ECE598: Representation Learning: Algorithms and Models Fall 2017

Lecture 6: Generative Models: Mixture of Gaussians

Lecturer: Prof. Pramod Viswanath Scribe: Moitreya Chatterjee, Tao Sun, WZ, Sep 14, 2017

What I cannot create, I do not understand - Richard Feynman.

6.1 Generative Models

In the context of probability and statistics, *Generative Models* are approaches that model all random variables associated with a phenomenon, both those that can be observed as well as the unobserved ones [Bis06].

Note: This is in contrast to *Discriminative Models* where only the dependence of the unobserved variables conditioned on the observed ones is modeled.

The contrast is exemplified by the following classification task. Let's assume that we are given data, $\mathcal{X}$, i.e. the observed variable and we want to determine its class label, $\mathcal{Y}$, the unobserved (target) variable. A generative classifier, such as Naive Bayes, makes use of the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$, i.e. $\mathrm{P}(\mathcal{X}, \mathcal{Y})$ to perform this inference. While a discriminative classifier, such as a Logistic Regression, would directly model the posterior class probabilities, i.e. $\mathrm{P}(\mathcal{Y}|\mathcal{X})$ to perform this inference.

Feynman's quote cited above embodies the spirit of generative models elegantly, suggesting that generative models give us a complete understanding of the distribution underlying the data.

6.2 Parametric vs Non-Parametric Generative Models

Statisticians represent *Generative Models*, either with functions which have a specific form and are defined by a set of parameters, say θ or using models that are free to learn any functional form depending on the data. The former class of models are called *Parametric Models*, for example a Gaussian (parameterized by μ and σ), while the latter category of models are often referred to as *Non-parametric Models*, of which the k-Nearest Neighbor is a popular example.

Note: Non-parametric models do not imply that the model is free from any parameters but that the set of parameters are flexible and not preset in advance.

This lecture primarily focuses on *Parametric* modeling techniques in the context of *Generative Models*.

6.3 Single Gaussian Model

Very often, statisticians model data from the real world using standard probability distributions. The Gaussian Distribution is a particularly popular choice, in this regard. This is so because of the following two reasons:

- Firstly, according to the *Central Limit Theorem*, the normalized sum of N, independent and identically distributed (i.i.d.) random variables, $\{x_i\}_{i=1}^{N}$ approaches a Gaussian distribution as $N \to \infty$ [Haj15].
- Secondly, the Gaussian distribution is parameterized by only two parameters, its mean μ and its variance σ^2.

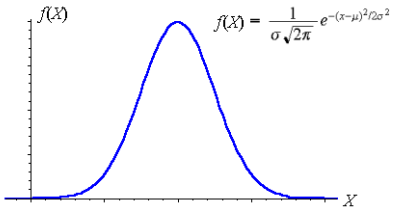

Figure 6.1: A Gaussian Distribution with mean μ and variance σ^2.

The probability density function of a Gaussian Random Variable $\mathcal{X} \sim \mathcal{N}(x; \mu, \sigma^2)$ is given by the following:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Figure 6.3 shows an example of a Gaussian Distribution.

Note: The density of the Gaussian Random Variable is symmetric.

6.4 Maximum Likelihood Estimate

Given a set of N observations, often times we assume them to be arising out of N independent and identically distributed (i.i.d.) random variables, say $\mathcal{X} = \{\mathcal{X}_i\}_{i=1}^{N}$, for the ease of modeling. Parametric Models seek to choose/tune their parameters so as to maximize the joint distribution of these random variables, i.e. the data. One common approach to this optimization task is to maximize the statistic known as *Maximum Likelihood*.

The Maximum Likelihood is defined as : $P(\mathcal{X}|\theta)$, where θ is the set of parameters of the model [Bis06]. Now, in the context of a model parameterized by a single Gaussian, where the data is assumed to be i.i.d. Gaussian, the computation becomes:

$$L = P(\mathcal{X}|\theta) = \prod_{i=1}^{N} P(\mathcal{X}_i|\theta),$$

where θ is assumed to be a vector of the mean and the variance of the distribution.

