STATS 319 Final Report: EM Algorithm

Wenlong Ji & Dileka Gunawardana

March 2024

1 Introduction

The field of statistics is based on the ability to extract meaning and stories about ourselves from data. In statistical inference settings where data regarding certain members of a population of interest is missing, we can look to the EM algorithm to provide a theoretically-backed solution. It uses a powerful iterative technique to cleverly cycle between the features of the data that we have available. In this report, we seek to discuss the following seminal paper by AP Dempster, NM Laird, and DB Rubin: "Maximum Likelihood From Incomplete Data Via the EM Algorithm".

Although the algorithm has alternative names such as the Baum-Welch algorithm in addition to variants in prior work, this paper was the first to combine previous ideas and theory along with a clear and understandable framework. We begin with a high-level overview of the algorithm, then discuss its mathematical formalization, the theory underlying its ability to work, several examples, and conclude with interesting variants.

2 High-Level Overview of the EM Algorithm

Before delving into the math behind the EM algorithm, it will be of use to understand the high-level intuition behind how it works. Thus, we ask the reader to consider an imaginary "dream world" far from our own, where an analyst has access to not only the underlying parametric model but also a large sample of data from any population of interest. In this "dreamworld", one could posit that these two quantities share a certain level of information. Specifically, given a large sample of data from the population of interest, the best guess for our parametric model's mean would be the sample mean. For example, if we assumed that the parametric form of the distribution underlying the population was Normal$(x, 5)$, then the most reasonable guess for x would be the sample mean based on the observed data. This is the intuition underlying the **M-Step**.

Likewise, given the underlying parametric model, the best guess for a missing data point for some member of the population would be the mean defined for the model. For example, if we knew that the distribution underlying the population was $\sim$ Normal$(10, 5)$, then the most reasonable guess for the missing data point would be 10. This is the intuition underlying the **E-step**. Of course in most real world applications, neither of these two quantities are available. The EM algorithm assumes one of these "dream-world" settings to get a best guess for the other setting and then repeats for the reverse. This two-step process is then iteratively repeated.

3 Missing Data

Before introducing a formalization of the EM algorithm, it is essential to formally define the notion of "missing data". There are numerous data analysis settings where one must account for certain members of the population who do not have corresponding data observed. Some simple **examples** include:

- A factory sensor that malfunctions for a brief period of time
- Census data that isn't able to access particularly rural communities
- Genetic studies where have only access to data on a subset of the genes

We can represent this notion mathematically by defining a many-one mapping $\mathcal{X} \to \mathcal{Y}$ where $X \in \mathcal{X}$ is our world of **complete data**, and $Y(X) \in \mathcal{Y}$ is the world of **data we observe** (so $y(x)$ is simply the complete data x with some of the points missing). We assume a family of sampling densities $f(x|\phi)$, so the distribution corresponding to the missing data actually observed integrates over every possible dataset $x \in \mathcal{X}$ that would lead to that observation y:

$$g(y|\phi) := \int_{\mathcal{X}(y)} f(x|\phi)dx$$

Thus, g is able to mathematically account for every possible complete data set that would've led to the data that we actually observed. In the setting of the factory sensor example introduced above, x would be the true data from time 0 time 10 that would've been observed had the sensor not malfunctioned. y is the data that is observed that excludes the time-window where the sensor was malfunctioning (i.e., time 0 to 9 if the sensor broke at time 9). Thus, g integrates over every possible value that the sensor could've observed from time 9 to time 10 where each possible value is weighted according to the density f.

4 Formalization of the EM Algorithm

Consider the following function Q:

$$Q(\phi'|\phi) := \mathbb{E}[\log f(x|\phi')|y, \phi] = \int_{\mathcal{X}(y)} \log f(x|\phi') f(x|\phi) dx$$

We can think of Q as weighting the possible points that the "complete data" can be based on how likely they are under our current guess of ϕ. Then, we will eventually want to choose the best possible parametrization choice for ϕ' by finding the one that maximizes this average density based on our current best guess for the parametrization.

