Expectation Maximization

- an approach to the iterative computation of maximum likelihood (ML) estimates
- useful in a variety of incomplete-data problems, where algorithms such as the Newton-Raphson method may turn out to be more complicated
- it is applied to *incomplete-data problems*, where ML estimation is made difficult by the absence of some part of data in a more familiar and simple data structure
- on each iteration of the EM algorithm, there are two steps, called, the expectation step (E-step) and the maximization step (M-step)
- because of this, the algorithm is named as EM by Dempster

Expectation Maximization

- it is closely related to the *ad hoc* approach to estimation with missing data,
- where the parameters are estimated after filling in initial values for the missing data
- the latter are then updated by their predicted values using these initial parameter estimates
- the parameters are then reestimated, and so on, proceeding iteratively until convergence

Expectation Maximization

- the basic idea of the EM algorithm is to associate with the given incomplete-data problem, a complete-data problem for which ML estimation is computationally more tractable
- the methodology of the EM algorithm then consists in reformulating the problem in terms of this more easily solvable complete-data problem
- establishing a relationship between the likelihoods of these two problems, and
- exploiting the simpler MLE computation of the complete-data problem in the M-step of the iterative computing algorithm

Expectation Maximization Algorithm

- the E-step consists in manufacturing data for the complete-data problem, using the observed data set of the incomplete-data problem and the current value of the parameters, so that
- the simpler M-step computation can be applied to this "completed" data set
- more precisely, it is the log likelihood of the complete-data problem that is "manufactured" in the E-step
- as it is based partly on unobservable data, it is replaced by its conditional expectation given the observed data, where this E-step is effected using the current fit for the unknown parameters

Expectation Maximization

- let, $\boldsymbol{Y}$ be the random vector corresponding to the observed data $\boldsymbol{y}$, having p.d.f. postulated as $g(\boldsymbol{y};\boldsymbol{\Psi})$, where
- $\boldsymbol{\Psi} = (\Psi_1, \dots, \Psi_d)^T$ is a vector of unknown parameters with parameter space $\boldsymbol{\Omega}$.
- the vector $\boldsymbol{\Psi}$ can be estimated by maximum likelihood approach
- the likelihood function for $\boldsymbol{\Psi}$ formed from the observed data $\boldsymbol{y}$ is given by

$$L(\boldsymbol{\Psi}) = g(\boldsymbol{y};\boldsymbol{\Psi})$$

- an estimate $\hat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$ can be obtained as a solution of the likelihood equation

$$\partial L\left(\boldsymbol{\Psi}\right)/\partial\boldsymbol{\Psi} = 0$$

Expectation Maximization

- or equivalently,

$$\partial \log L(\boldsymbol{\Psi}) / \partial \boldsymbol{\Psi} = 0$$

- the notion of incomplete data comes where the complete data may contain some variables that are never observable in a data sense
- within this framework, $\boldsymbol{x}$ denotes the vector containing the augmented or so-called complete data
- let $\boldsymbol{z}$ be the vector containing the additional data, referred to as the unobservable or missing data
- the p.d.f. of the random vector $\boldsymbol{X}$ corresponding to the complete-data vector $\boldsymbol{x}$ is denoted by $g_c(\boldsymbol{x}; \boldsymbol{\Psi})$

