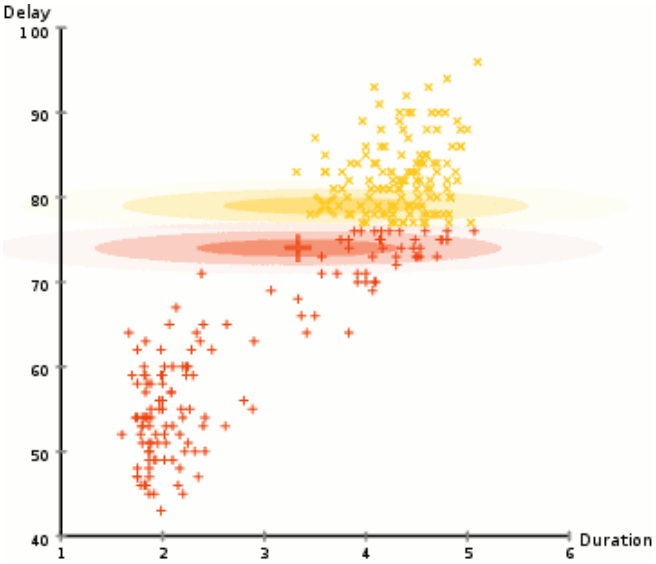
WIKIPEDIA

Expectation–maximization algorithm

In statistics, an **expectation–maximization** (**EM**) **algorithm** is an iterative method to find (local) maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Contents



EM clustering of Old Faithful eruption data. The random initial model (which, due to the different scales of the axes, appears to be two very flat and wide spheres) is fit to the observed data. In the first iterations, the model changes substantially, but then converges to the two modes of the geyser. Visualized using ELKI.

External links

History

The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin.[1] They pointed out that the method had been "proposed many times in special circumstances" by earlier authors. One of the earliest is the gene-counting method for estimating allele frequencies by Cedric Smith.[2] A very detailed treatment of the EM method for exponential families was published by Rolf Sundberg in his thesis and several papers[3][4][5] following his collaboration with Per Martin-Löf and Anders Martin-Löf.[6][7][8][9][10][11][12] The Dempster–Laird–Rubin paper in 1977 generalized the method and sketched a convergence analysis for a wider class of problems. Regardless of earlier inventions, the innovative Dempster–Laird–Rubin paper in the *Journal of the Royal Statistical Society* received an enthusiastic discussion at the Royal Statistical Society meeting with Sundberg calling the paper "brilliant". The Dempster–Laird–Rubin paper established the EM method as an important tool of statistical analysis.

The convergence analysis of the Dempster–Laird–Rubin algorithm was flawed and a correct convergence analysis was published by C. F. Jeff Wu in 1983.[13] Wu's proof established the EM method's convergence outside of the exponential family, as claimed by Dempster–Laird–Rubin.[13]

Introduction

The EM algorithm is used to find (local) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either missing values exist among the data, or the model can be formulated more simply by assuming the existence of further unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component to which each data point belongs.

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values, the parameters and the latent variables, and simultaneously solving the resulting equations. In statistical models with latent variables, this is usually impossible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

The EM algorithm proceeds from the observation that there is a way to solve these two sets of equations numerically. One can simply pick arbitrary values for one of the two sets of unknowns, use them to estimate the second set, then use these new values to find a better estimate of the first set, and then keep alternating between the two until the resulting values both converge to fixed points. It's not obvious that this will work, but it can be proven that in this context it does, and that the derivative of the likelihood is (arbitrarily close to) zero at that point, which in turn means that the point is either a maximum or a saddle point.[13] In general, multiple maxima may occur, with no guarantee that the global maximum will be found. Some likelihoods also have singularities in them, i.e., nonsensical maxima. For example, one of the *solutions* that may be found by EM in a mixture model involves setting one of the components to have zero variance and the mean parameter for the same component to be equal to one of the data points.

Description

Given the statistical model which generates a set $\mathbf{X}$ of observed data, a set of unobserved latent data or missing

values $\mathbf{Z}$, and a vector of unknown parameters $\boldsymbol{\theta}$, along with a likelihood function $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$, the maximum likelihood estimate (MLE) of the unknown parameters is determined by maximizing the marginal likelihood of the observed data

$$L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\, d\mathbf{Z}$$

However, this quantity is often intractable (e.g. if $\mathbf{Z}$ is a sequence of events, so that the number of values grows exponentially with the sequence length, the exact calculation of the sum will be extremely difficult).

The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying these two steps:

Expectation step (E step): Define $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ as the expected value of the log likelihood function of $\boldsymbol{\theta}$, with respect to the current conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ and the current estimates of the parameters $\boldsymbol{\theta}^{(t)}$:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \mathbf{E}_{\mathbf{Z}\mid\mathbf{X},\boldsymbol{\theta}^{(t)}}[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$$

Maximization step (M step): Find the parameters that maximize this quantity:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$$

The typical models to which EM is applied use $\mathbf{Z}$ as a latent variable indicating membership in one of a set of groups:

1. The observed data points $\mathbf{X}$ may be discrete (taking values in a finite or countably infinite set) or continuous (taking values in an uncountably infinite set). Associated with each data point may be a vector of observations.
2. The missing values (aka latent variables) $\mathbf{Z}$ are discrete, drawn from a fixed number of values, and with one latent variable per observed unit.
3. The parameters are continuous, and are of two kinds: Parameters that are associated with all data points, and those associated with a specific value of a latent variable (i.e., associated with all data points which corresponding latent variable has that value).