At each step $p = 1, 2, \ldots$ until we converge, the **EM algorithm** iterates between the following two steps:

1. **E-Step**: Compute $Q(\phi|\phi^{(p)})$
2. **M-Step**: Choose $\phi^{(p+1)}$ to be the value of $\phi \in \Omega$ that maximizes $Q(\phi|\phi^{(p)})$

We additionally mote that the EM algorithm has simple extensions to the **Bayesian** setting where we replace the MLE with the posterior mode in the M-Step.

5 Theory Underlying EM's Power

The main **idea** behind why the EM algorithm "works" is that every step of the algorithm essentially increases the value of the likelihood. Denote a step of EM by the function M, so $\phi^{(p+1)} = M(\phi^{(p)})$. Additionally, define the following three quantities:

- $k(x|y, \phi) := \frac{f(x|\phi)}{g(y|\phi)}$ is essentially how likely an option for the complete data set (x) is given the observed data (y) relative to all of the other options for the complete data
- $L(\phi) := \log g(y|\phi)$
- $H(\phi'|\phi) := \mathbb{E}[\log k(x|y, \phi')|y, \phi]$

Our goal is to show that $L(\phi)$ increases or stays the same with each step of the EM algorithm. It's clear from these definitions that $L(\phi) = \log f(x|\phi) - \log k(x|y, \phi)$. Taking expectations on both sides of this equation yields:

$$Q(\phi'|\phi) = L(\phi') + H(\phi'|\phi)$$

$$\mathbf{L}(\mathbf{M}(\phi)) - \mathbf{L}(\phi) = \{\mathbf{Q}(\mathbf{M}(\phi)|\phi) - \mathbf{Q}(\phi|\phi)\} + \{\mathbf{H}(\phi|\phi) - \mathbf{H}(\mathbf{M}(\phi)|\phi)\} \geq \mathbf{0}$$

The inequality above follows by the construction in the M-Step (for the first term involving Q) and Jensen's inequality (for the second term involving H).

6 Important Considerations with use of the EM Algorithm

The exciting theoretical underpinnings of the EM algorithm discussed above are based on two essential assumptions. We hi-light these assumptions below in order to help the reader understand the most appropriate settings for the EM algorithm:

- Parametric assumption on the population distribution

  Recall from the previous section that a major element the theory underlying the algorithm was an assumption on the parametric form of the distribution corresponding to the data (i.e., the f defined in the sampling densities $f(x|\phi)$, so they are of some form such as a normal, beta, gamma, etc.). Typically, we do

not have all of the information regarding our population of interest (including its parametric form) as this is often the motivation behind statistical inference in the first place. Thus, in most applications, we must settle for a parametric assumption that is reasonable given the information that we have regarding the population. However, it is important to note the possibility of a population distribution which does not follow the form of the parametric family being assumed, in which case the EM algorithm will not produce accurate results. As an example, consider the probability distribution associated with some population parameter of interest that is normal distributed (i.e., height). In this case, an exponential distribution assumption will lead to inaccurate estimates for the missing data points and inference. In settings where one is unsure of a reasonable parametric assumption, a sensitivity analysis may be of interest to consider how the results are affected by different parametric choices. If the results of the EM algorithm are similar across a class of reasonable parametric assumptions, then the setting may be appropriate.

- Missing data from population whose distribution is equivalent to that of the observed data

  Recall from the theory described above (notably in defining the Q function for the E-step) that it was assumed that the missing data had the same distribution as that of the observed data. Quite often, there is a reason why the data is missing that could be a result of that population's distribution being different. Let's reconsider two of the three missing data examples discussed above and whether or not the EM algorithm would be appropriate:

  - The factory sensor example (Ex 1) is a **valid** application of the EM algorithm because the sensor malfunction is a random event, so there's no reason for the data during this time period to be different from data collected while the sensor was functioning correctly
  - The census data example (Ex 2) is **not necessarily a valid** application of the EM algorithm. This is because the rural communities that we are missing data for likely have a very different population distribution. Thus, it's unfair to impute guesses for them based on data from less rural communities. A variety of demographic factors including average income, race, health, etc. are very different between these two groups. With that being said, there are corrections and extensions of the algorithm that can be considered in settings such as this example where we can account for the increased uncertainty.

