Minimax Estimation of Bandable Precision Matrices

Authored by:

Sahand Negahban
Addison Hu

Abstract

The inverse covariance matrix provides considerable insight for understanding statistical models in the multivariate setting. In particular, when the distribution over variables is assumed to be multivariate normal, the sparsity pattern in the inverse covariance matrix, commonly referred to as the precision matrix, corresponds to the adjacency matrix representation of the Gauss-Markov graph, which encodes conditional independence statements between variables. Minimax results under the spectral norm have previously been established for covariance matrices, both sparse and banded, and for sparse precision matrices. We establish minimax estimation bounds for estimating banded precision matrices under the spectral norm. Our results greatly improve upon the existing bounds; in particular, we find that the minimax rate for estimating banded precision matrices matches that of estimating banded covariance matrices. The key insight in our analysis is that we are able to obtain barely-noisy estimates of $k times k$ subblocks of the precision matrix by inverting slightly wider blocks of the empirical covariance matrix along the diagonal. Our theoretical results are complemented by experiments demonstrating the sharpness of our bounds.

1 Paper Body

Imposing structure is crucial to performing statistical estimation in the high-dimensional regime where the number of observations can be much smaller than the number of parameters. In estimating graphical models, a long line of work has focused on understanding how to impose sparsity on the underlying graph structure. Sparse edge recovery is generally not easy for an arbitrary distribution. However, for Gaussian graphical models, it is well-known that the graphical structure is encoded in the inverse of the covariance matrix ??1 = ?, commonly referred to as the precision matrix [12, 14, 3]. Therefore, accurate recovery of the precision matrix is paramount to understanding the structure of the graphical model. As a consequence, a great deal of work has focused on

sparse recovery of precision matrices under the multivariate normal assumption [8, 4, 5, 17, 16]. Beyond revealing the graph structure, the precision matrix also turns out to be highly useful in a variety of applications, including portfolio optimization, speech recognition, and genomics [12, 23, 18]. Although there has been a rich literature exploring the sparse precision matrix setting for Gaussian graphical models, less work has emphasized understanding the estimation of precision matrices under additional structural assumptions, with some exceptions for block structured sparsity [10] or bandability [1]. One would hope that extra structure should allow us to obtain more statistically efficient solutions. In this work, we focus on the case of bandable precision matrices, which capture ? Addison graduated from Yale in May 2017. Up-to-date contact information may be found at http: //huisaddison.com/.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

a sense of locality between variables. Bandable matrices arise in a number of time-series contexts and have applications in climatology, spectroscopy, fMRI analysis, and astronomy [9, 20, 15]. For example, in the time-series setting, we may assume that edges between variables Xi , Xj are more likely when i is temporally close to j, as is the case in an auto-regressive process. The precision and covariance matrices corresponding to distributions with this property are referred to as bandable, or tapering. We will discuss the details of this model in the sequel. Past work: Previous work has explored the estimation of both bandable covariance and precision matrices [6, 15]. Closely related work includes the estimation of sparse precision and covariance matrices [3, 17, 4]. Asymptotically-normal entrywise precision estimates as well as minimax rates for operator norm recovery of sparse precision matrices have also been established [16]. A line of work developed concurrently to our own establishes a matching minimax lower bound [13]. When considering an estimation technique, a powerful criterion for evaluating whether the technique performs optimally in terms of convergence rate is minimaxity. Past work has established minimax rates of convergence for sparse covariance matrices, bandable covariance matrices, and sparse precision matrices [7, 6, 4, 17]. The technique for estimating bandable covariance matrices proposed in [6] is shown to achieve the optimal rate of convergence. However, no such theoretical guarantees have been shown for the bandable precision estimator proposed in recent work for estimating sparse and smooth precision matrices that arise from cosmological data [15]. Of note is the fact that the minimax rate of convergence for estimating sparse covariance matrices matches the minimax rate of convergence of estimating sparse precision matrices. In this paper, we introduce an adaptive estimator and show that it achieves the optimal rate of convergence when estimating bandable precision matrices from the banded parameter space (3). We find, satisfyingly, that analogous to the sparse case, in which the minimax rate of convergence enjoys the same rate for both precision and covariance matrices, the minimax rate of convergence for estimating bandable precision matrices matches the minimax rate of convergence for estimating bandable covariance matrices that has been established in the literature [6]. Our contributions: Our goal is to estimate a banded precision matrix based on

