RESEARCH REVIEW JOURNALS
Volume-04
Issue-03
March -2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

A Study of Modern Sparse Estimation Analysis in Applied Statistics Lasso Penalty

[1]Varade Nitin Kumar Narahari & [2]Dr. Sudesh Kumar

[1]*Research Scholar OPJS University Churu Rajasthan*
[2]*Associate Professor OPJS University Churu Rajasthan*

ARTICLE DETAILS

Article History
Published Online: 13 March 2019

Keywords
Sparse Analysis, applied statistics, large covariance matrices, optimization problem, algorithm.

ABSTRACT

The paper proposes a new covariance estimator for large covariance matrices when the variables have a natural ordering. Using the Cholesky decomposition of the inverse, we impose a banded structure on the Cholesky factor, and select the bandwidth adaptively for each row of the Cholesky factor, using a novel penalty we call nested Lasso. This structure has more flexibility than regular banding, but, unlike regular Lasso applied to the entries of the Cholesky factor, results in a sparse estimator for the inverse of the covariance matrix. An iterative algorithm for solving the optimization problem is developed. The estimator is compared to a number of other covariance estimators and is shown to do best, both in simulations and on a real data example. Simulations show that the margin by which the estimator outperforms its competitors tends to increase with dimension.

1. Introduction

Estimating covariance matrices has always been an important part of multivariate analysis, and estimating large covariance matrices (where the dimension of the data p is comparable to or larger than the sample size n) has gained particular attention recently, since high-dimensional data are so common in modern applications (gene arrays, fMRI, spectroscopic imaging, and many others). There are many statistical methods that require an estimate of a covariance matrix. They include principal component analysis (PCA), linear and quadratic discriminant analysis (LDA and QDA) for classification, regression for multivariate normal data, inference about functions of the means of the components (e.g., about the mean response curve in longitudinal studies), and analysis of independence and conditional independence relationships between components in graphical models. Note that in many of these applications (LDA, regression, conditional independence analysis) it is not the population covariance itself that needs estimating, but its inverse Σ^{-1}, also known as the precision or concentration matrix. When p is small, an estimate of one of these matrices can easily be inverted to obtain an estimate of the other one; but when p is large, inversion is problematic, and it may make more sense to estimate the needed matrix directly.

Sparsity in the inverse is particularly useful in graphical models, since zeroes in the inverse imply a graph structure. Banerjee et al. (2006) and Yuan and Lin (2007), using different semi-definite programming algorithms, both achieve sparsity by penalizing the normal likelihood with an L1 penalty imposed directly on the elements of the inverse. This approach is computationally very intensive and does not scale well with dimension, but it is invariant under variable permutations. When a natural ordering of the variables is available, sparsity in the inverse is usually introduced via the modified Cholesky decomposition

$$\Sigma^{-1} = T^{\top} D^{-1} T.$$

Here T is a lower triangular matrix with ones on the diagonal, D is a diagonal matrix, and the elements below diagonal in the ith row of T can be interpreted as regression coefficients of the ith component on its predecessors; the elements of D give the corresponding prediction variances.

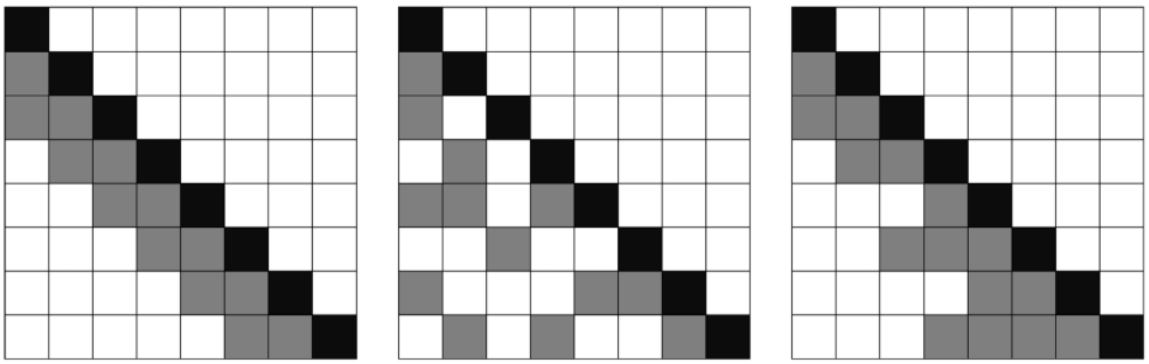

FIG. 1. The placement of zeros in the Cholesky factor T : (a) Banding; (b) Lasso penalty of Huang et al.; (c) Adaptive banding