7 Examples of the EM Algorithm

Here we discuss a few examples of using EM algorithm with missing data.

7.1 Missing Data Problem

Consider a five-category multinomial population, where the probability is $(\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$, and the complete data is $\mathbf{x} = (x_1, x_2, \cdots, x_5)$. Due to the missing data issue, we only observe $\mathbf{y} = (y_1, y_2, \cdots, y_4)$, where $y_1 = x_1 + x_2, y_2 = x_3, y_3 = x_4, y_4 = x_5$, and we hope to estimate π. The complete log-likelihood is

$$f(x|\pi) \propto x_1 \log(\frac{1}{2}) + (x_2 + x_5)\log(\frac{1}{4}\pi) + (x_3 + x_4)\log(\frac{1}{4}(1-\pi))$$

The EM steps produce an iterative estimate $\pi^{(p)}$:

1. Given the current estimate $\pi^{(p)}$, the estimates of x_1, x_2 is $x_1^{(p)} = \frac{1/2}{1/2+\pi^{(p)}/4} y_1$ and $x_2^{(p)} = \frac{\pi^{(p)}/4}{1/2+\pi^{(p)}/4} y_1$
2. Given the current estimate $\hat{x}_1^{(p)}, \hat{x}_2^{(p)}$, the maximum likelihood estimate is $\pi^{(p+1)} = \frac{x_2^{(p)}+x_5}{x_2^{(p)}+x_3+x_4+x_5}$.

The convergence of the EM iterations is shown in Table 1. After 4 iterations, the estimation error falls below 1e-4. In particular, the error shrinks with a factor of 0.13 in each iteration, indicating an exponential convergence behavior.

An alternative approach to estimate π in this model is to merge the first two categories and treat it as a four-category multinomial population with probability $(\frac{1}{2} + \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$. Under this model, we have complete data and therefore π can be simply estimated via maximizing the log-likelihood:

$$f(y|\pi) \propto y_1 \log(\frac{1}{2} + \frac{1}{4}\pi) + y_4 \log(\frac{1}{4}\pi) + (y_2 + y_3)\log(\frac{1}{4}(1-\pi)).$$

Taking the derivative of the log-likelihood gives us:

$$\frac{d}{d\pi} f(y|\pi) = \frac{y_1}{2+\pi} + \frac{y_4}{\pi} - \frac{y_2 + y_3}{1-\pi}.$$

Table 1: The convergence of EM algorithm in the multinomial model

p	$\pi^{(p)}$	$\pi^{(p)} - \pi^*$	$(\pi^{(p+1)} - \pi^*) \div (\pi^{(p)} - \pi^*)$
0	0·500000000	0·126821498	0·1465
1	0·608247423	0·018574075	0·1346
2	0·624321051	0·002500447	0·1330
3	0·626488879	0·000332619	0·1328
4	0·626777323	0·000044176	0·1328
5	0·626815632	0·000005866	0·1328
6	0·626820719	0·000000779	—
7	0·626821395	0·000000104	—
8	0·626821484	0·000000014	—

Solving $\frac{d}{d\pi} f(y|\pi) = 0$ is equivalent to solving a polynomial equation of degree three, which is often complicated. Instead, EM algorithm can provide a fast, intuitive, and easy-to-compute approach to approximate the MLE with small errors.

7.2 Finite Mixtures Model

Suppose we have observations $\mathbf{y} = (y_1, \cdots, y_n)$, and there exists a finite set of R states, such that each y_i is associated with a unique state (unobserved), denoted as $\mathbf{z} = (z_1, \cdots, z_n)$. Specifically, $z_i = k$ indicates y_i belongs to the k-th state. Assume the densities are $z_i \overset{\text{iid}}{\sim} v(\cdot|\phi)$, $y_i \overset{\text{iid}}{\sim} u(\cdot|z_i, \phi)$. The complete data likelihood is

$$\log f(\mathbf{y}, \mathbf{z}|\phi) = \sum_{i=1}^{n} (\log v(z_i|\phi) + \log u(y_i|z_i, \phi))$$