Now, since log is a monotonically increasing function, we can alternatively solve for the logarithm of the likelihood without altering the optimum set of parameters.

$$\log L = \sum_{i=1}^{N} \log \mathcal{N}(\mathcal{X}_i; \mu, \sigma^2)$$

$$= \sum_{i=1}^{N} [\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i - \mu)^2}{2\sigma^2}]$$

$$= N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

Optimizing the above equation for μ by differentiating it and setting it to 0, yields, $\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$. Likewise, the estimator for the variance is given by: $\hat{\sigma^2}_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu}_{ML})^2$.

Note: From the above derivation, we observe that the Maximum Likelihood estimate for both the mean and the variance are just the sample mean and the sample variance.

6.5 The Need for More Than One Gaussian

While the ease of modeling, is often a key reason why statisticians model data with one Gaussian but at times that turns out to be an over-simplification of the distribution underlying the data.

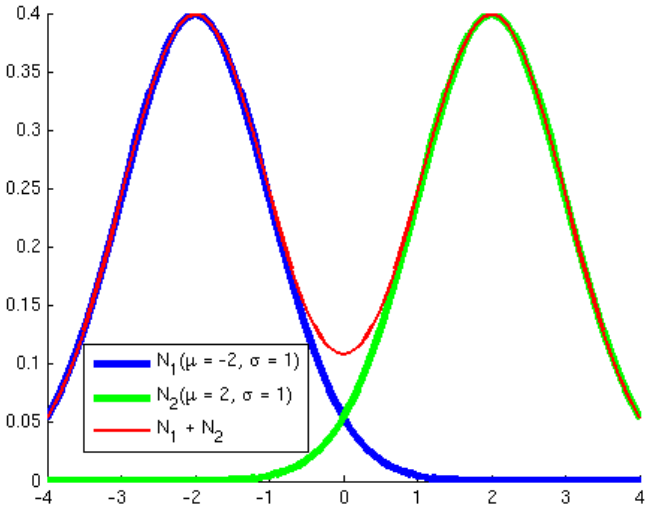

Figure 6.2: A Gaussian Distribution with mean μ and variance σ^2.

In Figure 6.5, assume that the red curve represents the distribution of the actual data. Now, clearly this data cannot be modeled by a single Gaussian. Two partially overlapping Gaussians (N_1 and N_2) seem to represent the data better. Such models are known as *Mixture Models*.

Thus, the likelihood of a sample $\mathcal{X}_i$, existing in such a space is given by:

$$P(\mathcal{X}_i|\theta) = \pi_{N_A} \mathcal{N}(x_i; \mu_A, \sigma_A^2) + \pi_{N_B} \mathcal{N}(x_i; \mu_B, \sigma_B^2),$$

where π_{N_A} and π_{N_B} are the *Mixing Probabilities* of the sample, representing how likely is the sample to have originated from one distribution over the other. They are constrained to sum to 1, i.e. $\pi_{N_A} + \pi_{N_B} = 1$. The other symbols carry their usual meaning.

6.5.1 Real World Example

Imagine in microscopic scale, a piece of graphene (a star material today) was deposited onto HfO_2 (a high-dielectric material widely used in semiconductor devices) substrate and the heights of the integrated system was measured with atomic force microscopy (AFM). The height distributions of graphene and the substrate follow Gaussian distributions respectively, with different means and possibly different variance. The height distribution of the integrated system obtained by imaging randomly chosen points on the sample surface shows a skew configuration. Now our question is: how to justify the distance between graphene and HfO_2? Or how to extract the difference between the means of the two hidden Gaussians with the measured samples?

As illustrated in Figure 6.1, this example is a mixture of Gaussians: distributions where we have several groups and the data within each group is normally distributed.

