Opinionated Lessons in Statistics

by Bill Press

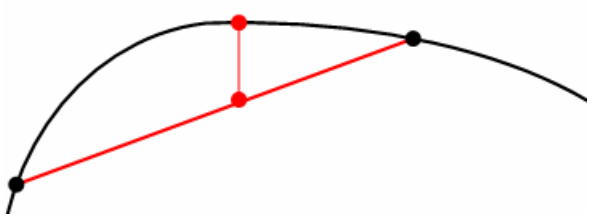
#30 Expectation Maximization (EM) Methods

Let's look at the theory behind EM methods:

Preliminary: Jensen's inequality

If a function is concave (downward), then

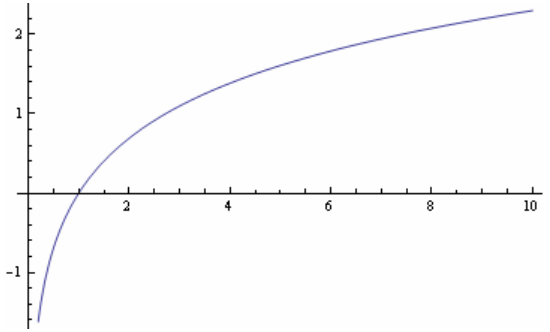
function(interpolation) ≥ interpolation(function)



Log is concave (downward). Jensen's inequality is thus:

$$\text{If} \qquad \sum_i \lambda_i = 1$$

$$\text{Then} \qquad \ln \sum_i \lambda_i Q_i \geq \sum_i \lambda_i \ln Q_i$$



This gets used a lot when playing with log-likelihoods. Proof of the EM method that we now give is just one example.

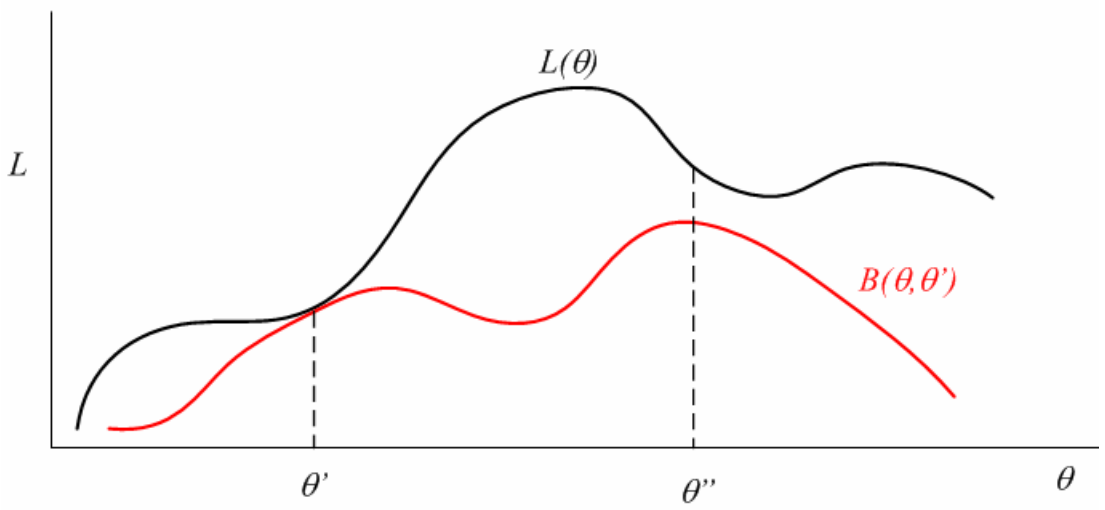
The basic EM theorem (Dempster, Laird, and Rubin):

$\mathbf{x}$ are the data
$\mathbf{z}$ are missing data or nuisance variables
$\boldsymbol{\theta}$ are parameters to be determined

