Chapter 1

Warped Mixture Models

> "What, exactly, is a cluster?"
>
> - Bernhard Schölkopf, personal communication

Previous chapters showed how the probabilistic nature of GPs lets us automatically include an appropriate amount of structure, and the correct kind, when building models of functions. In this chapter, we show how to take advantage of this property when composing GPs with other models, automatically balancing between the complexity of the GP and the other parts of the model.

This chapter considers a simple example: a Gaussian mixture model, warped by a draw from a GP. This novel model produces clusters with arbitrary shapes, depending on the warping. We call the proposed model the *infinite warped mixture model* (iWMM). Figure 1.3 shows a set of manifolds and datapoints sampled from the prior defined by this model. The probabilistic nature of the iWMM lets us automatically infer the number, dimension, and shape of a set of nonlinear manifolds, and summarize those manifolds in a low-dimensional latent space.

The work comprising the bulk of this chapter was done in collaboration with Tomoharu Iwata and Zoubin Ghahramani, and appeared in **?**. The main idea was born out of a conversation between Tomo and myself, and together we wrote almost all of the code as well as the paper. Tomoharu ran most of the experiments. Zoubin Ghahramani provided guidance and many helpful suggestions throughout the project.

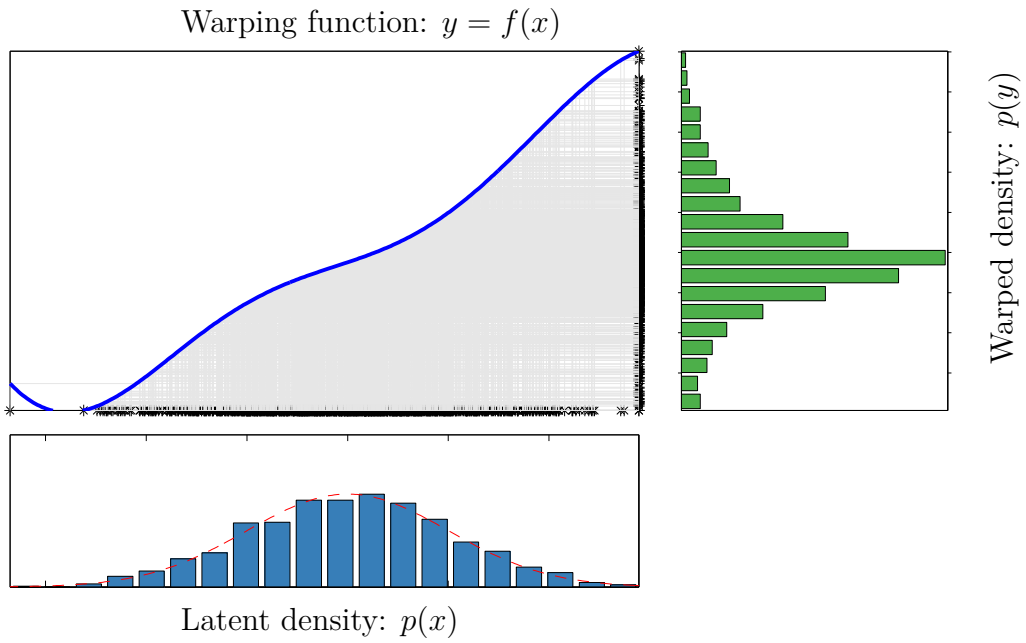


Figure 1.1 A draw from a one-dimensional Gaussian process latent variable model. *Bottom:* the density of a set of samples from a 1D Gaussian, specifying the distribution $p(x)$ in the latent space. *Top Right:* A function $y = f(x)$ drawn from a GP prior. Grey lines show points being mapped through f. *Left:* A nonparametric density $p(y)$ defined by warping the latent density through the sampled function.

1.1 The Gaussian process latent variable model

The iWMM can be viewed as an extension of the Gaussian process latent variable model (GP-LVM) (**?**), a probabilistic model of nonlinear manifolds. The GP-LVM smoothly warps a Gaussian density into a more complicated distribution, using a draw from a GP. Usually, we think of the Gaussian density as living in a "latent space" having Q dimensions, and the warped density living in the observed space having D dimensions.

