# Some Cautions on Applying the EM Algorithm to a Quality Assessment Application

**Lorrie L. Hoffman, Ph.D.**
**Professor and Chair**
**Department of Mathematics, Armstrong Atlantic State University**

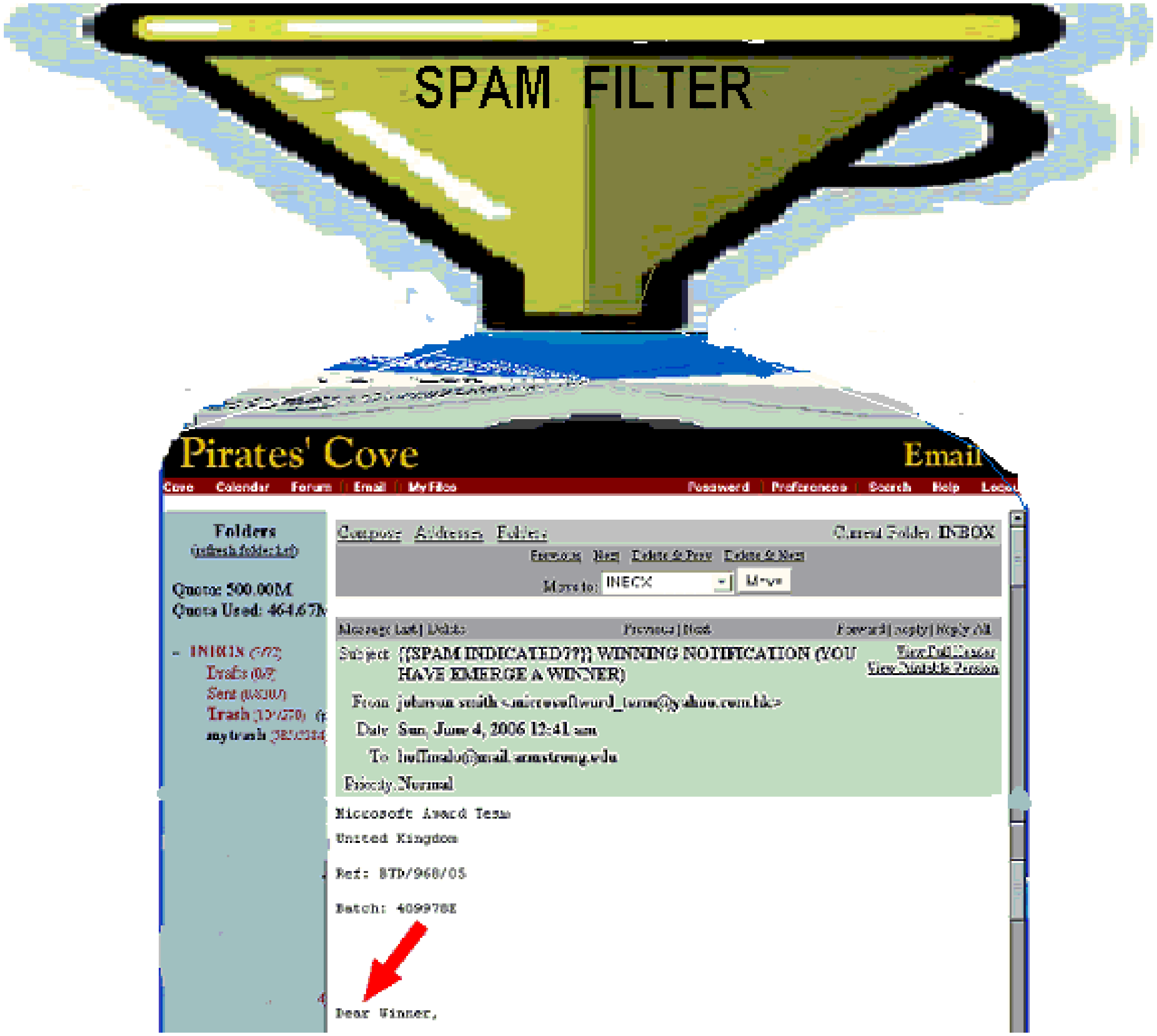
**The problem of handling missing data began to be extensively studied in the late 1970's. The mechanism of solution is inherently a multivariate one with the EM algorithm one potential approach. It has been just a decade ago that journals targeted at quality assessment wrote of future innovations in multivariate applications. Thus in a quality engineering environment, the act of addressing "missingness" in data collection and analysis is a rather new endeavor. Several articles using techniques like EM have recently appeared in the literature. Via an example dealing with proportion of defects, we explore the application of the EM algorithm that allows for use of more data than just the complete pairs. We illustrate the importance of the concept of "missing at random" and its effect on proper convergence to the maximum likelihood estimates.**