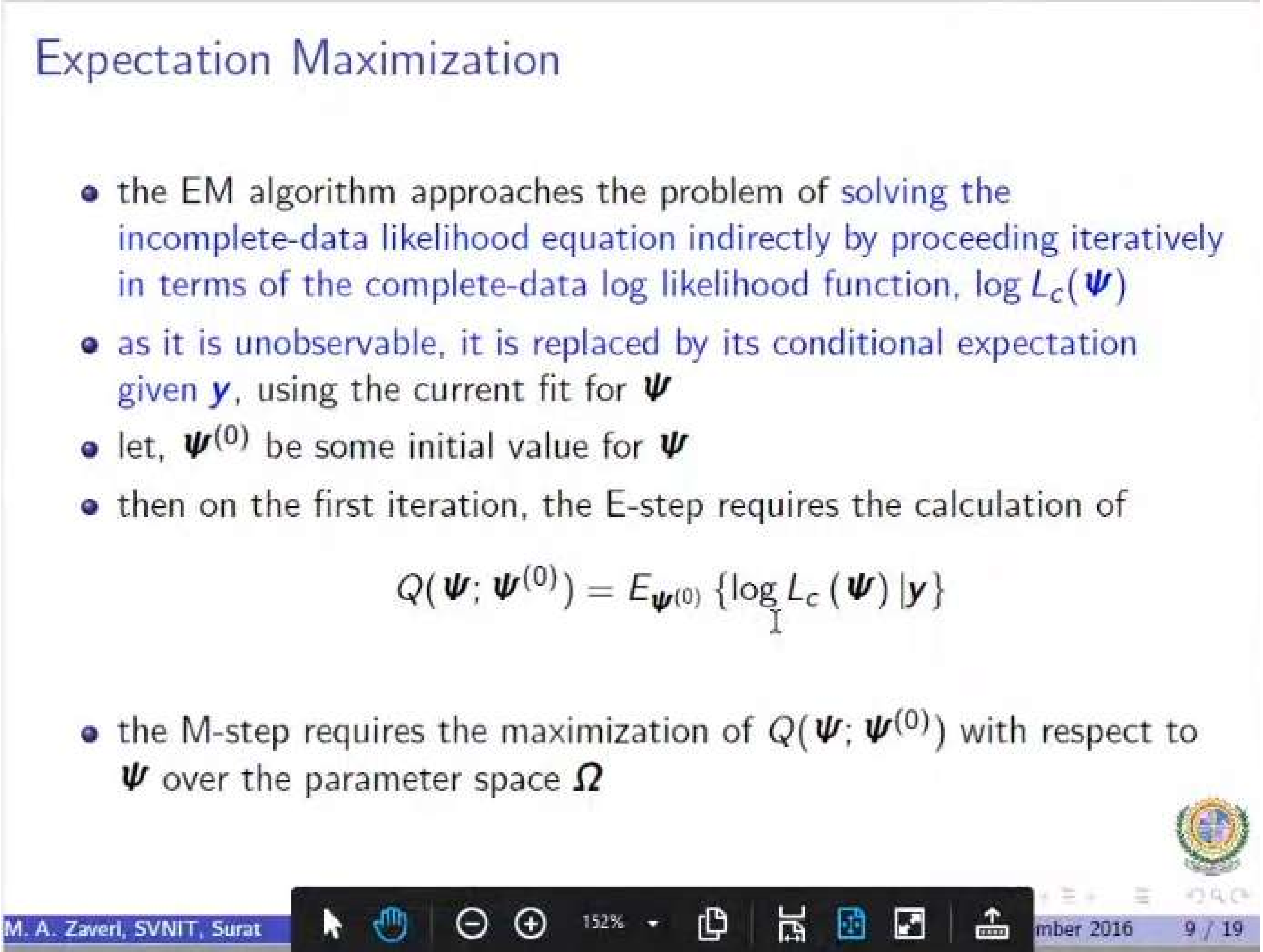
Expectation Maximization

$$\log L_c(\boldsymbol{\Psi}) = \log g_c(\boldsymbol{x}; \boldsymbol{\Psi})$$

- formally, two sample spaces $\mathcal{X}$ and $\mathcal{Y}$ and a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$
- instead of observing the complete-data vector $\boldsymbol{x}$ in $\mathcal{X}$, we observe the incomplete-data vector $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ in $\mathcal{Y}$
- it follows that

$$g(\boldsymbol{y}; \boldsymbol{\Psi}) = \int_{\mathcal{X}(\boldsymbol{y})} g_c(\boldsymbol{x}; \boldsymbol{\Psi}) d\boldsymbol{x}$$

where $\mathcal{X}(\boldsymbol{y})$ is the subset of $\mathcal{X}$ determined by the equation $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$



Expectation Maximization

- the EM algorithm approaches the problem of solving the incomplete-data likelihood equation indirectly by proceeding iteratively in terms of the complete-data log likelihood function, $\log L_c(\boldsymbol{\Psi})$
- as it is unobservable, it is replaced by its conditional expectation given $\boldsymbol{y}$, using the current fit for $\boldsymbol{\Psi}$
- let, $\boldsymbol{\Psi}^{(0)}$ be some initial value for $\boldsymbol{\Psi}$
- then on the first iteration, the E-step requires the calculation of

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)}) = E_{\boldsymbol{\Psi}^{(0)}} \{\log L_c(\boldsymbol{\Psi}) | \boldsymbol{y}\}$$

- the M-step requires the maximization of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)})$ with respect to $\boldsymbol{\Psi}$ over the parameter space $\boldsymbol{\Omega}$

Expectation Maximization

- choose $\boldsymbol{\Psi}^{(1)}$ such that

$$Q(\boldsymbol{\Psi}^{(1)}; \boldsymbol{\Psi}^{(0)}) \geq Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)})$$

  for all $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$
- the E and M steps are then carried out again, but this time with $\boldsymbol{\Psi}^{(0)}$ replaced by the current fit $\boldsymbol{\Psi}^{(1)}$
- on the $(k+1)$th iteration, the E and M steps are defined as follows:
  - **E-step** Calculate $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$, where

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}} \{\log L_c(\boldsymbol{\Psi}) | \boldsymbol{y}\}$$

  - **M-step** Choose $\boldsymbol{\Psi}^{(k+1)}$ to be any value of $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$ that maximizes $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$; that is

$$Q(\boldsymbol{\Psi}^{(k+1)}; \boldsymbol{\Psi}^{(k)}) \geq Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$$

    for all $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$

Expectation Maximization

- the E and M steps are alternated repeatedly until the difference

  $$L(\boldsymbol{\Psi}^{(k+1)}) - L(\boldsymbol{\Psi}^{(k)})$$

  changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k+1)})\}$
- it is shown that the (incomplete-data) likelihood function $L(\boldsymbol{\Psi})$ is not decreased after an EM iteration; that is,

  $$L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)})$$

  for $k = 0, 1, 2, \ldots$
- hence convergence must be obtained with a sequence of likelihood values that are bounded above

