EM Algorithm Derivation

Sarah Urbut

June 25, 2015

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 $\boldsymbol{\mu_1}$) above.
- π_k to represent the mixture proportions.

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:

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

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. This is equivalent the posterior probability that a data point j arose from component K. Note that the distribution $\hat{\boldsymbol{b_j}}|0, U_k + V_j)$ results from integrating over the uncertainty in $\boldsymbol{b}_j$ and thus represents the variance of the marginal distribution of $\hat{b}$, the T used in *Bovyetal*.

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 Wikipedia)

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 (please see section: Posteriors on genotype effect sizes for algebraic derivation).

2.2 M-Step

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

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

3 Derivation of Conditional Posterior

3.1 Posteriors on genotype effect sizes

By maximum likelihood in each tissue separately, we can easily obtain the estimates of the standardized genotype effect sizes, $\hat{\boldsymbol{b}}_j$, and their standard errors recorded on the diagonal of an $R \times R$ matrix noted $\hat{V}_j = \mathbb{V}(\hat{\boldsymbol{b}}_j)$. Using each pair of tissues, we can also fill the off-diagonal elements of $\hat{V}_j$.

If we now view $\hat{\boldsymbol{b}}_j$ and $\hat{V}_j$ as *observations* (i.e. known), we can write a new "likelihood" (using only the sufficient statistics):

$$\hat{\boldsymbol{b}}_j | \boldsymbol{b}_j \sim \mathcal{N}_R(\boldsymbol{b}_j, \hat{V}_j) \tag{5}$$

Let us imagine first that the prior on $\boldsymbol{b}_j$ is not a mixture but a single Normal:

$\boldsymbol{b}_j \sim \mathcal{N}_R(\boldsymbol{0}, U_0)$. As this prior is conjuguate to the "likelihood" above, the posterior simply is (see Wikipedia):

$$\boldsymbol{b}_j | \hat{\boldsymbol{b}}_j \sim \mathcal{N}_R(\boldsymbol{\mu}_{j1}, U_{j1})$$

where:

- $\boldsymbol{\mu}_{j1} = U_{j1}(\hat{V}_j^{-1}\hat{\boldsymbol{b}}_j)$;
- $U_j = (U_0^{-1} + \hat{V}_j^{-1})^{-1}$.

In practice however, we use a mixture, as defined at the beginning of the document.

Inside the sums, the posterior of the effect size for component k can be written as:

$$\begin{aligned} p(\boldsymbol{b}_j | \hat{\boldsymbol{b}}_j, \hat{V}_j, K) &= p(\boldsymbol{b}_j | \hat{\boldsymbol{b}}_j, \hat{V}_j, U_{0k}) \\ &\propto \mathcal{L}(\boldsymbol{b}_j) p(\boldsymbol{b}_j | U_{0k}) \\ &\propto \exp[(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j)^T \hat{V}_j^{-1} (\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j)] \ \exp(\boldsymbol{b}_j^T U_{0k}^{-1} \boldsymbol{b}_j) \\ &\propto \exp[\boldsymbol{b}_j^T (\hat{V}_j^{-1} + U_{0k}^{-1}) \boldsymbol{b}_j - \hat{\boldsymbol{b}}_j^T \hat{V}_j^{-1} \boldsymbol{b}_j - \boldsymbol{b}_j^T \hat{V}_j^{-1} \hat{\boldsymbol{b}}_j] \end{aligned} \tag{6}$$

Note that this also results from the fact that we assume that $\boldsymbol{b}_j \sim \mathcal{N}_R(\boldsymbol{0}, U_0)$ and thus that $\boldsymbol{\mu}_0$ is 0. Thus in the EM algorithm where we will keep updated the 'true μ_0 we can no longer consider this and so instead, we have:

$$
\begin{aligned}
p(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, K) &= p(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, U_{0k}) \\
&\propto \mathcal{L}(\boldsymbol{b}_j) p(\boldsymbol{b}_j|U_{0k}) \\
&\propto \exp[(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j)^T \hat{V}_j^{-1} (\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j)] \ \exp(\boldsymbol{b}_j - \boldsymbol{\mu}_0)^T U_{0k}^{-1} (\boldsymbol{b}_j - \boldsymbol{\mu}_0) \\
&\propto \exp[\boldsymbol{b}_j^T (\hat{V}_j^{-1} + U_{0k}^{-1}) \boldsymbol{b}_j - (\hat{\boldsymbol{b}}_j^T \hat{V}_j^{-1} + U_{0k}^{-1} \boldsymbol{\mu}_0) \boldsymbol{b}_j - \boldsymbol{b}_j^T (\hat{V}_j^{-1} \hat{\boldsymbol{b}}_j + U_{0k}^1 \boldsymbol{\mu}_0)]
\end{aligned} \tag{7}
$$

Defining $\Omega = (\hat{V}_j^{-1} + U_{0k}^{-1})^{-1}$ and noting that it is symmetric, we can use the property $\Omega^{-1}\Omega^T = I$ to factorize everything (and "complete the square"):

$$
p(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, U_{0k}) \propto \exp[(\boldsymbol{b}_j - \Omega(\hat{V}_j^{-1}\hat{\boldsymbol{b}}_j + U_{0k}^{-1}\boldsymbol{\mu}_0))^T \Omega^{-1} (\boldsymbol{b}_j - \Omega(\hat{V}_j^{-1}\hat{\boldsymbol{b}}_j + U_{0k}^{-1}\boldsymbol{\mu}_0))] \tag{8}
$$

