Transfer Learning for Covariance Matrix Estimation: Optimality and Adaptivity

T. Tony Cai[1], Dongwoo Kim[1], Yicheng Li[2], and Dongdong Xiang[2*]
[1]Department of Statistics and Data Science,
The Wharton School, University of Pennsylvania
[2]KLATASDS-MOE, School of Statistics,
East China Normal University

April 12, 2026

Abstract

Transfer learning, which leverages auxiliary data to improve inference in a target domain, has become a powerful tool in machine learning. In this work, we study minimax and adaptive estimation of large bandable covariance matrices under a transfer learning framework.

We first establish the minimax rate of convergence under the spectral norm and propose a rate-optimal estimator. Our analysis reveals a phase transition phenomenon that characterizes when and how auxiliary data improves estimation. We then address the problem of adaptation to unknown model parameters, deriving the adaptive rate of convergence. In contrast to conventional settings, we show that the cost of adaptation in the transfer learning can be substantial. To this end, we develop a fully data-driven adaptive estimator that achieves the adaptive rate up to a logarithmic factor.

Simulation studies support our theoretical results and demonstrate the practical effectiveness of the proposed adaptive procedure. We further illustrate its utility in a phoneme classification task, where incorporating related phonemes as auxiliary data significantly improves classification performance.

Keywords: Adaptive rate of convergence, Covariance matrix, High-dimensional statistics, Minimax rate of convergence, Transfer learning.

*Alphabetical order.

1 Introduction

In high-dimensional data analysis, understanding covariance structure is fundamental. Covariance matrices are central to regression, discriminant analysis, clustering, PCA, and graphical modeling. However, as dimensionality increases, accurate estimation becomes increasingly challenging, motivating extensive research into structured covariance models; for comprehensive reviews, see Cai et al. (2016), Cai (2017).

Among structured models, bandable covariance matrices—those with entries decaying away from the diagonal—have received significant attention due to their relevance in applications such as finance, biology, phonetics, climatology, and engineering (He & Chen 2016, Qiu & Chen 2012, Bien et al. 2016, Li et al. 2017, Lee et al. 2022, Qiu & Chen 2015, Yi & Zou 2013). Several estimation methods have been developed to exploit this structure, with minimax optimality (Bickel & Levina 2008, Cai et al. 2010, Cai & Yuan 2012).

Concurrently, *transfer learning*—which leverages information from auxiliary datasets to improve inference in a target domain—has become a powerful tool in modern machine learning. It has shown success in domains such as natural language processing, bioinformatics, medical imaging, and recommendation systems (Han et al. 2021, Petegrosso et al. 2017, Maqsood et al. 2019, Hu et al. 2019). For comprehensive surveys, see Weiss et al. (2016), Zhuang et al. (2021). Recent theoretical work has extended transfer learning to various statistical tasks, including nonparametric models, GLMs, multi-armed bandits, and functional estimation (Cai & Wei 2021, Cai & Pu 2022, Li et al. 2023, Cai, Cai & Li 2024, Cai, Kim & Pu 2024).

When transfer learning and covariance matrix estimation intersect, they create a synergy capable of addressing more complex and intriguing statistical problems. It enables more efficient use of related but distinct datasets, particularly when target data is limited or prohibitively expensive to collect. This synergy has proven useful in applications such as GTEx analysis, EEG-based BCI, and text classification (Li et al. 2022, Rodrigues et al. 2019, Raina et al. 2006), where transfer learning methods can substantially improve accuracy of covariance estimation.

1.1 Formulation of the Problem

Our discussion begins with an illustrative example of covariance matrix estimation within the transfer learning context, based on phoneme sound analysis. Phonemes—the fundamental units of sound in language—are typically represented by periodogram vectors, which capture sound intensity across a wide range of frequencies from short audio recordings of the corresponding phoneme (Hamooni & Mueen 2014).

Zooming in on a particular phoneme in English, say `ZH` as in "seizure", we consider a target sample $\{X_1^{(t)}, \ldots, X_{n_t}^{(t)}\}$ where each observation $X_i^{(t)} \in \mathbb{R}^d$ is independently and identically distributed. Within the context of our phoneme study, the dimensionality d corresponds to the number of frequencies chosen to extract periodogram vectors from the audio recordings. Our goal is to estimate the target covariance matrix $\Sigma^{(t)} \in \mathbb{R}^{d\times d}$ of the periodogram vectors for the `ZH` phoneme. This estimation problem naturally arises in applications such as phoneme classification tasks (Hastie et al. 1995, Bien et al. 2016).

A key characteristic of the periodogram vector is its inherent ordering from low to high frequencies, which induces a distinctive structure in the target covariance matrix $\Sigma^{(t)}$. Entries close to the diagonal reflect covariances between intensities at nearby frequencies and are expected to exhibit larger magnitudes. In contrast, entries farther from the diagonal—corresponding to covariances between intensities at widely separated frequencies—tend to decay toward zero. This structure is clearly visualized in Figure 7 in Section 4.

The described structure of the phoneme dataset aligns with a bandable structure, a concept formulated by Bickel & Levina (2008). 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:

$$\mathfrak{A}_d(B; u) \coloneqq \Big(b_{jk}\mathbb{1}\big(|j-k| > u\big)\Big)_{j,k\in[d]} \qquad \text{where } B = (b_{jk})_{j,k\in[d]} \in \mathbb{R}^{d\times d}.$$

Figure 2a in Section 2 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:

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

where $\mathbb{S}_d^+$ represents the collection of $d \times d$ symmetric and positive definite matrices and $\|\cdot\|_1$ indicates the matrix ℓ_1-norm.

The conventional learning framework centers around finding an effective method to estimate the target covariance matrix based on the available target sample. However, 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 (Garofolo et al. 1993). However, the ZH phoneme is the least frequently occurring sound in everyday English speech (Hamooni & Mueen 2014, Hayden 1950), making its recordings scarce and the collection of additional data prohibitively expensive.

