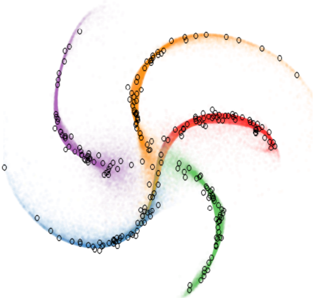
Warped Mixture Models

Tomoharu Iwata, David Duvenaud, Zoubin Ghahramani



Cambridge University
Computational and Biological Learning Lab

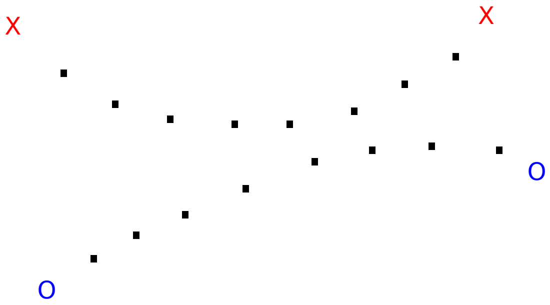
March 11, 2013

OUTLINE

- Motivation
- Gaussian Process Latent Variable Model
- Warped Mixtures
  - Generative Model
  - Inference
- Results
  - Generative Model
  - Inference
- Future Work:
  - Variational Inference
  - Semi-supervised learning
  - Other Latent Priors
- Life of a Bayesian Model

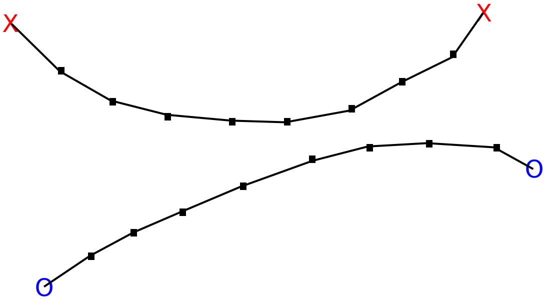
Motivation I: Manifold Semi-supervised Learning

- Most manifold learning algorithms start by constructing a graph locally.



Motivation I: Manifold Semi-supervised Learning

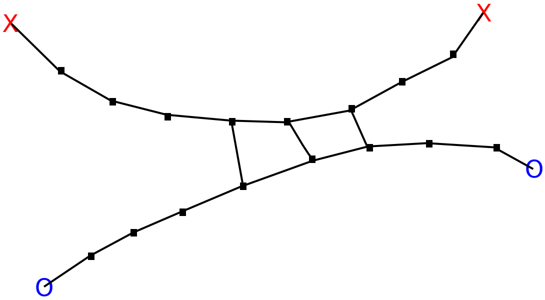
- Most manifold learning algorithms start by constructing a graph locally.



- Most don't update original topology to account for long-range structure or label information.

Motivation I: Manifold Semi-supervised Learning

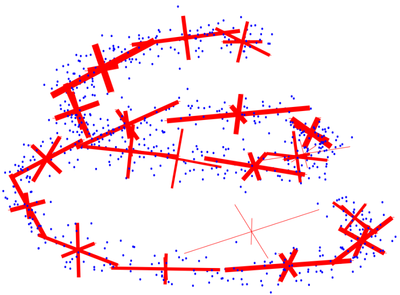
- Most manifold learning algorithms start by constructing a graph locally.

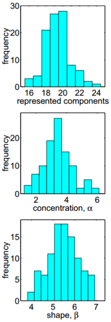


- Most don't update original topology to account for long-range structure or label information.
- Often hard to recover from a bad connectivity graph.

Motivation II: Infinite Gaussian Mixture Model

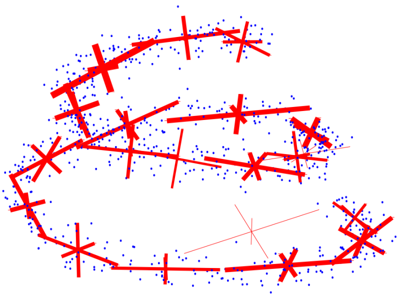
- Dirichelt Process prior on cluster weights.
- Recovers number of clusters automatically.
- Since each cluster must be Gaussian, number of clusters is often innapropriate

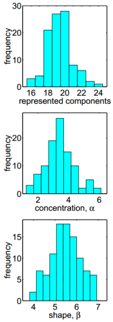




MOTIVATION II: INFINITE GAUSSIAN MIXTURE MODEL

- Dirichelt Process prior on cluster weights.
- Recovers number of clusters automatically.
- Since each cluster must be Gaussian, number of clusters is often innapropriate

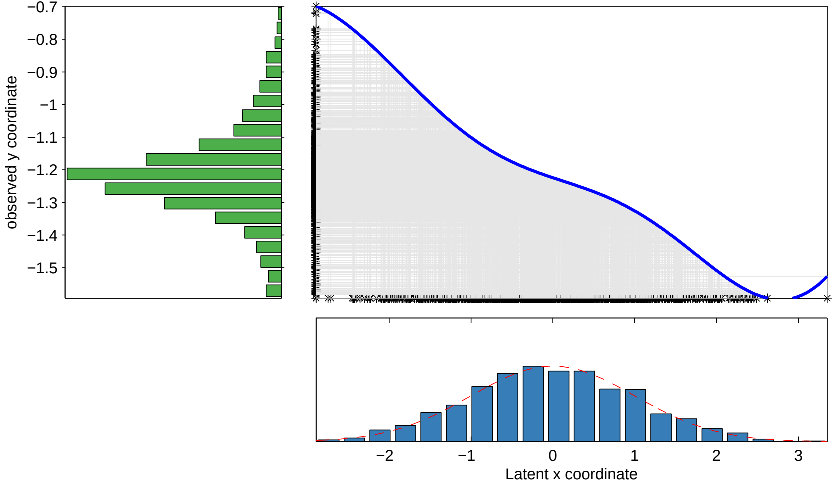




How to create nonparametric cluster shapes?

