Density Clustering
10/26-702 Spring 2014

1 Modes and Clusters

Let p be the density of $X \in \mathbb{R}^d$. Assume that p has modes $m_1, \ldots, m_{k_0}$ and that p is a *Morse function*, which means that the Hessian of p at each stationary point is non-degenerate. We can use the modes to define clusters as follows.

Given any point $x \in \mathbb{R}^d$, there is a unique gradient ascent path, or integral curve, passing through x that eventually leads to one of the modes. We define the clusters to be the "basins of attraction" of the modes, the equivalence classes of points whose ascent paths lead to the same mode. Formally, an *integral curve* through x is a path $\pi_x : \mathbb{R} \to \mathbb{R}^d$ such that $\pi_x(0) = x$ and

$$\pi_x'(t) = \nabla p(\pi_x(t)). \tag{1}$$

Integral curves never intersect (except at stationary points) and they partition the space.

Equation (1) means that the path π follows the direction of steepest ascent of p through x. The destination of the integral curve π through a (non-mode) point x is defined by

$$\text{dest}(x) = \lim_{t \to \infty} \pi_x(t). \tag{2}$$

It can then be shown that for all x, $\text{dest}(x) = m_j$ for some mode m_j. That is: all integral curves lead to modes. For each mode m_j, define the sets

$$\mathscr{A}_j = \Big\{ x : \ \text{dest}(x) = m_j \Big\}. \tag{3}$$

These sets are known as the *ascending manifolds*, and also known as the cluster associated with m_j, or the basin of attraction of m_j. The $\mathscr{A}_j$'s partition the space. See Figure 1. The collection of ascending manifolds is called the *Morse complex*.

Given data $X_1, \ldots, X_n$ we construct an estimate $\widehat{p}$ of the density. Let $\widehat{m}_1, \ldots, \widehat{m}_k$ be the estimated modes and let $\widehat{\mathscr{A}_1}, \ldots, \widehat{\mathscr{A}_k}$ be the corresponding ascending manifolds derived from $\widehat{p}$. The sample clusters $C_1, \ldots, C_k$ are defined to be $C_j = \{X_i : \ X_i \in \widehat{\mathscr{A}_j}\}$.

Recall that the kernel density estimator is

$$\widehat{p}(x) \equiv \widehat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{||x - X_i||}{h} \right) \tag{4}$$

where K is a smooth, symmetric kernel and $h > 0$ is the bandwidth.[1] The mean of the

[1] In general, we can use a bandwidth matrix H in the estimator, with $\widehat{p}(x) \equiv \widehat{p}_H(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)$ where $K_H(x) = |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}} x)$.

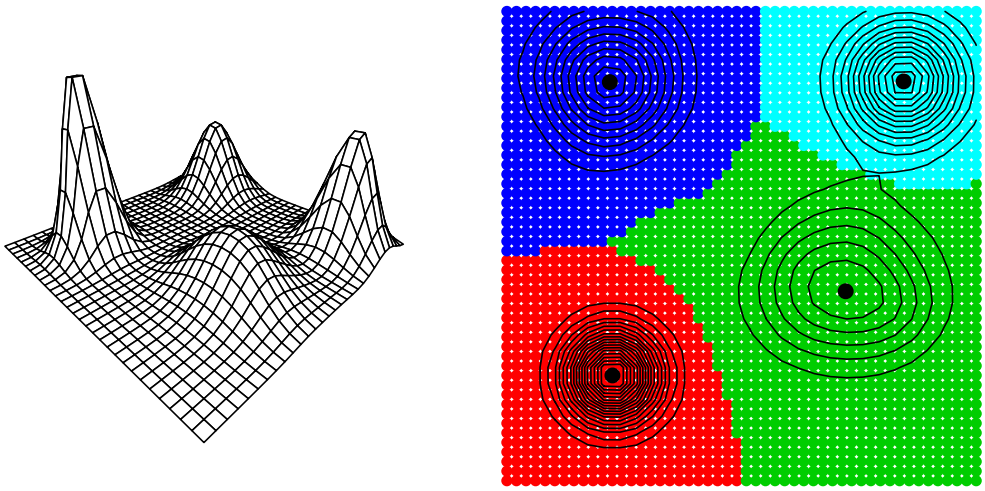

Figure 1: *The left plot shows a function with four modes. The right plot shows the ascending manifolds (basins of attraction) corresponding to the four modes.*

estimator is

$$p_h(x) = \mathbb{E}[\widehat{p}_h(x)] = \int K(t)p(x+th)dt. \tag{5}$$

To locate the modes of $\widehat{p}_h$ we use the *mean shift algorithm* which finds modes by approximating the steepest ascent paths. The algorithm is given in Figure 2. The result of this process is the set of estimated modes $\widehat{\mathcal{M}} = \{\widehat{m}_1, \ldots, \widehat{m}_k\}$. We also get the clustering for free: the mean shift algorithm shows us what mode each point is attracted to. See Figure 3.

