Submitted to the Annals of Statistics

TRANSFER LEARNING FOR COVARIANCE MATRIX ESTIMATION: OPTIMALITY AND ADAPTIVITY

BY T. TONY CAI [a,c], AND DONGWOO KIM[b]

Transfer learning, which leverages knowledge from an auxiliary source dataset to improve performance in a primary target domain, has emerged as a pivotal machine learning technique. In this paper, we consider minimax and adaptive estimation of large bandable covariance matrices within the transfer learning framework.

We first establish the minimax rate of convergence under the spectral norm and propose a rate-optimal estimation procedure. Our findings reveal intriguing phase transition phenomena that highlight the effectiveness of transfer learning and the use of source samples. We then address the problem of adaptation, establishing the adaptive rate of convergence up to a logarithmic factor. Our results demonstrate that, in sharp contrast to conventional settings, the cost of adaptation in transfer learning can be substantial in certain cases. We propose a novel data-driven algorithm that dynamically adapts to unknown model parameters. These theoretical insights are further validated by a simulation study, demonstrating the practicality and efficiency of the proposed adaptive algorithm.

1. Introduction. In the realm of high-dimensional data analysis, understanding the covariance structure is fundamental. Covariance matrices underpin numerous statistical methods, including regression analysis, discriminant and clustering analyses, principal component analysis, and Gaussian graphical models. However, their estimation in high dimensions poses formidable challenges, primarily due to the rapid degradation of sample covariance estimates. This issue has spurred extensive research efforts on estimation and inference for high-dimensional covariance matrices. See Cai, Ren and Zhou [8] for an in-depth review of large covariance matrix estimation and Cai [3] for a survey on global testing and large-scale multiple testing for covariance structures.

To address the challenges posed by high dimensionality and motivated by scientific applications, several classes of structured covariance matrices—including bandable, sparse, and spiked covariance matrices—have been studied in high-dimensional settings. In particular, estimation of bandable covariance matrices, which capture the intrinsic ordering and metric for variable indices that naturally emerge across various disciplines such as economics, biology, phonetics, climatology, and engineering, has received significant attention in the literature. Noteworthy examples include financial market analysis [17], prostate cancer assessment [34], phoneme examination [2], climate research [23], traffic prediction [21], and sonar spectrum inspection [35, 45]. Several methods have been developed to exploit the bandable structure, and minimax optimality has been investigated. See, for example, Bickel and Levina [1], Cai, Zhang and Zhou [11], and Cai and Yuan [10].

Concurrently, *transfer learning* has emerged as a pivotal technique in machine learning. It utilizes the knowledge acquired from an auxiliary source dataset to enhance the performance of statistical tasks in a primary target domain. This approach has found important applications in various fields, including large language models [15], bioinformatics [32, 38], medical imaging [27, 39], and recommendation systems [18, 30, 29]. It has thus established itself as a

MSC2020 subject classifications: Primary 62H12; secondary 62F12.

Keywords and phrases: Adaptive rate of convergence, Adaptivity, Bandable structure, Covariance matrix, High-dimensional statistics, Minimax rate of convergence, Phase transition, Transfer learning.

cornerstone for building efficient and high-performing machine learning models. For a more detailed exploration, see Weiss, Khoshgoftaar and Wang [44] and Zhuang et al. [47]. Inspired by these practical successes, there has been notable recent progress in the theoretical quantification of statistical transfer learning. This includes work on nonparametric classification [9, 19] and nonparametric regression [7, 26, 31], generalized linear models [25, 40], contextual multi-armed bandits [4], and functional mean estimation [6].

When these two powerful concepts intersect, they create a synergy capable of addressing more complex and intriguing statistical problems. Integrating covariance matrix estimation with transfer learning facilitates the utilization of different yet similar source datasets, proving invaluable when target observations are limited or prohibitively expensive to collect. Prominent examples include the analysis of Genotype-Tissue Expression (GTEx) data across various organs [22, 33], electroencephalogram (EEG)-based brain-computer interface (BCI) classification across different sessions and subjects [24, 28, 37, 46], and text classification using word co-occurrence patterns [36]. By transferring knowledge from source datasets, one can significantly enhance the estimation accuracy of the target covariance matrix.

1.1. *Formulation of the problem.* Our discussion begins with an illustrative example of covariance matrix estimation within the transfer learning context. We focus on phonemes, which are the fundamental units of sound in language. The dataset, meticulously curated by Hamooni and Mueen [14], comprises short audio recordings of 39 distinct English phonemes. Each data point is represented as a periodogram vector, capturing sound intensity across various frequencies.

Zooming in on a particular phoneme, say `ZH`, we are now given a target sample denoted as $\{X_1^{(t)}, \ldots, X_{n_t}^{(t)}\}$. These are d-dimensional random vectors that are independently and identically distributed. Within the context of our phoneme study, the dimensionality d corresponds to the maximum frequency of the periodogram. Each observation has a target mean vector $\mu^{(t)} \in {}^{d}$ and a target covariance matrix $\Sigma^{(t)} \in {}^{d\times d}$. The conventional learning framework centers around finding an effective method to estimate the target covariance matrix based on the available target sample.

An essential feature of the periodogram vector lies in its natural ordering from low to high frequencies, which imparts a distinctive structure to the target covariance matrix $\Sigma^{(t)}$. Entries near the diagonal denote covariances between intensities at proximate time points, and they are thus anticipated to possess greater magnitudes. Conversely, entries distant from the diagonal, representing covariances between widely separated time points, are expected to approach zero. This covariance pattern is illustrated in Figure A.3 of Section A.12 by Bien, Bunea and Xiao [2].

The described structure of the phoneme dataset aligns with a bandable structure, a concept formulated by Bickel and Levina [1] and Cai, Zhang and Zhou [11]. We examine the away-from-diagonal operator, denoted $\mathfrak{A}_d$, on the set of $d \times d$ matrices. Given a bandwidth $u > 0$, the operator is defined by:

$$\mathfrak{A}_d(B; u) := \Big(b_{jk}\big(|j-k| > u\big)\Big)_{j,k\in[d]} \qquad \text{where } B = (b_{jk})_{j,k\in[d]} \in {}^{d\times d}.$$

Figure 2a in Section 2.1 provides a visual representation of the away-from-diagonal operator $\mathfrak{A}_d$. It retains elements beyond a certain distance from the main diagonal. The bandwidth $u > 0$ given in the operator specifies this threshold of distance. Under the bandable structure, it is assumed that the magnitude of the away-from-diagonal part diminishes geometrically as the bandwidth increases. We define the class $\mathcal{B}_\alpha(M)$ for bandable covariance matrices characterized by a decay rate $\alpha > 0$ and radius $M > 0$ as follows:

$$\mathcal{B}_\alpha(M) := \big\{\Sigma \in {}_d^+ : \|\mathfrak{A}_d(\Sigma; u)\|_1 \le M u^{-\alpha} \text{ for any } u > 0\big\},$$

where $\overset{+}{d}$ represents the collection of $d \times d$ symmetric and positive definite matrices and $\|\cdot\|_1$ indicates the matrix ℓ_1-norm. We posit that the target covariance matrix conforms to the bandable structure, meaning $\Sigma^{(t)} \in \mathcal{B}_{\alpha_t}(M_t)$ for a target decay rate $\alpha_t > 0$ and a target radius $M_t > 0$.

For covariance matrix estimation, we would encounter a significant challenge within the conventional learning framework, particularly when it comes to the ZH phoneme. Data points for this phoneme are typically derived from continuous speech recordings, such as the TIMIT dataset provided by the Linguistic Data Consortium [13]. However, the phoneme ZH is the least frequently occurring sound in everyday English speech [14, 16], making its recordings scarce and the collection of additional data prohibitively expensive.

One way to address this limitation is to extend the conventional learning framework to include the transfer learning strategy. This approach allows us to utilize recordings of other phonemes that are different yet bear some similarities. For instance, both ZH and S phonemes are categorized as obstruent and fricative sounds in the hierarchy of English phonemes [12, 20]. Moreover, the phoneme S is significantly more common, ranking as the fifth most frequent sound in English speech [14, 16]. Consequently, we can leverage the similarity and larger sample size of the S phoneme to enhance the covariance matrix estimation for the ZH phoneme.

From a statistical perspective, in addition to having an independently and identically distributed target sample $\{X_1^{(t)}, \ldots, X_{n_t}^{(t)}\}$ from a target distribution Q, we have an auxiliary sample, $\{X_1^{(s)}, \ldots, X_{n_s}^{(s)}\}$, comprising d-dimensional random vectors that are independently and identically distributed from a source distribution P. The source sample is independent of the target sample and is also characterized by its source mean vector $\mu^{(s)} \in {}^{d}$ and source covariance matrix $\Sigma^{(s)} \in {}^{d\times d}$. While our objective remains the same to find an optimal estimation method for the target covariance matrix $\Sigma^{(t)}$, our methodology may benefit from leveraging the source sample alongside the target sample. Figure 1 gives an illustration of the transfer learning strategy.

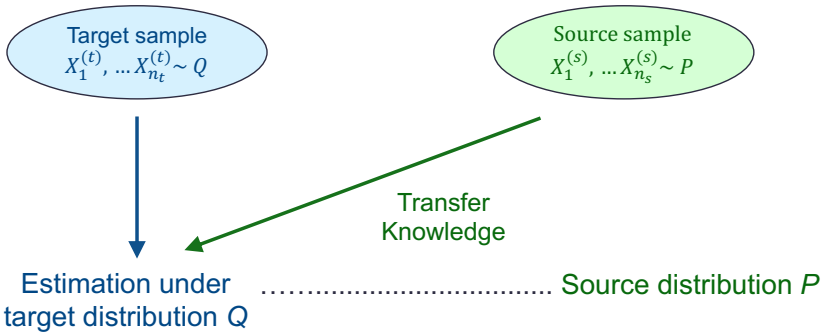


Fig 1: An illustration for transfer learning.

The success of transfer learning is contingent upon the establishment of a substantive and quantifiable relationship between the source and target samples. In this paper, we postulate that:

- The source covariance matrix $\Sigma^{(s)}$ exhibits a bandable structure as well. Assuming a source decay rate $\alpha_s > 0$ and a source radius $M_s > 0$, we have $\Sigma^{(s)} \in \mathcal{B}_{\alpha_s}(M_s)$.
- The disparity matrix, $\Delta^{(s)} := \Sigma^{(t)} - \Sigma^{(s)}$, representing the deviation between the target and source covariance matrices, is sparse in terms of the matrix ℓ_1-norm. For a given disparity threshold $\delta \geq 0$, we assume $\|\Delta^{(s)}\|_1 \leq \delta$.

1.2. *Main results and our contribution.* We first establish the minimax rate of convergence. By letting $n := n_t + n_s$ and $\alpha_{\max} := \alpha_t \vee \alpha_s$, the minimax optimal rate under the matrix ℓ_2-loss is expressed as:

$$\Phi_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d) := \left(n^{-2\alpha_{\max}/(2\alpha_{\max}+1)} \wedge \frac{d}{n}\right) + \frac{\log d}{n} + \left(n_t^{-2\alpha_t/(2\alpha_t+1)} \wedge \frac{d}{n_t} \wedge \delta^2\right) + \left(\frac{\log d}{n_t} \wedge \delta^2\right). \tag{1}$$

When there is no auxiliary source sample and no constraints on the disparity matrix, our setting simplifies to the conventional learning framework and the corresponding minimax rate $\Phi_{\alpha_t,\alpha_s,\infty}(n_t,0,d)$ reduces to the known result in the conventional setting as detailed in Cai, Zhang and Zhou [11]. Moreover, by comparing the minimax rate to the baseline performance, $\Phi_{\alpha_t,\alpha_s,\infty}(n_t,0,d)$, we can identify the necessary and sufficient conditions under which the source sample and transfer learning are indeed effective. This identification reveals interesting phase transition phenomena.