**Concepts Addressed by this paper:**

1. **Four different ways to handle missing data: 1) Listwise Deletion 2) Mean Imputation 3) EM algorithm 4) Direct Maximum Likelihood**
2. **The usually held belief that under Missing Completely at Random (MCAR) or Missing at Random (MAR) the order of "goodness" is: Direct ML > EM > mean imputation > listwise deletion**
3. **While it is true that the EM algorithm always converges, in the non-MAR it may not converge to the true estimate**
4. **Under non-MAR what is the "order of goodness" for the different approaches?**

**We explore these questions via an example.**

## AN EXAMPLE OF MARKER-BASED DIAGNOSTIC FILTERS:  SPAM ID-ing



Subject: {{SPAM INDICATED???}} WINNING NOTIFICATION (YOU HAVE EMERGE A WINNER)

From: johnson smith <microsoftword_team@yahoo.com.hk>

Date: Sun, June 4, 2006 12:41 am

To: huffmalo@mail.armstrong.edu

Priority: Normal

```
Microsoft Award Team

United Kingdom

Ref: BTD/968/05

Batch: 409978E

Dear Winner,
```

A simplistic explanation of an email SPAM filter is that it operates to detect defects (i.e. SPAM) in a two-stage manner: 1) it checks for keywords (corpus) known to be in good email and flags a new email as good if it contains those keywords and 2) it checks for keywords (corpus) known to be in SPAM email and marks a new email as SPAM if it contains those keywords.  As a rudimentary example of a SPAM filter consider one that is set to label p=proportion of good emails and 1-p=proportion of SPAM emails.  (Note: these summing to 1.0 is for ease in modeling.) For example, if 75% of the good email I receive contains the word “Armstrong” and 25% of the SPAM email I receive contains the word “Dear” then these are the keywords I would use to search each new email.

## The Statistical Model

**Here let**

$$X1 = \begin{cases} 1 & \text{if IDed as SPAM} \quad \text{with prob } 1-p \\ 0 & \text{if not IDed as not SPAM} \quad \text{with prob p} \end{cases} \qquad X2 = \begin{cases} 1 & \text{if IDed as good} \quad \text{with prob p} \\ 0 & \text{if not IDed as not SPAM} \quad \text{with prob } 1-p \end{cases}$$

**Let U = X1 and V = X1 + X2**

**f( $\vec{X}$ | $\vec{\Phi}$ ) $\underline{=}$ f(u, v | p ) = $p^{1+v-2u}$ $(1-p)^{1-v+2u}$ ; model is of the Regular exponential family since exp{log(p(1-p) +(v-2u)log(p/(1-p)} so K(u, v) is v-2u; <u, v> when both are observed is the complete vector.**

**Following the example, if an incoming email started “Dear Armstrong” then the data vector is: <1,2>, if it contained “Armstrong” but no “Dear” the data vector is: <0,1>, if it contained “Dear” but no “Armstrong” then the data vector is: <1,1> , and if it contained neither word: <0,0>. The probability of observing each of these vectors is:**

**and when p = .75**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **U** | | | ///// | | | **U** | | |
| | | **0** | **1** | | ///// | | | **0** | **1** | |
| | **0** | **p(1-p)** | **0** | **p(1-p)** | /////<br>///// | | **0** | **3/16** | **0** | **3/16** |
| **V** | **1** | **$p^2$** | **$(1-p)^2$** | **$p^2 + (1-p)^2$** | /////<br>///// | **V** | **1** | **9/16** | **1/16** | **10/16** |
| | **2** | **0** | **p(1-p)** | **p(1-p)** | /////<br>///// | | **2** | **0** | **3/16** | **3/16** |
| | | **p** | **(1-p)** | | ///// | | | **12/16** | **4/16** | |