2. Methods for penalized estimation of the Cholesky factor

For the sake of completeness, we start from a brief summary of the formal derivation of the Cholesky decomposition of Σ^{-1}. Suppose we have a random vector $X = (X_1, \ldots, X_p)^\top$, with mean 0 and covariance Σ. Let $X_1 = \varepsilon_1$ and, for $j > 1$, let

$$X_j = \sum_{l=1}^{j-1} \phi_{jl} X_l + \varepsilon_j, \tag{1}$$

where ϕ_{jl} are the coefficients of the best linear predictor of X_j from $X_1, \ldots, X_{j-1}$ and $\sigma_j^2 = \mathrm{Var}(\varepsilon_j)$ the corresponding residual variance. Let Φ be the lower triangular matrix with jth row containing the coefficients ϕ_{jl}, $l = 1, \ldots, j-1$, of the jth regression (1). Note that Φ has zeros on the diagonal. Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^\top$, and let $D = \mathrm{diag}(\sigma_j^2)$ be a diagonal matrix with σ_j^2 on the diagonal. Rewriting (1) in matrix form gives

$$\varepsilon = (I - \Phi)\mathbf{X}, \tag{2}$$

where I is the identity matrix. It follows from standard regression theory that the residuals are uncorrelated, so taking covariance of both sides of (2) gives

$$D = (I - \Phi)\Sigma(I - \Phi)^\top.$$

Letting $T = I - \Phi$, we can now write down the modified Cholesky decompositions Σ and Σ^{-1}

$$\Sigma = T^{-1} D (T^{-1})^\top, \qquad \Sigma^{-1} = T^\top D^{-1} T. \tag{3}$$

Note that the only assumption on X was mean 0; normality is not required to derive the Cholesky decomposition.

The natural question is how to estimate the matrices T and D from data. The standard regression estimates can be computed as long as p ≤ n, but in high dimensional situations one expects to do better by regularizing the coefficients in T in some way, for the same reasons one achieves better prediction from regularized regression [Hastie et al. (2001)]. If p>n, the regression problem becomes singular, and some regularization is necessary for the estimator to be well defined.

The negative log-likelihood of the data, up to a constant, is given by

$$\begin{aligned} \ell(\Sigma, \mathbf{x}_1, \ldots, \mathbf{x}_n) &= n \log |\Sigma| + \sum_{i=1}^{n} \mathbf{x}_i^\top \Sigma^{-1} \mathbf{x}_i \\ &= n \log |D| + \sum_{i=1}^{n} \mathbf{x}_i^\top T^\top D^{-1} T \mathbf{x}_i. \end{aligned} \tag{4}$$

The negative log-likelihood can be decomposed into

$$\ell(\Sigma, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{j=1}^{p} \ell_j(\sigma_j, \boldsymbol{\phi}_\mathbf{j}, \mathbf{x}_1, \ldots, \mathbf{x}_n),$$

Where,

$$\ell_j(\sigma_j, \boldsymbol{\phi}_\mathbf{j}, \mathbf{x}_1, \ldots, \mathbf{x}_n) = n \log \sigma_j^2 + \sum_{i=1}^{n} \frac{1}{\sigma_j^2} \left(x_{ij} - \sum_{l=1}^{j-1} \phi_{jl} x_{il} \right)^2. \tag{5}$$

Minimizing (4) is equivalent to minimizing each of the functions ℓ_j in (5), which is in turn equivalent to computing the best least squares, fit for each of the regressions (1).

Huang et al. (2006) proposed adding a penalty to (4) and minimizing

$$\ell(\Sigma, \mathbf{x}_1, \ldots, \mathbf{x}_n) + \lambda \sum_{j=2}^{p} P(\phi_j), \tag{6}$$

where the penalty P on the entries of $\phi_j = (\phi_{j1}, \ldots, \phi_{j,j-1})$ is

$$P(\phi_j) = \|\phi_j\|_d^d, \tag{7}$$

and $\|\cdot\|_d$ is the L_d vector norm with d = 1 or 2. The L_2 penalty (d = 2) does not result in a sparse estimate of the covariance, so we will not focus on it here. The L_1 penalty (d = 1), that is, the Lasso penalty, results in zeros irregularly placed in T as shown in Figure 1(b), which also does not produce a sparse estimate of Σ^{-1}. Again, minimizing (6) is equivalent to separately minimizing

$$\ell_j(\sigma_j, \boldsymbol{\phi_j}, \mathbf{x}_1, \ldots, \mathbf{x}_n) + \lambda P(\phi_j), \tag{8}$$