The EM steps are:

- **E step:** Estimate the hidden states z_i for given the current parameters $\phi^{(p)}$, i.e., compute $\mathbb{P}(z_i = k|\mathbf{y}, \phi^{(p)})$
- **M step:** Complete-data maximization with estimated states z_i, i.e., solve $\phi^{(p+1)} = \text{argmax}_\phi \, \mathbb{E}_{\mathbf{z}|\mathbf{y},\phi^{(p)}} \log f(\mathbf{y}, \mathbf{z}|\phi)$

A classical example is the **Gaussian Mixture Model (GMM)**. Where we assume the conditional distributions to be Gaussian.

$$z_i \overset{\text{iid}}{\sim} \text{Categorical}(\pi_1, \cdots, \pi_R), y_i|z_i = k \overset{\text{iid}}{\sim} \text{Normal}(\mu_k, \sigma_k^2)$$

For E-step, the posterior distribution for states $\mathbf{z}$ is

$$\omega_{k,i} = \mathbb{P}(z_i = k|\mathbf{y}, \phi) = \frac{\mathbb{P}(z_i = k|\phi)\mathbb{P}(y_i|z_i = k, \phi)}{\sum_{j=1}^{R} \mathbb{P}(z_i = j|\phi)\mathbb{P}(y_i|z_i = j, \phi)} = \frac{\pi_k \mathcal{N}(y_i|\mu_k, \sigma_k^2)}{\sum_{j=1}^{R} \pi_j \mathcal{N}(y_i|\mu_j, \sigma_j^2)}$$

For M-step, the expected likelihood is expressed as

$$\mathbb{E}_{\mathbf{z}|\mathbf{y},\phi}[\log f(\mathbf{y}, \mathbf{z}|\phi)] = \sum_{i=1}^{n} \sum_{k=1}^{R} \mathbb{P}(z_i = k|\mathbf{y}, \phi) \log \mathbb{P}(y_i, z_i = k|\phi)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{R} \omega_{k,i} \left[\log \pi_k - \frac{1}{2} \log(\sigma_k^2) - \frac{1}{2\sigma_k^2} (y_i - \mu_k)^2 \right] + C$$

The optimizer of M-step is then given by:

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \omega_{k,i}, \quad \hat{\mu}_k = \frac{\sum_{i=1}^{n} \omega_{k,i} y_i}{\sum_{i=1}^{n} \omega_{k,i}}, \quad \hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n} \omega_{k,i} (y_i - \hat{\mu}_k)^2}{\sum_{i=1}^{n} \omega_{k,i}}$$

The intuition behind the EM iterations:

- the E steps computes responsibilities $\omega_{k,i}$ of each point i to each state k.
- the M steps update the state probability, mean, and variance weighted by the computed responsibilities.
- This is often viewed as the soft assignment version of K-means algorithm, where we assign each point to the closest state, and update the state mean by the sample mean in each state.

