Fall 2009

The EM algorithm for group testing regression models under matrix pooling

Christopher R. Bilder
University of Nebraska-Lincoln, cbilder3@unl.edu

Boan Zhang
an.zhang@huskers.unl.edu

The EM algorithm for group testing regression models under matrix pooling

Technical report prepared by
Boan Zhang and Christopher R. Bilder
University of Nebraska-Lincoln
Department of Statistics
www.chrisbilder.com/grouptesting
Last updated: 9/15/2009

The array (matrix) group testing procedure is a non-hierarchical procedure proposed by Phatarfod and Sudbury (*Statistics in Medicine*, 1994). The procedure places individual specimens into a matrix-like grid, where specimens are pooled within each row and within each column. In situations where only one row (column) and one or more columns (rows) tests positive, the intersection points of positive rows and columns are positive provided there is no testing error. When more than one row and more than one column test positive, ambiguities arise on which of these individuals at the intersections led to the positive row and column test results. When testing errors are present, we may also have one or more rows testing positive and no columns testing positive (or vice versa). To clear these ambiguities, additional testing (usually on each individual) can be used to complete the decoding. For example, in the frequently used SA1 testing protocol of Phatarfod and Sudbury (1994), individuals who do not fall into the intersection points of a positive row and column are declared to be negative; individuals who do fall into the intersection points of a positive row and column are considered to be candidates for being positive, and they will receive individual retests.

Due to the dependence among the row and column group responses, the regression model fitting procedure of Vansteelandt et al. (2000) can not be used because it maximizes a binomial likelihood of independent *group* responses. Instead, the model fitting procedure proposed by Xie (2001) needs to be used in order to maximize a binomial likelihood written in terms of the independent *individual* responses. Let $\tilde{Y}_{ij}$ be a binary random variable (0 denotes negative, 1 denotes positive) for the true status of the individual in row i and column j ($i = 1, \ldots, I$ and $j = 1, \ldots, J$), and let $P(\tilde{Y}_{ij} = 1) = p_{ij} = \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij}) / [1 + \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})]$ for a $p\times1$ vector of covariates $\boldsymbol{x}_{ij}$ and a $p\times1$ vector of parameters $\boldsymbol{\beta}$. Simply, for one array, the likelihood function is $L(\boldsymbol{\beta} \mid \tilde{\boldsymbol{y}}) = \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{\tilde{y}_{ij}}(1-p_{ij})^{1-\tilde{y}_{ij}}$, where $\tilde{\boldsymbol{y}} = (\tilde{y}_{11},\ldots,\tilde{y}_{IJ})'$ is a $IJ\times1$ vector of individual responses. For more than one array, a third product can be taken in $L(\boldsymbol{\beta} \mid \tilde{\boldsymbol{y}})$ over $k = 1,\ldots, K$ arrays; only the one-array case will be discussed here for brevity.

Because the individual responses are not observed directly for some individuals and because testing error usually exists, the expectation-maximization (EM) algorithm is used to maximize the likelihood function. Initially, only the row and column responses are observed, so these will be conditioned on in

the E-Step. Let $\boldsymbol{R} = (R_1, \ldots, R_I)'$ and $\boldsymbol{C} = (C_1, \ldots, C_J)'$ be vectors of row and column responses, respectively. Retesting can also be carried out on some of the specimens to obtain additional information to fit the model. Experience shows that this additional information typically speeds up convergence for the EM algorithm in comparison to if retesting was not performed. Because these retests are generally performed on individuals exclusively, we will consider this case here only. Without loss of generality, we denote these observed individual responses as $\boldsymbol{Y}_Q = (Y_{ij})_{(i,j)\in Q}$ where Q is the index set pertaining to the individual tests. These individual responses Y_{ij} (0 denotes negative, 1 denotes positive) may or may not be equal to the true responses of $\tilde{Y}_{ij}$ due to the possibility of testing error. Under the frequently used SA1 testing protocol, individual testing is performed when $R_i = 1$ and $C_j = 1$. In this case, $Q = \{(s,t) \mid R_s = 1, C_t = 1, 1 \le s \le I, 1 \le t \le J\}$. Other protocols, like the one described in Kim et al. (2007), allow for retests of all specimens within a positive row (column) when all columns (rows) test negative. In the following discussion, we will consider the general case where some specimens receive individual tests. If there are no individual tests performed at all, we can let $Q = \emptyset$.

