EM Algorithm Derivation

Sarah Urbut

June 24, 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

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 Maximization step down.

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

3 Derivation of Conditional Posterior

3.1 Posteriors on genotype effect sizes

Also of interest are the posterior probabilities of the genotype effect sizes per gene-SNP pair in each tissue. This section was written with Sarah Urbut.

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

If we now view $\hat{\boldsymbol{b}}_{gp}$ and $\hat{V}_{gp}$ as *observations* (i.e. known), we can "forget" about the original data X_p, X_c and Y_g, and write a new "likelihood" (using only the sufficient statistics):

$$\hat{\boldsymbol{b}}_{gp}|\boldsymbol{b}_{gp} \sim_R (\boldsymbol{b}_{gp}, \hat{V}_{gp}) \tag{5}$$

Let us imagine first that the prior on $\boldsymbol{b}_{gp}$ is not a mixture but a single Normal: $\boldsymbol{b}_{gp} \sim_R (\mathbf{0}, U_0)$. As this prior is conjuguate to the "likelihood" above, the posterior simply is (see Wikipedia):

$$\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp} \sim_R (\boldsymbol{\mu}_{gp1}, U_{gp1})$$

where:

- $\boldsymbol{\mu}_{gp1} = U_{gp1}(\hat{V}_{gp}^{-1}\hat{\boldsymbol{b}}_{gp})$;
- $U_{gp1} = (U_0^{-1} + \hat{V}_{gp}^{-1})^{-1}$.

In practice however, we use a mixture (see **??**):

$$\boldsymbol{b}_{gp}|\boldsymbol{\eta}, \boldsymbol{\lambda}, _0 \sim \sum_{j=1}^{J}\sum_{l=1}^{L} \eta_j \lambda_l \ _R(\mathbf{0}, U_{0jl}) \tag{6}$$

for which the hyper-parameters are either fixed ($_0$) or estimated ($\boldsymbol{\eta}, \boldsymbol{\lambda}$) as described above using the full hierarchical model and the EM algorithm. We hence write $\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}$ for the "empirical Bayes" estimates of the priors $\boldsymbol{\eta}, \boldsymbol{\lambda}$.

A posterior we may well be interested in is $\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\Theta}_{-\pi_0}, v_{gp} = 1$, that is conditional on being the eQTN in at least one tissue, averaging over all configurations and the whole grid:

$$p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp} = 1) = \sum_j \sum_l (c_{gpj} = 1, d_{gpl} = 1|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp} = 1)\ p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, c_{gpj} = 1, d_{gpl} = 1 \tag{7}$$

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

$$\begin{aligned} p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, c_{gpj} = 1, d_{gpl} = 1) &= p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, U_{0jl}) \\ &\propto (\boldsymbol{b}_{gp}) p(\boldsymbol{b}_{gp}|U_{0jl}) \\ &\propto \exp[(\hat{\boldsymbol{b}}_{gp} - \boldsymbol{b}_{gp})^T \hat{V}_{gp}^{-1}(\hat{\boldsymbol{b}}_{gp} - \boldsymbol{b}_{gp})]\ \exp(\boldsymbol{b}_{gp}^T U_{0jl}^{-1}\boldsymbol{b}_{gp}) \\ &\propto \exp[\boldsymbol{b}_{gp}^T(\hat{V}_{gp}^{-1} + U_{0jl}^{-1})\boldsymbol{b}_{gp} - \hat{\boldsymbol{b}}_{gp}^T\hat{V}_{gp}^{-1}\boldsymbol{b}_{gp} - \boldsymbol{b}_{gp}^T\hat{V}_{gp}^{-1}\hat{\boldsymbol{b}}_{gp}] \end{aligned} \tag{8}$$

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

$$p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, U_{0jl}) \propto \exp[(\boldsymbol{b}_{gp} - \Omega\hat{V}_{gp}^{-1}\hat{\boldsymbol{b}}_{gp})^T\Omega^{-1}(\boldsymbol{b}_{gp} - \Omega\hat{V}_{gp}^{-1}\hat{\boldsymbol{b}}_{gp})] \tag{9}$$

For some configurations, U_{0jl} 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:

$$\begin{aligned}\Omega &= (V^{-1} + U^{-1})^{-1} \\ &= ((V^{-1} + U^{-1})UU^{-1})^{-1} \\ &= ((V^{-1}U + I)U^{-1})^{-1} \\ &= U(V^{-1}U + I)^{-1}\end{aligned} \tag{10}$$

Recognizing the kernel of a Normal distribution, we get:

$$\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, c_{gpj} = 1, d_{gpl} = 1 \ \sim \ _R(\boldsymbol{\mu}_{gp1jl}, U_{gp1jl}) \tag{11}$$

where

$$\begin{aligned}U_{gp1jl} &= \Omega \\ &= U_{0jl}\left(\hat{V}_{gp}^{-1}U_{0jl} + I\right)^{-1}\end{aligned} \tag{12}$$

and

$$\begin{aligned}\boldsymbol{\mu}_{gp1jl} &= \Omega\hat{V}_{gp}^{-1}\hat{\boldsymbol{b}}_{gp} \\ &= U_{gp1jl}\hat{V}_{gp}^{-1}\hat{\boldsymbol{b}}_{gp}\end{aligned} \tag{13}$$

