Explicit Methodology

Sarah Urbut

March 10, 2015

The goal of this document is to explicitly outline what I've done in extending the beta mixed prior model to a case in which we do not put a prior on the covariance matrix for $_j$ which explicitly recognizes the configuration model.

First, define our terms. 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 squared standard errors recorded on the diagonal of an $R \times R$ matrix noted $\hat{V}_j = \mathbb{V}(\hat{\boldsymbol{b}}_j)$.

The likelihood for this gene-snp pair is then:

$$\hat{\boldsymbol{b}}_j | \boldsymbol{b}_j \sim \mathcal{N}_R(\boldsymbol{b}_j, \hat{V}_j) \tag{1}$$

For all j gene-snp pairs, beta $\boldsymbol{b}_j$ represent the unknown standardized effect of a snp 'p' on gene 'g'.

$$\boldsymbol{b}_j | \boldsymbol{\pi}, \boldsymbol{U}_0 \sim \sum_{k,l} \pi_{k,l} \, \mathcal{N}_R(\boldsymbol{0}, \omega_l^2 U_k) \tag{2}$$

Where here I allow $\pi_{k,l}$ to represent the (unknown) prior weight on prior covariance matrix U_k and 'stretch factor' $\omega_{1...L}$. Here, I use two (0.1 and 0.075) and 14 matrices for U_k. See section on choice of covariance matrices. Furhtermore, we allow the latent variable z_j to indicate which combination of covariance matrix and stretch factor we are considering, thus z_j can take on KxL values $z_j = [1,1]...[k,l]$

We know that for a single normal, the posterior on $\boldsymbol{b}_j | \boldsymbol{U}_0$ is simply:

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

where:

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

Which leads us to a corresponding multivariate mixture posterior on $\boldsymbol{b}_{gp}$ as this prior is conjugate to likelihood.

$$\begin{aligned} p(\boldsymbol{b}_j | \hat{\boldsymbol{b}}_j, \hat{V}_j, \hat{\boldsymbol{\pi}} &= \sum_{k=1,l=1}^{K,L} p(\boldsymbol{b}_j | \hat{\boldsymbol{b}}_j, \hat{V}_j, k, l) \mathsf{P}(z_j = k, l | \hat{\boldsymbol{b}}_j, \hat{V}_j, \hat{\boldsymbol{\pi}}), \\ &= \sum_{k=1,l=1}^{K,L} (\boldsymbol{b}_j | \hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = k, l) \tilde{\pi}_{k,l} \end{aligned} \tag{3}$$

Where the posterior weight $\tilde{\pi}_{k,l}$ is simply

$$\tilde{\pi}_{k,l} = \frac{\Pr(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = k,l)\hat{\pi}_{kl}}{\sum_{k=1,l=1}^{K,L} \Pr(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, z_j = k,l)\hat{\pi}_{kl}} \tag{4}$$

Note also that $\hat{\pi}_{kl}$ represents the prior weights which are estimated hierarchically, using an EM algorithm, detailed in the corresponding section.

1 Choice of Covariance Matrices U_{kl}

Suppose that we form the following matrices to compute the relevant quantities:

- $\hat{\mathbf{B}}$, is the $J \times R$ matrix of standardized MLEs for each snp-gene pair across all $R = 43$ tissues;
- $\hat{\mathbf{SE}}$ is the corresponding $J \times R$ matrix of standard errors of the corresponding $\hat{\boldsymbol{b}}_j$ across all $R = 43$ tissues;
- $\mathbf{X_t}$ is the corresponding $J \times R$ matrix of t computed for each gene-snp pair statistics across all $R = 43$ tissues;
- $\mathbf{X_c}$ is the RxR covariance matrix of samples, computed by subtracting the column means for each tissue from X_t and computed as $\frac{1}{J} X_t^t X_t$
- $\mathbf{UDV^t}$ is the singular value decomposition of X_t, thus, U is the JxR matrix of eigenvectors of the 'feature covariance matrix' in its columns, d is the RxR diagonal matrix of singular values, and V^t is the RxR matrix with the eigenvectors of the tissue covariance matrix in its rows.
- $\mathbf{\Lambda F}$ is the sparse factor decomposition of ${\mathbf{X_c}}^t$, thus λ is JxQ matrix of factor loadings, where Q is the number of factors chosen and $\mathbf{F}$ is QxR matrix of factors, loosely corresponding to the 'eigenconfigurations' discussed in the 'next steps pdf'. Essentially, each factor may be composed of multiple tissues, analogous to the configuration model, in which a tissue can be active in multiple configurations and a Factor can, in turn, contain many active tissues. However, a Sparse Prior on the rows of λ means that a SNP is maximally active in one Factor.

For a given $\omega \in [0.075, 0.1]$, we specify 4 'types' of RxR prior correlation matrices $U_{k,l}$.

