Applied and Computational Harmonic Analysis 74 (2025) 101698

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

journal homepage: www.elsevier.com/locate/acha

Full Length Article

Data-driven optimal shrinkage of singular values under high-dimensional noise with separable covariance structure with application

Pei-Chun Su [a], Hau-Tieng Wu [a,b,*]

[a] Department of Mathematics, Duke University, Durham, NC, USA
[b] Department of Statistical Science, Duke University, Durham, NC, USA

ARTICLE INFO

Communicated by Radu Balan

Keywords:
Matrix denoising
Random matrix
High dimensional noise
Spike model
Separable covariance

ABSTRACT

We develop a data-driven optimal shrinkage algorithm, named *extended OptShrink* (eOptShrink), for matrix denoising with high-dimensional noise and a separable covariance structure. This noise is colored and dependent across samples. The algorithm leverages the asymptotic behavior of singular values and vectors of the noisy data's random matrix. Our theory includes the sticking property of non-outlier singular values, delocalization of weak signal singular vectors, and the spectral behavior of outlier singular values and vectors. We introduce three estimators: a novel rank estimator, an estimator for the spectral distribution of the pure noise matrix, and the optimal shrinker eOptShrink. Notably, eOptShrink does not require estimating the noise's separable covariance structure. We provide a theoretical guarantee for these estimators with a convergence rate. Through numerical simulations and comparisons with state-of-the-art optimal shrinkage algorithms, we demonstrate eOptShrink's application in extracting maternal and fetal electrocardiograms from single-channel trans-abdominal maternal electrocardiograms.

1. Introduction

We aim to denoise a $p \times n$ data matrix $\widetilde{S}$, comprised of $n \in \mathbb{N}$ noisy samples of dimension $p \in \mathbb{N}$. The data matrix is modeled as:

$$\widetilde{S} = S + Z = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i^\top + Z \in \mathbb{R}^{p \times n}, \tag{1}$$

where Z is a noise-only random matrix, potentially with a dependence structure that will be detailed later, S denotes a low-rank signal matrix with the singular value decomposition (SVD) $\sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i^\top$, where $r \geq 1$ is assumed to be small compared with p and n, $\mathbf{u}_i \in \mathbb{R}^p$ and $\mathbf{v}_i \in \mathbb{R}^n$ are left and right singular vectors respectively prescribing the signal, and $d_i > 0$ are the associated singular values describing signal strength that may depend on n. To simplify the discussion, we call the pair $\mathbf{u}_i$ and $\mathbf{v}_i$ the i-th signal and d_i the i-th signal strength hereafter. A commonly studied method for recovering S from $\widetilde{S}$ using SVD is known as *singular value shrinkage*. Initially mentioned in [26,46,33], this method involves selecting a nonlinear function $\varphi : [0, \infty) \to [0, \infty)$ to construct

* Corresponding author at: Department of Mathematics, Duke University, Durham, NC, USA.
E-mail address: hauwu@math.duke.edu (H.-T. Wu).

https://doi.org/10.1016/j.acha.2024.101698
Received 31 July 2023; Received in revised form 19 August 2024; Accepted 28 August 2024

Available online 12 September 2024

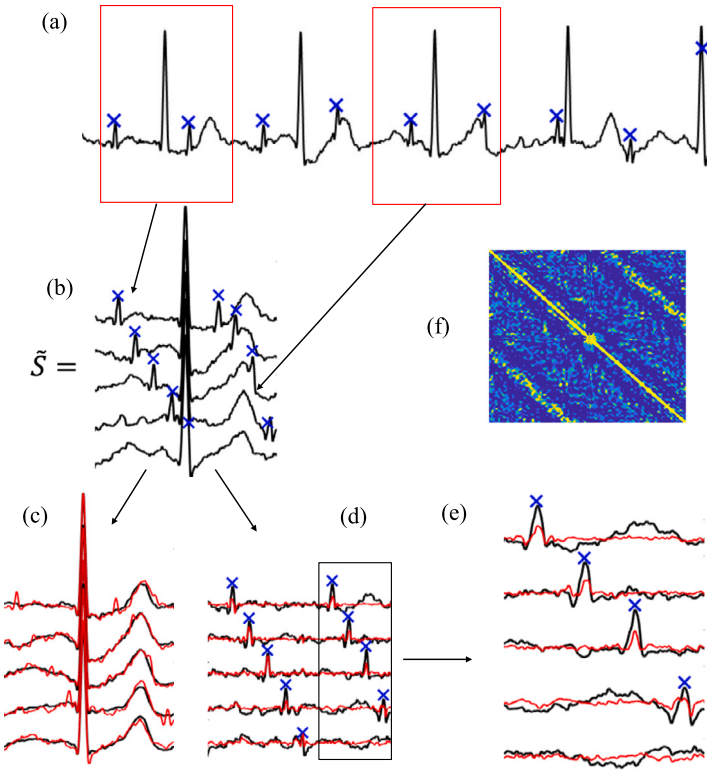

Fig. 1. An illustration of extracting fECG from the ta-mECG shown in (a), where the fetal R peaks labeled by experts are marked as blue crosses. (b) is the data matrix including the pieces truncated from the ta-mECG and aligned by the maternal R peaks. The results of TRAD and our eOptShrink (both with the operator norm) as the estimated mECG are shown as red and black curves shown in (c) respectively. By subtracting the estimated mECG from the ta-mECG, we obtain the estimated fECG shown in (d). The mECG estimation error could be visualized in the estimated fECG indicated by the black box, zoomed in (e). The covariance structure of the fECG as noise is shown in (f). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

$$\hat{S}_{\varphi} = \sum_{i=1}^{p\wedge n} \varphi(\widetilde{\lambda}_i)\widetilde{\xi}_i\widetilde{\zeta}_i^{\top} \tag{2}$$

as an estimate of S, where $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_{p\wedge n} \geq 0$ are the eigenvalues of $\widetilde{S}\widetilde{S}^{\top}$ and $\widetilde{S}^{\top}\widetilde{S}$ and $\{\widetilde{\xi}_i\}$ and $\{\widetilde{\zeta}_i\}$ are the left and right singular vectors of $\widetilde{S}$ respectively. φ is called a shrinker. Using a loss function $L_n : \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times n} \to \mathbb{R}_+$ to quantify the discrepancy between $\hat{S}_{\varphi}$ and S, the *optimal shrinker*, if exists, is defined as $\varphi^* := \arg\min_{\phi} \lim_{n\to\infty} L_n(\hat{S}_{\phi}, S)$. Common loss functions include the Frobenius norm and operator norm of $\hat{S}_{\varphi} - S$. This approach is referred to as *Optimal Shrinkage* (OS), as named in previous literature [23,31].

In this paper, we investigate the matrix denoising problem and develop the associated OS under the model (1) in the *high dimensional* setup, where $p = p(n)$ and $p/n \to \beta \in (0, \infty)$ as $n \to \infty$.[1] We begin by considering the white noise as a special example, where $Z = X$ in (1), with the entries of X being i.i.d. with zero mean, variance σ^2/n, $\sigma > 0$, and finite fourth moment. Asymptotically the empirical spectral distribution (ESD) of $ZZ^{\top}$ follows the Marchenko-Pastur (MP) law [45]. A phase transition occurs when the signal strength d_i exceeds a critical value, known as the *Baik-Ben Arous-Pèchè (BBP) phase transition* [9]. Due to these peculiar behaviors, the traditional SVD truncation scheme [33] needs to be modified to better recover S from $\widetilde{S}$. In [31], the optimal shrinker φ^* with various loss functions is derived using the closed form of the rightmost bulk edge λ_+. Here, φ^* is determined solely by σ and β. See an earlier work [53] as well. The OS approach under the white noise assumption (referred to as TRAD hereafter) has found wide applications, such as diffusion magnetic resonance imaging denoising [61], fetal electrocardiogram (fECG) extraction from the trans-abdominal maternal ECG (ta-mECG) [56], enhancing 3-D seismic data [5], ECG T-wave quality evaluation [57], otoacoustic emission signal denoising [44], stimulation artifact removal from intracranial electroencephalogram (EEG) [1], and cardiogenic artifact removal from EEG [16], among many others.

We illustrate the application of TRAD to the fECG extraction problem in Fig. 1. The ta-mECG recorded from the mother's abdomen during pregnancy is shown in Fig. 1(a). It is truncated into pieces marked by the red boxes and aligned based on the maternal R peaks. The fetal R peaks are labeled by blue crosses. Fig. 1(b) illustrates the associated data matrix $\widetilde{S}$, where the maternal ECG (mECG) is considered the signal and saved as matrix S, while the fECG and inevitable noise are jointly considered as the noise and saved as

[1] Note that in this paper we do not consider the related but different OS for the covariance structure.

matrix Z. The results of TRAD, represented by $\widehat{S}$, are shown in red curves in Fig. 1(c), representing the recovered mECG after denoise. In 1(d), $S-\widehat{S}$ is depicted, representing the recovered fECG. The decomposition of mECG and fECG is accomplished, allowing for clearer visualization of the fetal R peaks in Fig. 1(d). Noticeable poorly recovered segments, indicated by the black box, are magnified in Fig. 1(e). For further details, a literature review of relevant algorithms, and clinical applications, readers are referred to [56].

While TRAD has been successfully applied to various scenarios, the assumption of white noise is overly restrictive. Physiologically, the fECG (considered noise) typically deviates from white characteristics. Within a sample, the fECG exhibits dependencies due to electrophysiology, and across different samples, there is dependence due to the long-range nature of fECG recordings. Thus, in the fECG problem, noise manifests both intra-sample and inter-sample dependencies. Fig. 1(e) illustrates this, and Fig. 1(f) shows the nontrivial estimated covariance structure of the fECG. Another source of non-white noise arises from the application of different filters. Even if we assume the noise to be white, commonly employed filters can disrupt this white structure. Therefore, modifying the white noise model and TRAD to handle more complicated noise scenario is necessary.

A natural generalization is considering the high dimensional setting with a dependent noise structure in (1). This extends the traditional high dimensional spiked model [7,8,10]. It has been shown in [13] that if the ESD of singular values from Z converges to a non-random compactly supported probability measure, the top singular values of $\widetilde{S}$ and associated singular vectors, both left and right, are all biased from those of S. These biases converge to a closed form depending on the D-transform [13] and the Stieltjes transform [63] of the limiting singular value distribution of the noise matrix Z. Moreover, we also have the BBP transition. Under this setup, the optimal shrinker when L_n is the Frobenius norm loss is derived in [47] (See Proposition 2.6 below) under the following assumptions. First, the ESD of Z asymptotically converges to a non-random compactly supported probability measure μ_Z so that the derivative of the D-transform of μ_Z is $-\infty$ at the rightmost edge of the compact support. Second, a critical delocalization conjecture of singular vectors holds. The OS algorithm, named *OptShrink*, is provided in [47] by approximating the D-transform with respect to the truncated singular value distribution of $\widetilde{S}$, assuming knowledge of the rank. OptShrink is very general, but it is challenging to apply it directly to real-world problems without imposing some assumptions and knowing the rank. See [17] for an exploration of applying OptShrink to estimate diffusion magnetic resonance image.

Inspired by the OptShrink algorithm and driven by practical requirements and common dependence structures, this paper concentrates on noise exhibiting a separable covariance structure. [51]:

$$Z = A^{1/2} X B^{1/2}, \tag{3}$$

where X is a random matrix with independent entries and some moment conditions that will be specified later in Section 2.2, and A and B are $p\times p$ and $n\times n$ deterministic positive-definite matrices describing the colorness and dependence structure of the noise, respectively. This separable covariance structure has been studied and applied to problems like spatiotemporal analysis, wireless communication [27,34,32,25] and other applications. When $B=I_n$, it is known that the ESD of $ZZ^\top$ converges to the *deformed MP* law [45], and the largest eigenvalue distribution follows the Tracy-Widom law [58,59], commonly referred to as *edge universality*. Edge universality has been proved for $A=I_p$ [35,52] and general A [28,48,11,39,21,38] under various moment assumptions on the entries of X. For sample singular vectors, delocalization [38,49] and ergodicity [14] have been shown. When assuming a general separable covariance structure, the convergence of the ESD to a limiting law is shown in [51,62,66], edge universality and delocalization of eigenvectors are established in [21,65], and the local law (local estimates of the resolvents or Green's functions) of $ZZ^\top$ is recently proved in [64,65].

Building on the aforementioned theoretical results under various random matrix assumptions, significant efforts have been made to design matrix denoising algorithms for the separable covariance noise model. The special $B=I$ is discussed in [42] and the general B is discussed in [43]. They propose applying the whitening technique before applying TRAD; that is, performing TRAD on the whitened matrix $\widehat{A}^{-1/2}\widetilde{S}\widehat{B}^{-1/2}$, where $\widehat{A}$ and $\widehat{B}$ are the estimated A and B. This process yields $\widetilde{Q}$, from which S is estimated as $\widehat{A}^{1/2}\widetilde{Q}\widehat{B}^{1/2}$, effectively unwhitening the matrix. This approach works well when accurate estimates of A and B are obtainable, such as when both are diagonal. A fast numerical algorithm for λ_+ and the Stieltjes transform is available [40] when $Z=A^{1/2}XB^{1/2}$ and accurate estimates of the asymptotic spectral distribution of A and B are feasible. In [24], if the upper bound of r is known, d_i's are distinct and certain assumptions about the asymptotic ESD of Z hold, an algorithm named *ScreeNOT* (*Scree* reminds the commonly used Scree plot [15] in practice and *NOT* stands for Noise-adaptive Optimal Thresholding) is introduced. ScreeNOT evaluates the *optimal threshold* $\vartheta_{\mathrm{SN}}>0$ by optimizing the Frobenius norm of the recovered and clean data matrices using a *imputation* technique. The optimal threshold is applied to discard weak components close to the noise bulk edge, and approximates the underlying signal matrix, or denoise the matrix, by retaining the top $\widehat{r}_{\mathrm{SN}}$ singular values via $\widehat{S}_{\mathrm{SN}} := \sum_{i=1}^{\widehat{r}_{\mathrm{SN}}} \sqrt{\widetilde{\lambda}_i}\widetilde{\boldsymbol{\xi}}_i\widetilde{\boldsymbol{\eta}}_i^\top$ [24, (1.2)], where

$$\widehat{r}_{\mathrm{SN}} = |\{\widetilde{\lambda}_i | \widetilde{\lambda}_i > \vartheta_{\mathrm{SN}}\}|. \tag{4}$$

While the authors in [24] do not explicitly use $\widehat{r}_{\mathrm{SN}}$ as an estimated rank of the clean data matrix, it can be interpreted as such. The authors in [24] also provide a quantitatively interpretation of the Scree plot heuristic [15].

Inspired by the success and limitations of existing algorithms and their broad application in data science, we introduce a novel OS algorithm for matrix denoising tailored for real-world data within the separable covariance noise model (3). This algorithm, coined *extended OptShrink* (eOptShrink), builds upon OptShrink [47]. Our contributions are multifold. First, we extend the results in [12] and derive the asymptotic behavior of the outlier singular values and associated biased singular vectors of the noisy data matrix $\widetilde{S}=S+A^{1/2}XB^{1/2}$. This includes the BBP phase transition, the sticking properties of non-outlier singular values, and the delocalization of the non-outlier singular vectors, with an established convergence rate. These results align with those for deformed Wigner matrices [36,37], deformed rectangular matrices [13,19] and spiked covariance matrices [50,14,22,20], although they have

not yet been established for our setup to the best of our knowledge. Recall that delocalization is an essential condition needed for OptShrink [47]. Second, based on the developed theorems, we propose a fully data-driven OS algorithm leveraging these results. While OptShrink [47] typically requires spike rank information or an overestimated rank, our approach substitutes this requirement with a data-driven *effective rank* estimation from $\widetilde{S}$. Numerically, we show that the proposed approach precisely estimates the effective rank compared to other existing approaches, and the rank estimation has its own interest. Moreover, whereas OptShrink estimates the D-transform based on the "truncated" singular value distribution, our approach replaces this process with a novel method to estimate the spectral density distribution of Z. This method relies on accurately recovering the singular values of Z from those of $\widetilde{S}$ that are perturbed by signals. With these two components, we develop a more precise estimate of the Stieltjes transform and D-transform compared to existing imputation [24] and truncation [47] methods. Additionally, we extend OptShrink [47] to accommodate various loss functions beyond the Frobenius norm. Numerically, we evaluate the proposed algorithm by comparing it with existing algorithms [47,31,24] in simulated datasets and a real-world ta-mECG database. Overall, eOptShrink outperforms other algorithms.

The paper is structured as follows. Section 2 provides essential mathematical background, summarizes optimal shrinkage, and outlines the precise model assumptions. In Section 3, we present the main theoretical results. Section 4 introduces our proposed algorithm, eOptShrink. Section 5 conducts a series of numerical evaluations of eOptShrink, including comparisons with existing algorithms using both simulated and real databases. Additional numerical simulations, technical details, and proofs of theorems are included in the appendix.

