

Paired factor analysis for tree reconstruction

Kushal K Dey, Gao Wang

July 19, 2016

Let D_{nj} be the data corresponding to n -th sample and j -th gene. We assume for now that the data is Gaussian in its distribution. We assume there are K factors or nodes of the tree. We assume the model

$$E[D_{nj}|Z_n = (k_1, k_2), \lambda_n = q, F] = qF_{k_1,j} + (1 - q)F_{k_2,j}$$

We assume a prior on λ ,

$$Pr[\lambda_n = q] = \pi_q$$

Then we can write

$$Pr[D_n|Z_n = (k_1, k_2), F, s_{j=1,2,\dots,J}^2] = \sum_q \pi_q Pr[D_n|Z_n = (k_1, k_2), \lambda_n = q, F, s_{j=1,2,\dots,J}^2]$$

where s_j^2 is the variance of the j th feature.

We also assume the prior

$$Pr[Z_n = (k_1, k_2)] = \pi_{k_1, k_2} \quad k_1 < k_2$$

Then we can write

$$Pr[D_n|\pi, F] = \sum_{k_1 < k_2} \pi_{k_1, k_2} Pr[D_n|Z_n = (k_1, k_2), F, s_{j=1,2,\dots,J}^2]$$

We define the joint prior over the edges and the fraction of the edge represented as

$$\pi_{k_1, k_2, q} = \pi_{k_1, k_2} \pi_q \quad k_1 < k_2$$

The overall likelihood

$$L(\pi, F) = \prod_{n=1}^N Pr[D_n|\pi, F, s_{j=1,2,\dots,J}^2]$$

or we can write it as

$$L(\pi, F) = \prod_{n=1}^N \sum_{k_1 < k_2} \sum_q \left[\pi_{k_1, k_2, q} \times \prod_{j=1}^G N(D_{nj}; qF_{k_1, j} + (1 - q)F_{k_2, j}, s_j^2) \right]$$

$$\log L(\pi, F) = \sum_{n=1}^N \log \left(\sum_{k_1 < k_2} \sum_q \left[\pi_{k_1, k_2, q} \times \prod_{j=1}^G N(D_{nj}; qF_{k_1, j} + (1 - q)F_{k_2, j}, s_j^2) \right] \right)$$

This is the log likelihood we want to maximize and we need to return this log-likelihood.

We assume that q can take a finite set of values between 0 and 1, say $1/100, 2/100, \dots, 90/100, 1$.

Suppose we have run upto m iterations. For the $(m+1)$ th iteration, we have

$$\delta_{n,k_1,k_2,q}^{(m+1)} = Pr \left[Z_n = (k_1, k_2), \lambda_n = q | \pi^{(m)}, F^{(m)}, s_{j=1,2,\dots,J}^{(m)}, D_n \right]$$

$$\delta_{n,k_1,k_2,q}^{(m+1)} \propto Pr [Z_n = (k_1, k_2)] Pr [\lambda_n = q] Pr \left[D_n | \pi^{(m)}, F^{(m)}, s_{j=1,2,\dots,J}^{(m)}, Z_n = (k_1, k_2), \lambda_n = q \right]$$

$$\delta_{n,k_1,k_2,q}^{(m+1)} \propto \pi_{k_1,k_2,q}^{(m)} \prod_j N \left(D_{nj} | qF_{k_1,j}^{(m)} + (1-q)F_{k_2,j}^{(m)}, s_j^{(m)2} \right)$$

where $s_j^{(m)2}$ is the variance of the gene j .

We normalize δ so that

$$\sum_{k_1 < k_2} \sum_q \delta_{n,k_1,k_2,q}^{(m+1)} = 1 \quad \forall n$$

We define

$$\pi_{k_1,k_2,q}^{(m+1)} = \frac{1}{N} \sum_{n=1}^N \delta_{n,k_1,k_2,q}^{(m+1)}$$

We have therefore updated $\pi_{k_1,k_2,q}^{(m)}$ to $\pi_{k_1,k_2,q}^{(m+1)}$.

