On the issue of using low-rank covariance matrix in mash

Dongyue Xie

1 Model

Assume

$$\begin{aligned} \boldsymbol{x}_i &\sim N_R(\boldsymbol{\mu}_i, \boldsymbol{I}), \\ \boldsymbol{\mu}_i &\sim \sum_k \pi_k N_R(\boldsymbol{0}, \boldsymbol{U}_k), \end{aligned} \tag{1}$$

where $N_R(\cdot,\cdot)$ denotes a R-dimensional multivariate normal distribution. When using the model for multiple testing, mash first estimates $\boldsymbol{\pi}, \boldsymbol{U}$ using empirical Baes procedure then obtaines lfdr, lfsr and the other posterior summaries.

Using rank-1 $\boldsymbol{U}_k$ in above model breaks down one of the posterior summaries - lfsr. The issue generally apears when using rank-deficient covariance matrices. The assumption of rank-1 is not explicit specified in the mash model, but the ED algoithm preserves the rank of initializations. The initializations are generated using PCA hence most of them are rank-1 matrices. The TEEM algorithm does not preserve the rank but it truncates the eigenvalues and the estimated covraince matrices could still be low rank.

The simplest solution is to use full-ranked matrices as intializations. For example, generate covaraince matrices from a Wishart distribution as intializations. But this generally increase the computation cost and the estimated covariance matrices might be very close to singular.

In the following sections we discuss two possible solutions.

A more systematic solution to this problem is adding a full ranked matrix to $\boldsymbol{U}_k$, and the simplest one perhaps is $\sigma_k^2 \boldsymbol{I}$. In the next section we will give a Bayesian interpretation of the choice and describe an empirical Bayes procedure for determining σ_k^2.

2 Prior on covaraince matrix

Preferably, the parameter σ^2 should be data-dependent. Inspired by early works on empirical Bayes estimator of sample covariance matrix, we can employ the EB estimator of covariance matrix by introducing a prior on it. For more details, see Efron and Morris(1972), Efron and Morris(1976), and Haff(1980),

Define $\boldsymbol{T}_k = \boldsymbol{U}_k + \boldsymbol{I}$. When estimating $\boldsymbol{U}_k$ using EM algorithm, the last iteration gives

$$\hat{\boldsymbol{T}}_k^{\text{mle}} = \frac{\sum_i \gamma_{ik} \boldsymbol{x}_i \boldsymbol{x}_i^T}{n_k}, \tag{2}$$

where mle denotes the maximum likelihood estimator, γ_{ik} are posterior probabilities of $\boldsymbol{x}_i$ coming from mixture component k and $n_k = \sum_i \gamma_{ik}$. (In practice one would use a modified $\hat{\boldsymbol{T}}_k^{\text{mle}}$ by truncating eigenvalues to ensure $\boldsymbol{U}_k$ is a well-defined covariance matrix, and $\hat{\boldsymbol{U}}_k^{\text{mle}} = \hat{\boldsymbol{T}}_k^{\text{mle}} - \boldsymbol{I}$.) Most of the posterior probabilities γ_{ik} should be close to either 0 or 1 at the last iteration of EM. The reason is that assuming the model is correct, the posteriors should concentrate on the true mixture compoenent assignments. The number n_k can be regarded as the effective sampe size of mixture component k, i.e. the number of $\boldsymbol{x}_i$ that are from the mixture k. Therefore, we have the following approximation

$$\hat{\boldsymbol{T}}_k^{\text{mle}} = \frac{\sum_i \gamma_{ik} \boldsymbol{x}_i \boldsymbol{x}_i^T}{n_k} \approx \frac{1}{n_k} \sum_{i \in \mathcal{K}} \boldsymbol{x}_i \boldsymbol{x}_i^T, \tag{3}$$

where $\mathcal{K}$ is an index set of samples from mixture component k, so for $i \in \mathcal{K}$, $\boldsymbol{x}_i \sim N(\boldsymbol{0}, \boldsymbol{U}_k)$.

Let $\tilde{\boldsymbol{T}}_k = n_k \hat{\boldsymbol{T}}_k^{\text{mle}}$, then based on the approxiamation

$$\tilde{\boldsymbol{T}}_k \sim \mathcal{W}(\boldsymbol{T}_k, n_k), \tag{4}$$

where $\mathcal{W}$ denotes Wishart distribution.

Assume $\boldsymbol{T}_k$ has an inverse Wishart prior,

$$\boldsymbol{T}_k \sim \mathcal{W}^{-1}(\sigma_k^2 \boldsymbol{I}, \nu), \tag{5}$$

where ν is known integer and $\sigma_k^2 > 0$. The parameter ν is the degree of freemdom and $\nu \geq R$. Larger ν gives smaller prior variance. In practice, we shall set $\nu = R$ or $\nu = R + 1$.

Haff(1980) derived an estimator of σ_k^2 by maximizing the marginal likelihood $f(\tilde{\boldsymbol{T}}_k; \sigma_k^2)$. Directly maximizing $f(\tilde{\boldsymbol{T}}_k; \sigma_k^2)$ is non-trivial and the estimator instead maximizes the first order approximation of it. The empirical Bayes estiamtor of σ_k^2 is

$$\hat{\sigma}_k^2 = \frac{R\nu}{n_k + \nu} \frac{1}{\text{Tr}(\tilde{\boldsymbol{T}}_k^{-1})}. \tag{6}$$

The estimator $\hat{\sigma}_k^2$ is strictly postive since $\frac{R\nu}{n_k+\nu} > 0$ and $\tilde{\boldsymbol{T}}_k$ is positive definite.