- $U_{k=1,l=1,2} = \omega_l \ \mathbf{I}_R$
- $U_{k=2,l=1,2} = \omega_l \ \mathbf{X_c}$ The (naiively) estimated tissue covariance matrix
- $U_{k=3,l=1,2} = \omega_l \ \frac{1}{J} \ \mathbf{V}_{1..p} \ \mathbf{D}^2_{1..p} \ \mathbf{V}^t_{1..p}$ is the rank p eigenvector approximation of the tissue covariance matrices, i.e., the sum of the first p eigenvector approximations.

1.1 Factor Analysis Matrices

Let $\mathbf{\Lambda}$ represent the J x Q matrix of loadings, such that each SNP is maximally loaded on one $Factor$. Essentially, we have put a sparse prior on the rows of $\mathbf{\Lambda}$ such that each SNP can be a member of at most one factor class. The KxR matrix of factors then represents the matrix of 'eigenconfigurations' which indicates expression in a particular subset of tissues - each Factor may indicate activity in one, several or no tissues. As a critical difference with PCA, the factors can outnumber the tissues and each tissue can be a member of more than one factor.

- $U_{k=4:13,l=1,2} = \frac{1}{J} (\mathbf{F\Lambda})^t_q \ \mathbf{\Lambda F}_q$ corresponding to the sparse factor representation of the tissue covariance matrix (not the sum of the first q, as above)
- $U_{k=14,l=1,2} = \frac{1}{J} (\mathbf{F\Lambda})^t \ \mathbf{\Lambda F}$ is the sparse factor representation of the tissue covariance matrix, estimated using all q factors.

2 EM Algorithm Outline

Here the incomplete-data likelihood function is

$$L(\pi; \hat{\boldsymbol{b}}, \mathbf{z}) = P(\hat{\boldsymbol{b}}, \mathbf{z}|\theta) = \prod_{j=1}^{J} \sum_{k,l=1}^{KL} \pi_{kl} \Pr(\hat{\boldsymbol{b}}|z_j = [k,l]) \quad (5)$$

Now, in order to estimate the hierarchical prior weights $\pi_{k,l}$ we compute the KxL dimensional likelihood at each each gene snp pair j by evaluating the probability of observing

given that we know the true $\boldsymbol{b_j}$ arises from component k, l:

$$\begin{aligned} &\mathcal{L}(\pi_{\mathbf{kl}}; \hat{\boldsymbol{b}}_j, U_{0,k,l}\hat{V}_j) \\ &\qquad = \Pr(\hat{\boldsymbol{b}}_j|z_j = [k,l]) \\ &\qquad = \mathcal{N}_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, U_{0kl'} + \hat{V}_j) \end{aligned} \quad (6)$$

Which means we form a $J \times KL$ dimensional matrix entitled 'global.lik' in my .Rmd file,where in each row vector is the probability of the vector of observed MLEs given that the true j arose from element K, L, as specified by its corresponding prior covariance matrix $\mathbf{U_{0kl}}$. You simply compute the probability from an R dimensional multivariate normal with mean $\mathbf{0}$ and variance $\mathbf{U_{0kl}} + \hat{\mathbf{V}}_\mathbf{j}$. I treat each of the j rows as an i.i.d. sample from which to maximize the likelihood over using the mixEM algorithm.

In order to compare this with an 'intuitive estimate', I sum the columns and divide by the total likelihood of the dataset. Then, we can compare the estimates:

$$\hat{\pi}_{naiive.kl} = \frac{\sum_j \mathbf{L}(\pi; \hat{\boldsymbol{b}}_j, U_{k,l}\hat{V}_j)}{\sum_{j,k,l} \mathbf{L}(\pi; \hat{\boldsymbol{b}}_j, U_{k,l}\hat{V}_j)} \quad (7)$$

with the output of $mixEM$. These results are compared in the two bar plots entitled "mixEM estimated pi" and "naiiveWeights estimated pi" and attached at the end of this document.

3 Posterior Mean Plots

For each of the j pairs and each component k and l I can compute the posterior mean and covariance matrix using the formula for a single multivariate I store these in the objects $all.means$ and $all.covs$, where $all.means[j][k][l]$ corresponds to the posterior mean for the jth pair evaluated with prior covariance matrix U_{0kl} and results from the calculation: $\boldsymbol{\mu}_{jkl1} = \boldsymbol{U_{jkl1}}(\hat{\boldsymbol{V}}_\boldsymbol{j}^{-1}\hat{\boldsymbol{b}}_j)$ and $all.covs[j][k][l] = \boldsymbol{U_{jkl1}} = (\boldsymbol{U_{0kl}^{-1}} + \hat{\boldsymbol{V}}_\boldsymbol{j}^{-1})^{-1}$.

For each of the j pairs, I generate a corresponding posterior-weight matrix using $\tilde{\pi}_{k,l}$ as in (4) where I evaluate the probability of the data at each component $[kl]$ using the corresponding prior covariance matrix in computing $\mathcal{N}_R(\hat{\boldsymbol{b}}_j; \mathbf{0}, U_{0kl'} + \hat{V}_j)$ and weight the resulting likelihood by its EM estimated prior weight, and then divide by the corresponding sum over all prior weights and likelihoods. This is computed using the post.weight.mat function function I have written in R.

