Title
Comment: Ridge Regression and Regularization of Large Matrices

Journal
Technometrics, 62(4)

ISSN
0040-1706

Authors
Le, CM
Levin, K
Bickel, PJ
et al.

Publication Date
2020-10-01

DOI
10.1080/00401706.2020.1796815

Peer reviewed

Ridge Regression and Regularization of Large Matrices

Can M. Le
Department of Statistics, University of California, Davis
Keith Levin
Department of Statistics, University of Michigan
Peter J. Bickel
Department of Statistics, University of California, Berkeley and
Elizaveta Levina
Department of Statistics, University of Michigan

May 25, 2020

Abstract

We view ridge regression through the lens of eigenvalue shrinkage, and consider its influence on two modern problems in high-dimensional statistical inference: covariance estimation and community detection in networks.

1 Introduction

The enormous influence of the ridge penalty is well described in Trevor Hastie's excellent summary. Here we focus on one particular interpretation of ridge with deep connections to modern high-dimensional statistical inference: eigenvalue shrinkage. One interpretation of the ridge penalty is that it prevents singularity or ill-conditioning of the matrix $X^T X$ in linear regression: while $X^T X$ may or may not be invertible, the matrix $X^T X + \lambda I$, for $\lambda > 0$, always is, and it is always positive definite. One can thus view $X^T X + \lambda I$ and $(X^T X + \lambda I)^{-1}$ as regularized estimates of the corresponding population matrices. In its most general form, regularization moves an estimate towards a region in which the estimand is believed to be. This need not correspond to using a penalty, or equivalently a Bayesian prior, but can also be achieved by directly operating on a basic estimator; one such example is Steinian shrinkage, which moves the MLE of the mean towards 0.

The need for regularized estimation of matrices in high dimensions arises in multiple areas, including covariance estimation and network analysis. Two popular approaches are to regularize large matrices with a positive diagonal toward low-rankness and towards element-wise sparsity. These two approaches work somewhat against one another: the sparsest possible invertible estimator is diagonal, but it has full rank. On the other hand, a low rank matrix is unlikely to be sparse, as most entries are a function of the same few eigenvectors.

It is instructive to consider the effect of regularization on the estimated eigenvalues and eigenvectors. Ridge regression replaces each eigenvalue λ_i of $X^T X$ with $\lambda_i + \lambda$, making the matrix better-conditioned, since $(\lambda_1 + \lambda)/(\lambda_p + \lambda) < \lambda_1/\lambda_p$. We can think of this as regularizing towards sparsity, since we are shrinking towards the sparse identity matrix, and making the matrix full rank as a result. Ridge regularization does not change the eigenvectors of $X^T X$, which is necessary in some applications. Alternatively, regularization towards sparsity can be achieved by sparsifying the eigenvectors themselves, as in sparse principal components or sparse canonical correlation analysis [Jolliffe et al., 2003, Zou et al., 2006, Johnstone and Lu, 2009]. This is often done by imposing a LASSO or elastic net penalty on the eigenvector entries [Witten et al., 2009]. The choice of the algorithm may raise numerical issues, guiding the effectiveness of regularization [Journée et al., 2010].

2 Regularization in covariance estimation

A classical task in covariance estimation is to estimate $\Sigma \in \mathcal{R}^{p\times p}$ based on n i.i.d. observations $X_i \sim \mathcal{N}(0, \Sigma)$, collected in the columns of $X \in \mathcal{R}^{n\times p}$. A natural choice is the MLE, the sample covariance matrix $S = X^T X/n \in \mathcal{R}^{p\times p}$. However, unless $p \ll n$, S is not a consistent estimator and the eigenvalues of S are over-dispersed. For example, when $p/n \to \gamma \in (0, 1)$ and $\Sigma = I$, the largest eigenvalue of S converges to $1 + \sqrt{\gamma}$ rather than 1 [Marčenko and Pastur, 1967, Bai and Silverstein, 2010].

