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 right kind, when building models of functions. In this chapter, we'll show that we can also take advantage of this property when we compose GPs with other models. Besides simply learning the right amount of complexity of the function modeled by a GP, we can automatically trade off complexity between the GP and the other parts of the model.

We'll consider a simple example: a Gaussian mixture model, warped by a draw from a GP. This model will produce clusters with arbitrary shapes, depending on the warping. More generalliy, this model 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. 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 work comprising the bulk of this chapter was done in collaboration with Tomoharu Iwata and Zoubin Ghahramani, and appeared in (Iwata et al., 2013). Specifically, the main idea was borne out of a conversation between Tomo and myself, and together we wrote almost all of the code together as well as the paper. Tomo ran most of the experiments. Zoubin Ghahramani provided guidance and many helpful suggestions throughout the project.

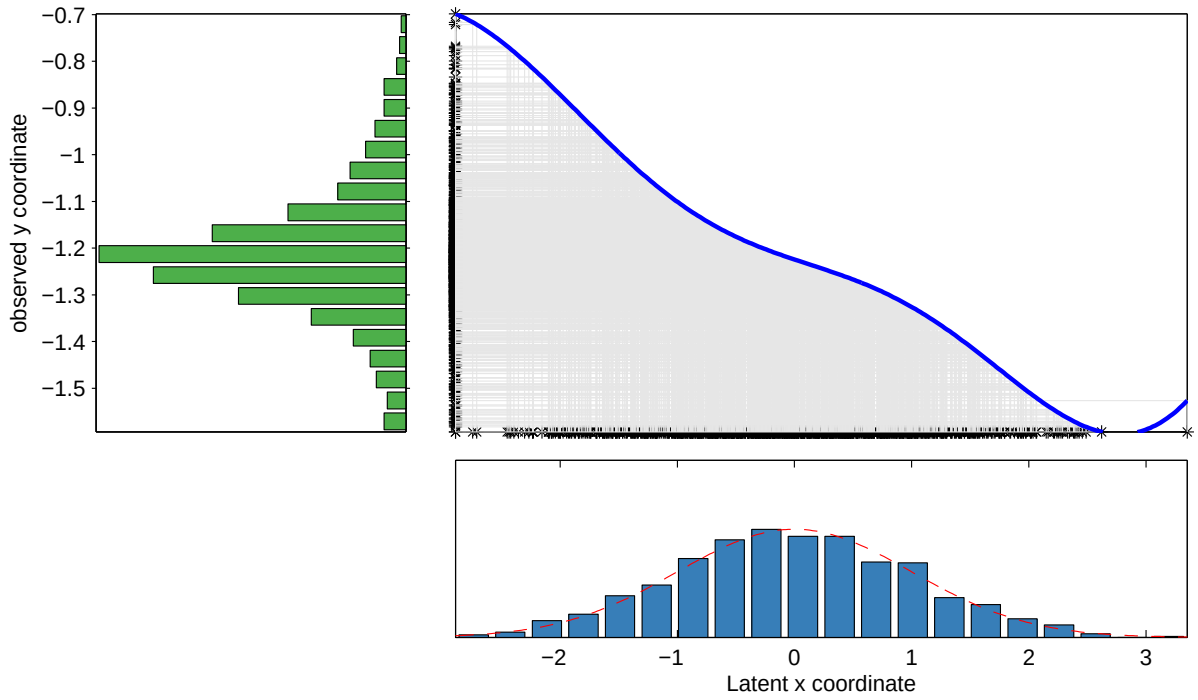


Fig. 1.1 A visual representation of the Gaussian process latent variable model (GP-LVM). *Bottom:* the density of a set of samples from a 1D Gaussian, specifying the distribution $p(\mathbf{X})$ in the latent space. *Top Right:* A function drawn from a GP prior. *Left:* A nonparametric density defined by warping the latent density through the sampled function.

1.1 The Gaussian Process Latent Variable Model

Besides being useful for modeling functions, a simple extension allows GPs to be useful for general density modeling, at the cost of inference becoming non-analytic. The GP-LVM can also be thought of as a method for modeling the covariance between rows of Y, using a number of parameters which grows only linearly with N.

The iWMM can be viewed as an extension to the 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.