In practice: for each of the j pairs, we can then compute the corresponding r dimensional vector of posterior means $\boldsymbol{\mu}_{j1}$ as simply the sum of each element of $all.means[j][k][l]$ weighted by its corresponding posterior weight. You can see in the code chunk that for pair j, I loop through each of the k, l component pairs and store the r dimensional row vector of weighted $\tilde{\pi}_{jkl}\boldsymbol{\mu}_{jkl}$ for each component in the $K \times L$ by R matrix $temp$. I then sum the columns to compute a vector of aggregated posterior means $\boldsymbol{\mu}_{j1}$ to produce an r dimensional row vector which I store in $post.means$. I complete this for all j gene SNP I then plot this aggregated posterior mean vector $\boldsymbol{\mu_{j1}}$ for 10 gene SNP pairs in the corresponding plots.

4 Figures

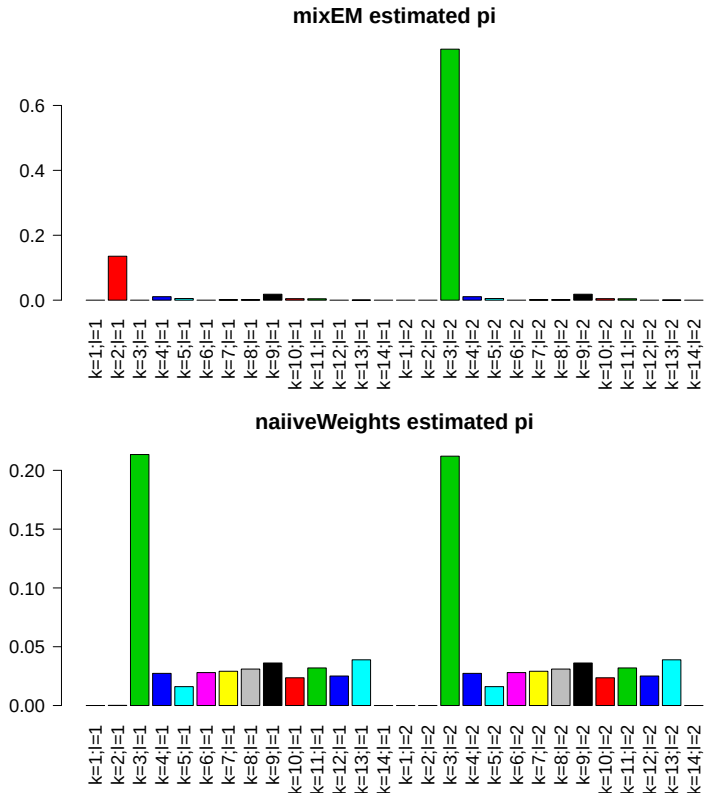

Figure 1: Comparing Estimation of Prior component weights, $\pi_{k,l}$ using mixEM and summing over one iteration of the likelihood

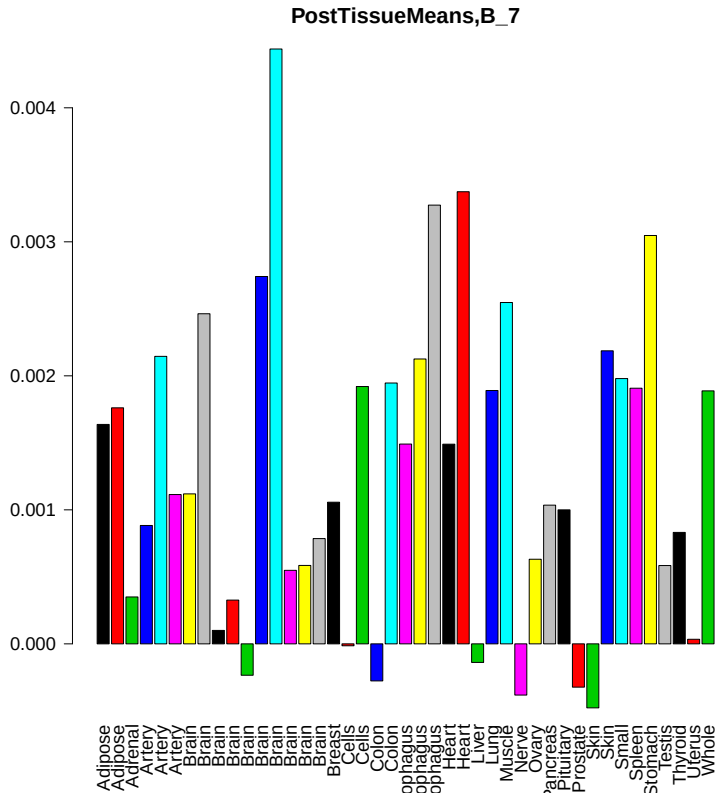

Figure 2: Weighted Posterior Mean Across All components for gene snp pair 7. See positive in most tissues