Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(514): 834–845. doi:10.1080/01621459.2015.1058265.

Convex Banding of the Covariance Matrix

Jacob Bien,

Cornell University, Department of Biological Statistics and Computational Biology, 1178 Comstock Hall, Cornell University, Ithaca, 14853 United States

Florentina Bunea, and

Cornell University, Statistical Science, Ithaca, 14850 United States

Luo Xiao

Johns Hopkins University, Department of Biostatistics, Baltimore, United States

Abstract

We introduce a new sparse estimator of the covariance matrix for high-dimensional models in which the variables have a known ordering. Our estimator, which is the solution to a convex optimization problem, is equivalently expressed as an estimator which tapers the sample covariance matrix by a Toeplitz, sparsely-banded, data-adaptive matrix. As a result of this adaptivity, the convex banding estimator enjoys theoretical optimality properties not attained by previous banding or tapered estimators. In particular, our convex banding estimator is minimax rate adaptive in Frobenius and operator norms, up to log factors, over commonly-studied classes of covariance matrices, and over more general classes. Furthermore, it correctly recovers the bandwidth when the true covariance is exactly banded. Our convex formulation admits a simple and efficient algorithm. Empirical studies demonstrate its practical effectiveness and illustrate that our exactly-banded estimator works well even when the true covariance matrix is only close to a banded matrix, confirming our theoretical results. Our method compares favorably with all existing methods, in terms of accuracy and speed. We illustrate the practical merits of the convex banding estimator by showing that it can be used to improve the performance of discriminant analysis for classifying sound recordings.

Keywords

High-dimensional; structured sparsity; adaptive; hierarchical group lasso

1 Introduction

The covariance matrix is one of the most fundamental objects in statistics, and yet its reliable estimation is greatly challenging in high dimensions. The sample covariance matrix is known to degrade rapidly as an estimator as the number of variables increases, as noted more than four decades ago (see, e.g., Dempster 1972), unless additional assumptions are placed on the underlying covariance structure (see Bunea & Xiao 2014 for an overview).

Indeed, the key to tractable estimation of a high-dimensional covariance matrix lies in exploiting any knowledge of the covariance structure. In this vein, this paper develops an estimator geared toward a covariance structure that naturally arises in many applications in which the variables are ordered. Variables collected over time or representing spectra constitute two subclasses of examples.

We observe an independent sample, $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^p$ of mean zero random vectors with true population covariance matrix $[\mathbf{X}_i \mathbf{X}_i^T] = \Sigma^*$. When the variables have a known ordering, it is often natural to assume that the dependence between variables is a function of the distance between variables' indices. Common examples are in stationary time series modeling, where $\Sigma_{j, (j+h)}^*$ depends only on h , the lag of the series, with $-K \leq h \leq K$, for some $K > 0$. One can weaken this assumption, allowing $\{\Sigma_{j, (j+h)}^* : -K \leq h \leq K\}$ to depend on j , while retaining the assumption that the bandwidth K does not depend on j . Another typical assumption is that $\Sigma_{jk}^* = \text{Cov}(X_j, X_k)$ decreases with $|j-k|$. A simple example is the moving-average process, where it is assumed that $\Sigma_{jk}^* = 0$ if $|j-k| > K$ for some bandwidth parameter K , and a covariance matrix Σ^* with this structure is called banded. Likewise, in a first-order autoregressive model, the elements decay exponentially with distance from the main diagonal, $\Sigma_{jk}^* \propto \beta^{|j-k|}$, where $|\beta| < 1$, justifying the term “approximately banded” used to describe such a matrix.

More generally, banded or approximately banded covariance matrices can be used to model any generic vector $(X_1, \dots, X_p)^T$ whose entries are ordered such that any entries that are more than K apart are uncorrelated (or at most very weakly correlated). This situation does not specify any particular decay or ordering of correlation strengths. It is noteworthy that K itself is unknown, and may depend on n or p or both.

A number of estimators have been proposed for this setting that outperform the sample covariance matrix, $\mathbf{S} = n^{-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$, where $\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^n \mathbf{X}_i$. In this paper we will focus on the squared Frobenius $\|\cdot\|_F^2$ and operator $\|\cdot\|_{op}$ norms of the difference between estimators and the population matrix, as measures of performance. Assuming that the random vectors are sampled from a multivariate normal distribution and that both the largest and smallest eigenvalues of Σ^* are bounded from above and below by some

constants, it has been shown that $\frac{1}{p} \|\mathbf{S} - \Sigma^*\|_F^2 = O_p\left(\frac{p}{n}\right)$ and that

$\|\mathbf{S} - \Sigma^*\|_{op} = O_p\left(p \sqrt{\frac{\log p}{n}}\right)$, neither of which can be close to zero for $p > n$ because the bounds are tight; see, for example, Bunea & Xiao (2014). This can be rectified when one uses, instead, estimators that take into account the structure of Σ^* . For instance, Bickel & Levina (2008) introduced a class of banding and tapering estimators of the form

$$\hat{\Sigma}_T = \mathbf{T} * \mathbf{S}, \quad (1.1)$$

where $\mathbf{T}$ is a Toeplitz matrix and $*$ denotes Schur multiplication. When the matrix $\mathbf{T}$ is of the form $\mathbf{T}_{jk} = 1\{|j - k| > K\}$, for a pre-specified K , one obtains what is referred to as the banded estimator. More general forms of $\mathbf{T}$ are allowed (Bickel & Levina 2008). The Frobenius and operator norm optimality of such estimators has been studied relative to classes of approximately banded population matrices discussed in detail in Section 4.2 below. Members of these classes are matrices with entries decaying with distance from the main diagonal at rate depending on the sample size n and a parameter $\alpha > 0$. The minimax lower bounds for estimation over these classes have been established, for both norms, in Cai et al. (2010). The banding estimator achieves them in both Frobenius norm (Bickel & Levina 2008) and operator norm (Xiao & Bunea 2014), and the same is true for a more general tapering estimator proposed in Cai et al. (2010). The common element of these estimators is that, while being minimax rate optimal, they are not minimax adaptive, in that their construction utilizes knowledge of α , which is typically not known in practice. Moreover, there is no guarantee that the banded estimators are positive definite, while this can be guaranteed via appropriate tapering.

Motivated by the desire to propose a rate-optimal estimator that does not depend on α , Cai & Yuan (2012) propose an adaptive estimator that partitions $\mathbf{S}$ into blocks of varying sizes, and then zeros out some of these blocks. They show that this estimator is minimax adaptive in operator norm, over a certain class of population matrices. The block-structure form of their estimator is an artifact of an interesting proof technique, which relates the operator norm of a matrix to the operator norm of a smaller matrix whose elements are the operator norms of blocks of the original matrix. The construction is tailored to obtaining optimality in operator norm, and the estimator leans heavily on the assumption of decaying covariance away from the diagonal. In particular, it automatically zeros out all elements far from the diagonal without regard to the data (so a covariance far from the diagonal could be arbitrarily large and still be set to zero). In this sense, the method is less data-adaptive than may be desirable and may suffer in situations in which the long-range dependence does not decay fast enough. In addition, as in the case of the banded estimators, this block-thresholded estimator cannot be guaranteed to be positive definite. If positive definiteness is desired, the authors note that one may project the estimator onto the positive semidefinite cone (however, this projection step would lose the sparsity of the original solution).

Other estimators with good practical performance, in both operator and Frobenius norm, have been proposed, notably the *banded Cholesky* and *nested lasso* methods of Rothman et al. (2010). These approaches involve regularizing the Cholesky factor via solving a series of least squares or weighted lasso problems. The resulting estimators are sparse and positive definite; however, to our knowledge a theoretical study of these estimators has not been conducted.