NOTATION: For any random variable X, denote $\widetilde{X}$ as the perturbed X, and $\widehat{X}$ as the estimator of X. C denotes a generic positive constant, whose value may change from one line to another. For sequences $\{a_n\}$ and $\{b_n\}$ indexed by n, $a_n = \mathrm{O}(b_n)$ means that $|a_n| \le C|b_n|$ for some constant $C > 0$ as $n \to \infty$ and $a_n = \mathrm{o}(b_n)$ means that $|a_n| \le c_n|b_n|$ for some positive sequence $c_n \downarrow 0$ as $n \to \infty$. We also use $a_n \lesssim b_n$ if $a_n = \mathrm{O}(b_n)$, $|a_n| \ll |b_n|$ if $a_n = \mathrm{o}(b_n)$, and $a_n \asymp b_n$ if $a_n = \mathrm{O}(b_n)$ and $b_n = \mathrm{O}(a_n)$. For a matrix M, $\|M\|$ means its operator norm and we may abuse the notation and write $M = \mathrm{O}(a)$ if $\|M\| = \mathrm{O}(a)$. I is reserved for the identity matrix of any dimension. For $a, b \in \mathbb{R}$, $a \vee b$ and $a \wedge b$ mean the maximal and minimal value of a and b respectively. We reserve $E + i\eta$ to indicate a point in $\mathbb{C}^+$, the upper half plan of $\mathbb{C}$. See Tables S.1 and S.2 for a list of notations.

2. Preliminaries

2.1. Background on random matrix theory

In this subsection, we review the necessary results for our exploration. The Stieltjes transform of a probability measure ν on $\mathbb{R}$ is $m_\nu(z) := \int \frac{1}{\lambda - z} d\nu(\lambda)$, where $z \in \mathbb{C}^+$. The ESD of an $n \times n$ symmetric matrix H is defined as $\pi_H := \frac{1}{n}\sum_{i=1}^n \delta_{\ell_i}$, where $\ell_1 \ge \ell_2 \ge \cdots \ge \ell_n$ are the eigenvalues of H and δ means the Dirac delta measure. Denote the common eigenvalues of $ZZ^\top$ and $Z^\top Z$ (respectively $\widetilde{S}\widetilde{S}^\top$ and $\widetilde{S}^\top\widetilde{S}$) by $\{\lambda_i\}_{i=1}^{p\wedge n}$ (respectively $\{\widetilde{\lambda}_i\}_{i=1}^{p\wedge n}$). For $z \in \mathbb{C}^+$, denote the Green functions (or resolvents) of $ZZ^\top$ and $Z^\top Z$ as

$$\mathcal{G}_1(z) := (ZZ^\top - zI_p)^{-1} \quad \text{and} \quad \mathcal{G}_2(z) := (Z^\top Z - zI_n)^{-1} \tag{5}$$

respectively. The Stieltjes transforms of ESDs of $ZZ^\top$ and $Z^\top Z$ are formulated respectively by the Green functions $\mathcal{G}_1(z)$ and $\mathcal{G}_2(z)$ via

$$m_1(z) := \frac{1}{p}\mathrm{Tr}\mathcal{G}_1(z) \quad \text{and} \quad m_2(z) := \frac{1}{n}\mathrm{Tr}\mathcal{G}_2(z). \tag{6}$$

Denote $\beta_n := \frac{p}{n}$. By a direct calculation, we have the relationship $m_2(z) = -\frac{1-\beta_n}{z} + \beta_n m_1(z)$. Similarly, we denote the Green functions of $\widetilde{S}\widetilde{S}^\top$ and $\widetilde{S}^\top\widetilde{S}$ as

$$\widetilde{\mathcal{G}}_1(z) := (\widetilde{S}\widetilde{S}^\top - zI_p)^{-1} \quad \text{and} \quad \widetilde{\mathcal{G}}_2(z) := (\widetilde{S}^\top\widetilde{S} - zI_n)^{-1} \tag{7}$$

respectively and the Stieltjes transforms of ESDs of $\widetilde{S}\widetilde{S}^\top$ and $\widetilde{S}^\top\widetilde{S}$ as

$$\widetilde{m}_1(z) := \frac{1}{p}\mathrm{Tr}\widetilde{\mathcal{G}}_1(z) \quad \text{and} \quad \widetilde{m}_2(z) := \frac{1}{n}\mathrm{Tr}\widetilde{\mathcal{G}}_2(z). \tag{8}$$

Next, we summarize existing knowledge about the behavior of $m_1(z)$ and $m_2(z)$ under mild assumptions. Assume that as $n \to \infty$, $\beta_n \to \beta \in (0, \infty)$ and π_A and π_B converge weakly to deterministic probability distributions $\rho_{A\infty} \ne \delta_0$ and $\rho_{B\infty} \ne \delta_0$. Also assume that the entries of X are independent with mean 0, the same variance and finite fourth moments. The first result is from [66, Chapter 4] (see also [18, Proposition 1.1] for a summary). It states that as $n \to \infty$, almost surely $\pi_{ZZ^\top}$ and $\pi_{Z^\top Z}$ converge to deterministic distributions, denoted as $\rho_{1\infty}$ and $\rho_{2\infty}$, respectively, with desirable properties. Without loss of generality, we detail this result assuming $\tau \le \beta_n \le 1$ for all n and $\beta_n \to \beta$, where $\tau \in (0, 1)$. The case $1 \le \beta_n \le \tau^{-1}$ holds similarly by taking a transpose. This result comes from analyzing the Stieltjes transforms of available ESDs. For any $z \in \mathbb{C}^+$, denote $(\mathsf{M}_{1\infty}(z), \mathsf{M}_{2\infty}(z)) \in \mathbb{C}^+ \times \mathbb{C}^+$ as the unique solution to the following system of self-consistent equations [66, Chapter 4]

$$\mathsf{M}_{1\infty}(z) = \beta \int \frac{x}{-z[1 + x\mathsf{M}_{2\infty}(z)]} \rho_{A\infty}(dx) \quad \text{and} \quad \mathsf{M}_{2\infty}(z) = \int \frac{x}{-z[1 + x\mathsf{M}_{1\infty}(z)]} \rho_{B\infty}(dx). \tag{9}$$

Instead of working with these two equations, it is sometimes more convenient to consider a function f_∞ on $\mathbb{C}^+ \times \mathbb{C}^+$ defined by

$$f_\infty(z,m) := -m + \int \frac{x}{-z + x\beta \int \frac{t}{1+tm}\rho_{A\infty}(dt)} \rho_{B\infty}(dx), \tag{10}$$

so that $\mathsf{M}_{2\infty}(z)$ is the unique solution to $f_\infty(z,m)=0$; that is, $f_\infty(z,\mathsf{M}_{2\infty}(z))=0$ for $z\in\mathbb{C}^+$. With the above quantities, we define

$$\begin{aligned} m_{1\infty}(z) &:= \int \frac{1}{-z[1+x\mathsf{M}_{2\infty}(z)]}\rho_{A\infty}(dx) \\ m_{2\infty}(z) &:= \int \frac{1}{-z[1+x\mathsf{M}_{1\infty}(z)]}\rho_{B\infty}(dx), \end{aligned} \tag{11}$$

where $m_{1\infty}$ is the Stieltjes transform of a deterministic probability measure $\mu_{1\infty}$ on $\mathbb{R}$ with support in $[0,\infty)$ and $m_{2\infty}$ is the Stieltjes transform of a deterministic probability measure $\mu_{2\infty}=\beta\mu_{1\infty}+(1-\beta)\delta_0$ on $\mathbb{R}$ with support in $[0,\infty)$. For all $E\in\mathbb{R}\backslash\{0\}$, $\lim_{\eta\downarrow 0} m_{1\infty}(E+i\eta)$ and $\lim_{\eta\downarrow 0} m_{2\infty}(E+i\eta)$ exist. Denote

$$m_{1\infty}(E) := \lim_{\eta\downarrow 0} m_{1\infty}(E+i\eta) \ \text{ and } \ m_{2\infty}(E) := \lim_{\eta\downarrow 0} m_{2\infty}(E+i\eta). \tag{12}$$

It has been shown [18, Theorem 3.1] that Im $m_{1\infty}(E)$ and Im $m_{2\infty}(E)$ are continuous on $E\in\mathbb{R}\backslash\{0\}$, and the probability measure $\mu_{1\infty}$ has a continuous density function $\rho_{1\infty}$ on $E\in(0,\infty)$. Furthermore, $\rho_{1\infty}$ is analytic for $E\in(0,\infty)$ where $\rho_{1\infty}(E)>0$ [18, Theorem 3.2]. Since $\mu_{2\infty}=\beta\mu_{1\infty}+(1-\beta)\delta_0$, we similarly have the continuous density function $\rho_{2\infty}$ on $\mathbb{R}\backslash\{0\}$ associated with the probability measure $\mu_{2\infty}$. See [18, Section 2] for other properties of $\mu_{1\infty}$. We also know that as $n\to\infty$, almost surely $\pi_{ZZ^\top}$ and $\pi_{Z^\top Z}$ converge weakly to $\rho_{1\infty}(x)dx$ and $\rho_{2\infty}(x)dx$, respectively [66, Chapter 4] (see also [18, Proposition 1.1]). Moreover, edge universality and the averaged local law have been established [65].

Several properties of $\mathsf{M}_{1\infty}$ and $\mathsf{M}_{2\infty}$ have been established as well. $\mathsf{M}_{1\infty}$ ($\mathsf{M}_{2\infty}$ respectively) is the Stieltjes transform of a Radon positive measure $q_{1\infty}$ ($q_{2\infty}$ respectively) on $\mathbb{R}$ with support in $[0,\infty)$ [18, Proposition 1.2]. For all $E\in\mathbb{R}\backslash\{0\}$, $\lim_{\eta\downarrow 0}\mathsf{M}_{1\infty}(E+i\eta)$ and $\lim_{\eta\downarrow 0}\mathsf{M}_{2\infty}(E+i\eta)$ exist. Denote

$$\mathsf{M}_{1\infty}(E) := \lim_{\eta\downarrow 0}\mathsf{M}_{1\infty}(E+i\eta) \ \text{ and } \ \mathsf{M}_{2\infty}(E) := \lim_{\eta\downarrow 0}\mathsf{M}_{2\infty}(E+i\eta).$$

Im $\mathsf{M}_{1\infty}(E)$ and Im $\mathsf{M}_{1\infty}(E)$ are continuous on $E\in\mathbb{R}\backslash\{0\}$ [18, Theorem 3.1]. The measures $q_{1\infty}$ and $q_{2\infty}$ have continuous derivatives on $\mathbb{R}\backslash\{0\}$, denoted as $\wp_{1\infty}$ and $\wp_{2\infty}$ respectively on $E\in\mathbb{R}\backslash\{0\}$, and we have [18, Theorem 3.1] $\mathtt{supp}(\rho_{1\infty})\cap(0,\infty)=\mathtt{supp}(\wp_{1\infty})\cap(0,\infty)=\mathtt{supp}(\wp_{2\infty})\cap(0,\infty)$.

Parallel results for finite n have been established [65, Theorem 2.3, Theorem 2.4, Lemma 2.5], using tools from [38,2,3]. Our algorithm and theory are based on these results. For any $z\in\mathbb{C}^+$, define $(\mathsf{M}_{1c}(z),\mathsf{M}_{2c}(z))\in\mathbb{C}^+\times\mathbb{C}^+$ as the unique solution to the system of self-consistent equations [65, Theorem 2.3]

$$\begin{aligned} \mathsf{M}_{1c}(z) &:= \beta_n \int \frac{x}{-z[1+x\mathsf{M}_{2c}(z)]}\pi_A(dx) \\ \mathsf{M}_{2c}(z) &:= \int \frac{x}{-z[1+x\mathsf{M}_{1c}(z)]}\pi_B(dx). \end{aligned}$$

Similarly, $\mathsf{M}_{2c}(z)$ is the unique solution to the function f defined on $\mathbb{C}^+\times\mathbb{C}^+$ by

$$f(z,m) := -m + \int \frac{x}{-z + x\beta_n \int \frac{t}{1+tm}\pi_A(dt)} \pi_B(dx)$$

for $z\in\mathbb{C}^+$. Define

$$\begin{aligned} m_{1c}(z) &:= \int \frac{1}{-z[1+x\mathsf{M}_{2c}(z)]}\pi_A(dx) \\ m_{2c}(z) &:= \int \frac{1}{-z[1+x\mathsf{M}_{1c}(z)]}\pi_B(dx), \end{aligned} \tag{13}$$

which are the Stieltjes transform of probability measures μ_{1c} and μ_{2c} with continuous density functions ρ_{1c} and ρ_{2c} [65, Theorem 2.3]. $\mathsf{M}_{1c}(z)$ and $\mathsf{M}_{2c}(z)$ are also the Stieltjes transforms of Radon positive measures with densities $\wp_{1c}$ and $\wp_{2c}$, having the same support as ρ_{1c} on $\mathbb{R}\backslash\{0\}$. It has been known that when $n\to\infty$, $m_{1c}(z)-m_1(z)$ and $m_{2c}(z)-m_2(z)$ converge to zero uniformly over properly chosen z [22, Theorem S.3.9(b)].

Next, we summarize relevant results for ρ_{1c}, ρ_{2c}, $\wp_{1c}$ and $\wp_{2c}$ when π_A and π_B are compactly supported, under mild assumptions, in the following lemma [65, Lemma 2.5].

Lemma 2.1. *Let A and B be deterministic symmetric matrices with eigenvalues $\sigma_1^a\geq\sigma_2^a\geq\ldots\geq\sigma_p^a$ and $\sigma_1^b\geq\sigma_2^b\geq\ldots\geq\sigma_n^b$ respectively and satisfy $\sigma_1^a\vee\sigma_1^b\leq\tau^{-1}$ and $\pi_A([0,\tau])\vee\pi_B([0,\tau])\leq 1-\tau$ for a small $\tau>0$. Then, $\mathtt{supp}(\rho_{1c})$ is compact with*

$$supp(\wp_{1c}) = supp(\wp_{2c}) = supp(\rho_{1c}) = \bigcup_{k=1}^{L}[e_{2k}, e_{2k-1}], \tag{14}$$

where $e_1 > e_2 > \ldots > e_{2L}$. *Moreover,* $(x, m) = (e_k, \mathsf{M}_{2c}(e_k))$ *are the real solutions to the equations*

$$f(x,m) = 0 \quad \textit{and} \quad \frac{\partial f}{\partial m}(x,m) = 0\,.$$

Finally, e_1 *is bounded,* $\mathsf{M}_{1c}(e_1) \in (-(\sigma_1^b)^{-1}, 0)$ *and* $\mathsf{M}_{2c}(e_1) \in (-(\sigma_1^a)^{-1}, 0)$.

Note that this result parallels the findings in [18, Proposition 3.3, Proposition 3.4, Theorem 3.3], which state that if $\mathtt{supp}(\rho_{A\infty}) \cap (0,\infty)$ and $\mathtt{supp}(\rho_{B\infty}) \cap (0,\infty)$ consist of K and $\tilde{K}$ connected components respectively, then $\mathtt{supp}(\rho_{1\infty}) \cap (0,\infty)$ consists of L connected components, where $L \leq K\tilde{K}$ depends only on $\rho_{A\infty}$ and $\rho_{B\infty}$. Additionally, $\mathtt{supp}(\rho_{A\infty})$ and $\mathtt{supp}(\rho_{B\infty})$ are compact if and only if $\mathtt{supp}(\rho_{1\infty})$ is compact, with $\mathtt{supp}(\rho_{1\infty}) = \bigcup_{k=1}^{L}[e_{2k,\infty}, e_{2k-1,\infty}]$, where $e_{1,\infty} > e_{2,\infty} > \ldots > e_{2L,\infty}$. Furthermore, $(x,m) = (e_k, \mathsf{M}_{2\infty}(e_k))$ are the real solutions to the equations $f_\infty(x,m) = 0$ and $\frac{\partial f_\infty}{\partial m}(x,m) = 0$. Finally, $e_{1,\infty}$ is bounded, and $\mathsf{M}_{1\infty}(e_{1,\infty}) \in (-(\max \mathtt{supp}(\rho_{B\infty}))^{-1}, 0)$ and $\mathsf{M}_{2\infty}(e_{1,\infty}) \in (-(\max \mathtt{supp}(\rho_{A\infty}))^{-1}, 0)$.

From now on, we will carry out our analysis following [38,14,65,22]. We refer to e_k as spectral edges or bulk edges. In this paper, we focus on the rightmost edge

$$\lambda_+ := e_1\,.$$

$[e_{2L}, e_1]$ can be viewed as the "spectral spreading" of the noise, and intuitively, the signal should be sufficiently strong compared to λ_+ "in some way" so that the signal can be observed. This "strength" will be precisely described below.