For the E-step, one needs to find $E(\tilde{Y}_{ij} \mid \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q) = P(\tilde{Y}_{ij} = 1 \mid \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q) \equiv \omega_{ij}$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Because one cannot write out a closed form expression of known quantities for these probabilities, one can employ a Gibbs sampling approach to estimate them. This involves successive sampling from the univariate conditional distribution of $\tilde{Y}_{ij}$ given $\boldsymbol{R} = \boldsymbol{r}$, $\boldsymbol{C} = \boldsymbol{c}$, $\boldsymbol{Y}_Q = \boldsymbol{y}_Q$ and all of the other true individual responses, and this sampling is done for each i and j. After a large enough set of samples is taken, all of the simulated $\tilde{y}_{ij}$ values for each i and j can be averaged over to find an estimate of ω_{ij}. The conditional probability expression and the algorithm itself are described more explicitly next.

For a given row and column combination (i, j), define $\tilde{\boldsymbol{Y}}_{-i,-j} = \{\tilde{Y}_{i',j'} : i' = 1,\ldots,I, j' = 1,\ldots,J, (i',j') \neq (i,j)\}$; i.e., all possible true individual response random variables excluding $\tilde{Y}_{ij}$. We can find an expression of $P(\tilde{Y}_{ij} = 1 \mid \tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)$ as

$$P(\tilde{Y}_{ij} = 1 \mid \tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q) = \frac{P(\tilde{Y}_{ij} = 1, \tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)}{P(\tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)}$$

First, to find the numerator,

$$
\begin{aligned}
&P(\tilde{Y}_{ij}=\tilde{y}_{ij},\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j},\boldsymbol{R}=\boldsymbol{r},\boldsymbol{C}=\boldsymbol{c},\boldsymbol{Y}_Q=\boldsymbol{y}_Q)\\
&=P(\boldsymbol{R}=\boldsymbol{r},\boldsymbol{C}=\boldsymbol{c},\boldsymbol{Y}_Q=\boldsymbol{y}_Q\mid\tilde{Y}_{ij}=\tilde{y}_{ij},\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j})P(\tilde{Y}_{ij}=\tilde{y}_{ij})P(\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j})\\
&=P(\boldsymbol{R}=\boldsymbol{r},\boldsymbol{C}=\boldsymbol{c}\mid\tilde{Y}_{ij}=\tilde{y}_{ij},\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j},\boldsymbol{Y}_Q=\boldsymbol{y}_Q)\times\\
&\quad P(\boldsymbol{Y}_Q=\boldsymbol{y}_Q\mid\tilde{Y}_{ij}=\tilde{y}_{ij},\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j})P(\tilde{Y}_{ij}=\tilde{y}_{ij})P(\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j})\\
&=P(\boldsymbol{R}=\boldsymbol{r}\mid\tilde{Y}_{ij}=\tilde{y}_{ij},\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j})P(\boldsymbol{C}=\boldsymbol{c}\mid\tilde{Y}_{ij}=\tilde{y}_{ij},\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j})\times\\
&\quad\left[\prod_{(s,t)\in Q}P(Y_{st}=y_{st}\mid\tilde{Y}_{st}=\tilde{y}_{st})\right]\times p_{ij}^{\tilde{y}_{ij}}(1-p_{ij})^{1-\tilde{y}_{ij}}\prod_{\substack{i'=1\\\{i'\neq i,\ j'\neq j\}}}^{I}\prod_{j'=1}^{J}p_{i'j'}^{\tilde{y}_{i'j'}}(1-p_{i'j'})^{1-\tilde{y}_{i'j'}}\ \text{(due to the usual}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{conditional assumption)}
\end{aligned}
$$

