# Supplementary material for “Asymptotic performance of PCA for high-dimensional heteroscedastic Data”

David Hong[1,*], Laura Balzano[2], Jeffrey A. Fessler[3]

*Department of Electrical Engineering and Computer Science,*
*University of Michigan, Ann Arbor, MI 48105, USA*

## Abstract

This supplement fleshes out several details in [5]. Section S1 relates the model (1) in the paper to spiked covariance models [1, 6]. Section S2 discusses interesting properties of the simplified expressions. Section S3 shows the impact of the parameters on the asymptotic PCA amplitudes and coefficient recovery. Section S4 contains numerical simulation results for PCA amplitudes and coefficient recovery, and Section S5 simulates complex-valued and Gaussian mixture data.

## S1. Relationship to spiked covariance models

The model (1) considered in the paper is similar in spirit to the generalized spiked covariance model of [1]. To discuss the relationship more easily, we will refer to the paper model (1) as the “inter-sample heteroscedastic model”. Both this and the generalized spiked covariance model generalize the Johnstone spiked covariance model proposed in [6]. In the Johnstone spiked covariance model [1], sample vectors $y_1, \dots, y_n \in \mathbb{C}^d$ are generated as

$$y_i = \operatorname{diag}(\alpha_1^2, \dots, \alpha_k^2, \underbrace{1, \dots, 1}_{d-k \text{ copies}})^{1/2} x_i, \tag{S1}$$

where $x_i \in \mathbb{C}^d$ are independent identically distributed (iid) vectors with iid entries that have mean $\mathrm{E}(x_{ij}) = 0$ and variance $\mathrm{E}|x_{ij}|^2 = 1$.

For normally distributed subspace coefficients and noise vectors, the inter-sample heteroscedastic model (1) is equivalent (up to rotation) to generating sample vectors $y_1, \dots, y_n \in \mathbb{C}^d$ as

$$y_i = \operatorname{diag}(\theta_1^2 + \eta_i^2, \dots, \theta_k^2 + \eta_i^2, \underbrace{\eta_i^2, \dots, \eta_i^2}_{d-k \text{ copies}})^{1/2} x_i, \tag{S2}$$

where $x_i \in \mathbb{C}^d$ are iid with iid normally distributed entries. (S2) generalizes the Johnstone spiked covariance model because the covariance matrix can vary across samples. Heterogeneity here is *across* samples; all entries $(y_i)_1, \dots, (y_i)_d$ within each sample $y_i$ have equal noise variance $\eta_i^2$.

The generalized spiked covariance model generalizes the Johnstone spiked covariance model differently. In the generalized spiked covariance model [1], sample vectors $y_1, \dots, y_n \in \mathbb{C}^d$ are generated as

$$y_i = \begin{bmatrix} \mathbf{\Lambda} & \\ & \mathbf{V}_{d-k} \end{bmatrix}^{1/2} x_i, \tag{S3}$$

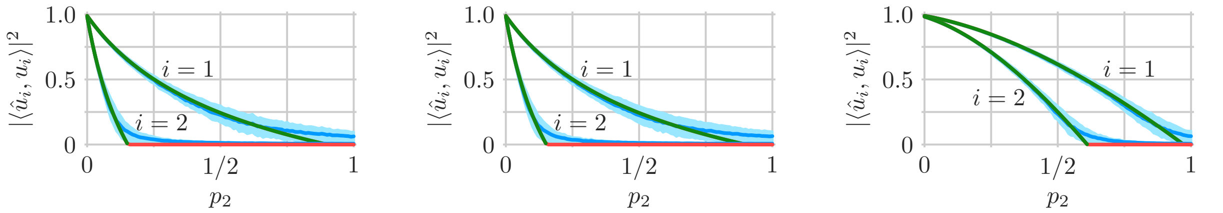
[*] Corresponding author

*Email addresses:* `dahong@umich.edu` (David Hong), `girasole@umich.edu` (Laura Balzano), `fessler@umich.edu` (Jeffrey A. Fessler)

[1] This work was supported by the National Science Foundation Graduate Research Fellowship [DGE #1256260].

[2] This work was supported by the ARO [W911NF-14-1-0634]; and DARPA [DARPA-16-43-D3M-FP-037].

[3] This work was supported by the UM-SJTU data science seed fund; and the NIH [U01 EB 018753].



**(a)** Deterministic inter-sample heteroscedastic model.

**(b)** Random inter-sample heteroscedastic model.

**(c)** Johnstone spiked covariance model with covariance (S6).

**Figure S1:** Simulated subspace recovery as a function of the contamination fraction $p_2$, the proportion of samples with noise variance $\sigma_2^2 = 3.25$, where the other noise variance $\sigma_1^2 = 0.1$ occurs in proportion $p_1 = 1 - p_2$. Subspace amplitudes are $\theta_1 = 1$ and $\theta_2 = 0.8$, and there are $10^4$ samples in $10^3$ dimensions. Simulation mean (blue curve) and interquartile interval (light blue ribbon) are shown with the asymptotic recovery (4) of Theorem 1 (green curve). The region where $A(\beta_i) \leq 0$ is the red horizontal segment with value zero (the prediction of Conjecture 1). Deterministic noise variances $\eta_1^2, \ldots, \eta_n^2$ are used for simulations in (a), random ones are used for those in (b), and (c) has data generated according to the Johnstone spiked covariance model with covariance matrix set as (S6).

where $x_i \in \mathbb{C}^d$ are iid with iid entries as in (S1), $\mathbf{\Lambda} \in \mathbb{C}^{k \times k}$ is a deterministic Hermitian matrix with eigenvalues $\alpha_1^2, \ldots, \alpha_k^2$, $\mathbf{V}_{d-k} \in \mathbb{R}^{(d-k)\times(d-k)}$ has limiting eigenvalue distribution $\nu$, and these all satisfy a few technical conditions [1]. All samples share a common covariance matrix, but the model allows, among other things, for heterogenous variance within the samples. To illustrate this flexibility, note that we could set

$$\mathbf{\Lambda} = \operatorname{diag}(\theta_1^2 + \eta_1^2, \ldots, \theta_k^2 + \eta_k^2), \qquad \mathbf{V}_{d-k} = \operatorname{diag}(\eta_{k+1}^2, \ldots, \eta_d^2). \tag{S4}$$

In this case, there is heteroscedasticity among the entries of each sample vector. Heterogeneity here is *within* each sample, not across them; recall that all samples have the same covariance matrix.

Therefore, for data with *intra*-sample heteroscedasticity, one should use the results of [1] and [10] for the generalized spiked covariance model. For data with *inter*-sample heteroscedasticity, one should use the new results presented in Theorem 1 of the paper [5] for the inter-sample heteroscedastic model. A couple variants of the inter-sample heteroscedastic model are also natural to consider in the context of spiked covariance models; the next two subsections discuss these.

### *S1.1. Random noise variances*

The noise variances $\eta_1^2, \ldots, \eta_n^2$ in the inter-sample heteroscedastic model (1) are deterministic. A natural variation could be to instead make them iid random variables defined as

$$\eta_i^2 = \begin{cases} \sigma_1^2 & \text{with probability } p_1, \\ \vdots & \\ \sigma_L^2 & \text{with probability } p_L, \end{cases} \tag{S5}$$

where $p_1 + \cdots + p_L = 1$. To ease discussion, this section will use the words "deterministic" and "random" before "inter-sample heteroscedastic model" to differentiate between the paper model (1) that has deterministic noise variances and its variant that instead has iid random noise variances (S5). In the random inter-sample heteroscedastic model, scaled noise vectors $\eta_1\varepsilon_1, \ldots, \eta_n\varepsilon_n$ are iid vectors drawn from a mixture. As a result, sample vectors $y_1, \ldots, y_n$ are also iid vectors with covariance matrix (up to rotation)

$$\mathrm{E}(y_i y_i^{\mathsf{H}}) = \operatorname{diag}(\theta_1^2 + \bar{\sigma}^2, \ldots, \theta_k^2 + \bar{\sigma}^2, \underbrace{\bar{\sigma}^2, \ldots, \bar{\sigma}^2}_{d-k \text{ copies}}), \tag{S6}$$

where $\bar{\sigma}^2 = p_1\sigma_1^2 + \cdots + p_L\sigma_L^2$ is the average variance.