For $z \in \mathbb{C}^+$, denote the D-transform of ρ_{1c} as

$$\mathcal{T}(z) := z m_{1c}(z) m_{2c}(z)\,. \tag{15}$$

The D-transform is a key tool in designing our algorithm, and here are some of its basic properties. Similar to the asymptotic case in (12), m_{1c} and m_{2c} can be extended from $\mathbb{C}^+$ to $x > \lambda_+$ via:

$$m_{1c}(x) = \int_0^{\lambda_+} \frac{\rho_{1c}(t)}{t-x}dt \quad \text{and} \quad m_{2c}(x) = \int_0^{\lambda_+} \frac{\rho_{2c}(t)}{t-x}dt\,. \tag{16}$$

Hence, for $x > \lambda_+$, $m_{1c}(x)$, $m_{2c}(x)$ and $\mathcal{T}(x)$ are well-defined. Moreover, by direct calculation, both $m_{1c}(x)$ and $m_{2c}(x)$ are negative and monotonically increasing over $x > \lambda_+$, making $\mathcal{T}(x)$ monotonically decreasing over $x > \lambda_+$. Hence, $m_{1c}(x)$, $m_{2c}(x)$ and $\mathcal{T}(x)$ are invertible over $x > \lambda_+$. This is a crucial property for studying signal deformation under high dimensional noise. See [13, Section 2.5] for more discussion on the D-transform. Denote the j-th *classical location* of the probability density ρ_{2c}, where $j = 1, \ldots, n$, as:

$$\gamma_j := \sup_x \left\{ \int_x^{+\infty} \rho_{2c}(x)dx > \frac{j-1}{n} \right\}. \tag{17}$$

In particular, $\gamma_1 = \lambda_+$.

2.2. Model assumption

In this subsection, we impose assumptions on the model (1). We follow [65,20] and consider the following bounded support condition to simplify the following discussion.

Definition 2.2. (Bounded support condition). A random matrix $X = [x_{ij}]_{\substack{i=1,\ldots,p \\ j=1,\ldots,n}}$ is said to have a bounded support $\phi_n > 0$ if $\max_{i,j} |x_{i,j}| \leq \phi_n$, where $\phi_n = n^{2/a-1/2}$ for some constant $a > 4$ is a deterministic parameter.

The bounded support assumption is introduced to simplify the discussion and can be easily removed. Recall that for a random matrix X whose entries are independent random variables fulfilling $\mathbb{E}x_{i,j} = 0$, $\mathbb{E}|x_{i,j}|^2 = n^{-1}$ and $\max_{i,j} \mathbb{E}|\sqrt{n}x_{i,j}|^a \leq C$ for $C > 0$ and $a > 4$, it can be truncated to a random matrix $\widetilde{X}$ with bounded support $\phi_n = n^{2/a-1/2+\varepsilon'}$, where $\varepsilon' > 0$ is a constant that can be arbitrarily small, fulfilling Assumption 2.3(i), so that $\mathbb{P}(\widetilde{X} = X) = 1 - O(n^{-a\varepsilon'})$. Refer to [6, Section 4.3.2] and [20, Corollary 3.19] for details. Consequently, all arguments and theoretical results with the bounded support condition can be extended to the setup without the bounded support condition with probability $1 - o(1)$.

Assumption 2.3. Fix a small constant $0 < \tau < 1$. We need the following assumptions for the model (1):

(i) (Assumption on x_{ij}). Suppose $X = [x_{ij}]_{\substack{i=1,\ldots,p \\ j=1,\ldots,n}} \in \mathbb{R}^{p\times n}$ has a bounded support ϕ_n and its entries are independent and satisfy

$$\max_{i,j} \left| \mathbb{E}x_{ij} \right| \leq n^{-2-\tau}\,, \ \max_{i,j} \left| \mathbb{E}|x_{ij}|^2 - n^{-1} \right| \leq n^{-2-\tau} \tag{18}$$

and

$$\max_{i,j} \mathbb{E}|\sqrt{n}x_{ij}|^4 \le C_4 \tag{19}$$

for a constant $C_4 > 0$.

(ii) (Assumptions on p/n). $p = p(n)$ satisfies

$$\tau < \frac{p}{n} < \tau^{-1}. \tag{20}$$

(iii) (Assumption on A and B). Assume A and B are deterministic symmetric matrices with eigendecompositions

$$A = Q^a \Sigma^a (Q^a)^\top \text{ and } B = Q^b \Sigma^b (Q^b)^\top \tag{21}$$

respectively, where $\Sigma^a = \mathrm{diag}(\sigma_1^a, \dots, \sigma_p^a)$, $\Sigma^b = \mathrm{diag}(\sigma_1^b, \dots, \sigma_n^b)$, $\sigma_1^a \ge \sigma_2^a \ge \dots \ge \sigma_p^a$, $\sigma_1^b \ge \sigma_2^b \ge \dots \ge \sigma_n^b$, $Q^a = (\mathbf{q}_1^a, \cdots, \mathbf{q}_p^a) \in O(p)$ and $Q^b = (\mathbf{q}_1^b, \cdots, \mathbf{q}_n^b) \in O(n)$. We assume that for all sufficiently large n,

$$1 + \mathsf{M}_{1c}(\lambda_+)\sigma_1^b \ge \tau \text{ and } 1 + \mathsf{M}_{2c}(\lambda_+)\sigma_1^a \ge \tau \tag{22}$$

and

$$\sigma_1^a \vee \sigma_1^b \le \tau^{-1} \text{ and } \pi_A([0,\tau]) \vee \pi_B([0,\tau]) \le 1 - \tau. \tag{23}$$

(iv) (Assumption on the signal strength). We assume

$$d_1 \ge d_2 \ge \cdots \ge d_r > 0 \tag{24}$$

for some $r \ge 1$ and $d_1 < \tau^{-1}$. We allow the signal strength d_i to depend on n but r is independent of n. Denote a fixed value $\alpha > 0$ as

$$\alpha := 1/\sqrt{\mathcal{T}(\lambda_+)}. \tag{25}$$

We allow the singular values d_k to depend on n under the condition that there exists an integer $1 \le r^+ \le r$, called the effective rank of $\widetilde{S}$, such that

$$d_k - \alpha > \phi_n + n^{-1/3} \text{ if and only if } 1 \le k \le r^+, \tag{26}$$

(v) (Assumption on the distribution of singular vectors). Let $G_{\mathbf{u}}^p \in \mathbb{R}^{p\times r}$ and $G_{\mathbf{v}}^n \in \mathbb{R}^{n\times r}$ be two independent matrices with i.i.d. entries distributed according to a fixed probability measure ν on $\mathbb{R}$ with mean zero and variance one, and satisfying the log-Sobolev inequality (see below). We assume that the left and right singular vectors, $\mathbf{u}_i \in \mathbb{R}^p$ and $\mathbf{v}_i \in \mathbb{R}^n$, are either

a. the i-th column of $\frac{1}{\sqrt{p}} G_{\mathbf{u}}^p$ and $\frac{1}{\sqrt{n}} G_{\mathbf{v}}^n$ respectively, or

b. the i-th columns obtained from the Gram-Schmidt (or QR factorization) of $G_{\mathbf{u}}^p$ and $G_{\mathbf{v}}^n$ respectively.

We comment on these assumptions. First, in *(iii)*, (22) guarantees a regular square-root behavior of the spectral densities ρ_{1c} and ρ_{2c} near λ_+ (See Lemma S.1.6), which is based on the behavior of $\mathsf{M}_{1c}(e_1)$ and $\mathsf{M}_{2c}(e_1)$ in Lemma 2.1, and rules out the existence of outlier eigenvalues of $ZZ^\top$. Therefore, it also makes sure that $\mathcal{T}(\lambda_+)$ is well-defined so as α in *(iv)*. (23) means that the spectra of A and B cannot concentrate at zero.

In *(iv)*, d_i fulfilling (26) is considered to be the singular value associated with a "sufficiently strong" signal. Recall that the signal is described by singular vectors in (1). Below, we will show that when a signal is sufficiently strong, it leads to a singular value outside the spectral bulk of Z. It is the reason that it is possible to study the signal. If d_i does not satisfy (26), they are considered "weak", associated with a "weak signal". In Theorem 3.2, we will show that weak signals will "stick" to the bulk edge of Z, which is related to the *BBP phase transition* [9]. This is why r^+ is called the *effective rank* of $\widetilde{S}$ in (26). See Figure S.1 for an illustration of the relationship of $\mathcal{T}$, α and λ_+. From now on, we call the top r^+ singular values of $\widetilde{S}$ outlier signal strengths or simply *outliers*, and the corresponding singular vectors the *outlier signals*, and call the $r^+ + 1, \dots, r$-th singular values of $\widetilde{S}$ non-outlier signal strengths or simply *non-outliers*, and the corresponding singular vectors the *non-outlier signals*. Similarly, we call the top r^+ eigenvalues of $\widetilde{S}\widetilde{S}^\top$ *outliers*, and call the $r^+ + 1, \dots, r$-th eigenvalues of $\widetilde{S}\widetilde{S}^\top$ *non-outliers*.

The random signal setup in *(v)* is the same as that in [13] and others. The log-Sobolev inequality (see [4, Section 2.3.2] for the definition) implies that entries of $\mathbf{u}_i$ and $\mathbf{v}_i$ have sub-Gaussian tails, which leads to the desired concentration property [12, Proposition 6.2] that is summarized in Lemma S.1.5. If $\mathbf{u}_i$ and $\mathbf{v}_i$ have deterministic entries and Z is random but has a bi-unitarily invariant distribution, then we are in the same setting as the second model of (v) [13, Remark 2.6]. However, we know that Z is bi-unitarily invariant if A and B are both unitarily invariant, which is not applicable to our case. How to handle the deterministic signal setup will be explored in our future work.

2.3. Optimal shrinkers for various loss functions

Recall the definition of asymptotic loss and optimal shrinker provided in [31, Definitions 1 and 2].

Definition 2.4 *(Asymptotic loss).* Let $\mathcal{L} := \{L_{p,n} | p, n \in \mathbb{N}\}$ be a family of loss functions, where each $L_{p,n} : M_{p\times n} \times M_{p\times n} \to [0,\infty)$ is a loss function obeying that $\widehat{S} \to L_{p,n}(S,\widehat{S})$ is continuous and $L_{p,n}(S,S) = 0$. Suppose $p = p(n)$ and $\lim_{n\to\infty} p(n)/n \to \beta > 0$. Let $\varphi : [0,\infty) \to [0,\infty)$ be a nonlinear function and consider $\widehat{S}_\varphi$ to be the singular shinkage estimate (2). When $\lim_{n\to\infty} L_{p,n}$ exists, we define the asymptotic loss of the shrinker φ with respect to $L_{p,n}$ with the signal $\mathbf{d} = (d_1, \cdots, d_r)$ as $L_\infty(\varphi|\mathbf{d}) = \lim_{n\to\infty} L_{p,n}(S, \widehat{S}_\varphi)$, where S is defined in (1).

Definition 2.5 *(Optimal shrinker).* Let L_∞ and φ be as defined in Definition 2.4. If a shrinker φ^* has an asymptotic loss that satisfies $L_\infty(\varphi^*|\mathbf{d}) \le L_\infty(\varphi|\mathbf{d})$ for any other shrinker φ, any $r \ge 1$, and any $\mathbf{d} \in \mathbb{R}^r$, then we say that φ^* is uniquely asymptotically admissible (or simply "optimal") for the loss family $\mathcal{L}$ and that class of shrinkers.

From now on, we denote the optimal shrinker of $\widetilde{\lambda}_i$ as

$$\varphi_i^* := \varphi^*(\widetilde{\lambda}_i) \tag{27}$$

for convenience. In [31, Sections IV. A and C], the optimal shrinkers under different loss functions were computed. When the loss function is the operator norm, the optimal shrinker was proved to be $\varphi_i^* = d_i$ when $p = n$ in [31, Section IV.B]. Later in [41, Lemma 5.1], it is shown that $\varphi_i^* = d_i\sqrt{\frac{a_{1,i}\wedge a_{2,i}}{a_{1,i}\vee a_{2,i}}}$ when $p \ne n$, where $a_{1,i} := \lim_{n\to\infty}\langle \mathbf{u}_i, \widetilde{\boldsymbol{\xi}}_i\rangle^2$ and $a_{2,i} := \lim_{n\to\infty}\langle \mathbf{v}_i, \widetilde{\boldsymbol{\zeta}}_i\rangle^2$. We list the results here for readers' convenience.

Proposition 2.6 *([31,41]). When $d_i \ge \alpha$, the optimal shrinker is $\varphi_i^* = d_i\sqrt{a_{1,i}a_{2,i}}$, $\varphi_i^* = d_i\sqrt{\frac{a_{1,i}\wedge a_{2,i}}{a_{1,i}\vee a_{2,i}}}$ and $\varphi_i^* = d_i\big(\sqrt{a_{1,i}a_{2,i}} - \sqrt{(1-a_{1,i})(1-a_{2,i})}\,\big)$ when the Frobenius norm, operator norm and nuclear norm are considered in the loss function respectively. When $d_i < \alpha$, for any loss function, we have $\varphi_i^* = 0$.*

With Proposition 2.6, if we can estimate d_i, $a_{1,i}$ and $a_{2,i}$ using the eigenstructure of the noisy matrix $\widetilde{S}\widetilde{S}^\top$, we could obtain the desired optimal shrinkers. For example, as shown in [13], if the ESD of Z converges almost surely weakly to a compactly supported probability measure, then when $d_i > \alpha$, we have that $d_i = \frac{1}{\sqrt{\mathcal{T}(\widetilde{\lambda}_i)}}$, $a_{1,i} = \frac{m_{1c}(\widetilde{\lambda}_i)}{d_i^2\mathcal{T}'(\widetilde{\lambda}_i)}$ and $a_{2,i} = \frac{m_{2c}(\widetilde{\lambda}_i)}{d_i^2\mathcal{T}'(\widetilde{\lambda}_i)}$, based on which the optimal shrinker with respect to the Frobenius norm is derived by replacing $d_i\sqrt{a_{1,i}a_{2,i}}$ with the corresponding values in [47]. Moreover, as shown in [31], when the noise is white with $\sigma > 0$, and X has independent entries with zero mean, unit variance, and finite fourth moment, if $d_i > \beta^{1/4}$, we have $\widetilde{\lambda}_i > \sigma(1+\sqrt{\beta})$ as $n \to \infty$. Denote $y_i = \widetilde{\lambda}_i/\sigma^2$. We have $d_i = \ell_i$, $a_{1,i} = \frac{\ell_i^4-\beta}{\ell_i^4+\beta\ell_i^2}$ and $a_{2,i} = \frac{\ell_i^4-\beta}{\ell_i^4+\ell_i^2}$, where $\ell_i := \frac{1}{\sqrt{2}}\sqrt{y_i-\beta-1+\sqrt{(y_i-\beta-1)^2-4\beta}}$. In this special case, when $\widetilde{\lambda}_i > \sigma(1+\sqrt{\beta})$, the optimal shrinker has the closed form $\varphi_i = \frac{\sigma}{y_i}\sqrt{(y_i^2-\beta-1)^2-4\beta}$, $\varphi_i = \sigma\ell_i$ and $\varphi_i = \frac{\sigma}{\ell_i^2 y_i}(\ell_i^4 - \beta - \sqrt{\beta}\ell_i y_i)$ respectively when the Frobenius norm, operator norm, and nuclear norm respectively are considered, and when $\widetilde{\lambda}_i \le \sigma(1+\sqrt{\beta})$, $\varphi_i^* = 0$ for any loss function. The rank can be decided by how many $\widetilde{\lambda}_i > \sigma(1+\sqrt{\beta})$. If σ is unknown, it is suggested in [30], among others, e.g., [53], to estimate σ by $\check{\sigma}(\widetilde{S}) := s_{med}/\sqrt{\mu_\beta}$, where s_{med} is a median singular value of $\widetilde{S}$ and μ_β is the median of the MP distribution. To extend OptShrink under our noise setup (3), we study the asymptotic behavior of singular values and vectors and their relation with d_i, $a_{1,i}$ and $a_{2,i}$ with a convergence rate.

3. Main results

In this section, we state the main theoretical results about the biased singular values and vectors, including the limiting behavior and associated convergence rate. These results form the foundation of the proposed eOptShrink algorithm and could be of its own interest. To simplify the presentation of our results and their proofs, we apply the notion of stochastic domination, which is a systematic framework to state results of the form "ξ is bounded by ζ with high probability up to a small power of n" [29].

Definition 3.1 *(Stochastic domination).* Let $\xi = \left(\xi^{(n)}(u) : n \in \mathbb{N}, u \in U^{(n)}\right)$ and $\zeta = \left(\zeta^{(n)}(u) : n \in \mathbb{N}, u \in U^{(n)}\right)$ be two families of non-negative random variables, where $U^{(n)}$ is a possibly n-dependent parameter set. We say ξ is stochastically dominated by ζ, uniformly in u, if for any fixed (small) $\epsilon > 0$ and (large) $D > 0$, $\sup_{u\in U^{(n)}} \mathbb{P}\left[\xi^{(n)}(u) > n^\epsilon \zeta^{(n)}(u)\right] \le n^{-D}$ for large enough $n \ge n_0(\epsilon, D)$, and we shall use the notation $\xi \prec \zeta$ or $\xi = O_\prec(\zeta)$. Throughout this paper, the stochastic domination will always be uniform in all parameters that are not explicitly fixed, such as matrix indices, and z that takes values in some compact set. Note that $n_0(\epsilon, D)$ may depend on quantities that are explicitly constant, such as τ in Assumption 2.3. Moreover, we say an event Ξ holds with high probability if for any constant $D > 0$, $\mathbb{P}(\Xi) \ge 1 - n^{-D}$, when n is sufficiently large.