The disparity threshold δ emerges as a crucial factor in phase transition. Our finding indicates that transfer learning loses its efficacy when δ exceeds a specific value. When δ is below this critical value, the source sample size n_s and the auto-smoothing effect become instrumental. Specifically, transfer learning proves advantageous with sufficient source samples, while limited source samples necessitate the auto-smoothing benefit. Finally, an increase in source sample size does not always enhance the performance of transfer learning. We identify a secondary threshold for δ, below which a larger source sample size meaningfully contributes to more effective transfer learning.

The minimax rate is achieved by our novel approach, termed the blockwise tridiagonal estimator. This method selectively utilizes either the target or pooled sample covariance matrix, depending on the similarity between the target and source population covariance matrices. The blockwise tridiagonal operator is then applied to the chosen sample covariance matrix, retaining only the main, super, and subdiagonals from a blockwise standpoint. This approach is logical, considering the bandable structure of the covariance matrix. In the end, the blockwise tridiagonal estimator attains optimality by carefully choosing the appropriate block size or bandwidth.

Although the blockwise tridiagonal estimator attains the minimax optimal rate, it depends on the decay rates α_t and α_s as well as the disparity threshold δ, which are typically unknown in practice. The second critical insight of this paper addresses the challenge of adaptation to unknown parameters. We establish the following optimal *adaptive rate of convergence* in the case where the model parameters are unknown:

$$\Phi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d) := \left(n^{-2\alpha_{\max}/(2\alpha_{\max}+1)} \wedge \frac{d}{n}\right) + \frac{\log d}{n} + \left(\varphi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d) \wedge \frac{d}{n_t} \wedge \delta^2\right) + \left(\frac{\log d}{n_t} \wedge \delta^2\right), \tag{2}$$

where it is further defined that $\alpha_{\min} := \alpha_t \wedge \alpha_s$ and

$$\varphi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d) := \begin{cases} n_t^{-2\alpha_t/(2\alpha_t+1)} & \text{if } n_s \le n_t, \\ n_t^{-2\alpha_{\min}/(2\alpha_{\min}+1)} & \text{if } n_s > n_t. \end{cases} \tag{3}$$

It is worth noting that the adaptive rate of convergence, as given in Equation (2), is uniformly slower than the minimax rate presented in Equation (1). This inherent disparity is an unavoidable aspect of adaptivity, as dictated by the principle of the adaptive rate. A fully data-driven algorithm cannot achieve universal optimality across all models; it must inevitably

sacrifice some degree of optimality for certain models and regimes. Furthermore, the adaptive rate essentially represents the minimum *cost of adaptation* associated with the absence of knowledge of model parameters. An adaptive estimator that exceeds the adaptive rate for certain models must fall short in others. Crucially, any performance gains from the adaptive rate are consistently outweighed by the losses.

The cost of adaptation could be significant in certain settings. However, possessing additional information can refine the scope of the problem and potentially eliminate these costs. For instance, as previously discussed, if it is known that $n_s = 0$ and $\delta = \infty$, the problem simplifies to the conventional learning framework under which adaptation can be realized at no extra cost (see Cai and Yuan [10], for example). Beyond this basic case, we further explore scenarios where adaptivity can be realized without additional costs.

The adaptive rate in Equation (2) is attained within a factor of $\log n$ by our proposed algorithm, termed the adaptive tridiagonal block-thresholding estimator. This algorithm is entirely data-driven and adjusts dynamically to unknown model parameters. The primary challenges are optimal bandwidth selection and reduction of potential bias when incorporating the source sample. The introduction of the superbanding operator addresses the former by partitioning the high-dimensional bias into more manageable lower-dimensional segments. Meanwhile, the latter challenge is tackled through disparity matrices and blockwise thresholding, both of which are meticulously designed to minimize the adaptation costs.

1.3. *Organization and notation.* The rest of the paper is organized as follows. We conclude this section by introducing the basic notation. Section 2 explores optimal transfer learning for covariance matrix estimation, presenting the blockwise tridiagonal estimator and establishing the minimax rate of convergence. Section 3 focuses on adaptivity, introducing the adaptive tridiagonal block-thresholding algorithm and establishing the adaptive rate of convergence. Section 4 carries out numerical experiments to evaluate the effectiveness of the proposed adaptive algorithm in various settings. Section 5 explores possible directions of future research. Finally, the proofs of the main theorems and technical results are presented in the Supplementary Material [5].

Throughout the paper, the primary asymptotic components are the sample sizes (n_t, n_s), the dimensionality (d) and the disparity threshold (δ), while all other variables are treated as constants. We conform to the conventional big-Oh(O), big-Omega(Ω) and big-Theta(Θ) notations. For conciseness, we occasionally substitute these notations with the symbols $\lesssim$, $\gtrsim$ and $\asymp$, respectively. Additionally, we adhere to the convention of little-oh(o) notation and simply write $a \ll b$ or $b \gg a$ to signify that $a = o(b)$.

In addition, let us introduce several notations pertinent to covariance matrices. We define the pooled covariance matrix $\Sigma \in {}^{d\times d}$, which represents the weighted average of the population covariance matrices, $\Sigma^{(t)}$ and $\Sigma^{(s)}$. Concurrently, we define another disparity matrix $\Delta \in {}^{d\times d}$ as the deviation of the target covariance matrix $\Sigma^{(t)}$ from the pooled covariance matrix Σ. By letting $n := n_t + n_s$, these are mathematically represented as:

$$\Sigma := \frac{n_t}{n}\Sigma^{(t)} + \frac{n_s}{n}\Sigma^{(s)}, \qquad \text{and} \qquad \Delta := \Sigma^{(t)} - \Sigma = \frac{n_s}{n}\Delta^{(s)}.$$

We proceed to define the sample counterparts of the covariance matrices. The target sample covariance matrix $\check{\Sigma}^{(t)}$ and the source sample covariance matrix $\check{\Sigma}^{(s)}$ are defined that for each group indicator $g \in \{t, s\}$,

$$\check{\Sigma}^{(g)} := \frac{1}{n_g}\sum_{i=1}^{n_g}\big(X_i^{(g)} - \bar{X}^{(g)}\big)\big(X_i^{(g)} - \bar{X}^{(g)}\big)^\top \qquad \text{where } \bar{X}^{(g)} := \frac{1}{n_g}\sum_{i=1}^{n_g} X_i^{(g)}.$$

Subsequently, we define the pooled sample covariance matrix $\check{\Sigma}$ as a weighted combination of the target and source sample covariance matrices:

$$\check{\Sigma} := \frac{n_t}{n}\check{\Sigma}^{(t)} + \frac{n_s}{n}\check{\Sigma}^{(s)}.$$

Finally, it is apparent to define the sample disparity matrices $\check{\Delta}^{(s)}$ and $\check{\Delta}$ as:

$$\check{\Delta}^{(s)} := \check{\Sigma}^{(t)} - \check{\Sigma}^{(s)}, \qquad \text{and} \qquad \check{\Delta} := \check{\Sigma}^{(t)} - \check{\Sigma} = \frac{n_s}{n}\check{\Delta}^{(s)}.$$

2. Optimal transfer learning. In this section, our goal is to present an optimal method for estimating the target covariance matrix $\Sigma^{(t)}$ within the transfer learning framework. In particular, we first introduce a novel algorithm and evaluate its maximal risk in terms of the squared matrix ℓ_2-norm, denoted as $\|\cdot\|_2$, in Section 2.1. Subsequently, in Section 2.2, we validate the optimality of the proposed approach by deriving a matching lower bound, thereby establishing the minimax rate of convergence.

To explicate the concept of optimality in covariance matrix estimation, let us introduce a statistical model, denoted by $\mathcal{P}_{\alpha_t,\alpha_s,\delta} = \mathcal{P}_{\alpha_t,\alpha_s\delta}(M_t, M_s, \rho)$, which represents the collection of joint distributions for the target and source observations. In particular, the target and source samples are independently generated from the procedure outlined in Section 1.1. Moreover, they conform to the uniform sub-Gaussianity assumption described as follows:

ASSUMPTION 1 (Uniform sub-Gaussianity). The target observations $X_1^{(t)}, \ldots, X_{n_t}^{(t)}$ and the source observations $X_1^{(s)}, \ldots, X_{n_s}^{(s)}$ are uniformly sub-Gaussian random vectors with a standard-deviation proxy $\rho > 0$. More specifically, we assume:

$$\left.\begin{aligned} \big\|a^\top\big(X_1^{(t)} - X_1^{(t)}\big)\big\|_{\psi_2} \\ \big\|a^\top\big(X_1^{(s)} - X_1^{(s)}\big)\big\|_{\psi_2} \end{aligned}\right\} \le \rho, \qquad \text{for any unit vector } a \in {}^d,$$

where we denote by $\|\cdot\|_{\psi_2}$ the sub-Gaussian norm or Orlicz ψ_2-norm. As per Vershynin [42] and Wainwright [43], the sub-Gaussian norm of a random variable Z is defined by

$$\|Z\|_{\psi_2} := \inf\big\{r > 0 : e^{Z^2/r^2} \le 1\big\}.$$

2.1. *Estimation method and minimax upper bound.* We begin by introducing an operator on matrices, called *blockwise tridiagonal operator* denoted by $\mathfrak{T}_d$. This operator is defined for an integer bandwidth $u > 0$ and can be expressed as follows. Here, $\lfloor x \rfloor$ represents the greatest integer not exceeding $x \in$.

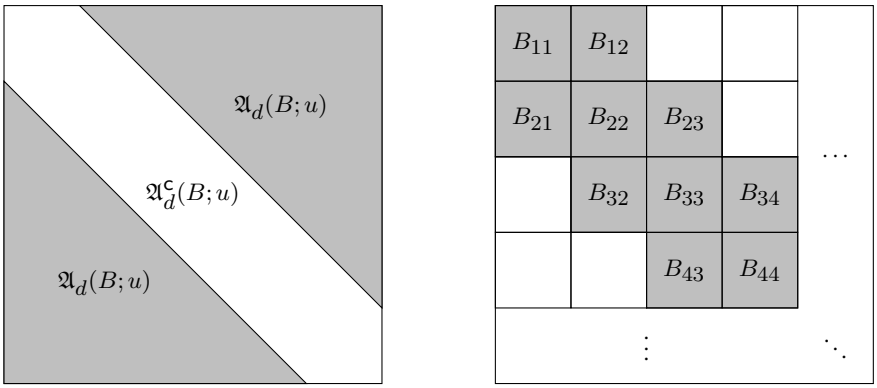
$$\mathfrak{T}_d(B; u) := \Big(b_{jk}\big(|u_j - u_k| \le 1\big)\Big)_{j,k\in[d]} \quad \text{where} \quad \begin{cases} B = (b_{jk})_{j,k\in[d]} \in {}^{d\times d}, \\ u_j := 1 + \big\lfloor (j-1)/u \big\rfloor, \ (j \in [d]). \end{cases}$$

Given a matrix $B = (b_{jk})_{j,k\in[d]} \in {}^{d\times d}$ and a bandwidth $u > 0$, we can segment B into block matrices of size $u \times u$. These blocks are sequentially denoted as B_{xy} $(x, y = 1, 2, \ldots)$ such that the block B_{xy} consists of every (j,k)-th entry, b_{jk}, satisfying $(u_j, u_k) = (x, y)$. It is essential to note that the terminal blocks along each row and column may be smaller than $u \times u$, contingent upon the relationship between the bandwidth u and the dimension d.