However, it is possible to apply EM to other sorts of models.

The motive is as follows. If the value of the parameters $\boldsymbol{\theta}$ is known, usually the value of the latent variables $\mathbf{Z}$ can be found by maximizing the log-likelihood over all possible values of $\mathbf{Z}$, either simply by iterating over $\mathbf{Z}$ or through an algorithm such as the Baum–Welch algorithm for hidden Markov models. Conversely, if we know the value of the latent variables $\mathbf{Z}$, we can find an estimate of the parameters $\boldsymbol{\theta}$ fairly easily, typically by simply grouping the observed data points according to the value of the associated latent variable and averaging the values, or some function of the values, of the points in each group. This suggests an iterative algorithm, in the case where both $\boldsymbol{\theta}$ and $\mathbf{Z}$ are unknown:

1. First, initialize the parameters $\boldsymbol{\theta}$ to some random values.
2. Compute the probability of each possible value of $\mathbf{Z}$, given $\boldsymbol{\theta}$.
3. Then, use the just-computed values of $\mathbf{Z}$ to compute a better estimate for the parameters $\boldsymbol{\theta}$.
4. Iterate steps 2 and 3 until convergence.

The algorithm as just described monotonically approaches a local minimum of the cost function.

Properties

Speaking of an expectation (E) step is a bit of a misnomer. What are calculated in the first step are the fixed, data-dependent parameters of the function Q. Once the parameters of Q are known, it is fully determined and is maximized in the second (M) step of an EM algorithm.

Although an EM iteration does increase the observed data (i.e., marginal) likelihood function, no guarantee exists that the sequence converges to a maximum likelihood estimator. For multimodal distributions, this means that an EM algorithm may converge to a local maximum of the observed data likelihood function, depending on starting values. A variety of heuristic or metaheuristic approaches exist to escape a local maximum, such as random-restart hill climbing (starting with several different random initial estimates $\theta^{(t)}$), or applying simulated annealing methods.

EM is especially useful when the likelihood is an exponential family: the E step becomes the sum of expectations of sufficient statistics, and the M step involves maximizing a linear function. In such a case, it is usually possible to derive closed-form expression updates for each step, using the Sundberg formula (published by Rolf Sundberg using unpublished results of Per Martin-Löf and Anders Martin-Löf).[4][5][8][9][10][11][12]

The EM method was modified to compute maximum a posteriori (MAP) estimates for Bayesian inference in the original paper by Dempster, Laird, and Rubin.

Other methods exist to find maximum likelihood estimates, such as gradient descent, conjugate gradient, or variants of the Gauss–Newton algorithm. Unlike EM, such methods typically require the evaluation of first and/or second derivatives of the likelihood function.

Proof of correctness

Expectation-maximization works to improve $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ rather than directly improving $\log p(\mathbf{X} \mid \boldsymbol{\theta})$. Here is shown that improvements to the former imply improvements to the latter.[14]

For any $\mathbf{Z}$ with non-zero probability $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$, we can write

$$\log p(\mathbf{X} \mid \boldsymbol{\theta}) = \log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \log p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}).$$

We take the expectation over possible values of the unknown data $\mathbf{Z}$ under the current parameter estimate $\boldsymbol{\theta}^{(t)}$ by multiplying both sides by $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)})$ and summing (or integrating) over $\mathbf{Z}$. The left-hand side is the expectation of a constant, so we get:

$$\begin{aligned} \log p(\mathbf{X} \mid \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}), \end{aligned}$$

where $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ is defined by the negated sum it is replacing. This last equation holds for every value of $\boldsymbol{\theta}$ including $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$,

$$\log p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}),$$

and subtracting this last equation from the previous equation gives

$$\log p(\mathbf{X} \mid \boldsymbol{\theta}) - \log p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}),$$

However, Jensen's inequality tells us that $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \geq H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})$, so we can conclude that

$$\log p(\mathbf{X} \mid \boldsymbol{\theta}) - \log p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}).$$

In words, choosing $\boldsymbol{\theta}$ to improve $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ causes $\log p(\mathbf{X} \mid \boldsymbol{\theta})$ to improve at least as much.

As a maximization–maximization procedure

The EM algorithm can be viewed as two alternating maximization steps, that is, as an example of coordinate descent.[15][16] Consider the function:

$$F(q, \theta) := \mathrm{E}_q[\log L(\theta; x, Z)] + H(q),$$

where q is an arbitrary probability distribution over the unobserved data z and $H(q)$ is the entropy of the distribution q. This function can be written as

$$F(q, \theta) = -D_{\mathrm{KL}}\left(q \parallel p_{Z|X}(\cdot \mid x; \theta)\right) + \log L(\theta; x),$$

where $p_{Z|X}(\cdot \mid x; \theta)$ is the conditional distribution of the unobserved data given the observed data x and D_{KL} is the Kullback–Leibler divergence.