A modified version of the algorithm is the blurred mean-shift algorithm (Carreira-Perpinan, 2006). Here, we use the data as the mesh and we replace the data with the mean-shifted data at each step. This converges very quickly but must be stopped before everything converges to a single point; see Figures 4 and 5.

What we are doing is tracing out the *gradient flow*. The flow lines lead to the modes and they define the clusters. In general, a flow is a map $\phi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ such that $\phi(x,0) = x$ and $\phi(\phi(x,t),s) = \phi(x,s+t)$. The latter is called the semi-group property.

2 Choosing the Bandwidth

As usual, choosing a good bandwidth is crucial. You might wonder if increasing the bandwidth, decreases the number of modes. Silverman (1981) showed that the answer is yes if you use a Normal kernel.

Mean Shift Algorithm

1. Input: $\widehat{p}(x)$ and a mesh of points $A = \{a_1, \ldots, a_N\}$ (often taken to be the data points).
2. For each mesh point a_j, set $a_j^{(0)} = a_j$ and iterate the following equation until convergence:

$$a_j^{(s+1)} \longleftarrow \frac{\sum_{i=1}^n X_i K\left(\frac{||a_j^{(s)} - X_i||}{h}\right)}{\sum_{i=1}^n K\left(\frac{||a_j^{(s)} - X_i||}{h}\right)}.$$

3. Let $\widehat{\mathscr{M}}$ be the unique values of the set $\{a_1^{(\infty)}, \ldots, a_N^{(\infty)}\}$.
4. Output: $\widehat{\mathscr{M}}$.

Figure 2: *The Mean Shift Algorithm.*

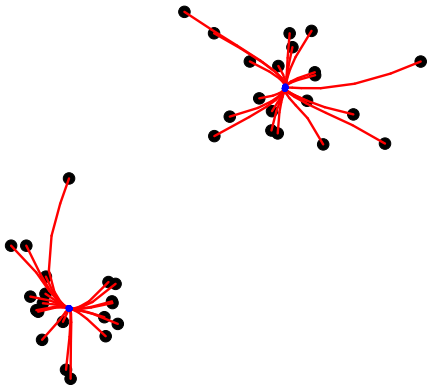

Figure 3: A simple example of the mean shift algorithm.

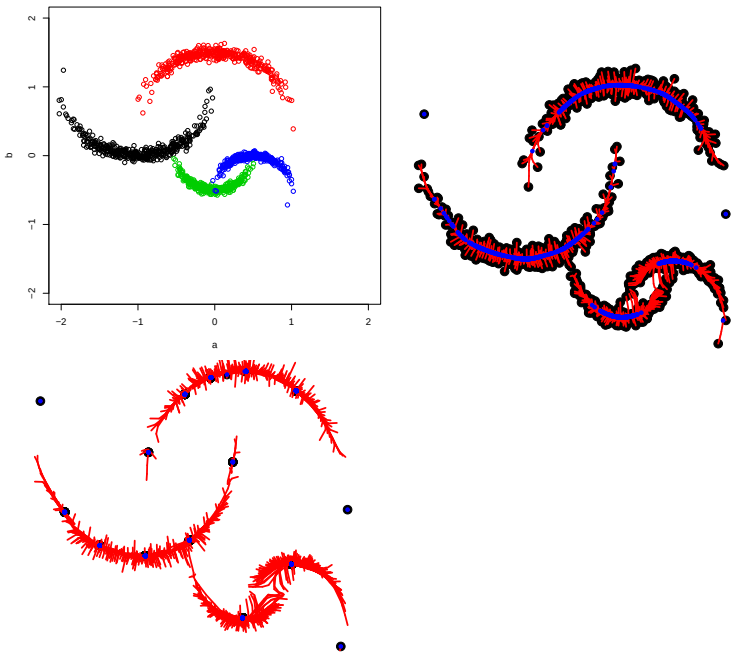

Figure 4: The crescent data example. Top left: data. Top right: a few steps of mean-shift. Bottom left: a few steps of blurred mean-shift.

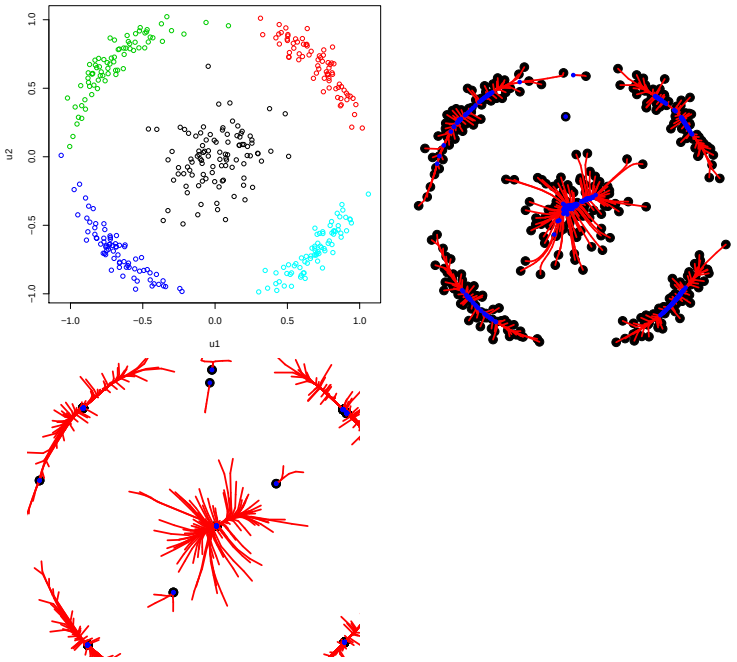

Figure 5: The Broken Ring example. Top left: data. Top right: a few steps of mean-shift. Bottom left: a few steps of blurred mean-shift.