One natural approach, in the spirit of Steinian shrinkage, is to adjust the spectrum of S while keeping its eigenspace unchanged, yielding *orthogonally invariant* estimators of the form $Q \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)Q^T$, where Q is the matrix of eigenvectors of S and the $\hat{\lambda}_i$ are modified eigenvalues of S. The matrix $X^T X_\lambda I$ used by ridge is an orthogonally invariant estimator, with a constant added to each eigenvalue of S. Ledoit and Wolf [2004] proposed another, $(1 - \rho)S + \rho\nu I$, where $\rho \in [0, 1]$ and $\nu \geq 0$ are optimized with respect to

Frobenius error and estimated from data. Subsequent work has extended these ideas via Rao-Blackwellization [Chen et al., 2010], better estimation of the optimal ρ and ν [Fisher and Sun, 2011] and relaxing the normality assumptions [Touloumis, 2015].

Unfortunately, linear shrinkage fails to capture the nonlinear over-dispersion of the sample eigenvalues predicted by random matrix theory. Early on, Stein [1986] proposed nonlinear shrinkage of the eigenvalues as an improvement over the MLE S under the entropy loss $L(\hat{\Sigma}, \Sigma) = \operatorname{tr} \hat{\Sigma}\Sigma^{-1} - \log\det \hat{\Sigma}\Sigma^{-1}$. Won et al. [2013] minimize this loss subject to a condition number constraint to ensure numerical stability. Since many loss functions (e.g., operator and Frobenius norms) are rotationally invariant, Donoho et al. [2018] derived eigenvalue shrinkage procedures for a variety of loss functions, and established asymptotic optimality among a family of rotationally invariant estimators under the spiked covariance model Bai and Silverstein [2010], Yao et al. [2015]. El Karoui [2008], Ledoit and Wolf [2012] proposed adjusting the sample eigenvalues based on the functional equation relating the limiting spectral distribution to its Stieltjes transform.

Orthogonally invariant shrinkage of the covariance matrix is illustrated in Figure 1. The data are $n = 1000$ samples from $\mathcal{N}(0, \Sigma)$ with covariance $\Sigma = \operatorname{diag}(4, 2, 1, \ldots, 1) \in \mathcal{R}^{p\times p}$, and $p = 250$. The first plot shows the over-dispersed eigenvalues of the sample covariance, and the next two plots results of two different orthogonally invariant shrinkage estimators.

Orthogonally invariant estimators do not address the fact that eigenvectors of S are not consistent in high dimensions, either [Johnstone and Lu, 2009]. Regularization of the whole matrix rather than only its spectrum has largely focused on imposing sparsity or other structural assumptions. Tapering or banding the covariance matrix [Wu and Pourahmadi,

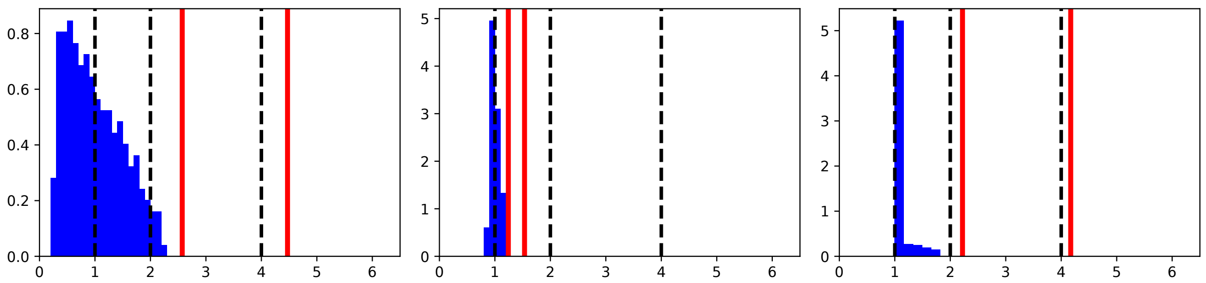

Figure 1: Eigenvalue shrinkage. Each plot shows true population eigenvalues (vertical dashed lines), the two largest sample eigenvalues (vertical red lines), and the histogram of the remaining $p-2$ sample eigenvalues. Left: sample covariance matrix. Middle: linear shrinkage toward the identity by Fisher and Sun [2011]. Right: optimal nonlinear shrinkage by Donoho et al. [2018].

2003, Bickel and Levina, 2008a] is especially suitable when the variables have a spatial or temporal ordering; order-invariant analogues can be obtained by thresholding [Bickel and Levina, 2008b, Rothman et al., 2009, Cai and Liu, 2011]; see Cai et al. [2016] for review. There is also a large literature on imposing sparsity on the precision matrix, with connections to Gaussian graphical model estimation.

