


Statistical properties of large data sets with linear latent features

Philipp Fleig ^{*}

Department of Physics & Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Ilya Nemenman 

Department of Physics, Emory University, Atlanta, Georgia 30322, USA;

Department of Biology, Emory University, Atlanta, Georgia 30322, USA;

and Initiative in Theory and Modeling of Living Systems, Atlanta, Georgia 30322, USA



(Received 16 January 2022; revised 4 May 2022; accepted 23 May 2022; published 5 July 2022)

Analytical understanding of how low-dimensional latent features reveal themselves in large-dimensional data is still lacking. We study this by defining a probabilistic linear latent features model with additive noise and by analytically and numerically computing the statistical distributions of pairwise correlations and eigenvalues of the data correlation matrix. This allows us to resolve the latent feature structure across a wide range of data regimes set by the number of recorded variables, observations, latent features, and the signal-to-noise ratio. We find a characteristic imprint of latent features in the distribution of correlations and eigenvalues and provide an analytic estimate for the boundary between signal and noise, even in the absence of a spectral gap.

DOI: [10.1103/PhysRevE.106.014102](https://doi.org/10.1103/PhysRevE.106.014102)

I. INTRODUCTION

Massively parallel experiments are now standard in science. They record the state of the system through $N \gg 1$ variables x_i , $i = 1 \dots N$. These variables could be positions of particles, agents, or tracers in dusty plasmas [1], soft matter [2], insect swarming [3], and dynamical systems [4]. They can be field values at different spatial points in fluids [5], climate data [6], or activity of “nodes” in gene expression networks [7], neural recordings [8], postures [9,10], biodiversity [11], ecology [12], etc. Crucially, the number of recorded variables is often larger than the number of true (latent) degrees of freedom in the system [13–15]. This allows to use correlations among the measured variables to detect the latent ones and to use the latter in models of the system [13,16–18].

Physical systems are nonstationary, which limits the number of times T that they can be realistically measured, and one often has $T \sim N$. Hence, statistical fluctuations are large, and in the limit $T \leq N$ the data correlation matrix even becomes degenerate. Once latent features are extracted from such undersampled data (typically by using the Principal Components Analysis method [19]), one uses the Random Matrix Theory [20], and, in particular, the Marčenko-Pastur (MP) eigenvalue density of a pure noise correlation matrix [21] to identify those features which can be trusted. Specifically, one calculates the upper and the lower bounds of eigenvalues expected by pure chance from T measurements of N independent variables, and only eigenvalues outside this interval (and their eigenvectors) are deemed statistically significant.

This approach assumes that signal-induced correlations among the variables do not influence the spectrum of noise-induced correlations. This has never been proven and, as

we will show, is, in fact, incorrect. More generally, we are not aware of results to produce the eigenvalue density of the correlation matrix when the correlations come from the sampling noise *and* from true low-dimensional latent signals with the latter having a known distribution (although see work on spiked covariance matrix models [22–25]). Even statistics of the entries of the correlation matrix (rather than of its eigenvalues) have not been reported in this case. In this paper, we close these gaps and calculate—numerically and analytically, using the Random Matrix Theory methods—statistical properties of correlation matrices for data sets with low-dimensional latent features structure. We show that the distribution of pairwise correlations and the spectra of their eigenvalues carry signatures of the number of latent features, allowing one not only to choose rigorously, which of the principal components are above the noise floor, but also to see if the overall model of latent features plus noise is accurate for a data set.

We analyze two commonly occurring limits. First is the *classical statistics* limit, where the number of latent features m can be of similar size as N , whereas both N and m are much smaller than the number of observations T . Second is the *intensive* limit, where the ratio of number of variables to observations N/T is finite, whereas m is much smaller than both N and T . We leave the *extensive* limit where m grows with N for later work. We believe that our results are an important step in the development of analytical tools for understanding large data sets.

II. THE MODEL AND ITS LIMITS

We consider observations produced by a probabilistic matrix model combining stochastic latent signals and uncorrelated noise,

$$\mathbf{X} = \mathbf{UV} + \sigma\mathbf{R}. \quad (1)$$

^{*}Corresponding author: fleig@sas.upenn.edu

The component matrices \mathbf{U} and \mathbf{V} have dimensions $T \times m$ and $m \times N$, respectively. Thus, m latent features get randomly sampled T times (matrix \mathbf{U}), and each of the N measured variables is a random linear combination of the latent features (matrix \mathbf{V}). We assume $m \leq T, N$ throughout this paper such that the rank of the signal matrix \mathbf{UV} is equal to m , and the features can be estimated from the data. Such factor models are commonly used to model latent structures in real data [26,27].

We choose that the entries of \mathbf{U} and \mathbf{V} are Gaussian random variables with zero mean and variances σ_U^2 and σ_V^2 , respectively,

$$U_{t\mu} \sim \mathcal{N}(0, \sigma_U^2), \quad V_{\mu n} \sim \mathcal{N}(0, \sigma_V^2), \quad (2)$$

$$t = 1, \dots, T, \quad \mu = 1, \dots, m, \quad n = 1, \dots, N. \quad (3)$$

We make this choice for analytic tractability. However, we note that many random matrix results are universal and hold to a certain degree independent of the specific distribution of data [20]. Whereas we do not know for sure, we expect most results that we present here to hold similarly for non-Gaussian (but finite variance) distributions of \mathbf{U} and \mathbf{V} . Note also that in applications to real data the means and variances of the entries may need to be matched to those of the measured variables. Finally, the elements of the noise matrix \mathbf{R} are independent and identically distributed (i.i.d.) unit variance Gaussian random variables, so that the noise in every observation has variance σ^2 .

From Eqs. (1) and (2), the elements of the signal matrix \mathbf{UV} are a sum of m products of two Gaussian variables with variances σ_U^2 and σ_V^2 . In Appendix A, we derive the probability density of these entries and show that their variance is

$$\sigma_{UV}^2 = m\sigma_U^2\sigma_V^2. \quad (4)$$

As expected from addition of independent random variables, each latent component adds $\sigma_U^2\sigma_V^2$ to the variance of the observations.

Furthermore, for $m \gg 1$, the probability density of \mathbf{X} approaches a Gaussian with zero mean, see Fig. 4. This allows us to define a Gaussian signal-to-noise (SNR) ratio $\text{SNR} \equiv \sigma_{UV}^2/\sigma^2 = m\sigma_U^2\sigma_V^2/\sigma^2$. Since our goal is to calculate properties of the data matrix independent of the units of each variable, we normalize the data matrix,

$$\tilde{\mathbf{X}} \equiv \mathbf{X}/\sigma_X, \quad \sigma_X^2 \equiv \sigma_{UV}^2 + \sigma^2 = \sigma^2(1 + \text{SNR}). \quad (5)$$

This normalization by an expected standard deviation is different from subtracting empirical means and standardizing by an empirical standard deviation. However, we expect the difference to be $\sim T^{-1/2}$ and, thus, negligible in what follows.

We now focus on the normalized empirical covariance matrix (NECM),

$$\mathbf{C} = \frac{1}{T} \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} = \frac{1}{T} (\tilde{\mathbf{UV}})^T (\tilde{\mathbf{UV}}) + \tilde{\sigma}^2 \mathbf{R}^T \mathbf{R} + \tilde{\sigma} (\tilde{\mathbf{UV}})^T \mathbf{R} + \tilde{\sigma} \mathbf{R}^T \tilde{\mathbf{UV}}, \quad (6)$$

as well as the matrix of correlation coefficients,

$$c_{pq} = \frac{C_{pq}}{\sqrt{C_{pp}C_{qq}}}. \quad (7)$$

To explore different regimes of the problem, we define

$$q \equiv N/T, \quad q_U \equiv m/T, \quad \text{and} \quad q_{VT} \equiv m/N. \quad (8)$$

Only two of these parameters are independent. These parameters emerge naturally in our theoretical analysis Appendix C 1. Then the classical statistics and the intensive limits introduced above become

$$\text{Classical stats.: } q \rightarrow 0, \quad q_U \rightarrow 0, \quad q_{VT} = \text{const}, \quad (9)$$

$$\text{Intensive: } q = \text{const}, \quad q_U \rightarrow 0, \quad q_{VT} \rightarrow 0, \quad (10)$$

together with $\text{SNR} = \text{const}$ in both limits. Notably, q_U^{-1} gives the number of observations available per latent feature to be learned. Since the parameter q_U is small in both limits, the latent features are sampled well, even if the measured variables (controlled by q) may not be. Note that many modern neuroscience (and other empirical) data sets and their models fall into one of these limits [13–15,28].

III. DENSITY OF PAIRWISE CORRELATIONS

The first observable statistics we calculate is the density of correlations c_{pq} of the standardized variables $\tilde{\mathbf{X}}$. Our goal is to analyze the dependence of the density of the matrix entries c_{pq} on m , T , and the noise strength. The numerator in the correlation matrix in Eq. (7) has three contributions: $(\mathbf{UV})^T (\mathbf{UV})$ from the pure latent features signal, $\mathbf{R}^T \mathbf{R}$ from the pure noise, and two cross terms between the signal and the noise, e.g., $(\mathbf{UV})^T \mathbf{R}$. Each term is analyzed separately in Appendix B and reduced to correlations between independent Gaussian vectors. Such correlations are distributed according to the symmetric Beta distribution [29],

$$\text{pdf}(r) = \text{Beta}(r; \alpha, \alpha; \ell = -1; s = 2), \quad (11)$$

where (pdf) is the probability density function, the location ℓ and scale s of the Beta distribution are set such that correlations fall on the interval $[-1, 1]$. The shape parameter α is determined individually for the signal-signal, noise-noise, and signal-noise contributions,

$$\alpha_s = \frac{m-1}{2}, \quad \alpha_n = \frac{T-1}{2}, \quad \alpha_{\text{sn}} = \frac{m^{1/2}T^{1/2} - 1}{2}. \quad (12)$$

The sum of Beta distributions is well approximated by a single Beta distribution [30]. This allows us to approximate the distribution of entries of the full correlation matrix by a single Beta distribution of the form Eq. (11). In Appendix B, we show that, in the limit when contributions of $O(T^{-1/2})$ and $O(m^{-1/2})$ can be neglected, the parameter of the approximated distribution is

$$\alpha \approx \frac{[\sqrt{m(1+1/\text{SNR})}^{-1} + \sqrt{T(1+\text{SNR})}^{-1}]^{-2} - 1}{2}. \quad (13)$$

Notably, the shape of the distribution depends on m through α . Thus, the number of latent dimensions can be estimated from the empirical distribution of the correlation coefficients.

