

# Explicit Methodology

Sarah Urbut

April 14, 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,  $\hat{\mathbf{b}}_j$ , and their squared standard errors recorded on the diagonal of an  $R \times R$  matrix noted  $\hat{V}_j = \mathbb{V}(\hat{\mathbf{b}}_j)$ .

The likelihood for this gene-snp pair is then:

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

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

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

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. Furthermore, 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  $K \times L$  values  $z_j = [1, 1] \dots [k, l]$

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

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

where:

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

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

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

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

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

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  $\hat{\mathbf{b}}_j$  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  $R \times R$  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  $J \times R$  matrix of eigenvectors of the 'feature covariance matrix' in its columns,  $d$  is the  $R \times R$  diagonal matrix of singular values, and  $V^t$  is the  $R \times R$  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  $\lambda$  is  $J \times Q$  matrix of factor loadings, where  $Q$  is the number of factors chosen and  $\mathbf{F}$  is  $Q \times R$  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  $\lambda$  means that a SNP is maximally active in one Factor.

For a given  $\omega \in [0.075, 0.1]$ , we specify 4 'types' of  $R \times R$  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 (naively) estimated tissue covariance matrix
- $U_{k=3, l=1, 2} = \omega_l \frac{1}{J} \mathbf{V}_{1..p} \mathbf{D}_{1..p}^2 \mathbf{V}_{1..p}^t$  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  $J \times Q$  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  $K \times R$  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.

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

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

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

given that we know the true  $\mathbf{b}_j$  arises from component  $k, l$ :

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

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  $\mathbf{U}_{0\mathbf{kl}}$ . You simply compute the probability from an  $R$  dimensional multivariate normal with mean  $\mathbf{0}$  and variance  $\mathbf{U}_{0\mathbf{kl}} + \hat{\mathbf{V}}_j$ . 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:

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

with the output of *mixEM*. These results are compared in the two bar plots entitled “mixEM estimated pi” and “naiveWeights 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  $j$ th pair evaluated with prior covariance matrix  $U_{0kl}$  and results from the calculation:  $\boldsymbol{\mu}_{jkl1} = \mathbf{U}_{jkl1} (\hat{\mathbf{V}}_j^{-1} \hat{\mathbf{b}}_j)$  and *all.covs*[ $j$ ][ $k$ ][ $l$ ] =  $\mathbf{U}_{jkl1} = (\mathbf{U}_{0kl}^{-1} + \hat{\mathbf{V}}_j^{-1})^{-1}$ .

For each of the  $j$  pairs, I generate a corresponding posterior-weight matrix using  $\tilde{\pi}_{k,l}$  as in (??) where I evaluate the probability of the data at each component  $[kl]$  using the corresponding prior covariance matrix in computing  $\mathcal{N}_R(\hat{\mathbf{b}}_j; \mathbf{0}, U_{0kl} + \hat{V}_j)$  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 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

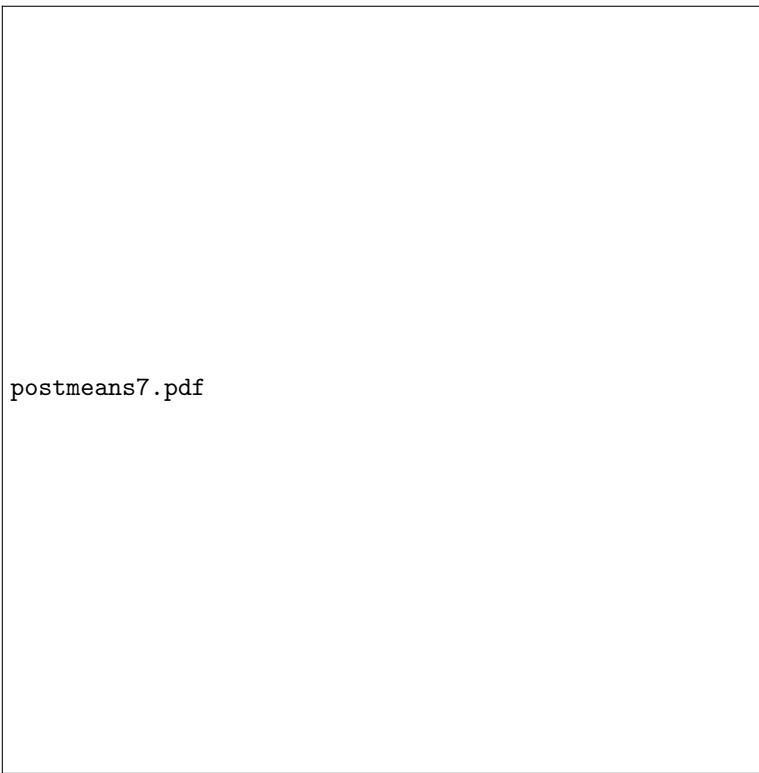


Figure 2: Weighted Posterior Mean Across All components for gene snp pair 7. See positive in most tissues