n i.i.d. observations. We consider a parameter space of precision matrices ? with a power law decay structure nearly identical to the bandable covariance matrices considered for covariance matrix estimation [6]. We present a simple-to-implement algorithm for estimating the precision matrix. Furthermore, we show that the algorithm is minimax optimal with respect to the spectral norm. The upper and lower bounds given in Section 3 together imply the following optimal rate of convergence for estimating bandable precision matrices under the spectral norm. Informally, our results show the following bound for recovering a banded precision matrix with bandwidth k. Theorem 1.1 (Informal). The minimax risk for estimating the precision matrix ? over the class P? given in (3) satisfies:

2 k + log p

?

inf sup E ? ? ? ? (1) ? n ? P? ? k as defined in Equation (7). where this bound is achieved by the tapering estimator ? An important point to note, which is shown more precisely in the sequel, is that the rate of convergence as compared to sparse precision matrix recovery is improved by a factor of min(k log(p), k 2). We establish a minimax upper bound by detailing an algorithm for obtaining an estimator given observations x1 , . . . , xn and a pre-specified bandwidth k, and studying the resultant estimator?s risk properties under the spectral norm. We show that an estimator using our algorithm with the optimal choice of bandwidth attains the minimax rate of convergence with high probability. To establish the optimality of our estimation routine, we derive a minimax lower bound to show that the rate of convergence cannot be improved beyond that of our estimator. The lower bound is established by constructing subparameter spaces of (3) and applying testing arguments through Le Cam?s method and Assouad?s lemma [22, 6]. To supplement our analysis, we conduct numerical experiments to explore the performance of our estimator in the finite sample setting. The numerical experiments confirm that even in the finite sample case, our proposed estimator exhibits the minimax rate of convergence. 2

The remainder of the paper is organized as follows. In Section 2, we detail the exact model setting and introduce a blockwise inversion technique for precision matrix estimation. In Section 3, theorems establishing the minimaxity of our estimator under the spectral norm are presented. An upper bound on the estimator?s risk is given in high probability with the help of a result from set packing. The minimax lower bound is derived by way of a testing argument. Both bounds are accompanied by their proofs. Finally, in Section 4, our estimator is subjected to numerical experiments. Formal proofs of the theorems may be found in the longer version of the paper [11]. Notation: We will now collect notation that will be used throughout the remaining sections. Vectors will be denoted as lower-case x while matrices are upper-case A. The spectral or operator norm of a matrix is defined to be kAk = supx6=0,y6=0 hAx, yi while the matrix ‘1 norm of a symmetric matrix Pm A ? Rm?m is defined to be kAk1 = maxj i=1 —Aij —.

2

Background and problem set-up

In this section we present details of our model and the estimation procedure. If one considers observations of the form x1 , . . . , xn ? Rp drawn from a distribution with precision matrix ?p?p and zero mean, the goal then is to estimate the unknown matrix ?p?p based on the observations {xi }ni=1 . Given a random sample of p-variate observations x1 , . . . , xn drawn from a multivariate distribution with population covariance ? = ?p?p , our procedure is based on a tapering estimator derived from blockwise estimates for estimating the precision matrix ?p?p = ??1 . The maximum likelihood estimator of ? is n

