SIAM J. MATH. DATA SCI.
Vol. 5, No. 1, pp. 222–250

Optimally Weighted PCA for High-Dimensional Heteroscedastic Data*

David Hong†, Fan Yang‡, Jeffrey A. Fessler§, and Laura Balzano§

Abstract. Modern data are increasingly both high-dimensional and heteroscedastic. This paper considers the challenge of estimating underlying principal components from high-dimensional data with noise that is heteroscedastic across samples, i.e., some samples are noisier than others. Such heteroscedasticity naturally arises, e.g., when combining data from diverse sources or sensors. A natural way to account for this heteroscedasticity is to give noisier blocks of samples less weight in PCA by using the leading eigenvectors of a weighted sample covariance matrix. We consider the problem of choosing weights to optimally recover the underlying components. In general, one cannot know these optimal weights since they depend on the underlying components we seek to estimate. However, we show that under some natural statistical assumptions the optimal weights converge to a simple function of the signal and noise variances for high-dimensional data. Surprisingly, the optimal weights are not the inverse noise variance weights commonly used in practice. We demonstrate the theoretical results through numerical simulations and comparisons with existing weighting schemes. Finally, we briefly discuss how estimated signal and noise variances can be used when the true variances are unknown, and we illustrate the optimal weights on real data from astronomy.

Key words. principal component analysis, large-dimensional data, heterogeneous quality, optimal weighting

MSC code. 62H25

DOI. 10.1137/22M1470244

1. Introduction.

Principal Component Analysis (PCA) is a fundamental technique for discovering underlying components in data and is a workhorse method for analyzing modern *high-dimensional* data. However, conventional PCA does not recover underlying principal components well when the data has *heteroscedastic* noise, as is common in practice. In particular, its performance can degrade substantially when the noise is heteroscedastic across samples, i.e., some samples are noisier than others. PCA suffers from treating all the samples

* Received by the editors January 10, 2022; accepted for publication (in revised form) October 11, 2022; published electronically March 30, 2023.
https://doi.org/10.1137/22M1470244
Funding: The first author was supported in part by NSF Graduate Research Fellowship DGE 1256260, NSF grant ECCS-1508943, NSF BIGDATA grant IIS 1837992, the Dean's Fund for Postdoctoral Research of the Wharton School, and NSF Mathematical Sciences Postdoctoral Research Fellowship DMS 2103353. The second author was supported in part by the Wharton Dean's Fund for Postdoctoral Research. The third author was supported in part by the UM-SJTU data science seed fund, NSF grant IIS 1838179, and NIH grant U01 EB 018753. The fourth author was supported in part by DARPA-16-43-D3M-FP-037, NSF CAREER award CCF-1845076, NSF grant ECCS-1508943, and ARO YIP award W911NF1910027.

† Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA, 19104 USA (dahong67@wharton.upenn.edu).

‡ Yau Mathematical Sciences Center, Tsinghua University, and Beijing Institute of Mathematical Sciences and Applications, Beijing, 100084 China (fyangmath@mail.tsinghua.edu.cn).

§ Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48109 USA (fessler@umich.edu, girasole@umich.edu).

uniformly with performance held back by the noisiest samples, as was rigorously characterized in [17]. Weighted PCA addresses this shortcoming by giving less weight to lower quality samples. This naturally raises a crucial question: how should the weights be chosen? Namely, *what are the optimal weights?*

This paper addresses this question by rigorously deriving optimal weights that are simple functions of the signal and noise variances. Surprisingly, they are not the inverse noise variance weights that are commonly used in practice. We now elaborate in more detail.

1.1. High-dimensional and heteroscedastic data. Modern applications of PCA span numerous and diverse areas across all of engineering and the sciences, ranging from medical imaging [3, 38] to cancer data classification [41], genetics [31], and environmental sensing [36, 47], to name just a few. Increasingly, the number of features measured is comparable to or even larger than the number of samples, i.e., the data are *high-dimensional*. Traditional asymptotic analysis of the performance of methods as the number of samples grows (with a fixed number of features) do not apply well to such settings. Modern data analysis needs new theory and methods for the high-dimensional regime where both the number of features and number of samples are large [24].

Modern datasets are also frequently composed of samples with heteroscedastic (i.e., heterogeneous) noise. In particular, we consider noise that is *heteroscedastic across samples*, namely, some samples are noisier than others. Such data arises naturally when samples are obtained at varying times or by varying means or equipment. For example, in the field of analytical chemistry, [10] considers spectrophotometric data obtained from averages taken over varying windows of time; samples from shorter windows are noisier. As another example, in the field of air quality monitoring, samples come from various sources: government agencies provide low-noise data obtained from carefully operated instruments, while individuals provide noisier data obtained from cheaper and easy-to-setup sensors [39, 45]. As a final example, in the field of astronomy, measurements of astronomical objects such as stars and quasars can have various levels of noise due to atmospheric and detector effects that vary from object to object [4, 43, 44]. More generally, modern big data analysis is often performed using datasets built up by combining myriad sources, so one can expect that data with heteroscedastic noise will be the norm. Modern data analysis needs PCA methods that effectively account for this type of heteroscedasticity. Indeed, such methods may also unlock new opportunities to effectively leverage new sources of data with heteroscedastic noise.

1.2. Weighted PCA. Weighted PCA accounts for heteroscedastic noise by giving smaller weight to noisier samples. Analogous to unweighted PCA, the principal components are the leading eigenvectors of the *weighted sample covariance matrix*

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} := \sum_{\ell=1}^{L} w_\ell \boldsymbol{Y}_\ell \boldsymbol{Y}_\ell^{\mathsf{H}},$$

where $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_L$ are L blocks of samples with associated noise variances $v_1, \dots, v_L > 0$, the superscript H denotes the Hermitian transpose, and $w_1, \dots, w_L \geq 0$ are the weights. Existing choices for the weights include the following:

- **Uniform weights** ($w_\ell = 1$): these weights correspond to unweighted PCA and can be a natural choice when the noise is close to homoscedastic. However, its performance degrades with increasing noise heteroscedasticity, as was shown in [17, Theorem 2].
- **Binary weights** ($w_\ell = 1$ for less noisy blocks and $w_\ell = 0$ for the rest): these weights correspond to performing unweighted PCA using only less noisy blocks of samples and are a natural choice when some samples are much noisier than the rest. The idea is to exclude noisier samples that do more harm than good. However, doing so also omits any useful information that was in the excluded samples. How to decide if a block of samples is better to include or exclude can be unclear.
- **Inverse noise variance weights** ($w_\ell = 1/v_\ell$): these weights whiten the noise, making it homoscedastic, and can be interpreted as a maximum likelihood weighting [49]. They are a natural way to account for noise heteroscedasticity while using all the samples and are commonly used in practice.

It has been unclear which of these existing options to choose and whether any are optimal.

1.3. Contribution of this paper. The main contribution of this paper is *optimal weights* for high-dimensional heteroscedastic data, that are rigorously derived under some natural statistical assumptions. Roughly, we show that when both dimensions of the data are large, the optimal weights converge to the following simple *asymptotic optimal weights*:

$$w_\ell = \frac{1}{v_\ell} \frac{1}{1 + v_\ell/\lambda},$$

where v_ℓ is the noise variance and λ is the signal variance. Notably, these weights are inverse noise variance weights scaled by a simple term that depends on the noise-to-signal ratio v_ℓ/λ. See section 2 for the precise statement of the result and section 6 for its proof.

Naturally, one wonders how well these results apply for data with finitely many samples and features. Numerical simulations in section 3 illustrate that the optimal weights in finite dimensions (which are a function of the random signal coefficients and noise) concentrate around the asymptotic optimal weights as the data grows in size. As a result, the asymptotic optimal weights are often close to the optimal weights in finite dimensions when the dimensions are large enough.

We also compare the asymptotic optimal weights with the existing weights above: uniform, binary, and inverse noise variance weights. In particular, we consider how close they are to the optimal weights in finite dimensions (subsection 4.1), how well they perform in finite dimensions (subsection 4.2), and in what regimes they achieve positive asymptotic recovery (subsection 4.3). Overall, the asymptotic optimal weights outperform the existing weights.

One also wonders how to calculate the asymptotic optimal weights when the signal and noise variances are unknown. Naturally, one might consider simply using estimators of these variances. We explain that the resulting estimated weights are also asymptotically optimal as long as the estimators are consistent, and we give an example of such estimators (section 7).

Finally, we illustrate the asymptotic optimal weights on real data (section 8). The data are quasar spectra measured by the Sloan Digital Sky Survey and have heteroscedastic noise. The example exhibits some of the main themes of the paper and illustrates the potential for optimally weighted PCA to improve performance in real data.

1.4. Related works. Previous work on PCA for noise that is heteroscedastic (whether across samples or otherwise) have addressed various important questions, as elaborated below. However, to the best of our knowledge, the important question of optimal weighting was not previously considered. This paper rigorously answers the question of optimal weighting for noise that is heteroscedastic across samples. It will be interesting for future works to consider this question for other forms of heteroscedastic noise.

Some of our other papers considered various aspects of noise that is heteroscedastic across samples. In particular, [16, 17] derive the asymptotic performance of unweighted PCA and characterize the impact of heteroscedasticity. See [17, sections 1.3 and 2.3] for a discussion of the connections to previous analyses of PCA for homoscedastic noise (such as [22, 35, 37]), and see [17, section S1] for a discussion of the connections to spiked covariance models. Alternatively, [18, 19] consider a probabilistic PCA approach, where the noise heteroscedasticity is modeled via the statistical likelihood. The resulting method is not a weighted PCA. Instead, one must solve a challenging optimization problem, and [18, 19] develop several algorithms for this purpose.

A closely related model for heterogeneous data arises in the context of low-rank clutter estimation for RADAR. In this setting, the noise is homoscedastic but the clutter signal has heterogeneous strengths, i.e., the clutter covariances are a common low-rank matrix scaled by heterogeneous power factors. Maximum likelihood estimation of the common low-rank matrix and the power factors involves solving a challenging optimization problem, and [8, 9, 11, 42] develop efficient algorithms for this purpose. The estimation performance is analyzed in [6], and [1] considers maximum a posteriori estimation. This heterogeneous signal strength model is related to the heteroscedastic noise model in the present paper through a straightforward rescaling of the data: scaling each sample by the inverse of its power factor yields the model in the present paper. Thus, the optimally weighted PCA developed here can be straightforwardly modified to apply to the heterogeneous signal strength model; see supplementary material SM1 for details.

Several recent works develop PCA variants for high-dimensional data with noise that is heteroscedastic across features. In contrast to the samplewise heteroscedasticity we consider, featurewise heteroscedasticity produces a nonuniform bias along the diagonal of the covariance matrix that skews its eigenvectors even with infinitely many samples. An approach based on spectral shrinkage with noise whitening (to make it homoscedastic) is developed in [29]; the noise is whitened by weighting the features by their inverse noise variance. Whitening both the features and samples is considered in [30], and whitening in the context of linearly transformed signals is considered in [14]. Alternatively, [50] addresses the bias in the covariance matrix by iteratively replacing its biased diagonal entries using low-rank approximation. Estimating the number of underlying principal components is another important problem in this setting, and recent works [20, 26, 28] have developed new methods for tackling this challenge under heteroscedastic noise. Estimated principal components are also often combined with estimates of associated signal variances to obtain estimates of an underlying signal matrix or covariance. For homoscedastic noise, existing works have made tremendous progress on how to estimate these signal variances to optimize various objectives, typically by applying a carefully designed shrinkage to the eigenvalues of the sample covariance matrix; see, e.g., [15] and the references therein. A few recent works address this question in the context of heteroscedastic noise:

reference [29] derives optimal shrinkages for use with whitening, [30] derives optimal spectral denoisers, and [34] derives an optimal data-driven shrinkage.

Many works have considered weighted PCA methods in general; see [25, section 14.2.1] for a survey of some of these works. For example, [12, sections 5.4–5.5] discusses weighting features by inverse noise variance weights to account for featurewise heteroscedasticity. Weighting both samples and features is proposed in [10] for analyzing spectrophotometric data from scanning wavelength kinetics experiments; the weights are again inverse noise variance. Similar schemes have also been proposed in metabolomics [21] and astronomy [4, 43], to name just a few areas. Weighting data by inverse noise variance weights has been a recurring theme.

Overall, previous work on PCA for heteroscedastic noise made significant progress on various important questions. However, the important question of optimal weighting was not previously considered. This paper addresses that question for noise that is heteroscedastic across samples.

1.5. Organization of this paper. Section 2 states our main result: optimal weights and performance for high-dimensional data with heteroscedastic noise. Section 3 performs numerical simulations in finite dimensions, and section 4 compares the asymptotic optimal weights with existing weighting schemes: inverse noise variance weighted PCA, PCA using only a single block of the data, and unweighted PCA. Section 5 compares optimally weighted PCA with some additional methods. Section 6 proves the main result. The optimal weights depend on the signal and noise variances. Section 7 describes how estimates of these variances can be used when the true variances are unknown. Section 8 illustrates optimally weighted PCA on real data coming from astronomy. Codes for reproducing the figures in this paper are available online at https://gitlab.com/dahong/optimally-weighted-pca-heteroscedastic-data.

For readers mostly interested in understanding the underlying theory and proofs of the main result, we suggest starting with sections 2 and 6. For readers mostly interested in using optimally weighted PCA, we suggest starting with sections 2–5, 7, and 8.

2. Main result: Optimal weights and performance.

We begin by making precise the notion of optimal weights and optimal performance. Consider a dataset $\boldsymbol{Y}$ having k underlying orthonormal components $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$, where $\boldsymbol{Y}$ is made of L blocks $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_L$ of samples with heteroscedastic noise. Then, given weights $w_1, \dots, w_L \geq 0$, the $\boldsymbol{w}$-weighted PCA estimate of the ith component $\boldsymbol{u}_i$ from $\boldsymbol{Y}$, denoted $\hat{\boldsymbol{u}}_i(\boldsymbol{w}, \boldsymbol{Y})$, is

$$\hat{\boldsymbol{u}}_i(\boldsymbol{w}, \boldsymbol{Y}) := i\text{th leading eigenvector of the weighted sample covariance } \sum_{\ell=1}^{L} w_\ell \boldsymbol{Y}_\ell \boldsymbol{Y}_\ell^{\mathsf{H}}. \tag{2.1}$$

A natural way to measure the performance of the estimate, i.e., how well $\hat{\boldsymbol{u}}_i(\boldsymbol{w}, \boldsymbol{Y})$ recovers the ith component $\boldsymbol{u}_i$, is by the square inner product $r_i(\boldsymbol{w}, \boldsymbol{Y})$ given by

$$r_i(\boldsymbol{w}, \boldsymbol{Y}) := |\boldsymbol{u}_i^{\mathsf{H}} \hat{\boldsymbol{u}}_i(\boldsymbol{w}, \boldsymbol{Y})|^2. \tag{2.2}$$

Finally, optimal weights $\boldsymbol{w}_i^\star(\boldsymbol{Y})$ and the optimal performance $r_i^\star(\boldsymbol{Y})$ for the ith component are defined by

$$\boldsymbol{w}_i^\star(\boldsymbol{Y}) \in \operatorname*{argmax}_{\boldsymbol{w}} r_i(\boldsymbol{w}, \boldsymbol{Y}), \qquad r_i^\star(\boldsymbol{Y}) = \max_{\boldsymbol{w}} r_i(\boldsymbol{w}, \boldsymbol{Y}). \tag{2.3}$$

Note that the performance (2.2) depends on the underlying component $\boldsymbol{u}_i$. However, in practice, $\boldsymbol{u}_i$ is of course unknown so the optimization (2.3) cannot be done. Fortunately, as our main result below shows, the optimal weights $\boldsymbol{w}_i^\star(\boldsymbol{Y})$ and optimal performance $r_i^\star(\boldsymbol{Y})$ can be predicted when the data (a) satisfies some natural statistical assumptions, and (b) grows large in size, i.e., under the following setting.