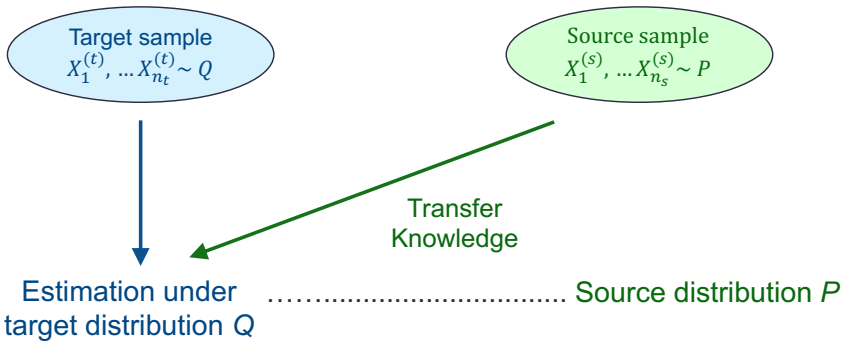


Figure 1: An illustration for transfer learning

One way to address this limitation is to extend the conventional learning framework through a transfer learning strategy. This approach allows us to utilize recordings of other phonemes that are different yet bear some similarities. For instance, the S phoneme, as in "see", is categorized as an obstruent fricative in the hierarchy of English phonemes, like the ZH phoneme (Dekel et al. 2004, Lee & Hon 1989). Moreover, the S phoneme is significantly more common, ranking as the fifth most frequent sound in English speech (Hamooni & Mueen 2014, Hayden 1950). 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 covariance matrix $\Sigma^{(s)} \in \mathbb{R}^{d\times d}$. While our objective remains the same, namely 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 illustrates the transfer learning strategy.

The effectiveness of transfer learning depends on a meaningful and quantifiable relationship between the source and target samples. In this paper, we assume that:

- Both the target and source covariance matrices are bandable. Specifically, for target parameters (α_t, M_t) and source parameters (α_s, M_s), we assume $\Sigma^{(t)} \in \mathcal{B}_{\alpha_t}(M_t)$ and $\Sigma^{(s)} \in \mathcal{B}_{\alpha_s}(M_s)$.

- The disparity matrix $\Delta^{(s)} \coloneqq \Sigma^{(t)} - \Sigma^{(s)}$, which measures the deviation between the target and source covariance matrices, is sparse in the matrix ℓ_1-norm. For a given disparity threshold $\delta \geq 0$, we assume $\|\Delta^{(s)}\|_1 \leq \delta$.

More general disparity conditions and approximately bandable misspecification are discussed in the supplementary material.

Let $\mathcal{P}_{\alpha_t,\alpha_s,\delta}(M_t, M_s, \rho)$ denote the collection of joint distributions of the target and source samples satisfying these assumptions together with the uniform sub-Gaussian condition in Assumption 2.1 with parameter ρ for both samples. We focus on the bandable class as it serves as the canonical model for ordered variables and provides a transparent setting to study the interactions between structural decay, domain disparity, and relative sample sizes.

To establish minimax optimality, we evaluate estimators $\widehat{\Sigma}^{(t)}$ that utilize both target and source samples under the squared spectral-norm loss. The minimax risk is defined as:

$$\inf_{\widehat{\Sigma}^{(t)}} \sup_{\mathbb{P}\in\mathcal{P}_{\alpha_t,\alpha_s,\delta}(M_t,M_s,\rho)} \mathbb{E}\,|\,\widehat{\Sigma}^{(t)} - \Sigma^{(t)}\,|\,_2^2.$$

Our first main result characterizes the sharp rate of this risk.

1.2 Main Results and Our Contribution

Under the above model and loss, our first main result shows that the minimax risk satisfies the following rate. By letting $n \coloneqq n_t + n_s$ and $\alpha_{\max} \coloneqq \alpha_t \vee \alpha_s$, the minimax optimal rate

is expressed as:

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

That is, Equation (1) gives the order of the minimax risk for estimating the target covariance matrix in the transfer learning setting under the squared spectral-norm loss.

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 et al. (2010). 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.

While we have established the minimax rate of convergence, it remains to construct an optimal estimator that does not rely on the unknown parameters α_t, α_s, or the disparity threshold δ, which are typically inaccessible in practice. The second major contribution of this paper is to address the challenge of adaptation to such unknown parameters. Notably, and in sharp contrast to the conventional learning setting (Cai & Yuan 2012), we show that adaptation necessarily incurs a cost in the transfer learning framework. Specifically, we establish the following *adaptive rate of convergence* when model parameters are unknown:

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

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

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

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. Any kind of adaptive estimator that performs better than 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 adaptive rate in Equation (2) is achieved up to a logarithmic factor by our proposed algorithm, the adaptive tridiagonal block-thresholding estimator. This fully data-driven procedure dynamically adapts to unknown model parameters without requiring prior knowledge. We further demonstrate its practical effectiveness through both simulation studies and a real-data application. The simulation results validate the theoretical phase transition phenomena: transfer learning improves estimation when the disparity is small, but offers little benefit when the disparity is large. Additionally, we apply the adaptive estimator to a phoneme classification task using the TIMIT dataset. Incorporating auxiliary data from related phonemes substantially enhances classification accuracy in both QDA and LDA, underscoring the practical utility of our proposed method.

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 simulation study and real-data analysis 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.

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 $\Sigma \in \mathbb{R}^{d\times d}$, which represents the weighted average of the population covariance matrices, $\Sigma^{(t)}$ and $\Sigma^{(s)}$. Concurrently, we define another disparity matrix $\Delta \in \mathbb{R}^{d\times d}$ as the deviation of the target covariance matrix $\Sigma^{(t)}$ from the pooled covariance matrix Σ. By letting $n \coloneqq n_t + n_s$, these are mathematically represented as:

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

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 $g \in \{t, s\}$,

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

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

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

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

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

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. Recall that $\mathcal{P}_{\alpha_t,\alpha_s,\delta}(M_t, M_s, \rho)$ denotes the collection of joint distributions of the target and source samples such that $\Sigma^{(t)} \in \mathcal{B}_{\alpha_t}(M_t)$, $\Sigma^{(s)} \in \mathcal{B}_{\alpha_s}(M_s)$, $\|\Sigma^{(t)} - \Sigma^{(s)}\|_1 \le \delta$, and both samples satisfy the following uniform sub-Gaussianity condition:

Assumption 2.1 (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:

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

where we denote by $\|\cdot\|_{\psi_2}$ the sub-Gaussian norm or Orlicz ψ_2-norm. As per Wainwright (2019) and Vershynin (2018), the sub-Gaussian norm of a random variable Z is defined by

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

We are now ready to state the optimality theory for estimating the target covariance matrix under the transfer learning framework. The following theorem characterizes the minimax rate of convergence under the squared spectral-norm loss. In particular, it highlights how the interaction between the sample sizes (n_t, n_s), the dimension (d), and the disparity level (δ) governs the fundamental limits of transfer learning.