**The grid on the right was used to generate the complete data n = 32 vectors in Table 1.**

# Dealing with Missing Data

Consider the sample data in Table 1 that is “generated” from the model shown above. For ease of computation we only allowed V to be missing and in each of the cases investigated exactly 50% “missingness”. A purely illustrative look since with all of U present we might use: $\hat{p} = 1 - \overline{U}$

## On the E(stimation) and M(aximization) Approach (finding pem)

ASSUMES: $\vec{X}$ is a data vector; $\vec{\Phi}$ is the parameter vector ; f( $\vec{X} | \vec{\Phi}$ ) is the pdf or pmf i.e.

$\int_{-\infty}^{t} f(\vec{x} | \vec{\Phi}) d\vec{x}$ = P(X < t) where f( $\vec{X} | \vec{\Phi}$ ) is a member of the regular exponential family i.e.

f( $\vec{X} | \vec{\Phi}$ ) = exp{p( $\vec{\Phi}$ )K( $\vec{X}$ ) + S( $\vec{X}$ ) +q( $\vec{\Phi}$ )} Note: K( $\vec{X}$ ) can be shown to be a complete sufficient statistic for estimating $\vec{\Phi}$ ; good properties MVUE

let L( $\vec{X}$ , $\vec{\Phi}$ ) = $\prod_{i=1}^{n} f(x_i | \vec{\Phi})$ viewed as a function of $\vec{\Phi}$ when maximized the function of $\vec{X}$ is called the MLE. MLE is a function of K( $\vec{X}$ )

Let $\vec{x}$ represent the hypothetical complete data collected and $\vec{y}$ be the data we were able to observe. Let f( $\vec{x} | \vec{\Phi}$ ) be from a regular exponential family. Let

$\vec{t}^{(p)}$ = E(t( $\vec{X}$ )| $\vec{y}$ , $\vec{\Phi}^{(p)}$ ) and let $\vec{\Phi}^{(p+1)}$ solve $\vec{t}^{(p)}$ = E(t( $\vec{X}$ )| $\vec{\Phi}$ ) then

L( $\vec{X}$ , $\vec{\Phi}^{(p+1)}$ ) $\geq$ L( $\vec{X}$ , $\vec{\Phi}^{(p)}$ ) and $\lim_{p \to \infty} \Phi^{(p)}$ exists.

Explanation: $\vec{t}^{(p)}$ = E(t( $\vec{x}$ )| $\vec{y}$ , $\vec{\Phi}^{(p)}$ ) and

this is the E (or estimation) step

Let $\vec{\Phi}^{(p+1)}$ be the solution of $\vec{t}^{(p)}$= E(t( $\vec{X}$ )| $\vec{\Phi}$ )

this is the M (or maximization) step

Known as the EM algorithm

**How? 1) E-step: look at the distribution of missing variable(s) conditioned on the observed variable(s) and use the mean of that distribution as an estimate for the missing data; 2) M-step: now compute your parameter as usual from the "filled in" data then do it again.**

**For our SPAM Example (recall to simplify the example we have only v missing):**
**What will we use as an estimate for v?**

**Here,** $f(v|u) = \dfrac{p^{1+v-2u}\,(1-p)^{1-v+2u}}{p^{1-u}\,(1-p)^{u}}$ **so** $E(V|U=0) = p$ **and** $E(V|U=1) = p+1$ **used for the E-step.**

**Find the Likelihood**

**L(.) = f(U1, V1| p)*f(U2, V2| p)*…*f(Un,Vn|p) =** $p^{\,n+\sum_{i=1}^{n} v_i - 2\sum_{i=1}^{n} u_i}\,(1-p)^{\,n-\sum_{i=1}^{n} v_i + 2\sum_{i=1}^{n} u_i}$

**dlogL(.)/dp set to zero yields an estimate for p of ½ + ½** $\bar{v}$ **-** $\bar{u}$ **. Needed for the M-step.**

**-----------------------------------------------------------------**

**Examples of Computations for Case 3 of Not Missing at Random shown in Table 1:**

**E-step: plist = .781 =** $p^{(0)}$ **is substituted for missing V values**
**M-step:** $p^{(1)}$ **= 1/2 + 1/2 * [(9*.781 + 7 * (1 + .781) + 11)/32] – 8/32 = .7265; do it again.**