Theorem 1 (Silverman 1981) *Let $\widehat{p}_h$ be a kernel density estimator using a Gaussian kernel. Then the number of modes of $\widehat{p}_h$ is a non-increasing function of h.*

We still need a way to pick h. We can use cross-validation as before. One could argue that we should choose h so that we estimate the gradient $g(x) = \nabla p(x)$ well since the clustering is based on the gradient flow.

How can we estimate the loss of the gradient? Consider, first the scalar case. Note that

$$\int (\widehat{p}' - p')^2 = \int (\widehat{p}')^2 - 2\int \widehat{p}p' + \int (p')^2.$$

We can ignore the last term. The first term is known. To estimate the middle term, we use integration by parts to get

$$\int \widehat{p}p' = -\int p''p$$

suggesting the cross-validation estimator

$$\int (\widehat{p}'(x))^2 dx + \frac{2}{n}\sum_i \widehat{p}_i''(X_i)$$

where $\widehat{p}_i''$ is the leave-one-out second derivative. More generally, by repeated integration by parts, we can estimate the loss for the r^{th} derivative by

$$\text{CV}_r(h) = \int (\widehat{p}(r)(x))^2 dx - \frac{2}{n}(-1)^r \sum_i \widehat{p}(2r)_i(X_i).$$

Let's now discuss estimating derivatives more generally following Chacon and Duong (2013). Let

$$\widehat{p}_H(x) = \frac{1}{n}\sum_{i=1}^n K_H(x - X_i)$$

where $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$. Let $D = \partial/\partial x = (\partial/\partial x_1, \ldots, \partial/\partial x_d)$ be the gradient operator. Let $H(x)$ be the Hessian of $p(x)$ whose entries are $\partial^2 p/(\partial x_j \partial x_k)$. Let

$$D^{\otimes r}p = (Dp)^{\otimes r} = \partial^r p/\partial x^{\otimes r} \in \mathbb{R}^{d^r}$$

denote the r^{th} derivatives, organized into a vector. Thus

$$D^{\otimes 0}p = p, \quad D^{\otimes 1}p = Dp, \quad D^{\otimes 2}p = \text{vec}(H)$$

where vec takes a matrix and stacks the columns into a vector.

The estimate of $D^{\otimes r}p$ is

$$\widehat{p}^{(r)}(x) = D^{\otimes r}\widehat{p}_H(x) = \frac{1}{n}\sum_{i=1}^n D^{\otimes r}K_H(x - X_i) = \frac{1}{n}\sum_{i=1}^n |H|^{-1/2}(H^{-1/2})^{\otimes r}D^{\otimes r}K(H^{-1/2}(x - X_i).$$

The integrated squared error is

$$L = \int ||D^{\otimes r}\widehat{p}_H(x) - D^{\otimes r}p(x)||^2 dx.$$

Chacon, Duong and Wand shows that $\mathbb{E}[L]$ is minimized by choosing H so that each entry has order $n^{-2/(d+2r+4)}$ leading to a risk of order $O(n^{-4/(d+2r+4)})$. In fact, it may be shown that

$$\begin{aligned}\mathbb{E}[L] = \frac{1}{n}|H|^{-1/2}\text{tr}((H^{-1})^{\otimes r}R(D^{\otimes r}K)) - \frac{1}{n}\text{tr}R^*(K_H \star K_H, D^{\otimes r}p) \\ + \text{tr}R^*(K_H \star K_H, D^{\otimes r}p) - 2\text{tr}R^*(K_H, D^{\otimes r}p) + \text{tr}R(D^{\otimes r}p)\end{aligned}$$

where

$$R(g) = \int g(x)g^T(x)dx$$

$$R^*(a,g) = \int (a \star g)(x)g^T(x)dx$$

and $(a \star g)$ is componentwise convolution.