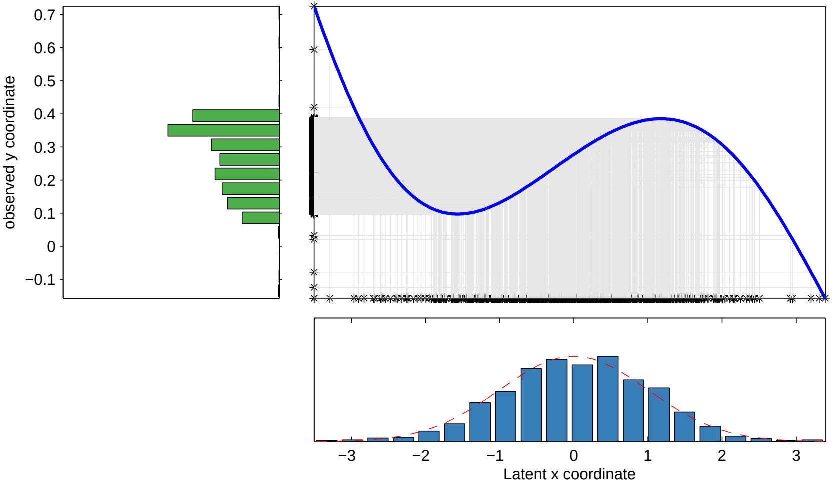
Gaussian Process Latent Variable Model

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$, latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$. $\mathbf{y} = f(\mathbf{x})$.



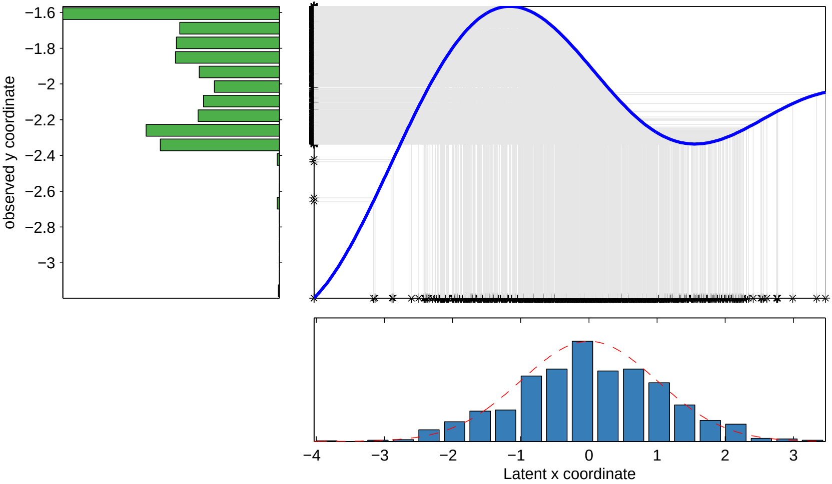
Gaussian Process Latent Variable Model

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$, latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$. $\mathbf{y} = f(\mathbf{x})$.



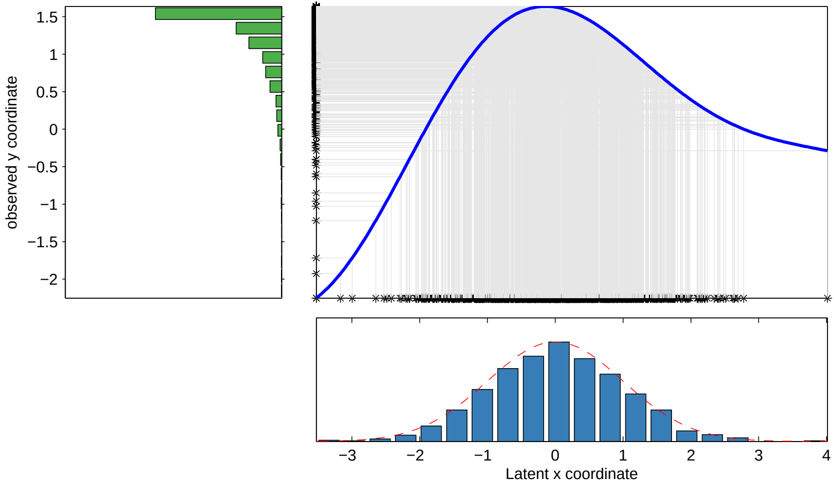
Gaussian Process Latent Variable Model

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$, latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$. $\mathbf{y} = f(\mathbf{x})$.



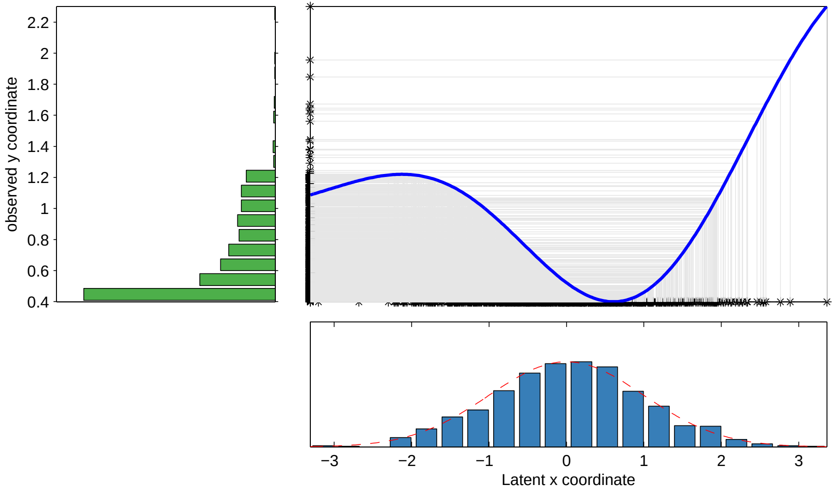
Gaussian Process Latent Variable Model

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$, latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$. $\mathbf{y} = f(\mathbf{x})$.