Furthermore, also inside the sums of **??**, the posterior mixture weight for component j, l, here noted $\tilde{w}_{jl}$, can be written as:

$$\begin{aligned}\tilde{w}_{jl} &= (c_{gpj} = 1, d_{gpl} = 1|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp} = 1) \\ &= \frac{\hat{\eta}_j\hat{\lambda}_l \ p(\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, c_{gpj} = 1, d_{gpl} = 1)}{p(\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp} = 1)}\end{aligned} \tag{14}$$

We recognize the marginal likelihood, and thus the Bayes factor **??**, leading to:

$$\tilde{w}_{jl} = \frac{\hat{\eta}_j\hat{\lambda}_{l \ gpjl}}{\sum_{j'}\sum_{l'}\hat{\eta}_{j'}\hat{\lambda}_{l' \ gpj'l'}} \tag{15}$$

However, here, thinking about the implementation (i.e. to avoid reloading all the BFs when computing the posteriors), we will rather 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 $\hat{\boldsymbol{b}}_{gp}$ is random and, using the law of total expectation with $\boldsymbol{b}_{gp}$ as well as 2 and then **??**, we obtain:

$$\begin{aligned}
[\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, c_{gpj}=1, d_{gpl}=1] =& _{\boldsymbol{b}_{gp}} [[\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, c_{gpj}=1, d_{gpl}=1, \boldsymbol{b}_{gp}]] \\
=& _{\boldsymbol{b}_{gp}} [\boldsymbol{b}_{gp}|c_{gpj}=1, d_{gpl}=1] \\
=& \mathbf{0}
\end{aligned} \tag{16}$$

Now using the law of total variance with $\boldsymbol{b}_{gp}$ as well as 2 and **??**, we obtain:

$$\begin{aligned}
[\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, c_{gpj}=1, d_{gpl}=1] =& _{\boldsymbol{b}_{gp}} [[\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, c_{gpj}=1, d_{gpl}=1, \boldsymbol{b}_{gp}]] \\
+& _{\boldsymbol{b}_{gp}} [[\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, c_{gpj}=1, d_{gpl}=1, \boldsymbol{b}_{gp}]] \\
=& _{\boldsymbol{b}_{gp}} [\hat{V}_{gp}] \\
+& _{\boldsymbol{b}_{gp}} [\boldsymbol{b}_{gp}|c_{gpj}=1, d_{gpl}=1] \\
=& \hat{V}_{gp} + U_{0jl}
\end{aligned} \tag{17}$$

Therefore the posterior weight is:

$$\tilde{w}_{jl} = \frac{\hat{\eta}_j \hat{\lambda}_l \; _R(\hat{\boldsymbol{b}}_{gp}; \mathbf{0}, U_{0jl} + \hat{V}_{gp})}{\sum_{j'} \sum_{l'} \hat{\eta}_{j'} \hat{\lambda}_{l'} \; _R(\hat{\boldsymbol{b}}_{gp}; \mathbf{0}, U_{0j'l'} + \hat{V}_{gp})} \tag{18}$$

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

In the end, the posterior of the effect size is:

$$\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp}=1 \; \sim \; \sum_{j=1}^{J} \sum_{l=1}^{L} \tilde{w}_{jl} \; \boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, c_{gpj}=1, d_{gpl}=1 \tag{19}$$

In practice, we may be interested in reporting the grand mean of this posterior, noted $\boldsymbol{\mu}_{gp1}$, and its grand covariance matrix, noted U_{gp1}.

Let us introduce Z_{jl} the indicator variable equal to 1 if the variable $\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp}=1$ comes from the j,l-th component. Using the law of total expectations by conditioning on Z_{jl}, the grand mean simply is a weighted sum of **??**:

$$\begin{aligned}
\boldsymbol{\mu}_{gp1} &= [\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp}=1] \\
&=_{Z_{jl}} [[\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, c_{gpj}=1, d_{gpl}=1]] \\
&=_{Z_{jl}} [\boldsymbol{\mu}_{gp1jl}] \\
&= \sum_{j,l} \tilde{w}_{jl} \; \boldsymbol{\mu}_{gp1jl}
\end{aligned} \tag{20}$$

Similarly, for the grand covariance matrix, we can use the law of total covariances:

$$
\begin{aligned}
U_{gp1} &= [\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp} = 1] \\
&= {}_{Z_{jl}}\left[[\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp} = 1]\right] - {}_{Z_{jl}}\left[[\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}, v_{gp} = 1]\right] \\
&= {}_{Z_{jl}}\left[U_{gp1jl}\right] + {}_{Z_{jl}}\left[\boldsymbol{\mu}_{gp1jl}\right] \\
&= \sum_{j,l} \tilde{w}_{jl}\, U_{gp1jl} + \sum_{j,l} \tilde{w}_{jl}\, (\boldsymbol{\mu}_{gp1jl} - \boldsymbol{\mu}_{gp1})(\boldsymbol{\mu}_{gp1jl} - \boldsymbol{\mu}_{gp1})^T
\end{aligned}
\tag{21}
$$