1X ? = (? ?)(xl ? x ?)¿ ? ?ij)1?i,j?p = (xl ? x n

(2)

l=1

? is the empirical mean of the vectors xi . We will construct estimators of the precision matrix where x ? along the diagonal, and averaging over the resultant subblocks. ? = ??1 by inverting blocks of ? Throughout this paper we adhere to the convention that ?ij refers to the ij th element in a matrix ?. Consider the parameter space F? , with associated probability measure P? , given by: ( ) X ?1 ?? F? = F? (M0 , M) = ? : max {—?ij — : —i ? j— ? k} ? M k for all k, ?i (?) ? [M0 , M0] j

i

(3) where ?i (?) denotes the ith eigenvalue of ?, with ?i ? ?j for all i ? j. We also constrain ? ¿ 0, M ¿ 0, M0 ¿ 0. Observe that this parameter space is nearly identical to that given in Equation (3) of [6]. We take on an additional assumption on the minimum eigenvalue of ? ? F? , which is used in the technical arguments where the risk of estimating ? under the spectral norm is bounded in terms of the error of estimating ? = ??1 . Observe that the parameter space intuitively dictates that the magnitude of the entries of ? decays in power law as we move away from the diagonal. As with the parameter space for bandable covariance matrices given in [6], we may understand ? in (3) as a rate of decay for the precision entries ?ij as they move away from the diagonal; it can also be understood in terms of the smoothness parameter in nonparametric estimation [19]. As will be discussed in Section 3, the optimal choice of k depends on both n and the decay rate ?. 2.1

Estimation procedure

We now detail the algorithm for obtaining minimax estimates for bandable ?, which is also given as pseudo-code2 in Algorithm 1. The algorithm is inspired by the tapering procedure introduced by Cai, Zhang, and Zhou [6] in the case of covariance matrices, with modifications in order to estimate the precision matrix. Estimating 2

In the pseudo-code, we adhere to the NumPy convention (1) that arrays are zero-indexed, (2) that slicing an array arr with the operation arr[a:b] includes the element indexed at a and excludes the element indexed at b, and (3) that if b is greater than the length of the array, only elements up to the terminal element are included, with no errors.

3

the precision matrix introduces new difficulties as we do not have direct access to the estimates of elements of the precision matrix. For a given integer

k, 1 ? k ? p, we construct a tapering estimator as follows. First, we calculate the maximum likelihood estimator for the covariance, as given in Equation (2). Then, for all integers 1 ? m ? l ? p and m ? 1, we define the matrices with square blocks of size at most 3m along the diagonal: ? (3m) = (? ? ?ij 1{l ? m ? i ¡ l + 2m, l ? m ? j ¡ l + 2m})p?p l?m

(4)

? (3m) , we replace the nonzero block with its inverse to obtain ? ? (3m) . For a given l, we For each ? l?m l?m refer to the individual entries of this intermediate matrix as follows: l ? (3m) = (? ? ?ij 1{l ? m ? i ¡ l + 2m, l ? m ? j ¡ l + 2m})p?p (5) l?m

(3m)

? For each l, we then keep only the central m ? m subblock of ? l?m to obtain the blockwise estimate (m) ? ?l : ? (m) = (? ? ? l 1{l ? i ¡ l + m, l ? j ¡ l + m})p?p (6) l

ij

Note that this notation allows for l ¡ 0 and l + m ¿ p; in each case, this out-of-bounds indexing allows us to cleanly handle corner cases where the subblocks are smaller than m ? m. For a given bandwidth k (assume k is divisible by 2), we calculate these blockwise estimates for both m = k and m = k2 . Finally, we construct our estimator by averaging over the block matrices: ? ? p p X X 2 k/2) (k) (?k = ? ? ? ? ? ? (7) ? ? ? l l k k l=1?k

l=1? /2

k 2

We note that within entries of the diagonal, each entry is effectively the sum of k2 estimates, and as we move from k2 to k from the diagonal, each entry is progressively the sum of one fewer entry. Therefore, within k2 of the diagonal, the entries are not tapered; and from k2 to k of the diagonal, the entries are linearly tapered to zero. The analysis of this estimator makes careful use of this tapering schedule and the fact that our estimator is constructed through the average of block matrices of size at most k ? k. 2.2