Theorem 2.2. *One can consistently estimate the target covariance matrix if and only if at least one of the following conditions holds:*

Assumption (a) $\log d \ll n_t$.

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

Under either condition, the optimal convergence rate under the squared spectral-norm loss is given by

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

Remark 2.3. These results extend beyond the exact bandable model and the matrix ℓ_1-disparity condition. Under weaker assumptions, we replace the ℓ_1 constraint with an upper-tail stability (UTS) condition and allow for approximately bandable misspecification. In this broader setting, the same upper bound holds, supplemented only by a term quantifying the misspecification level.

We maintain the exact ℓ_1-based formulation in the main text to keep the phase-transition phenomena transparent. The more general statements are deferred to Subsection A.1 in the supplementary material for pedagogical clarity, rather than any limitation in the theory itself. In the remainder of the paper, we retain the class $\mathcal{P}_{\alpha_t,\alpha_s,\delta}$ to streamline the presentation.

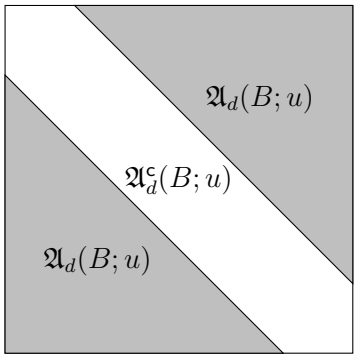
Due to space constraints, the proof of Theorem 2.2 is deferred to the supplementary material. We instead introduce the key technical ingredient for establishing the upper

bound: the *blockwise tridiagonal operator*, denoted by $\mathfrak{T}_d$, which was previously employed in high-dimensional covariance estimation by Cai & Zhang (2016). This operator is defined for an integer bandwidth $u > 0$ and can be expressed as follows. Let $\lfloor x \rfloor$ denote the greatest integer not exceeding $x \in \mathbb{R}$.

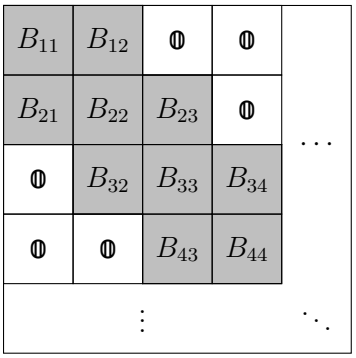
$$\mathfrak{T}_d(B; u) := \Big(b_{jk} \mathbb{1}\big(|u_j - u_k| \le 1\big) \Big)_{j,k \in [d]} \quad \text{where} \quad \begin{cases} B = (b_{jk})_{j,k \in [d]} \in \mathbb{R}^{d \times d}, \\ u_j := 1 + \big\lfloor (j-1)/u \big\rfloor, \ (j \in [d]). \end{cases}$$

Intuitively, when the matrix B is partitioned into $u \times u$ blocks, the operator $\mathfrak{T}_d$ retains only the blockwise main diagonal and the immediate super- and sub-diagonal blocks, resulting in a block-tridiagonal structure. A visual illustration of this mechanism is provided in Figure 2b.

By applying the blockwise tridiagonal operator $\mathfrak{T}_d$ with a carefully chosen bandwidth, we can establish the upper bound in the minimax rate of Theorem 2.2. Since the bandable structure implies decay of the away-from-diagonal parts, it is natural to truncate the sample covariance matrices using $\mathfrak{T}_d$. Although full sample covariance matrices perform poorly in high-dimensional settings, the low-dimensional blocks retained by $\mathfrak{T}_d$ can yield accurate local estimates, provided that the bandwidth is selected appropriately. Detailed derivations are provided in the supplementary material.



(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

Figure 2: A visual representation of important matrix operators

Given the extensive literature on covariance matrix estimation under the conventional setting, it is instructive to compare our minimax result with known rates. Notably, the transfer learning framework reduces to the conventional framework when $n_s = 0$ and $\delta = \infty$,

as the source sample is absent and the source population provides no information about the target covariance structure. In this case, Cai et al. (2010) established that the minimax rate of convergence is given by

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

The optimal convergence rate established in Theorem 2.2 also provides insight into the effectiveness of the transfer learning. A natural question is to identify the conditions under which transfer learning yields tangible benefits, and to quantify the extent of such improvements. Notably, the analysis reveals a phase transition phenomenon, where the utility of transfer learning undergoes a sharp change depending on the disparity level (δ) and the source sample size (n_s). The main findings are concisely summarized in Table 1.

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

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

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 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. This also highlights a safety property at the minimax level: as the disparity increases into the strong-disparity

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="2">No</td></tr>
<tr><td>with auto-smoothing benefit</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Ample source sampling regime ($n_s \gg n_t$)</td><td></td><td>Yes / Yes</td><td>Yes / No</td></tr>
</table>

regime, the transfer benchmark degrades back to the target-only minimax rate Φ^{LE}, rather than becoming worse than the conventional baseline. 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 $n_s \gg n_t$ holds, 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 & Pu (2022) 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 Section 2 establishes the minimax rate of convergence, the analysis critically depends on the knowledge of key model parameters, including the decay rates α_t and α_s, as well as the disparity threshold δ. Although this assumption facilitates theoretical insight, these parameters are typically unknown in practice, which limits the practical applicability of the proposed framework.

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 & Yuan 2012).

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:

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

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

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

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, for any constant $\zeta > 0$, we can identify a generic constant $c_2 > 0$ and parameters $\alpha_0 > 0$ and $\delta_0 \geq 0$, defined in the proof in the supplementary material, such that:

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

Although the optimal rate is given by Φ as discussed in Section 2, Proposition 3.1 delivers a sobering message: no adaptive estimator can 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 Subsection 3.1, we introduce an adaptive estimator and assess its maximal risk in terms of the squared matrix ℓ_2-norm. Subsequently, in Subsection 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*.