(S6) is a spiked covariance matrix and the samples $y_1, \dots, y_n$ are iid vectors, and so it could be tempting to think that the data can be equivalently generated from the Johnstone spiked covariance model with covariance matrix (S6). However this is not true. The PCA performance of the random inter-sample heteroscedastic model is similar to that of the deterministic version and is different from that of the Johnstone spiked covariance model with covariance matrix (S6). Figure S1 illustrates the distinction in numerical simulations. In all simulations, we drew $10^4$ samples from a $10^3$ dimensional ambient space, where the subspace amplitudes were $\theta_1 = 1$ and $\theta_2 = 0.8$. Two noise variances $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 3.25$ have proportions $p_1 = 1 - p_2$ and $p_2$. In Figure S1a, data are generated according to the deterministic inter-sample heteroscedastic model. In Figure S1b, data are generated according to the random inter-sample heteroscedastic model. In Figure S1c, data are generated according to the Johnstone spiked covariance model with covariance matrix (S6).

Figures S1a and S1b demonstrate that data generated according to the inter-sample heteroscedastic model have similar behavior whether the noise variances $\eta_1^2, \dots, \eta_n^2$ are set deterministically or randomly as (S5). The similarity is expected because the random noise variances in the limit will equal $\sigma_1^2, \dots, \sigma_L^2$ in proportions approaching $p_1, \dots, p_L$ by the law of large numbers. Thus data generated with random noise variances should have similar asymptotic PCA performance as data generated with deterministic noise variances.

Figures S1b and S1c demonstrate that data generated according to the random inter-sample heteroscedastic model behave quite differently from data generated according to the Johnstone spiked covariance model, even though both have iid sample vectors with covariance matrix (S6). To understand why, recall that in the random inter-sample heteroscedastic model, the noise standard deviation $\eta_i$ is shared among the entries of the scaled noise vector $\eta_i \varepsilon_i$. This induces statistical dependence among the entries of the sample vector $y_i$ that is not eliminated by whitening with $\mathrm{E}(y_i y_i^{\mathsf{H}})^{-1/2}$. Whitening a sample vector $y_i$ generated according to the Johnstone spiked covariance model, on the other hand, produces the vector $x_i$ that has iid entries by definition. Thus, the random inter-sample heteroscedastic model is not equivalent to the Johnstone spiked covariance model. One should use Theorem 1 in the paper [5] to analyze asymptotic PCA performance in this setting rather than existing results for the Johnstone spiked covariance model [2, 3, 7–9].

### *S1.2. Row samples*

In matrix form, the inter-sample heteroscedastic model can be written as

$$\mathbf{Y} = (y_1, \dots, y_n) = \mathbf{U}\boldsymbol{\Theta}\mathbf{Z}^{\mathsf{H}} + \mathbf{E}\mathbf{H} \in \mathbb{C}^{d \times n},$$

where

$\mathbf{Z} = (z^{(1)}, \dots, z^{(k)}) \in \mathbb{C}^{n \times k}$ is the coefficient matrix,

$\mathbf{E} = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{C}^{d \times n}$ is the (unscaled) noise matrix,

$\mathbf{H} = \operatorname{diag}(\eta_1, \dots, \eta_n) \in \mathbb{R}_+^{n \times n}$ is a diagonal matrix of noise standard deviations.

Samples in the paper [5] are the columns $y_1, \dots, y_n$ of the data matrix $\mathbf{Y}$, but one could alternatively form samples from the rows

$$y^{(i)} = \begin{bmatrix} (y_1)_i \\ \vdots \\ (y_n)_i \end{bmatrix} = \mathbf{Z}^* \boldsymbol{\Theta} u^{(i)} + \mathbf{H}\varepsilon^{(i)}, \tag{S7}$$

where $u^{(i)} = ((u_1)_i, \dots, (u_n)_i)$ and $\varepsilon^{(i)} = ((\varepsilon_1)_i, \dots, (\varepsilon_n)_i)$ are the $i$th rows of $\mathbf{U}$ and $\mathbf{E}$, respectively. Row samples (S7) are exactly the columns of the transposed data matrix $\mathbf{Y}^\top$ and so row samples have the same PCA amplitudes as column samples; principal components and score vectors swap.

In (S7), noise heteroscedasticity is within each row sample $y^{(i)}$ rather than across row samples $y^{(1)}, \dots, y^{(d)}$, and so one might think that the row samples could be equivalently generated from the generalized spiked covariance model (S3) with a covariance similar to (S4). However, the row samples are neither independent nor identically distributed; $\mathbf{U}$ induces dependence across rows as well as variety in their distributions. As a result, the row samples do not match the generalized spiked covariance model.

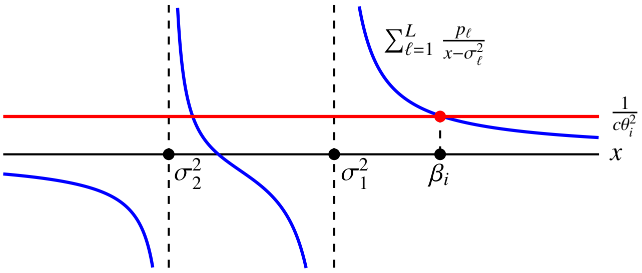


**Figure S2:** Location of the largest real root $\beta_i$ of $B_i(x)$ for two noise variances $\sigma_1^2 = 2$ and $\sigma_2^2 = 0.75$, occurring in proportions $p_1 = 70\%$ and $p_2 = 30\%$, where the sample-to-dimension ratio is $c = 1$ and the subspace amplitude is $\theta_i = 1$.

One could make $\mathbf{U}$ random according to the "i.i.d. model" of [2]. As noted in Remark 1, Theorem 1 from the paper [5] still holds and the asymptotic PCA performance is unchanged. For such $\mathbf{U}$, the row samples $y^{(1)}, \dots, y^{(d)}$ are now identically distributed but they are still not independent; dependence arises because $\mathbf{Z}$ is shared. To remove the dependence, one could make $\mathbf{Z}$ deterministic and also design it so that the row samples are iid with covariance matrix matching that of (S3), but doing so no longer matches the inter-sample heteroscedastic model. It corresponds instead to having deterministic coefficients associated with a random subspace. Thus to analyze asymptotic PCA performance for row samples one should still use Theorem 1 in the paper [5] rather than existing results for the generalized spiked covariance model [1, 10].