Implementation details

The naive algorithm performs O(p + k) inversions of square matrices with size at most 3k. This method can be sped up considerably through an application of the Woodbury matrix identity and the Schur complement relation [21, 2]. Doing so reduces the computational complexity of the algorithm from O(pk 3) to O(pk 2). We discuss the details of modified algorithm and its computational complexity below. ? (3m) and are interested in obtaining ? ? (3m) . We observe that the nonzero block Suppose we have ? l?m l?m+1 ? (3m) corresponds to the inverse of the nonzero block of ? ? (3m) , which only differs by one of ? l?m+1

l?m+1

? (3m) , the matrix for which the inverse of the nonzero block corresponds row and one column from ? l?m (3m) ? ? (3m) , ? ? (3m) to ? , which we have already computed. We may understand the movement from ? l?m (3m)

l?m

l?m

(3m)

? ? to ? l?m+1 (to which we already have direct access) and ?l?m+1 as two rank-1 updates. Let us view (3m) (3m) ? ? the nonzero blocks of ? l?m , ?l?m as the block matrices:

A ? R1?1 B ? R1?(3m?1) 3m ? NonZero(?l?m) = B ¿ ? R(3m?1)?1 C ? R(3m?1)?(3m?1)

? ? R1?(3m?1) A? ? R1?1 B (3m) ? NonZero(?l?m) = ? ¿ B ? R(3m?1)?1 C? ? R(3m?1)?(3m?1) ? 3m , ? ? (3m) , we may trivially compute C ?1 as The Schur complement relation tells us that given ? l?m l?m follows:

?1 ? ¿ B C? CB C ?1 = C? ?1 + B ¿ A?1 B = C? ? (8) ? ¿ A + B CB 4

Algorithm 1 Blockwise Inversion Technique ? k) function F IT B LOCKWISE(?, ? ? ? 0p?p for l ? [1 ? k, p) do ? ?? ? + B LOCK I NVERSE(?, ? k, l) ? end for for l ? [1 ? bk/2c, p) do ? ?? ? ? B LOCK I NVERSE(?, ? bk/2c, l) ? end for ? return ? end function ? m, l) function B LOCK I NVERSE(?, . Obtain 3m ? 3m block inverse. s ? max{l ? m, 0} f ? min{p, l + 2m}

?1 ? M ? ?[s:f, s:f] . Preserve central m ? m block of inverse. s ? m + min{l ? m, 0} N ? M [s:s+m, s:s+m] . Restore block inverse to appropriate indices. s ? max{l, 0} f ? min{l + m, p} P [s:f, s:f] = N return P end function

by the Woodbury matrix identity, which gives an efficient algorithm for computing the inverse of a matrix subject to a low-rank (in this case, rank-1) perturbation. This allows us to move from the inverse of a matrix in R3m?3m to the inverse of a matrix in R(3m?1)?(3m?1) where a row and column have been removed. A nearly identical argument allows us to move from the R(3m?1)?(3m?1) matrix to an R3m?3m matrix where a row and column have been appended, which gives us the desired block ? (3m) . of ? l?m+1 With this modification to the algorithm, we need only compute the inverse of a square matrix of width 2m at the beginning of the routine; thereafter, every subsequent block inverse may be computed through simple rank one matrix updates. 2.3

Complexity details

We now detail the factor of k improvement in computational complexity provided through the application of the Woodbury matrix identity and the Schur complement relation introduced in Section 2.2. Recall that the naive implementation of Algorithm 1 involves O(p + k) inversions of square matrices of size at most 3k, each of which cost O(k 3). Therefore, the overall complexity of the naive algorithm is O(pk 3), as k ¡ p. Now, consider the Woodbury-Schur-improved algorithm. The initial single inversion of a 2k ? 2k matrix costs O(k 3). Thereafter, we perform O(p + k) updates of the form given in Equation (8). These updates simply require vector matrix operations. Therefore, the update complexity on each iteration is O(k 2). It follows that the overall complexity of the amended algorithm is O(pk 2).