where

$$\prod_{\substack{i'=1\\\{i'\neq i,\ j'\neq j\}}}^{I}\prod_{j'=1}^{J}p_{i'j'}^{\tilde{y}_{i'j'}}(1-p_{i'j'})^{1-\tilde{y}_{i'j'}}$$

denotes the product is taken over all combinations of $i' = 1,\ldots, I$ and $j' = 1, \ldots, J$ except the (i, j) combination. Then

$$
\begin{aligned}
&P(\tilde{Y}_{ij}=\tilde{y}_{ij},\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j},\boldsymbol{R}=\boldsymbol{r},\boldsymbol{C}=\boldsymbol{c},\boldsymbol{Y}_Q=\boldsymbol{y}_Q)\\
&=\left[\prod_{i=1}^{I}P(R_i=r_i\mid\tilde{Y}_{i1}=\tilde{y}_{i1},\ldots,\tilde{Y}_{iJ}=\tilde{y}_{iJ})\right]\left[\prod_{j=1}^{J}P(C_j=c_j\mid\tilde{Y}_{1j}=\tilde{y}_{1j},\ldots,\tilde{Y}_{Ij}=\tilde{y}_{Ij})\right]\times\\
&\quad\left[\prod_{(s,t)\in Q}P(Y_{st}=y_{st}\mid\tilde{Y}_{st}=\tilde{y}_{st})\right]\times\left[\prod_{i=1}^{I}\prod_{j=1}^{J}p_{ij}^{\tilde{y}_{ij}}(1-p_{ij})^{1-\tilde{y}_{ij}}\right]
\end{aligned}
$$

due to the independence among row responses and among column responses. Noting that $\tilde{Y}_{ij}=1$ is in the numerator of $P(\tilde{Y}_{ij}=1\mid\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j},\boldsymbol{R}=\boldsymbol{r},\boldsymbol{C}=\boldsymbol{c},\boldsymbol{Y}_Q=\boldsymbol{y}_Q)$, we find that if $(i, j) \in Q$:

$$
\begin{aligned}
&P(\tilde{Y}_{ij}=1,\tilde{\boldsymbol{Y}}_{-i,-j}=\tilde{\boldsymbol{y}}_{-i,-j},\boldsymbol{R}=\boldsymbol{r},\boldsymbol{C}=\boldsymbol{c},\boldsymbol{Y}_Q=\boldsymbol{y}_Q)\\
&=\left[\prod_{i=1}^{I}P(R_i=r_i\mid\tilde{Y}_{i1}=\tilde{y}_{i1},\ldots,\tilde{Y}_{iJ}=\tilde{y}_{iJ})\right]\left[\prod_{j=1}^{J}P(C_j=c_j\mid\tilde{Y}_{1j}=\tilde{y}_{1j},\ldots,\tilde{Y}_{Ij}=\tilde{y}_{Ij})\right]\times\\
&\quad\left[\prod_{(s,t)\in Q}P(Y_{st}=y_{st}\mid\tilde{Y}_{st}=\tilde{y}_{st})\right]\times\left[\prod_{i=1}^{I}\prod_{j=1}^{J}p_{ij}^{\tilde{y}_{ij}}(1-p_{ij})^{1-\tilde{y}_{ij}}\right]\\
&=\left[\prod_{\substack{i'=1\\i'\neq i}}^{I}P(R_{i'}=r_{i'}\mid\tilde{Y}_{i'1}=\tilde{y}_{i'1},\ldots,\tilde{Y}_{i'J}=\tilde{y}_{i'J})\right]\left[\prod_{\substack{j'=1\\j'\neq j}}^{J}P(C_{j'}=c_{j'}\mid\tilde{Y}_{1j'}=\tilde{y}_{1j'},\ldots,\tilde{Y}_{Ij'}=\tilde{y}_{Ij'})\right]\times\\
&\quad\left[\prod_{(s,t)\in Q\setminus\{(i,j)\}}P(Y_{st}=y_{st}\mid\tilde{Y}_{st}=\tilde{y}_{st})\right]\times\left[\prod_{\substack{i'=1\\\{i'\neq i,\ j'\neq j\}}}^{I}\prod_{j'=1}^{J}p_{i'j'}^{\tilde{y}_{i'j'}}(1-p_{i'j'})^{1-\tilde{y}_{i'j'}}\right]\times\\
&\quad P(R_i=r_i\mid\tilde{Y}_{i1}=\tilde{y}_{i1},\ldots,\tilde{Y}_{ij}=1,\ldots,\tilde{Y}_{iJ}=\tilde{y}_{iJ})\times P(C_j=c_j\mid\tilde{Y}_{1j}=\tilde{y}_{1j},\ldots,\tilde{Y}_{ij}=1,\ldots,\tilde{Y}_{Ij}=\tilde{y}_{Ij})\times\\
&\quad P(Y_{ij}=y_{ij}\mid\tilde{Y}_{ij}=1)\times p_{ij}\\
&=\left[\prod_{\substack{i'=1\\i'\neq i}}^{I}P(R_{i'}=r_{i'}\mid\tilde{Y}_{i'1}=\tilde{y}_{i'1},\ldots,\tilde{Y}_{i'J}=\tilde{y}_{i'J})\right]\left[\prod_{\substack{j'=1\\j'\neq j}}^{J}P(C_{j'}=c_{j'}\mid\tilde{Y}_{1j'}=\tilde{y}_{1j'},\ldots,\tilde{Y}_{Ij'}=\tilde{y}_{Ij'})\right]\times\\
&\quad\left[\prod_{(s,t)\in Q\setminus\{(i,j)\}}P(Y_{st}=y_{st}\mid\tilde{Y}_{st}=\tilde{y}_{st})\right]\times\left[\prod_{\substack{i'=1\\\{i'\neq i,\ j'\neq j\}}}^{I}\prod_{j'=1}^{J}p_{i'j'}^{\tilde{y}_{i'j'}}(1-p_{i'j'})^{1-\tilde{y}_{i'j'}}\right]\times\\
&\quad P(R_i=r_i\mid\tilde{R}_i=1)\times P(C_j=c_j\mid\tilde{C}_j=1)\times P(Y_{ij}=y_{ij}\mid\tilde{Y}_{ij}=1)\times p_{ij}
\end{aligned}
$$

where $(s,t) \in Q \setminus \{(i,j)\}$ means all indices in Q except for (i, j) and $\tilde{R}_i$ and $\tilde{C}_j$ are the true values for R_i and C_j, respectively; if $(i, j) \notin Q$:

$$
\begin{aligned}
&P(\tilde{Y}_{ij} = 1, \tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q) \\
&= \left[\prod_{i=1}^{I} P(R_i = r_i \mid \tilde{Y}_{i1} = \tilde{y}_{i1}, \ldots, \tilde{Y}_{iJ} = \tilde{y}_{iJ})\right]\left[\prod_{j=1}^{J} P(C_j = c_j \mid \tilde{Y}_{1j} = \tilde{y}_{1j}, \ldots, \tilde{Y}_{Ij} = \tilde{y}_{Ij})\right] \times \\
&\quad \left[\prod_{(s,t)\in Q} P(Y_{st} = y_{st} \mid \tilde{Y}_{st} = \tilde{y}_{st})\right] \times \left[\prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{\tilde{y}_{ij}}(1-p_{ij})^{1-\tilde{y}_{ij}}\right] \\
&= \left[\prod_{\substack{i'=1 \\ i' \neq i}}^{I} P(R_{i'} = r_{i'} \mid \tilde{Y}_{i'1} = \tilde{y}_{i'1}, \ldots, \tilde{Y}_{i'J} = \tilde{y}_{i'J})\right]\left[\prod_{\substack{j'=1 \\ j' \neq j}}^{J} P(C_{j'} = c_{j'} \mid \tilde{Y}_{1j'} = \tilde{y}_{1j'}, \ldots, \tilde{Y}_{Ij'} = \tilde{y}_{Ij'})\right] \times \\
&\quad \left[\prod_{(s,t)\in Q} P(Y_{st} = y_{st} \mid \tilde{Y}_{st} = \tilde{y}_{st})\right] \times \left[\prod_{\substack{i'=1 \\ \{i' \neq i,\ j' \neq j\}}}^{I}\prod_{j'=1}^{J} p_{i'j'}^{\tilde{y}_{i'j'}}(1-p_{i'j'})^{1-\tilde{y}_{i'j'}}\right] \times \\
&\quad P(R_i = r_i \mid \tilde{R}_i = 1) \times P(C_j = c_j \mid \tilde{C}_j = 1) \times p_{ij}.
\end{aligned}
$$

We can see from the above equations the contributions that the individual retests have on the probabilities. For large sensitivities and specificities, they contribute values close to 0 or 1.

Second, to find the denominator, note that

$$
\begin{aligned}
&P(\tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q) \\
&= P(\tilde{Y}_{ij} = 0, \tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q) + \\
&\quad P(\tilde{Y}_{ij} = 1, \tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)
\end{aligned}
$$

Using results from $P(\tilde{Y}_{ij} = \tilde{y}_{ij}, \tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)$, we can write the probability for $(i, j) \in Q$ as

$$
\begin{aligned}
&P(\tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q) \\
&= \left[\prod_{\substack{i'=1 \\ i' \neq i}}^{I} P(R_{i'} = r_{i'} \mid \tilde{Y}_{i'1} = \tilde{y}_{i'1}, \ldots, \tilde{Y}_{i'J} = \tilde{y}_{i'J})\right]\left[\prod_{\substack{j'=1 \\ j' \neq j}}^{J} P(C_{j'} = c_{j'} \mid \tilde{Y}_{1j'} = \tilde{y}_{1j'}, \ldots, \tilde{Y}_{Ij'} = \tilde{y}_{Ij'})\right] \times \\
&\quad \left[\prod_{(s,t)\in Q\setminus\{(i,j)\}} P(Y_{st} = y_{st} \mid \tilde{Y}_{st} = \tilde{y}_{st})\right] \times \left[\prod_{\substack{i'=1 \\ \{i' \neq i,\ j' \neq j\}}}^{I}\prod_{j'=1}^{J} p_{i'j'}^{\tilde{y}_{i'j'}}(1-p_{i'j'})^{1-\tilde{y}_{i'j'}}\right] \times \\
&\quad \Big\{P(R_i = r_i \mid \tilde{Y}_{i1} = \tilde{y}_{i1}, \ldots, \tilde{Y}_{ij} = 0, \ldots, \tilde{Y}_{iJ} = \tilde{y}_{iJ}) \times P(C_j = c_j \mid \tilde{Y}_{1j} = \tilde{y}_{1j}, \ldots, \tilde{Y}_{ij} = 0, \ldots, \tilde{Y}_{Ij} = \tilde{y}_{Ij}) \times \\
&\qquad P(Y_{ij} = y_{ij} \mid \tilde{Y}_{ij} = 0)(1-p_{ij}) + P(R_i = r_i \mid \tilde{R}_i = 1) \times P(C_j = c_j \mid \tilde{C}_j = 1) \times P(Y_{ij} = y_{ij} \mid \tilde{Y}_{ij} = 1) \\
&\qquad \times p_{ij}\Big\}
\end{aligned}
$$

and for $(i, j) \notin Q$:

$$
P(\tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \mathrm{Y_Q} = y_Q)
$$

$$= \left[\prod_{\substack{i'=1\\ i'\neq i}}^{I} P(R_{i'} = r_{i'} \mid \tilde{Y}_{i'1} = \tilde{y}_{i'1},\ldots,\tilde{Y}_{i'J} = \tilde{y}_{i'J})\right]\left[\prod_{\substack{j'=1\\ j'\neq j}}^{J} P(C_{j'} = c_{j'} \mid \tilde{Y}_{1j'} = \tilde{y}_{1j'},\ldots,\tilde{Y}_{Ij'} = \tilde{y}_{Ij'})\right]\times$$
$$\prod_{(s,t)\in Q} P(Y_{st} = y_{st} \mid \tilde{Y}_{st} = \tilde{y}_{st}) \times \left[\prod_{\substack{i'=1\\ \{i'\neq i,\ j'\neq j\}}}^{I}\prod^{J}_{j'=1} p_{i'j'}^{\tilde{y}_{i'j'}}(1-p_{i'j'})^{1-\tilde{y}_{i'j'}}\right]\times$$
$$\left\{P(R_i = r_i \mid \tilde{Y}_{i1} = \tilde{y}_{i1},\ldots,\tilde{Y}_{ij} = 0,\ldots,\tilde{Y}_{iJ} = \tilde{y}_{iJ}) \times P(C_j = c_j \mid \tilde{Y}_{1j} = \tilde{y}_{1j},\ldots,\tilde{Y}_{ij} = 0,\ldots,\tilde{Y}_{Ij} = \tilde{y}_{Ij}) \times \right.$$
$$\left. (1-p_{ij}) + P(R_i = r_i \mid \tilde{R}_i = 1) \times P(C_j = c_j \mid \tilde{C}_j = 1) \times p_{ij}\right\}$$

Then for $(i, j) \in Q$:

$$P(\tilde{Y}_{ij} = 1 \mid \tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)$$
$$= \frac{P(R_i = r_i \mid \tilde{R}_i = 1) \times P(C_j = c_j \mid \tilde{C}_j = 1) \times P(Y_{ij} = y_{ij} \mid \tilde{Y}_{ij} = 1) p_{ij}}{\{P(R_i = r_i \mid \tilde{Y}_{i1} = \tilde{y}_{i1},\ldots,\tilde{Y}_{ij} = 0,\ldots,\tilde{Y}_{iJ} = \tilde{y}_{iJ}) \times P(C_j = c_j \mid \tilde{Y}_{1j} = \tilde{y}_{1j},\ldots,\tilde{Y}_{ij} = 0,\ldots,\tilde{Y}_{Ij} = \tilde{y}_{Ij}) \times$$
$$P(Y_{ij} = y_{ij} \mid \tilde{Y}_{ij} = 0)(1-p_{ij}) + P(R_i = r_i \mid \tilde{R}_i = 1) \times P(C_j = c_j \mid \tilde{C}_j = 1) \times P(Y_{ij} = y_{ij} \mid \tilde{Y}_{ij} = 1) p_{ij}\}$$

and for $(i, j) \notin Q$:

$$P(\tilde{Y}_{ij} = 1 \mid \tilde{\boldsymbol{Y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)$$
$$= \frac{P(R_i = r_i \mid \tilde{R}_i = 1) \times P(C_j = c_j \mid \tilde{C}_j = 1) \times p_{ij}}{\left\{P(R_i = r_i \mid \tilde{Y}_{i1} = \tilde{y}_{i1},\ldots,\tilde{Y}_{ij} = 0,\ldots,\tilde{Y}_{iJ} = \tilde{y}_{iJ}) \times P(C_j = c_j \mid \tilde{Y}_{1j} = \tilde{y}_{1j},\ldots,\tilde{Y}_{ij} = 0,\ldots,\tilde{Y}_{Ij} = \tilde{y}_{Ij}) \times (1-p_{ij})\right.}$$
$$\left. + P(R_i = r_i \mid \tilde{R}_i = 1) \times P(C_j = c_j \mid \tilde{C}_j = 1) \times p_{ij}\right\}$$

