Lecture 20: Spectral Methods

Akshay Krishnamurthy
akshay@cs.umass.edu

November 13, 2017

1 Recap

Last time we saw the spectral clustering algorithm, which used the eigenvectors of some matrix to identify clusters in data. Today we'll see another *spectral* method for learning latent variable models. As in spectral clustering, these methods are based on extracting eigenvectors for certain matrices build from the data.

2 Gaussian Mixture Model

For simplicity we will focus on a simple Gaussian Mixture Model. Consider a mixture of k spherical gaussians in $\mathbb{R}^d$ which is the following generative process. Let $w \in \Delta([k])$ denote a distribution and let $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ be the mean vectors. Each point x_i is generated by first choosing a component $h_i \sim w$ and then $x_i \sim \mathcal{N}(\mu_{h_i}, I)$. We are given n samples $x_1, \ldots, x_n$ drawn according to this process.

Such models are called *latent variable models* since the component assignments h_i are latent or unobserved. The standard approach of maximum likelihood estimation is typically intractable in a latent variable model due to this unobservability. Specifically, the negative log likelihood here is

$$\mathcal{L}(w, \mu_1, \ldots, \mu_k; x_{1:n}) = \sum_{i=1}^{n} -\log\left(\sum_{j=1}^{k} w_j \mathcal{N}(x_i; \mu_j, I)\right)$$

In general this is a non-convex optimization, in part because there is interplay between the mixture weights w and the parameter μ. Or maybe more obviously, there are clearly many optima since if we permute the means and the weights in the same way we get the same log-likelihood. Just to contrast if we just had one component, the log-likelihood would be

$$\mathcal{L}(\mu; x_{1:n}) = \sum_{i=1}^{n} \log(\mathcal{N}(x_i; \mu_j, I)) = \sum_{i=1}^{n} \|x_i - \mu_j\|_2^2/2 + \frac{d}{2}\log(2\pi)$$

which is clearly a convex function of μ.

Given that in the mixture model the log-likelihood is non-convex, how should we go about estimating the parameters? Probably in 689 you saw the EM algorithm, which is an iterative method for parameter estimation in latent variable models that is based on optimizing the log-likelihood. This method does have some theoretical guarantees, but today we'll discuss another approach based on the *method of moments*.

2.1 Method of Moments

At a high level, the method of moments amounts to solving a system of polynomial equations based on the moments in the data. The idea is that we want to find a distribution P_θ where $\theta = (w, \mu_1, \ldots, \mu_k)$ such that the true moments:

$$\forall j \in \mathbb{N}, \mathbb{E}_{x \sim P_\theta}\left(\bigotimes_{i=1}^{j} x\right)$$

agree with the observed moments in the data. Here I am using $\otimes$ to denote outer product, which is a tensor with j modes and d dimensions per mode. Intuitively this make sense since the parameters of the model influence the moments.

As an example, let us calculate the moments for the Gaussian Mixture Model with parameters $w, \mu_1, \ldots, \mu_k$.

Lemma 1. *For the GMM, we have*

$$M_1 \triangleq \mathbb{E}[x] = \sum_{i=1}^{k} w_i \mu_i$$

$$M_2 \triangleq \mathbb{E}[x \otimes x] - I = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i^T$$

$$M_3 \triangleq \mathbb{E}[x \otimes x \otimes x] - \sum_{i=1}^{d} (\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x]) = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$$

Proof. The first moment is obvious. For the second moment

$$\mathbb{E}[x \otimes x] = \sum_{i=1}^{k} w_i \mathbb{E}_{z \sim \mathcal{N}(0,I)} (\mu_i + z) \otimes (\mu_i + z) = \sum_{i=1}^{k} w_i (\mu_i \otimes \mu_i + I)$$

and the conclusion follows. Similarly for the third moment

$$\begin{aligned} \mathbb{E}[x \otimes x \otimes x] &= \sum_{i=1}^{k} w_i \mathbb{E}_{z \sim \mathcal{N}(0,I)} (\mu_i + z) \otimes (\mu_i + z) \otimes (\mu_i + z) \\ &= \sum_{i} w_i \mu_i \otimes \mu_i \otimes \mu_i + \sum_{i=1}^{k} \sum_{j} w_i \mu_i \otimes e_j \otimes e_j + e_j \otimes \mu_i \otimes e_j + e_j \otimes e_j \otimes \mu_i \end{aligned}$$