Numerical validation of this result in the pure signal limit $\text{SNR} \rightarrow \infty$ is shown in Fig. 1. For small m , the distribution distinctly changes shape as m varies. When m becomes comparable to N , i.e., $q_{VT} \rightarrow 1$, the distribution approaches a Gaussian, making it difficult to infer the precise value of

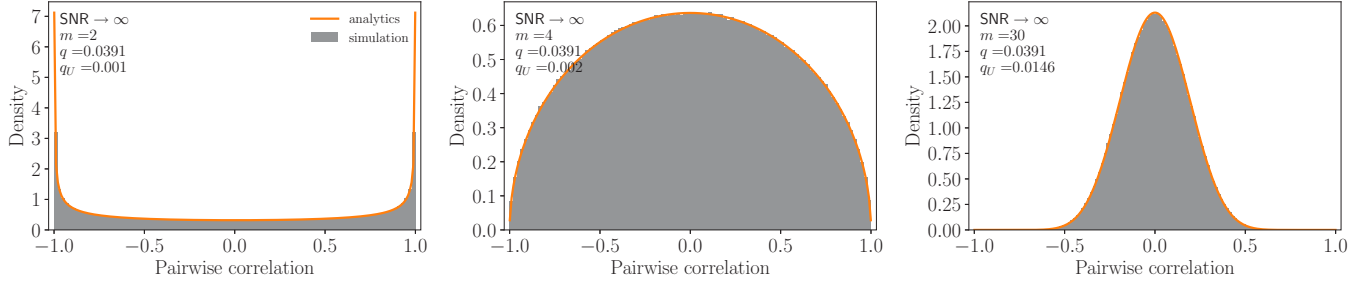


FIG. 1. Distribution of pairwise correlations in the pure signal limit $\text{SNR} \rightarrow \infty$, for $m = 2, 4$, and 30 latent features. Orange: analytical form (a symmetric Beta distribution). Gray: simulated data. Each simulation has $N = 80$ variables and $T = 2048$ observations and constitutes 1000 independent model realizations.

m from its shape. The quality of the analytic approximation increases with smaller q_U , i.e., when more observations per latent feature are available. The analytic approximation is also exact in the large noise limit $\text{SNR} \rightarrow 0$. However, deviations appear for finite SNR when m is small, see Fig. 5.

IV. EIGENVALUE DENSITY

We compute the eigenvalue density of NECM \mathbf{C} , cf. Eq. (6) from its Stieltjes transform $g_N(z) = N^{-1} \text{Tr}(z\mathbf{I} - \mathbf{C})^{-1}$, where z is a complex number. We denote the large- N limit of g_N by g_C [20]. The eigenvalue density is obtained from the Sokhotski-Plemelj formula,

$$\rho(\lambda) = \frac{1}{\pi} \lim_{\eta \rightarrow 0^+} \text{Im } g_C(z = \lambda - i\eta), \quad (14)$$

where Im denotes the imaginary part.

Full details of the computations are in Appendix C. Briefly, we consider the eigenvalue density in the classical and intensive limits, Eqs. (9) and (10). This allows us to simplify the calculations, neglecting cross terms between the signal $\mathbf{U}\mathbf{V}$ and the noise \mathbf{R} , cf. Appendix C 3. Then, in the *classical statistics* limit, the Stieltjes transform satisfies the third order polynomial equation,

$$ag_C^3 + bg_C^2 + cg_C + d = 0, \quad (15)$$

with

$$a = \frac{qz}{1 + \text{SNR}}, \quad (16)$$

$$b = -\frac{qq_{V^T}z}{\text{SNR}} + \frac{(q_{V^T} - 1)q + 1}{1 + \text{SNR}} - z, \quad (17)$$

$$c = \frac{(q - 1)q_{V^T}}{\text{SNR}} + q_{V^T}z(1 + \text{SNR}^{-1}) - q_{V^T} + 1, \quad (18)$$

$$d = -q_{V^T}(1 + \text{SNR}^{-1}). \quad (19)$$

Whereas one can solve this cubic equation analytically, the results are unwieldy, allowing for little insight. Instead we rely on analytic approximations in the two noise limits as well as on numerics. Taking the zero noise limit, $\text{SNR} \rightarrow \infty$, the equation reduces to a quadratic polynomial, which we solve and evaluate Eq. (14) to find the eigenvalue density,

$$\rho^\infty(\lambda) = \frac{\sqrt{(\lambda - \lambda_-^\infty)(\lambda_+^\infty - \lambda)}}{2\pi\lambda\bar{\sigma}_X^{-2}q_{V^T}^{-1}} + (1 - q_{V^T})\delta(\lambda), \quad (20)$$

where $\bar{\sigma}_X^2 \equiv \sigma_X^2/m = \sigma_U^2\sigma_V^2$. The Dirac- δ function represents the $N - m$ eigenvalues of the NECM that are trivially zero. The m nontrivial eigenvalues lie in a finite interval with bounds,

$$\lambda_\pm^\infty = \bar{\sigma}_X^{-2}(1 \pm \sqrt{q_{V^T}})^2. \quad (21)$$

The density vanishes everywhere else. For finite SNR we solve Eq. (15) numerically. A comparison of eigenvalue densities in the different noise regimes, including the MP density [21] representing the pure noise limit $\text{SNR} \rightarrow 0$ is shown in Fig. 2(a), top.

In Appendix C 4 a, we derive an analytic approximation for the eigenvalue bounds λ_\pm^{SNR} at finite SNR, given by a weighted average of the zero noise bounds λ_\pm^∞ and those of the MP density $\lambda_\pm^{\text{MP}} = (1 \pm \sqrt{q})^2$ [31],

$$\lambda_\pm^{\text{SNR}} \approx (1 + \text{SNR}^{-1})^{-1}\lambda_\pm^\infty + (1 + \text{SNR})^{-1}\lambda_\pm^{\text{MP}}. \quad (22)$$

In Fig. 2(a), bottom, we compare this approximation and the true numerically computed bounds at different values of the SNR. The approximation is good everywhere with the largest deviation at $\text{SNR} \sim 10^{-1}$. The part of the eigenvalue spectrum associated with the pure latent feature signal lies outside of the interval $(1 + \text{SNR})^{-1}[\lambda_-^{\text{MP}}, \lambda_+^{\text{MP}}]$. Eigenvalues within this interval, shown as a striped band, correspond to noise. When the SNR is increased the noise range is shifted to the left compared to the MP range due to the presence of the latent features signal renormalizing the NECM. Thus, using the naïve MP bounds to reject eigenvalues as noise—a common procedure in data analysis—is incorrect.

A different picture emerges in the *intensive* limit. The equation for g_C is a lengthy sixth order polynomial, cf. Eq. (C102). For the $\text{SNR} \rightarrow \infty$ limit, we find the following analytic expression for the density:

$$\rho^\infty(\lambda) = \frac{\sqrt{(\lambda - \lambda_-^\infty)(\lambda_+^\infty - \lambda)}}{2\pi\lambda\bar{\sigma}_X^{-2}(1 + q)q_{V^T}^{-1}} + (1 - q_{V^T})\delta(\lambda), \quad (23)$$

with the eigenvalue bounds,

$$\lambda_\pm^\infty = \bar{\sigma}_X^{-2}(\sqrt{1 + q} \pm \sqrt{q_{V^T}})^2. \quad (24)$$

The eigenvalue density at different levels of noise is shown in Fig. 2(b), top. For large SNR, there is a gap in the density between the eigenvalues corresponding to noise and those corresponding to signal. The gap closes at lower SNR, and the combined density converges to the MP density for $\text{SNR} \rightarrow 0$. In the intensive limit, the approximate expression Eq. (22)

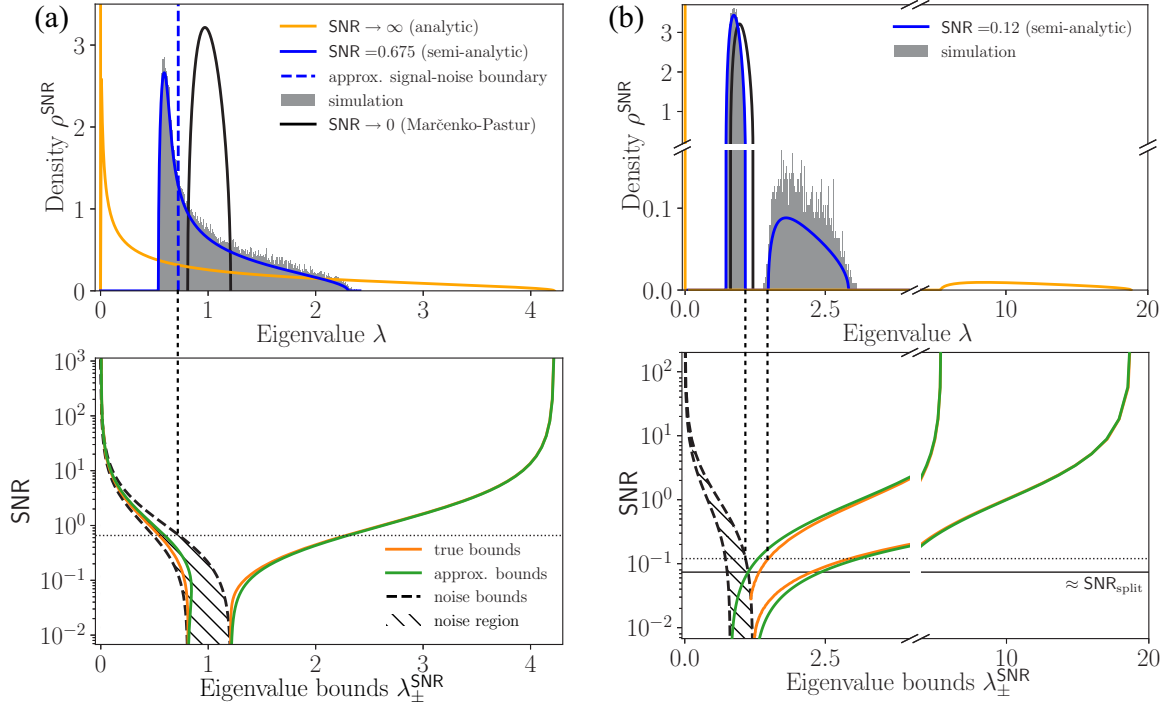


FIG. 2. Eigenvalue density and bounds as a function of SNR in the classical statistics and the intensive limit. (a) *Classical statistics* limit ($q = 0.01$, $q_{vT} = 0.9$) at three levels of SNR. Top plot: zero noise analytic density, Eq. (20), (yellow); large noise limit given by the Marčenko-Pastur distribution (black); and intermediate noise semianalytic density (blue) with the numerical simulation (gray) for comparison. Vertical dashed blue line is the approximate boundary between noise (left) and signal (right). Bottom: eigenvalue bounds $\lambda_{\pm}^{\text{SNR}}$ as a function of the SNR. True bounds obtained as numerical solution of Eq. (15) (orange) and approximate bounds given by the analytic expression in Eq. (22) (green). The approximate noise region is striped. Horizontal dotted line indicates the SNR value of the blue density curve, and the dashed line connecting the plots indicates the signal-noise boundary. (b) *Intensive* limit ($q = 0.01$, $q_{vT} = 0.09$) with plots analogous to (a). Top: zero noise density, Eq. (23) (yellow); for intermediate noise (blue), the left bump corresponds to noise, and the right bump corresponds to the latent signal. Bottom: for $\text{SNR} \gtrsim \text{SNR}_{\text{split}}$ (horizontal solid black line), the density splits into two bumps. Simulations constitute 360 independent realisations with $N = 300$ and $\sigma_U^2 = \sigma_V^2 = 1$.

for the bounds of the signal part of the density at finite SNR is good, showing the largest deviation for $\text{SNR} \sim 10^{-1}$, cf. Fig. 2(b), bottom. In Appendix C 4 b, we estimate that the eigenvalue density splits at $\text{SNR}_{\text{split}} \approx (\lambda_+^{\text{MP}} - \lambda_-^{\text{MP}})/\lambda_{\infty}$. The gap appears because having more observed variables (and, hence, more data to characterize the latent components) makes it easier to distinguish the signal from the noise. Thus, even if individual variables cannot be sampled well when $T < N$, high-throughput data sets still have value if $N \gg m$.