3

Rate optimality under the spectral norm

Here we present the results that establish the rate optimality of the above estimator under the spectral norm. For symmetric matrices A, the spectral norm, which corresponds to the largest singular value of A, coincides with the ‘2 -operator norm. We establish optimality by first deriving an upper bound 5

in high probability using the blockwise inversion estimator defined in Section 2.1. We then give a matching lower bound in expectation by carefully constructing two sets of multivariate normal distributions and then applying Assouad?s lemma and Le Cam?s method. 3.1

Upper bound under the spectral norm

In this section we derive a risk upper bound for the tapering estimator defined in (7) under the operator norm. We assume the distribution of the xi ?s is subgaussian; that is, there exists ? ¿ 0 such that:

t2 ? P —v¿ (xi ? E xi)— ¿ t ? e? 2 (9) for all t ¿ 0 and kvk2 = 1. Let P? = P? (M0 , M, ?) denote the set of distributions of xi that satisfy (3) and (9). ? k , defined in (7), of the precision matrix ?p?p with p ¿ Theorem 3.1. The tapering estimator ? 1

n 2?+1 satisfies:

2

k + log p

?

?2? sup P ? ? ? ? C + Ck = O p?15

k n P?

(10)

with k = o(n), log p = o(n), and a universal constant C ¿ 0. 1 ? =? ? k with k = n 2?+1 In particular, the estimator ? satisfies:

2

2? log p

?

? 2?+1 sup P ? ? ? ? Cn + C = O p?15

k n P?

(11)

1

Given the result in Equation (10), it is easy to show that setting k = n 2?+1 yields the optimal rate by balancing the size of the inside-taper and outside-taper terms, which gives Equation (11). The proof of this theorem, which is given in the supplementary material, relies on the fact that when we invert a 3k ? 3k block, the difference between the central k ? k block and the corresponding k ? k block which would have been obtained by inverting the full matrix has a negligible contribution to the risk. As a result, we are able to take concentration bounds on the operator norm of subgaussian matrices, customarily used for bounding the norm of the difference of covariance matrices, and apply them instead to differences of precision matrices to obtain our result. The key insight is that we can relate the spectral norm of a k ? k subblock produced by our estimator to the spectral norm of the corresponding k ? k subblock of the covariance matrix, which allows us to apply concentration bounds from classical random matrix theory. Moreover, it turns out that if we apply the tapering schedule induced by the construction of our estimator to the population parameter ? ? F? , we may express the tapered population ? as a sum of block matrices in exactly the same way that our estimator is expressed as a sum of block matrices. In particular, the tapering schedule is presented next. Suppose

a population precision matrix ? ? F? . Then, we denote the tapered version of ? by ?A , and construct: ?A = (?ij ? vij)p?p ?B = (?ij ? (1 ? vij))p?p where the tapering coefficients are given by: ? ? ?1 vij = —i?j— k/2 ? ?0

for —i ? j— ¡ k2 for k2 ? —i ? j— ¡ k for —i ? j— ? k

We then handle the risk of estimating the inside-taper ?A and the risk of estimating the outside-taper ?B separately. Because our estimator and the population parameter are both averages over k ? k block matrices along the diagonal, we may then take a union bound over the high probability bounds on the spectral norm deviation for the k ? k subblocks to obtain a high probability bound on the risk of our estimator. We refer the reader to the longer version of the paper for further details [11]. 6

3.2

Lower bound under the spectral norm

In Section 3.1, we established Theorem 3.1, which states that our estimator achieves the rate of 2? 1 convergence n? 2?+1 under the spectral norm by using the optimal choice of k = n 2?+1 . Next we demonstrate a matching lower bound, which implies that the upper bound established in Equation (11) is tight up to constant factors. Specifically, for the estimation of precision matrices in the parameter space given by Equation (3), the following minimax lower bound holds. Theorem 3.2. The minimax risk for estimating the precision matrix ? over P? under the operator norm satisfies:

2 2? log p

?

inf sup E ? ? ? ? cn? 2?+1 + c (12) ? P? n ? As in many information theoretic lower bounds, we first identify a subset of our parameter space that captures most of the complexity of the full space. We then establish an information theoretic limit on estimating parameters from this subspace, which yields a valid minimax lower bound over the original set. Specifically, for our particular parameter space F? , we identify two subparameter spaces, F11 , F12 . The first, F11 , is a collection of 2k matrices with varying levels of density. To this collection, we 2? apply Assouad?s lemma obtain a lower bound with rate n? 2?+1 . The second, F12 , is a collection of diagonal matrices, to which we apply Le Cam?s method to derive a lower bound with rate logn p . The rate given in Theorem 3.2 is therefore a lower bound on minimax rate for estimating the union (F11 ? F12) = F1 ? F? . The full details of the subparameter space construction and derivation of lower bounds may be found in the full-length version of the paper [11].