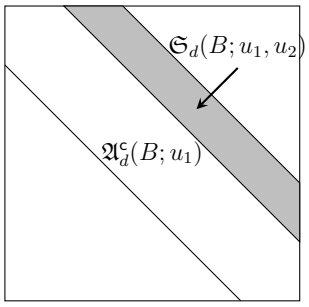
3.1 Adaptive Tridiagonal Block-thresholding Estimator

We will establish an adaptive method for estimating bandable covariance matrices within the transfer learning framework. Before diving into the primary algorithm, we define

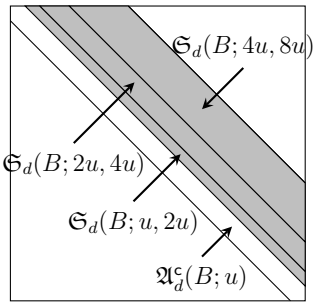
an important operator known as the superbanding operator, denoted by $\mathfrak{S}_d$. Given two bandwidths $u_2 > u_1 > 0$, this operator is formulated as:

$$\mathfrak{S}_d(B; u_1, u_2) := \Big(b_{jk} \mathbb{1}\big(u_1 < k - j \le u_2\big) \Big)_{j,k\in[d]} \qquad \text{where } B = \big(b_{jk}\big)_{j,k\in[d]} \in \mathbb{R}^{d\times d}.$$

Recall that the banding operator $\mathfrak{A}_d^{\mathfrak{c}}$ 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 $\mathfrak{A}_d^{\mathfrak{c}}(\,\cdot\,; u_1)$. 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 $\mathfrak{A}_d(\,\cdot\,; u_1)$.

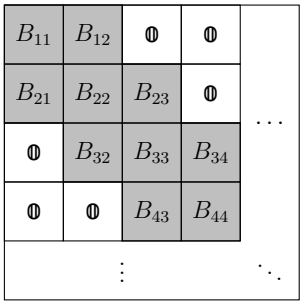


(a) The output, $\mathfrak{S}_d(B; u_1, u_2)$, of the superbanding operator

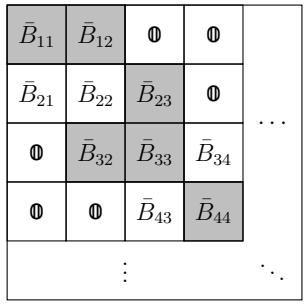


(b) Approximation and partition of the away-from-diagonal region

Figure 3: The superbanding operator $\mathfrak{S}_d$



(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

Figure 4: The tridiagonal block-thresholding operator $\overline{\mathfrak{T}}_d$

Plus, 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 $B \in \mathbb{R}^{d\times d}$ 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 $R > 0$:

$$\bar{B}_{ij} := B_{ij}\mathbb{1}\big(\|B_{ij}\|_2 \geq R\big) \qquad \text{for each sub-matrix } B_{ij} \text{ of } \mathfrak{T}_d(B;u).$$

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:

$$\widehat{u} := \operatorname*{argmin}_{u\in\mathcal{U}} \left(\sum_{\ell=1}^{\ell_u} \|\mathfrak{S}_d(\check{\Sigma}^{(t)}; 2^{\ell-1}u, 2^{\ell}u)\|_2 + \lambda\rho^2 \sqrt{\frac{u+\log d}{n_t}} \right),$$

where we define

$$u_{\min} = \log d, \qquad u_{\max} = \big((n_t \log^{-1} n_t) \wedge d\big) \vee \log d,$$
$$\mathcal{U}_{\mathrm{dy}} := \{2^k u_{\min} : k \in \mathbb{Z}_{\geq 0},\, 2^k u_{\min} \leq u_{\max}\}, \qquad \mathcal{U} := (\mathcal{U}_{\mathrm{dy}} \cup \{d, u_{\max}\}) \cap [u_{\min}, u_{\max}],$$

and

$$\ell_u := \max\big(\{0\} \cup \{\ell \in \mathbb{Z}^+ : 2^{\ell-1}u \leq n_t \log^{-1} n_t\}\big).$$

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:

$$\widehat{v} := \operatorname*{argmin}_{v \in \mathcal{V}} \left(\sum_{\ell=1}^{\ell_v} \|\mathfrak{S}_d(\check{\Sigma}; 2^{\ell-1}v, 2^{\ell}v)\|_2 + \tau\rho^2 \sqrt{\frac{v + \log d}{n}} \right),$$

where

$$\begin{aligned}
&v_{\min} = \log d, \qquad v_{\max} = (n \log^{-1} n \wedge d) \vee \log d, \\
&\mathcal{V}_{\mathrm{dy}} := \{2^k v_{\min} : k \in \mathbb{Z}_{\geq 0},\, 2^k v_{\min} \leq v_{\max}\}, \qquad \mathcal{V} := (\mathcal{V}_{\mathrm{dy}} \cup \{d, v_{\max}\}) \cap [v_{\min}, v_{\max}], \\
&\ell_v := \max\big(\{0\} \cup \{\ell \in \mathbb{Z}^+ : 2^{\ell-1}v \leq n \log^{-1} n\}\big).
\end{aligned}$$

4: For any large enough constant $\xi_1 > 0$, solve the following optimization problem:

$$\widehat{w} := \operatorname*{argmin}_{w \in \mathcal{W}} \left(\sum_{m=1}^{\ell_w} \|\mathfrak{S}_d(\check{\Delta}; 2^{m-1}w, 2^{m}w)\|_2 + \xi_1\rho^2 \sqrt{\frac{w + \log nd}{n_t}} \right),$$

where

$$\begin{aligned}
&w_{\min} = \log d, \qquad w_{\max} = (n_t \log^{-1} n_t \wedge d) \vee \log d, \\
&\mathcal{W}_{\mathrm{dy}} := \{2^k w_{\min} : k \in \mathbb{Z}_{\geq 0},\, 2^k w_{\min} \leq w_{\max}\}, \qquad \mathcal{W} := (\mathcal{W}_{\mathrm{dy}} \cup \{d, w_{\max}\}) \cap [w_{\min}, w_{\max}], \\
&\ell_w := \max\big(\{0\} \cup \{m \in \mathbb{Z}^+ : 2^{m-1}w \leq n_t \log^{-1} n_t\}\big).
\end{aligned}$$

5: Output the final estimator:

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

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

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

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.

We now introduce two procedures: Algorithm 1 relies solely on the target sample, and Algorithm 2 leverages pooled data from both the target and source samples. In both algorithms, the superbanding operator $\mathfrak{S}_d$ plays a key role in selecting the optimal bandwidth for the blockwise tridiagonal operator $\mathfrak{T}_d$. The idea is to decompose the away-from-diagonal region into smaller dyadic segments. Additionally, the tridiagonal block-thresholding operator $\overline{\mathfrak{T}}_d$ is essential in Algorithm 2 to mitigate the risk of incorporating a poorly aligned or adversarial source distribution. Finally, we carefully combine both algorithms to construct the *adaptive tridiagonal block-thresholding estimator*.