We denote $\{\boldsymbol{\xi}_i\}_{i=1}^p$ and $\{\boldsymbol{\zeta}_i\}_{i=1}^n$ as the left and right singular vectors of the noise matrix Z, respectively, and $\{\widetilde{\boldsymbol{\xi}}_i\}_{i=1}^p$ and $\{\widetilde{\boldsymbol{\zeta}}_i\}_{i=1}^n$ as the left and right singular vectors of $\widetilde{S}$, respectively, which can be viewed as a perturbation of Z by adding S. For $x > \alpha$, denote $\theta(x)$, $a_1(x)$ and $a_2(x)$ by

$$\theta(x) := \mathcal{T}^{-1}(x^{-2}),\ a_1(x) = \frac{m_{1c}(\theta(x))}{x^2\mathcal{T}'(\theta(x))},\ a_2(x) = \frac{m_{2c}(\theta(x))}{x^2\mathcal{T}'(\theta(x))}. \tag{28}$$

Clearly, on (λ_+, ∞), $\mathcal{T}(x)$ is monotonically decreasing as $x \to \infty$, so $\theta(x)$ is monotonically increasing as $x \to \infty$. These terms are used to estimate the signal strength d_i and inner products of the clean and noisy left singular vectors $\langle \mathbf{u}_i, \widetilde{\boldsymbol{\xi}}_j \rangle$ and clean and noisy right singular vectors $\langle \mathbf{v}_i, \widetilde{\boldsymbol{\zeta}}_j \rangle$, which we will detail in the following theorems.

3.1. Results of singular values

We first state the results for singular values. Define

$$\mathbb{O}_+ = \{1, \cdots, r^+\} \ \text{ and } \ \Delta(d_i) := |d_i - \alpha|^{1/2}. \tag{29}$$

In other words, the outlier singular values are indexed by $\mathbb{O}_+$. The followings are our main theorems, and their proofs are postponed to the supplementary material. Part of the proof of the following theorems is motivated by [65] and [22], where the focus is the covariance matrix analysis. The main difference comes from the fact that in general, the covariance model in [65,22] cannot be directly decomposed into the summation of two independent matrices, like S and $Z = A^{1/2}XB^{1/2}$ in our model, so the results cannot be directly applied. We first state the location of the outlier eigenvalues and the first few non-outlier eigenvalues of $\widetilde{S}\widetilde{S}^\top$.

Theorem 3.2. *Suppose Assumption 2.3 holds. Then we have for $1 \le i \le r^+$,*

$$|\widetilde{\lambda}_i - \theta(d_i)| \prec \phi_n \Delta(d_i)^2 + n^{-1/2}\Delta(d_i). \tag{30}$$

Furthermore, for a fixed integer $\varpi > r^+$, we have for $r^+ + 1 \le i \le \varpi$,

$$|\widetilde{\lambda}_i - \lambda_+| \prec \phi_n^2 + n^{-2/3}. \tag{31}$$

The above theorem gives the absolute error bounds for the locations of the outlier eigenvalues and the top finite non-outlier eigenvalues of $\widetilde{S}\widetilde{S}^\top$ and implies the occurrence of the BBP transition [9]. When $\phi_n \le n^{-1/3}$, the right-hand side of (31) becomes $n^{-2/3}$; that is, the non-outliers associated with weak signals deviate from the bulk edge with the order of the Tracy-Widom law [60].

Next, we study the *eigenvalue sticking*, which states how the non-outlier eigenvalues of $\widetilde{S}\widetilde{S}^\top$ "stick" to the eigenvalues of $ZZ^\top$.

Theorem 3.3. *Suppose Assumption 2.3 holds and assume*

$$\alpha_+ := \min_{1 \le i \le r} \Delta(d_i)^2 \ge n^{\epsilon}(\phi_n + n^{-1/3}) \tag{32}$$

for a small positive constant $\epsilon < \frac{1}{2} - \frac{2}{a}$. Fix a small constant $c > 0$. We have

$$\left|\widetilde{\lambda}_{r^+ + i} - \lambda_i\right| \prec \frac{1}{n\alpha_+} + n^{-3/4} + i^{1/3}n^{-5/6} + n^{-1/2}\phi_n + i^{-2/3}n^{-1/3}\phi_n^2 \tag{33}$$

for all $1 \le i \le cn$. If either (a) $\mathbb{E}x_{ij}^3 = 0$ for all $1 \le i \le p$ and $1 \le j \le n$, or (b) either A or B is diagonal, we have a stronger bound

$$\left|\widetilde{\lambda}_{r^+ + i} - \lambda_i\right| \prec \frac{1}{n\alpha_+} \tag{34}$$

for all $1 \le i \le \tau n$.

Theorem 3.3 establishes the absolute error bound for the non-outlier eigenvalues of $\widetilde{S}\widetilde{S}^\top$ with respect to the eigenvalues of $ZZ^\top$. The parameter α_+ impacts how strongly non-outlier eigenvalues of $\widetilde{S}\widetilde{S}^\top$ stick to eigenvalues of $ZZ^\top$. When (32) holds for $r^+ + 1 \le i \le \varpi$, where $\varpi > r^+$ is a fixed integer, the bound in (33) is more precise than that in (31), particularly when $\phi_n \gg n^{-1/3}$. When (32) fails, for example, when $d_i = \alpha$ for some i, we can only control the behavior of $\widetilde{\lambda}_i$ for $i = r^+ + 1, \ldots, r$ using (31). When $\alpha_+ \gg n^{-1/3}$ and $\phi_n \le n^{-1/6}$, the bounds in (33) and (34) are much tighter than $n^{-2/3}$.

3.2. Results of singular vectors

Next, we discuss the singular vectors. We first show that when $i \notin \mathbb{O}_+$, the non-outlier singular vectors $\widetilde{\boldsymbol{\xi}}_i$ and $\widetilde{\boldsymbol{\zeta}}_i$ are roughly uniformly distributed and approximately orthogonal to the spaces spanned by $\{\mathbf{u}_j\}_{j=1}^r$ and $\{\mathbf{v}_j\}_{j=1}^r$, respectively. This *delocalization* is crucial for designing eOptShrink. When $i \in \mathbb{O}_+$ and the signal is strong but not too strong, a similar behavior is observed.

Theorem 3.4. *Suppose Assumption 2.3 holds. Denote $\eta_i := n^{-3/4} + n^{-5/6}i^{1/3} + n^{-1/2}\phi_n$ and $\varkappa_i := i^{2/3}n^{-2/3}$. For any sufficiently small constant $c > 0$ and $r^+ + 1 \le i \le cn$, we have for $j = 1, \ldots, r$,*

$$|\langle \mathbf{u}_j, \widetilde{\boldsymbol{\xi}}_i \rangle|^2 \vee |\langle \mathbf{v}_j, \widetilde{\boldsymbol{\zeta}}_i \rangle|^2 \prec \frac{n^{-1} + \phi_n^3 + \eta_i\sqrt{\varkappa_i}}{\Delta(d_j)^4 + \phi_n^2 + \varkappa_i}. \tag{35}$$

If either A or B is diagonal, the right hand side of (35) becomes $\frac{n^{-1}+\phi_n^3}{\Delta(d_j)^4+\phi_n^2+\varkappa_i}$. *Moreover, fix a constant* $\widetilde{\tau} \in (0, 1/9)$ *such that* $n^{\widetilde{\tau}}(\phi_n + n^{-1/3}) \to 0$ *as* $n \to \infty$. *If* $i \in \mathbb{O}_+$ *satisfies*

$$\Delta(d_i)^2 \le n^{\widetilde{\tau}}(\phi_n + n^{-1/3}), \tag{36}$$

we have for $j = 1, \dots, r$,

$$|\langle \mathbf{u}_j, \widetilde{\boldsymbol{\xi}}_i \rangle|^2 \vee |\langle \mathbf{v}_j, \widetilde{\boldsymbol{\zeta}}_i \rangle|^2 \prec n^{3\widetilde{\tau}} \left(\frac{n^{-1} + \phi_n^3 + \eta_i \sqrt{\varkappa_i}}{\Delta(d_j)^4 + \phi_n^2 + \varkappa_i} \right). \tag{37}$$

To appreciate this theorem, assume $\phi_n \le n^{-1/3}$ and d_j, $1 \le j \le r^+$, satisfies $d_j - \alpha \gtrsim 1$. Then, for all $r^+ + 1 \le i \le n^{1/4}$, $\widetilde{\boldsymbol{\xi}}_i$ and $\widetilde{\boldsymbol{\zeta}}_i$ are delocalized with $|\langle \mathbf{u}_j, \widetilde{\boldsymbol{\xi}}_i \rangle|^2 \vee |\langle \mathbf{v}_j, \widetilde{\boldsymbol{\zeta}}_i \rangle|^2 \prec n^{-1}$. When d_j is close to the threshold α, for a fixed finite i so that $\widetilde{\lambda}_i$ is near the edge λ_+, (35) gives $|\langle \mathbf{u}_j, \widetilde{\boldsymbol{\xi}}_i \rangle|^2 \vee |\langle \mathbf{v}_j, \widetilde{\boldsymbol{\zeta}}_i \rangle|^2 \prec \frac{n^{-1}}{|d_j - \alpha|^2 + n^{-2/3}}$; that is, the delocalization bound changes from the optimal order n^{-1} to $n^{-1/3}$ as d_j approaches α.

Next, we state the behavior of the outlier singular vectors. For any $\mathbf{A} \subset \mathbb{O}_+$, define

$$\nu_i(\mathbf{A}) := \begin{cases} \min_{j \notin \mathbf{A}} |d_j - d_i| & \text{if } i \in \mathbf{A} \\ \min_{j \in \mathbf{A}} |d_j - d_i| & \text{if } i \notin \mathbf{A} \end{cases} \tag{38}$$

and two projections,

$$\mathcal{P}_{\mathbf{A}} := \sum_{k \in \mathbf{A}} \widetilde{\boldsymbol{\xi}}_k \widetilde{\boldsymbol{\xi}}_k^\top \quad \text{and} \quad \mathcal{P}'_{\mathbf{A}} := \sum_{k \in \mathbf{A}} \widetilde{\boldsymbol{\zeta}}_k \widetilde{\boldsymbol{\zeta}}_k^\top. \tag{39}$$

Theorem 3.5. *Suppose Assumption 2.3 holds. Fix any* $\boldsymbol{A} \subset \mathbb{O}_+$, *we have*

$$\begin{aligned} \left| \langle \mathbf{u}_i, \mathcal{P}_{\boldsymbol{A}} \mathbf{u}_j \rangle - \delta_{ij} \mathbb{1}(i \in \boldsymbol{A}) a_1(d_i) \right| \vee \left| \langle \mathbf{v}_i, \mathcal{P}'_{\boldsymbol{A}} \mathbf{v}_j \rangle - \delta_{ij} \mathbb{1}(i \in \boldsymbol{A}) a_2(d_i) \right| \\ \prec Q(i, j, \boldsymbol{A}, n) \end{aligned} \tag{40}$$

for all $i, j = 1, \cdots, r$, *where*

$$\begin{aligned} Q(i,j,\boldsymbol{A},n) = {} & \mathbb{1}(i \in \boldsymbol{A}, j \notin \boldsymbol{A}) \Delta(d_i) \left[\sqrt{\frac{\phi_n^2}{\nu_j(\boldsymbol{A})}} + \frac{\psi_1(d_j)\Delta(d_j)}{\nu_j(\boldsymbol{A})} \right] \\ & + \mathbb{1}(i \notin \boldsymbol{A}, j \in \boldsymbol{A}) \Delta(d_j) \left[\sqrt{\frac{\phi_n^2}{\nu_i(\boldsymbol{A})}} + \frac{\psi_1(d_i)\Delta(d_i)}{\nu_i(\boldsymbol{A})} \right] \\ & + \sqrt{R(i,\boldsymbol{A}) R(j,\boldsymbol{A})}, \end{aligned}$$

$\psi_1(d_i) := \phi_n + \frac{n^{-1/2}}{\Delta(d_i)}$, *and*

$$R(i,\boldsymbol{A}) := \mathbb{1}(i \in \boldsymbol{A}) \psi_1(d_i) + \mathbb{1}(i \notin \boldsymbol{A}) \frac{\phi_n^2}{\nu_i(\boldsymbol{A})} + \frac{\psi_1^2(d_i)\Delta^2(d_i)}{\nu_i^2(\boldsymbol{A})}. \tag{41}$$

Theorem 3.5 establishes the relationship between the outlier singular vectors of $\widetilde{S}$ and S. To read this seemingly complicated bound, we take a look at all quantities one by one. The quantity $\nu_i(\mathbf{A})$ represents the spectral gap, with $\nu_i(\{i\})$ being the spectral gap of d_i. Also, $\nu_i(\{j\}) = \nu_j(\{i\})$ is the distance between d_i and d_j. The quantity $\psi_1(d_i)$ represents the "minimal required spectral gap" for recovering a singular vector even if it is strong. As $n \to \infty$, $\psi_1(d_i) \to 0$ implying that $R(i, \mathbf{A}) \to 0$ and hence $Q(i, j, \mathbf{A}, n) \to 0$, provided $\nu_i(\mathbf{A})$ and $\nu_j(\mathbf{A})$ are away from zero, since Assumption 2.3 ensures bounded singular values.

Next, take the left singular vectors $\mathbf{u}_i$ and $\widetilde{\boldsymbol{\xi}}_i$ as an example. Let $\mathbf{A} = \{i\}$. From (40), since $\langle \mathbf{u}_i, \mathcal{P}_{\{i\}} \mathbf{u}_i \rangle = |\langle \mathbf{u}_i, \widetilde{\boldsymbol{\xi}}_i \rangle|^2$ and $Q(i, i, \{i\}, n) = R(i, \{i\}) = \psi_1(d_i) + \frac{\psi_1^2(d_i)\Delta^2(d_i)}{\nu_i^2(\{i\})}$, we get:

$$\left| |\langle \mathbf{u}_i, \widetilde{\boldsymbol{\xi}}_i \rangle|^2 - a_1(d_i) \right| \vee \left| |\langle \mathbf{v}_i, \widetilde{\boldsymbol{\zeta}}_i \rangle|^2 - a_2(d_i) \right| \prec \psi_1(d_i) + \frac{\psi_1^2(d_i)\Delta^2(d_i)}{\nu_i^2(\{i\})}. \tag{42}$$

When d_i is well-separated from other signals in that $\nu_i(\{i\}) \gg \psi_1(d_i) = \phi_n + \frac{n^{-1/2}}{\Delta(d_i)}$, the error term converges to 0, meaning the biases of $\widetilde{\boldsymbol{\xi}}_i$ recovering $\mathbf{u}_i$ and $\widetilde{\boldsymbol{\zeta}}_i$ recovering $\mathbf{v}_i$ are determined by $\sqrt{a_1(d_i)}$. If d_i is sufficiently strong, a direct calculation using (15), (16),

Algorithm 1 eOptShrink.

Input: $\widetilde{S} = \sum_{i=1}^{p\wedge n} \sqrt{\widetilde{\lambda}_i}\widetilde{\xi}_i\widetilde{\zeta}_i^\top$, a constant $c = \min(\frac{1}{2.01}, \frac{1}{\log(\log n)})$, and the desired loss function.
Compute:
(i) $\widehat{\lambda}_+$ in (48) and $\widehat{r}^+$ in (49).
(ii) $\widehat{F}_{\mathrm{e}}(x)$ in (54) and $\widehat{d}_{\mathrm{e},j}$, $\widehat{a}_{\mathrm{e},1,j}$ and $\widehat{a}_{\mathrm{e},2,j}$ for $1 \le j \le \widehat{r}^+$ in (57).
(iii) $\widehat{\varphi}_{\mathrm{e},i}$ for $1 \le i \le \widehat{r}^+$ in (58) for the associated norm.
Output: The estimator of the clean data matrix $\widehat{S}_{\widehat{\varphi}} = \sum_{i=1}^{\widehat{r}^+} \widehat{\varphi}_{\mathrm{e},i}\widetilde{\xi}_i\widetilde{\zeta}_i^\top$.

and (30) shows $a_1(d_i) = \frac{m_{1c}(\theta(d_i))}{d_i^2\mathcal{T}'(\theta(d_i))} \approx 1$ and $a_2(d_i) \approx 1$, leading to $|\langle \mathbf{u}_i, \widetilde{\xi}_i\rangle|^2 \approx 1$ and $|\langle \mathbf{v}_i, \widetilde{\zeta}_i\rangle|^2 \approx 1$. Similarly, for $\mathbf{A} = \{k\}$, where $k \neq i$, we get $Q(i,i,\{k\},n) = R(i,\{k\}) = \frac{\phi_2^2}{\nu_i(\{k\})} + \frac{\psi_1^2(d_i)\Delta^2(d_i)}{\nu_i^2(\{k\})}$, leading to

$$|\langle \mathbf{u}_i, \widetilde{\xi}_k\rangle|^2 \vee |\langle \mathbf{v}_i, \widetilde{\zeta}_k\rangle|^2 \prec \frac{\phi_2^2}{\nu_i(\{k\})} + \frac{\psi_1^2(d_i)\Delta^2(d_i)}{\nu_i^2(\{k\})}, \tag{43}$$

which is small when $\nu_i(\{k\}) = |d_i - d_k| \gg \psi_1(d_i) = \phi_n + \frac{n^{-1/2}}{\Delta(d_i)}$; that is, $\widetilde{\xi}_k$ and $\mathbf{u}_i$, as well as $\widetilde{\zeta}_k$ and $\mathbf{v}_i$, are almost orthogonal. In the last example, if there are two close consecutive singular values $d_i \ge d_{i+1}$, but away from others; that is, $\nu_i(\{i\}) \ll \psi_1(d_i)$ and $\nu_i(\{i,i+1\}) \gg \psi_1(d_i)$, recovering $\mathbf{u}_i$ and $\mathbf{u}_{i+1}$ is challenging. For $\mathbf{A} := \{i, i+1\}$, we have $Q(i,i,\mathbf{A},n) = R(i,\mathbf{A}) = \psi_1(d_i) + \frac{\psi_1^2(d_i)\Delta^2(d_i)}{\nu_i^2(\mathbf{A})}$, leading to

$$\left|\langle \mathbf{u}_i, \mathcal{P}_{\mathbf{A}}\mathbf{u}_i\rangle - a_1(d_i)\right| \vee \left|\langle \mathbf{v}_i, \mathcal{P}'_{\mathbf{A}}\mathbf{v}_i\rangle - a_2(d_i)\right| \prec \psi_1(d_i) + \frac{\psi_1^2(d_i)\Delta^2(d_i)}{\nu_i^2(\mathbf{A})}, \tag{44}$$

and the error term converges to 0. This means $\mathbf{u}_i$ can be well recovered by a vector in the space spanned by $\widetilde{\xi}_i$ and $\widetilde{\xi}_{i+1}$.