A standard definition of the GP-LVM is as follows:

$$\text{latent coordinates } \mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^\mathsf{T} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{x}|0, \mathbf{I}_Q) \tag{1.1}$$

$$\text{warping functions } \boldsymbol{f} = (f_1, f_2, \ldots, f_D)^\mathsf{T} \overset{\text{iid}}{\sim} \mathcal{GP}(0, \text{SE} + \text{WN}) \tag{1.2}$$

$$\text{observed datapoints } \mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)^\mathsf{T} = \boldsymbol{f}(\mathbf{X}) \tag{1.3}$$

Under the GP-LVM, the probability of observations given the latent coordinates

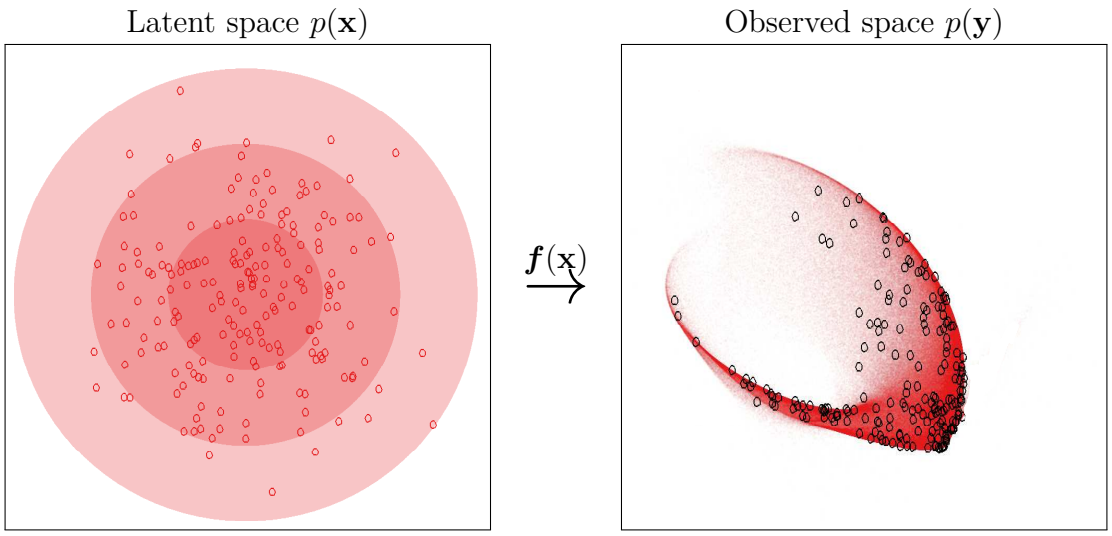


Figure 1.2 A draw from a multi-dimensional Gaussian process latent variable model. *Left:* Isocontours and samples from a 2D Gaussian, specifying the distribution $p(\mathbf{x})$ in the latent space. *Right:* Observed density $p(\mathbf{y})$ has a nonparametric shape, defined by warping the latent density through a function drawn from a GP prior.

integrating out the mapping functions, is simply a product of multivariate normals:

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}) = \prod_{d=1}^{D} p(\mathbf{Y}_{:,d}|\mathbf{X},\boldsymbol{\theta}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{Y}_{:,d}|0,\mathbf{K}_{\boldsymbol{\theta}}) \tag{1.4}$$

$$= (2\pi)^{-\frac{DN}{2}} |\mathbf{K}|^{-\frac{D}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{Y}^{\mathsf{T}}\mathbf{K}^{-1}\mathbf{Y})\right), \tag{1.5}$$

where $\boldsymbol{\theta}$ are the kernel parameters and $\mathbf{K}$ is the Gram matrix $k(\mathbf{X},\mathbf{X})$.

Typically, the GP-LVM is used for dimensionality reduction or visualization, and the latent coordinates are set by maximizing (1.5). In that setting, the Gaussian prior density on $\mathbf{x}$ is essentially a regularizer which keeps the latent coordinates from spreading arbitrarily far apart. In contrast, we integrate out the latent coordinates, and the iWMM places a more flexible parameterization on $p(\mathbf{x})$ than a single isotropic Gaussian. Just as the GP-LVM can be viewed as a manifold learning algorithm, the iWMM can be viewed as learning a set of manifolds, one for each cluster.

1.2 The infinite warped mixture model

This section defines in detail the infinite warped mixture model (iWMM). Like the GP-LVM, the iWMM assumes a smooth nonlinear mapping from a latent density to an