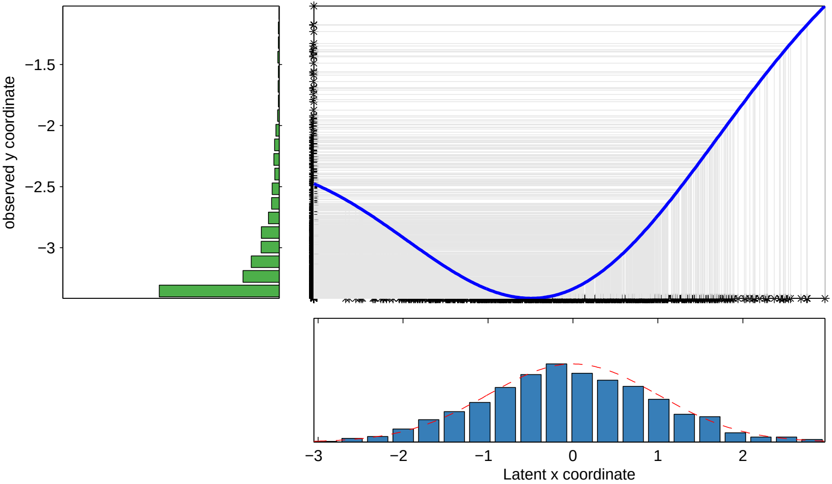
Gaussian Process Latent Variable Model

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$, latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$. $\mathbf{y} = f(\mathbf{x})$.



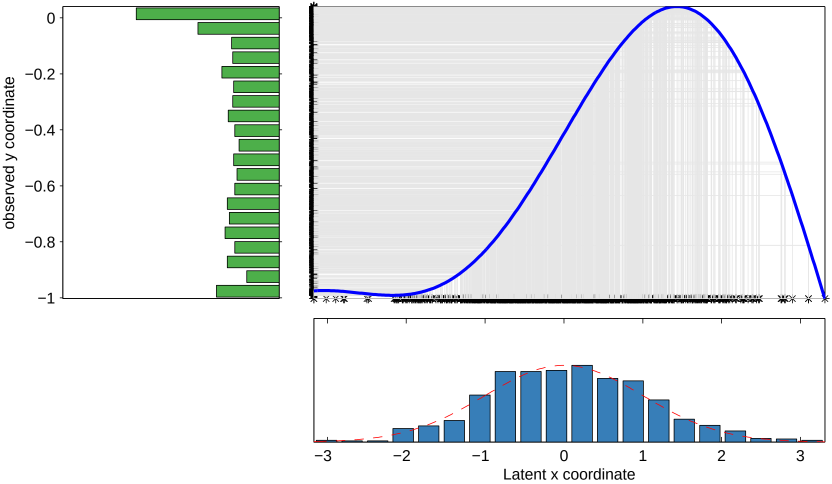
Gaussian Process Latent Variable Model

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$, latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$. $\mathbf{y} = f(\mathbf{x})$.



GAUSSIAN PROCESS LATENT VARIABLE MODEL

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$, latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$. $\mathbf{y} = f(\mathbf{x})$.



Gaussian Process Latent Variable Model

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$,
have latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$.
$\mathbf{y}_d = f_d(\mathbf{x})$, where each $f_d(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K})$.
Mapping marginal likelihood:

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = (2\pi)^{-\frac{DN}{2}} |\mathbf{K}|^{-\frac{D}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{Y}^\top \mathbf{K}^{-1} \mathbf{Y})\right)$$

Prior on x, treated mainly as a regularizer:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, I)$$

Can be interpreted as a density model.

Gaussian Process Latent Variable Model

Suppose observations $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top$ where $\mathbf{y}_n \in \mathbb{R}^D$,
have latent coordinates $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, where $\mathbf{x}_n \in \mathbb{R}^Q$.
$\mathbf{y}_d = f_d(\mathbf{x})$, where each $f_d(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K})$.
Mapping marginal likelihood:

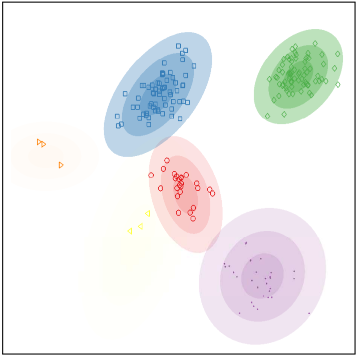
$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = (2\pi)^{-\frac{DN}{2}} |\mathbf{K}|^{-\frac{D}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{Y}^\top \mathbf{K}^{-1} \mathbf{Y})\right)$$

Prior on x, treated mainly as a regularizer:

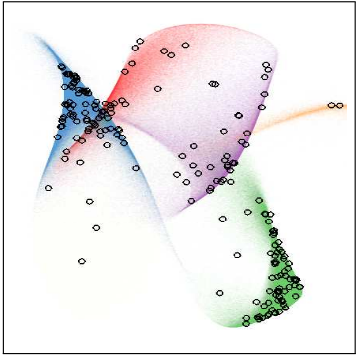
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, I)$$

Can be interpreted as a density model.
Can give warped densities; how to get clusters?

Warped Mixture Model



Latent space



Observed space

A sample from the iWMM prior:

- Sample a latent mixture of Gaussians.
- Warp the latent mixture to produce non-Gaussian manifolds in observed space.

Some areas with almost no density; some edges and peaks.

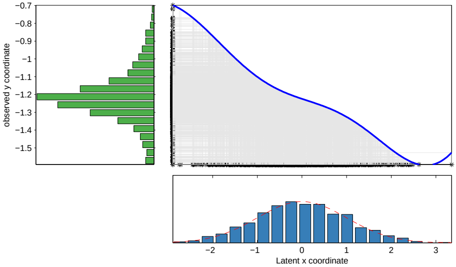
Warped Mixture Model

- An extension of GP-LVM, where $p(x)$ is a mixture of Gaussians.
- Or: An extension of iGMM, where mixture is warped.
- Given mixture assignments, likelihood has only two parts: GP-LVM and GMM

$$p(\mathbf{Y}|\mathbf{X},\mathbf{Z},\boldsymbol{\theta}) = \underbrace{(2\pi)^{-\frac{DN}{2}}|\mathbf{K}|^{-\frac{D}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\top})\right)}_{\text{GP-LVM Likelihood}}$$

$$\times \underbrace{\prod_i \sum_{c=1}^{\infty} \lambda_c \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_c,\mathbf{R}_c^{-1}) I(\mathbf{x}_i \in \mathbf{Z}_c)}_{\text{Mixture of Gaussians Likelihood}}$$

INFERENCE



Find posterior over latent X's.