V. DISCUSSION

We calculated statistical properties of data with latent linear features, including the density of pairwise correlations and the density of eigenvalues of the NECM. We identify two important insights. First, by looking at the distribution of the correlations and their eigenvalues, one can understand whether the latent features model is a reasonable model for the data at hand. Second, whereas this is not the main focus of our paper, our results also allow to estimate the number of latent features from the statistics of the data (in the future, it might be interesting to relate them to other related mathematical results and practical algorithms [32,33]). Importantly, even if the eigenvalue density does not have a prominent gap, one can understand that the underlying model has a latent structure,

which manifests as a distortion of the MP sea of eigenvalues. This is because our signal matrix in Eq. (1) is stochastic, in contrast to spiked covariance models, where deterministic perturbations appear as δ functions in the spectrum and are detectable as true outliers [22–25]. Second, since the spectrum of noise correlations in our model is shifted compared to the MP model, one should not use a simple cutoff at the right edge of the MP density to distinguish statistically significant principal components.

A lot of ink has been expended discussing relative advantages and disadvantages of measuring a few variables well many times (biophysical approach) over measuring many variables infrequently and with high noise (high-throughput biology) [34,35]. We find that in the intensive limit the many measured variables lead to separation of the noise and the signal eigenvalues, resulting in a potentially more accurate inference. However, the number of observations and their quality contribute to the SNR, a high value of which is also required for the opening of the signal-noise gap. Thus, the quality and the quantity of measurements all contribute to the value of data in very specific ways, which we now understand.

Our analysis involved approximations. The first was in neglecting the signal-noise contributions in the computation of the eigenvalue density. Including these contributions would make the polynomial equation for the Stieltjes transform

substantially more complicated. However, we do not expect this leading to significant qualitative changes to the structure of the eigenvalue density in the limits considered. An additional limitation is that the approximation for the bounds of the eigenvalue density Eq. (22) is strictly only valid in the extreme limits $\text{SNR} \rightarrow 0$ and $\text{SNR} \rightarrow \infty$. In deriving the analytic density of correlations, we assumed $q_U \rightarrow 0$ in accordance with Eqs. (9) and (10) and worked to the leading order in T . We expect the fits to improve if these assumptions are removed, in particular, in the regime of finite SNR, cf. Fig. 5.

Finally, to connect our results with real data, additional steps are required. For example, methods to estimate the SNR from the data and to determine whether the Gaussian assumption for the distribution of the noise and the latent components is valid will need to be developed. Furthermore, for our model and its extensions, it is also important to calculate the expected overlap of empirical eigenvectors with their true values.

ACKNOWLEDGMENTS

We thank M. Potters for insightful comments. I.N. thanks the Aspen Center for Physics, partially funded by NSF Grant No. PHY-1607611 for hospitality. P.F. thanks M. Kramar for discussions and support. This work was supported, in part, by the Simons Foundation Grants No. 400425 (P.F.), No. 827661 (I.N.) and by NSF Grants No. BCS-1822677, No. PHY-2014173, and No. PHY-2010524 (I.N.).

APPENDIX A: DATA DISTRIBUTION FOR THE LATENT FEATURE MODEL WITH NO NOISE, ITS VARIANCE, AND LARGE- m LIMIT

Each entry X_{ij} of the latent features data matrix \mathbf{UV} is given by the sum of m products of two i.i.d. Gaussian random variables $u \sim \mathcal{N}(0, \sigma_U^2)$ and $v \sim \mathcal{N}(0, \sigma_V^2)$,

$$X_{ij} \sim \sum_{\mu=1}^m uv. \quad (\text{A1})$$

The product $x = uv$ is distributed according to the normal product distribution [36],

$$x \sim \frac{K_0\left(\frac{|x|}{\sigma_U \sigma_V}\right)}{\pi \sigma_U \sigma_V}, \quad (\text{A2})$$

where K_ν is the modified Bessel function of the second kind,

$$K_\nu(x) = \frac{\Gamma(\nu + \frac{1}{2})(2x)^\nu}{\sqrt{\pi}} \int_0^\infty dq \frac{\cos(q)}{(x^2 + q^2)^{\nu+1/2}}. \quad (\text{A3})$$

To derive the probability density of the latent feature model entries X_{ij} , we first compute the characteristic function φ_x by taking the Fourier transform of the normal product distribution. We then use the fact that the characteristic function φ_X of the sum of m products x is given by $\varphi_X = (\varphi_x)^m$. The inverse Fourier transform of φ_X then yields the sought after probability density.

Specifically, the characteristic function φ_x of the normal product distribution is

$$\begin{aligned} \varphi_x(t) &= \mathbb{E}(e^{itx}) = \int_{-\infty}^{\infty} dx \frac{K_0\left(\frac{|x|}{\sigma_U \sigma_V}\right)}{\pi \sigma_U \sigma_V} e^{itx} \\ &= \int_{-\infty}^{\infty} dx \frac{K_0(|x|)}{\pi} e^{it\sigma_U \sigma_V x} \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} dx \int_0^\infty dq \frac{\cos(q)}{\sqrt{|x|^2 + q^2}} e^{it\sigma_U \sigma_V x} \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} dx \int_0^\infty dq \frac{\cos(xq)}{\sqrt{1 + q^2}} e^{it\sigma_U \sigma_V x} \\ &= \frac{1}{\pi} \int_0^\infty dq \frac{1}{\sqrt{1 + q^2}} \\ &\quad \times \int_{-\infty}^{\infty} \frac{dx}{2\pi} (e^{ix(q + \sigma_U \sigma_V t)} + e^{ix(\sigma_U \sigma_V t - q)}) \\ &= \int_0^\infty dq \frac{1}{\sqrt{1 + q^2}} [\delta(\sigma_U \sigma_V t + q) + \delta(\sigma_U \sigma_V t - q)] \\ &= \frac{1}{\sqrt{1 + \sigma_U^2 \sigma_V^2 t^2}} \end{aligned} \quad (\text{A4})$$

for $t \in \mathbb{R} \setminus \{0\}$, and $\delta(\cdot)$ is the Dirac- δ function.

The characteristic function φ_X of the sum of m products x is given by

$$\varphi_X = (\varphi_x)^m = (1 + \sigma_U^2 \sigma_V^2 t^2)^{-(m/2)}. \quad (\text{A5})$$

Finally, performing the inverse transformation we obtain the probability density function of the sum,

$$\begin{aligned} \text{pdf}(X) &= \int_{-\infty}^{\infty} \frac{dt}{2\pi} e^{-itX} \frac{1}{(1 + \sigma_U^2 \sigma_V^2 t^2)^{m/2}} \\ &= \int_{-\infty}^{\infty} \frac{dt}{2\pi} e^{-itX} \\ &\quad \times \int_0^\infty dq \frac{\delta(\sigma_U \sigma_V t + q) + \delta(\sigma_U \sigma_V t - q)}{(1 + q^2)^{m/2}} \\ &= \frac{1}{\sigma_U \sigma_V} \int_0^\infty dq \\ &\quad \times \int_{-\infty}^{\infty} \frac{dt}{2\pi} e^{-(itX/\sigma_U \sigma_V)} \frac{\delta(t + q) + \delta(t - q)}{(1 + q^2)^{m/2}} \\ &= \frac{1}{\pi \sigma_U \sigma_V} \int_0^\infty dq \frac{\cos\left(\frac{q|X|}{\sigma_U \sigma_V}\right)}{(1 + q^2)^{m/2}} \\ &= \left[\frac{|X|}{\sigma_U \sigma_V}\right]^{m-1} \frac{1}{\pi \sigma_U \sigma_V} \int_0^\infty dq \frac{\cos(q)}{\left(\frac{|X|^2}{\sigma_U^2 \sigma_V^2} + q^2\right)^{m/2}} \\ &= \left[\frac{|X|}{2}\right]^{(m-1)/2} \frac{K_{(m-1)/2}\left(\frac{|X|}{\sigma_U \sigma_V}\right)}{(\sigma_U \sigma_V)^{(m+1)/2} \sqrt{\pi} \Gamma\left(\frac{m}{2}\right)}. \end{aligned} \quad (\text{A6})$$

Since the probability density function of X is symmetric around zero, the mean of the distribution vanishes

$$\mu_{UV} = \int_{-\infty}^{\infty} dX X \text{pdf}(X) = 0. \quad (\text{A7})$$

The variance is

$$\begin{aligned}
 \sigma_{UV}^2 &= \int_{-\infty}^{\infty} dX X^2 \text{pdf}(X) \\
 &= \frac{1}{2^{(m-1)/2} \sqrt{\pi} \Gamma\left(\frac{m}{2}\right)} \\
 &\quad \times \int_{-\infty}^{\infty} \frac{dX}{\sigma_U \sigma_V} \left[\frac{|X|}{\sigma_U \sigma_V} \right]^{(m-1)/2} |X|^2 K_{(m-1)/2}\left(\frac{|X|}{\sigma_U \sigma_V}\right) \\
 &= \frac{2\sigma_U^2 \sigma_V^2}{2^{(m-1)/2} \sqrt{\pi} \Gamma\left(\frac{m}{2}\right)} \int_0^{\infty} dX |X|^{(m+3)/2} K_{(m-1)/2}(|X|). \tag{A8}
 \end{aligned}$$

The integral above can be evaluated in terms of generalized hypergeometric functions [37]. We present the calculation for when m is even in detail,

$$\int_0^{\infty} dX X^{\alpha-1} K_{\nu}(X) = [\Sigma(\nu, \alpha; Z) + \Sigma(-\nu, \alpha; Z)]_0^{\infty}, \tag{A9}$$

where

$$\begin{aligned}
 \Sigma(\nu, \alpha; Z) &\equiv -\frac{2^{\nu-1} \pi Z^{\alpha-\nu} \csc(\pi \nu)}{(\nu - \alpha) \Gamma(1 - \nu)} \\
 &\quad \times {}_1F_2\left(\frac{\alpha - \nu}{2}; 1 - \nu, \frac{\alpha - \nu}{2} + 1; \frac{Z^2}{4}\right), \tag{A10}
 \end{aligned}$$

with parameters,

$$\alpha \equiv \frac{m+5}{2} \quad \text{and} \quad \nu \equiv \frac{m-1}{2}, \tag{A11}$$

and ${}_1F_2$ is the generalized hypergeometric function,

$${}_1F_2(a_1; b_1, b_2; z) = \sum_{k=0}^{\infty} \frac{(a_1)_k z^k}{(b_1)_k (b_2)_k k!}. \tag{A12}$$

In the expression above, $(\cdot)_k$ is the Pochhammer symbol, and $\csc(\cdot)$ is the cosecant. Since m is even, we also have $\nu \notin \mathbb{Z}$. Putting everything together, we obtain the following expression for the variance:

$$\begin{aligned}
 \sigma_{UV}^2 &= \frac{[\Sigma(\nu, \alpha; Z) + \Sigma(-\nu, \alpha; Z)]_0^{\infty}}{2^{(m-3)/2} \sqrt{\pi} \Gamma\left(\frac{m}{2}\right) \sigma_U^{-2} \sigma_V^{-2}} \\
 &= \lim_{Z \rightarrow \infty} \frac{\Sigma(\nu, \alpha; Z) + \Sigma(-\nu, \alpha; Z)}{2^{(m-3)/2} \sqrt{\pi} \Gamma\left(\frac{m}{2}\right) \sigma_U^{-2} \sigma_V^{-2}}, \tag{A13}
 \end{aligned}$$

where we have used the fact that the numerator after the first equality vanishes at $Z = 0$. We can evaluate the limit $Z \rightarrow \infty$, on the right-hand side numerically as shown in Fig. 3 and find that the variance of the latent feature data values is

$$\sigma_{UV}^2 = m \sigma_U^2 \sigma_V^2. \tag{A14}$$

This is in agreement with the intuition that every latent dimension contributes its own variance to the variance of the data.