To estimate the loss, we expand L as

$$L = \int ||D^{\otimes r}\widehat{p}_H(x)||^2 dx - 2\int \langle D^{\otimes r}\widehat{p}_H(x), D^{\otimes r}p(x)\rangle dx + \text{constant}.$$

Using some high-voltage calculations, Chacon and Duong (2013) derived the following leave-one-out approximation to the first two terms:

$$\text{CV}_r(H) = (-1)^r|H|^{-1/2}(\text{vec}(H^{-1})^{\otimes r})^T B(H)$$

where

$$B(H) = \frac{1}{n^2}\sum_{i,j} D^{\otimes 2r}\overline{K}(H^{-1/2}(X_i - X_j)) - \frac{2}{n(n-1)}\sum_{i\neq j} D^{\otimes 2r}K(H^{-1/2}(X_i - X_j))$$

and $\overline{K} = K \star K$ In practice, the minimization is easy if we restrict to matrices of the form $H = h^2 I$.

3 Theoretical Analysis

How well can we estimate the modes? Here we derive the rate of convergence. For simplicity, let's focus on a simple case. Assume that p has one mode m and that p is locally quadratic around m. We will use the analysis in Donoho and Liu (1991).

Let $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$, let m_h maximize p_h and let $\widehat{m}_h$ maximize $\widehat{p}_h$. Note that

$$\widehat{p}_h(\widehat{m}_h) \geq \widehat{p}_h(m_h).$$

Thus

$$\begin{aligned}
&[\widehat{p}_h(\widehat{m}_h) - p_h(\widehat{m}_h)] - [\widehat{p}_h(m_h) - p_h(m_h)] \\
&\qquad \geq [\widehat{p}_h(m_h) - p_h(\widehat{m}_h)] - [\widehat{p}_h(m_h) - p_h(m_h)] = p_h(m_h) - p_h(\widehat{m}_h) \\
&\qquad \geq c||m_h - \widehat{m}_h||^2
\end{aligned}$$

for some $c > 0$, where we used the fact that p_h is locally quadratic around m_h. We conclude that

$$\sqrt{nh^d}||m_h - \widehat{m}_h||^2 \preceq Z_n(\widehat{m}_h) - Z_n(m_h)$$

where

$$Z_n(t) = \sqrt{nh^d}(\widehat{p}_h(t) - p_h(t)).$$

We know that, for each t, $Z_n(t)$ converges to a Normal and hence, $Z_n(t) = O_P(1)$. In fact, $\sup_t |Z_n(t)| = O_P(1)$. It follows that

$$||m_h - \widehat{m}_h|| = O_P\left(\frac{1}{nh^d}\right)^{1/4}.$$

Using a Taylor series, it may be shown that $||m_h - m|| = O(h)$. Hence,

$$||m - \widehat{m}_h|| = O_P\left(\frac{1}{nh^d}\right)^{1/4} + O(h).$$

We see that the optimal h is $h \asymp n^{-1/(4+d)}$ giving

$$||m - \widehat{m}_h|| = O_P\left(n^{-\frac{1}{4+d}}\right).$$

Romano (1988) showed that the minimax rate is $n^{-\frac{p-1}{2p+d}}$ assuming p bounded derivatives in a neighborhood around the mode. Our analysis gives the minimax rate for $p = 2$.

4 Level Sets

An alternative to mode clustering is *level set clustering*. Let $L_t = \{x : \ p(x) > t\}$ denote an upper level set of p. Suppose that L_t can be decomposed into finitely many disjoint sets: $L_t = C_1 \bigcup \cdots \bigcup C_{k_t}$. We call $\mathscr{C}_t = \{C_1, \ldots, C_{k_t}\}$ the level set clusters at level t.

Let $\mathscr{C} = \bigcup_{t \geq 0} \mathscr{C}_t$. The clusters in $\mathscr{C}$ form a tree: if $A, B \in \mathscr{C}$, the either (i) $A \subset B$ or (ii)$B \subset A$ or (iii) $A \cap B = \emptyset$. We call $\mathscr{C}$ the *level set cluster tree*.

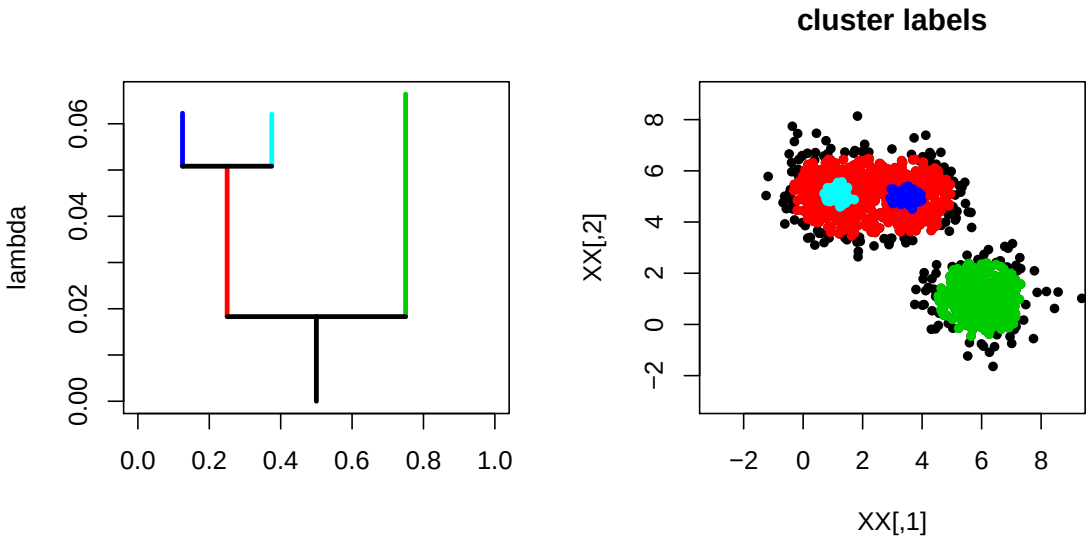

Figure 6: DeBaClR in two dimensions.