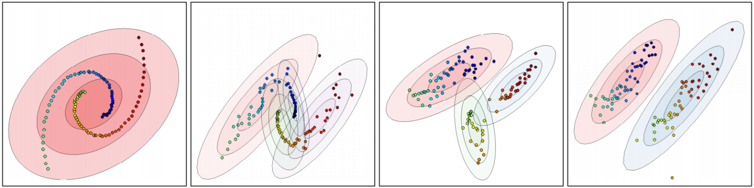
Many ways to do inference; high dimension of latent space means derivatives helpful.

Our scheme:

- Alternate:
  1. Sampling latent cluster assignments
  2. Updating latent positions and GP hypers with HMC
- No cross-validation, but HMC params annoying to set
- Show demo!

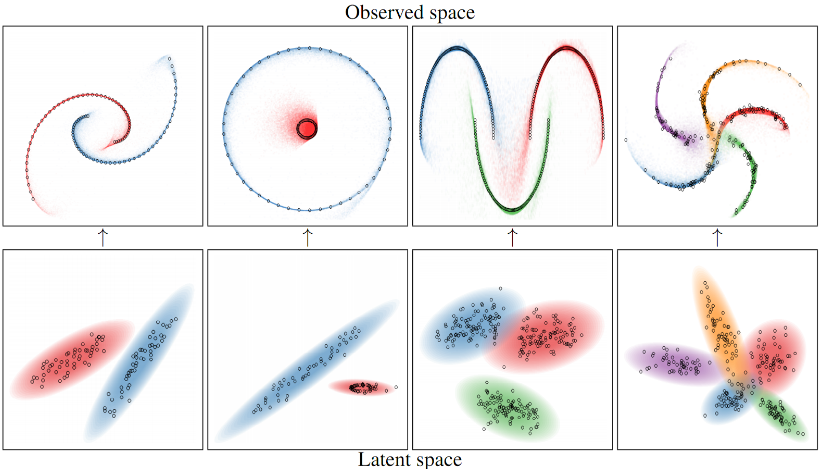
Inference: Mixing

- Changing number of clusters helps mixing.



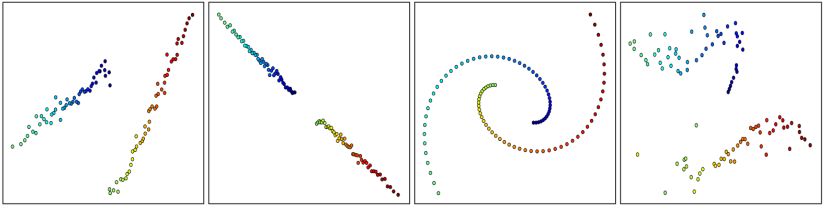
(a) 1 (b) 500 (c) 1800 (d) 3000

Density Results



- Automatically reduces latent dimension, separately per-cluster!
- Wishart prior may be causing problems.

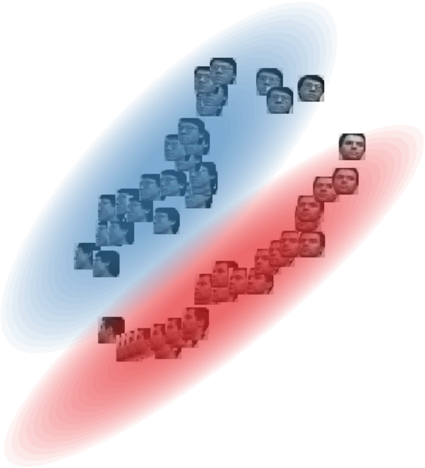
Latent Visualization



(a) iWMM (b) WM (c) GPLVM (d) BGPLVM

- Hard to summarize posterior which is symmetric - average for now.
- VB might address problem of summarizing posterior.

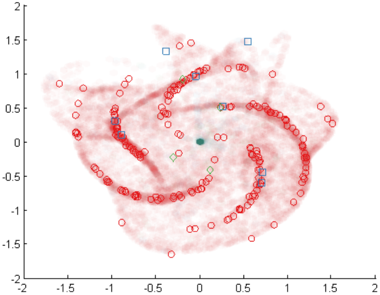
Latent Visualization: UMIST Faces dataset



Captures number, dimension, relationship between manifolds.

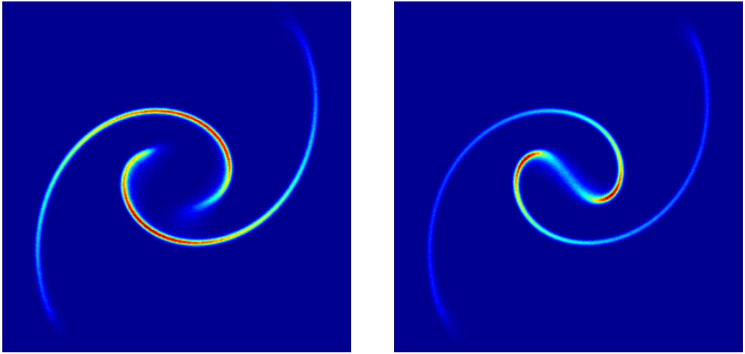
The Warped Density Model

- What if we take density model of GP-LVM seriously?
- Why not just warp one Gaussian?
- Even one latent Gaussian can be made fairly flexible.



The Warped Density Model

Even one latent Gaussian can be made fairly flexible, but must place some mass between clusters.



(a) iWMM (b) WM

Also easier to interpret latent clusters.

RESULTS

Evaluated iWMM as a density model, as well as a clustering model.