The adaptive tridiagonal block-thresholding estimator, as we describe in Theorem 3.2, is entirely data-driven. This includes Algorithms 1 and 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. An extension paralleling Remark 2.3 is also available for the adaptive estimator; we defer it to the supplementary material.

Theorem 3.2 (Adaptive tridiagonal block-thresholding estimator)**.** *Consider the output of Algorithms 1 and 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:

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

provided the following assumption holds:

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

When conditions $n_s = 0$ and $\delta = \infty$ are met, the transfer learning paradigm simplifies to the conventional learning framework. 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 & Yuan (2012) 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.

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:

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

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.

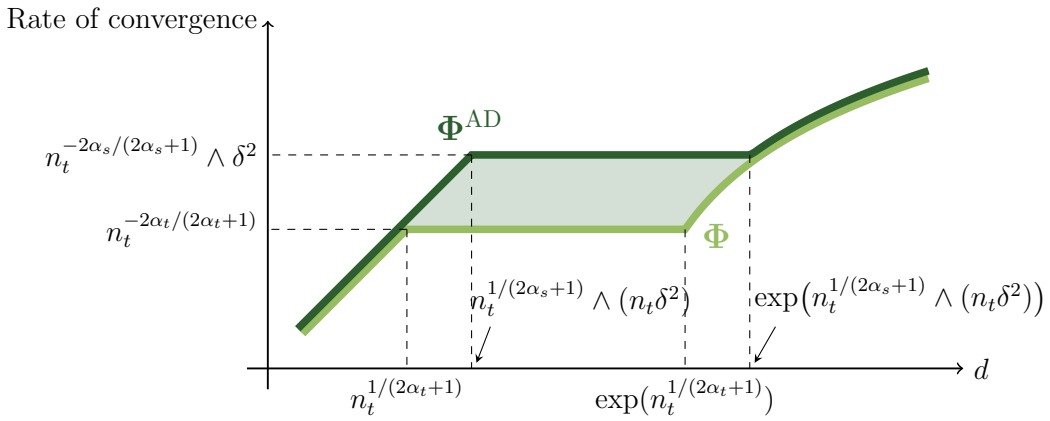


Figure 5: Comparison between the adaptive rate Φ^{AD} and the minimax rate Φ

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 the supplementary material.

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 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:*

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

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):

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

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

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

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 by Tsybakov (1998).

(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 $\log n$.

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

(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.

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

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 insights into the feasibility of applying our estimator in practical contexts.

(Scenario I: $n_s \leq n_t$)
This regime parallels the conventional learning framework, particularly when $n_s = 0$ and $\delta = \infty$. Adaptivity is achieved without additional cost, simply because the source sample is not enough. However, this setting limits the main advantage of transfer learning—leveraging a larger sample size.

(Scenario II: $\alpha_t \leq \alpha_s$)
When the source covariance decays at least as fast as the target, we can bypass the auto-unsmoothing effect. Since $\alpha_{\min} = \alpha_t$, the adaptive rate φ^{AD} matches the optimal rate:

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

(Scenario III: $\delta \leq C_\delta n_t^{-\alpha_t/(2\alpha_t+1)}$)
When the source and target covariance matrices are closely aligned, auto-unsmoothing has no impact. Adaptivity thus follows cost-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 $\log d \geq C_d n_t^{1/(2\alpha_s+1)}$)
When the dimension d is too small or too large relative to the target sample size n_t, the decay rate becomes irrelevant in determining the convergence rate. Here, the cost of adaptation disappears, regardless of the source sample size n_s.

Finally, we note that in the least favorable regime, the cost of adaptation can also manifest itself as negative transfer, meaning that a transfer-based procedure performs worse than the target-only procedure that ignores the source sample. For example, the adaptive

rate $n_t^{-2\alpha_s/(2\alpha_s+1)}$ can be slower than the target-only rate $n_t^{-2\alpha_t/(2\alpha_t+1)}$. Therefore, one should not expect a fully data-driven transfer rule to be both adaptive and uniformly protected against such deterioration.

4 Numerical Experiments

This section evaluates the numerical performance of the adaptive estimator, as depicted in Theorem 3.2. Subsection 4.1 examines its behavior under varying source sample sizes (n_s) and disparity levels (δ), confirming consistency with our theoretical findings. Subsection 4.2 demonstrates its practical utility in a phoneme classification task. Supplementary Material also contains additional experiments that examine robustness beyond the baseline setting, including more general source-target disparity structures, misspecification of exact bandability, and non-Gaussian sampling.

4.1 Simulation Study

We consider the target and source covariance matrices as follows:

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

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

In this context, $\xi_{jk}^{(t)}$ and $\xi_{jk}^{(s)}$ $(1 \leq j < k \leq d)$ 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 spectral-norm loss. The comprehensive results are visually represented in Figure 6, which unveils several intriguing patterns and corroborates our theoretical findings.