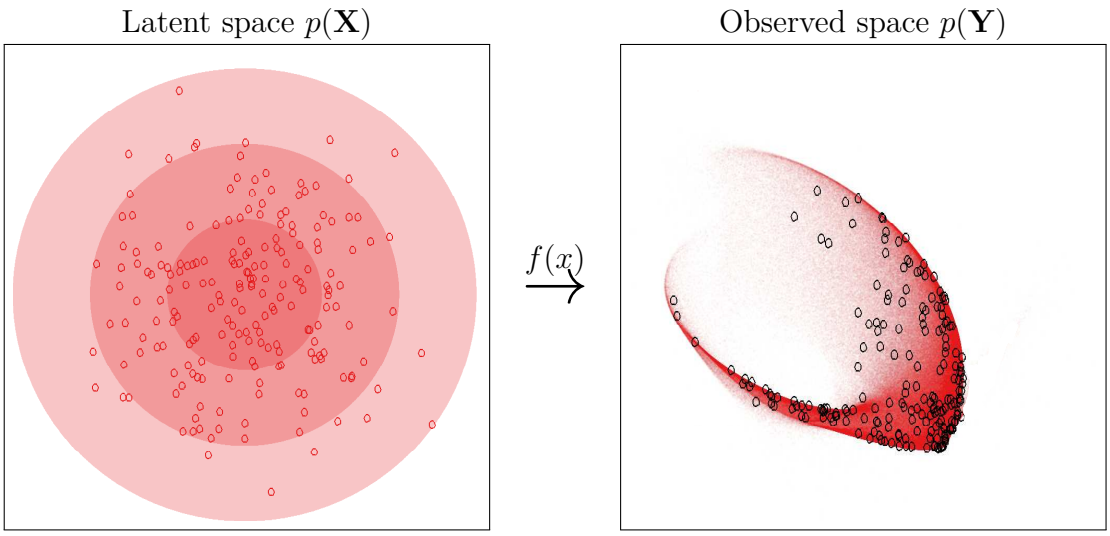


Fig. 1.2 A visual representation of the Gaussian process latent variable model. *Left:* Isocontours and samples from a 2D Gaussian, specifying the distribution $p(\mathbf{X})$ in the latent space. *Right:* Density and samples from a nonparametric density defined by warping the latent density through a function drawn from a GP prior.

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

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

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

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

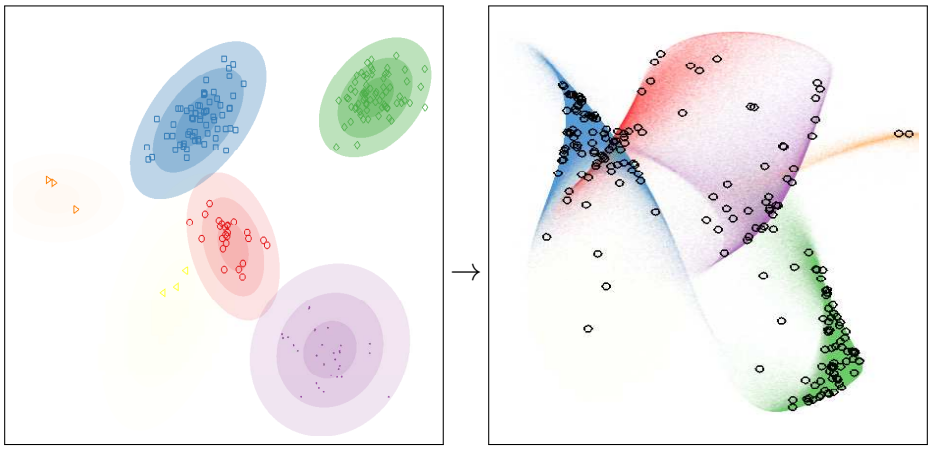
Under the GP-LVM, the probability of observations given the latent coordinates integrating out the mapping functions, is simply a product of multivariate normals:

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

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

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'll integrate out the latent coordinates, and place a more flexible parameterization on $p(\mathbf{x})$ than a single isotropic Gaussian. Just as the



Latent space Observed space

Fig. 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 non-Gaussian manifolds in the observed space.

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

In this section, we define in detail the infinite warped mixture model (iWMM). Just 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 infinite mixture of Gaussians (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 $\mathbf{R}_c$ 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 infinite Gaussian mixture model (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.

