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

Kushal K Dey
University of Chicago, Dept. of Statistics,
5220 S Kenwood Ave,
Chicago, IL, 60615

Abstract

Shrinkage estimation of correlation matrices provides well conditioned and more accurate estimators of the population correlation matrix, specially 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 from the data. We compare the performance of our method under different (n, p) scenarios with respect to the standard approaches and also provides a demonstration of the performance of our approach on a single cell RNA-seq data from mouse pre-implantation and we show the method indeed extracts useful biological features from the data. The methods are implemented in a Github R package **CorShrink**.

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

1 Introduction

Estimation the covariance or correlation matrix of the variables is a common practice for researchers interested in a broad spectrum of statistical applications, ranging from understanding the relationship among variables, perform classification or regression and even form groups or clusters or features. The most common choice of an estimator is the sample covariance or correlation matrix which is also the Maximum Likelihood estimate. While this estimator works fine for $n > p$ cases, it has extremely high approximation error with respect to the population covariance/correlation matrix 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 [1] [2] in particular under $n << p$ scenarios. This approach was further developed and generalized by Schäfer and Strimmer, 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 papers used a single target for shrinking, but a multi-target covariance shrinkage approach was recently proposed - see Lancewicki and Aladjem 2014 [3].

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. We adaptively determine the amount of shrinkage by optimally determining the shrinkage weights for each target and assuming that the set of targets cover the range of variation of the data well. We will discuss about the noise structure and the model fit in more details in the Methods section. We also perform comparisons of our model performance with respect to the Schäfer and Strimmer approach and the Graphical LASSO algorithm developed by Friedman et al [9] for sparse representation of the correlation and primarily inverse correlation matrices used for building causal networks. We show that *CorShrink* performs marginally better than the Schäfer and Strimmer shrinkage in terms of eigenspace approximation to the population covariance when $n << p$, and both *CorShrink* and Shafer-Strimmer method perform much better as correlation shrinkage methods compared to GLASSO. We also show an application our method on a single cell mouse pre-implantation RNA-seq data due to Deng et al. 2014 [7].

2 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 shrunken 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, $\cdots$, 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 $R = ((r_{ij}))_{i,j=1,2,\cdots,P}$, P being the number of features, calculated over N data samples.

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

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

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

We assume Normal distribution for the ρ,

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

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 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 [6] 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 $Dir(10,1,,1,\cdots,1)$.
- *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 [6], 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 [12] [**?**].

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

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

and then 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.

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

Then the mean field distribution for π is given by

$$\begin{aligned}
\log q^{\star}(\pi) &= E_{Z,\xi}\left[\log p(\hat{\rho},Z,\pi,\xi)\right] && (6)\\
&= E_Z\left[\sum_{k=1}^{K}(\alpha_k-1)\log(\pi_k) + \sum_{i=1}^{J}\sum_{j<i}\sum_{k=1}^{K} z_{ij:k}\log(\pi_k)\right] + const. && (7)\\
&= \sum_{k=1}^{K}\left[(\alpha_k-1) + \sum_{i=1}^{J}\sum_{j<i}\delta_{ij:k}\right]\log(\pi_k) && (8)\\
& && (9)
\end{aligned}$$

So the variational distribution for π is of the form

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

The variational distribution of the latent variable Z is obtained similarly

$$\begin{aligned}
\log q^{\star}(Z) &= E_{\pi,\xi}\left[\log p(\hat{\rho},Z,\pi,\xi)\right] && (11)\\
&= E_{\pi,\xi}\left[\log p(Z|\pi) + \log p(\hat{\rho}|Z,\xi,\pi)\right] && (12)\\
&= \sum_{i=1}^{P}\sum_{j<i}\sum_{k=1}^{K} z_{ij:k}E_{\pi}\left(\log(\pi_k)\right) + \sum_{i=1}^{P}\sum_{j<i}\sum_{k=1}^{K} z_{ij:k}\left[\frac{1}{2}E_{\xi}\left[\log\frac{1}{\xi_k}\right] - \frac{\hat{\rho}_{ij}^2}{2}E_{\xi}\left[\frac{1}{\xi}\right]\right] && (13)\\
& && (14)
\end{aligned}$$

