Sparse Covariance and Precision Random Design Regression

Xi Fang, Steven Winter, Adam B Kashlak*

Abstract Linear regression for high dimensional data is an inherently challenging problem with many solutions generally involving some structural assumption on the model such as lasso's sparsity in the parameter vector. Considering the random design setting, we apply a different sparsity assumption: sparsity in the covariance or precision matrix of the predictors. Thus, we propose a new regression estimator by first applying methods for estimating a sparse covariance or precision matrix. This matrix is then incorporated into the estimator for the regression parameters. We mainly compare this methodology against the classic ridge or Tikhonov regularization method.

1 Introduction

Linear regression is a backbone of statistical methodology. The classical least squares approach has a simple and elegant theory, but fails in the high dimensional setting where the number of parameters p is greater than the sample size n. High dimensional datasets have led to over 40 years of research resulting in methods such as ridge regression, lasso, elastic net, SCAD, and many others [4]. In this work, we contribute to the compendium of such methods by constructing an estimator for high dimensional regression models making use of sparse covariance and sparse precision matrix estimators in the random design setting.

Sparsity is not new to linear regression. The renowned lasso estimator [18] is one of the most important recent contributions to statistics and mathematics. However,

Xi Fang
University of Alberta, Edmonton, Alberta Canada T6G 2G1, e-mail: xfang@ualberta.ca

Steven Winter
University of Alberta, Edmonton, Alberta Canada T6G 2G1, e-mail: szwinter@ualberta.ca

Adam B Kashlak (corresponding author)
University of Alberta, Edmonton, Alberta Canada T6G 2G1, e-mail: kashlak@ualberta.ca

the assumption of lasso is sparsity in the parameters and is used for model selection. In contrast, we consider sparsity in the covariance or precision matrix of the random design matrix X. That is, we assume most off-diagonal entries to be zero and construct a regression estimator under this assumption.

Remark 1 (Sparse Covariance and Precision Matrices). Though being inverses of one another, a sparse covariance and a sparse precision matrix result in two different implications for the underlying data. The covariance matrix considers the marginal correlation between each pair of random variables. A zero entry implies that there is no linear relation between these two variables and in the Gaussian case implies pairwise independence. The precision matrix considers the conditional correlation structure of the data. A zero entry implies that the two variables are uncorrelated—independent when Gaussian—conditioned on the remaining random variables. Thus, the precision matrix defines a network structure and is useful in the study of Gaussian graphical models. Sparse covariance and precision matrices arise in many high dimensional datasets such as genomics, climate, and socioeconomics.

2 Estimator Construction

We begin with a set of n predictor-response pairs (y_i, x_i), $i = 1, \ldots, n$ with $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ and assume that $x_{1j}, \ldots, x_{nj}$ for $j = 1, \ldots, p$ and the $y_1, \ldots, y_n$ are centred as is common in the penalized regression literature. The standard theory of the least squares estimator when $p < n$ for the linear model,

$$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

with unknown parameters β and mean-zero errors ε_i, yields the *ordinary least squares* (OLS) estimator $\hat{\beta}^{\text{ols}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ for design matrix X with ijth entry x_{ij} and $Y = (y_1, \ldots, y_n)$. Under random design [3, 10], the rows of X are treated as iid random vectors, and the least squares loss $L_{\text{ols}}(\tilde{\beta}) = \mathrm{E}\|Y - X\tilde{\beta}\|_{\ell^2}^2$ is minimized by $\beta = (\mathrm{E}[X^{\mathrm{T}}X])^{-1}\mathrm{E}\left(X^{\mathrm{T}}Y\right)$ where we write $\Sigma = n^{-1}\mathrm{E}[X^{\mathrm{T}}X]$, the $p \times p$ covariance matrix for the rows of X. We denote the ith row of X to be x_i.