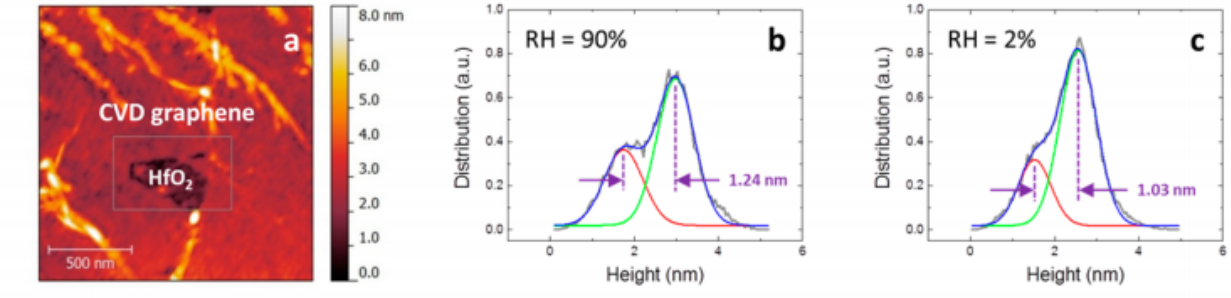

Figure 6.3: Atomic force microscopy (AFM) height measurements of graphene on HfO_2 under different relative humidity (RH). (a) AFM image of graphene on HfO_2. (b) Height distribution for the indicated rectangular region in (a) under RH = 90%. (c) Height distribution for the same region under RH = 2%. (Figure adapted from [OMS+15])

6.6 Method of Moments for Parameter Estimation

The above examples justify our intuition that the expressivity of the mixture models is strictly better than that of the individual Gaussians. In practice, however, one needs to learn the parameters of these underlying models to be able to obtain a better understanding and a compact representation of the data. So now the the key question that arises is given a set of examples, distributed as mixture of Gaussians, *how do we estimate all the parameters of the model*?

One of the earliest approaches to address this question of estimating parameters was proposed by [Pea94]. His proposed approach was presented for the task of dissecting a frequency curve into N Normal (Gaussian) curves and used the full set of examples in the dataset for the estimation task.

Figure 6.6 shows the plot of a function. Now, the N^{th} order moment of this function is given by:

$$M_N = \int x^N y \, \mathrm{dx}$$

Solving this integral corresponds to obtaining a scaled representation of the area under the curve.

Pearson proposed that in order to solve for the parameters of the mixture model, we could utilize these moment equations [Pea94]. According to his proposed algorithm, one needed as many moment

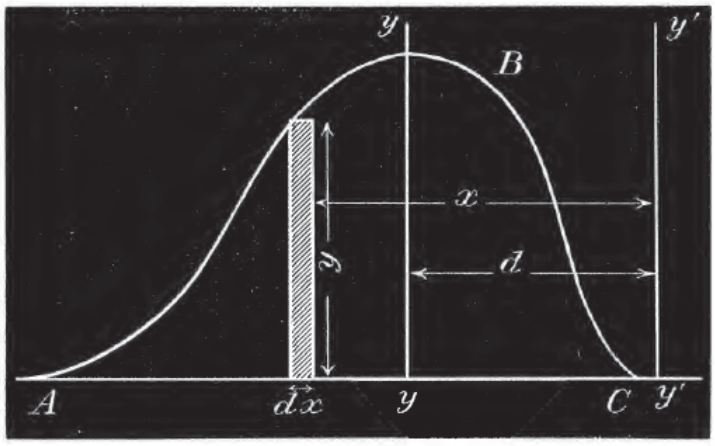

Figure 6.4: A Curve showing the computation procedure of moments.

equations as there are unknowns, to recover the parameters. Solving the system typically results in more than one solution for the system of equations. Thus, one additional higher moment equation is needed to establish the final result. He used this algorithm to recover the parameters of a 2-Gaussian Mixture Model from 6 (5+1) moment equations. Mathematical details regarding this procedure can be found in Section 6.10.1.

