1 Outline

- **Coding** Run the codes for the variational models 1 and 2 using chunks from Matthew's ash code
- **Results** : Simulation study (show the comparative performance of this method with the Strimmer approach). compare the largest eigenvalues and the trends for different choices of n/p. (Eigenvalues comparison plots + Eigenvector distance plot + a table of distance between original and the projected from the various approaches)
- **Results**: Real data application (Deng et al correlation network)- before and after shrinkage. (If possible, we talk briefly about causation networks too). First do PCA to scale down the data to a small number of dimensions and then use this method on the samples to build the correlation network.
- **Results**: Tree construction on the samples along the developmental phases by applying PCA on the S matrix and by applying PCA on the shrunken S matrix.
- **Discussion**: How this can be used as part of linear regression framework.
- **Discussion**: Why variational framework was used (size of the data and speed of computation), alternate versions based on MCMC can also be recommended when the number of features to handle is small.
- **Discussion**: Other types of shrinkage methods that can be used (Global local shrinkage priors etc etc). Discuss the flexibility of the prior chosen.
- **Introduction**: Depends on what results we get, will be frame accordingly

2 Correlation Matrix Shrinkage Model

Let $X_{N\times J}$ be a data matrix where N denotes the number of samples and J denotes the number of features. Let $W = ((\rho_{ij}))_{i,j=1,2,\cdots,J}$ denote the population correlation matrix

$$W = E\left(X_{n*}X_{n*}^T\right) \forall n \tag{1}$$

Let us denote the sample correlation matrix by $R = ((r_{ij}))_{i,j=1,2,\cdots,J}$ where

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

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

$$s_{ij} = \frac{1}{n}\sum_{n=1}^{N}\left(x_{ni} - \bar{x}_i\right)\left(x_{nj} - \bar{x}_j\right)$$

$$s_{ii} = \frac{1}{n}\sum_{n=1}^{N}\left(x_{ni} - \bar{x}_i\right)^2$$

The sample correlation matrix R is widely used in various applied domains of statistics as an approximation of the unknown W and in cases where $n >> p$, the approximation works well. However, in recent times, we come across many example applications stemming from genetics, text mining, signal processing and many other applications, where $n << p$. In such cases, the sample correlation matrix no longer performs well as an estimate of the population correlation matrix. The biggest constraint for the sample correlation matrix is that it has rank n, which may be way smaller than the rank of the population correlation matrix. To solve this problem, shrinkage methods have been proposed in the literature to shrink the sample correlation matrix to the population correlation matrix. In this paper, we propose adaptive shrinkage techniques to shrink the sample correlation matrix to the population correlation matrix.

We propose the following model.

For any two features i and j with $i < j$, we define a binary size K latent variable vector $Z_{ij\star}$

$$Pr\left[Z_{ijk} = 1\right] = \pi_k \qquad\qquad j < i$$

$$Pr\left[Z|\pi\right] = \prod_{i=1}^{J}\prod_{j<i}\prod_{k=1}^{K}\pi_k^{Z_{ijk}}$$

We define a Dirichlet distribution prior for the parameter π.

$$Pr\left(\pi|\alpha\right) \propto \prod_{k=1}^{K}\pi_k^{\alpha_k - 1}$$

We define latent variables ρ_{ij}, such that

$$Pr\left(\rho|Z,\pi\right) \propto \prod_{i=1}^{J}\prod_{j<i}\prod_{k=1}^{K}\left[N\left(\rho_{ij} : 0, \sigma_k^2\right)\right]^{Z_{ijk}}$$

Then we define

$$Pr\left(\hat{\rho}|\rho\right) = \prod_{i=1}^{J}\prod_{j<i} N\left(\hat{\rho}_{ij}|\rho_{ij}, s_{ij}^2\right)$$