The need for a clustering model that doesn't have a fixed set of cluster shapes is motivated by fact that 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. If we care about getting the number of clusters right, then we need a flexible model of cluster shapes.

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, if we introduce uncertainty about any of these parameters, or add extra latent variables, then there's usually no analytic method to integrate out these extra parameters. In this section, we'll outline how we can infer all the parameters in the iWMM given only a set of observations $\mathbf{Y}$. Details can be found in **??**.

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 cluster memberships $\mathbf{Z}$ using collapsed Gibbs sampling, integrating out the mixture parameters. The conditional probability of $\mathbf{Z}$ given $\mathbf{X}$ is given by (**??**).

2. Given the cluster assignments $\mathbf{Z}$, we sample the latent coordinates $\mathbf{X}$ and kernel parameters $\boldsymbol{\theta}$ using Hamiltonian Monte Carlo.

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 such as (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006).

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, 1.5 and 1.7.

1.4 Related Work

The GP-LVM is effective as a nonlinear latent variable model 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 prior which penalizes a high-dimensional latent space was introduced by Geiger et al. (2009), 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. Because each latent cluster has a different set of parameters, the effective dimension of each cluster can vary, allowing manifolds of differing dimension in the observed space. This ability is demonstrated in figure 1.5b.

The iWMM can also be viewed as a generalization of the mixture of probabilistic principle component analyzers (Tipping and Bishop, 1999), or mixture of factor analyzers (Ghahramani and Beal, 2000), where the linear mapping of the mixtures is generalized to a nonlinear mapping by Gaussian processes, and number of components is infinite.

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 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. In contrast, the iWMM infers such parameters automatically. One of the main advantages of the iWMM over these methods is that there is 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 high-dimensional feature space defined by the kernel mapping of the observed space. However, the kernel GMM

uses a fixed nonlinear mapping function, with no guarantee that the latent points will be well-modeled by a GMM. In contrast, the iWMM infers the mapping function such that the latent co-ordinates will be well-modeled by a mixture of Gaussians.

1.5 Experimental Results

1.6 sec:iwmm-experiments

1.6.1 Clustering Faces

We first 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 capture a person turning his head to the right.

