Singular vectors of sums of rectangular random matrices and optimal estimation of high-rank signals: The extensive spike model

Itamar D. Landau [1,*] Gabriel C. Mel,[2] and Surya Ganguli[1]
[1]*Department of Applied Physics, Stanford University, Stanford, California 94305, USA*
[2]*Neuroscience Graduate Program, Stanford University, Stanford, California 94305, USA*

Across many disciplines spanning from neuroscience and genomics to machine learning, atmospheric science, and finance, the problems of denoising large data matrices to recover hidden signals obscured by noise, and of estimating the structure of these signals, is of fundamental importance. A key to solving these problems lies in understanding how the singular value structure of a signal is deformed by noise. This question has been thoroughly studied in the well-known spiked matrix model, in which data matrices originate from low-rank signal matrices perturbed by additive noise matrices, in an asymptotic limit where matrix size tends to infinity but the signal rank remains finite. We first show, strikingly, that the singular value structure of large finite matrices (of size ∼1000) with even moderate-rank signals, as low as 10, is not accurately predicted by the finite-rank theory, thereby limiting the application of this theory to real data. To address these deficiencies, we analytically compute how the singular values and vectors of an arbitrary *high*-rank signal matrix are deformed by additive noise. We focus on an asymptotic limit corresponding to an *extensive* spike model, in which *both* the signal rank and the size of the data matrix tend to infinity at a constant ratio. We map out the phase diagram of the singular value structure of the extensive spike model as a joint function of signal strength and rank. We further exploit these analytics to derive optimal rotationally invariant denoisers to recover the hidden *high*-rank signal from the data, as well as optimal invariant estimators of the signal covariance structure. Our extensive-rank results yield several conceptual differences compared to the finite-rank case: (1) as signal strength increases, the singular value spectrum does not directly transition from a unimodal bulk phase to a disconnected phase, but instead there is a bimodal connected regime separating them; (2) the signal singular vectors can be partially estimated *even* in the unimodal bulk regime, and thus the transitions in the data singular value spectrum do not coincide with a detectability threshold for the signal singular vectors, unlike in the finite-rank theory; (3) signal singular values interact nontrivially to generate data singular values in the extensive-rank model, whereas they are noninteracting in the finite-rank theory; and (4) as a result, the more sophisticated data denoisers and signal covariance estimators we derive, which take into account these nontrivial extensive-rank interactions, significantly outperform their simpler, noninteracting, finite-rank counterparts, even on data matrices of only moderate rank. Overall, our results provide fundamental theory governing how high-dimensional signals are deformed by additive noise, together with practical formulas for optimal denoising and covariance estimation.

I. INTRODUCTION

Estimating structure in high-dimensional data from noisy observations constitutes a fundamental problem across many disciplines, especially in the age of big data. A common scenario is that such data are presented as a large matrix. Such matrices could contain, for example, the observed time series of many recorded neurons in neuroscience, the expression level of many genes across many conditions in genomics, or the time series of many stock prices in finance. Given such

data matrices, one often wishes to (1) understand the structure of the data via its singular value decomposition, (2) denoise the data in order to find clean signals hidden in the data, and (3) estimate the covariance structure of these clean hidden signals. These hidden signals could correspond, for example, to temporally correlated cell assemblies in neuroscience, gene modules in genomics, or sectors of correlated stocks in finance.

These three problems of data understanding, data denoising, and signal-covariance estimation raise fundamental new challenges in the era of big data, where the number of observations (e.g., the length of time series or the number of conditions) is often comparable to the number of variables (e.g., the number of recorded neurons, genes, or stock prices). As a result, tools from random matrix theory (RMT) designed for this high-dimensional regime have grown in prominence across a wide range of disciplines, including neuroscience [1], psychology [2], genetics [3], finance [4],

*idlandau@stanford.edu

machine learning [5–10], atmospheric science [11,12], wireless communications [13], integrated energy systems [14], magnetic resonance imaging (including spectroscopy [15,16], diffusion [17], and functional MRI [18,19]).

In this work we develop RMT tools in order to quantitatively study the basic question of how the singular value decomposition (SVD) of an arbitrary high-dimensional hidden signal matrix is deformed under additive observation noise. Based on this understanding of the relation between the data and signal SVDs, we go on to derive both optimal denoisers of the data to recover the hidden signal as well as optimal estimators for the signal covariance.

An influential line of prior related research has studied spiked matrix models, focusing on an asymptotic limit in which the size of the hidden signal matrix tends to infinity but its rank remains finite [20–27]. These works consider the addition of a random noise matrix to the signal to generate a data matrix, and they study how the singular values and singular vectors of the data are related to those of the signal. The finite number of signal eigenvalues or singular values constitute a set of "spikes" in the signal spectrum, hence the name spiked matrix model.

A key observation in these models is that the addition of noise to the signal yields a data matrix that (1) has inflated singular values relative to the signal and (2) has singular vectors that are rotated relative to the signal. For the finite-rank rectangular spiked matrix model, both the degree of singular-value inflation and the angle of the singular vector rotation can be explicitly computed [22,23]. Notably, in this finite-rank regime, the multiple spikes do not interact as they get deformed from signal to data. This means that to predict the mapping from a given signal singular value to the corresponding data singular value, as well as the angle between a data and signal singular vector, one needs to know only the noise distribution and the singular value of the signal spike in question; one does not need to know all the other signal singular values. This underlying simplicity in the relation between signal and data spectral structure implies that one can optimally denoise data and optimally estimate signal covariance, by applying shrinkage functions that act independently, albeit nonlinearly, on each singular value or eigenvalue of the data [23,28,29]. The idea is that these shrinkage functions that shrink data singular values independently partially reverse the independent singular-value inflation and compensate for the independent singular-vector rotation, both due to additive noise.

In this work, however, we demonstrate that the assumptions and consequences of the finite-rank model may constitute a significant limitation for the practical application of this theory and its associated estimation techniques. For example, below we will see that the spectral structure of random matrices of large size (e.g., 1000×500) and of even moderate rank (e.g., 10) cannot be accurately modeled by the finite-rank spiked matrix model.

This lack of numerical accuracy of the finite-rank theory for large but finite-size matrices of moderate rank could have a significant impact on the three problems of spectral understanding, data denoising, and signal covariance estimation across the empirical sciences, where the effective rank of signals is expected to vary significantly, and sometimes even

be quite high. Therefore, it is imperative to develop a theory that more accurately describes data containing higher-rank signals. We develop that theory by generalizing the finite-rank theory to an *extensive-rank* theory in which the rank of the signal matrix is proportional to the size of the signal and data matrices, working in an asymptotic limit where *both* the size and rank approach infinity.

We note that it is not immediately obvious how to extend existing finite-rank results to the extensive regime. The finite-rank theory [20–22] makes use of algebraic formulas for matrices with low-rank perturbations that do not directly generalize, and so one must resort to more elaborate tools from RMT and free probability. Along these lines, powerful theoretical methods have been developed in recent years for studying the eigendecomposition of sums of square Hermitian matrices [30], and deriving techniques for optimally estimating arbitrary square-symmetric matrices from noisy observations [31–36].

However, the situation for rectangular matrices, relevant to data from many fields including neuroscience, genomics and finance, lags behind that of square matrices. While the singular value spectrum of sums of rectangular matrices has been calculated [37–40], and a few works have studied optimal denoising of rectangular matrices under a known (usually Gaussian) prior [41–43]. there are currently no methods for determining the deformation of the singular vectors of a rectangular signal matrix due to an additive noise matrix.

The outline of our paper is as follows. In Sec. II we motivate our work with an illustrative numerical study of the spiked matrix model, showing that the finite-rank theory fails to accurately predict the outlier singular values and singular vector deformations in data matrices containing even moderate-rank signals. In Sec. III we introduce tools from RMT that we will need to derive our results, including Hermitianization, block matrix resolvents, block Stieltjes transforms and their inversion formulas, and block $\mathcal{R}$ transforms. In Sec. IV we study how the singular values and singular vectors of an arbitrary rectangular signal matrix are deformed under the addition of a noise matrix to generate a data matrix. To do so, we derive a subordination relation that relates the resolvent of the Hermitianization of a data matrix to that of its hidden signal matrix in Sec. IV A. We next employ this subordination relation to derive expressions for the overlap between data singular vectors and the signal singular vectors in Sec. IV B. We then apply these results to study the extensive spike model in which the rank of the signal spike is assumed to grow linearly with the number of variables (and observations) in Sec. IV C. There we map out the phase diagram of the SVD as a joint function of signal strength and rank ratio. Intriguingly, we find that certain transitions in the singular value spectrum of the data do *not* coincide with the detectability of the signal, as they do in the finite-rank model. Finally, in Sec. V we exploit the expressions for singular vector overlaps in order to derive optimal rotationally invariant estimators for both data denoising (Sec. V A) and signal-covariance estimation (Sec. V B). We find that unlike in the finite-rank model, in the extensive-rank model signal singular values interact nontrivially to generate data singular values. Therefore, we obtain more sophisticated optimal data denoisers and signal-covariance estimators that take into ac-

count these nontrivial extensive-rank interactions, and which furthermore significantly outperform their simpler, noninteracting, finite-rank counterparts.

We note that recently a set of partially overlapping results appeared on a preprint server [44]. In our discussion section, we describe the relation and additional contributions of our work relative to that of [44].

II. A MOTIVATION: INADEQUACIES OF THE FINITE-RANK SPIKED MATRIX MODEL

Let Y be an $N_1 \times N_2$ signal matrix. We can think of each of the N_1 rows of Y as a variable, and each of the N_2 columns as a distinct experimental condition or time point, with Y_{ij} representing the clean, uncorrupted value of variable i under condition j. Now consider a noisy data matrix R, given by

$$R = Y + X, \tag{1}$$

where X is a random $N_1 \times N_2$ additive noise matrix. X is assumed to have well-defined limiting singular value spectrum in the limit of large N_1 with fixed aspect ratio, $c = N_1/N_2$. Furthermore we assume the probability distribution $P_X(X)$ over X is rotationally invariant, meaning $P_X(X) = P_X(O_1 X O_2)$, where O_1 and O_2 are orthogonal matrices of size $N_1 \times N_1$ and $N_2 \times N_2$, respectively. These assumptions guarantee the asymptotic freeness of X and Y. For a general definition of freeness, see [40].

We are interested in understanding the relationship between the singular value decomposition (SVD) of the data matrix R and the SVD of the clean signal matrix Y. In general we will write the SVD of the data as

$$R = \hat{U}_1 \hat{S} \hat{U}_2^T = \sum_k \hat{s}_k \hat{\boldsymbol{u}}_{1k} \hat{\boldsymbol{u}}_{2k}^T, \tag{2}$$

where each $\hat{U}_a$, for $a = 1, 2$, is an $N_a \times N_a$ matrix with orthonormal columns, $\hat{\boldsymbol{u}}_{ak}$ for $k = 1, \ldots, N_a$, and $\hat{S}$ is a diagonal $N_1 \times N_2$ matrix with $\hat{s}_k$ along the diagonal.

As a motivating example, we will study a version of the spiked matrix model [20–22] in which the signal matrix Y is given by

$$Y = s U_1 U_2^T = s \sum_{k=1}^K \boldsymbol{u}_{1k} \boldsymbol{u}_{2k}^T, \tag{3}$$

where each U_a, for $a = 1, 2$, is an $N_a \times K$ matrix with orthonormal columns, $\boldsymbol{u}_{ak}$ for $k = 1, \ldots, N_a$, and s is the signal strength. This signal model can be thought of as a rank K spike of strength s in that its singular value spectrum has K singular values all equal to s.

In the finite-rank setting, where K remains finite as $N_1, N_2 \to \infty$, there is a signal-detectability phase transition [20,22] in the singular value structure of the data matrix R. For $s < s_{crit}$, where s_{crit} is a critical signal strength that depends on the singular value spectrum of the noise matrix X, the entire signal in Y is swamped by the additive noise X and cannot be seen in the data R. More precisely, in the large-size limit, when $s < s_{crit}$ the singular value spectrum of the data R is *identical* to the singular value spectrum of the noise X. Furthermore, *no* left (right) singular vector of the data matrix R has an $O(1)$ overlap with the K-dimensional signal subspace

corresponding to the column space of U_1 (U_2). However, for $s > s_{crit}$ the singular value spectrum of the data R is now not only composed of a noise bulk, identical to the spectrum of X, as before, but also acquires K outlier singular values all equal to $\hat{s}$. The location of the data spike at $\hat{s}$ occurs at a slightly larger value than the signal spike at s. This reflects singular-value inflation in the data R relative to the signal Y, due to the addition of noise X. Furthermore, each singular vector of the data R corresponding to an outlier singular value acquires a nontrivial $O(1)$ overlap with the K-dimensional signal subspace of Y even in the asymptotic limit $N_1, N_2 \to \infty$.

The location of the outlier data singular values and their corresponding singular-vector overlaps with the signal subspace have been calculated for *finite K* and general rotationally invariant noise matrices X [22]. In the special case where the elements of X are i.i.d. Gaussian, explicit formulas can be derived (see Appendix A for a review). This signal-detectability phase transition in the finite-rank spiked model is depicted in Fig. 1 for an i.i.d. Gaussian noise matrix X.

Notably, according to the finite-rank theory, the K spikes do not interact. More precisely, above the critical signal strength, in the large-size limit, the K identical singular values of Y are all predicted to map to K identical outlier singular values of the data matrix R. Furthermore, the overlaps of the K corresponding data singular vectors with the signal subspace are predicted to be identical and completely independent of the finite value of K (see [45], however, for finite-size fluctuations in the square-symmetric spiked covariance model). More generally, if the signal Y consists of K *different* rank 1 spikes each with a unique signal strength s_l for $l = 1, \ldots, K$, the corresponding location of the data spike $\hat{s}_l$ can be computed by inserting each s_l into a single local singular value inflation function $\hat{s}(s)$ (depicted in Fig. 1), without considering the location of any other signal spike $s_{l'}$ for $l' \neq l$. In this precise sense, at finite K the spikes do not interact; one need not consider the position of any other signal spikes to compute how any one signal spike is inflated to a data spike. The same noninteracting picture is true for singular vector overlaps (Fig. 1(b)).

This lack of interaction between different spikes in the signal as they are corrupted to generate data spikes allows optimal denoising operations based on the finite-rank theory to be remarkably simple. For example, estimators for both data denoising [23,28,29], which corresponds to trying to directly estimate the signal Y given the corrupted data R, and covariance estimation [46], which corresponds to estimating the true covariance matrix $C = YY^T$ from the data R, both involve applying a *single* shrinkage function, which nonlinearly modifies each data singular value of R in a manner that acts *independently* of any other singular value. This shrinkage function, applied to each data singular value $\hat{s}$, in a sense optimally undoes the singular-value inflation $s \to \hat{s}$ and compensates for the singular-vector rotation $\boldsymbol{u}_a \to \hat{\boldsymbol{u}}_a$ which arises in going from signal Y to data $R = Y + X$. Moreover, the reason the shrinkage can act independently on each data singular value is directly related to the property of the finite-rank theory that each signal singular value is inflated *independently* through the same inflation function, while each signal singular vector is rotated *independently* through the same random rotation.
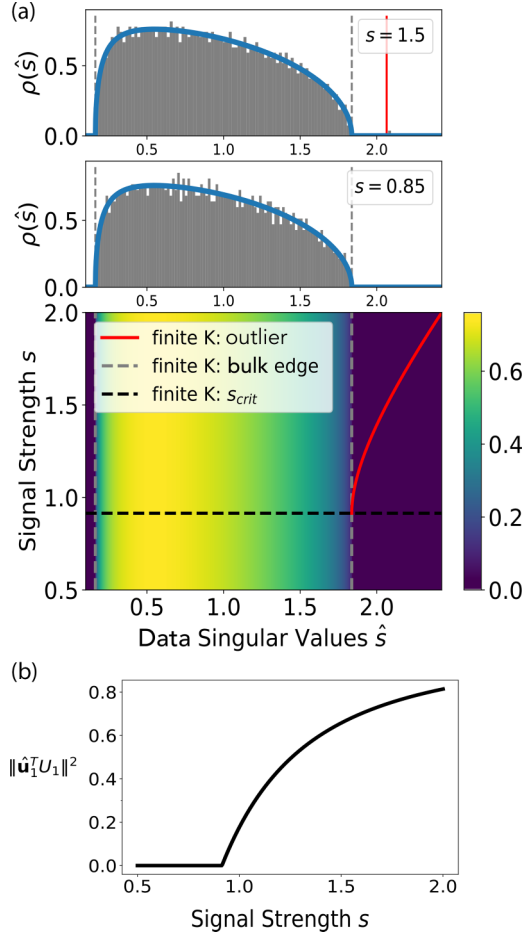
FIG. 1. Background: Signal-detectability phase transition in the finite-rank spiked matrix model. (a) The bottom panel (color online) shows the singular value spectrum of the spiked matrix model given by the finite-rank theory in the asymptotic limit with aspect ratio, $c = \frac{N_1}{N_2} = 0.7$ (see Appendix A for formulas). The singular value of the data matrix, R, is in the x axis, and the strength of the single nonzero singular value of the signal matrix, Y, is in the y axis. The "bulk" spectrum of the data is identical to the spectrum of the noise matrix, X. The bounds of that spectrum are the vertical dashed gray lines. Above the critical signal, $s_{crit} = c^{1/4}$ (black horizontal), the data have an outlier singular value shown as a solid red curve. The top two panels show theory curves corresponding to two horizontal slices, with $s = 0.85, 1.5$, together with a histogram of singular values each of a single instantiation with $N_2 = 2000$. The top panel has a single outlier eigenvalue very close to the theory prediction. The panel below shows a data spectrum that is indistinguishable from noise. (b) The overlap of the top left singular vector of the data with the left singular subspace of the signal, given by the finite-rank theory. The overlap becomes nonzero at exactly the same critical signal, s_{crit}, at which an outlier singular value appears in the data. X is Gaussian i.i.d. with variance $1/N_2$ throughout.