However, this is hard to scale for algorithms with more than a handful of parameters or when the system resides in a high-dimensional space. Besides, the solutions to such system of equations is also pretty involved and increases in complexity with higher order moments. Typically, to estimate the m^{th} moment to ϵ accuracy, we need $(\frac{1}{\epsilon})^m$ samples, a fairly large number for a high m. To allay these concerns, we delve into another, more efficient algorithm called *Expectation-Maximization*.

6.7 Expectation-Maximization Algorithm

The *Expectation-Maximization* algorithm (EM-Algorithm) is an iterative computational technique for computing the maximum likelihood of the data in contexts where there are hidden random variables, in addition to observed data and unknown parameters [Haj15].

Let θ be the set of parameters to be estimated, Y be the set of observed data, Z be the unobserved or hidden component of the data, X be the complete data defined as X = (Y, Z), h(.) be a function that maps the complete data to its observed part, i.e. Y = h(X), $p_{cd}(X|\theta)$ be the likelihood of the complete data, then the EM-Algorithm involves the following two steps:

- E-Step: In this step, we compute:

$$Q(\theta|\theta^k) = E[\log p_{cd}(X|\theta)|Y, \theta^k]$$

- M-Step: In this step, we compute:

$$\theta^{(k+1)} \leftarrow \arg\max_{\theta} Q(\theta|\theta_k)$$

The algorithm is initialized with some values of the parameter θ and then the E and the M steps are undertaken iteratively. The algorithm is known to converge to stationary points, relatively quickly in several scenarios.

6.8 Applying EM to estimate the Parameters of a GMM

Now, let us set up the algorithm to estimate the parameters of a 2-Gaussian Mixture Model using the EM setting.

Say the data is generated from a mixture of 2-Gaussians distributed on 1-dimension. Then the set of all data points $\mathcal{X} = \{x_i\}_{i=1}^N$ and each of the points is distributed as:

$$P(x_n) = \sum_{k=1}^{2} \pi_k \mathcal{N}(x_n|\mu_k, \sigma_k),$$

where μ_k, σ_k denote the parameters of the k^{th} Gaussian and π_k is the mixing probability associated with the corresponding Gaussian.

So the log-likelihood of the complete data is given by:

$$\log P(X|\theta) = \sum_{n=1}^{N} \log\{\sum_{k=1}^{2} \pi_k \mathcal{N}(x_n|\mu_k, \sigma_k)\}$$

In EM-algorithm, we seek to maximize this quantity. So, taking the derivative w.r.t. μ and setting it to 0, we have:

$$0 = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \sigma_k)}{\sum_{j=1}^{2} \pi_j \mathcal{N}(x_n|\mu_j, \sigma_j)} (\sigma_k)^{-1}(x_n - \mu_k)$$

Let us call the highlighted part $\gamma(z_{nk})$, or the *Responsibility* of distribution k towards sample n.

Thus, taking μ_k over to the other side of the above equation leads to the solution:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n,$$

where we define:

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

Similarly, taking the derivative w.r.t. σ_k, we get:

$$\sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)^2$$

To solve for the mixing probabilities π_k, we can make use of the constraint $\sum_k \pi_k = 1$ and combine it with the optimization objective $\log P(X|\theta)$ and obtain a Lagrangian. Solving this Lagrangian yields that:

$$\pi_k = \frac{N_k}{N}$$

Note: Estimating the parameters involves computing the responsibilities, which is why the above equations may not be thought of as closed form solutions and one needs to use the EM algorithm to solve them.

So we start with an initial estimate of the parameters and undertake the following steps [Bis06].

- **E-Step** : With the current estimate of the parameters, estimate the responsibility, $\gamma(z_{nk})$.
- **M-Step** : With current estimate of the responsibility, compute the parameters of the GMM Model.

A more general treatment of the EM algorithm and the mathematical details and its derivation can be found in

6.9 EM Convergence on Mixtures of two Gaussians

The EM algorithm is based on the maximum likelihood principle, which is only guaranteed to find stationary points. However, these stationary points may be far away from the optimal points.