**So, does this process converge? Yes, as proven by Dempster, Laird and Rubin (1977), but unfortunately not to the correct value of the parameter p. What is the limit as** $n \rightarrow \infty$ **of** $p^{(n)}$ **?**

$p^{(n)}$ **= .5 + .5 * [(9 *** $p^{(n-1)}$ **+ 7 * ( 1 +** $p^{(n-1)}$**) +11 ) / 32 ] - .25 so .53125** $\sum_{k=0}^{n}(.25)^{k}$ **+** $(.25)^{n}$ ***** $p^{(0)}$ **= .708**

$p^{(n)}$ **-** $p^{(n-1)}$ **= .53125 *** $(.25)^{n}$ **- .4190625 < 0 for n**$\geq$**1 so our estimate of p is monotonically decreasing.**

**So , is L(** $\vec{X}$ **,** $\vec{\Phi}^{(p+1)}$ **)** $\geq$ **L(** $\vec{X}$ **,** $\vec{\Phi}^{(p)}$ **) ? Viewing L(.) =** $p^{32\hat{p}}\,(1-p)^{32(1-\hat{p})}$ **with** $\hat{p}$ **monotonically decreasing and p>.5, L(.) is monotonically increasing.**

## On the Direct Maximum Likelihood Approach (finding pml)

**L(.) =** $\left( \dfrac{p^{n_c + \sum_{i=1}^{n_c} v_i - 2\sum_{i=1}^{n_c} u_i} (1-p)^{n_c - \sum_{i=1}^{n_c} v_i + 2\sum_{i=1}^{n_c} u_i}}{p^{n_c - \sum_{i=1}^{n_c} u_i} (1-p)^{\sum_{i=1}^{n_c} u_i}} \right) * p^{n_a - \sum_{i=1}^{n_a} u_i} (1-p)^{\sum_{i=1}^{n_a} u_i}$ **where $n_a$ counts all of the data vectors**

**and $n_c$ counts the complete data vectors.**

**Taking a logarithm and then a derivative and setting to zero yields:**

**pml =** $\left( n_a + \sum_{i=1}^{n_c} v_i - \sum_{i=1}^{n_c} u_i - \sum_{i=1}^{n_a} u_i \right) \Big/ (n_a + n_c)$

## On Listwise Deletion (finding plist)

**Calculate ½ + ½ $\bar{v}$ - $\bar{u}$ using only vectors having both u and v (complete cases).**

## On Mean Imputation (finding pmean)

**Calculate $\bar{u}$ and $\bar{v}$ from complete cases and substitute these for missing values then calculate ½ + ½ $\bar{v}$ - $\bar{u}$**

|  |  |  |  | P(v miss\|u.v)=P(vmiss\|u) |  | Case 1 |  | Case 2 |  | Case 3 |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| None Missing |  | MCAR |  | MisAR |  | notMisAR |  | notMisAR |  | notMisAR |  |
| U | V | U | V | U | V | U | V | U | V | U | V |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | . | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | . | 0 | 0 | 0 | . | 0 | 0 |
| 0 | 0 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | 0 |
| 0 | 0 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | 0 |
| 0 | 0 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | . | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | . | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | . | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | . | 0 | . | 0 | 1 | 0 | 1 | 0 | . |
| 0 | 1 | 0 | . | 0 | . | 0 | 1 | 0 | 1 | 0 | . |
| 0 | 1 | 0 | . | 0 | . | 0 | . | 0 | 1 | 0 | . |
| 0 | 1 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | . |
| 0 | 1 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | . |
| 0 | 1 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | . |
| 0 | 1 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | . |
| 0 | 1 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | . |
| 0 | 1 | 0 | . | 0 | . | 0 | . | 0 | . | 0 | . |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | . | 1 | 1 | 1 | . |
| 1 | 1 | 1 | . | 1 | 1 | 1 | . | 1 | . | 1 | . |
| 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | . |
| 1 | 2 | 1 | 2 | 1 | 2 | 1 | . | 1 | . | 1 | . |
| 1 | 2 | 1 | . | 1 | 2 | 1 | . | 1 | . | 1 | . |
| 1 | 2 | 1 | . | 1 | 2 | 1 | . | 1 | . | 1 | . |
| 1 | 2 | 1 | . | 1 | 2 | 1 | . | 1 | . | 1 | . |