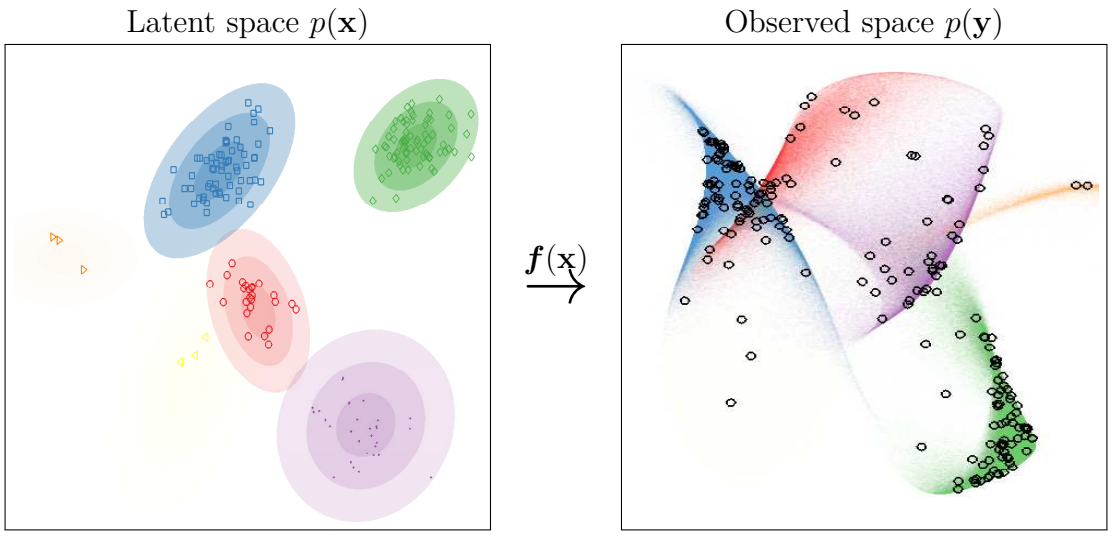


Figure 1.3 A sample from the iWMM prior. *Left:* In the latent space, a mixture distribution is sampled from a Dirichlet process mixture of Gaussians. *Right:* The latent mixture is smoothly warped to produce a set of non-Gaussian manifolds in the observed space.

observed density. The difference is that the iWMM assumes that the latent density is an infinite Gaussian mixture model (iGMM) (**?**):

$$p(\mathbf{x}) = \sum_{c=1}^{\infty} \lambda_c \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \mathbf{R}_c^{-1}) \tag{1.6}$$

where λ_c, $\boldsymbol{\mu}_c$ and $\mathbf{R}_c$ is the mixture weight, mean, and precision matrix of the c^{th} mixture component.

The iWMM can be seen as a generalization of either the GP-LVM or the iGMM. To be precise, the iWMM with a single fixed spherical Gaussian density on the latent coordinates corresponds to the GP-LVM, while the iWMM with fixed direct mapping function $f_d(\mathbf{x}) = x_d$ and $Q = D$ corresponds to the iGMM.

A flexible model of cluster shapes is required to correctly estimate the number of clusters, if those clusters do not happen to be Gaussian. For example, a mixture of Gaussians fit to a single non-Gaussian cluster (such as one that is curved or heavy-tailed) will report that the data contains many Gaussian clusters.

1.3 Inference

As discussed in **??**, one of the main advantages of GP priors is that, given inputs $\mathbf{X}$, outputs $\mathbf{Y}$ and kernel parameters $\boldsymbol{\theta}$, we can analytically integrate over functions mapping $\mathbf{X}$ to $\mathbf{Y}$. However, inference becomes more difficult if we introduce uncertainty about the kernel parameters, or the input locations $\mathbf{X}$. This section outlines how to infer all parameters in the iWMM given only a set of observations $\mathbf{Y}$. Details can be found in appendix **??**.

By placing conjugate priors on the parameters of the Gaussian mixture components, we can analytically integrate out the cluster shapes, given the assignments of points to clusters. The only remaining variables to infer are the latent points $\mathbf{X}$, the cluster assignments $\mathbf{z}$, and the kernel parameters $\boldsymbol{\theta}$. We can obtain samples from their posterior $p(\mathbf{X}, \mathbf{z}, \boldsymbol{\theta}|\mathbf{Y})$ by iterating two steps:

1. Given the latent points $\mathbf{X}$, we sample the discrete cluster memberships $\mathbf{z}$ using collapsed Gibbs sampling, integrating out the mixture parameters (**??**).

2. Given the cluster assignments $\mathbf{z}$, we sample the continuous latent coordinates $\mathbf{X}$ and kernel parameters $\boldsymbol{\theta}$ using Hamiltonian Monte Carlo (HMC) (**?**, chapter 30). The relevant equations are given by **????????**.

The complexity of each iteration of HMC is dominated by the $\mathcal{O}(N^3)$ computation of $\mathbf{K}^{-1}$. This complexity could be improved by making use of an inducing-point approximation (**??**).