The level sets can be estimated in the obvious way: $\widehat{L}_t = \{x : \ \widehat{p}_h(x) > t\}$. How do we decompose $\widehat{L}_t$ into its connected components? This can be done as follows. For each t let

$$\mathscr{X}_t = \{X_i : \ \widehat{p}_h(X_i) > t\}.$$

Now construct a graph G_t where each $X_i \in \mathscr{X}_t$ is a vertex and there is an edge between X_i and X_j if and only if $||X_i - X_j|| \leq \epsilon$ where $\epsilon > 0$ is a tuning parameter. G_t is a called a Rips graphs. The clusters at level t are estimated by taking the connected components of the graph G_t.

In R, the package *denpro* does these calculations. A Python package, called DeBaCl, written by Brian Kent, can be found at

`http://www.brianpkent.com/projects.html`.

Fabrizio Lecci has written an R implementation called DeBaClR which we hope to post soon. Two examples are shown in Figures 6 and 7.

5 Persistence

Consider a smooth density p with $M = \sup_x p(x) < \infty$. The t-level set clusters are the connected components of the set $L_t = \{x : \ p(x) \geq t\}$. Suppose we find the upper level sets $L_t = \{x : \ p(x) \geq t\}$ as we vary t from M to 0. *Persistent homology* measures how the topology of L_t varies as we decrease t. In our case, we are only interested in the modes, which

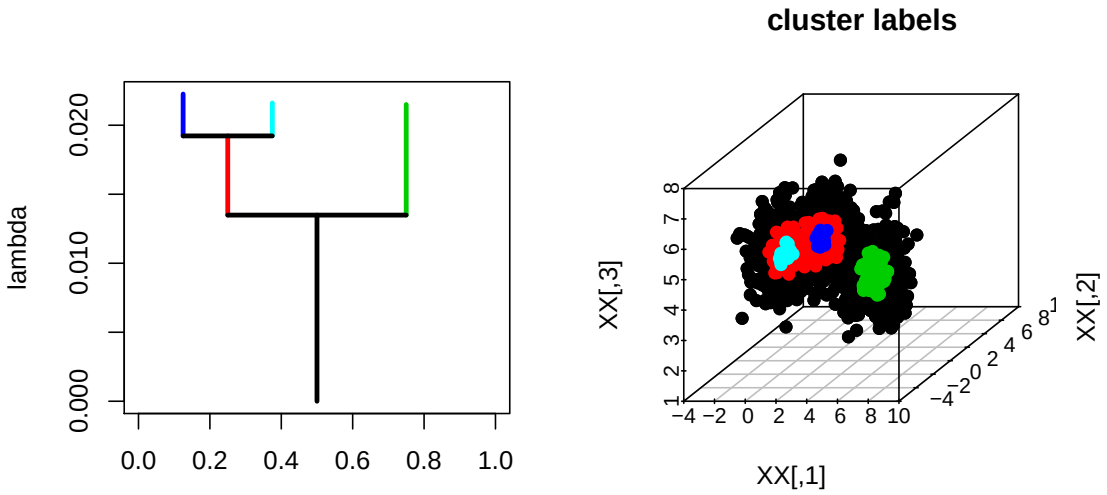

Figure 7: DeBaClR in three dimensions.

correspond to the zeroth order homology. (Higher order homology refers to holes, tunnels etc.)

Imagine setting $t = M$ and then gradually decreasing t. Whenever we hit a mode, a new level set cluster is born. As we decrease t further, some clusters may merge and we say that one of the clusters (the one born most recently) has died. See Figure 8.

In summary, each mode m_j has a death time and a birth time denoted by (d_j, b_j). (Note that the birth time is larger than the death time because we start at high density and move to lower density.) The modes can be summarized with a persistence diagram where we plot the points $(d_1, b_1), \ldots, (d_k, b_k)$ in the plane. See Figure 8. Points near the diagonal correspond to modes with short lifetimes. We might kill modes with lifetimes smaller than the bootstrap quantile ϵ_α defined by

$$\epsilon_\alpha = \inf\left\{z : \frac{1}{B}\sum_{b=1}^{B} I\left(||\widehat{p}_h^{*b} - \widehat{p}_h||_\infty > z\right) \leq \alpha\right\}. \tag{6}$$

Here, $\widehat{p}_h^{*b}$ is the density estimator based on the b^{th} bootstrap sample. This corresponds to killing a mode if it is in a $2\epsilon_\alpha$ band around the diagonal. See Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan and Singh (2013). Note that the starting and ending points of the vertical bars on the level set tree are precisely the coordinates of the persistence diagram.