In this work, however, we find that the assumptions and resulting consequences of the finite-rank theory may constitute a significant limitation for the practical application of this model both to explain the properties of noise corrupted data as well as to optimally denoise such data. To illustrate, we test the finite-rank theory for various values of K, with N_1

and N_2 fixed. In Fig. 2 we show simulation results in which we find substantial deviations between simulations and finite-rank theory predictions, for both the location of the leading data singular value outlier and the data-signal singular-vector overlap, for K as small as 10 with $N_1 = 1000$. Thus, even for moderate numbers of spikes and relatively large matrices, the finite-rank theory cannot explain the SVD of the data well (though as mentioned above, see [45], for finite-size fluctuations in the square-symmetric case). As a consequence, as we will show below, typical denoising techniques, which depend crucially on the predicted singular structure of the data, perform poorly, even for moderate K.

Thus, motivated by the search for better denoisers of higher-rank data, we extend the finite-rank theory to a completely different asymptotic limit of extensive rank, in which the rank K of the data is proportional to the number of variables N_1 as both become large. We show that our extensive-rank theory both (1) more accurately explains the SVD of large data matrices of even moderate rank and (2) provides better denoisers in these cases than the finite-rank theory. And, interestingly, our extensive-rank theory reveals qualitatively unique phenomena that do not occur at finite rank, including highly nontrivial interactions between the extensive number of signal singular values, as they become corrupted to generate data singular values, under additive noise.

III. MATHEMATICAL PRELIMINARIES

We review some basic concepts from random matrix theory and introduce notation. Let M be an $N \times N$ Hermitian matrix M. We denote by $G_M(z)$ the matrix resolvent of M:

$$G_M(z) := (zI - M)^{-1}. \tag{4}$$

we define the normalized trace operator τ as

$$\tau[M] := \frac{1}{N}\text{Tr}[M]. \tag{5}$$

The Stieltjes transform $g_M(z)$ is the normalized trace of $G_M(z)$:

$$g_M(z) := \tau[(zI - M)^{-1}]. \tag{6}$$

In this work we will be interested in the singular values and vectors of rectangular matrices. In order to apply Hermitian matrix methods to a rectangular matrix $R \in \mathbb{R}^{N_1 \times N_2}$, we will work with its Hermitianization, which we denote with boldface throughout,

$$\boldsymbol{R} := \begin{bmatrix} 0 & R \\ R^T & 0 \end{bmatrix}, \tag{7}$$

which is an $N \times N$ Hermitian matrix, with $N = N_1 + N_2$. The eigenvalues and eigenvectors of $\boldsymbol{R}$ can be written $\pm s, \frac{1}{\sqrt{2}}\begin{pmatrix} \boldsymbol{u_1} \\ \pm \boldsymbol{u_2} \end{pmatrix}$, where s is a singular value of R, and $\boldsymbol{u_1}, \boldsymbol{u_2}$ are the corresponding left and right singular vectors. This will allow us to extract information about the singular value decomposition of a rectangular matrix R from the eigendecomposition of the Hermitian matrix $\boldsymbol{R}$.

Hermitianization leads naturally to a Hermitian *block resolvent*, which is a function of two complex scalars z_1 and z_2,
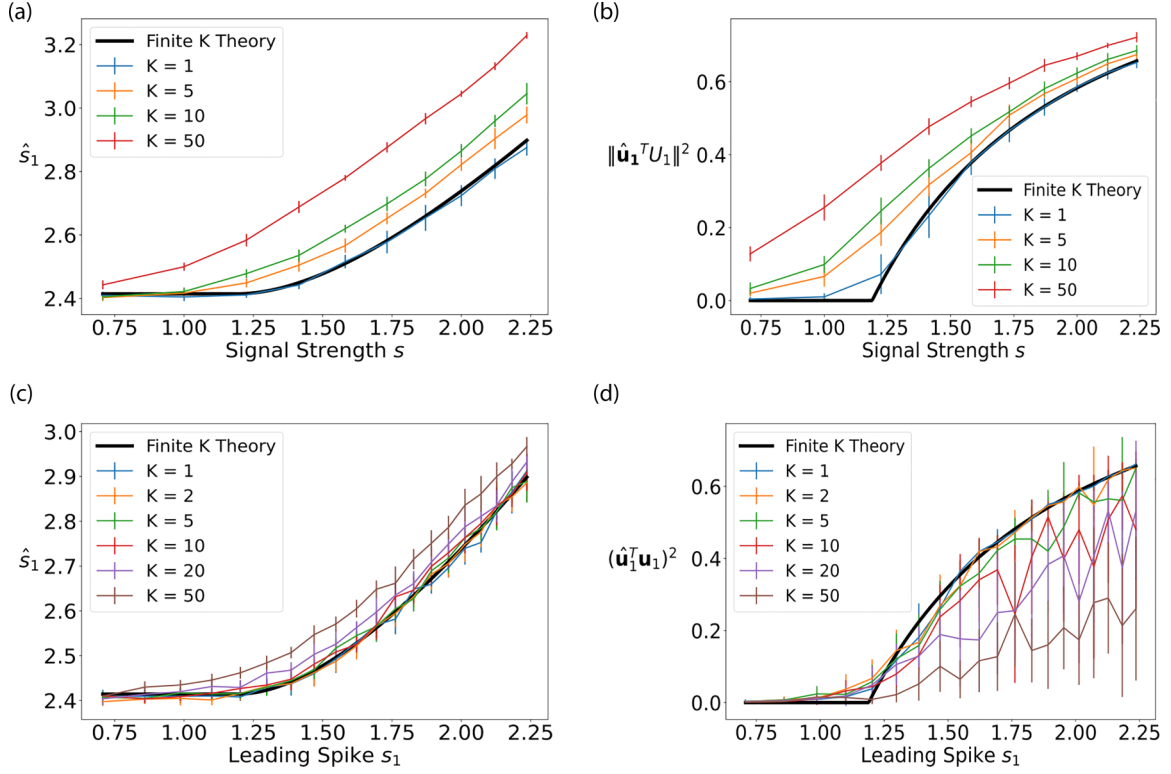
FIG. 2. Finite-rank theory fails to capture the singular value structure of the spiked rectangular matrix model. (a) Singular-value inflation, i.e., leading data singular value, $\hat{s}_1$, as a function of signal singular value, s_1, for spikes of various ranks K. Black shows the finite-rank theory (which is independent of the rank of a spike). Matrix size in this and all subsequent panels is $N_1 = 1000$, $N_2 = 500$. Numerical results are presented as mean and standard deviation over 10 instantiations for each value of b and s. (b) Singular-vector rotation, i.e., overlap of leading data left singular vector, $\hat{\mathbf{u}}_1$, with the K-dimensional left singular space of the signal, U_1. (c) To break the degeneracy of the spikes with rank $K > 1$ in (a), we consider a single leading signal spike with singular value s along with $K - 1$ spikes drawn independently and uniformly in $[0, s]$. We plot the leading data singular value as a function of s for various K, compared to finite-rank theory (black) (d). For the same signal model as in (c), we plot the overlap of the leading data left singular vector with the leading signal singular vector as function of signal strength s, for different K and for the finite-rank theory (black). We see that the finite-rank theory incorrectly estimates both the singular values and singular vectors of signals of even moderate rank K. See Appendix A for finite-rank theory formulas.

rather than one:

$$G^R(z) := \begin{bmatrix} z_1 I_{N_1} & -R \\ -R^T & z_2 I_{N_2} \end{bmatrix}^{-1}, \quad (8)$$

where $z = (z_1, z_2)$ is a complex vector. This block resolvent can be computed explicitly, with each block written in terms of a standard square-matrix resolvent:

$$G^R(z) = \begin{bmatrix} z_2 G_{RR^T}(z_1 z_2) & R G_{R^T R}(z_1 z_2) \\ R^T G_{RR^T}(z_1 z_2) & z_1 G_{R^T R}(z_1 z_2) \end{bmatrix}. \quad (9)$$

Analogously, we define the *block* Stieltjes transform $g^R(z)$ as the two-element complex vector consisting of the normalized traces of each diagonal block of G^R:

$$g_1^R(z) = \tau_1\left[G_{11}^R(z)\right] = z_2 g_{RR^T}(z_1 z_2), \quad (10a)$$

$$g_2^R(z) = \tau_2\left[G_{22}^R(z)\right] = z_1 g_{R^T R}(z_1 z_2). \quad (10b)$$

Here we have introduced notation for the blockwise normalized traces:

$$\tau_a(M) := \frac{1}{N_a} \text{Tr}[M_{aa}], \quad (11)$$

where M_{aa} is the ath diagonal block of size $N_a \times N_a$.

Notationally, we write the block vectors and block matrices g^R and G^R in bold, while we indicate the component blocks in standard roman font, with the indices a, b for both scalar, g_a^R, and matrix G_{ab}^R blocks, with $a, b \in \{1, 2\}$. We will also use the fact that the eigenvalues of RR^T and $R^T R$ differ by exactly $|N_1 - N_2|$ zeros, implying the two elements of g^R are related by $g_2^R(z) = \frac{z_1}{z_2} c g_1^R(z) + \frac{1-c}{z_2}$.

Each element $g_a^R(z)$ can be written in terms of the corresponding singular value density:

$$g_1^R(z_1, z_2) = \int_{-\infty}^{+\infty} \frac{z_2}{z_1 z_2 - s^2} \rho_1^R(s) \, ds, \quad (12a)$$

$$g_2^R(z_1, z_2) = \int_{-\infty}^{+\infty} \frac{z_1}{z_1 z_2 - s^2} \rho_2^R(s) \, ds, \quad (12b)$$

where $\rho_a^R(s)$ denotes the singular value distribution of R, accounting for N_a singular values. Note that for nonzero s with finite singular value density, $\rho_2^R(s) = c \rho_1^R(s)$.

The special case in which the two arguments are equal, $z_1 = z_2 = z$, will be important, and so we abbreviate: $g^R(z) := g^R(z, z)$.

We can write an inversion relation for the singular value densities using the Sokhotski-Plemelj theorem, which states,

$\lim_{\eta \to 0^+} \text{Im}[\int \frac{f(x)}{x - i\eta} dx] = \pi f(0)$. Applying this theorem to $f(x) = \frac{z}{z+x} \rho_a^R(x)$ yields

$$\rho_a^R(s) = \frac{2}{\pi} \lim_{\eta \to 0} \text{Im}[g_a^R(s - i\eta)]. \quad (13)$$

Finally, we define the *block* $\mathcal{R}$ transform:

$$\boldsymbol{\mathcal{R}}^R(\boldsymbol{t}) = (\boldsymbol{g}^R)^{-1}(\boldsymbol{t}) - \frac{1}{\boldsymbol{t}}, \quad (14)$$

where $\boldsymbol{t} \in \mathbb{C}^2$ is in the range of $\boldsymbol{g}^R$; we denote by $(\boldsymbol{g}^R)^{-1}$ the functional inverse of the block Stieltjes transform $\boldsymbol{g}^R$, satisfying $(\boldsymbol{g}^R)^{-1}[\boldsymbol{g}^R(z)] = z$; and $1/\boldsymbol{t}$ is the componentwise multiplicative inverse of $\boldsymbol{t}$.

The block $\mathcal{R}$ transform will arise naturally in our calculation of the subordination relation for the sum of free rectangular matrices, $R = Y + X$, and as we shall verify, it is additive for independent, rotationally invariant matrices:

$$\boldsymbol{\mathcal{R}}^R(\boldsymbol{t}) = \boldsymbol{\mathcal{R}}^Y(\boldsymbol{t}) + \boldsymbol{\mathcal{R}}^X(\boldsymbol{t}). \quad (15)$$

IV. THE SINGULAR VALUE DECOMPOSITION OF SUMS OF RECTANGULAR MATRICES

In this section we characterize how an additive noise matrix X deforms the singular values and vectors of a signal matrix Y to generate singular values and vectors of the data matrix $R = Y + X$ [see (1) and following text]. We consider general signal matrices of the form

$$Y = U_1 S U_2^T, \quad (16)$$

where each U_a is an $N_a \times N_a$ orthonormal matrix ($a = 1, 2$), and S is $N_1 \times N_2$ diagonal matrix.

We begin by deriving an asymptotically exact subordination formula relating the block resolvents (9) of R and Y in the limit $N_1, N_2 \to \infty$ with the aspect ratio $c = N_1/N_2$ fixed. From this, we extract both the singular value spectrum of R, as well as the overlaps between the singular vectors of R and those of the signal matrix, Y.

A. A subordination relation for the sum of rectangular matrices

Exploiting the rotational invariance of $P_X(X)$, we first calculate the block resolvent of R as an expectation over arbitrary rotations of the noise X. Thus, we write $R = Y + O_1 X O_2^T$, where O_a are Haar-distributed orthogonal $N_a \times N_a$ matrices. We can write the Hermitianization (7) of $\boldsymbol{R}$ in terms of the Hermitianized X and Y:

$$\boldsymbol{R} = \boldsymbol{Y} + \bar{\boldsymbol{O}} X \bar{\boldsymbol{O}}^T, \quad (17)$$

where we have written $\bar{\boldsymbol{O}} = \begin{bmatrix} O_1 & 0 \\ 0 & O_2 \end{bmatrix}$.

The main result of this section is the following subordination relation for the expectation of the block resolvent $\boldsymbol{G}^R$, taken over the random block-orthogonal matrix $\bar{\boldsymbol{O}}$:

$$\mathbb{E}_{\bar{\boldsymbol{O}}}[\boldsymbol{G}^R(z)] = \boldsymbol{G}^Y(z - \boldsymbol{\mathcal{R}}^X(\boldsymbol{g}^R(z))). \quad (18)$$

As mentioned above, this notation refers to the special case in which the argument to $\boldsymbol{G}^R$ is the two-dimensional complex vector with equal arguments, $z_1 = z_2 = z$. Note that the argument to $\boldsymbol{G}^Y$, by a slight abuse of notation, is the

vector $z - \mathcal{R}_a^X(\boldsymbol{g}^R(z))$ for $a = 1, 2$. In Appendix B we present the detailed derivation for this case, which is sufficient for computing the singular values and associated singular-vector overlaps. We provide a sketch of the calculation here. The general case follows.

We first write the analog of a partition function,

$$\mathcal{Z}^R(\boldsymbol{Y}) = \det(z I - \boldsymbol{R})^{-1/2}, \quad (19)$$

and observe that we can write the desired matrix inverse as a derivative of the corresponding free energy:

$$\boldsymbol{G}^R(z) = 2 \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{Y}} \log \mathcal{Z}^R(\boldsymbol{Y}). \quad (20)$$

We would like to average this over the block-orthogonal matrix $\bar{\boldsymbol{O}}$, yielding a "quenched" average free energy. In Appendix B 1 we show that in the large N limit, the quenched and annealed averages are equivalent. In short, viewing $\log \mathcal{Z}^R$ as a function of $\bar{\boldsymbol{O}}$, we find it has a Lipschitz constant proportional to $1/\sqrt{N}$, and then use the concentration of measure of the orthogonal group, $\mathbb{SO}(N)$, with additional concentration inequalities to show that

$$\lim_{N \to \infty} \frac{1}{N} \mathbb{E}_{\bar{\boldsymbol{O}}}[\log \mathcal{Z}^R(\boldsymbol{Y})] = \lim_{N \to \infty} \frac{1}{N} \log \mathbb{E}_{\bar{\boldsymbol{O}}}[\mathcal{Z}^R(\boldsymbol{Y})]. \quad (21)$$

We can therefore calculate our desired block resolvent as

$$\mathbb{E}_{\bar{\boldsymbol{O}}}[\boldsymbol{G}^R(z)] = 2 \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{Y}} \log \mathbb{E}_{\bar{\boldsymbol{O}}}[\mathcal{Z}^R(\boldsymbol{Y})]. \quad (22)$$

We proceed by writing the determinant as a Gaussian integral,

$$\mathcal{Z}^R(\boldsymbol{Y}) = \int \frac{d\boldsymbol{v}}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2} \boldsymbol{v}^T (z I - \boldsymbol{R}) \boldsymbol{v}\right], \quad (23)$$

and then we substitute $\boldsymbol{R} = \boldsymbol{Y} + \bar{\boldsymbol{O}} X \bar{\boldsymbol{O}}^T$, extract terms that do not depend on $\bar{\boldsymbol{O}}$, and take the expectation of the terms that do, which yields an intermediate integral,

$$I^X(\boldsymbol{v}) \equiv \mathbb{E}_{\bar{\boldsymbol{O}}}[e^{\frac{1}{2} \boldsymbol{v}^T \bar{\boldsymbol{O}} X \bar{\boldsymbol{O}}^T \boldsymbol{v}}]. \quad (24)$$

This integral is analogous to the Harish-Chandra-Itzykson-Zuber (HCIZ) or spherical integral, which appears in the calculation of the subordination relation for sums of square-symmetric matrices [33,34,36]. We compute this "block-spherical" integral asymptotically in Appendix C and highlight key points of the calculation here.

First, we observe the key difference between our calculation and the square-symmetric case. In the square-symmetric case, the expectation is over a single Haar-distributed orthogonal matrix that rotates $\boldsymbol{v}$ arbitrarily, and so the expectation depends only on the norm of $\boldsymbol{v}$. In our rectangular case, however, $\bar{\boldsymbol{O}}$ has two blocks, and they rotate the N_1- and N_2-dimensional blocks of $\boldsymbol{v}$ separately, so that $I^X(\boldsymbol{v})$ depends on the norms of each of these two blocks. Therefore, we define the two-component vector, $\boldsymbol{t}$ with components

$$t_a := \frac{1}{N_a} \|\boldsymbol{v}_a\|^2. \quad (25)$$

We calculate the expectation (24) by performing an integral over an arbitrary N-dimensional vector, while enforcing blockwise norm constraints using the Fourier representation of the delta function, and introducing integration variables, q_1 and q_2.