## S2. Additional properties

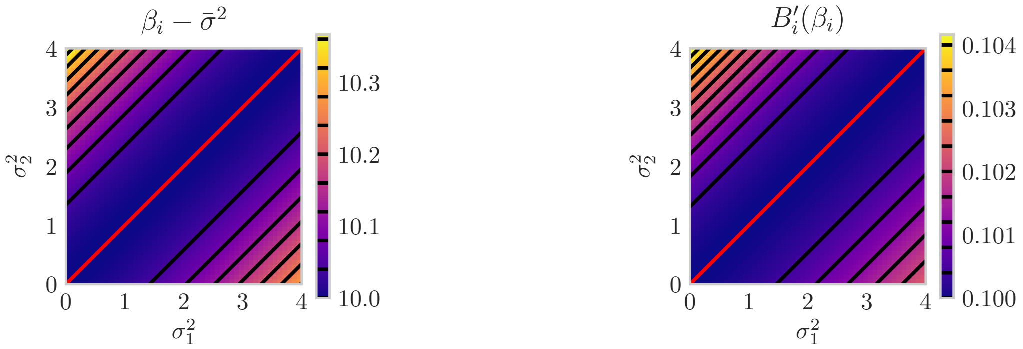
This section highlights a few additional properties of $\beta_i$, $B_i'(\beta_i)$ and $A(\beta_i)$ that lend deeper insight into how they vary with the noise variances $\sigma_1^2, \dots, \sigma_L^2$.

### *S2.1. Expressing $A(\beta_i)$ in terms of $\beta_i$ and $B_i'(\beta_i)$*

We can rewrite $A(\beta_i)$ in terms of $\beta_i$ and $B_i'(\beta_i)$ as follows:

$$\begin{aligned}
A(\beta_i) &= 1 - c\sum_{\ell=1}^{L} \frac{p_\ell \sigma_\ell^4}{\left(\beta_i - \sigma_\ell^2\right)^2} = 1 - c\sum_{\ell=1}^{L} p_\ell \left\{1 - \frac{-2\beta_i\sigma_\ell^2 + \beta_i^2}{\left(\beta_i - \sigma_\ell^2\right)^2}\right\} \\
&= 1 - c\sum_{\ell=1}^{L} p_\ell \left\{1 - \frac{-2\beta_i\sigma_\ell^2 + 2\beta_i^2 - \beta_i^2}{\left(\beta_i - \sigma_\ell^2\right)^2}\right\} \\
&= 1 - c\sum_{\ell=1}^{L} p_\ell \left\{1 + \beta_i^2 \frac{1}{\left(\beta_i - \sigma_\ell^2\right)^2} - 2\beta_i \frac{1}{\beta_i - \sigma_\ell^2}\right\} \\
&= 1 - c\sum_{\ell=1}^{L} p_\ell - c\beta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell}{\left(\beta_i - \sigma_\ell^2\right)^2} + 2c\beta_i \sum_{\ell=1}^{L} \frac{p_\ell}{\beta_i - \sigma_\ell^2} \\
&= 1 - c - c\beta_i^2 \left\{\frac{1}{c\theta_i^2} B_i'(\beta_i)\right\} + 2c\beta_i \left\{\frac{1 - B_i(\beta_i)}{c\theta_i^2}\right\} \\
&= 1 - c - \frac{\beta_i}{\theta_i^2}\{\beta_i B_i'(\beta_i) - 2\}, \qquad \text{(S8)}
\end{aligned}$$

since $B_i(\beta_i) = 0$. Thus we focus on properties of $\beta_i$ and $B_i'(\beta_i)$ for the remainder of Section S2; (S8) relates them back to $A(\beta_i)$.



**(a)** $\beta_i - \bar{\sigma}^2$ over $\sigma_1^2$ and $\sigma_2^2$.

**(b)** $B_i'(\beta_i)$ over $\sigma_1^2$ and $\sigma_2^2$.

**Figure S3:** Illustration of $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$ as a function of two noise variances $\sigma_1^2$ and $\sigma_2^2$. The level curves are along lines parallel to $\sigma_1^2 = \sigma_2^2$ for all values of sample-to-dimension ratio $c$, proportions $p_1$ and $p_2$, and subspace amplitude $\theta_i$

### *S2.2. Graphical illustration of $\beta_i$*

Note that $\beta_i$ is the largest solution of

$$\frac{1}{c\theta_i^2} = \sum_{\ell=1}^{L} \frac{p_\ell}{x - \sigma_\ell^2}, \tag{S9}$$

because $\beta_i$ is the largest real root of $B_i(x)$. Figure S2 illustrates (S9) for two noise variances $\sigma_1^2 = 2$ and $\sigma_2^2 = 0.75$, occurring in proportions $p_1 = 70\%$ and $p_2 = 30\%$, where the sample-to-dimension ratio is $c = 1$ and the subspace amplitude is $\theta_i = 1$. The plot is a graphical representation of $\beta_i$ and gives a way to visualize the relationship between $\beta_i$ and the model parameters. Observe, for example, that $\beta_i$ is larger than all the noise variances and that increasing $\theta_i$ or $c$ amounts to moving the horizontal red line down and tracking the location of the intersection.

### *S2.3. Level curves*

Figure S3 shows $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$ as functions (implicitly) of $L = 2$ noise variances $\sigma_1^2$ and $\sigma_2^2$, where

$$\bar{\sigma}^2 = p_1\sigma_1^2 + \cdots + p_L\sigma_L^2$$

is the average noise variance. Figure S3 illustrates that lines parallel to the diagonal $\sigma_1^2 = \sigma_2^2$ are level curves for both $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$. This is a general phenomenon: lines parallel to the diagonal $\sigma_1^2 = \cdots = \sigma_L^2$ are level curves of both $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$ for all sample-to-dimension ratios $c$, proportions $p_1, \dots, p_L$ and subspace amplitudes $\theta_i$.

To show this fact, note that $\beta_i - \bar{\sigma}^2$ is the largest real solution to

$$0 = B_i(x + \bar{\sigma}^2) = 1 - c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell}{x - (\sigma_\ell^2 - \bar{\sigma}^2)}, \tag{S10}$$

because $0 = B_i(\beta_i)$. Changing the noise variances to $\sigma_1^2 + \Delta, \dots, \sigma_L^2 + \Delta$ for some $\Delta$ also changes the average noise variance to $\bar{\sigma}^2 + \Delta$ and so $\sigma_\ell^2 - \bar{\sigma}^2$ remains unchanged. As a result, the solutions to (S10) remain unchanged.

Similarly, note that

$$B_i'(\beta_i) = c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell}{(\beta_i - \sigma_\ell^2)^2} = c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell}{\{(\beta_i - \bar{\sigma}^2) - (\sigma_\ell^2 - \bar{\sigma}^2)\}^2} \tag{S11}$$

remains unchanged when changing the noise variances to $\sigma_1^2 + \Delta, \dots, \sigma_L^2 + \Delta$.

Thus we conclude from (S10) and (S11) that lines parallel to $\sigma_1^2 = \cdots = \sigma_L^2$ are level curves for both $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$. The line $\sigma_1^2 = \cdots = \sigma_L^2$ in particular minimizes the value of both, as was established in the proof of Theorem 2.

### *S2.4. Hessians along the line $\sigma_1^2 = \cdots = \sigma_L^2$*

We consider $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$ as functions (implicitly) of the noise variances $\sigma_1^2, \ldots, \sigma_L^2$. To denote derivatives more clearly, we denote the $i$th noise variance as $v_i = \sigma_i^2$.

Written in this notation, we have

$$0 = 1 - c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell}{\beta_i - v_\ell}, \tag{S12}$$

$$B_i'(\beta_i) = c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell}{(\beta_i - v_\ell)^2}. \tag{S13}$$

Taking the total derivative of (S12) with respect to $v_s$ and $v_t$ and solving for $\partial^2\beta_i/(\partial v_t \partial v_s)$ yields an initially complicated expression, but evaluating it on the line $v_1 = \cdots = v_L$ vastly simplifies it, yielding:

$$\frac{\partial^2(\beta_i - \bar{\sigma}^2)}{\partial v_t \partial v_s} = \frac{2}{c\theta_i^2}(p_s\delta_{s,t} - p_s p_t). \tag{S14}$$

where $\delta_{s,t} = 1$ if $s = t$ and 0 otherwise. Notably, $\bar{\sigma}^2 = p_1 v_1 + \cdots + p_L v_L$ has zero Hessian everywhere.

