Non-asymptotic error controlled sparse high dimensional precision matrix estimation

Adam B Kashlak
Mathematical & Statistical Sciences
University of Alberta
Edmonton, Canada, T6G 2G1
March 27, 2019

Abstract

Estimation of a high dimensional precision matrix is a critical problem to many areas of statistics including Gaussian graphical models and inference on high dimensional data. Working under the structural assumption of sparsity, we propose a novel methodology for estimating such matrices while controlling the false positive rate–percentage of matrix entries incorrectly chosen to be non-zero. We specifically focus on false positive rates tending towards zero with finite sample guarantees. This methodology is distribution free, but is particularly applicable to the problem of Gaussian network recovery. We also consider applications to constructing gene networks in genomics data.

1 Introduction

Attempting to estimate the graphical structure of a high dimensional network is problemsome when edges are rare but the number of nodes is large. In standard statistical classification problems, we fix a palatable false positive rate and aim to recover as many true positives as possible. Thus, we treat support recovery as a binary classification problem. For $p$ nodes, the number of potential edges to consider is on the order of $p^2$. With a sample size $n < p$, there are too many parameters to accurately estimate. However, when trying to classify edges as significant or not under the assumption of sparsity, the assumption that most edges are not significant, if we have even a small false positive rate, then this will result in many erroneous connections potentially obscuring follow up research. For example, a genomics study considering the conditional correlation structure of, say, 2000 genes will have to consider over two million potential edges. A standard 1% false positive rate will result in tens of thousands of erroneous connections.

To address this issue, we consider the extreme estimation setting of recovering the support of a precision matrix $\Omega$ in the high dimensional setting, $p \gg n$, assuming $\Omega$ belongs to a class of sparse positive definite matrices, and for false positive rates $\alpha \to 0^+$. Making use of debiased estimators (Bühlmann, 2013; Javanmard & Montanari, 2014; Van de Geer et al., 2014; Jankova & Van De Geer, 2015), the non-asymptotic results of Kashlak & Kong (2017), and a clever subsampling methodology, we can achieve finite sample guarantees in this extreme setting.

There has been much past work on covariance and precision matrix estimation much of which is summarized in the survey Fan et al. (2016). Most estimators for high dimensional precision matrices are based on $\ell^1$ penalization including the graphical lasso (Friedman et al., 2008), the CLIME and ACLIME estimators (Cai et al., 2011, 2016), and the debiased estimator of Jankova & Van De Geer (2015) used for constructing confidence sets and running hypothesis tests. These articles all rely on high dimensional asymptotics, which is that estimation is successful in the limit as $p$ and $n$ grow to infinity together generally such that $(\log p)/n$, or some variant thereof, is $o(1)$. Our work differs as it considers guaranteed results for controlling the support recovery of $\Omega$ as the false positive rate $\alpha \to 0$ for fixed finite $p \gg n$, which is asymptotic in $\alpha$, a controllable
tuning parameter, instead of in \( n \) and \( p \), which are generally fixed in the real world experimental setting.

Our method can be deconstructed into two steps. The first is to find a bias-corrected initial estimate \( \hat{\Omega}_0 \) for the precision matrix \( \Omega \). The second is to construct a ball in the operator norm topology corresponding to the target false positive rate and then search this ball for a sparse estimator. In actuality, given a false positive rate of \( \alpha \), we step down in a binary fashion constructing successive false positive rates of \( \gamma^{-1} \), \( \gamma^{-2} \),..., and a sequence of estimators tending towards one with the desired \( \alpha \). Depending on the sample size, we can randomly partition our data set into smaller sets, apply this two-step procedure to each subsample in parallel, and combine the results for improved performance. In step two, we use a binary search procedure which converges rapidly minimizing the computational burden.

Our search methodology is effectively a method for controlled thesholding of the initial debiased estimator. The literature on thresholding for precision matrices is quite light when compared to thresholding covariance matrices (Bickel & Levina, 2008a,b; Rothman et al., 2009; Cai & Liu, 2011; Kashlak & Kong, 2017). The challenge is that unlike the unbiased empirical covariance estimator, there is no unbiased empirical precision estimator. Debiasing such precision matrix estimators allows for thresholding as mentioned in Jankova & Van De Geer (2015). Furthermore, Jankova & Van De Geer (2015) proposed an entrywise thresholding method given some false positive rate \( \alpha \), which while working well in some high dimensional asymptotic sense, proved to not achieve the target false positive rate on simulated data.

2 Methodology

2.1 Notation

In this article, we primarily make use of the family of Schatten norms. For \( \Omega \in \mathbb{R}^{p \times p} \) with \( i,j \)th entry \( \Omega_{i,j} \) and eigenvalues \( \lambda_1, \ldots, \lambda_p \), we define the trace norm \( \| \Omega \|_1 = \sum_{i=1}^{p} |\lambda_i| \), the Hilbert-Schmidt norm \( \| \Omega \|_2 = (\sum_{i=1}^{p} |\lambda_i|^2)^{1/2} = (\sum_{j=1}^{p} \Omega_{i,j}^2)^{1/2} \), and the operator norm \( \| \Omega \|_{\infty} = \max_{i=1,\ldots,p} |\lambda_i| \). We will also use the standard \( \ell^p \) norms applied to vectors in \( \mathbb{R}^p \) and entrywise to matrices in \( \mathbb{R}^{p \times p} \).