When $p > n$, the standard covariance estimator $\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} x_i^{\mathrm{T}} x_i$ is known to be far from Σ and furthermore not full rank and thus not invertible. The classic ridge regression solution—also called Tikhonov regularization—considers the ℓ^2 penalized least squares loss

$$L_{\text{R}}(\tilde{\beta}) = \mathrm{E}\|Y - X\tilde{\beta}\|_{\ell^2}^2 + \lambda\|\tilde{\beta}\|_{\ell^2}^2$$

with solution $\hat{\beta}^{\text{R}} = (X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}Y$ for $p \times p$ identity matrix I_p. This estimator is known to *shrink* the values of $\hat{\beta}^{\text{R}}$ towards zero adding some bias for a sizeable reduction in the estimator's variance. In Section 3 below, we compute the ridge estimator via the `glmnet` package in R [9].

Let $r = \min\{n, p\}$ be the rank of X. The singular value decomposition for the $n \times p$ design matrix can be written as $X = UDV^{\mathrm{T}}$ where D is the $r \times r$ diagonal matrix of singular values, and U and V are $n \times r$ and $p \times r$ matrices, respectively. Thus, for high dimensional data, $r = n$ and U is orthonormal whereas $V^{\mathrm{T}}V = I$ but $VV^{\mathrm{T}} \neq I$ and is, in fact, a projection onto the n-dimensional row space of X, a linear subspace of $\mathbb{R}^p$. Under this decomposition and the notion of a pseudo-inverse, the OLS estimator can be extended to high dimensional data as

$$\hat{\beta}^{\mathrm{ols}} = VD^{-2}V^{\mathrm{T}}VDU^{\mathrm{T}}Y = VD^{-1}U^{\mathrm{T}}Y.$$

Similarly, the ridge estimator becomes $\hat{\beta}^{\mathrm{R}} = V(D^2 + \lambda I)^{-1}DU^{\mathrm{T}}Y$. Note that the squared singular values $d_1^2, \ldots, d_n^2$ on the diagonal D^2 are the estimated non-zero eigenvalues for the covariance matrix Σ. Hence, this regularization method is augmenting the eigenvalues by adding λ to each to get $d_i^2 + \lambda$ for $i = 1, \ldots, n$ and just λ for $i = n+1, \ldots, p$. Therefore, the ridge estimator is, in fact, shrinking the estimated eigenvalues for the precision matrix Σ^{-1} to zero as $\lambda \to \infty$. This, in turn, takes $\hat{\beta}^{\mathrm{R}}$ to zero.

The ridge estimator replaces $X^{\mathrm{T}}X$ with $X^{\mathrm{T}}X + \lambda I$. Building off of this inspiration, our proposed methodology is to replace $X^{\mathrm{T}}X$ with a sparse covariance estimator, or replace the undefined $(X^{\mathrm{T}}X)^{-1}$ with a sparse precision matrix estimator.

2.1 *Replacing* $X^{\mathrm{T}}X$

Under the sparsity assumption stated in the introduction for either the covariance or precision matrix, we could consider an alternative estimator being inspired by the Naive-Bayes method. Specifically, we could compute $\hat{\Sigma}^{\mathrm{diag}}$ being the empirical covariance estimator from before with all off-diagonal entries set to zero. Then as this matrix is invertible, we can consider the estimator $\hat{\beta}^{\mathrm{NB}} = (n\hat{\Sigma}^{\mathrm{diag}})^{-1}X^{\mathrm{T}}Y$ where $\hat{\Sigma}^{\mathrm{diag}}$ is multiplied by n to undo the normalization in the covariance estimator. If we were to normalize the data such that $n\hat{\Sigma}^{\mathrm{diag}} = I_p$, then our estimator would be merely $X^{\mathrm{T}}Y$ or equivalently the ℓ^2 inner products between Y and each of the p columns of X.[1] However, the removal of all off-diagonal entries may be too extreme of a methodology. Instead, we relax away from such a diagonal-only estimator by considering sparse estimators for Σ and Σ^{-1} in the following subsections. However, we first take a look at the implications of replacing $X^{\mathrm{T}}X$ with a different positive definite matrix M.