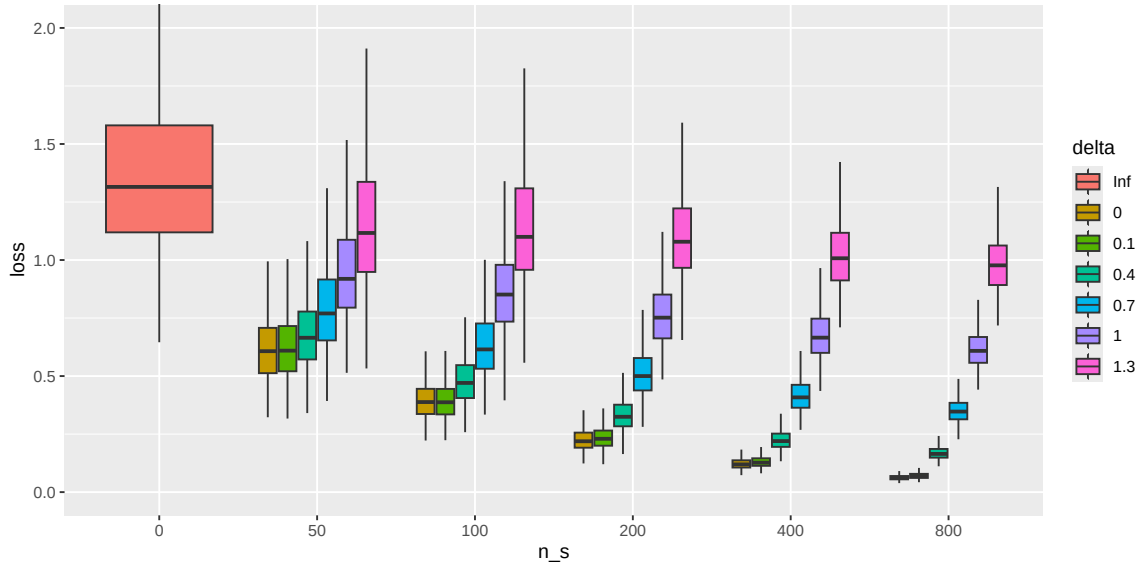


Figure 6: The results of numerical experiments under $\alpha_t = 2$ and $\alpha_s = 1$

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 of the supplementary material.

Our theoretical framework suggests that adaptivity across the full spectrum of mod-

els 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.

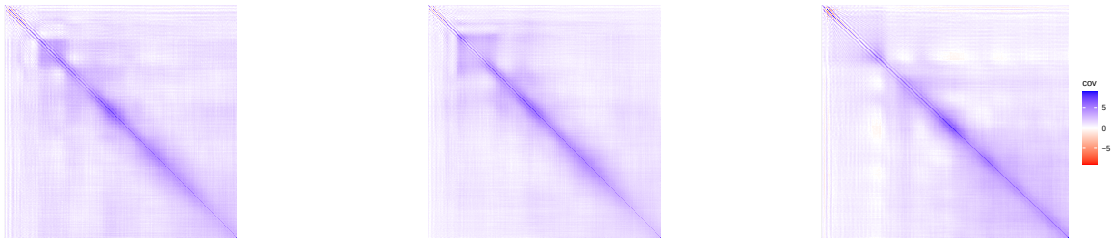
Conversely, when δ is set below 1.3, Figure 6 exhibits an evident improvement relative to the baseline scenario, where $(n_s, \delta) = (0, \infty)$. This improvement amplifies as the source sample size n_s expands or as the disparity threshold δ diminishes. Remarkably, the model's performance at $\delta = 0.1$ 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.

Practical guidance. A reasonable workflow is the following. One should first assess whether the source sample is indeed related to the target population. If so, a conservative approach is to compute the normalized sample-level discrepancy score $\widehat{\delta}_{\mathrm{op}} := \frac{\|\check{\Sigma}^{(s)} - \check{\Sigma}^{(t)}\|_{\mathrm{op}}}{\|\check{\Sigma}^{(t)}\|_{\mathrm{op}}}$, where $\check{\Sigma}^{(t)}$ and $\check{\Sigma}^{(s)}$ denote the target and source sample covariance matrices. Smaller values of $\widehat{\delta}_{\mathrm{op}}$ indicate that the source covariance is close to the target covariance relative to the target scale. In practice, a very conservative, problem-dependent cutoff level such as 0.6 can be used as a screening rule; we include this threshold only as a conservative illustration drawn from our diagnostic experiments. When $\widehat{\delta}_{\mathrm{op}} < 0.6$ and simple exploratory diagnostics such as correlation heatmaps or leading-eigenvalue decay show broadly compatible structure, the transfer version of the adaptive tridiagonal block-thresholding estimator is a reasonable default. However, in other cases, a target-only estimator is safer. This guideline is necessarily heuristic rather than definitive, because our adaptive lower-bound theory shows that no data-driven rule can both exploit the source sample and uniformly avoid negative transfer over the whole model class.

4.2 Application to Discriminant Analysis of Phoneme Dataset

In this section, we apply our adaptive tridiagonal block-thresholding estimator to a real-world dataset. Since the true covariance matrix is generally unknown in practice, directly assessing the accuracy of our estimate is infeasible. However, any method that relies on covariance estimation is expected to benefit from a more accurate covariance estimator when a similar but distinct source sample is available. To evaluate the practical utility of our estimator, we employ discriminant analysis in a classification problem, using classification accuracy as an indirect measure of performance. Specifically, we anticipate higher accuracy in discriminant analysis when incorporating a source sample and applying our covariance estimator, as knowledge transfer should lead to a more accurate estimate of the covariance matrix.

We consider a binary classification problem as described in Hastie et al. (2009). The dataset is derived from the TIMIT database, a widely used resource for speech recognition research, provided by the Linguistic Data Consortium (Garofolo et al. 1993). The processed phoneme benchmark considered here follows the version used in Hastie et al. (1995). It contains short sound recordings of male speakers pronouncing one of two similar-sounding phonemes. The objective is to develop a classifier for automatically labeling these sounds. The predictor vectors, x, are log-periodograms, which represent the log-intensity of the sound recordings across $d = 256$ frequencies. Each log-periodogram is labeled with its corresponding phoneme, y.

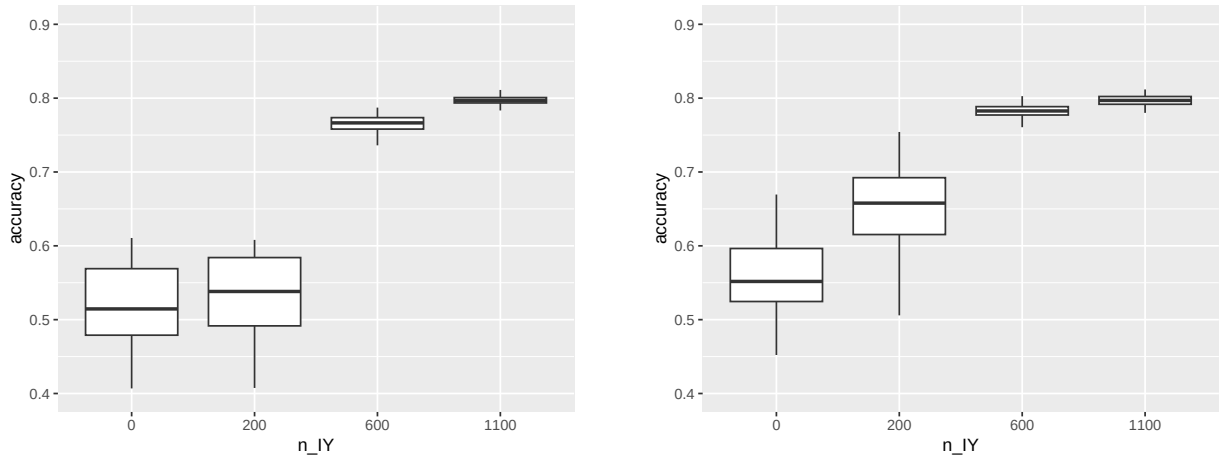


(a) `AA` (b) `AO` (c) `IY`

Figure 7: Sample Covariance Matrices of Three Phoneme Sounds

Our analysis focuses on three phonemes: `AA` as the vowel in "dark", `AO` as the first vowel in "water" and `IY` as the vowel in "she". Specifically, we develop a classifier to distinguish between `AA` and `AO` phonemes. They sound very similar and are difficult to differentiate,