Posterior predictive density

One disadvantage of this model class is that its predictive density has no closed form. To approximate the predictive density, we sample latent points from the posterior on the latent density, and map them through warpings drawn from the corresponding posterior density. The Gaussian noise added to each observation means that each sample adds a Gaussian to the Monte Carlo estimate of the predictive density. Details can be found in **??**. This procedure was used to generate the plots of posterior density in figures 1.3, 1.4 and 1.6.

1.4 Related work

The literature on manifold learning, clustering and dimensionality reduction is extensive. This section highlights some of the most relevant related work.

Extensions of the GP-LVM

The GP-LVM has been used effectively in a wide variety of applications (**???**). The latent positions $\mathbf{X}$ in the GP-LVM are typically obtained by maximum a posteriori estimation or variational Bayesian inference (**?**), placing a single fixed spherical Gaussian prior on $\mathbf{x}$.

A regularized extension of the GP-LVM which allows estimation of the dimension of the latent space was introduced by **?**, in which the latent variables and their intrinsic dimensionality are simultaneously optimized. The iWMM can also infer the intrinsic dimensionality of nonlinear manifolds: inferring the Gaussian covariance for each latent cluster allows the variance of irrelevant dimensions to become small. The marginal likelihood of the latent Gaussian mixture will favor using as few dimensions as possible to describe each cluster. In fact, because each latent cluster has a different set of parameters, each cluster can have a different effective dimension. This allows the iWMM to model manifolds of differing dimension in the observed space, as demonstrated in figure 1.4(b).

? considered several modifications of the GP-LVM which model the latent density using a mixture of Gaussians centered around the latent points. They approximated the observed density $p(\mathbf{Y})$ by another mixture of Gaussians, obtained by moment-matching the latent Gaussians after they had been warped into the observed space. Training was done by maximizing a leave-some-out predictive density. This method had poor predictive performance compared to simple baselines, and was not have a generative clustering model.

Related linear models

The iWMM can also be viewed as a generalization of the mixture of probabilistic principle component analyzers (**?**), or the mixture of factor analyzers (**?**), where the linear mapping of the mixtures is generalized to a nonlinear mapping by Gaussian processes, and the number of components is infinite.

Non-probabilistic methods

There exist non-probabilistic clustering methods which can find clusters with complex shapes, such as spectral clustering (**?**) and nonlinear manifold clustering (**??**). Spectral clustering finds clusters by first forming a similarity graph, then finding a low-dimensional latent representation using the graph, and finally, clustering the latent coordinates via k-means. The performance of spectral clustering depends on parameters which are usually set manually, such as the number of clusters, the number of neighbors, and the variance parameter used for constructing the similarity graph. The iWMM infers such parameters automatically, and has no need to construct a similarity graph.

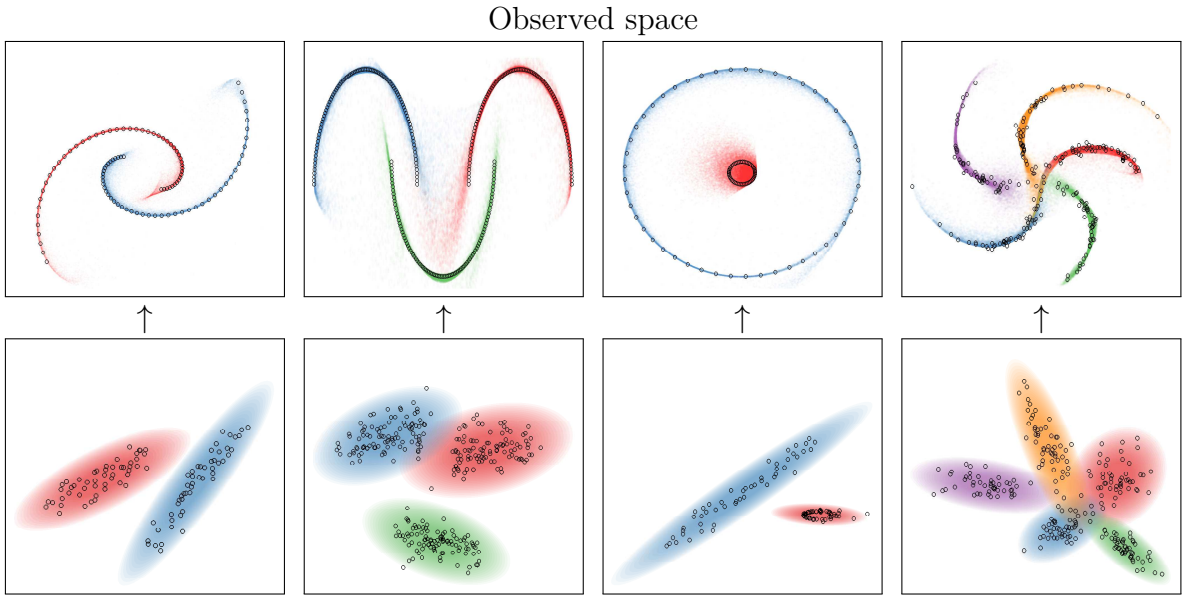
The kernel Gaussian mixture model (**?**) can also find non-Gaussian shaped clusters. This model estimates a GMM in the implicit infinite-dimensional feature space defined by the kernel mapping of the observed space. However, the kernel parameters must be set by cross-validation. In contrast, the iWMM infers the mapping function such that the latent coordinates will be well-modeled by a mixture of Gaussians.