1.1 The convex banding estimator

In this work we introduce a new covariance estimator with the goal of bridging several gaps in the existing literature. Our contributions are as follows:

1. We construct a new estimator that is sparsely-banded and positive definite with high probability. Our estimator is the solution to a convex optimization problem, and, by construction, has a data-dependent bandwidth. We call our estimation procedure *convex banding*.
2. We propose an efficient algorithm for constructing this estimator and show that it amounts to the tapering of the sample covariance matrix by a data-dependent matrix. In contrast to previous tapering estimators, which require a fixed tapering matrix, this data-dependent tapering allows our estimator to adapt to the unknown bandwidth of the true covariance matrix.
3. We show that our estimator is minimax rate adaptive (up to logarithmic factors) with respect to the Frobenius norm over a new class of population matrices that we term *semi-banded*. This class generalizes those of banded or approximately banded matrices. This establishes our estimator as the first with proved minimax rate adaptivity in Frobenius norm (up to logarithmic factors) over the previously studied covariance matrix classes. Moreover, members of the newly introduced class do not require entries to decay with the distance between their indices. This extends the scope of banded covariance estimators. We also show that our estimator is minimax rate optimal (again, up to logarithmic factors) and adaptive with respect to the operator norm over a class of matrices with elements close to banded matrices, with bandwidth that can grow with n , p or both, at an appropriate rate. Moreover, we show that our estimator recovers the underlying sparsity pattern with high probability.

An unusual (and favorable) aspect of our estimator is that it is simultaneously sparse and positive definite with high probability—a property not shared by any other method with comparable theoretical guarantees.

The precise definition of our convex banding estimator and a discussion of the algorithm are given in Sections 2 and 3 below. Our target Σ^* either has all elements beyond a certain bandwidth K being zero or it is close (in a sense defined precisely in Section 4) to a K -banded matrix. We therefore aim at constructing an estimator that will zero out all covariances that are beyond a certain data-dependent bandwidth. If we regard the elements we would like to set to zero as a group, it is natural to consider a penalized estimator, with a penalty that zeros out groups. The most basic penalty that sets groups of parameters to zero simultaneously, without any other restrictions on these parameters, is known as the Group Lasso (Yuan & Lin 2006), a generalization of the Lasso (Tibshirani 1996). Zhao et al. (2009) show that by taking a hierarchically-nested group structure, one can develop penalties that set one group of parameters to zero whenever another group is zero. Penalties that render various hierarchical sparsity patterns have been proposed and studied in, for instance,

Jenatton et al. (2010), Radchenko & James (2010), Bach et al. (2012), Bien et al. (2013). The most common applications considered in these works are to regression models.

The convex banding estimator employs a new hierarchical group structure that is tailored to covariance matrices with a banded or semi-banded structure; its optimal properties cannot be obtained from simple extensions of any of the existing related penalties. We discuss this in Sections 2.1 and 4. We also provide a connection between our convex banded estimator and tapering estimators. Section 3.1 shows that our estimator can also be written in the form (1.1), but where $\mathbf{T}$ is a data-driven, sparse tapering matrix, with entries given by a data-dependent recursion formula, not a pre-specified, non-random, tapering function. This representation has both practical and theoretical implications: it allows us to compute our estimator efficiently, relate our estimator to previous banded or tapered estimators, and establish that it is banded and positive definite with high probability. These issues are treated in Sections 3.1 and 4.3, respectively.

In Section 4.1, we prove that, when the population covariance matrix is itself a banded matrix, our estimator recovers the sparsity pattern, under minimal signal strength conditions, which are made possible by the fact that we employ a hierarchical penalty. In Section 4.2 we show that our convex banding estimator is minimax rate adaptive, in both Frobenius and operator norms, up to multiplicative logarithmic factors, over appropriately defined classes of population matrices. In Section 5, we perform a thorough simulation study that investigates the sharpness of our bounds and demonstrates that our estimator compares favorably to previous methods. We conclude with a real data example in which we show that using our estimate of the covariance matrix leads to improved classification accuracy in both quadratic and linear discriminant analysis.

2 The definition of the convex banding estimator

To describe our estimator, we must begin by defining a set of groups that will induce the desired banded-sparsity pattern. We define

$$g_\ell = \{jk \in [p]^2 : |j - k| \geq p - \ell\}$$

to be the two right triangles of indices formed by taking the $\ell + 1$ indices that are farthest from the main diagonal of a $p \times p$ matrix. See the left panel of Figure 1 for a depiction of g_3 when $p = 5$. For notational ease we will denote, as above, $jk = (j, k)$, and $[p]^2 = \{1, \dots, p\} \times \{1, \dots, p\}$.

We will also find it useful to express these groups as a union of subdiagonals s_m :

$$g_\ell = \bigcup_{m=1}^{\ell} s_m \text{ where } s_m = \{jk : |j - k| = p - m\}.$$

For example, $g_1 = s_1 = \{(1, p), (p, 1)\}$ and, at the opposite extreme, $g_{p-1}^c = s_p$ is the diagonal. While the indexing of g_ℓ and s_m may at first seem “backwards” (in the sense that we count them from the outside-in rather than from the diagonal-out), our indexing is natural here

because $|s_m| = 2m$ and g_ℓ consists of two equilateral triangles with side-lengths of ℓ elements. The right panel of Figure 1 depicts the nested group structure: $g_1 \subset \dots \subset g_{p-1}$, where this largest group contains all off-diagonal elements.

The following notation is used in the definition of our estimator below. Given a subset of matrix indices $g \subseteq [p]^2$ of a $p \times p$ matrix Σ , let $\Sigma_g \in \mathbb{R}^{|g|}$ be the vector with elements $\{\Sigma_{jk} : jk \in g\}$. For a given non-negative sequence of weights $w_{\ell m}$ with $1 \leq m \leq \ell \leq p-1$, discussed in the following sub-section, and for a given $\lambda \geq 0$, let $\hat{\Sigma}$ be defined as the solution to the following convex optimization problem:

$$\hat{\Sigma} = \arg \min_{\Sigma} \left\{ \frac{1}{2} \|\Sigma - \mathbf{S}\|_F^2 + \lambda \|\Sigma\|_{2,1}^* \right\}, \quad (2.1)$$

where $\|\cdot\|_F$ is the Frobenius norm and

$$\|\Sigma\|_{2,1}^* = \sum_{\ell=1}^{p-1} \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|\Sigma_{s_m}\|_2^2} = \sum_{\ell=1}^{p-1} \|(\mathbf{W}^{(\ell)} * \Sigma)_{g_\ell}\|_2. \quad (2.2)$$

Our penalty term is a weighted group lasso, using $\{g_\ell : 1 \leq \ell \leq p-1\}$ as the group structure. For the second equality, we express the penalty as the elementwise product, denoted by $*$, of Σ with a sequence of weight matrices, $\mathbf{W}^{(\ell)}$, that are Toeplitz with $\mathbf{W}_{s_m}^{(\ell)} = w_{\ell m} \mathbf{1}_{\{m \leq \ell\}} \cdot \mathbf{1}_{2m}$.

Remark 1

Problem 2.1 is strictly convex, so $\hat{\Sigma}$ is the unique solution.

Remark 2

As the tuning parameter λ is increased, subdiagonals of $\hat{\Sigma}$ become exactly zero. As long as $w_{\ell m} > 0$ for all $1 \leq m \leq \ell \leq p-1$, the hierarchical group structure ensures that a subdiagonal will only be set to zero if all elements farther from the main diagonal are already zero. Thus,

we get an estimated bandwidth, $\hat{K}$, which satisfies $\sum_{g_{p-\hat{K}-1}} = 0$ and $\sum_{g_{p-\hat{K}}} \neq 0$ (see Theorem 2 and Corollary 1 for details).

We refer to $\hat{\Sigma}$ as the *convex banding* of the matrix $\mathbf{S}$. We show in Section 4.3 that $\hat{\Sigma}$ is positive definite with high probability. Empirically, we find that positive definiteness holds except in very extreme cases. Moreover, in Section 4.3 we propose a different version of convex banding that guarantees positive definiteness:

$$\tilde{\Sigma} = \arg \min_{\tilde{\Sigma} \in \delta \mathbf{I}_p} \left\{ \frac{1}{2} \|\tilde{\Sigma} - \mathbf{S}\|_F^2 + \lambda \|\tilde{\Sigma}\|_{2,1}^* \right\}. \quad (2.3)$$

Of course when $\lambda_{\min}(\hat{\Sigma}) \geq \delta$, the two estimators coincide.

2.1 The weight sequence $w_{\ell m}$

Since for each ℓ the weight $w_{\ell m}$ penalizes s_m , $1 \leq m \leq \ell$ and the subdiagonals s_m increase in size with m , we want to give the largest subdiagonals the largest weight. We will therefore choose weights with $w_{\ell m} = \max_{1 \leq m \leq \ell} w_{\ell m}$ and will sometimes write $w_{\ell} = w_{\ell \ell}$. We consider three choices of weights that satisfy this property.