The algorithm in Section 2.3 makes use of hard thresholding. For \( \Omega \in \mathbb{R}^{p \times p} \), we define the hard thresholding operator \( \varphi : \mathbb{R}^{p \times p} \times \mathbb{R}^+ \rightarrow \mathbb{R}^{p \times p} \), which returns a matrix with \( i,j \)th entry

\[
\varphi(\Omega; t)_{i,j} = \begin{cases} 
0 & \text{if } |\Omega_{i,j}| < t \\
\Omega_{i,j} & \text{if } |\Omega_{i,j}| \geq t
\end{cases}
\]

which simply removes entries from \( \Omega \) with magnitude less than \( t \).

2.2 Initial Estimator

Let \( X_1, \ldots, X_n \in \mathbb{R}^p \) be an independent and identically distributed collection of mean zero random vectors with common unknown positive definite covariance matrix \( \Sigma \) and corresponding precision matrix \( \Omega = \Sigma^{-1} \). Of course, we require the covariance and its inverse to exist, but put no further assumptions on the distribution of the \( X_i \) at the moment except to assume that \( \Omega \) is sparse. Specifically,

\[
\Omega \in \mathcal{U}(\kappa, \delta) = \left\{ A \in \mathbb{R}^{p \times p} : A = A^T, \sum_{i=1}^{p} 1(A_{i,j} \neq 0) \leq \kappa \text{ for } j = 1, \ldots, p, \text{ if } A_{i,j} \neq 0 \text{ then } |A_{i,j}| \geq \delta \right\}
\]

where \( \mathcal{U}(\kappa, \delta) \) is the class of sparse \( \mathbb{R}^{p \times p} \) matrices with no more than \( \kappa \) entries in each row or column non-zero and with non-zero entries bounded away from zero allowing them to be detectable. In practice, we attempt to recover the support of the normalized matrix \( \Omega \) with diagonal entries of 1 to avoid scale issues.
Our approach is similar to Kashlak & Kong (2017) who attempt to recover the support of a covariance matrix $\Sigma$ by starting with an initial estimator, constructing a confidence ball, and searching this ball for a sparser estimator. However, whereas this preceding work can use the empirical estimator,

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} X_i X_i^T,$$

as an unbiased initial estimator for controlled thresholding, we cannot construct an unbiased estimator for the precision matrix $\Omega$ in the $p > n$ setting.

A standard estimator for sparse precision matrices is the graphical lasso (Friedman et al., 2008), which is based on an $\ell_1$ penalized maximum likelihood under the Gaussian distribution:

$$\hat{\Omega}^{GL} = \arg \min_{\Theta \in \mathbb{R}^{p \times p}} \left\{ \text{tr} (\hat{\Sigma} \Theta) - \log \det(\Theta) + \lambda \sum_{i=1}^{p} \sum_{j \neq i} |\Theta_{i,j}| \right\}$$

for some tuning parameter $\lambda > 0$. The graphical lasso is debiased in Jankova & Van De Geer (2015), which uses a correction factor based on the subgradient of the above optimization extending the work of Van de Geer et al. (2014) that debiases the classic lasso estimator for linear regression. The resulting debiased estimator is

$$\hat{\Omega}^{deGL} = 2\hat{\Omega}^{GL} - \hat{\Omega}^{GL} \hat{\Sigma} \hat{\Omega}^{GL}.$$

Alternatively, the CLIME method (Cai et al., 2011) solves a constrained $\ell_1$ optimization problem to find $\hat{\Omega}^{CL}$.

$$\min_{\Theta \in \mathbb{R}^{p \times p}} \|\Theta\|_{\ell_1} \text{ such that } \max_{i,j} |\hat{\Sigma} \Theta - I_p|_{i,j} \leq \lambda$$

with a more sophisticated version ACLIME (Cai et al., 2016) adapting to individual entries.

A different regularized estimator is the ridge estimator,

$$\hat{\Omega}^{RD} = \arg \min_{\Theta \in \mathbb{R}^{p \times p}} \left( \|\hat{\Sigma} \Theta - I_p\|_2^2 + \lambda \|\Theta\|_2^2 \right),$$

whose closed form solution is $\hat{\Omega}^{RD} = (\hat{\Sigma} + \lambda I_p)^{-1}$. Consider the singular value decomposition $X = UDV^T$ where $U \in \mathbb{R}^{n \times r}, D \in \mathbb{R}^{r \times r}$, and $V \in \mathbb{R}^{p \times r}$ for $r = \min(n, p)$. In Bühlmann (2013), a bias corrected estimator for ridge regression is proposed. This is achieved by finding some other estimator, which is used to correct for the projection bias. We can apply the same method to the ridge precision matrix estimator to get

$$\hat{\Omega}^{deRD} = \hat{\Omega}^{RD} + P_0 \hat{\Omega},$$

where $\hat{\Omega}$ is some other estimate for $\Omega$ and $P_0 = VV^T - \text{diag}(VV^T)$ where $VV^T$ is the projection in $\mathbb{R}^p$ onto an $n$ dimensional subspace spanned by the columns of $X$, and $P_0$ is that projection with the diagonal entries set to zero. Bühlmann (2013) uses the graphical lasso estimator for $\Omega$.