The posterior distribution of $\boldsymbol{T}_k$ is

$$\boldsymbol{T}_k | \tilde{\boldsymbol{T}}_k \sim \mathcal{W}^{-1}(\tilde{\boldsymbol{T}}_k + \sigma_k^2 \boldsymbol{I}, n_k + \nu), \tag{7}$$

The posterior mean of $\boldsymbol{T}_k$ is

$$\mathbb{E}(\boldsymbol{T}_k | \tilde{\boldsymbol{T}}_k, \sigma_k^2) = \frac{\tilde{\boldsymbol{T}}_k + \sigma_k^2 \boldsymbol{I}}{n_k + \nu - R - 1}. \tag{8}$$

The empirical Bayes estiamtor of $\boldsymbol{T}_k$ is the posterior mean with estimated σ_k^2,

$$\hat{\boldsymbol{T}}_k^{\text{eb}} = \frac{\tilde{\boldsymbol{T}}_k + \hat{\sigma}_k^2 \boldsymbol{I}}{n_k + \nu - R - 1}. \tag{9}$$

To get the EB estiamtor of $\boldsymbol{U}_k$, we shall write

$$\begin{aligned} \hat{\boldsymbol{T}}_k^{\text{eb}} &= \hat{\boldsymbol{U}}_k^{\text{eb}} + \boldsymbol{I}, \\ \tilde{\boldsymbol{T}}_k &= n_k \hat{\boldsymbol{T}}_k^{\text{mle}} = n_k(\hat{\boldsymbol{U}}_k^{\text{mle}} + \boldsymbol{I}). \end{aligned} \tag{10}$$

Subsitituing (10) into (9), we have

$$\hat{\boldsymbol{U}}_k^{\text{eb}} = \frac{n_k}{n_k + \nu - R - 1}\hat{\boldsymbol{U}}_k^{\text{mle}} + \frac{R + 1 - \nu + \hat{\sigma}_k^2}{n_k + \nu - R - 1}\boldsymbol{I}. \tag{11}$$

When $\nu = R + 1$, we have

$$\hat{\boldsymbol{U}}_k^{\text{eb}} = \hat{\boldsymbol{U}}_k^{\text{mle}} + \frac{\hat{\sigma}_k^2}{n_k}\boldsymbol{I} = \hat{\boldsymbol{U}}_k^{\text{mle}} + \frac{R(R+1)}{(n_k + R + 1)\text{Tr}\left((\hat{\boldsymbol{T}}_k^{\text{mle}})^{-1}\right)}\boldsymbol{I} \tag{12}$$

3 Lower bound of variance

We start with a simple one dimensional case. Assume $x_i \sim N(0, 1 + \sigma^2)$, our goal is to estimate σ^2 and possibley set $\hat{\sigma}^2$ to be the largest value of it consistent with data.

The log likelihood of $l(\sigma^2)$ is

$$l(\sigma^2) = -\frac{n}{2}\log(1 + \sigma^2) - \frac{ns^2}{2(1 + \sigma^2)}, \tag{13}$$

where $s^2 = \sum_i x_i^2/n$, the MLE of $(1 + \sigma^2)$.

Then we have

$$l(\sigma^2) - l(0) = -\frac{n}{2}\left(\log(1 + \sigma^2) - \frac{\sigma^2}{1 + \sigma^2}s^2\right). \tag{14}$$

When $s^2 = 1$(truncated) or equivalently $\hat{\sigma}^2 = 0$, and σ^2 is very small, using the approximation $\log(1 + \sigma^2) \approx \sigma^2$, we have

$$l(\sigma^2) - l(0) \approx -\frac{n}{2}\frac{(\sigma^2)^2}{1 + \sigma^2} \approx -\frac{n}{2}(\sigma^2)^2. \tag{15}$$

Ae rule of thumb is that 2 units drop in log likelihood gives 95% confidence intervals. For example, $l(\hat{\theta}) - 2 \approx l(\hat{\theta} + 2 * se(\hat{\theta}))$.

Solving for $l(\sigma^2) - l(0) = -2$, we have $\sigma^2 \approx \frac{2}{\sqrt{n}}$. This suggests that we can set the lower bound of σ^2 to be $\frac{2}{\sqrt{n}}$.

Now going back to the multivariate case. Let the eigendecomposition of $\boldsymbol{U}_k$ be $\boldsymbol{Q}_k\boldsymbol{\Lambda}_k\boldsymbol{Q}^T$, and $\boldsymbol{U}_k + \boldsymbol{I} = \boldsymbol{Q}_k(\boldsymbol{\Lambda}_k + \boldsymbol{I})\boldsymbol{Q}_k^T$. Then

$$\begin{aligned} \boldsymbol{x}_i &\sim N(\boldsymbol{0}, \boldsymbol{U}_k + \boldsymbol{I}), \\ \Longrightarrow \boldsymbol{Q}_k^T\boldsymbol{x}_i &\sim N(\boldsymbol{0}, \boldsymbol{\Lambda}_k + \boldsymbol{I}). \end{aligned} \tag{16}$$

This suggests that we can set the eigenvalues of $\hat{\boldsymbol{U}}_k$ to be at least $\frac{2}{\sqrt{n_k}}$.