Chapter 1

Warped Mixture Models

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

Previous chapters showed how the probabilistic nature of GPs sometimes allows the automatic determination of the appropriate structure when building models of functions. One can also take advantage of this property when composing GPs with other models, automatically trading-off complexity between 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 (density manifolds) having arbitrary nonparametric shapes. We call the proposed model the *infinite warped mixture model* (iWMM). 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 Iwata et al. (2013). The main idea was born out of a conversation between Tomoharu and myself, and together we wrote almost all of the code as well as the paper. Tomoharu ran most of the experiments, and Zoubin Ghahramani provided guidance and many helpful suggestions throughout the project.

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) (Lawrence, 2004), a probabilistic model of nonlinear manifolds. The GP-LVM

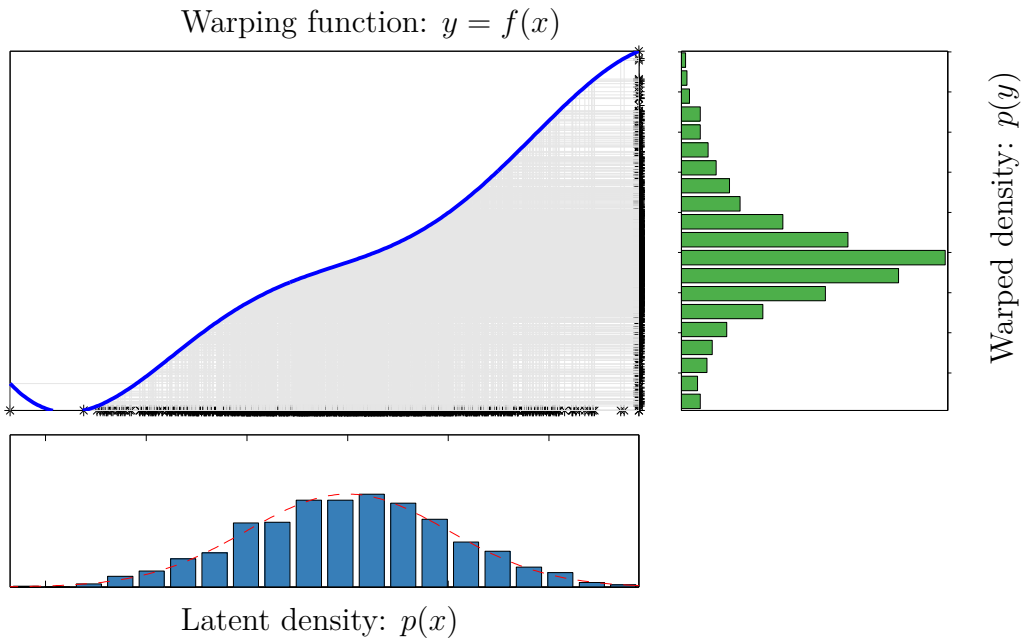


Figure 1.1: A draw from a one-dimensional Gaussian process latent variable model. *Bottom:* the density of a set of samples from a 1D Gaussian specifying the distribution $p(x)$ in the latent space. *Top left:* A function $y = f(x)$ drawn from a GP prior. Grey lines show points being mapped through f.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

smoothly warps a Gaussian density into a more complicated distribution, using a draw from a GP. Usually, we say that the Gaussian density is defined in a "latent space" having Q dimensions, and the warped density is defined in the "observed space" having D dimensions.