To compute the integral, we make a saddle-point approximation in the asymptotic limit of large N. Appealingly, we find the saddle-point conditions are of the form

$$t_1 = q_2^* g_{XX^T}(q_1^* q_2^*), \tag{26a}$$

$$t_2 = q_1^* g_{X^T X}(q_1^* q_2^*). \tag{26b}$$

That is, the block Stieltjes transform, $\boldsymbol{g}^X(\boldsymbol{q}^*) = \boldsymbol{t}$, arises naturally, and the saddle point of the block-spherical integral (24) is its functional inverse evaluated at the vector of blockwise norms of $\boldsymbol{v}$.

Inserting the saddle-point solution, we find that asymptotically

$$I^X(\boldsymbol{v}) = \exp\left[\frac{N}{2} H^X(\boldsymbol{t})\right], \tag{27}$$

where, for a neighborhood of values of $\boldsymbol{t}$ around 0, the saddle-point free energy itself, $H^X(\boldsymbol{t})$, has gradient with elements proportional to the block $\mathcal{R}$ transform (14):

$$\frac{dH^X(\boldsymbol{t})}{dt_a} = \frac{N_a}{N} \mathcal{R}_a^X(\boldsymbol{t}). \tag{28}$$

Thus, the block $\mathcal{R}$ transform arises via the antiderivative of the logarithm of the block-spherical integral, analogously to the regular $\mathcal{R}$ transform in the case of square-symmetric matrices.

Note that given the definition in 24, it is straightforward to see that $I^R(\boldsymbol{v}) = I^Y(\boldsymbol{v}) I^X(\boldsymbol{v})$, and thus $H^R(\boldsymbol{t}) = H^Y(\boldsymbol{t}) + H^X(\boldsymbol{t})$. Therefore, we have established the additivity of the block $\mathcal{R}$ transform as well.

Continuing with the derivation of the subordination relation (18), we next substitute the result for $I^X(\boldsymbol{v})$ back into the Gaussian integral over $\boldsymbol{v}$ (23), and then we introduce another pair of integration variables, $\hat{\boldsymbol{t}}$, in order to decouple $\boldsymbol{v}$ from its blockwise norms, $\boldsymbol{t}$. Performing the Gaussian integral we find

$$\mathbb{E}_{\bar{\partial}}[\mathcal{Z}^R(Y)] \propto \int dt \, d\hat{t} \exp\left(\frac{N}{2} P^{X,Y}(\boldsymbol{t}, \hat{\boldsymbol{t}})\right), \tag{29}$$

with

$$P^{X,Y}(\boldsymbol{t}, \hat{\boldsymbol{t}}) := -\frac{1}{N}(N_1 t_1 \hat{t}_1 + N_2 t_2 \hat{t}_2) + H^X(\boldsymbol{t})$$
$$-\frac{1}{N} \log \det \boldsymbol{G}^Y(z - \hat{\boldsymbol{t}}). \tag{30}$$

Note that the block resolvent of Y arises here naturally as a function of the two-element vector, $z - \hat{\boldsymbol{t}}$, despite the fact that we set out to find $\boldsymbol{G}^R$ evaluated at the point (z, z).

The integrals over $\boldsymbol{t}$ and $\hat{\boldsymbol{t}}$ yield an additional pair of saddle-point conditions. The first requires $\hat{\boldsymbol{t}}^* = \mathcal{R}^X(\boldsymbol{t}^*)$ and combining with the second gives

$$\boldsymbol{t}^* = \boldsymbol{g}^Y(z - \mathcal{R}^X(\boldsymbol{t}^*)). \tag{31}$$

We have thus found the desired annealed free energy, $2 \log \mathbb{E}_{\bar{\partial}}[\mathcal{Z}^R] = N P^{X,Y}(\boldsymbol{t}^*, \mathcal{R}^X(\boldsymbol{t}^*))$ [see (30)].

We next take the derivative with respect to Y (see Appendix B for a more careful treatment), which gives

$$\mathbb{E}_{\bar{\partial}}[\boldsymbol{G}^R(z)] = \boldsymbol{G}^Y(z - \mathcal{R}^X(\boldsymbol{t}^*)). \tag{32}$$

Finally to find $\boldsymbol{t}^*$, we take the blockwise normalized traces to find $\boldsymbol{g}^R(z) = \boldsymbol{g}^Y(z - \mathcal{R}^X(\boldsymbol{t}^*)) = \boldsymbol{t}^*$, and that completes the derivation of the block resolvent subordination relation (18).

We note that the saddle-point condition (31) turns out to be the subordination relation for the block Stieltjes transform:

$$\boldsymbol{g}^R(z) = \boldsymbol{g}^Y\left(z - \mathcal{R}^X(\boldsymbol{g}^R(z))\right). \tag{33}$$

Note that while $\boldsymbol{g}^R$ is evaluated at the scalar point (z, z), the argument to $\boldsymbol{g}^Y$ is the vector subordination function $\boldsymbol{\zeta} \equiv z - \mathcal{R}^X(\boldsymbol{g}^R(z))$ whose components are distinct in general.

The singular value spectrum of the sum of rectangular matrices can thus be obtained by first finding the block Stieltjes transform, either by employing the additivity of the block $\mathcal{R}$ transform or by solving the subordination relation (33), and then using the inversion relation (13).

B. Deformation of singular vectors due to additive noise

Turning now to the singular vectors of the data matrix $R = \hat{U}_1 \hat{S} \hat{U}_2^T = Y + X$, we quantify the effect of the noise, X, on the signal, $Y = U_1 S U_2^T$, via the matrix of squared overlaps between the clean singular vectors of the signal, U_a with $a = 1, 2$ for left and right, respectively, and the noise-corrupted singular vectors of the data, $\hat{U}_a$, written as $(\hat{U}_a^T U_a)^2$.

In the noiseless case $X = 0$, one has $(\hat{U}_a^T U_a)^2 = I_{N_a}$, signifying perfect correspondence between signal and data singular vectors. In the presence of substantial noise, the overlaps of a signal singular vector are generically distributed over order N_a data singular vectors and are of order $1/N_a$; therefore we define the rescaled expected square overlap between a given singular vector, $\hat{\boldsymbol{u}}_a$ of R with corresponding singular value $\hat{s}$, and a given singular vector, $\boldsymbol{u}_a$ of Y, with corresponding singular value s, where once again $a = 1, 2$ for left and right singular vectors, respectively:

$$\Phi_a(\hat{s}, s) = N_a \mathbb{E}\left[\left(\hat{\boldsymbol{u}}_a^T \boldsymbol{u}_a\right)^2\right]. \tag{34}$$

To see how to obtain the expected square overlaps from the block resolvent, $\boldsymbol{G}^R(z)$, we write each of the diagonal blocks, $G_{aa}(z)^R$ (9), in terms of their eigendecomposition, and multiply on both sides by a "target" singular vector of Y, e.g., $\boldsymbol{u}_a$ with associated singular value s:

$$\boldsymbol{u}_a^T G_{aa}^R(z) \boldsymbol{u}_a = \sum_{k=1}^{N_a} \frac{z}{z^2 - \hat{s}_k^2} \left(\boldsymbol{u}_a^T \hat{\boldsymbol{u}}_{ak}\right)^2. \tag{35}$$

If we choose $z = \hat{s} - i\eta$ where $\rho_a^R(\hat{s}) \sim O(1)$, with $N_a^{-1} \ll \eta \ll 1$, and take the imaginary part, then we get a weighted average of the square overlaps of a macroscopic number of singular vectors of R, $\hat{\boldsymbol{u}}_{ak}$, that have singular values close to $\hat{s}$, with the target singular vector $\boldsymbol{u}_a$, each weighted by $\pi/2\rho_a^R(\hat{s}_k)$. If we first take the limit of large N_a and then take $\eta \to 0$ we obtain the expectation

$$\lim_{\eta \to 0} \frac{2}{\pi} \text{Im}\left[\boldsymbol{u}_a^T G_{aa}^R(\hat{s} - i\eta) \boldsymbol{u}_a\right] \to \rho_a^R(\hat{s}) \Phi_a(\hat{s}, s). \tag{36}$$

Now we use the subordination relation (18) to replace the resolvent of R with the resolvent of Y: $G_{aa}^R(z) = G_{aa}^Y(\boldsymbol{\zeta}(z))$ where we have written the two-component vector

$$\boldsymbol{\zeta}(z) = z - \mathcal{R}^X(\boldsymbol{g}^R(z)). \tag{37}$$

Since u_a is an eigenvector of $G_{aa}^Y(\zeta_1, \zeta_2)$ with eigenvalue $\frac{\zeta_b}{\zeta_1\zeta_2 - s^2}$ where $b = 2$ for $a = 1$ and $b = 1$ for $a = 2$, we find

$$\Phi_1(\hat{s}, s) = \frac{2}{\pi \rho_1^R(\hat{s})} \lim_{\eta \to 0} \operatorname{Im} \frac{\zeta_2(\hat{s} - i\eta)}{\zeta_1(\hat{s} - i\eta)\zeta_2(\hat{s} - i\eta) - i\eta - s^2},$$
(38a)

$$\Phi_2(\hat{s}, s) = \frac{2}{\pi \rho_2^R(\hat{s})} \lim_{\eta \to 0} \operatorname{Im} \frac{\zeta_1(\hat{s} - i\eta)}{\zeta_1(\hat{s} - i\eta)\zeta_2(\hat{s} - i\eta) - i\eta - s^2}.$$
(38b)

These expressions can be written in terms of the real and imaginary parts of the block $\mathcal{R}$ transform of the noise X. In the following section we provide simplified expressions for the important case of Gaussian noise.

1. Arbitrary signal with Gaussian noise

We show in Appendix D that the block $\mathcal{R}$ transform of an $N_1 \times N_2$ (with $c = \frac{N_1}{N_2}$) Gaussian matrix with i.i.d. entries of variance σ^2/N_2 is

$$\boldsymbol{R}^X(\boldsymbol{t}) = \sigma^2 \begin{pmatrix} t_2 \\ ct_1 \end{pmatrix}.$$
(39)

Note that from the definition of the $\mathcal{R}$ transform, one can find that $\mathcal{R}_2^A(\boldsymbol{t})t_2 = c\mathcal{R}_1^A(\boldsymbol{t})t_1$, for any rectangular A with aspect ratio c, and (39) is the only pair of linear functions of $\boldsymbol{t}$ that satisfies this constraint.

We substitute (39) into the block Stieltjes transform subordination relation yielding $\boldsymbol{g}^R(z) = \boldsymbol{g}^Y(\boldsymbol{\zeta})$ with $\zeta_1 = z - \sigma^2 g_2^R(z)$ and $\zeta_2 = z - c\sigma^2 g_1^R(z)$, and then use the identity $g_2^R(z) = cg_1^R(z) + \frac{1-c}{z}$ (for arbitrary rectangular R, the spectra of RR^T and R^TR differ only by a set of 0 eigenvalues). Then using the definition $g_1^Y(\boldsymbol{\zeta}) = \zeta_2 g_{YY^T}(\zeta_1\zeta_2)$, we arrive at

$$g_1^R(z) = \zeta_2(z)g_{YY^T}\left(\zeta_2(z)\left(\zeta_2(z) - \sigma^2\frac{1-c}{z}\right)\right),$$
(40)

with

$$\zeta_2(z) := z - c\sigma^2 g_1^R(z).$$
(41)

This is a self-consistency equation for the block Stieltjes transform of R, $\boldsymbol{g}^R(z)$, that depends on the noise variance σ^2, the aspect ratio c, and the standard Stieltjes transform of the signal covariance, $g_{YY^T}(z)$.

Once this equation is solved, the singular vector overlaps can be obtained as well. We introduce notation for the real and imaginary parts of the block Stieltjes transform, $g_1^R(\hat{s}) = h_1^R + if_1^R$, where we assume that the spectral density at $\hat{s}$ is finite. Then we insert this into (39) to get the real and imaginary parts of the block $\mathcal{R}$ transform of X. After defining, for notational ease,

$$\nu(z) := \operatorname{Re}\zeta_2(z) = z - c\sigma^2 h_1^R(z),$$
(42)

we can finally simplify the overlaps (38) for the case of Gaussian noise:

$$\Phi_1(\hat{s}, s) = \frac{\nu(\hat{s})\mathcal{B}(\hat{s}) - c\sigma^2\mathcal{A}(\hat{s}, s)}{[\mathcal{A}(\hat{s}, s)]^2 + [f_1^R\mathcal{B}(\hat{s})]^2}$$
(43a)

$$\Phi_2(\hat{s}, s) = \frac{[\nu(\hat{s}) - \sigma^2\frac{1-c}{\hat{s}}]\mathcal{B}(\hat{s}) - c\sigma^2\mathcal{A}(\hat{s}, s)}{[\mathcal{A}(\hat{s}, s)]^2 + [f_1^R\mathcal{B}(\hat{s})]^2},$$
(43b)

where we have

$$\mathcal{A}(\hat{s}, s) = \nu(\hat{s})\left[\nu(\hat{s}) - \sigma^2\frac{1-c}{\hat{s}}\right] - \left[s^2 + c^2\sigma^4\left(f_1^R\right)^2\right],$$
(44a)

$$\mathcal{B}(\hat{s}) = 2c\sigma^2\left[\nu(\hat{s}) - \sigma^2\frac{1-c}{2\hat{s}}\right].$$
(44b)

Formula (43a) is confirmed in Fig. 3, which shows the left singular vector overlaps between data and signal, when the signal, Y, is Gaussian as well. The bottom color plot can be thought of as an input-output map for the singular value structure under additive noise. It shows that a signal singular vector associated with a given singular value s undergoes, loosely speaking, both "diffusion" and "inflation," aligning partially with data singular vectors across a range of singular values with a peak associated with larger singular values, $\hat{s} > s$. In the upper three panels we observe that individual overlaps are not self-averaging; a smooth overlap function emerges only when one averages either over many overlaps within a range of singular values, or over many instantiations.

We stress that these formulas for the overlap of data singular vectors with signal singular vectors do not depend directly on the unobserved signal Y. Rather, they depend only on the noise variance and the block Stieltjes transform, $g_1^R(z)$, of the noisy data matrix, R. Furthermore, $g_1^R(z)$ can be estimated empirically via kernel methods for the empirical spectral density and its Hilbert transform [32,35,36]. This suggests that significant information about the structure of the unobserved extensive signal can be inferred from noisy empirical data, and this will lay the foundation for the optimal estimators derived below.

C. SVD of the extensive spike model

We now return to the spiked matrix model $R = Y + X$, with signal $Y = sU_1U_2^T$, where s is a scalar, and U_a are $N_a \times K$ matrices with orthogonal columns. But now we assume the rank of the spike grows linearly with the number of rows at a fixed rank ratio, b, i.e., $K = bN_1$, while the aspect ratio $c = N_1/N_2$ is fixed as before. We will assume the elements of the noise matrix X are i.i.d. Gaussian: $X_{ij} \sim \mathcal{N}(0, \frac{1}{N_2})$. In the following we first discuss the singular values and then the singular vectors of the extensive-rank model.

1. Singular value spectrum of the extensive spike model

YY^T has K eigenvalues equal to s^2 and $N_1 - K$ zero eigenvalues. Its Stieltjes transform is therefore found to be

$$g_{YY^T}(z) = \frac{z + (b-1)s^2}{z(z - s^2)}.$$
(45)

We can now make use of the self-consistency equation for $g_1^R(z)$ (40). Momentarily writing g in place of $g_1^R(z)$ and simplifying, we find

$$\left[\left(\zeta_2 - \frac{1-c}{z}\right)g - 1\right]\left[\left(\zeta_2 - \frac{1-c}{z}\right)\zeta_2 - s^2\right] = bs^2,$$
(46)

where we write $\zeta_2 = z - c\sigma^2 g$ as above. This is a quartic polynomial for $g = g_1^R(z)$. We solve this numerically for z near the real line in order to find the density of singular values of
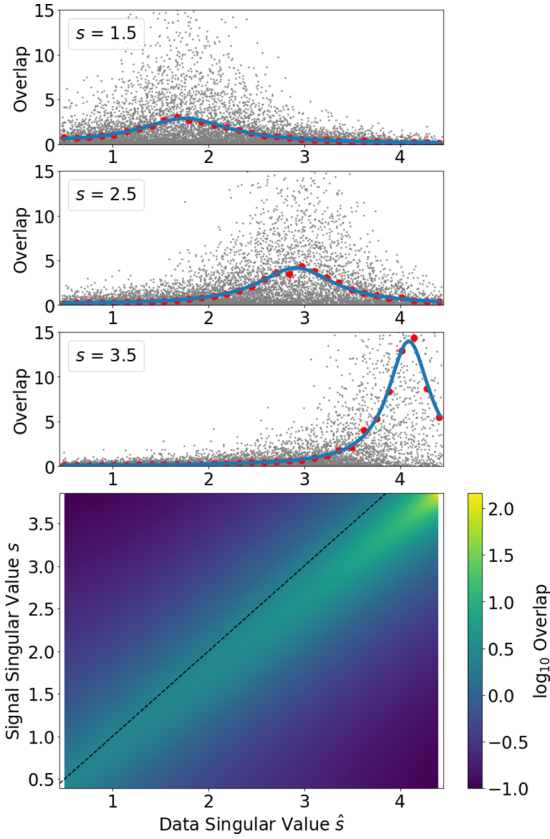
FIG. 3. Singular vector overlaps of sums of Gaussians. Bottom plot (color online) shows the theoretical prediction of the logarithm of the overlap, $\log_{10} \Phi_1(\hat{s}, s)$ (Eq. (43a)), between the left singular vectors of $R = Y + X$ and those of Y, as a bivariate function of the associated singular values $\hat{s}$ of R and s of Y. Dashed line is identity $s = \hat{s}$. The signal Y and noise X are both rectangular Gaussian matrices with aspect ratio $c = N_1/N_2 = 3/2$. Elements of Y are i.i.d. with variance σ_y^2/N_2 where $\sigma_y^2 = 3$. Elements of X are i.i.d. with variance σ_x^2/N_2 with $\sigma_x^2 = 1$. Top three panels show singular vector overlaps for three horizontal slices associated with three fixed "target" signal singular values $s = 1.5, 2.5, 3.5$, for 10 realizations of random matrices with $N_1 = 1500$ and $N_2 = 1000$. Each gray dot denotes an overlap between a left singular vector of R with singular value $\hat{s}$ (position on x axis) with the left singular vector of Y with singular value *closest* to s. Red dots reflect binning the singular values of R from all 10 realizations, with number of bins set to $\sqrt{N_2}$ giving bin width ≈ 0.13. Blue is the theoretical prediction from $\Phi_1(\hat{s}, s)$ in (43a). Note that as the signal singular value s increases, $\Phi_1(\hat{s}, s)$ as a function of $\hat{s}$ becomes more concentrated about a value *larger* than s. This reflects the fact that singular vector structure in the signal Y at singular value s is mapped to singular vector structure in the data R at larger singular values $\hat{s}$, due to singular value inflation under the addition of noise X.