Likewise, taking the total derivative of (S13) with respect to $v_s$ and $v_t$ yields an initially complicated expression that is again vastly simplified by evaluating it on the line $v_1 = \cdots = v_L$, yielding:

$$\frac{\partial^2 B_i'(\beta_i)}{\partial v_t \partial v_s} = \frac{2}{(c\theta_i^2)^4}(p_s\delta_{s,t} - p_s p_t). \tag{S15}$$

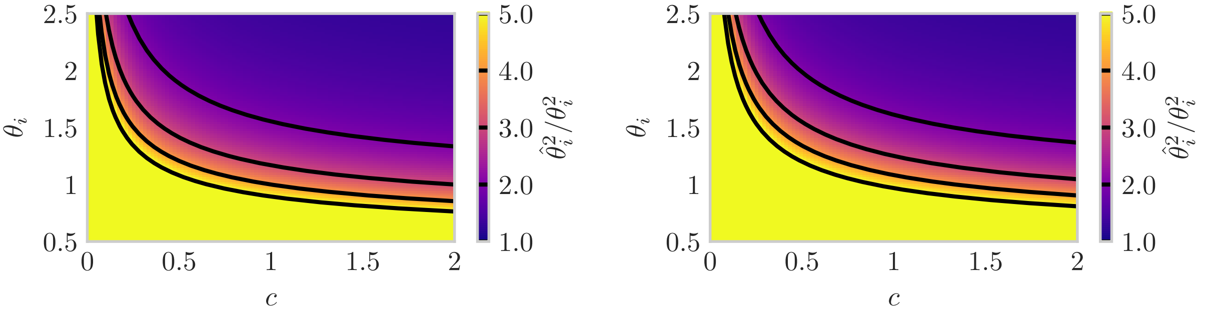
(S14) and (S15) show that the Hessian matrices for $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$ are both scaled versions of the matrix

$$\mathbf{H} = \underbrace{\begin{bmatrix} p_1 & & \\ & \ddots & \\ & & p_L \end{bmatrix}}_{\text{diag}(p)} - \underbrace{\begin{bmatrix} p_1 \\ \vdots \\ p_L \end{bmatrix} \begin{bmatrix} p_1 & \cdots & p_L \end{bmatrix}}_{pp^\top} \tag{S16}$$

on the line $v_1 = \cdots = v_L$. The (scaled) Hessian matrix (S16) is a rank one perturbation by $-pp^\top$ of diag($p$), and so its eigenvalues downward interlace with those of diag($p$) (see Theorem 8.1.8 of [4]). Namely, $\mathbf{H}$ has eigenvalues $\lambda_1, \ldots, \lambda_L$ satisfying

$$\lambda_1 \leq p_{(1)} \leq \lambda_2 \leq \cdots \leq \lambda_L \leq p_{(L)},$$

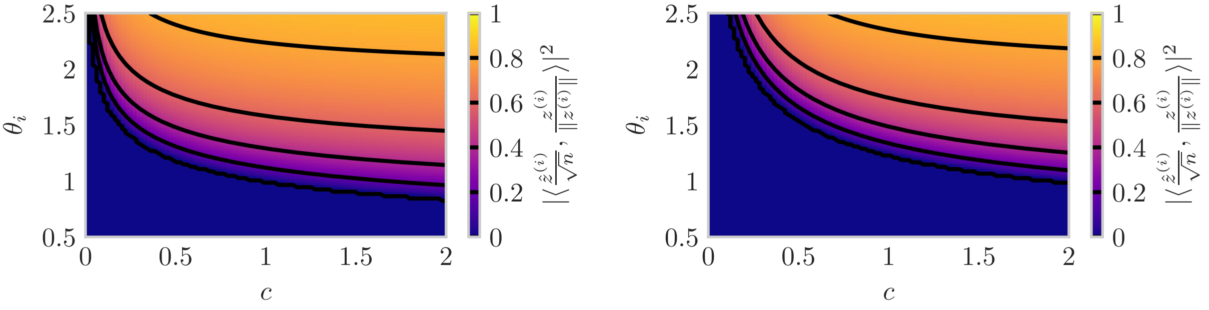
where $p_{(1)}, \ldots, p_{(L)}$ are the proportions in increasing order. The vector $\mathbf{1}$ of all ones, i.e., the vector in the direction of $v_1 = \cdots = v_L$, is an eigenvector of $\mathbf{H}$ with eigenvalue zero; note that $\mathbf{H1} = \text{diag}(p)\mathbf{1} - pp^\top\mathbf{1} = p - p = 0$. This eigenvalue is less than $p_{(1)} > 0$ and so $\lambda_1 = 0$ and $\lambda_2, \ldots, \lambda_L \geq p_{(1)} > 0$. Hence the Hessians of $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$ are both zero in the direction of the line $v_1 = \cdots = v_L$ and positive definite in other directions. This property provides deeper insight into the fact that $\beta_i - \bar{\sigma}^2$ and $B_i'(\beta_i)$ are minimized on the line $\sigma_1^2 = \cdots = \sigma_L^2$, as was established in the proof of Theorem 2.



**(a)** Homoscedastic noise with $\sigma_1^2 = 1$.

**(b)** Heteroscedastic noise with $p_1 = 80\%$ of samples at $\sigma_1^2 = 0.8$ and $p_2 = 20\%$ of samples at $\sigma_2^2 = 1.8$.

**Figure S4:** Asymptotic amplitude bias (2) of the *i*th PCA amplitude as a function of sample-to-dimension ratio $c$ and subspace amplitude $\theta_i$ with average noise variance equal to one. Contours are overlaid in black. The contours in (b) are slightly further up and to the right than in (a); more samples are needed to reduce the positive bias.



**(a)** Homoscedastic noise with $\sigma_1^2 = 1$.

**(b)** Heteroscedastic noise with $p_1 = 80\%$ of samples at $\sigma_1^2 = 0.8$ and $p_2 = 20\%$ of samples at $\sigma_2^2 = 1.8$.

**Figure S5:** Asymptotic coefficient recovery (5) of the *i*th score vector as a function of sample-to-dimension ratio $c$ and subspace amplitude $\theta_i$ with average noise variance equal to one. Contours are overlaid in black and the region where $A(\beta_i) \leq 0$ is shown as zero (the prediction of Conjecture 1). The phase transition in (b) is further right than in (a); more samples are needed to recover the same strength signal.

## S3. Impact of parameters: amplitude and coefficient recovery

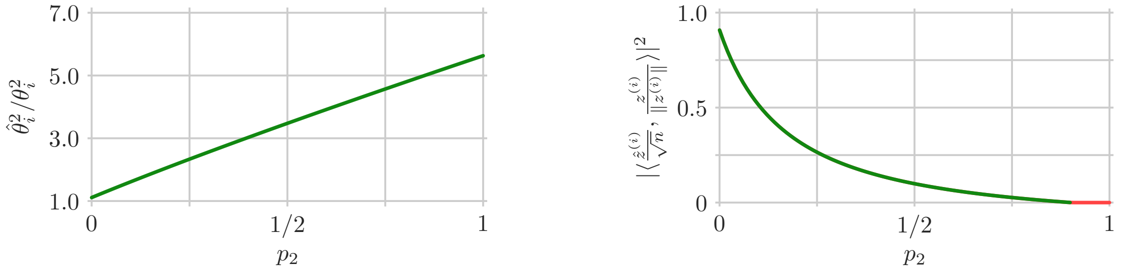
Section 3 of [5] discusses how the asymptotic subspace recovery (4) of Theorem 1 depends on the model parameters: sample-to-dimension ratio $c$, subspace amplitudes $\theta_1, \ldots, \theta_k$, proportions $p_1, \ldots, p_L$ and noise variances $\sigma_1^2, \ldots, \sigma_L^2$. This section shows that the same phenomena occur for the asymptotic PCA amplitudes (2) and coefficient recovery (5). For the asymptotic PCA amplitudes, we consider the ratio $\hat{\theta}_i^2/\theta_i^2$. As discussed in Remark 4, the asymptotic PCA amplitude $\hat{\theta}_i$ is positively biased relative to the subspace amplitude $\theta_i$, and so the almost sure limit of $\hat{\theta}_i^2/\theta_i^2$ is greater than one, with larger values indicating more bias.