A generative definition of the GP-LVM is:

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

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

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

Under the GP-LVM, the probability of observations $\mathbf{Y}$ given the latent coordinates

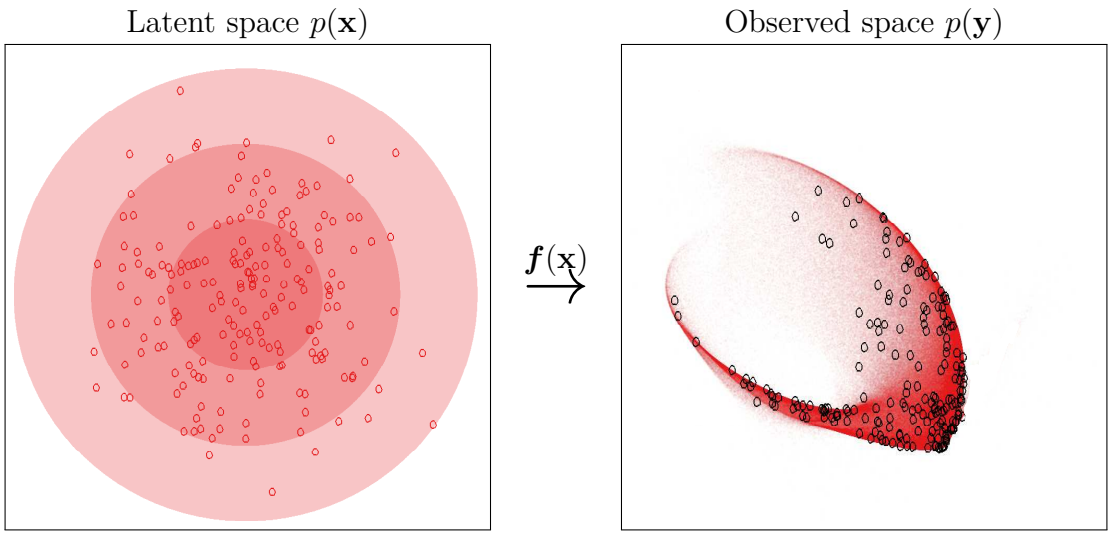


Figure 1.2: A draw from a two-dimensional Gaussian process latent variable model. *Left:* Isocontours and samples from a 2D Gaussian, specifying the distribution $p(\mathbf{x})$ in the latent space. *Right:* The observed density $p(\mathbf{y})$ has a nonparametric shape, defined by warping the latent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

$\mathbf{X}$, integrating over the mapping functions $\boldsymbol{f}$ is simply a product of GP likelihoods:

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

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

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

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. One can also approximately integrate out $\mathbf{X}$, which is the approach taken in this chapter.

1.2 The infinite warped mixture model

This section defines the infinite warped mixture model (iWMM). Like the GP-LVM, the iWMM assumes a smooth nonlinear mapping from a latent density to an observed density. The only difference is that the iWMM assumes that the latent density is an