Find $\boldsymbol{\theta}$ that maximizes the log-likelihood of the data:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &\equiv \ln P(\mathbf{x}|\boldsymbol{\theta}) \\
&= \ln \left[\sum_{\mathbf{z}} P(\mathbf{x}|\mathbf{z}\boldsymbol{\theta}) P(\mathbf{z}|\boldsymbol{\theta}) \right] \qquad \text{marginalize over z} \\
&= \ln \left[\sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') \frac{P(\mathbf{x}|\mathbf{z}\boldsymbol{\theta}) P(\mathbf{z}|\boldsymbol{\theta})}{P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}')} \right] - \ln P(\mathbf{x}|\boldsymbol{\theta}') + L(\boldsymbol{\theta}') \\
&\geq \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') \ln \left[\frac{P(\mathbf{x}|\mathbf{z}\boldsymbol{\theta}) P(\mathbf{z}|\boldsymbol{\theta})}{P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') P(\mathbf{x}|\boldsymbol{\theta}')} \right] + L(\boldsymbol{\theta}') \\
&= \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') \ln \left[\frac{P(\mathbf{x}\mathbf{z}|\boldsymbol{\theta})}{P(\mathbf{z}\mathbf{x}|\boldsymbol{\theta}')} \right] + L(\boldsymbol{\theta}') \\
&\equiv B(\boldsymbol{\theta}, \boldsymbol{\theta}') \quad \text{for any } \boldsymbol{\theta}', \text{ a bound on } L(\boldsymbol{\theta})
\end{aligned}
$$

Notice that at $\boldsymbol{\theta} = \boldsymbol{\theta}'$ we have $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}')$,
so the bound touches the actual likelihood:



So, if we maximize $B(\boldsymbol{\theta}, \boldsymbol{\theta}')$ over $\boldsymbol{\theta}$, we are guaranteed that the new max $\boldsymbol{\theta}''$ will increase $L(\boldsymbol{\theta})$. This can terminate only by converging to (at least a local) max of $L(\boldsymbol{\theta})$ (Can you see why?)

And it works whether the maximization step is local or global.

So the general EM algorithm repeats the maximization:

$$\boldsymbol{\theta}'' = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') \ln \left[\frac{P(\mathbf{xz}|\boldsymbol{\theta})}{P(\mathbf{zx}|\boldsymbol{\theta}')} \right]$$

$$= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') \ln \left[P(\mathbf{xz}|\boldsymbol{\theta}) \right]$$

$$= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') \ln \left[P(\mathbf{x}|\mathbf{z}\boldsymbol{\theta}) P(\mathbf{z}|\boldsymbol{\theta}) \right]$$

each form is sometimes useful

sometimes (missing data) no dependence on z

sometimes (nuisance parameters) a uniform prior

computing this is the E-step

This is an expectation that can often be computed in some better way than literally integrating over all possible values of z.

maximizing this is the M-step

This is a general way of handling missing data or nuisance parameters if you can estimate the probability of what is missing, given what you see (and a parameters guess).

Might not be instantly obvious how GMM fits this paradigm!

$\mathbf{z}$ (missing) is the assignment of data points to components
$\boldsymbol{\theta}$ consists of the $\boldsymbol{\mu}$s and $\boldsymbol{\Sigma}$s

$$P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') \to p_{nk}$$

$$\sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}\boldsymbol{\theta}') \ln\left[P(\mathbf{x}|\boldsymbol{\theta})\right] \to -\sum_{n,k} p_{nk}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)\cdot\boldsymbol{\Sigma}_k^{-1}\cdot(\mathbf{x}_n - \boldsymbol{\mu}_k) - \ln\det\boldsymbol{\Sigma}_k\right]$$

Showing that this is maximized by the previous re-estimation formulas for μ and Σ is a multidimensional (fancy) form of the theorem that the mean is the measure of central tendency that minimizes the mean square deviation.

See Wikipedia: Expectation-Maximization Algorithm for detailed derivation.

The next time we see an EM method will be when we discuss Hidden Markov Models. The "Baum-Welch re-estimation algorithm" for HMMs is an EM method.