Nonparametric cluster shapes

To the best of our knowledge, the only other Bayesian clustering method with nonparametric cluster shapes is that of **?**, who for one-dimensional data introduce a nonparametric model of *unimodal* clusters, where each cluster's density function strictly decreases away from its mode.

Deep Gaussian processes

An elegant way to construct a GP-LVM with a more structured latent density $p(\mathbf{x})$ is to use another GP-LVM to model the latent coordinates $\mathbf{X}$. This latent GP-LVM can have another GP-LVM modeling its latent space, etc. This is exactly the model class considered by **?**, who also test to what extent each layer's latent representation implicitly forms clusters. They found that when modeling MNIST hand-written digits, nearest-neighbour classification performed best in the 4th layer of a 5-layer deep nested GP-LVM, suggesting that the latent density might have been implicitly forming clusters at that level.



(a) 2-curve (b) 3-semi (c) 2-circle (d) Pinwheel

Figure 1.4 *Top row:* The observed unlabeled data points (black), and the cluster densities inferred by the iWMM (colors). *Bottom row:* Latent coordinates and Gaussian components from a single sample from the posterior. Each point in the latent space corresponds to a point in the observed space.

1.5 Experimental results

1.5.1 Synthetic datasets

Figure 1.4 demonstrates the proposed model on four synthetic datasets. None of these four datasets can be appropriately clustered by Gaussian mixture models (GMM). For example, consider the 2-curve data shown in Figure 1.4(a), where 100 data points lie in each of two curved lines in a two-dimensional observed space. A GMM with two components cannot separate the two curved lines, while a GMM with many components could separate the two lines only by breaking each line into many clusters. In contrast, the iWMM represents the two non-Gaussian clusters in the observed space by two Gaussian-shaped clusters in the latent space. Figure 1.4(b) shows a simple extension to three non-linearly-seperable clusters.

Figure 1.4(c) shows an interesting manifold learning challenge: a dataset consisting of two concentric circles. The outer circle is modeled in the latent space by a Gaussian with one effective degree of freedom. This linear topology is fit to the outer circle in the

observed space by bending the two ends until they cross over. In contrast, the sampler fails to discover the 1D topology of the inner circle, modeling it with a 2D manifold instead. This example demonstrates that each cluster in the iWMM manifold can have a different effective dimension.

Figure 1.4(d) shows a five-armed variant of the pinwheel dataset of **?**, generated by warping a mixture of Gaussians into a spiral. This generative process closely matches the assumptions of the iWMM. Unsurprisingly, the iWMM is able to recover an anologous latent structure, and its predictive density follows the observed data manifolds.

1.5.2 Clustering face images

We also examined our model's ability to model images without pre-processing. We constructed a dataset consisting of 50 greyscale 32x32 pixel images of two individuals from the UMIST faces dataset (**?**). Both series of images show a person turning his head to the right to varying degrees.

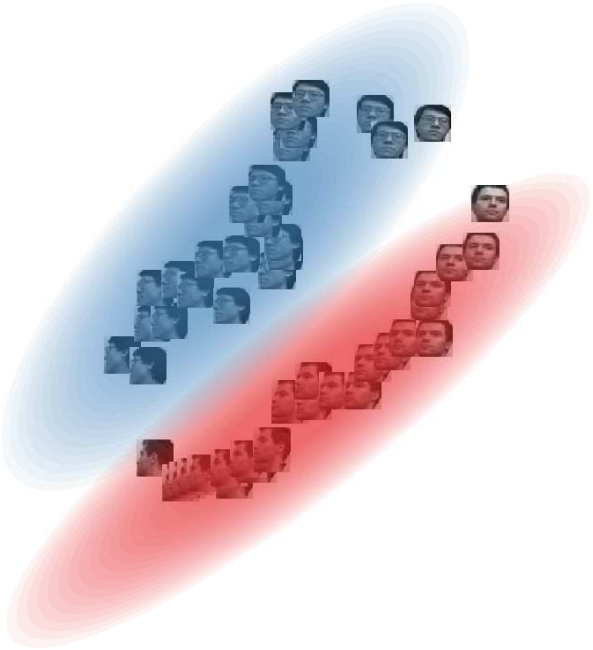


Figure 1.5 A sample from the 2-dimensional latent space when modeling a series of 32x32 face images. Images are rendered at their latent 2D coordinates. Our model correctly discovers that the data consists of two separate manifolds, both approximately one-dimensional, which both share the same head-turning structure.