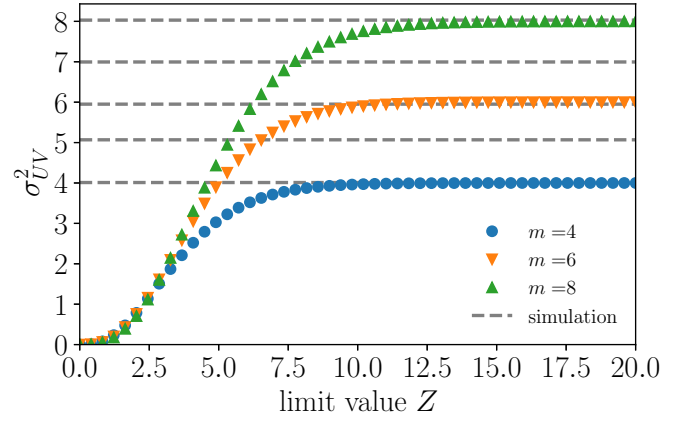


FIG. 3. Comparison between the variances computed from simulated data of the latent feature model for $m = 4, 5, 6, 7, 8$ (dashed gray lines) and the numerically evaluated limit expression for σ_{UV}^2 in Eq. (A13) as a function of the limit value Z for even m values. We have chosen $\sigma_U^2 = \sigma_V^2 = 1$.

We note that, for large values of the number of latent features m , the distribution (A6) becomes normal, in agreement with the law of large numbers,

$$\text{pdf}(X_{ij}) = \frac{1}{\sqrt{2\pi\sigma_{UV}^2}} e^{-(X^2/2\sigma_{UV}^2)}. \tag{A15}$$

Crucially, the variance of X_{ij} remains dependent on m . Figure 4 compares exact analytical expression of the probability distribution and its Gaussian approximation to numerical simulations.

As a final note, if we were interested in the distribution of data with noise, we would need to convolve the density in Eq. (A6) with the Gaussian density of the noise.

APPENDIX B: PROBABILITY DENSITY OF THE CORRELATION COEFFICIENTS

For our latent features model with noise, here we calculate the probability distribution of entries in the empirical data correlation matrix. Before doing this, a few notes are in order. First, the correlations depend on the basis in which variables are measured, becoming a diagonal matrix in the special case when the measured variables are the principal axes of the data cloud. Thus, to make statements independent of the basis, we consider the distribution of typical correlations, or correlations in the basis random with respect to the principal axes of the data. For a given realization, the N -dimensional data cloud is typically anisotropic, with $m < N$ long directions dominated by the latent feature signal and $N - m$ short directions dominated by noise. When $N \gg m$, principal axes of the data cloud do not align with the measured variables for the vast majority of random rotations, and correlations between any random pair of variables have contributions from all latent dimensions. Thus, we expect the number of latent dimensions to be imprinted in the distribution of the elements of the correlation matrix so that the statistics of the elements carries information about the underlying structure of the model.

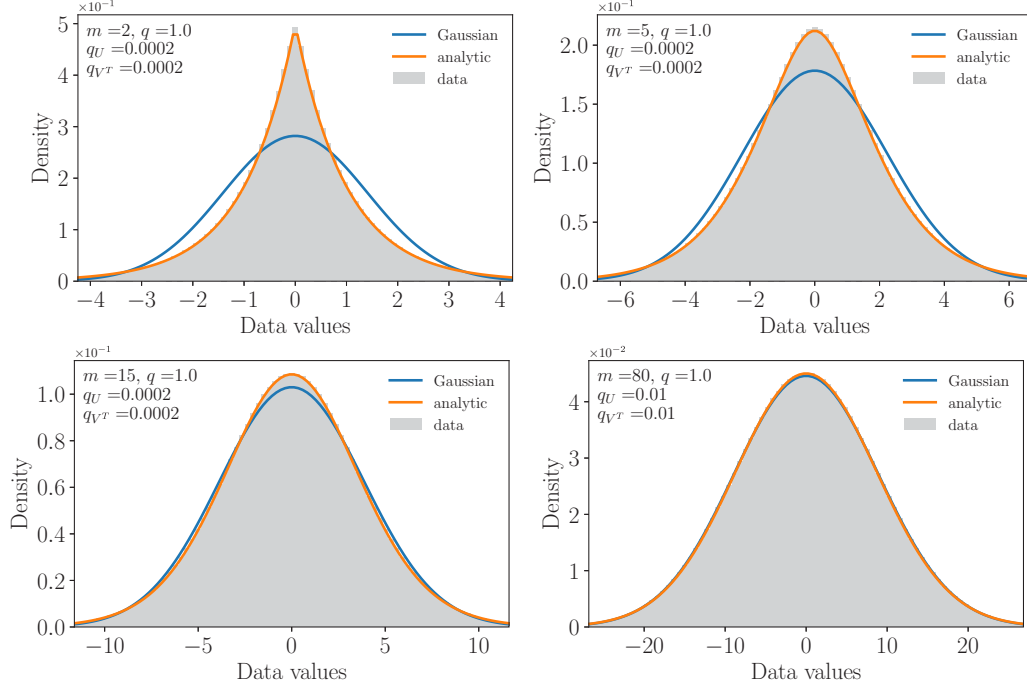


FIG. 4. Comparison of simulated data (gray) and the analytical distribution (orange). In the limit of large m , the distribution approaches a Gaussian form (blue). Simulated data constitute a single realization of the model with $\sigma_U^2 = \sigma_V^2 = 1$.

1. Preliminaries: Density of the correlation coefficient of two random Gaussian variables

The correlation coefficient of two independent zero-mean variables x and y sampled T times is

$$r = \frac{1}{T} \sum_t \frac{x_t y_t}{\sigma_x \sigma_y}, \quad (\text{B1})$$

where the vectors' components are mutually independent, i.i.d. random variables. The correlation coefficient is distributed according to [29]

$$\text{pdf}(r) = \frac{\Gamma(\frac{T}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{T-1}{2})} (1-r^2)^{(T-3)/2}. \quad (\text{B2})$$

This can be rewritten in terms of a Beta distribution,

$$\text{Beta}(x; \alpha, \beta) = \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad (\text{B3})$$

where $x \in [0, 1]$ and $\text{B}(\alpha, \beta)$ is the Beta function. Specifically, the density of correlations is given by the symmetric Beta distribution,

$$\text{pdf}(r) = \text{Beta}(r; \alpha, \alpha; \ell = -1, s = 2), \quad (\text{B4})$$

where the location ℓ and scale s are set such that the density is defined on the interval of correlation values $[-1, 1]$, and

$$\alpha = \frac{T-1}{2}. \quad (\text{B5})$$

We also note that the variance of a symmetric Beta distribution with the scale $s = 2$ is

$$\text{var} = \frac{s^2}{4(2\alpha+1)} = \frac{1}{2\alpha+1}. \quad (\text{B6})$$

2. Density of correlations in the latent feature model

There are multiple contributions to the correlations among the measured variables. We compute them individually and then combine the contributions. We find that each contribution is distributed according to a symmetric Beta distribution. To obtain the overall density, we approximate the sum of Beta distributions by a single Beta distribution, the parameter of which is obtained by matching the variance to the sum of the variances of the individual components. To perform these analyses, we only keep terms to the leading order in the $\text{SNR} \rightarrow 0$ or the $\text{SNR} \rightarrow \infty$ limit. Furthermore, we assume that q_U is small in accordance with the classical and intensive regimes limits.

We start with the pure noise contribution to the correlations,

$$(c_R)_{pq} = \frac{1}{T} \sum_t \frac{R_{pt}^T R_{tq}}{\sigma_p^n \sigma_q^n}, \quad (\text{B7})$$

$$(\sigma_q^n)^2 = \frac{1}{T} \sum_t R_{qt}^T R_{tq}. \quad (\text{B8})$$

The expression on the right-hand side is the correlation coefficient between two random Gaussian variables. Using Eq. (B4), we arrive at

$$\text{pdf}[(c_R)_{pq}] = \text{Beta}[(c_R)_{pq}; \alpha_n, \alpha_n; -1, 2], \quad p \neq q, \quad (\text{B9})$$

with

$$\alpha_n = \frac{T-1}{2}, \quad (\text{B10})$$

and the variance of this density is

$$\text{var}_n = T^{-1}. \quad (\text{B11})$$

Next we compute the density of the pure signal contribution,

$$(c_{UV})_{pq} = \frac{1}{T\sigma_p^s\sigma_q^s} \sum_t \left(\sum_\mu V_{p\mu} U_{\mu t} \right) \left(\sum_\nu U_{t\nu} V_{\nu q} \right), \quad (\text{B12})$$

$$(\sigma_p^s)^2 = \frac{1}{T} \sum_t \left(\sum_\mu V_{p\mu} U_{\mu t} \right) \left(\sum_\nu U_{t\nu} V_{\nu p} \right), \quad (\text{B13})$$

and similarly for σ_q . Rearranging, we find

$$(c_{UV})_{pq} = \frac{1}{\sigma_p^s\sigma_q^s} \sum_{\mu\nu} V_{p\mu} V_{\nu q} \left(\frac{1}{T} \sum_t U_{\mu t} U_{t\nu} \right), \quad (\text{B14})$$

$$(\sigma_p^s)^2 = \sum_{\mu\nu} V_{p\mu} V_{\nu p} \left(\frac{1}{T} \sum_t U_{\mu t} U_{t\nu} \right). \quad (\text{B15})$$

The expression in parentheses of both of the equations above is a (co)-variance of Gaussian random numbers. For $\mu = \nu$, it follows the scaled χ^2 distribution with T degrees of freedom. For $\mu \neq \nu$, it is given by a rescaled version of the distribution in Eq. (A6) with T instead of m . Crucially, the variance of either is $1/T$. Thus, in the limit $q \rightarrow 0$, the terms in parentheses are $\sigma_U^2 \delta_{\mu\nu} + O(T^{-1/2})$ where the correction $O(T^{-1/2})$ is probabilistic but will be neglected in what follows. We get

$$(c_{UV})_{pq} = \sigma_U^2 \sum_{\mu\nu} \frac{V_{p\mu} V_{\nu q} \delta_{\mu\nu}}{\sigma_p^s\sigma_q^s} = m\sigma_U^2 \left(\frac{1}{m} \sum_\mu \frac{V_{p\mu} V_{\mu q}}{\sigma_p^s\sigma_q^s} \right), \quad (\text{B16})$$

$$(\sigma_p^s)^2 = m\sigma_U^2 \left(\frac{1}{m} \sum_\mu V_{p\mu}^2 \right), \quad (\text{B17})$$

We see that the sought after correlation is a correlation coefficient between Gaussian variables but with m samples instead of T . Using again Eq. (B4), we write

$$\text{pdf}[(c_{UV})_{pq}] = \text{Beta}[(c_{UV})_{pq}; \alpha_s, \alpha_s; -1, 2], \quad p \neq q, \quad (\text{B18})$$

with parameter,

$$\alpha_s = \frac{m-1}{2}. \quad (\text{B19})$$

We remind the reader that Eq. (B18) holds to $O(T^{-1/2})$. The variance of this density is

$$\text{var}_s = m^{-1}. \quad (\text{B20})$$

This expression agrees with numerical simulations very well, cf. Fig. 1.