Then the steps in the EM algorithm may be viewed as:

Expectation step: Choose q to maximize F:
$q^{(t)} = \arg\max_q \ F(q, \theta^{(t)})$
Maximization step: Choose θ to maximize F:
$\theta^{(t+1)} = \arg\max_\theta \ F(q^{(t)}, \theta)$

Applications

EM is frequently used for data clustering in machine learning and computer vision. In natural language processing, two prominent instances of the algorithm are the Baum–Welch algorithm for hidden Markov models, and the inside-outside algorithm for unsupervised induction of probabilistic context-free grammars.

EM is frequently used for parameter estimation of mixed models,[17][18] notably in quantitative genetics.[19]

In psychometrics, EM is almost indispensable for estimating item parameters and latent abilities of item response theory models.

With the ability to deal with missing data and observe unidentified variables, EM is becoming a useful tool to price and manage risk of a portfolio.

The EM algorithm (and its faster variant ordered subset expectation maximization) is also widely used in medical image reconstruction, especially in positron emission tomography, single photon emission computed tomography, and x-ray computed tomography. See below for other faster variants of EM.

In structural engineering, the Structural Identification using Expectation Maximization (STRIDE)[20] algorithm is an

output-only method for identifying natural vibration properties of a structural system using sensor data (see Operational Modal Analysis).

Filtering and smoothing EM algorithms

A Kalman filter is typically used for on-line state estimation and a minimum-variance smoother may be employed for off-line or batch state estimation. However, these minimum-variance solutions require estimates of the state-space model parameters. EM algorithms can be used for solving joint state and parameter estimation problems.

Filtering and smoothing EM algorithms arise by repeating this two-step procedure:

E-step
Operate a Kalman filter or a minimum-variance smoother designed with current parameter estimates to obtain updated state estimates.

M-step
Use the filtered or smoothed state estimates within maximum-likelihood calculations to obtain updated parameter estimates.

Suppose that a Kalman filter or minimum-variance smoother operates on measurements of a single-input-single-output system that possess additive white noise. An updated measurement noise variance estimate can be obtained from the maximum likelihood calculation

$$\widehat{\sigma}_v^2 = \frac{1}{N}\sum_{k=1}^{N}(z_k - \widehat{x}_k)^2,$$

where $\widehat{x}_k$ are scalar output estimates calculated by a filter or a smoother from N scalar measurements z_k. The above update can also be applied to updating a Poisson measurement noise intensity. Similarly, for a first-order auto-regressive process, an updated process noise variance estimate can be calculated by

$$\widehat{\sigma}_w^2 = \frac{1}{N}\sum_{k=1}^{N}(\widehat{x}_{k+1} - \widehat{F}\widehat{x}_k)^2,$$

where $\widehat{x}_k$ and $\widehat{x}_{k+1}$ are scalar state estimates calculated by a filter or a smoother. The updated model coefficient estimate is obtained via

$$\widehat{F} = \frac{\sum_{k=1}^{N}(\widehat{x}_{k+1} - \widehat{F}\widehat{x}_k)}{\sum_{k=1}^{N}\widehat{x}_k^2}.$$

The convergence of parameter estimates such as those above are well studied.[21][22][23][24]

Variants

A number of methods have been proposed to accelerate the sometimes slow convergence of the EM algorithm, such as those using conjugate gradient and modified Newton's methods (Newton–Raphson).[25] Also, EM can be used with constrained estimation methods.

Parameter-expanded expectation maximization (PX-EM) algorithm often provides speed up by "us[ing] a

`covariance adjustment' to correct the analysis of the M step, capitalising on extra information captured in the imputed complete data".[26]

Expectation conditional maximization (ECM) replaces each M step with a sequence of conditional maximization (CM) steps in which each parameter θ_i is maximized individually, conditionally on the other parameters remaining fixed.[27] Itself can be extended into the *Expectation conditional maximization either (ECME)* algorithm.[28]

This idea is further extended in *generalized expectation maximization (GEM)* algorithm, in which is sought only an increase in the objective function F for both the E step and M step as described in the As a maximization-maximization procedure section.[15] GEM is further developed in a distributed environment and shows promising results.[29]

It is also possible to consider the EM algorithm as a subclass of the **MM** (Majorize/Minimize or Minorize/Maximize, depending on context) algorithm,[30] and therefore use any machinery developed in the more general case.

α-EM algorithm

The Q-function used in the EM algorithm is based on the log likelihood. Therefore, it is regarded as the log-EM algorithm. The use of the log likelihood can be generalized to that of the α-log likelihood ratio. Then, the α-log likelihood ratio of the observed data can be exactly expressed as equality by using the Q-function of the α-log likelihood ratio and the α-divergence. Obtaining this Q-function is a generalized E step. Its maximization is a generalized M step. This pair is called the α-EM algorithm [31] which contains the log-EM algorithm as its subclass. Thus, the α-EM algorithm by Yasuo Matsuyama is an exact generalization of the log-EM algorithm. No computation of gradient or Hessian matrix is needed. The α-EM shows faster convergence than the log-EM algorithm by choosing an appropriate α. The α-EM algorithm leads to a faster version of the Hidden Markov model estimation algorithm α-HMM. [32]

Relation to variational Bayes methods

EM is a partially non-Bayesian, maximum likelihood method. Its final result gives a probability distribution over the latent variables (in the Bayesian style) together with a point estimate for θ (either a maximum likelihood estimate or a posterior mode). A fully Bayesian version of this may be wanted, giving a probability distribution over θ and the latent variables. The Bayesian approach to inference is simply to treat θ as another latent variable. In this paradigm, the distinction between the E and M steps disappears. If using the factorized Q approximation as described above (variational Bayes), solving can iterate over each latent variable (now including θ) and optimize them one at a time. Now, k steps per iteration are needed, where k is the number of latent variables. For graphical models this is easy to do as each variable's new Q depends only on its Markov blanket, so local message passing can be used for efficient inference.