The blockwise tridiagonal operator $\mathfrak{T}_d$ only retains elements satisfying $|u_j - u_k| \le 1$, effectively preserving the main, super, and subdiagonals in a blockwise perspective. any blocks beyond these specified diagonals are nullified by the operator $\mathfrak{T}_d$. This mechanism is graphically represented in Figure 2b. Consequently, the matrix, $\mathfrak{T}_d(B; u)$, produced by the

blockwise tridiagonal operator can be viewed as a tridiagonal matrix from a block matrix standpoint.

To facilitate subsequent discussions, we might extend the integer bandwidth. When the bandwidth $u > 0$ is specified as a non-integer value, we will simply use its integer part, $\lfloor u \rfloor$, in the application of the blockwise tridiagonal operator. From now on, we will assume a general bandwidth $u > 0$ when considering the blockwise tridiagonal operator $\mathfrak{T}_d(\,\cdot\,; u)$.



(a) The output, $\mathfrak{A}_d(B;u)$, of the away-from-diagonal operator

(b) The output, $\mathfrak{T}_d(B;u)$, of the blockwise tridiagonal operator

Fig 2: A visual representation of important matrix operators

We are ready to introduce the estimator for the target covariance matrix $\Sigma^{(t)}$. The key concept involves employing the blockwise tridiagonal operator $\mathfrak{T}_d$ with a suitable bandwidth. Given that the bandable structure implies a geometric decay of the away-from-diagonal part, it is logical to truncate the sample covariance matrices using the blockwise tridiagonal operator. While the sample covariance matrices tend to underperform in high-dimensional scenarios, the individual blocks generated by the operator $\mathfrak{T}_d$ can preserve a low-dimensional nature, provided the bandwidth is judiciously selected.

It is crucial to recognize in transfer learning that the inclusion of the source sample does not invariably enhance performance. There should be a risk that the source sample may not align with the target sample or could be adversarially generated to compromise the accuracy of statistical inference. An intuitive approach involves selective inclusion based on any measure of similarity between the target and source samples. This strategy has been validated as effective in various statistical challenges, including the contextual multi-armed bandit [4] and the functional mean estimation [6]. Within our framework, the similarity is quantified by the disparity threshold, δ. We can opt for either the target sample covariance matrix $\check{\Sigma}^{(t)}$ or the pooled sample covariance matrix $\check{\Sigma}$ inside the blockwise tridiagonal operator $\mathfrak{T}_d$, depending on the magnitude of δ. The estimator suggested in Theorem 2.1 synthesizes these concepts and achieves the rate Φ, as specified in Equation (1).

THEOREM 2.1 (Block tridiagonal estimator). *Consider blockwise tridiagonal estimator* $\widehat{\Sigma}^{(t)}$ *for the target covariance matrix* $\Sigma^{(t)}$ *as follows:*

$$\widehat{\Sigma}^{(t)} := \begin{cases} \mathfrak{T}_d(\check{\Sigma}^{(t)}; u^*) & \text{if } n_t\delta^2 > (n_t^{1/(2\alpha_t+1)} \wedge d) \vee \log d, \\ \mathfrak{T}_d(\check{\Sigma}; v^*) & \text{if } n_t\delta^2 \le (n_t^{1/(2\alpha_t+1)} \wedge d) \vee \log d, \end{cases}$$

where the optimal bandwidths $u^*, v^* > 0$ are judiciously selected such that:

$$u^* \asymp n_t^{1/(2\alpha_t+1)} \wedge d \qquad \textit{and} \qquad v^* \asymp n^{1/(2\alpha_{\max}+1)} \wedge d.$$

The maximal risk of the blockwise tridiagonal estimator $\widehat{\Sigma}^{(t)}$ now satisfies:

$$\sup_{\in \mathcal{P}_{\alpha_t,\alpha_s,\delta}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \lesssim \Phi_{\alpha_t,\alpha_s,\delta}(n_t, n_s, d),$$

as long as either of the following assumptions is true:

Assumption (a) $\log d \le c_1 n_t$ *for a small enough constant* $c_1 > 0$.
Assumption (b) $\log d \le c_2 n$ *and* $\delta \le c_2$ *for a small enough constant* $c_2 > 0$.

According to the discussion following Theorem 2.2 , which presents a matching lower bound, consistent estimation is feasible only if either Assumption (a*) or Assumption (b*) does hold. Therefore, to ensure reliable estimation, it is essential to include analogous assumptions, such as Assumption (a) or Assumption (b), in Theorem 2.1.

It should be emphasized that the transfer learning framework under the specific conditions, $n_s = 0$ and $\delta = \infty$, reduces to the conventional learning framework, as the source samples are no longer available and the source population conveys no information to the target covariance matrix. In such a scenario, the blockwise tridiagonal estimator introduced in Theorem 2.1 is given by:

$$\widehat{\Sigma}^{(t)} = \mathfrak{T}_d(\check{\Sigma}^{(t)}; u^*) \qquad \text{where } u^* \asymp n_t^{1/(2\alpha_t+1)} \wedge d.$$

Given that the conventional learning problem for covariance matrix estimation has been extensively studied, it presents an intriguing opportunity to compare our estimator with those introduced in preceding research. Our blockwise tridiagonal estimator exhibits a comparable structure with the tapering estimator proposed by Cai, Zhang and Zhou [11]. The principal distinction lies in the thresholding type: our estimator employs a hard-thresholding strategy, in contrast to the soft-thresholding approach utilized by the tapering estimator. Nonetheless, both estimators achieve an identical rate of convergence, represented by $\Phi_{\alpha_t,\alpha_s,\infty}(n_t, 0, d)$, when operating under the optimal bandwidth u^* of Theorem 2.1. Additionally, Cai, Zhang and Zhou [11] demonstrate that $\Phi_{\alpha_t,\alpha_s,\infty}(n_t, 0, d)$ is the minimax rate of convergence within the conventional learning framework.

$$\Phi_{\alpha_t,\alpha_s,\infty}(n_t, 0, d) \asymp \left(n_t^{-2\alpha_t/(2\alpha_t+1)} \wedge \frac{d}{n_t} \right) + \frac{\log d}{n_t}. \tag{4}$$

Bickel and Levina [1] also introduced an analogous estimator of hard-thresholding type, known as the banding estimator. Instead of the blockwise tridiagonal operator $\mathfrak{T}_d$, it utilizes the banding operator, denoted by $\mathfrak{A}_d^{\mathsf{c}}$ and represented as the white area in Figure 2a. The rationale behind their estimator is that the banding operator is complementary to the away-from-diagonal operator $\mathfrak{A}_d$, satisfying the equation $\mathfrak{A}_d(B; u) + \mathfrak{A}_d^{\mathsf{c}}(B; u) = B$ for any square matrix $B \in {}^{d\times d}$ and bandwidth $u > 0$. However, the bandwidth they advocate in the banding estimator is not optimal, resulting in a uniformly slower rate of convergence.

2.2. *Matching lower bound.* The following theorem characterizes a lower bound for the minimax risk in estimating the target covariance matrix under the transfer learning paradigm.

THEOREM 2.2 (Mimimax lower bound). *Given any estimator $\widehat{\Sigma}^{(t)}$ of the target covariance matrix $\Sigma^{(t)}$, we have*

$$\sup_{\in\mathcal{P}_{\alpha_t,\alpha_s,\delta}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \gtrsim \left(n^{-2\alpha_{\max}/(2\alpha_{\max}+1)} \wedge \frac{d}{n}\right) + \left(\frac{\log d}{n} \wedge 1\right) + \left(n_t^{-2\alpha_t/(2\alpha_t+1)} \wedge \frac{d}{n_t} \wedge \delta^2\right) + \left(\frac{\log d}{n_t} \wedge \delta^2 \wedge 1\right).$$

According to Theorem 2.2, a consistent estimation is achievable only if at least one of the following assumptions holds:

Assumption (a*) $\log d \ll n_t$
Assumption (b*) $\log d \ll n$ and $\delta \ll 1$.

When either of these assumptions is fulfilled, the lower bound presented in Theorem 2.2 coincides exactly with the upper bound in Theorem 2.1. The upper and lower bounds together yield the minimax rate of convergence, $\Phi_{\alpha_t,\alpha_s,\delta}(n_t, n_s, d)$, as defined in Equation (1).

An essential inquiry arises concerning the effectiveness of the transfer learning and the source sample. Our objective is to rigorously analyze the conditions under which, and the extent to which, the transfer learning proves beneficial. What makes this analysis particularly compelling is the presence of a phase transition under which the efficacy of transfer learning undergoes a significant shift. The results are concisely encapsulated in Table 1.

TABLE 1
Does transfer learning offer benefits? If so, does increasing the source sample size offer benefits?

<table>
<tr><th colspan="2"></th><th>Minimal disparity model ($\delta \ll \delta_1^*$)</th><th>Moderate disparity model ($\delta \gtrsim \delta_1^*$ and $\delta \ll \delta_2^*$)</th><th>Strong disparity model ($\delta \gtrsim \delta_2^*$)</th></tr>
<tr><td rowspan="2">Limited source sampling regime ($n_s \lesssim n_t$)</td><td>without auto-smoothing benefit</td><td colspan="2">No</td><td rowspan="3">No</td></tr>
<tr><td>with auto-smoothing benefit</td><td colspan="2" rowspan="2">Yes / No</td></tr>
<tr><td>Ample source sampling regime ($n_s \gg n_t$)</td><td>Yes / Yes</td></tr>
</table>

The disparity threshold δ plays a crucial role in the phase transition since it serves as a key measure of similarity between the target and source covariance matrices. Let us define two critical points $\delta_1^* \lesssim \delta_2^*$ by

$$\delta_1^* := \left(n^{-\frac{\alpha_{\max}}{2\alpha_{\max}+1}} \wedge \sqrt{\frac{d}{n}}\right) + \sqrt{\frac{\log d}{n}} \quad \text{and} \quad \delta_2^* := \left(n_t^{-\frac{\alpha_t}{2\alpha_t+1}} \wedge \sqrt{\frac{d}{n_t}}\right) + \sqrt{\frac{\log d}{n_t}}.$$

Depending on the level of the disparity threshold δ, we naturally classify our model into three distinct categories: the minimal, the moderate and the strong disparity models as outlined in Table 1.

Recall from Equation (4) that $\Phi^{\mathrm{LE}} := \Phi_{\alpha_t,\alpha_s,\infty}(n_t, 0, d)$ accords with the minimax rate of convergence in the conventional learning scenario. It is noteworthy that the minimax rate Φ must be no slower than Φ^{LE}, as it is possible to construct an estimator that relies solely on the target sample. Hence, the rate Φ^{LE} serves as a baseline and can be regarded as the least effective rate of convergence within the transfer learning framework.

Our model experiences a notable transition at the critical point $\delta \asymp \delta_2^*$, which dictates the potential benefits of transfer learning compared to conventional learning. Specifically, under the strong disparity model, characterized by $\delta \gtrsim \delta_2^*$, the minimax rate of convergence aligns exactly with Φ^{LE}. Despite the availability of the source sample, the disparity threshold δ is too substantial to facilitate the effective transfer of knowledge. Conversely, in the context of the minimal or moderate disparity model, characterized by $\delta \ll \delta_2^*$, incorporating the source sample presents an opportunity to enhance the accuracy of estimating the target covariance matrix.

