# mash No Baseline

Sarah Urbut

August 10, 2016

## Contents

# 1 Purpose

The purpose of this document is to propose a method for extending `mash` to estimate 'true' effects across conditions in a setting in which no obvious baseline exists. We assume that we observe noisy, uncentered averages $\hat{\boldsymbol{C}}_{jr}$ in each of $R$ conditions, and seek to estimate the underlying true 'deviations' from average measurement across conditions and can be seen as the effects in `mash`.

Here, the use of bold-face notation indicates a vector, while matrix quantities are typeset in capital but unboldface letters.

# 2 Defining the Old 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.*

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

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

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

Now, we observe for each gene $j$ a vector of uncentered noisy average feature expression $\hat{\boldsymbol{C}}$ across R conditions:

$$\hat{\boldsymbol{C}}|\boldsymbol{C} \sim N_R(\boldsymbol{C}, \hat{V}) \tag{3}$$

where the 'true' uncentered averages $\boldsymbol{C}$ can be written as follows:

$$\boldsymbol{C}|\mu, \boldsymbol{v} = \mu\mathbf{1} + \boldsymbol{v} \tag{4}$$

Where $\mu$ is a scalar that is the mean of the 'true' uncentered averages $\boldsymbol{C}$.

$\boldsymbol{v}$ is a zero-centered mixture of multivariate normals:

$$\boldsymbol{v}|\boldsymbol{\pi}, \mathbf{U} \sim \sum_{\mathbf{k},\mathbf{l}} \pi_{\mathbf{k},\mathbf{l}}\, N_{\mathbf{R}}(\mathbf{0}, \omega_{\mathbf{l}} \mathbf{U}_{\mathbf{k}}) \tag{5}$$

Critically, our quantity of interest now, $\boldsymbol{v}$ represents the true 'deviations' from average gene expression across each condition and can be seen as the effects in `mash`.

# 3 Applications

We will again apply a two-step process to our selection of covariance matrices, where we select a set of denoised 'pattern' matrices $U_k$ by using the EM algorithm on the max effects across conditions, and then expanding this list by a fixed grid of scalar weights $\omega_l$ such that we conclude with a list of $P = KxL$ covariance matrices $\Sigma$. We can then:

- estimate the $P$ prior weights $\boldsymbol{\pi}$ on this fixed P-list of covariance matrices from a training matrix of randomly selected feature expression measurements across conditions
- compute the posterior distribution $\boldsymbol{v}|L\hat{\boldsymbol{C}}, \boldsymbol{s_j}$

  Let

  $$L\boldsymbol{C} = L\mu\mathbf{1} + L\boldsymbol{v} \tag{6}$$

  L is the $RxR$ centering matrix $L_r = I_r - \frac{1}{r}\mathbf{1}\mathbf{1}^\top$ which removes the mean of each R column vector.

  Then :

  $$\begin{aligned} L\boldsymbol{C} &= L\mu\mathbf{1} + L\boldsymbol{v} \\ L\boldsymbol{C} &= 0 + L\boldsymbol{v} \\ L\hat{\boldsymbol{C}} &= L\boldsymbol{v} + E \end{aligned} \tag{7}$$

  Where $E \sim \mathcal{N}(0, L\hat{V}L')$

## 3.1 Selecting The Covariance Matrices

We initiate our set of covariance matrices for the denoising step as before in `mash`, where now we compute the empirical covariance matrices and a variety of dimensional reductions on the feature-centered JxR *matrix* of maximum average, $L\hat{C}'$ instead of $\hat{C}$ alone. In practice, we actually use the matrix of maximum uncentered $T$ statistics. Three critical things to note:

1. Here, L will be $RxR$ because we need $U_k$ to be RxR

2. When denoising with Bovy, our previous approach used the matrix of maximum $T$ statistics $T_{Mxr}$ to both initialize and train the BovyEM. Now, we will initalize with the MXR matrix of t($L_{R,R}T'$) (or alternatively, $TL$) and train the EM on the MxR-1 matrix of maximum $t(L_{R-1,R}T')$
3. When choosing $\omega$, we will use the diagonal of $LVL'$, where $V$ is $D(s.j^2)$, and select from the MxR matrix of centered T statistics and their centered standard errors.