Geometric interpretation

In information geometry, the E step and the M step are interpreted as projections under dual affine connections, called the e-connection and the m-connection; the Kullback–Leibler divergence can also be understood in these terms.

Examples

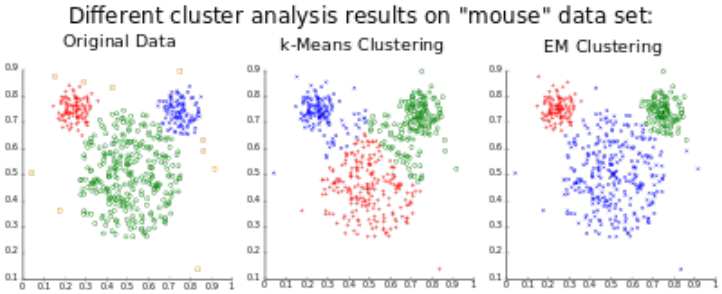
Gaussian mixture

Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ be a sample of n independent observations from a mixture of two multivariate normal distributions of dimension d, and let $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ be the latent variables that determine the component from which the observation originates.[16]

$$X_i \mid (Z_i = 1) \sim \mathcal{N}_d(\boldsymbol{\mu}_1, \Sigma_1) \text{ and}$$
$$X_i \mid (Z_i = 2) \sim \mathcal{N}_d(\boldsymbol{\mu}_2, \Sigma_2),$$

where

$$\mathbf{P}(Z_i = 1) = \tau_1 \text{ and}$$
$$\mathbf{P}(Z_i = 2) = \tau_2 = 1 - \tau_1.$$



Comparison of k-means and EM on artificial data visualized with ELKI. Using the variances, the EM algorithm can describe the normal distributions exactly, while k-means splits the data in Voronoi-cells. The cluster center is indicated by the lighter, bigger symbol.

The aim is to estimate the unknown parameters representing the *mixing* value between the Gaussians and the means and covariances of each:

$$\theta = (\boldsymbol{\tau}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2),$$

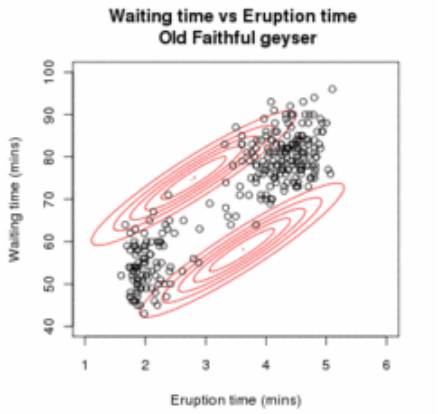
where the incomplete-data likelihood function is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j \, f(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j),$$

and the complete-data likelihood function is

$$L(\theta; \mathbf{x}, \mathbf{z}) = p(\mathbf{x}, \mathbf{z} \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{2} [f(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j)\tau_j]^{\mathbb{I}(z_i = j)},$$

or



An animation demonstrating the EM algorithm fitting a two component Gaussian mixture model to the Old Faithful dataset. The algorithm steps through from a random initialization to convergence.

$$L(\theta; \mathbf{x}, \mathbf{z}) = \exp\left\{ \sum_{i=1}^{n} \sum_{j=1}^{2} \mathbb{I}(z_i = j) \left[\log \tau_j - \tfrac{1}{2} \log |\Sigma_j| - \tfrac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j) - \tfrac{d}{2} \log(2\pi) \right] \right\},$$

where $\mathbb{I}$ is an indicator function and f is the probability density function of a multivariate normal.

In the last equality, for each i, one indicator $\mathbb{I}(z_i = j)$ is equal to zero, and one indicator is equal to one. The inner sum thus reduces to one term.

E step

Given our current estimate of the parameters $\theta^{(t)}$, the conditional distribution of the Z_i is determined by Bayes

theorem to be the proportional height of the normal density weighted by τ:

$$T_{j,i}^{(t)} := \mathrm{P}(Z_i = j \mid X_i = \mathbf{x}_i; \theta^{(t)}) = \frac{\tau_j^{(t)}\ f(\mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}{\tau_1^{(t)}\ f(\mathbf{x}_i; \boldsymbol{\mu}_1^{(t)}, \Sigma_1^{(t)}) + \tau_2^{(t)}\ f(\mathbf{x}_i; \boldsymbol{\mu}_2^{(t)}, \Sigma_2^{(t)})}.$$

These are called the "membership probabilities", which are normally considered the output of the E step (although this is not the Q function of below).