Given that the disparity threshold is sufficiently small as in the minimal or moderate disparity models, the size of the source sample becomes a key factor in this enhancement. We explore the necessary and sufficient conditions for the enhancement in two separate scenarios: the ample and the limited source sampling regimes as presented in Table 1. Additionally, a comparative analysis of these regimes reveals another phase transition phenomenon.

Within the ample source sampling regime where $n_s \gg n_t$ does hold, it is readily verifiable that the minimax rate of convergence diminishes more rapidly than the baseline rate Φ^{LE}. In other words, a substantial size of the source sample is conducive to effective transfer learning, aligning with our expectations under the minimal or moderate disparity model. On the contrary, under the limited source sampling regime, defined by $n_s \lesssim n_t$, the effectiveness of the source sample may be compromised. Typically, the minimax rate of convergence aligns with the baseline rate Φ^{LE}. Nevertheless, the transfer learning can still benefit from the scarce source sample if the following extra conditions are satisfied:

- The source covariance matrix exhibits an expedited decay, as indicated by $\alpha_s > \alpha_t$.
- The dimensionality is neither too small nor large, as $n_t^{1/(2\alpha_s+1)} \ll d \ll \exp(n_t^{1/(2\alpha_t+1)})$.

These conditions give rise to the *auto-smoothing* effect, which allows the target covariance matrix to be analyzed as if it possesses the source decay rate α_s rather than the target decay rate α_t. If the source covariance matrix exhibits an expedited decay, the auto-smoothing now facilitates the more accurate estimation for the target covariance matrix. Auto-smoothing is akin to the concept of transferable smoothness described by Cai and Pu [7] in the context of transfer learning for nonparametric regression. The dimensionality condition is equally critical, ensuring that the improved decay rate effectively aids in estimation. Therefore, those extra conditions are both essential and adequate to gain benefit from the auto-smoothing and to ensure effective transfer learning under the limited source sampling regime as outlined in Table 1.

Moving forward, our analysis now centers on regimes and models where transfer learning proves advantageous. Pertaining to the magnitude of this advantage, a final phase transition is observed at the critical point, $\delta \asymp \delta_1^*$. Within the moderate disparity model, satisfying $\delta \gtrsim \delta_1^*$, the minimax rate of convergence is simply established as δ^2. Although the established rate indicates an enhancement over the baseline rate Φ^{LE}, the expansion of the source sample size fails to accelerate the convergence rate as long as other conditions remain the same. On the other hand, within the minimal disparity model, characterized by $\delta \ll \delta_1^*$, an increase in the source sample size yields additional benefits. It can be demonstrated that the minimax rate of convergence precisely matches the most efficient rate Φ^{ME}, defined as follows. By leveraging a model with a negligible disparity threshold, we can extract greater benefits from the larger source sample, leading to enhanced performance of estimation.

$$\Phi^{\mathrm{ME}} := \left(n^{-2\alpha_{\max}/(2\alpha_{\max}+1)} \wedge \frac{d}{n} \right) + \frac{\log d}{n}.$$

Let us further examine the case where the minimax rate coincides with the most effective rate Φ^{ME}. It is immediate that $\Phi^{\mathrm{ME}} = \Phi_{\alpha_{\max},\alpha_{\max},0}(n_t, n_s, d)$ holds, suggesting that the target

and the source samples could be considered as originating from the same population. This representation is underscored by the auto-smoothing effect, which arises from the condition $\alpha_t = \alpha_s = \alpha_{\max}$. As we have discussed before, this effect equalizes the decay rates of the target and source covariance matrices, ensuring both exhibit the fastest decay rate, $\alpha_{\max}$. Not only that, another condition $\delta = 0$ indicates that the source and target covariance matrices are identical, effectively recasting the transfer learning framework into the conventional learning framework with augmented sample size, $n = n_t + n_s$. While the source population may differ from the target population, to estimate the target covariance matrix, the minimax rate of convergence guarantees that they are fundamentally equivalent.

3. Adaptive transfer learning. While we have demonstrated in Section 2 that the blockwise tridiagonal estimator, as described in Theorem 2.1, achieves the minimax convergence rate Φ, its practical use is limited. The implementation of the blockwise tridiagonal estimator necessitates knowledge of the model parameters, including decay rates α_t and α_s as well as the disparity threshold δ, which are typically unknown in practice.

In this section, we aim to introduce a data-driven algorithm that automatically adapts to unknown parameters over a broad range of parameter spaces, particularly $\alpha_t, \alpha_s > 0$ and $\delta \geq 0$. However, it is important to recognize that in adaptive transfer learning, there is no free lunch. The following proposition claims that no estimator can simultaneously attain both adaptivity and optimality within the transfer learning context. This stands in stark contrast to the conventional learning framework, where adaptive and optimal methods have been already suggested (e.g. Cai and Yuan [10]).

PROPOSITION 3.1 (No free lunch in adaptive transfer learning). *Consider an estimator $\widehat{\Sigma}^{(t)}$ whose maximal risk is non-decreasing as a function of the source sample size:*

$$n_s \mapsto \sup_{\in \mathcal{P}_{\alpha_t,\alpha_s,\delta}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \text{ is monotonically decreasing.}$$

Suppose that there exists a generic constant $C_1 > 0$ satisfying:

$$\sup_{\in \mathcal{P}_{\alpha_t,\alpha_s,\delta}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \leq C_1 \Phi_{\alpha_t,\alpha_s,\delta}(n_t, n_s, d), \tag{5}$$

where the model $\mathcal{P}_{\alpha_t,\alpha_s,\delta}$ and regime (n_t, n_s, d) meet the subsequent conditions:

$$\begin{cases} n_s > n_t, \quad 0 < \alpha_s < \alpha_t, \quad \delta \gg n_t^{-\alpha_t/(2\alpha_t+1)}, \\ n_t^{1/(2\alpha_t+1)} \ll d \ll \exp\big(n_t^{1/(2\alpha_s+1)} \wedge (n_t\delta^2)\big). \end{cases} \tag{6}$$

In this case, we can identify a generic constant $c_2 > 0$ such that:

$$\sup_{\in \mathcal{P}_{\alpha_0,\alpha_0,\delta_0}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \geq c_2 n_t^{2\alpha_s\zeta/(2\alpha_s+1)} \Phi_{\alpha_0,\alpha_0,\delta_0}(n_t, n_s^{1+\zeta}, d). \tag{7}$$

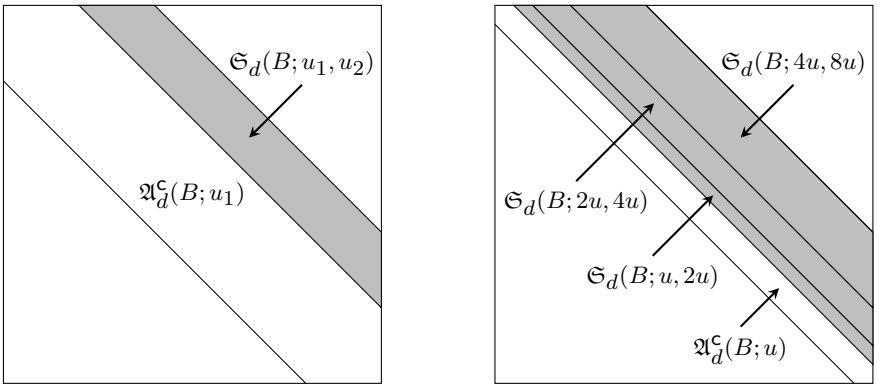
Although the optimal rate is given by Φ as discussed in Section 2, Proposition 3.1 delivers a sobering message: no adaptive estimator cannot attain the optimal rate Φ. More specifically, if any estimator $\widehat{\Sigma}^{(t)}$ achieves the minimax rate under a model $\mathcal{P}_{\alpha_t,\alpha_s,\delta}$ and regime (n_t, n_s, d) satisfying Equation (6), the same estimator $\widehat{\Sigma}^{(t)}$ should be sub-optimal under another model $\mathcal{P}_{\alpha_0,\alpha_0,\delta_0}$ and regime $(n_t, n_s^{1+\zeta}, d)$ for any constant $\zeta > 0$. In summary, Proposition 3.1 highlights the inherent trade-offs between adaptivity and optimality within the transfer learning framework. Any adaptive estimator needs to compromise the optimal rate Φ or pay the *cost of adaptation*.

Having acknowledged this trade-off, our focus naturally shifts to the heart of the matter: how can we minimize the compromise on the optimal rate Φ or the cost of adaptation in the context of adaptive transfer learning? As an initial step, in Section 3.1, we introduce an adaptive estimator and assess its maximal risk in terms of the squared matrix ℓ_2-norm. Subsequently, in Section 3.2, we rigorously demonstrate that our proposed algorithm effectively minimizes the cost of adaptation. By synthesizing insights from both sections, we establish the *adaptive rate of convergence*.

3.1. *Adaptive tridiagonal block-thresholding estimator.* We will establish an adaptive method for estimating bandable covariance matrices within the transfer learning framework. Before diving into the primary algorithm, we define an important operator known as the superbanding operator, denoted by $\mathfrak{S}_d$. Given two bandwidths $u_2 > u_1 > 0$, this operator is formulated as:

$$\mathfrak{S}_d(B; u_1, u_2) := \Big(b_{jk}\big(u_1 < k - j \le u_2\big)\Big)_{j,k\in[d]} \qquad \text{where } B = \big(b_{jk}\big)_{j,k\in[d]} \in {}^{d\times d}.$$

Recall that the banding operator $\mathfrak{A}_d^{\mathsf{c}}$ only maintains some specific regions around the main diagonal, as illustrated in Figure 3a. The superbanding operator, $\mathfrak{S}_d(\,\cdot\,; u_1, u_2)$, is configured to pull out the superdiagonal of bandwidth $(u_2 - u_1)$, positioned just above and to the right of the region pulled by the banding operator $\mathfrak{A}_d^{\mathsf{c}}(\,\cdot\,; u_1)$. This functionality is graphically represented in Figure 3a, where the operator $\mathfrak{S}_d(\,\cdot\,; u_1, u_2)$ preserves the elements within the grey region while nullifying those in the white region. It is worth noting that the superbanding operator $\mathfrak{S}_d(\,\cdot\,; u_1, u_2)$ preserves a specific segment of the region that is also preserved by the away-from-diagonal operator $\mathfrak{A}_d(\,\cdot\,; u_1)$.



(a) The output, $\mathfrak{S}_d(B; u_1, u_2)$, of the superbanding operator

(b) Approximation and partition of the away-from-diagonal region

Fig 3: A key concept for adaptation: the superbanding operator $\mathfrak{S}_d$

Returning to the main problem of adaptivity, the blockwise tridiagonal estimator outlined in Theorem 2.1 encounters two fundamental challenges in practice. The initial one involves selecting the optimal bandwidths, a procedure intricately linked to the decay rates α_t and α_s. This link becomes particularly outstanding when addressing the bias term or the away-from-diagonal matrix $\mathfrak{A}_d(\Sigma^{(t)}; u)$ for a bandwidth $u > 0$. Given that the decay rates are typically unknown in real-world scenarios, a data-driven method is required to control the bias. However, the away-from-diagonal matrix $\mathfrak{A}_d(\Sigma^{(t)}; u)$ intrinsically possesses a high-dimensional

nature, unlike the blockwise tridiagonal matrix $\mathfrak{T}_d(\Sigma^{(t)}; u)$. Directly replacing it with sample covariance matrices thus leads to suboptimal performance.