### *S3.1. Impact of sample-to-dimension ratio c and subspace amplitude $\theta_i$*

As in Section 3.1, we vary the sample-to-dimension ratio $c$ and subspace amplitude $\theta_i$ in two scenarios:

a) there is only one noise variance fixed at $\sigma_1^2 = 1$

b) there are two noise variances $\sigma_1^2 = 0.8$ and $\sigma_2^2 = 1.8$ occurring in proportions $p_1 = 80\%$ and $p_2 = 20\%$.



**(a)** Asymptotic amplitude bias (2).

**(b)** Asymptotic coefficient recovery (5).

**Figure S6:** Asymptotic amplitude bias (2) and coefficient recovery (5) of the $i$th PCA amplitude and score vector as functions of the contamination fraction $p_2$, the proportion of samples with noise variance $\sigma_2^2 = 3.25$, where the other noise variance $\sigma_1^2 = 0.1$ occurs in proportion $p_1 = 1 - p_2$. The sample-to-dimension ratio is $c = 10$ and the subspace amplitude is $\theta_i = 1$. The region where $A(\beta_i) \leq 0$ is the red horizontal segment in (b) with value zero (the prediction of Conjecture 1).

Both scenarios have average noise variance 1. Figures S4 and S5 show analogous plots to Figure 1 but for the asymptotic PCA amplitudes (2) and coefficient recovery (5), respectively.

As was the case for Figure 1 in Section 3.1, decreasing the subspace amplitude $\theta_i$ degrades both the asymptotic amplitude performance (i.e., increases bias) shown in Figure S4 and the asymptotic coefficient recovery shown in Figure S5, but the lost performance could be regained by increasing the number of samples. Furthermore, both the asymptotic amplitude performance shown in Figure S4 and the asymptotic coefficient recovery shown in Figure S5 decline when the noise is heteroscedastic. Though the difference is subtle for the asymptotic amplitude bias, the contours move up and to the right in both cases. This degradation is consistent with Theorem 2; PCA performs worse on heteroscedastic data than it does on homoscedastic data of the same average noise variance and more samples or a larger subspace amplitude are needed to compensate.

### *S3.2. Impact of proportions $p_1, \dots, p_L$*

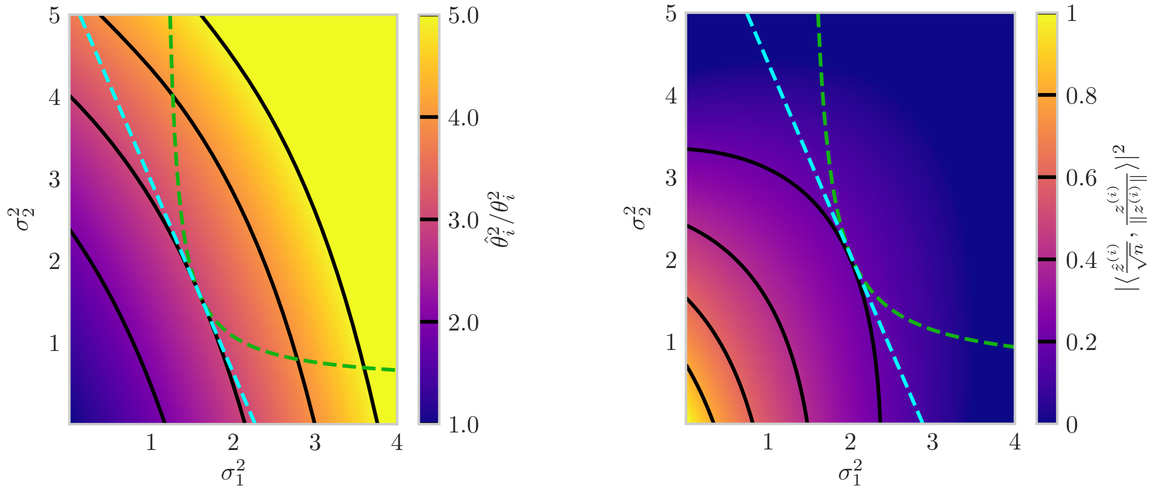
As in Section 3.2, we consider two noise variances $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 3.25$ occurring in proportions $p_1 = 1 - p_2$ and $p_2$, where the sample-to-dimension ratio is $c = 10$ and the subspace amplitude is $\theta_i = 1$. Figure S6 shows analogous plots to Figure 2 but for the asymptotic PCA amplitudes (2) and coefficient recovery (5). As was the case for Figure 2 in Section 3.2, performance generally degrades in Figure S6 as $p_2$ increases and low noise samples with noise variance $\sigma_1^2$ are traded for high noise samples with noise variance $\sigma_2^2$. The performance is best when $p_2 = 0$ and all the samples have the smaller noise variance $\sigma_1^2$, i.e., there is no contamination.

### *S3.3. Impact of noise variances $\sigma_1^2, \dots, \sigma_L^2$*

As in Section 3.3, we consider two noise variances $\sigma_1^2$ and $\sigma_2^2$ occurring in proportions $p_1 = 70\%$ and $p_2 = 30\%$, where the sample-to-dimension ratio is $c = 10$ and the subspace amplitude is $\theta_i = 1$. Figure S7 shows analogous plots to Figure 3 but for the asymptotic PCA amplitudes (2) and coefficient recovery (5). As was the case for Figure 3 in Section 3.3, performance typically degrades with increasing noise variances. The contours in Figure S7b are also generally horizontal for small $\sigma_1^2$ and vertical for small $\sigma_2^2$. They indicate that when the gap between the two largest noise variances is "sufficiently" wide, the asymptotic coefficient recovery is roughly determined by the largest noise variance. This property mirrors the asymptotic subspace recovery and occurs for similar reasons, discussed in detail in Section 3.3. Along each dashed cyan line in Figure S7, the average noise variance is fixed and the best performance for both the PCA amplitudes and coefficient recovery again occurs when $\sigma_1^2 = \sigma_2^2 = \bar{\sigma}^2$, as was predicted by Theorem 2. Along each dashed green curve in Figure S7, the average inverse noise variance is fixed and the best performance for both the PCA amplitudes and coefficient recovery again occurs when $\sigma_1^2 = \sigma_2^2$, as was predicted in Remark 6.