R (see Appendix E for the polynomial coefficients and details of numerical solution).

For strong signal s, the spectrum in the extensive case differs from the finite rank case most clearly in that singular values reflecting the signal do not concentrate at a single data singular value. Rather (see Fig. 4 top) for sufficiently strong signal s, the presence of noise blurs the signal singular values into a continuous bulk that is disconnected from the noise
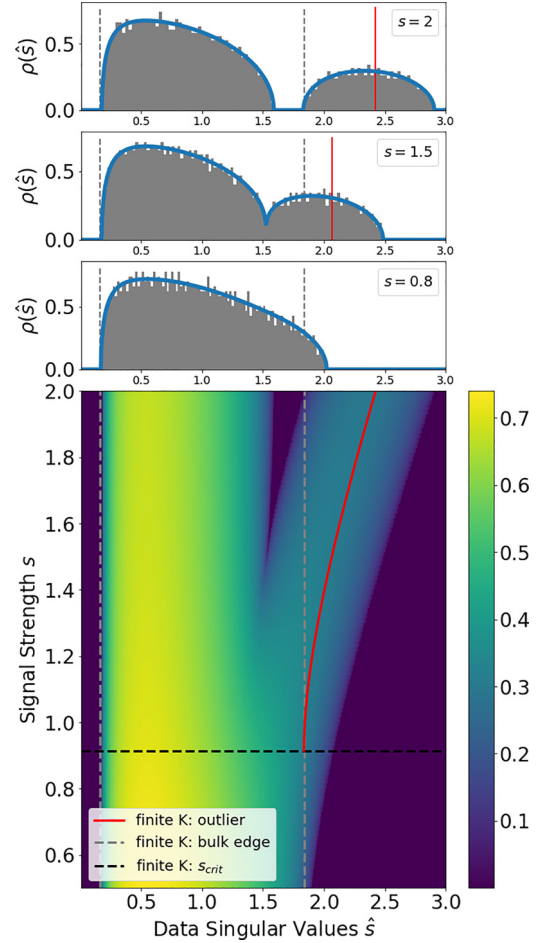

FIG. 4. Signal-strength transition in the SV density of the extensive spike model. Each row of the bottom color map shows theoretical predictions for the singular value density of the extensive spike model, $\rho_1^R(\hat{s})$, corresponding to different signal strengths s (along y axis) at a fixed rank ratio of $b = K/N_1 = 0.25$. In all panels the aspect ratio is fixed to $c = N_1/N_2 = 0.7$. Features of the finite-rank spike model are shown as lines for comparison. The horizontal black dashed line indicates the threshold signal strength s_{crit} above which the finite-rank model acquires an outlier singular value. The red curve indicates the position of this outlier singular value. The vertical gray dashed lines indicate the edges of the bulk spectrum of the finite-rank model. The top three panels, corresponding to horizontal slices of the color maps, plot the singular value density at three different signal strengths $s = 0.8, 1.5$, and 2. Solid blue curves indicate theoretical predictions from numerically solving (46), while gray histograms indicate the spectral density from a single realization with $N_2 = 2000$. For comparison, the red vertical spike indicates the position of outlier singular value in the finite-rank theory, while the gray dashed spike indicates the edge of the noise bulk in this theory. Together these panels demonstrate that as s increases, the singular value density undergoes first a crossover from a unimodal to a bimodal regime, and then a phase transition from a connected to a disconnected phase.

bulk. This signal bulk appears near the single outlier predicted by the finite-rank theory, but has significant spread.

At weak signals s there is a single, unimodal bulk spectrum, just as in the finite-rank setting, but in contrast, these weak signals make their presence felt by extending the leading edge of the bulk beyond the edge of the spectrum predicted by the
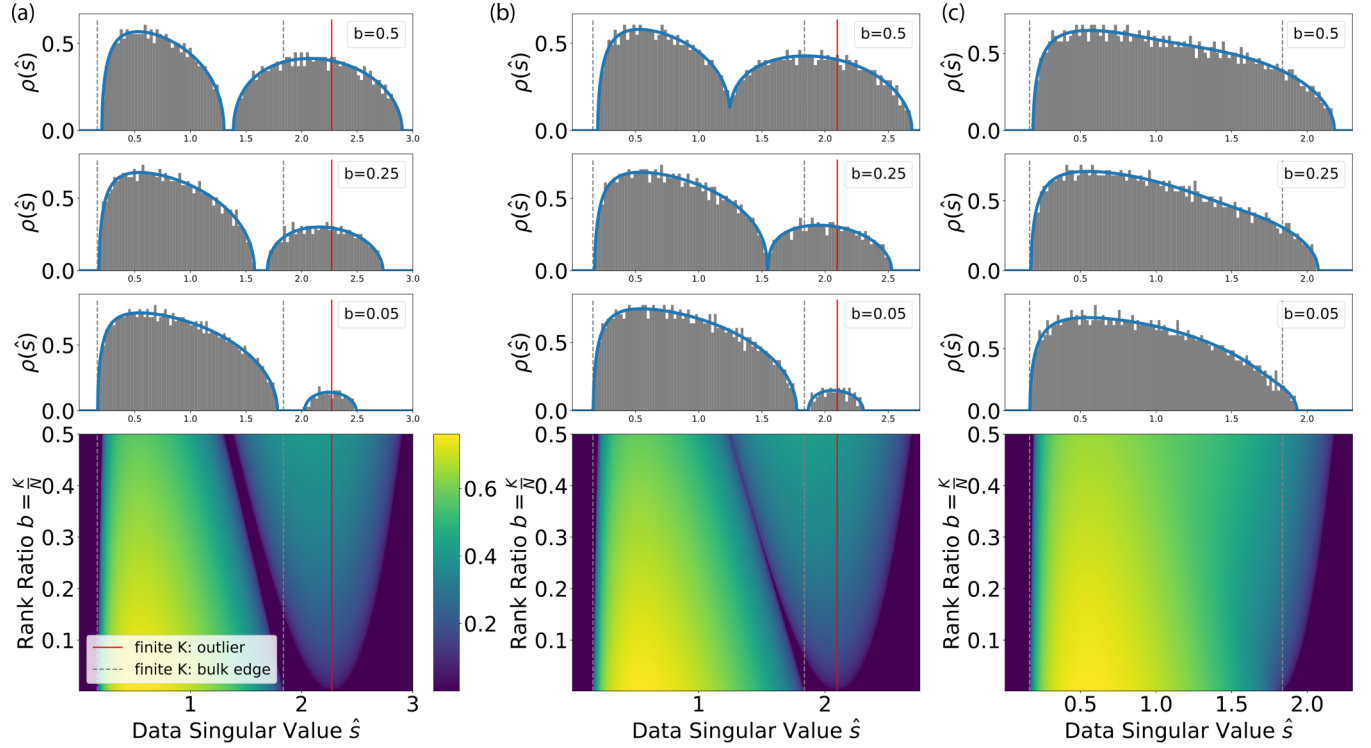
FIG. 5. Rank-ratio transitions in the SV density of the extensive spike model. Each row of the bottom color maps show theoretical predictions for the singular value density $\rho_1^R(\hat{s})$ corresponding to different rank ratios b (y axis) at a fixed signal strength, and all color maps have same color scale. Top three panels indicate matching theory (blue curves) and empirics of a single realization (gray histograms) for three rank ratios $b = 0.05, 0.25, 0.5$, and the aspect ratio is fixed to $c = N_1/N_2 = 0.7$ with $N_2 = 2000$. Comparisons to the finite-rank theory are shown using the same conventions as in Fig. 4. (a) Results for $s = 1.8$, illustrating that for sufficiently strong signal the singular value density remains in the connected phase for all values of b. (b) Results for $s = 1.55$, illustrating that for intermediate signal strengths the density undergoes a transition from disconnected to connected as the rank ratio b increases. (c) Results for $s = 0.9$, illustrating that for subthreshold signals the density remains connected for all b. $s_{crit} = c^{1/4} \approx 0.915$ throughout.

finite-rank theory, *even* when the signal strength s is below the critical signal strength s_{crit} predicted by finite-rank model (Fig. 4 third panel).

At intermediate signal strength s, the singular value distribution exhibits a connected bimodal regime not present in the finite-rank model (Fig. 4 second panel).

Thus, as s increases, we see two qualitative changes: first a crossover from a single unimodal bulk to a single bimodal bulk, and then from one connected bulk to two disconnected bulks. This final splitting of the signal bulk from the noise bulk is a phase transition as the block Stieltjes transform goes from having a single branch cut to two disjoint branch cuts. This transition happens at significantly larger signal s than the signal-detectability phase transition in the finite-rank regime (Fig. 4 bottom).

In the limit of low rank (small b) the spectrum approaches the finite-rank theory as expected (Appendix A). Interestingly, we find that as a function of rank ratio b, there are three distinct regimes. For sufficiently strong signals (Fig. 5(a)), the signal bulk remains disjoint from the noise bulk for all b. For intermediate signals (Fig. 5(b)), the two bulks merge but the spectrum remains bimodal for all b. Finally, for weak signals (Fig. 5(c)), there is a single connected bulk for all b.

2. Singular vector subspace overlap in the extensive spike model

We now turn to the singular vectors of the extensive spike model. For simplicity we focus on the left singular vectors.

Since the K nonzero singular values of the signal are degenerate, the only meaningful overlap to study is a subspace overlap, or the projection of the data singular vectors, $\hat{\boldsymbol{u}}_{1k}$, onto the entire subspace defined by U_1. Therefore we compute

$$\left\|\hat{\boldsymbol{u}}_{1k}^T U_1\right\|^2 = \sum_{m=1}^{K} \left(\hat{\boldsymbol{u}}_{1k}^T \boldsymbol{u}_{1m}\right)^2. \tag{47}$$

Since this is an extensive sum, we expect that it is self-averaging and should be well predicted by $b\Phi_1(\hat{s}_k, s)$, where Φ is defined in (43).

After solving (46) for the block Stieltjes transform of R, we insert the result in (43) to find $\Phi_1(\hat{s}, s)$. In Appendix A we return to the simulation results presented in Fig. 2 and show that the extensive-rank theory predicts both the leading outlier singular value and the subspace overlap of the corresponding singular vector, even when the finite-rank theory fails.

In Fig. 6 we explore the phase diagram of the extensive-rank model and successfully confirm the predictions of the extensive-rank theory for singular vector overlaps by comparing these predictions to numerical simulations. For strong signal s (Fig. 6 top panel), the overlap of the data singular vectors with the true signal subspace is reasonably approximated by the finite-rank theory [22]. However, for moderate signals (Fig. 6 second panel) the data singular vectors interact, competing for the signal subspace. Singular vectors associated
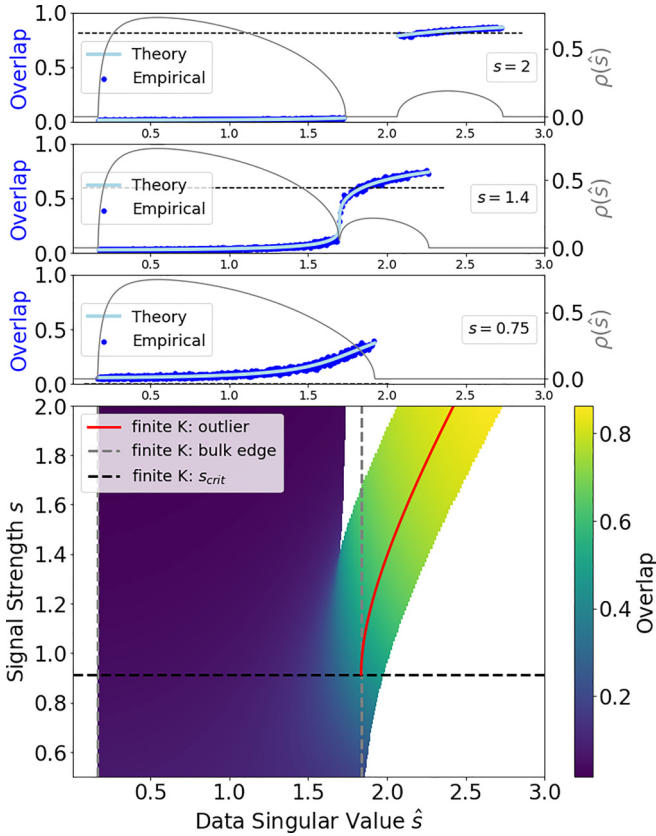
FIG. 6. Singular vector overlaps in the extensive spike model. Each row of the bottom color plot shows the theoretical prediction for the overlap of a left singular vector with singular value $\hat{s}$ of the data matrix $R = Y + X$, with the entire K-dimensional signal subspace of Y (i.e., the squared norm of the projection of the singular vector onto this subspace). The prediction is given by $b\Phi_1(\hat{s}, s)$, using (43) after numerically solving for $g_1^R(\hat{s})$ from (46). Different rows along the y axis correspond to different signal strengths s for Y. Comparisons to the finite-rank theory are shown using the same conventions as in Fig. 4. The top three panels show horizontal slices for $s = 0.75, 1.4, 2.0$. Solid gray curves indicate the singular value density of the data matrix R in the extensive-rank model. Blue dots indicate numerical calculations for the overlap for a single realization with $N_2 = 2000$. Solid light blue lines through the blue dots indicate matching theoretical predictions for this overlap. For comparison, the horizontal dashed line indicates the overlap predicted by the finite-rank theory (which depends only on s and not $\hat{s}$). The third panel indicates that the signal subspace of Y is detectable in the top data singular vectors of R, *even* at small signal strengths s *below* the transition in the singular value density of R from unimodal to bimodal. The aspect ratio is $c = N_1/N_2 = 0.7$, while the rank ratio for Y is fixed at $b = K/N_1 = 0.1$.

with the leading edge of the signal bulk have higher subspace overlap with the signal, while those at the lower edge overlap less. Perhaps most intriguingly, even for weak signals below the finite-rank phase transition at $s = s_{crit}$ the top data singular vectors still overlap significantly with the signal subspace (Fig. 6 third panel). Note, this overlap is nontrivial and $O(1)$ *even* when the singular value spectrum of the data is in the unimodal bulk regime.

We observe that the extensive spike model exhibits a singular value inflation in its *singular-vector overlaps*. Not only
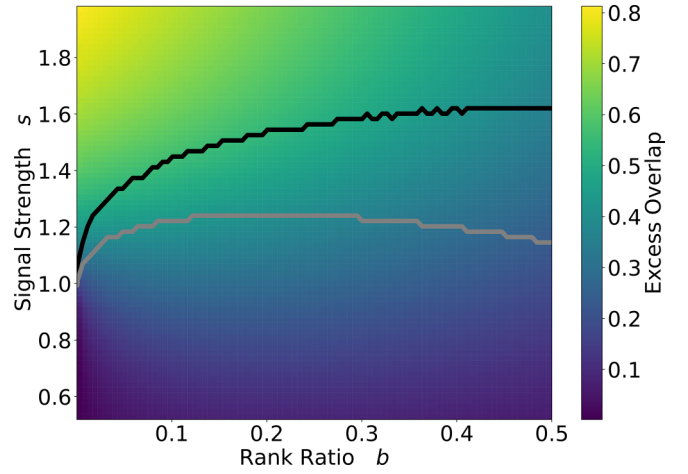

FIG. 7. Singular vector overlaps disregard singular value phases. Two-dimensional phase diagram shows the average "excess" subspace overlap (48) of the top b fraction of data singular vectors with a signal of strength s (y axis) and rank ratio b (x axis). The (lower) gray line separates the unimodal and bimodal regimes of the SV spectrum, and the (upper) black line separates the connected phase from the disconnected phase. The singular vector overlap does not respect the boundaries of the SV spectrum. The signal impacts the data via significant overlaps with the signal subspace well below the boundary between unimodal and bimodal regimes. Aspect ratio $c = 0.7$.

are data singular values larger than the corresponding signal singular values, just as in the finite-rank model, but also the singular-vector overlap peaks at the upper edge of the data singular values.

Figure 7 summarizes the results of this section with a two-dimensional phase diagram in the signal-strength vs rank (s-b) plane. It shows the boundaries between three regimes of the singular value spectrum: unimodal, bimodal, and disconnected. Additionally, the color map shows the average *excess* signal subspace overlap of the singular vectors associated with the top b fraction of singular values. Since, by chance, any random vector is expected to have an overlap b with the signal subspace, we compute the excess overlap as $\tilde{\Phi}(\hat{s}, s) = b(\Phi(\hat{s}, s) - 1)$. We then average the excess overlap across the singular vectors associated with the top b singular values, that is,

$$\int_t^\infty \tilde{\Phi}(\hat{s}, s)\rho_1^R(\hat{s})\, d\hat{s}, \tag{48}$$

where t is given by $b = \int_t^\infty \rho_1^R(\hat{s})\, d\hat{s}$.