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

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

The model implies that we shrink the $\hat{\rho}_{ij}$ to 0 but the amount of shrinkage is decided globally by s_{ij}^2 and locally by the posterior distribution of $Z_{ijk}|\hat{\rho_{ij}}$ and the choices of the σ_k. We propose two different prior assumptions on the σ_k.

$$Pr\left(\hat{\rho}|Z,\pi\right) \propto \prod_{i=1}^{J}\prod_{j<i}\prod_{k=1}^{K}\left[N\left(\hat{\rho}_{ij} : 0, \sigma_k^2 + s_{ij}^2\right)\right]^{Z_{ijk}}$$

This reduces the problem to a single latent variable Z. The parameters of the model are π and σ_k. We present two approaches, one with a fixed grid of user defined values for σ_k and the other where we have a relatively flat prior on the σ_k.

The Fisher's z-transform is a variance stabilizing transform and if all the samples are independent, for n large, the asymptotic variance of $\hat{\rho}_{ij}$ would be equal to $s_{ij}^2 = s^2 = \frac{1}{n-3}$ for all i and j. We define a refined parameter

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

- **Model 1**: σ_k or ξ_k are fixed apriori, usually uniformly placed on a grid of values. Here we take the grid as in the adaptive shrinkage framework proposed in Stephens 2016 (ashr paper).

- **Model 2**: We introduce an additional layer of randomness by assuming a distribution on ξ_k. For the sake of technical advantage we assume an Inverse-Gamma distribution on ξ_k, and we choose the prior paramaters of the Inverse Gamma distribution such that it is very flat and allows for a wide range of possible values of ξ_k and hence σ_k. It must be emphasized here that ξ_k's are bounded below by $\frac{1}{n-3}$ while an Inverse gamma distribution has support from 0 to ∞. However we consider n to be moderately large, so $\frac{1}{n-3}$ would be a small number, which together with the flat shape of the inverse gamma prior will nullify the effect of the lower bound.

For model inference, we use a Variational EM algorithm. Let us consider the complete probability distribution involving the data, latent variables, parameters etc .

$$p(\hat{\rho}, Z, \pi, \xi) = p(\hat{\rho}|Z, \xi, \pi)p(Z|\pi)p(\pi)p(\xi)$$

$$\log p(\hat{\rho}, Z, \pi, \xi) = \log p(\hat{\rho}|Z, \xi, \pi) + \log p(Z|\pi) + \log p(\pi) + \log p(\xi)$$

We consider the mean field variational distribution over the latent variable Z and the parameters π and σ are given by

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

where $\xi = (\xi_1, \xi_2, \cdots, \xi_K)$.

$$\log q^{\star}(\pi) = E_{Z,\xi}\left[\log p(\hat{\rho}, Z, \pi, \xi)\right] \tag{2}$$

$$= E_Z\left[\sum_{k=1}^{K}(\alpha_k - 1)\log(\pi_k) + \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} z_{ijk}\log(\pi_k)\right] + const. \tag{3}$$

$$= \left[\sum_{k=1}^{K}(\alpha_k - 1)\log(\pi_k) + \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} E_Z(z_{ijk})\log(\pi_k)\right] + const \tag{4}$$

$$= \left[\sum_{k=1}^{K}(\alpha_k - 1)\log(\pi_k) + \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} \delta_{ijk}\log(\pi_k)\right] \tag{5}$$

$$= \sum_{k=1}^{K}\left[(\alpha_k - 1) + \sum_{i=1}^{J}\sum_{j<i}\delta_{ijk}\right]\log(\pi_k) \tag{6}$$

$$\tag{7}$$

So, we get

$$\pi \sim Dir\left(\pi|\beta_1, \beta_2, \cdots, \beta_K\right)$$

where $\beta_k = \alpha_k + \sum_{i=1}^{J}\sum_{j<i}\delta_{ijk}$. Now we determine the variational distribution of the latent variable Z.