Finally, for the signal-noise cross terms in the correlation, we have

$$\begin{aligned} (c_{(UV)^{TR}})_{pq} &= \frac{1}{T\sigma_p^s\sigma_q^n} \sum_t \sum_\mu V_{p\mu} U_{\mu t} R_{tq} \\ &= \frac{1}{\sigma_p^s\sigma_q^n} \sum_\mu V_{p\mu} \left(\frac{1}{T} \sum_t U_{\mu t} R_{tq} \right). \end{aligned} \quad (\text{B21})$$

For the quantity in parentheses in Eq. (B21), we define

$$r_{\mu p} \equiv \frac{1}{T} \sum_t U_{\mu t} R_{tq}. \quad (\text{B22})$$

This is a covariance between two independent Gaussian random numbers and again follows a rescaled form of the distribution in Eq. (A6) with variance $\sigma_U^2 \sigma_q^{n2} T^{-1}$. Since T is large, the distribution approaches a Gaussian, and we further define $r'_{\mu q} \equiv \sigma_U \sigma_q^{n2} T^{-1/2} r'_{\mu q}$ such that $r'_{\mu q}$ is a unit Gaussian random variable. Thus, we obtain

$$\begin{aligned} (c_{(UV)^{TR}})_{pq} &= \frac{\sigma_U m \sigma_q^{n2} T^{-1/2}}{\sigma_p^s\sigma_q^n} \left(\frac{1}{m} \sum_\mu V_{p\mu} r'_{\mu q} \right) \\ &= m^{1/2} T^{-1/2} \left(\frac{1}{m} \sum_\mu \frac{V_{p\mu} r'_{\mu q}}{\frac{1}{m} \sum_\mu V_{p\mu}^2} \right), \end{aligned} \quad (\text{B23})$$

where we have extracted the factor of m to highlight that the expression in parentheses is the correlation between Gaussian random numbers. From this, using Eq. (B4), we conclude that

$$\begin{aligned} \text{pdf}[(c_{(UV)^{TR}})_{pq}] &= \text{Beta}[(c_{(UV)^{TR}})_{pq}; \alpha_{\text{sn}}, \alpha_{\text{sn}}; -1, 2], \\ p &\neq q, \end{aligned} \quad (\text{B24})$$

with parameter,

$$\alpha_{\text{sn}} = \frac{m^{1/2} T^{1/2} - 1}{2}. \quad (\text{B25})$$

The variance of this density is

$$\text{var}_{\text{sn}} = m^{-1/2} T^{-1/2} = \sqrt{\text{var}_s \text{var}_n}. \quad (\text{B26})$$

An analogous expression holds for the $R^T UV$ contribution.

The empirical correlation matrix is given by

$$c_{pq} = \frac{1}{T} \sum_t \frac{X_{pt} X_{tq}}{\sigma_p^{\text{sn}} \sigma_q^{\text{sn}}}, \quad (\text{B27})$$

where

$$(\sigma_p^{\text{sn}})^2 = (\sigma_p^s)^2 + (\sigma_p^n)^2. \quad (\text{B28})$$

Using Eqs (B7), (B12), and (B21), the correlation matrix can be written as a weighted sum of the three types of contributions,

$$\begin{aligned} c_{pq} &= \frac{\sigma_p^s\sigma_q^s}{\sigma_p^{\text{sn}}\sigma_q^{\text{sn}}} (c_{UV})_{pq} + \frac{\sigma_p^s\sigma_q^n}{\sigma_p^{\text{sn}}\sigma_q^{\text{sn}}} (c_{(UV)^{TR}})_{pq} \\ &+ \frac{\sigma_p^n\sigma_q^s}{\sigma_p^{\text{sn}}\sigma_q^{\text{sn}}} (c_{R^T UV})_{pq} + \frac{\sigma_p^n\sigma_q^n}{\sigma_p^{\text{sn}}\sigma_q^{\text{sn}}} (c_R)_{pq}. \end{aligned} \quad (\text{B29})$$

Each term on the right-hand side of this equation follows a Beta distribution as computed above. However, the α parameter of each distribution is modified by the corresponding weight in the above sum. Consequently, the variance of each distribution is rescaled by the weight,

$$\text{var}'_s = \frac{\sigma_p^s\sigma_q^s}{\sigma_p^{\text{sn}}\sigma_q^{\text{sn}}} \text{var}_s, \quad (\text{B30})$$

$$\text{var}'_n = \frac{\sigma_p^n\sigma_q^n}{\sigma_p^{\text{sn}}\sigma_q^{\text{sn}}} \text{var}_n, \quad (\text{B31})$$

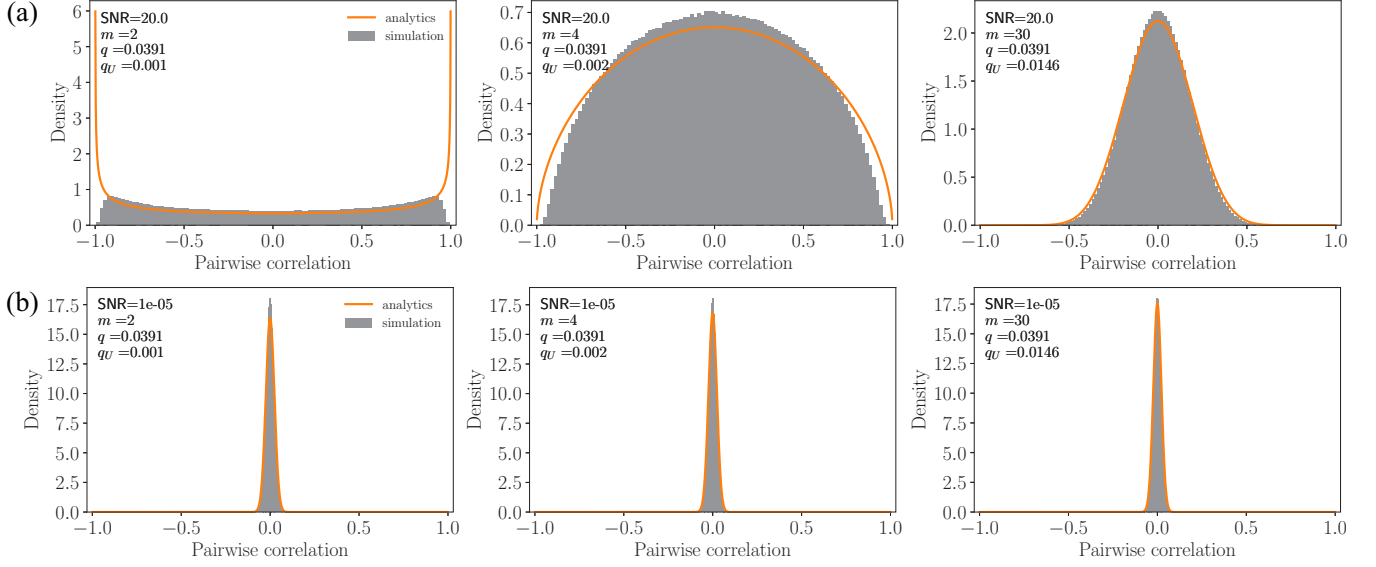


FIG. 5. Distribution of pairwise correlations in the regimes of finite and small signal-to-noise ratio with $m = 2, 4$ and 30 latent features. Analytic form (orange) and simulated data (gray). (a) $\text{SNR} = 20$ and (b) $\text{SNR} = 10^{-5}$ (large noise limit). Each simulation is run with $N = 80$ variables and $T = 2048$ observations and constitutes 1000 independent model realizations.

$$\text{var}'_{\text{sn}} = \frac{\sigma_p^s \sigma_q^n}{\sigma_p^{\text{sn}} \sigma_q^{\text{sn}}} \text{var}_{\text{sn}}. \quad (\text{B32})$$

To determine an expression for the combined distribution of signal and noise correlations, we make use of the observation that the sum of Beta distributions can be well approximated by a single Beta distribution [38]. We determine the parameters of the Beta distribution by adding the means and variances of the distributions in the sum and analytically match the parameter of the single Beta distribution.

The means of the Beta distributions in Eqs. (B9), (B18), and (B24) are zero and, thus, the mean of the density of the combined contributions is also zero. Taking the sum of variances we obtain

$$\text{var} = \text{var}'_s + \text{var}'_n + \text{var}'_{\text{sn}} + \text{var}'_{\text{ns}}. \quad (\text{B33})$$

In the limit when T and m are large enough such that contributions of $O(T^{-1/2})$ and $O(m^{-1/2})$ can be neglected, we have the following convergence of the empirical quantities,

$$(\sigma_p^s)^2 \rightarrow m \sigma_U^2 \sigma_V^2, \quad (\text{B34})$$

$$(\sigma_p^n)^2 \rightarrow \sigma^2, \quad (\text{B35})$$

$$(\sigma_p^{\text{sn}})^2, (\sigma_p^{\text{ns}})^2 \rightarrow m \sigma_U^2 \sigma_V^2 + \sigma^2. \quad (\text{B36})$$

Consequently, the variances of the contributions take the form

$$\text{var}'_s \rightarrow \frac{m^{-1}}{1 + \text{SNR}^{-1}}, \quad (\text{B37})$$

$$\text{var}'_n \rightarrow \frac{T^{-1}}{1 + \text{SNR}}, \quad (\text{B38})$$

$$\text{var}'_{\text{sn}}, \text{var}'_{\text{ns}} \rightarrow \frac{m^{-1/2} T^{-1/2}}{\sqrt{1 + \text{SNR}} \sqrt{1 + \text{SNR}^{-1}}}, \quad (\text{B39})$$

Thus, in this limit, the variance of the Beta distribution, Eq. (B33), is of the form

$$\text{var} \approx \left(\frac{m^{-1/2}}{\sqrt{1 + \text{SNR}^{-1}}} + \frac{T^{-1/2}}{\sqrt{1 + \text{SNR}}} \right)^2. \quad (\text{B40})$$

Finally, from the relation in Eq. (B6), we obtain the parameter α of the sought after Beta distribution,

$$\alpha = \frac{\text{var}^{-1} - 1}{2}. \quad (\text{B41})$$

A comparison between the analytic form of the density and simulated data is shown in Fig. 1 for $\text{SNR} \rightarrow \infty$ and in Fig. 5 for finite SNR and $\text{SNR} \rightarrow 0$. In the extreme noise limits, the analytic form closely matches the simulation. In the large noise limit of $\text{SNR} \rightarrow 0$, shown in Fig. 5(b), the density is close to a Gaussian because the number of observations T is large. In the regime of finite SNR, shown in Fig. 5(a), deviations between the analytic form and the simulation appear for small values of m . We expect that these deviations will disappear by removing the various approximations made in the above analytic derivation.

