CorShrink: A random multi-target adaptive shrinkage method for correlation matrix estimation

Abstract

Shrinkage estimation of correlation matrices provides well conditioned and more accurate estimators of the population correlation matrix, particularly in 'small n, large p' contexts. In this paper, we propose a multi target shrinkage method called *CorShrink* where the targets are noisy versions of identity matrix and which learns adaptively the optimal amount of shrinkage to these targets from the data. For different (n, p) scenarios, we compare the performance of our method with respect to the standard approaches. We also provide a demonstration of how it can be used to understand and interpret gene-gene correlation structure in single cell RNA-seq data. The methods are implemented in R package **CorShrink** available from Github (`https://github.com/kkdey/CorShrink`).

Keywords: shrinkage, covariance matrix estimation, variational EM, single cell RNA-seq

1 Introduction

Estimating the covariance or correlation matrix of variables is a common practice among researchers interested in a broad spectrum of statistical applications, ranging from understanding the relationship among variables, performing classification or regression and even form groups or clusters or features. The most common choice for an estimator is the sample covariance or correlation matrix which is also the Maximum Likelihood estimate. While this estimator works well for $n > p$ scenarios, it has extremely high approximation error due to its low rank structure when $n << p$.

In 2003, Ledoit and Wolf proposed an estimator that is well conditioned and has much lesser approximation error than the sample correlation matrix, especially under $n << p$ scenarios [7] [8]. This approach was further generalized by Schäfer and Strimmer (2005) [10] [11], who besides proposing new shrinkage estimators, also provided analytic calculation of the optimal shrinkage intensity. The idea was to fit 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 minimize the expected error of approximation of the shrunken estimate. The above methods used a single target for shrinking, but a multi-target covariance shrinkage approach was recently proposed - see Lancewicki and Aladjem (2014) [6].

In this paper, we propose three versions of an alternative method called *CorShrink* which assumes multiple targets T_1, T_2, $\cdots$, T_K, all of which are noisy versions of the identity matrix and the noise variation increases with each $k = 1, 2, \cdots, K$. We adaptively determine the amount of shrinkage by optimally determining the shrinkage weights for each target and selecting the set of targets to cover the range of variation of the data well. We explain the noise structure and the model fit in greater details in the next section. We also perform comparisons of our model performance with respect to the Schäfer and Strimmer approach (Schäfer and Strimmer (2005) [10]) and the Graphical LASSO algorithm developed by Friedman et al (2008) [5] for sparse representation of the correlation and primarily inverse correlation matrices used in building causal networks. We show that *CorShrink* performs better than the Schäfer and Strimmer shrinkage in terms of flexibility in choosing the amount

of shrinkage when $n << p$, and both *CorShrink* and Schäfer-Strimmer method perform much better as correlation shrinkage methods compared to GLASSO. We also show an application our method on a single cell RNA-seq data from mouse pre-implantation embryos due to Deng et al (2014) [3].

2 Methods and Materials

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

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

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.

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

We define latent variables ρ_{ij}, such that they are normally distributed centered around 0.

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

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

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

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

The assumption of mixture normal distribution of ρ_{ij} with each component centered around 0 implies that the population correlation matrix would be a mixture too and each component would be centered around an identity matrix with some noise variation determined by $\sigma_k, k = 1,2,\cdots,K$.

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_k values, selected such that it covers the span of the variation of the data well. Here we propose to use a similar grid for σ_k values as suggested in Stephens (2016) [12] for modeling false discovery rates. We add a component $\sigma_k = 0$ 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_k 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 with different choices, we finally settled with the default Dirichlet prior as $Dir(10,1,,1,\cdots,1)$.
- *CorShrink-VEM2*: We no longer assume the σ_k values to be fixed. Rather we assume that they come from an Inverse-Gamma distribution. We assume $Inv-Gamma(\varepsilon,\varepsilon)$ distributions which are relatively non-informative in order to make the choice of σ more flexible. For our applications in this paper, we 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 (2016) [12], which fits an EM algorithm. For the *CorShrink-VEM* and *CorShrink-VEM2* models, we used Mean Field Variational EM approach to estimate the model parameters. Variational methods are faster than MCMC methods as they often provide analytic updates of the parameters over iterations, thereby ensuring faster computation (see Beal and Ghahramani (2003) [1] and Blei et al (2016) [2].

