Covariance Regularization by Thresholding

Mustafa Mughal

April 21st, 2022

University of Toronto, mustafa.mughal@mail.utoronto.ca

Abstract: This paper will highlight the importance of regularizing a sample covariance matrix as well as some usage in the field of statistics today. There are several ways in which a matrix can be regularized, however, this paper will focus on thresholding and comparing results with other methods such as Banding and Ledoit-Wolf shrinkage. Given a square matrix M with n observations, a covariance matrix in this form will struggle in performance when working with a considerably high number of predictors compared to the number of observations. Hence the need to regularize this matrix to maintain efficient performance through regularizing by thresholding. Results will be shown for positive definite covariance matrices with MVN variables and will demonstrate how to select the choice of threshold.

1. Introduction and Motivation

A covariance matrix is defined to be a square, symmetric matrix that describes the covariance between 2 or more random variables and where the diagonals correspond to the variance of each random variable. This matrix captures the way in which all variables in a data set change together. Covariance matrices are key in many statistical topics of interest such as principle components analysis, linear and quadratic classification (LDA and QDA) and all forms of regression. Throughout studies and experiments conducted today, high dimensional data settings or wide big data, are of great interest for research where conditions exist such that $p >> n$. In these settings, empirical covariance matrices struggle in terms of performance as they are ill-conditioned and are overall poor estimators of the population given samples of size n from p-variate MVN or Gaussian distribution predictors [1]. That is, when there is an exceedingly high number of predictors as there are observations, the sample empirical covariance matrix struggles in various performance measures which will be shown throughout this report. Therefore, it is evident that there does arise a need to improve upon a standard sample covariance matrix for more consistent results in high dimensions and one of the ways is through regularizing the matrix. The covariance matrix of interest throughout this paper will be in the form of the following for a model with n observations and where $k = 1, ..., n$. The covariance matrix will be symmetric and along the diagonals will be the variance of each random variable in the data set used to predict the outcome variable:

$$\hat{\sum} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})^T \tag{1}$$

1.1. *Regularization*

Regularization has a multiple difference connotations when working with matrices. However in this case, due to the nature of the high dimensional environment such that $p >> n$, regularizing a covariance matrix in this manner will essentially mean to reduce the degree of freedom d of an estimator which can often be accomplished by reducing the number of parameters in the given model [1]. This technique gives rise to methods such as LASSO and RIDGE regression which are the more simpler methods to reduce model complexity. These methods regularize by setting model coefficients to 0 or close to 0 respectively, which reduces the number of parameters in the model. Additionally, regularization is a technique used to help reduce over fitting as well as it can lead to sparse matrices and lower degree of freedom for a model [1]. Ideally, there should be $n > 10p$ where n is the number of observations and p is the number of predictors in a model, if the variable distribution is MVN/ Gaussian. If the distribution is not MVN/ Gaussian $n > 30p$ would be desired. These requirements ensure good conditioning of the covariance matrix for these models. Similarly, it can be shown that $n < 5p$ can lead to over-fitting and poor conditioning of the covariance matrix. It is important to note that one should not be regularizing a model too much since it could lead to under-fitting and poor performance as it may cut off importance predictors, so there is a need to find the correct balance of regularization such that neither over-fitting or under-fitting is present.

2. Thresholding

A proposed regularization method to improve covariance matrix performance is by way of thresholding. This is an efficient and permutation-invariant method to regularize a covariance matrix, where the model does not assume any spatial relationships between any of its predictors. Thresholding additionally provides consistency in the operator norm and uniformly over a set of sparse matrices which are often used in big data settings. A sparse matrix is defined to be a matrix primarily composed of coefficients set to 0 and has several practical uses throughout the world today. When working with big data, it can be extremely computationally expensive to use all of it all the time, which is why sparse matrices are so useful in this cases as they get rid of unwanted or unnecessary information upon request. With practical uses in machine learning, recommendation systems with large catalogues such as Netflix and image processing with a large amount of black pixels, sparse matrices are used to efficiently provide data and information that can readily be available for use. Finally thresholding can also obtain explicit convergence in the operator norm. This is of great importance since convergence in operator norm implies convergence in eigenvalues, which is a requirement for Principal Component Analysis (PCA). Therefore thresholding can be used in PCA applications which in of itself has multiple uses in the world today in fields such as IT with applications such as facial recognition, image processing and computational rendering.