1. (Non-hierarchical) Group lasso penalty—Letting $w_{\ell m} = 1_{\{\ell=m\}} \sqrt{2\ell}$, for $1 \leq m \leq \ell \leq p-1$ yields

$$\left\| \sum_{\ell=1}^{p-1} \left\| \sum_{m=1}^{\ell} s_m \right\|_2^* \right\|_2 = \sum_{\ell=1}^{p-1} \sqrt{2\ell} \left\| \sum_{m=1}^{\ell} s_m \right\|_2. \quad (2.4)$$

This is a traditional group lasso penalty that acts on subdiagonals, with weights $w_{\ell} = \sqrt{|s_{\ell}|}$. This penalty is not hierarchical, as each subdiagonal may be set to zero independently of the others. In Section 4.1 we show how failing to use a hierarchical penalty requires more stringent minimum signal conditions in order to correctly recover the true bandwidth of Σ^* . However, if the interest is in accurate estimation in Frobenius or operator norm of population matrices that are close to banded, we show in Section 4.2.1 that this estimator is minimax rate optimal, up to logarithmic factors, despite the fact that in finite samples it may fail to have the correct sparsity pattern.

2. Basic hierarchical penalty—If $w_{\ell m} = \sqrt{2\ell}$, for all m and ℓ the penalty term becomes

$$\left\| \sum_{\ell=1}^{p-1} \left\| \sum_{m=1}^{\ell} s_m \right\|_2^* \right\|_2 = \sum_{\ell=1}^{p-1} \sqrt{2\ell} \left\| \sum_{m=1}^{\ell} s_m \right\|_2, \quad (2.5)$$

which employs the same weight, $\sqrt{2\ell}$, as above, but now for a triangle g_{ℓ} . Recalling that $|g_{\ell}| = \ell(\ell+1)$, we note that this does not follow the common principle guiding weight choices in the (non-overlapping) group lasso literature of using $\sqrt{|g_{\ell}|}$. In Section 4.1, we show this choice of weights leads to consistent bandwidth selection under minimal conditions on the strength of the signal; however, we shall show that a more refined choice ensures rate optimal estimation of Σ^* with respect to either Frobenius or operator norm. The problem with this simple choice of weights stems from the fact that subdiagonals far from the main diagonal (small m) are excessively penalized (since they appear in $p-m$ terms). To balance this overaggressive enforcement of hierarchy, one desires weights that decay with decreasing m within a fixed group g_{ℓ} and yet still exhibit the $\sqrt{2\ell}$ growth on w_{ℓ} .

3. General hierarchical penalty—Based on the considerations above, we take the following choice of weights:

$$w_{\ell m} = \frac{\sqrt{2\ell}}{\ell - m + 1}, \text{ for } 1 \leq m \leq \ell, 1 \leq \ell \leq p - 1. \quad (2.6)$$

We will show in Sections 4.2.1 and 4.2.2 that the corresponding estimator is minimax rate adaptive, in Frobenius and operator norm, up to logarithmic factors, over appropriately defined classes of population covariance matrices.

Remark 3: We re-emphasize the difference between the weighting schemes considered above. The estimators corresponding to (2.5) and (2.6) will always impose the sparsity structure of a banded matrix, a fact that is apparent from Theorem 2 below. In contrast, the group lasso estimator (2.4), which is performed on each sub-diagonal separately, may fail to recover this pattern.

3 Computation and properties

The most common approach to solving the standard group lasso is blockwise coordinate descent (BCD). Tseng (2001) proves that when the nondifferentiable part of a convex objective is separable, BCD is guaranteed to solve the problem. While this separable structure is not present in (2.1), Jenatton et al. (2010) show that the dual of this problem does, meaning that BCD on the dual can be used.

Theorem 1

A dual of (2.1) is given by

$$\underset{\mathbf{A}^{(\ell)} \in \mathbb{R}^{p \times p}}{\text{Minimize}} \frac{1}{2} \left\| \mathbf{S} - \lambda \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)} \right\|_F^2 \quad \text{s.t.} \quad \|\mathbf{A}_{g_\ell}^{(\ell)}\|_2 \leq 1, \mathbf{A}_{g_{\ell^c}}^{(\ell)} = 0 \text{ for } 1 \leq \ell \leq p-1. \quad (3.1)$$

In particular, given a solution to the dual, $(\hat{\mathbf{A}}^{(1)}, \dots, \hat{\mathbf{A}}^{(p-1)})$, the solution to (2.1) is given by the primal-dual relation, $\hat{\Sigma} = \mathbf{S} - \lambda \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)}$.

Proof—See Section A.1 of the supplementary materials.

Algorithm 1 gives a BCD algorithm for solving (3.1), which by the primal-dual relation in turn gives a solution to (2.1). The blocks correspond to each dual variable matrix. The update over each $\mathbf{A}^{(\ell)}$ involves projection of a matrix $\hat{\mathbf{R}}^{(\ell)}$ (defined in Algorithm 1) onto an ellipsoid, which amounts to finding $\hat{\nu}_\ell$, a root of the univariate function $h_\ell(\nu) - \lambda^2$, where

$$h_\ell(\nu) = \sum_{m=1}^{\ell} \frac{w_{\ell m}^2}{(w_{\ell m}^2 + \nu)^2} \|\hat{\mathbf{R}}_{s_m}^{(\ell)}\|^2. \quad (3.2)$$

We explain in Section A.2 of the supplementary materials the details of ellipsoid projection and observe that we can get $\hat{\nu}_\ell$ in closed form for all but the last $\hat{K}$ values of $\hat{\mathcal{K}}$ where $\hat{K}$ is the bandwidth of $\hat{\Sigma}$. A remarkable feature of our algorithm is that only a single pass of BCD is required to reach convergence. This property is proved in Jenatton et al. (2011) and is a direct consequence of the nested group structure. When we use the simple weights (2.5), in which $w_{\ell m} = w_\ell$ Algorithm 1 becomes extraordinarily simple and transparent:

1. Initialize $\hat{\Sigma} \leftarrow \mathbf{S}$
2. For $\ell = 1, \dots, p-1$: $\hat{\Sigma}_{g_\ell} \leftarrow (1 - \lambda_{w_\ell} / \|\hat{\Sigma}_{g_\ell}\|_2) + \hat{\Sigma}_{g_\ell}$.

In words, we start with the sample covariance matrix, and then work from the corners of the matrix inward toward the main diagonal. At each step, the next largest triangle-pair is group-soft-thresholded. If a triangle is ever set to zero, it will remain zero for all future steps. We will show that this simple weighting scheme admits exact bandwidth and pattern recovery in Section 4.1.

Algorithm 1 BCD on dual of Problem (2.1).

Inputs: $\mathbf{S}$, λ , and weights matrices, $\mathbf{W}^{(\ell)}$. Initialize $\hat{\mathbf{A}}^{(\ell)} = 0$ for all ℓ

For $\ell = 1, \dots, p-1$:

- Compute $\hat{\mathbf{R}}^{(\ell)} \leftarrow \mathbf{S} - \lambda \sum_{\ell'=1}^{p-1} \mathbf{W}^{(\ell')} * \hat{\mathbf{A}}^{(\ell')}$
- For $m \in \mathcal{L}_{\text{set}}$ $\hat{\mathbf{A}}_{sm}^{(\ell)} \leftarrow \frac{w_{\ell m}}{\lambda(w_{\ell m}^2 + \max\{\hat{\nu}_\ell, 0\})} \hat{\mathbf{R}}_{sm}^{(\ell)}$ where $\hat{\nu}_\ell$ satisfies $\lambda^2 = h_\ell(\hat{\nu}_\ell)$, as in (3.2).

$\{\hat{\mathbf{A}}^{(\ell)}\}$ is a solution to (3.1) and $\hat{\mathbf{R}}^{(p)} = \mathbf{S} - \lambda \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)}$ is the solution to (2.1).

3.1 Convex banding as a tapering estimator

The next result shows that $\hat{\Sigma}$ can be regarded as a tapering estimator with a data-dependent tapering matrix. In Section 4, we will see that, in contrast to estimators that use a fixed tapering matrix, our estimator adapts to the unknown bandwidth of the true matrix Σ^* .