With $P(\phi_1) = 0.$

We propose replacing the L_1 penalty $\lambda \sum_{l=1}^{j-1} |\phi_{jl}|$ with a new nested Lasso penalty,

$$J_0(\phi_j) = \lambda\left(|\phi_{j,j-1}| + \frac{|\phi_{j,j-2}|}{|\phi_{j,j-1}|} + \frac{|\phi_{j,j-3}|}{|\phi_{j,j-2}|} + \cdots + \frac{|\phi_{j,1}|}{|\phi_{j,2}|}\right), \tag{9}$$

where we define $0/0 = 0.$ The effect of this penalty is that if the lth variable is not included in the jth regression $(\phi_{jl} = 0)$, then all the subsequent variables (l − 1 through 1) are also excluded, since giving them nonzero coefficients would result in an infinite penalty. Hence, the j th regression only uses $k_j \leq j - 1$ closest predecessors of the jth coordinate, and each regression has a different order kj .

However, the nested Lasso penalty is of independent interest and may be used in other contexts, for example, for group variable selection. To address the scaling issue in general, we propose two easy modifications of the penalty (9):

$$J_1(\phi_j) = \lambda\left(\frac{|\phi_{j,j-1}|}{|\hat{\phi}^*_{j,j-1}|} + \frac{|\phi_{j,j-2}|}{|\phi_{j,j-1}|} + \frac{|\phi_{j,j-3}|}{|\phi_{j,j-2}|} + \cdots + \frac{|\phi_{j,1}|}{|\phi_{j,2}|}\right), \tag{10}$$

‘

$$J_2(\phi_j) = \lambda_1 \sum_{t=1}^{j-1} |\phi_{j,t}| + \lambda_2 \sum_{t=1}^{j-2} \frac{|\phi_{j,t}|}{|\phi_{j,t+1}|}, \tag{11}$$

3. Numerical results

In this section we compare adaptive banding to other methods of regularizing the inverse. Our primary comparison is with the Lasso method of Huang et al. (2006) and with nonadaptive banding of Bickel and Levina (2007); these methods are closest to ours and also provide a sparse estimate of the Cholesky factor. As a benchmark, we also include the shrinkage estimator of Ledoit and Wolf (2003), which does not depend on the order of variables.

Simulation data: Simulations were carried out for three different covariance models. The first one has a tri-diagonal Cholesky factor and, hence, a tridiagonal inverse:

$$\boldsymbol{\Sigma}_1 : \phi_{j,j-1} = 0.8; \qquad \phi_{j,j'} = 0, \qquad j' < j - 1; \qquad \sigma_j^2 = 0.01.$$

The second one has entries of the Cholesky factor exponentially decaying as one moves away from the diagonal. Its inverse is not sparse, but instead has many small entries:

$$\boldsymbol{\Sigma}_2 : \phi_{j,j'} = 0.5^{|j-j'|}, \qquad j' < j; \qquad \sigma_j^2 = 0.01.$$

Both these models were considered by Huang et al. (2006), and similar models were also considered by Bickel and Levina (2007). In both Σ_1 and Σ_2, all the rows have the same structure, which favors regular non-adaptive banding.

To test the ability of our algorithm to adapt, we also considered the following structure:

$$\Sigma_3 : k_j \sim U(1, \lceil j/2 \rceil); \qquad \phi_{j,j'} = 0.5, \qquad k_j \le j' \le j-1;$$

$$\phi_{j,j'} = 0, \qquad j' < k_j; \qquad \sigma_j^2 = 0.01.$$

Here $U(k_1, k_2)$ denotes an integer selected at random from all integers from k_1 to k_2. For moderate values of p, this structure is stable, and this is what we generate for p = 30 in the simulations below. For larger p, some realizations can generate a poorly conditioned true covariance matrix, which is not a problem in principle, but makes computing performance measures awkward. To avoid this problem, we divided the variables for p = 100 and p = 200 into 3 and 6 independent blocks, respectively, and generated a random structure from the model described above for each of the blocks. We will refer to all these models as Σ_3. The structure of Σ_3 should benefit more from adaptive banding.

For each of the covariance models, we generated n = 100 training observations, along with a separate set of 100 validation observations. We considered three different values of p: 30, 100 and 200, and two different distributions: normal and multivariate t with 3 degrees of freedom, to test the behavior of the estimator on heavy-tailed data. The estimators were computed on the training data, with tuning parameters for all methods selected by maximizing the likelihood on the validation data. Using these values of the tuning parameters, we then computed the estimated covariance matrix on the training data and compared it to the true covariance matrix.