$$\log q^{\star}(Z) = E_{\pi,\xi}\left[\log p(\hat{\rho}, Z, \pi, \xi)\right] \tag{8}$$

$$= E_{\pi,\xi}\left[\log p(Z|\pi) + \log p(\hat{\rho}|Z, \xi, \pi)\right] \tag{9}$$

$$= E_{\pi,\xi}\left[\sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} z_{ijk}\log(\pi_k) + \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} z_{ijk}\left[-\log(\xi_k) - 0.5\log(2\pi) - \frac{\hat{\rho}_{ij}^2}{2\xi_k}\right]\right] \tag{10}$$

$$= \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} z_{ijk}E_{\pi}\left(\log(\pi_k)\right) + \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} z_{ijk}\left[\frac{1}{2}E_{\xi}\left[\log\frac{1}{\xi_k}\right] - \frac{\hat{\rho}_{ij}^2}{2}E_{\xi}\left[\frac{1}{\xi}\right]\right] \tag{11}$$

$$\tag{12}$$

If $\xi_k \sim Inv - Gamma(a, b)$, then $\frac{1}{\xi_k} \sim Gamma(a, b)$, and it can be shown that

$$E_{\xi_k}\left[\log\frac{1}{\xi_k}\right] = E_U\left[\log(U)\right]$$

where $U \sim Gamma(a, b)$. It can be checked that

$$E_U\left[\log(U)\right] = -\log(b) + \psi(a)$$

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

$$E_{\xi_k}\left[\frac{1}{\xi_k}\right] = \frac{\nu_{1k}}{\nu_{2k}}$$

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

$$\log q^{\star}(Z) \propto \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} z_{ijk}\left\{\psi(\beta_k) - \psi(\sum_{l=1}^{K}\beta_l) + 0.5 \times (\psi(\nu_{1k}) - \log(\nu_{2k})) - \frac{\hat{\rho}_{ij}^2}{2}\frac{\nu_{1k}}{\nu_{2k}}\right\}$$

$$q^{\star}(Z) = \prod_{i=1}^{J}\prod_{j<i}\prod_{k=1}^{K}\delta_{ijk}^{Z_{ijk}}$$

where

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

Finally we derive the variational distribution for ξ_k. In the mean field definition, we assume that the variational distribution of all the ξ_k are independent.

$$\begin{aligned}
\log q^{\star}(\xi_k) &= E_{Z,\pi,\xi\neq k}\left[\log p(\hat{\rho}, Z, \pi, \sigma^2)\right] && (13)\\
&= E_{Z,\pi,\xi\neq k}\left[\log p(\hat{\rho}|Z, \sigma, \pi) + \log p(\sigma^2)\right] && (14)\\
&= -\sum_{i=1}^{J}\sum_{j<i} E_Z(Z_{ijk})\left\{0.5 * \log(\xi_k) + \frac{\hat{\rho_{ij}}^2}{2\xi_k}\right\} - (a+1)log(\xi_k) - \frac{b}{\xi_k} && (15)\\
&= -log(\xi_k)\left\{(a+1) + 0.5 \times \sum_{i=1}^{J}\sum_{j<i}\delta_{ijk}\right\} - \frac{0.5 \times \sum_{i=1}^{J}\sum_{j<i}\delta_{ijk}\hat{\rho}_{ij}^2 + b}{\xi_k} && (16)\\
& && (17)
\end{aligned}$$

So, we get

$$q^{\star}(\xi_k) \sim Inv - Gamma\left(a + 0.5 \times \sum_{i=1}^{J}\sum_{j<i}\delta_{ijk}, 0.5 \times \sum_{i=1}^{J}\sum_{j<i}\delta_{ijk}\hat{\rho}_{nj}^2 + b\right)$$

$$\nu_{1k} = a + 0.5 \times \sum_{i=1}^{J}\sum_{j<i}\delta_{ijk}$$

$$\nu_{2k} = 0.5 \times \sum_{i=1}^{J}\sum_{j<i}\delta_{ijk}\hat{\rho}_{nj}^2 + b$$