### 2.3 Binary Search

Given an initial estimator from the previous setting henceforth denoted $\hat{\Omega}_0$ for simplicity of notation and some $\gamma > 1$ such that $\alpha = \gamma^{-m}$ for some positive integer $m$ corresponding to the number of iterates of the below algorithm, we aim to construct an estimator $\Omega_m$ with false positive rate $\alpha = \gamma^{-m}$ by carefully thresholding the entries in $\hat{\Omega}_0$ so that

$$\frac{\left| \{(i,j) : (\hat{\Omega}_m)_{i,j} \neq 0 \text{ and } (\Omega)_{i,j} = 0, i \neq j \} \right|}{\left| \{(i,j) : (\Omega)_{i,j} = 0, i \neq j \} \right|} \approx \alpha,$$

which is that the desired false positive rate is achieved. The two extreme estimators are the initial estimator $\hat{\Omega}_0$ and the diagonal matrix $\Omega_\infty$ with diagonal entries coinciding with $\hat{\Omega}_0$ and off-diagonal entries set to zero. These correspond to the 100% and 0% false positive cases, respectively. For the remainder, we normalize $\hat{\Omega}_0$ to have unit diagonal thus making $\Omega_\infty = I_p$. 

3
For values of $\alpha$ tending towards zero, we consider the operator norm balls centred at $I_p$ being $B_\alpha = \{ \Pi \in \mathbb{R}^{p \times p} : \|\Pi - I_p\|_\infty \leq \|\hat{\Omega}_m - I_p\|_\infty \}$ where $\|\cdot\|_\infty$ refers to the operator norm for $\ell^2(\mathbb{R}^p) \to \ell^2(\mathbb{R}^p)$, which is the principal eigenvalue. This motivates the following algorithm:

**Constructing an error controlled estimator**

Begin with an initial estimator from Section 2.2 denoted $\hat{\Omega}_0$.

Given the $s$th iterated estimator $\hat{\Omega}_s$, we construct the $(s + 1)$th estimator

1. Compute $r_s = \|\hat{\Omega}_s - I_p\|_\infty$.
2. Compute $r'_s = r_s \gamma^{-1/2}$.
3. Find $t_{s+1} = \min(t)$ such that $\|\varphi(\hat{\Omega}_s; t) - I_p\|_\infty \leq r'_s$
   where $\varphi$ is the hard threshold operator.
4. Set $\hat{\Omega}_{s+1} = \varphi(\hat{\Omega}_s; t_{s+1})$
5. Repeat step 2 with $\hat{\Omega}_{s+1}$ until $s = m$

In this algorithm, we quickly locate the densest estimator close to $I_p$, alternatively being the sparsest estimator close to $\hat{\Omega}_0$, by using a binary search procedure. Given the $s$th iterated matrix as a starting point, we set the smallest half of the non-zero entries in magnitude in $\hat{\Omega}_s$ to zero, compute the distance to $I_p$ and then if the distance is greater than $r'_s$, we remove half of the remaining entries whereas if the distance is less than $r'_s$, we reintroduce half of the removed entries.

**Remark 2.1.** For choice of $\gamma > 1$ and number of steps $m$, values of $\gamma$ closer to 1 result in smaller steps adding stability but requiring a larger $m$ to achieve a small false positive rate. In the simulations of Section 3.1, we use $\gamma = 2$.

For the choice of initial estimator from Section 2.2, the best performance in simulated data experiments on multivariate normal data was achieved by beginning the above procedure with the debiased graphical lasso estimator of Jankova & Van De Geer (2015). This is mainly because their estimator follows the requirements of the theorems in the subsequent section as long as the data under analysis is sub-Gaussian. However, good performance is still observed when the data is sub-exponential as can be seen in the supplementary material.

**Remark 2.2.** A similar search algorithm is proposed in Kashlak & Kong (2017). The main differences are that in the cited work, the operator norm balls are confidence sets centred about the identity matrix $I_p$. Furthermore, the radii are reduced by a factor of $\gamma^{-1}$ in that previous work due to the low rank structure of the empirical covariance estimator. Specifically, $E\|\hat{\Sigma}\| = O(p)$ rather than $O(p^{1/2})$.

### 2.4 Theoretical Guarantees

The reason for shrinking the radius by $\gamma^{-1/2}$ in the above algorithm comes from applying tools from random matrix theory (Tao, 2012) to sparse matrices—see the supplementary material for proofs. The following result states that when $\Omega$ is sufficiently sparse and $p \gg n$, we can reduce the false positives by $\gamma^{-1}$ by shrinking the radius of an operator norm ball centred around the $p$-dimensional identity matrix $I_p$ by $\gamma^{-1/2}$.

