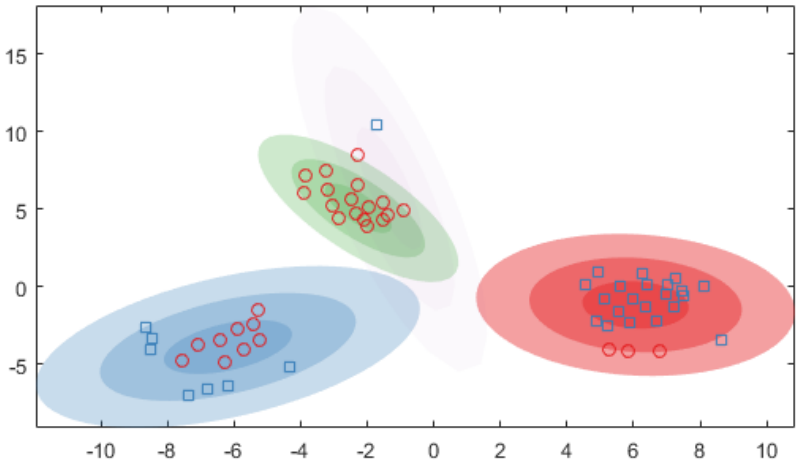
Figure 5: Variation of the parameter γ_j for a 3D Latent Space as MCMC progresses for the double Swiss roll dataset in a multiview setting.

4 Conclusion and Future Work

We have presented a novel algorithm for doing nonparametric Bayesian clustering of multimodal data, through a shared latent space. Our model can be seen as a generalization of either the infinite warped Gaussian mixture model or an extension of the soft-shared GP-LVM. We show how both the presence of a Dirichlet Process prior as in the former, and a multimodal approach as taken in the latter, allows us to better represent latent spaces for real-world high dimensional datasets. A Gaussian prior density on the latent space acts as a regularizer, and completely integrating out the latent space using the Dirichlet process prior allows for an even more flexible parametrization than a single isotropic Gaussian (Iwata et al., 2013). Converting datasets into a multiview format through domain expert knowledge can be seen as a means of providing a strong supervision in the right direction for learning an effective latent space for our data.

The applications of iWMMM, other than those demonstrated above, are plenty. Since it is based on GP-LVM, it can be used for density estimation in both the latent as well as observed spaces, and for visualization of multimodal high-dimensional data. It can also be used for missing value imputations across observation spaces, using mapping to the shared latent space as a common ground to fantasize data in any of the other observation spaces (Damianou, 2015). This model can have interesting applications in cognitive science as well. For instance, children have been noted to show the shape bias, wherein they generalize information about objects by its shape, as opposed to othebe seen to have observations in at least two views: the shape view and the word (noun) view. The development of a shape bias could be seen as the noun view anchoring the shape view through a shared latent space over which the infant attempts to learn natural categories of objects. iWMMM can also be used to reconcile segmented feature spaces for a set of entities. For instance, CrossCat is a nonparametric Bayes algorithm to infer structure in feature spaces, and cluster entities with each of the clustered feature spaces (Mansinghka et al., 2016). These multiple clusterings can be difficult to reconcile, and iWMMM provides a principled method of combining them, by interpreting every feature space cluster as a separate observation space, independent of each other given some latent space. iWMMM can also be used in increasingly important cross-modal challenges within Computer Vision, such as those in multimodal scene representation (Castrejón et al., 2016) where images compose one observation space, and word-vectors compose another, or in zero-shot learning through cross-modal transfer (Socher et al., 2013), and other avenues in transfer learning.

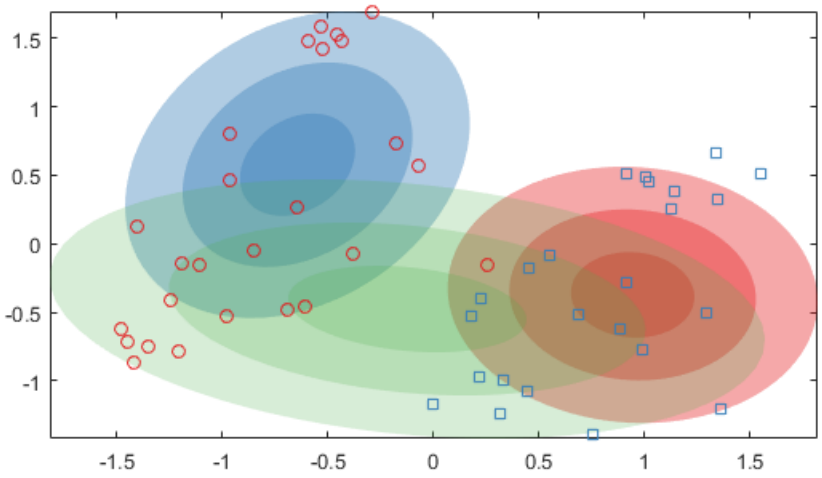
There are also some interesting ways in which the iWMMM can be extended. First, in terms of inference, the algorithm is rather slow, in particular as the number of data points increase, making it unscalable. Techniques from variational inference can be used here to fasten MCMC inference (Damianou, 2015). Second, in terms of representation, a single Gaussian process prior may be



(a) Points in the Latent Space

(b) Faces plotted in Latent Space

Figure 6: A posterior sample for the Yale faces dataset in the multiview setting, when latent space is of 2 dimensions. Note that faces of both views (left/right) have been placed tightly adjacent in this plot.