Theorem 2—The convex banding estimator, $\hat{\Sigma}$, can be written as a tapering estimator with a Toeplitz, data-dependent tapering matrix, $\hat{\Sigma} = \hat{\mathbf{T}} * \mathbf{S}$:

$$\hat{\mathbf{T}}_{sm} = \begin{cases} \mathbf{1}_m & \text{for } m=p \text{ (diagonal)} \\ \prod_{\ell=m}^{p-1} \frac{[\hat{\nu}_\ell] +}{w_{\ell m}^2 + [\hat{\nu}_\ell] +} \mathbf{1}_m & \text{for } 1 \leq m \leq p-1 \end{cases}$$

where $\hat{\nu}_\ell$ satisfies $\lambda^2 = \sum_{m=1}^{\ell} \frac{w_{\ell m}^2}{(w_{\ell m}^2 + \hat{\nu}_\ell)^2} \|\hat{\mathbf{R}}_{sm}^{(\ell)}\|^2$ and $\mathbf{1}_m \in \mathbb{R}^m$ is the vector of ones.

Proof: See Section A.3 of supplementary materials. $\square$

The following result shows that, as desired, our estimator is banded.

Corollary 1— $\hat{\Sigma}$ is a banded matrix with bandwidth $\hat{K} = p - 1 - \max\{\ell: \hat{\nu}_\ell \leq 0\}$.

Proof: By definition, $\hat{\nu}_{p-1-\hat{K}} \leq 0$ and $\hat{\nu}_{p-\hat{K}}, \dots, \hat{\nu}_{p-1} > 0$. It follows from the theorem that $\hat{\mathbf{T}}_{sm} = 0.1_m$ for all $m \leq p - 1 - \hat{K}$ and $[\hat{\mathbf{T}}_{sm}]_i > 0$ for $p - \hat{K} \leq m \leq p - 1$. $\square$

4 Statistical properties of convex banding

In this section, we study the statistical properties of the convex banding estimator. We begin by stating two assumptions that specify the general setting of our theoretical analysis.

Assumption 1

Let $\mathbf{X} = (X_1, \dots, X_p)^T$. Assume $\mathbf{X} = 0$ and denote $\mathbf{X}\mathbf{X}^T = \Sigma^*$. We assume that each X_j is marginally sub-Gaussian: $\exp(tX_j / \sqrt{\Sigma_{jj}^*}) \leq \exp(Ct^2)$ for all $t \geq 0$ and for some constant $C > 0$ that is independent of j . Moreover, $\max_j |\Sigma_{jj}^*| \leq M$, for some constant $M > 0$.

Assumption 2

The dimension p can grow with n at most exponentially in n : $\gamma_0 \log n \leq \log p \leq \gamma n$, for some constants $\gamma_0 > 0$, $\gamma > 0$.

The requirement that p grow at least polynomially fast as a function of n is not necessary but it simplifies the results allowing us to write $\log p$ instead of $\log[\max(p, n)]$. In Section 4.1, we prove that our estimator recovers the true bandwidth of Σ^* with high probability assuming the nonzero values of Σ^* are large enough. We will demonstrate that our estimator can detect lower signals than what could be recovered by an estimator that does not enforce hierarchy, re-emphasizing the need for a hierarchical penalty. In Section 4.2, we show that our estimator is minimax adaptive, up to multiplicative logarithmic terms, with respect to both the operator and Frobenius norms. Our results hold either with high probability or in expectation, and are established over classes of population matrices defined in Sections 4.2.1 and 4.2.2, respectively.

We begin by introducing a random set, on which all our results hold. Fixing $x > 0$, let

$$\mathcal{A}_x = \left\{ \max_{1 \leq i, j \leq p} |S_{ij} - \Sigma_{ij}^*| \leq x \sqrt{\log p / n} \right\}. \quad (4.1)$$

The following lemma shows that this set has high probability.

Lemma 1

Under Assumptions 1 and 2, there exists a constant $c > 0$ such that for sufficiently large $x > 0$, $\mathbb{P}(\mathcal{A}_x) \geq 1 - c/p$.

Proof—See Section A.4 of the supplementary materials, proof of (A.4.1) of Theorem A.1. $\square$

Remark 4

- i. Similar results exist in the literature (see, e.g., Ravikumar et al. 2011). Lemma 3 in Bickel & Levina (2008) is proved under a Gaussianity assumption coupled with the assumption that $\|\Sigma^*\|_{op}$ is bounded. Whereas inspection of the proof shows that the latter is not needed, we cannot quote this result directly for other types of design. The commonly employed assumptions of sub-Gaussianity are placed on the entire vector $\mathbf{X} = (X_1, \dots, X_p)^T$ and postulate that there exists $\tau > 0$ such that $\mathbb{P}\{|\mathbf{v}^T(\mathbf{X} - (\mathbf{X}))| > t\} \leq \exp\{-t^2/(2\tau)\}$, for all $t > 0$ and $\|\mathbf{v}\|_2 = 1$, (see, e.g., Cai & Zhou 2012). However, if such a τ exists, then $\|\Sigma^*\|_{op} \leq \tau$. Lemma 1 above shows that a bound on $\|\Sigma^*\|_{op}$ can be avoided in the probability bounds regarding $\max_{ij} |S_{ij} - \sum_{ij}^*|$, and the distributional assumption can be weakened to marginals.
- ii. In the classical case in which p does not depend on n , one can modify the definition of the set $\mathcal{A}_x$ by replacing the factor $\log(p)$ by $\log[\max(n, p)]$, and the result of Lemma 1 will continue to hold with probability $1 - c/\max(n, p)$, for a possibly different constant c .

4.1 Exact bandwidth recovery

Suppose Σ^* has bandwidth K , that is, for $L = p - K - 1$, we have $\sum_{g_L}^* = 0$ and $\sum_{s_{L+1}}^* \neq 0$ (see right panel of Figure 1). We prove in this section that under mild conditions our estimator $\hat{\Sigma}$ correctly recovers K with high probability. The next theorem expresses the intuitive result that if λ is chosen sufficiently large, we will not over-estimate the true bandwidth.

Theorem 3—If $\lambda \geq x \sqrt{\log p/n}$ and $w_\ell = \sqrt{2\ell}$, then $\hat{K} \leq K$, on the set $\mathcal{A}_x$.

Proof: See Section A.5 of the supplementary materials. $\square$

For our estimator to be able to detect the full band, we must require that the “signal” be sufficiently large relative to the noise. In the next theorem we measure the size of the signal by the ℓ_2 norm of each sub-diagonal (scaled in proportion to the square root of its size, $w_\ell = \sqrt{2\ell}$) and the size of the noise by λ .

Theorem 4—If $\min_{\ell \geq L+1} \|\sum_{s_\ell}^* \|_2 / w_\ell > 2\lambda$, where $\lambda \geq x \sqrt{\log p / n}$ and $w_\ell = \sqrt{2\ell}$, then $K \leq \hat{K}$, on the set $\mathcal{A}_x$.

Proof: See Section A.5 of the supplementary materials. $\square$

In typical high-dimensional statistical problems, the support set can generically be any subset and therefore one must require that each element in the support set be sufficiently large on its own to be detected. By contrast, in our current context we know that the support set is of the specific form $\{L+1, \dots, p-1\}$, for some unknown L . Thus, as long as the signal is sufficiently large at $L+1$, one might expect that the signal could be weaker at subsequent elements of the support. In the next theorem we demonstrate this phenomenon

by showing that when $\sum_{s_{L+1}}^*$ exceeds the threshold given in the previous theorem, it may “share the wealth” of this excess, relaxing the requirement on the size of $\sum_{s_{L+2}}^*$.

Theorem 5 (“Share the wealth”)—Suppose (for some $\gamma > 0$)

$$\|\sum_{s_{L+1}}^* \|_2 = (2+\gamma)\lambda w_{L+1}, \quad \min_{\ell \geq L+3} \|\sum_{s_\ell}^* \|_2 / w_\ell > 2\lambda, \quad \|\sum_{s_{L+2}}^* \|_2 > \lambda w_{L+2} (1 + \sqrt{1-\gamma^2}) 1_{\{\gamma < 1\}}$$

where $\lambda \geq x \sqrt{\log p / n}$ and $w_\ell = \sqrt{2\ell}$ and $w_{\ell+1} > 0$, then $K \leq \hat{K}$ on the set $\mathcal{A}_x$.