In this subsection, we are particularly interested in discussing EM convergence on mixture of only two Gaussians, since it is one of the simplest model. We will try to recover parameters for the two Gaussians (μ_1, μ_2, Σ_1, Σ_2) and the weight parameter (w). In total, there are 5 parameters to be estimated.

Initially, Pearson *et.al.*, [Pea94] proposed to use 5 moments to identify the parameters. 5 moments will give 5 equations and there are 5 parameters to be solved. However, there are 2 valid solutions for the 5 equations. Pearson *et.al.* then selected the one whose 6-th moment was closest to the observed empirical 6-th moment. The method works well in the crab dataset.

However, Pearson *et.al.*, did not give any analysis on the above method. It isn't proved to be generalized to other examples. Besides, no efficiency matter has been discussed. In order to tackle the problems, Hardt and Price [HP15] propose to give a computationally efficient estimator to solve the problem. It achieves an optimal bound on the number of samples. It turns out that their proposed method and Pearson's method from more than 100 year ago are very close. Hardt and Price's results can be viewed as showing that Pearson's method is actually an optimal solution to the problem, which has not yet been shown at that time. Hardt and Price also find out that $\Theta(\sigma^{12})$ samples are necessary and sufficient to estimate each parameter to constant additive distance, where σ^2 is the overall variance of the mixture. The estimation method is computationally efficient and can be generalized to d-dimensional with a necessary factor of $O(\log d)$ in the sample requirement.

In the work by Xu *et.al.* [XHM16], global convergence properties of EM algorithm on mixture of two Gaussians are studied. This paper, in particular, has discussed two popular models. They are: 1) Samples x are from the mixture distribution $0.5N(-\theta^*, \Sigma) + 0.5N(\theta^*, \Sigma)$, where Σ is a known covariance matrix and θ^* is the unknown parameter that we're estimating; 2) Samples x are from the mixture distribution $0.5N(\mu_1^*, \Sigma) + 0.5N(\mu_2^*, \Sigma)$, where Σ is a known covariance matrix and (μ_1^*, μ_2^*) is the unknown parameter that we're estimating. Xu *et.al.* has proved, in that paper, convergence for the sequence of iterates for Population EM from each model: 1)the sequence of $\theta^{(t)}$ converges to either θ^*, $-\theta^*$ or 0; 2) the sequence $(\mu_1^{(t)}, \mu_2^{(t)})$ converges to either (μ_1^*, μ_2^*), (μ_2^*, μ_1^*)

or $((\mu_1^* + \mu_2^*)/2, (\mu_1^* + \mu_2^*)/2)$. The paper has also proved that the limits of the Sample-based EM iterates converge in probability to the unknown parameters of interest, as long as Sample-based EM is initialized at points where Population EM would converge to these parameters as well.

Daskalakis *et.al.* [DTZ16] have shown that the population version of EM, where the algorithm is given access to infinitely many samples from the mixture, converges geometrically to the correct mean vectors. In addition, they provide simple, closed-form expressions for the convergence rate. As the title of the paper, the paper gives an illustration of mixture of two one-dimensional Gaussians. It has been proved that ten steps of the EM algorithm initialized at infinity result in less than 1% error estimation of the means. While in the finite sample regime, we show that, under a rondom initialization, $O(\sqrt{d/\epsilon^2})$ samples suffice to compute the unknown vectors to within ϵ in Mahalanobis distance, where d is the dimension. In particular, the error rate of the EM based estimator is $O(\sqrt{d/n})$ where n is the number of samples, which is optimal up to logarithmic factors.

In this class, we thus explored an assortment of tools and algorithms for generative modeling using parametric techniques.