This paper will further emphasises the importance and benefits of thresholding by comparing it to other similar covariance regularization methods and techniques such as the Ledoit-Wolf shrinkage. Before exploring the simulation results, it is necessary to explain how thresholding works when applied to a sample covariance matrix. The following Frobenius matrix norm, or sometimes called the Euclidean norm, is a matrix norm used to measure the distances between a set of matrices [2].

$$|||M||_F^2 = \sum_{i,j} m_{ij}^2 = tr(MM^T) \tag{2}$$

We can see that the frobenius matrix norm sums and squares all elements in matrix M which is equivalent to taking the trace of matrix M multiplied by its transpose since the matrix is square.

Lets now define the threshold operator $T_s(M)$, which reads as matrix M thresholded at s. As previously mentioned, this function will be permutation-invariant and will maintain positive definiteness given that both inequalities in (4) holds.

$$T_s(M) = [m_{ij} I(|m_{ij}| \geq s)] \tag{3}$$

$$||T_s - T_0|| \leq \varepsilon \quad and \quad \lambda_{min} > \varepsilon \tag{4}$$

Lets now show that positive definiteness holds when (4) holds for (3). It is clear that T_0 is the original Matrix M, since the absolute value of all entries in the matrix will be greater than or equal to 0 as we will only consider matrices with entries in the subset of $\mathbb{R}$.

For all vectors v such that $||v||_2 = 1$, the following holds.

$$\begin{gathered} v^T T_s(M) v \\ \geq v^T M v - \varepsilon \\ \geq \lambda_{min}(M) - \varepsilon \\ > 0 \end{gathered}$$

Therefore $T_s(M)$ is positive definite.

Hence, the threshold operator that this paper will cover will only deal with positive definite matrices as there may be some implications and inconsistencies if it is not positive definite. Positive definite matrices are of great interest throughout several fields such as physics, mathematics, statistics and computer science.

2.1. Selecting threshold s

When thresholding a matrix, it is important to select an optimal threshold s that helps to best regularize the covariance matrix such that a balance between over-fitting and under-fitting is achieved. Similar to other dimension reducing methods, the optimal parameter can be found using cross validation. For example in ridge regression $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{d} \beta_j^2$, the lambda tuning parameter can be optimally selected via cross validation. In thresholding, minimizing the following risk function can result in a strong performing s threshold value. The sample n is split into $n_1 = n(1 - \frac{1}{logn})$ and $n_2 = \frac{n}{logn}$ and then the process is repeated N times.

We will also consider the two empirical covariance matrices $\hat{\sum}_{1,v}$ and $\hat{\sum}_{2,v}$ following the v^{th} split. Recall that $T_0(M) = M$ so we have $T_0(\hat{\Sigma}_{2,v}) = \hat{\Sigma}_{2,v}$.

$$\hat{R}(s) = \frac{1}{N}\sum_{v=1}^{N} ||T_s(\hat{\sum}_{1,v}) - T_0(\hat{\sum}_{2,v})||_F^2$$

$$= \frac{1}{N}\sum_{v=1}^{N} ||T_s(\hat{\sum}_{1,v}) - \hat{\sum}_{2,v}||_F^2 \tag{5}$$

