

1 Methods and Materials

As a simple version of a shrinkage estimator for a covariance matrix, one fits a convex combination of the empirical sample covariance matrix (S) along with a chosen target matrix T , which can be chosen to be an identity matrix or constant correlation matrix. The mixing proportion δ in the convex combination $\delta T + (1 - \delta)S$ is usually selected to maximize the expected accuracy of the shrunk estimator. In our approach, we shrink the correlations to 0, implying the target matrix is the identity matrix. However, instead of a single target T , we assume multiple random targets T_1, T_2, \dots, T_k , all centered around the same identity correlation matrix but each with different degrees of noise variation, usually increasing with k . The belief is that such an approach would adaptively decide on the amount of shrinkage without requiring to follow a Cross Validation approach.

Let us denote the sample correlation matrix by $R = ((r_{ij}))_{i,j=1,2,\dots,P}$, P being the number of features, calculated over N data samples.

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \quad s_{ij} = \frac{1}{n} \sum_{n=1}^N (x_{ni} - \bar{x}_i)(x_{nj} - \bar{x}_j) \quad (1)$$

where s_{ij} is the sample covariance between the vectors $x_{*,i}$ and $x_{*,j}$.

We propose the following model

For any two features i and j with $i < j$, we define a binary size K latent variable vector $((Z_{ij:k}))$ where $Z_{ij:k}$ takes the values 1 with probability π_k and 0 otherwise.

$$Pr[Z|\pi] = \prod_{i=1}^P \prod_{j<i} \prod_{k=1}^K \pi_k^{Z_{ij:k}} \quad (2)$$

We define latent variables ρ_{ij} , such that

$$Pr(\rho|Z, \pi) = \prod_{i=1}^P \prod_{j<i} \prod_{k=1}^K [N(\rho_{ij} : 0, \sigma_k^2)]^{Z_{ij:k}} \quad (3)$$

We assume Normal distribution for the ρ ,

$$Pr(\hat{\rho}_{ij}|\rho) = \prod_{i=1}^P \prod_{j<i} N\left(\hat{\rho}_{ij}|\rho_{ij}, s_n^2 = \frac{1}{n-3}\right) \quad (4)$$

where $\hat{\rho}_{ij}$ are the Fisher's z-scores of the sample correlations r_{ij} given by

$$\hat{\rho}_{ij} = \frac{1}{2} \log \left(\frac{1 + r_{ij}}{1 - r_{ij}} \right) \quad (5)$$

The model implies that we shrink the $\hat{\rho}_{ij}$ to 0 but the amount of shrinkage is decided both by the number of independent samples $s_n^2 = \frac{1}{n-3}$ and also by σ_k .

We propose three different models depending on our assumptions on π and σ .

- *CorShrink-ML*: We choose a fixed grid of σ values, selected such that it covers the span of the variation of the data well. Here we propose to use a similar grid (with minor adjustments) as suggested in Stephens 2016 [?] for modeling false discovery rates. We add a component with $\sigma_k = 0$ that represents the null component of the prior. We fit the mixing proportions π of the components using EM algorithm.
- *CorShrink-VEM*: We use the same grid of σ values as in the *CorShrink-ML* model, but now we assume a Dirichlet prior on π , that puts a high weight on the null component and treats the other components equivalently. From performance comparisons on simulated data, we assumed the default Dirichlet prior to be $Dir(10, 1, \dots, 1)$.
- *CorShrink-VEM2*: We additionally assume the σ values to be not fixed but to come from a Inverse-Gamma distribution. We assume $Inv - Gamma(\varepsilon, \varepsilon)$ distributions which are relatively non-informative in order to make the choice of σ very flexible. For our applications in this paper, assume ε to be 0.01.

The estimation of π for the *CorShrink-ML* model was performed using the **ashr** package due to Matthew Stephens [?], which fits an EM algorithm. For the *CorShrink-VEM* and *CorShrink-VEM2*, we use Mean Field Variational EM models to estimate the model parameters. Variational methods are faster than MCMC methods as they often provide analytic updates to parameters thereby ensuring faster computation [?] [?].

For *CorShrink-VEM2* model where π and σ are both random, we first perform a change of variables

$$\xi_k = \sigma_k^2 + \frac{1}{n-3}$$

Suppose the priors on π and ξ are

$$\pi \sim Dir(\alpha_1, \alpha_2, \dots, \alpha_K) \quad \xi_k \sim Inv - Gamma(a, b) \quad \forall k$$

and then define the mean field variational distribution on the latent variable Z and the parameters π and $\xi_1, \xi_2, \dots, \xi_K$ as follows.