Importantly, the figure demonstrates that in contrast to the finite-rank setting, the transitions in the data singular value spectrum of the data do *not* coincide with the detectability of the signal. Rather, the alignment of the data singular vectors with the signal subspace is a smooth function of both signal strength s and rank ratio b, and nonzero excess overlap can occur even in the unimodal regime.

V. OPTIMAL ROTATIONALLY INVARIANT ESTIMATORS

We now consider two estimation problems given noisy observations, $R = Y + X$: (1) denoising R in order to optimally

reconstruct Y and (2) estimation of the true signal covariance, $C = YY^T$. We focus on the case where both signal Y and noise X are rotationally invariant ($P_Y(M) = P_Y(O_1MO_2)$ for arbitrary orthogonal matrices O_1, O_2, and similarly for X). In this setting it is natural to consider *rotationally invariant* estimators F that transform consistently with rotations of the data: $F(O_1RO_2) = O_1F(R)O_2$ [47,48]. Such F can alter only the singular values of R while leaving the singular vectors unchanged. More generally, when Y is not rotationally invariant, our results yield the best estimator that modifies only singular values of R.

Our problem thus reduces to determining optimal shrinkage functions for the singular values. In the finite-rank case, distinct singular values and their associated singular vectors of Y respond independently to noise, so the optimal shrinkage of $\hat{s}$ depends only on $\hat{s}$ [23,29,46]. As we show below, this is no longer the case in the extensive-rank regime. The optimal shrinkage for each singular value generally depends on the entire data singular value spectrum.

A. Denoising rectangular data

We first derive a minimal mean-square error (MMSE) denoiser to reconstruct the rotationally invariant signal, Y, from the noisy data, R. Under the assumption of rotational invariance, the denoised matrix is constrained to have the same singular vectors as the data R and thus takes the form $\tilde{Y} = \hat{U}_1 \phi(\hat{S}) \hat{U}_2^T$. The MSE can be written

$$\mathcal{E} = \frac{1}{N_1N_2} \text{Tr}(Y - \tilde{Y})(Y - \tilde{Y})^T$$

$$= \frac{1}{N_1N_2} \sum_m s_m^2 + \phi^2(\hat{s}_m) - 2\phi(\hat{s}_m)\hat{u}_{1m}^T Y \hat{u}_{2m}. \quad (49)$$

Minimizing with respect to $\phi(\hat{s}_m)$ gives the optimal shrinkage function,

$$\phi^*(\hat{s}_m) = \hat{u}_{1m}^T Y \hat{u}_{2m}, \quad (50)$$

which appears to require knowledge of the very matrix being estimated, namely, Y. However, in the large-size limit it is possible to estimate $\phi^*(\hat{s}_m)$ via the resolvent $G^R(z)$. We first write

$$\text{Tr}[YG^R(z)]_{11} = \text{Tr}[YR^T G_{RR^T}(z^2)]$$

$$= \sum_l \frac{\hat{s}_l}{z^2 - \hat{s}_l^2} \hat{u}_{1l}^T Y \hat{u}_{2l}. \quad (51)$$

As z is brought toward the singular value $\hat{s}_m$ the sum is increasingly dominated by the contribution from $\hat{u}_{1m}^T Y \hat{u}_{2m} = \phi^*(\hat{s}_m)$. We find

$$\phi^*(\hat{s}) = \frac{2}{\pi \rho_1^R(\hat{s})} \lim_{\eta \to 0} \text{Im}\{\tau_1[YG^R(\hat{s} - i\eta)]\}. \quad (52)$$

We next apply the subordination relation (18), yielding a product of Y with a Y resolvent, whose trace is readily found:

$$\tau_1[YG^R(z)] = \tau_1[YG^Y(\zeta)] = \zeta_1 g_1^Y(\zeta) - 1, \quad (53)$$

where $\zeta_a(z) = z - \mathcal{R}_a^X(g^R(z))$, and we have used the identity $\tau[CG_C(z)] = zg_C(z) - 1$ for arbitrary symmetric C.

Since $g_1^Y(\zeta) = g_1^R(z)$ (33), we obtain

$$\phi^*(\hat{s}) = \frac{2}{\pi \rho_1^R(\hat{s})} \lim_{\eta \to 0} \text{Im}\{\zeta_1(\hat{s} - i\eta)g_1^R(\hat{s} - i\eta)\}, \quad (54)$$

which depends only on the block Stieltjes transform of the empirical data matrix, R, and the block $\mathcal{R}$ transform of the noise, X. Importantly, the dependence on the unknown signal Y is gone, making this formula amenable to practical applications, at least when the noise distribution of X is known.

For i.i.d. Gaussian noise with known variance, $\frac{\sigma^2}{N_2}$, we have $\mathcal{R}_1^X(\boldsymbol{g}) = \sigma^2 g_2$ and the general relation $g_2(z) = cg_1(z) + \frac{1-c}{z}$, so (54) simplifies considerably. Writing the real and imaginary parts, $g_1^R(z) = h_1^R(z) + if_1^R(z)$, we obtain the following simple expression depending only on the variance of the noise, and the Hilbert transform of the observed data spectral density:

$$\phi^*(\hat{s}) = \hat{s} - 2c\sigma^2 h_1^R(\hat{s}) - \sigma^2 \frac{1-c}{\hat{s}}. \quad (55)$$

This expression for the Gaussian case was derived previously in [41].

Figure 8 compares (55) to the optimal shrinkage found based on the finite-rank theory [29]. The extensive-rank formulas recover many more significant singular values (Fig. 8(a)). Moreover the mean-square error of $Y^* = \hat{U}_1 \phi^*(\hat{S}) \hat{U}_2$ is superior to that of the finite-rank denoiser, steadily improving as a function of the signal rank, while the finite-rank denoiser worsens (Fig. 8(b)). In fact, for our simulations with $N_1 = 1000$ and $N_2 = 500$, the extensive-rank denoiser outperformed the finite-rank denoiser for all $K > 5$, across the range of signal strengths tested. Finally, given an estimate of the noise variance σ^2, we are able to numerically estimate $g_1^R(\hat{s})$ with kernel methods (Appendix F) and compute an empirical shrinkage function that is very close to the theoretical optimum (Fig. 8(c)).

B. Estimating the signal covariance

We now derive an MMSE-optimal rotationally invariant estimator for the signal covariance, $C = YY^T$. Just as in [31,33], and similarly to our results in the previous section, the optimal estimator is given by $C^* = \hat{U}_1, \psi^*(\hat{S})\hat{U}_1^T$, where

$$\psi^*(\hat{s}_l) = \hat{u}_{1l}^T C \hat{u}_{1l}. \quad (56)$$

We observe that the top-left block of the square of the Hermitianization $\boldsymbol{Y}$ is given by C, and so

$$[\boldsymbol{Y}^2\boldsymbol{G}^R(z)]_{11} = \sum_l \frac{z}{z^2 - \hat{s}_l^2} \hat{u}_{1l}^T C \hat{u}_{1l}. \quad (57)$$

Thus, we can calculate the optimal shrinkage function by the inversion relation (13):

$$\psi^*(\hat{s}) = \frac{2}{\pi \rho_1^R(\hat{s})} \lim_{\eta \to 0} \text{Im}\{\tau_1[\boldsymbol{Y}^2\boldsymbol{G}^R(\hat{s} - i\eta)]\}. \quad (58)$$

Now we apply the subordination relation, $\boldsymbol{G}^R(z) = \boldsymbol{G}^Y(\boldsymbol{\zeta}(z))$ with $\zeta_a = z - \mathcal{R}_a^X(\boldsymbol{g}^R(z))$, which gives $\boldsymbol{Y}^2\boldsymbol{G}^R(z) = \boldsymbol{Y}^2\boldsymbol{G}^Y(\boldsymbol{\zeta})$, which has top-left block $\zeta_2 YY^T G_{YY^T}(\zeta_1\zeta_2)$.

Again, using the identity $\tau[CG_C(z)] = zg_C(z) - 1$ for arbitrary symmetric C, we have

$$\tau_1[\boldsymbol{Y}^2\boldsymbol{G}^R(z)] = \zeta_2[\zeta_1 g_1^Y(\zeta_1\zeta_2) - 1]. \quad (59)$$
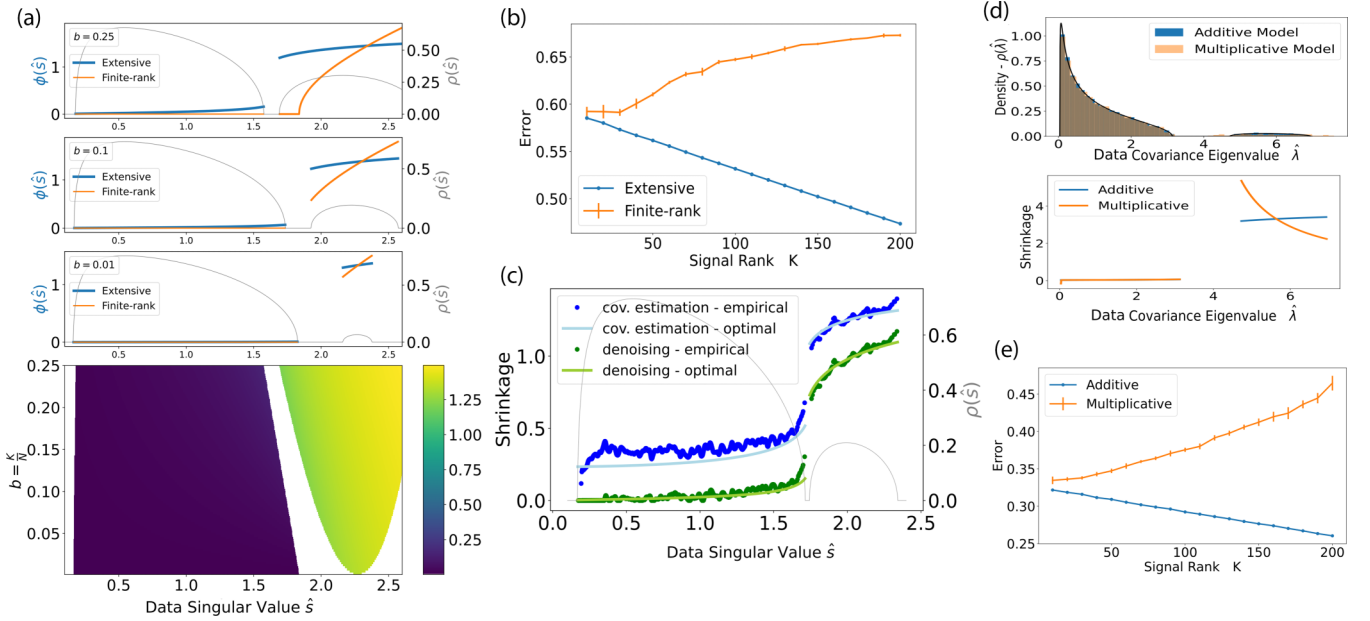
FIG. 8. Optimal denoising of extensive spikes. (a) Each row of the bottom color map shows the optimal shrinkage function $\phi^*(\hat{s})$ (55) for denoising data from the extensive spike model. Different rows on the y axis correspond to different rank ratios $b = {}^K/_{N_1}$ of the signal Y, while the signal strength s of Y is fixed at $s = 1.8$ and the aspect ratio is fixed at $c = 0.7$. The top three panels show horizontal slices with $b = 0.25, 0, 1, 0.01$. Blue (darker) curves indicate the optimal shrinkage function for the extensive-rank model, while orange (lighter) curves indicate the optimal shrinkage function for the finite-rank model [29] (7) (which does not depend on b). These panels indicate that the optimal shrinkage function for the extensive-rank model balances singular values more than that of the finite-rank model by more (less) aggressively shrinking larger (smaller) singular values. (b) Comparison of mean-square error in rectangular data denoising of K spikes, as a function of K for fixed signal strength $s = 2$, using the optimal shrinkage function for the finite-rank model (orange) vs that of the extensive spike model (blue). Even at small spike numbers of $K = 10$ for $N_1 = 1000 \times N_2 = 500$ sized data matrices, the extensive denoiser outperforms the finite-rank denoiser, and at larger K the extensive (finite-rank) denoiser gets better (worse). (c) Empirical shrinkage and comparison of optimal shrinkage function for two different errors: in blue (darker) denoising the rectangular signal matrix, $\phi^*(\hat{s})$ (55), and in green (lighter) estimating the $N_1 \times N_1$ signal covariance matrix $\sqrt{\psi^*(\hat{s})}$ (61), for signal strength $s = 1.5$ and rank ratio $b = 0.1$, with aspect ratio $c = 0.7$. Lighter curves show the theoretical optima found using Eq. (46). Darker dots show empirical shrinkage obtained via kernel estimation (see Appendix F) of the block Stieltjes transform from the data singular values with $N_2 = 2000$. (d) Comparison between multiplicative model (spiked covariance) and additive model (spiked rectangular model) with $K = 50$. *Top:* Eigenvalue spectra of data covariance (RR^T) for multiplicative model and additive model. *Bottom:* Optimal shrinkage for covariance estimation under the wrong model. Data spectrum generated by additive model, shrinkage function of multiplicative model [31] (13) vs the correct, additive model. (e) Mean-square error in covariance estimation as a function of K using the multiplicative model (orange) vs the correct, additive model (blue). Throughout (d) and (e), $s = 2$ with $N_1 = 1000$ and $N_2 = 1500$, and the multiplicative model is displayed in orange (lighter) and the additive model is displayed in blue (darker).

We therefore conclude for general noise matrix, X:

$$\psi^*(\hat{s}) = \frac{2}{\pi \rho_1^R(\hat{s})} \lim_{\eta \to 0} \text{Im}\big[\zeta_2(\hat{s} - i\eta)$$

$$\times \big(\zeta_1(\hat{s} - i\eta) g_1^R(\hat{s} - i\eta) - 1\big)\big] \qquad (60)$$

Once again, for i.i.d. Gaussian noise with known variance, $\frac{\sigma^2}{N_2}$, our estimator (60) simplifies considerably. Using the optimal shrinkage function found above for rectangular denoising, $\phi^*(\hat{s}) = \hat{s} - 2c\sigma^2 h_1^R - \sigma^2 \frac{1-c}{\hat{s}}$, where h_1^R is the real part of $g_1^R(\hat{s})$, we finally obtain

$$\psi^*(\hat{s}) = \phi^*(\hat{s}) \bigg(\phi^*(\hat{s}) + \sigma^2 \frac{1-c}{\hat{s}} \bigg)$$

$$- c\sigma^2 \big(c\sigma^2 |g_1^R(\hat{s})|^2 - 1 \big). \qquad (61)$$

Just as in optimal data denoising, we find that given an estimate of the noise variance, σ^2, the optimal shrinkage for covariance estimation depends only on the spectral density of

R and its Hilbert transform, which can be estimated directly from data.

In Fig. 8 we show the optimal shrinkage function for the extensive spike model, and demonstrate that it can be approximated given only an estimate of the noise variance and the empirical data matrix, R (Fig. 8(c)). We find that the optimal singular value shrinkage of singular values derived for covariance estimation (61), $\sqrt{\psi^*(\hat{s})}$, is substantially different than $\phi^*(\hat{s})$ (55) obtained for denoising the rectangular signal (Fig. 8(c)). The denoising shrinkage suppresses the noise more aggressively, but suppresses the signal singular values more as well.

Finally, we compare the shrinkage obtained from assuming a multiplicative form of noise instead of the additive spiked rectangular model studied here. In the finite-rank regime, the spiked rectangular model can be instead modeled as a multiplicative model with data arising form a spiked covariance. Concretely, the data in the multiplicative model are generated as $R_{mult} = \sqrt{C_{mult}}X$, i.e., each column is sampled

from a spiked covariance: $C_{mult} = YY^T + I$. In the finite-rank regime, with Gaussian noise, the two models yield identical spectra and covariance-eigenvector overlaps. The optimal shrinkage for covariance estimation for the multiplicative model for arbitrary C_{mult} has previously been reported [31] Eq. (13) for Gaussian noise and [33] Eq. (IV.8) for more general noise], and here we consider the impact of employing the multiplicative shrinkage formula on data generated from the additive spiked rectangular model. We observe (Fig. 8(d) top) that for small rank ratio ($b = 0.05$) the two models give fairly similar eigenvalue distributions. Nevertheless, applying the optimal multiplicative shrinkage on the additive model data gives poor results: the shrinkage obtained is nonmonotonic in the data eigenvalue (Fig. 8(d) bottom). Furthermore, the mean-square error in covariance estimation obtained with the multiplicative shrinkage worsens as a function of rank (Fig. 8(e)).

VI. DISCUSSION

While one approach to estimation depends on prior information about the structure of the signal (such as sparsity of singular vectors, for example), we have followed a line of work on rotationally invariant estimation that assumes there is no special basis for either the signal or the noise [47,48]. In this approach, knowledge of the expected deformation of the singular value decomposition (SVD) of the data due to noise allows for the explicit calculation of optimal estimators.

In the case of finite-rank signals, where the impact of additive noise on singular values and vectors is known [21,22], formulas for optimal shrinkage for both denoising [23,28,29] and covariance estimation [46] have been found. For extensive-rank signals, however, while formulas for the singular value spectrum of the free sum of rectangular matrices are known [37,38,40], there are no prior results for the singular vectors of sums of generic rectangular matrices (though see [44] for contemporaneous results).

Even in the setting of square, Hermitian matrices, results on eigenvectors of sums are relatively new [30,31]. Recent work derived a subordination relation for the product of square symmetric matrices and applied it to a "multiplicative" noise model in which each observation of high-dimensional data is drawn independently from some unknown, potentially extensive-rank, covariance matrix [33]. In that context, knowledge of the overlaps of the data covariance with the unobserved population covariance is sufficient to enable the construction of an optimal rotationally invariant estimator [31,33,35,36].

