Inferring Latent Social Networks from Stock Holdings

Manual for the EM Algorithm

Harrison Hong* Jiangmin Xu†

September, 2017

*Columbia University, NBER, CAFR (e-mail: hh2679@columbia.edu),
†Guanghua School of Management, Peking University (e-mail: jiangminxu@gsm.pku.edu.cn)

1 Quick Overview of the EM Algorithm Program

The whole EM algorithm is run via the Matlab file `mainprogram.m`, which has two main parts.

The first part imports the data, and sets up the initial parameter values, the parameter bounds and the convergence criteria to be used in the main algorithm.

The second part is the execution of the main EM algorithm, which is an iterative loop that:

(I) Starts with the initial parameter value θ_0 as the value for θ'

(II) Calls up the Matlab function file `Qfun.m` to obtain the function $Q(\theta, \theta')$ as specified in Equation (23) of the main paper (which is a function of θ). This function file `Qfun.m`:

(1) Calls up the function file `ellk.m` in a loop of $\{i, k\}$ to obtain the value of $\sum_{y_{i,k}=0}^{+\infty} \log\big(f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)\big) f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$ for all the investor-city $\{i, k\}$ pairs. The function file `ellk.m` calls up two files, `fwyjoint.m` and `fycond.m`:

(i) The function file `fwyjoint.m` obtains $\log\big(f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)\big)$, which is a function of θ, for the given investor-city $\{i, k\}$ pair and value of $y_{i,k}$. The file `fwyjoint.m` will call up two files, `fwcond.m` and `fymarg.m`:

(a) The function file `fwcond.m` obtains $f(\mathbf{w}_{i,k}|y_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)$ as a function of θ for the given investor-city $\{i, k\}$ pair and value of $y_{i,k}$.

(b) The function file `fymarg.m` obtains $\rho(y_{i,k}|\theta)$ as a function of θ for the given investor-city $\{i, k\}$ pair and value of $y_{i,k}$.

then `fwyjoint.m` computes $f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) = f(\mathbf{w}_{i,k}|y_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) \times \rho(y_{i,k}|\theta)$ as a function of θ for the given investor-city $\{i, k\}$ pair and value of $y_{i,k}$.

(ii) The function file `fycond.m` obtains the value of $f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$ for the given investor-city $\{i, k\}$ pair and value of $y_{i,k}$, under the current parameter estimates θ'. It will call up the file `fwyjoint.m` for usage as well.

(2) Obtains

$$Q(\theta, \theta') = \sum_{i=1}^{N} \sum_{k=1}^{K} \left[\sum_{y_{i,k}=0}^{+\infty} \log \left(f(\mathbf{w}_{i,k}, y_{i,k} | \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta) \right) f(y_{i,k} | \mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta') \right]$$

once the loop of $\{i, k\}$ finishes for `ellk.m` (i.e. once the value of $\sum_{y_{i,k}=0}^{+\infty} \log \left(f(\mathbf{w}_{i,k}, y_{i,k} | \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta) \right) f(y_{i,k} | \mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta')$ is obtained for every investor-city $\{i, k\}$ pair.)

(III) Maximizes the function $Q(\theta, \theta')$ obtained from `Qfun.m` over θ, to obtain the new parameter estimate $\hat{\theta}$.

(IV) Checks for convergence of the EM algorithm. If the convergence criterion is reached, the algorithm concludes and outputs the parameter estimate $\hat{\theta}$ from the final iteration as the final estimate $\hat{\theta}_{MLE}$. Otherwise, it sets $\theta' = \hat{\theta}$ and goes back to repeat steps (II) and (III).

2 EM Algorithm, Detailed Explanation

2.1 Initialization of Parameter Values

An initial value for the parameter vector θ' is needed in order to start the EM algorithm. The initialization is as follows:

- Initialization of $\{a_i, b_k, \omega_k\}$: We set the initial value of a_i to be 50 for each i (i.e. our starting point is the simple case of equal gregariousness), set the initial value of b_k to be equal to the number of stocks headquartered in city k divided by the total number of stocks, and set the initial value of ω_k to be 1.2 for each k such that the starting values of ω_k are close to the Poisson null benchmark (i.e. $\omega = 1$).