Let M be a positive definite symmetric matrix with eigen-decomposition WSW^{T} for W the orthonormal matrix of eigenvectors and S the diagonal matrix of eigenvalues. We wish to write a new regression estimator $\hat{\beta}^{\mathrm{M}} = M^{-1}X^{\mathrm{T}}Y$. However, this will initially fail as the first n eigenvectors in W will (most likely) not coincide with the n left singular vectors V of X resulting in a nonsensical estimator. Thus, we have to

[1] The estimator $X^{\mathrm{T}}Y$ occurs in practice in orthogonal experimental designs when X is chosen such that $X^{\mathrm{T}}X = I_p$ assuming $p < n$. [19]

rotate the entire problem. Let W_n be the $p \times n$ matrix consisting of the first n columns of W, and let S_n be the $n \times n$ diagonal matrix with the n principal eigenvalues of M on the diagonal. Replacing

$$X \Rightarrow Z := XVW_n^{\mathrm{T}} \text{ and } \beta \Rightarrow \beta^\star := W_nV^{\mathrm{T}}\beta \tag{1}$$

gives the rotated model $Y = X\beta + \varepsilon = Z\beta^\star + \varepsilon$. The new estimator making use of M is

$$\hat{\beta}^{\mathrm{M}} = M^{-1}Z^{\mathrm{T}}Y = W_nS_n^{-1}DU^{\mathrm{T}}Y, \tag{2}$$

which is an estimator for the rotated parameter vector $\beta^\star$. Note that in Section 3, we compare a variety of such estimators in mean squared error. As such transformations as in Equation 1 are isometries, we can still compare mean squared errors estimated over many random simulations as well as the mean squared prediction error for the forest fire and Arizona crime data.

Remark 2. In the above Equations 1 and 2, we could instead consider $p \times p$ orthonormal matrices $\tilde{W}$ and $\tilde{V}$ being the eigenvectors of M and $X^{\mathrm{T}}X$, respectively. Computationally, the resulting estimator will be equivalent as the eigenvalues corresponding to those $p-n$ additional columns will be zero. Hence, any rotation in those directions will not affect the estimator. The above formulation is more computationally efficient by ignoring these extraneous directions.

2.1.1 Sparse Covariance Estimation

There is a vast literature on sparse covariance matrix estimators for high dimensional data. Two broad approaches are penalized estimators [2, 16] and threshold estimators [1, 17, 5]. The latter methods apply a threshold function entrywise to the off-diagonal entries of the empirical covariance matrix, which effectively sets entries below a specified threshold to zero. A threshold is typically chosen via cross validation. As sample sizes are typically small and cross validation is furthermore computationally expensive, a threshold can also be selected by choosing a suitable individual false positive rate $\alpha \in [0,1]$ being the probability that an off-diagonal entry is falsely included in the support of the estimator—i.e. the probability that the ijth entry in the estimator is not zero given that $\Sigma_{ij} = 0$ [14]. This α acts as a regularization parameter. Indeed, this estimator allows us to relax away from the above naive-Bayes estimator, which would correspond to $\alpha = 0$, by increasing α to allow for a few off-diagonal entries to be non-zero. Such estimators are computed via the R package `sparseMatEst` [13].

Remark 3 (Positive Definiteness). Even though the empirical covariance estimator is positive semi-definite, a thresholded covariance estimator may no longer be. To rectify this problem, let $\hat{\Sigma}^{(\alpha)}$ be a sparse covariance estimator with false positive rate α, and denote the eigenvalues of $\hat{\Sigma}^{(\alpha)}$ in decreasing order to be $\lambda_1^{(\alpha)} \geq \ldots \geq \lambda_n^{(\alpha)}$. Then assuming $\lambda_n^{(\alpha)} < 0$ we add $\{|\lambda_n^{(\alpha)}| + \lambda_1^{(\alpha)}/100\}I_p$ to $\hat{\Sigma}^{(\alpha)}$ to make the new estimator positive definite with a condition number of 100.