It can be shown that

$$E_{\xi_k}\left[\log\frac{1}{\xi_k}\right] = -\log(\nu_{2k}) + \psi(\nu_{1k}) \qquad E_{\xi_k}\left[\frac{1}{\xi_k}\right] = \frac{\nu_{1k}}{\nu_{2k}} \qquad E_{\pi}\left(\log(\pi_k)\right) = \psi(\beta_k) - \psi(\sum_{l=1}^{K}\beta_l) \qquad (15)$$

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

$$q^{\star}(Z) = \prod_{i=1}^{P}\prod_{j<i}\prod_{k=1}^{K}\delta_{ij:k}^{Z_{ij:k}} \qquad \delta_{ij:k} \propto \exp\left(\left\{\psi(\beta_k) - \psi(\sum_{l=1}^{K}\beta_l) + 0.5\times\left(\psi(\nu_{1k}) - \log(\nu_{2k})\right) - \frac{\hat{\rho}_{ij}^2}{2}\frac{\nu_{1k}}{\nu_{2k}}\right\}\right) \qquad (16)$$

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

$$q(Z, \pi) = q(Z) q(\pi) \tag{17}$$

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

$$q^{\star}(Z) = \prod_{i=1}^{P} \prod_{j<i} \prod_{k=1}^{K} \delta_{ij:k}^{Z_{ij:k}} \qquad \delta_{ij:k} \propto \exp\left(\left\{ \psi(\beta_k) - \psi(\sum_{l=1}^{K} \beta_l) + 0.5 \times \left(\log \frac{1}{\xi_k} \right) - \frac{\hat{\rho}_{ij}^2}{2} \frac{1}{\xi_k} \right\} \right) \tag{18}$$

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.

3 Results

3.1 Simulation Results

We begin by illustrating the performance of the *CorShrink* method on simulated data. We randomly generated covariance matrices of dimensions p varying in the range 100, 500 and 1000 and then generated n samples from a Multivariate Normal distribution centered at 0 and with the above generated covariance matrix. We considered four choices of $n = 5, 10, 50, 200$ that spans all three scenarios - $n << p$, $n < p$ and $n > p$. We then performed shrinkage on the sample covariance matrix using the three versions of *CorShrink*, GLASSO at 4 different regularization parameter values ranging from low to high shrinkage and the Schäfer-Strimmer method [9] [10] [4].

In Figure 1, we compare the eigenvalues of the shrunk covariance matrices under different shrinkage schemes and different choices of n for $p = 100$. Results corresponding to other choices of p can be checked here [?Link]. Figure 1 shows that the trends of the eigenvalues from the *CorShrink-ML*, *CorShrink-VEM2* and Schäfer-Strimmer shrinkage methods are consistently close to the population eigenvalues across all four choices of n. The *CorShrink-VEM* version performs well for the $n << p$ scenarios, however its performance is not so good for moderate to large values of n. This is probably due to the fact that for larger data, the strong weight of the Dirichlet hyperprior on the null component and the fixed grid of component variances of underlying mixture model on Fisher Z-scores makes the model inflexible to adapt itself to data, a problem that is solved in VEM2 when the component variances are more flexibly chosen by the model. GLASSO for low shrinkage (regularization parameter $\rho = 0.05$) is very close to the sample covariance matrix and for high shrinkage ($\rho = 10$) provides a matrix close to diagonal and therefore is a bad fit to the population covariance matrix. The $\rho = 0.5$ or $\rho = 1$ provide slightly better fit to the population covariance matrix in terms of eigenvalue patterns, but noticeably, for $n << p$ cases, the top eigenvalues (with the highest magnitude) remain very close to that of the sample covariance matrix despite increasing the level of shrinkage and it is the lower order eigenvalues that adapt more rapidly with increasing ρ. Here we must emphasize that usually, the top few eigenvalues are of principal interest to researchers interested in lower dimensional representation, and under $n << p$ scenario, the three versions of the *CorShrink* approach and the Schäfer-Strimmer shrinkage method are more effective than GLASSO in mapping the top eigenvalues close to the ones from the population covariance matrix.