Here we have to multiply out all the terms and whenever there is just one z in the outer product we get zero. If there are two zs then, applying expectation they become the identity, which we rewrite as $\sum_j e_j \otimes e_j$. □

The point here is that the weights w and the means μ_i are hidden in the moment tensors. Of course we cannot recover the parameters from just the first moment, since we have too many degrees of freedom. Even with the second moment it seems tricky. While we expect the matrix to be symmetric, if the μs are not orthogonal, it's not clear how we can extract them from the matrix. If they were orthogonal, we could take an eigendecomposition but even here we might lose some information due to the symmetries. Just for intuition if $\mu_1, \ldots, \mu_k$ were orthogonal, then collecting them a columns of matrix $V \in \mathbb{R}^{d \times k}$ and with Λ a diagonal matrix with w_i on the diagonal, we can write

$$M_2 = V \Lambda V^T.$$

This is almost the eigendecomposition. One issue is that we did not ask for μs to be unit normed, so once we do that we will lose the w_i information. Similarly if μ is an eigenvector than so is $-\mu$ so we might lose the sign information. This information should be recoverable by looking at the first moment in the orthogonal case, but let's turn to a much more general approach based on the third moment.

3 Tensors and decompositions

We focus only on symmetric third-order tensors. An orthogonal decomposition of a symmetric 3-tensor $T \in \mathbb{R}^{d \times d \times d}$ is a collection of orthonormal vectors $v_1, \ldots, v_k$ and positive scalars $\lambda_i > 0$ such that

$$T = \sum_{i=1}^{k} \lambda_i v_i \otimes v_i \otimes v_i$$

This first thing to observe since we have a third order tensor, we preserve sign information. We cannot flip the sign of the vector v_i without also flipping the sign of λ_i (which is why we require it to be positive). In general tensor decompositions are very delicate and a lot of intuition from the matrix case breaks down. But since we are working with orthogonal decompositions things will be much simpler.

Fixed point characterization. Like we saw last time in the matrix case, we can view a tensor T as a map $T : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$T(u) = T(I, u, u) = \sum_{i=1}^{d} \sum_{j,k} T_{i,j,k}(e_j^T u)(e_k^T u)e_i$$

This collapses two modes of the tensor by taking tensor-vector product with u. The fixed point characterization of a tensor eigenvector is

$$T(u) = \lambda u$$

It is easily verified that if T has orthogonal decomposition then the vectors in the decompositions are the eigenvectors. However since the map $T(u)$ is non-linear, even with multiplicity, a linear combination of eigenvectors will in general no longer be an eigenvector. However if λ_i, v_i are the eigenvalue/vector pairs, then for any subset $S \subset [k]$ the vector $u = \sum_{i \in S} v_i/\lambda_i$ will be an eigenvector. Thus the decompositions are not necessarily unique.

These combinations are in a sense spurious and they are not the robust fixed points of the map $T(I, u, u)$, although they are legitimate fixed points. Instead, we say that a unit vector u is a robust eigenvector if there exists some $\epsilon > 0$ such that for all $\theta \in \{u' \mid \|u' - u\| \leq \epsilon\}$ repeated iteration of

$$\bar{\theta} \to \frac{T(I, \bar{\theta}, \bar{\theta})}{\|T(I, \bar{\theta}, \bar{\theta})\|}$$

starting from θ converges to u. If T has orthogonal decomposition, then these are precisely the robust eigenvectors.