2.1.2 Sparse Precision Estimation

The most famous method of sparse precision matrix estimation is the graphical lasso [8], but other regularized estimators also exist [6]. Unlike for covariance matrices, threshold estimation of the precision matrix is more challenging as there is no unbiased estimator for Σ^{-1} threshold. However, [12] applies the same idea of individual false positive rate control by thresholding the debiased glasso estimator of [11]. This precision matrix estimation method is also implemented in the R package `sparseMatEst` [13].

3 Numerical Results

3.1 Simulated Data

In this section, we test the following estimators of the form $\hat{\beta}^{\mathrm{M}} = M^{-1}Z^{\mathrm{T}}Y$ from Equation 2. For M, we consider threshold based sparse covariance estimators from [14] and the standard ridge estimator. For M^{-1}, we consider threshold based sparse precision estimators from [12] as well as the graphical lasso [8]. To gauge the success of each estimator, we estimate the normalized mean squared error,

$$\mathrm{MSE}(\tilde{\beta}) = \|\tilde{\beta} - \beta\|_{\ell^2}^2 / \|\beta\|_{\ell^2}^2,$$

over 100 replications for $\beta = (1, \ldots, 1)$, the rows of X being iid $\mathcal{N}(0, \Sigma)$ for some sparse Σ discussed below, and $Y = X\beta + \varepsilon$ with iid $\varepsilon_i \sim \mathcal{N}(0, 4)$.

Figure 1 contains the results—estimated log base-2 mean squared errors—for such simulations for Σ tridiagonal with main diagonal 1 and off-diagonal entries 0.4 and contains results for Σ banded with main diagonal 1 and three off-diagonals with values $3/4$, $1/2$, and $1/4$. Since these methods normalize the variance of the predictors before penalizing, we only consider settings where all diagonal entries are 1. For all methods, many choices of the tuning parameter—$\alpha \in [0, 1]$ for sparse covariance and $\lambda \geq 0$ for ridge and glasso—were considered and the best was taken. Hence, Figure 1 displays results for optimal choice in tuning parameter. In both cases, the performance of the sparse covariance methodology was on par with that of ridge regression, and both of these outperformed the graphical lasso.

This methodology was also tested for sparse precision matrices—i.e. rerunning the above simulations but specifying Σ^{-1} to be tri-diagonal or banded as opposed to Σ. In that setting, the sparse precision methodology performed much more poorly than ridge regression. Hence, those results are not included. The answer as to why the sparse matrix-based regression estimator succeeds for covariance matrices but not for precision matrices remains illusive. However, good performance of the precision estimator is observed in Section 3.3.

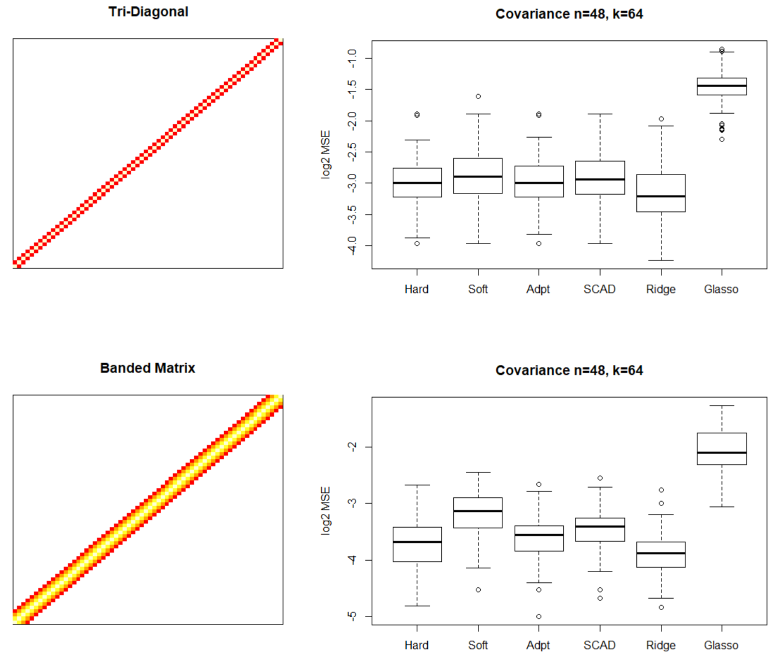

Fig. 1 Top Row: Results of methods given a tridiagonal covariance matrix. Bottom Row: Results of methods given a banded covariance matrix. Here, $n = 48$, $p = 64$, and the plots were the result of 100 replications with $\beta = (1, \ldots, 1)$.