Denote the resulting $P(\tilde{Y}_{ij} = 1 \mid \tilde{\boldsymbol{Y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)$ by γ_{ij}. Thus, $\tilde{Y}_{ij} \mid \left(\tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i,-j}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q\right) \sim$ Bernoulli(γ_{ij}), where γ_{ij} can not be calculated directly because it depends on unknown individual responses. Note that for the case of no individual retests, we shall use the formula for $(i,j) \notin Q$ for all i and j.

We need to calculate ω_{ij} for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ at each E-step of the EM algorithm. Through using B Gibbs samples, we can estimate it through $\hat{\omega}_{ij} = (B-a)^{-1}\sum_{b=a+1}^{B} \tilde{y}_{ij}^{(b)}$, where $\tilde{y}_{ij}^{(b)}$ is the b^{th} simulated value from $\tilde{Y}_{ij} \mid \left(\tilde{\boldsymbol{Y}}_{-i,-j} = \tilde{\boldsymbol{y}}_{-i-j}^{(b)}, \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q\right) \sim$ Bernoulli($\gamma_{ij}^{(b)}$) and a is a sufficiently long burn-in period. Note that $\tilde{\boldsymbol{y}}_{-i-j}^{(b)}$ results from the most current simulated (or initialized) values of $\tilde{y}_{11},\ldots,\tilde{y}_{IJ}$. We now define the EM algorithm formally here:

1) Initialize $\tilde{y}_{11}^{(0)} = 0, \ldots, \tilde{y}_{IJ}^{(0)} = 0$.
2) E-Step
   a) Find the first sample $\tilde{y}_{11}^{(1)},\ldots,\tilde{y}_{IJ}^{(1)}$ where each is simulated from the corresponding Bernoulli($\gamma_{ij}^{(1)}$) distribution that uses the most updated $\tilde{\boldsymbol{y}}_{-i,-j}$.
   b) Find the second sample $\tilde{y}_{11}^{(2)},\ldots,\tilde{y}_{IJ}^{(2)}$ where each is simulated from the corresponding Bernoulli($\gamma_{ij}^{(2)}$) distributions that uses the most updated $\tilde{\boldsymbol{y}}_{-i,-j}$.

c) Continue this process for $b = 3, \ldots, B$ sets of samples.

d) Calculate $\widehat{\omega}_{ij} = (B-a)^{-1}\sum_{b=a+1}^{B}\tilde{y}_{ij}^{(b)}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

3) M-Step

a) Maximize the expected value of the conditional log-likelihood function

$$E\left[\log(L(\boldsymbol{\beta} \mid \tilde{\boldsymbol{Y}})) \mid \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q\right] = \sum_{i=1}^{I}\sum_{j=1}^{J}\widehat{\omega}_{ij}\log(p_{ij}) + (1-\widehat{\omega}_{ij})\log(1-p_{ij})$$

to find an estimate of $\boldsymbol{\beta}$ where $E(\tilde{Y}_{ij} \mid \boldsymbol{R} = \boldsymbol{r}, \boldsymbol{C} = \boldsymbol{c}, \boldsymbol{Y}_Q = \boldsymbol{y}_Q)$ has been substituted by $\widehat{\omega}_{ij}$.

4) Repeat steps 2 and 3 until convergence is reached when $\left|\left(\hat{\beta}_d^{(r)} - \hat{\beta}_d^{(r-1)}\right) / \hat{\beta}_d^{(r-1)}\right| < \varepsilon$ for all $d = 0, \ldots, p-1$, where $\hat{\boldsymbol{\beta}}^{(r)}$ is the r^{th} estimate of $\boldsymbol{\beta}$ and $\varepsilon > 0$.