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

$$\xi_k = \sigma_k^2 + \frac{1}{n-3} \tag{6}$$

Suppose the priors on π and ξ are

$$\pi \sim Dir\,(\alpha_1,\alpha_2,\cdots,\alpha_K) \qquad\qquad \xi_k \sim Inv-Gamma(a,b)\ \ \forall\, k \tag{7}$$

Here $(\alpha_1,\alpha_2,\cdots,\alpha_K) = (10,1,1,\cdots,1)$ and $a = b = 0.01$. Next we define the mean field variational distribution on the latent variable Z and the parameters π and $\xi_1,\xi_2,\cdots,\xi_K$ as follows.

$$q(Z,\pi,\xi) = q(Z)q(\pi)\prod_{k=1}^{K} q(\xi_k) \tag{8}$$

Then the mean field distribution for π is given by

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

$$= E_Z\left[\sum_{k=1}^{K}(\alpha_k-1)\log(\pi_k) + \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} z_{ij:k}\log(\pi_k)\right] + const. \tag{10}$$

$$= \sum_{k=1}^{K}\left[(\alpha_k-1) + \sum_{i=1}^{J}\sum_{j<i}\delta_{ij:k}\right]\log(\pi_k) \tag{11}$$

$$\tag{12}$$

So the variational distribution for π is of the form

$$\pi \sim Dir(\pi|\beta_1, \beta_2, \cdots, \beta_K) \qquad \beta_k = \alpha_k + \sum_{i=1}^{J} \sum_{j<i} \delta_{ij:k} \tag{13}$$

The variational distribution of the latent variable Z is obtained similarly

$$\begin{aligned}
\log q^\star(Z) &= E_{\pi,\xi}\left[\log p(\hat{\rho}, Z, \pi, \xi)\right] && (14)\\
&= E_{\pi,\xi}\left[\log p(Z|\pi) + \log p(\hat{\rho}|Z, \xi, \pi)\right] && (15)\\
&= \sum_{i=1}^{P}\sum_{j<i}\sum_{k=1}^{K} z_{ij:k} E_\pi\left(\log(\pi_k)\right) + \sum_{i=1}^{P}\sum_{j<i}\sum_{k=1}^{K} z_{ij:k}\left[\frac{1}{2}E_\xi\left[\log\frac{1}{\xi_k}\right] - \frac{\hat{\rho}_{ij}^2}{2}E_\xi\left[\frac{1}{\xi}\right]\right] && (16)\\
& && (17)
\end{aligned}$$

It can be shown that

$$E_{\xi_k}\left[\log\frac{1}{\xi_k}\right] = -\log(\nu_{2k}) + \psi(\nu_{1k}) \qquad E_{\xi_k}\left[\frac{1}{\xi_k}\right] = \frac{\nu_{1k}}{\nu_{2k}} \tag{18}$$

$$E_\pi\left(\log(\pi_k)\right) = \psi(\beta_k) - \psi(\sum_{l=1}^{K}\beta_l) \tag{19}$$

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

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

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

For *CorShrink-VEM* model, the σ_k or $\xi_k = \sigma_k^2 + \frac{1}{n-3}$ are fixed and the variational distribution is of the form

$$q(Z, \pi) = q(Z)q(\pi) \tag{22}$$

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

$$q^\star(Z) = \prod_{i=1}^{P}\prod_{j<i}\prod_{k=1}^{K} \delta_{ij:k}^{Z_{ij:k}} \qquad \delta_{ij:k} \propto \exp\left(\left\{\psi(\beta_k) - \psi(\sum_{l=1}^{K}\beta_l) + 0.5 \times \left(\log\frac{1}{\xi_k}\right) - \frac{\hat{\rho}_{ij}^2}{2}\frac{1}{\xi_k}\right\}\right) \tag{23}$$

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 continue updating the parameters until convergence.

It must be stressed that the normal assumption in Equation 4 is reasonable only for moderately large n. Also, ξ_k's are bounded below by $\frac{1}{n-3}$ which we ignore in defining an Inverse Gamma distribution on the $\xi_k, k = 1, 2, \cdots, K$. Therefore we do not recommend our models, especially the *CorShrink-VEM2* model, for very small sample size n. Having said that, the initialization using *CorShrink-VEM* fixes the ξ_k initial values to be $> \frac{1}{n-3}$ by fixing the grid of σ_k and as a result, the final estimates usually adjust themselves automatically to the lower bound. In case they violate, we forcibly set them to the lower bound value.