6.10 *Additional reading

6.10.1 Applying Method of Moments to Two Gaussian Mixtures

The earliest work of using MoM to solve the two Gaussian mixture problem was proposed by Karl Pearson more than 120 years ago[Pea94]. The target distribution has the following form:

$$p(x) = \frac{c}{\sigma_1\sqrt{2\pi}} \exp\Big(-\frac{(x-\beta_1)^2}{2\sigma_1^2}\Big) + \frac{1-c}{\sigma_2\sqrt{2\pi}} \exp\Big(-\frac{(x-\beta_2)^2}{2\sigma_2^2}\Big)$$

the parameter set $\widehat{\theta} = (c, \beta_1, \sigma_1^2, \beta_2, \sigma_2^2)$ has 5 elements (here we use β_1 and β_2 for the distribution means, in order not to be confused with the distribution moment representation μ), we need to match the first 5 moments to gain sufficient conditions to solve for the 5 parameters.

The first 5 sample moments are calculated as

$$m_1 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)})^1, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)})^2, \quad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)})^3, \quad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)})^4, \quad m_5 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)})^5 \tag{6.1}$$

The first 5 distribution moments are calculated as

$$\begin{aligned}
\mu_1(\widehat{\theta}) &= c\beta_1 + (1-c)\beta_2 \\
\mu_2(\widehat{\theta}) &= c(\beta_1^2 + \sigma_1^2) + (1-c)(\beta_2^2 + \sigma_2^2) \\
\mu_3(\widehat{\theta}) &= c(\beta_1^3 + 3\beta_1\sigma_1^2) + (1-c)(\beta_2^3 + 3\beta_2\sigma_2^2) \\
\mu_4(\widehat{\theta}) &= c(\beta_1^4 + 6\beta_1^2\sigma_1^2 + 3\sigma_1^4) + (1-c)(\beta_2^4 + 6\beta_2^2\sigma_2^2 + 3\sigma_2^4) \\
\mu_5(\widehat{\theta}) &= c(\beta_1^5 + 10\beta_1^3\sigma_1^2 + 15\beta_1\sigma_1^4) + (1-c)(\beta_2^5 + 10\beta_2^3\sigma_2^2 + 15\beta_2\sigma_2^4)
\end{aligned} \tag{6.2}$$

By equating the corresponding element in Equation (6.11) and (6.12) and rearrange the variables, we obtain a 9th order polynomial equation. The detailed form of the equation could be found in [Pea94], in which some re-parametrization was performed.

As long as the polynomial equation is obtained, the real roots of the polynomial could be found using any proper numerical solvers, and the results may not be unique. In case of multiple candidate roots for the polynomial equation, the 6th order theoretical moments need to be calculated with the obtained parameter sets. The parameter set whose 6th order theoretical moment closest to the observed 6th order sample moment should be chosen as the final one.

6.10.2 Optimal Bound on Number of Samples Needed

By MoM Pearson successfully solved the 2 Gaussian mixture model. However, a general justification of the accuracy and efficiency of the approach concerning the requred number of samples, is lacked. Regarding the problem of 2 Gaussian mixtures, Hardt et al.[HP15] recently presented the upper and lower bounds giving a computationally efficient moment-based estimator with an optimal convergence rate. The one dimensional case turns out to be closely related to Pearson's approach, showing Pearson's original estimator to be an optimal solution to the problem he proposed.

Theorem 6.1. *Denoting by σ^2 the overall variance of the mixture, $\Theta(\sigma^{12})$ samples are necessary and sufficient to estimate each parameter to some constant error. [HP15, Har14]*

Bibliography

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[DTZ16] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. *arXiv preprint arXiv:1609.00368*, 2016.

[Haj15] Bruce Hajek. *Random processes for engineers*. Cambridge University Press, 2015.

[Har14] Moritz Hardt. Pearson's polynomial. http://blog.mrtz.org/2014/04/22/pearsons-polynomial.html, 2014.

[HP15] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760. ACM, 2015.

[OMS+15] Eric J Olson, Rui Ma, Tao Sun, Mona A Ebrish, Nazila Haratipour, Kyoungmin Min, Narayana R Aluru, and Steven J Koester. Capacitive sensing of intercalated h2o molecules using graphene. *ACS applied materials & interfaces*, 7(46):25804–25812, 2015.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[XHM16] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.