Variational Characterization. We may also generalize the Rayleigh quotient to

$$\frac{T(u, u, u)}{(u^T u)^{3/2}}$$

Theorem 2. *Let T have orthogonal decomposition and consider*

$$\max_u T(u, u, u) \ \textit{s.t.} \ \|u\| \leq 1$$

The stationary points are eigenvectors of T and a stationary point is an isolated local maximizer if and only if $u = v_i$ for some v_i in the orthogonal decomposition.

The reduction. Why does looking at the third-order tensor help us in identifying the parameters in the GMM? The idea is that after suitable transformation, we can construct a 3-tensor $\tilde{M}_3$ that has orthogonal decomposition whose components are closely related to the means and whose eigenvalues are essentially the weights. However, as we have expressed M_3 there is no requirement that μ_is are orthogonal, so we need to *whiten*. First let $W \in \mathbb{R}^{d \times k}$ be a matrix such that

$$W^T M_2 W = I_{k \times k}$$

In the population case we can take eigendecomposition of $M_2 = UDU^T$ and write $W = UD^{-1/2}$, where $U \in \mathbb{R}^{d \times k}$. Now write $\tilde{\mu}_i = \sqrt{w_i} W^T \mu_i$ and observe that

$$W^T M_2 W = \sum_{i=1}^{k} W^T (\sqrt{w_i}\mu_i)(\sqrt{w_i}\mu_i^T)W = \sum_{i=1}^{k} \tilde{\mu}_i \tilde{\mu}_i^T = I$$

Since we have shown that the sum of k outer products is the identity matrix, this implies that the vectors $\tilde{\mu}_i$ must be orthonormal.

Now multiply each mode of M_3 by W to define $\tilde{M}_3 = M_3(W, W, W) \in \mathbb{R}^{k \times k \times k}$ as

$$\tilde{M}_3 = \sum_{i=1}^{k} w_i \bigotimes_{j=1}^{3} (W^T \mu_i) = \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \tilde{\mu}_i \otimes \tilde{\mu}_i \otimes \tilde{\mu}_i$$

This matrix $\tilde{M}_3$ has orthogonal decomposition that is the $\tilde{\mu}_i$s. Moreover, we have the following theorem.

Theorem 3. *Under a non-degeneracy condition that $\mu_1, \dots \mu_k$ are linearly independent and $w_i > 0$ are strictly positive, the set of robust eigenvectors of $\tilde{M}_3$ are exactly $\{\tilde{\mu}_1, \dots, \tilde{\mu}_k\}$ and the eigenvalues are $1/\sqrt{w_i}$. Moreover the true means μ_i can be easily recovered by $(W^T)^\dagger \tilde{\mu}_i / \sqrt{w_i}$.*

Thus we see how in the population limit we can use the tensor decomposition to extract the means μ_i and the weights w_i.

4 Perturbation analysis of tensor power method

So far our discussion has focused on the population case, where we can build the moments M_1, M_2, M_3 exactly. What should we do when there is noise? Also while we have said that the orthogonal decomposition of $\tilde{M}_3$ reveals the parameters, how do we compute the orthogonal decomposition. The idea is to do a form of *robust power iteration*, which is something you might have seen for computing eigenvectors in a matrix.

Let us assume we have a three-tensor $\hat{T}$ that we have estimated from the data. The idea is to use the fixed point characterization to find the robust eignevectors. Specifically, we draw θ_0 uniformly from the unit sphere and repeat the iteration

$$\theta_{t+1} \leftarrow \frac{\tilde{T}(I, \theta_t, \theta_t)}{\|\tilde{T}(I, \theta_t, \theta_t)\|}$$

Do this with many different random initializations and pick the best one to get $\hat{\theta}$. Then set $\hat{\lambda} = T(\hat{\theta}, \hat{\theta}, \hat{\theta})$ and deflate by updating $\tilde{T} \leftarrow \tilde{T} - \hat{\lambda} \hat{\theta}^{\otimes 3}$. Then repeat the whole thing with this deflated $\tilde{T}$.