In the next section, we discuss applications of the three *CorShrink* models on simulated data and a single cell RNA-seq data with cells drawn from from mouse pre-implantation embryo due to Deng et al (2014) [3].

3 Results

3.1 Simulation Results

We begin by illustrating the performance of the *CorShrink* method on simulated data. We tested our method on randomly generated covariance matrices of varying dimensions and also on structured matrices- for e.g., block diagonal matrices. We first demonstrate results from applying our models on a block diagonal matrix of dimension 100. We assumed the block diagonal matrix to be of the form

$$\begin{pmatrix} 0.2\mathbf{I} + 0.8\mathbf{e}\mathbf{e}^T & 0 \\ 0 & 0.7\mathbf{I} + 0.3\mathbf{e}\mathbf{e}^T \end{pmatrix} \tag{24}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{e}$ is the vector of all 1's. We considered four choices of number of samples drawn, $n = 5, 10, 50, 200$ spanning all three scenarios - $n << p$, $n < p$ and $n > p$. We then estimated the correlation matrix using three versions of *CorShrink*, GLASSO at four different regularization parameter values ranging from low to high shrinkage and the Schäfer-Strimmer method (see Friedman et al (2008) [5], Witten et al (2010) [13] and Schäfer-Strimmer (2005) [10]).

The image plots in Figure 1 (a) show that sample correlation fails to approximate the population correlation matrix well in regions of low correlation. The Schäfer-Strimmer method, in trying to shrink towards the identity matrix, shrinks the correlations in high correlation regions excessively. The GLASSO method for 0.05 is very close to the sample covariance matrix, whereas for the other choices, it tends to shrink the correlations too much, demonstrating the sensitivity of GLASSO to the choice of the regularization parameter. Due to the flexibility of the *CorShrink* methods to choose a target out of a range of possible targets (all of which are noisy versions of identity matrix), it adaptively learns to shrink the correlations. As a result, they enforce more shrinkage on low correlation regions and less shrinkage on the high correlation values compared to the Schäfer-Strimmer method and the GLASSO methods. In terms of eigenvalue comparisons in Figure 1 (b), *CorShrink* and Schäfer-Strimmer shrinkage methods act similarly at the top eigenvalues, but at