To address this challenge, let us approximate the away-from-diagonal matrix $\mathfrak{A}_d(\Sigma^{(t)}; u)$ with the superbanding matrix $\mathfrak{S}_d(\Sigma^{(t)}; u, 2^{\ell^*}u)$ for a suitably chosen integer $\ell^* \geq 0$. We then divide it into smaller and more manageable segments by doubling the bandwidth of the superbanding operator. This strategy not only encompasses the approximated area but also reduces the approximation error when substituting with sample covariance matrices. The outlined approach is both visually and mathematically elaborated in Figure 3b and the following equation:

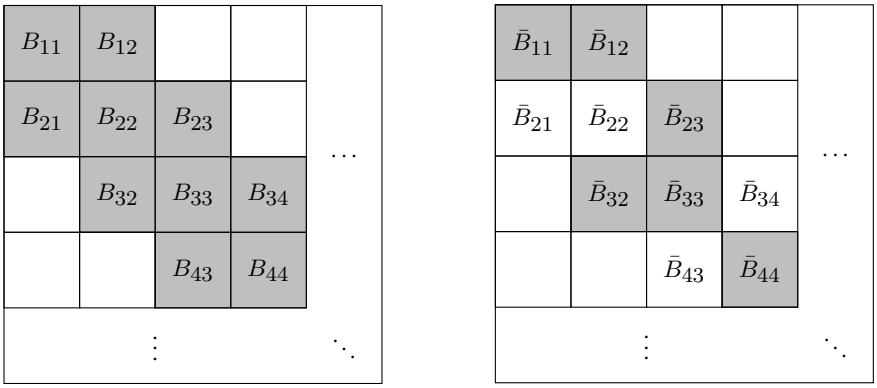
$$\begin{aligned}\mathfrak{A}_d(\Sigma^{(t)}; u) &\approx \mathfrak{S}_d(\Sigma^{(t)}; u, 2^{\ell^*}u) + \mathfrak{S}_d^{\top}(\Sigma^{(t)}; u, 2^{\ell^*}u) \\ &= \sum_{\ell=1}^{\ell^*} \Big(\mathfrak{S}_d(\Sigma^{(t)}; 2^{\ell-1}u, 2^{\ell}u) + \mathfrak{S}_d^{\top}(\Sigma^{(t)}; 2^{\ell-1}u, 2^{\ell}u) \Big).\end{aligned}$$

On the other hand, as illustrated in Theorem 2.1, we have developed two blockwise tridiagonal estimators: one originated from the target sample covariance matrix and the other from the pooled sample covariance matrix. The final estimator requires a selection technique based on the disparity threshold δ, which introduces the second major challenge concerning adaptivity. Since the parameter δ is practically unknown, a data-driven strategy is necessary to manage the disparity between the target and source covariance matrices. Block-thresholding serves as a viable solution to achieve this goal.

We present the tridiagonal block-thresholding operator, denoted by $\overline{\mathfrak{T}}_d$, which is a block-thresholding variant of the blockwise tridiagonal operator $\mathfrak{T}_d$. To formally define this operator, we consider the blockwise tridiagonal matrix $\mathfrak{T}_d(B; u)$ for a matrix $B \in {}^{d\times d}$ and a bandwidth $u > 0$. In accordance with its sub-matrix representation depicted in Figure 4a, we establish the thresholded sub-matrices for a given threshold $R > 0$:

$$\bar{B}_{ij} := B_{ij}\big(\|B_{ij}\|_2 \geq R\big) \qquad \text{for each sub-matrix } B_{ij} \text{ of } \mathfrak{T}_d(B; u).$$

The tridiagonal block-thresholding operator $\overline{\mathfrak{T}}_d(\,\cdot\,; u, R)$ is now depicted in Figure 4b. It mirrors the structure of the blockwise tridiagonal operator $\mathfrak{T}_d$, with the distinction that each sub-matrix B_{ij} of $\mathfrak{T}_d(B; u)$ is substituted by its thresholded counterpart $\bar{B}_{ij}$. The tridiagonal block-thresholding operator reduces certain sub-matrices to zero, which are depicted as white squares in Figure 4b. This contrasts with the representation of the blockwise tridiagonal operator in Figure 4a, where all sub-matrices are gray squares.



(a) Revisited: the output, $\mathfrak{T}_d(B; u)$, of the blockwise tridiagonal operator

(b) The output, $\overline{\mathfrak{T}}_d(B; u, R)$, of the tridiagonal block-thresholding operator

Fig 4: A key concept for adaptation: the tridiagonal block-thresholding operator $\overline{\mathfrak{T}}_d$

Algorithm 1 Adaptive blockwise tridiagonal estimator for conventional learning

Require: Target sample $\{X_1^{(t)}, \ldots, X_{n_t}^{(t)}\}$.

1: Compute the target sample covariance matrix $\check{\Sigma}^{(t)}$.

2: For any large enough constant $\lambda > 0$, solve the following optimization problem:

$$\widehat{u} := \underset{u_{\min} \le u \le u_{\max}}{\operatorname{argmin}} \left(\sum_{\ell=1}^{\ell_u} \|\mathfrak{S}_d(\check{\Sigma}^{(t)}; 2^{\ell-1}u, 2^{\ell}u)\|_2 + \lambda \rho^2 \sqrt{\frac{u + \log d}{n_t}} \right),$$

where we define $u_{\min} = \log d$, $u_{\max} = \big((n_t \log^{-1} n_t) \wedge d\big) \vee \log d$ and

$$\ell_u := \max\{\ell \in {}^+ : 2^{\ell-1}u \le n_t \log^{-1} n_t\}.$$

3: Output the final estimator:

$$\widehat{\Sigma}^{(t)} := \mathfrak{T}(\check{\Sigma}^{(t)}; \widehat{u}).$$

Algorithm 2 Adaptive tridiagonal block-thresholding estimator for transfer learning

Require: Target sample $\{X_1^{(t)}, \ldots, X_{n_t}^{(t)}\}$ and source sample $\{X_1^{(s)}, \ldots, X_{n_s}^{(s)}\}$.

1: Compute the target and source sample covariance matrices, $\check{\Sigma}^{(t)}$ and $\check{\Sigma}^{(s)}$.

2: Define the pooled sample covariance matrix $\check{\Sigma}$ and sample disparity matrix $\check{\Delta}$ by:

$$\check{\Sigma} := \frac{n_t}{n}\check{\Sigma}^{(t)} + \frac{n_s}{n}\check{\Sigma}^{(s)} \qquad \text{and} \qquad \check{\Delta} := \check{\Sigma}^{(t)} - \check{\Sigma}.$$

3: For any large enough constant $\tau > 0$, solve the following optimization problem:

$$\widehat{v} := \underset{v_{\min} \le v \le v_{\max}}{\operatorname{argmin}} \left(\sum_{\ell=1}^{\ell_v} \|\mathfrak{S}_d(\check{\Sigma}; 2^{\ell-1}v, 2^{\ell}v)\|_2 + \tau \rho^2 \sqrt{\frac{v + \log d}{n}} \right),$$

where it is defined that $v_{\min} = \log d$, $v_{\max} = (n \log^{-1} n \wedge d) \vee \log d$ and

$$\ell_v := \max\{\ell \in {}^+ : 2^{\ell-1}u \le n \log^{-1} n\}.$$

4: For any large enough constants ξ_1, solve another optimization problem:

$$\widehat{w} := \underset{w_{\min} \le w \le w_{\max}}{\operatorname{argmin}} \left(\sum_{\ell=1}^{\ell_w} \|\mathfrak{S}_d(\check{\Delta}; 2^{m-1}w, 2^{m}w)\|_2 + \xi_1 \rho^2 \sqrt{\frac{w + \log nd}{n_t}} \right),$$

where it is defined that $w_{\min} = \log d$, $w_{\max} = (n_t \log^{-1} n_t \wedge d) \vee \log d$, and

$$\ell_w := \max\{m \in {}^+ : 2^{m-1}w \le n_t \log^{-1} n_t\}.$$

5: Output the final estimator:

$$\widehat{\Sigma}^{(t)} := \begin{cases} \mathfrak{T}_d(\check{\Sigma}; \widehat{v}) + \overline{\mathfrak{T}}_d(\check{\Delta}; \widehat{w}, R_{\widehat{w}}) & \text{if } \log d \le n_t \log^{-1} n_t, \\ \mathfrak{T}_d(\check{\Sigma}; \widehat{v}) & \text{if } \log d > n_t \log^{-1} n_t. \end{cases}$$

where the threshold $R_w > 0$ corresponding to the bandwidth $w > 0$ is defined as:

$$R_w := \xi_2 \rho^2 \sqrt{\frac{w + \log nd}{n_t}} \qquad \text{for any large enough constant } \xi_2 > 0.$$

Building on the adaptation principles previously discussed, we shall introduce two distinct methodologies for the adaptive estimation of the target covariance matrix: Algorithm 1 and Algorithm 2. The necessity for two separate algorithms stems from the fact that integration of the source sample does not always enhance performance in transfer learning. Analogous to Theorem 2.1, Algorithm 1 exclusively leverages the target sample, while Algorithm 2 utilizes both the target and source samples.

Central to both algorithms is an optimization step for identifying the adaptive bandwidth. In Algorithm 1, we establish the adaptive bandwidth $\widehat{u}$ for the target sample covariance matrix. In Algorithm 2, we identify two adaptive bandwidths, $\widehat{v}$ and $\widehat{w}$, for the pooled sample covariance matrix and the sample disparity matrix, respectively. The objective function for each optimization consists of two components: the bias term, approximated and segmented by the superbanding matrices, and the penalty term, which increases as the bandwidth expands. Notably, these optimization problems are tractable, as the effective search space for each objective function is the set of discrete integer bandwidths up to n_t or n.

On the other hand, the tridiagonal block-thresholding operator $\overline{\mathfrak{T}}_d$ is vital in Algorithm 2, as opposed to Algorithm 1. Such a distinction arises because Algorithm 2 aims to integrate the source sample, which requires addressing the potential bias in the process. Consequently, we add an extra step to apply the tridiagonal block-thresholding operator to the sample disparity matrix. It automatically adjusts for bias only when the disparity exceeds a predefined threshold.

Ultimately, given that the integration of the source sample does not consistently improve performance in transfer learning, we will carefully combine Algorithm 1 and Algorithm 2. As outlined in Theorem 3.2, this combination must be adaptive as well, independent of unknown model parameters. The final estimator is called the adaptive tridiagonal block-thresholding estimator.

THEOREM 3.2 (Adaptive tridiagonal block-thresholding estimator). *Consider the output of Algorithm 1 and Algorithm 2, denoted by* $\widehat{\Sigma}^{(t)}_{\mathrm{ADC}}$ *and* $\widehat{\Sigma}^{(t)}_{\mathrm{ADT}}$*, respectively. The following estimator is then introduced:*

$$\widehat{\Sigma}^{(t)} := \begin{cases} \widehat{\Sigma}^{(t)}_{\mathrm{ADC}} & \text{if } n_s \le n_t \text{ and } \log d \le n_t \log^{-1} n_t, \\ \widehat{\Sigma}^{(t)}_{\mathrm{ADT}} & \text{otherwise.} \end{cases}$$

The maximal risk of this estimator $\widehat{\Sigma}^{(t)}$ *is bounded from above as follows:*

$$\sup_{\in \mathcal{P}_{\alpha_t,\alpha_s,\delta}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \lesssim \left(n^{-2\alpha_{\max}/(2\alpha_{\max}+1)} \wedge \frac{d}{n}\right) + \frac{\log d}{n} + \left(\varphi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t, n_s, d) \wedge \frac{d}{n_t} \wedge \delta^2\right) + \left(\frac{\log n \log d}{n_t} \wedge \delta^2\right),$$

provided either of the following assumptions is true:

Assumption (c) $\log d \le c_1 n$ *and* $\log n \le c_1 n_t$ *for a small enough constant* $c_1 > 0$.
Assumption (d) $\delta \le C_2$ *for a large enough constant* $C_2 > 0$.