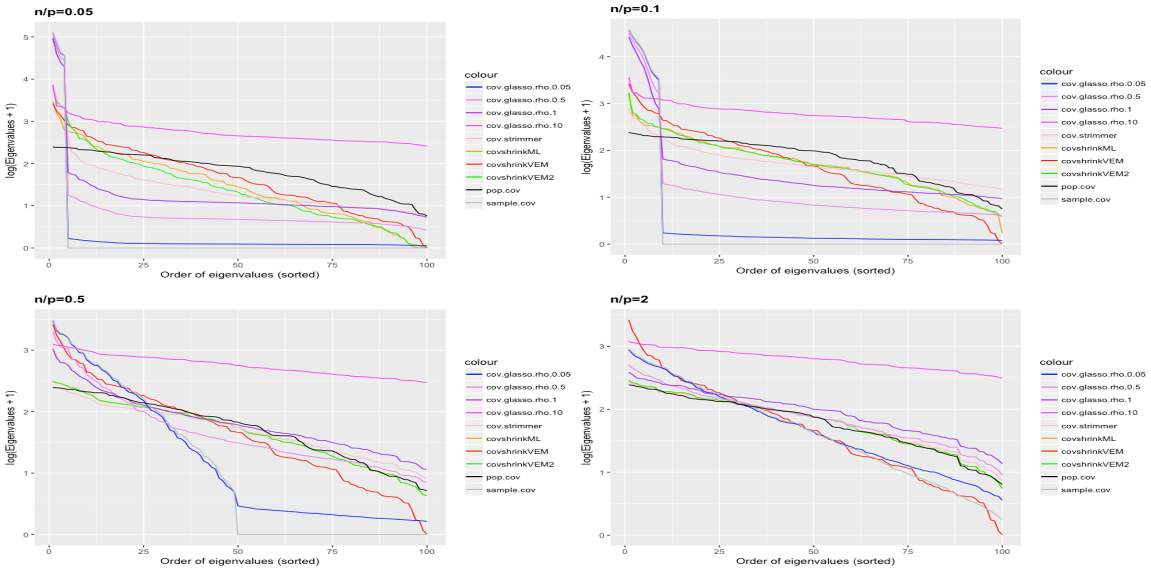

Figure 1: Distribution of the sorted eigenvalues for the shrunk covariance matrices due to the three versions of *CorShrink*- namely *CorShrinkML*, *CorShrinkVEM* and *CorShrinkVEM2*, three versions of GLASSO for three regularization parameters ρ, varying from 0.05, 0.5 and 1 [low to high shrinkage] and the Schäfer-Strimmer shrinkage method, along the distributions of the eigenvalues for the sample covariance and the population covariance matrices.

Besides the eigenvalue trends and the top few eigenvalues, another important consideration in comparing these shrinkage methods is how close the eigenvectors from the shrunk covariance matrices are with respect to the population covariance matrix. Table **??** present the average distance between the top 5 eigenvectors of the each shrinkage method with respect to the population covariance. Again we find that the Schäfer-Strimmer, *CorShrink-ML* and *CorShrink-VEM2* produce shrunk covariance matrices closest to the population covariance matrix in terms of the eigen-spaces corresponding to the top eigenvalues. In Figure 2, we plot the distribution of the correlations from the shrunk matrices obtained using different shrinkage methods . We observed that the distribution is more concentrated around 0 for the *CorShrink* models when compared to the GLASSO and Schäfer-Strimmer methods. Additionally, *CorShrink* retains some correlation values with large magnitudes. This characteristic of the *CorShrink* approach would ensure a sparse representation when used in building correlation networks.