- Initialization of $\{\alpha, \boldsymbol{\beta}\}$: We estimate a simple OLS model of $w_{i,j,k} = \alpha + \boldsymbol{\beta}\boldsymbol{X}_j + u_j$ and obtain the OLS estimates $\{\hat{\alpha}_{OLS}, \hat{\boldsymbol{\beta}}_{OLS}\}$ as the initial values for $\{\alpha, \boldsymbol{\beta}\}$. This can be carried out quickly using any statistical package such as Stata.

- Initialization of $\{\lambda_0, \lambda_1\}$: We set 0.01 to be the initial value for both λ_0 and λ_1. These initial values are small in order to give our data the full flexibility in determining the estimates for those two parameters.

We use θ_0 to denote this initial value for θ'.

2.2 The E-Step

- As described in Section 2.3.2 of our main paper, the E-step of our EM algorithm boils down to finding the function $Q(\theta, \theta')$ as specified in Equation (23) of the main paper given the observable data $\left\{\mathbf{w}_{i,k} = \{w_{i,j,k}\}_{j \in J_k}, \{\boldsymbol{X}_j\}\right\}$ and the current parameter estimates θ'.

- The function $Q(\theta, \theta')$ is

$$Q(\theta, \theta') = \sum_{i=1}^{N} \sum_{k=1}^{K} \Bigg[\sum_{y_{i,k}=0}^{+\infty} \log \big(f(\mathbf{w}_{i,k}, y_{i,k} | \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta) \big) f(y_{i,k} | \mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta') \Bigg] \quad \text{(M.1)}$$

so that for each i and k, the key thing is to numerically compute the expression

$$\sum_{y_{i,k}=0}^{+\infty} \log \big(f(\mathbf{w}_{i,k}, y_{i,k} | \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta) \big) f(y_{i,k} | \mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta') \quad \text{(M.2)}$$

To do this numerically, we will need to replace the $+\infty$ by a finite positive value $\bar{Y}$ as the upper limit of the summation. We set $\bar{Y} = 50$.[1] Now since this is a summation over the $y_{i,k}$ space from 0 to $\bar{Y}$, it means for each value of $y_{i,k}$ starting from 0, we will compute the value for

$$q(\theta, \theta', y_{i,k}) = \log \big(f(\mathbf{w}_{i,k}, y_{i,k} | \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta) \big) f(y_{i,k} | \mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta') \quad \text{(M.3)}$$

then we will sum up these values of $q(\theta, \theta', y_{i,k})$ to obtain the value for (M.2).

- For each value of $y_{i,k}$, the computation of $q(\theta, \theta', y_{i,k})$ consists of two parts: (A) computing $\log \big(f(\mathbf{w}_{i,k}, y_{i,k} | \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta) \big)$, and (B) computing $f(y_{i,k} | \mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta')$.

  (A) The computation of $\log \big(f(\mathbf{w}_{i,k}, y_{i,k} | \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta) \big)$ consists of the following steps:

  (1) Obtain the density $f(\mathbf{w}_{i,k} | y_{i,k}, \{\boldsymbol{X}_j\}_{j \in \mathcal{J}_k}, \theta) = \prod_{j=1}^{J_k} f(w_{i,j,k} | y_{i,k}, \boldsymbol{X}_j, \theta)$ as a function of parameter vector θ. This step is done in Matlab function file `fwcond.m`, which does two things: (a) for each stock j headquartered in city k, obtain $f(w_{i,j,k} | y_{i,k}, \boldsymbol{X}_j, \theta)$ as given in Equation (15) of the main paper by plugging in

[1] We tested with other values for $\bar{Y}$, including 30, 80, 100. The result changes little for values of $\bar{Y}$ greater than 30. This is because the value of the marginal density for $y_{i,k}$ (the negative binomial density) becomes very small for large values of y under any reasonable parameter values for a_i, b_k, ω_k. We choose $\bar{Y} = 50$ instead of larger values of 80 or 100 for the purpose of speeding up the computation of the algorithm. Choosing larger values for $\bar{Y}$ would lead to little change in the parameter estimate but a significant increase in the computation time.

the values of the data $w_{i,j,k}$ and $\boldsymbol{X}_j$ and the value of $y_{i,k}$; (b) do the multiplication $\prod_{j=1}^{J_k} f(w_{i,j,k}|y_{i,k}, \boldsymbol{X}_j, \theta)$ over all the stocks headquartered in city k.