We define the parameter

$$\theta := (\pi_{k_1,k_2,q}, F, s_{j=1,2,\dots,J})$$

We define the complete loglikelihood

$$\log L_c(\theta; D, Z, \lambda) = \log \pi_{k_1,k_2,q} + \log L(D|Z, \lambda, q, F)$$

We take the expectation of this quantity with respect to $[Z, \lambda | D, \theta^{(m)}]$.

$$Q(\theta | \theta^{(m)}) \propto - \sum_{n=1}^N \sum_{k_1 < k_2} \sum_q \delta_{n,k_1,k_2,q}^{(m+1)} \sum_j \left[\log s_j + \frac{(D_{nj} - qF_{k_1,j} - (1-q)F_{k_2,j})^2}{2s_j^2} \right]$$

We try to maximize this quantity with respect to F , So, we can take derivative with respect to F and try to solve the resulting normal equation.

This equation, conditional on $[Z, \lambda | D, \theta^{(m)}]$, can be written as

$$D_{N \times J} = L_{N \times K} F_{K \times J} + E_{N \times J}$$

where

$$e_{nj} \sim N(0, s_j^2)$$

We define

$$D'_{nj} := \frac{D_{nj}}{s_j}$$

If we consider finding the factors on a gene by gene basis, we do not need to worry about s_j .

$$L_{nk} = \begin{cases} q \text{ or } (1-q) & \lambda_n = q \\ 0 & \text{o.w.} \end{cases}$$

We have

$$\begin{aligned} E_{Z,\lambda|D,\theta^{(m)}} [L_{nk}] &= \sum_q \sum_{k_2 > k} q \delta_{n,k,k_2,q}^{(m+1)} + \sum_q \sum_{k_1 < k} (1-q) \delta_{n,k_1,k,q}^{(m+1)} \\ E_{Z,\lambda|D,\theta^{(m)}} [L_{nk}^2] &= \sum_q \sum_{k_2 > k} q^2 \delta_{n,k,k_2,q}^{(m+1)} + \sum_q \sum_{k_1 < k} (1-q)^2 \delta_{n,k_1,k,q}^{(m+1)} \end{aligned}$$

Also for any $k \neq l$,

$$E_{Z,\lambda|D,\theta^{(m)}} [L_{nk} L_{nl}] = \sum_q q(1-q) \delta_{n,k,l,q}^{(m+1)}$$

We use these to solve for the equation

$$[E_{Z,\lambda|D,\theta^{(m)}} (L^T L)] F \approx [E_{Z,\lambda|D,\theta^{(m)}} (L)]^T D$$

The solution therefore is

$$F \approx [E_{Z,\lambda|D,\theta^{(m)}} (L^T L)]^{-1} [E_{Z,\lambda|D,\theta^{(m)}} (L)]^T D$$

For $W = L^T L$

$$W_{kl} = \sum_n L_{kn} L_{nl}$$

$$E_{Z,\lambda|D,\theta^{(m)}} (W_{kl}) = \sum_n E_{Z,\lambda|D,\theta^{(m)}} (L_{kn} L_{nl})$$

We use the definition of $E_{Z,\lambda|D,\theta^{(m)}} [L_{nk}^2]$ and $E_{Z,\lambda|D,\theta^{(m)}} [L_{nk} L_{nl}]$ from above to solve this linear system.

In the same way as we computed F by solving for the normal equation obtained from taking derivative of the function $Q(\theta|\theta^{(m)})$, we take derivative of the latter with respect to s_j^2 to obtain EM updates of the genes variance terms. O taking derivative, we obtain the estimates as

$$\hat{s}_j^2 = \sum_{n=1}^N \sum_{k_1 < k_2} \sum_q \delta_{n,k_1,k_2,q}^{(m+1)} (D_{nj} - qF_{k_1,j} - (1-q)F_{k_2,j})^2$$