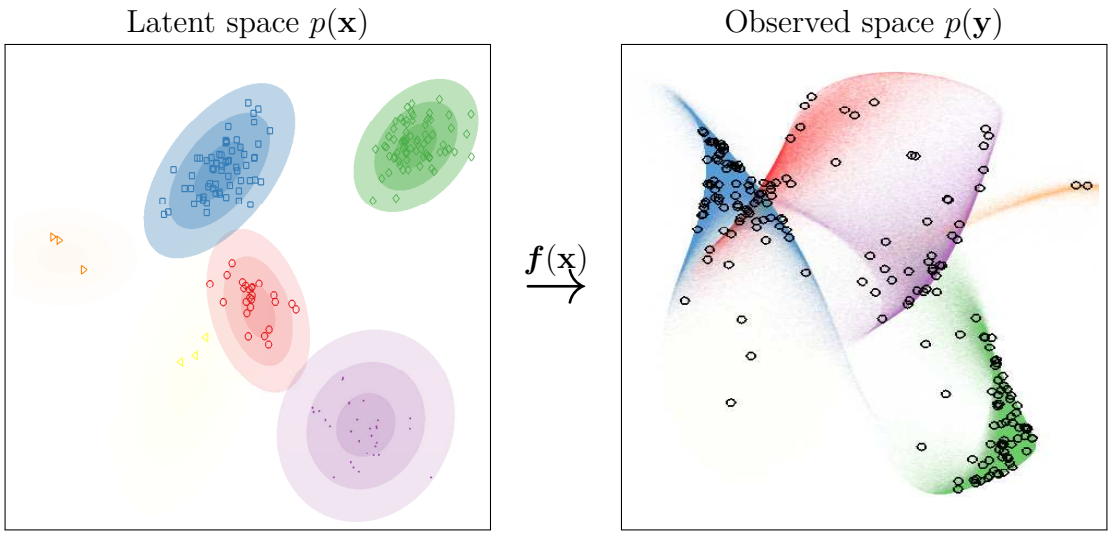


Figure 1.3: A sample from the iWMM prior. *Left:* In the latent space, a mixture distribution is sampled from a Dirichlet process mixture of Gaussians. *Right:* The latent mixture is smoothly warped to produce a set of non-Gaussian manifolds in the observed space.

infinite Gaussian mixture model (iGMM) (Rasmussen, 2000):

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

where λ_c, $\boldsymbol{\mu}_c$ and $\mathbf{R}_c$ denote 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: The iWMM with a single fixed spherical Gaussian density on the latent coordinates $p(\mathbf{x})$ corresponds to the GP-LVM, while the iWMM with fixed mapping $\mathbf{y} = \mathbf{x}$ and $Q = D$ corresponds to the iGMM.

If the clusters being modeled do not happen to have Gaussian shapes, a flexible model of cluster shapes is required to correctly estimate the number of clusters. 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 $\boldsymbol{\theta}$, one can analytically integrate over functions mapping

$\mathbf{X}$ to $\mathbf{Y}$. However, inference becomes more difficult when one introduces uncertainty about the kernel parameters or the input locations $\mathbf{X}$. This section outlines how to compute approximate posterior distributions over all parameters in the iWMM given only a set of observations $\mathbf{Y}$. Further details can be found in appendix **??**.

We first place conjugate priors on the parameters of the Gaussian mixture components, allowing analytic integration over latent 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 a sample of the latent points $\mathbf{X}$, sample the discrete cluster memberships $\mathbf{z}$ using collapsed Gibbs sampling, integrating out the iGMM parameters (**??**).

2. Given the cluster assignments $\mathbf{z}$, sample the continuous latent coordinates $\mathbf{X}$ and kernel parameters $\boldsymbol{\theta}$ using Hamiltonian Monte Carlo (HMC) (MacKay, 2003, 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 (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006).

Posterior predictive density

One disadvantage of the GP-LVM is that its predictive density has no closed form, and the iWMM inherets this problem. To approximate the predictive density, we first sample latent points, then sample warpings of those points into the observed space. The Gaussian noise added to each observation by the WN kernel component means that each sample adds a Gaussian to the Monte Carlo estimate of the predictive density. Details can be found in appendix **??**. 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 (Lawrence, 2004; Lawrence and Urtasun, 2009; Salzmann et al., 2008). The latent positions $\mathbf{X}$ in the GP-LVM are typically obtained by maximum a posteriori estimation or variational Bayesian inference (Titsias and Lawrence, 2010), placing a single fixed spherical Gaussian prior on $\mathbf{x}$.

A regularized extension of the GP-LVM that allows estimation of the dimension of the latent space was introduced by Geiger et al. (2009), in which the latent variables and their intrinsic dimensionality were simultaneously optimized. The iWMM can also infer the intrinsic dimensionality of nonlinear manifolds: the Gaussian covariance parameters for each latent cluster allow 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. Because each latent cluster has a different set of parameters, each cluster can have a different effective dimension in the observed space, as demonstrated in figure 1.4(c).

Nickisch and Rasmussen (2010) 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 a second mixture of Gaussians, obtained by moment-matching the density obtained by warping each latent Gaussian into the observed space. Because their model was not generative, training was done by maximizing a leave-some-out predictive density. This method had poor predictive performance compared to simple baselines.