An alternative family of approaches uses the geometry of the space of positive definite matrices to impose regularization through curvature, rather than explicit shrinkage [Smith, 2005]. These ideas have primarily been applied to computing matrix averages. Schwartzman [2016] showed how different choices of matrix geometry give rise to different choices of matrix means. Lodhia et al. [2019] showed how the "geometrical" regularization imposed by the matrix harmonic mean can outperform the arithmetic mean in some high-dimensional settings. In summary, a range of covariance regularization ideas directly descend from the ridge penalty, and all rely on the broad idea of regularization to improve the estimator given by $X^T X$ in high dimensions.

3 Regularization in network analysis

Network analysis studies interactions between entities, represented as a graph and typically encoded by an adjacency matrix A, a binary $n \times n$ matrix with $A_{ij} = 1$ if there is an edge from node i to node j. A is typically modeled as random with expectation given by the probability matrix $P = \mathbb{E}\, A$. In many applications, a single adjacency matrix is observed, and structural assumptions must be imposed on P to facilitate inference. These assumptions typically posit that P has low-dimensional latent structure. Under a popular model called the inhomogeneous Erdős-Rényi graph [Bollobas et al., 2007], all edges are independent, and thus all information about the latent structure is contained in P.

Regularization of eigenvalues and eigenvectors in network analysis has several motivations. A low-rank assumption on P leads naturally to thresholding small eigenvalues to zero. Another common assumption is community structure in the network (which implies low rank). Community structure is often observed in real-world social networks, with nodes partitioned into groups according to similarity of their connectivity patterns. A popular, tractable, and by now well-understood model for networks with communities is the stochastic block model [SBM; Holland et al., 1983, Abbe, 2018]. Under the SBM, P is block-constant, and the probability of connection between two nodes is fully determined by their community memberships. In a model with K communities, the leading K eigenvectors of P contain all the information about community structure. Spectral clustering [von

Luxburg, 2007] is popular and successful in practice, but it requires that the K leading eigenvectors of A being close to those of P. This can be established by first showing that A concentrates well around P, and then using the Weyl's inequality and the Davis-Kahan theorem to conclude that the eigenvalues and eigenvectors of A and P are close.

Concentration has been extensively studied in random matrix theory. For the inhomogeneous Erdős-Rényi random networks, the matrix Bernstein inequality gives $\|A - \mathbb{E}\, A\| = O(\sqrt{d \log n})$ with high probability if the maximum expected degree $d = \max_i \mathbb{E} \sum_{j=1}^{n} A_{ij}$ grows at least as $\log n$ [Oliveira, 2010]. The optimal bound if d grows at least as fast as $\log n$ is $\|A - \mathbb{E}\, A\| = (2 + o(1))\sqrt{d}$ [Benaych-Georges et al., 2017]. For the normalized Laplacian (defined by $L = D^{-1/2} A D^{1-/2}$ where D is the diagonal matrix of node degrees $d_i = \sum_{j=1}^{n} A_{ij}$ on the diagonal), which often performs better in practice by reducing degree heterogeneity, concentration follows directly from the concentration of the adjacency matrix and node degrees.

In the sparse case, meaning the average degree grows slower than $\log n$, spectral clustering is known to perform poorly due to high degree variance. Several regularization approaches have been proposed, including reducing the influence of high-degree nodes [Chin et al., 2015, Le et al., 2017] and adding a small quantity to either the diagonal of A or to every element of A prior to clustering [Chaudhuri et al., 2012, Amini et al., 2013]. As with ridge, these methods shrink eigenvalues, helping those corresponding to informative eigenvectors stay at the top of the spectrum, and therefore allowing spectral clustering to recover communities. The goal of regularization here is to restore concentration of the adjacency matrix or its Laplacian around their expectation, even in the setting where $d = O(1)$

[Chin et al., 2015, Le et al., 2017].

The effect of regularization is illustrated in Figure 2 with networks generated from a SBM with $n = 100$ nodes and $K = 2$ communities, with the first 50 nodes assigned to the first community and the rest to the second. The probability of an edge within the same community is 0.05, and between different communities 0.01. Figure 2 shows the first two eigenvectors of the Laplacian before regularization (left) and after adding $0.1d/n$ to every entry of the adjacency matrix (right). Spectral clustering clearly fails without regularization (mislabeling 49% of the nodes), but after regularization, communities are evident in the signs of entries of the second eigenvector (clustering error is reduced to 9%). As with covariance estimation, the core ridge idea of shrinking eigenvalues has found uses in modern applications far beyond its original design.