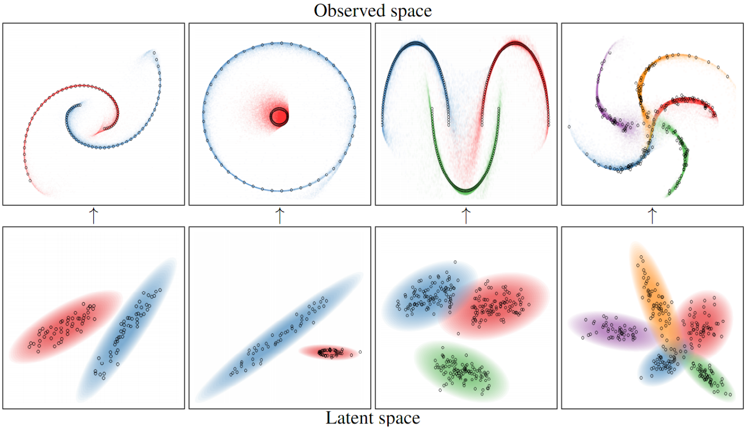
Table 2: Average test log likelihood for evaluating density estimation performance.

	2-curve	2-circle	3-semi	Pinwheel	Iris	Glass	Wine	Vowel
KDE	-2.652	-1.490	-0.295	-0.921	-1.644	3.376	-4.101	5.863
iGMM	-3.632	-1.794	-2.312	-1.920	-1.485	3.455	-3.771	-0.642
WM ($Q = 2$)	-1.212	-0.884	-0.627	-0.747	-1.647	5.473	**-3.197**	5.999
WM ($Q = D$)	-1.212	-0.884	-0.627	-0.747	-1.394	6.005	-4.630	0.705
iWMM ($Q = 2$)	**-1.190**	**-0.833**	**-0.081**	**-0.574**	-1.433	5.995	-3.475	**6.391**
iWMM ($Q = D$)	**-1.190**	**-0.833**	**-0.081**	**-0.574**	**-0.959**	**6.653**	-5.221	1.779

Table 3: Rand index for evaluating clustering performance.

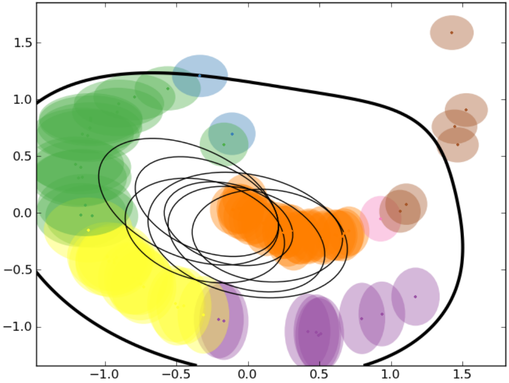
	2-curve	2-circle	3-semi	Pinwheel	Iris	Glass	Wine	Vowel
iGMM	0.544	0.815	0.732	0.813	0.776	0.618	0.712	0.759
iWMM ($Q = 2$)	**0.644**	**0.847**	**1.000**	**0.953**	0.776	0.657	0.666	0.660
iWMM ($Q = D$)	**0.644**	**0.847**	**1.000**	**0.953**	0.776	**0.675**	**0.748**	**0.773**

LIMITATIONS



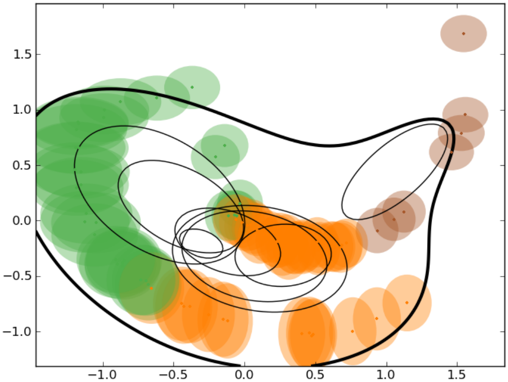
- $\mathcal{O}(N^3)$ runtime
- Stationary kernel means diffculty modeling clusters of different sizes.

Future Work: Variational Inference



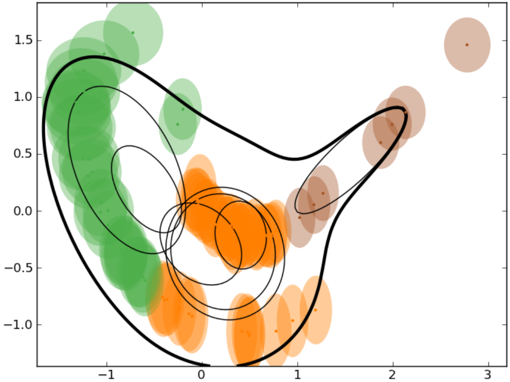
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



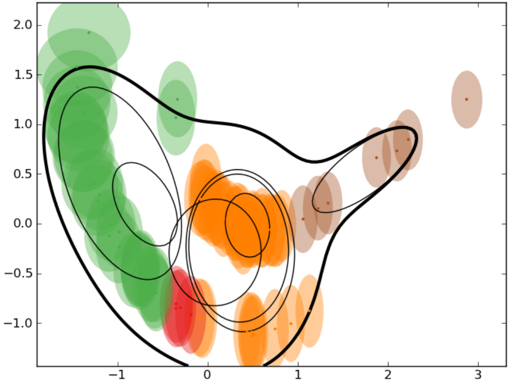
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



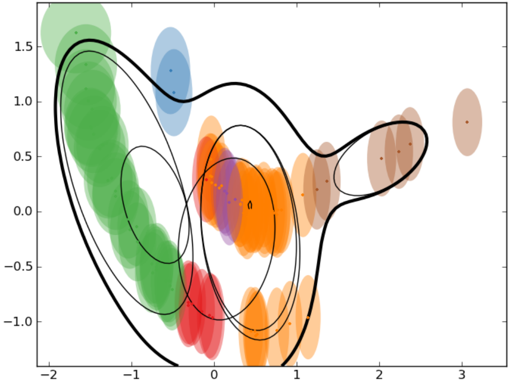
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