even for human listeners, given the short duration of the recordings. We randomly split the data into training and test sets, with the training set containing AA samples of $n_{\texttt{AA}} = 69$ and AO samples of $n_{\texttt{AO}} = 102$.



(a) QDA Classifiers

(b) LDA Classifiers

Figure 8: Accuracy of Discriminant Analysis of Phoneme Dataset

Note that the training set is relatively small compared to the dimension $d = 256$, which hinders both the accurate estimation of the covariance matrix and the performance of discriminant analysis. One way to mitigate the effects of limited sample size is to integrate similar but distinct sounds. In particular, the IY sound is classified as a sonorant vowel, like both AA and AO (Dekel et al. 2004, Lee & Hon 1989). Figure 7 supports the use of the adaptive tridiagonal block-thresholding estimator, as all covariance matrices exhibit a bandable structure, and IY shows some degree of similarity with both AA and AO.

With 1,163 auxiliary IY samples available, we utilize subsets of these to enhance covariance matrix estimation for Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA). Specifically, we randomly sample IY sounds at sizes $n_{\texttt{IY}} \in \{0, 200, 600, 1100\}$. Note that $n_{\texttt{IY}} = 0$ represents the conventional estimator, which lacks access to auxiliary data.

For each random split, LDA utilizes the pooled sample covariance matrix of the AA and AO training data, while QDA employs class-specific sample covariance matrices. When $n_{\texttt{IY}} > 0$, we compute the sample covariance of the IY subsample and provide it as the source input to the adaptive tridiagonal block-thresholding estimator—applied once for the pooled covariance in LDA and separately for the two class-specific matrices in QDA. Class means

and priors are consistently estimated from the AA/AO training data alone. The resulting covariance estimates are then plugged into the standard LDA or QDA decision rules.

Theoretical support for this approach is provided in Proposition E.1 of Supplementary Material E, which links operator-norm covariance error to excess classification risk in binary Gaussian LDA. The classification accuracy of these models is presented in Figure 8. While both classifiers perform only marginally better than a random baseline when $n_{\texttt{IY}} = 0$, both accuracy and stability improve significantly as more IY samples are incorporated.

5 Discussions

This paper explores minimax and adaptive estimation of high-dimensional bandable covariance matrices within a transfer learning framework. We establish the minimax rate of convergence under squared spectral-norm loss and introduce a blockwise tridiagonal estimator that achieves optimality. This minimax rate reveals intriguing phase transition phenomena, characterizing the regimes where source samples significantly enhance target-domain inference.

Conversely, our analysis of adaptive convergence rates shows that—unlike conventional learning—the cost of adaptation in transfer learning can be substantial in certain regimes. To address this, we develop a novel data-driven algorithm that automatically adapts to unknown parameters.

From a practical perspective, the adaptive transfer estimator is a robust default when the source is scientifically related to the target and exploratory summaries indicate similar covariance structures, as adaptation often incurs no additional cost in favorable regimes. However, our adaptive lower-bound theory demonstrates that no universally valid rule can exploit the source sample while uniformly avoiding negative transfer. When source and target domains appear clearly mismatched, a target-only estimator remains the safer choice.

Our study focuses on estimation under the squared spectral-norm loss. While conventional bandable covariance estimation has been investigated under other measures—such as the matrix ℓ_1-norm, Frobenius norm, and general Bregman divergences—extending these to the transfer learning framework represents a compelling direction for future work. Similarly, extending these results to other structures, such as sparse and spiked covariance matrices,

or to the estimation of bandable precision matrices, remains an open and promising area of research.

Finally, we identify robustness to heavy-tailed or contaminated observations as a critical future direction. Although simulations in the supplementary material suggest the proposed procedure remains empirically competitive beyond Gaussian settings, a rigorous extension would require robust estimators tailored to the transfer setting. Potential avenues include matrix-truncation and spectrum-truncation approaches (Minsker & Wei 2018, Ke et al. 2019), or contamination-robust procedures under Huber's model (Chen et al. 2018, Avella-Medina et al. 2018). Retaining sharp upper bounds and adaptive guarantees in these robust-transfer settings remains a substantial methodological challenge.

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

SUPPLEMENTARY MATERIAL

Title: Supplement to "Transfer Learning for Covariance Matrix Estimation: Optimality and Adaptivity."

This supplementary material provides 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 contents are organized as follows. Supplementary Material A introduces fundamental notation, develops weaker-assumption and misspecified-model extensions through Proposition A.1 and Theorems A.3 and A.4, and contains the proofs of Theorems A.5, 2.2, 3.2 and 3.3 and Proposition 3.1. Supplementary Material B and Supplementary Material C provide proofs of the technical results for optimality and adaptivity, respectively. Supplementary Material D contains additional numerical experiments. Supplementary Material E gives a preliminary classification guarantee for plug-in LDA.

References

Avella-Medina, M., Battey, H. S., Fan, J. & Li, Q. (2018), 'Robust estimation of high-dimensional covariance and precision matrices', *Biometrika* **105**(2), 271–284.

Bickel, P. J. & Levina, E. (2008), 'Regularized estimation of large covariance matrices', *The Annals of Statistics* **36**(1), 199–227.

Bien, J., Bunea, F. & Xiao, L. (2016), 'Convex Banding of the Covariance Matrix', *Journal of the American Statistical Association* **111**(514), 834–845.

Cai, C., Cai, T. T. & Li, H. (2024), 'Transfer learning for contextual multi-armed bandits', *The Annals of Statistics* **52**(1), 207–232.