We have derived analogous results for signals with additive noise: we have computed an asymptotically exact subordination relation for the block resolvent of the free sum of rectangular matrices, i.e., for the resolvent of the Hermitianization of the sum in terms of the resolvents of the Hermitianization of the summands. From the subordination relation, we derived the expected overlap between singular vectors of the sum and singular vectors of the summands. These overlaps quantify how singular vectors are deformed by additive noise. We have calculated separate optimal nonlinear singular-value shrinkage expressions for signal denoising and for covariance estimation. Under the assumption of i.i.d.

Gaussian noise these shrinkage functions depend only on the noise variance and the empirical data singular value density, which we have shown can be estimated by kernel methods.

We have applied our results in order to study the extensive spike model. We found a significant improvement in estimating signals with even fairly low rank ratios, over methods that are based on the finite-rank theory. Our results may have significant impact on ongoing research questions around spiked matrix models [24–27], such as the question of the detectability of spikes or optimal estimates for the number of spikes, for example.

The subordination relation derived here is closely related to operator-valued free probability, which provides a systematic calculus for block matrices with orthogonally or unitarily invariant blocks, such as the 2×2-block Hermitianizations $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{R}$. In that approach, spectral properties of a matrix are encoded via 2×2 operator-valued Stieltjes and $\mathcal{R}$ transforms, whose diagonal elements correspond exactly to the block Stieltjes and $\mathcal{R}$ transforms defined here. A fundamental result in this context is an additive subordination relation for the operator-valued Stieltjes transform, which is an identical formula to (33) [40].

We comment briefly on our derivation of the block resolvent subordination, which is summarized in Sec. IV A and treated fully in Appendix B. First, we note that previous work derived resolvent subordination relations for square symmetric matrices using the replica method [33,34,36]. These works assume the replicas decouple, which results in a calculation that is equivalent to computing the annealed free energy. Here we used concentration of measure arguments to prove that the annealed approximation is asymptotically correct (Appendix B 1).

In the course of our derivation of the subordination relation we encountered the expectation over arbitrary block-orthogonal rotations of the Hermitianization of the noise matrix [Eq. (24) and Appendix C], which we called a "block spherical integral." As noted above, this integral plays an analogous role to the HCIZ spherical integral which appears in the derivation of the subordination relation of square symmetric matrices [36]. In that setting, the logarithm of the rank-1 spherical integral yields the antiderivative of the standard $\mathcal{R}$ transform for square symmetric matrices [49]. To our knowledge, the particular block spherical integral in our work (Appendix C) has not been studied previously. In fact, it is very closely related to the rectangular spherical integral, whose logarithm is the antiderivative of the so-called *rectangular* $\mathcal{R}$ transform [37]. In our setting, two such rectangular spherical integrals are coupled, and the logarithm of the result is the antiderivative of the *block* $\mathcal{R}$ transform (14) (up to componentwise proportionality constants related to the aspect ratio). While the *rectangular* $\mathcal{R}$ transform is additive, its relationship to familiar RMT objects such as the Stieltjes transform is quite involved. In contrast, the block $\mathcal{R}$ transform that arises from the block spherical integral is a natural extension of the scalar $\mathcal{R}$ transform, with a simple definition in terms of the functional inverse of the block Stieltjes transform. Furthermore, as mentioned above, the block $\mathcal{R}$ transform is essentially a form of the more general operator $\mathcal{R}$ transform from operator-valued free probability. This formulation is appealing because it provides a direct link between

a new class of spherical integrals and operator-valued free probability.

We stress that even under the assumption of Gaussian i.i.d. noise, the optimal estimators we obtained in Sec. V are not quite *bona fide* empirical estimators, as they depend on an estimate of the noise variance. This may not be a large obstacle, but we leave it for future work. We do note that while under the assumption of finite-rank signals, appropriate noise estimates can be obtained straightforwardly, for example, from the median data singular value (see [28], for example), this is no longer the case in the extensive regime that we study. In empirical contexts in which one has access to multiple noisy instantiations of the same underlying signal, however, a robust estimate of the noise variance may be readily available.

Other recent work has also studied estimation problems in the extensive-rank regime. Reference [50] studied the distribution of pairwise correlations in the extensive regime. Reference [41] studied optimal denoising under a known, factorized extensive-rank prior, and arrived at the same shrinkage function we find for the special case of Gaussian i.i.d. noise ([55]). References [43] and [42] studied both denoising and matrix factorization (dictionary learning) with known, extensive-rank prior.

Finally, recently a preprint [44] presented work partially overlapping with ours. They derived the subordination relation for the resolvent of Hermitianizations as well as the optimal rotationally invariant data denoiser, and additionally established a relationship between the rectangular spherical integral and the asymptotic mutual information between data and signal. However, unlike our work, this contemporaneous work (1) does not calculate the optimal estimator of the signal covariance; (2) does not explore the phase diagram of extensive spike model and its associated conceptual insights about the decoupling of singular value phases from singular vector detectability that occurs at extensive but not finite rank; (3) does not extensively numerically explore the inaccuracy and inferior data-denoising and signal-estimation performance of the finite-rank model compared to the extensive-rank model, a key motivation for extensive rank theory; (4) at a technical level [44] follows the approach of [33] using a decoupled replica approach yielding an annealed approximation, whereas we prove the annealed approximation is accurate using results from concentration of measure; and (5) also at a technical level [44] employs the rectangular spherical integral resulting in rectangular $\mathcal{R}$ transforms, whereas we introduce the block spherical integral yielding the block $\mathcal{R}$ transform, thereby allowing us to obtain simpler formulas.

We close by noting that our results for optimal estimators depend on the assumption of rotational (orthogonal) invariance. Extending this work to derive estimators for extensive-rank signals with structured priors is an important topic for future study. The rectangular subordination relation and the resulting formulas for singular vector distortion due to additive noise hold for arbitrary signal matrices. These may prove to be of fundamental importance from the perspective of signal estimation in the regime of high-dimensional statistics, as any attempt to estimate the structure of a signal in the presence of noise must overcome both the distortion of the signal's singular value spectrum *and* the deformation of the signal's singular vectors.

APPENDIX A: FINITE-RANK THEORY FOR THE SPIKED MATRIX MODEL

We review formulas from [22] for the finite-rank spiked matrix model, $R = sU_1U_2^T + X$, where the U_a are $N_a \times K$ with orthonormal columns, and X is a random $N_1 \times N_2$ matrix with well-defined singular value spectrum in the large-size limit with fixed aspect ratio $c = N_1/N_2$.

In the case where the noise X is i.i.d. Gaussian with variance $1/N_2$, the critical signal strength below which the signal is undetectable is $s_{crit} = c^{1/4}$. The top K singular values of R are given by

$$\hat{s}_{l \leqslant K} = \begin{cases} s\sqrt{\left(1 + \frac{c}{s^2}\right)\left(1 + \frac{1}{s^2}\right)} & \text{for } s > s_{crit} \\ 1 + \sqrt{c} & \text{otherwise.} \end{cases} \quad \text{(A1)}$$

In Fig. 9 we show that in the limit of small rank ratio, the singular value density obtained from the extensive-rank theory approaches this result from finite-rank theory. For the square-symmetric setting, see also [45] for derivation of the fluctuations around this asymptotic limit, which take the form of the eigenvalues of a $K \times K$ random matrix.
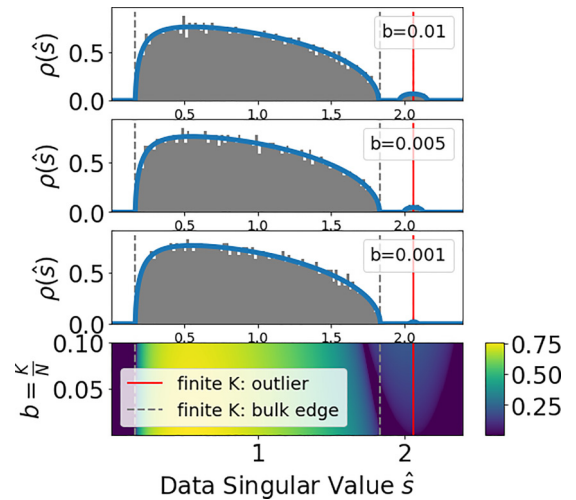

FIG. 9. Singular value density of extensive spike model in the low-rank limit. Color plot shows singular value density of extensive spike model with aspect ratio $c = N_1/N_2$ and signal strength $s = 1.5$. The rank ratio, $b = K/N_1$, varies along the y axis. Top three panels show horizontal slices for $b = 0.001, 0.005, 0.01$ together with empirical histograms of individual model instantiations with $N_2 = 2000$. The extensive-rank theory converges to the finite-rank theory as b gets small.
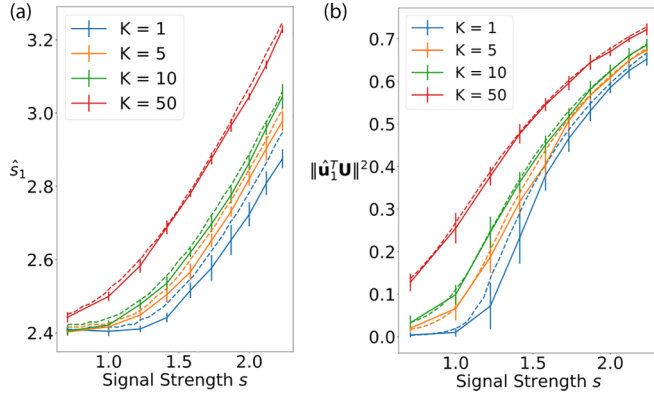
FIG. 10. Extensive-rank theory captures the singular value structure of the spiked rectangular matrix model. Compare Fig. 2. (a) Leading data singular value, $\hat{s}_1$, as a function of signal singular value, s_1, for various ranked spikes. (b) Projection of leading data left singular vector, $\mathbf{u}_1$, on the K-dimensional left singular space of the signal. Dashed lines show extensive-rank theory. The two panels match Figs. 2(a) and 2(b), with $N_1 = 1000$ and $N_2 = 500$ and numerical results presented as mean and standard deviation over 10 instantiations for each value of b and s. This figure shows that the extensive-rank theory captures the deviations from finite-rank theory at finite N_1 and N_2.

The overlaps of the corresponding singular vectors, $\hat{\boldsymbol{u}}_{al}$ for $l = 1, \dots, K$, with the signal subspaces, U_a, for $a = 1, 2$ are given by

$$\left\| \hat{\boldsymbol{u}}_{1l}^T U_1 \right\|^2 = \begin{cases} \frac{s^4 - c}{s^4 + cs^2} & \text{for } s > s_{crit} \\ 0 & \text{otherwise} \end{cases}, \quad \text{(A2a)}$$

$$\left\| \hat{\boldsymbol{u}}_{2l}^T U_2 \right\|^2 = \begin{cases} \frac{s^4 - c}{s^4 + s^2} & \text{for } s > s_{crit} \\ 0 & \text{otherwise} \end{cases}. \quad \text{(A2b)}$$

In Fig. 10 we return to the simulation results from Fig. 2 displaying both the leading singular value and the overlap for various values of K, together with the theory results from the extensive-rank theory.

For a generic noise matrix, X, with block Stieltjes transform, $\boldsymbol{g}^X(z)$, [22] defines the D transform, which in our notation is the product of the elements of $\boldsymbol{g}^X(z)$ (where the argument has $z_1 = z_2 = z$):

$$D_X(z) = g_1^X(z) g_2^X(z). \quad \text{(A3)}$$

Then the critical signal satisfies

$$D_X(x_+) = \frac{1}{s_{crit}^2}, \quad \text{(A4)}$$

where x_+ is the supremum of the support of the singular value spectrum of X.

For suprathreshold signals, $s > s_{crit}$, the data singular value outlier, $\hat{s}$, satisfies

$$D_X(\hat{s}) = \frac{1}{s^2}, \quad \text{(A5)}$$

and the two overlaps, corresponding to blocks $a = 1$ and $a = 2$ for left and right singular vectors, respectively, are given by

$$\left\| \hat{\boldsymbol{u}}_a^T U_a \right\|^2 = \frac{-2 g_a^X(\hat{s})}{D_X(\hat{s}) D_X'(\hat{s})}. \quad \text{(A6)}$$

APPENDIX B: DERIVATION OF THE BLOCK-RESOLVENT SUBORDINATION RELATION

Here we calculate the asymptotic subordination relation (18), found in Sec. IV A, for the block resolvent of the free sum of rectangular matrices $R = Y + O_1 X O_2^T$, and O_a Haar-distributed orthogonal matrices of size N_a for $a = 1, 2$. We write $N = N_1 + N_2$ and study the large N limit with fixed aspect ratio $c = N_1/N_2$. For notational ease we introduce the ratio of each block's size to the entire matrix:

$$\beta_a := \frac{N_a}{N}. \quad \text{(B1)}$$

We begin by writing

$$\boldsymbol{M} := zI - \boldsymbol{R} = zI - \left(Y + \bar{\boldsymbol{O}} X \bar{\boldsymbol{O}}^T \right), \quad \text{(B2)}$$

where $\bar{\boldsymbol{O}} = \begin{bmatrix} O_1 & 0 \\ 0 & O_2 \end{bmatrix}$. Next we define the partition function, $\mathcal{Z}^R(\boldsymbol{Y}) := (\det \boldsymbol{M})^{-1/2}$, which we can write as a Gaussian integral:

$$\mathcal{Z}^R(\boldsymbol{Y}) = \int \frac{d\boldsymbol{v}}{\sqrt{2\pi}^N} \exp\left(-\frac{1}{2} \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v} \right). \quad \text{(B3)}$$

We also define a corresponding free energy

$$\mathcal{F}^R(\boldsymbol{Y}) := 2 \log \mathcal{Z}^R(\boldsymbol{Y}), \quad \text{(B4)}$$

and the desired block resolvent is $\boldsymbol{G}^R(z) = \boldsymbol{M}^{-1} = \frac{d}{d\boldsymbol{Y}} \mathcal{F}^R(\boldsymbol{Y})$.

Prior work on the case of square symmetric matrices has employed the replica trick to compute this quenched average [33,34,36,44]. In our notation, this amounts to approximating $\log \mathcal{Z}^R = \lim_{n \to 0} \frac{(\mathcal{Z}^R)^n - 1}{n}$, and then computing $\mathbb{E}_{\bar{\boldsymbol{O}}}[\boldsymbol{G}^R(z)] = \lim_{n \to 0} \mathbb{E}_{\bar{\boldsymbol{O}}}[(\mathcal{Z}^R)^{n-1} \frac{d\mathcal{Z}^R}{d\boldsymbol{Y}}]$ via n Gaussian integrals. Prior work has assumed that the replicas do not couple, which effectively amounts to computing the annealed average, $\log \mathbb{E}_{\bar{\boldsymbol{O}}}[\mathcal{Z}^R(\boldsymbol{Y})]$.

Instead, we show in Appendix B 1 using concentration inequalities that the annealed calculation is in fact asymptotically exact. In particular, as $N \to \infty$,

$$\mathbb{E}_{\bar{\boldsymbol{O}}} \left[\frac{2}{N} \log \mathcal{Z}^R(\boldsymbol{Y}) \right] \to \frac{2}{N} \log \mathbb{E}_{\bar{\boldsymbol{O}}}[\mathcal{Z}^R(\boldsymbol{Y})]. \quad \text{(B5)}$$

Writing out the expectation and separating factors that depend on $\bar{\boldsymbol{O}}$, we have

$$\mathbb{E}_{\bar{\boldsymbol{O}}}[\mathcal{Z}^R(\boldsymbol{Y})] = \int \frac{d\boldsymbol{v}}{\sqrt{2\pi}^N} e^{-\frac{1}{2} \boldsymbol{v}^T (zI - \boldsymbol{Y}) \boldsymbol{v}}$$
$$\times \mathbb{E}_{\bar{\boldsymbol{O}}}[e^{\frac{1}{2} \boldsymbol{v}^T \bar{\boldsymbol{O}} X \bar{\boldsymbol{O}}^T \boldsymbol{v}}]. \quad \text{(B6)}$$

The expectation over $\bar{\boldsymbol{O}}$ on the right hand side is a rank-1 block spherical integral. In Appendix C, we derive an asymptotic expression for the expectation, which depends only on, $\mathcal{R}^X$, the block $\mathcal{R}$ transform of the noise matrix X, and the blockwise norms of the vector, $\boldsymbol{v}$. Introducing the two-element vector, $\boldsymbol{t}$ whose ath entry is $\frac{1}{N_a} \|\boldsymbol{v}_a\|^2$, we have

$$\mathbb{E}_{\bar{\boldsymbol{O}}}[e^{\frac{1}{2} \boldsymbol{v}^T \bar{\boldsymbol{O}} X \bar{\boldsymbol{O}}^T \boldsymbol{v}}] = \exp\left(\frac{N}{2} H^X(\boldsymbol{t}) \right), \quad \text{(B7)}$$

where in anticipation of a saddle-point condition below, we write $H^X(t)$ as a contour integral within $\mathbb{C}^2$ from 0 to t:

$$H^X(t) := \int_0^t dw \cdot [\beta \odot \mathcal{R}^X(w)], \qquad (B8)$$

where $\beta = \frac{1}{N}\binom{N_a}{N_b}$ and $\odot$ is elementwise product. This gives

$$\mathbb{E}_{\bar{O}}[\mathcal{Z}^R(Y)] = \int \frac{dv}{(\sqrt{2\pi})^N} e^{-\frac{1}{2}v^T(zI-Y)v} \exp\left(\frac{N}{2}H^X(t)\right). \qquad (B9)$$

In order to decouple v from t, we introduce integration variables and Fourier expressions for the delta-function constraints $\delta(N_a t_a - \|v_a\|^2)$:

$$1 = \int dt_a \int \frac{d\hat{t}_a}{4\pi i} \exp\left(-\frac{1}{2}\hat{t}_a(N_a t_a - \|v_a\|^2)\right). \qquad (B10)$$

We now have

$$\mathbb{E}_{\bar{O}}[\mathcal{Z}^R(Y)] = \int \left(\prod_a \frac{dt_a \, d\hat{t}_a}{4\pi i} e^{-\frac{1}{2}N_a t_a \hat{t}_a}\right) \exp\left(\frac{N}{2}H^X(t)\right)$$

$$\times \int \frac{dv}{(\sqrt{2\pi})^N} \exp\left(-\frac{1}{2}v^T(zI - \bar{T} - Y)v\right), \qquad (B11)$$

where we have introduced the diagonal $N \times N$ matrix, $\bar{T}$, which has $\hat{t}_1$ along the first N_1 diagonal elements followed by $\hat{t}_2$ along the remaining N_2 elements.