3.2 Forest Fire Data

For a first real data application, we consider the mean squared prediction errors (MSPE) for different regression methods on the Portuguese forest fire data [7] available online on the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). We aim to predict the $\log(\text{Area Burned})$ based on a variety of indices and weather measurements: the coordinates of the location of the fire, the Fine Fuel Moisture Code (FFMC), the Duff Moisture Code (DMC), the Drought Code (DC), the Initial Spread Index (ISI), the outside temperature, the relative humidity, and the wind speed. For more details, see [7]. Only fires whose total area burned was greater than zero were considered due to the log-transform, which was necessary due to the extreme skewness of the `area` data. Thus, we have $n = 270$ and $p = 9$. Though, this is not high dimensional data, there is strong collinearity among the predictors warranting the use of shrinkage estimators.

To compute the MSPE, we randomly split the data into training and testing sets of sample size $n_{\text{train}} = 225$ and $n_{\text{test}} = 45$, respectively, to fit the model and then compute

$$\text{MSPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (A_i - \hat{A}_i)^2$$

for A_i the total log-area burned for the ith observation of the test set and $\hat{A}_i$ the ith predicted value. This was averaged over 1000 replications with randomly selected training and testing sets.

Figure 2 displays the results of our methods. The support of the replacement matrix for $X^{\mathrm{T}}X$ is considered on the left for different values of α. Note that for $\alpha = 0.5$, most of the off-diagonal entries have already been removed. The MSPE on the right is considered for sparse covariance matrices with four different types of thresholds: Hard, Soft, Adaptive Lasso, and SCAD thresholding. More details on these can be found in [17, 14]. Ridge regression is also included whereas sparse precision and glasso methods are excluded due to their poor performance on simulated data. Most notably, the methods all return similar MSPE for this dataset, but hard and scad thresholding are the most robust with respect to choice in tuning parameter compared to the other methods.

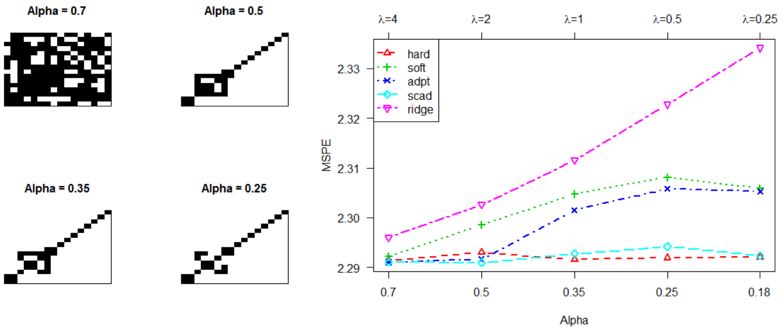

Fig. 2 On the left, the support in black of the thresholded $X^{\mathrm{T}}X$ matrix for different α for the forest fire data [7]. On the right, the mean squared prediction error computed over 1000 replications for 5 different estimators with 5 different values of their respective tuning parameter—α on the bottom for the sparse covariance and λ on the top for ridge regression.

3.3 *Arizona Crime Data*

We secondly repeat the previous analysis on the *Communities and Crime Data Set* also from the UCI Machine Learning Repository and discussed in [15]. This dataset contains potential predictors of violent crime collected across the USA. For the sake of our methodology, we only considered the $n = 20$ observations taken from the state of Arizona. There are $p = 99$ predictors in this dataset. The dataset was randomly split into $n_{\text{train}} = 13$ and $n_{\text{test}} = 7$ and the MSPE was computed over 1000 replications.