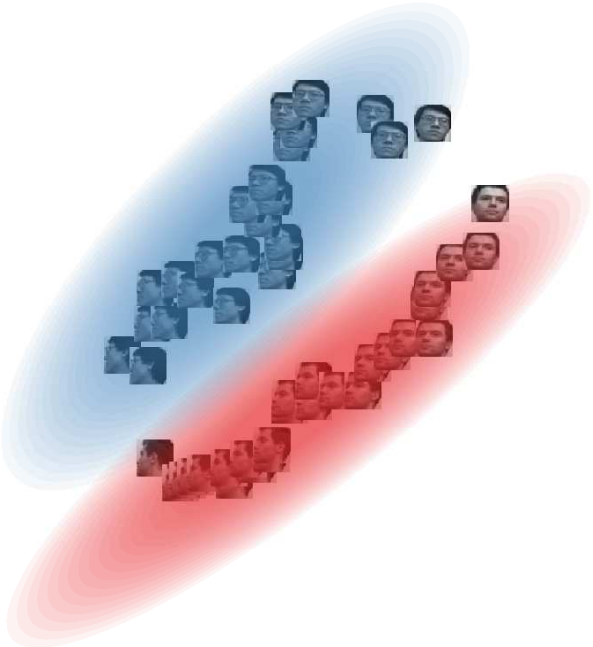


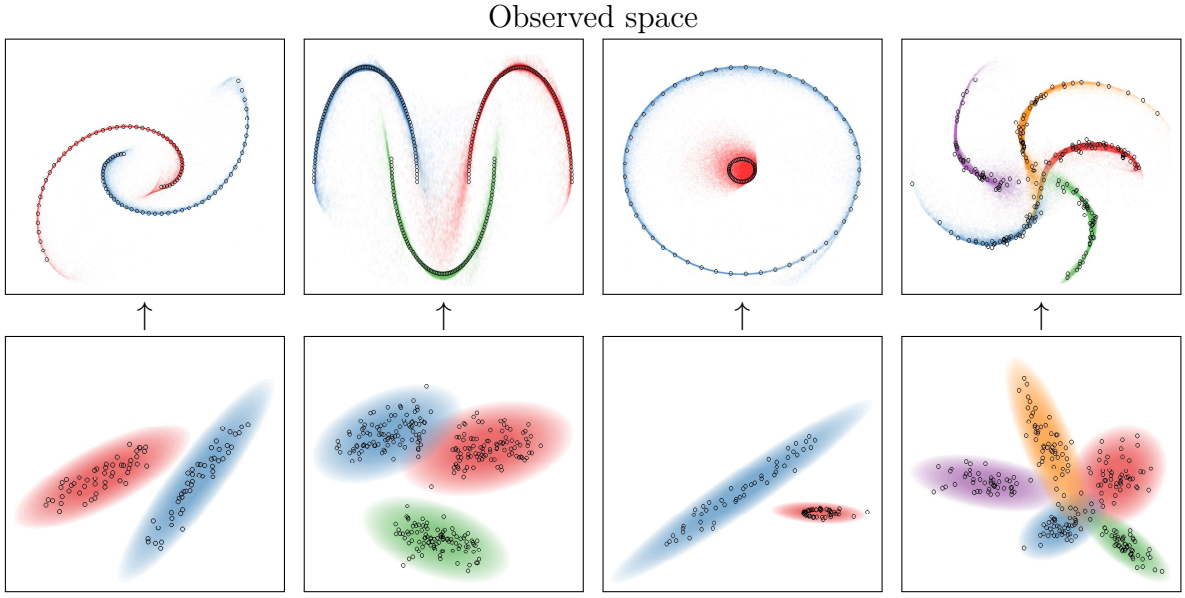
Fig. 1.4 A sample from the 2-dimensional latent space when modeling a series of 32x32 face images. Our model correctly discovers that the data consists of two separate manifolds, both approximately one-dimensional, which share the same head-turning structure.

Figure 1.4 shows a sample from the posterior over the latent coordinates, as well as the density model. 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.

1.6.2 Synthetic Datasets

Next, we demonstrate the proposed model on the four synthetic datasets shown in Figure 1.5. None of these four datasets can be appropriately clustered by Gaussian



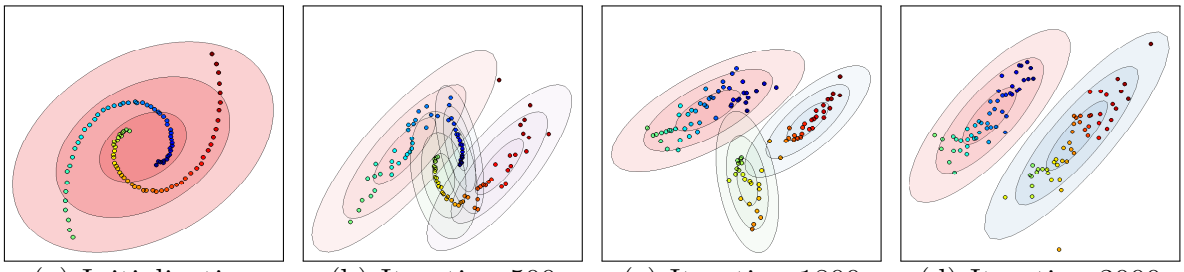
(a) 2-curve (b) 3-semi (c) 2-circle (d) Pinwheel

Fig. 1.5 Top row: The observed, unlabeled data points, and the clusters inferred by the iWMM. Bottom row: Latent coordinates and Gaussian components, shown for a single sample from the posterior. Each point in the latent space corresponds to a point in the observed space. This figure is best viewed in color.

mixture models (GMM). For example, consider the 2-curve data shown in Figure 1.5 (a), where 100 data points lie in one 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, with the iWMM, the two non-Gaussian-shaped clusters in the observed space were represented by two Gaussian-shaped clusters in the latent space, as shown at the bottom row of Figure 1.5 (a). The iWMM separated the two curved lines by nonlinearly warping two Gaussians from the latent space to the observed

space.

Figure 1.5 (c) shows an interesting manifold learning challenge: a dataset consisting of two circles. The outer circle is modeled in the latent space by a Gaussian with effectively one degree of freedom. This linear topology fits the outer circle in the observed space by bending the two ends until they overlap. 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.