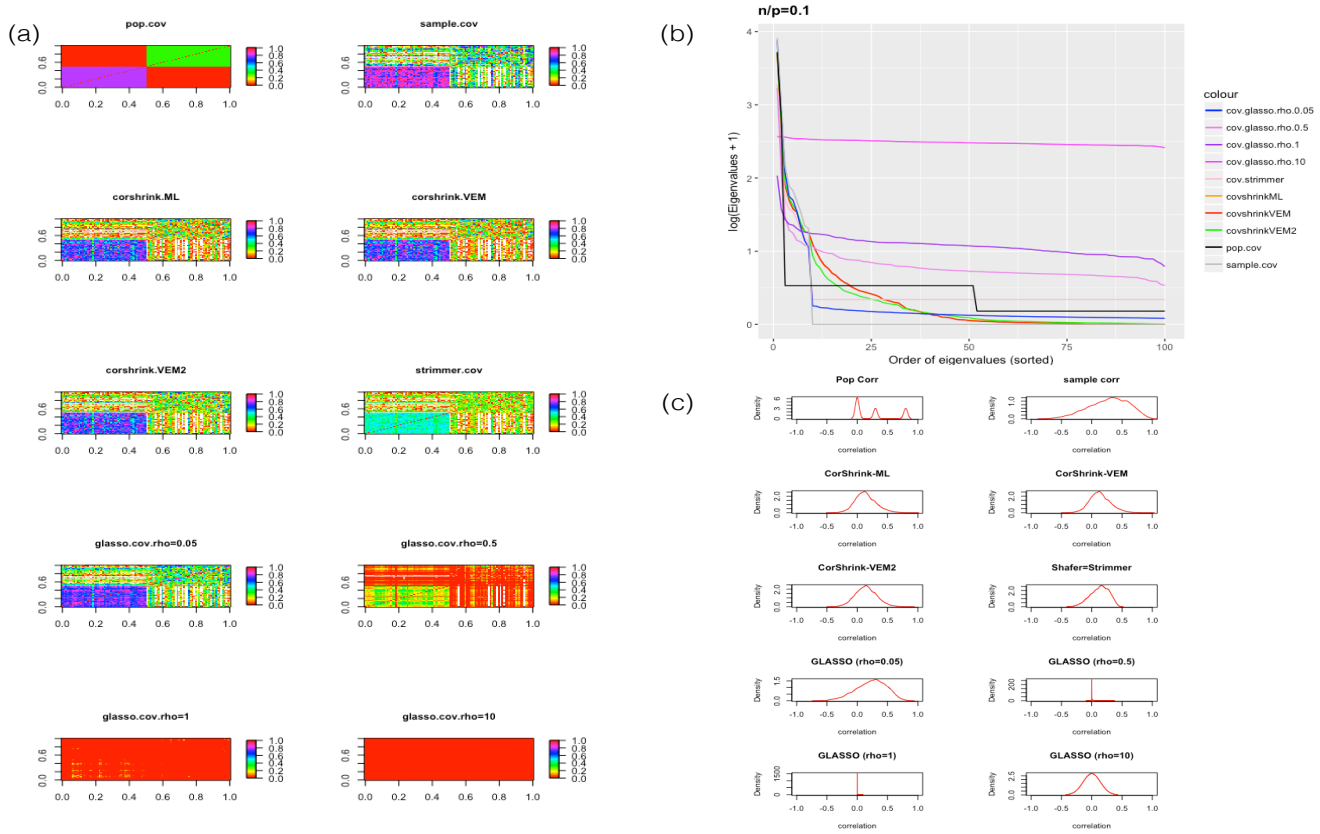

Figure 1: Corresponding to 3 versions of *CorShrink*, GLASSO under 4 values of regularization parameters, sample correlation and the population correlation, we show the image plots of the correlation structure in (a). In (b), we display the eigenvalue trends ordered from highest to lowest values for different shrinkage methods and we check how they relate to the original trend for the population correlation matrix. (c) We plot the distributional patterns of the estimated correlations from the different shrinkage methods and compare them with that of the sample correlation and the population correlation

the lower eigenvalues, the eigenvalue trends become flat for the Schäfer-Strimmer shrinkage, while that of a *CorShrink* model forms more of an asymptote. Figure 1 (c) shows the distributional profile of the correlation estimates from the different shrinkage methods. Expectedly the distributions for the shrinkage methods are more concentrated around 0 compared to sample correlation matrix. While all the methods fail to detect the three peaks of the original matrix, *CorShrink* and Schäfer-Strimmer methods show a small bump in the distribution in the region of second mode of the original matrix. The peak extraction power is much better when n is not too small compared to p. To see this, the reader can check the correlation distribution plot for $n = 50$ and $p = 100$ [see `https://github.com/kkdey/CorShrink-paper/blob/master/figures/correlation_distribution/correlation_distribution_0_5.png`].

In the above simulation study, we had a specific block structure to the covariance matrix which resulted in sharp falls initially and flat regions subsequently in the eigenvalue trends. To make a more general comparison, we randomly generated a covariance matrix (using **clusterGeneration** package by Qiu and Joe (2015) [9]) and estimated the corresponding population correlation matrix via different methods . The results are illustrated in Fig 2 for $n << p$ scenarios. The *CorShrink* eigenvalue trends seem to follow the population eigenvalue trends closely. Noticeably the top eigenvalues from our approach and the Schäfer-Strimmer approach are much closer to the top eigenvalues of the original matrix when compared to estimates from the GLASSO or the sample

correlation matrix. For GLASSO, it seems that increasing the regularization for low n does not affect the top eigenvalues as much as it affects the lower order eigenvalues. This is important since it is usually the top eigenvalues that are of primary interest to researchers in analyzing the structure of dependence.