Proof: See Section A.5 of the supplementary materials. $\square$

When $\gamma = 0$, Theorem 5 reduces to Theorem 4. However, as γ increases, the required size of

$\|\sum_{s_{L+2}}^* \|_2$ decreases without preventing the bandwidth from being misestimated. In fact, for

$\gamma \leq 1$, there is no requirement on $\sum_{s_{L+2}}^*$ for bandwidth recovery. This robustness to individual subdiagonals being small is a direct result of our method being hierarchical.

Remark 5

- i. By Lemma 1, $\mathbb{P}(\mathcal{A}_x) > 1 - c/p$, under Assumption 1 and 2. Thus, all the results of this section hold with this probability.
- ii. Theorems 3 and 4 apply to all three weighting schemes considered in Section 2.1. Theorem 5’s additional requirement of positivity excludes the “Group lasso” weights (2.4), which is non-hierarchical and therefore is unable to “share the wealth.”

4.2 Minimax adaptive estimation

4.2.1 Frobenius norm rate optimality—In this section we show that the estimator $\hat{\Sigma}$, with weights given by either (2.4) or (2.6), is minimax adaptive, up to multiplicative logarithmic factors, with respect to the Frobenius norm, over a class of population

covariance matrices that generalizes both the class of K -banded matrices and previously studied classes of approximately banded matrices. We begin by stating Theorem 6, which is a general oracle inequality from which adaptivity to the optimal minimax rate will follow as immediate corollaries.

For any $\mathbf{B} \in \mathbb{R}^{p \times p}$, let $L(\mathbf{B})$ be an integer such that $\mathbf{B}_{g_{L(\mathbf{B})}} = 0$ and $\mathbf{B}_{s_{L(\mathbf{B})+1}} \neq 0$, that is $\mathbf{B}$ has bandwidth $K(\mathbf{B}) = p - 1 - L(\mathbf{B})$. Let $\mathcal{S}_p$ be the class of all $p \times p$ positive definite covariance matrices. In Theorem A.2 of Section A.6 of the supplementary materials, we

provide a deterministic upper bound on $\|\hat{\Sigma} - \Sigma^*\|_F^2$ that indicates what terms need to be controlled in order to obtain optimal stochastic performance of our estimator. The following theorem, which is the main result of this section, is a consequence of this theorem, and shows that for appropriate choices of the weights and tuning parameter, the proposed estimator achieves the best bias-variance trade-off both in probability and in expectation, up to small additive terms, which are the price paid for adaptivity.

Theorem 6: Suppose $\Sigma^* \in \mathcal{S}_p$ and $\max_j |\Sigma_{jj}^*| \leq M$ for some constant M . If the weights are given by either (2.4) or (2.6) and $\lambda = x \sqrt{\log p/n}$, then on the set $\mathcal{A}_x$ the convex banding estimator of (2.1) satisfies

$$\|\hat{\Sigma} - \Sigma^*\|_F^2 \leq \inf_{\mathbf{B} \in \mathbb{R}^{p \times p}} \left\{ \|\Sigma^* - \mathbf{B}\|_F^2 + 16x^2 \frac{K(\mathbf{B})p \log p}{n} \right\} + x^2 \frac{p \log p}{n}.$$

Moreover, for x sufficiently large,

$$\|\hat{\Sigma} - \Sigma^*\|_F^2 \lesssim \inf_{\mathbf{B} \in \mathbb{R}^{p \times p}} \left\{ \|\Sigma^* - \mathbf{B}\|_F^2 + \frac{K(\mathbf{B})p \log p}{n} \right\} + p/n$$

where the symbol $\lesssim$ is used to denote an inequality that holds up to positive multiplicative constants that are independent of n and p .

Proof: See Section A.6 of the supplementary materials. $\square$

For any given $K \geq 1$, the class of K -banded matrices is

$$\mathcal{B}(K) = \mathcal{B}(K, M) =: \left\{ \Sigma \in \mathcal{S}_p : \sum_{g_{p-K-1}} = 0, \sum_{s_{p-K}} \neq 0 \text{ and } \max_j |\Sigma_{jj}| \leq M \right\}.$$

Motivated by Theorem 6 above, we define the following general class of covariance matrices, henceforth referred to as K -semi-banded:

$$\mathcal{G}(K) = \mathcal{G}(K, M) := \left\{ \Sigma \in \mathcal{S}_p : \max_j |\Sigma_{jj}| \leq M \text{ and } \|\Sigma - \mathbf{B}\|_F^2 \lesssim pK \log p/n \text{ for some } \mathbf{B} \in \mathcal{B}(K, M) \right\}, \quad (4.2)$$

for any $K \geq 1$. Trivially, $\mathcal{B}(K) \subset \mathcal{G}(K)$, for any $K \geq 1$. Notice that a semi-banded covariance matrix does not have any explicit restrictions on the order of magnitude of $\|\hat{\Sigma}\|_{op}$; nor does it require exact zero entries.

Theorem 7 below gives the minimax lower bound, in probability, for estimating a covariance matrix $\Sigma^* \in \mathcal{B}(K)$, in Frobenius norm, from a sample $\mathbf{X}_1, \dots, \mathbf{X}_n$. Let $\mathbb{P}_K$ be the class of probability distributions of p -dimensional vectors satisfying Assumption 1 and with covariance matrix in $\mathcal{B}(K)$.

Theorem 7: If $K \leq \sqrt{n}$, there exist absolute constants $c > 0$ and $\beta \in (0, 1)$ such that

$$\inf_{\Sigma \in \mathbb{P}_K} \sup P(\|\hat{\Sigma} - \Sigma^*\|_F^2 > cpK/n) \geq \beta. \quad (4.3)$$

To the best of our knowledge, the minimax lower bound for Frobenius norm estimation over $\mathcal{B}(K)$ is new. The proof of the existing related result in Cai et al. (2010), established for approximately banded matrices, or that in Cai & Zhou (2012), for *approximately sparse* matrices, could be adapted to this situation, but the resulting proof would be involved, as it would require re-doing all their calculations. We provide in Section A.7 of the supplementary materials a different and much shorter proof, tailored to our class of covariance matrices.

From Theorems 6 and 7 above we conclude that convex banding with either the (non-hierarchical) group lasso penalty, (2.4), or the general hierarchical penalty, (2.6), achieves, adaptively, the minimax rate, up to log terms, not only over $\mathcal{B}(K)$, but over the larger class $\mathcal{G}(K)$ of semi-banded matrices. We summarize this below.

Corollary 2: Suppose $\Sigma^* \in \mathcal{G}(K)$. The convex banding estimator of (2.1), with weights given by either (2.4) or (2.6), and with $\lambda = x \sqrt{\log p/n}$ satisfies

$$\|\hat{\Sigma} - \Sigma^*\|_F^2 \lesssim pK \log p/n$$

on the set $\mathcal{A}_x$ and also $\|\hat{\Sigma} - \Sigma^*\|_F^2 \lesssim pK \log p/n$.

To the best of our knowledge, this is the first estimator which, in particular, has been proven to be minimax adaptive over the class of exactly banded matrices, up to logarithmic factors.

It should be emphasized that, despite certain similarities, the problem we are studying is different from that of estimating the autocovariance matrix of a stationary time series based on the observation of a single realization of the series. For a discussion of this distinction and a treatment of this latter problem, see Xiao et al. (2012).

Remark 6: An immediate consequence of Theorem A.2 of the supplementary materials is that for the basic hierarchical penalty, (2.5), we have, on the set $\mathcal{A}_x$,

$\|\hat{\Sigma} - \Sigma^*\|_F^2 \lesssim pK^2 \log p/n$, for $\Sigma^* \in \mathcal{B}(K)$, which is a sub-optimal rate. Therefore, if Frobenius norm rate optimality is of interest and one desires a banded estimator, then the simple weights (2.5) do not suffice, and one must resort to the more general ones given by (2.6).