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

Fig. 1.6 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.

1.6.3 Mixing

An interesting side-effect of learning the number of latent clusters is that this added flexibility can help the sampler escape local minima, helping the sampler to mix properly. Figure 1.6 shows the samples of the latent coordinates and clusters of the iWMM over time, when modeling the 2-curve data. 1.6(a) shows the latent coordinates initialized at the observed coordinates, starting with one latent component. At the 500th iteration 1.6(b), each curved line is modeled by two components. At the 1800th iteration 1.6(c), the left curved line is modeled by a single component. At the 3000th iteration 1.6(d), the right curved line is also modeled by a single component, and the dataset is appropriately clustered. This configuration was relatively stable, and a similar state was found at the 5000th iteration.

(a) iWMM (b) GP-LVM

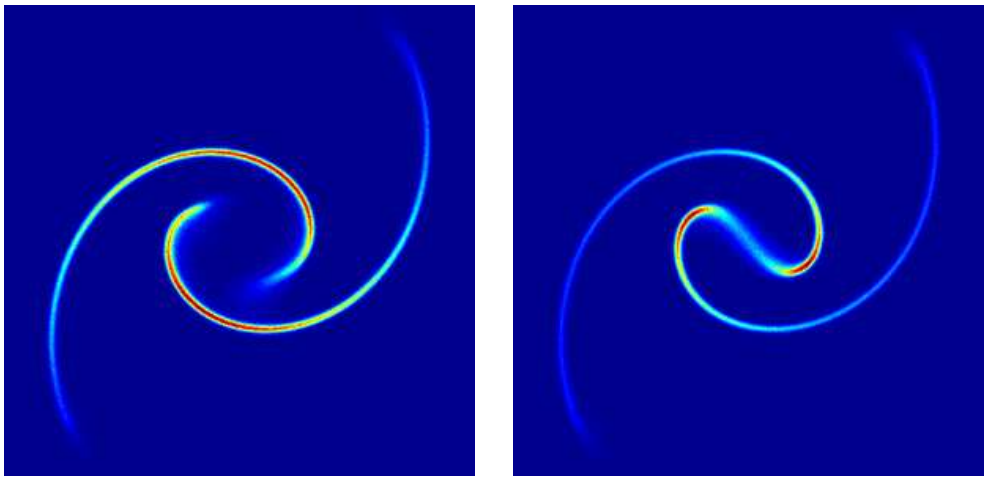


Fig. 1.7 *Left:* The posterior density in the observed space with the 2-curve data inferred by the iWMM. *Right:* The posterior inferred by the iWMM with one component, a model equivalent to the GP-LVM.

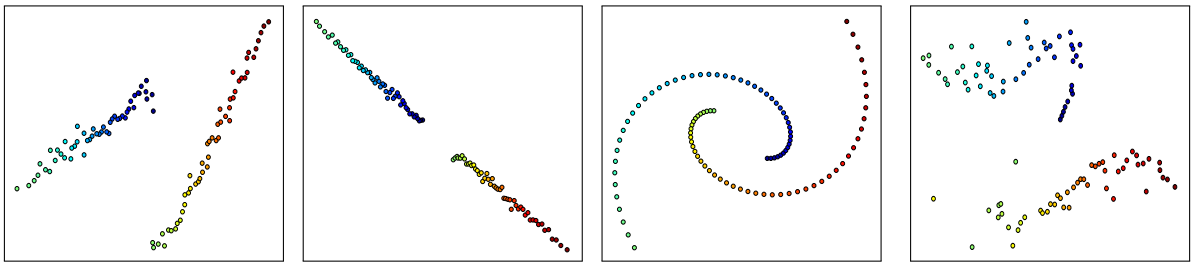
1.6.4 Density Estimation

Figure 1.7(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 with the density of data shown in Figure 1.5(a).

This result can be compared to a special case of our model, which uses only a single Gaussian to model the latent coordinates instead of an infinite GMM. Figure 1.7(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.