Related linear models

The iWMM can also be viewed as a generalization of the mixture of probabilistic principle component analyzers (Tipping and Bishop, 1999), or the mixture of factor analyzers (Ghahramani and Beal, 2000), where the linear mapping is replaced by a draw from a GP, 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 (Ng et al., 2002) and nonlinear manifold clustering (Cao and Haralick, 2006; Elhamifar and Vidal, 2011). Spectral clustering finds clusters by first forming a similarity graph, then finding a low-dimensional latent rep-

resentation 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 (Wang et al., 2003) 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 Rodríguez and Walker (2012), who for one-dimensional data introduce a nonparametric model of *unimodal* clusters, where each cluster's density function strictly decreases away from its mode.

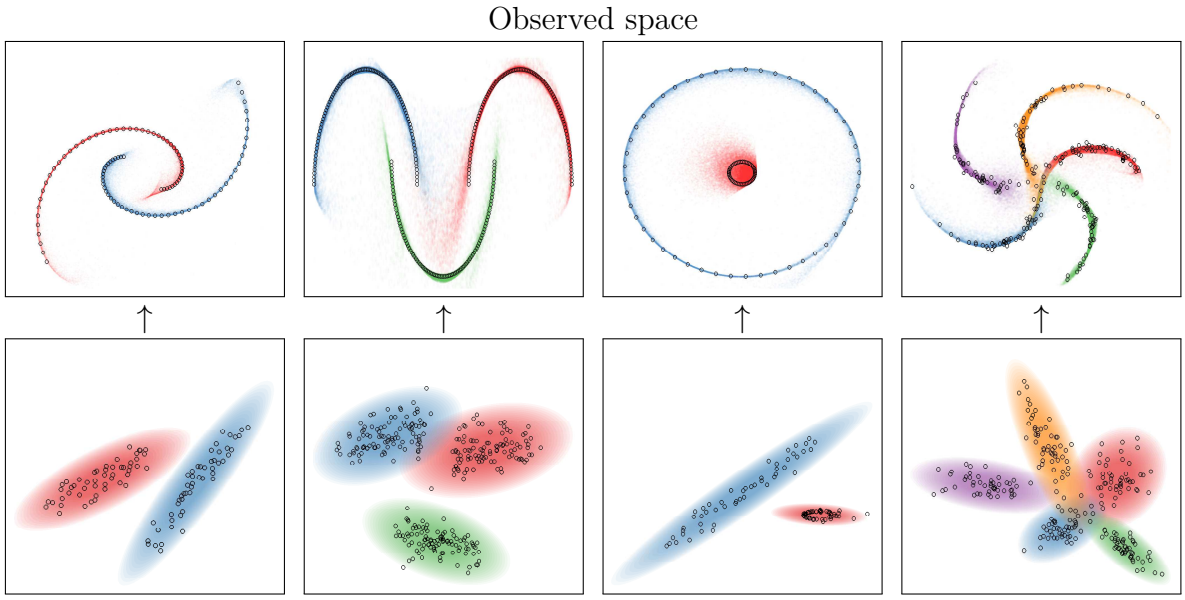
Deep Gaussian processes

An elegant way to construct a GP-LVM having a more structured latent density $p(\mathbf{x})$ is to use a second GP-LVM to model the latent coordinates $\mathbf{X}$. This latent GP-LVM can have a third GP-LVM modeling its latent density, etc. This model class was considered by Damianou and Lawrence (2013), who also tested to what extent each layer's latent representation grouped points having the same label. 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.

1.5 Experimental results

1.5.1 Synthetic datasets

Figure 1.4 demonstrates the proposed model on four synthetic datasets. None of these 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



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

Figure 1.4: *Top row:* Observed unlabeled data points (black), and 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.

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 similar three-cluster example.

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 can have a different effective dimension.

Figure 1.4(d) shows a five-armed variant of the pinwheel dataset of Adams and Ghahramani (2009), 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 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 (Graham and Allinson, 1998). Both series of images show a person turning his head to the right.