Finally, compare (42) in Theorem 3.4 and (37) when $i \in \mathbb{O}_+$ and $\Delta(d_i)^2 = d_i - \alpha \le n^{\widetilde{\tau}}(\phi_n + n^{-1/3})$. To simplify the discussion, assume d_i is well-separated from other signals in that $\nu_i(\{i\}) \gg \psi_1(d_i)$, and let $\phi_n \lesssim n^{-1/3}$. With the lower bound of $d_i - \alpha$ in (26) when $i \in \mathbb{O}_+$, we get $n^{-1/3} < \Delta(d_i)^2 \le n^{-1/3+\widetilde{\tau}}$. Without loss of generality, assume $\Delta(d_i)^2 = n^{-1/3+\widetilde{\tau}}$ and compare the error bounds in (42) and (37) for $\widetilde{\tau} \in (0, 1/9)$. These conditions give $\psi_1(d_i) \asymp \phi_n + n^{-1/3-\widetilde{\tau}/2}$. Using the definitions of $a_1(x)$ and $a_2(x)$ in (28) and approximations of integral transforms in (S.1.8), (S.1.11), and (S.1.12), we find $|a_1(d_i)| \asymp |a_2(d_i)| \asymp n^{-1/3+\widetilde{\tau}}$. Above approximations of $\Delta(d_i)$, $\psi(d_i)$, $a_1(d_i)$, and $a_2(d_i)$ and (42) give us the following bound $|\langle \mathbf{u}_i, \widetilde{\xi}_i\rangle|^2 \vee |\langle \mathbf{v}_i, \widetilde{\zeta}_i\rangle|^2 \prec n^{-1/3+\tau} + \phi_n$. Moreover, together with $i = j \in \mathbb{O}_+$, (37) gives us $|\langle \mathbf{u}_i, \widetilde{\xi}_i\rangle|^2 \vee |\langle \mathbf{v}_i, \widetilde{\zeta}_i\rangle|^2 \prec n^{-1/3+\widetilde{\tau}}$. Therefore, for $i \in \mathbb{O}_+$, if the signal is strong but not too strong, (37) provides a smaller absolute error bound compared to (42).

4. Proposed eOptShrink algorithm

We are ready to introduce a data-driven algorithm to estimate the optimal shrinker φ, called *extended OptShrink* (eOptShrink). eOptShrink involves three main steps. Using the delocalization and bias estimates of singular vectors in Theorems 3.4 and 3.5 and the sticking result from Theorem 3.3, we show that under mild conditions, we can accurately estimate λ_k for the top finite k eigenvalues of $ZZ^\top$ when n is sufficiently large and r^+ is known, leading to a more precise estimate of $\pi_{ZZ^\top}$. In practice r and r^+ are unknown, and based on the established theory, estimating r can be challenging. However, we show that r^+ can be accurately estimated by approximating λ_+. The pseudocode of eOptShrink is summarized in Algorithm 1. The Matlab implementation is available at https://github.com/PeiChunSu/eOptShrink. Below, we detail the algorithm and its theoretical support, along with an asymptotic convergence rate.

4.1. Existing imputation approach

We review an imputation scheme proposed in [24]. Considering the square root behavior $\rho_{1c}(x) \asymp (\lambda_+ - x)^{1/2}$ as $x \to \lambda_+$ for $x \in \mathbb{R}$ when $Z = A^{1/2}X$ [55], we have, for a fixed large integer ϖ and sufficiently large p and n,

$$\frac{\ell - 1}{p} = \int_{\gamma_\ell}^{\lambda_+} \rho_{1c}(z)dz = C' \int_{\gamma_\ell}^{\lambda_+} (\lambda_+ - z)^{1/2}dz = \frac{2C'}{3}(\lambda_+ - \gamma_\ell)^{3/2} \tag{45}$$

for $1 \le \ell \le \varpi$ and $C' > 0$, where γ_ℓ is the classical location defined in (17). Since γ_ℓ can be approximated by λ_ℓ (see Lemma S.1.10 for details), this leads to an estimate of the distance between the j-th and ℓ-th eigenvalues, where $1 \le j, \ell \le \varpi$,

$$\lambda_\ell - \lambda_j \approx C'' \left[\left(\frac{j-1}{p}\right)^{2/3} - \left(\frac{\ell - 1}{p}\right)^{2/3}\right], \tag{46}$$

for some unknown $C'' > 0$. In [24], by fixing an integer k so that $r \le k$ and $2k + 1 < \varpi$ for a large constant ϖ, C'' is estimated by

$$\check{C}'' := \frac{\widetilde{\lambda}_{k+1} - \widetilde{\lambda}_{2k+1}}{(2k/p)^{2/3} - (k/p)^{2/3}}$$

and the j-th eigenvalue, where $j = 1, \dots, k$, as a missing value is imputed by

$$\check{\lambda}_j := \widetilde{\lambda}_{k+1} + \frac{1 - (\frac{j-1}{k})^{2/3}}{2^{2/3} - 1}(\widetilde{\lambda}_{k+1} - \widetilde{\lambda}_{2k+1}),$$

where $k = 4r$ is suggested in [24, p33]. The cumulative distribution function (CDF) of $\pi_{ZZ^\top}$ is estimated by

$$\widehat{F}_{\texttt{imp}}(x) := \frac{1}{p}\sum_{j=k+1}^{p} \mathbb{1}(\widetilde{\lambda}_j \le x) + \frac{1}{p}\sum_{j=1}^{k} \mathbb{1}(\check{\lambda}_j \le x). \tag{47}$$

With $\widehat{F}_{\texttt{imp}}$, the matrix denoising algorithm, ScreeNOT, is given in (4).

4.2. Proposed data-driven optimal shrinker, eOptShrink

4.2.1. Step 1: estimate r^+

We first estimate λ_+ and use it to estimate r^+. Based on the discussion for (46) and the sticking behavior in (33), we modify the estimator $\check{C}''$ for the constant C'' by constructing

$$\widehat{C} := \frac{\widetilde{\lambda}_{k+r^++1} - \widetilde{\lambda}_{2k+r^++1}}{(2k/p)^{2/3} - (k/p)^{2/3}}.$$

Since r^+ is unknown but fixed, set $k = \lfloor n^c \rfloor \gg r^+$, where $\lfloor\cdot\rfloor$ is the floor function, for a small fixed constant $c > 0$, and construct an estimator of λ_+ as

$$\widehat{\lambda}_+ := \widetilde{\lambda}_{\lfloor n^c \rfloor + 1} + \frac{1}{2^{2/3} - 1}\left(\widetilde{\lambda}_{\lfloor n^c \rfloor + 1} - \widetilde{\lambda}_{2\lfloor n^c \rfloor + 1}\right). \tag{48}$$

Then we estimate r^+ based on Theorem 3.2 and set

$$\widehat{r}^+ = \left|\{\widetilde{\lambda}_i | \widetilde{\lambda}_i / \widehat{\lambda}_+ - 1 > n^{-1/3}\}\right|. \tag{49}$$

The following theorems guarantee the performance of $\widehat{\lambda}_+$ and $\widehat{r}^+$.

Theorem 4.1. *Suppose Assumption 2.3 and (32) hold and $c \in (0,1)$. We have*

$$|\widehat{\lambda}_+ - \lambda_+| \prec \frac{1}{n\alpha^+} + n^{-\min\{\frac{2}{3}+\frac{c}{3}, \frac{3}{4}, \frac{5}{6}-\frac{c}{3}, \frac{4}{3}-\frac{4c}{3}\}} + n^{-1/2}\phi_n + n^{-1/3-2c/3}\phi_n^2. \tag{50}$$

Theorem 4.2. *Suppose Assumption 2.3 and (32) hold true for some ε such that $n^\varepsilon(\phi_n + n^{-1/3}) > n^{-1/6}$. Denote the event $\Xi(r^+) := \{\widehat{r}^+ = r^+\}$ and let $0 < c < 1/2$. Then $\Xi(r^+)$ is an event with high probability.*

4.2.2. Step 2: estimate the CDF of the eigenvalue distribution of $ZZ^\top$

We achieve this goal by modifying the eigenvalue distribution of $\widetilde{S}\widetilde{S}^\top$. Similar to the idea of the truncated spectrum in [47], we omit the first $\widehat{r}^+$ eigenvalues of $\widetilde{S}\widetilde{S}^T$. Moreover, as discussed in the last section, we estimate C'' via $\widehat{C}'' := \frac{\widetilde{\lambda}_{\lfloor n^c \rfloor + \widehat{r}^+ + 1} - \widetilde{\lambda}_{2\lfloor n^c \rfloor + \widehat{r}^+ + 1}}{(2\lfloor n^c \rfloor/p)^{2/3} - (\lfloor n^c \rfloor/p)^{2/3}}$, and thus we reconstruct the $(\widehat{r}^+ + 1)$th to the $(\widehat{r}^+ + \lfloor n^c \rfloor)$th eigenvalues by

$$\widehat{\lambda}_j := \widetilde{\lambda}_{\lfloor n^c \rfloor + \widehat{r}^+ + 1} + \frac{1 - \left(\frac{j - \widehat{r}^+ - 1}{\lfloor n^c \rfloor}\right)^{2/3}}{2^{2/3} - 1}\left(\widetilde{\lambda}_{\lfloor n^c \rfloor + \widehat{r}^+ + 1} - \widetilde{\lambda}_{2\lfloor n^c \rfloor + \widehat{r}^+ + 1}\right), \tag{51}$$

where $1 \le j \le \lfloor n^c \rfloor$ and c is a small fixed positive constant. By (33) and the fact that $|\gamma_{\lfloor n^c \rfloor} - \gamma_{\lfloor n^c \rfloor + r^+ + 1}| \asymp n^{-2/3}$, we immediately have the following comparison for $\check{\lambda}_j$ and $\widehat{\lambda}_j$.

Theorem 4.3. *Suppose Assumption 2.3 and (32) hold true and $c \in (0, 1/2)$. With high probability, for $j = 1, \dots, \lfloor n^c \rfloor$, we have*

$$|\check{\lambda}_j - \lambda_j| \prec \frac{1}{n\alpha_+} + n^{-1/2}\phi_n + n^{-2/3} + n^{-1/3}\phi_n^2, \tag{52}$$

$$|\widehat{\lambda}_{j+\widehat{r}^+} - \lambda_j| \prec \frac{1}{n\alpha^+} + n^{-1/2}\phi_n + n^{-\min\{\frac{2}{3}+\frac{c}{3}, \frac{3}{4}, \frac{5}{6}-\frac{c}{3}, \frac{4}{3}-\frac{4c}{3}\}} + n^{-1/3-2c/3}\phi_n^2. \tag{53}$$

Recall that $\phi_n = n^{2/a-1/2}$ with $a > 4$, which corresponds to the entries of random matrix X before truncation having finite a-moment. Note that the difference between the right-hand side of (52) and (53) is in their last two terms $n^{-2/3} + n^{-1/3}\phi_n^2$ and

$n^{-\min\{\frac{2}{3}+\frac{c}{3},\frac{3}{4},\frac{5}{6}-\frac{c}{3},\frac{4}{3}-\frac{4c}{3}\}}+n^{-1/3-2c/3}\phi_n^2$. When $4<a\le 6$, both $n^{-2/3}+n^{-1/3}\phi_n^2$ and $n^{-\min\{\frac{2}{3}+\frac{c}{3},\frac{3}{4},\frac{5}{6}-\frac{c}{3},\frac{4}{3}-\frac{4c}{3}\}}+n^{-1/3-2c/3}\phi_n^2$ are dominated by $n^{-1/2}\phi_n$, and our estimator $\widehat{\lambda}_{j+\widehat{r}^+}$ has a similar error bound on absolute error when compared with $\check{\lambda}_j$. When $a>6$, both $n^{-2/3}+n^{-1/3}\phi_n^2$ and $n^{-\min\{\frac{2}{3}+\frac{c}{3},\frac{3}{4},\frac{5}{6}-\frac{c}{3},\frac{4}{3}-\frac{4c}{3}\}}+n^{-1/3-2c/3}\phi_n^2$ become more dominant. Since $n^{-\min\{\frac{2}{3}+\frac{c}{3},\frac{3}{4},\frac{5}{6}-\frac{c}{3},\frac{4}{3}-\frac{4c}{3}\}}<n^{-2/3}$ for $c\in(0,1/2)$, our estimator $\widehat{\lambda}_{j+\widehat{r}^+}$ exhibits a lower upper bound on absolute error compared with $\check{\lambda}_j$. Finally, we propose to estimate the CDF of $\pi_{ZZ^\top}$ by

$$\widehat{F}_{\mathrm{e}}(x):=\frac{1}{p-\widehat{r}^+}\left(\sum_{j=\widehat{r}^++1}^{\lfloor n^c\rfloor+\widehat{r}^+}\mathbb{1}(\widehat{\lambda}_j\le x)+\sum_{j=\lfloor n^c\rfloor+\widehat{r}^++1}^{p}\mathbb{1}(\widetilde{\lambda}_j\le x)\right). \tag{54}$$

4.2.3. Step 3: estimate the optimal shrinker

With $\widehat{F}_{\mathrm{e}}$, we now state our optimal shrinker with the proposed rank estimator. The CDF of the ESD of $ZZ^\top$ is estimated by (54) with r^+ estimated by $\widehat{r}^+$ in (49). Consider the following "discretization" of the associated quantities. For $1\le i\le\widehat{r}^+$, denote the estimators of $m_{1c}(\widetilde{\lambda}_i)$ and $m_{2c}(\widetilde{\lambda}_i)$ as

$$\begin{aligned}\widehat{m}_{e,1,i}&:=\int\frac{d\widehat{F}_e(x)}{x-\widetilde{\lambda}_i}=\frac{1}{p-\widehat{r}^+}\left(\sum_{j=\widehat{r}^++1}^{\lfloor n^c\rfloor+\widehat{r}^+}\frac{1}{\widehat{\lambda}_j-\widetilde{\lambda}_i}+\sum_{j=\lfloor n^c\rfloor+\widehat{r}^++1}^{p}\frac{1}{\widetilde{\lambda}_j-\widetilde{\lambda}_i}\right),\\ \widehat{m}_{e,2,i}&:=-\frac{1-\beta_n}{\widetilde{\lambda}_i}+\beta_n\widehat{m}_{e,1,i}.\end{aligned} \tag{55}$$

Similarly, denote the discretization of $m'_{1c}(\widetilde{\lambda}_i)$ and $m'_{2c}(\widetilde{\lambda}_i)$ as

$$\begin{aligned}\widehat{m}'_{e,1,i}&=\frac{1}{p-\widehat{r}^+}\left(\sum_{j=\widehat{r}^++1}^{\lfloor n^c\rfloor+\widehat{r}^+}\frac{1}{(\widehat{\lambda}_j-\widetilde{\lambda}_i)^2}+\sum_{j=\lfloor n^c\rfloor+\widehat{r}^++1}^{p}\frac{1}{(\widetilde{\lambda}_j-\widetilde{\lambda}_i)^2}\right),\\ \widehat{m}'_{e,2,i}&=\frac{1-\beta_n}{\widetilde{\lambda}_i^2}+\beta_n\widehat{m}'_{e,1,i}.\end{aligned}$$