3. Simulations and Results

The simulations process will follow a 100 Monte Carlo repetitions, generating new test and training data sets each iteration. Furthermore, we will explore the performance measure of the sample covariance matrix as itself, then when regularized using thresholing and also with Ledoit-Wolf shrinkage. The decision to limit the simulation to just these performance measures and not being able to implement banding is purely based on computational burden and efficiency, as my machine is limited in that regard. We can see from the results the improvement that thresholding makes compared to the non regularized standard empirical covariance matrix and even Ledoit-Wolf shrinkage in this scenario. Using library climate, the data set used is the ”profile demo”, which as its name implies is a brief demographic sample of the entire data set for performance purposes. In a linear model, all continuous variables are used to predict TEMP. Furthermore, the sample covariance matrix in this simulation is not ordered and non-permuted, which helps to further emphasise the benefits of thresholding as Ledoit-Wolf struggles in this environment on a non-ordered matrix. After running the simulations, the following errors and results are recorded in the following tables. Lets take a look at the produced matrices from each method of threshold, Ledoit-Wolf shrinkage and the sample respectively.

```
> threshold_cov
```

	PRES	HGHT	TEMP	DWPT	RELH	MIXR	DRCT	SKNT	THTA	THTE	THTV
PRES	1	0	1	1	1	0	1	0	0	0	0
HGHT	0	1	0	0	0	0	0	0	1	1	1
TEMP	1	0	0	0	0	0	0	0	0	0	0
DWPT	1	0	0	0	0	0	0	0	0	0	0
RELH	1	0	0	0	0	0	0	0	0	0	0
MIXR	0	0	0	0	0	0	0	0	0	0	0
DRCT	1	0	0	0	0	0	1	0	0	0	0
SKNT	0	0	0	0	0	0	0	0	0	0	0
THTA	0	1	0	0	0	0	0	0	1	1	1
THTE	0	1	0	0	0	0	0	0	1	1	1
THTV	0	1	0	0	0	0	0	0	1	1	1

```
> LW_shrinkage
```

	PRES	HGHT	TEMP	DWPT	RELH	MIXR	DRCT	SKNT	THTA	THTE	THTV
PRES	6.259922e+13	-5.650705e+12	1.084414e+10	1.822969e+10	1.838316e+10	6.274335e+08	3.470086e+10	4.819102e+09	-9.114907e+10	-8.941267e+10	-9.105174e+10
HGHT	-5.650705e+12	2.261209e+14	-3.141529e+11	-5.281237e+11	-5.325789e+11	-1.817660e+10	-1.005500e+12	-1.396308e+11	2.640974e+12	2.590671e+12	2.638154e+12
TEMP	1.084414e+10	-3.141529e+11	6.240478e+13	1.013533e+09	1.022055e+09	3.488454e+07	1.929071e+09	2.679107e+08	-5.067317e+09	-4.970775e+09	-5.061905e+09
DWPT	1.822969e+10	-5.281237e+11	1.013533e+09	6.240588e+13	1.718142e+09	5.864219e+07	3.243106e+09	4.503960e+08	-8.518847e+09	-8.356556e+09	-8.509750e+09
RELH	1.838316e+10	-5.325789e+11	1.022055e+09	1.718142e+09	6.240591e+13	5.913528e+07	3.270556e+09	4.542025e+08	-8.590820e+09	-8.427165e+09	-8.581647e+09
MIXR	6.274335e+08	-1.817660e+10	3.488454e+07	5.864219e+07	5.913528e+07	6.240418e+13	1.116145e+08	1.550100e+07	-2.931901e+08	-2.876042e+08	-2.928770e+08
DRCT	3.470086e+10	-1.005500e+12	1.929071e+09	3.243106e+09	3.270556e+09	1.116145e+08	6.241035e+13	8.576418e+08	-1.622155e+10	-1.591266e+10	-1.620423e+10
SKNT	4.819102e+09	-1.396308e+11	2.679107e+08	4.503960e+08	4.542025e+08	1.550100e+07	8.576418e+08	6.240430e+13	-2.252529e+09	-2.209631e+09	-2.250125e+09
THTA	-9.114907e+10	2.640974e+12	-5.067317e+09	-8.518847e+09	-8.590820e+09	-2.931901e+08	-1.622155e+10	-2.252529e+09	6.244678e+13	4.179270e+10	4.255860e+10
THTE	-8.941267e+10	2.590671e+12	-4.970775e+09	-8.356556e+09	-8.427165e+09	-2.876042e+08	-1.591266e+10	-2.209631e+09	4.179270e+10	6.244517e+13	4.174808e+10
THTV	-9.105174e+10	2.638154e+12	-5.061905e+09	-8.509750e+09	-8.581647e+09	-2.928770e+08	-1.620423e+10	-2.250125e+09	4.255860e+10	4.174808e+10	6.244669e+13

```
> sample_cov
```