**Theorem 2.1** (Controlled False Positives). Let $\Omega \in \mathcal{U}(\kappa, \delta)$ with $\kappa = O(p^\nu)$ for $\nu < 1/2$, and $\|\Omega\|_\infty = o(p^{1/2})$. For some false positive rate $\alpha = \gamma^{-s}$ with $s \in \mathbb{Z}^+$, let the bias of the initial estimator be $\|\text{bias}(\hat{\Omega})\|_\infty = \|E\hat{\Omega} - \Omega\|_\infty = o(\gamma^{s}p^{1/2})$. Then,

$$(a) \quad \frac{E\|\hat{\Omega}_{s+1} - \Omega\|_\infty}{E\|\hat{\Omega}_s - \Omega\|_\infty} = \frac{1 + o(1)}{\gamma^{-1/2} + o(1)}$$

$$(b) \quad \frac{E\|\hat{\Omega}_{s+1} - I_p\|_\infty}{E\|\hat{\Omega}_s - I_p\|_\infty} = \frac{1 + o(1)}{\gamma^{-1/2} + o(1)}$$

**Remark 2.3.** To control the false positive rate, we require a few assumptions in the theorem statement. Namely, the number of non-zero entries per row $\kappa$ and the operator norm of $\Omega$ cannot grow at a rate faster than $p^{1/2}$ as $p$ increases. In the simulations of Section 3.1, we have the much nicer setting where these quantities remain bounded as $p$ increases.
Remark 2.4. The accuracy of the conclusion of Theorem 2.1 is improved as $p$ increases for a fixed sample size $n$. This fact motivates the subsampling methodology in Section 2.5. Secondly, as the false positive rate decreases, the bias is allowed to be larger. Hence, we can choose $\gamma^{-s}$ based on $p$. This is because as $s$ increases, we threshold more aggressively effectively fighting against increases in the bias.

As noted in Remark 2.1, we chose the method of Jankova & Van De Geer (2015) for our simulations in Section 3.1. It is shown in their work that this debiased estimator has the following maximal entrywise bias, $b_{\text{max}} = \max_{i,j} \{|\text{bias}(\hat{\Omega})|\} = O(\kappa \log(p)n^{-1/2})$ or $O(\kappa^{3/2} \log(p)n^{-1/2})$ depending on specific assumptions on the sub-Gaussian nature of the data. Hence, $\|\text{bias}(\hat{\Omega})\|_{\infty} = o(b_{\text{max}}p^{1/2}\gamma^{-s})$. As we have control over the false positive rate $\gamma^{-s}$, we can choose this to satisfy the conditions of the above theorem. Namely, considering false positive rates less than $p^{-1/2}$ or $p^{-3/4}$, which are generally of more interest than large false positive rates.

Further considering the debiased estimator described in Jankova & Van De Geer (2015) and given the assumptions made in that article, we can recover the support of the matrix asymptotically as $n,p$ increase and $\alpha \to 0$. Indeed, Lemma 9 from Jankova & Van De Geer (2015) and equivalently Theorem 1 from Ravikumar et al. (2011) assume an irrepresentability condition common in the lasso literature and get convergence rates of the graphical lasso estimator depending on the tail behaviour of the random vectors $X_1, \ldots, X_n$. Thus, the graphical lasso estimator asymptotically recovers the support. Further assuming sub-Gaussian tails for the $X_i$, Jankova & Van De Geer (2015) shows that the remainder term in the debiased graphical lasso estimator is asymptotically negligible. Thus, thresholding the debiased estimator will in turn re-recover the support. Similarly, the literature on sparse covariance matrix estimation generally requires sub-Gaussian tails for asymptotic support recovery (Bickel & Levina, 2008a,b; Rothman et al., 2009; Cai & Liu, 2011; Kashlak & Kong, 2017).

Beyond distributional assumptions, a quick calculation can demonstrate that the true positive probability is necessarily greater than the false positive probability. For some threshold $t_s \in [0,1]$ corresponding to a false positive rate of $\alpha = \gamma^{-s}$, we have assuming symmetry of the distribution of $\hat{\Omega}_{i,j}$ that

$$P(\text{true positive}) = P\left( |\hat{\Omega}_{i,j}| > t_s \mid \hat{\Omega}_{i,j} \neq 0 \right)$$

$$= 2P \left( \hat{\Omega}_{i,j} - \hat{\Omega}_{i,j} > t_s - \hat{\Omega}_{i,j} \mid \hat{\Omega}_{i,j} \neq 0 \right)$$

$$= P \left( |\hat{\Pi}_{i,j}| > t'_s \right) = P(\text{false positive for } t'_s)$$

which is that the probability of a true positive at threshold $t_s$ corresponds to the probability of a false positive at threshold $t'_s > t_s$. Thus, this method is guaranteed to return at least as many true positives proportionally as false positives and generally, as will be seen in Section 3.1, performs much better.