The adaptive tridiagonal block-thresholding estimator, as we described in Theorem 3.2, is entirely data-driven. This includes Algorithm 1 and Algorithm 2, along with the criterion for their selection. As evidenced by Theorem 3.2, this estimator achieves the adaptive rate Φ^{AD}, up to a factor of $\log n$, as explicated in Equations (2) and (3). Our next step involves a thorough analysis of the adaptive rate Φ^{AD} and its comparison with the corresponding minimax rate Φ.

When both conditions $n_s = 0$ and $\delta = \infty$ are met, the transfer learning paradigm simplifies to the conventional learning framework, as we have discussed following Theorem 2.1. In this situation, the adaptive tridiagonal block-thresholding estimator in Theorem 3.2 corresponds to the result of Algorithm 1. Moreover, as indicated by Theorem 3.2, this estimator achieves the minimax rate of convergence under the conventional learning framework:

$$\Phi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\infty}(n_t, 0, d) \asymp \Phi_{\alpha_t,\alpha_s,\infty}(n_t, 0, d).$$

Given the comprehensive analysis of the adaptive covariance matrix estimation in the conventional learning setting, it offers a valuable chance to benchmark our estimator against those established in prior studies. Cai and Yuan [10] have conducted an in-depth study on the adaptive estimation of bandable covariance matrices within the same context, culminating in the development of a block-thresholding estimator. It has been demonstrated to be data-driven and achieve the optimal rate over a broad collection of bandable covariance matrices in the conventional setting. However, its accessibility is hindered by the complex nature of the sub-matrix constructions. On the contrary, the proposed Algorithm 1 produces the blockwise tridiagonal matrix with a more straightforward interpretation, even though the optimal bandwidth is selected via a sophisticated optimization process. Therefore, the adaptive tridiagonal block-thresholding estimator represents a logical and more user-friendly advancement in the realm of transfer learning.

Fig 5: Comparison between the adaptive rate Φ^{AD} and the minimax rate Φ

It is significant to highlight that the adaptive rate Φ^{AD} is slower than the minimax rate Φ if and only if all the conditions specified in Equation (6) hold. When our model and regime satisfy these conditions, both rates can be expressed in the following manner:

$$\Phi_{\alpha_t,\alpha_s,\delta}(n_t, n_s, d) \asymp n_t^{-2\alpha_t/(2\alpha_t+1)} + \left(\frac{\log d}{n_t} \wedge \delta^2\right),$$

$$\Phi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t, n_s, d) \asymp \left(n_t^{-2\alpha_s/(2\alpha_s+1)} \wedge \frac{d}{n_t} \wedge \delta^2\right).$$

In the context at hand, Figure 5 illustrates the discrepancy between the adaptive rate Φ^{AD} and the minimax rate Φ, with the former depicted in dark green curve and the latter in light green curve. The graphical representation considers the dimension d as a variable while the sizes of the target and the source sample, n_t and n_s, remain constant. As we have discussed in

Proposition 3.1, this discrepancy represents an essential compromise to tackle the challenges of adaptivity within our transfer learning framework. The shaded green region in Figure 5 emphasizes that the adaptive tridiagonal block-thresholding estimator falls short of achieving the optimal rate of convergence. This region therefore serves as a visual representation of the cost of adaptation, an indispensable investment to ensure that our estimation remains flexible across a diverse spectrum of models.

We turn our attention to the underlying reasons behind the observed discrepancy between the adaptive rate Φ^{AD} and the minimax rate Φ. This discrepancy hinges on the term φ^{AD}, as defined in Equation (3). Upon closer examination, we find that φ^{AD} gives a sub-optimal rate when there are more source observations than target observations:

$$\varphi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d) = n_t^{-2\alpha_{\min}/(2\alpha_{\min}+1)}, \qquad \text{if } n_s > n_t.$$

In this scenario, integrating the source sample becomes essential by adaptively handling the disparity matrix $\Delta^{(s)}$ to achieve the optimal rate of convergence. Nonetheless, this dynamic process introduces an intriguing phenomenon known as *auto-unsmoothing*, as reflected in the behavior of φ^{AD}. Remarkably, the target covariance matrix behaves as if it possesses the decay rate of $\alpha_{\min}$, even if its true decay rate α_t may be more rapid than $\alpha_{\min}$. This stands in stark contrast to the auto-smoothing effect delineated in Section 2.2.

The condition $\alpha_s < \alpha_t$, as expressed in Equation (6), now becomes pivotal. It serves as the gateway to incurring the cost of adaptation through auto-unsmoothing. The remaining conditions in Equation (6), regarding the disparity threshold δ and dimension d, reserve the impact of auto-unsmoothing and induce the discrepancy between two rates, Φ^{AD} and Φ.

3.2. *Adaptive lower bound.* We are well aware, thanks to Proposition 3.1, that no adaptive estimator can achieve optimality; it is inevitable to compromise the optimal rate Φ or pay the cost of adaptation. It is now reasonable to ask whether this cost is effectively minimized by the adaptive tridiagonal block-thresholding estimator. Astonishingly, while there may exist estimators that outperform the proposed approach, it comes at a considerable expense. The subsequent theorem provides a mathematical rationale for this assertion.

THEOREM 3.3 (Adaptive lower bound). *Consider an estimator $\widehat{\Sigma}^{(t)}$ whose maximal risk is non-decreasing as a function of the source sample size:*

$$n_s \mapsto \sup_{\in\mathcal{P}_{\alpha_t,\alpha_s,\delta}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \text{ is monotonically decreasing.} \tag{8}$$

For this estimator, we can assert the existence of two sufficiently small constants $c_1, c_2 > 0$ such that the subsequent statement is true: if the estimator $\widehat{\Sigma}^{(t)}$ attains the following maximal risk under some model $\mathcal{P}_{\alpha_t,\alpha_s,\delta}$ and some regime (n_t, n_s, d):

$$\sup_{\in\mathcal{P}_{\alpha_t,\alpha_s,\delta}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \le c_1 \Phi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d), \tag{9}$$

then we can identify another model $\mathcal{P}_{\alpha_0,\alpha_0,\delta_0}$ such that

$$\sup_{\in\mathcal{P}_{\alpha_0,\alpha_0,\delta_0}} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \ge c_2 n_t^{2\alpha_s\zeta/(2\alpha_s+1)} \Phi^{\mathrm{AD}}_{\alpha_0,\alpha_0,\delta_0}(n_t, n_s^{1+\zeta}, d), \tag{10}$$

under the regime $(n_t, n_s^{1+\zeta}, d)$ for any constant $\zeta > 0$.

In light of Theorem 3.3, any sensible estimator for the target covariance matrix is expected to meet the monotonicity assumption as outlined in Equation (8). According to the premise in Equation (9), within certain models and regimes, this estimator could realize a more rapid rate

of convergence compared to the adaptive rate Φ^{AD} at a minimum by the order of a constant. However, as stipulated by Equation (10), within an alternate model and regime, the given estimator must yield a slower rate of convergence relative to the adaptive rate Φ^{AD} at least by some polynomial order of n_t. In summary, any reasonable attempt to refine the adaptive rate Φ^{AD} comes with considerable trade-offs.

By integrating the insights from Theorem 3.3 with those from Theorem 3.2, the rate Φ^{AD} can be rightfully termed as the adaptive rate of convergence over the class of entire models, $\{\mathcal{P}_{\alpha_t,\alpha_s,\delta} : \alpha_t, \alpha_s > 0,\, \delta \geq 0\}$, in accordance with the definition provided by Tsybakov [41].

(i) An estimator $\widehat{\Sigma}^{(t)}$ attains the convergence rate Φ^{AD}. In particular, Theorem 3.2 introduces the adaptive tridiagonal block-thresholding estimator, which achieves the rate Φ^{AD} up to the factor of $\log n$.

$$\sup_{\substack{\alpha_t,\alpha_s>0\\ \delta\geq 0}} \sup_{\in \mathcal{P}_{\alpha_t,\alpha_s,\delta}} \left(\Phi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d)\right)^{-1} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \leq O(\log n).$$

(ii) We consider an alternative rate of convergence, denoted as $\widetilde{\Phi}^{\mathrm{AD}}$, such that the mapping $n_s \mapsto \widetilde{\Phi}^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d)$ is monotonically decreasing and the following is satisfied. These assumptions suggest the potential to devise a novel methodology for estimating the target covariance matrix that achieves the given rate $\widetilde{\Phi}^{\mathrm{AD}}$ across the entire spectrum of models.

$$\sup_{\substack{\alpha_t,\alpha_s>0\\ \delta\geq 0}} \inf_{\widehat{\Sigma}^{(t)}} \sup_{\in \mathcal{P}_{\alpha_t,\alpha_s,\delta}} \left(\widetilde{\Phi}^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d)\right)^{-1} \|\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\|_2^2 \leq O(1).$$

Suppose that within a certain model $\mathcal{P}_{\alpha_t,\alpha_s,\delta}$, this alternative rate $\widetilde{\Phi}^{\mathrm{AD}}$ is more rapid rate than the adaptive rate Φ^{AD}. From a mathematical standpoint,

$$\frac{\widetilde{\Phi}^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d)}{\Phi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s,d)} \to 0, \qquad \text{as } n_t, n_s \to \infty. \tag{11}$$

In this context, Theorem 3.3 guarantees the identification of an alternative model $\mathcal{P}_{\alpha_0,\alpha_0,\delta_0}$ where the alternative rate $\widetilde{\Phi}^{\mathrm{AD}}$ is slower than the adaptive rate Φ^{AD}. In particular, any attempt to refine the adaptive rate Φ^{AD} for some model $\mathcal{P}_{\alpha_t,\alpha_s,\delta}$ comes with considerable trade-offs. This includes a disproportionately larger loss in the model $\mathcal{P}_{\alpha_0,\alpha_0,\delta_0}$ compared to the gains in the model $\mathcal{P}_{\alpha_t,\alpha_s,\delta}$ under the identical regime:

$$\frac{\widetilde{\Phi}^{\mathrm{AD}}_{\alpha_0,\alpha_0,\delta_0}(n_t,n_s^{1+\zeta},d)}{\Phi^{\mathrm{AD}}_{\alpha_0,\alpha_0,\delta_0}(n_t,n_s^{1+\zeta},d)} \cdot \frac{\widetilde{\Phi}^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s^{1+\zeta},d)}{\Phi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t,n_s^{1+\zeta},d)} \to \infty, \qquad \text{as } n_t, n_s \to \infty, \tag{12}$$

for any constant $\zeta > \alpha_s^{-1}(2\alpha_t+1)^{-1}(\alpha_t - \alpha_s)$. In fact, we have $\zeta > 0$ because $\alpha_t > \alpha_s$ is a necessary condition for the model $\mathcal{P}_{\alpha_t,\alpha_s,\delta}$ to fulfill the assumption in Equation (11).

The concept of adaptive rate of convergence introduces a distinct form of optimality among adaptive estimators. We can now assert that the cost of adaptation associated with the adaptive tridiagonal block-thresholding estimator, as introduced in Theorem 3.2, is inherently minimal. Any effort to reduce this cost within a particular model inevitably results in a significantly increased cost within another model. In conclusion, the suggested estimator stands out as a superior estimation method within our framework, supported by the adaptive rate of convergence, Φ^{AD}.