APPENDIX C: SPECTRUM OF THE NORMALIZED EMPIRICAL COVARIANCE MATRIX

To compute the eigenvalue density of the NECM, \mathbf{C} , we use methods of Random Matrix Theory [39]. The standard approach is to compute the finite size Stieltjes transform,

$$g_{\mathbf{C}}^N(z) = \frac{1}{N} \text{Tr}(z\mathbf{I} - \mathbf{C})^{-1}, \quad (\text{C1})$$

where \mathbf{I} is the identity matrix, $z \in \mathbb{C}$, and $g_{\mathbf{C}}^N$ is a complex function. In the limit of large matrices—large N or thermodynamic limit—the finite size Stieltjes transform becomes $g_{\mathbf{C}}(z)$. Then the eigenvalue density is obtained as the imaginary part

of the limit of the Stieltjes transform,

$$\rho(\lambda) = \frac{1}{\pi} \lim_{\eta \rightarrow 0^+} \text{Im } \mathfrak{g}(z = \lambda - i\eta), \quad (\text{C2})$$

where Im denotes the imaginary part.

We start with writing again the definition of the NECM, which differs from the correlation matrix only by $O(T^{-1/2})$,

$$\begin{aligned} \mathbf{C} &= \frac{1}{T} \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} = \frac{1}{T} (\mathbf{UV} + \sigma \mathbf{R})^T (\mathbf{UV} + \sigma \mathbf{R}) \\ &= \frac{1}{T} [(\widetilde{\mathbf{UV}})^T (\widetilde{\mathbf{UV}}) + \tilde{\sigma}^2 \mathbf{R}^T \mathbf{R} \\ &\quad + \tilde{\sigma} (\widetilde{\mathbf{UV}})^T \mathbf{R} + \tilde{\sigma} \mathbf{R}^T \widetilde{\mathbf{UV}}]. \end{aligned} \quad (\text{C3})$$

The NECM contains three different contributions: the $(\mathbf{UV})^T (\mathbf{UV})$ from the pure latent feature signal, $\mathbf{R}^T \mathbf{R}$ from pure noise, and two terms of the type $(\mathbf{UV})^T \mathbf{R}$, which are cross terms between the latent signal and the noise. Each contribution is an $N \times N$ random matrix. Critical to computing the eigenvalue density of random matrices is the concept of *matrix freeness* [40], which is the generalization of statistical independence to matrices. The eigenvalue spectrum of sums and products of free matrices can be computed from spectra of summands and factors using the \mathcal{R} and the \mathcal{S} -transforms, which are related to the Stieltjes transform \mathfrak{g} and are additive and multiplicative, respectively. The signal-signal and the noise-noise contributions in the NECM definition are certainly free with respect to each other. We will argue in Appendix C 3 that, in our regimes of interest [the zero-noise limit ($\text{SNR} \rightarrow \infty$), the classical statistics limit from Eq. (9), and intensive limit from Eq. (10)], the cross-term contributions are negligible so that we can drop them and approximate the NECM as

$$\mathbf{C} \approx \frac{(\mathbf{UV})^T (\mathbf{UV}) + \sigma^2 \mathbf{R}^T \mathbf{R}}{\sigma_X^2 T} := \mathbf{C}_{\widetilde{\mathbf{UV}}} + \mathbf{C}_{\tilde{\sigma} \mathbf{R}}, \quad (\text{C4})$$

so that free matrix theory applies.

1. Parametrizing the random matrix problem and the large matrix limit

To calculate the spectrum of the signal-signal contribution to the NECM,

$$\mathbf{C}_{\widetilde{\mathbf{UV}}} = \frac{1}{\sigma_X^2 T} (\mathbf{UV})^T (\mathbf{UV}), \quad (\text{C5})$$

we note that, assuming $m < T, N$, this $N \times N$ matrix is of rank m . Thus, we can work in the basis, where

$$\mathbf{C}_{\widetilde{\mathbf{UV}}} = \begin{pmatrix} \mathbf{H}_{\widetilde{\mathbf{UV}}} & 0 \\ 0 & 0 \end{pmatrix}, \quad (\text{C6})$$

and

$$\mathbf{H}_{\widetilde{\mathbf{UV}}} = \frac{1}{\sigma_X^2 T} (\mathbf{U}^T \mathbf{U})(\mathbf{V} \mathbf{V}^T). \quad (\text{C7})$$

There are m nontrivial eigenvalues associated with \mathbf{H} , whereas the remaining $N - m$ eigenvalues are zero. The finite size Stieltjes transform, $g_{\mathbf{C}}^N = N^{-1} \text{Tr}(z \mathbf{I} - \mathbf{C}_{\widetilde{\mathbf{UV}}})^{-1}$ is then of the

form

$$\begin{aligned} g_{\mathbf{C}_{\widetilde{\mathbf{UV}}}}^N(z) &= \frac{1}{N} \left(m \frac{1}{m} \sum_{\mu=1}^m \frac{1}{z - \lambda_{\mu}} + \frac{N - m}{z} \right) \\ &= \frac{1}{N} \left(m h_{\mathbf{H}_{\widetilde{\mathbf{UV}}}}^m(z) + \frac{N - m}{z} \right), \end{aligned} \quad (\text{C8})$$

where λ_{μ} 's are the m eigenvalues of $\mathbf{H}_{\widetilde{\mathbf{UV}}}$ and $h_{\mathbf{H}_{\widetilde{\mathbf{UV}}}}^m(z)$ is its finite size Stieltjes transform.

Now we note that $\mathbf{H}_{\widetilde{\mathbf{UV}}}$ in Eq. (C7) is the product of two white Wishart matrices,

$$\mathbf{H}_{\widetilde{\mathbf{UV}}} = \frac{N}{\sigma_X^2} \mathbf{W}_U \mathbf{W}_{V^T}, \quad (\text{C9})$$

where

$$\mathbf{W}_Y = \frac{1}{T} \mathbf{Y}^T \mathbf{Y}, \quad (\text{C10})$$

is the Wishart matrix, and \mathbf{Y} is a $T \times N$ matrix with i.i.d. standard normal entries. The key parameter characterizing such standard \mathbf{W}_Y is the ratio of the number of columns to that of rows,

$$q \equiv \frac{N}{T}. \quad (\text{C11})$$

Since \mathbf{U} and \mathbf{V}^T are $T \times m$ and $N \times m$ matrices, respectively, a natural characterisation of $\mathbf{H}_{\widetilde{\mathbf{UV}}}$ is then,

$$q \equiv \frac{N}{T}, \quad q_U \equiv \frac{m}{T}, \quad q_{V^T} \equiv \frac{m}{N}, \quad (\text{C12})$$

with $q q_{V^T} = q_U$ so that there are only two independent parameters.

It is now convenient to define

$$\sigma_X^2 = m \left(\sigma_U^2 \sigma_V^2 + \frac{\sigma^2}{m} \right) \equiv m \tilde{\sigma}_X^2, \quad (\text{C13})$$

where we used Eq. (A14) so that Eq. (C9) becomes

$$\mathbf{H}_{\widetilde{\mathbf{UV}}} = \frac{1}{q_{V^T} \tilde{\sigma}_X^2} \mathbf{W}_U \mathbf{W}_{V^T}. \quad (\text{C14})$$

In the following, we only consider the limit of large matrices. Here T, N, m , and σ^2 go to infinity in such a way that q, q_{V^T} , and SNR are all constant. Then, in the thermodynamic limit the finite size Stieltjes transform in Eq. (C8) becomes

$$g_{\mathbf{C}_{\widetilde{\mathbf{UV}}}} = q_{V^T} \mathfrak{h} + \frac{1 - q_{V^T}}{z}, \quad (\text{C15})$$

where $g_{\mathbf{C}_{\widetilde{\mathbf{UV}}}}$ and \mathfrak{h} are the large matrices limits of the Stieltjes transforms of $g_{\mathbf{C}_{\widetilde{\mathbf{UV}}}}^N$ and $h_{\mathbf{H}_{\widetilde{\mathbf{UV}}}}^m$, respectively.

2. The spectrum of $\mathbf{C}_{\widetilde{\mathbf{UV}}}$

We now compute the eigenvalue density of $\mathbf{C}_{\widetilde{\mathbf{UV}}}$. The first step is to compute the Stieltjes transform \mathfrak{h} . From Eq. (C14), it is clear that this reduces to the problem of computing the eigenvalue spectrum of a product of two Wishart matrices.

The spectrum of a product of two free matrices can be computed with the help of the \mathcal{S} -transform, which is defined for a random matrix \mathbf{A} as

$$\mathcal{S}_{\mathbf{A}}(t) = \frac{t+1}{t\mathcal{T}_{\mathbf{A}}^{-1}(t)}, \quad (\text{C16})$$

where $\mathcal{T}_{\mathbf{A}}^{-1}(t)$ is the functional inverse of the \mathcal{T} -transform $\mathcal{T}_{\mathbf{A}}(z)$. In turn, the \mathcal{T} -transform is related to the Stieltjes transform of \mathbf{A} through the relation,

$$\mathcal{T}_{\mathbf{A}}(z) = z\mathfrak{g}_{\mathbf{A}}(z) - 1. \quad (\text{C17})$$

Crucially, for free matrices \mathbf{A} and \mathbf{B} , the \mathcal{S} -transform is multiplicative,

$$\mathcal{S}_{\mathbf{AB}}(t) = \mathcal{S}_{\mathbf{A}}(t)\mathcal{S}_{\mathbf{B}}(t), \quad (\text{C18})$$

and, furthermore, for a scalar a ,

$$\mathcal{S}_{a\mathbf{A}}(t) = a^{-1}\mathcal{S}_{\mathbf{A}}(t). \quad (\text{C19})$$

For the white Wishart matrix, Eq. (C10), the \mathcal{S} -transform is known to be [39]

$$\mathcal{S}_{\mathbf{W}_V}(t) = \frac{1}{1+qt}. \quad (\text{C20})$$

Thus, we only need to use the multiplicative property of the \mathcal{S} -transform to compute the signal-signal contributions to the NECM. Specifically,

$$\mathcal{S}_{\mathbf{H}_{UV}}(t) = q_{V^T}\bar{\sigma}_X^2\mathcal{S}_{\mathbf{W}_U}\mathcal{S}_{\mathbf{W}_{V^T}} = \frac{q_{V^T}\bar{\sigma}_X^2}{(1+q_U t)(1+q_{V^T} t)}. \quad (\text{C21})$$

Equation (C16) then yields

$$\mathcal{T}_{\mathbf{H}_{UV}}^{-1}(t) = \frac{t+1}{t\mathcal{S}_{\mathbf{H}_{UV}}(t)} = \frac{t+1}{t} \frac{(1+q_U t)(1+q_{V^T} t)}{q_{V^T}\bar{\sigma}_X^2}. \quad (\text{C22})$$

We now solve the equation for the functional inverse $\mathcal{T}^{-1}[\mathcal{T}(z)] = z$, using the definition of the \mathcal{T} -transform Eq. (C17) and dividing by a common factor of z . We obtain a cubic equation for the Stieltjes transform \mathfrak{h} ,

$$\mathfrak{h}^3 z^2 q_U q_{V^T} + \mathfrak{h}^2 z [q_{V^T}(1-q_U) + q_U(1-q_{V^T})] + \mathfrak{h}[(1-q_U)(1-q_{V^T}) - zq_{V^T}\bar{\sigma}_X^2] + q_{V^T}\bar{\sigma}_X^2 = 0. \quad (\text{C23})$$

Finally, we divide by $q_{V^T}\bar{\sigma}_X^2$ to obtain

$$\mathfrak{h}^3 \frac{z^2 q_U}{\bar{\sigma}_X^2} + \mathfrak{h}^2 \frac{z}{\bar{\sigma}_X^2} (1+q-2q_U) + \mathfrak{h} \left(\frac{q_{V^T}^{-1} - q - 1 + q_U}{\bar{\sigma}_X^2} - z \right) + 1 = 0. \quad (\text{C24})$$

Similar equations for the Stieltjes transform of the product of two random matrices have been stated in Refs. [41–43]. Their polynomials differ from Eq. (C24) in detail because we consider the Stieltjes transform of the covariance matrix including a theoretical normalisation factor.