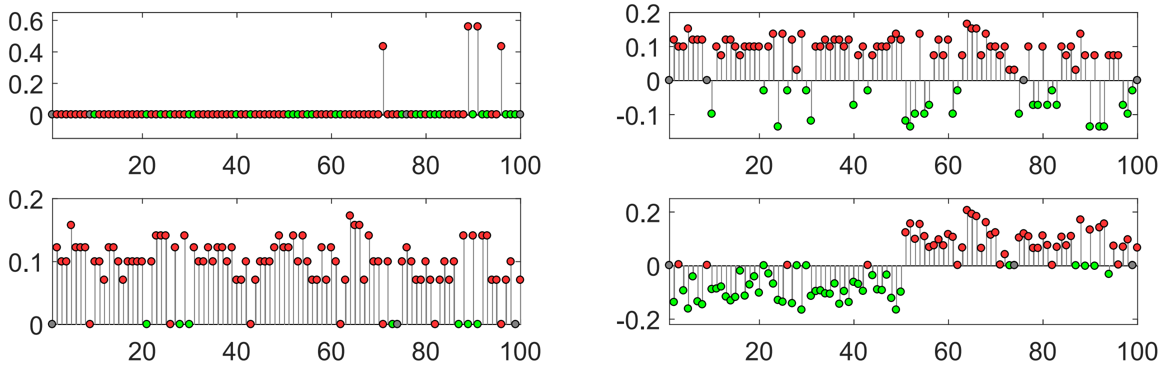

Figure 2: Two leading eigenvectors of the Laplacian (left, $\lambda_1 = \lambda_2 = 1$) and regularized Laplacian (right, $\lambda_1 = 1, \lambda_2 = 0.92$), for an SBM with two communities.

References

E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(16-480):1–86, 2018.

A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. Fitting community models to large sparse networks. *Annals of Statistics*, 41(4):2097–2122, 2013.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

F. Benaych-Georges, C. Bordenave, and A. Knowles. Spectral radii of sparse random matrices. *arXiv:1704.02945*, 2017.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008a.

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008b.

B. Bollobas, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures and Algorithms*, 31:3–122, 2007.

T. T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

T. T. Cai, Z. Ren, and H. H. Zhou. Estimating structured high-dimensional covariance and precision matrices:optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10:1–59, 2016.

K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 23:35.1 – 35.23, 2012.

Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10), 2010.

P. Chin, A. Rao, and V. Vu. Stochastic block model and community detection in the sparse graphs : A spectral algorithm with optimal rate of recovery. *Proceedings of Machine Learning Research*, 40:391–423, 2015.

D. Donoho, M. Gavish, and I. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, 46(4):1742–1778, 2018.

N. El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36(6):2757–2790, 2008.

T. J. Fisher and X. Sun. Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics & Data Analysis*, 55(5): 1909–1918, 2011.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, June 2009.

I. T. Jolliffe, N T. Trendafilov, and M. Uddin. A modified principal component technique based on the LASSO. *Jour. Comp. Graph. Statist.*, 12(3):531–547, 2003.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, 11:517–553, Mar 2010.

C. M. Le, E. Levina, and R. Vershynin. Concentration and regularization of random graphs. *Random Structures and Algorithms*, 51(3):538–561, 2017.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

A. Lodhia, K. Levin, and E. Levina. Matrix means and a novel high-dimensional shrinkage phenomenon. *arXiv:1910.07434*, 2019.

V. A. Marčenko and L. A. Pastur. Distributions of eigenvalues of some sets of random matrices. *Matematicheskii Sbornik*, 1:507–536, 1967.

Roberto Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv:0911.0600*, 2010.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

A. Schwartzman. Lognormal distributions and geometric averages of symmetric positive definite matrices. *International Statistical Review*, 84(3):456–486, 2016.

S. T. Smith. Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630, 2005.

C. Stein. Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, 34:1373–1403, 1986.

A. Touloumis. Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Computational Statistics & Data Analysis*, 83:251–261, 2015.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

J.-H. Won, J. Lim, S.-J. Kim, and B. Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society Series B*, 75(3):427–450, 2013.

W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

J. Yao, S. Zheng, and Z. Bai. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal components analysis. *Journal of Computational and Graphical Statistics*, 15:265–286, 2006.