On the other hand, exploring scenarios where adaptation incurs no cost presents an intriguing avenue of research. As previously established, the adaptive rate Φ^{AD} coincides with

the minimax rate Φ if and only if at least one condition specified in Equation (6) is violated. Consequently, we can investigate several sufficient conditions and examine their practical implications. This study not only enhances our theoretical understanding but also provides valuable insights into the feasibility of applying our estimator in various practical contexts.

(Scenario I: $n_s \leq n_t$)
This regime would be perceived as an extension of the conventional learning framework which parallels the transfer learning framework under the supplementary conditions of $n_s = 0$ and $\delta = \infty$. Analogous to the conventional learning framework, we achieve the adaptivity without incurring additional costs, as the integration of the source sample is unnecessary. Nevertheless, the condition $n_s \leq n_t$ restricts the core potential of transfer learning—a framework that leverages an augmented sample size to enhance estimation accuracy.

(Scenario II: $\alpha_t \leq \alpha_s$)
Within this model, the adaptive incorporation of the source sample remains unaffected by auto-unsmoothing. In particular, due to the condition $\alpha_{\min} = \alpha_t$, we can treat the target covariance matrix as possessing its true decay rate, α_t. As a consequence, the term φ^{AD} in the adaptive rate consistently yields the optimal rate, allowing us to benefit from cost-free adaptivity. This model is particularly pragmatic when the source covariance matrix demonstrates faster decay as one moves away from the diagonal.

$$\varphi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t, n_s, d) = n_t^{-2\alpha_t/(2\alpha_t+1)}, \qquad \text{under Scenario II.}$$

(Scenario III: $\delta \leq C_\delta n_t^{-\alpha_t/(2\alpha_t+1)}$ for any constant $C_\delta > 0$)
This model is applicable when one reasonably assumes that the target covariance matrix aligns closely with the source covariance matrix. In this context, auto-unsmoothing does not adversely impact adaptive estimation. Exploiting the proximity between the target and source covariance matrices, the target covariance matrix is invariably treated as possessing its true decay rate, α_t. As a result, adaptivity can be achieved for free.

$$\varphi^{\mathrm{AD}}_{\alpha_t,\alpha_s,\delta}(n_t, n_s, d) \wedge \delta^2 \asymp n_t^{-2\alpha_t/(2\alpha_t+1)} \wedge \delta^2, \qquad \text{under Scenario III.}$$

(Scenario IV: $d \leq C_d n_t^{1/(2\alpha_t+1)}$ or $\log d \geq C_d n_t^{1/(2\alpha_s+1)}$ for any constant $C_d > 0$)
In this regime, the dimension d is either too small or too large relative to the target sample size. Analogous to other scenarios, there is no cost of adaptation here, as the decay rate does not influence the optimal rate of convergence. Importantly, this sufficient condition is independent of the source sample size. The cost of adaptation arises primarily from the auto-unsmoothing effect, which is closely tied to the target covariance matrix.

4. Numerical experiments. In this section, we evaluate the numerical performance and practical applicability of the data-driven adaptive estimator in Theorem 3.2, which is formulated based on Algorithm 1 and Algorithm 2. Notably, this estimator is characterized by its computational efficiency and straightforward implementation. We substantiate these characteristics through an extensive simulation study.

We consider the target and source covariance matrices as follows:

$$\Sigma^{(t)} := \left(\sigma^{(t)}_{jk}\right)_{j,k\in[d]} \qquad \text{where } \sigma^{(t)}_{jk} := \begin{cases} 1 & \text{if } j = k, \\ C_t|j-k|^{-3}\xi^{(t)}_{jk} & \text{if } j \neq k, \end{cases}$$

$$\Sigma^{(s)} := \left(\sigma^{(s)}_{jk}\right)_{j,k\in[d]} \qquad \text{where } \sigma^{(s)}_{jk} := \begin{cases} 1 & \text{if } j = k, \\ \sigma^{(t)}_{jk} + C_s|j-k|^{-2}\xi^{(s)}_{jk} & \text{if } j \neq k. \end{cases}$$

In this context, $\xi_{jk}^{(t)}$ and $\xi_{jk}^{(s)}$ $(1 \le j < k \le d)$ represent independent Rademacher variables with $\xi_{jk}^{(t)} = \xi_{kj}^{(t)}$ and $\xi_{jk}^{(s)} = \xi_{kj}^{(s)}$ $(j, k \in [d])$. The constant C_t is set at 0.1, while the other constant $C_s > 0$ is adjusted in accordance with the disparity threshold $\delta := \|\Sigma^{(t)} - \Sigma^{(s)}\|_1$. Notably, the target covariance matrix $\Sigma^{(t)}$ exhibits a target decay rate of $\alpha_t = 2$, whereas the source covariance matrix $\Sigma^{(s)}$ presents a source decay rate of $\alpha_s = 1$. This configuration merits careful attention as the condition $\alpha_s < \alpha_t$ introduces additional challenges, including the cost of adaptation, in the estimation of the target covariance matrix.

The target sample size n_t and the dimensionality d remain fixed at $(n_t, d) = (50, 50)$, reflecting their non-critical role in the effectiveness of transfer learning. Our investigation instead includes various combinations of the source sample size n_s and the disparity threshold δ. The source sample size is assessed at $n_s \in \{50, 100, 200, 400, 800\}$, and the disparity threshold is evaluated at $\delta \in \{0, 0.1, 0.4, 0.7, 1, 1.3\}$. Importantly, we introduce the scenario $(n_s, \delta) = (0, \infty)$ for comparative analysis. This scenario represents the conventional learning paradigm and thus, establishes the baseline performance for our simulation study.

For each specified configuration, we repeat 1,000 simulations to generate target and source samples from mean-zero Gaussian distributions. Upon each simulated dataset, we employ the adaptive tridiagonal block-thresholding estimator, as described in Theorem 3.2, and calculate its squared matrix ℓ_2-loss. The comprehensive results are visually represented in Figure 6, which unveils several intriguing patterns and corroborates our theoretical findings.

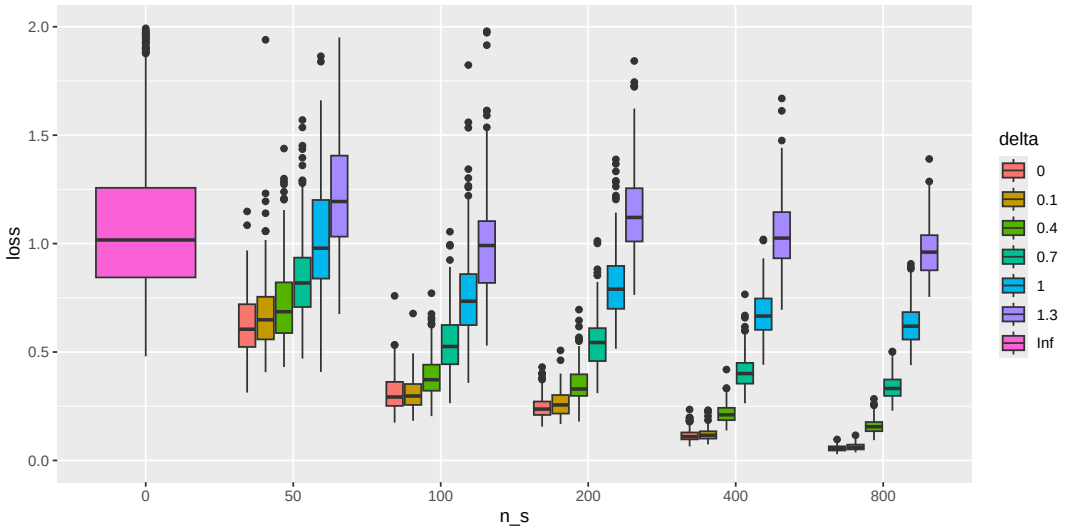


Fig 6: The results of numerical experiments under $\alpha_t = 2$ and $\alpha_s = 1$

First of all, it is observed that when $\delta = 1.3$, transfer learning fails to demonstrate any enhancement over the baseline scenario, where $(n_s, \delta) = (0, \infty)$. This lack of improvement persists despite an increase in the source sample size n_s. Due to the substantial discrepancy between the target and source covariance matrices, the source sample fails to provide any pertinent information for the estimation of the target covariance matrix. This phenomenon can be interpreted within the strong disparity model, which inherently limits the utility of transfer learning as detailed in Table 1.

Our theoretical framework suggests that adaptivity across the full spectrum of models requires a compromise on the optimal rate of convergence within certain models. In particular, the model characterized by $\delta = 1.3$ is quite close to the compromised models, as detailed in

Equation (6). Nevertheless, as depicted in Figure 6, the performance of the adaptive tridiagonal block-thresholding estimator under $\delta = 1.3$ does not markedly deteriorate compared to the baseline scenario of $(n_s, \delta) = (0, \infty)$. This observation implies that, in practical terms, the cost associated with adaptation may not be as substantial as theorized.

Conversely, when δ is set below 1.3, Figure 6 exhibits an evident improvement relative to the baseline scenario, where $(n_s, \delta) = (0, \infty)$. This improvement amplifies as the source sample size n_s expands or as the disparity threshold δ diminishes. Remarkably, the model's performance at $\delta = 0.1$ is nearly indistinguishable from that at $\delta = 0$, suggesting that the source sample generated with $\delta = 0.1$ is as potent as one with $\delta = 0$. These empirical findings are consistent with our theoretical assertions and validate in practice the effectiveness of the adaptive tridiagonal block-thresholding estimator.

5. Discussions. This paper explores minimax and adaptive estimation of high-dimensional bandable covariance matrices within the transfer learning framework. We establish the minimax rate of convergence under the squared spectral norm loss and introduce a blockwise tridiagonal estimator that achieves the optimal rate. The minimax rate reveals intriguing phase transition phenomena, highlighting the effectiveness of transfer learning, and the utilization of source samples. Conversely, the adaptive rate of convergence shows that, unlike in conventional learning where the minimax optimal rate can be adaptively achieved, the cost of adaptation in transfer learning can be substantial in certain cases. Furthermore, we develop a novel data-driven algorithm that automatically adapts to unknown parameters.

Our study focuses on estimation under the squared spectral norm loss. In conventional settings, bandable covariance matrix estimation has been investigated under other error measures, including the matrix ℓ_1 norm, Frobenius norm, and more general Bregman divergences. It would be interesting to explore the estimation of bandable covariance matrices under these error measures within the transfer learning framework.

As mentioned in the introduction, several classes of structured covariance matrices, including sparse and spiked covariance matrices, have been studied in the conventional high-dimensional settings. Extending transfer learning to the estimation of these types of covariance matrices is a promising area for future research. Another interesting direction is the estimation of the bandable precision matrix, the inverse of the covariance matrix, within the transfer learning framework.

These topics are left for future investigation. We anticipate that the framework, insights, and theoretical results presented in this paper will serve as valuable resources for subsequent studies in this domain.

Funding. The research was supported in part by NIH grants R01-GM123056 and R01-GM129781.