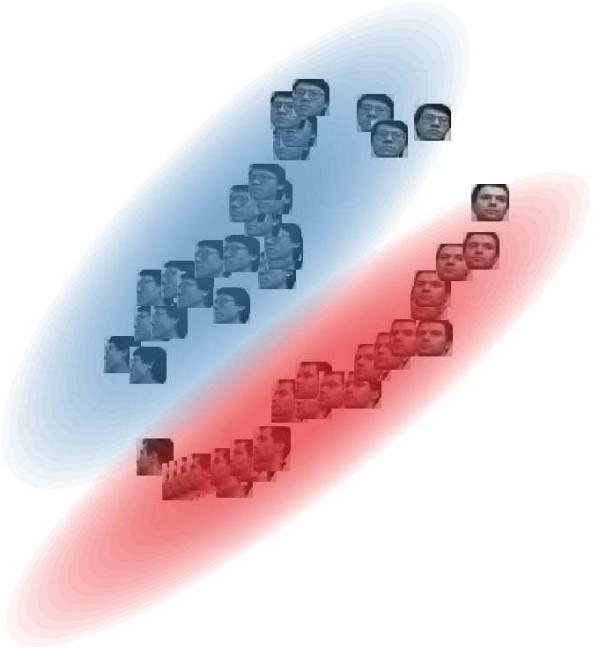


Figure 1.5: A sample from the 2-dimensional latent space modeling a series of face images. Images are rendered at their latent 2D coordinates. The iWMM reports that the data consists of two separate manifolds, both approximately one-dimensional, which both share the same head-turning structure.

Figure 1.5 shows a sample from the posterior over latent coordinates and density, with each image rendered at its location in the latent space. The model has recovered three 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(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 seperation

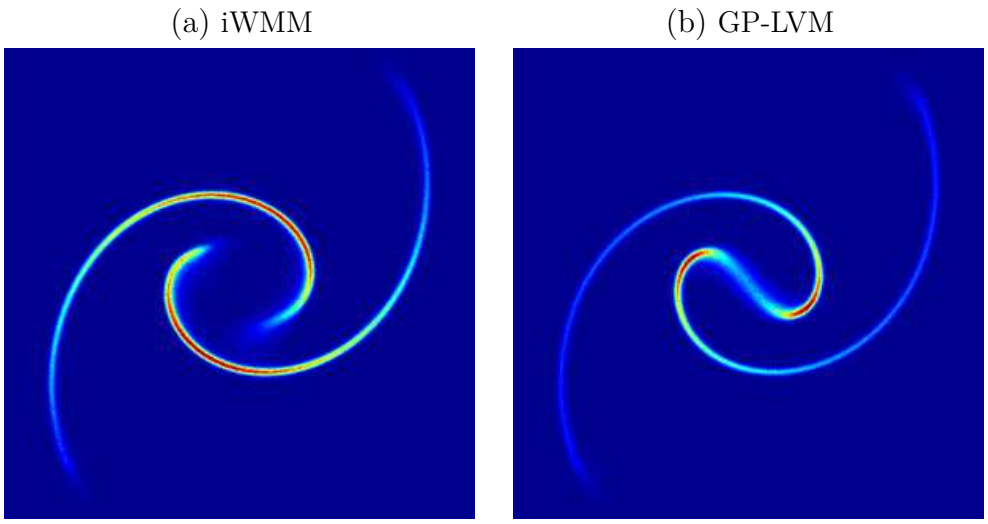


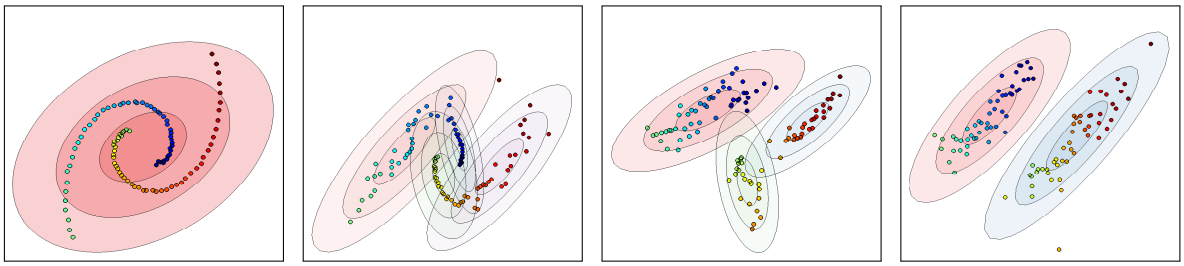
Figure 1.6: *Left:* Posterior density inferred by the iWMM in the observed space, on the 2-curve data. *Right:* Posterior density inferred by the iWMM with one component, a model equivalent to a fully-Bayesian GP-LVM.