### *S3.4. Impact of adding data*

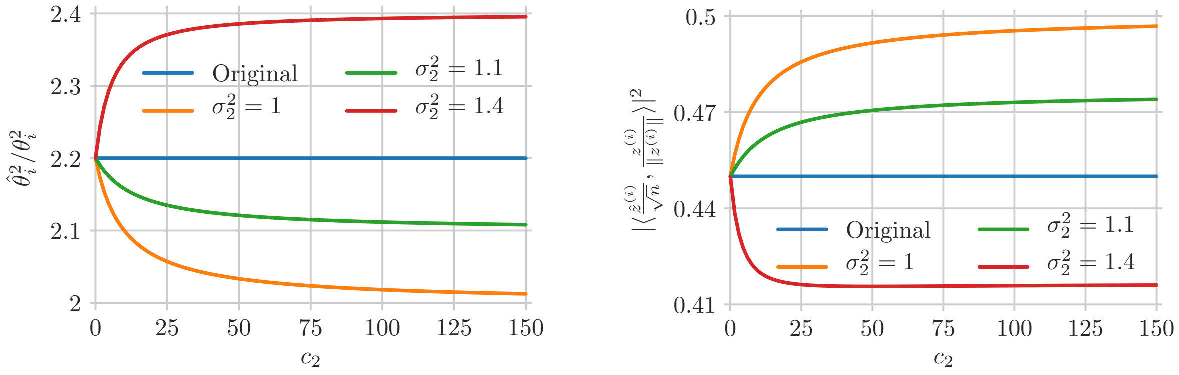
As in Section 3.4, we consider adding data with noise variance $\sigma_2^2$ and sample-to-dimension ratio $c_2$ to an existing dataset that has noise variance $\sigma_1^2 = 1$, sample-to-dimension ratio $c_1 = 10$ and subspace amplitude $\theta_i = 1$ for the $i$th



**(a)** Asymptotic amplitude bias (2).

**(b)** Asymptotic coefficient recovery (5).

**Figure S7:** Asymptotic amplitude bias (2) and coefficient recovery (5) of the $i$th PCA amplitude and score vector as functions of noise variances $\sigma_1^2$ and $\sigma_2^2$ occurring in proportions $p_1 = 70\%$ and $p_2 = 30\%$, where the sample-to-dimension ratio is $c = 10$ and the subspace amplitude is $\theta_i = 1$. Contours are overlaid in black and the region where $A(\beta_i) \leq 0$ is shown as zero in (b), matching the prediction of Conjecture 1. Along each dashed cyan line, the average noise variance is fixed and the best performance occurs when $\sigma_1^2 = \sigma_2^2 = \bar{\sigma}^2$. Along each dashed green curve, the average inverse noise variance is fixed and the best performance again occurs when $\sigma_1^2 = \sigma_2^2$.



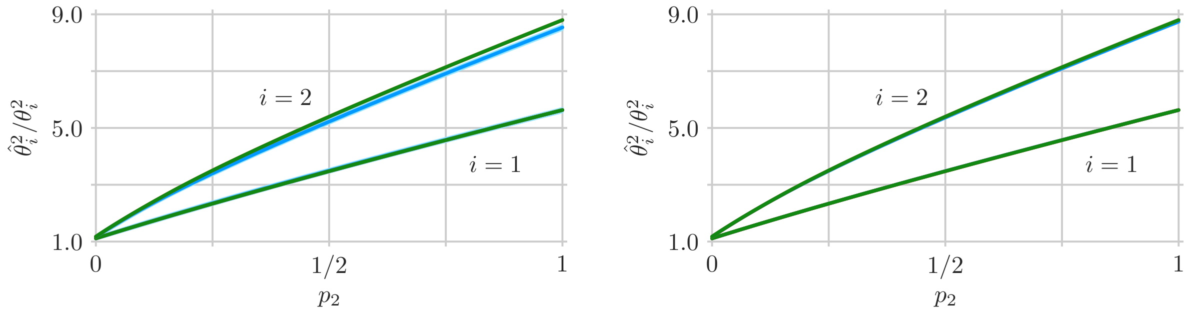
**(a)** Asymptotic amplitude bias (2).

**(b)** Asymptotic coefficient recovery (5).

**Figure S8:** Asymptotic amplitude bias (2) and coefficient recovery (5) of the $i$th PCA amplitude and score vector for samples added with noise variance $\sigma_2^2$ and samples-per-dimension $c_2$ to an existing dataset with noise variance $\sigma_1^2 = 1$, sample-to-dimension ratio $c_1 = 10$ and subspace amplitude $\theta_i = 1$.

component. The combined dataset has a sample-to-dimension ratio of $c = c_1 + c_2$ and is potentially heteroscedastic with noise variances $\sigma_1^2$ and $\sigma_2^2$ appearing in proportions $p_1 = c_1/c$ and $p_2 = c_2/c$.

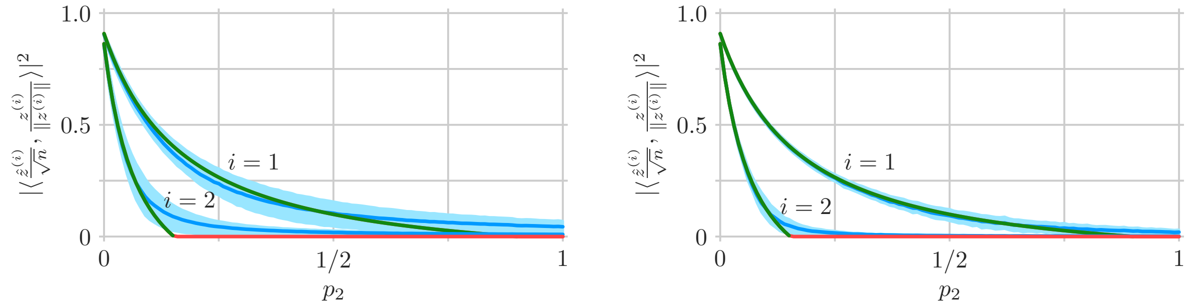
Figure S8 shows analogous plots to Figure 4 in Section 3.4 but for the asymptotic PCA amplitudes (2) and coefficient recovery (5). As was the case for Figure 4, the orange curves show the recovery when $\sigma_2^2 = 1 = \sigma_1^2$ and illustrate the benefit we would expect for homoscedastic data: increasing the samples per dimension improves recovery. The green curves show the performance when $\sigma_2^2 = 1.1 > \sigma_1^2$; as before, these samples are “slightly” noisier and performance improves for any number added. Finally, the red curves show the performance when $\sigma_2^2 = 1.4 > \sigma_1^2$.



**(a)** $10^3$ samples in $10^2$ dimensions.

**(b)** $10^4$ samples in $10^3$ dimensions.

**Figure S9:** Simulated amplitude bias (2) as a function of the contamination fraction $p_2$, the proportion of samples with noise variance $\sigma_2^2 = 3.25$, where the other noise variance $\sigma_1^2 = 0.1$ occurs in proportion $p_1 = 1 - p_2$. The sample-to-dimension ratio is $c = 10$ and the subspace amplitudes are $\theta_1 = 1$ and $\theta_2 = 0.8$. Simulation mean (blue curve) and interquartile interval (light blue ribbon) are shown with the asymptotic bias (2) of Theorem 1 (green curve). Increasing data size from (a) to (b) results in even smaller interquartile intervals, indicating concentration to the mean, which is converging to the asymptotic bias.



**(a)** $10^3$ samples in $10^2$ dimensions.

**(b)** $10^4$ samples in $10^3$ dimensions.

**Figure S10:** Simulated coefficient recovery (5) as a function of the contamination fraction $p_2$, the proportion of samples with noise variance $\sigma_2^2 = 3.25$, where the other noise variance $\sigma_1^2 = 0.1$ occurs in proportion $p_1 = 1 - p_2$. The sample-to-dimension ratio is $c = 10$ and the subspace amplitudes are $\theta_1 = 1$ and $\theta_2 = 0.8$. Simulation mean (blue curve) and interquartile interval (light blue ribbon) are shown with the asymptotic recovery (5) of Theorem 1 (green curve). The region where $A(\beta_i) \leq 0$ is the red horizontal segment with value zero (the prediction of Conjecture 1). Increasing data size from (a) to (b) results in smaller interquartile intervals, indicating concentration to the mean, which is converging to the asymptotic recovery.