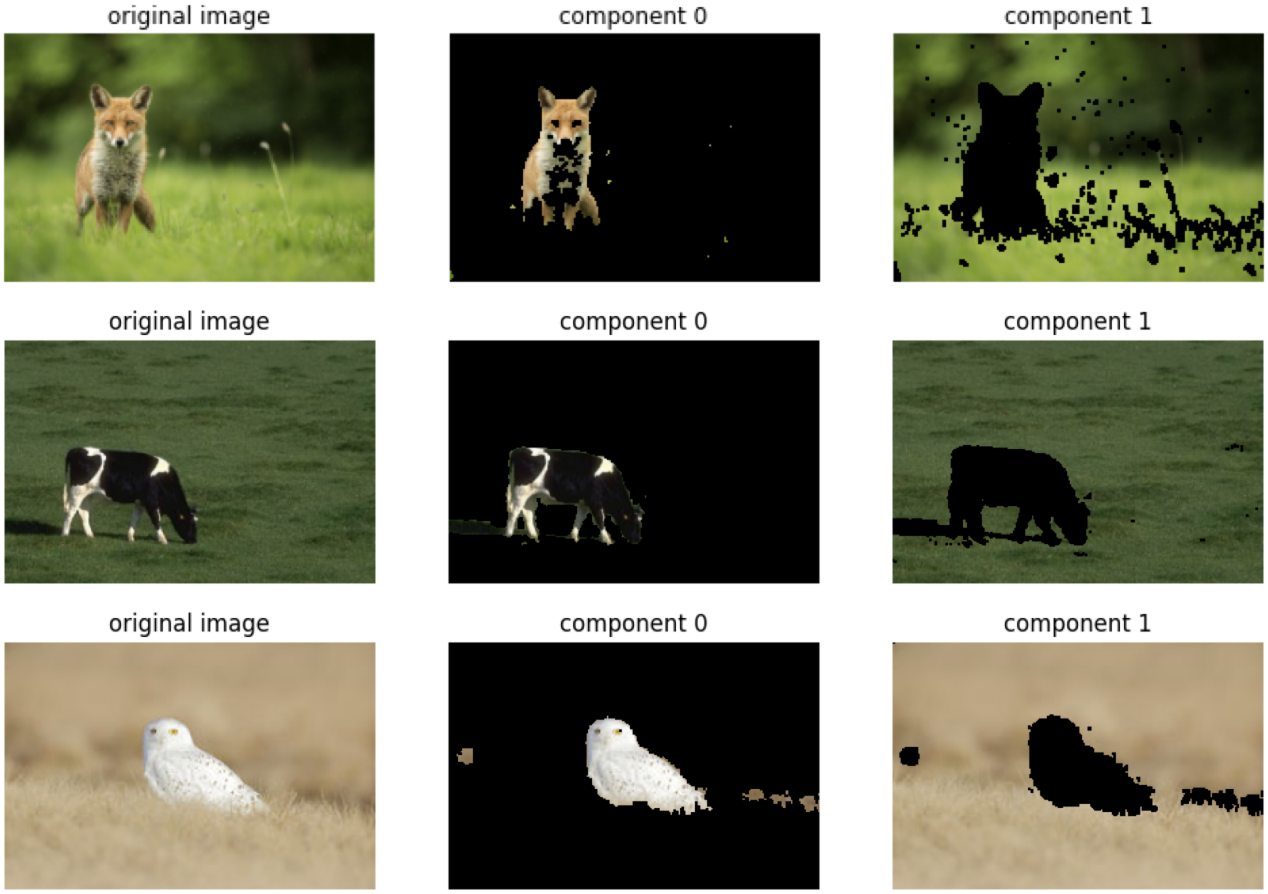

Figure 1. Examples of image segmentation using EM algorithm on Gaussian Mixture Models. Taken from Professor Scott Linderman's Stats 305C material.

Although the mechanism of GMM is quite simple, it has been widely used in many real-world applications:

1. Speaker Identification: GMM can be used for speaker identification systems. Each speaker is modeled using a GMM trained on their voice data. During identification, the likelihood of the input speech belonging to each speaker's GMM is calculated, and the speaker with the highest likelihood is identified.

2. Image Segmentation: GMM can be applied to image segmentation tasks. Each pixel in the image is represented by a feature vector (e.g., color, texture). The GMM is trained on these feature vectors, and each Gaussian component represents a different segment or object in the image. The pixels are then assigned to the most likely Gaussian component, resulting in a segmented image.

3. Anomaly Detection: GMM can be used for anomaly detection in various domains, such as network intrusion detection or fraud detection. The GMM is trained on normal data, and during inference, the likelihood of a new data point belonging to the learned distribution is calculated. If the likelihood falls below a certain threshold, the data point is considered an anomaly.

8 Development of EM Algorithm.

While the EM algorithm provides convenient solutions for many simple problems, it has certain limitations and challenges. In this section, we will explore some variants of the EM algorithm that address these limitations and discuss the historical development of the algorithm. Despite its effectiveness, the EM algorithm has some drawbacks that have led to the development of various variants [Gupta et al., 2011]. Let's discuss a few of these variants and their motivations.