(2) Obtain the density $\rho(y_{i,k}|\theta) = \frac{\Gamma(y_{i,k}+\zeta_{i,k})}{\Gamma(\zeta_{i,k})\,\Gamma(y_{i,k}+1)} \left(\frac{1}{\omega_k}\right)^{\zeta_{i,k}} \left(\frac{\omega_k - 1}{\omega_k}\right)^{y_{i,k}}$ as a function of parameter vector θ by plugging in the value of $y_{i,k}$. This step is done in Matlab function file `fymarg.m`.

(3) Obtain the joint density $f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) = f(\mathbf{w}_{i,k}|y_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta) \times \rho(y_{i,k}|\theta)$ as a function of parameter vector θ by multiplying $f(\mathbf{w}_{i,k}|y_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)$ and $\rho(y_{i,k}|\theta)$ just obtained from the previous two steps (1) and (2). Then take the natural log to obtain $\log\big(f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)\big)$. This step is done in Matlab function `fwyjoint.m`.

(B) The computation of $f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$ is done in the Matlab function file `fycond.m` and consists of the following steps:

(1) Use the joint density $f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)$ obtained from step (3) of part (A) above, but now plugging in the current parameter estimates θ' for θ to obtain the value of $f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$, i.e. this gives a definitive value, not a function of θ any more.

(2) Use the same density $f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$ from the previous step, but now evaluate this density for each value of $y_{i,k}$ (i.e. we will obtain a value for the density $f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$ for each value of $y_{i,k} = 0, 1, ..., \bar{Y}$), and then sum up these density values to obtain the value of

$$f(\mathbf{w}_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta') = \sum_{y_{i,k}=0}^{\bar{Y}} f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$$

(3) Use the density value $f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$ obtained from the previous step (1), and the density value $f(\mathbf{w}_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$ obtained from the previous step (2), to obtain the value of $f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta') = \dfrac{f(\mathbf{w}_{i,k}\,,\, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')}{f(\mathbf{w}_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')}$.

As a result, for each value of $y_{i,k}$, once we have computed $\log\big(f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)\big)$ from step (A) (which is a function of θ for a given $y_{i,k}$) and $f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$ from step (B) (which is a definitive value for a given $y_{i,k}$), we can obtain

$$q(\theta, \theta', y_{i,k}) = \log\big(f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)\big) f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')$$

as a function of θ.

- We then compute $q(\theta, \theta', y_{i,k})$ (as a function of θ) for every $y_{i,k}$ value by executing steps (A) and (B) repeatedly, and once this is done we do the summation

$$\sum_{y_{i,k}=0}^{\bar{Y}} q(\theta, \theta', y_{i,k}) = \sum_{y_{i,k}=0}^{\bar{Y}} \log\big(f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)\big) f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta') \quad \text{(M.4)}$$

  This whole step is carried out in Matlab function `ellk.m`.

- Finally, we loop over all the investor-city $\{i, k\}$ pairs and execute the above step repeatedly for every investor i in every city k using the data $\big\{\mathbf{w}_{i,k} = \{w_{i,j,k}\}_{j\in J_k}, \{\boldsymbol{X}_j\}\big\}$. This gives us the desired function $Q(\theta, \theta')$:

$$Q(\theta, \theta') = \sum_{i=1}^{N}\sum_{k=1}^{K}\Bigg[\sum_{y_{i,k}=0}^{\bar{Y}} \log\big(f(\mathbf{w}_{i,k}, y_{i,k}|\{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta)\big) f(y_{i,k}|\mathbf{w}_{i,k}, \{\boldsymbol{X}_j\}_{j\in\mathcal{J}_k}, \theta')\Bigg] \quad \text{(M.5)}$$

  which is a deterministic function of θ that can be maximized. This is the final part of the E-step and is carried out in Matlab Function `Qfun.m`.