For some configurations, U_{0k} may not be positive-definite, and thus not invertible. We therefore need to avoid writing Ω as a function of U^{-1} in favor of U:

$$
\begin{aligned}
\Omega &= (V^{-1} + U^{-1})^{-1} \\
&= ((V^{-1} + U^{-1})UU^{-1})^{-1} \\
&= ((V^{-1}U + I)U^{-1})^{-1} \\
&= U(V^{-1}U + I)^{-1}
\end{aligned} \tag{9}
$$

Recognizing the kernel of a Normal distribution, we get:

$$
\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, K \ \sim \ \mathcal{N}_R(\boldsymbol{\mu}_{j1k}, U_{j1k}) \tag{10}
$$

where

$$
\begin{aligned}
U_{j1k} &= \Omega \\
&= U_{0k} \left(\hat{V}_j^{-1} U_{0k} + I\right)^{-1}
\end{aligned} \tag{11}
$$

and

$$
\begin{aligned}
\boldsymbol{\mu}_{j1k} &= \Omega \hat{V}_j^{-1} \hat{\boldsymbol{b}}_j \\
&= U_{j1k} \hat{V}_j^{-1} \hat{\boldsymbol{b}}_j
\end{aligned} \tag{12}
$$

and in the case when we do not assume $\boldsymbol{\mu}_0 = 0$, we have

$$
\begin{aligned}
\boldsymbol{\mu}_{j1k} &= \Omega(\hat{V}_j^{-1}\hat{\boldsymbol{b}}_j + U_{0k}^{-1}\boldsymbol{\mu}_0)) \\
&= U_{j1k}(\hat{V}_j^{-1}\hat{\boldsymbol{b}}_j + U_{0k}^{-1}\boldsymbol{\mu}_0))
\end{aligned} \tag{13}
$$

To compute the posterior weights, we will exploit the fact that the marginal likelihood corresponds to a Normal density when the conditional likelihood is Normal with known

variance and the prior of its mean is also Normal (see Berger, 1985, example 1 in section 4.2). This means that we only need the mean and covariance matrix of this Normal density.

With a small abuse of notation, let us consider below that $\hat{\boldsymbol{b}}_j$ is random and, using the law of total expectation with $\boldsymbol{b}_j$ as well as 5, we obtain:

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{b}}_j|\hat{V}_j, K] &= \mathbb{E}_{\boldsymbol{b}_j}[\mathbb{E}[\hat{\boldsymbol{b}}_j|\hat{V}_j, K, \boldsymbol{b}_j]] \\ &= \mathbb{E}_{\boldsymbol{b}_j}[\boldsymbol{b}_j|K] \\ &= \mathbf{0}\end{aligned} \tag{14}$$

Now using the law of total variance with $\boldsymbol{b}_j$ as well as 5 and **??**, we obtain:

$$\begin{aligned}\mathbb{V}[\hat{\boldsymbol{b}}_j|\hat{V}_j, K] &= \mathbb{E}_{\boldsymbol{b}_j}[\mathbb{V}[\hat{\boldsymbol{b}}_j|\hat{V}_j, K, \boldsymbol{b}_j]] \\ &\quad + \mathbb{V}_{\boldsymbol{b}_j}[\mathbb{E}[\hat{\boldsymbol{b}}_j|\hat{V}_j, K, \boldsymbol{b}_j]] \\ &= \mathbb{E}_{\boldsymbol{b}_j}[\hat{V}_j] + \mathbb{V}_{\boldsymbol{b}_j}[\boldsymbol{b}_j|k] \\ &= \hat{V}_j + U_{0k}\end{aligned} \tag{15}$$

Therefore the posterior weight is:

$$\tilde{w}_{jl} = \frac{\hat{\pi_k}\, \mathcal{N}_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, U_{0k} + \hat{V}_j)}{\sum_k \hat{\pi}_k\, \mathcal{N}_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, U_{0k} + \hat{V}_j)} \tag{16}$$

The notation $\mathcal{N}_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, U_{0k} + \hat{V}_j)$ means that we calculate the multivariate Normal density with mean $\mathbf{0}$ and covariance matrix $U_{0k} + \hat{V}_j$ at the point $\hat{\boldsymbol{b}}_j$.

It is quite straightforward from this document that 13 is equivalent to $\boldsymbol{b}_{jk}$ (the conditional component-specific posterior mean) and that 11 is equivalent to B_{jk} ((the conditional component-specific posterior covariance matrix). Furthermore, μ_{0k} is equivalent to the $\boldsymbol{m_k}$ (representing the component-specific "true" mean) and U_{0k} is U_k the same (the component specific 'true covariance of the underlying effect).

4 Translation Algebra

For the authors:

$$\begin{aligned}\boldsymbol{b}_{jk} &= U_k[U_k + \hat{V}_j]^{-1}\hat{\boldsymbol{b}}_j \\ B_{jk} &= U_k - U_k[U_k + \hat{V}_j]^{-1}U_k\end{aligned} \tag{17}$$

For us:

$$\begin{aligned} \boldsymbol{b}_{jk} &= U_k \left(\hat{V}_j^{-1} U_k + I \right)^{-1} \hat{V}_j^{-1} \hat{\boldsymbol{b}}_j \\ B_{jk} &= U_k \left(\hat{V}_j^{-1} U_k + I \right)^{-1} \end{aligned} \tag{18}$$