4

Experimental results

We implemented the blockwise inversion technique in NumPy and ran simulations on synthetic datasets. Our experiments confirm that even in the finite sample case, the blockwise inversion technique achieves the theoretical rates. In the experiments, we draw observations from a multivariate normal distribution with precision parameter ? ? F? , as defined in (3). Following [6], for given constants ?, ?, p, we consider precision matrices ? = (?ij)1?i,j?p of the form:

1 for 1 ? i = j ? p ?ij = (13) ?—i ? j—???1 for 1 ? i 6= j ? p Though the precision matrices considered in our experiments are Toeplitz, our estimator does not take advantage of this knowledge. We choose ? = 0.6 to ensure that the matrices generated are non-negative definite. 1

In applying the tapering estimator as defined in (7), we choose the bandwidth to be k = bn 2?+1 c, which gives the optimal rate of convergence, as established in Theorem 3.1. In our experiments, we varied ?, n, and p. For our first set of experiments, we allowed ? to take on values in {0.2, 0.3, 0.4, 0.5}, n to take values in {250, 500, 750, 1000}, and p to take values in {100, 200, 300, 400}. Each setting was run for five trials, and the averages are plotted with error bars to show variability between experiments. We observe in Figure 1a that the spectral norm error increases linearly as log p increases, confirming the logn p term in the rate of convergence. Building upon the experimental results from the first set of simulations, we provide an additional sets of trials for the ? = 0.2, p = 400 case, with n ? {11000, 3162, 1670}. These sample sizes were chosen so that in Figure 1b, there is overlap between the error plots for ? = 0.2 and the other ? regimes3 . As with Figure 1a, Figure 1b confirms the minimax rate of convergence given in Theorem 2? 3.1. Namely, we see that plotting the error with respect to n? 2?+1 results in linear plots with almost 3

For the ? = 0.2, p = 400 case, we omit the settings where n ? {250, 500, 750} from Figure 1b to improve the clarity of the plot.

7
Setting: n = 1000
8
0.025
? = 0. 2 ? = 0. 3 ? = 0. 4 ? = 0. 5
Spectral Norm Error
6
0.020
Mean Spectral Norm
7
? = 0. 2 ? = 0. 3 ? = 0. 4 ? = 0. 5
Setting: p = 400
0.015
5 4
0.010
3 2
0.005
1 0 4.6
4.8
5.0
5.2 5.4 log(p)
5.6
5.8
0.0000.02
6.0

0.04
0.06
0.08?
n ?2? + 1 2
0.10
0.12
0.14
2?
(b) Mean spectral norm error as n? 2?+1 changes.
(a) Spectral norm error as log p changes.

Figure 1: Experimental results. Note that the plotted error grows linearly as a function of log p and 2? n? 2?+1 , respectively, matching the theoretical results; however, the linear relationship is less clear in the ? = 0.2 case, due to the subtle interplay of the error terms. identical slopes. We note that in both plots, there is a small difference in the behavior for the case ? = 0.2. This observation can be attributed to the fact that for such a slow decay of the precision matrix bandwidth, we have a more subtle interplay between the bias and variance terms presented in the theorems above.