SUPPLEMENTARY MATERIAL

Supplement to "Transfer Learning for Covariance Matrix Estimation: Optimality and Adaptivity"
(doi: 0000000; .pdf). This supplementary material contains the complete proofs of the main theorems and technical results presented in the paper titled "Transfer Learning for Covariance Matrix Estimation: Optimality and Adaptivity." The structure of the material is as follows. Supplementary Material A begins by introducing some fundamental notations and then handles the proof of main theorems: Theorem 2.1–2.2, Proposition 3.1 and Theorem 3.2–3.3. Next, Supplementary Material B covers the proof of technical results concerning the optimality arguments: Proposition A.1–A.2, Lemma A.4 and Lemma B.1–B.8. Lastly, Supplementary Material C includes the proofs of technical results relevant to adaptivity arguments such as Proposition A.5–A.8, Proposition C.1–C.5 and Lemma C.6–C.8.

REFERENCES

[1] BICKEL, P. J. and LEVINA, E. (2008). Regularized Estimation of Large Covariance Matrices. *The Annals of Statistics* **36** 199–227. https://doi.org/10.1214/009053607000000758

[2] BIEN, J., BUNEA, F. and XIAO, L. (2016). Convex Banding of the Covariance Matrix. *Journal of the American Statistical Association* **111** 834–845. https://doi.org/10.1080/01621459.2015.1058265

[3] CAI, T. T. (2017). Global Testing and Large-Scale Multiple Testing for High-Dimensional Covariance Structures. *Annual Review of Statistics and Its Application* **4** 423–446. https://doi.org/10.1146/annurev-statistics-060116-053754

[4] CAI, C., CAI, T. T. and LI, H. (2024). Transfer Learning for Contextual Multi-Armed Bandits. *The Annals of Statistics* **52** 207–232. https://doi.org/10.1214/23-AOS2341

[5] CAI, T. T. and KIM, D. (2024). Supplement to "Transfer Learning for Covariance Matrix Estimation: Optimality and Adaptivity".

[6] CAI, T. T., KIM, D. and PU, H. (2024). Transfer Learning for Functional Mean Estimation: Phase Transition and Adaptive Algorithms. *The Annals of Statistics* **52** 654–678. https://doi.org/10.1214/24-AOS2362

[7] CAI, T. T. and PU, H. (2022). Transfer Learning for Nonparametric Regression: Non-Asymptotic Minimax Analysis and Adaptive Procedure. *Technical report*.

[8] CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating Structured High-Dimensional Covariance and Precision Matrices: Optimal Rates and Adaptive Estimation. *Electronic Journal of Statistics* **10** 1–59. https://doi.org/10.1214/15-EJS1081

[9] CAI, T. T. and WEI, H. (2021). Transfer Learning for Nonparametric Classification: Minimax Rate and Adaptive Classifier. *Annals of Statistics* **49** 100–128. https://doi.org/10.1214/20-AOS1949

[10] CAI, T. T. and YUAN, M. (2012). Adaptive Covariance Matrix Estimation through Block Thresholding. *The Annals of Statistics* **40** 2014–2042. https://doi.org/10.1214/12-AOS999

[11] CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics* **38** 2118–2144. https://doi.org/10.1214/09-AOS752

[12] DEKEL, O., KESHET, J. and SINGER, Y. (2004). An Online Algorithm for Hierarchical Phoneme Classification. In *Proceedings of the First International Conference on Machine Learning for Multimodal Interaction*. *MLMI'04* 146–158. Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30568-2_13

[13] GAROFOLO, J. S., LAMEL, L. F., FISHER, W. M., FISCUS, J. G., PALLETT, D. S. and DAHLGREN, N. L. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus. https://doi.org/10.35111/17gk-bn40

[14] HAMOONI, H. and MUEEN, A. (2014). Dual-Domain Hierarchical Classification of Phonetic Time Series. In *2014 IEEE International Conference on Data Mining* 160–169. https://doi.org/10.1109/ICDM.2014.92

[15] HAN, W., PANG, B. and WU, Y. (2021). Robust Transfer Learning with Pretrained Language Models through Adapters. https://doi.org/10.48550/arXiv.2108.02340

[16] HAYDEN, R. E. (1950). The Relative Frequency of Phonemes in General-American English. *WORD* **6** 217–223. https://doi.org/10.1080/00437956.1950.11659381

[17] HE, J. and CHEN, S. X. (2016). Testing Super-Diagonal Structure in High Dimensional Covariance Matrices. *Journal of Econometrics* **194** 283–297. https://doi.org/10.1016/j.jeconom.2016.05.007

[18] HU, G., ZHANG, Y. and YANG, Q. (2019). Transfer Meets Hybrid: A Synthetic Approach for Cross-Domain Collaborative Filtering with Text. In *The World Wide Web Conference*. *WWW '19* 2822–2829. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3308558.3313543

[19] KPOTUFE, S. and MARTINET, G. (2021). Marginal Singularity and the Benefits of Labels in Covariate-Shift. *The Annals of Statistics* **49** 3299–3323. https://doi.org/10.1214/21-AOS2084

[20] LEE, K. F. and HON, H. W. (1989). Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37** 1641–1648. https://doi.org/10.1109/29.46546

[21] LEE, K., LEE, K. and LEE, J. (2022). Estimation of Conditional Mean Operator under the Bandable Covariance Structure. *Electronic Journal of Statistics* **16** 1253–1302. https://doi.org/10.1214/22-EJS1981

[22] LI, S., CAI, T. T. and LI, H. (2022). Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control. *Journal of the American Statistical Association* **0** 1–13. https://doi.org/10.1080/01621459.2022.2044333

[23] LI, Y., DING, A. A. and DY, J. (2017). Rate Optimal Estimation for High Dimensional Spatial Covariance Matrices. In *Proceedings of the Ninth Asian Conference on Machine Learning* **77** 208–223. PMLR.

[24] LI, F., XIA, Y., WANG, F., ZHANG, D., LI, X. and HE, F. (2020). Transfer Learning Algorithm of P300-EEG Signal Based on XDAWN Spatial Filter and Riemannian Geometry Classifier. *Applied Sciences* **10** 1804. https://doi.org/10.3390/app10051804

[25] LI, S., ZHANG, L., CAI, T. T. and LI, H. (2023). Estimation and Inference for High-Dimensional Generalized Linear Models with Knowledge Transfer. *Journal of the American Statistical Association* **0** 1–12. https://doi.org/10.1080/01621459.2023.2184373

[26] MA, C., PATHAK, R. and WAINWRIGHT, M. J. (2023). Optimally Tackling Covariate Shift in RKHS-based Nonparametric Regression. *The Annals of Statistics* **51** 738–761. https://doi.org/10.1214/23-AOS2268

[27] MAQSOOD, M., NAZIR, F., KHAN, U., AADIL, F., JAMAL, H., MEHMOOD, I. and SONG, O.-Y. (2019). Transfer Learning Assisted Classification and Detection of Alzheimer's Disease Stages Using 3D MRI Scans. *Sensors* **19** 2645. https://doi.org/10.3390/s19112645

[28] NGUYEN, C. H., KARAVAS, G. K. and ARTEMIADIS, P. (2017). Inferring Imagined Speech Using EEG Signals: A New Approach Using Riemannian Manifold Features. *Journal of Neural Engineering* **15** 016002. https://doi.org/10.1088/1741-2552/aa8235

[29] PAN, W., XIANG, E. and YANG, Q. (2012). Transfer Learning in Collaborative Filtering with Uncertain Ratings. *Proceedings of the AAAI Conference on Artificial Intelligence* **26** 662–668. https://doi.org/10.1609/aaai.v26i1.8197

[30] PAN, W., XIANG, E., LIU, N. and YANG, Q. (2010). Transfer Learning in Collaborative Filtering for Sparsity Reduction. *Proceedings of the AAAI Conference on Artificial Intelligence* **24** 230–235. https://doi.org/10.1609/aaai.v24i1.7578

[31] PATHAK, R., MA, C. and WAINWRIGHT, M. (2022). A New Similarity Measure for Covariate Shift with Applications to Nonparametric Regression. In *Proceedings of the 39th International Conference on Machine Learning* **162** 17517–17530. PMLR.

[32] PETEGROSSO, R., PARK, S., HWANG, T. H. and KUANG, R. (2017). Transfer Learning across Ontologies for Phenome–Genome Association Prediction. *Bioinformatics* **33** 529–536. https://doi.org/10.1093/bioinformatics/btw649

[33] PIERSON, E., CONSORTIUM, T. G., KOLLER, D., BATTLE, A. and MOSTAFAVI, S. (2015). Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLOS Computational Biology* **11** e1004220. https://doi.org/10.1371/journal.pcbi.1004220

[34] QIU, Y. and CHEN, S. X. (2012). Test for Bandedness of High-Dimensional Covariance Matrices and Bandwidth Estimation. *The Annals of Statistics* **40** 1285–1314. https://doi.org/10.1214/12-AOS1002

[35] QIU, Y. and CHEN, S. X. (2015). Bandwidth Selection for High-Dimensional Covariance Matrix Estimation. *Journal of the American Statistical Association* **110** 1160–1174. https://doi.org/10.1080/01621459.2014.950375

[36] RAINA, R., NG, A. Y. and KOLLER, D. (2006). Constructing Informative Priors Using Transfer Learning. In *Proceedings of the 23rd International Conference on Machine Learning*. *ICML '06* 713–720. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1143844.1143934

[37] RODRIGUES, P. L. C., JUTTEN, C. and CONGEDO, M. (2019). Riemannian Procrustes Analysis: Transfer Learning for Brain–Computer Interfaces. *IEEE Transactions on Biomedical Engineering* **66** 2390–2401. https://doi.org/10.1109/TBME.2018.2889705

[38] SCHWEIKERT, G., RÄTSCH, G., WIDMER, C. and SCHÖLKOPF, B. (2008). An Empirical Analysis of Domain Adaptation Algorithms for Genomic Sequence Analysis. In *Advances in Neural Information Processing Systems* **21**. Curran Associates, Inc.

[39] SHIN, H.-C., ROTH, H. R., GAO, M., LU, L., XU, Z., NOGUES, I., YAO, J., MOLLURA, D. and SUMMERS, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging* **35** 1285–1298. https://doi.org/10.1109/TMI.2016.2528162

[40] TIAN, Y. and FENG, Y. (2022). Transfer Learning Under High-Dimensional Generalized Linear Models. *Journal of the American Statistical Association* **0** 1–14. https://doi.org/10.1080/01621459.2022.2071278

[41] TSYBAKOV, A. B. (1998). Pointwise and Sup-Norm Sharp Adaptive Estimation of Functions on the Sobolev Classes. *The Annals of Statistics* **26** 2420–2469. https://doi.org/10.1214/aos/1024691478

[42] VERSHYNIN, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108231596

[43] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108627771

[44] WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A Survey of Transfer Learning. *Journal of Big Data* **3** 9. https://doi.org/10.1186/s40537-016-0043-6

[45] YI, F. and ZOU, H. (2013). SURE-tuned Tapering Estimation of Large Covariance Matrices. *Computational Statistics & Data Analysis* **58** 339–351. https://doi.org/10.1016/j.csda.2012.09.007

[46] ZANINI, P., CONGEDO, M., JUTTEN, C., SAID, S. and BERTHOUMIEU, Y. (2018). Transfer Learning: A Riemannian Geometry Framework With Applications to Brain–Computer Interfaces. *IEEE Transactions on Biomedical Engineering* **65** 1107–1116. https://doi.org/10.1109/TBME.2017.2742541

[47] ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., XIONG, H. and HE, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE* **109** 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Department of Statistics and Data Science
The Wharton School, University of Pennsylvania
265 South 37th Street
Philadelphia, Pennsylvania 19104
USA
E-mail: [a]*tcai@wharton.upenn.edu*
[b]*dongwooo@wharton.upenn.edu*
URL: [c]*http://www-stat.wharton.upenn.edu/~tcai/*