1 The EM Algorithm

1.1 Purpose

The purpose of this document is to express a way of selecting a set of covariance matrices for fitting a mixture of multivariate normals. Previously, we have used a fixed grid of U_k to represent the prior covariance matrices on the vector of 'true' effects across tissues, $\boldsymbol{b}$. We then estimate the weights on each of these matrices π hierarchically, using the EM algorithm. Here, we propose to estimates these covariance matrices simultaneously, thus reflecting the ideal patterns of covariance present in the data.

1.2 Defining the Model

For a given gene-snp pair, $\boldsymbol{b}$ represents the R vector of unknown standardized effect. We model the prior distribution from which $\boldsymbol{b}$ is drawn as a mixture of multivariate *Normals*.

$$\boldsymbol{b}|\boldsymbol{\pi}, \mathbf{U} \sim \sum_{\mathbf{k,l}} \pi_{\mathbf{k,l}} \, N_{\mathbf{R}}(\mathbf{0}, \omega_{\mathbf{l}} \mathbf{U_k}) \tag{1}$$

- Choice of U_k determines the direction, while ω_l determines the 'stretch' or 'tails' of each distribution
- $\pi_{k,l}$ to represent the (unknown) prior weight on prior covariance matrix $U_{k,l}$
- Use the EM algorithm to estimate the optimal combination of weights: How often does this particular pattern of sharing occur in the data?
- Now, we will also the EM algorithm to estimate these prior covariance matrices, thus simultaneously inferring both the patterns and the relative frequencies that maximize the likelihood of the data set

Furthermore, for a given gene-snp pair, the Likelihood on $\boldsymbol{b}$:

$$\hat{\boldsymbol{b}}|\boldsymbol{b} \sim N_R(\boldsymbol{b}, \hat{V}) \tag{2}$$

We know that for a single multivariate *Normal* the posterior on $\boldsymbol{b}|U_0$ is simply:

$$\boldsymbol{b}|\hat{\boldsymbol{b}} \sim N_R(\boldsymbol{\mu}_1, U_1)$$

where:

- $\boldsymbol{\mu}_1 = U_1(\hat{V}^{-1}\hat{\boldsymbol{b}})$;
- $U_1 = (U_0^{-1} + \hat{V}^{-1})^{-1}$.

If we added the subscript k, then for each component, we have a component specific posterior covariance matrix U_{1k} and a component specific posterior mean $\boldsymbol{\mu_{1k}}$, as in the JxKxR

`all.arrays$post.means`

object, where the [j,k,] element represents the posterior mean for the jth snp across all R tissues.

2 Algorithm

2.1 E-Step

For a data set with J gene snp pairs and K components:

- U_k to represent the 'true' covariance matrix of effects,
- B_{jk} to represent the RxR posterior conditional covariance matrix for each gene-snp pair at each component (U_1 above)
- $\boldsymbol{b_{jk}}$ to represent the R-dimensional posterior mean for each gene-snp pair at each component (analogous

In the E- step, using the same notation as the authors Bovy et al where q_{jk} represents the latent 'label' of each gene-snp pair according to its membership:

$$\begin{aligned} q_{jk} &= \frac{\pi_k N(\hat{\boldsymbol{b}}|0, U_k + V_j)}{\sum_k \pi_k N(\hat{\boldsymbol{b_j}}|0, U_k + V_j)} \\ \boldsymbol{b_{jk}} &= B_{jk}(U_k^{-1}\boldsymbol{m_k} + \hat{V}^{-1}\hat{\boldsymbol{b}}) \\ B_{jk} &= (U_k^{-1} + \hat{V_j^{-1}})^{-1} \end{aligned} \tag{3}$$

Quite simply, the latent indicator label is simply the likelihood at a particular component times the current update of the prior weight π_k at that component, divided by the marginal probability of observing that gene snp pair.

The current component specific posterior covariance B_{jk} is the posterior covariance matrix of a single multivariate normal distribution and the current component-specific posterior mean $\boldsymbol{b_{jk}}$ is the posterior mean of a single multivariate normal.

`(see https://en.wikipedia.org/wiki/Conjugate_prior)`

Note that this is slightly different than our expression for the posterior conditional mean $\boldsymbol{\mu_{1k}}$ above because in the previous model, we assumed that each $\boldsymbol{b_j} \sim N(0, U_k)$ (i.e., the prior mean was $\mathbf{0}$) where here we estimate the underlying mean for each component, $\boldsymbol{m_k}$ using the EM algorithm and so we need to use the full formula for a multivariate normal with known residual matrix V_j [in Wiki notation, Σ]: $\left(\left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\hat{\mathbf{b}}\right), \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right.$, where $\hat{\mathbf{b}}$ is the vector of MLEs for a given gene-snp pair).

2.2 M-Step

Now, let $q_k = \sum_j q_{j,k}$. We can write the following expectation step down.

$$\begin{aligned} \pi_k &= \frac{1}{J}\sum_j q_{jk} \\ \boldsymbol{m_k} &= \frac{1}{q_k}\sum_j \boldsymbol{b_{jk}} \\ U_k &= \frac{1}{q_k}\sum_j q_{jk}[(\boldsymbol{m_k} - \boldsymbol{b_{jk}})(\boldsymbol{m_j} - \boldsymbol{b_{jk}})^T + B_{jk}] \end{aligned} \tag{4}$$