This E step corresponds with setting up this function for Q:

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta^{(t)}}\left[\log L(\theta; \mathbf{x}, \mathbf{Z})\right] \\ &= \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta^{(t)}}\left[\log \prod_{i=1}^{n} L(\theta; \mathbf{x}_i, Z_i)\right] \\ &= \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta^{(t)}}\left[\sum_{i=1}^{n} \log L(\theta; \mathbf{x}_i, Z_i)\right] \\ &= \sum_{i=1}^{n} \mathrm{E}_{Z_i|\mathbf{X};\theta^{(t)}}\left[\log L(\theta; \mathbf{x}_i, Z_i)\right] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{2} P(Z_i = j \mid X_i = \mathbf{x}_i; \theta^{(t)}) \log L(\theta_j; \mathbf{x}_i, j) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{2} T_{j,i}^{(t)} \left[\log \tau_j - \tfrac{1}{2}\log|\Sigma_j| - \tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_j) - \tfrac{d}{2}\log(2\pi)\right]. \end{aligned}$$

The expectation of $\log L(\theta; \mathbf{x}_i, Z_i)$ inside the sum is taken with respect to the probability density function $P(Z_i \mid X_i = \mathbf{x}_i; \theta^{(t)})$, which might be different for each $\mathbf{x}_i$ of the training set. Everything in the E step is known before the step is taken except $T_{j,i}$, which is computed according to the equation at the beginning of the E step section.

This full conditional expectation does not need to be calculated in one step, because τ and $\boldsymbol{\mu}/\Sigma$ appear in separate linear terms and can thus be maximized independently.

M step

$Q(\theta \mid \theta^{(t)})$ being quadratic in form means that determining the maximizing values of θ is relatively straightforward. Also, τ, ($\boldsymbol{\mu}_1$,Σ_1) and ($\boldsymbol{\mu}_2$,Σ_2) may all be maximized independently since they all appear in separate linear terms.

To begin, consider τ, which has the constraint $\tau_1 + \tau_2$=1:

$$\begin{aligned} \boldsymbol{\tau}^{(t+1)} &= \arg\max_{\boldsymbol{\tau}} Q(\theta \mid \theta^{(t)}) \\ &= \arg\max_{\boldsymbol{\tau}} \left\{ \left[\sum_{i=1}^{n} T_{1,i}^{(t)}\right] \log \tau_1 + \left[\sum_{i=1}^{n} T_{2,i}^{(t)}\right] \log \tau_2 \right\}. \end{aligned}$$

This has the same form as the MLE for the binomial distribution, so

$$\tau_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{j,i}^{(t)}}{\sum_{i=1}^{n} (T_{1,i}^{(t)} + T_{2,i}^{(t)})} = \frac{1}{n} \sum_{i=1}^{n} T_{j,i}^{(t)}.$$

For the next estimates of ($\boldsymbol{\mu}_1$,Σ_1):

$$\begin{aligned}(\boldsymbol{\mu}_1^{(t+1)}, \Sigma_1^{(t+1)}) &= \underset{\boldsymbol{\mu}_1, \Sigma_1}{\arg\max} Q(\theta \mid \theta^{(t)}) \\ &= \underset{\boldsymbol{\mu}_1, \Sigma_1}{\arg\max} \sum_{i=1}^{n} T_{1,i}^{(t)} \left\{ -\tfrac{1}{2} \log |\Sigma_1| - \tfrac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1) \right\}\end{aligned}.$$

This has the same form as a weighted MLE for a normal distribution, so

$$\boldsymbol{\mu}_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} T_{1,i}^{(t)}} \text{ and } \Sigma_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^\top}{\sum_{i=1}^{n} T_{1,i}^{(t)}}$$

and, by symmetry,

$$\boldsymbol{\mu}_2^{(t+1)} = \frac{\sum_{i=1}^{n} T_{2,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} T_{2,i}^{(t)}} \text{ and } \Sigma_2^{(t+1)} = \frac{\sum_{i=1}^{n} T_{2,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})^\top}{\sum_{i=1}^{n} T_{2,i}^{(t)}}.$$

Termination

Conclude the iterative process if $E_{Z|\theta^{(t)},\mathbf{x}}[\log L(\theta^{(t)}; \mathbf{x}, \mathbf{Z})] \leq E_{Z|\theta^{(t-1)},\mathbf{x}}[\log L(\theta^{(t-1)}; \mathbf{x}, \mathbf{Z})] + \varepsilon$ for ε below some preset threshold.

Generalization

The algorithm illustrated above can be generalized for mixtures of more than two multivariate normal distributions.

Truncated and censored regression

The EM algorithm has been implemented in the case where an underlying linear regression model exists explaining the variation of some quantity, but where the values actually observed are censored or truncated versions of those represented in the model.[33] Special cases of this model include censored or truncated observations from one normal distribution.[33]

Alternatives

EM typically converges to a local optimum, not necessarily the global optimum, with no bound on the convergence rate in general. It is possible that it can be arbitrarily poor in high dimensions and there can be an exponential number of local optima. Hence, a need exists for alternative methods for guaranteed learning, especially in the high-dimensional setting. Alternatives to EM exist with better guarantees for consistency, which are termed *moment-based approaches*[34] or the so-called *spectral techniques*[35][36]. Moment-based approaches to learning the parameters of a probabilistic model are of increasing interest recently since they enjoy guarantees such as global

convergence under certain conditions unlike EM which is often plagued by the issue of getting stuck in local optima. Algorithms with guarantees for learning can be derived for a number of important models such as mixture models, HMMs etc. For these spectral methods, no spurious local optima occur, and the true parameters can be consistently estimated under some regularity conditions.