|  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| pall | .750 | plist | 0.750 | plist | 0.625 | plist | 0.844 | plist | 0.844 | plist | 0.781 |
|  |  | pmean | 0.750 | pmean | 0.875 | pmean | 0.719 | pmean | 0.781 | pmean | 0.594 |
|  |  | pem | 0.750 | pem | 0.750 | pem | 0.771 | pem | 0.792 | pem | 0.708 |
|  |  | pml | 0.750 | pml | 0.750 | pml | 0.771 | pml | 0.792 | pml | 0.708 |

**Table 1. Listwise Deletion (plist). Mean Substitution (pmean), EM alogorithm (pem), Direct Maxium Likelihood (pml) Approaches for finding parameter p in SPAM example for various missing patterns (MCAR is missing completely at random; MAR is missing at random; notMisAR is not missing at random.**

## Discussion and Conclusions

**By comparing the 4 approaches over the 5 cases (MCAR, MisAR, notMisAR-case1, notMisAR-case2, notMisAR-case3) shown in Table 1, we observe the following:**
**1. For this example, for MCAR, Direct ML and EM and mean imputation and listwise deletion do equally well at discerning that p = .75.**
**2. As proven elsewhere, Direct ML and EM correctly estimate p = .75 in the MisAR (missing at random) situation. The order of "goodness" is: Direct ML = EM > mean imputation = listwise deletion**
**3. Neither mean imputation nor listwise deletion do well in this MisAR situation.**
**4. None of the approaches find the true p = .75 for this example in any of the 3 cases of notMisAR. So even though it has been proven that the EM algorithm will converge, the process is not converging to the true p = .75 value.**
**5. For this example with notMisAR-case1, Direct ML and EM tie for being closest to the true p = .75.**
**6. For this example with notMisAR-case2, mean imputation wins for being closest to the true p =.75.**
**7. For this example with notMisAR-case3, listwise deletion wins for being closest to the true p =.75.**
**8. In this example in all 5 cases the Direct ML and the EM give the same answer and never are "very far" from p = .75.**

**The practical conclusion from this study is that we would use the easier to apply formula of the Direct ML when we have data vectors from our SPAM filter process. In the event that we have results only from the "SPAM" check and not from the "good" check for some of the evaluated email, we could use those incomplete vectors along with the complete vectors to discern the underlying mechanism used for filtering. That is, we could reverse-engineer to understand the sensitivity of the double-filtering. In the case with p = .75 we would then know that <0, 1> meaning <notSPAM, good email> will amount to p*p = .56 or about 56% of all email allowed into the INBOX and <1, 2> meaning <SPAM, good> and <0, 0> meaning <notSPAM, notgood> will amount to 2p*(1-p) = .37 or about 37% labeled as "suspicious" and <1, 1> meaning <SPAM, notgood> amounting to about 7%.**

## References

Allison, P.D., *Missing Data*, Sage Publications, Thousand Oaks, CA, 2002.

Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, Vol. 39, 1-22, 1977.

"General FAQ #25: Handling missing or incomplete data," Information Technology Services, University of Texas at Austin, May 10, 2004. http://www.utexas.edu/its/rc/answers/general/gen25.html. Accessed at 4:40 PM on 5/25/2006.

Hoffman, L.L., *Missing Data in Growth Curves*, Ph.D. Thesis, University of Iowa, available through http://wwwlib.umi.com/disserations/gateway , as AAT8128410, 1981.

Kanter, A., "Bayesian spam filters use math that works like magic," in CyberSpeak column, USA Today, 9/17/2004. http://www.usatoday.com/tech/columnist/andrewkantor/2004-09-17-kantor_x.htm . Accessed at 9:03 AM on 6/5/2006.

Montgomery, D.C., and Woodall, W. H. (editors), "A Discussion on Statistically-Based Process Monitoring and Control," *Journal of Quality Technology*, Vol. 29, No. 2, 121-162, 1997.

Ng, H.K.T., Chan, P.S., and Balakrishnan, N., "Optimal Progressive Censoring Plans for the Weibull Distribution," *Technometrics*, vol. 46, no. 4, 470 - 481, 2004.