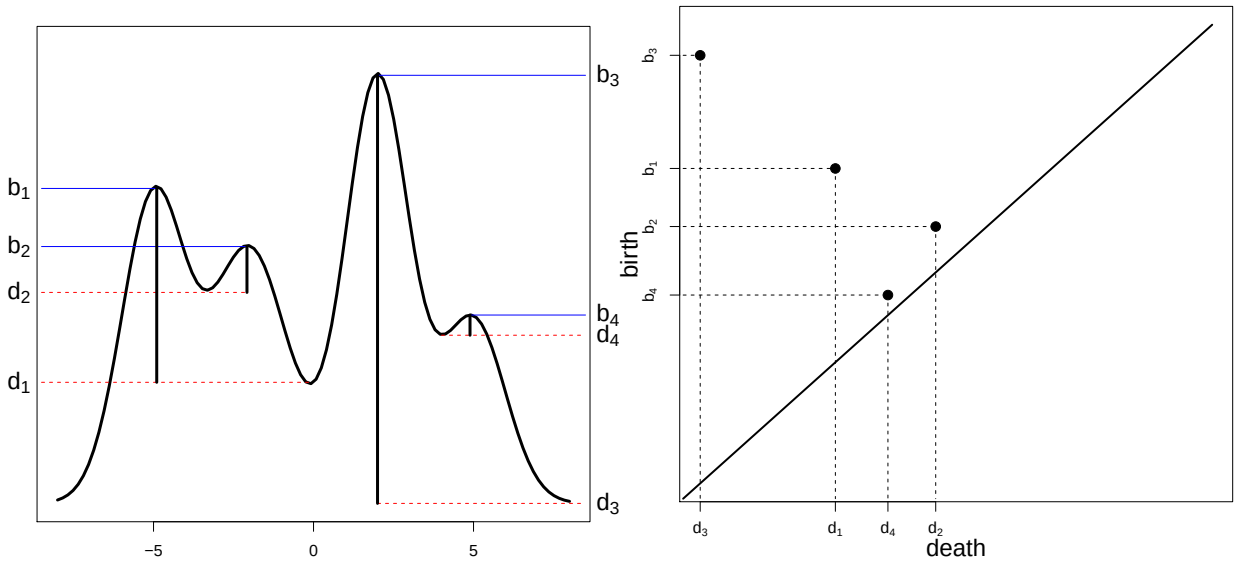

Figure 8: *Starting at the top of the density and moving down, each mode has a birth time b and a death time d. The persistence diagram (right) plots the points $(d_1,b_1),\ldots,(d_4,b_4)$. Modes with a long lifetime are far from the diagonal.*

6 Mixtures of Normals

Another method for clustering is based on mixtures. You covered this in 701. A mixture of d-dimensional multivariate Gaussians is

$$p(x) = \sum_{j=1}^{k} \pi_j \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right\}.$$

There are $\frac{kd(d+3)}{2}+k-1$ parameters to estimate. This is usually done by maximum likelihood. But maximizing the likelihood is NP hard. Instead, one usually uses the EM algorithm and hopes for the best. We can then assign observations to clusters since

$$\mathbb{P}(X_i \text{ is from cluster } j) = \frac{\pi_j\phi_j(X_i)}{\sum_s \pi_s\phi_s(X_i)}.$$

There are problems with mixtures models which we now discuss.

Computation. Finding the mle is NP-hard.

Infinite Likelihood. Let $p_\psi(x) = \sum_{j=1}^k \pi_j\phi(x;\mu_j,\sigma_j^2)$, be a mixture of Gaussians. Let $\mathscr{L}(\psi) = \prod_{i=1}^n p_\psi(X_i)$ be the likelihood function based on a sample of size n. Then $\sup_\psi \mathscr{L}(\psi) = \infty$. To see this, set $\mu_j = X_1$ for some j. Then $\phi(X_1;\mu_j,\sigma_j^2) = (\sqrt{2\pi}\sigma_j)^{-1}$. Now let $\sigma_j \to 0$. We have $\phi(X_1;\mu_j,\sigma_j^2) \to \infty$. Therefore, the log-likelihood is unbounded. This behavior is very

different from a typical parametric model. Fortunately, if we define the maximum likelihood estimate to be a mode of $\mathscr{L}(\psi)$ in the interior of the parameter space, we get a well-defined estimator.

Multimodality of the Density. Consider the mixture of two Gaussians

$$p(x) = (1-\pi)\phi(x;\mu_1,\sigma^2) + \pi\phi(x;\mu_0,\sigma^2).$$

You would expect $p(x)$ to be multimodal but this is not necessarily true. The density $p(x)$ is unimodal when $|\mu_1 - \mu_2| \leq 2\sigma$ and bimodal when $|\mu_1 - \mu_2| > 2\sigma$. One might expect that the maximum number of modes of a mixture of k Gaussians would be k. However, there are examples where a mixture of k Gaussians has more than k modes. In fact, Edelsbrunner, Fasy and Rote (2012) show that the relationship between the number of modes pf p and the number of components in the mixture is very complex.