There are many criteria one can use to evaluate covariance matrix estimation, for example, any one of the matrix norms can be calculated for the difference (L_1, L_2, L_∞, or Frobenius norm). There is no general agreement on which loss to use in which situation.

$$\Delta_{\mathrm{KL}}(\Sigma, \hat{\Sigma}) = \mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) - \ln|\hat{\Sigma}^{-1}\Sigma| - p. \tag{12}$$

Another popular loss is the entropy loss for the covariance matrix, which was used by Huang et al. (2006). The entropy loss is the same as the Kullback–Leibler loss except the roles of the covariance matrix and its inverse are switched.

The results for the normal data and the three models are summarized in Table 1, which gives the average losses and the corresponding standard errors over 50 replications. The NA values for the sample appear when the matrix is singular. The J_0 penalty has been omitted because it is dominated by J_1 and J_2.

In general, we see that banding and adaptive banding perform better on all three models than the sample, Ledoit–Wolf's estimator and Lasso. On Σ_1 and Σ_2, as expected, banding and adaptive banding are very similar (particularly once standard errors are taken into account); but on Σ_3, adaptive banding does better, and the larger p, the bigger the difference. Also, for normal data the J_2 penalty always dominates J_1, though they are quite close.

TABLE 1 Multivariate normal simulations for models Σ_1 (banded Cholesky factor), Σ_2 (nonsparseCholesky factor with elements decaying exponentially as one moves away from the diagonal) and Σ_3 (sparse Cholesky factor with variable length rows).

p	Sample	Ledoit–Wolf	Lasso	J_1	J_2	Banding
			Σ_1			
30	8.38(0.14)	3.59(0.04)	1.26(0.04)	0.79(0.02)	0.64(0.02)	0.63(0.02)
100	NA	29.33(0.12)	6.91(0.11)	2.68(0.04)	2.21(0.03)	2.21(0.03)
200	NA	90.86(0.19)	14.57(0.13)	5.10(0.06)	4.35(0.05)	4.34(0.05)
			Σ_2			
30	8.38(0.14)	3.59(0.02)	2.81(0.04)	1.42(0.03)	1.32(0.02)	1.29(0.03)
100	NA	18.16(0.02)	16.12(0.09)	5.01(0.07)	4.68(0.06)	4.55(0.05)
200	NA	40.34(0.02)	32.84(0.11)	9.88(0.06)	9.28(0.06)	8.95(0.06)
			Σ_3			
30	8.68(0.12)	171.31(1.00)	4.62(0.07)	3.26(0.05)	3.14(0.06)	3.82(0.05)
100	NA	945.65(2.15)	35.60(0.71)	11.82(0.13)	11.24(0.12)	14.34(0.09)
200	NA	1938.32(3.04)	118.84(1.54)	23.30(0.17)	22.70(0.16)	29.50(0.14)

To test the behavior of the methods with heavy-tailed data, we also performed simulations for the same three covariance models under the multivariate t3 distribution (the heaviest-tail t distribution with finite variance). These results are given in Table 2. All methods perform worse than they do for normal data, but banding and adaptive banding still do better than other methods.

Because the standard errors are larger, it is harder to establish a uniform winner among J1, J2 and banding, but generally these results are consistent with results obtained for normal data.

TABLE 2 Multivariate t3 simulations for models $\Sigma_1, \Sigma_2, \Sigma_3$

p	Sample	Ledoit–Wolf	Lasso	J_1	J_2	Banding
			Σ_1			
30	30.33(0.65)	9.22(0.65)	7.60(0.74)	4.32(0.21)	3.68(0.19)	4.22(0.60)
100	NA	58.24(2.61)	38.99(1.44)	15.58(0.78)	13.85(0.72)	13.74(0.72)
200	NA	139.21(3.02)	111.62(2.73)	31.45(1.80)	28.22(1.71)	27.95(1.70)
			Σ_2			
30	30.33(0.65)	6.20(0.15)	8.44(0.20)	5.91(0.24)	5.21(0.22)	5.23(0.24)
100	NA	24.37(0.67)	31.92(0.83)	21.76(0.76)	18.87(0.71)	19.33(0.85)
200	NA	50.40(1.41)	64.28(1.98)	44.58(2.00)	38.46(1.75)	39.81(1.98)
			Σ_3			
30	30.77(0.74)	199.73(4.32)	14.48(0.40)	11.47(0.44)	11.57(0.47)	11.69(0.39)
100	NA	1061.54(12.62)	82.05(1.47)	43.38(1.14)	45.01(1.13)	42.78(1.04)
200	NA	2182.54(21.29)	182.82(9.51)	87.5(2.75)	91.25(2.79)	85.65(2.49)