As before, performance degrades when adding a small number of these noisier samples. However, unlike subspace recovery, performance degrades when adding any amount of these samples. In the limit $c_2 \to \infty$, the asymptotic amplitude bias is $1 + \sigma_2^2/\theta_i^2$ and the asymptotic coefficient recovery is $1/(1 + \sigma_2^2/\theta_i^2)$; neither has perfect recovery in the limit when added samples are noisy.

## S4. Numerical simulation: amplitude and coefficient recovery

Section 4 of [5] shows that the asymptotic subspace recovery (4) of Theorem 1 is meaningful for practical settings with finitely many samples in a finite-dimensional space. This section shows that the same is true for the asymptotic PCA amplitudes (2) and coefficient recovery (5). For the asymptotic PCA amplitudes, we again consider the ratio $\hat{\theta}_i^2/\theta_i^2$. As discussed in Remark 4, the asymptotic PCA amplitude $\hat{\theta}_i$ is positively biased relative to the subspace amplitude $\theta_i$, and so the almost sure limit of $\hat{\theta}_i^2/\theta_i^2$ is greater than one, with larger values indicating more bias.

As in Section 4, this section simulates data according to the model described in Section 2.1 for a two-dimensional subspace with subspace amplitudes $\theta_1 = 1$ and $\theta_2 = 0.8$, two noise variances $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 3.25$, and a sample-to-dimension ratio of $c = 10$. We sweep the proportion of high noise points $p_2$ from zero to one, setting $p_1 = 1 - p_2$ as in Section 4. The first simulation considers $n = 10^3$ samples in a $d = 10^2$ dimensional ambient space ($10^4$ trials). The second increases these to $n = 10^4$ samples in a $d = 10^3$ dimensional ambient space ($10^3$ trials). All simulations generate data from the standard normal distribution, i.e., $z_{ij}, \varepsilon_{ij} \sim \mathcal{N}(0, 1)$. Figures S9 and S10 show analogous plots to Figure 5 but for the asymptotic PCA amplitudes (2) and coefficient recovery (5), respectively.

As was the case for Figure 5 in Section 4, both Figures S9 and S10 illustrate the following general observations:

a) the simulation mean and almost sure limit generally agree in the smaller simulation of $10^3$ samples in a $10^2$ dimensional ambient space

b) the smooth simulation mean deviates from the non-smooth almost sure limit near the phase transition

c) the simulation mean and almost sure limit agree better for the larger simulation of $10^4$ samples in a $10^3$ dimensional ambient space

d) the interquartile intervals for the larger simulations are roughly half those of the smaller simulations, indicating concentration to the means.

In fact, the amplitude bias in Figure S9 and the coefficient recovery in Figure S10 both have significantly better agreement with their almost sure limits than the subspace recovery in Figure 5 has with its almost sure limit. The amplitude bias in Figure S9, in particular, is tightly concentrated around its almost sure limit (2). Furthermore, Figure S10 demonstrates good agreement with Conjecture 1, providing evidence that there is indeed a phase transition below which the coefficients are also not recovered.

## S5. Additional numerical simulations

Section 4 of [5] and Section S4 provide numerical simulation results for real-valued data generated using normal distributions. This section illustrates the generality of the model in Section 2.1 by showing analogous simulation results for circularly symmetric complex normal data in Figure S11 and for a mixture of Gaussians in Figure S12. As before, we show the results of two simulations for each setting. The first simulation considers $n = 10^3$ samples in a $d = 10^2$ dimensional ambient space ($10^4$ trials). The second increases these to $n = 10^4$ samples in a $d = 10^3$ dimensional ambient space ($10^3$ trials).

Figure S11 mirrors Sections 4 and S4 and simulates data according to the model described in Section 2.1 for a two-dimensional subspace with subspace amplitudes $\theta_1 = 1$ and $\theta_2 = 0.8$, two noise variances $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 3.25$, and a sample-to-dimension ratio of $c = 10$. We again sweep the proportion of high noise points $p_2$ from zero to one, setting $p_1 = 1 - p_2$. The only difference is that Figure S11 generates data from the standard *complex* normal distribution, i.e., $z_{ij}, \varepsilon_{ij} \sim \mathcal{CN}(0, 1)$.

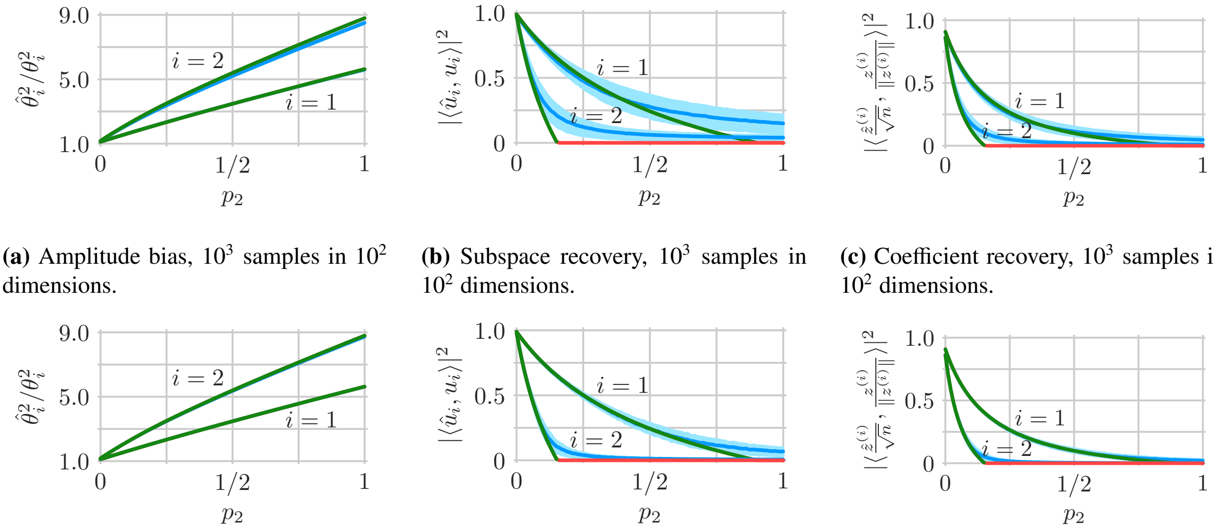
Figure S12 instead simulates a *homoscedastic* setting of the model described in Section 2.1 over a range of noise distributions, all *mixtures* of Gaussians. As before, we consider a two-dimensional subspace with subspace amplitudes $\theta_1 = 1$ and $\theta_2 = 0.8$, and a sample-to-dimension ratio of $c = 10$. Figure S12 generates coefficients $z_{ij} \sim \mathcal{N}(0, 1)$ from the standard normal distribution and generates noise entries $\varepsilon_{ij}$ from the Gaussian mixture model

$$\varepsilon_{ij} \sim \begin{cases} \mathcal{N}\left(0, \lambda_1^2/\sigma^2\right) & \text{with probability } p_1, \\ \mathcal{N}\left(0, \lambda_2^2/\sigma^2\right) & \text{with probability } p_2, \end{cases}$$

where $\lambda_1^2 = 0.1$ and $\lambda_2^2 = 3.25$, and the *single* noise variance is set to

$$\sigma^2 = p_1\lambda_1^2 + p_2\lambda_2^2. \tag{S17}$$

Each scaled noise entry $\eta_i\varepsilon_{ij} = \sigma\varepsilon_{ij}$ is a mixture of two Gaussian distributions with variances $\lambda_1^2$ and $\lambda_2^2$. We sweep the mixture probability $p_2$ from zero to one, setting $p_1 = 1 - p_2$. Thus, Figure S12 illustrates performance over a range of noise distributions. The noise variance (S17) in Figure S12 matches the average noise variance in Figure S11