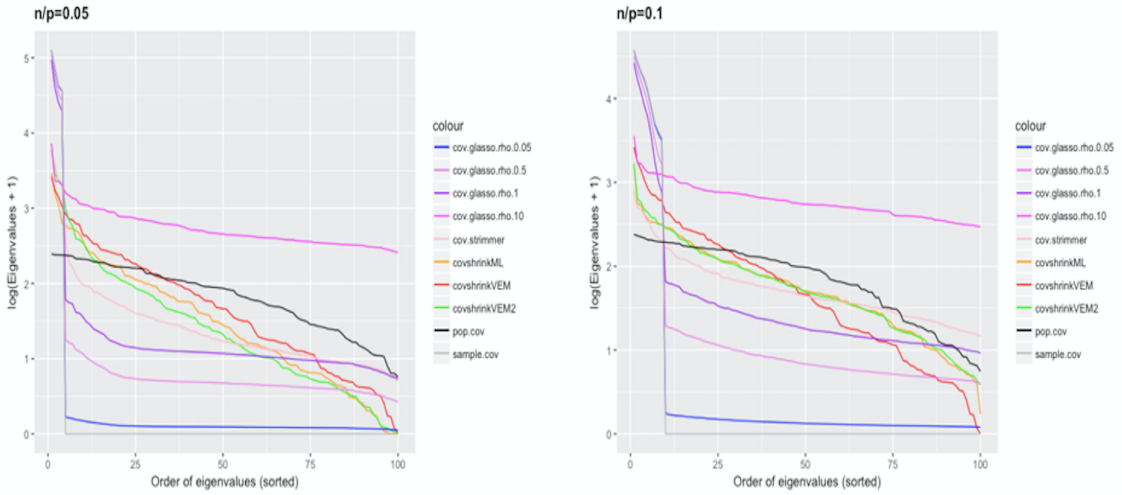

Figure 2: The eigenvalue trends from highest to lowest eigenvalue for $n = 5, p = 100$ and $n = 10, p = 100$ for a randomly generated covariance matrix with a smoothly decaying trend in eigenvalues from highest to lowest.

3.2 Real data application

Deng et al (2014) [3] collected single-cell expression data of mouse preimplantation embryos from the zygote to blastocyst stage, with a number cells sequenced at each stage. In total, 259 cells were sequenced and the RNA-seq data was collected over 22431 genes. Transcription factor genes were found to play an important role in clustering the developmental phases from a previous study by Dey et al (2016) [4]. 1281 genes from the Deng et al data matched with the Transcription Factor database at `http://www.bioguo.org/AnimalTFDB/species.php?spe=Mus_musculus` (see Zhang et al (2012) [15]). We applied the *CorShrink-ML* method both ways - on cells and on genes with the aim to determine cell-cell or gene-gene grouping patterns in the data.

We first apply our methods on the covariance matrix of the 259 cells. In this case, the genes may be treated as samples, but they are correlated due to having same transcription factor or being part of the same pathway and this makes it hard to decide on the effective number of independent samples n to use in the model for shrinking correlations. We performed PCA and applied broken stick model to determine first 9 PCs as signal and used $n = 9$ in our model. Figure 3 shows the image plots of the sample correlation matrix and the *CorShrink-ML* method along with the PC1 vs PC2 plots applied to the estimated matrices respectively. The image plot for the *CorShrink-ML* estimate appears less noisy than the sample correlation image plot and the blocks of (8cell,16 cell) and the blastocysts are more easily discernible in our case. For the PCA plot, the distinction between the cell development phases is still discernible after shrinkage but expectedly, the relative distances between the phases has shrunk. The results for the *CorShrink-VEM* and *CorShrink-VEM2* were very similar to that of *CorShrink-ML* method.

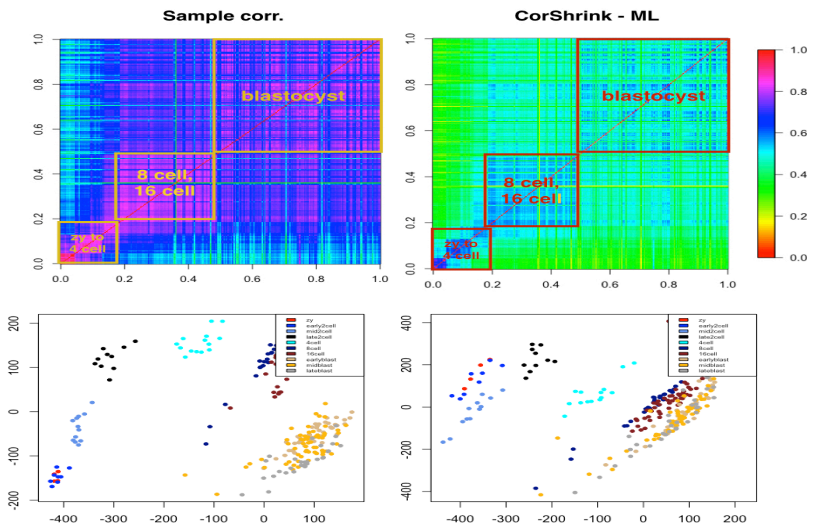

Figure 3: The image plot and the PC1 vs PC2 plot of the sample correlation matrix and the estimated correlation matrix obtained using *CorShrink-ML* method.