Figure 1.5 shows a sample from the posterior over latent coordinates as well as the density model, with each image rendered at its location in the latent space. The model has recovered three relevant, interpretable features of the dataset: First, that there are two distinct faces. Second, that each set of images lies approximately along a smooth one-dimensional manifold. Third, that the two manifolds share roughly the same structure: the front-facing images of both individuals lie close to one another, as do the side-facing images.

1.5.3 Density estimation

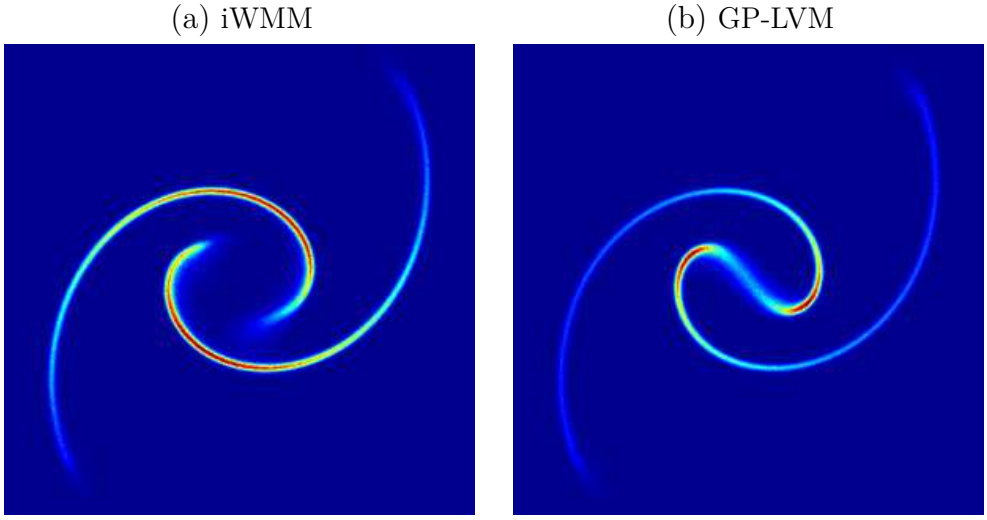
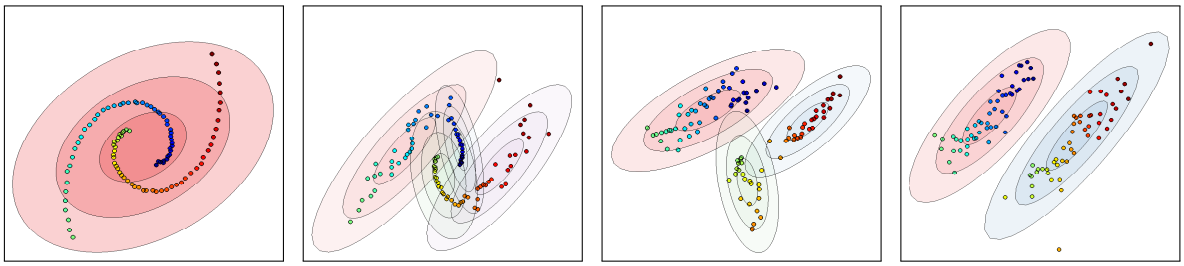


Figure 1.6 *Left:* Posterior density inferred by the iWMM in the observed space, on the 2-curve data. *Right:* Posterior density inferred by the iWMM with one component, a model equivalent to a fully-Bayesian GP-LVM.

Figure 1.6(a) shows the posterior density in the observed space inferred by the iWMM on the 2-curve data, computed using 1000 samples from the Markov chain. The two separate manifolds of high density implied by the two curved lines was recovered by the iWMM. Note also that the density along the manifold varies along with the density of data, shown in figure 1.4(a).

This result can be compared to a special case of our model with only a single Gaussian in the latent space, equivalnt to a fully-Bayesian GP-LVM. Figure 1.6(b) shows that the single-cluster variant of the iWMM posterior is forced to place significant density connecting the two clusters, since it has to reproduce the observed density manifold by warping a single Gaussian.