The next step is to solve Eq. (C24) analytically in the classical statistics limit and the intensive limit. We remind the reader that for the pure signal contribution we work in the zero-noise limit $\text{SNR} \rightarrow \infty$ such that

$$\bar{\sigma}_X^2 \equiv \frac{\sigma_X^2}{m} = \sigma_U^2 \sigma_V^2 (1 + \text{SNR}^{-1}) = \sigma_U^2 \sigma_V^2. \quad (\text{C25})$$

a. Classical statistics limit

In the *classical statistics* limit Eq. (9), the polynomial equation for the Stieltjes transform, Eq. (C24) becomes

$$\mathfrak{h}^2 \frac{z}{\bar{\sigma}_X^2} + \mathfrak{h} \left(\frac{q_{V^T}^{-1} - 1}{\bar{\sigma}_X^2} - z \right) + 1 = 0. \quad (\text{C26})$$

The discriminant is

$$\Delta = z^2 - 2 \frac{1+q_{V^T}^{-1}}{\bar{\sigma}_X^2} z + \left(\frac{q_{V^T}^{-1} - 1}{\bar{\sigma}_X^2} \right)^2, \quad (\text{C27})$$

and the roots of the discriminant are

$$\lambda_{\pm}^{\infty} = \bar{\sigma}_X^{-2} (1 \pm \sqrt{q_{V^T}^{-1}})^2. \quad (\text{C28})$$

We, thus, obtain

$$\mathfrak{h}_{\pm} = \frac{-\frac{q_{V^T}^{-1}-1}{\bar{\sigma}_X^2} + z \pm \sqrt{(z-\lambda_{-}^{\infty})(z-\lambda_{+}^{\infty})}}{2z\bar{\sigma}_X^{-2}}. \quad (\text{C29})$$

To obtain $\mathfrak{g}_{\mathbf{C}_{UV}}$, we now need to add the contribution of the zero eigenvalues,

$$\begin{aligned} \mathfrak{g}_{\mathbf{C}_{UV}} &= q_{V^T} \mathfrak{h}_{\pm} + \frac{1-q_{V^T}}{z} \\ &= \frac{1-q_{V^T}}{2z} + \frac{q_{V^T}}{2\bar{\sigma}_X^{-2}} \pm \frac{\sqrt{(z-\lambda_{-}^{\infty})(z-\lambda_{+}^{\infty})}}{2zq_{V^T}^{-1}\bar{\sigma}_X^{-2}}. \end{aligned} \quad (\text{C30})$$

We are now ready to obtain the eigenvalue density as in Eq. (C2). Whereas this is a standard calculation [44], we summarize it here for the reader's benefit. The second term on the right-hand side of Eq. (C30) is real, does not contribute to the imaginary part, and we ignore it. For the first and the third terms, we multiply the numerators and the denominators by $z^* = \lambda + i\eta$. The imaginary part of the first term is then,

$$\text{Im} \left(\frac{1-q_{V^T}}{2z} \right) = \frac{(1-q_{V^T})\eta}{2(\eta^2 + \lambda^2)} = \frac{(1-q_{V^T})\pi}{2} \delta_{\eta}(\lambda), \quad (\text{C31})$$

where we have used the definition of the Lorentz curve $\delta_{\eta}(\lambda) = \pi^{-1}\eta/(\eta^2 + \lambda^2)$. For the third term, the crucial step is to rewrite the square root using the relation,

$$\sqrt{a+ib} = P+iQ, \quad (\text{C32})$$

where a and b are real, $b \neq 0$ and

$$\begin{aligned} P &= \frac{1}{\sqrt{2}} \sqrt{\sqrt{a^2+b^2}+a}, \\ Q &= \frac{\text{sgn}(b)}{\sqrt{2}} \sqrt{\sqrt{a^2+b^2}-a}, \end{aligned} \quad (\text{C33})$$

where $\text{sgn}(x) = 1$ for $x > 0$ and -1 for $x < 0$ [45]. For the argument of the square root in the third term of Eq. (C30), we find

$$\begin{aligned} a &= \lambda^2 - \eta^2 + \lambda_{+}^{\infty} \lambda_{-}^{\infty} - (\lambda_{+}^{\infty} + \lambda_{-}^{\infty})\lambda, \\ b &= (-2\lambda + \lambda_{+}^{\infty} + \lambda_{-}^{\infty})\eta. \end{aligned} \quad (\text{C34})$$

The imaginary part of the third term takes the form

$$\begin{aligned} \text{Im}\left(\pm \frac{\sqrt{(z - \lambda_-^\infty)(z - \lambda_+^\infty)}}{2zq_{V^T}^{-1}\bar{\sigma}_X^{-2}}\right) &= \pm \frac{\text{Im}(z^*[P + iQ])}{2q_{V^T}^{-1}\bar{\sigma}_X^{-2}|z|^2} = \pm \frac{1}{2q_{V^T}^{-1}\bar{\sigma}_X^{-2}} \left(\frac{\eta}{\eta^2 + \lambda^2} P + \frac{\lambda}{\eta^2 + \lambda^2} Q \right) \\ &= \pm \frac{1}{2q_{V^T}^{-1}\bar{\sigma}_X^{-2}} \left(\pi \delta_\eta(\lambda) P + \frac{\lambda}{\eta^2 + \lambda^2} Q \right). \end{aligned} \quad (\text{C35})$$

The final step to evaluate Eq. (C2) and to obtain the eigenvalue density is to take the limit $\eta \rightarrow 0^+$. In this limit, the Lorentz curve in Eq. (C31) converges to the Dirac- δ function. Combining Eqs. (C31) and (C35) yields

$$\begin{aligned} \rho^\infty(\lambda) &= \frac{1}{\pi} \lim_{\eta \rightarrow 0^+} \text{Im } \mathbf{g}_{C_{\bar{v}}} \\ &= \pm \frac{\lim_{\eta \rightarrow 0^+} P}{2q_{V^T}^{-1}\bar{\sigma}_X^{-2}} \delta(\lambda) \pm \frac{\lim_{\eta \rightarrow 0^+} Q}{2\pi\lambda\bar{\sigma}_X^{-2}q_{V^T}^{-1}} + \frac{1 - q_{V^T}}{2} \delta(\lambda), \end{aligned} \quad (\text{C36})$$

with

$$\lim_{\eta \rightarrow 0^+} P = \sqrt{\lambda_+^\infty \lambda_-^\infty} = \bar{\sigma}_X^{-2} (1 - q_{V^T}^{-1}), \quad (\text{C37})$$

where we have used the expression for the zero noise eigenvalue bounds in Eq. (C28), and

$$\begin{aligned} \lim_{\eta \rightarrow 0^+} Q &= \frac{\text{sgn}(b)}{\sqrt{2}} \sqrt{2|(\lambda - \lambda_-^\infty)(\lambda_+^\infty - \lambda)|} \\ &= \text{sgn}(b) \sqrt{(\lambda - \lambda_-^\infty)(\lambda_+^\infty - \lambda)}, \end{aligned} \quad (\text{C38})$$

when $\lambda \in [\lambda_-^\infty, \lambda_+^\infty]$, and the expression vanishes elsewhere. The \pm signs in Eq. (C36) are chosen such as to obtain a physically meaningful eigenvalue density. Finally, we find the following form of the eigenvalue density:

$$\rho^\infty(\lambda) = \frac{\sqrt{(\lambda - \lambda_-^\infty)(\lambda_+^\infty - \lambda)}}{2\pi\lambda\bar{\sigma}_X^{-2}q_{V^T}^{-1}} + (1 - q_{V^T})\delta(\lambda), \quad (\text{C39})$$

with $\bar{\sigma}_X^2 \equiv \sigma_X^2/m = \sigma_U^2\sigma_V^2$. We note that in Ref. [46] an expression for an eigenvalue density was given in the special case when $T = N = m$ and not including our theoretical normalization factor.

b. Intensive limit

For the *intensive limit* Eq. (10), the polynomial equation for the Stieltjes transform Eq. (C24) becomes

$$\mathfrak{h}^2 \frac{z}{\bar{\sigma}_X^2} (1 + q) + \mathfrak{h} \left(\frac{q_{V^T}^{-1} - q - 1}{\bar{\sigma}_X^2} - z \right) + 1 = 0. \quad (\text{C40})$$

The discriminant is

$$\Delta = z^2 - 2 \frac{q_{V^T}^{-1} + q + 1}{\bar{\sigma}_X^2} z + \left(\frac{q_{V^T}^{-1} - q - 1}{\bar{\sigma}_X^2} \right)^2. \quad (\text{C41})$$

The roots of the discriminant are

$$\lambda_\pm^\infty = \bar{\sigma}_X^{-2} (\sqrt{1 + q} \pm \sqrt{q_{V^T}^{-1}})^2. \quad (\text{C42})$$

Then the solution of the quadratic equation is

$$\mathfrak{h}_\pm = \frac{-\frac{q_{V^T}^{-1} - q - 1}{\bar{\sigma}_X^2} + z \pm \sqrt{(z - \lambda_-^\infty)(z - \lambda_+^\infty)}}{2z\bar{\sigma}_X^{-2}(1 + q)}. \quad (\text{C43})$$

Following a calculation analogous to the classical limit, we now add the contribution of the zero eigenvalues and then determine the density of the eigenvalues. We find

$$\rho^\infty(\lambda) = \frac{\sqrt{(\lambda - \lambda_-^\infty)(\lambda_+^\infty - \lambda)}}{2\pi\lambda\bar{\sigma}_X^{-2}(1 + q)q_{V^T}^{-1}} + (1 - q_{V^T})\delta(\lambda), \quad (\text{C44})$$

with $\bar{\sigma}_X^2 \equiv \sigma_X^2/m = \sigma_U^2\sigma_V^2$.