Setting. We will assume the following setting throughout the remainder of this paper.

(a) The noisy data blocks $\boldsymbol{Y}_1 \in \mathbb{C}^{d\times n_1}, \dots, \boldsymbol{Y}_L \in \mathbb{C}^{d\times n_L}$ are generated from the components $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k \in \mathbb{C}^d$ with corresponding signal variances $\lambda_1 > \cdots > \lambda_k > 0$ as follows:

$$\boldsymbol{Y}_\ell = \boldsymbol{F}\boldsymbol{Z}_\ell + \boldsymbol{E}_\ell \in \mathbb{C}^{d\times n_\ell} \quad \text{for } \ell = 1, \dots, L, \tag{2.4}$$

where

- $\boldsymbol{F} := [\sqrt{\lambda_1}\boldsymbol{u}_1, \dots, \sqrt{\lambda_k}\boldsymbol{u}_k] \in \mathbb{C}^{d\times k}$ is a deterministic factor matrix common to all the blocks,
- $\boldsymbol{Z}_\ell \in \mathbb{C}^{k\times n_\ell}$ is a coefficient matrix with IID entries having zero mean and unit variance,
- $\boldsymbol{E}_\ell \in \mathbb{C}^{d\times n_\ell}$ is a noise matrix with IID entries having zero mean and variance $v_\ell > 0$,

and the noise entries further satisfy a technical condition: bounded ath moment with $a > 4$, i.e., $\exists_{a>4}$ s.t. $\mathbb{E}|(\boldsymbol{E}_\ell)_{i,j}|^a < \infty$.[1] Note that this model also includes real-valued data with real-valued coefficients and noise.

(b) The number of features d and numbers of samples $n_1, \dots, n_L$ all grow towards infinity but with fixed aspect ratios $n_\ell/d = c_\ell > 0$. This asymptotic regime captures datasets where the number of features and samples are roughly comparable, as is common in modern big data settings.

Note that under the model (2.4), the optimal weights and performance (2.3) are random quantities so their convergence will be probabilistic. Specifically, the convergence holds with probability one, i.e., it is *almost sure convergence*, which we will denote by $\xrightarrow{\text{a.s.}}$.

We are now ready to state the main result on the optimal weights and performance.

Theorem 2.1 (asymptotic optimal weights and performance). *The optimal weights $\boldsymbol{w}_i^\star(\boldsymbol{Y})$ and corresponding optimal performance $r_i^\star(\boldsymbol{Y})$ converge as*

$$\boldsymbol{w}_i^\star(\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \bar{\boldsymbol{w}}_i^\star := \left(\frac{1}{v_1}\frac{1}{1+v_1/\lambda_i}, \dots, \frac{1}{v_L}\frac{1}{1+v_L/\lambda_i}\right) \quad \textit{up to scaling,} \tag{2.5}$$

$$r_i^\star(\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \bar{r}_i^\star := \textit{the unique solution } x \in (0,1) \textit{ of } \sum_{\ell=1}^{L} \frac{c_\ell}{v_\ell/\lambda_i}\frac{1-x}{v_\ell/\lambda_i + x} = 1, \tag{2.6}$$

except when $\sum_{\ell=1}^{L} c_\ell(\lambda_i/v_\ell)^2 \le 1$, in which case $\boldsymbol{u}_i$ is asymptotically unrecoverable by any weighted PCA, i.e., $r_i(\boldsymbol{w}, \boldsymbol{Y}) \xrightarrow{\text{a.s.}} 0$ for all weights $\boldsymbol{w}$.

[1] This technical condition on the noise is satisfied by numerous distributions including the sub-Gaussian and sub-Exponential families [46, Propositions 2.5.2 and 2.7.1].

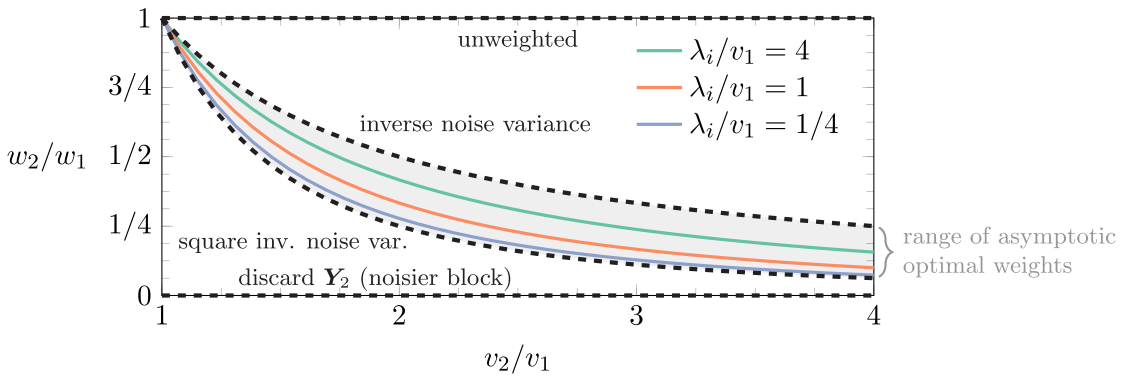


Figure 2.1. *Relative weight w_2/w_1 given by the optimal weights (2.5) as a function of the relative noise variance v_2/v_1 for various signal-to-noise ratios λ_i/v_1. The optimal weights downweight noisier data more aggressively than inverse noise variance weights, but also do not discard noisier data. They lie in the region between inverse and square inverse noise variance weights.*

Remark 2.2 (optimal weights downweight more than inverse noise variance weights). The optimal weights (2.5) are not the inverse noise variance weights that are commonly used. As illustrated in Figure 2.1, optimal weights downweight noisier data more aggressively, but never discard data. Specifically, when the signal-to-noise ratio λ_i/v_ℓ is small, the optimal weights are square inverse noise variance weights up to scale. As λ_i/v_ℓ grows, the optimal weights gradually become less aggressive and approach inverse noise variance weights.

Remark 2.3 (using estimated signal and noise variances). The optimal weights (2.5) depend on the noise variances $\boldsymbol{v}$ and on the signal component variance λ_i. In some settings, these parameters are known, e.g., from calibration data. When they are unknown, one may estimate them using existing ideas and approaches, then plug them in to obtain estimated weights. As discussed in section 7, these estimated weights are also asymptotically optimal as long as the variance estimates are consistent.

Remark 2.4 (heterogeneous signal strengths). The result may at first appear to be limited to data with homogeneous signal strengths. However, it generalizes straightforwardly to the case where the signal in each block is scaled by an associated signal strength, as arises, e.g., in RADAR applications [8, 9, 42]. Simply preweight the data to recover the model (2.4); see supplementary material SM1 for a detailed description.

Remark 2.5 (handling potentially degenerate cases). A careful reader may note the subtle and technical point that there may exist degenerate choices of $\boldsymbol{w}$ for which $r_i(\boldsymbol{w}, \boldsymbol{Y})$ is not well defined, e.g., if the ith leading eigenvector becomes undefined due to eigenvalue multiplicity. At such points, we define $r_i(\boldsymbol{w}, \boldsymbol{Y})$ by its lim sup over $\boldsymbol{w}$. Doing so makes $r_i(\boldsymbol{w}, \boldsymbol{Y})$ upper semicontinuous in $\boldsymbol{w}$ and avoids degenerate situations where its maximum does not exist.

Remark 2.6 (nonorthogonality of estimated components). Since the optimal weights (2.5) are component-specific, the components $\hat{\boldsymbol{u}}_1(\bar{\boldsymbol{w}}_1^\star, \boldsymbol{Y}), \ldots, \hat{\boldsymbol{u}}_k(\bar{\boldsymbol{w}}_k^\star, \boldsymbol{Y})$ estimated by optimally weighted PCA may not be orthogonal in practice. In applications where orthogonality is crucial, one option is to sacrifice componentwise optimality and use a single set of weights, e.g., that just optimizes recovery of the weakest component or that optimizes some appropriate

overall metric of performance. Alternatively, in many such cases, the principal subspace is of greater interest than the individual components; in these cases, one could orthogonalize the components, e.g., via Gram–Schmidt.

Remark 2.7 (phase transition). Analogous to unweighted PCA under homoscedastic noise, optimally weighted PCA exhibits a phase transition between settings with zero asymptotic performance and those with nonzero asymptotic performance. As described in Theorem 2.1, optimally weighted PCA has nonzero asymptotic performance when $\sum_{\ell=1}^{L} c_\ell(\lambda_i/v_\ell)^2 > 1$ (or in other words, $\lambda_i > (\sum_{\ell=1}^{L} c_\ell/v_\ell^2)^{-1/2}$). Notably, if any weighting scheme has nonzero asymptotic performance, then optimally weighted PCA does too, as illustrated in subsection 4.3.

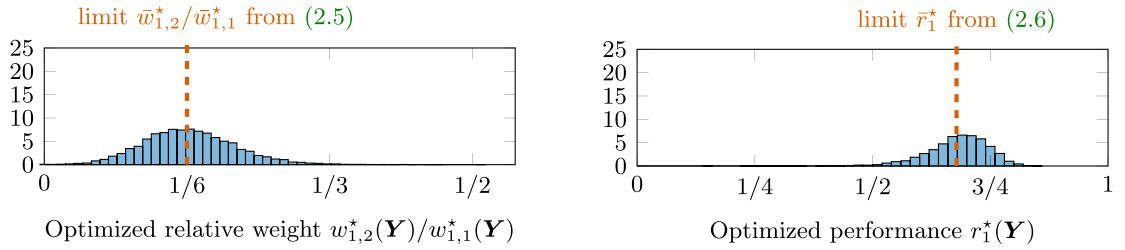
Before proving the main result (Theorem 2.1) in section 6, we provide some more intuition about it through numerical simulations in finite dimensions (section 3) and comparisons with existing weighting schemes (section 4).

3. Numerical simulation.

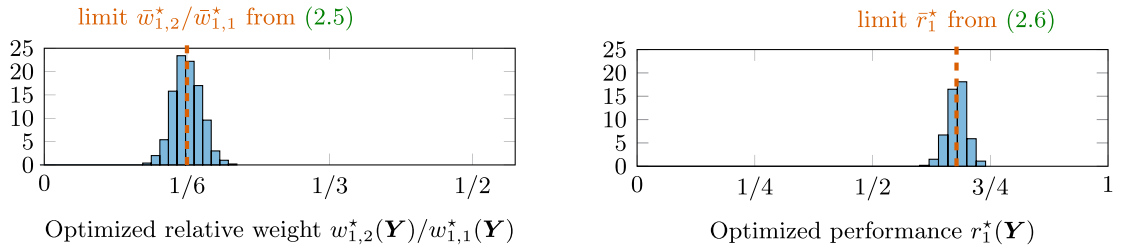
This section performs numerical simulations in finite dimensions. Specifically, we generate $L = 2$ blocks of data $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ according to the model (2.4) with

- $k = 1$ component $\boldsymbol{u}_1 \in \mathbb{R}^d$ uniformly drawn from the unit sphere,
- Gaussian coefficients $(\boldsymbol{Z}_\ell)_{ij} \overset{\text{IID}}{\sim} \mathcal{N}(0,1)$ and noise entries $(\boldsymbol{E}_\ell)_{ij} \overset{\text{IID}}{\sim} \mathcal{N}(0, v_\ell)$,
- component variance $\lambda_1 = 1$ and noise variances $\boldsymbol{v} = (1, 3)$.

Figure 3.1 shows the nonasymptotic empirical distributions of the optimal weights $\boldsymbol{w}_1^\star(\boldsymbol{Y})$ and the corresponding optimal performance $r_1^\star(\boldsymbol{Y})$ from (2.3) obtained using the true underlying



(a) Nonasymptotic distributions for $d = 100$, $n_1 = 400$ and $n_2 = 800$.



(b) Nonasymptotic distributions for $d = 1000$, $n_1 = 4000$ and $n_2 = 8000$.

Figure 3.1. *Nonasymptotic empirical distributions of optimal weights $\boldsymbol{w}_1^\star(\boldsymbol{Y})$ and optimal performance $r_1^\star(\boldsymbol{Y})$ from (2.3) for an illustrative example with two blocks of data $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ generated with noise variances $v_1 = 1$ and $v_2 = 3$, and with one underlying component having variance $\lambda_1 = 1$.*

component $\boldsymbol{u}_1$. Since the weights are meaningful only up to scale, we show the optimal relative weight $w^\star_{1,2}(\boldsymbol{Y})/w^\star_{1,1}(\boldsymbol{Y})$, where $w^\star_{i,\ell}(\boldsymbol{Y})$ is the ℓth entry of $\boldsymbol{w}^\star_i(\boldsymbol{Y})$. Similarly, $\bar{w}^\star_{i,\ell}$ denotes the ℓth entry of $\bar{\boldsymbol{w}}^\star_i$.

Note first that the nonasymptotic distributions for both the optimal weights $\boldsymbol{w}^\star_1(\boldsymbol{Y})$ and the optimal performance $r^\star_1(\boldsymbol{Y})$ are generally centered around their respective theoretical limits $\bar{\boldsymbol{w}}^\star_1$ and $\bar{r}^\star_1$ from (2.5) and (2.6). Moreover, they concentrate as the data grows in size from Figure 3.1a to Figure 3.1b. This illustrates the almost sure convergence of the nonasymptotic optimal weights and performance to their limits.

Naturally, one also wonders whether the asymptotic results of Theorem 2.1 can be used to choose optimal weights or predict optimal performance for real data, which are finite-dimensional. These experiments demonstrate that the asymptotic optimal weights (2.5) and performance (2.6) can indeed be applied to choose weights that are often close to optimal for finite-dimensional data and to predict their corresponding performance.

4. Comparison with existing weighting schemes. This section compares the asymptotic optimal weights of (2.5) with existing weighting schemes. To ease discussion, we will focus on the case with only two blocks $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, i.e., $L = 2$, but the same insights apply more generally. The weighting schemes considered are as follows:

(4.1)	inverse noise variance:	$\boldsymbol{w} = (1/v_1, 1/v_2)$,
(4.2)	only use $\boldsymbol{Y}_1$:	$\boldsymbol{w} = (1, 0)$,
(4.3)	only use $\boldsymbol{Y}_2$:	$\boldsymbol{w} = (0, 1)$,
(4.4)	unweighted:	$\boldsymbol{w} = (1, 1)$,

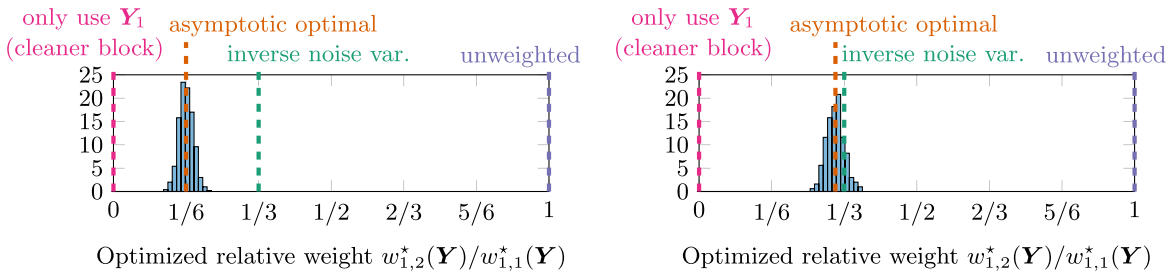
where the weights are, as always, meaningful only up to scale. Note that using only $\boldsymbol{Y}_1$ or $\boldsymbol{Y}_2$ are both special cases of general binary weights that discard blocks of data.