Relation to previously studied classes of matrices: The class of semi-banded matrices is more general than the previously studied class of approximately banded matrices considered in Cai et al. (2010):

$\mathcal{C}_\alpha = \left\{ \Sigma \in \mathcal{S}_p : |\sum_{ij} \Sigma_{ij}| \leq M_1 |i - j|^{-(\alpha+1)} \text{ for all } i \neq j \text{ and } \lambda_{\max}(\Sigma) \leq M_0 \right\}$ Bickel & Levina (2008) and Cai & Yuan (2012) consider a closely related class,

$\mathcal{D}_\alpha = \left\{ \Sigma \in \mathcal{S}_p : \max_j \sum_{i: |i-j| > k} |\Sigma_{ij}| \leq M_1 |i - j|^{-\alpha}, \text{ for all } k > 0, 0 < M_0^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M_0 \right\}$

Interestingly, the class of approximately banded matrices $\mathcal{C}_\alpha$ does not include, in general, the class of exactly banded matrices $\mathcal{B}(K)$, as membership in these classes forces neighboring entries to decay rapidly, a restriction that is not imposed on elements of $\mathcal{B}(K)$. Moreover, the largest eigenvalue of a K -banded matrix can be of order K (e.g., consider a block diagonal matrix with p/K blocks of size K , with fixed within-block correlations), and hence cannot be bounded by a constant when K is allowed to grow with n and p . Therefore, $\mathcal{C}_\alpha$ is different from, but not necessarily more general than, $\mathcal{B}(K)$, for arbitrary values of K .

In contrast, the class $\mathcal{G}(K)$ introduced in (4.2) above is more general than $\mathcal{B}(K)$, for all K ,

and also contains $\mathcal{C}_\alpha$ for an appropriate value of K . Let $K_\alpha \lesssim \left(\frac{n}{\log p}\right)^{\frac{1}{2\alpha+2}}$. Then, it is immediate to see that, for any $\alpha > 0$, $\mathcal{C}_\alpha \subset \mathcal{G}(K_\alpha)$. To gain further insight into particular types of matrices in $\mathcal{G}(K_\alpha)$, notice that it contains, for instance, matrices in which neighboring elements do not have to start decaying immediately, but only when they are far enough from the main diagonal. Specifically, for some positive constants $M_1 < M_0$ and any $\alpha > 0$, let

$\mathcal{G}_\alpha = \left\{ \Sigma \in \mathcal{S}_p : |\sum_{ij} \Sigma_{ij}| \leq M_0, \text{ if } |i - j| \leq K_\alpha \text{ and } |\sum_{ij} \Sigma_{ij}| \leq M_1 |i - j - K_\alpha|^{-(\alpha+1)}, \text{ if } |i - j| > K_\alpha \right\}$

Then we have, for any fixed $\alpha > 0$ and appropriately adjusted constants, that

$\mathcal{C}_\alpha \subset \mathcal{G}_\alpha \subset \mathcal{G}(K_\alpha)$ and $\mathcal{B}(K_\alpha) \subset \mathcal{G}_\alpha \subset \mathcal{G}(2K_\alpha)$. The following corollary gives the rate of estimation over $\mathcal{G}(K_\alpha)$.

Corollary 3: Suppose $\Sigma^* \in \mathcal{G}(K_\alpha)$. The convex banding estimator of (2.1), with weights given by either (2.4) or (2.6), and with $\lambda = x \sqrt{\log p/n}$ where x is as in Theorem 6 satisfies.

$$(i) \|\hat{\Sigma} - \Sigma^*\|_F^2/p \lesssim \left(\frac{\log p}{n}\right)^{\frac{2\alpha+1}{2\alpha+2}} \text{ on the set } \mathcal{A}_x \text{ and } (ii) \|\hat{\Sigma} - \Sigma^*\|_F^2/p \lesssim \left(\frac{\log p}{n}\right)^{\frac{2\alpha+1}{2\alpha+2}}$$

The upper bound in (ii), up to logarithmic factors, is the minimax lower bound derived in Cai et al. (2010) over $\mathcal{C}_\alpha$ and $\mathcal{D}_\alpha$. Therefore, Corollary 3 shows that the minimax rate, with respect to the Frobenius norm, can be achieved, *adaptively* over $\alpha > 0$, not only over the classes of approximately banded matrices, but also over the larger class of semi-banded matrices $\mathcal{G}(K_\alpha)$. In contrast, the two existing estimators with established minimax rates of convergence in Frobenius norm require knowledge of α prior to estimation. This is the case for the banding estimator of Bickel & Levina (2008) and the tapering estimator of Cai et al. (2010), over the class $\mathcal{D}_\alpha$. Adaptivity to the minimax rate, over $\mathcal{D}_\alpha$, with respect to the Frobenius norm, is alluded to in Section 5 of Cai & Yuan (2012), who proposed a block-thresholding estimator, but a formal statement and proof are left open.

4.2.2 Operator norm rate optimality—In this section we study the operator norm optimality of our estimator over classes of matrices that are close, in operator norm, to banded matrices, with bandwidths that might increase with n or

$$p:\mathcal{G}_{op}(K) = \left\{ \Sigma \in \mathcal{S}_p : \|\Sigma - \mathbf{B}\|_{op} \lesssim K \sqrt{\log p/n} \text{ for some } \mathbf{B} \in \mathcal{B}(K) \right\}.$$

Minimax lower bounds in operator norm over the classes of approximately sparse matrices and approximately banded matrices are provided in Cai & Zhou (2012) and Cai et al. (2010). Moreover, an inspection of the proof of Theorem 3 in Cai et al. (2010) shows that the proof of the bound stated in Theorem 8 above follows immediately from it, by the same arguments, by letting their k be our K . We therefore state the bound for clarity and completeness, and omit the proof. The following theorem gives the minimax lower bound, in probability, for estimating a covariance matrix $\Sigma^* \in \mathcal{G}_{op}(K)$, in operator norm, from a sample $\mathbf{X}_1, \dots, \mathbf{X}_n$. Let $\mathbb{P}_K$ be the class of probability distributions of p -dimensional vectors satisfying Assumption 1 and with covariance matrix in $\mathcal{B}(K)$.

Theorem 8: If $K < \sqrt{n}$, there exist absolute constants $c > 0$ and $\beta \in (0, 1)$ such that

$$\inf_{\Sigma^* \in \mathcal{G}_{op}(K)} \sup_{\mathbb{P}_K} P \left(\|\hat{\Sigma} - \Sigma^*\|_{op} > cK/\sqrt{n} \right) \geq \beta \quad (4.4)$$

Having established the best possible rate, we next show that our proposed estimator achieves adaptively, up to logarithmic factors, the target rate.

Theorem 9: The convex banding estimator of (2.1) with weights given by either (2.4) or (2.6), and with $\lambda = 2x \sqrt{\log p/n}$, where x is as in Theorem 6, satisfies

$$\|\hat{\Sigma} - \Sigma^*\|_{op} \lesssim K \sqrt{\log p/n}, \text{ on the set } \mathcal{A}_x,$$

for any $\Sigma^* \in \mathcal{G}_{op}(K)$ that meets the signal strength condition

$$\min_{\ell > L(\Sigma^*)} \|\Sigma_{s_\ell}^*\|_2 / \sqrt{2\ell} \geq c \text{ for some constant } c > 0.$$

Proof: See Section A.8 of the supplementary materials. $\square$

Remark 7: We show in Section A.9 of the supplementary materials that if the signal strength condition fails and $p > n$, then the estimator $\hat{\Sigma}$ may not even be consistent in operator norm.

Our focus in this section was on the behavior of our estimator over a class that generalizes that of banded matrices, such as $\mathcal{G}_{op}$, since no minimax adaptive estimators in operator norm have been constructed, to the best of our knowledge, for this class. Our estimator is however slightly sub-optimal, in operator norm, over different classes, for instance over the class of approximately banded matrices. It is immediate to show that it achieves, adaptively, the slightly suboptimal rate $O\{(\log p/n)^{\frac{\alpha}{2\alpha+2}}\}$ over the class $\mathcal{D}_\alpha$, when a signal strength condition similar to that given in Theorem 9 is satisfied. However, the minimax optimal rate over this class is $O\{(\log p/n)^{\frac{\alpha}{2\alpha+1}}\}$, and has been shown to be achieved, non-adaptively, by the tapering estimator of Cai et al. (2010) and in Xiao & Bunea (2014), by the banding estimator. For the latter case, a SURE-type bandwidth selection is proposed, as an alternative to the usual cross validation-type criteria, for the practical bandwidth choice. Moreover, the optimal rate of convergence in operator norm can be achieved, adaptively, by a block-thresholding estimator proposed by Cai & Zhou (2012). For these reasons, we do not further pursue the issue of operator norm minimax adaptive estimation over $\mathcal{D}_\alpha$ in this work.