For $1\le i\le\widehat{r}^+$, the estimators of the D-transform $\mathcal{T}(\widetilde{\lambda}_i)$ and its derivative $\mathcal{T}'(\widetilde{\lambda}_i)$ are

$$\widehat{\mathcal{T}}_{e,i}=\widetilde{\lambda}_i\widehat{m}_{e,1,i}\widehat{m}_{e,2,i},\quad \widehat{\mathcal{T}}'_{e,i}=\widehat{m}_{e,1,i}\widehat{m}_{e,2,i}+\widetilde{\lambda}_i\widehat{m}'_{e,1,i}\widehat{m}_{e,2,i}+\widetilde{\lambda}_i\widehat{m}'_{e,2,i}\widehat{m}_{e,1,i}, \tag{56}$$

and the estimators of d_i, $a_{1,i}=\langle\mathbf{u}_i,\widetilde{\boldsymbol{\xi}}_i\rangle^2$, and $a_{2,i}=\langle\mathbf{v}_i,\widetilde{\boldsymbol{\zeta}}_i\rangle^2$ are

$$\widehat{d}_{e,i}=\sqrt{\frac{1}{\widehat{\mathcal{T}}_{e,i}}},\quad \widehat{a}_{e,1,i}=\frac{\widehat{m}_{e,1,i}}{\widehat{d}_{e,i}^2\widehat{\mathcal{T}}'_{e,i}},\ \text{ and }\ \widehat{a}_{e,2,i}=\frac{\widehat{m}_{e,2,i}}{\widehat{d}_{e,i}^2\widehat{\mathcal{T}}'_{e,i}}. \tag{57}$$

As a result, we estimate the optimal shrinker φ_i^* in Proposition 2.6 by

$$\begin{aligned}&\widehat{\varphi}_{\mathrm{e},i}=\widehat{d}_{\mathrm{e},i}\sqrt{\widehat{a}_{\mathrm{e},1,i}\widehat{a}_{\mathrm{e},2,i}}, && \text{(Frobenius norm)}\\ &\widehat{\varphi}_{\mathrm{e},i}=\widehat{d}_{e,i}\sqrt{\frac{\widehat{a}_{e,1,i}\wedge\widehat{a}_{e,2,i}}{\widehat{a}_{e,1,i}\vee\widehat{a}_{e,2,i}}}, && \text{(Operator norm)}\\ &\widehat{\varphi}_{\mathrm{e},i}=\widehat{d}_{\mathrm{e},i}\left(\sqrt{\widehat{a}_{\mathrm{e},1,i}\widehat{a}_{\mathrm{e},2,i}}-\sqrt{(1-\widehat{a}_{\mathrm{e},1,i})(1-\widehat{a}_{\mathrm{e},2,i})}\right), && \text{(Nuclear norm)}\end{aligned} \tag{58}$$

for $1\le i\le\widehat{r}^+$, and $\widehat{\varphi}_{\mathrm{e},i}=0$ otherwise. The following theorem provides the convergence guarantee of the proposed estimator.

Theorem 4.4. *Suppose Assumption 2.3 and (32) hold true for some $\varepsilon>0$ such that $n^{\epsilon}(\phi_n+n^{-1/3})>n^{-1/6}$, and $c\in(0,1/2)$. For all three types of loss functions mentioned in Proposition 2.6, for $1\le i\le\widehat{r}^+$, conditional on $\Xi(r^+)$, we have $|\varphi_i^*-\widehat{\varphi}_{e,i}|\prec\phi_n+n^{-1/2}/\Delta(d_i)$.*

If we replace $\widehat{F}_e(x)$ by

$$\widehat{F}_{\mathrm{T}}(x):=\frac{1}{p-\widehat{r}^+}\sum_{i=\widehat{r}^++1}^{p}\mathbb{1}(\widetilde{\lambda}_i\le x), \tag{59}$$

then by Theorem 4.2, eOptShrink with the Frobenius norm loss is reduced to OptShrink proposed in [47]. It is possible to estimate φ_i^* by $\widehat{F}_{\mathrm{T}}(x)$ or $\widehat{F}_{\mathtt{imp}}(x)$. We denote that resulting estimates of φ_i^* as $\widehat{\varphi}_{\mathrm{T},i}$ and $\widehat{\varphi}_{\mathtt{imp},i}$, respectively. In the next section, we numerically show that using $\widehat{F}_{\mathrm{e}}(x)$ results in a lower estimation error compared to either using $\widehat{F}_{\mathrm{T}}(x)$ or $\widehat{F}_{\mathtt{imp}}(x)$.

4.2.4. Selection of c in practice

Recall that in *Step 1*, we necessitate a constant c such that $\lfloor n^c \rfloor \gg r^+$ to compute the estimator of bulk edge $\widehat{\lambda}_+$ and the estimated effective rank $\widehat{r}^+$. Conclusions from Theorems 4.1 and 4.2 necessitate $c \in (0, 1/2)$ for our estimators to converge to the ground truth when n is sufficiently large. Similarly, in *Step 2*, based on Theorem 4.1, 4.2, and 4.3, we require $c \in (0, 1/2)$ to ensure that our estimator $\widehat{\lambda}_{j+\widehat{r}^+}$ has a smaller asymptotic error rate than $\check{\lambda}_j$ for accurately estimating λ_j, where $j = 1, \ldots, \lfloor n^c \rfloor$. It is worth noting that while we have convergence rate in the established theorems, when n is finite, it is not clear how to choose the optimal c that minimizes the right-hand side of (53). The situation is worse in practical scenarios when n is not significantly large. For example, when $n = 300$, a small c may fail to guarantee $\lfloor n^c \rfloor \gg r^+$. A conservative consideration is choosing a c close to $1/2$. With this approach, computing $\widehat{\lambda}_{j+\widehat{r}^+}$ for $\lfloor n^c \rfloor$ times becomes non-optimal and time-consuming for large n. To balance various practical situations, it is logical to opt for a small $c > 0$ when n is large and a large $c < 1/2$ when n is small. Hence, we propose considering $c = \frac{1}{2.01} \wedge \frac{1}{\log(\log(n))}$, ensuring the convergence conditions mentioned above while computing $\widehat{\lambda}_{j+\widehat{r}^+}$ for $\lfloor n^{(1/2.01)\wedge(1/\log(\log(n)))} \rfloor$ iterations. The selection of the term $1/\log(\log(n))$ is motivated by computational considerations. Specifically, $\frac{1}{\log(\log(n))} > \frac{1}{2.01}$ for $n < 1743$, while $\frac{1}{\log(\log(n))} < \frac{1}{2.01}$ for $n \geq 1743$. Consequently, this choice ensures a reasonable matrix size to illustrate scenarios when either $n^{1/2.01}$ or $n^{1/\log(\log(n))}$ is dominant. In the subsequent numerical simulations, we vary n within the range of 300 to 2100 to illustrate the performance of this choice of constant c. It is important to emphasize that this choice of c is not necessarily optimal, but in practice, we found it to work effectively.

5. Numerical evaluation

We assess the performance of eOptShrink through numerical simulations involving various types of noise and the single-channel fECG extraction problem. In all our findings, we use interquartile range error bars or present means with standard deviations to establish performance superiority between methods. We use a paired t-test to compare the performance of the two methods, considering $p < 0.005$ as statistically significant.

5.1. Simulated signals

We consider different types of noise. Suppose $X \in \mathbb{R}^{p\times n}$ has i.i.d. entries from a Student's t-distribution with 10 degrees of freedom, normalized such that $\mathbb{E}X_{ij}^2 = 1/n$. Set $A = \frac{1}{L_A} Q_A D_A Q_A^T \in \mathbb{R}^{p\times p}$, where $D_A = \mathrm{diag}\{\ell_1, \ell_2, \ldots, \ell_p\}$, $Q_A \in O(p)$ is generated by the QR decomposition of a random $p \times p$ matrix independent of X, and $L_A = \sum_{i=1}^p \ell_i$ is a normalizing factor. The same method is used to generate $B = \frac{1}{L_B} Q_B D_B Q_B^T \in \mathbb{R}^{n\times n}$, which is assumed to be independent of A and X. We consider three types of noise. The first one is the white noise (called TYPE1 below), where $D_A = I_p$ and $D_B = I_n$. The second one has a separable covariance structure (called TYPE2 below) with a gap in the limiting distribution, where $D_A = \mathrm{diag}\left\{ \sqrt{1 + 9 \times \frac{1}{p}}, \sqrt{1 + 9 \times \frac{2}{p}}, \cdots, \sqrt{1 + 9 \times \frac{p-1}{p}}, \sqrt{10} \right\}$ and $D_B = \mathrm{diag}\left\{ \sqrt{10 + \frac{1}{n}}, \sqrt{10 + \frac{2}{n}}, \cdots, \sqrt{10 + \frac{\lfloor n/4 \rfloor}{n}}, \sqrt{0.3}, \cdots, \sqrt{0.3}, \sqrt{0.3} \right\}$. The third one (called TYPE3 below) has a more complicated separable covariance structure with $D_A = \mathrm{diag}\left\{ \exp(\frac{1}{p}), \exp(\frac{2}{p}) \cdots, \exp(\frac{p-1}{p}), \exp(1) \right\}$ and $D_B = \mathrm{diag}\left\{ 1.1 + \sin(4\pi(\frac{1}{n})), 1.1 + \sin(4\pi(\frac{2}{n})), \cdots, 1.1 + \sin(4\pi(\frac{n-1}{n})), \ 1.1 + \sin(4\pi) \right\}$. The signal matrix is $S = \sum_{i=1}^r d_i \mathbf{u}_i \mathbf{v}_i^\top$, where $r = 10$, the d_i are i.i.d. uniformly sampled from $[0, 4]$ and ordered $d_1 \geq d_2 \ldots \geq d_{10}$, and the left and right singular vectors are generated by the QR decomposition of two independent random matrices. We independently realize $\tilde{S} = S + A^{1/2} X B^{1/2}$ 100 times for different n, noise types and $p/n = 0.5$ or 1. Results comparing different algorithms are provided. More simulations are provided in Section S.7 in the supplementary material.

To compare the performance of different algorithms, we need to calculate α (25). For Type1 noise, α is determined by $\widehat{\alpha} := (p/n)^{1/4}$ as described in the paragraph after Proposition 2.6. For TYPE2 and TYPE3 noise, directly calculation of α is challenging, so we use a numerical approach. For the chosen p/n, construct $Z = A^{1/2} X B^{1/2} \in \mathbb{R}^{p'\times n'}$ with large n' and $p'/n' = p/n$, and denote the eigenvalues of $Z^\top Z$ as $\{\lambda_i\}_{i=1}^{n'}$. Then, calculate $\widehat{\alpha} := 1/\sqrt{\lambda_1 \check{m}_1(\lambda_1) \check{m}_2(\lambda_1)}$, where $\check{m}_1(x) = \frac{1}{p'-1} \left(\sum_{j=2}^{p'} \frac{1}{\lambda_j - x} \right)$ and $\check{m}_2(x) = \frac{1}{n'-1} \left(\sum_{j=2}^{n'} \frac{1}{\lambda_j - x} \right)$. By Lemma S.1.10 and Theorem S.1.15, $|\widehat{\alpha} - \alpha| \lesssim (n')^{-1/3}$, which is sufficiently small for large n'. We set $n' = 10000$ and independently construct $\widehat{\alpha}$ for 100 times. For TYPE2 noise, we have $\widehat{\alpha} = 1.3495 \pm 0.0290$ when $p/n = 0.5$, and $\widehat{\alpha} = 1.6515 \pm 0.0180$ when $p/n = 1$, and for TYPE3 noise, we have $\widehat{\alpha} = 1.5242 \pm 0.0320$ when $p/n = 0.5$, and $\widehat{\alpha} = 1.8115 \pm 0.0348$ when $p/n = 1$, where we show the mean $\pm$ standard deviation. In subsequent steps, α is determined by the mean of the 100 calculations, and for simplicity we still denote it as $\widehat{\alpha}$ afterwards.

5.1.1. Visualization of ESDs and thresholds

For TRAD, the noise level is estimated using $\check{\sigma}(\widetilde{S})$ for fair comparison. For ScreeNOT, we use the ground truth rank, set $k = 4r$ and call the estimated rank by the hard threshold ϑ_{SN} in (4) the *ScreeNOT rank*. Fig. 2 illustrates ESDs of $\widetilde{S} = S + A^{1/2} X B^{1/2}$ for all three types of noise when $n = 2000$. An obvious gap is present in the bulks associated with TYPE2 noise when $p/n = 1$. The black line indicates the estimated bulk edge from (48), which separates the noise and signal for all three types of noise. The yellow line, the estimated bulk edge by TRAD using $\check{\sigma}(\widetilde{S})(1 + \sqrt{p/n})$, separates noise and signal well for TYPE1 noise but not for TYPE2 and TYPE3 noise. The red line indicates the ScreeNOT rank, which discards weak components of signals that are very close to the noise bulk edge. For TYPE1 noise, the black and yellow lines align around $1 + \sqrt{p/n}$.

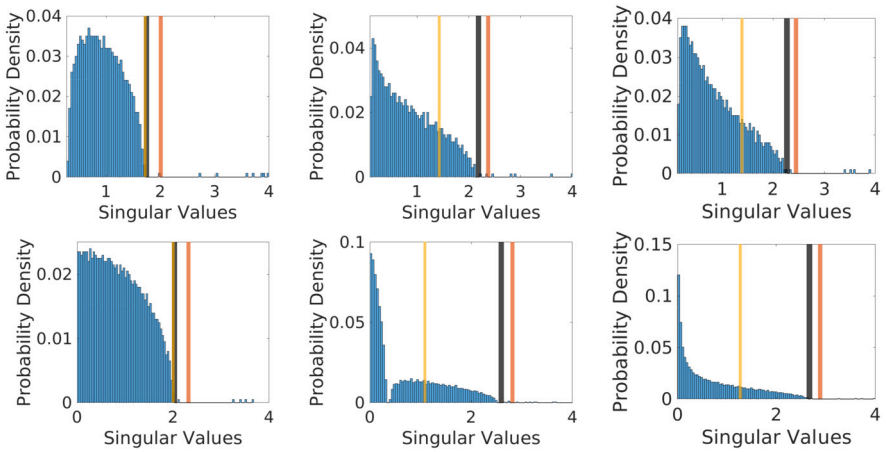

Fig. 2. ESD of $\widetilde{S}$ when $n = 2000$. The black and yellow lines are the estimated bulk edge from (48) and TRAD respectively. The red line is ϑ in (4) by ScreeNOT. The first row has $p/n = 0.5$ and the second row has $p/n = 1$. The first, second, and third columns are TYPE1, TYPE2, and TYPE3 noise respectively.

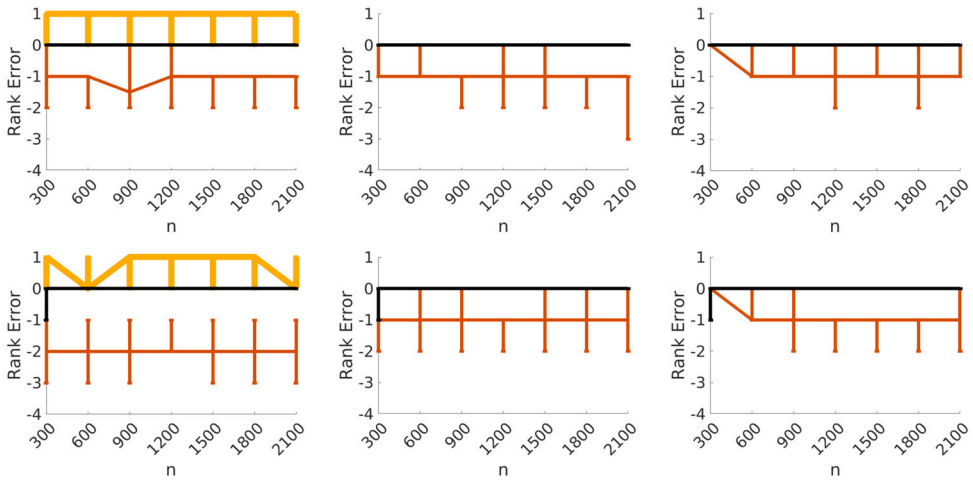

Fig. 3. A comparison of different rank estimators, where we show $\hat{r}^+ - r^+$. The black (red and yellow respectively) lines are errors of rank estimator from our estimator (49) (ScreeNot and TRAD respectively). The first row has $p/n = 0.5$ and the second row has $p/n = 1$. The first, second, and third columns are TYPE1, TYPE2, and TYPE3 noise respectively.

5.1.2. Rank estimation

To evaluate rank estimation, the ground truth is determined using $\hat{\alpha}$, with r^+ determined by $r_*^+ := \#\{i | d_i - \hat{\alpha} > n^{-3/10}\}$, where $n^{-3/10}$ accounts for the 10 degrees of freedom of entries of X. Given that $\hat{\alpha}$ is sufficiently close to α, we can safely assume that $r^+ = r_*^+$ and check the performance of our rank estimator $\hat{r}^+$. Fig. 3 compares the estimated rank using (49), the ScreeNOT rank, and the rank estimated by TRAD, which counts the eigenvalues of $\widetilde{S}\widetilde{S}^\top$ exceeding $\check{\sigma}(\widetilde{S})(1 + \sqrt{p/n})$. TRAD consistently overestimates the rank for TYPE2 and TYPE3 noise and occasionally for TYPE1 noise. ScreeNOT rank often underestimates the rank, resulting in larger errors than our approach, and it shows no improvement as n grows. Our approach outperforms the others. When n is small, like 300, our rank estimator sometimes underestimates the rank due to the $n^{-1/3}$ gap in (49), which excludes eigenvalues of $\widetilde{S}\widetilde{S}^\top$ that lack significant information.