4.1. Comparison of weights. This section compares how close the various weighting schemes are to the distribution of the actual empirically optimized weights $\boldsymbol{w}^\star_i(\boldsymbol{Y})$ from (2.3) in an illustrative example. As in section 3, $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ are generated from the model (2.4) with Gaussian coefficients and noise, and a single signal component $\boldsymbol{u}_1 \in \mathbb{R}^d$ drawn uniformly from the unit sphere. The noise variances are $\boldsymbol{v} = (1, 3)$, and the data sizes are $d = 1000$, $n_1 = 4000$, and $n_2 = 8000$.

Figure 4.1 shows the nonasymptotic distribution of the empirically optimized weights (2.3) with the asymptotic optimal weights $\bar{\boldsymbol{w}}^\star_1$ from (2.5) and the existing weights (4.1)–(4.4) overlaid. Figure 4.1a shows a case with a moderate signal component variance $\lambda_1 = 1$, and Figure 4.1b shows a case with a strong signal component $\lambda_1 = 30$. Since the weights are meaningful only up to scale, we show the relative weight $w^\star_{1,2}(\boldsymbol{Y})/w^\star_{1,1}(\boldsymbol{Y})$ given to the noisier block $\boldsymbol{Y}_2$, where $w^\star_{i,\ell}(\boldsymbol{Y})$ is the ℓth entry of $\boldsymbol{w}^\star_i(\boldsymbol{Y})$, as in Figure 3.1.

We make the following observations from Figure 4.1:

- As before, the optimal weights are centered around the asymptotic optimal weights.
- When the signal component variance is moderate, the optimal weights do not overlap with the existing weighting schemes. They more aggressively downweight the noisier block than inverse noise variance weights but also do not discard the noisier block.
- When the signal component variance is large, the optimal weights overlap with inverse noise variance weights.



(a) Moderate signal component variance $\lambda_1 = 1$.

(b) Large signal component variance $\lambda_1 = 30$.

Figure 4.1. *Comparison of weighting schemes for an illustrative example with two blocks of data* $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ *and* $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ *generated with noise variances* $v_1 = 1$ *and* $v_2 = 3$*, where* $d = 1000$*,* $n_1 = 4000$*, and* $n_2 = 8000$*. The nonasymptotic distributions of empirically optimized weights* $\boldsymbol{w}^\star_1(\boldsymbol{Y})$ *from* (2.3) *are shown as histograms, with the asymptotic optimal weights* $\bar{\boldsymbol{w}}^\star_1$ *from* (2.5) *and the existing weights* (4.1)–(4.4) *overlaid as lines (the weights* (4.3) *that only use* $\boldsymbol{Y}_2$ *do not appear since they are at infinity).*

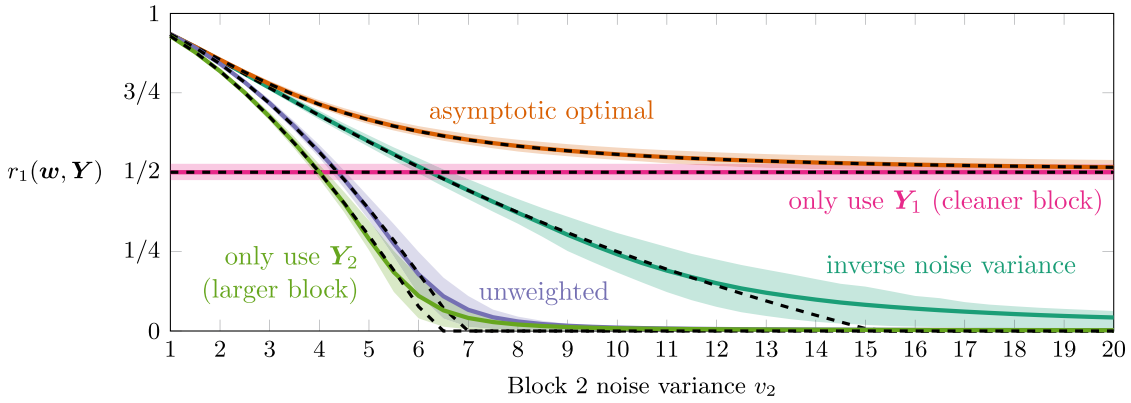


Figure 4.2. *Performance comparison of weighting schemes* (2.5) *and* (4.1)–(4.4) *for an illustrative example with two blocks of data* $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ *and* $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ *with* $d = 10^3$*,* $n_1 = 10^3$*,* $n_2 = 10^4$*, and signal component variance* $\lambda_1 = 2$*. The first block has noise variance* $v_1 = 1$*, while the second block noise variance ranges from* $v_2 = 1$ *to* $v_2 = 20$*. For each weighting scheme, the solid colored curve is the average from* 400 *trials, the ribbon indicates the corresponding interquartile interval, and the dashed black curve is the asymptotic performance from Theorem* 2.1 *and Proposition* 4.1*.*

4.2. Comparison of performance. This section compares the various weighting schemes in terms of their performance $r_i(\boldsymbol{w}, \boldsymbol{Y})$ in finite dimensions. As in section 3, $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ are generated from the model (2.4) with Gaussian coefficients and noise, and a single signal component $\boldsymbol{u}_1 \in \mathbb{R}^d$ drawn uniformly from the unit sphere. The signal component variance is $\lambda_1 = 2$, and the data sizes are $d = 10^3$, $n_1 = 10^3$, and $n_2 = 10^4$. The first noise variance is $v_1 = 1$ and the second noise variance ranges from $v_2 = 1$ to $v_2 = 20$. Figure 4.2 shows the nonasymptotic distribution of performance for the asymptotic optimal weights (2.5) and the existing weights (4.1)–(4.4).

We make the following observations from Figure 4.2:

- Across the entire sweep, the asymptotic optimal weights are generally best.
- The performance of the asymptotic optimal weights is well predicted by the asymptotic performance from Theorem 2.1.

- When v_2 is small, there is a lot of clean data coming from $\boldsymbol{Y}_2$ since n_2 is fairly large. All weighting schemes do well except the scheme of using only $\boldsymbol{Y}_1$.
- As v_2 grows, $\boldsymbol{Y}_2$ becomes noisier and the methods that use $\boldsymbol{Y}_2$ degrade in performance. Asymptotic optimal weighting degrades the most slowly/gracefully.
- As v_2 continues to grow, all methods that use $\boldsymbol{Y}_2$ eventually do worse than using only $\boldsymbol{Y}_1$ and hit zero asymptotic performance, *except for the asymptotic optimal weighting*.
- When v_2 is large, asymptotic optimal weighting performs similarly to using only $\boldsymbol{Y}_1$.
- Asymptotic optimal weights naturally transition from using $\boldsymbol{Y}_2$ to largely ignoring $\boldsymbol{Y}_2$ without any tuning or manual choice of a "cutoff."
- Unweighted PCA and inverse noise variance weighted PCA sometimes perform worse when given more data; in some cases using only $\boldsymbol{Y}_1$ or $\boldsymbol{Y}_2$ was better. In contrast, asymptotic optimal weighting always performs better than using only $\boldsymbol{Y}_1$ or only $\boldsymbol{Y}_2$. Moreover, it always uses all data blocks, i.e., all weights are nonzero. With optimal weighting, more data can only help, it never hurts.

Figure 4.2 overlaid the asymptotic performance for each weighting scheme. For optimally weighted PCA, this limit was given in (2.6). The following proposition provides an analogous result for the existing weighting schemes.

Proposition 4.1 (asymptotic performance of the existing weighting schemes). *The weighting schemes* (4.1)–(4.4) *have corresponding performance converging as*

$$\text{(4.5)}\qquad \text{inverse noise variance}: \quad r_i(\boldsymbol{w},\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \max\left(0, \frac{c - \bar{v}^2/\lambda_i^2}{c + \bar{v}/\lambda_i}\right),$$

$$\text{(4.6)}\qquad \text{only use } \boldsymbol{Y}_1: \quad r_i(\boldsymbol{w},\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \max\left(0, \frac{c_1 - v_1^2/\lambda_i^2}{c_1 + v_1/\lambda_i}\right),$$

$$\text{(4.7)}\qquad \text{only use } \boldsymbol{Y}_2: \quad r_i(\boldsymbol{w},\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \max\left(0, \frac{c_2 - v_2^2/\lambda_i^2}{c_2 + v_2/\lambda_i}\right),$$

$$\text{(4.8)}\qquad \text{unweighted}: \quad r_i(\boldsymbol{w},\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \max\left(0, \frac{A(\beta_i)}{\beta_i B_i'(\beta_i)}\right),$$

where $c := c_1 + \cdots + c_L$, $\bar{v} := (p_1/v_1 + \cdots + p_L/v_L)^{-1}$, $p_\ell := c_\ell/c$,

$$A(x) := 1 - \sum_{\ell=1}^{L} \frac{c_\ell v_\ell^2}{(x - v_\ell)^2}, \qquad B_i(x) := 1 - \lambda_i \sum_{\ell=1}^{L} \frac{c_\ell}{x - v_\ell},$$

and β_i *is the largest real root of the rational function* B_i.

Proposition 4.1 is a by-product of Lemma 6.2 in our proof of Theorem 2.1, with some parts shown previously and some shown in this paper. Specifically, (4.6) and (4.7) are exactly the well-studied homoscedastic case [22, 23, 35, 37], since the noise is homoscedastic when using only $\boldsymbol{Y}_1$ or $\boldsymbol{Y}_2$. For unweighted PCA (4.8), [17] derived the performance for the case where $A(\beta_i) > 0$ and conjectured the behavior for $A(\beta_i) \le 0$. Finally, for the performance of inverse noise variance weights (4.5), closely related results were contemporaneously derived in the recent work [29].

4.3. Comparison of phase transitions. The asymptotic performances (2.6) and (4.5)–(4.8) of the various weighting schemes exhibit phase transitions between settings with zero asymptotic performance and those with nonzero asymptotic performance. Namely, each scheme has nonzero asymptotic performance for data parameters $\boldsymbol{c}$, $\boldsymbol{v}$, and λ_i in an associated regime. Figure 4.3 compares these regimes.

We make the following observations from Figure 4.3:

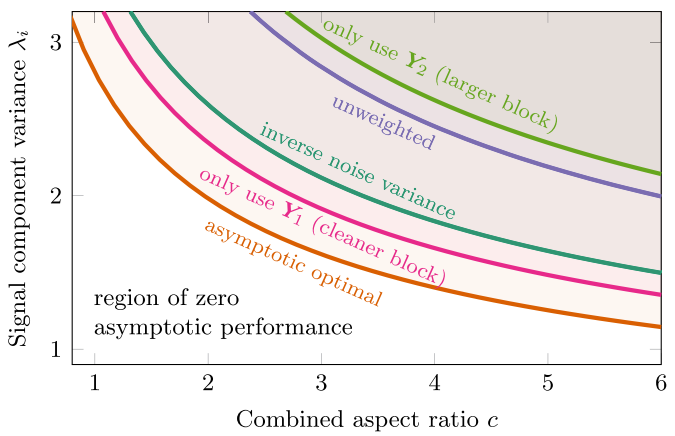
- None of the existing schemes dominate the rest with respect to nonzero asymptotic performance. In some cases one is better than another.
- Asymptotic optimal weighting dominates all of them. Whenever nonzero asymptotic performance is possible for one of the existing schemes, it is also possible for asymptotic optimal weighting.
- Asymptotic optimal weighting also achieves nonzero asymptotic performance in settings where all of the existing schemes have zero asymptotic performance.

For optimally weighted PCA, the condition defining the regime is given in Theorem 2.1. The following proposition gives analogous conditions for the existing weighting schemes and follows straightforwardly from Proposition 4.1.

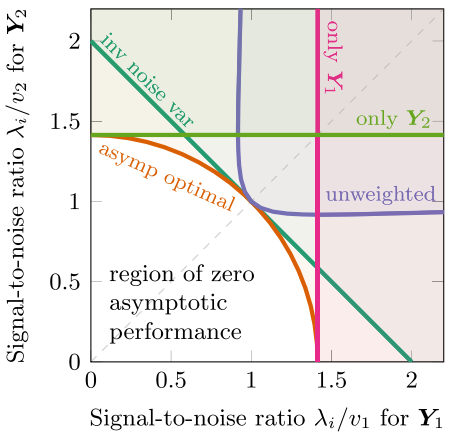
Proposition 4.2 (phase transitions of existing weighting schemes). *The asymptotic performance of the weighting schemes are nonzero if and only if, respectively,*

$$\begin{aligned}
&\text{inverse noise variance:} && c\ \cdot(\lambda_i/\bar{v}\ )^2 > 1,\\
&\text{only use } \boldsymbol{Y}_1\text{:} && c_1\cdot(\lambda_i/v_1)^2 > 1,\\
&\text{only use } \boldsymbol{Y}_2\text{:} && c_2\cdot(\lambda_i/v_2)^2 > 1,\\
&\text{unweighted :} && A(\beta_i) > 0,
\end{aligned}$$

where c, $\bar{v}$, A, and β_i are as in Proposition 4.1.



(a) With respect to combined aspect ratio c for entire dataset $\boldsymbol{Y}$ and signal component variance λ_i, for blocks with associated aspect ratios $\boldsymbol{c} = c\cdot(1/11, 10/11)$ and variances $\boldsymbol{v} = (1, 5)$.



(b) With respect to signal-to-noise ratios λ_i/v_1 and λ_i/v_2, for blocks with associated aspect ratios $\boldsymbol{c} = (1/2, 1/2)$.

Figure 4.3. *Comparison of phase transitions from Theorem 2.1 and Proposition 4.2 for an illustrative example of two blocks of data $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ with variances v_1 and v_2 and signal component variance λ_i. For each weighting scheme, asymptotic performance is zero below the phase transition and nonzero above it. In all cases, the shading goes up and to the right.*

5. Comparison with additional methods. This section compares the performance of optimally weighted PCA with additional PCA methods designed for some form of heteroscedastic noise. Specifically, we consider the following iterative methods:

- HePPCAT [19] is a probabilistic PCA approach that accounts for noise with samplewise heteroscedasticity by modeling the heteroscedasticity in the statistical likelihood. We used 1000 iterations.
- HeteroPCA [50] addresses the bias in the diagonal of the covariance matrix caused by noise with featurewise heteroscedasticity. It does so by iteratively replacing the biased entries using low-rank approximation. We used 100 iterations.

Figure 5.1 shows the nonasymptotic distribution of performance for these methods for the setup considered in subsection 4.2: $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ are generated from the model (2.4) with Gaussian coefficients and noise, and a single signal component $\boldsymbol{u}_1 \in \mathbb{R}^d$ drawn uniformly from the unit sphere. The signal component variance is $\lambda_1 = 2$, and the data sizes are $d = 10^3$, $n_1 = 10^3$, and $n_2 = 10^4$. The first noise variance is $v_1 = 1$ and the second noise variance ranges from $v_2 = 1$ to $v_2 = 20$.