Figure 3 displays the results for the precision matrix estimator, which performed better than the covariance-based approach. This is reasonable as many predictors may be correlated—e.g. number of homeless shelters and number of vacant houses—but conditionally uncorrelated—e.g. taking median income into account. Here, all precision thresholding methods had very similar MSPE. In contrast, the ridge estimator either performed better or worse depending on choice of λ; though the scale of the vertical axis indicates that ridge regression only achieves a slightly better MSPE after careful tuning of λ.

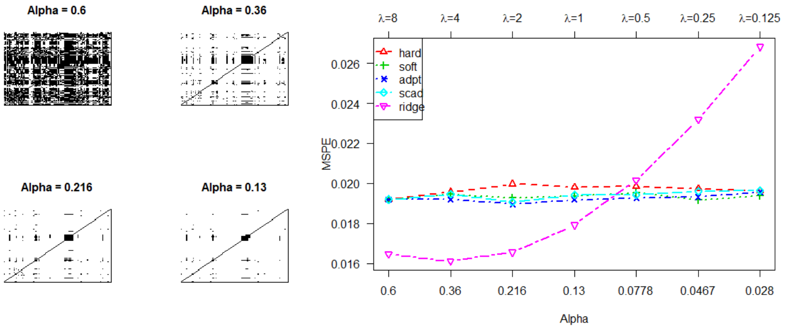

Fig. 3 On the left, the support in black of the thresholded inverse $X^{\mathrm{T}}X$ estimator for different α for the Arizona crime data [15]. On the right, the mean squared prediction error computed over 1000 replications for 7 different estimators with 7 different values of their respective tuning parameter—α on the bottom for the sparse covariance and λ on the top for ridge regression.

4 Discussion

In this article, we proposed an alternative estimator for the parameters of a high dimensional regression model under the random design setting where it is assumed that the rows of the design matrix have a sparse covariance or sparse precision structure. Such structural assumptions do occur in real data problems and are distinct from the usual notation of regression sparsity—that is, sparsity in the parameter vector β.

The result of multiple simulation experiments, beyond what is detailed in Section 3, indicate that using a sparse covariance estimator in place of $X^{\mathrm{T}}X$ can achieve similar but no superior results to that of standard ridge regression. Replacing $(X^{\mathrm{T}}X)^{-1}$ with a sparse precision estimator or the classic graphical lasso estimator did not yield good performance in contrast to ridge regression for simulated data. However, we did see strong performance on the Arizona crime dataset.

The success of this methodology does warrant further investigations into such methods considering how they can be improved and if there are scenarios where they can outperform standard ridge regression. Even though the performance of ridge regression was comparable to our methodology, it did not perform significantly better. Also, our method appears more robust to choice of tuning parameter meaning one can achieve similar performance to ridge regression without the need to carefully tune λ.

Acknowledgments

The authors would like to thank Dr. Xu (Sunny) Wang from Wilfrid Laurier University and Dr. Yan Yuan from the University of Alberta for organizing the special session on Interdisciplinary Data Analysis of High-Dimensional Multimodal Data at AMMCS 2019 where this work was presented. We would also like to thank the comments of the anonymous reviewers who helped improve this work.

References

[1] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

[2] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[3] L. Breiman and D. Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381): 131–136, 1983.

[4] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

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

[6] T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[7] P. Cortez and A. d. J. R. Morais. A data mining approach to predict forest fires using meteorological data. 2007.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[9] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[10] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1, 2012.

[11] J. Jankova and S. Van De Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

[12] A. B. Kashlak. Non-asymptotic error controlled sparse high dimensional precision matrix estimation. *arXiv preprint arXiv:1903.10988*, 2019.

[13] A. B. Kashlak. *sparseMatEst: Sparse Matrix Estimation and Inference*, 2019. URL https://CRAN.R-project.org/package=sparseMatEst. R package version 1.0.0.

[14] A. B. Kashlak and L. Kong. A concentration inequality based methodology for sparse covariance estimation. *arXiv preprint arXiv:1705.02679*, 2017.

[15] M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[16] A. J. Rothman. Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740, 2012.

[17] 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.

[18] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[19] C. J. Wu and M. S. Hamada. *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons, 2011.