$$q(Z, \pi, \xi) = q(Z)q(\pi) \prod_{k=1}^K q(\xi_k)$$

Then the mean field distribution for π is given by

$$\log q^*(\pi) = E_{Z, \xi} [\log p(\hat{\rho}, Z, \pi, \xi)] \quad (6)$$

$$= E_Z \left[\sum_{k=1}^K (\alpha_k - 1) \log(\pi_k) + \sum_{i=1}^J \sum_{j < i} \sum_{k=1}^K z_{ij:k} \log(\pi_k) \right] + \text{const.} \quad (7)$$

$$= \sum_{k=1}^K \left[(\alpha_k - 1) + \sum_{i=1}^J \sum_{j < i} \delta_{ij:k} \right] \log(\pi_k) \quad (8)$$

$$(9)$$

So the variational distribution for π is of the form

$$\pi \sim \text{Dir}(\pi | \beta_1, \beta_2, \dots, \beta_K) \quad \beta_k = \alpha_k + \sum_{i=1}^J \sum_{j < i} \delta_{ij:k} \quad (10)$$

The variational distribution of the latent variable Z is obtained similarly

$$\log q^*(Z) = E_{\pi, \xi} [\log p(\hat{\rho}, Z, \pi, \xi)] \quad (11)$$

$$= E_{\pi, \xi} [\log p(Z | \pi) + \log p(\hat{\rho} | Z, \xi, \pi)] \quad (12)$$

$$= \sum_{i=1}^P \sum_{j < i} \sum_{k=1}^K z_{ij:k} E_{\pi} (\log(\pi_k)) + \sum_{i=1}^P \sum_{j < i} \sum_{k=1}^K z_{ij:k} \left[\frac{1}{2} E_{\xi} \left[\log \frac{1}{\xi_k} \right] - \frac{\hat{\rho}_{ij}^2}{2} E_{\xi} \left[\frac{1}{\xi} \right] \right] \quad (13)$$

$$(14)$$

It can be shown that

$$E_{\xi_k} \left[\log \frac{1}{\xi_k} \right] = -\log(v_{2k}) + \psi(v_{1k}) \quad E_{\xi_k} \left[\frac{1}{\xi_k} \right] = \frac{v_{1k}}{v_{2k}} \quad (15)$$

$$E_{\pi} (\log(\pi_k)) = \psi(\beta_k) - \psi\left(\sum_{l=1}^K \beta_l\right) \quad (16)$$

where ψ represents the digamma function. Using all of the above results, we get the following distribution of Z ,

$$q^*(Z) = \prod_{i=1}^P \prod_{j < i} \prod_{k=1}^K \delta_{ij:k}^{Z_{ij:k}} \quad (17)$$

where

$$\delta_{ij:k} \propto \exp \left(\left\{ \psi(\beta_k) - \psi\left(\sum_{l=1}^K \beta_l\right) + 0.5 \times (\psi(v_{1k}) - \log(v_{2k})) - \frac{\hat{\rho}_{ij}^2}{2} \frac{v_{1k}}{v_{2k}} \right\} \right) \quad (18)$$

For *CorShrink-VEM* model, the σ_k or $\xi_k = \sigma_k + \frac{1}{n-3}$ are fixed and the variational distribution is of the form

$$q(Z, \pi) = q(Z)q(\pi) \quad (19)$$

The variational distribution is same as in *CorShrink-VEM2* model, whereas the variational distribution of Z can be achieved similarly as follows

$$q^*(Z) = \prod_{i=1}^P \prod_{j < i} \prod_{k=1}^K \delta_{ij:k}^{Z_{ij:k}} \quad (20)$$

where

$$\delta_{ij:k} \propto \exp \left(\left\{ \psi(\beta_k) - \psi\left(\sum_{l=1}^K \beta_l\right) + 0.5 \times \left(\log \frac{1}{\xi_k} \right) - \frac{\hat{\rho}_{ij}^2}{2} \frac{1}{\xi_k} \right\} \right) \quad (21)$$

The *CorShrink-VEM2* model is flexible in choice of π and ξ_k 's, however it also has the problem of hitting a local maxima and the σ_k 's for multiple k 's to converge to same point. In order to counter that, we initialize the parameters first using the *CorShrink-VEM* model that assumes a fixed grid of well spread out ξ_k values. Post the initialization, we apply the *CorShrink-VEM2* model to the parameters.

The other point to note is that actually the ξ_k 's are bounded below by $\frac{1}{n-3}$ which we ignore in defining an Inverse Gamma distribution on the ξ . This is a compromise for very small n and we do therefore do not recommend the use of *CorShrink-VEM2* for very small n values. Having said that, the initialization using *CorShrink-VEM* fixes the ξ_k initial values to be $> \frac{1}{n-3}$ and we usually find that the final estimates would automatically adjust themselves to the lower bound and in case they violate, we forcibly set them to the lower bound value.

In the next section, we discuss the applications of these three models on simulated and a real data drawn from single cell mouse embryo pre-implantation data.