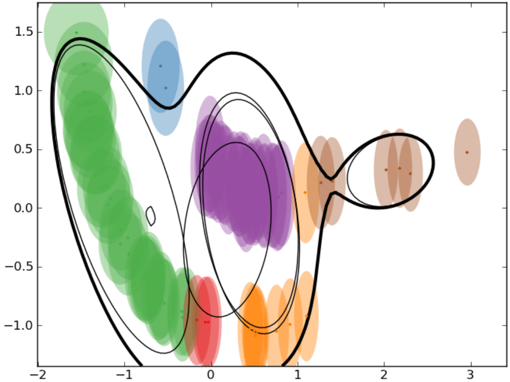
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



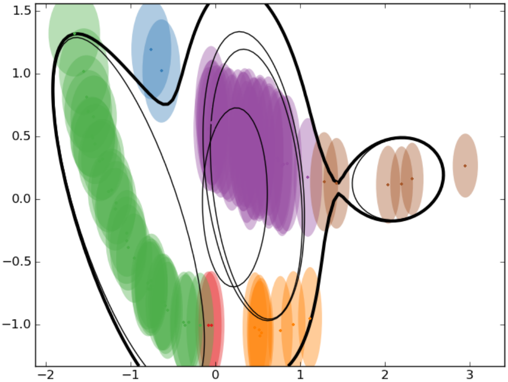
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

FUTURE WORK: VARIATIONAL INFERENCE



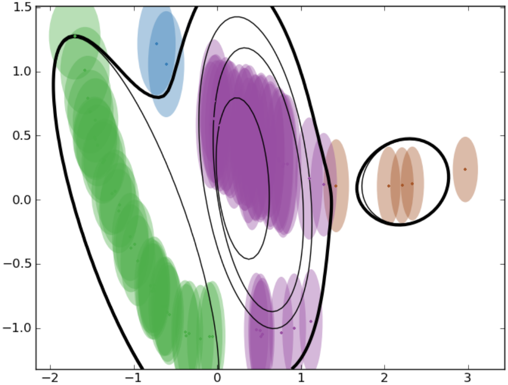
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



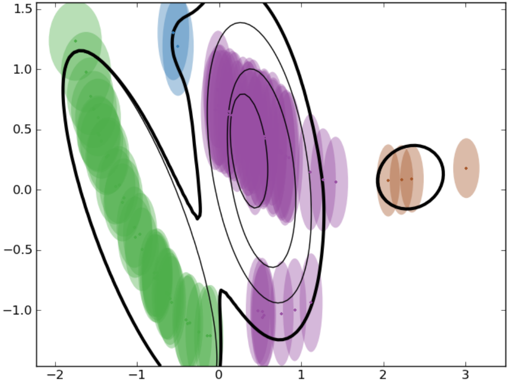
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

FUTURE WORK: VARIATIONAL INFERENCE



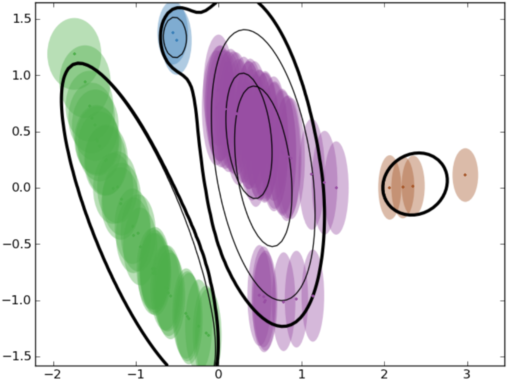
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



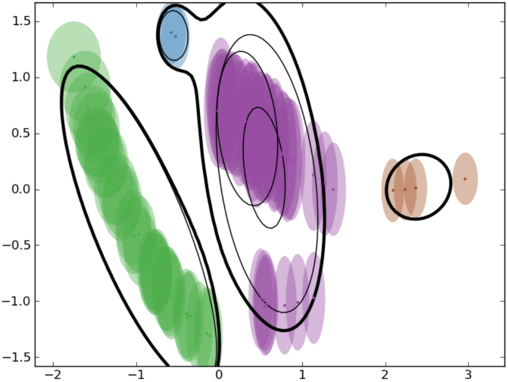
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



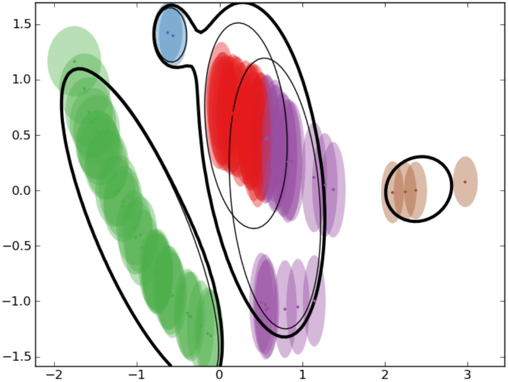
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

FUTURE WORK: VARIATIONAL INFERENCE



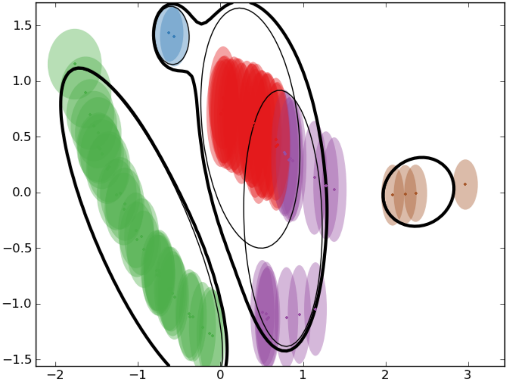
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



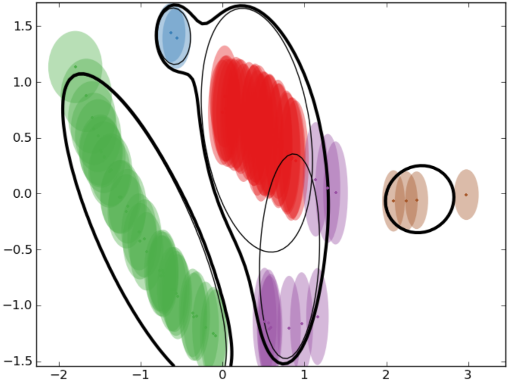
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