We make the following observations from Figure 5.1:

- HePPCAT accounts for the heterogeneous quality of the data blocks, and it performs very similarly to optimally weighted PCA in this case (the curves overlap). Note that HePPCAT involves solving a nonconvex optimization problem, and it currently lacks a guarantee of convergence to a global optimizer. In contrast, optimally weighted PCA can be computed simply and reliably via the well-studied singular value decomposition.
- HeteroPCA is also designed to account for heteroscedastic noise, but does so primarily for featurewise heteroscedasticity. It treats the samples uniformly and performs very similarly to unweighted PCA in this case (the curves overlap); it performs worse than optimally weighted PCA and is even eventually worse than using only $\boldsymbol{Y}_1$. This example highlights how samplewise and featurewise heteroscedasticity in the noise differ. They have qualitatively different impacts that seem to call for distinct approaches.

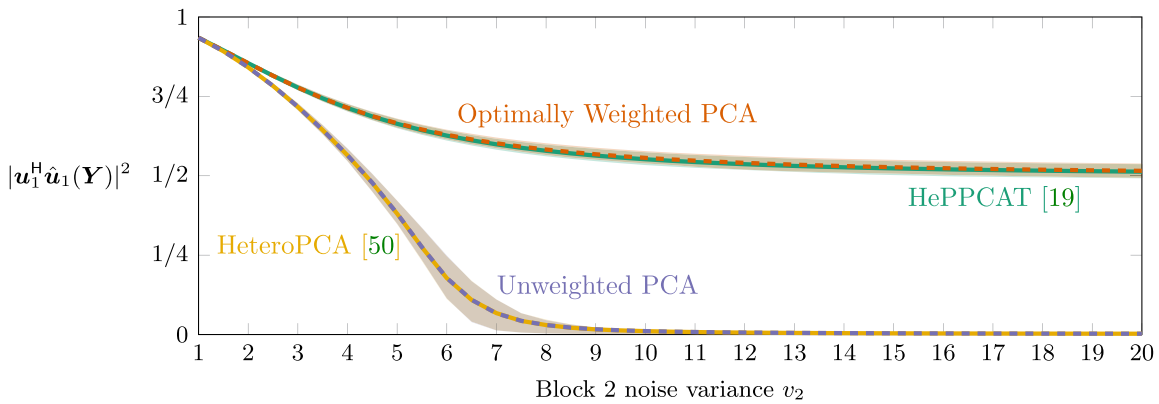


Figure 5.1. *Performance comparison with additional methods for the illustrative example in Figure 4.2: two blocks of data $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$ with $d = 10^3$, $n_1 = 10^3$, $n_2 = 10^4$, and signal component variance $\lambda_1 = 2$. The first block has noise variance $v_1 = 1$, while the second block noise variance ranges from $v_2 = 1$ to $v_2 = 20$. For each method, the colored curve is the average from 400 trials, and the ribbon indicates the corresponding interquartile interval.*

6. Proof of main result. Theorem 2.1 states that unless $\sum_{\ell=1}^{L} c_\ell(\lambda_i/v_\ell)^2 \le 1$, the optimal weights $\boldsymbol{w}_i^\star(\boldsymbol{Y})$ and corresponding optimal performance $r_i^\star(\boldsymbol{Y})$ for the ith component converge almost surely to $\bar{\boldsymbol{w}}_i^\star$ (up to scale) and $\bar{r}_i^\star$ in the right-hand sides of (2.5) and (2.6).

Namely, $\bar{\boldsymbol{w}}_i^\star$ and $\bar{r}_i^\star$ are the result of first optimizing (with respect to the weights $\boldsymbol{w}$) then taking the limit (as the data $\boldsymbol{Y}$ grows in size). Unfortunately, $\boldsymbol{w}_i^\star(\boldsymbol{Y})$ and $r_i^\star(\boldsymbol{Y})$ are complicated functions, making it challenging to directly analyze their limit. So, we instead first take the limit then optimize. More precisely, using aslim to denote almost sure limits and writing $\boldsymbol{Y}^{(d)}$ to make the limits more explicit, we first derive

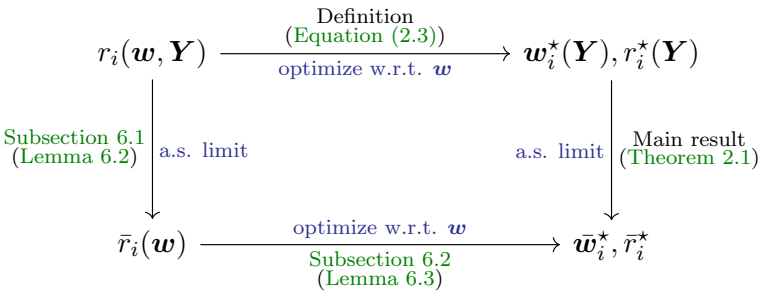
$$\operatorname*{argmax}_{\boldsymbol{w}} \bar{r}_i(\boldsymbol{w}), \quad \text{where} \quad \bar{r}_i(\boldsymbol{w}) = \operatorname*{aslim}_{d\to\infty} r_i(\boldsymbol{w}, \boldsymbol{Y}^{(d)}),$$

then use that result to obtain the result we want, i.e.,

$$\operatorname*{aslim}_{d\to\infty} \boldsymbol{w}_i^\star(\boldsymbol{Y}^{(d)}), \quad \text{where} \quad \boldsymbol{w}_i^\star(\boldsymbol{Y}^{(d)}) = \operatorname*{argmax}_{\boldsymbol{w}} r_i(\boldsymbol{w}, \boldsymbol{Y}^{(d)}).$$

The following diagram illustrates the approach:

(6.1)



For this approach to work, the diagram must commute, i.e., the optimizer of the almost sure limit (which we derive) must match the almost sure limit of the optimizer (which we want). The following lemma states a suitable sufficient condition under which this happens, i.e., the maximizer of the limit is the limit of the maximizer. This lemma may be proved using techniques and results from variational analysis, e.g., [40, Chapter 7]. For convenience, supplementary material SM2 also provides an elementary, self-contained, and concise proof.

Lemma 6.1 (diagram commutes). *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be compact, $f_n : \mathcal{X} \to \mathbb{R}$ be a sequence of functions, and $f : \mathcal{X} \to \mathbb{R}$ be a function, such that on $\mathcal{X}$,*

1. *each f_n has a maximum,*
2. *f_n converges uniformly to f,*
3. *f is continuous, and*
4. *f has a unique maximizer.*

Then the maximum of f_n and the set of maximizers of f_n both converge, i.e.,

$$\max_{\boldsymbol{x}\in\mathcal{X}} f_n(\boldsymbol{x}) \to \max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}), \qquad \operatorname*{argmax}_{\boldsymbol{x}\in\mathcal{X}} f_n(\boldsymbol{x}) \to \operatorname*{argmax}_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}),$$

where the set convergence is with respect to the Hausdorff distance.

With Lemma 6.1 in hand, it now remains to (a) derive the almost sure limit $\bar{r}_i(\boldsymbol{w})$ of $r_i(\boldsymbol{w}, \boldsymbol{Y})$, and show that the convergence is uniform in $\boldsymbol{w}$ (Lemma 6.2 in subsection 6.1) and (b) optimize $\bar{r}_i(\boldsymbol{w})$ and show that the optimizer is unique (up to scaling) except when $\sum_{\ell=1}^{L} c_\ell (\lambda_i / v_\ell)^2 \leq 1$ (Lemma 6.3 in subsection 6.2).

6.1. Almost sure limit of the performance. This section derives the almost sure limit of the performance $r_i(\boldsymbol{w}, \boldsymbol{Y})$, where the convergence is uniform in the weights $\boldsymbol{w}$.

Lemma 6.2 (almost sure limit of the performance). *For $i = 1, \ldots, k$,*

$$r_i(\boldsymbol{w}, \boldsymbol{Y}) \quad \xrightarrow{\text{a.s.}} \quad \bar{r}_i(\boldsymbol{w}) := \max\left(0, \frac{1}{\beta_{i,\boldsymbol{w}}} \frac{A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}\right), \tag{6.2}$$

where the convergence is uniform with respect to $\boldsymbol{w}$ on $\mathbb{R}^L_{\geq 0} \setminus \{\boldsymbol{0}_L\}$,

$$A_{\boldsymbol{w}}(x) := 1 - \sum_{\ell=1}^{L} \frac{c_\ell w_\ell^2 v_\ell^2}{(x - w_\ell v_\ell)^2}, \qquad B_{i,\boldsymbol{w}}(x) := 1 - \lambda_i \sum_{\ell=1}^{L} \frac{c_\ell w_\ell}{x - w_\ell v_\ell}, \tag{6.3}$$

and $\beta_{i,\boldsymbol{w}}$ is the largest real root of $B_{i,\boldsymbol{w}}$.

The remainder of this subsection proves Lemma 6.2. After defining some notations, we derive the limit of the singular values, then derive the limit of $r_i(\boldsymbol{w}, \boldsymbol{Y})$ in two regimes (above and below the phase transition), and finally derive the above algebraic form. There are several other ways to structure these derivations; see, e.g., [7]. The approach we take here carefully combines the general perturbation approach of [5] with celebrated random matrix theoretic results on local laws [13, 27, 48] (reviewed in supplementary material SM3.1) to obtain uniform convergence for the singular values and vectors. The algebraic form is derived following the approach in [17]. Throughout the proof, we postpone some detailed calculations to the supplement (supplement.pdf [local/web 475KB]).

6.1.1. Notation and preliminaries. Let $\hat{\theta}_{i,\boldsymbol{w}}$, $\hat{\boldsymbol{u}}_{i,\boldsymbol{w}}$, and $\hat{\boldsymbol{q}}_{i,\boldsymbol{w}}$ denote, respectively, the ith singular value, ith left singular vector, and ith right singular vector of the normalized and weighted data matrix,

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{w}} := \frac{1}{\sqrt{n}} [\sqrt{w_1}\, \boldsymbol{Y}_1, \ldots, \sqrt{w_L}\, \boldsymbol{Y}_L] = \boldsymbol{U} \boldsymbol{\Theta} \widetilde{\boldsymbol{Q}}_{\boldsymbol{w}}^{\mathsf{H}} + \widetilde{\boldsymbol{E}}_{\boldsymbol{w}}, \tag{6.4}$$

where $\boldsymbol{U} := [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$, $\boldsymbol{\Theta} := \operatorname{diag}(\theta_1, \ldots, \theta_k)$ has diagonal entries $\theta_i := \sqrt{\lambda_i}$, and

$$\widetilde{\boldsymbol{Q}}_{\boldsymbol{w}} := \frac{1}{\sqrt{n}} [\sqrt{w_1}\, \boldsymbol{Z}_1, \ldots, \sqrt{w_L}\, \boldsymbol{Z}_L]^{\mathsf{H}}, \qquad \widetilde{\boldsymbol{E}}_{\boldsymbol{w}} := \frac{1}{\sqrt{n}} [\sqrt{w_1}\, \boldsymbol{E}_1, \ldots, \sqrt{w_L}\, \boldsymbol{E}_L],$$

are normalized and weighted coefficients and noise. We indicate that these are functions of the weights via the subscript $\boldsymbol{w}$, and omit the dependence of the singular values and vectors on the data (for brevity). We also omit the index for d; all limits are as $d \to \infty$ unless otherwise specified.

Note that the left singular vectors of $\widetilde{\boldsymbol{Y}}_{\boldsymbol{w}}$ are exactly the eigenvectors of the weighted sample covariance, so the performance can be written as $r_i(\boldsymbol{w}, \boldsymbol{Y}) = |\boldsymbol{u}_i^{\mathsf{H}} \hat{\boldsymbol{u}}_{i,\boldsymbol{w}}|^2$. The noise

matrix $\widetilde{\boldsymbol{E}}_{\boldsymbol{w}}$ satisfies the usual random matrix theoretic conditions, and is well known to have a singular value distribution that converges weakly almost surely to a nonrandom compactly supported measure $\mu_{\boldsymbol{w}}$ whose Stieltjes transform

$$m_{\mathrm{MP}}(\boldsymbol{w},\zeta) := \int \frac{1}{\zeta^2 - t^2} \mathrm{d}\mu_{\boldsymbol{w}}(t) \tag{6.5}$$

is the unique solution to the generalized Marchenko–Pastur equation

$$\frac{1}{m_{\mathrm{MP}}(\boldsymbol{w},\zeta)} = \zeta^2 - \sum_{\ell=1}^{L} \frac{p_\ell w_\ell v_\ell}{1 - w_\ell v_\ell m_{\mathrm{MP}}(\boldsymbol{w},\zeta)/c} \tag{6.6}$$

for which $\operatorname{Im} m_{\mathrm{MP}}(\boldsymbol{w},\zeta) < 0$ for ζ^2 in the upper half complex plane [33], where $c := c_1 + \cdots + c_L$ and $p_\ell := c_\ell / c$.

Moreover, the operator norm of the noise matrix converges to the upper-edge of $\mu_{\boldsymbol{w}}$ (see supplementary material SM3.2 for detailed derivation), i.e.,

$$\|\widetilde{\boldsymbol{E}}_{\boldsymbol{w}}\|_{\mathrm{op}} \underset{\boldsymbol{w}\in\Delta_L}{\overset{\text{a.s.}}{\rightrightarrows}} b_{\boldsymbol{w}} := \sup\{\text{support of } \mu_{\boldsymbol{w}}\}, \tag{6.7}$$

where $\Delta_L := \{\boldsymbol{w} \in \mathbb{R}^L_{\geq 0} : w_1 + \cdots + w_L = 1\}$ is the probability simplex, and $\overset{\text{a.s.}}{\rightrightarrows}$ denotes almost sure uniform convergence.