See also

- mixture distribution
- compound distribution
- density estimation
- total absorption spectroscopy
- The EM algorithm can be viewed as a special case of the majorize-minimization (MM) algorithm.[37]

References

1. Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society, Series B*. **39** (1): 1–38. JSTOR 2984875 (https://www.jstor.org/stable/2984875). MR 0501537 (https://www.ams.org/mathscinet-getitem?mr=0501537).
2. Ceppelini, R.M. (1955). "The estimation of gene frequencies in a random-mating population". *Ann. Hum. Genet*. **20** (2): 97–115. doi:10.1111/j.1469-1809.1955.tb01360.x (https://doi.org/10.1111%2Fj.1469-1809.1955.tb01360.x). PMID 13268982 (https://pubmed.ncbi.nlm.nih.gov/13268982).
3. Sundberg, Rolf (1974). "Maximum likelihood theory for incomplete data from an exponential family". *Scandinavian Journal of Statistics*. **1** (2): 49–58. JSTOR 4615553 (https://www.jstor.org/stable/4615553). MR 0381110 (https://www.ams.org/mathscinet-getitem?mr=0381110).
4. Rolf Sundberg. 1971. *Maximum likelihood theory and applications for distributions generated when observing a function of an exponential family variable*. Dissertation, Institute for Mathematical Statistics, Stockholm University.
5. Sundberg, Rolf (1976). "An iterative method for solution of the likelihood equations for incomplete data from exponential families". *Communications in Statistics – Simulation and Computation*. **5** (1): 55–64. doi:10.1080/03610917608812007 (https://doi.org/10.1080%2F03610917608812007). MR 0443190 (https://www.ams.org/mathscinet-getitem?mr=0443190).
6. See the acknowledgement by Dempster, Laird and Rubin on pages 3, 5 and 11.
7. G. Kulldorff. 1961. *Contributions to the theory of estimation from grouped and partially grouped samples*. Almqvist & Wiksell.
8. Anders Martin-Löf. 1963. "Utvärdering av livslängder i subnanosekundsområdet" ("Evaluation of sub-nanosecond lifetimes"). ("Sundberg formula")
9. Per Martin-Löf. 1966. *Statistics from the point of view of statistical mechanics*. Lecture notes, Mathematical Institute, Aarhus University. ("Sundberg formula" credited to Anders Martin-Löf).

10. Per Martin-Löf. 1970. *Statistika Modeller (Statistical Models): Anteckningar från seminarier läsåret 1969–1970 (Notes from seminars in the academic year 1969-1970), with the assistance of Rolf Sundberg.* Stockholm University. ("Sundberg formula")
11. Martin-Löf, P. The notion of redundancy and its use as a quantitative measure of the deviation between a statistical hypothesis and a set of observational data. With a discussion by F. Abildgård, A. P. Dempster, D. Basu, D. R. Cox, A. W. F. Edwards, D. A. Sprott, G. A. Barnard, O. Barndorff-Nielsen, J. D. Kalbfleisch and G. Rasch and a reply by the author. *Proceedings of Conference on Foundational Questions in Statistical Inference* (Aarhus, 1973), pp. 1–42. Memoirs, No. 1, Dept. Theoret. Statist., Inst. Math., Univ. Aarhus, Aarhus, 1974.
12. Martin-Löf, Per (1974). "The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data". *Scand. J. Statist*. **1** (1): 3–18.
13. Wu, C. F. Jeff (Mar 1983). "On the Convergence Properties of the EM Algorithm" (https://doi.org/10.1214/aos/1176346060). *Annals of Statistics*. **11** (1): 95–103. doi:10.1214/aos/1176346060 (https://doi.org/10.1214%2Faos%2F1176346060). JSTOR 2240463 (https://www.jstor.org/stable/2240463). MR 0684867 (https://www.ams.org/mathscinet-getitem?mr=0684867).
14. Little, Roderick J.A.; Rubin, Donald B. (1987). *Statistical Analysis with Missing Data* (https://archive.org/details/statisticalanaly00litt). Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons. pp. 134 (https://archive.org/details/statisticalanaly00litt/page/n145)–136. ISBN 978-0-471-80254-9.
15. Neal, Radford; Hinton, Geoffrey (1999). Michael I. Jordan (ed.). *A view of the EM algorithm that justifies incremental, sparse, and other variants* (ftp://ftp.cs.toronto.edu/pub/radford/emk.pdf) (PDF). *Learning in Graphical Models*. Cambridge, MA: MIT Press. pp. 355–368. ISBN 978-0-262-60032-3. Retrieved 2009-03-22.
16. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2001). "8.5 The EM algorithm". *The Elements of Statistical Learning* (https://archive.org/details/elementsstatisti00thas_842). New York: Springer. pp. 236 (https://archive.org/details/elementsstatisti00thas_842/page/n237)–243. ISBN 978-0-387-95284-0.
17. Lindstrom, Mary J; Bates, Douglas M (1988). "Newton—Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data". *Journal of the American Statistical Association*. **83** (404): 1014. doi:10.1080/01621459.1988.10478693 (https://doi.org/10.1080%2F01621459.1988.10478693).
18. Van Dyk, David A (2000). "Fitting Mixed-Effects Models Using Efficient EM-Type Algorithms". *Journal of Computational and Graphical Statistics*. **9** (1): 78–98. doi:10.2307/1390614 (https://doi.org/10.2307%2F1390614). JSTOR 1390614 (https://www.jstor.org/stable/1390614).
19. Diffey, S. M; Smith, A. B; Welsh, A. H; Cullis, B. R (2017). "A new REML (parameter expanded) EM algorithm for linear mixed models" (https://doi.org/10.1111/anzs.12208). *Australian & New Zealand Journal of Statistics*. **59** (4): 433. doi:10.1111/anzs.12208 (https://doi.org/10.1111%2Fanzs.12208).
20. Matarazzo, T. J., and Pakzad, S. N. (2016). “STRIDE for Structural Identification using Expectation Maximization: Iterative Output-Only Method for Modal Identification.” Journal of Engineering Mechanics.http://ascelibrary.org/doi/abs/10.1061/(ASCE)EM.1943-7889.0000951