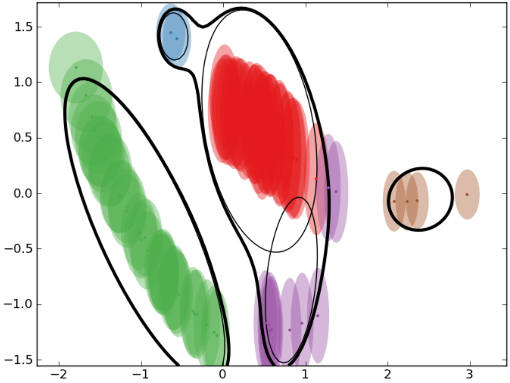
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



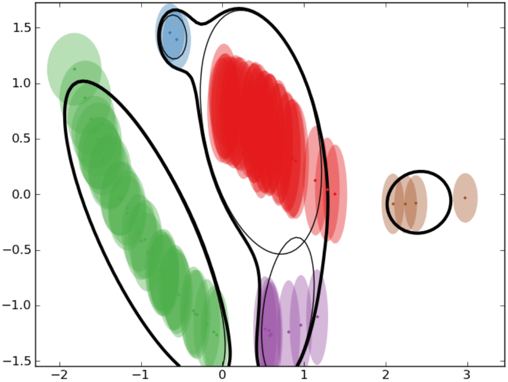
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



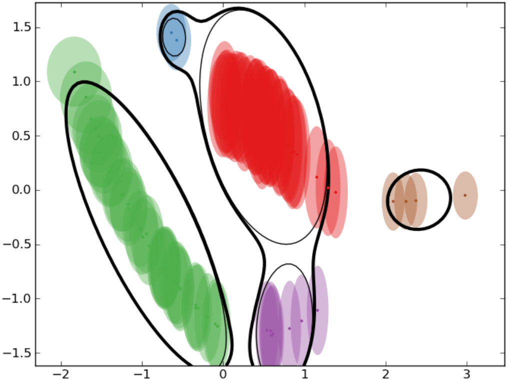
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



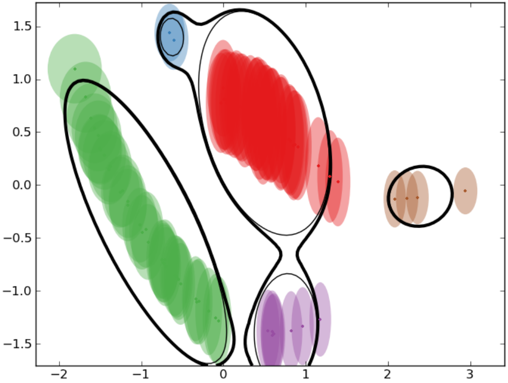
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



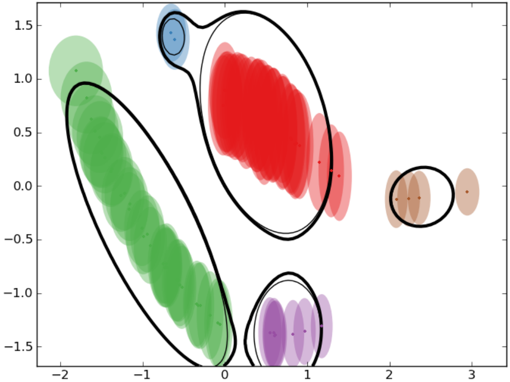
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



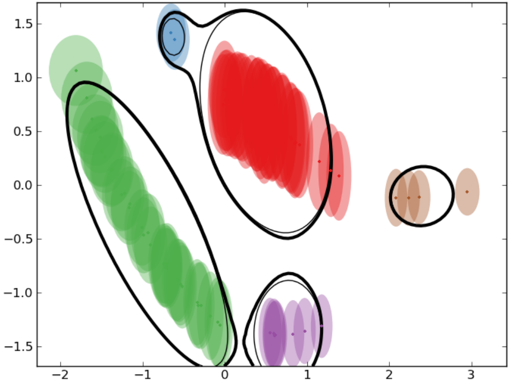
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



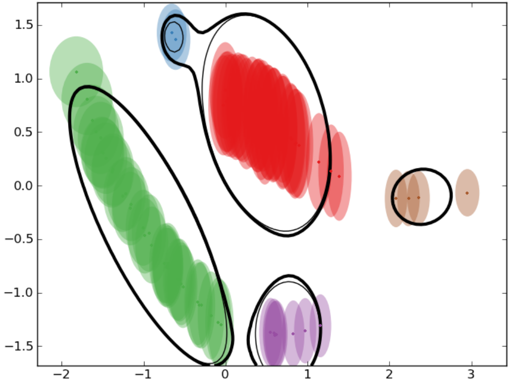
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



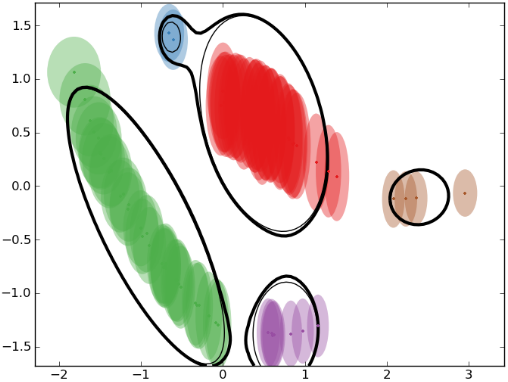
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



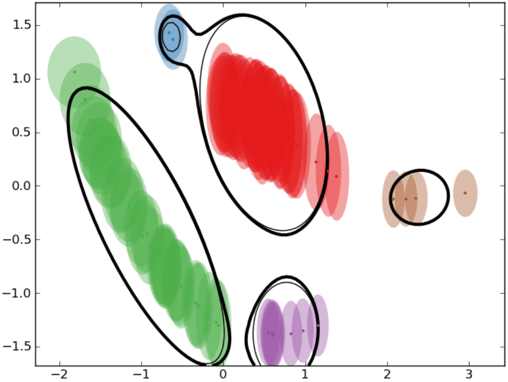
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