6.1.2. Limits of the singular values. Using a similar argument as [5, section 4], one can show that any singular value of $\widetilde{\boldsymbol{Y}}_{\boldsymbol{w}}$ that does not tend to a limit greater than $b_{\boldsymbol{w}}$ tends to $b_{\boldsymbol{w}}$, so we focus on singular values with limits greater than $b_{\boldsymbol{w}}$. Following the approach of [5, section 4], this section studies those singular values through the matrix-valued function

$$\boldsymbol{M}(\boldsymbol{w},\zeta) := \begin{bmatrix} \boldsymbol{U} & \\ & \widetilde{\boldsymbol{Q}}_{\boldsymbol{w}} \end{bmatrix}^{\mathsf{H}} \boldsymbol{G}(\boldsymbol{w},\zeta) \begin{bmatrix} \boldsymbol{U} & \\ & \widetilde{\boldsymbol{Q}}_{\boldsymbol{w}} \end{bmatrix} - \begin{bmatrix} & \boldsymbol{\Theta}^{-1} \\ \boldsymbol{\Theta}^{-1} & \end{bmatrix}, \tag{6.8}$$

where $\boldsymbol{G}(\boldsymbol{w},\zeta)$ is the resolvent (or Green's function) defined as

$$\boldsymbol{G}(\boldsymbol{w},\zeta) := \left(\zeta \boldsymbol{I}_{d+n} - \begin{bmatrix} & \widetilde{\boldsymbol{E}}_{\boldsymbol{w}} \\ \widetilde{\boldsymbol{E}}_{\boldsymbol{w}}^{\mathsf{H}} & \end{bmatrix} \right)^{-1}. \tag{6.9}$$

The link to singular values is made through an extension of [5, Lemma 4.1] that incorporates the weights (see supplementary material SM3.3 for detailed derivation). It states that

$$\forall_{\zeta > \|\widetilde{\boldsymbol{E}}_{\boldsymbol{w}}\|_{\mathrm{op}}} \quad \zeta \text{ is a singular value of } \widetilde{\boldsymbol{Y}}_{\boldsymbol{w}} \iff \det \boldsymbol{M}(\boldsymbol{w},\zeta) = 0, \tag{6.10}$$

so we instead study $\boldsymbol{M}$ in the limit. A careful application of anisotropic local laws [13, 27, 48] (see supplementary material SM3.4 for detailed calculations) yields that for any $\tau > 0$,

$$\boldsymbol{M}(\boldsymbol{w},\zeta) \underset{(\boldsymbol{w},\zeta)\in\Omega(\tau)}{\overset{\text{a.s.}}{\rightrightarrows}} \overline{\boldsymbol{M}}(\boldsymbol{w},\zeta) := \begin{bmatrix} \varphi_{1,\boldsymbol{w}}(\zeta)\boldsymbol{I}_k & \\ & \varphi_{2,\boldsymbol{w}}(\zeta)\boldsymbol{I}_k \end{bmatrix} - \begin{bmatrix} & \boldsymbol{\Theta}^{-1} \\ \boldsymbol{\Theta}^{-1} & \end{bmatrix}, \tag{6.11}$$

where $\Omega(\tau) := \{(\boldsymbol{w}, \zeta) \in \Delta_L \times \mathbb{C} : (\mathrm{Re}\zeta, \mathrm{Im}\zeta) \in [b_{\boldsymbol{w}} + \tau, \infty) \times [-1, 1]\}$ and

$$\varphi_{1,\boldsymbol{w}}(\zeta) := \zeta m_{\mathrm{MP}}(\boldsymbol{w}, \zeta) = \int \frac{\zeta}{\zeta^2 - t^2} \, \mathrm{d}\mu_{\boldsymbol{w}}(t), \qquad \varphi_{2,\boldsymbol{w}}(\zeta) := \sum_{\ell=1}^{L} \frac{p_\ell w_\ell}{\zeta - w_\ell v_\ell \varphi_{1,\boldsymbol{w}}(\zeta)/c}. \tag{6.12}$$

Finally, we apply [5, Lemma A.1] with (6.10) and (6.11) in the same way as [5, section 4]; straightforward calculations (see supplementary material SM3.5) verify that $\varphi_{1,\boldsymbol{w}}$ and $\varphi_{2,\boldsymbol{w}}$ satisfy the conditions of [5, Lemma A.1]. Moreover, noting that these arguments extend to uniform convergence in $\boldsymbol{w} \in \Delta_L$ yields that

$$\hat{\theta}_{i,\boldsymbol{w}} \underset{\boldsymbol{w} \in \Delta_L}{\overset{\text{a.s.}}{\rightrightarrows}} \bar{\theta}_{i,\boldsymbol{w}}, \quad \text{where} \quad \bar{\theta}_{i,\boldsymbol{w}} := \begin{cases} \rho_{i,\boldsymbol{w}} & \text{if } \theta_i > \tilde{\theta}_{\boldsymbol{w}}, \\ b_{\boldsymbol{w}} & \text{otherwise,} \end{cases} \tag{6.13}$$

where $D_{\boldsymbol{w}}(\zeta) := \varphi_{1,\boldsymbol{w}}(\zeta)\varphi_{2,\boldsymbol{w}}(\zeta)$ for $\zeta > b_{\boldsymbol{w}}$, $\rho_{i,\boldsymbol{w}} := D_{\boldsymbol{w}}^{-1}(1/\theta_i^2)$, and $\tilde{\theta}_{\boldsymbol{w}}^2 := 1/D_{\boldsymbol{w}}(b_{\boldsymbol{w}}^+)$. Here $D_{\boldsymbol{w}}^{-1}$ denotes the inverse function of $D_{\boldsymbol{w}}$, and $f(b^+) := \lim_{\zeta \to b^+} f(\zeta)$ is the limit from above.

6.1.3. Performance above the phase transition. This section derives the limit of the performance $r_i(\boldsymbol{w}, \boldsymbol{Y})$ above the phase transition. Namely, for any $\nu > 0$, we prove uniform convergence with respect to $\boldsymbol{w}$ over the domain $\mathcal{W}_>(\nu) := \{\boldsymbol{w} \in \Delta_L : \bar{\theta}_{i,\boldsymbol{w}} > b_{\boldsymbol{w}} + \nu\}$.

Following the approach of [5, section 5], we study $r_i(\boldsymbol{w}, \boldsymbol{Y}) = |\boldsymbol{u}_i^{\mathsf{H}} \hat{\boldsymbol{u}}_{i,\boldsymbol{w}}|^2$ through the following extension of [5, Lemma 5.1] (see supplementary material SM3.6 for detailed derivation):

$$\forall_{\boldsymbol{w} \in \Delta_L} \quad \hat{\theta}_{i,\boldsymbol{w}} > \|\widetilde{\boldsymbol{E}}_{\boldsymbol{w}}\|_{\mathrm{op}} \quad \Longrightarrow \quad \left\{ \; 0 = \boldsymbol{M}(\boldsymbol{w}, \hat{\theta}_{i,\boldsymbol{w}}) \begin{bmatrix} \boldsymbol{\Theta} \widetilde{\boldsymbol{Q}}_{\boldsymbol{w}}^{\mathsf{H}} \hat{\boldsymbol{q}}_{i,\boldsymbol{w}} \\ \boldsymbol{\Theta} \boldsymbol{U}^{\mathsf{H}} \hat{\boldsymbol{u}}_{i,\boldsymbol{w}} \end{bmatrix} \text{ and} \tag{6.14a}$$

$$1 = \chi_1(\boldsymbol{w}) + \chi_2(\boldsymbol{w}) + 2\,\mathrm{Re}\,\chi_3(\boldsymbol{w}) \; \right\}, \tag{6.14b}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{w}} := (\hat{\theta}_{i,\boldsymbol{w}}^2 \boldsymbol{I}_d - \widetilde{\boldsymbol{E}}_{\boldsymbol{w}} \widetilde{\boldsymbol{E}}_{\boldsymbol{w}}^{\mathsf{H}})^{-1}$ and

$$\begin{aligned} \chi_1(\boldsymbol{w}) &:= \sum_{j_1, j_2 = 1}^{k} \theta_{j_1} \theta_{j_2} \cdot (\tilde{\boldsymbol{q}}_{j_1,\boldsymbol{w}}^{\mathsf{H}} \hat{\boldsymbol{q}}_{i,\boldsymbol{w}}) \cdot (\tilde{\boldsymbol{q}}_{j_2,\boldsymbol{w}}^{\mathsf{H}} \hat{\boldsymbol{q}}_{i,\boldsymbol{w}})^* \cdot \boldsymbol{u}_{j_2}^{\mathsf{H}} \hat{\theta}_{i,\boldsymbol{w}}^2 \boldsymbol{\Gamma}_{\boldsymbol{w}}^2 \boldsymbol{u}_{j_1}, \\ \chi_2(\boldsymbol{w}) &:= \sum_{j_1, j_2 = 1}^{k} \theta_{j_1} \theta_{j_2} \cdot \quad (\boldsymbol{u}_{j_1}^{\mathsf{H}} \hat{\boldsymbol{u}}_{i,\boldsymbol{w}}) \cdot \quad (\boldsymbol{u}_{j_2}^{\mathsf{H}} \hat{\boldsymbol{u}}_{i,\boldsymbol{w}})^* \cdot \tilde{\boldsymbol{q}}_{j_2,\boldsymbol{w}}^{\mathsf{H}} \widetilde{\boldsymbol{E}}_{\boldsymbol{w}}^{\mathsf{H}} \boldsymbol{\Gamma}_{\boldsymbol{w}}^2 \widetilde{\boldsymbol{E}}_{\boldsymbol{w}} \tilde{\boldsymbol{q}}_{j_1,\boldsymbol{w}}, \\ \chi_3(\boldsymbol{w}) &:= \sum_{j_1, j_2 = 1}^{k} \theta_{j_1} \theta_{j_2} \cdot \quad (\boldsymbol{u}_{j_1}^{\mathsf{H}} \hat{\boldsymbol{u}}_{i,\boldsymbol{w}}) \cdot (\tilde{\boldsymbol{q}}_{j_2,\boldsymbol{w}}^{\mathsf{H}} \hat{\boldsymbol{q}}_{i,\boldsymbol{w}})^* \cdot \boldsymbol{u}_{j_2}^{\mathsf{H}} \hat{\theta}_{i,\boldsymbol{w}} \boldsymbol{\Gamma}_{\boldsymbol{w}}^2 \widetilde{\boldsymbol{E}}_{\boldsymbol{w}} \tilde{\boldsymbol{q}}_{j_1,\boldsymbol{w}}. \end{aligned} \tag{6.15}$$

It follows from (6.13) that almost surely, eventually, for all $\boldsymbol{w} \in \mathcal{W}_>(\nu)$, $\hat{\theta}_{i,\boldsymbol{w}} > \|\widetilde{\boldsymbol{E}}_{\boldsymbol{w}}\|_{\mathrm{op}}$ and hence the condition of (6.14) holds. It remains to study the limits of χ_1, χ_2, and χ_3.

Carefully applying (6.11) and (6.13) to (6.14a) in a similar way as [5, section 5] yields that (see supplementary material SM3.7 for detailed calculations), for any $\nu > 0$,

$$\chi_1(\boldsymbol{w}) - \theta_i^2\left[+\frac{\varphi_{1,\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})}{2\rho_{i,\boldsymbol{w}}} - \frac{\varphi'_{1,\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})}{2}\right]|\boldsymbol{u}_i^{\mathsf{H}}\hat{\boldsymbol{u}}_{i,\boldsymbol{w}}|^2\frac{\varphi_{2,\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})}{\varphi_{1,\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})} \underset{\boldsymbol{w}\in\mathcal{W}_>(\nu)}{\overset{\text{a.s.}}{\rightrightarrows}} 0, \tag{6.16a}$$

$$\chi_2(\boldsymbol{w}) - \theta_i^2\left[-\frac{\varphi_{2,\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})}{2\rho_{i,\boldsymbol{w}}} - \frac{\varphi'_{2,\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})}{2}\right]|\boldsymbol{u}_i^{\mathsf{H}}\hat{\boldsymbol{u}}_{i,\boldsymbol{w}}|^2 \underset{\boldsymbol{w}\in\mathcal{W}_>(\nu)}{\overset{\text{a.s.}}{\rightrightarrows}} 0, \tag{6.16b}$$

$$\chi_3(\boldsymbol{w}) \underset{\boldsymbol{w}\in\mathcal{W}_>(\nu)}{\overset{\text{a.s.}}{\rightrightarrows}} 0. \tag{6.16c}$$

Finally, applying (6.16) to (6.14b) yields that (see supplementary material SM3.8 for detailed calculations), for any $\nu > 0$,

$$r_i(\boldsymbol{w},\boldsymbol{Y}) = |\boldsymbol{u}_i^{\mathsf{H}}\hat{\boldsymbol{u}}_{i,\boldsymbol{w}}|^2 \underset{\boldsymbol{w}\in\mathcal{W}_>(\nu)}{\overset{\text{a.s.}}{\rightrightarrows}} -\frac{2\varphi_{1,\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})}{\theta_i^2 D'_{\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})}. \tag{6.17}$$

6.1.4. Performance below the phase transition. This section bounds the limit of the performance $r_i(\boldsymbol{w},\boldsymbol{Y})$ below the phase transition. Namely, for any $\nu > 0$, we derive a uniform bound with respect to $\boldsymbol{w}$ over the domain $\mathcal{W}_\le(\nu) := \Delta_L \setminus \mathcal{W}_>(\nu) = \{\boldsymbol{w}\in\Delta_L : \bar{\theta}_{i,\boldsymbol{w}} \le b_{\boldsymbol{w}} + \nu\}$.

Following the approach of [7], we study $r_i(\boldsymbol{w},\boldsymbol{Y}) = |\boldsymbol{u}_i^{\mathsf{H}}\hat{\boldsymbol{u}}_{i,\boldsymbol{w}}|^2$ by first obtaining the following deterministic bound (see supplementary material SM3.9 for detailed calculations):

$$|\boldsymbol{u}_i^{\mathsf{H}}\hat{\boldsymbol{u}}_{i,\boldsymbol{w}}|^2 \le -\nu\cdot\operatorname{Im}\left\{\zeta_{i,\boldsymbol{w}}^{-1}\left[\widetilde{\boldsymbol{M}}(\boldsymbol{w},\zeta_{i,\boldsymbol{w}}) - \widetilde{\boldsymbol{M}}(\boldsymbol{w},\zeta_{i,\boldsymbol{w}})\left[\boldsymbol{M}(\boldsymbol{w},\zeta_{i,\boldsymbol{w}})\right]^{-1}\widetilde{\boldsymbol{M}}(\boldsymbol{w},\zeta_{i,\boldsymbol{w}})\right]_{ii}\right\}, \tag{6.18}$$

where $\zeta_{i,\boldsymbol{w}}^2 := \hat{\theta}_{i,\boldsymbol{w}}^2 + \imath\nu$ and

$$\widetilde{\boldsymbol{M}}(\boldsymbol{w},\zeta) := \begin{bmatrix}\boldsymbol{U} & \\ & \widetilde{\boldsymbol{Q}}_{\boldsymbol{w}}\end{bmatrix}^{\mathsf{H}}\boldsymbol{G}(\boldsymbol{w},\zeta)\begin{bmatrix}\boldsymbol{U} & \\ & \widetilde{\boldsymbol{Q}}_{\boldsymbol{w}}\end{bmatrix}. \tag{6.19}$$

Next, by standard calculations (see supplementary material SM3.10), we have that almost surely, eventually, the following bounds hold for all $\boldsymbol{w}\in\mathcal{W}_\le(\nu)$:

$$|\zeta_{i,\boldsymbol{w}}^{-1}| \le \widetilde{C}_1, \qquad \left\|\widetilde{\boldsymbol{M}}(\boldsymbol{w},\zeta_{i,\boldsymbol{w}})\right\|_{\text{op}} \le \widetilde{C}_2, \qquad \|\boldsymbol{M}(\boldsymbol{w},\zeta_{i,\boldsymbol{w}})^{-1}\|_{\text{op}} \le \widetilde{C}_3\nu^{-1/2}, \tag{6.20}$$

where $\widetilde{C}_1$, $\widetilde{C}_2$, and $\widetilde{C}_3$ do not depend on ν or $\boldsymbol{w}$.

Finally, applying (6.20) to (6.18) yields that for any $\nu > 0$, almost surely, eventually,

$$\sup_{\boldsymbol{w}\in\mathcal{W}_\le(\nu)} r_i(\boldsymbol{w},\boldsymbol{Y}) = \sup_{\boldsymbol{w}\in\mathcal{W}_\le(\nu)} |\boldsymbol{u}_i^{\mathsf{H}}\hat{\boldsymbol{u}}_{i,\boldsymbol{w}}|^2 \le \widetilde{C}_4(\nu + \nu^{1/2}), \tag{6.21}$$

where $\widetilde{C}_4 := \widetilde{C}_1\max(\widetilde{C}_2, \widetilde{C}_2^2\widetilde{C}_3)$ does not depend on ν.

6.1.5. Uniform convergence and algebraic form of performance. Noting that $\nu > 0$ can be arbitrarily small in (6.17) and (6.21) yields uniform convergence across $\boldsymbol{w}\in\Delta_L$, i.e.,

$$r_i(\boldsymbol{w},\boldsymbol{Y}) \underset{\boldsymbol{w}\in\Delta_L}{\overset{\text{a.s.}}{\rightrightarrows}} \bar{r}_i(\boldsymbol{w}) := \begin{cases} -\frac{2\varphi_{1,\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})}{\theta_i^2 D'_{\boldsymbol{w}}(\rho_{i,\boldsymbol{w}})} & \text{if } \theta_i > \tilde{\theta}_{\boldsymbol{w}}, \\ 0 & \text{otherwise.}\end{cases} \tag{6.22}$$

Since $r_i(\boldsymbol{w}, \boldsymbol{Y})$ is scale invariant, i.e., $r_i(\gamma \boldsymbol{w}, \boldsymbol{Y}) = r_i(\boldsymbol{w}, \boldsymbol{Y})$ for any $\gamma > 0$, it immediately follows that the convergence is also uniform over $\mathbb{R}^L_{\geq 0} \setminus \{\mathbf{0}_L\}$ as well.