Cai, T. T. (2017), 'Global Testing and Large-Scale Multiple Testing for High-Dimensional Covariance Structures', *Annual Review of Statistics and Its Application* **4**(Volume 4, 2017), 423–446.

Cai, T. T., Kim, D. & Pu, H. (2024), 'Transfer learning for functional mean estimation: Phase transition and adaptive algorithms', *The Annals of Statistics* **52**(2), 654–678.

Cai, T. T. & Pu, H. (2022), 'Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure', *Technical report* .

Cai, T. T., Ren, Z. & Zhou, H. H. (2016), 'Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation', *Electronic Journal of Statistics* **10**(1), 1–59.

Cai, T. T. & Wei, H. (2021), 'Transfer Learning for Nonparametric Classification: Minimax Rate and Adaptive Classifier', *The Annals of Statistics* **49**(1), 100–128.

Cai, T. T. & Yuan, M. (2012), 'Adaptive covariance matrix estimation through block thresholding', *The Annals of Statistics* **40**(4), 2014–2042.

Cai, T. T. & Zhang, A. (2016), 'Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data', *Journal of Multivariate Analysis* **150**, 55–74.

Cai, T. T., Zhang, C.-H. & Zhou, H. H. (2010), 'Optimal rates of convergence for covariance matrix estimation', *The Annals of Statistics* **38**(4), 2118–2144.

Chen, M., Gao, C. & Ren, Z. (2018), 'Robust covariance and scatter matrix estimation under huber's contamination model', *The Annals of Statistics* **46**(5).

Dekel, O., Keshet, J. & 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, Springer-Verlag, Berlin, Heidelberg, pp. 146–158.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. & Dahlgren, N. L. (1993), 'TIMIT Acoustic-Phonetic Continuous Speech Corpus'.

Hamooni, H. & Mueen, A. (2014), Dual-Domain Hierarchical Classification of Phonetic Time Series, *in* '2014 IEEE International Conference on Data Mining', pp. 160–169.

Han, W., Pang, B. & Wu, Y. (2021), 'Robust Transfer Learning with Pretrained Language Models through Adapters'.

Hastie, T., Buja, A. & Tibshirani, R. (1995), 'Penalized Discriminant Analysis', *The Annals of Statistics* **23**(1), 73–102.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York, NY.

Hayden, R. E. (1950), 'The Relative Frequency of Phonemes in General-American English', *WORD* **6**(3), 217–223.

He, J. & Chen, S. X. (2016), 'Testing super-diagonal structure in high dimensional covariance matrices', *Journal of Econometrics* **194**(2), 283–297.

Hu, G., Zhang, Y. & Yang, Q. (2019), Transfer Meets Hybrid: A Synthetic Approach for Cross-Domain Collaborative Filtering with Text, *in* 'The World Wide Web Conference', WWW '19, Association for Computing Machinery, New York, NY, USA, pp. 2822–2829.

Ke, Y., Minsker, S., Ren, Z., Sun, Q. & Zhou, W.-X. (2019), 'User-friendly covariance estimation for heavy-tailed distributions', *Statistical Science* **34**(3).

Lee, K.-F. & Hon, H.-W. (1989), 'Speaker-independent phone recognition using hidden Markov models', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37**(11), 1641–1648.

Lee, K., Lee, K. & Lee, J. (2022), 'Estimation of conditional mean operator under the bandable covariance structure', *Electronic Journal of Statistics* **16**(1), 1253–1302.

Li, S., Cai, T. T. & Li, H. (2022), 'Transfer Learning in Large-Scale Gaussian Graphical Models with False Discovery Rate Control', *Journal of the American Statistical Association* **0**(0), 1–13.

Li, S., Zhang, L., Cai, T. T. & Li, H. (2023), 'Estimation and Inference for High-Dimensional Generalized Linear Models with Knowledge Transfer', *Journal of the American Statistical Association* **0**(0), 1–12.

Li, Y., Ding, A. A. & Dy, J. (2017), Rate Optimal Estimation for High Dimensional Spatial Covariance Matrices, *in* 'Proceedings of the Ninth Asian Conference on Machine Learning', Vol. 77, PMLR, pp. 208–223.

Maqsood, M., Nazir, F., Khan, U., Aadil, F., Jamal, H., Mehmood, I. & Song, O.-y. (2019), 'Transfer Learning Assisted Classification and Detection of Alzheimer's Disease Stages Using 3D MRI Scans', *Sensors* **19**(11), 2645.

Minsker, S. & Wei, X. (2018), 'Estimation of the covariance structure of heavy-tailed distributions'.
URL: *http://arxiv.org/abs/1708.00502*

Petegrosso, R., Park, S., Hwang, T. H. & Kuang, R. (2017), 'Transfer learning across ontologies for phenome–genome association prediction', *Bioinformatics* **33**(4), 529–536.

Qiu, Y. & Chen, S. X. (2012), 'Test for bandedness of high-dimensional covariance matrices and bandwidth estimation', *The Annals of Statistics* **40**(3), 1285–1314.

Qiu, Y. & Chen, S. X. (2015), 'Bandwidth Selection for High-Dimensional Covariance Matrix Estimation', *Journal of the American Statistical Association* **110**(511), 1160–1174.

Raina, R., Ng, A. Y. & Koller, D. (2006), Constructing informative priors using transfer learning, *in* 'Proceedings of the 23rd International Conference on Machine Learning', ICML '06, Association for Computing Machinery, New York, NY, USA, pp. 713–720.

Rodrigues, P. L. C., Jutten, C. & Congedo, M. (2019), 'Riemannian Procrustes Analysis: Transfer Learning for Brain–Computer Interfaces', *IEEE Transactions on Biomedical Engineering* **66**(8), 2390–2401.

Tsybakov, A. B. (1998), 'Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes', *The Annals of Statistics* **26**(6), 2420–2469.

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Weiss, K., Khoshgoftaar, T. M. & Wang, D. (2016), 'A survey of transfer learning', *Journal of Big Data* **3**(1), 9.

Yi, F. & Zou, H. (2013), 'SURE-tuned tapering estimation of large covariance matrices', *Computational Statistics & Data Analysis* **58**, 339–351.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. & He, Q. (2021), 'A Comprehensive Survey on Transfer Learning', *Proceedings of the IEEE* **109**(1), 43–76.