**(a)** Amplitude bias, $10^3$ samples in $10^2$ dimensions.

**(b)** Subspace recovery, $10^3$ samples in $10^2$ dimensions.

**(c)** Coefficient recovery, $10^3$ samples in $10^2$ dimensions.

**(d)** Amplitude bias, $10^4$ samples in $10^3$ dimensions.

**(e)** Subspace recovery, $10^4$ samples in $10^3$ dimensions.

**(f)** Coefficient recovery, $10^4$ samples in $10^3$ dimensions.

**Figure S11:** Simulated complex-normal PCA performance as a function of the contamination fraction $p_2$, the proportion of samples with noise variance $\sigma_2^2 = 3.25$, where the other noise variance $\sigma_1^2 = 0.1$ occurs in proportion $p_1 = 1 - p_2$. The sample-to-dimension ratio is $c = 10$ and the subspace amplitudes are $\theta_1 = 1$ and $\theta_2 = 0.8$. Simulation mean (blue curve) and interquartile interval (light blue ribbon) are shown with the almost sure limits of Theorem 1 (green curve). The region where $A(\beta_i) \leq 0$ is shown as red horizontal segments with value zero (the prediction of Conjecture 1).

as we sweep $p_2$. However, Figures S12 and S11 differ because Figure S12 simulates a *homoscedastic* setting while Figure S11 simulates a *heteroscedastic* setting. Figure S12 also differs from Figure S1b that simulates data from the random inter-sample heteroscedastic model of Section S1.1. While both simulate (scaled) noise from a mixture model, scaled noise entries $\eta_i \varepsilon_{ij}$ in Figure S12 are all iid. Scaled noise entries $\eta_i \varepsilon_{ij}$ in the random inter-sample heteroscedastic model are independent only across samples; they are *not* independent within each sample. Figure S12 is instead more like Figure S1c that simulates data from the Johnstone spiked covariance model. See Section S1.1 for a comparison of these models.

As was the case for (real-valued) standard normal data in Sections 4 and S4, Figures S11 and S12 illustrate the following general observations:

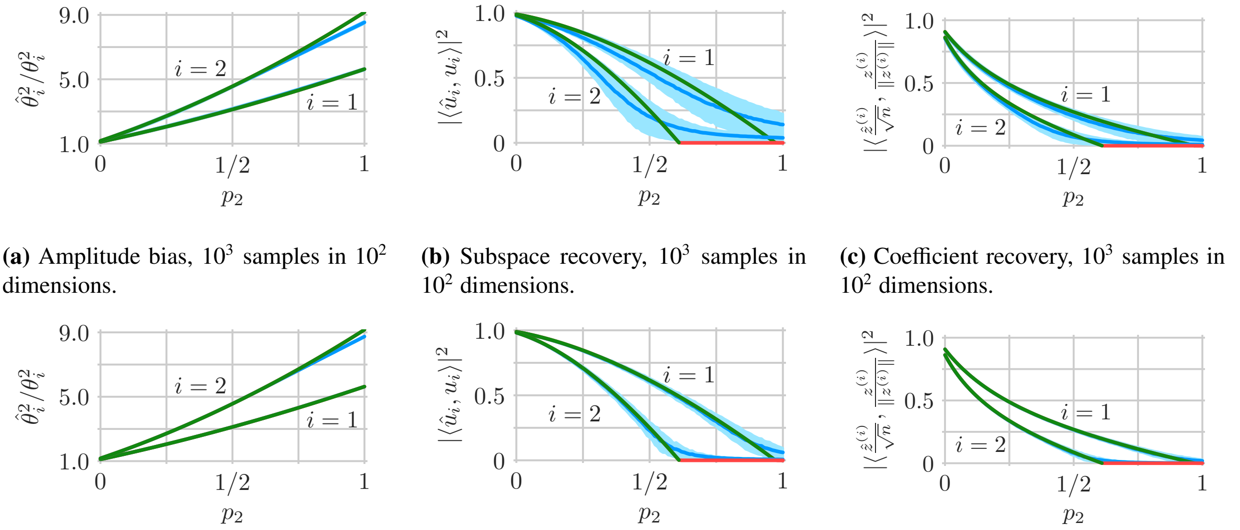
a) the simulation means and almost sure limits generally agree in the smaller simulations of $10^3$ samples in a $10^2$ dimensional ambient space

b) the smooth simulation means deviate from the non-smooth almost sure limits near the phase transitions

c) the simulation means and almost sure limits agree better for the larger simulations of $10^4$ samples in a $10^3$ dimensional ambient space

d) the interquartile intervals for the larger simulations are roughly half those of the smaller simulations, indicating concentration to the means.

The agreement between simulations and almost sure limits demonstrated in both Figures S11 and S12 highlights the generality of the model considered in [5]: it allows for both complex-valued data and non-Gaussian distributions. In both cases, the asymptotic results of Theorem 1 remain meaningful for practical settings with finitely many samples in a finite-dimensional space.



**(a)** Amplitude bias, $10^3$ samples in $10^2$ dimensions.

**(b)** Subspace recovery, $10^3$ samples in $10^2$ dimensions.

**(c)** Coefficient recovery, $10^3$ samples in $10^2$ dimensions.

**(d)** Amplitude bias, $10^4$ samples in $10^3$ dimensions.

**(e)** Subspace recovery, $10^4$ samples in $10^3$ dimensions.

**(f)** Coefficient recovery, $10^4$ samples in $10^3$ dimensions.

**Figure S12:** Simulated mixture model PCA performance as a function of the mixture probability $p_2$, the probability that a scaled noise entry $\eta_i \varepsilon_{ij}$ is Gaussian with variance $\lambda_2^2 = 3.25$, where it is Gaussian with variance $\lambda_1^2 = 0.1$ otherwise, i.e., with probability $p_1 = 1 - p_2$. The sample-to-dimension ratio is $c = 10$ and the subspace amplitudes are $\theta_1 = 1$ and $\theta_2 = 0.8$. Simulation mean (blue curve) and interquartile interval (light blue ribbon) are shown with the almost sure limits of Theorem 1 (green curve). The region where $A(\beta_i) \leq 0$ is shown as red horizontal segments with value zero (the prediction of Conjecture 1).

## References

[1] Z. Bai, J. Yao, On sample eigenvalues in a generalized spiked population model, J. Multivariate Anal. 106 (2012) 167–177.

[2] F. Benaych-Georges, R.R. Nadakuditi, The singular values and vectors of low rank perturbations of large rectangular random matrices, J. Multivariate Anal. 111 (2012) 120 – 135.

[3] M. Biehl, A. Mietzner, Statistical mechanics of unsupervised structure recognition, J. Phys. A 27 (1994) 1885–1897.

[4] G. Golub, C. Van Loan, Matrix Computations, Johns Hopkins University Press.

[5] D. Hong, L. Balzano, J. A. Fessler, Asymptotic performance of PCA for high-dimensional heteroscedastic data, under review, 2018.

[6] I.M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statist. 29 (2001) 295–327.

[7] I.M. Johnstone, A.Y. Lu, On consistency and sparsity for principal components analysis in high dimensions, J. Amer. Statist. Assoc. 104 (2009) 682–693.

[8] B. Nadler, Finite sample approximation results for principal component analysis: A matrix perturbation approach, Ann. Statist. 36 (2008) 2791–2817.

[9] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statist. Sinica 17 (2007) 1617–1642.

[10] J. Yao, S. Zheng, Z. Bai, Large Sample Covariance Matrices and High-Dimensional Data Analysis, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 2015.