21. Einicke, G. A.; Malos, J. T.; Reid, D. C.; Hainsworth, D. W. (January 2009). "Riccati Equation and EM Algorithm Convergence for Inertial Navigation Alignment". *IEEE Trans. Signal Process*. **57** (1): 370–375. Bibcode:2009ITSP...57..370E (https://ui.adsabs.harvard.edu/abs/2009ITSP...57..370E). doi:10.1109/TSP.2008.2007090 (https://doi.org/10.1109%2FTSP.2008.2007090).
22. Einicke, G. A.; Falco, G.; Malos, J. T. (May 2010). "EM Algorithm State Matrix Estimation for Navigation". *IEEE Signal Processing Letters*. **17** (5): 437–440. Bibcode:2010ISPL...17..437E (https://ui.adsabs.harvard.edu/abs/2010ISPL...17..437E). doi:10.1109/LSP.2010.2043151 (https://doi.org/10.1109%2FLSP.2010.2043151).
23. Einicke, G. A.; Falco, G.; Dunn, M. T.; Reid, D. C. (May 2012). "Iterative Smoother-Based Variance Estimation". *IEEE Signal Processing Letters*. **19** (5): 275–278. Bibcode:2012ISPL...19..275E (https://ui.adsabs.harvard.edu/abs/2012ISPL...19..275E). doi:10.1109/LSP.2012.2190278 (https://doi.org/10.1109%2FLSP.2012.2190278).
24. Einicke, G. A. (Sep 2015). "Iterative Filtering and Smoothing of Measurements Possessing Poisson Noise". *IEEE Transactions on Aerospace and Electronic Systems*. **51** (3): 2205–2011. Bibcode:2015ITAES..51.2205E (https://ui.adsabs.harvard.edu/abs/2015ITAES..51.2205E). doi:10.1109/TAES.2015.140843 (https://doi.org/10.1109%2FTAES.2015.140843).
25. Jamshidian, Mortaza; Jennrich, Robert I. (1997). "Acceleration of the EM Algorithm by using Quasi-Newton Methods". *Journal of the Royal Statistical Society, Series B*. **59** (2): 569–587. doi:10.1111/1467-9868.00083 (https://doi.org/10.1111%2F1467-9868.00083). MR 1452026 (https://www.ams.org/mathscinet-getitem?mr=1452026).
26. Liu, C (1998). "Parameter expansion to accelerate EM: The PX-EM algorithm". *Biometrika*. **85** (4): 755–770. CiteSeerX 10.1.1.134.9617 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.134.9617). doi:10.1093/biomet/85.4.755 (https://doi.org/10.1093%2Fbiomet%2F85.4.755).
27. Meng, Xiao-Li; Rubin, Donald B. (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework". *Biometrika*. **80** (2): 267–278. doi:10.1093/biomet/80.2.267 (https://doi.org/10.1093%2Fbiomet%2F80.2.267). MR 1243503 (https://www.ams.org/mathscinet-getitem?mr=1243503). S2CID 40571416 (https://api.semanticscholar.org/CorpusID:40571416).
28. Liu, Chuanhai; Rubin, Donald B (1994). "The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence". *Biometrika*. **81** (4): 633. doi:10.1093/biomet/81.4.633 (https://doi.org/10.1093%2Fbiomet%2F81.4.633). JSTOR 2337067 (https://www.jstor.org/stable/2337067).
29. Jiangtao Yin; Yanfeng Zhang; Lixin Gao (2012). "Accelerating Expectation-Maximization Algorithms with Frequent Updates" (http://rio.ecs.umass.edu/mnilpub/papers/cluster2012-yin.pdf) (PDF). *Proceedings of the IEEE International Conference on Cluster Computing*.
30. Hunter DR and Lange K (2004), A Tutorial on MM Algorithms (http://www.stat.psu.edu/~dhunter/papers/mmtutorial.pdf), The American Statistician, 58: 30-37
31. Matsuyama, Yasuo (2003). "The α-EM algorithm: Surrogate likelihood maximization using α-logarithmic information measures". *IEEE Transactions on Information Theory*. **49** (3): 692–706. doi:10.1109/TIT.2002.808105 (https://doi.org/10.1109%2FTIT.2002.808105).
32. Matsuyama, Yasuo (2011). "Hidden Markov model estimation based on alpha-EM algorithm: Discrete and continuous alpha-HMMs". *International Joint Conference on Neural Networks*: 808–816.