2.5 Subsampling

As Theorem 2.1, becomes more accurate for large $p$, we can run the above methodology in parallel by randomly partitioning the sample of size $n$ into subsamples of size $n/k$. As a result, we run our method in parallel $k$ times returning estimators $\hat{\Omega}^{(1)}, \ldots, \hat{\Omega}^{(k)}$. For a false positive rate of $0 < \alpha < 1$ and the assumption from Theorem 2.1 that $\Omega \in \mathcal{U}(\kappa, \delta)$ with $\kappa = O(p^\nu)$ and $\nu < 1/2$, then the expected number of false positive recoveries is $\alpha[p(p+1)/2 - Cp^{1+\nu}] = O(\alpha p^n)$. After splitting the sample into $k$ disjoint pieces, we can run the above algorithm in parallel to construct $k$ independent estimators $\hat{\Omega}^{(1)}, \ldots, \hat{\Omega}^{(k)}$. The probability of recovering the same false positive entry in $d$ or more of the $\hat{\Omega}^{(i)}$ is a binomial tail sum: $\sum_{i=d}^k \binom{k}{i} \alpha^i (1-\alpha)^{k-i}$. When $\alpha < 0.001$, then choosing $d = 2$ is generally sufficient. Indeed, we can bound this binomial tail sum by the
Kullback-Leibler divergence. For $B \sim \text{Binomial}(k, \alpha)$,

$$P(B \geq d) \leq \exp \left\{ -d \log \left( \frac{d/k}{\alpha} \right) - (k - d) \log \left( \frac{1 - d/k}{1 - \alpha} \right) \right\} \leq k^d \left( \frac{\alpha}{d} \right)^d \left( \frac{1 - \alpha}{k - d} \right)^{k-d} \leq \frac{k^d}{d^d (k - d)^{k - d}} \alpha^d \leq 2^k \alpha^d.$$ 

If we want a target false positive rate of $2^{-10}$, we can choose $\alpha$ such that $2^k \alpha^d = 2^{-10}$ or $\alpha = 2^{-(10+k)/d}$ to achieve the same false positive rate as when not subsampling. For example, if we take $d = 2$ and keep $k < 10$, then we can relax the false positive rate for each individual subsampled estimator.

This addition of subsampling to the methodology allows for faster runtimes, as the $k$ estimators can be computed in parallel with fewer iterations, and also increases accuracy as the target false positive rate decreases below $p^{-1}$ as will be seen in Section 3.1.

3 Applications

3.1 Simulated Data

We test our methodology for three different graphical structures. In the first case, $\Omega$ tridiagonal–i.e. $\Omega_{i,i} = 1$, $\Omega_{i,j} = 1/3$ if $|i - j| = 1$, $\Omega_{i,j} = 0$ if $|i - j| > 1$. This is the sparsest connected graph structure with $p - 1$ edges on $p$ nodes. In the second case, $\Omega$ represents a binary tree graph with depth $d$ resulting in $d(d + 1)/2$ nodes and $d(d - 1)$ edges, which is roughly $2p$ edges for $p$ nodes. Similar to case one, diagonal entries are set to 1 and non-zero off-diagonal entries are set to $1/3$.

In the third case, $\Omega$ represents a disjoint collection of $k$ complete graphs on $d$ nodes. Thus, there are $kd$ nodes and $kd(d + 1)/2$ edges, and $\Omega$ is a block diagonal matrix with blocks $\frac{1}{3}I_d + \frac{2}{3}I_d$ with $I_d$ being the $d \times d$ matrix of all 1’s and $I_d$, the $d \times d$ identity matrix.

For simulations, the sample size was fixed at $n = 50$. The dimensions considered are $p = 496, 1128, 2080$ as they correspond to binary trees of depth 31, 47, 64, respectively, and correspond to block diagonal matrices with $62, 141, 260$ $(8 \times 8)$-blocks, respectively. In all nine cases considered, the non-zero off-diagonal entries are all set to $1/3$ with the diagonal entries set to 1.

In Figure 1, we plot the empirical true and false positive rates on the vertical axis against the target false positive rate on the horizontal axis. The target false positive rate acts as a tunable parameter. Three types of matrices, three dimensions, and three values of the graphical lasso regularization parameter $\lambda = 0.5, 1, 2$ are considered. Ideally, the observed false positive rates will be close to the solid black line, which is where the observed and target rates coincide. This occurs as the dimension increases in all three cases. For the sparser models–tridiagonal and binary tree structures–we are able to maintain a true positive rate above $20\%$ as the false positive rate is taken to zero. Performance is much worse when trying to recover the block diagonal matrix. For the lowest dimension, $p = 496$, setting the penalization parameter $\lambda = 0.5$ gives the best results whereas in the higher dimensions $\lambda = 1$ gives the best performance albeit only slightly. Hence, it appears that for choosing an initial estimator, the graphical lasso penalization parameter can be set to 1 regardless of dimension and true $\Omega$ for this methodology. On this note, we also considered $\lambda = 0.25$ but the performance was terrible and thus it is not included in the figures. For an alternative look at the simulations, receiver operating characteristic curves are included in the supplementary material. The same simulations were also run on multivariate Laplace data with details in the supplementary material. In short, choosing $\lambda = 1$ allowed for the empirical false positive rate to stay close to the target rate whereas $\lambda = 2$ did not perform as well.