# 4 Likelihood

Now we will replace the RxR matrix $L$ with the R-1xR matrix $L*$, effectively removing a data point from the observed uncentered statistics, such that the rank of the marginal variance of $w$ is guaranteed to be equal to the dimension of $w$.

Now for each gene J at each component k, integrating over $\boldsymbol{v}$,

$$\begin{aligned} L_{R-1,R}\boldsymbol{C} &\sim \mathcal{N}(0, L_{R-1,R}U_k L'_{R-1,R}) \\ L_{R-1,R}\hat{\boldsymbol{C}} &\sim \mathcal{N}(0, L_{R-1,R}U_k L'_{R-1,R} + L_{R-1,R}\hat{V}L'_{R-1,R}) \end{aligned} \tag{8}$$

And thus we can use the Bovy et al algorithm invoked in both the Extreme Deconvolution package and in 'Sarah's MixEm' where:

$$T_{jp} = L_{R-1,R}U_k L'_{R-1,R} + L_{R-1,R}\hat{V}_j L'_{R-1,R} \tag{9}$$

For each gene, and $w_j = L_{R-1,R}\hat{\boldsymbol{C}}_j$.

Recall that our previous approach was simplified by the fact that $\boldsymbol{w}_j$ was simply $\hat{\boldsymbol{b}_j}$ and the projection matrix was simply the $I_r$ identity matrix. Our inference on $\boldsymbol{b}$ was analogous to their inference on $_j$.

As before, we are interested in returning the prior covariance $U_k$ matrices of the 'true' deviations $\boldsymbol{v}$, which we will then rescale by choosing a set of $\omega$ that are appropriate to $L\hat{\boldsymbol{C}}$ to comprise a set of $P = KxL$ prior covariance matrices $\Sigma$.

and choose the set of $\pi$ that maximizes compute the following likelihood at each of the P components:

$$L_{R-1,R}\hat{\boldsymbol{C}}_j \sim \mathcal{N}(0, L_{R-1,R}\Sigma_p L'_{R-1,R} + L_{R-1,R}\hat{V}_j L'_{R-1,R}) \tag{10}$$

## 5 Posteriors

Now, as before we can compute a posterior distribution such that:

$$\boldsymbol{v}|L_{R-1,R}\hat{\boldsymbol{C}}, \pi, \Sigma, \boldsymbol{s} \sim N(\mu^1, U^1) \tag{11}$$

Where at each of the P components for each gene J

$$\begin{aligned} \mu^1_{jp} &= \Sigma_p L'_{R-1,R} T^{-1}_{jp} L_{R-1,R} \hat{\boldsymbol{C}}_j \\ U^1_{jp} &= \Sigma_p - \Sigma_p L'_{R-1,R} T^{-1}_{jp} L_{R-1,R} \Sigma_p \end{aligned} \tag{12}$$

## 6 Differences required over `mash` implementation

- We will now work with a matrix of observed column-centered gene averages, $L\hat{C}'$ in order to:
  1. initialize our choice of $U_k$;
  2. choose the maxes by which to denoise,
  3. choose our set of scales, $\omega_l$
  4. compute our hierarchical weights, $\boldsymbol{\pi}_p$ as well as our posteriors.
- It is critical to note that here **L will need to be RxR because $U_k$ must be RxR**
- The new distribution we seek to estimate for each j is then $v|L_{R-1,R}\hat{\boldsymbol{C}}, \boldsymbol{s}_j$
- To choose the maxes, I think we ought to use a $w_j$ cutoff since computing the univariate lfsr on $w_j$ and the diagonal of $LVL'$ assumes that $LVL'$ is diagonal when we know it cannot be.