1.6.5 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, 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 sep-



(a) iWMM (b) iWMM ($C = 1$) (c) GP-LVM (d) Bayesian GP-LVM

Fig. 1.8 Latent coordinates of the 2-curve data, estimated by four different methods.

arated, and they form two straight lines. This result indicates that in some cases, the iWMM can recover the topology of the data before it has been warped into a manifold. For comparison, Figure 1.8 (b) shows the latent coordinates estimated by the iWMM when forced to use a single cluster: the latent coordinates lie in two sections of a single straight line. Figure 1.8 (c) and (d) show the latent coordinates estimated by the GP-LVM when optimizing or integrating out the latent coordinates, respectively. Recall that the iWMM ($C = 1$) is a more flexible model than the GP-LVM, since the GP-LVM enforces a spherical covariance in the latent space. These methods did not unfold the two curved lines, since the effective dimension of their latent representation is fixed beforehand. In contrast, the iWMM effectively formed a low-dimensional representation in the latent space.

Regardless of the dimension of the latent space, the iWMM will tend to model each cluster with as low-dimensional a Gaussian as possible. This is because, if the data in a cluster can be made to lie in a low-dimensional plane, a narrowly-shaped Gaussian will assign the latent coordinates much higher likelihood than a spherical Gaussian.

1.6.6 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. In each experiment, 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 compares the clustering performance of the iWMM with the iGMM, quan-

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

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**

tified by the Rand index (Rand, 1971), which measures the correspondence between inferred clusters and true clusters. The iGMM is another probabilistic generative model commonly used for clustering, which can be seen as a special case of the iWMM in which the Gaussian clusters are not warped. These experiments demonstrate the extent to which nonparametric cluster shapes allow a mixture model to recover more meaningful clusters.

Table 1.3 lists average test log likelihood, comparing the proposed models with kernel density estimation (KDE), and the infinite Gaussian mixture model (iGMM). In KDE, the kernel width is estimated by maximizing the leave-one-out log densities. 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 iWMMs. We also include the $Q = 2$ case in an attempt to characterize how much modeling power is lost by forcing the latent representation to be visualizable. The proposed models achieved high test likelihoods compared with the KDE and the iGMM.

Table 1.3 Average test log-likelihood 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

Source Code

Code to reproduce all the above experiments is available at github.com/duvenaud/warped-mixtures.

1.7 Conclusion

In this chapter, we 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 data contours. We then introduced an efficient sampler for this model which integrates out both the cluster parameters and the warping function exactly. We further demonstrated that allowing non-parametric cluster shapes improves clustering performance over the Dirichlet process Mixture of Gaussians.

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.

Since the proposed model is generative, it can be used for density estimation as well as clustering. It can also be extended to handle missing data, integrate with other probabilistic models, and use other families of distributions for the latent components.

1.8 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). 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 aren't 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 of section 1.8 would be possible, automatically finding the topology of the data manifold.

References

R.P. Adams and Z. Ghahramani. Archipelago: nonparametric Bayesian semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009. (page 11)

W. Cao and R. 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 Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011. (page 11)

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

A. Geiger, R. Urtasun, and T. 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)

Z. 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 7)

Tomoharu Iwata, David Duvenaud, and Zoubin Ghahramani. Warped mixtures for nonparametric cluster shapes. Bellevue, Washington, July 2013. URL `http://arxiv.org/pdf/1206.1846`. (page 1)

N.D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 16:329–336, 2004. (page 6)

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

S.N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, pages 223–238, 1998.

R.M. Neal. Density modeling and clustering using dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629, 2003. (page 13)

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

H. Nickisch and C. Rasmussen. Gaussian mixture modeling with Gaussian process latent variable models. *Pattern Recognition*, pages 272–282, 2010.

J. Quiñonero-Candela and C.E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005. (pages 5 and 13)

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

C.E. Rasmussen. The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 12(5.2):2, 2000. (page 4)

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*, volume 38. The MIT Press, Cambridge, MA, USA, 2006.

M. Salzmann, R. Urtasun, and P. 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)

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 2006. (pages 5 and 13)

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. (page 13)

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

M. Titsias and N. Lawrence. Bayesian Gaussian process latent variable model. *AISTATS*, 2010. (page 6)

J. Wang, J. Lee, and C. Zhang. Kernel trick embedded Gaussian mixture model. In *Algorithmic Learning Theory*, pages 159–174. Springer, 2003. (page 6)

Y. Zhang and C. Sutton. Quasi-Newton Markov chain Monte Carlo. *Advances in Neural Information Processing Systems*, pages 2393–2401, 2011. (page 13)