In Figure 2, we consider the tridiagonal and binary tree models estimated by the standard approach as before and by subsampling as in Section 2.5 with $k = 2$. By subsampling, we can extend the false positive rate from around $\alpha = p^{-1}$, where the standard method begins to breakdown, all the way to around $\alpha = p^{-2}$. Hence, we can have a controlled reduction in the
empirical false positive rate to effectively zero, as the precision matrix only has $p^2$ entries, using the subsampling method. In Table 1, we tabulate empirically achieved true and false positive rates from the cross validated graphical lasso (Friedman et al., 2018) and ACLIME estimators (Pang et al., 2016). These estimators were computed by splitting the $n = 50$ sample in half and optimizing the tuning parameters with respect to the operator norm distance. The graphical lasso estimator’s achieved false positive was roughly halved as the dimension doubled. The ACLIME estimator much more aggressively penalized the matrix entries. The threshold method of Jankova & Van De Geer (2015) was also tested, which for a false positive rate $\alpha$, sets the individual entries in the debiased estimator $\hat{\Omega}$ to zero if
\[
\hat{\Omega}_{i,j} < \Phi^{-1}\left(1 - \frac{\alpha}{p(p - 1)}\right) \frac{\hat{v}_{i,j}}{\sqrt{n}}
\]
where $\hat{\Sigma}$ is the empirical covariance matrix and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. A figure in line with Figure 1 for this method is included in the supplementary material. In short, this approach should work asymptotically as $n,p \to \infty$. In our simulations, the empirically achieved false positive rate was not close to the target false positive rate.

### 3.2 Geonomics data

We apply our methodology to the dataset of gene expressions for a small round blue cell tumours microarray experiment from Khan et al. (2001), which is also analyzed in other works (Rothman et al., 2009; Cai & Liu, 2011). The data set consists of a training set of 64 vectors containing 2308 gene expressions. The data contains four types of tumours. Considering this $X \in \mathbb{R}^{64 \times 2308}$, we construct error controlled precision matrix estimators using the algorithm from Section 2.3 choosing the debiased estimator of Jankova & Van De Geer (2015) as the initial estimator and for graphical lasso penalization parameter $\lambda = 0.5, 1, 2$.

In Figure 3, we plot the cardinality of the support of the estimators for the three penalization parameters as the false positive rate ranges from $2^{-1}$ to $2^{-10}$. For $\lambda = 0.5$, the method plateaus quickly returning a small support of 74 nonzero off-diagonal entries. The other two lines continue to decay towards zero off-diagonal entries.

In Figure 4, we see kernel density plots of the number of non-zero entries per row aggregated over all 2308 rows. In these plots, $\lambda = 0.5, 1, 2$ and $\alpha = 0.5, 0.001$ are considered. Comparing the top and bottom row of plots, we see that there are many rows with approx 500 nonzero entries when $\alpha = 0.5$, which quickly drops to fewer than 10 nonzero entries per row once $\alpha = 0.001$. 

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Cross Validated GLASSO} & \text{Cross Validated ACLIME} \\
\text{False Pos} & \text{True Pos} & \text{False Pos} & \text{True Pos} \\
\hline
p = 496 & 1128 & 2080 & 496 & 1128 & 2080 \\
\text{TriDiag} & 9.53 & 5.27 & 3.15 & 48.2 & 36.3 & 29.9 \\
\text{BinTree} & 9.71 & 5.22 & 3.15 & 47.3 & 36.9 & 29.8 \\
\text{BlkDiag} & 10.11 & 5.37 & 3.20 & 14.1 & 8.6 & 5.5 \\
\hline
\end{array}
\]
Figure 1: The achieved true and false positives plotted against the target false positive rate displayed on the log-log scale for multivariate Gaussian data. The rows from top to bottom correspond to the tridiagonal, binary tree, and block diagonal matrices. The columns from left to right correspond to dimensions 496, 1128, and 2080.
Figure 2: The achieved true and false positives plotted against the target false positive rate displayed on the log-log scale as in Figure 1. The red lines are for the standard approach while blue lines are for subsampling by dividing the set in half.

Figure 3: The log_{10} number of non-zero entries in the 2308 × 2308 precision estimator for graphical lasso penalizations of 0.5, 1, 2 and for false positive rates from 2^{-1} to 2^{-10}. 

Figure 4: Density plots of the number of nonzero entries per row for precision estimators with, from left to right, graphical lasso penalizations of 0.5, 1, 2 and with false positive rates 0.5, top row, and 0.001, bottom row.

4 Supplementary Material

Supplementary material available at Biometrika online includes proofs for the main results and auxiliary results. Receiver operating characteristic curves for the data in Figure 1 are included, and additional simulations for exponential data and the method from Jankova & Van De Geer (2015) are also included.

References

BICKEL, P. J. & LEVINA, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 2577–2604.

BICKEL, P. J. & LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 199–227.

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* 19, 1212–1242.

CAI, T. & LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 106, 672–684.

CAI, T., LIU, W. & LUO, X. (2011). A constrained 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* 106, 594–607.