Sections 5.1.1 and 5.1.2 demonstrate that TRAD consistently performs best with TYPE1 noise due to its closed form for λ_+ and optimal shrinkers. However, TRAD's performance deteriorates with TYPE2 or TYPE3 noise because its closed form becomes invalid, leading to significant errors. ScreeNOT, which does not detect weak signals or perform shrinkage, also shows high errors for both noise types. eOptShrink matches TRAD's performance with TYPE1 noise when n is large and outperforms all methods with TYPE2 and TYPE3 noise. Overall, these simulations confirm that (49) accurately estimates r^+, validating the use of $\hat{r}^+$ (49) in eOptShrink.

5.1.3. Optimal shrinkage estimation via $\widehat{F}_{\mathrm{T}}(x)$, $\widehat{F}_{\mathrm{imp}}(x)$ and $\widehat{F}_{\mathrm{e}}(x)$

We evaluate the effectiveness of the proposed pseudo-distribution $\widehat{F}_{\mathrm{e}}$ compared to existing distributions $\widehat{F}_{\mathrm{T}}(x)$ and $\widehat{F}_{\mathrm{imp}}(x)$ for estimating optimal shrinkage in Proposition 2.6 via estimating d_{r^+} and $\sqrt{a_{1,r^+}a_{2,r^+}}$. For $\widehat{F}_{\mathrm{imp}}(x)$ in (47), we set $k = 2r$ based on oracle rank information. For a fair comparison, we adjust c so that $\lfloor n^c \rfloor = 2r$ when modifying the top $2\hat{r}^+$ eigenvalues of $\widetilde{S}\widetilde{S}^\top$ in $\widehat{F}_{\mathrm{e}}(x)$ (54).

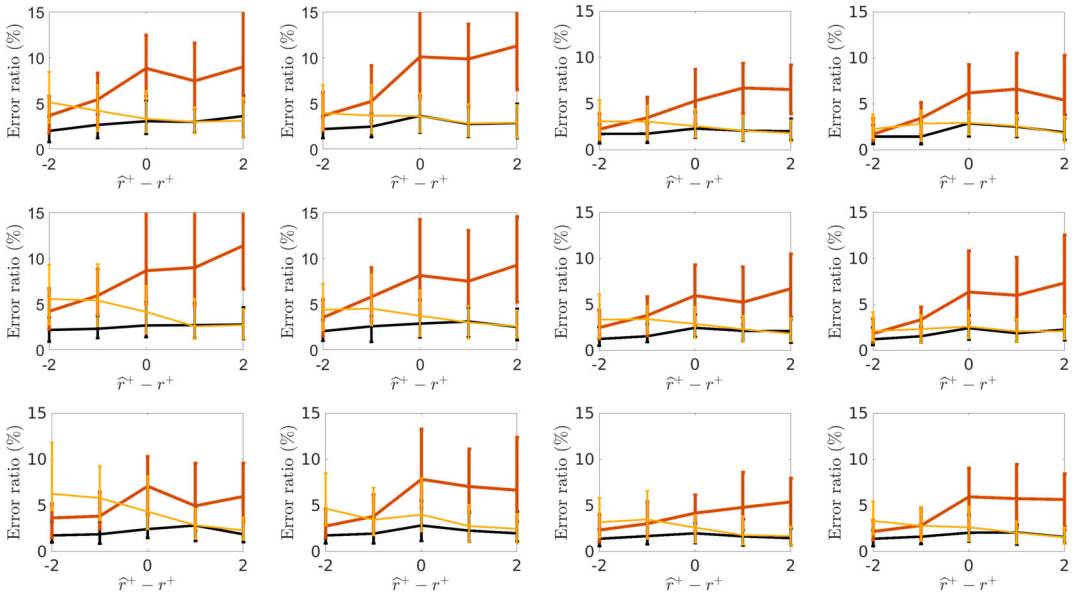

Fig. 4. Interquartile errorbars of error ratios of estimating $d_{\min\{r^+,\widehat{r}^+\}}$ by using $\widehat{F}_{\mathrm{e}}(x)$, $\widehat{F}_{\mathrm{T}}(x)$ and $\widehat{F}_{\mathrm{imp}}(x)$, shown in black, yellow, and red lines respectively. If the corresponding error ratio is too high, the associated curve is not totally plotted to enhance the visualization. From the first row to the third row: TYPE1, TYPE2, and TYPE3 noise. From the first column to the fourth column: $p/n = 0.5$ and $n = 300$, $p/n = 1$ and $n = 300$, $p/n = 0.5$ and $n = 600$, and $p/n = 1$ and $n = 600$.

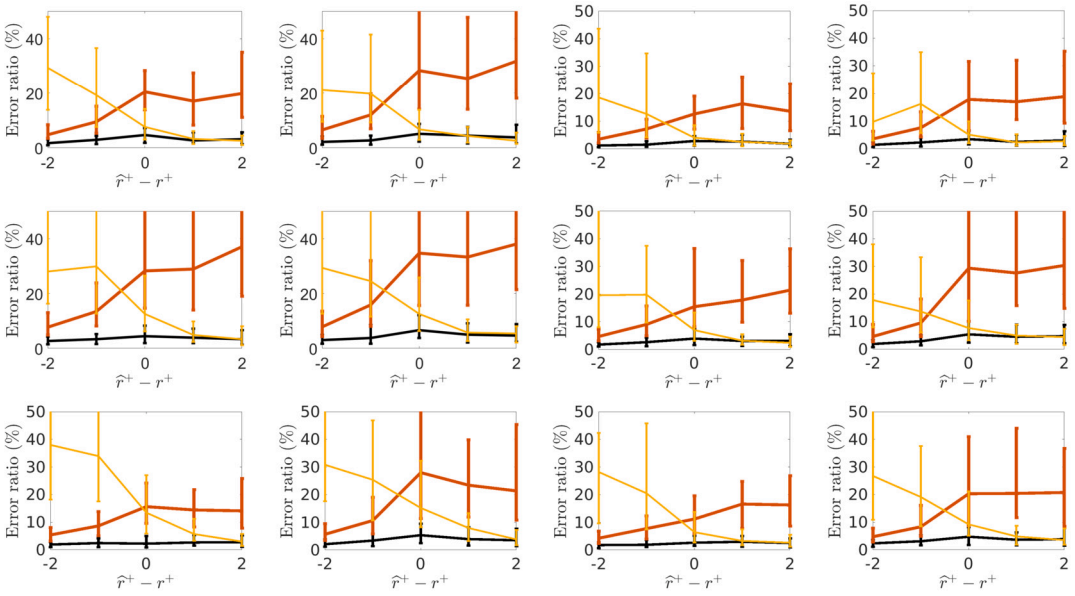

Fig. 5. Interquartile errorbars of error ratios of estimating $\sqrt{a_{1,\min\{r^+,\widehat{r}^+\}}a_{2,\min\{r^+,\widehat{r}^+\}}}$ using $\widehat{F}_{\mathrm{e}}(x)$, $\widehat{F}_{\mathrm{T}}(x)$ and $\widehat{F}_{\mathrm{imp}}(x)$, shown in black, yellow, and red lines respectively. If the corresponding error ratio is too high, the associated curve is not totally plotted to enhance the visualization. From the first row to the third row: TYPE1, TYPE2, and TYPE3 noise. From the first column to the fourth column: $p/n = 0.5$ and $n = 300$, $p/n = 1$ and $n = 300$, $p/n = 0.5$ and $n = 600$, and $p/n = 1$ and $n = 600$.

As shown in the last simulation, our rank estimator is precise for large n, but less reliable for small n. Since $\widehat{F}_{\mathrm{T}}(x)$ and $\widehat{F}_{\mathrm{e}}(x)$ depend on the estimated rank $\widehat{r}^+$, we examine the impact of erroneous rank estimates by considering cases where $\widehat{r}^+ - r^+ = -2, \ldots, +2$. Figs. 4 and 5 illustrate the error ratios for estimating $d_{\min\{r^+,\widehat{r}^+\}}$ and $\sqrt{a_{1,\min\{r^+,\widehat{r}^+\}}a_{2,\min\{r^+,\widehat{r}^+\}}}$ with $n = 300$ and $n = 600$ using different pseudo-distributions. The r^+th singular value, being the smallest "strong" one, is the most challenging one to recover. We conduct 100 simulations for each n. Our approach consistently shows the lowest error ratios with a statistical significance, and the error decreases as n grows from 300 to 600. This indicates that eOptShrink is robust to slightly errors in rank estimation.

Table 1
A comparison of the Frobenius norm loss relative to the ground truth between eOptShrink and OptShrink [47] across various selected ranks. The matrix size is $p = n = 1500$. All results are presented as mean ± standard deviation. The first column lists the noise type, while the second column identifies the algorithm used. The third through seventh columns show results for different selected ranks in both algorithms: $\widehat{r}^+ - 1$, $\widehat{r}^+$, $\widehat{r}^+ + 1$, the ground truth rank r, and the ground truth effective rank r^+. A dagger next to the mean indicates that, for the same noise type and selected rank, eOptShrink's error is significantly lower than OptShrink's. An asterisk indicates that the associated algorithm's error for the selected rank is significantly higher than the error using the ground truth effective rank r^+ (shown in the seventh column).

Noise	Algorithm	$\widehat{r}^+ - 1$	$\widehat{r}^+$	$\widehat{r}^+ + 1$	r	r^+
TYPE1	eOptShrink	$7.742^{*} \pm 0.578$	7.494 ± 0.523	$7.492^{\dagger} \pm 0.522$	$7.494^{\dagger} \pm 0.522$	7.493 ± 0.522
	OptShrink	$7.750^{*} \pm 0.586$	7.494 ± 0.523	7.494 ± 0.522	$7.507^{*} \pm 0.511$	7.493 ± 0.522
TYPE2	eOptShrink	$9.143^{*\dagger} \pm 0.644$	8.799 ± 0.687	8.797 ± 0.684	$8.812^{*\dagger} \pm 0.665$	8.799 ± 0.687
	OptShrink	$9.163^{*} \pm 0.653$	8.799 ± 0.687	8.799 ± 0.684	$8.894^{*} \pm 0.613$	8.798 ± 0.687
TYPE3	eOptShrink	$9.334^{*\dagger} \pm 0.823$	9.119 ± 0.777	$9.126^{\dagger} \pm 0.772$	$9.183^{*\dagger} \pm 0.731$	9.120 ± 0.777
	OptShrink	$9.352^{*} \pm 0.838$	9.118 ± 0.777	$9.133^{*} \pm 0.765$	$9.292^{*} \pm 0.668$	9.117 ± 0.777

5.2. Comparison of eOptShrink and OptShrink

Following the rank estimation comparison in Subsection 5.1.2 and the pseudo-distribution comparison in Subsection 5.1.3, we now compare the proposed eOptShrink algorithm with the Frobenius norm loss and the OptShrink algorithm from [47] by reporting the Frobenius norms between the cleaned matrix and the ground truth matrix S as a measure of error.

OptShrink uses an estimated effective rank [47], but to our knowledge how to choose rank for OptShrink is unclear. We suggest to use $\widehat{r}^+$ as the estimated effective rank for OptShrink. To check its efficiency and sensitivity to such rank estimate, we test various ranks, including $\widehat{r}^+ - 1$, $\widehat{r}^+$, $\widehat{r}^+ + 1$, r, and r^+, where r^+ is numerically determined as described in Section 5.1.2. We also assess the sensitivity of eOptShrink to rank estimation by varying the rank used in step (i) of Algorithm 1. Results with $p = n = 1500$ are shown in Table 1.

Using the true effective rank r^+, the estimated $\widehat{r}^+$, or a slight overestimate ($\widehat{r}^+ + 1$), both eOptShrink and OptShrink yield the lowest errors with no significant differences, except for the case with rank $\widehat{r}^+ + 1$ under TYPE3 noise, where OptShrink's error is significantly higher than that of eOptShrink. When using $\widehat{r}^+ - 1$, or the true rank r, which is typically much larger than r^+, the error increases significantly for both algorithms. However, in these cases, eOptShrink's error is notably lower than that of OptShrink, except when using $\widehat{r}^+ - 1$ under TYPE1 noise. These results, combined with Subsections 5.1.2 and 5.1.3, highlight the importance of accurate rank estimation and suggest that replacing $\widehat{F}_{\mathrm{T}}(x)$ in OptShrink with the proposed $\widehat{F}_{\mathrm{e}}(x)$ offers potential benefits.

5.3. Fetal ECG extraction problem

In our previous work [56], TRAD is the critical step of the algorithm to recover the mECG when we have only one or two ta-mECG channels. The algorithm is composed of two steps. The first step is mainly for two channels, and we ignore it when we only have one channel. The second step is composed of two substeps. Step 2-1 is designed to detect the maternal R peaks from the single channel ta-mECG, which is not the concern of OS. Step 2-2 is mainly illustrated in Fig. 1, where we view fECG as noise and mECG as the signal, and OS is applied to recover mECG from the ta-mECG. As is mentioned in Introduction, fECG, when viewed as noise, is not white, and there is a dependence among segments. Thus, it is natural to consider replacing TRAD in [56] by OptShrink or eOptShrink. We consider a semi-real simulated database and a real-world database from *2013 PhysioNet/Computing in Cardiology Challenge* [54], abbreviated as CinC2013.

5.3.1. Semi-real simulated database

The semi-real ta-mECG data is constructed from the Physikalisch-Technische Bundesanstalt (PTB) Database https://physionet.org/physiobank/database/ptbdb/, abbreviated as `PTBDB` following the same way detailed in [56]. The database contains 549 recordings from 290 subjects (one to five recordings for one subject) aged 17 to 87 with the mean age 57.2. 52 out of 290 subjects are healthy. Each recording includes simultaneously recorded conventional 12 lead and the Frank lead ECGs. Each signal is digitalized with the sampling rate 1000 Hz. More technical details can be found online. Take 57-second Frank lead ECGs from a healthy recording, denoted as $V_x(t)$, $V_y(t)$ and $V_z(t)$ at time $t \in \mathbb{R}$, as the maternal vectocardiogram (VCG). Take $(\theta_{xy}, \theta_z) = (\frac{\pi}{4}, \frac{\pi}{4})$, and the simulated mECG is created by $\mathrm{mECG}(t) = (V_x(t)\cos\theta_{xy} + V_y(t)\sin\theta_{xy})\cos\theta_z + V_z(t)\sin\theta_z$. We create 40 mECGs. The *simulated fECG* of healthy fetus are created from another 40 recordings from healthy subjects, where 114-second V2 and V4 recordings are taken. The simulated and simulated fECG come from different subjects. The simulated fECG then are resampled at 500 Hz. As a result, the simulated fECG has about double the heart rate compared with the simulated mECG if we consider both signals sampled at 1000 Hz. The amplitude of the simulated fECG is normalized to the same level of simulated mECG and then multiplied by $0 < R < 1$ shown in the second column of Table 2 to make the amplitude relationship consistent with the usual situation of real ta-mECG signals. We generate 40 simulated healthy fECGs. The clean simulated ta-mECG is generated by directly summing simulated mECG and fECG. We then create a simulated noise starting with a random vector $\mathbf{x} = (x_1, x_2, x_3, \ldots)$ with i.i.d. entries with student t-10 distribution. The noise is then created and denotes as $\mathbf{z}$ with the entries $z_i = (1 + 0.5\sin((i \mod 500)/500))(x_i + x_{i+1} + x_{i+2})$. The final simulated ta-mECG is generated by adding the created noise to the clean simulated ta-mECG according to the desired SNR ratio shown in the first column

Table 2
The comparison of RMSE for the mECG morphology of different algorithms applied to the simulated ta-mECG database. R is the simulated fECG amplitude. All results are presented as mean ± standard deviation. Three different estimated rank, $\hat{r}^+ - 1$, $\hat{r}^+$, and $\hat{r}^+ + 1$ are selected for OptShrink. The star next to the mean stands for the smaller RMSE with statistical significance when comparing with TRAD and OptShrink using $\hat{r}^+ - 1$ and $\hat{r}^+ + 1$ as the estimated rank by the paired t-test. The asterisk next to the mean stands for the smaller RMSE with statistical significance when comparing with OptShrink using $\hat{r}^+$ as the estimated rank by the paired t-test. Optimal shrinkers with respect to Frobenius norm loss are applied for both TRAD and eOptShrink.