(a) Points in the first 2 dimensions of Latent Space

(b) Faces plotted in Latent Space projected onto a 2D space using Random Projections

Figure 7: A posterior sample for the Yale faces dataset in the multiview setting, when latent space is of 8 dimensions. Note that faces of both views (left/right) have been placed tightly adjacent in this plot.

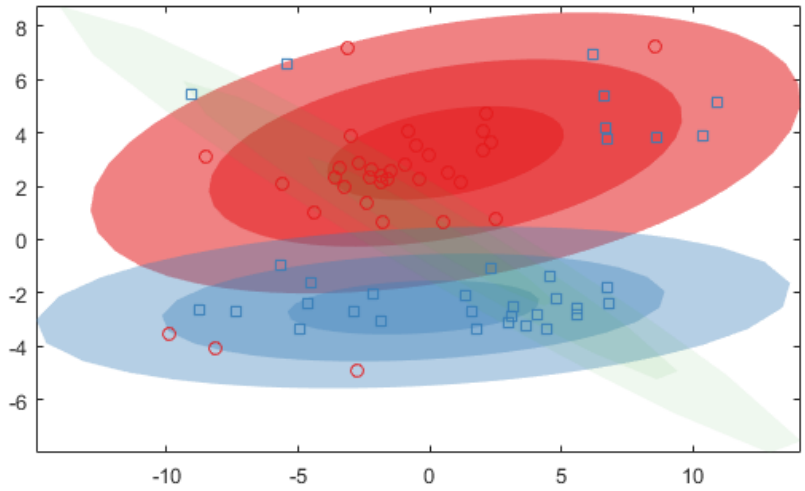


Figure 8: A posterior sample for the Yale faces dataset in the singleview setting, when latent space is of 2 dimensions.

insufficient to capture cluster shapes that are highly warped. There are multiple ways to address this problem, one of them being to consider a separate GP mapping for every cluster. Another method could be the composition of multiple Gaussian processes in succession, inspired by the expressive power of composing perceptrons, that has advanced the field of deep learning. Such Deep Gaussian Processes would be able to fit to increasingly more complicated nonparametrically shaped clusters (Damianou & Lawrence, 2013). In the multiview setting, Lawrence & Moore (2007) have attempted to hand design a hierarchically arranged deep latent space, but this creates several issues, in particular that we cannot integrate over the entire latent space, which hinders in the propagation of uncertainty. Thus, it would be interesting to allow integration over hierarchically instantiated deep latent spaces, such as by imposing generic priors on tree-structures of latent spaces, like the nested Chinese restaurant process prior (Blei at al., 2004), or the Pitman Yor diffusion tree prior(Knowles & Ghahramani, 2015). This would enable multiple modalities to arrange themselves in a hierarchy of shared latent spaces, permitting a very autonomous spread of supervision across the modalities. For instance, two visual modalities can supervise each other through a more closely shared latent space (analogous to the brain involuntarily reconciling the two views perceived through two human eyes into a 3D composition of the world), as opposed to when supervising an aural modality (analogous to a more high-level and voluntary reconciliation of cues of speech seen through the eyes and heard through the ears, by the brain).

Acknowledgments

The author would like to thank Joshua B. Tenenbaum, Atabak Ashfaq, and Ishaan Grover for enlightening discussions that were immensely helpful. The author would also like to acknowledge the use of iWMM's codebase, which their authors have made readily available at `http://github.com/duvenaud/warped-mixtures`, and was built upon in this paper to a multimodal setting with soft-sharing of the latent space.

References

Blei, D. M., Griffiths, T. L., Jordan, M. I. & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16, 17.

Castrejón, L., Aytar, Y., Vondrick, C., Pirsiavash, H., & Torralba, A. *Learning Aligned Cross-Modal Representations from Weakly Aligned Data.*

Damianou, A. C., & Lawrence, N. D. (2013, August). Deep Gaussian Processes. In *AISTATS* (pp. 207-215).

Damianou, A. (2015). *Deep Gaussian processes and variational propagation of uncertainty* (Doctoral dissertation, University of Sheffield).

Damianou, A., Lawrence, N. D., & Ek, C. H. (2016). Multi-view Learning as a Nonparametric Nonlinear Inter-Battery Factor Analysis. *arXiv preprint arXiv:1604.04939.*

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B, 195*(2), 216-222.

Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence, 23*(6), 643-660.

Iwata, T., Duvenaud, D., & Ghahramani, Z. (2013). Warped Mixtures for Nonparametric Cluster Shapes. In *Uncertainty in Artificial Intelligence* (p. 311).

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science, 10*(3), 307-321.

Knowles, D. A., & Ghahramani, Z. (2015). Pitman Yor Diffusion Trees for Bayesian Hierarchical Clustering. *IEEE transactions on pattern analysis and machine intelligence, 37*(2), 271-289.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems, 16*(3), 329-336.

Lawrence, N. D., & Moore, A. J. (2007, June). Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning* (pp. 481-488). ACM.

Lee, K. C., Ho, J., & Kriegman, D. J. (2005). Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence, 27*(5), 684-698.

Mansinghka, V., Shafto, P., Jonas, E., Petschulat, C., Gasner, M., & Tenenbaum, J. B. (2016). CrossCat: A fully Bayesian nonparametric method for analyzing heterogeneous, high dimensional data. In *Journal of Machine Learning Research 17*(138), 1-49.

Rasmussen, C. E. (1999, December). The infinite Gaussian mixture model. In *NIPS* (Vol. 12, pp. 554-560).

Rasmussen, C. E., Williams, C. K. I. (2006). *Gaussian processes for machine learning.*

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems* (pp. 935-943).