4. Conclusion

The study presented a new covariance estimator for ordered variables with a banded structure, which, by selecting the bandwidth adaptively for each row of the Cholesky factor, achieves more flexibility than regular banding but still preserves sparsity in the inverse. Adaptive banding is achieved using a novel nested Lasso penalty, which takes into account the ordering structure among the variables. The estimator has been shown to do well both in simulations and a real data example. Zhao et al. (2006) proposed a related penalty, the composite absolute penalty (CAP), for handling hierarchical structures in variables. However, Zhao et al. (2006) only considered a hierarchy with two levels, while,in our setting, there are essentially p − 1 hierarchical levels; hence, it is not clear how to directly apply CAP without dramatically increasing the number of tuning parameters.The theoretical properties of the estimator are a subject for future work. The nested Lasso penalty is not convex in the parameters; it is likely that the theory developed by Fan and Li (2001) for non-convex penalized maximum likelihood estimation can be extended to cover the nested Lasso (it is not directly applicable since our penalty cannot be decomposed into a sum of identical penalties on the individual coefficients). However, that theory was developed only for the case offixed p, n → ∞, and the more relevant analysis for estimation of large covariance matrices would be under the assumption p → ∞, n → ∞, with p growing at a rate equal to or possibly faster than that of n, as was done for the banded estimator by Bickel and Levina (2007). Another interesting question for future work is extending this idea to estimators invariable under variable permutations.

References

1. ADAM, B., QU, Y., DAVIS, J., WARD, M., CLEMENTS, M., CAZARES, L., SEMMES, O., SCHELLHAMMER, P., YASUI, Y., FENG, Z. and WRIGHT, G. (2002). Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Research 62 3609–3614.
2. ANDERSON, T. W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York. MR0091588 BANERJEE, O., D'ASPREMONT, A. and EL GHAOUI, L. (2006). Sparse covariance selection via robust maximum likelihood estimation. In Proceedings of ICML.
3. BICKEL, P. J. and LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes," and some alternatives when there are many more variables than observations. Bernoulli 10 989–1010.
4. MR2108040 BICKEL, P. J. and LEVINA, E. (2007). Regularized estimation of large covariance matrices. Ann. Statist. To appear.
5. DIGGLE, P. and VERBYLA, A. (1998). Nonparametric estimation of covariance structure in longitudinal data. Biometrics 54 401–415.
6. DJAVAN, B., ZLOTTA, A., KRATZIK, C., REMZI, M., SEITZ, C., SCHULMAN, C. and MARBERGER, M. (1999). Psa, psa density, psa density of transition zone, free/total psa ratio, and psa velocity for early detection of prostate cancer in men with serum psa 2.5 to 4.0 ng/ml. Urology 54 517–522.
7. FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. J. Econometrics. To appear.
8. FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 1348–1360. MR1946581
9. FRIEDMAN, J., HASTIE, T., HÖFLING, H. G. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. Ann. Appl. Statist. 1 302–332.
10. FU, W. (1998). Penalized regressions: The bridge versus the lasso. J. Comput. Graph. Statist. 7 397–416. MR1646710
11. FURRER, R. and BENGTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. J. Multivariate Anal. 98 227–255. MR2301751
12. HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). The Elements of Statistical Learning. Springer, Berlin. MR1851606
13. HUANG, J., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. Biometrika 93 85–98. MR2277742
14. JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. Ann. Statist. 29 295–327. MR1863961
15. JOHNSTONE, I. M. and LU, A. Y. (2007). Sparse principal components analysis. J. Amer. Statist. Assoc. To appear.

16. LEDOIT, O. and WOLF, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88 365–411. MR2026339
17. MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). Multivariate Analysis. Academic Press, New York. MR0560319
18. PANNEK, J. and PARTIN, A. (1998). The role of psa and percent free psa for staging and prognosis prediction in clinically localized prostate cancer. Semin. Urol. Oncol. 16 100–105.
19. POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. Biometrika 86 677–690.
20. MR1723786 SMITH, M. and KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. J. Amer. Statist. Assoc. 97 1141–1153. MR1951266
21. STAMEY, T., JOHNSTONE, I., MCNEAL, J., LU, A. and YEMOTO, C. (2002). Preoperative serum prostate specific antigen levels between 2 and 22 ng/ml correlate poorly with post-radical prostatectomy cancer morphology: Prostate specific antigen cure rates appear constant between 2 and 9 ng/ml. J. Urol. 167 103–111