2.3 The M-Step

We maximize the function $Q(\theta, \theta')$ computed from the E-step over the parameter vector θ, and obtain the new estimate $\hat{\theta}$.

1. To ensure the stability and the speed of the maximization process, we set up lower and

upper bounds for the parameters. The lower bounds for $\{a_i, b_k, \omega_k, \alpha, \boldsymbol{\beta}, \lambda_0, \lambda_1\}$ are $\underline{L} = \{0.01$ for all a_i, 0.001 for all b_k, 1.0 for all ω_k, -10 for α and $\boldsymbol{\beta}$, 0.0001 for λ_0 and $\lambda_1\}$; The upper bounds $\overline{H} = \{1000$ for all a_i, 0.9 for all b_k, 10 for all ω_k, 10 for α and $\boldsymbol{\beta}$, 10 for λ_0 and $\lambda_1\}$.

2. Use the Matlab build-in routine `fmincon` with the associated interior-point algorithm to carry out the maximization.[2] Since `fmincon` is a minimization routine in Matlab, we negate the function $Q(\theta, \theta')$ and minimize $-Q(\theta, \theta')$ over θ using `fmincon`, which is equivalent to maximizing $Q(\theta, \theta')$ over θ. In the `fmincon` routine, we also input $\underline{L}$ and $\overline{H}$ as the lower bound and the upper bound for the parameter vector θ, and input $\sum_k b_k = 1$ as an equality constraint (recall this is the normalization condition).

3. Once the execution of `fmincon` on $-Q(\theta, \theta')$ finishes (converges), it gives the new estimate $\hat{\theta}$, i.e. $\hat{\theta} = \arg\max_{\theta} Q(\theta, \theta')$. The convergence criterion set for `fmincon` is that the first-order optimality measure is less than 10^{-4}.[3]

2.4 Execution of the Whole EM algorithm

The EM algorithm starts with the initial parameter values θ_0 as θ', and repeats the E-step and M-step recursively until the algorithm converges. The whole EM algorithm is executed as follows:

- **Iteration 1**: Let $\theta' = \theta_0$ (the initial parameter values), use it to carry out the E-step to obtain the function $Q(\theta, \theta_0)$, and then carry out the M-step on $Q(\theta, \theta_0)$ to obtain the new parameter estimate $\hat{\theta}_1$.

- **Iteration s, $s \geq 2$**: Let $\theta' = \hat{\theta}_{s-1}$ (i.e. the parameter estimate obtained from the previous iteration step $s-1$), use it to carry out the E-step to obtain the function $Q(\theta, \hat{\theta}_{s-1})$, and then carry out the M-step on $Q(\theta, \hat{\theta}_{s-1})$ to obtain the new parameter estimate $\hat{\theta}_s$.

[2] See `https://www.mathworks.com/help/optim/ug/fmincon.html`.
[3] See `https://www.mathworks.com/help/optim/ug/first-order-optimality-measure.html`.

- **Convergence of the EM algorithm**: At the end of any iteration step s, evaluate the difference $D_s = |Q(\hat{\theta}_s, \hat{\theta}_s) - Q(\hat{\theta}_{s-1}, \hat{\theta}_{s-1})|$.[4] If $D_s \leq \epsilon$, we conclude the iterative process and $\hat{\theta}_s$ is our final parameter estimate. Otherwise go to iteration $s+1$ and repeat the process.

  The ϵ is the convergence criterion parameter that is chosen to be arbitrarily small. In practice we set $\epsilon = 10^{-3}$. We tried $\epsilon = 10^{-4}$ as well; it does not change the final parameter estimate in any meaningful way (i.e. changes are in the order of 10^{-4}) but it increases the amount of computation time significantly, hence we adopt for $\epsilon = 10^{-3}$ as the convergence criterion.

The whole EM algorithm is run via the Matlab file `mainprogram.m`.

[4]The meaning of $Q(\hat{\theta}_s, \hat{\theta}_s)$ is to get the value of the function $Q(\theta, \theta')$ by plugging in the value of the estimate $\hat{\theta}_s$ for both θ' and θ. Similarly for the meaning of $Q(\hat{\theta}_{s-1}, \hat{\theta}_{s-1})$.