Nonidentifiability. A model $\{p_\theta(x) : \theta \in \Theta\}$ is identifiable if

$$\theta_1 \neq \theta_2 \quad \text{implies} \quad P_{\theta_1} \neq P_{\theta_2}$$

where P_θ is the distribution corresponding to the density p_θ. Mixture models are nonidentifiable in two different ways. First, there is nonidentifiability due to permutation of labels. For example, consider a mixture of two univariate Gaussians,

$$p_{\psi_1}(x) = 0.3\phi(x;0,1) + 0.7\phi(x;2,1)$$

and

$$p_{\psi_2}(x) = 0.7\phi(x;2,1) + 0.3\phi(x;0,1),$$

then $p_{\psi_1}(x) = p_{\psi_2}(x)$ even though $\psi_1 = (0.3, 0.7, 0, 2, 1)^T \neq (0.7, 0.3, 2, 0, 1)^T = \psi_2$. This is not a serious problem although it does contribute to the multimodality of the likelihood.

A more serious problem is local nonidentifiability. Suppose that

$$p(x;\eta,\mu_1,\mu_2) = (1-\eta)\phi(x;\mu_1,1) + \eta\phi(x;\mu_2,1). \tag{7}$$

When $\mu_1 = \mu_2 = \mu$, we see that $p(x;\eta,\mu_1,\mu_2) = \phi(x;\mu)$. The parameter η has disappeared. Similarly, when $\eta = 1$, the parameter μ_2 disappears. This means that there are subspaces of the parameter space where the family is not identifiable. This local nonidentifiability causes many of the usual theoretical properties— such as asymptotic Normality of the maximum likelihood estimator and the limiting χ^2 behavior of the likelihood ratio test— to break down. For the model (7), there is no simple theory to describe the distribution of the likelihood ratio test for $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. The best available theory is very complicated. However, some progress has been made lately using ideas from algebraic geometry (Yamazaki and Watanabe 2003, Watanabe 2010).

The lack of local identifiabilty causes other problems too. For example, we usually have that the Fisher information is non-zero and that $\widehat{\theta} - \theta = O_P(n^{-1/2})$ where $\widehat{\theta}$ is the maximum

likelihood estimator. Mixture models are, in general, irregular: they do not satisfy the usual regularity conditions that make parametric models so easy to deal with. Here is an example from Chen (1995).

Consider a univariate mixture of two Gaussians distribution:

$$p_\theta(x) = \frac{2}{3}\phi(x; -\theta, 1) + \frac{1}{3}\phi(x; 2\theta, 1).$$

Then it is easy to check that $I(0) = 0$ where $I(\theta)$ is the Fisher information. Moreover, no estimator of θ can converge faster than $n^{-1/4}$ if the number of components is not known in advance. Compare this to a Normal family $\phi(x;\theta,1)$ where the Fisher information is $I(\theta) = n$ and the maximum likelihood estimator converges at rate $n^{-1/2}$. Moreover, the distribution of the mle is not even well understood for mixture models. The same applies to the likelihood ratio test.

Nonintinuitive Group Membership. Our motivation for studying mixture modes in this chapter was clustering. But one should be aware that mixtures can exhibit unexpected behavior with respect to clustering. Let

$$p(x) = (1-\pi)\phi(x; \mu_1, \sigma_1^2) + \pi\phi(x; \mu_2, \sigma_2^2).$$

Suppose that $\mu_1 < \mu_2$. We can classify an observation as being from cluster 1 or cluster 2 by computing the probability of being from the first or second component, denoted $Z = 0$ and $Z = 1$. We get

$$\mathbb{P}(Z = 0|X = x) = \frac{(1-\pi)\phi(x; \mu_1, \sigma_1^2)}{(1-\pi)\phi(x; \mu_1, \sigma_1^2) + \pi\phi(x; \mu_2, \sigma_2^2)}.$$

Define $Z(x) = 0$ if $\mathbb{P}(Z = 0|X = x) > 1/2$ and $Z(x) = 1$ otherwise. When σ_1 is much larger than σ_2, Figure 9 shows $Z(x)$. We end up classifying all the observations with large X_i to the leftmost component. Technically this is correct, yet it seems to be an unintended consequence of the model and does not capture what we mean by a cluster.