	PRES	HGHT	TEMP	DWPT	RELH	MIXR	DRCT	SKNT	THTA	THTE	THTV
PRES	140151.7492	-3324899.31	8620.83487	13554.54333	13031.89231	504.287603	14816.62179	2192.179487	-44800.8864	-43388.5418	-44717.587
HGHT	-3324899.3090	96500987.59	-184656.65705	-310627.03333	-313379.08974	-10683.606859	-593400.05128	-82402.211538	1558813.3974	1529251.3929	1557157.340
TEMP	8620.8349	-184656.66	559.56692	850.81750	789.74038	32.476128	629.74808	102.241026	-2189.5980	-2098.1997	-2184.137
DWPT	13554.5433	-310627.03	850.81750	1354.49000	1322.48333	48.973917	1196.52500	225.266667	-3970.5392	-3832.8433	-3962.415
RELH	13031.8923	-313379.09	789.74038	1322.48333	1414.92308	48.282692	1289.71795	247.961538	-4212.7808	-4077.3096	-4204.723
MIXR	504.2876	-10683.61	32.47613	48.97392	48.28269	2.120797	40.03737	2.570513	-129.4283	-123.4896	-129.063
DRCT	14816.6218	-593400.05	629.74808	1196.52500	1289.71795	40.037372	10939.57692	525.692308	-11715.5712	-11608.2186	-11710.185
SKNT	2192.1795	-82402.21	102.24103	225.26667	247.96154	2.570513	525.69231	211.230769	-1378.6904	-1371.2923	-1378.487
THTA	-44800.8864	1558813.40	-2189.59801	-3970.53917	-4212.78077	-129.428340	-11715.57115	-1378.690385	28586.0977	28235.3053	28567.809
THTE	-43388.5418	1529251.39	-2098.19974	-3832.84333	-4077.30962	-123.489622	-11608.21859	-1371.292308	28235.3053	27901.1603	28218.042
THTV	-44717.5873	1557157.34	-2184.13705	-3962.41500	-4204.72308	-129.063026	-11710.18462	-1378.486538	28567.8091	28218.0421	28549.585

It is clearly shown that the thresholding covariance matrix estimator is now a sparse matrix consisting of only 0 and 1. This helps to omit variables that do not seem to help predict the response variable TEMP. As for the Ledoit-Wold regularization method, it produces extremely high or low values for each element throughout the matrix. That may be due to its performance when working with non-ordered data.

Next, Table 1 will compare the standard errors of the sample empirical covariance matrix, the thresholded covariance matrix and the Ledoit-Wolf covariance matrix after the 100 Monte Carlo repetitions.

Errors	Sample	Thresholding	Ledoit-Wolf
std err	8790882	0.4124442	2.649829e+13
avg err	8890747	0.4316020	2.828139e+13

Table 1. Average and standard errors for matrix 1-norm and Frobenius-norm for Sample, Thresholded and Ledoit-Wolf covariance matrices

Clearly, it is shown that thresholding produces the best performing covariance matrix by its extremely low error values compared to those of the standard sample covariance matrix and the Ledoit-Wolf method. The latter two show signs of poor conditioning, as is to be expected when working with a non-permuted and non-ordered matrix in a high-dimensional environment [3]. It is shown that Thresholding works excellently and is a simple method to apply to large covariance matrices, irregardless of its ordering. Part of this is due to transforming the covariance matrix into a sparse matrix, which helps to improve performances as earlier explained in an efficient manner. Lets now take a look at some more matrix performance measures starting with the 1-norm which calculates the maximum column sum of a matrix M. This is a method frequently used in machine learning and can help to keep the coefficients of a model small and reduce overall model complexity. Next, the Frobenius matrix norm will also be examined for all 3 covariance matrices in Table 2. Then, the results will be further explained as the importance of the frobenius norm was mentioned to be the measurement of the distanced between sets of matrices. This will highlight the consistency of thresholding over a set of sparse matrices, unlike the other methods it is compared to. The MCSE for both the 1-norm and frobenius-norm will be displayed for each of the 3 empirical matrices.