The proof concludes by deriving the algebraic description (6.2) of $\bar{r}_i(\boldsymbol{w})$. Following the approach of [17, section 5.2], we change variables to

$$\psi_{\boldsymbol{w}}(\zeta) := \frac{c\zeta}{\varphi_{1,\boldsymbol{w}}(\zeta)}, \tag{6.23}$$

and observe that, analogously to [17, section 5.3],

$$D_{\boldsymbol{w}}(\zeta) = \frac{1 - B_{i,\boldsymbol{w}}(\psi_{\boldsymbol{w}}(\zeta))}{\theta_i^2}, \qquad \frac{D'_{\boldsymbol{w}}(\zeta)}{\zeta} = -\frac{2c}{\theta_i^2} \frac{B'_{i,\boldsymbol{w}}(\psi_{\boldsymbol{w}}(\zeta))}{A_{\boldsymbol{w}}(\psi_{\boldsymbol{w}}(\zeta))}, \tag{6.24}$$

$\psi_{\boldsymbol{w}}(b^+_{\boldsymbol{w}}) = \alpha_{\boldsymbol{w}}$ and $\psi_{\boldsymbol{w}}(\rho_{i,\boldsymbol{w}}) = \beta_{i,\boldsymbol{w}}$ when $\theta_i^2 > \tilde{\theta}^2_{\boldsymbol{w}}$, where $\alpha_{\boldsymbol{w}}$ and $\beta_{i,\boldsymbol{w}}$ are the largest real roots of $A_{\boldsymbol{w}}$ and $B_{i,\boldsymbol{w}}$, respectively. See supplementary material SM3.11 for the detailed derivations.

Even though $\psi_{\boldsymbol{w}}(\rho_i)$ is defined only when $\theta_i^2 > \tilde{\theta}^2_{\boldsymbol{w}}$, the largest real roots $\alpha_{\boldsymbol{w}}$ and $\beta_{i,\boldsymbol{w}}$ are always defined and always larger than $\max_\ell(w_\ell v_\ell)$. Thus

$$\theta_i^2 > \tilde{\theta}^2_{\boldsymbol{w}} = \frac{\theta_i^2}{1 - B_{i,\boldsymbol{w}}(\psi_{\boldsymbol{w}}(b^+_{\boldsymbol{w}}))} \Leftrightarrow B_{i,\boldsymbol{w}}(\alpha_{\boldsymbol{w}}) < 0 \Leftrightarrow \alpha_{\boldsymbol{w}} < \beta_{i,\boldsymbol{w}} \Leftrightarrow A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) > 0, \tag{6.25}$$

where the final two equivalences hold because $A_{\boldsymbol{w}}(x)$ and $B_{i,\boldsymbol{w}}(x)$ are both strictly increasing functions for $x > \max_\ell(w_\ell v_\ell)$ and $A_{\boldsymbol{w}}(\alpha_{\boldsymbol{w}}) = B_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) = 0$.

Finally, using (6.24) and (6.25) to rewrite (6.22) concludes the proof of Lemma 6.2.

6.2. Optimization of the almost sure limit. This section optimizes $\bar{r}_i(\boldsymbol{w})$ and shows that the maximizer is unique (up to scaling).

Lemma 6.3 (optimization of the almost sure limit). *The asymptotic performance $\bar{r}_i(\boldsymbol{w})$ is continuous and is maximized as*

$$\{\gamma \bar{\boldsymbol{w}}^\star_i : \gamma > 0\} = \operatorname*{argmax}_{\boldsymbol{w} \in \mathbb{R}^L_{\geq 0} \setminus \{\mathbf{0}_L\}} \bar{r}_i(\boldsymbol{w}), \qquad \bar{r}^\star_i = \max_{\boldsymbol{w} \in \mathbb{R}^L_{\geq 0} \setminus \{\mathbf{0}_L\}} \bar{r}_i(\boldsymbol{w}), \tag{6.26}$$

except when $\sum_{\ell=1}^L c_\ell (\lambda_i / v_\ell)^2 \leq 1$, in which case $\bar{r}_i(\boldsymbol{w}) = 0$ for all weights $\boldsymbol{w} \in \mathbb{R}^L_{\geq 0} \setminus \{\mathbf{0}_L\}$.

The remainder of this subsection proves Lemma 6.3. A major challenge for the proof is that $\bar{r}_i(\boldsymbol{w})$ is defined implicitly via the root $\beta_{i,\boldsymbol{w}}$, and setting the gradient equal to zero to obtain local maxima yields a complicated nonlinear system of equations to solve. Surprisingly, the system turns out to have a simple solution that we derive by carefully exploiting the structure of the system. Moreover, we show that the solution is globally optimal, obtaining the optimal weights and their corresponding performance.

Before deriving the gradient, note that $\bar{r}_i(\boldsymbol{w})$ is not always differentiable everywhere due to its truncation at zero. However, it can be rewritten as

$$\bar{r}_i(\boldsymbol{w}) = \max\left(0, \tilde{r}_i(\boldsymbol{w})\right), \quad \text{where} \quad \tilde{r}_i(\boldsymbol{w}) := \frac{1}{\beta_{i,\boldsymbol{w}}} \frac{A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}, \tag{6.27}$$

so the problem reduces to maximizing the differentiable function $\tilde{r}_i(\boldsymbol{w})$ then checking whether it is positive. Furthermore, $\tilde{r}_i(\boldsymbol{w})$ achieves its maximum over the feasible region $\mathbb{R}^L_{\geq 0} \setminus \{\mathbf{0}_L\}$. To see why, note that $\tilde{r}_i(\boldsymbol{w})$ is scale-invariant, i.e.,

$$\forall_{\gamma>0} \quad \tilde{r}_i(\gamma \boldsymbol{w}) = \frac{1}{\beta_{i,\gamma\boldsymbol{w}}} \frac{A_{\gamma\boldsymbol{w}}(\beta_{i,\gamma\boldsymbol{w}})}{B'_{i,\gamma\boldsymbol{w}}(\beta_{i,\gamma\boldsymbol{w}})} = \frac{1}{\gamma\beta_{i,\boldsymbol{w}}} \frac{A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{(1/\gamma)B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})} = \tilde{r}_i(\boldsymbol{w})$$

since $A_{\gamma\boldsymbol{w}}(x) = A_{\boldsymbol{w}}(x/\gamma)$, $B_{i,\gamma\boldsymbol{w}}(x) = B_{i,\boldsymbol{w}}(x/\gamma)$, and $B'_{i,\gamma\boldsymbol{w}}(x) = (1/\gamma)B'_{i,\boldsymbol{w}}(x/\gamma)$, resulting additionally in $\beta_{i,\gamma\boldsymbol{w}} = \gamma\beta_{i,\boldsymbol{w}}$. Thus, $\tilde{r}_i(\boldsymbol{w})$ can equivalently be maximized over the compact set $\Delta_L := \{\boldsymbol{w} \in \mathbb{R}^L_{\geq 0} : w_1 + \cdots + w_L = 1\}$ and hence achieves its maximum.

Next, note that the feasible region $\mathbb{R}^L_{\geq 0} \setminus \{\mathbf{0}_L\}$ is not open, so we partition it into $2^L - 1$ sets according to which weights are zero. Namely, consider partitions of the form

$$\mathcal{P}_{\mathcal{L}} := \{\boldsymbol{w} \in \mathbb{R}^L_{\geq 0} : \forall_{\ell \in \mathcal{L}}\ w_\ell = 0, \forall_{\ell \notin \mathcal{L}}\ w_\ell > 0\} \quad \text{for } \mathcal{L} \subset \{1, \ldots, L\} \text{ a proper subset.}$$

Since $\tilde{r}_i(\boldsymbol{w})$ achieves its maximum, a maximizer exists within at least one of the partitions. Moreover, since $\tilde{r}_i(\boldsymbol{w})$ is differentiable, $\tilde{r}_i(\boldsymbol{w})$ is maximized at a critical point of a partition. It remains to identify and compare the critical points of all the partitions $\mathcal{P}_{\mathcal{L}}$.

First consider $\mathcal{P}_\emptyset$, i.e., the set of positive weights $w_1, \ldots, w_L > 0$, and let $\tilde{w}_j := 1/w_j$. This parameterization ends up simplifying the manipulations. Differentiating $\tilde{r}_i(\boldsymbol{w})$ and (6.3) with respect to $\tilde{w}_j$ yields

$$\frac{\partial \tilde{r}_i(\boldsymbol{w})}{\partial \tilde{w}_j} = \tilde{r}_i(\boldsymbol{w}) \left[-\frac{1}{\beta_{i,\boldsymbol{w}}} \frac{\partial \beta_{i,\boldsymbol{w}}}{\partial \tilde{w}_j} + \frac{1}{A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})} \frac{\partial A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\partial \tilde{w}_j} - \frac{1}{B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})} \frac{\partial B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\partial \tilde{w}_j} \right], \tag{6.28a}$$

$$\frac{\partial A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\partial \tilde{w}_j} = A'_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) \frac{\partial \beta_{i,\boldsymbol{w}}}{\partial \tilde{w}_j} + 2 \frac{c_j v_j^2}{(\beta_{i,\boldsymbol{w}} \tilde{w}_j - v_j)^3} \beta_{i,\boldsymbol{w}}, \tag{6.28b}$$

$$\frac{\partial B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\partial \tilde{w}_j} = B''_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) \frac{\partial \beta_{i,\boldsymbol{w}}}{\partial \tilde{w}_j} - 2\lambda_i \frac{c_j \tilde{w}_j}{(\beta_{i,\boldsymbol{w}} \tilde{w}_j - v_j)^3} \beta_{i,\boldsymbol{w}} + \lambda_i \frac{c_j}{(\beta_{i,\boldsymbol{w}} \tilde{w}_j - v_j)^2}, \tag{6.28c}$$

$$0 = \frac{\partial B_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\partial \tilde{w}_j} = B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) \frac{\partial \beta_{i,\boldsymbol{w}}}{\partial \tilde{w}_j} + \lambda_i \frac{c_j}{(\beta_{i,\boldsymbol{w}} \tilde{w}_j - v_j)^2} \beta_{i,\boldsymbol{w}}, \tag{6.28d}$$

where one must carefully account for the fact that $A_{\boldsymbol{w}}(x)$, $B_{i,\boldsymbol{w}}(x)$, and $\beta_{i,\boldsymbol{w}}$ are functions of $\tilde{w}_j$. The problem now is to set $\partial \tilde{r}_i(\boldsymbol{w})/\partial \tilde{w}_j$ equal to zero and carefully exploit the structure of the system (6.28) to solve it.

In particular, rewriting (6.28b) and (6.28c) in terms of $\partial \beta_{i,\boldsymbol{w}}/\partial \tilde{w}_j$ using (6.28d) yields

$$\frac{\partial A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\partial \tilde{w}_j} = \left[A'_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) - \frac{2B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\lambda_i} \frac{v_j^2}{\beta_{i,\boldsymbol{w}} \tilde{w}_j - v_j} \right] \frac{\partial \beta_{i,\boldsymbol{w}}}{\partial \tilde{w}_j}, \tag{6.29a}$$

$$\frac{\partial B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\partial \tilde{w}_j} = \left[B''_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) + 2B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) \frac{\tilde{w}_j}{\beta_{i,\boldsymbol{w}} \tilde{w}_j - v_j} \right] \frac{\partial \beta_{i,\boldsymbol{w}}}{\partial \tilde{w}_j} - \frac{B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{\beta_{i,\boldsymbol{w}}} \frac{\partial \beta_{i,\boldsymbol{w}}}{\partial \tilde{w}_j}. \tag{6.29b}$$

Substituting (6.29a) and (6.29b) into (6.28a) then rearranging yields

$$\frac{\partial \tilde{r}_i(\boldsymbol{w})}{\partial \tilde{w}_j} = \frac{2}{\lambda_i \beta_{i,\boldsymbol{w}}} \frac{\partial \beta_{i,\boldsymbol{w}}}{\partial \tilde{w}_j} \left[\lambda_i \Delta_{i,\boldsymbol{w}} - \frac{\lambda_i \beta_{i,\boldsymbol{w}} \tilde{r}_i(\boldsymbol{w}) \tilde{w}_j + v_j^2}{\beta_{i,\boldsymbol{w}} \tilde{w}_j - v_j} \right], \tag{6.30}$$

where the following term is independent of j:

$$\Delta_{i,\boldsymbol{w}} := \frac{1}{2} \frac{A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})} \left[\frac{A'_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})} - \frac{B''_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})}{B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})} \right].$$

Since $\beta_{i,\boldsymbol{w}} > \max_\ell(w_\ell v_\ell) > 0$, it follows from (6.28d) that $\partial \beta_{i,\boldsymbol{w}} / \partial \tilde{w}_j \neq 0$, so it follows from (6.30) that $\partial \tilde{r}_i(\boldsymbol{w}) / \partial \tilde{w}_j$ is zero exactly when

$$\lambda_i \Delta_{i,\boldsymbol{w}} = \frac{\lambda_i \beta_{i,\boldsymbol{w}} \tilde{r}_i(\boldsymbol{w}) \tilde{w}_j + v_j^2}{\beta_{i,\boldsymbol{w}} \tilde{w}_j - v_j}. \tag{6.31}$$

Rearranging (6.27) and substituting (6.31) yields

$$\begin{aligned} 0 = A_{\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) - \tilde{r}_i(\boldsymbol{w}) \beta_{i,\boldsymbol{w}} B'_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}}) &= 1 - \sum_{\ell=1}^{L} \frac{c_\ell (v_\ell^2 + \lambda_i \beta_{i,\boldsymbol{w}} \tilde{r}_i(\boldsymbol{w}) \tilde{w}_\ell)}{(\beta_{i,\boldsymbol{w}} \tilde{w}_\ell - v_\ell)^2} \\ &= 1 - \Delta_{i,\boldsymbol{w}} \lambda_i \sum_{\ell=1}^{L} \frac{c_\ell}{\beta_{i,\boldsymbol{w}} \tilde{w}_\ell - v_\ell} = 1 - \Delta_{i,\boldsymbol{w}} (1 - B_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})) = 1 - \Delta_{i,\boldsymbol{w}}, \end{aligned}$$

so $\Delta_{i,\boldsymbol{w}} = 1$. Substituting into (6.31) and solving for $\tilde{w}_j$ yields

$$w_j = \frac{1}{\tilde{w}_j} = \frac{(1 - \tilde{r}_i(\boldsymbol{w})) \beta_{i,\boldsymbol{w}}}{v_j (1 + v_j / \lambda_i)} = \frac{\gamma_{i,\boldsymbol{w}}}{v_j (1 + v_j / \lambda_i)}, \tag{6.32}$$

where the constant $\gamma_{i,\boldsymbol{w}} := (1 - \tilde{r}_i(\boldsymbol{w})) \beta_{i,\boldsymbol{w}}$ is (a) independent of j, (b) parameterizes the ray of critical points in $\mathcal{P}_\emptyset$, and (c) can be chosen freely, e.g., as unity yielding $\bar{\boldsymbol{w}}_i^\star$ from (2.5).