(a) Initialization (b) Iteration 500 (c) Iteration 1800 (d) Iteration 3000

Figure 1.7 The inferred infinite GMMs over iterations in the two-dimensional latent space with the iWMM using the 2-curve data. Labels indicate the number of iterations of the sampler, and the color of each point represents its ordering in the observed coordinates.

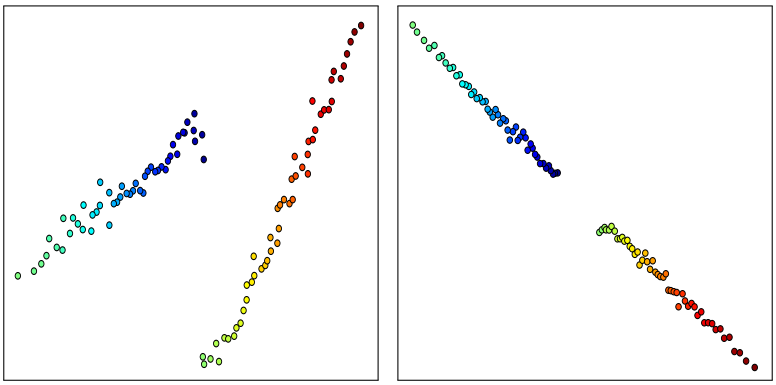
1.5.4 Mixing

An interesting side-effect of learning the number of latent clusters is that this added flexibility can help the sampler escape local minima. Figure 1.7 shows the samples of the latent coordinates and clusters of the iWMM over a single Markov chain modeling the 2-curve data. Figure 1.7(a) shows the latent coordinates initialized at the observed coordinates, starting with one latent component. After 500 iterations, each curved line was modeled by two components. After 1800 iterations, the left curved line was modeled by a single component. After 3000 iterations, the right curved line was also modeled by a single component, and the dataset was appropriately clustered. This configuration was relatively stable, and a similar state was found at the 5000th iteration.

1.5.5 Visualization

Next, we briefly investigate the potential of the iWMM for low-dimensional visualization of data. Figure 1.8(a) shows the latent coordinates obtained by averaging over 1000 samples from the posterior of the iWMM. Because rotating the latent coordinates does not change their probability, simple averaging may not be an adequate way to summarize the posterior. However, we show this result in order to show the characteristics of latent coordinates obtained by the iWMM. The estimated latent coordinates are clearly separated, and they form two straight lines. This result is an example of the iWMM recovering the original topology of the data before it was warped.

For comparison, Figure 1.8(b) shows the latent coordinates estimated by the iWMM when forced to use a single cluster (again, equivalent to a fully-Bayesian GP-LVM). In this case, the latent coordinates lie in two sections of a single straight line.



(a) iWMM (b) iWMM ($C = 1$)

Figure 1.8 Latent coordinates of the 2-curve data, estimated by two different methods.

Regardless of the dimension of the latent space, the iWMM will tend to model each cluster with as low-dimensional a Gaussian as possible, since a narrowly-shaped Gaussian will assign the latent coordinates much higher likelihood than a spherical Gaussian.

1.5.6 Clustering performance

We more formally evaluated the density estimation and clustering performance of the proposed model using four real datasets: iris, glass, wine and vowel, obtained from LIBSVM multi-class datasets (**?**), in addition to the four synthetic datasets shown above: 2-curve, 3-semi, 2-circle and pinwheel (**?**). The statistics of these datasets are summarized in Table 1.1. In each experiment, we show the results of ten-fold cross-validation.

Table 1.1 Statistics of the datasets used for evaluation.

	2-curve	3-semi	2-circle	pinwheel	iris	glass	wine	vowel
samples: N	100	300	100	250	150	214	178	528
dimension: D	2	2	2	2	4	9	13	10
num. clusters: C	2	3	2	5	3	7	3	11

Results in bold are not significantly different from the best performing method in each column according to a paired t-test.

Table 1.2 compares the clustering performance of the iWMM with the iGMM, quantified by the Rand index (**?**), which measures the correspondence between inferred clusters and true clusters. Since the manifold on which the observed data lies can be at most D-dimensional, we set the latent dimension Q equal to the observed dimension D in

Table 1.2 Average Rand index for evaluating clustering performance.

	2-curve	3-semi	2-circle	Pinwheel	Iris	Glass	Wine	Vowel
iGMM	0.52	0.79	0.83	0.81	0.78	0.60	0.72	**0.76**
iWMM(Q=2)	**0.86**	**0.99**	**0.89**	**0.94**	**0.81**	**0.65**	0.65	0.50
iWMM(Q=D)	**0.86**	**0.99**	**0.89**	**0.94**	0.77	0.62	**0.77**	**0.76**

iWMMs. We also included the $Q = 2$ case in an attempt to characterize how much modeling power is lost by forcing the latent representation to be visualizable.