8.1 Stationary Points and Global Optimization

One limitation of the EM algorithm is that it only finds stationary points of the likelihood function, which may not necessarily be the global maximum. To overcome this, one approach is to use EM in conjunction with a global optimizer to explore the parameter space more efficiently [Ali et al., 2005]. By combining EM with

techniques such as simulated annealing or genetic algorithms, the chances of finding the global maximum can be improved.

8.2 Computational Tractability

Another challenge with the EM algorithm is that the required computations may not be tractable, especially for complex models or large datasets. To address this, several variants have been proposed:

- Generalized EM: In this variant, the M-step is modified to only ensure that the likelihood is increasing, rather than maximizing it completely. This can be achieved using techniques such as gradient ascent or Newton's method.
- MCMC-EM: Markov Chain Monte Carlo (MCMC) methods can be used to approximate the E-step when the expectation is intractable. By sampling from the posterior distribution of the latent variables, the expectation can be estimated more efficiently.

8.3 Convergence Speed

The convergence speed of the EM algorithm can be slow in some cases, which has led to the development of acceleration techniques such as Aitken's acceleration [Meilijson, 1989]: it accelerates the convergence of EM by using a Taylor expansion to find the optimal step size. It can be seen as an analogy to Newton's method for EM.

8.4 Maximum Likelihood Estimation Alternatives

In some situations, the maximum likelihood estimate may not be the desired output. For example, one may prefer a posterior distribution to compute the mean. Along this idea, a few variants of EM are:

- Stochastic EM [Celeux, 1985]: In this variant, a random sample is drawn in the E-step to produce a posterior distribution for the parameters. This introduces stochasticity into the algorithm and can help escape local optima.
- Data Augmentation [Tanner and Wong, 1987]: This variant randomizes both the E-step and the M-step. In the M-step, parameters are drawn from a posterior distribution that incorporates prior knowledge. The original M-step corresponds to finding the posterior mode if a non-informative prior is assumed.

8.5 Historical Notes

The EM algorithm has a rich history, with its ideas being used implicitly in various contexts before its formal introduction. Here are some notable historical developments:

- In 1886, Newcomb [Newcomb, 1886] considered the estimation of parameters of a mixture of two univariate normals, which can be seen as the earliest example of an EM-type algorithm.
- In 1958, Hartley [Hartley, 1958] presented the main ideas of EM, rooted in the special case of counting data. He recognized the potential of the algorithm and expressed satisfaction in seeing its widespread application.
- In the 1960s and 1970s, Baum and Welch developed the Baum-Welch algorithm [Baum et al., 1970, Welch, 2003] for fitting hidden Markov models (HMMs). This algorithm is essentially an application of the EM algorithm to HMMs.
- The EM algorithm was formally introduced and named by Dempster, Laird, and Rubin in their seminal paper [Dempster et al., 1977] in 1977. They generalized the algorithm to solve arbitrary maximum likelihood problems with missing or latent data.

The quote by Hartley, *"I felt like the old minstrel who has been singing his song for 18 years and now finds, with considerable satisfaction, that his folklore is the theme of an overpowering symphony"*, beautifully captures the journey of the EM algorithm from its early roots to its widespread recognition and application.

References

[Ali et al., 2005] Ali, M. M., Khompatraporn, C., and Zabinsky, Z. B. (2005). A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. *Journal of global optimization*, 31:635–672.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171.

[Celeux, 1985] Celeux, G. (1985). The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly*, 2:73–82.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the sems algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

[Gupta et al., 2011] Gupta, M. R., Chen, Y., et al. (2011). Theory and use of the em algorithm. *Foundations and Trends® in Signal Processing*, 4(3):223–296.

[Hartley, 1958] Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, 14(2):174–194.

[Meilijson, 1989] Meilijson, I. (1989). A fast improvement to the em algorithm on its own terms. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 51(1):127–138.

[Newcomb, 1886] Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American journal of Mathematics*, pages 343–366.

[Tanner and Wong, 1987] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

[Welch, 2003] Welch, L. R. (2003). Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4):10–13.