Solving (6.32) for $\beta_{i,\boldsymbol{w}}$, substituting into $0 = B_{i,\boldsymbol{w}}(\beta_{i,\boldsymbol{w}})$, and rearranging yields that the corresponding $\tilde{r}_i(\bar{\boldsymbol{w}}_i^\star)$ is a root of

$$R_{i,\emptyset}(x) := 1 - \sum_{\ell=1}^{L} \frac{c_\ell}{v_\ell / \lambda_i} \frac{1 - x}{v_\ell / \lambda_i + x}.$$

Since $R_{i,\emptyset}(x)$ increases from negative infinity to one as x increases from $-\min_\ell(v_\ell)/\lambda_i$ to one, it has exactly one real root in that domain. In particular, this root is the largest real root since $R_{i,\emptyset}(x) \geq 1$ for $x \geq 1$. Furthermore, $\tilde{r}_i(\bar{\boldsymbol{w}}_i^\star)$ increases continuously to one as $c := c_1 + \cdots + c_L$ increases to infinity, so $\tilde{r}_i(\bar{\boldsymbol{w}}_i^\star)$ must be the largest real root.

Likewise, the critical points of other partitions $\mathcal{P}_\mathcal{L}$ are given by setting the positive weights proportional to $\bar{\boldsymbol{w}}_i^\star$ with the corresponding $\tilde{r}_i(\bar{\boldsymbol{w}}_i^\star)$ given by the largest real root of

$$R_{i,\mathcal{L}}(x) := 1 - \sum_{\ell \notin \mathcal{L}} \frac{c_\ell}{v_\ell / \lambda_i} \frac{1 - x}{v_\ell / \lambda_i + x}.$$

For $\mathcal{L}_1 \subset \mathcal{L}_2$ a proper subset, the largest real root of $R_{i,\mathcal{L}_1}$ is greater than that of $R_{i,\mathcal{L}_2}$ since $R_{i,\mathcal{L}_1}(x) < R_{i,\mathcal{L}_2}(x)$ for any $x \in (-\min_\ell(v_\ell)/\lambda_i, 1)$. As a result, $\tilde{r}_i(\boldsymbol{w})$ is maximized in $\mathcal{P}_\emptyset$.

Finally, we check when $\tilde{r}_i(\bar{\boldsymbol{w}}_i^\star)$ is positive. Recalling that $R_{i,\emptyset}(x)$ is an increasing function on $x \in (0,1)$ and noting that $R_{i,\emptyset}(0) = 1 - \sum_{\ell=1}^L c_\ell(\lambda_i/v_\ell)^2$ yields that $\tilde{r}_i(\bar{\boldsymbol{w}}_i^\star)$ is positive if and only if $\sum_{\ell=1}^L c_\ell(\lambda_i/v_\ell)^2 > 1$. When it is positive, the maximizers are given by the critical points above; otherwise, $\bar{r}_i(\boldsymbol{w}) = 0$ for all $\boldsymbol{w}$. This concludes the proof of Lemma 6.3.

7. When signal and noise variances are unknown. The asymptotic optimal weights from the main result (Theorem 2.1) depend on both the signal component variance λ_i and the noise variances $\boldsymbol{v}$, but one or both of these variances may be unknown in some settings. Of course, one could estimate these variances using existing ideas and approaches, then plug them into (2.5) in Theorem 2.1. The question then is, are these resulting estimated weights also asymptotically optimal? Fortunately, it is straightforward to see that the answer is yes under natural conditions on the variance estimates. The following proposition makes this statement precise; it follows straightforwardly from Lemmas 6.2 and 6.3.

Proposition 7.1 (asymptotic optimality with estimated variances). *Suppose $\hat{\lambda}_i(\boldsymbol{Y})$ and $\hat{\boldsymbol{v}}(\boldsymbol{Y})$ are consistent estimates of λ_i and $\boldsymbol{v}$, i.e.,*

$$\hat{\lambda}_i(\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \lambda_i, \qquad \hat{\boldsymbol{v}}(\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \boldsymbol{v}. \tag{7.1}$$

Let $\hat{\boldsymbol{w}}_i^\star(\boldsymbol{Y})$ be the estimated weights obtained by plugging $\hat{\boldsymbol{v}}(\boldsymbol{Y})$ and $\hat{\lambda}_i(\boldsymbol{Y})$ into (2.5) in Theorem 2.1, i.e.,

$$\hat{\boldsymbol{w}}_i^\star(\boldsymbol{Y}) := \left(\frac{1}{\hat{v}_1(\boldsymbol{Y})} \frac{1}{1 + \hat{v}_1(\boldsymbol{Y})/\hat{\lambda}_i(\boldsymbol{Y})}, \dots, \frac{1}{\hat{v}_L(\boldsymbol{Y})} \frac{1}{1 + \hat{v}_L(\boldsymbol{Y})/\hat{\lambda}_i(\boldsymbol{Y})} \right). \tag{7.2}$$

Then the estimated weights and their corresponding performance converge to the asymptotic optimal weights and performance, i.e.,

$$\hat{\boldsymbol{w}}_i^\star(\boldsymbol{Y}) \xrightarrow{\text{a.s.}} \bar{\boldsymbol{w}}_i^\star, \qquad r_i(\hat{\boldsymbol{w}}_i^\star(\boldsymbol{Y}), \boldsymbol{Y}) \xrightarrow{\text{a.s.}} \bar{r}_i^\star. \tag{7.3}$$

Namely, the estimated weights are asymptotically optimal.

Thus, optimal weighting only needs consistent estimates of λ_i and $\boldsymbol{v}$ that may be obtained using any one of various existing approaches; which one is most appropriate will depend on the specific application. Here we consider a simple pair of estimators as an illustrative example.

Example 7.2 (variance estimators). As an illustrative example, consider the following simple estimators for the signal and noise variances:

$$\hat{\lambda}_i(\boldsymbol{Y}) := \Xi\left(\hat{\lambda}_i^{(\text{inv})}(\boldsymbol{Y}; \hat{\boldsymbol{v}}(\boldsymbol{Y})); \left[\sum_{\ell=1}^L \frac{p_\ell}{\hat{v}_\ell(\boldsymbol{Y})} \right]^{-1} \right), \quad \hat{\boldsymbol{v}}(\boldsymbol{Y}) := \left(\frac{\|\boldsymbol{Y}_1\|_{\text{F}}^2}{d\,n_1}, \dots, \frac{\|\boldsymbol{Y}_L\|_{\text{F}}^2}{d\,n_L} \right), \tag{7.4}$$

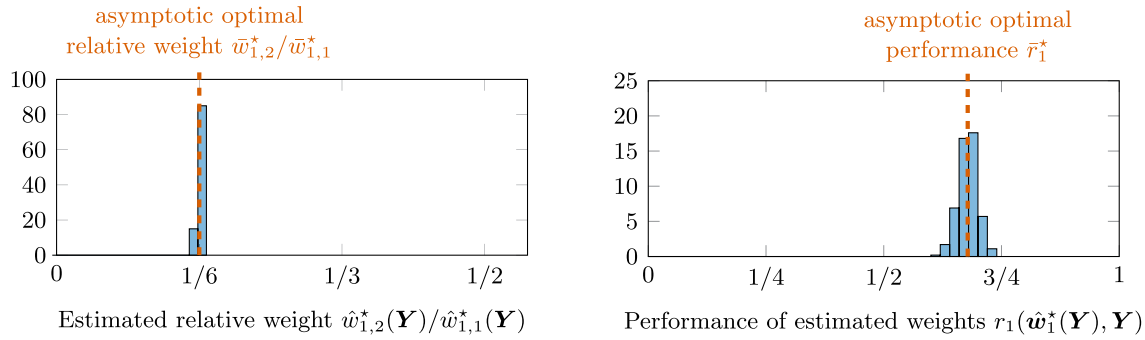


Figure 7.1. *Nonasymptotic empirical distributions of estimated weights* $\hat{\boldsymbol{w}}^\star_1(\boldsymbol{Y})$ *from* (7.2) *and corresponding performance* $r_1(\hat{\boldsymbol{w}}^\star_1(\boldsymbol{Y}), \boldsymbol{Y})$ *for an illustrative example with two blocks of data* $\boldsymbol{Y}_1 \in \mathbb{R}^{d\times n_1}$ *and* $\boldsymbol{Y}_2 \in \mathbb{R}^{d\times n_2}$*, where the estimated weights are computed using the signal and noise variance estimators* (7.4) *from Example* 7.2*. The data blocks are generated with noise variances* $v_1 = 1$ *and* $v_2 = 3$*, one underlying component having variance* $\lambda_1 = 1$*, and dimensions* $d = 1000$*,* $n_1 = 4000$*, and* $n_2 = 8000$*.*

where $p_\ell := c_\ell/c$, $c := c_1 + \cdots + c_L$, and

$$\hat{\lambda}_i^{(\mathrm{inv})}(\boldsymbol{Y}; \boldsymbol{v}) := i\text{th leading eigenvalue of } \sum_{\ell=1}^{L} \left(\frac{1/v_\ell}{n_1/v_1 + \cdots + n_L/v_L} \right) \boldsymbol{Y}_\ell \boldsymbol{Y}_\ell^{\mathsf{H}}, \tag{7.5}$$

$$\Xi(\lambda; v) := \text{the larger root of the quadratic polynomial } (x + v/c)(x + v) - \lambda x. \tag{7.6}$$

It is straightforward to verify with standard techniques (see supplementary material SM4 for details) that these estimators are consistent as long as $c > (\bar{v}/\lambda_i)^2$, i.e., when inverse noise variance weighting is above the phase transition. Thus, by Proposition 7.1, the resulting estimated weights and their corresponding performance converge to the asymptotic optimal weights and performance.

Figure 7.1 illustrates the nonasymptotic behavior of these estimated weights in numerical simulations. The data is generated as in section 3, with dimensionality $d = 1000$ and block sizes $\boldsymbol{n} = (4000, 8000)$. The estimated weights and their performance generally concentrate around the asymptotic optimal weights and performance. Moreover, the estimated weights achieve performance closely matching the nonasymptotic empirically optimized weights in Figure 3.1b.

8. Illustration on real data from astronomy. This section illustrates optimally weighted PCA on real data from astronomy. In particular, we consider quasar spectra from the Sloan Digital Sky Survey (SDSS) Data Release 16 [2] using the associated DR16Q quasar catalog [32]. Each spectrum is a vector of flux measurements across wavelengths for a particular quasar, and comes with associated noise variance estimates across wavelengths. The noise is heteroscedastic across both quasars and wavelengths, but here we focus on a subset that is somewhat homoscedastic across wavelengths. Supplementary material SM5 describes the details of the subset selected and the preprocessing performed (filtering, interpolation, centering, normalization).

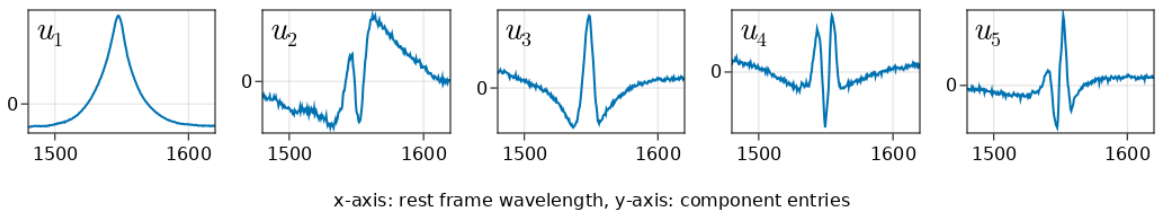
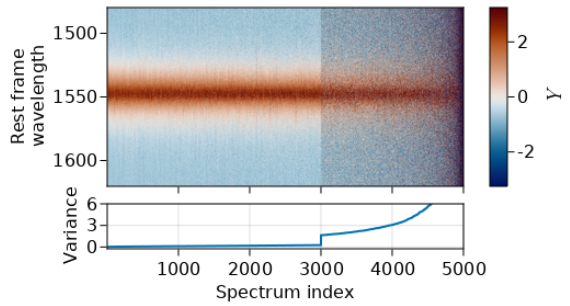
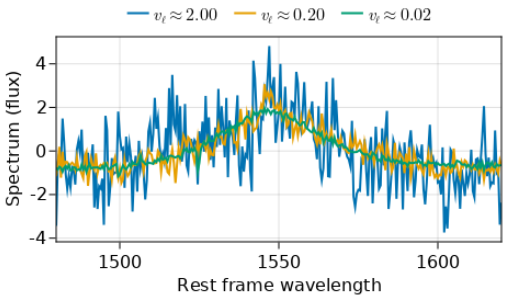


Figure 8.1. *Ground truth components computed from* 5000 *samples with smallest variances.*



(a) Data matrix $\boldsymbol{Y} \in \mathbb{R}^{d \times n}$ (top) and associated noise variances $\boldsymbol{v} \in \mathbb{R}^{n}_{\geq 0}$ (bottom).



(b) Example spectra with associated high, medium, and low noise variances.

Figure 8.2. *Quasar spectra dataset and example spectra.*

The resulting data has $d = 281$ wavelengths measured for $n^{(\text{full})} = 10459$ spectra, yielding a data matrix $\boldsymbol{Y}^{(\text{full})} \in \mathbb{R}^{d \times n^{(\text{full})}}$ with a vector of associated variances $\boldsymbol{v}^{(\text{full})} \in \mathbb{R}^{n^{(\text{full})}}_{\geq 0}$. Figure 8.1 shows plots of components $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_5 \in \mathbb{R}^d$ computed via an unweighted PCA on the 5000 samples from $\boldsymbol{Y}^{(\text{full})}$ with smallest variances. We regard these components as "ground truth."

To evaluate the various weighting schemes, we formed a test dataset $\boldsymbol{Y} \in \mathbb{R}^{d \times n}$ containing $n = 5000$ samples by combining the 3000 samples with the smallest variances (a subset of the 5000 samples used to produce ground truth) and the 2000 samples with the largest variances. This provides a dataset with noise heteroscedasticity across samples, shown as a heatmap in Figure 8.2a. Figure 8.2b shows a few sample spectra from the dataset; note that they have a common shape but have varying levels of noise.

We computed the leading singular vectors $\hat{\boldsymbol{u}}_1(\boldsymbol{w}_1, \boldsymbol{Y}), \ldots, \hat{\boldsymbol{u}}_5(\boldsymbol{w}_5, \boldsymbol{Y})$ via unweighted PCA, inverse variance weighted PCA, and optimally weighted PCA; the optimally weighted PCA has component-specific weights, so its weights are not the same across the components.[2] We used the provided noise variances, and estimated the signal variances using the estimator from Example 7.2 (with the provided noise variances substituted). Table 8.1 shows the recovery of the ground truth singular vectors calculated from the "clean" samples above.

[2]While the components $\hat{\boldsymbol{u}}_1(\hat{\boldsymbol{w}}_1^\star(\boldsymbol{Y}), \boldsymbol{Y}), \ldots, \hat{\boldsymbol{u}}_5(\hat{\boldsymbol{w}}_5^\star(\boldsymbol{Y}), \boldsymbol{Y})$ from optimally weighted PCA are not guaranteed to be orthogonal in general (due to the component-specific weighting), they were approximately orthogonal here. In particular, $\max_{i \neq j} |\hat{\boldsymbol{u}}_i(\hat{\boldsymbol{w}}_i^\star(\boldsymbol{Y}), \boldsymbol{Y})^{\mathsf{H}} \hat{\boldsymbol{u}}_j(\hat{\boldsymbol{w}}_j^\star(\boldsymbol{Y}), \boldsymbol{Y})|^2 \approx 0.00013$.