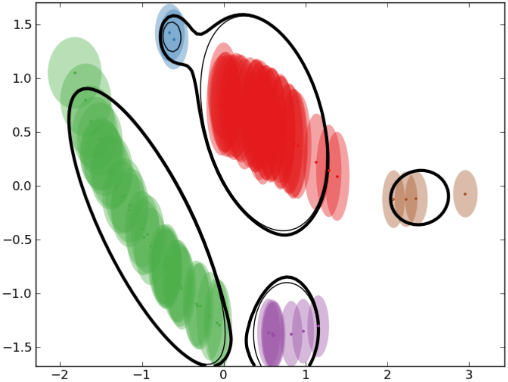
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



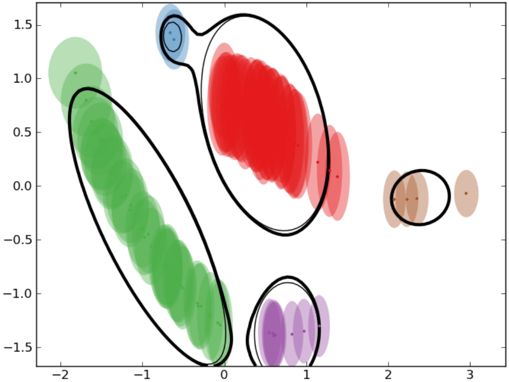
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



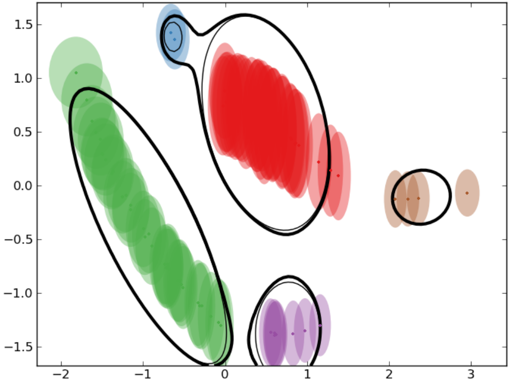
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



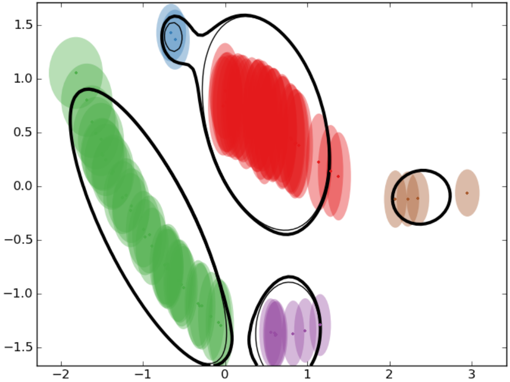
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



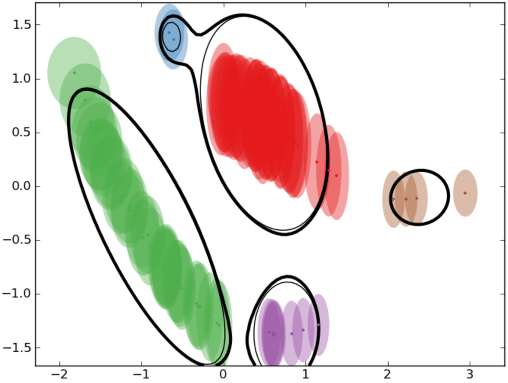
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



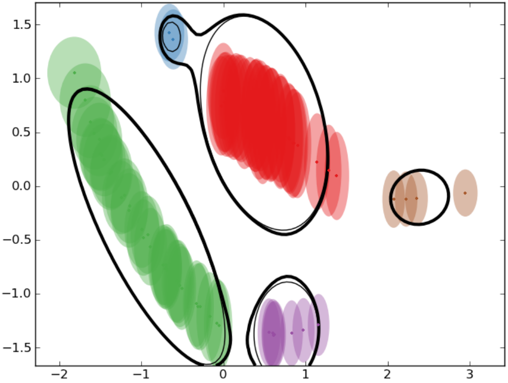
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

FUTURE WORK: VARIATIONAL INFERENCE



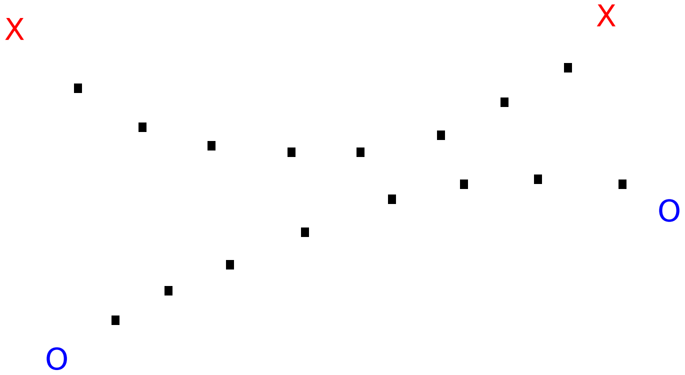
- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Variational Inference



- Joint work with James Hensman.
- Optimization instead of integration.
- SVI could allow large datasets.
- Non-convex optimization is hard; harder than mixing?

Future Work: Semi-supervised learning



- Spread labels along regions of high density.

Other Priors on Latent Densities

Density model is separate from warping model.

- Hierarchical clustering (bio applications)
- Deep Gaussian Processes

Life of a Bayesian Model

- Write down generative model.
- Sample from it to see if it looks reasonable.
- Fiddle with sampler for a month.
- Maybe years later, a decent inference scheme comes out.
- Modeling decisions are in principle separate from inference scheme
- Can verify approximate inference schemes on examples.
- Modeling sophistication is far ahead of inference sophistication

Life of a Bayesian Model

- Write down generative model.
- Sample from it to see if it looks reasonable.
- Fiddle with sampler for a month.
- Maybe years later, a decent inference scheme comes out.
- Modeling decisions are in principle separate from inference scheme
- Can verify approximate inference schemes on examples.
- Modeling sophistication is far ahead of inference sophistication

Thanks!