of the density into two unconnected manifolds was recovered by the iWMM. Also note 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 fully-Bayesian GP-LVM, equivalent to a special case of our model having only a single Gaussian in the latent space. Figure 1.6(b) shows that the GP-LVM places significant density connecting the two clusters, since it has to reproduce the observed density manifold by warping a single Gaussian.

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.

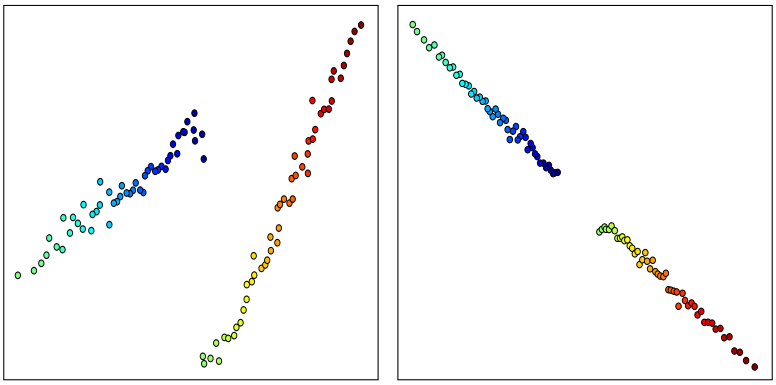


(a) Initialization (b) Iteration 500 (c) Iteration 1800 (d) Iteration 3000

Figure 1.7: Latent coordinates and densities of the iWMM, plotted throughout one run of a Markov chain.

1.5.5 Visualization

Next, we briefly investigate the utility of the iWMM for low-dimensional visualization of data. Figure 1.8(a) shows the latent coordinates obtained by averaging over 1000



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

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

samples from the posterior of the iWMM. The estimated latent coordinates are clearly separated, forming 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 fully-Bayesian GP-LVM, in which case the latent coordinates lie in two sections of a single straight line.

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 (Chang and Lin, 2011), in addition to the four synthetic datasets shown above: 2-curve, 3-semi, 2-circle and pinwheel (Adams and Ghahramani, 2009). The statistics of these datasets are summarized in Table 1.1. For each experiment,

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

we show the results of ten-fold cross-validation. Results in bold are not significantly different from the best performing method in each column according to a paired t-test.

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**

Table 1.2 compares the clustering performance of the iWMM with the iGMM, quantified by the Rand index (Rand, 1971), 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. 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 $\mathbf{y} = \mathbf{x}$. 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), the iGMM, and the fully-Bayesian GP-LVM. For KDE, the kernel width was estimated by maximizing the leave-one-out 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
GP-LVM(Q=2)	−1.02	−0.36	**−0.78**	**−0.78**	−1.91	**5.70**	**−1.95**	**6.04**
GP-LVM(Q=D)	−1.02	−0.36	**−0.78**	**−0.78**	−1.86	**5.59**	−2.89	−0.29
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 GP-LVM performed competitively with the iWMM, although it never significantly outperformed the iWMM having the same latent dimension.

The sometimes large differences between performance in the $D = 2$ case and the $D = Q$ case of these two methods may be attributed to the fact that when the observed dimension is high, many samples are required from the latent distribution to produce accurate estimates of the posterior predictive density at the test locations. This difficulty might be resolved by using a warping with back-constraints (Lawrence, 2006), which 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 also introduced a sampler for this model which integrates out both the cluster parameters and the warping function exactly at each step.