This algorithm has a guarantee that is similar to Davis-Kahan and Weyl's theorem, if we write $\tilde{T} = T + E$ where T has orthgonal decomposition $\lambda_1, \dots, \lambda_k$ and $v_1, \dots, v_k$ then the algorithm finds vectors $\hat{v}_i, \hat{\lambda}_i$ such that

$$\|v_i - \hat{v}_i\| \leq \frac{8\|E\|}{\lambda_i}, \qquad |\lambda_j - \hat{\lambda}_j| \leq 5\|E\| \qquad \|T - \sum_j \hat{\lambda}_j \hat{v}_j^{\otimes 3}\| \leq 55\|E\|.$$

This requires some further assumptions and setting the parameters correctly and so one, so I am not stating the formal theorem here. The main point is, now we can analyze these *spectral methods* for latent variable models. Specifically we can use matrix and tensor concentration inequalities to understand what happens when we whiten M_3. This can lead to precise sample complexity guarantees for learning mixtures of gaussians. One thing to note here is that these methods require a robust notion of linear independence, in that they require $\lambda_{\min}(T)$ to be large. Or rather the sample complexity will depend on $\lambda_{\min}^{-1}$. Thus this method does not work when $k \gg d$, which could happen in some applications. I believe there is some work on generalizing to $k \geq d$, I will try to find some pointers.

Note that this approach also applies to a number of other latent variable models including HMMs, LDA, GMMs with non-spherical covariance etc.

5 Other results on GMMs

Let me mention some other related results on GMMs.

TCS results. Separating gaussian mixtures has received much attention from the theoretical computer science community. I think the culmination is a paper due to Moitra and Valiant. First, in the case where $k \geq d$ they prove that in general there is an exponential lower bound on the sample complexity

Theorem 4. *There exists two GMMs F, F' of k-components both of which are well behaved (large w_i, separated components etc.) each such that $TV(F, F') \leq O(\exp(-k/30))$, but for which F is not a good estimate for F' (in the sense that $|w_i - w_i'| \geq 1/4$ or $TV(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_i' \Sigma_i')) \geq 1/4$).*

This shows that in the high-component case (which is not considered by the spectral method above), it may require $\exp(k)$ samples to learn a mixture of gaussians.

In the same paper by Moitra and Valiant they also give an algorithm that learns a mixture of k components (without any other assumptions) with computational and sample complexity that are polynomial in n and $1/\epsilon$, but could have exponential dependence on k (which by above is necessary).

EM Analysis. Here we consider maximizing the log likelihood

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \sum_j w_j \mathcal{N}(x_i; \mu_j, I)$$

The EM algorithm iterates by maximizing

$$\theta^{t+1} = \operatorname*{argmax}_{\theta} Q(\theta | \theta^t) = \frac{1}{n} \sum_{i=1}^{n} \sum_j p(h_i = j | w^t, \mu^t, x_i) \log w_j \mathcal{N}(x_i; \mu_j, I)$$

where $p(h_i = j | w^t, \mu^t, x_i)$ is the fractional memberships for point i in component j using the parameters w^t, μ^t.

For a simple mixture of two gaussians, Balakrishnan, Wainwright and Yu prove that the EM algorithm, when initialized from a reasonable location converges with linear rate

$$\|\mu^t - \mu^\star\|_2 \leq \kappa^t \|\mu^0 - \mu^\star\|_2 + O(\sqrt{d \log(1/\delta)/n})$$

under some assumptions.

The main advantage of EM is that it is more robust to model misspecification. The maximum likelihood approach makes sense even when the data is not actually generated from a mixture model (or in general from the model that you are using). In general, the MLE computes the parametric distribution that is closest in KL-divergence to the observed empirical distribution. In a sense this is like moving to the agnostic learning case.

Unfortunately spectral methods can perform quite poorly under model misspecification since they rely more heavily on the modeling assumptions (e.g., to construct the moments etc.). One lesson here is that optimization-based methods may be more effective in practice since they make sense even when the modeling assumptions are not satisfied.