5

Discussion

In this paper we have presented minimax upper and lower bounds for estimating banded precision matrices after observing n samples drawn from a p-dimensional subgaussian distribution. Furthermore, we have provided a computationally efficient algorithm that achieves the optimal rate of convergence for estimating a banded precision matrix under the operator norm. Theorems 3.1 and 3.2 together establish that the minimax rate of convergence for estimating precision matrices over the parameter 2? space F? given in Equation (3) is n? 2?+1 + logn p , where ? dictates the bandwidth of the precision matrix. The rate achieved in this setting parallels the results established for estimating a bandable covariance matrix [6]. As in that result, we observe that different regimes dictate which term dominates in the 1 2? rate of convergence. In the setting where log p is of a lower order than n 2?+1 , the n? 2?+1 term dominates, and the rate of convergence is determined by the smoothness parameter ?. However, when 1 log p is much larger than n 2?+1 , p has a much greater influence on the minimax rate of convergence. Overall, we have shown the performance gains that may be obtained through added structural constraints. An interesting line of future work will be to explore algorithms that uniformly exhibit a smooth transition between fully banded models and sparse models on the precision matrix. Such methods could adapt to the structure and allow for mixtures between banded and sparse precision matrices. Another interesting direction would be in understanding how dependencies between the n observations will influence the error rate of the estimator. Finally, the results presented here apply to the case of subgaussian random variables. Unfortunately, moving away from the Gaussian setting in general breaks the connection between precision matrices and graph structure. Hence, a fruitful line of work will be to also develop methods that can be applied to estimating the banded graphical model

structure with general exponential family observations. Acknowledgements We would like to thank Harry Zhou for stimulating discussions regarding matrix estimation problems. SN acknowledges funding from NSF Grant DMS 1723128.

8

2 References

[1] P. J. Bickel and Y. R. Gel. Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(5):711?728, 2011. [2] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge University Press, Cambridge, UK, 2004. [3] T. T. Cai, W. Liu, and X. Luo. A Constrained L1 Minimization Approach to Sparse Precision Matrix Estimation. arXiv:1102.2233 [stat], February 2011. arXiv: 1102.2233. [4] T. T. Cai, W. Liu, and H. H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. Ann. Statist., 44(2):455?488, 04 2016. [5] T. T. Cai, Z. Ren, H. H. Zhou, et al. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Electronic Journal of Statistics, 10(1):1?59, 2016. [6] T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics, 38(4):2118?2144, August 2010. [7] T. T. Cai and H. H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. Ann. Statist., 40(5):2389?2420, 10 2012. [8] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. Biostatistics, 2007. [9] K. J. Friston, P. Jezzard, and R. Turner. Analysis of functional mri time-series. Human brain mapping, 1(2):153?171, 1994. [10] M. J. Hosseini and S.-I. Lee. Learning sparse gaussian graphical models with overlapping blocks. In Advances in Neural Information Processing Systems, pages 3808?3816, 2016. [11] A. J. Hu and S. N. Negahban. Minimax Estimation of Bandable Precision Matrices. arXiv, 2017. arXiv: 1710.07006v1. [12] S. L. Lauritzen. Graphical Models. Oxford Statistical Science Series. Clarendon Press, Oxford, 1996. [13] K. Lee and J. Lee. Estimating Large Precision Matrices via Modified Cholesky Decomposition. arXiv:1707.01143 [stat], July 2017. arXiv: 1707.01143. [14] N. Meinshausen and P. B?hlmann. High-dimensional graphs and variable selection with the Lasso. Annals of Statistics, 34:1436?1462, 2006. [15] N. Padmanabhan, M. White, H. H. Zhou, and R. O?Connell. Estimating sparse precision matrices. Monthly Notices of the Royal Astronomical Society, 460(2):1567?1576, 2016. [16] Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. The Annals of Statistics, 43(3):991?1026, June 2015. [17] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics, 2:494?515, 2008. [18] G. Saon and J. T. Chien. Bayesian sensing hidden markov models for speech recognition. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5056?5059, May 2011. [19] A. B. Tsybakov.

Introduction to Nonparametric Estimation. Springer Publishing Company, Incorporated, 1st edition, 2008. [20] H. Visser and J. Molenaar. Trend estimation and regression analysis in climatological time series: an application of structural time series models and the kalman filter. Journal of Climate, 8(5):969?979, 1995. [21] M. A. Woodbury. Inverting modified matrices. Statistical Research Group, Memo. Rep. no. 42. Princeton University, Princeton, N. J., 1950. [22] B. Yu. Assouad, Fano and Le Cam. In Festschrift for Lucien Le Cam, pages 423?435. Springer-Verlag, Berlin, 1997. [23] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. Biometrika, 94(1):19?35, 2007.

9