CAI, T., LIU, W. & ZHOU, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics* 44, 455–488.

FAN, J., LIAO, Y. & LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal* 19, C1–C32.
Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441.

Friedman, J., Hastie, T. & Tibshirani, R. (2018). *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. R package version 1.10.

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

Javanmard, A. & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* 15, 2869–2909.

Kashlak, A. B. & Kong, L. (2017). A concentration inequality based methodology for sparse covariance estimation. *arXiv preprint arXiv:1705.02679*.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine* 7, 673–679.

Pang, H., Qi, D., Liu, H. & Vanderbei, R. (2016). *fastclime: A Fast Solver for Parameterized LP Problems, Constrained Li Minimization Approach to Sparse Precision Matrix Estimation and Dantzig Selector*. R package version 1.4.1.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B. et al. (2011). High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics* 5, 935–980.

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

Tao, T. (2012). *Topics in random matrix theory*, vol. 132. American Mathematical Soc.

Van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42, 1166–1202.

## 5 Proofs

The theory behind the methodology in this article comes from redoing the results from random matrix theory, which generally achieves an operator norm of $O(p^{1/2})$, under the assumption that only $\alpha \in (0, 1)$ of the entries are non-zero. This gives similar results but with an operator norm of $O(\alpha^{1/2}p^{1/2})$.

We first prove a lazy version of Hoeffding’s concentration inequality.

**Lemma 5.1.** For some $M > 0$, let $Z_1, \ldots, Z_p \in [-M, M]$ be independent mean zero random variables not necessarily equally distributed, and let $b_1, \ldots, b_p \sim \text{Bernoulli}(\alpha)$ also independent of the $Z_i$. Then,

$$\Pr \left( \sum_{i=1}^{p} b_i Z_i \geq tM(\alpha p)^{1/2} \right) \leq 2e^{-t^2/4}.$$

*Proof.* The proof is more or less the same as that of the standard Hoeffding inequality. Without loss of generality, let $M = 1$. For any $i = 1, \ldots, p$, we can bound the variance of $b_i Z_i$ by

$$\text{Var}(b_i Z_i) = E(b_i^2) E(Z_i^2) \leq \alpha.$$
Thus, we have a bound on the cumulant generating function of \(b_iZ_i\) that \(\log \text{E}e^{|b_iZ_i|} \leq \alpha \lambda^2\) and \(\log \text{E}e^{\lambda \sum_{i=1}^{p} b_iZ_i} \leq \alpha \lambda^2\). Therefore, via Chernoff’s inequality, we have for any \(\lambda > 0\) that
\[
P\left( \sum_{i=1}^{p} b_iZ_i \geq t \right) \leq e^{-\lambda \sum_{i=1}^{p} b_iZ_i} \leq e^{-\lambda t} \text{E} \exp \left( \lambda \sum_{i=1}^{p} b_iZ_i \right) \leq \exp \left( -\lambda t + \alpha \lambda^2 \right)
\]
Minimizing the righthand side over \(\lambda\) gives \(\lambda = t/(2\alpha p)\) and finally that \(P(\sum_{i=1}^{p} b_iZ_i \geq t) \leq e^{-t^2/(4\alpha p)}\). Running the proof for \(\sum_{i=1}^{p} b_iZ_i \leq -t\) gives the reverse inequality. Combining those gives the desired result. Lastly, to adjust for \(M \neq 1\), replace \(Z_i\) with \(2Z_i/M\) in the above result. \(\square\)

Using the lazy version of Hoeffding’s inequality, we can get concentration of the operator norm for a sparse random matrix using a modification of the standard \(\varepsilon\)-net argument as presented in Tao (2012) Section 2.3.

**Theorem 5.1.** Let \(A \in \mathbb{R}^{p \times p}\) be a symmetric random matrix with entries \(A_{i,j} = 0\), \(|A_{i,j}| < 1\), and \(A_{i,i} = A_{j,j}\). Let \(B \in \{0,1\}^{p \times p}\) be a symmetric Bernoulli random matrix with iid entries such that \(P(B_{i,j} = 1) = \alpha \in [0,1]\). If the lower triangular entries of \(A\) are independent then,
\[
\text{E} \|A \circ B\|_\infty = O((\alpha p)^{1/2})
\]
where \(A \circ B\) denotes the entrywise or Hadamard product of \(A\) and \(B\).

