**Part I: Describe the difference between EM and SEM**

*When EM is used and how it works*

The expectation maximisation algorithm (EM) is an iterative technique used to find maximum likelihood parameters of a model in such cases, where the equations cannot be solved directly. In complete data problems, maximum likelihood estimation is based on solving the first-order-condition in order to estimate the parameters. However, if our statistical model involves latent variables, this is typically not possible. Especially for incomplete-data problems the EM algorithm is therefore heavily used.

The EM algorithm works as follows: Suppose we want to find the ML estimates for some parameters (Θ) in an incomplete-data problem. Then the ML estimate of those parameters is determined by the marginal likelihood of the observed data. Typically this quantity is not solvable by normal MLE and we use the EM algorithm to find the ML estimate of the marginal likelihood. Hereby, the algorithm iteratively applies two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step it calculates the expected value of the log likelihood function w.r.t. the conditional distribution of misingdata|observeddata and for the current estimate for the parameters. For that we start with an initial value Θ(0):

Q(Θ| Θ(t)) = EYmis|Yobs, Θ (t)[log L(Θ; Yobs, Ymis)] = ∫logL(Θ|Y) \* f(Ymis|Yobs, Θ = Θ(t))dYmis

In the M-step, the parameter vector Θ is updated by finding the parameters that maximize the conditional expectation from the E-step:

Θ(t+1) = argmax Q(Θ| Θ(t))

These updated parameter estimates are used in the next E-step to determine the distribution of the latent variables/missing data. The algorithm iterates between the E-step and the M-step until some stopping criterion is fulfilled.

Alternatives to EM: *moment-based approaches* or the so-called *spectral technique* have better convergence in high-dimensional cases*.*

*Advantages & disadvantages of EM for missing data problems*

+ by each iteration we improve: Q(Θ(t+1)|Θ(t)) ≥ Q(Θ|Θ(t)), for all t

+ works well in practise and quite easy to understand

+ easily implemented for cases from the exponential family since it uses identical computational method as MLE in the M-step.

- EM does not necessarily converge to a global optimum

- convergence can be slow, especially in high dimensional cases (because of many local optima)

- works well for exponential family but less well for other classes of problems

- does not provide the asymptotic variance-covariance matrix of the maximum likelihood estimator (in contrast, this is a natural byproduct of some other methods, e.g. Newton-Rhapson).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*How SEM works*

The SEM algorithm (supplemented EM) is procedure that supplements EM and obtains numerically stable asymptotic VCOV matrices for parameters. By applying SEM, we take use of the code for EM itself as well as the code for computing the complete-data VCOV matrix and code for standard matrix operations. Basically, the algorithm “uses the fact that the rate of convergence of EM is governed by the fractions of missing information”. It has been shown that the VCOV matrix can be computed from the complete-data VCOV matrix plus a term representing the increase in variance due to the missing data. In the case of a single parameter:

V = Vc + ∆V,where ∆V = [r/(1-r)]\*Vc and r is the rate of convergence of EM

In the case of multiple parameters, we can write:

V = I-1oc + ∆V, where ∆V = I-1ocDM(I - DM)-1, Ioc = E[Io(Θ |Y)|Yobs, Θ]│Θ = Θ\* andIo(Θ |Y) is the complete-data observed information matrix (which would be difficult to evaluate)

The SEM algorithm therefore consists of three computational parts:

1.) Evaluation of I-1oc (= Expectation over the conditional distribution f(Ymis|Yobs, Θ) of observed information matrix Io(Θ |Yobs) = )

For exponential family cases: Io(Θ|Y) = Io(Θ|S(Y)) is a linear function of S(Y), where S(Y) is “a vector of complete-data sufficient statistics”. Thus Ioc = Io(Θ\*|S\*(Yobs)). We obtain S\*(Yobs) = E[S(Y)|Yobs, Θ\*] at the last E-step. Consequently, we obtain I-1oc just by plugging in S\*(Yobs) for S(Y) in Io(Θ\*|S\*(Yobs)) = Ioc. (For non-exponential family cases, the approach is a bit different but nevertheless solvable)

2.) Computation of DM (= )│ Θ= Θ\*, “rate of convergence matrix” )

All quantities here can be obtained by sing only the EM-code! We have for each element in DM

rij = = [Θj(t+1)(i) – Θj\*] / [Θi(t) – Θi\*] for j=1,…d

So, we obtain rij after having computed Θ\* by EM. However, different ways to obtain DM exist.

Note: when some variables have no missing values, EM will converge in one step to Θ\* (independent from the starting value choice).

3.) Evaluation of V = I-1oc + ∆V

Having computed I-1oc and ∆V we obtain the VCOV matrix V. However, this matrix can be (numerically) asymmetric. This is however an advantage of SEM since we can use the (a)symmetry of V to check diagnostics of EM algorithm:

If V is asymmetric, this indicates a programming error in either EM or SEM code.

If V is symmetric but not positive definite, this indicates that EM has not converged only to a saddle point. Hence, by SEM we can monitor convergence of EM to a (local) maximum.

Notes:

* Agile-SEM is a development of SEM that controls the numerical noise intensity on a parameter basis🡪best convergence properties, accuracy and efficiency + requires fewer tuning parameters.
* SEM is established in most opensource and commercial statistical softwares (e.g. ecoML in R)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Advantages of SEM over other techniques of VCOV estimation and general advantages*

Before Meng & Rubin proposed their SEM algorithm (supplemented EM), there existed other techniques for obtaining the asymptotic VCOV matrix after running EM. However, these techniques have major problems that SEM does not have.

Advantages of SEM:

+ symmetry of VCON matrix can be used as diagnostics for EM (to check for needed number of iterations & monitor EM-convergence to (local) maximum)

+ Advantage over bootstrap & jackknife techniques: SEM is also applicable to non-iid cases, under complicated missing data patterns and under several levels of randomness

[+ applicable to multiparameter cases]

+ SEM converges at different steps for different elements of DM, which suggests that VCOV-matrix-computation by other methods (e.g. Newton-Rhapson) is not as accurate.

+ automatic “self-adjustment” 🡪 stability [??-->p.908]

+ The computation of I-1oc is very accurate and SEM obtains typically quite stable and accurate values for DM and hence for ∆V. Consequently, the VCOV matrix V is calculated typically quite accurate as well.

+ parallel computing is possible and can make SEM even faster than EM [p.908]

Disadvantages of SEM:

- (without parallel computing,) SEM needs (d+1)/2 more time than EM (d = size of the dxd VCOV matrix)

- storage for matrices needed (but not a big problem nowadays)