A more interesting application of our model lies in estimating the gene-gene correlation matrix from the 259 samples which corresponds to $n << p$ scenario. To do so, we first eliminated genes which have shown non-zero expression in less than 10 samples. This was aimed at removing expression spiking bias due to sequencing or library preparation biases. After filtering, we were left with 933 genes on which we applied the **CorShrink-ML** model. From Figure 4, we find that **CorShrink-ML** tends to shrink the low to medium correlations significantly while retaining the very strong correlations. We observed the pattern of gene expression across samples for the genes that appear highly correlated among each other in the heatmap (corresponding to upper right corner of the heatmap in Figure 4). We also performed functional annotation enrichment for these genes relative to all other genes that formed the background. The top GO annotations enriched in these genes corresponded to hippocampus (GO:0021766), limbic system(GO:0021761) and hind brain development(GO:0030902) (see the full list of enriched genes at `https://github.com/kkdey/CorShrink-paper/blob/master/utilities/high_cor_genes_deng_tf.txt`).

4 Discussion

Our goal here is to highlight the potential of the *CorShrink* models as covariance or correlation shrinkage methods which has more flexibility in choosing the amount of shrinkage for the high and low sample correlations, compared to the Schäfer Strimmer shrinkage approach (see Figure 1). Also our methods outperform GLASSO as a covariance shrinkage estimator irrespective of the choice of regularization parameter used for the latter (see Figure 1 and Figure 2).

In terms of computational time, the *CorShrink-ML* method is faster than the *CorShrink-VEM* and *CorShrink-VEM2* methods. For instance, the time taken to run *CorShrink-ML*, *CorShrink-VEM* and *CorShrink-VEM2* on the 259 samples were 9 seconds, 44 seconds and 3.1minutes respectively.

This method also opens other areas of applications and extensions that we intend to pursue in future. The *CorShrink* models can be easily extended to partial correlation and partial covariance matrices and also facilitate efficient computation of the inverse correlation and covariance matrices. One can