The integral over v is a Gaussian integral with inverse covariance $(zI - \bar{T} - Y)$ (which is positive-definite for sufficiently large z). Crucially, this covariance is exactly, $G^Y(z - \hat{t})$ the block resolvent of Y with a shifted argument, $z - \hat{t}$. Note that the block resolvent, as a function of two complex numbers, has emerged here in our calculation.

The result is the inverse square root of the determinant:

$$\int dv \, e^{-\frac{1}{2}v^T(zI-Y-\bar{T})v} \propto \det(zI - Y - \bar{T})^{-\frac{1}{2}}. \qquad (B12)$$

Thus, ignoring proportionality constants we have

$$\mathbb{E}_{\bar{O}}[\mathcal{Z}^R(Y)] \propto \int dt \, d\hat{t} \exp\left(\frac{N}{2}P^{X,Y}(t,\hat{t})\right), \qquad (B13)$$

with

$$P^{X,Y}(t,\hat{t}) := -\beta_1 t_1 \hat{t}_1 - \beta_2 t_2 \hat{t}_2 + H^X(t)$$
$$-\frac{1}{N}\log\det(zI - Y - \bar{T}), \qquad (B14)$$

where, remember, $\beta_a := N_a/N$.

We expect this integral to concentrate around its saddle point in the large-size limit. We find that taking the derivative of $P^{X,Y}(t,\hat{t})$ with respect to t_a gives the following appealing saddle-point condition for $\hat{t}$:

$$\hat{t} = R^X(t). \qquad (B15)$$

In order to take the derivatives with respect to $\hat{t}_a$, we find it helpful to write out N_2 singular values s_m of Y (including $N_2 - N_1$ zeros when $N_2 > N_1$). Then $(zI - Y - \bar{T})$ decouples into

2×2 matrices of the form $\begin{bmatrix} z - \hat{t}_1 & -s_m \\ -s_m & z - \hat{t}_2 \end{bmatrix}$, and that allows us to write

$$\det(zI - Y - \bar{T}) = (z - \hat{t}_1)^{(N_1 - N_2)}$$
$$\times \prod_{m=1}^{N_2} \left[(z - \hat{t}_1)(z - \hat{t}_2) - s_m^2\right].$$

Then we find that taking the derivative of (B14) gives the final saddle-point condition:

$$t_1 = (z - \hat{t}_2)g_{YY^T}((z - \hat{t}_1)(z - \hat{t}_2)), \qquad (B16)$$
$$t_2 = (z - \hat{t}_1)g_{Y^TY}((z - \hat{t}_1)(z - \hat{t}_2)). \qquad (B17)$$

We can write this concisely in vector notation:

$$t^* = g^Y(z - R^X(t^*)). \qquad (B18)$$

Thus, asymptotically, the desired free energy is $\mathbb{E}_{\bar{O}}[\mathcal{F}^R(Y)] = NP^{X,Y}(t^*, \mathcal{R}^X(t^*))$.

Informally, to derive the matrix subordination relation, we differentiate $\mathcal{F}^R$, which, from (B14), yields $(zI - \bar{T} - Y)^{-1} = G^Y(z - R(t^*))$. But we argued above that $\frac{d}{dY}\mathcal{F}^R(Y) = G^R(z)$, which gives the subordination relation.

More formally, consider a Hermitian test matrix, A, with a bounded spectral distribution, and then observe that $\frac{1}{N}\frac{d}{dy}\log\det(M + yA) = \tau[AM^{-1}]$. Thus, we substitute $Y \to Y + yA$ into the expression for $\mathcal{F}^R$ (B14) and differentiate to find

$$\lim_{N \to \infty} \tau[A\mathbb{E}_{\bar{O}}[G^R(z)]] = \lim_{N \to \infty} \tau[AG^Y(z - \mathcal{R}^X(t^*))]. \qquad (B19)$$

Using A proportional to either $\begin{bmatrix} I_{N_1} & 0 \\ 0 & 0 \end{bmatrix}$ or $\begin{bmatrix} 0 & 0 \\ 0 & I_{N_2} \end{bmatrix}$, we can now take the normalized blockwise traces of both sides, yielding

$$g^R(z) = g^Y(z - \mathcal{R}^X(t^*)). \qquad (B20)$$

Thus, comparing to (B18) we have $t^* = g^R(z)$, and (B20) becomes the subordination relation for the block Stieltjes transform. Substituting $t^* = g^R(z)$ into (B19), we obtain the desired resolvent relation

$$\tau[A\mathbb{E}_{\bar{O}}[G^R(z)]] = \tau[AG^Y(z - \mathcal{R}^X(g^R(z)))], \qquad (B21)$$

for all Hermitian test matrices A with bounded spectrum, or as written informally in the main text, $\mathbb{E}_{\bar{O}}[G^R(z)] = G^Y(z - \mathcal{R}^X(g^R(z)))$.

1. Proof that the annealed free energy asymptotically equals the quenched free energy

Suppose A, B are Hermitian matrices with bounded spectrum. Define the function

$$f(O) := \frac{1}{N}\log\det[zI - (A + OBO^T)], \qquad (B22)$$

for arbitrary orthogonal $O \in \mathbb{SO}(N)$. For sufficiently large z, the matrix in the determinant is always positive, and this is a smooth function on $\mathbb{SO}(N)$ bounded above and below by constants $c_\pm := \log[z \pm (\|A\|_{op} + \|B\|_{op})]$, where $\|\cdot\|_{op}$ is the operator norm. For such z, we prove the following

Lipschitz bound below (see Sec. B 1 a for proof):

$$|f(O_1) - f(O_2)| \leqslant \frac{\mu}{\sqrt{N}} \|O_1 - O_2\|_2, \qquad \text{(B23)}$$

where $\mu := \pi \|B\|_{op} e^{-c_-}$ and $\| \cdot \|_2$ is the Euclidean norm $\|X\|_2 = \sqrt{\text{Tr}[X^T X]}$.

In particular, we will be interested in the case that the orthogonal matrix O is block diagonal with blocks $O_a \in \mathbb{SO}(N_a)$, and thus O is a member of the product space $\mathbb{SO}(N_1) \times \mathbb{SO}(N_2)$ with $N_1 + N_2 = N$. The group $\mathbb{SO}(N_a)$ with Haar measure and Hilbert-Schmidt metric obeys a logarithmic Sobolev inequality with constant $\frac{4}{N_a - 2}$, so the product space has Sobolev constant $\max_a \frac{4}{N_a - 2} = \frac{4}{\gamma N - 2}$, where $\gamma := \min(\frac{N_1}{N}, \frac{N_2}{N})$ ([51], Theorems 5.9, 5.16), and we can apply Theorem 5.5 of [51], yielding

$$\mathbb{P}\left[|f(O) - \mathbb{E}_{\bar{O}}[f(O)]| \geqslant \frac{\mu}{\sqrt{N}} r \right] \leqslant 2 \exp\left(-(\gamma N - 2) \frac{r^2}{8} \right), \qquad \text{(B24)}$$

for all $r \geqslant 0$.

Writing $H := \mathbb{E}_O[f(O)]$, and defining $M := z - A - OBO^T$ (so that $\det M = e^{Nf(O)}$), this implies

$$\mathbb{P}[\det(M) \geqslant e^{NH + \sqrt{N}\mu r}] \leqslant 2e^{-(\gamma N - 2)\frac{r^2}{8}}. \qquad \text{(B25)}$$

Since $\det(M) \leqslant e^{Nc_+}$, we can upper bound the expectation:

$$\mathbb{E}_O[\det(M)] \leqslant (1 - 2e^{-(\gamma N - 2)\frac{r^2}{8}})e^{NH + \sqrt{N}\frac{\pi\mu}{2}r}$$
$$+ 2e^{-(\gamma N - 2)\frac{r^2}{8}} e^{Nc_+}. \qquad \text{(B26)}$$

Choosing $r = \sqrt{8c_+/\gamma}$, we find that $\frac{1}{N} \log \mathbb{E}_O[\det(M)]$ is less than or equal to

$$\frac{1}{N} \log[(1 - 2e^{-(N-2)c^+/\gamma})e^{NH + \sqrt{N}\frac{\pi\mu}{2}r} + 2e^{2c_+ + \gamma}] \xrightarrow{N \to \infty} H$$

$$= \frac{1}{N}\mathbb{E}_O[\log \det(M)],$$

which shows that the limiting annealed average is less than or equal to the limiting quenched average. We could obtain a lower bound via a similar argument, but we have directly via Jensen's inequality that the quenched average is less than or equal to the annealed average, $\frac{1}{N}\mathbb{E}_O[\log \det(M)] \leqslant \frac{1}{N} \log \mathbb{E}_O[\det(M)]$, so in the limit they are equal:

$$\lim_{N \to \infty} \frac{1}{N}\mathbb{E}_O[\log \det(M)] = \lim_{N \to \infty} \frac{1}{N} \log \mathbb{E}_O[\det(M)]. \quad \text{(B27)}$$

a. Lipschitz bound

To prove (B23), note that the gradient of f (B22) is

$$\nabla_O f(O) = -2\frac{1}{N} M^{-1} OB, \qquad \text{(B28)}$$

where, as above, $M = z - (A + OBO^T)$. Thus, the ordinary Euclidean norm of the gradient is

$$\|\nabla_O f(O)\|_2 = 2\frac{1}{N} \|M^{-1} OB\|_2 \qquad \text{(B29)}$$

$$= \frac{2}{N} \sqrt{\text{Tr}[M^{-2} OB^2 O^T]}. \qquad \text{(B30)}$$

M^{-2} and OB^2O^T are positive definite Hermitian matrices, so $\text{Tr}[M^{-2}OB^2O^T] \leqslant N\|B^2\|_{op}\|M^{-2}\|_{op}$. From M's definition we have $\|M^{-2}\|_{op} \leqslant [z - (\|A\|_{op} + \|B\|_{op})]^{-2} = e^{-2c_-}$, and so

$$\|\nabla_O f(O)\|_2 \leqslant \frac{2\|B\|_{op}}{e^{c_-} \sqrt{N}}. \qquad \text{(B31)}$$

This shows that f changes by at most $\frac{2\|B\|_{op}}{e^{c_-}\sqrt{N}}$ times the geodesic distance on the group: $|f(O_1) - f(O_2)| \leqslant \frac{2\|B\|_{op}}{e^{c_-}\sqrt{N}} d(O_1, O_2)_{\mathbb{SO}(N)}$. The geodesic distance is upper bounded by $\pi/2$ times the Euclidean distance ([51], p. 159), so

$$|f(O_1) - f(O_2)| \leqslant \frac{\pi\|B\|_{op}}{e^{c_-}\sqrt{N}} \|O_1 - O_2\|_2. \qquad \text{(B32)}$$

APPENDIX C: RANK-1 BLOCK SPHERICAL HCIZ INTEGRAL

In this section we introduce the "block spherical integral," which extends the HCIZ integral to the setting of Hermitianizations of rectangular matrices.

We consider an $N_1 \times N_2$ matrix, X, with $N = N_1 + N_2$ and consider the limit of large N with fixed $c = N_1/N_2$. For notational ease we will introduce

$$\beta_a := \frac{N_a}{N}, \qquad \text{(C1)}$$

for both $a = 1, 2$.

In the general-rank setting we write

$$I^X(\boldsymbol{T}) := \mathbb{E}_{\bar{O}}\left[\exp\left(\frac{N}{2} \text{Tr} \boldsymbol{T} \bar{\boldsymbol{O}} X \bar{\boldsymbol{O}}^T \right) \right], \qquad \text{(C2)}$$

where $\bar{\boldsymbol{O}} = \begin{bmatrix} O_1 & 0 \\ 0 & O_2 \end{bmatrix}$ is a block-orthogonal matrix, i.e., both O_a are Haar-distributed $N_a \times N_a$ matrices, and $\boldsymbol{T}$ is an arbitrary $N \times N$ matrix.

We here solve the rank-1 case, which arises in Appendix B in the calculation of the subordination relation. In order to match the normalization there, we write $T = \frac{1}{N}\boldsymbol{v}\boldsymbol{v}^T$, where the individual elements, v_i, are $O(1)$. We have

$$I^X(\boldsymbol{T}) := \mathbb{E}_{\bar{O}}\left[\exp\left(\frac{1}{2} \boldsymbol{v}^T \bar{\boldsymbol{O}} X \bar{\boldsymbol{O}}^T \boldsymbol{v} \right) \right]. \qquad \text{(C3)}$$

We write $\boldsymbol{v} = \begin{pmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \end{pmatrix}$ in block form, and observe that the block-orthogonal $\bar{\boldsymbol{O}}$ preserves the within-block norms of $\boldsymbol{v}$. Therefore, we define

$$\boldsymbol{w}_a := O_a^T \boldsymbol{v}_a, \qquad \text{(C4)}$$

for $a = 1, 2$, and perform integrals over arbitrary $\boldsymbol{w} = \begin{pmatrix} \boldsymbol{w_1} \\ \boldsymbol{w_2} \end{pmatrix}$ while enforcing norm constraints within blocks. We define the two-component vector $\boldsymbol{t}$:

$$t_a := \frac{1}{N_a} \|\boldsymbol{v}_a\|^2 \qquad \text{(C5)}$$

Then we can write

$$I^X(\boldsymbol{T}) = \frac{Z(\boldsymbol{t}, X)}{Z(\boldsymbol{t}, 0)}, \qquad \text{(C6)}$$

where we have defined

$$Z(\boldsymbol{t}, X) := \int \frac{d\boldsymbol{w}}{(2\pi)^{N/2}} \exp\left(\frac{1}{2}\boldsymbol{w}^T X \boldsymbol{w}\right)$$
$$\times \prod_{a=1,2} \delta(\|\boldsymbol{w}_a\|^2 - N_a t_a). \qquad (C7)$$

We calculate $Z(\boldsymbol{t}, X)$ by using the Fourier representation of the delta function, over the imaginary axis: $\delta(x) = \int_{-i\infty}^{+i\infty} \frac{\exp(-qx/2)}{4\pi i} dq$. This gives

$$Z(\boldsymbol{t}, X) := \int \left(\prod_{a=1,2} \frac{dq_a}{4\pi i} e^{\frac{1}{2}(N_a q_a t_a)}\right)$$
$$\times \int \frac{d\boldsymbol{w}}{(2\pi)^{N/2}} e^{-\frac{1}{2}\boldsymbol{w}^T(\bar{Q}-X)\boldsymbol{w}}, \qquad (C8)$$

where we have introduced the $N \times N$ diagonal matrix, $\bar{Q}$, which has q_1 on its first N_1 diagonal elements, and q_2 on the remaining N_2 elements.

The Gaussian integral over $\boldsymbol{w}$ now yields $\det(\bar{Q} - X)^{-1/2}$. Writing N_2 singular values of X as x_m (which includes $N_2 - N_1$ zeros in the case $N_2 > N_1$), we can write

$$\det(\bar{Q} - X) = q_1^{(N_1-N_2)} \prod_{m=1}^{N_2} (q_1 q_2 - x_m^2). \qquad (C9)$$

Thus, at this stage we have

$$Z(\boldsymbol{t}, X) := \int_{-i\infty}^{i\infty} \frac{dq_1 \, dq_2}{(4\pi i)^2} \exp\left[\frac{N}{2} F^X(\boldsymbol{t}, \boldsymbol{q})\right] \qquad (C10)$$

with

$$F^X(\boldsymbol{t}, \boldsymbol{q}) = \beta_1 q_1 t_1 + \beta_2 q_2 t_2 + (\beta_2 - \beta_1) \log q_1$$
$$- \frac{1}{N} \sum_{m=1}^{N_2} \log(q_1 q_2 - x_m^2). \qquad (C11)$$

To find the saddle point, we take partial derivatives with respect to q_1 and q_2, and find

$$t_1 = q_2^* g_{XX^T}(q_1^* q_2^*), \qquad (C12)$$

$$t_2 = q_1^* g_{X^T X}(q_1^* q_2^*). \qquad (C13)$$

For notational clarity, in this section we define the functional inverse of the block Stieltjes transform, $\mathcal{B}^X(\boldsymbol{t}) := (\boldsymbol{g}^X)^{-1}(\boldsymbol{t})$, satisfying

$$\mathcal{B}^X(\boldsymbol{g}^X(\boldsymbol{z})) = \boldsymbol{z}. \qquad (C14)$$

In the limit of large z_1, z_2, we have $g_1^X(\boldsymbol{z}) \approx \frac{1}{z_2}$ and $g_2^X(\boldsymbol{z}) \approx \frac{1}{z_1}$, and therefore for small t_1, t_2 we have $\mathcal{B}_1^X(\boldsymbol{t}) \approx \frac{1}{t_2}$ and $\mathcal{B}_2^X(\boldsymbol{t}) \approx \frac{1}{t_1}$. Generally, the functional inverse, $\mathcal{B}^X(\boldsymbol{t})$ exists for $\boldsymbol{t}$ with sufficiently small norm.