MCSE	Sample	Thresholding	Ledoit-Wolf
1-norm	2298420	0.06263571	1.001692e+13
F-Norm	2090138	0.06460255	1.241864e+13

Table 2. MCSE for matrix 1-norm and Frobenius-norm for Sample, Thresholded and Ledoit-Wolf covariance matrices.

Here in Table 2, we can see similar results as thresholding vastly outperforms both the sample and Ledoit-Wolf covariance matrices by a significant margin, further demonstrating the power of thresholding. The simulation is limited however in terms of complexity and run time as on average can take up to 5 minutes to fully compile and complete the 100 iteration Monte Carlo simulation. This may be due to having to compute several matrix operations and loop them 100 times, which places a burden on the machine running the code. However, conclusions can still be made from this simulation. Thresholding is an efficient and simple method to apply in order to regularize a covariance matrix. It makes matrices sparse which eases burden on computation, as well as being able to perform regardless of data not being permuted or ordered [2]. This can not be said for Ledoit-Wolf as it struggles when data is not ordered. These results further verify that thresholding is permutation-invariant and does have consistency over a set of sparse matrices as well as matrix operator norm performance measures. Differences between this simulation and the ones conducted by P.J. Bickel and E. Levina are mostly due to hardware limitations as well as perhaps different algorithms when computing the threshold and Ledoit-Wolf estimators. Furthermore, this simulation selects the optimal threshold s from a predetermined set of values ranging from 0 to 5000, then chosen by minimizing the risk function given by (5). This is also another potential factor as to why results differ where perhaps selecting by cross validation or increasing the range of possible s could have provided a more accurate result. There would simply be too much of a strain on computational power on the machine running this simulation. That being stated, these results from this simulation still show the application and benefits of regularizing a sample covariance matrix for use in high dimension settings.

4. Conclusion

To conclude, the thresholding estimator of the sample covariance matrix is an excellent performing matrix when working in a large data set such as CLIMATE. Thresholding allows for simple and quick implementation when applied to large covariance matrices of varying sizes [4]. Consistency in matrix operator norms as well as on a set of sparse matrices implies that eigenvalues and eigenvectors converge as well. This is important since it allows for use in PCA applications which are frequently used in applied statistics and research in the field today. Although the simulation was computationally expensive, that is largely due to calculating several matrix norms at once as well as for computing multiple matrices and the Ledoit-Wolf shrinkage estimates as well. Being able to work on non ordered and non permuted data proves as an advantage over other regularization methods such as banding and Ledoit-Wolf shrinkage [3], showing the versatility of thresholding. A potential downside is that there maybe loss of positive definiteness but since we limit our model and simulations by assuming (4) holds, this will not be a problem in this case but it can be when working in more of a general environment without this limitation.

References

1. Abadie, A., ; Kasy, M. (2017). (tech.). *Choosing among regularized estimators in empirical economics*. Cambridge, MA: Harvard. 3-48 https://scholar.harvard.edu/files/kasy/files/riskml.pdf.
2. Bickel, P. J., ; Levina, E. (2008). *Covariance Regularization By Thresholding* (6th ed., Vol. 36, Ser. 2577-2604). Institute of Mathematical Sciences. doi:10.1214/08-AOS600
3. Ledoit, O., Wolf, M. (2020) . *The Power of (Non-)Linear Shrinkage: A Review and Guide to Covariance Matrix Estimation* in SSRN, University of Zurich 2020, pp. 3-15. http://dx.doi.org/10.2139/ssrn.3384500
4. Pourahmadi, M. (2013). Banding, Tapering, and Thresholding. In M. Pourahmadi (Ed.), High-Dimensional Covariance Estimation (pp. 141–151). essay, Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118573617