Algorithm for dr-tree reconstruction

Kushal K Dey, Gao Wang

July 19, 2016

Let D_{nj} be the data corresponding to nth sample and jth 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

$$E\left[D_{nj}|Z_n = (k_1, k_2), \lambda_n = q, F\right] = qF_{k1,j} + (1-q)F_{k2,j}$$

We assume a prior on λ,

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

Then we can write

$$Pr\left[D_n|Z_n = (k_1, k_2), F\right] = \sum_q \pi_q Pr\left[D_n|Z_n = (k_1, k_2), \lambda_n = q, F\right]$$

We also assume the prior

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

Then we can write

$$Pr\left[D_n|\pi, F\right] = \sum_{k_1<k_2} \pi_{k_1,k_2} Pr\left[D_n|Z_n = (k_1, k_2), F\right]$$

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

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

The overall likelihood

$$L(\pi, F) = \prod_{n=1}^{N} Pr\left[D_n|\pi, F\right]$$

or we can write it as

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

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

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 $1/100, 2/100, \cdots, 90/100, 1$.

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

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

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

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

where s_j^2 is the estimated variance of the gene j.

We normalize δ so that

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

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

$$\theta := (\pi_{k_1,k_2,q}, F)$$

We define the complete loglikelihood

$$logL_c\left(\theta; D, Z, \lambda\right) = log\pi_{k_1,k_2,q} + logL(D|Z, \lambda, q, F)$$

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

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

We try to maximize this quantity with respect to F,

Conditional on $\left[Z, \lambda|D, \theta^{(m)}\right]$, we can write

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

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.

$$\begin{aligned} L_{nk} &= q \;\; or \;\; (1-q) \qquad\quad \lambda_n = q \\ &= 0 o.w \end{aligned}$$

We have

$$E_{Z,\lambda|D,\theta^{(m)}}\left[L_{nk}\right] = \sum_q \sum_{k_2>k} q\delta^{(m+1)}_{n,k,k_2,q} + \sum_q \sum_{k_1<k} (1-q)\delta^{(m+1)}_{n,k1,k,q}$$

$$E_{Z,\lambda|D,\theta^{(m)}}\left[L^2_{nk}\right] = \sum_q \sum_{k_2>k} q^2\delta^{(m+1)}_{n,k,k_2,q} + \sum_q \sum_{k_1<k} (1-q)^2\delta^{(m+1)}_{n,k1,k,q}$$

Also for any $k \neq l$,

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

We use these to solve for the equation

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

The solution therefore is

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

For $W = L^T L$

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

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

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