Expectation Maximization

- another way of expressing is that $\boldsymbol{\Psi}^{(k+1)}$ belongs to

$$\mathcal{M}(\boldsymbol{\Psi}^{(k)}) = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$$

  which is the set of points that maximize $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)})$
- it is not necessary to specify the exact mapping from $\mathcal{X}$ to $\mathcal{Y}$, nor
- the corresponding representation of the incomplete-data density g in terms of the complete-data density g_c
- all that is necessary is the specification of the complete-data vector $\boldsymbol{x}$ and the conditional density of $\boldsymbol{X}$ given the observed data vector $\boldsymbol{y}$
- specification of this conditional density is needed in order to carry out the E-step

Expectation Maximization for MAP estimation

- the EM algorithm is modified to produce the maximum a posteriori (MAP) estimate in incomplete-data problems
- the computation of the MAP estimate in a Bayesian framework via the EM algorithm, corresponding to some prior density
- $p(\boldsymbol{\Psi})$ for $\boldsymbol{\Psi}$ is described as follows
- let, the incomplete and complete data posterior densities for $\boldsymbol{\Psi}$ be given by
- $p(\boldsymbol{\Psi}|\boldsymbol{y})$ and $p(\boldsymbol{\Psi}|\boldsymbol{x})$, respectively
- then the MAP estimate of $\boldsymbol{\Psi}$ is the value of $\boldsymbol{\Psi}$ that maximizes the log (incomplete-data) posterior density

$$\log p(\boldsymbol{\Psi}|\boldsymbol{y}) = \log L(\boldsymbol{\Psi}) + \log p(\boldsymbol{\Psi})$$

here $p(\cdot)$ is being used as a generic symbol for a p.d.f.

Expectation Maximization for MAP estimation

- the EM algorithm is implemented as follows to compute the MAP estimate
  - **E-step** on the $(k+1)$th iteration,
  - calculate the conditional expectation of the log complete-data posterior density given the observed data vector $\boldsymbol{y}$, using the current MAP estimate $\boldsymbol{\Psi}^{(k)}$ of $\boldsymbol{\Psi}$
  - that is, calculate

$$E_{\boldsymbol{\Psi}^{(k)}}\left\{\log p\left(\boldsymbol{\Psi}|\boldsymbol{x}\right)|\boldsymbol{y}\right\} = Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) + \log p(\boldsymbol{\Psi})$$

  - **M-step** choose $\boldsymbol{\Psi}^{(k+1)}$ to maximize over $\boldsymbol{\Psi}^{(k)} \in \boldsymbol{\Omega}$
- the E-step is effectively the same as seen earlier
- the M-step differs in that the objective function for the maximization process is equal to $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ augmented by the log prior density, $\log p(\boldsymbol{\Psi})$
- the presence of this latter term as the result of the imposition of a Bayesian prior for $\boldsymbol{\Psi}$

Expectation Maximization for MAP estimation

- the EM algorithm has several properties relative to other iterative algorithms such as Newton-Raphson and Fisher's scoring method for finding MLEs
- some of its advantages compared to its competitors are as follows:
  - the EM algorithm is numerically stable with each EM iteration increasing the likelihood (except at a fixed point of the algorithm)
  - under favourable conditions, the EM algorithm has reliable global convergence, i.e.
  - starting from an arbitrary point $\boldsymbol{\Psi}^{(0)}$ in the parameter space, convergence is nearly always to a local maximizer
  - the EM algorithm is easily implemented, because
  - the E-step of each iteration only involves taking expectations over complete-data conditional distributions and
  - the M-step of each iteration only requires complete-data ML estimation, which is often in simple closed form

Expectation Maximization for MAP estimation

- the EM algorithm is easy to program, since no evaluation of the likelihood nor its derivatives is involved
- the cost per iteration is generally low, which can offset the larger number of iterations needed for the EM algorithm compared to other competing procedures
- by watching the monotone increase in likelihood over iterations, it is easy to monitor convergence
- the EM algorithm can be used to provide estimated values of the 'missing' data

Expectation Maximization for MAP estimation

- Some of the criticisms (disadvantages) of the EM algorithm are as follows:
    - the EM algorithm may converge slowly even in some seemingly inocuous problems and in problems where there is too much 'incomplete information'
    - in some problems, the E-step may be analytically intractable
    - the EM algorithm like the Newton-type methods does not guarantee convergence to the global maximum when there are multiple maxima
    - in this case, the estimate obtained depends upon the initial value

Expectation Maximization for MAP estimation

- in general, no optimization algorithm is guaranteed to converge to a global or local maximum, and
- the EM algorithm is not magical to this regard

Thank You