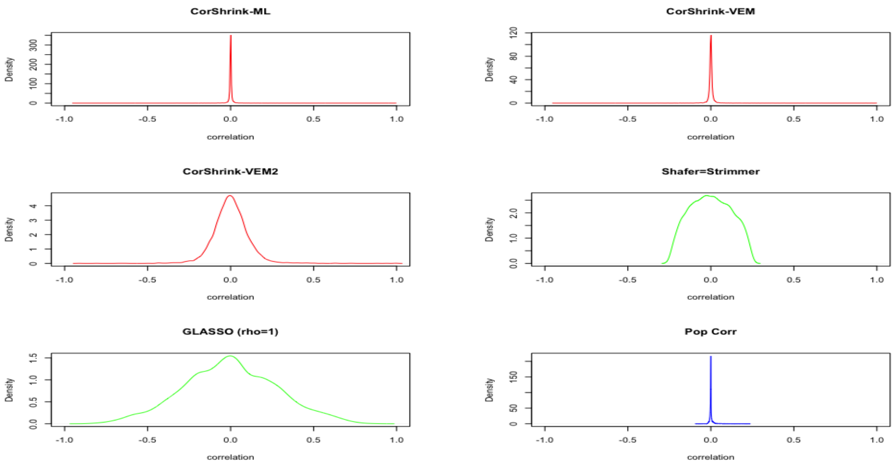

Figure 2: Overall distribution of the correlations after shrinking the covariance matrix using the three models of *CorShrink*, Schäfer-Strimmer method and the GLASSO method for regularization parameter of 1 (which was found to be the best fit in terms of eigenvalue patterns out of the four choices considered). The correlation distributions are compared with the actual correlation distribution from the population covariance matrix.

Table 1: Average Distance of the first 5 eigenvectors of the shrunk covariance matrices from different shrinkage methods versus population covariance matrices.

Methods	$n/p = 0.05$	$n/p = 0.1$	$n/p = 0.5$	$n/p = 2$
CorShrink-ML	0.030	0.034	0.033	0.036
CorShrink-VEM	0.037	0.039	0.057	0.062
CorShrink-VEM2	0.031	0.046	0.043	0.037
Schäfer-Strimmer	0.037	0.054	0.051	0.044
GLASSO ($\rho = 0.05$)	0.083	0.085	0.085	0.069
GLASSO ($\rho = 0.1$)	0.078	0.083	0.084	0.061
GLASSO ($\rho = 0.5$)	0.064	0.081	0.081	0.057
GLASSO ($\rho = 1$)	0.063	0.079	0.080	0.057
Sample cov	0.083	0.084	0.085	0.076

4 Discussion

Our goal here is to highlight the potential of the *CorShrink* models in performing adaptive correlation and covariance shrinkage that has comparable to better performance over the Schäfer Strimmer shrinkage approach in terms of eigen-space and eigenvalue patterns comparisons and that outperforms GLASSO as a correlation shrinkage method irrespective of the choice of regularization parameter used for the latter (see Figure 1 and Table **??**). In terms of computational time, the computation time for the *CorShrink-ML* method is comparable with GLASSO under medium to high shrinkage and Schäfer Strimmer method, while *CorShrink-VEM* and *CorShrink-VEM2* are slower in comparison. For instance, the time taken to run *CorShrink-ML*, *CorShrink-VEM* and *CorShrink-VEM2* on the Deng et al samples data were ?? , ?? and ?? seconds whereas that for GLASSO ($\rho = 1$) and Schäfer-Strimmer methods were ?? .

The *CorShrink* can be easily extended to partial correlation and partial covariance matrices and also leads to efficient computation of the inverse correlation and covariance matrices. The latter would allow one to build causal networks based on the *CorShrink* models and compare them to the GLASSO based causal networks. An example of a causal network on the samples of the Deng et al data using the three shrinkage approaches *CorShrink-ML*, Schäfer-Strimmer and GLASSO ($\rho = 1$) is provided here. This method also opens other areas of applications and extensions that we intend to pursue in future. One can combine Linear Discriminant Analysis and Multiple regression problem with the shrunk covariance matrices obtained from the *CorShrink* approaches, in the same way the Schäfer Strimmer method has been used in these domains [8] [4]. In this paper, we did not consider any additional structure on the covariance matrices. But for structured covariance matrices, one would want to pool the knowledge of the 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 could were fitted on simulated data. The Deng et al single cell data [7] is available as a R data package with instructions for downloading and loading into R at `https://github.com/kkdey/singleCellRNASeqMouseDeng2014`.

References

[1] 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.

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

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

[4] 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.

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

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

[7] 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.

[8] 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.

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

[10] 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.

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

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