Thus, the saddle-point condition for $Z(\boldsymbol{t}, X)$ can be written succinctly as $\boldsymbol{q}^* = \mathcal{B}^X(\boldsymbol{t})$.

Finally, we find the asymptotic value of $I^X(\boldsymbol{v})$ (C10) by also solving the saddle point for $Z(\boldsymbol{t}, 0)$. For $X = 0$ we have $g_{XX^T} z = z^{-1}$, so that the saddle-point condition for $Z(\boldsymbol{t}, 0)$ is simply $q_a^* = t_a^{-1}$. This yields $F^0(\boldsymbol{t}, \boldsymbol{q}^*) = \sum_a \beta_a(1 + \log t_a)$.

We therefore arrive at our asymptotic approximation for the rank-1 block spherical integral:

$$I^X(\boldsymbol{T}) = \exp\frac{N}{2} H^X(\boldsymbol{t}), \qquad (C15)$$

where we have

$$H^X(\boldsymbol{t}) = \sum_{a=1,2} \beta_a[t_a \mathcal{B}_a^X(\boldsymbol{t}) - \log t_a - 1]$$
$$- \frac{1}{N} \log \det(\bar{\boldsymbol{\mathcal{B}}}^X(\boldsymbol{t}) - X), \qquad (C16)$$

where we have written $\bar{\boldsymbol{\mathcal{B}}}^X(\boldsymbol{t})$ to indicate the $N \times N$ diagonal matrix with $\mathcal{B}_1^X(\boldsymbol{t})$ along the top N_1 diagonal elements, and $\mathcal{B}_2^X(\boldsymbol{t})$ along the remaining N_2.

We observe an appealing relationship between the rank-1 block spherical integral and the block $\mathcal{R}$ transform. By the saddle-point conditions, the partial derivatives of $F^X(\boldsymbol{t}, \boldsymbol{q}^*)$ with respect to q_a are zero. Therefore the gradient of H^X with respect to $\boldsymbol{t}$ treats $\mathcal{B}^X$ as constant, and we have simply

$$\frac{dH^X(\boldsymbol{t})}{dt_a} = \beta_a\left(\mathcal{B}_a^X(\boldsymbol{t}) - \frac{1}{t_a}\right) = \beta_a \mathcal{R}_a^X(\boldsymbol{t}). \qquad (C17)$$

We therefore write $H^X(\boldsymbol{t})$ as a contour integral in $\mathbb{C}^2$:

$$H^X(\boldsymbol{t}) = \int_0^t d\boldsymbol{w} \cdot (\boldsymbol{\beta} \odot \mathcal{R}^X(\boldsymbol{t})), \qquad (C18)$$

where $\odot$ is the elementwise product and $\boldsymbol{\beta} = \binom{\beta_1}{\beta_2} = \frac{1}{N}\binom{N_1}{N_2} = \frac{1}{1+c}\binom{c}{1}$.

APPENDIX D: THE BLOCK $\mathcal{R}$ TRANSFORM OF GAUSSIAN NOISE

In this Appendix we calculate the block $\mathcal{R}$ transform for the $N_1 \times N_2$ (with $c = {}^{N_1}/_{N_2}$) matrix X with i.i.d. Gaussian elements: $X_{ij} \sim \mathcal{N}(0, \frac{\sigma^2}{N_2})$.

For notational clarity, here we write the functional inverse of the block Stieltjes transform for any rectangular matrix, A, as $\mathcal{B}^A(\boldsymbol{t}) := (\boldsymbol{g}^A)^{-1}(\boldsymbol{t})$, satisfying

$$\mathcal{B}^A(\boldsymbol{g}^A(\boldsymbol{z})) = \boldsymbol{z}. \qquad (D1)$$

Note that from the definition of $\boldsymbol{g}^A(\boldsymbol{z})$ one can find a relationship between the two elements of the inverse block Stieltjes transform:

$$t_2 \mathcal{B}_2^A(\boldsymbol{t}) = c t_1 \mathcal{B}_1^A(\boldsymbol{t}), \qquad (D2)$$

where c is the aspect ratio of A.

The block $\mathcal{R}$ transform is defined as

$$\mathcal{R}^A(\boldsymbol{t}) = \mathcal{B}^A(\boldsymbol{t}) - \frac{1}{\boldsymbol{t}}, \qquad (D3)$$

where the multiplicative inverse, $^1/_{\boldsymbol{t}}$, is elementwise.

To find $\mathcal{B}^X$, we observe that in general the product of the two elements of $\boldsymbol{g}^X(\boldsymbol{z})$ is a scalar function that depends only on the product of the elements of $\boldsymbol{z}$; that is, $g_1^X(\boldsymbol{z})g_2^X(\boldsymbol{z}) = z_1 z_2 g_{XX^T}(z_1 z_2) g_{X^T X}(z_1 z_2)$. Therefore, we define

$$\Lambda_X(z) := z g_{XX^T}(z) g_{X^T X}(z). \qquad (D4)$$

We can find $\mathcal{B}^X(t)$ by first inverting $\Lambda_X(z)$, and then

$$\mathcal{B}_1^X(t)\mathcal{B}_2^X(t) = \Lambda_X^{-1}(t_1 t_2). \tag{D5}$$

For the Gaussian matrix, X, we have

$$g_{XX^T}(z) = \frac{z + \sigma^2(1+c) - \sqrt{(z - x_+^2)(z - x_-^2)}}{2z\sigma^2} \tag{D6}$$

with

$$x_\pm = \sigma(1 \pm \sqrt{c}). \tag{D7}$$

From there we find

$$\Lambda_X(z) = \frac{z - \sigma^2(1+c) + \sqrt{(z - x_+^2)(z - x_-^2)}}{2c\sigma^4}. \tag{D8}$$

Some further algebra yields

$$\Lambda_X^{-1}(t) = c\sigma^4 t + \frac{1}{t} + \sigma^2(1+c). \tag{D9}$$

Thus, we have

$$\mathcal{B}_1^X(t)\mathcal{B}_2^X(t) = c\sigma^4 t_1 t_2 + \frac{1}{t_1 t_2} + \sigma^2(1+c). \tag{D10}$$

Using the general relationship between elements of $\mathcal{B}(t)$ (D2) yields a quadratic equation for $\mathcal{B}_1^X$. We choose the root that yields $\mathcal{B}_1^X \approx 1/t_1$ in the large t limit, and then use (D2) to find $\mathcal{B}_2^X$, finally arriving at

$$\mathcal{B}_1^X(t) = \frac{1}{t_1} + \sigma^2 t_2, \tag{D11}$$

$$\mathcal{B}_2^X(t) = \frac{1}{t_2} + c\sigma^2 t_1. \tag{D12}$$

Finally this yields for the block $\mathcal{R}$ transform:

$$\boldsymbol{R}^X(t) = \sigma^2\left(\frac{t_2}{ct_1}\right). \tag{D13}$$

As a side note we point out that from the relationship between elements of the block Stieltjes inverse (D2) it follows that

$$t_2 \mathcal{R}_2^A(t) = ct_1 \mathcal{R}_1^A(t), \tag{D14}$$

for all rectangular A with aspect ratio c.

APPENDIX E: BLOCK STIELTJES TRANSFORM AND SINGULAR VALUE DENSITY OF THE EXTENSIVE SPIKED MODEL

We report the quartic equation from Sec. IV C for the first element of the block Stieltjes transform of the $N_1 \times N_2$ exten-

sive spiked model, $R = sU_1U_2 + X$, with rank ratio $b = K/N_1$ and aspect ratio $c = N_1/N_2$. For notational simplicity, in this section we write $g := g_1^R(z)$. Multiplying out (46) yields a quartic:

$$Ag^4 + Bg^3 + Cg^2 + Dg + E = 0 \tag{E1}$$

with

$$A = -c^3, \tag{E2}$$

$$B = c^2\left(3z - 2\frac{1-c}{z}\right), \tag{E3}$$

$$C = c\left[-3z^2 - \left(\frac{1-c}{z}\right)^2 - 5c + 4 + s^2\right], \tag{E4}$$

$$D = z^3 + (4c - 2 - s^2)z + (1 - 2c + s^2)\frac{1-c}{z}, \tag{E5}$$

$$E = -z^2 - c + 1 + (1-b)s^2. \tag{E6}$$

In order to obtain the singular value density $\rho_1^R(\hat{s})$ numerically, we use the roots method of the NumPy polynomial class in Python (3.9.7), to solve with $z = \hat{s} - 10^{-7}i$, and select the root with the largest imaginary part. To our knowledge there is no guarantee that the root with largest imaginary part is the correct root, but we find this works in practice.

APPENDIX F: KERNEL ESTIMATES OF EMPIRICAL SPECTRAL DENSITIES

In order to employ our optimal estimators in empirical settings (Sec. V), we need to be able to estimate the block Stieltjes transform, $g_1^R(\hat{s})$, from data. Developing optimal algorithms to achieve this is left for future work, but here we use a technique inspired by [36] and [35] to demonstrate proof of principle. We use this kernel on the extensive spike model in Fig. 8(c).

Given N_1 data singular values, $\{\hat{s}_m\}$ (assuming $N_1 < N_2$ without loss of generality), we define a smoothed block Stieltjes transform:

$$\tilde{g}_1^R(z) := \frac{1}{N_1}\sum_{m=1}^{N_1}\frac{z}{z^2 - \hat{s}_m^2 - i\eta_m}, \tag{F1}$$

where η_m is a local bandwidth term given by

$$\eta_m = \frac{\hat{s}_m}{N^{1/2}}. \tag{F2}$$

[1] O. I. Rumyantsev, J. A. Lecoq, O. Hernandez, Y. Zhang, J. Savall, R. Chrapkiewicz, J. Li, H. Zeng, S. Ganguli, and M. J. Schnitzer, Fundamental bounds on the fidelity of sensory cortical coding, Nature (London) **580**, 100 (2020).

[2] A. M. Saxe, J. L. McClelland, and S. Ganguli, A mathematical theory of semantic development in deep neural networks, Proc. Natl. Acad. Sci. USA **116**, 11537 (2019).

[3] F. Luo, J. Zhong, Y. Yang, and J. Zhou, Application of random matrix theory to microarray data for discovering functional gene modules, Phys. Rev. E **73**, 031924 (2006).

[4] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, T. Guhr, and H. E. Stanley, Random matrix approach to cross correlations in financial data, Phys. Rev. E **65**, 066126 (2002).

[5] J. Pennington and P. Worah, Nonlinear random matrix theory for deep learning, J. Stat. Mech.: Theory Exp. (2019) 124005.

[6] J. Pennington, S. Schoenholz, and S. Ganguli, Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice, in *Advances in Neural Information Processing Systems* (Curran Associates Inc., Long Beach, CA, USA, 2017).

[7] J. Pennington, S. S. Schoenholz, and S. Ganguli, The emergence of spectral universality in deep networks, in *Artificial Intelligence and Statistics (AISTATS)* (PMLR, 2018).

[8] A. K. Lampinen and S. Ganguli, An analytic theory of generalization dynamics and transfer learning in deep linear networks, in *International Conference on Learning Representations (ICLR)* (ICLR, 2018).

[9] C. H. Martin and M. W. Mahoney, Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning, J. Mach. Learn. Res. **22**, 7479 (2021).

[10] A. Wei, W. Hu, and J. Steinhardt, More than a toy: Random matrix models predict how real-world neural representations generalize, in *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 162 (PMLR, 2022), pp. 23549–23588.

[11] M. S. Santhanam and P. K. Patra, Statistics of atmospheric correlations, Phys. Rev. E **64**, 016102 (2001).

[12] E. F. Santos, A. L. Barbosa, and P. J. Duarte-Neto, Global correlation matrix spectra of the surface temperature of the oceans from random matrix theory to Poisson fluctuations, Phys. Lett. A **384**, 126689 (2020).

[13] A. M. Tulino and S. Verdú, Random matrix theory and wireless communications, Found. Trends Commun. Inf. Theory **1**, 1 (2004).

[14] D. Zhu, B. Wang, H. Ma, and H. Wang, Evaluating the vulnerability of integrated electricity-heat-gas systems based on the high-dimensional random matrix theory, CSEE J. Power Energy Syst. **6**, 878 (2019).

[15] J. Mosso, D. Simicic, K. Simsek, R. Kreis, C. Cudalbu, and I. O. Jelescu, MP-PCA denoising for diffusion MRS data: Promises and pitfalls, NeuroImage **263**, 119634 (2022).

[16] W. T. Clarke and M. Chiew, Uncertainty in denoising of MRSI using low-rank methods, Magn. Reson. Med. **87**, 574 (2022).

[17] C. M. W. Tax, M. Bastiani, J. Veraart, E. Garyfallidis, and M. Okan Irfanoglu, What's new and what's next in diffusion MRI preprocessing, NeuroImage **249**, 118830 (2022).

[18] R. Bansal and B. S. Peterson, Use of random matrix theory in the discovery of resting state brain networks, Magn. Reson. Imaging **77**, 69 (2021).

[19] W. Zhu, X. Ma, X. H. Zhu, K. Ugurbil, W. Chen, and X. Wu, Denoise functional magnetic resonance imaging with random matrix theory based principal component analysis, IEEE Trans. Biomed. Eng. **69**, 3377 (2022).

[20] J. Baik, G. B. Arous, and S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, Ann. Probab. **33**, 1643 (2005).

[21] P. Loubaton and P. Vallet, Almost sure localization of the eigenvalues in a Gaussian information plus noise model—Application to the spiked models, Electron. J. Probab. **16**, 1934 (2011).

[22] 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**, 120 (2012).

[23] A. A. Shabalin and A. B. Nobel, Reconstruction of a low-rank matrix in the presence of Gaussian noise, J. Multivariate Anal. **118**, 67 (2013).

[24] A. El Alaoui and M. I. Jordan, Detection limits in the high-dimensional spiked rectangular model, in *Proceedings of the 31st Conference On Learning Theory*, Proceedings of Machine Learning Research, Vol. 75 (PMLR, 2018), pp. 410–438.

[25] J. Barbier, N. Macris, and C. Rush, All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation, *Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2020), Vol. 33, pp. 14915–14926.

[26] B. Aubin, B. Loureiro, A. Maillard, F. Krzakala, and L. Zdeborová, The spiked matrix model with generative priors, in *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., 2019).

[27] 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. Am. Stat. Assoc. **118**, 374 (2021).

[28] M. Gavish and D. L. Donoho, The optimal hard threshold for singular values is $4\sqrt{3}$, IEEE Trans. Inf. Theory **60**, 5040 (2014).

[29] M. Gavish and D. L. Donoho, Optimal shrinkage of singular values, IEEE Trans. Inf. Theory **63**, 2137 (2017).

[30] R. Allez and J.-P. Bouchaud, Eigenvector dynamics under free addition, Random Matrices Theory Appl. **03**, 1450010 (2014).

[31] O. Ledoit and S. Péché, Eigenvectors of some large sample covariance matrix ensembles, Probab. Theory Relat. Fields **151**, 233 (2011).

[32] O. Ledoit and M. Wolf, Nonlinear shrinkage estimation of large-dimensional covariance matrices, Ann. Stat. **40**, 1024 (2012).

[33] J. Bun, R. Allez, J.-P. Bouchaud, and M. Potters, Rotationally invariant estimator for general noisy matrices, IEEE Trans. Inf. Theory **62**, 7475 (2016).

[34] J. Bun, J.-P. Bouchaud, and M. Potters, Cleaning large correlation matrices: Tools from random matrix theory, Phys. Rep. **666**, 1 (2017).

[35] O. Ledoit and M. Wolf, Analytical nonlinear shrinkage of large-dimensional covariance matrices, Ann. Stat. **48**, 3043 (2020).

[36] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, 2020).

[37] F. Benaych-Georges, Rectangular random matrices, related convolution, Probab. Theory Relat. Fields **144**, 471 (2009).

[38] R. Speicher, C. Vargas, and T. Mai, Free deterministic equivalents, rectangular random matrix models, and operator-valued free probability theory, arXiv:1110.1237.

[39] F. Benaych-Georges, Rectangular R-transform as the limit of rectangular spherical integrals, J. Theor. Probab. **24**, 969 (2011).

[40] J. A. Mingo and R. Speicher, *Free Probability and Random Matrices, Fields Institute Monographs*, Vol. 35 (Springer, 2017).

[41] E. Troiani, V. Erba, F. Krzakala, A. Maillard, and L. Zdeborová, Optimal denoising of rotationally invariant rectangular matrices, *Mathematical and Scientific Machine Learning* (PMLR, 2022), pp. 97–112.

[42] J. Barbier and N. Macris, Statistical limits of dictionary learning: Random matrix theory and the spectral replica method, Phys. Rev. E **106**, 024136 (2022).

[43] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising, J. Stat. Mech. (2022) 083301.

[44] F. Pourkamali and N. Macris, Rectangular rotational invariant estimator for general additive noise matrices, arXiv:2304.12264.

[45] Z. Bai and J.-F. Yao, Central limit theorems for eigenvalues in a spiked population model, Ann. Inst. H. Poincaré Probab. Statist. **44**, 447 (2008).

[46] D. Donoho, M. Gavish, and I. Johnstone, Optimal shrinkage of eigenvalues in the spiked covariance model, Ann. Stat. **46**, 1742 (2018).

[47] A. Takemura, An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population, Tsukuba J. Math. **8**, 367 (1984).

[48] C. Stein, Lectures on the theory of estimation of many parameters, J. Sov. Math. **34**, 1373 (1986).

[49] B. Collins, Moments and cumulants of polynomial random variables on unitarygroups, the Itzykson-Zuber integral, and free probability, Int. Math. Res. Notices **2003**, 953 (2003).

[50] P. Fleig and I. Nemenman, Statistical properties of large data sets with linear latent features, Phys. Rev. E **106**, 014102 (2022).

[51] E. S. Meckes, *The Random Matrix Theory of the Classical Compact Groups*, Cambridge Tracts in Mathematics (Cambridge University Press, Cambridge, 2019).