**Proof.** For now, we consider \(A\) and \(B\) as iid ensembles—i.e. remove the \(A_{i,j} = A_{j,i}\) and \(B_{i,j} = B_{j,i}\) condition—and adjust for the symmetry at the end of the proof. Let \(\Pi = A \circ B\). For any vector \(v \in \mathbb{R}^p\) such that \(\|v\|_2 = 1\), we have that \(\|\Pi v\|_2 = (\sum_{i=1}^{p}(\sum_{j=1}^{p} A_{i,j}B_{i,j}v_j)^2)^{1/2}\). Therefore, from the proof of Lemma 5.1, independence of entries, and \(\|v\|_2 = 1\),
\[
\log \text{E} e^{\lambda \|\Pi v\|_2} = \sum_{i=1}^{p} \log \text{E} e^{\lambda \sum_{j=1}^{p} A_{i,j}B_{i,j}v_j} \leq \alpha \lambda^2.
\]
Thus, for any arbitrary unit vector \(v\), we can apply Lemma 5.1 to get
\[
P\left( \|\Pi v\|_2 > t(\alpha p)^{1/2} \right) \leq 2e^{-t^2/4}.
\]
To extend to \(\|\Pi\|_\infty\), we construct a maximal \((1/2)\)-net of the unit sphere \(S = \{v : \|v\|_2 = 1\}\). That is, let \(V\) be the maximal set of points in \(S\) such that for any \(u, v \in V\), \(\|u - v\|_2 \geq 0.5\). Let \(v^* \in S\) be the vector such that \(\|\Pi\|_\infty = \|\Pi v^*\|_2\), which exists via compactness. Thus, there exists a \(v_0 \in V\) such that \(\|v^* - v_0\|_2 < 0.5\) as otherwise, the set \(V\) would not be maximal. Furthermore, \(\|\Pi(v^* - v_0)\|_2 \leq 2\|\Pi\|_\infty\) and thus \(\|\Pi v_0\|_2 \geq \|\Pi\|_\infty/2\). Therefore, by the union bound,
\[
P\left( \|\Pi\|_\infty > t(\alpha p)^{1/2} \right) \leq P\left( \bigcup_{v \in V} \left\{ \|\Pi v\|_2 > \frac{t}{2}(\alpha p)^{1/2} \right\} \right) \leq 2|V|e^{-t^2/16}.
\]
From Lemma 2.3.4 of Tao (2012), \(|V| = (2C)^p\) for some absolute constant \(C > 0\). Thus, for \(t\) large enough, the righthand side becomes negligible.

Now, considering \(A\) and \(B\) symmetric with zero diagonal as in the theorem statement, we have that \(\|\Pi\|_\infty \leq \|\Pi_{\text{low}}\|_\infty + \|\Pi_{\text{up}}\|_\infty\) for \(\Pi_{\text{low}}\) and \(\Pi_{\text{up}}\) the strict lower and upper triangular portions of \(\Pi\). Hence,
\[
P\left( \|\Pi\|_\infty > t(\alpha p)^{1/2} \right) \leq P\left( \|\Pi_{\text{low}}\|_\infty > t(\alpha p)^{1/2}/2 \right) + P\left( \|\Pi_{\text{up}}\|_\infty > t(\alpha p)^{1/2}/2 \right) \leq 4|V|e^{-t^2/32}.
\]
\(\square\)

**Remark 5.2.** The constants in the above proof are not necessarily optimal, but sufficient to justify our approach to controlled precision matrix estimation.
Therefore, the same simulations as in Section 3.1 were rerun replacing the multivariate Gaussian distribution with the multivariate Laplace distribution, and are displayed in Figure 6. As expected, the true positive rate is not as large as in the Gaussian setting. When the penalization parameter for the graphical lasso is set to $\lambda = 2$, we see that our methodology does not maintain the desired false positive rate as $\alpha \to 0$. However, for $\lambda = 1$, the empirical false positive rate does approximately track with the desired false positive rate in the three simulation settings.
Figure 5: Empirical Receiver operating characteristic curves displayed on the log-log scale. The rows from top to bottom correspond to the tridiagonal, binary tree, and block diagonal matrices. The columns from left to right correspond to dimensions 496, 1128, and 2080.
Figure 6: The achieved true and false positives plotted against the target false positive rate displayed on the log-log scale for multivariate Laplace data. The rows from top to bottom correspond to the tridiagonal, binary tree, and block diagonal matrices. The columns from left to right correspond to dimensions 496, 1128, and 2080.
6.3 Asymptotic Threshold

In the work of Jankova & Van De Geer (2015), a thresholding method is proposed for the debiased graphical lasso estimator making use of the normal distribution function. Namely,

\[ \hat{\Omega}_{i,j} < \Phi^{-1} \left( 1 - \frac{\alpha}{p(p-1)} \right) \frac{\hat{\Sigma}_{i,j}}{\sqrt{n}} \]

where \( \hat{\Sigma} \) is the empirical covariance matrix and \( \Phi(\cdot) \) is the cumulative distribution function for the standard normal distribution. The same simulations as in Section 3.1 were run on this method for graphical lasso penalization parameters of \( \lambda = 2^{-2}, 2^{-3}, 2^{-2}, 2^{-1}, 1, 2 \). The larger values of \( \lambda \) resulted in all of the off-diagonal entries being set to zero. Hence, Figure 7 displays the false and true positive rates for only \( \lambda = 1/16, 1/8 \). This method should work asymptotically as \( n, p \to \infty \). However, for our specific choices of \( n \) and \( p \), this method failed to achieve anything close to the target false positive rate.
Figure 7: Achieved true and false positive rates plotted against the target rate and displayed on the log-log scale. The rows from top to bottom correspond to the tridiagonal, binary tree, and block diagonal matrices. The columns from left to right correspond to dimensions 496, 1128, and 2080.