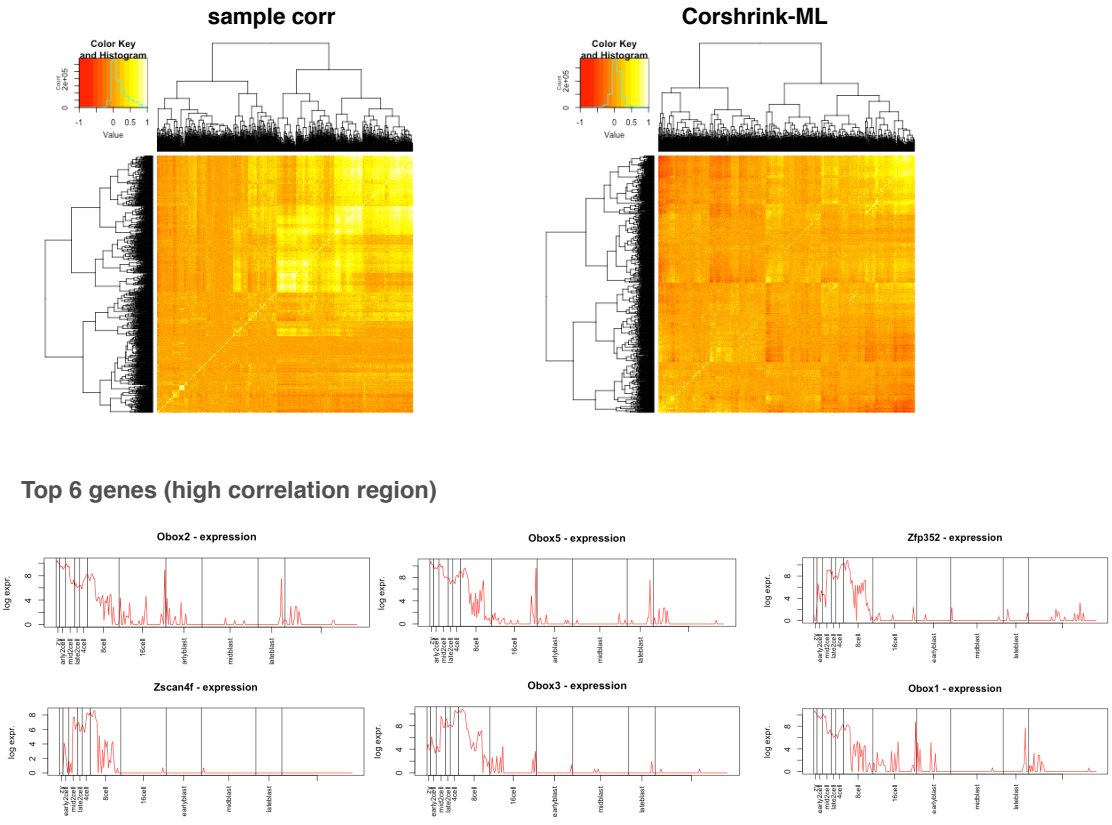

Figure 4: The heatmap representation of the sample correlation matrix and the *CorShrink-ML* estimated correlation matrix for the TF genes in Deng et al 2014 data [3]. We investigated the high correlation region at the top right corner of the heatmap and extracted top 6 most highly correlated genes. All of them showed decreasing trends of log expression from early stages of development (zygote, 2 cell) to the late stages.

create causal networks based on the *CorShrink* models and compare them to the GLASSO based causal networks. Additionally, one can combine Linear Discriminant Analysis and Multiple regression problem with the estimated covariance matrices obtained from the *CorShrink* models, in the same way the Schäfer Strimmer method has been used in these domains [14] [10]. Also for structured covariance matrices, one would want to pool the knowledge of the underlying structure into the shrinkage method. For example, for a block covariance matrix, it makes more sense to apply *CorShrink* separately on each block and pool the blocks together.

The codes to fit the *CorShrink* models on data are implemented in an R package **CorShrink** which is available on Github at `https://github.com/kkdey/CorShrink`. It also contains a README demonstrating how these models were fitted on simulated data. The Deng et al single cell data [3] is available as a R data package with instructions for downloading and loading into R at `https://github.com/kkdey/singleCellRNASeqMouseDeng2014`. The scripts to recreate the results and the figures are available in the Github repo `https://github.com/kkdey/CorShrink-paper`.

References

[1] Beal MJ, Ghahramani Z. The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures *Bayesian Statistics*, 7.

[2] Blei DM, Kucukelbir A, McAuliffe JD. 2016. Variational Inference: A Review for Statisticians. https://arxiv.org/pdf/1601.00670.

[3] Deng Q, Ramskold D, Reinius B, Sandberg R. 2014. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science*. 343 (6167) 193-196.

[4] Dey KK, Hsiao CJ, Stephens M. 2016. Clustering RNA-seq expression data using grade of membership models *http://biorxiv.org/content/early/2016/05/03/051631*

[5] Friedman J, Hastie T, Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 9.3.

[6] Lancewicki T. and Aladjem M. 2014. Multi-Target Shrinkage Estimation for Covariance Matrices. *IEEE Transactions on Signal Processing*. 62 (24), 6380-6390

[7] Ledoit O. and Wolf M. 2003. "Improved estimation of the covariance matrix of stock returns with an application to portofolio selection. *Journal of Empirical Finance*. 10 (5): 603?621.

[8] Ledoit O. and Wolf M. 2004. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*. 30 (4): 110?119.

[9] Qiu W, Joe H. 2015. clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). R package version 1.3.4.

[10] Schäfer J and Strimmer K. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol*.4.32.

[11] Schäfer J and Strimmer K. 2005. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*. 21: 754-764.

[12] Stephens M. 2016. False discovery rates: a new deal. *Biostatistics* Advance Access.

[13] Witten DM, Friedman JH, Simon N. 2010. New Insights and Faster Computations for the Graphical Lasso. *Journal of Computational and Graphical Statistics*, 20, 4, 892?900.

[14] Xu P., Brock GN, Parrish RS. 2009. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics & Data Analysis*. 53.5.

[15] Zhang HM, Chen H, Liu W, Liu H, Gong J, Wang H and Guo AY. (2012 database issue) *Nucl. Acids Res*. doi: 10.1093/nar/gkr965