SNR	R	TRAD	OptShrink			eOptShrink
			$r^+ - 1$	$\hat{r}^+$	$\hat{r}^+ + 1$	
1 dB	1/4	10.841 ± 1.87	14.718 ± 2.227	6.414⋆ ± 1.233	6.676 ± 1.257	6.413⋆⋆ ± 1.231
	1/6	8.859 ± 1.509	11.810 ± 2.239	5.386⋆ ± 1.018	5.594 ± 1.048	5.373⋆⋆ ± 1.006
	1/8	7.696 ± 1.299	10.084 ± 2.067	4.789⋆ ± 0.908	4.907 ± 0.904	4.777⋆⋆ ± 0.899
0 dB	1/4	11.185 ± 2.014	15.331 ± 2.407	6.517⋆ ± 1.177	6.772 ± 1.248	6.517⋆± 1.177
	1/6	9.192 ± 1.642	12.380 ± 2.318	5.507⋆ ± 1.007	5.674 ± 1.018	5.507⋆± 1.007
	1/8	7.962 ± 1.387	10.844 ± 1.703	4.936⋆ ± 0.919	5.039 ± 0.895	4.932⋆⋆ ± 0.915
−1 dB	1/4	11.675 ± 2.121	15.823 ± 3.059	6.771⋆ ± 1.184	6.976 ± 1.236	6.771⋆ ± 1.184
	1/6	9.526 ± 1.766	12.834 ± 2.583	5.658⋆ ± 0.971	5.793 ± 1.002	5.658⋆ ± 0.971
	1/8	8.289 ± 1.483	10.933 ± 2.386	5.016⋆ ± 0.905	5.122 ± 0.899	5.016⋆ ± 0.905

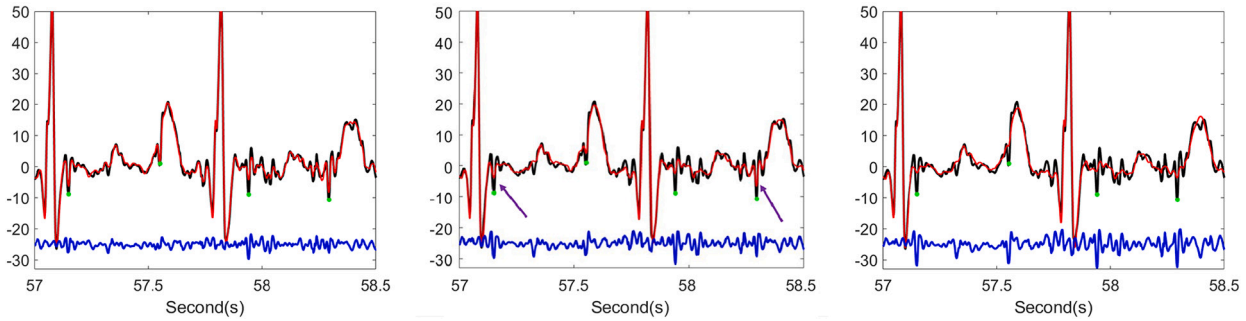

Fig. 6. A comparison of mECG and fECG recovery using different algorithms (recording 1 and channel 1 from CinC2013) is illustrated. The left panel depicts results obtained by TRAD, the middle one by OptShrink, and the right one by eOptShrink. The original ta-mECG is depicted in black, with the recovered mECG overlaid in red, and the fetal R peaks labeled by experts are indicated as green dots. The difference between the ta-mECG and the recovered mECG is displayed in blue, representing the recovered fetal ECG. The purple arrows highlight ventricular activity residue, a limitation of OptShrink due to the lack of a precise estimator of effective rank.

of Table 2. As a result, for each combination of the fECG amplitude R and SNR, we acquire 40 recordings of 57 seconds simulated ta-mECG signals with the sampling rate 1000 Hz, and each recording has 5.7×10^4 data points.

Assuming the covariance structure of the summation of fECG and background noise fits the separable covariance model $Z = A^{1/2} X B^{1/2}$, we apply our eOptShrink algorithm to each recording in the simulated database in Step 2-2 and compare its performance with TRAD and OptShrink. We test OptShrink's sensitivity to the estimated effective rank by setting the rank estimator to $\hat{r}^+ - 1$, $\hat{r}^+$, and $\hat{r}^+ + 1$. For a fair comparison, both TRAD and eOptShrink are applied using Frobenius norm loss, as OptShrink is based on this loss function. We compute the averaged root mean square error (RMSE) of the recovered mECG compared to the ground truth across 5.7×10^4 data points per recording, with results aggregated over 40 recordings. The results are shown in Table 2.

Our results reveal that eOptShrink consistently achieves the lowest RMSE compared to TRAD and OptShrink with statistical significance, especially when the estimated effective rank is set to $\hat{r}^+ - 1$ or $\hat{r}^+ + 1$. When compared to OptShrink with the effective rank set to $\hat{r}^+$, eOptShrink shows a lower RMSE for all values of R when SNR= 1 dB, and for $R = 1/8$ when SNR= 0 dB. For other SNR and R values, there is no statistically significant difference. Notably, the choice of $\hat{r}^+$ results in the smallest RMSE with statistical significance. This finding suggests the importance of an accurate effective rank estimation for OptShrink and that our estimator $\hat{r}^+$ is an appropriate choice. Furthermore, using $\widehat{F}_e(x)$ instead of $\widehat{F}_T(x)$ improves RMSE. Overall, lower SNRs lead higher RMSEs across all amplitude ratios R and higher fECG amplitudes result in higher noise levels for mECG recovery, leading to higher RMSEs across all SNRs for all methods.

5.3.2. CinC2013 database

Each recording in CinC2013 includes four ta-mECG channels and simultaneously recorded directly contacted fECG, all resampled at 1000 Hz and lasting for 1 minute. For each channel of every recording, we execute the fECG extraction algorithm using various OS algorithms in Step 2-2. In Fig. 6, we compare the recovered mECG and the detected fetal R peak locations using TRAD, OptShink, and eOptShrink. While OptShrink [47] claims to mitigate rank overestimation effects, we use an overestimated rank $\hat{r}^+ + 1$ for OptShrink, and apply both TRAD and eOptShrink with the Frobenius norm loss. eOptShrink provides a more accurate recovery of mECG *morphology* and fECG. Notably, ventricular activity residues observed in the fECG recovered by OptShrink, indicated by purple

arrows, are absent in eOptShrink, leading to a better evaluation of maternal and fetal heart electrophysiology. The clinical implications will be discussed in future work.

Data availability

The database is publicly available.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.acha.2024.101698.

References

[1] Sankaraleengam Alagapan, Hae Won Shin, Flavio Fröhlich, Hau-Tieng Wu, Diffusion geometry approach to efficiently remove electrical stimulation artifacts in intracranial electroencephalography, J. Neural Eng. 16 (3) (2019) 036010.
[2] Johannes Alt, Singularities of the density of states of random Gram matrices, Electron. Commun. Probab. 22 (ne) (2017) 1–13.
[3] Johannes Alt, László Erdős, Torben Krüger, Local law for random Gram matrices, Electron. J. Probab. 22 (ne) (2017) 1–41.
[4] W. Greg Anderson, Alice Guionnet, Ofer Zeitouni, An Introduction to Random Matrices, Number 118, Cambridge University Press, 2010.
[5] Rasoul Anvari, Mokhtar Mohammadi, Amin Roshandel Kahoo, Enhancing 3-d seismic data using the t-svd and optimal shrinkage of singular value, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (1) (2018) 382–388.
[6] Zhidong Bai, Jack W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, vol. 20, Springer, 2010.
[7] Zhidong Bai, Jian-feng Yao, Central limit theorems for eigenvalues in a spiked population model, Ann. I.H.P. Probab. Stat. 44 (3) (2008) 447–474.
[8] Zhidong Bai, Jianfeng Yao, On sample eigenvalues in a generalized spiked population model, J. Multivar. Anal. 106 (2012) 167–177.
[9] Jinho Baik, Gérard Ben Arous, Sandrine Péché, et al., Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, Ann. Probab. 33 (5) (2005) 1643–1697.
[10] Jinho Baik, Jack W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, J. Multivar. Anal. 97 (6) (2006) 1382–1408.
[11] Zhigang Bao, Guangming Pan, Wang Zhou, Universality for the largest eigenvalue of sample covariance matrices with general population, Ann. Stat. 43 (1) (2015) 382–421.
[12] Florent Benaych-Georges, Alice Guionnet, Mylène Maida, Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices, Electron. J. Probab. 16 (2011) 1621–1662.
[13] Florent Benaych-Georges, Raj Rao Nadakuditi, The singular values and vectors of low rank perturbations of large rectangular random matrices, J. Multivar. Anal. 111 (2012) 120–135.
[14] Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, Jun Yin, On the principal components of sample covariance matrices, Probab. Theory Relat. Fields 164 (1–2) (2016) 459–552.
[15] Raymond B. Cattell, The scree test for the number of factors, Multivar. Behav. Res. 1 (2) (1966) 245–276.
[16] Neng-Tai Chiu, Stephanie Huwiler, M. Laura Ferster, Walter Karlen, Hau-Tieng Wu, Caroline Lustenberger, Get rid of the beat in mobile eeg applications: a framework towards automated cardiogenic artifact detection and removal in single-channel eeg, Biomed. Signal Process. Control 72 (2022) 103220.
[17] Lucilio Cordero-Grande, Daan Christiaens, Jana Hutter, Anthony N. Price, Jo V. Hajnal, Complex diffusion-weighted image estimation via matrix recovery under general noise models, Neuroimage 200 (2019) 391–404.
[18] Romain Couillet, Walid Hachem, Analysis of the limiting spectral measure of large random matrices of the separable covariance type, Random Matrices: Theory Appl. 3 (04) (2014) 1450016.
[19] Xiucai Ding, High dimensional deformed rectangular matrices with applications in matrix denoising, Bernoulli 26 (1) (2020) 387–417.
[20] Xiucai Ding, Spiked sample covariance matrices with possibly multiple bulk components, Random Matrices: Theory Appl. 10 (01) (2021) 2150014.
[21] Xiucai Ding, Fan Yang, A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices, Ann. Appl. Probab. 28 (3) (2018) 1679–1738.
[22] Xiucai Ding, Fan Yang, Spiked separable covariance matrices and principal components, Ann. Stat. 49 (2) (2021) 1113–1138.
[23] David L. Donoho, Matan Gavish, Iain M. Johnstone, Optimal shrinkage of eigenvalues in the spiked covariance model, Ann. Stat. 46 (4) (2018) 1742.
[24] David L. Donoho, Matan Gavish, Elad Romanov, ScreeNOT: exact mse-optimal singular value thresholding in correlated noise, arXiv preprint, arXiv:2009.12297, 2020.
[25] Mathias Drton, Satoshi Kuriki, Peter Hoff, Existence and uniqueness of the Kronecker covariance mle, Ann. Stat. 49 (5) (2021) 2721–2754.
[26] Carl Eckart, Gale Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (3) (1936) 211–218.
[27] Bradley Efron, Are a set of microarrays independent of each other?, Ann. Appl. Stat. 3 (3) (2009) 922.
[28] Noureddine El Karoui, Tracy–Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices, Ann. Probab. 35 (2) (2007) 663–714.
[29] László Erdős, Antti Knowles, Horng-Tzer Yau, Averaging fluctuations in resolvents of random band matrices, Ann. Henri Poincaré 14 (8) (2013) 1837–1926.
[30] Matan Gavish, David L. Donoho, The optimal hard threshold for singular values is $4/\sqrt{3}$, IEEE Trans. Inf. Theory 60 (8) (2014) 5040–5053.
[31] Matan Gavish, David L. Donoho, Optimal shrinkage of singular values, IEEE Trans. Inf. Theory 63 (4) (2017) 2137–2152.
[32] David Gerard, Peter Hoff, Equivariant minimax dominators of the mle in the array normal model, J. Multivar. Anal. 137 (2015) 32–49.
[33] Gene Golub, William Kahan, Calculating the singular values and pseudo-inverse of a matrix, J. Soc. Ind. Appl. Math., Ser. B Numer. Anal. 2 (2) (1965) 205–224.
[34] Peter D. Hoff, Limitations on detecting row covariance in the presence of column covariance, J. Multivar. Anal. 152 (2016) 249–258.
[35] Iain M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Stat. 29 (2) (2001) 295–327.
[36] Antti Knowles, Jun Yin, The isotropic semicircle law and deformation of Wigner matrices, Commun. Pure Appl. Math. 66 (11) (2013) 1663–1749.
[37] Antti Knowles, Jun Yin, The outliers of a deformed Wigner matrix, Ann. Probab. 42 (5) (2014) 1980–2031.
[38] Antti Knowles, Jun Yin, Anisotropic local laws for random matrices, Probab. Theory Relat. Fields 169 (1) (2017) 257–352.
[39] Ji Oon Lee, Kevin Schnelli, Tracy–Widom distribution for the largest eigenvalue of real sample covariance matrices with general population, Ann. Appl. Probab. 26 (6) (2016) 3786–3839.
[40] William Leeb, Rapid evaluation of the spectral signal detection threshold and Stieltjes transform, Adv. Comput. Math. 47 (4) (2021) 1–29.
[41] William Leeb, Optimal singular value shrinkage for operator norm loss: extending to non-square matrices, Stat. Probab. Lett. 186 (2022) 109472.
[42] William Leeb, Elad Romanov, Optimal spectral shrinkage and pca with heteroscedastic noise, IEEE Trans. Inf. Theory 67 (5) (2021) 3009–3037.
[43] William E. Leeb, Matrix denoising for weighted loss functions and heterogeneous signals, SIAM J. Math. Data Sci. 3 (3) (2021) 987–1012.
[44] Tzu-Chi Liu, Yi-Wen Liu, Hau-Tieng Wu, Denoising click-evoked otoacoustic emission signals by optimal shrinkage, J. Acoust. Soc. Am. 149 (4) (2021) 2659–2670.
[45] Vladimir A. Marčenko, Leonid Andreevich Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR Sb. 1 (4) (1967) 457.
[46] Leon Mirsky, Symmetric gauge functions and unitarily invariant norms, Q. J. Math. 11 (1) (1960) 50–59.

[47] Raj Rao Nadakuditi, Optshrink: an algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage, IEEE Trans. Inf. Theory 60 (5) (2014) 3002–3018.
[48] Alexei Onatski, The Tracy–Widom limit for the largest eigenvalues of singular complex Wishart matrices, Ann. Appl. Probab. 18 (2) (2008) 470–490.
[49] Sean O'Rourke, Van Vu, Ke Wang, Eigenvectors of random matrices: a survey, J. Comb. Theory, Ser. A 144 (2016) 361–442.
[50] Paul Debashis, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Stat. Sin. (2007) 1617–1642.
[51] Paul Debashis, Jack W. Silverstein, No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix, J. Multivar. Anal. 100 (1) (2009) 37–57.
[52] Natesh S. Pillai, Jun Yin, Universality of covariance matrices, Ann. Appl. Probab. 24 (3) (2014) 935–1001.
[53] Andrey A. Shabalin, Andrew B. Nobel, Reconstruction of a low-rank matrix in the presence of gaussian noise, J. Multivar. Anal. 118 (2013) 67–76.
[54] Ikaro Silva, Joachim Behar, Reza Sameni, Tingting Zhu, Julien Oster, Gari D. Clifford, George B. Moody, Noninvasive fetal ecg: the physionet/computing in cardiology challenge 2013, in: Computing in Cardiology 2013, IEEE, 2013, pp. 149–152.
[55] Jack W. Silverstein, Sang-Il Choi, Analysis of the limiting spectral distribution of large dimensional random matrices, J. Multivar. Anal. 54 (2) (1995) 295–309.
[56] Pei-Chun Su, Stephen Miller, Salim Idriss, Piers Barker, Hau-Tieng Wu, Recovery of the fetal electrocardiogram for morphological analysis from two trans-abdominal channels via optimal shrinkage, Physiol. Meas. 40 (11) (2019) 115005.
[57] Pei-Chun Su, Elsayed Z. Soliman, Hau-Tieng Wu, Robust t-end detection via t-end signal quality index and optimal shrinkage, Sensors 20 (24) (2020) 7052.
[58] Craig A. Tracy, Harold Widom, Level-spacing distributions and the airy kernel, Commun. Math. Phys. 159 (1) (1994) 151–174.
[59] Craig A. Tracy, Harold Widom, On orthogonal and symplectic matrix ensembles, Commun. Math. Phys. 177 (3) (1996) 727–754.
[60] Craig A. Tracy, Harold Widom, Distribution functions for largest eigenvalues and their applications, arXiv preprint, arXiv:math-ph/0210034, 2002.
[61] Jelle Veraart, Dmitry S. Novikov, Daan Christiaens, Benjamin Ades-Aron, Jan Sijbers, Els Fieremans, Denoising of diffusion mri using random matrix theory, Neuroimage 142 (2016) 394–406.
[62] Lili Wang, Debashis Paul, Limiting spectral distribution of renormalized separable sample covariance matrices when $p/n \to 0$, J. Multivar. Anal. 126 (2014) 25–52.
[63] D.V. Widder, The Stieltjes transform, Trans. Am. Math. Soc. 43 (1) (1938) 7–60.
[64] Haokai Xi, Fan Yang, Jun Yin, Convergence of eigenvector empirical spectral distribution of sample covariance matrices, Ann. Stat. 48 (2) (2020) 953–982.
[65] Fan Yang, Edge universality of separable covariance matrices, Electron. J. Probab. 24 (2019) 1–57.
[66] Lixin Zhang, Spectral analysis of large dimensional random matrices, PhD thesis, National University of Singapore, 2007.