Table 8.1

Recoveries $r_i(\boldsymbol{w}, \boldsymbol{Y})$ for unweighted PCA ($w_\ell = 1$), inverse variance weighted PCA ($w_\ell = 1/v_\ell$), and optimally weighted PCA. Higher is better, best value (up to rounding) is shown in bold.

Component	1	2	3	4	5
Unweighted PCA	0.003	0.307	0.009	0.004	0.018
Inverse variance weighted PCA	**1.000**	0.903	0.915	0.817	0.811
Optimally weighted PCA	**1.000**	**0.920**	**0.934**	**0.884**	**0.880**

The following observations apply to this data and also summarize some of the main themes of this paper:

- unweighted PCA performs poorly for heteroscedastic data,
- inverse variance weighted PCA performs much better than unweighted PCA, and
- optimally weighted PCA performs even better than inverse variance weighted PCA.

Similar comparisons for these leading components occurred for many of the other test datasets we tried. The comparisons were less consistent for components 6 and on; optimal weights were sometimes better and inverse noise variance weights were sometimes better. This inconsistent behavior is potentially due to the data not matching the model closely enough or perhaps the ground truth needing to be chosen differently. A detailed investigation of this phenomenon is beyond our present scope.

Overall, this example with quasar spectra coming from astronomy illustrates the potential for optimally weighted PCA to improve the recovery of underlying principal components from real data that has heteroscedastic noise.

9. Conclusion.

This paper derived asymptotic optimal weights for weighted PCA when the data is high-dimensional with noise that is heteroscedastic across samples. The optimal weights are a simple function of the signal and noise variances, and are not the inverse noise variance weights commonly used in practice. Numerical simulations illustrated that the asymptotic optimal weights are often close to the optimal weights in finite dimensions when the dimensions are large enough. Comparisons of the asymptotic optimal weights with existing weighting schemes illustrated that the asymptotic optimal weights (a) are generally closer to the optimal weights in finite dimensions, (b) appropriately combine all the data to achieve the best performance, and (c) achieve nonzero asymptotic performance in the widest range of settings. Additional simulations illustrated that optimally weighted PCA compares favorably with other PCA methods designed for some form of heteroscedastic noise. Finally, we briefly discussed how one can use estimated signal and noise variances when the true variances are unknown, and illustrated the optimal weights on real data from astronomy.

Overall, optimally weighted PCA is a simple, principled, and promising method for estimating underlying principal components from high-dimensional data with noise that is heteroscedastic across samples. However, many open questions remain. While the asymptotic optimal performance (2.6) is often close to the performance of optimally weighted PCA in finite dimensions, it would be useful to also derive higher-order asymptotics for the performance as well as nonasymptotic characterizations. These results would provide refined estimates of the performance. Another interesting variation of the asymptotic regime is to allow the number of blocks L to grow with d, potentially with $n_1, \ldots, n_L = \mathrm{O}(1)$. This regime may better capture

datasets where each sample has its own associated noise variance (e.g., like the astronomy data in section 8) or where the block structure is unknown. While the proof of Theorem 2.1 does not seem to readily generalize to such cases, we conjecture that the optimal weights will still have the same form and that estimates of the signal and noise variances can still be used when the true variances are unknown. Another interesting direction is to study whether optimally weighted PCA is optimal across not just weighted PCA but also across more general classes of methods, e.g., by deriving fundamental bounds on the achievable performance. Finally, it would be interesting to study how optimally weighted PCA might be combined with other techniques to handle settings where the noise is not just heteroscedastic across samples but also across features.

Acknowledgments. The authors thank Raj Rao Nadakuditi, Romain Couillet, and Edgar Dobriban for helpful discussions regarding the singular values and vectors of low rank perturbations of large random matrices. The authors also thank Dejiao Zhang for helpful comments about the form of the optimal weights. The authors also thank the editor and the anonymous referees for their helpful suggestions that led to various improvements throughout this paper.

The example quasar spectra were provided by the Sloan Digital Sky Survey [2, 32]. Funding for the Sloan Digital Sky Survey IV was been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions.

SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics — Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

REFERENCES

[1] R. B. Abdallah, A. Breloy, M. N. E. Korso, and D. Lautru, *Bayesian signal subspace estimation with compound Gaussian sources*, Signal Process., 167 (2020), 107310, https://doi.org/10.1016/j.sigpro.2019.107310.

[2] R. Ahumada, C. A. Prieto, A. Almeida, F. Anders, et al., *The 16th data release of the Sloan Digital Sky Surveys: First release from the APOGEE-2 southern survey and full release of eBOSS spectra*, Astrophys. J. Suppl. Ser., 249 (2020), 3, https://doi.org/10.3847/1538-4365/ab929e.

[3] B. A. Ardekani, J. Kershaw, K. Kashikura, and I. Kanno, *Activation detection in functional MRI using subspace modeling and maximum likelihood estimation*, IEEE Trans. Med. Imaging, 18 (1999), pp. 101–114, https://doi.org/10.1109/42.759109.

[4] S. Bailey, *Principal component analysis with noisy and/or missing data*, Publ. Astron. Soc. Pac., 124 (2012), pp. 1015–1023, https://doi.org/10.1086/668105.

[5] F. Benaych-Georges and R. R. Nadakuditi, *The singular values and vectors of low rank perturbations of large rectangular random matrices*, J. Multivariate Anal., 111 (2012), pp. 120–135, https://doi.org/10.1016/j.jmva.2012.04.019.

[6] O. Besson, *Bounds for a mixture of low-rank compound-Gaussian and white Gaussian noises*, IEEE Trans. Signal Process., 64 (2016), pp. 5723–5732, https://doi.org/10.1109/tsp.2016.2603965.

[7] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Isotropic local laws for sample covariance and generalized Wigner matrices*, Electron. J. Probab., 19 (2014), pp. 1–53, https://doi.org/10.1214/ejp.v19-3054.

[8] A. Breloy, G. Ginolhac, F. Pascal, and P. Forster, *Clutter subspace estimation in low rank heterogeneous noise context*, IEEE Trans. Signal Process., 63 (2015), pp. 2173–2182, https://doi.org/10.1109/tsp.2015.2403284.

[9] A. Breloy, G. Ginolhac, F. Pascal, and P. Forster, *Robust covariance matrix estimation in heterogeneous low rank context*, IEEE Trans. Signal Process., 64 (2016), pp. 5794–5806, https://doi.org/10.1109/tsp.2016.2599494.

[10] R. N. Cochran and F. H. Horne, *Statistically weighted principal component analysis of rapid scanning wavelength kinetics experiments*, Anal. Chem., 49 (1977), pp. 846–853, https://doi.org/10.1021/ac50014a045.

[11] A. Collas, F. Bouchard, A. Breloy, G. Ginolhac, C. Ren, and J.-P. Ovarlez, *Probabilistic PCA from heteroscedastic signals: Geometric framework and application to clustering*, IEEE Trans. Signal Process., 69 (2021), pp. 6546–6560, https://doi.org/10.1109/tsp.2021.3130997.

[12] J.-C. Deville and E. Malinvaud, *Data analysis in official socio-economic statistics*, J. R. Stat. Soc. Ser. A, 146 (1983), pp. 335–361, https://doi.org/10.2307/2981452.

[13] X. Ding and F. Yang, *A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices*, Ann. Appl. Probab., 28 (2018), pp. 1679–1738, https://doi.org/10.1214/17-aap1341.

[14] E. Dobriban, W. Leeb, and A. Singer, *Optimal prediction in the linearly transformed spiked model*, Ann. Statist., 48 (2020), pp. 491–513, https://doi.org/10.1214/19-aos1819.

[15] D. Donoho, M. Gavish, and I. Johnstone, *Optimal shrinkage of eigenvalues in the spiked covariance model*, Ann. Statist., 46 (2018), pp. 1742–1778, https://doi.org/10.1214/17-aos1601.

[16] D. Hong, L. Balzano, and J. A. Fessler, *Towards a theoretical analysis of PCA for heteroscedastic data*, in Proceedings of the 54th Allerton Conference on Communication, Control, and Computing, IEEE, 2016, pp. 496–503, https://doi.org/10.1109/allerton.2016.7852272.

[17] D. Hong, L. Balzano, and J. A. Fessler, *Asymptotic performance of PCA for high-dimensional heteroscedastic data*, J. Multivariate Anal., 167 (2018), pp. 435–452, https://doi.org/10.1016/j.jmva.2018.06.002.

[18] D. Hong, L. Balzano, and J. A. Fessler, *Probabilistic PCA for heteroscedastic data*, in Proceedings of the 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, IEEE, 2019, pp. 26–30, https://doi.org/10.1109/camsap45676.2019.9022436.

[19] D. Hong, K. Gilman, L. Balzano, and J. A. Fessler, *HePPCAT: Probabilistic PCA for data with heteroscedastic noise*, IEEE Trans. Signal Process., 69 (2021), pp. 4819–4834, https://doi.org/10.1109/tsp.2021.3104979.

[20] D. Hong, Y. Sheng, and E. Dobriban, *Selecting the Number of Components in PCA via Random Signflips*, preprint, http://arxiv.org/abs/2012.02985v1, 2020.

[21] J. J. Jansen, H. C. J. Hoefsloot, H. F. M. Boelens, J. van der Greef, and A. K. Smilde, *Analysis of longitudinal metabolomics data*, Bioinformatics, 20 (2004), pp. 2438–2446, https://doi.org/10.1093/bioinformatics/bth268.

[22] I. M. Johnstone and A. Y. Lu, *On consistency and sparsity for principal components analysis in high dimensions*, J. Amer. Statist. Assoc., 104 (2009), pp. 682–693, https://doi.org/10.1198/jasa.2009.0121.

[23] I. M. Johnstone and D. Paul, *PCA in high dimensions: An orientation*, Proc. IEEE, 106 (2018), pp. 1277–1292, https://doi.org/10.1109/jproc.2018.2846730.

[24] I. M. Johnstone and D. M. Titterington, *Statistical challenges of high-dimensional data*, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., 367 (2009), pp. 4237–4253, https://doi.org/10.1098/rsta.2009.0159.
[25] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer-Verlag, New York, 2002, https://doi.org/10.1007/b98835.
[26] Z. T. Ke, Y. Ma, and X. Lin, *Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis*, J. Amer. Statist. Assoc., (2021), pp. 1–19, https://doi.org/10.1080/01621459.2021.1933497.
[27] A. Knowles and J. Yin, *Anisotropic local laws for random matrices*, Probab. Theory Related Fields, 169 (2016), pp. 257–352, https://doi.org/10.1007/s00440-016-0730-4.
[28] B. Landa, T. T. C. K. Zhang, and Y. Kluger, *Biwhitening Reveals the Rank of a Count Matrix*, preprint, http://arxiv.org/abs/2103.13840v2, 2021.
[29] W. Leeb and E. Romanov, *Optimal spectral shrinkage and PCA with heteroscedastic noise*, IEEE Trans. Inform Theory, 67 (2021), pp. 3009–3037, https://doi.org/10.1109/tit.2021.3055075.
[30] W. E. Leeb, *Matrix denoising for weighted loss functions and heterogeneous signals*, SIAM J. Math. Data Sci., 3 (2021), pp. 987–1012, https://doi.org/10.1137/20m1319577.
[31] J. T. Leek, *Asymptotic conditional singular value decomposition for high-dimensional genomic data*, Biometrics, 67 (2010), pp. 344–352, https://doi.org/10.1111/j.1541-0420.2010.01455.x.
[32] B. W. Lyke, A. N. Higley, J. N. McLane, D. P. Schurhammer, et al., *The Sloan Digital Sky Survey quasar catalog: Sixteenth data release*, Astrophys. J. Suppl. Ser., 250 (2020), 8, https://doi.org/10.3847/1538-4365/aba623.
[33] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Math USSR-Sb., 1 (1967), pp. 457–483, https://doi.org/10.1070/sm1967v001n04abeh001994.
[34] R. R. Nadakuditi, *OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage*, IEEE Trans. Inform Theory, 60 (2014), pp. 3002–3018, https://doi.org/10.1109/tit.2014.2311661.
[35] B. Nadler, *Finite sample approximation results for principal component analysis: A matrix perturbation approach*, Ann. Statist., 36 (2008), pp. 2791–2817, https://doi.org/10.1214/08-aos618.
[36] S. Papadimitriou, J. Sun, and C. Faloutsos, *Streaming pattern discovery in multiple time-series*, in Proceedings of the 31st International Conference on Very Large Data Bases, ACM, 2005, pp. 697–708, http://www.vldb.org/archives/website/2005/program/paper/thu/p697-papadimitriou.pdf.
[37] D. Paul, *Asymptotics of sample eigenstructure for a large dimensional spiked covariance model*, Statist. Sinica, 17 (2007), pp. 1617–1642, http://www3.stat.sinica.edu.tw/statistica/J17N4/J17N418/J17N418.html.
[38] H. Pedersen, S. Kozerke, S. Ringgaard, K. Nehrke, and W. Y. Kim, *k-t PCA: Temporally constrained k-t BLAST reconstruction using principal component analysis*, Magn. Reson. Med., 62 (2009), pp. 706–716, https://doi.org/10.1002/mrm.22052.
[39] PurpleAir, *Real Time Air Quality Monitoring*, https://www2.purpleair.com.
[40] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer, Berlin Heidelberg, 1998, https://doi.org/10.1007/978-3-642-02431-3.
[41] N. Sharma and K. Saroha, *A novel dimensionality reduction method for cancer dataset using PCA and feature ranking*, in Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2015, pp. 2261–2264, https://doi.org/10.1109/icacci.2015.7275954.
[42] Y. Sun, A. Breloy, P. Babu, D. P. Palomar, F. Pascal, and G. Ginolhac, *Low-complexity algorithms for low rank clutter parameters estimation in radar systems*, IEEE Trans. Signal Process., 64 (2016), pp. 1986–1998, https://doi.org/10.1109/tsp.2015.2512535.
[43] O. Tamuz, T. Mazeh, and S. Zucker, *Correcting systematic effects in a large set of photometric light curves*, Mon. Not. R. Astron. Soc., 356 (2005), pp. 1466–1470, https://doi.org/10.1111/j.1365-2966.2004.08585.x.
[44] P. Tsalmantza and D. W. Hogg, *A data-driven model for spectra: Finding double redshifts in the Sloan Digital Sky Survey*, Astrophys. J., 753 (2012), 122, https://doi.org/10.1088/0004-637x/753/2/122.
[45] US Environmental Protection Agency, *Air Quality System Data Mart*, https://www.epa.gov/airdata.
[46] R. Vershynin, *High-Dimensional Probability*, Cambridge University Press, Cambridge, 2018, https://doi.org/10.1017/9781108231596.

[47] G. S. Wagner and T. J. Owens, *Signal detection using multi-channel seismic data*, Bull. Seismol. Soc. Amer., 86 (1996), pp. 221–231, https://doi.org/10.1785/bssa08601a0221.
[48] H. Xi, F. Yang, and J. Yin, *Convergence of eigenvector empirical spectral distribution of sample covariance matrices*, Ann. Statist., 48 (2020), pp. 953–982, https://doi.org/10.1214/19-aos1832.
[49] G. Young, *Maximum likelihood estimation and factor analysis*, Psychometrika, 6 (1941), pp. 49–53, https://doi.org/10.1007/bf02288574.
[50] A. R. Zhang, T. T. Cai, and Y. Wu, *Heteroskedastic PCA: Algorithm, optimality, and applications*, Ann. Statist., 50 (2022), pp. 53–80, https://doi.org/10.1214/21-aos2074.