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 showed 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 (Teh et al., 2006), or a Dirichlet diffusion tree (Neal, 2003). Another straightforward extension of our model would be making inference more scalable by using sparse Gaussian processes (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006) or more advanced Hamiltonian Monte Carlo methods (Zhang and Sutton, 2011).

A finite cluster count model

Miller and Harrison (2013) 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 still allows efficient Gibbs sampling, called the mixture of finite mixtures. Replacing the Dirichlet process with a mixture of finite mixtures could improve the consistency properties of the iWMM.

Semi-supervised learning

A straightforward 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 are.

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 applicable, automatically discovering the topology of the data manifold.

References

Ryan P. Adams and Zoubin Ghahramani. Archipelago: Nonparametric Bayesian semi-supervised learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1–8. ACM, 2009. (pages 8 and 12)

Wenbo Cao and Robert Haralick. Nonlinear manifold clustering by dimensionality. In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 920–924. IEEE, 2006. (page 6)

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. (page 12)

Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013. (page 7)

Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In *Advances in Neural Information Processing Systems*, pages 55–63, 2011. (page 6)

Andreas Geiger, Raquel Urtasun, and Trevor Darrell. Rank priors for continuous non-linear dimensionality reduction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 880–887. IEEE, 2009. (page 6)

Zoubin Ghahramani and M.J. Beal. Variational inference for Bayesian mixtures of factor analysers. *Advances in Neural Information Processing Systems*, 12:449–455, 2000. (page 6)

Daniel B Graham and Nigel M Allinson. Characterizing virtual eigensignatures for general purpose face recognition. *Face Recognition: From Theory to Applications*, 163:446–456, 1998. (page 9)

Tomoharu Iwata, David Duvenaud, and Zoubin Ghahramani. Warped mixtures for nonparametric cluster shapes. Bellevue, Washington, July 2013. (page 1)

Neil D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, pages 329–336, 2004. (pages 1 and 6)

Neil D. Lawrence. Local distance preservation in the GP-LVM through back constraints. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 513–520, 2006. (page 13)

Neil D. Lawrence and Raquel Urtasun. Non-linear matrix factorization with Gaussian processes. In *Proceedings of the 26th International Conference on Machine Learning*, pages 601–608, 2009. (page 6)

David J.C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University press, 2003. (page 5)

Jeffrey W. Miller and Matthew T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems 26*, pages 199–206. Curran Associates, Inc., 2013. (page 14)

Radford M. Neal. Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629, 2003. (page 14)

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856, 2002. (page 6)

Hannes Nickisch and Carl E. Rasmussen. Gaussian mixture modeling with Gaussian process latent variable models. *Pattern Recognition*, pages 272–282, 2010. (page 6)

Joaquin Quiñonero-Candela and Carl E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6: 1939–1959, 2005. (pages 5 and 14)

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, pages 846–850, 1971. (page 12)

Carl E. Rasmussen. The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 2000. (page 4)

Carlos E. Rodríguez and Stephen G. Walker. Univariate Bayesian nonparametric mixture modeling with unimodal kernels. *Statistics and Computing*, pages 1–15, 2012. (page 7)

Mathieu Salzmann, Raquel Urtasun, and Pascal Fua. Local deformation models for monocular 3D shape recovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. (page 6)

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 2006. (pages 5 and 14)

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. (page 14)

Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999. (page 6)

Michalis Titsias and Neil D. Lawrence. Bayesian Gaussian process latent variable model. *International Conference on Artificial Intelligence and Statistics*, 2010. (page 6)

Jingdong Wang, Jianguo Lee, and Changshui Zhang. Kernel trick embedded Gaussian mixture model. In *Algorithmic Learning Theory*, pages 159–174. Springer, 2003. (page 7)

Yichuan Zhang and Charles A. Sutton. Quasi-Newton methods for Markov chain Monte Carlo. *Advances in Neural Information Processing Systems*, pages 2393–2401, 2011. (page 14)