4.3 Positive definite, banded covariance estimation

In this section, we study the positive definiteness of our estimator. Empirically, we have observed $\hat{\Sigma}$ to be positive definite except in quite extreme circumstances (e.g., very low n settings). We begin by proving that $\hat{\Sigma}$ is positive definite with high probability assuming certain conditions.

Theorem 10—Under Assumptions 1 and 2, the convex banding estimator with weights given by either (2.4) or (2.6) and with $\lambda = 2x \sqrt{\log p/n}$ (for appropriately large x) has minimum eigenvalue at least $C' K \sqrt{\log p/n} > 0$ with probability larger than $1 - c/p$, for some $c > 0$, provided that (i) $\lambda_{\min}(\Sigma^*) \geq 2C' K \sqrt{\log p/n}$ and (ii) $\min_{\ell > L(\Sigma^*)} \|\Sigma_{s_\ell}^*\|_2 / \sqrt{2\ell} \geq c'$ for some constant $c' > 0$.

Proof: See Section A.10 of the supplementary materials. $\square$

Although we find empirically that $\hat{\Sigma}$ is reliably positive definite, it is reasonable to desire an estimator that is *always* positive definite. In such a case, we suggest using the estimator $\tilde{\Sigma}$ defined in (2.3). Algorithm 1 of Section A.10 of the supplementary materials provides a BCD approach similar to Algorithm 1.

Corollary 4—Under the assumptions of Theorem 10, $\tilde{\Sigma} = \hat{\Sigma}$ with high probability as long as we choose $\delta \leq C' K \sqrt{\frac{\log p}{n}}$.

We conclude this section by noting that, in light of the corollary, $\tilde{\Sigma}$ has all the same properties as $\hat{\Sigma}$ under the additional assumptions of Theorem 10.

5 Empirical study

In this section, we study the empirical performance of our estimator. In Section 5.1, we investigate how our estimator's performance depends on K , n , and p , and we compare our method to several existing methods for banded covariance estimation. Section 5.2 considers using the convex banding estimator and several other banded estimators within the context of a classification problem.

5.1 Simulations

5.1.1 Convergence rate dependence on K —We fix $n = 100$, and for $p \in \{500, 1000, 2000\}$ and ten equally spaced values of K between 10 and 500, we simulate data according to an instance of a moving-average(K) process. In particular, we draw n multivariate normal vectors with covariance matrix $\Sigma_{ij}^* = (1 - |i - j|/K) \cdot 1\{|i - j| \leq K\}$. We draw 20 samples of size n to compute the average scaled squared Frobenius norm, $\|\hat{\Sigma} - \Sigma^*\|_F^2/p$, and operator norm, $\|\hat{\Sigma} - \Sigma^*\|_{op}$. Figure 2 shows how the squared Frobenius norm varies with K when we take $\lambda=2 \sqrt{\log(p)/n}$. It is seen to grow approximately linearly in K in agreement with our bound for this quantity given in Corollary 2. In the right panel, we scale the error by $\log p$, which appears on the right-hand side of Corollary 2. The fact that with this scaling the curves become more aligned shows that the p dependence of the bound holds (approximately). Also, we see that the behavior matches that of the bounds most closely for large p and small K . In Section A.11 of the supplementary materials we show that the empirical behavior of our estimator in operator norm likewise corroborates the rate derived in Theorem 9. In Figure 3 we repeat the same simulation for $\hat{\Sigma}$ with the simple weighting scheme, $w_{\ell_m} = \sqrt{2\ell}$. We find that this weighting scheme performs empirically worse than the general weighting scheme of (2.6); this is also seen in further simulations included in Section A.11 of the supplementary materials.

5.1.2 Convergence rate dependence on n —To investigate the dependence on sample size, we vary n along an equally-spaced grid from 10 to 500, fixing $K = 50$ and taking, as before, $p \in \{500, 1000, 2000\}$. Simulating as before and taking, again $\lambda = 2 \sqrt{\log(p)/n}$, Figure 4 exhibits the $1/n$ dependence suggested by the bound in Corollary 2. In Section A.11 of the supplementary materials we observe the $1/\sqrt{n}$ behavior for operator norm in agreement with Theorem 9.

5.1.3 Comparison to other banded estimators—In this section, we compare the performance of the following methods that perform banded covariance estimation: (i) **Hier Band**: our method with weights given in (2.6) and no eigenvalue constraint; (ii) **Simple Hier Band**: our method with weights given in (2.5) and no eigenvalue constraint; (iii) **GP Band**: group lasso without hierarchy (weighting scheme given in (2.4)) and no eigenvalue constraint; (iv) **Block Thresh**: Cai & Yuan (2012)’s block-thresholding approach (our implementation); (v) **BandChol**: Rothman et al. (2010)’s banding of the Cholesky factor approach via solving successive regression problems; (vi) **Banding**: Bickel & Levina (2008)’s $\mathbf{T} * \mathbf{S}$ estimator with $\mathbf{T}_{jk} = 1_{\{|j-k| \leq K\}}$.

Each method has a single tuning parameter that is varied to its best-case value. In the case of **Banding**, this makes it equivalent to an oracle-like procedure in which all non-signal elements are set to zero and all signal elements are left unshrunk.

While we focus on estimation error as a means of comparison, it should be noted that these methods could be evaluated in other ways as well. For example, **Banding** and **Block Thresh** frequently lead to estimators that are not positive semidefinite. This means that these covariance estimates cannot be directly used in other procedures requiring a covariance matrix (see, e.g., Section 5.2 for an example where not being positive definite would be problematic).

We simulate four scenarios for the basis of our comparison: (i) **MA(5)** with $p = 200$, $n = 100$. See Section 5.1.1 for description. (ii) **MA(80)** with $p = 500$, $n = 100$. See Section 5.1.1 for description. (iii) **CY**: A setting used in Cai & Yuan

(2012) in which Σ^* is only approximately banded: $\Sigma_{ij}^* = 0.6|i - j|^{-2}U_{ij}$, where $U_{ij} = U_{ji} \sim \text{Uniform}(0, 1)$. We take $p = 200$, $n = 100$. (iv) **Increasing operator norm**: We produce a sequence of Σ^* having increasing operator norm. In particular, we take a block-diagonal matrix with p/K blocks of size K : $\Sigma = I_{p/K} \otimes B_K$, each block is

$B_K = (0.8)1_K 1_K^T + (0.2)I_K$; we take $p = 1000$, $n = 100$, and vary K between 20 and 400. We can make several observations from this comparison. We find that the hierarchical group lasso methods outperform the regular group lasso. This is to be expected since in our simulation scenarios, the true covariance matrix is banded (or approximately banded). As noted before, since we report the best performance of each method over all values of the tuning parameter, **Banding** is effectively an “oracle” method in that the banding is being performed exactly where it should be for **MA(5)**. So it is no surprise that this method does so well in this scenario. However, in the approximately banded scenario it

suffers, likely because it does not do as well as methods that do shrink the nonzero values. We find that the `<m>Block Thresh</m>` does not perform very well in any of these scenarios (though Figure 5 does show it substantially improves on the sample covariance matrix). Finally, we find that the `<m>BandChol</m>` method performs well (although not shown here, we also observed that the related Nested Lasso method of Rothman et al. (2010) performs quite well).

Figure 6 shows the results for the third scenario, in which we consider a sequence of Σ^* 's that are block diagonal with operator norm increasing linearly in K . Upon examination of the `<m>Block Thresh</m>` method's estimates, we find that the reason this estimator performs poorly in the regime of large K is that it sets all elements to zero that are sufficiently far from the main diagonal without regard to their observed size in $\mathbf{S}$ (this is referred to as step b in their paper). `<m>Banding</m>` performs very well in terms of operator norm, which makes sense since Σ^* is a banded matrix and so simple banding with the correct bandwidth is essentially an oracle procedure.