33. Wolynetz, M.S. (1979). "Maximum likelihood estimation in a linear model from confined and censored normal data". *Journal of the Royal Statistical Society, Series C*. **28** (2): 195–206. doi:10.2307/2346749 (https://doi.org/10.2307%2F2346749). JSTOR 2346749 (https://www.jstor.org/stable/2346749).
34. Pearson, Karl (1894). "Contributions to the Mathematical Theory of Evolution" (https://doi.org/10.1098/rsta.1894.0003). *Philosophical Transactions of the Royal Society of London A*. **185**: 71–110. Bibcode:1894RSPTA.185...71P (https://ui.adsabs.harvard.edu/abs/1894RSPTA.185...71P). doi:10.1098/rsta.1894.0003 (https://doi.org/10.1098%2Frsta.1894.0003). ISSN 0264-3820 (https://www.worldcat.org/issn/0264-3820). JSTOR 90667 (https://www.jstor.org/stable/90667).
35. Shaban, Amirreza; Mehrdad, Farajtabar; Bo, Xie; Le, Song; Byron, Boots (2015). "Learning Latent Variable Models by Improving Spectral Solutions with Exterior Point Method" (https://www.cc.gatech.edu/~bboots3/files/SpectralExteriorPoint-NIPSWorkshop.pdf) (PDF). *UAI*: 792–801.
36. Balle, Borja Quattoni, Ariadna Carreras, Xavier (2012-06-27). *Local Loss Optimization in Operator Models: A New Insight into Spectral Learning*. OCLC 815865081 (https://www.worldcat.org/oclc/815865081).
37. Lange, Kenneth. "The MM Algorithm" (http://www.stat.berkeley.edu/~aldous/Colloq/lange-talk.pdf) (PDF).

Further reading

- Hogg, Robert; McKean, Joseph; Craig, Allen (2005). *Introduction to Mathematical Statistics*. Upper Saddle River, NJ: Pearson Prentice Hall. pp. 359–364.
- Dellaert, Frank (2002). "The Expectation Maximization Algorithm". CiteSeerX 10.1.1.9.9735 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.9735). gives an easier explanation of EM algorithm as to lowerbound maximization.
- Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer. ISBN 978-0-387-31073-2.
- Gupta, M. R.; Chen, Y. (2010). "Theory and Use of the EM Algorithm". *Foundations and Trends in Signal Processing*. **4** (3): 223–296. CiteSeerX 10.1.1.219.6830 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.219.6830). doi:10.1561/2000000034 (https://doi.org/10.1561%2F2000000034). A well-written short book on EM, including detailed derivation of EM for GMMs, HMMs, and Dirichlet.
- Bilmes, Jeff (1998). "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models". CiteSeerX 10.1.1.28.613 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.613). includes a simplified derivation of the EM equations for Gaussian Mixtures and Gaussian Mixture Hidden Markov Models.
- McLachlan, Geoffrey J.; Krishnan, Thriyambakam (2008). *The EM Algorithm and Extensions* (2nd ed.). Hoboken: Wiley. ISBN 978-0-471-20170-0.

External links

- Various 1D, 2D and 3D demonstrations of EM together with Mixture Modeling (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_2D_PointSegmentation_EM_Mixture) are provided as part of the paired SOCR activities and applets. These applets and

activities show empirically the properties of the EM algorithm for parameter estimation in diverse settings.

- k-MLE: A fast algorithm for learning statistical mixture models (https://arxiv.org/abs/1203.5181)
- Class hierarchy in (https://github.com/l-/CommonDataAnalysis)C++ (GPL) including Gaussian Mixtures
- The on-line textbook: Information Theory, Inference, and Learning Algorithms (http://www.inference.phy.cam.ac.uk/mackay/itila/), by David J.C. MacKay includes simple examples of the EM algorithm such as clustering using the soft *k*-means algorithm, and emphasizes the variational view of the EM algorithm, as described in Chapter 33.7 of version 7.2 (fourth edition).
- Variational Algorithms for Approximate Bayesian Inference (http://www.cse.buffalo.edu/faculty/mbeal/papers/beal03.pdf), by M. J. Beal includes comparisons of EM to Variational Bayesian EM and derivations of several models including Variational Bayesian HMMs (chapters (http://www.cse.buffalo.edu/faculty/mbeal/thesis/index.html)).
- The Expectation Maximization Algorithm: A short tutorial (http://www.seanborman.com/publications/EM_algorithm.pdf), A self-contained derivation of the EM Algorithm by Sean Borman.
- The EM Algorithm (http://pages.cs.wisc.edu/~jerryzhu/cs838/EM.pdf), by Xiaojin Zhu.
- EM algorithm and variants: an informal tutorial (https://arxiv.org/abs/1105.1476) by Alexis Roche. A concise and very clear description of EM and many interesting variants.