Improper Posteriors. Bayesian inference is based on the posterior distribution $p(\psi|X_1,\ldots,X_n) \propto \mathscr{L}(\psi)\pi(\psi)$. Here, $\pi(\psi)$ is the prior distribution that represents our knowledge of ψ before seeing the data. Often, the prior is improper, meaning that it does not have a finite integral. For example, suppose that $X_1,\ldots,X_n \sim N(\mu,1)$. It is common to use an improper prior $\pi(\mu) = 1$. This is improper because

$$\int \pi(\mu)d\mu = \infty.$$

Nevertheless, the posterior $p(\mu|\mathscr{D}_n) \propto \mathscr{L}(\mu)\pi(\mu)$ is a proper distribution, where $\mathscr{L}(\mu)$ is the data likelihood of μ. In fact, the posterior for μ is $N(\overline{X}, 1/\sqrt{n})$ where $\overline{x}$ is the sample mean. The posterior inferences in this case coincide exactly with the frequentist inferences. In many parametric models, the posterior inferences are well defined even if the prior is improper and usually they approximate the frequentist inferences. Not so with mixtures. Let

$$p(x;\mu) = \frac{1}{2}\phi(x; 0, 1) + \frac{1}{2}\phi(x; \mu, 1). \tag{8}$$

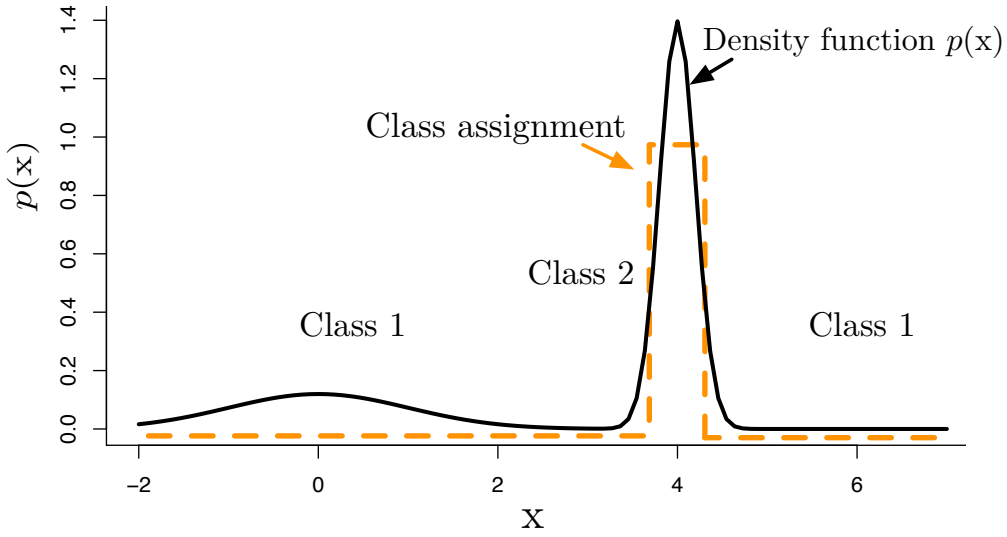

Figure 9: Mixtures are used as a parametric method for finding clusters. Observations with $x = 0$ and $x = 6$ are both classified into the first component.

If $\pi(\mu)$ is improper then so is the posterior. Moreover, Wasserman (2000) shows that the only priors that yield posteriors in close agreement to frequentist methods are data-dependent priors.

Use With Caution. Mixture models can have very unusual and unexpected behavior. This does not mean that we should not use mixture modes. Indeed, mixture models are extremely useful. However, when you use mixture models, it is important to keep in mind that many of the properties of models that we often take for granted, may not hold.

References

Arias-Castro, E., Mason, D. and Pelletier, B. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. Unpublished Manuscript, 2013.

Chacon, J. Clusters and water flows: a novel approach to modal clustering through Morse theory. arXiv preprint arXiv:1212.1384, 2012.

Chacon, J. and Monfort, P. A comparison of bandwidth selectors for mean shift clustering. arXiv preprint arXiv:1310.7855, 2013.

Chacon, J. and Duong, T. Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. Test, 19(2):375-398, 2010.

Chacon, J. and Duong, T. and Wand, M. Asymptotics for general multivariate kernel density

derivative estimators. Statistica Sinica, 21:807-840, 2011.

Chacon, J. and Duong, T. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. Electronic Journal of Statistics, 7:1935-2524, 2013.

Chazal, F., Guibas, L.J., Oudot, S.Y. and Skraba, P. Persistence-based clustering in riemannian manifolds. In Proceedings of the 27th annual ACM symposium on Computational geometry, pages 97-106. ACM, 2011.

Comaniciu, D. and Meer, P. Mean shift: a robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(5):603-619, may 2002. ISSN 0162-8828. doi: 10.1109/34.1000236.

Donoho, D. and Liu, R. Geometrizing rates of convergence, III. The Annals of Statistics, pages 668-701, 1991.

Carreira-Perpinan, M. (2006). Fast nonparametric clustering with Gaussian blurring mean-shift. Proceedings of the 23rd international conference on Machine learning. 153–160.

Silverman, B. (1981). Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society. Series B (Methodological), pages 97-99, 1981.

Fasy, Lecci, Rinaldo, Wasserman, Balakrishnan and Singh (2013). Statistical Inference For Persistent Homology. arXiv:1303.7117.

Wasserman, L. (2000). Asymptotic inference for mixture models by using data-dependent priors. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62, 159–180.