5.2 An application to discriminant analysis of phoneme data

In this section, we develop an application of our banded estimator. In general, one would expect any procedure that requires an estimate of the covariance matrix to benefit from convex banding when the true covariance matrix is banded (or approximately banded). Consider the classification setting in which we observe n i.i.d. pairs, $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of predictors and y_i labels the class of the i th point. In quadratic discriminant analysis (QDA), one assumes that $\mathbf{x}_i/y_i = k \sim N_p(\boldsymbol{\mu}^{*(k)}, \Sigma^{*(k)})$. The QDA classification rule is to assign $\mathbf{x} \in \mathbb{R}^p$ to the class k maximizing $\hat{\mathbb{P}}(y=k|\mathbf{x})$. Here $\hat{\mathbb{P}}$ denotes the estimate of $\mathbb{P}(y=k|\mathbf{x})$ given by replacing the parameters $\boldsymbol{\mu}^{*(k)}$, $\Sigma^{*(k)}$, and $\mathbb{P}(y=k)$ with their maximum likelihood estimates.

To demonstrate the use of our covariance estimate in QDA, we consider a binary classification problem described in Hastie et al. (2009). The dataset consists of short sound recordings of male voices saying one of two similar sounding phonemes, and the goal is to build a classifier for automatic labeling of the sounds. The predictor vectors, $\mathbf{x}$, are log-periodograms, representing the (log) intensity of the recordings across $p = 256$ frequencies. Because the predictors are naturally ordered, one might expect a regularized estimate of the within-class covariance matrix to be appropriate, and inspection of the sample covariance matrices (see Figure A.3 in Section A.12 of the supplementary materials) supports this. There are $n_1 = 695$ and $n_2 = 1022$ phonemes in the two classes. We randomly split the data into equal sized training and test sets. On the training set, five-fold cross validation is

performed to select tuning parameters (λ_1, λ_2) and then the convex banding estimates, $\sum_{\lambda_1}^{(1)}$ and $\sum_{\lambda_2}^{(2)}$, are used in place of the sample covariances to form the basis of the prediction rule, $\hat{\mathbb{P}}(y=k|\mathbf{x})$. The left panel of Figure 7 shows that QDA can be substantially improved by using a regularized estimate (and, in particular, that convex banding does quite well). Linear discriminant analysis (LDA) is like QDA but assumes a common covariance matrix between classes, $\Sigma^{*(k)} = \Sigma^*$, which is typically estimated as

$\mathbf{S}_w = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)\mathbf{S}^{(1)} + (n_2 - 1)\mathbf{S}^{(2)}]$, where $\mathbf{S}^{(k)} = (n_k - 1)^{-1} \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}}^{(k)})(\mathbf{x}_i - \bar{\mathbf{x}}^{(k)})^T$. The right panel of Figure 7 shows that LDA does better than the QDA methods, as expected in this regime of p and n (see, e.g., Cheng 2004) and can itself be improved by a regularized version of $\mathbf{S}_w$, in which we replace $\mathbf{S}^{(k)}$ in the above expression with $\hat{\sum}_{\lambda_k}^{(k)}$ (again, cross validation is performed to select the tuning parameters).

6 Conclusion

We have introduced a new kind of banding estimator of the covariance matrix that has strong practical and theoretical performance. Formulating our estimator through a convex optimization problem admits precise theoretical results and efficient computational algorithms. We prove that our estimator is minimax rate adaptive in Frobenius norm over the class of semi-banded matrices introduced in Section 4.2.1 and minimax rate adaptive in operator norm over the class of banded matrices, up to multiplicative logarithmic factors. Both classes of matrices over which optimality is established are allowed to have bandwidth growing with n , p or both. The construction of adaptive estimators over these classes, with established rate optimality is, to the best of our knowledge, new. Moreover, the proposed estimators recover the true bandwidth with high probability, for truly banded matrices, under minimal signal strength conditions. In contrast to existing estimators, our proposed estimator, which enjoys all the above theoretical properties, is guaranteed to be both exactly banded, and positive definite, with high probability. We also propose a version of the estimator, that is guaranteed to be positive definite in finite samples. Simulation studies show that the hierarchically banded estimator is a strong and fast competitor of the existing estimators. Through a data example, we demonstrate how our estimator can be effectively used to improve classification error of both quadratic and linear discriminant analysis. Indeed, we expect that many common statistical procedures that require a covariance matrix may be improved by using our estimator when the assumption of semi-bandedness is appropriate. An `R` (R Core Team 2013) package, named `hierband`, will be made available, implementing our estimator.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Jacob Bien was supported by National Science Foundation grant DMS-1405746, Florentina Bunea by National Science Foundation grant DMS-10007444, and Luo Xiao by National Institute of Neurological Disorders and Stroke Grant R01NS060910. We thank Adam Rothman for providing `R` code for the Nested Lasso method and also Xiaohan Yan and Guo Yu for finding typos in an earlier version.

References

Bach F, Jenatton R, Mairal J, Obozinski G. Structured sparsity through convex optimization. *Statistical Science*. 2012; 27(4):450–468.

- Bickel PJ, Levina E. Regularized estimation of large covariance matrices. *The Annals of Statistics*. 2008;199–227.
- Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. *The Annals of Statistics*. 2013; 41(3):1111–1141. [PubMed: 26257447]
- Bunea F, Xiao L. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. To appear in *Bernoulli*. 2014 arXiv:1212.5321.
- Cai TT, Yuan M. Adaptive covariance matrix estimation through block thresh-olding. *The Annals of Statistics*. 2012; 40(4):2014–2042.
- Cai TT, Zhang CH, Zhou HH. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*. 2010; 38(4):2118–2144.
- Cai TT, Zhou HH. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*. 2012; 40(5):2389–2420.
- Cheng Y. Asymptotic probabilities of misclassification of two discriminant functions in cases of high dimensional data. *Statistics & Probability Letters*. 2004; 67(1):9–17.
- Dempster AP. Covariance selection. *Biometrics*. 1972; 28(1):157–175.
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Second. Springer Verlag; New York: 2009.
- Jenatton R, Audibert J, Bach F. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*. 2011; 12:2777–2824.
- Jenatton R, Mairal J, Obozinski G, Bach F. Proximal methods for sparse hierarchical dictionary learning. *Proceedings of the International Conference on Machine Learning (ICML)*. 2010
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2013. URL: <http://www.R-project.org/>
- Radchenko P, James GM. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*. 2010; 105(492):1541–1553.
- Ravikumar P, Wainwright MJ, Raskutti G, Yu B, et al. High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*. 2011; 5:935–980.
- Rothman A, Levina E, Zhu J. A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*. 2010; 97(3):539.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc B*. 1996; 58:267–288.
- Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*. 2001; 109(3):475–494.
- Xiao H, Wu WB, et al. Covariance matrix estimation for stationary time series. *The Annals of Statistics*. 2012; 40(1):466–493.
- Xiao L, Bunea F. On the theoretic and practical merits of the banding estimator for large covariance matrices. arXiv preprint. 2014 arXiv:1402.0844.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*. 2006; 68:49–67.
- Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*. 2009; 37(6A):3468–3497.

Figure 1.

(Left) The group g_3 ; (Middle) the nested groups in the penalty (2.1); (Right) a K-banded matrix where $K = 2$ and $p = 6$. Note that $L = p - K - 1 = 2$ counts the number of

subdiagonals that are zero. This can be expressed as $\sum_{g_L} = 0$ or $\sum_{s_\ell} = 0$ for $1 \leq \ell \leq L$.

Figure 2.

Convergence in Frobenius norm. Monte Carlo estimate of (Left) $E\|\hat{\Sigma} - \Sigma^*\|_F^2/p$ and (Right) $E\|\hat{\Sigma} - \Sigma^*\|_F^2/(p \log p)$ as a function of K , the bandwidth of Σ^*

Figure 3.
Comparison of our procedure with simple and general weights.

Figure 4.

(Left:) For large n , expected $E\|\hat{\Sigma} - \Sigma^*\|_F^2/p$ is seen to decay like n^{-1} as can be seen from -1 slope of log-log plot (gray lines are of slope -1). (Right:) Scaling this quantity by $\log p$ aligns the curves for large n . Both of these phenomena are suggested by Corollary 2.

Figure 5. Comparison of methods in three settings in terms of (Top) squared Frobenius norm, relative to $\|S - \Sigma^*\|_F^2$ and (Bottom) operator norm relative to $\|S - \Sigma^*\|_{op}$.

Figure 6.
Comparison of methods in terms of (Left) squared Frobenius norm and (Right) operator norm in a setting where $\|\Sigma^*\|_{\text{op}}$ is increasing.

Figure 7. Test set misclassification error rates of discriminant analysis of phoneme data. For both QDA and LDA, using a regularized covariance matrix improves the misclassification error rate.