These experiments were designed to measure the extent to which nonparametric cluster shapes helped to estimate meaningful clusters. To eliminate any differences due to different inference procedures, we used identical code for the iGMM and iWMM, the only difference being that the warping function was set to the identity $\mathbf{x} = \mathbf{y}$. Both variants of the iWMM usually outperformed the iGMM on this measure.

1.5.7 Density estimation

Next, we compared the iWMM in terms of predictive density against kernel density estimation (KDE) and the (iGMM). For KDE, the kernel width was estimated by maximizing the leave-one-out density.

Although all these methods are consistent density estimators, we expect the iWMM to have an advantage for finite datasets, since its density contours can follow the data density. Table 1.3 lists average test log likelihoods.

Table 1.3 Average test log-likelihoods for evaluating density estimation performance.

	2-curve	3-semi	2-circle	Pinwheel	Iris	Glass	Wine	Vowel
KDE	−2.47	−0.38	−1.92	−1.47	**−1.87**	1.26	−2.73	**6.06**
iGMM	−3.28	−2.26	−2.21	−2.12	−1.91	3.00	**−1.87**	−0.67
iWMM(Q=2)	**−0.90**	**−0.18**	**−1.02**	**−0.79**	**−1.88**	**5.76**	**−1.96**	**5.91**
iWMM(Q=D)	**−0.90**	**−0.18**	**−1.02**	**−0.79**	**−1.71**	**5.70**	−3.14	−0.35

The iWMM usually achieved higher test likelihoods than the KDE and the iGMM. The sometimes large differences between performance in the $D = 2$ case and the $D = Q$ case may be attributed to the fact that when the latent dimension is high, it requires many samples from the latent distribution to produce an accurate estimate of the posterior density at the test locations. This difficulty in inference might be solved by using a warping with back-constraints (**?**), that would allow a more direct evaluation of the density at a given point in the observed space.

Source code

Code to reproduce all the above figures and experiments is available at
http://www.github.com/duvenaud/warped-mixtures.

1.6 Conclusions

This chapter introduced a simple generative model of non-Gaussian density manifolds which can infer nonlinearly separable clusters, low-dimensional representations of varying dimension per cluster, and density estimates which smoothly follow the contours of each cluster. We then introduced a sampler for this model which integrates out both the cluster parameters and the warping function exactly.

Non-probabilistic methods such as spectral clustering can also produce nonparametric cluster shapes, but usually lack principled methods for setting kernel parameters, the number of clusters, and the implicit dimension of the learned manifolds, other than by cross-validation. This chapter shows that using a fully generative model allows most model choices to be determined automatically.

Many methods have been proposed which can perform some combination of clustering, manifold learning, density estimation and visualization. We demonstrated that a simple but flexible probabilistic generative model can perform well at all these tasks.

1.7 Future work

More sophisticated latent density models

The Dirichlet process mixture of Gaussians in the latent space of our model could easily be replaced by a more sophisticated density model, such as a hierarchical Dirichlet process (**?**), or a Dirichlet diffusion tree (**?**). Another straightforward extension of our model would be making inference more scalable by using sparse Gaussian processes (**??**) or more advanced Hamiltonian Monte Carlo methods (**?**).

A finite cluster count model

? note that the Dirichlet process assumes infinitely many clusters, and that estimates of the number of clusters in a dataset based on Bayesian inference are inconsistent under this model. They propose a consistent alternative which also allows efficient Gibbs sampling, called the mixture of finite mixtures. Replacing the Dirichlet process with

a mixture of finite mixtures would improve the conistency properties of the iWMM in most applications.

Semi-supervised learning

An interesting but more complex extension of the iWMM would be a semi-supervised version of the model. The iWMM could allow label propagation along regions of high density in the latent space, even if the individual points in those regions are stretched far apart along low-dimensional manifolds in the observed space. Another natural extension would be to allow a separate warping for each cluster, producing a mixture of warped Gaussians, rather than a warped mixture of Gaussians.

Learning the topology of data manifolds

Some datasets naturally live on manifolds which are not simply connected. For example, motion capture data or video of a person walking in a circle naturally lives on a torus, with one coordinate specifying the phase of the person's step, and another specifying how far around the circle they've walked.

As shown in **??**, using structured kernels to specify the warping of a latent space gives rise to interesting topologies on the observed density manifold. If a suitable method for computing the marginal likelihood of a GP-LVM is available, an automatic search similar to that described in section 1.7 would be possible, automatically finding the topology of the data manifold.