3. Approximation to neglect the signal-noise cross terms

Now we explore when the contribution of the signal-noise cross terms to the NECM can be neglected. Specifically, we will show that it can be performed if $q_U \rightarrow 0$ (that is, the number of measurements is much larger than the number of latent features), which we always assume. To show this, we compute the eigenvalue bounds $\lambda_\pm^{\text{signal noise}}$ of the signal-noise contribution and compare their scaling with T to the scaling of the pure signal and the pure noise eigenvalue bounds.

For the pure signal contribution, the previous section shows that the eigenvalue bounds λ_\pm^∞ are $\bar{\sigma}_X^{-2} \sim O(T^0)$. The pure noise eigenvalue bounds, given by the Marčenko-Pastur bounds, scale as

$$\lambda_\pm^{\text{MP}} \sim 1 \pm T^{-1/2}, \quad (\text{C45})$$

where 1 is due to self-correlations. On the other hand, the signal-noise cross terms do not have self-correlations, and, thus, we expect their bounds to scale as

$$\lambda_\pm^{\text{signal-noise}} \sim T^{-1/2}, \quad (\text{C46})$$

becoming negligible for $T \rightarrow \infty$. In, Appendix C3 a we show this analytically in the classical statistics limit. We have not been able to achieve similar results more generally. However, since $q_U \rightarrow 0$ also in the intensive limit, we expect similar results to hold there too. To show this, we resort to numerical simulations.

Specifically, we numerically estimate the Jensen-Shannon divergence between the numerically evaluated eigenvalue densities of the NECM, computed with and without the signal-noise cross terms. To obtain a perceptually intuitive measure of the difference between these distributions, we convert the Jensen-Shannon divergence to the effective sensitivity index d' —the distance between the means of two unit variance normal distribution with the same Jensen-Shannon divergence as the two analyzed eigenvalue spectra. We investigate the dependence of d' on various choices of our model parameters. The comparison between the spectra of the full and the approximate NECM is shown in Fig. 6. We observe that the sensitivity index reaches a maximum for $\text{SNR} \sim 10^{-1}$ and falls off in the limits of small or large SNR where the noise or the signal dominate, respectively. Crucially, the maximum

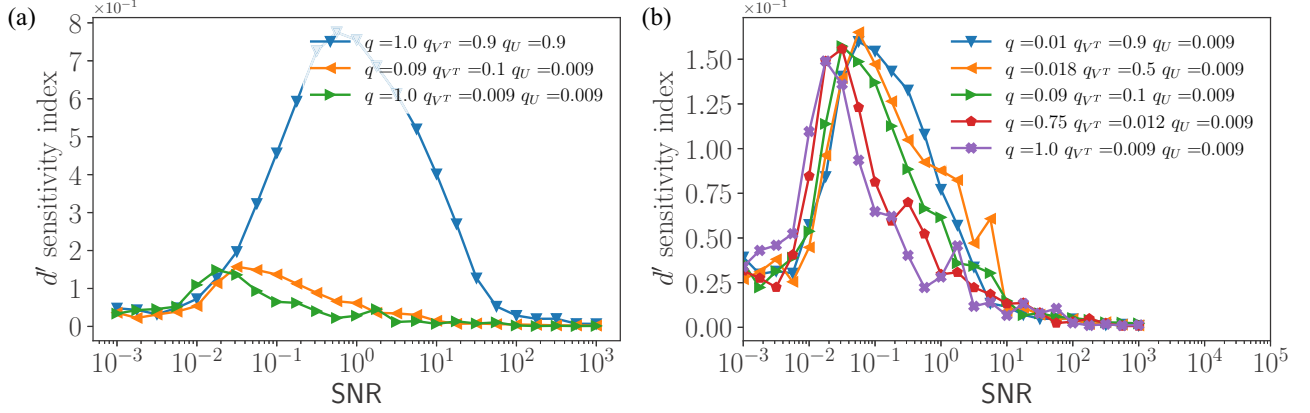


FIG. 6. Difference between the eigenvalue density of the NECM spectrum with and without the signal-noise cross terms quantified by d' . (a) Classical limit (orange), intensive limit (green), and neither of the two limits (blue). (b) Magnified view of d' in the limits of interest: classical limit (blue, orange, and green) and intensive limit (red and purple). Eigenvalue densities are computed from 120 realizations of the random matrix model.

value of d' is small for $q_U \rightarrow 0$. Thus, neglecting the cross-term contributions to the NECM spectrum in our limits of interest is warranted.

a. Scaling behavior of the signal-noise eigenvalue bounds in the classical limit

We now derive the scaling of the signal-noise eigenvalue spectrum bounds in Eq. (C46) in the classical statistics limit. From the NECM in Eq. (C3), the signal-noise cross terms are of the form

$$\mathbf{M} \equiv \frac{\tilde{\sigma}}{T} (\widetilde{\mathbf{U}\mathbf{V}})^T \mathbf{R}. \quad (\text{C47})$$

To compute the spectrum of this matrix, we use the following trick. The singular values of \mathbf{M} are equal to the square roots of the nonzero eigenvalues of its square,

$$\mathbf{M}^2 = \mathbf{M}\mathbf{M}^T = \frac{\tilde{\sigma}^2}{T^2} (\widetilde{\mathbf{U}\mathbf{V}})^T \mathbf{R}\mathbf{R}^T \widetilde{\mathbf{U}\mathbf{V}}. \quad (\text{C48})$$

In turn, the nonzero eigenvalues of this $N \times N$ matrix, are equal to the eigenvalues of the $T \times T$ matrix,

$$\widehat{\mathbf{M}}^2 \equiv q^2 \tilde{\sigma}^2 \frac{\mathbf{R}\mathbf{R}^T}{N} \frac{(\widetilde{\mathbf{U}\mathbf{V}})(\widetilde{\mathbf{U}\mathbf{V}})^T}{N}. \quad (\text{C49})$$

We note that the right-hand side of the above equation is a product of the $T \times T$ dual correlation matrices $\mathbf{C}_{\tilde{\sigma}\mathbf{R}^T}$ and $\mathbf{C}_{(\widetilde{\mathbf{U}\mathbf{V}})^T}$. To compute the spectrum of the product, we employ the \mathcal{S} -transform formalism as explained above. The first step is to obtain the \mathcal{S} -transforms of the dual correlation matrices, which we compute from the Stieltjes transform [39]. For the noise part, we have

$$\mathfrak{g}_{\mathbf{C}_{\mathbf{R}^T}}(z) = q^2 \mathfrak{g}_{\mathbf{C}_{\mathbf{R}}}(qz) + \frac{1-q}{z}. \quad (\text{C50})$$

For the signal part, we have

$$\mathfrak{g}_{\mathbf{C}_{(\widetilde{\mathbf{U}\mathbf{V}})^T}}(z) = q^2 \mathfrak{g}_{\mathbf{C}_{\widetilde{\mathbf{U}\mathbf{V}}}}(qz) + \frac{1-q}{z}. \quad (\text{C51})$$

For the noise Wishart matrix, the Stieltjes transform is [cf. Eq. (C20)],

$$\mathcal{S}_{\mathbf{C}_{\mathbf{R}^T}} = \frac{1}{1 + q^{-1}t}. \quad (\text{C52})$$

Including the renormalized noise strength $\tilde{\sigma}$ and the additional factor of q from Eq. (C49), by using the scaling relation Eq. (C19), the \mathcal{S} -transform is

$$\mathcal{S}_{\mathbf{C}_{q\tilde{\sigma}\mathbf{R}^T}} = \frac{q^{-2}\tilde{\sigma}^{-2}}{1 + q^{-1}t}. \quad (\text{C53})$$

Next, we write the \mathcal{S} -transform of the pure signal part. Evaluating Eq. (C51) using Eq. (C30), we find

$$\mathfrak{g}_{\mathbf{C}_{(\widetilde{\mathbf{U}\mathbf{V}})^T}} = \frac{\frac{2-q-qq_{VT}}{\tilde{\sigma}_X^2} + q^2 q_{VT} z \pm qq_{VT} \sqrt{(qz - \lambda_+^\infty)(qz - \lambda_-^\infty)}}{2z\tilde{\sigma}_X^{-2}}, \quad (\text{C54})$$

from which we obtain the following equation for the Stieltjes transform:

$$\left(2z\tilde{\sigma}_X^{-2} \mathfrak{g}_{\mathbf{C}_{(\widetilde{\mathbf{U}\mathbf{V}})^T}} - \frac{2-q-qq_{VT}}{\tilde{\sigma}_X^2} - q^2 q_{VT} z \right)^2 - q^2 q_{VT}^2 (qz - \lambda_+^\infty)(qz - \lambda_-^\infty) = 0. \quad (\text{C55})$$

Using the relation, $\mathcal{T} = z\mathfrak{g} - 1$, we find the equation for the \mathcal{T} -transform,

$$\mathcal{T}_{\mathbf{C}_{(\widetilde{\mathbf{U}\mathbf{V}})^T}}^2 + \mathcal{T}_{\mathbf{C}_{(\widetilde{\mathbf{U}\mathbf{V}})^T}} (-zq^2 q_{VT} \tilde{\sigma}_X^2 + qq_{VT} + q) + q^2 q_{VT} = 0. \quad (\text{C56})$$

Interpreted as an equation for the functional inverse transform \mathcal{T}^{-1} , this becomes

$$t^2 + t(-\mathcal{T}_{\mathbf{C}_{(\widetilde{\mathbf{U}\mathbf{V}})^T}}^{-1} q^2 q_{VT} \tilde{\sigma}_X^2 + qq_{VT} + q) + q^2 q_{VT} = 0. \quad (\text{C57})$$

Now solving for the functional inverse transform, we find

$$\mathcal{T}_{\mathbf{C}_{(\widetilde{\mathbf{U}\mathbf{V}})^T}}^{-1} = \frac{(t+q)(t+qu)}{tqq_U \tilde{\sigma}_X^2}, \quad (\text{C58})$$

from which we determine the \mathcal{S} -transform,

$$\mathcal{S}_{\mathbf{C}_{\widehat{\mathbf{U}}\widehat{\mathbf{V}}}} = \frac{t+1}{t\mathcal{T}_{\mathbf{C}_{\widehat{\mathbf{U}}\widehat{\mathbf{V}}}}^{-1}} = \frac{qq_U\bar{\sigma}_X^2(t+1)}{(t+q)(t+q_U)}. \quad (\text{C59})$$

The \mathcal{S} -transform of the product now reads

$$\begin{aligned} \mathcal{S}_{\widehat{\mathbf{M}}^2} &= \mathcal{S}_{\mathbf{C}_{\widehat{\mathbf{U}}\widehat{\mathbf{V}}}} \mathcal{S}_{\mathbf{C}_{q\bar{\sigma}\mathbf{R}}} \\ &= \frac{t+1}{t} \frac{tqq_U\bar{\sigma}_X^2}{(t+q)(t+q_U)} \frac{q^{-2}\bar{\sigma}^{-2}}{1+q^{-1}t}. \end{aligned} \quad (\text{C60})$$

From this we read off the inverse transform,

$$\begin{aligned} \mathcal{T}_{\widehat{\mathbf{M}}^2}^{-1} &= \frac{(t+q)(t+q_U)(1+q^{-1}t)}{tq^{-1}q_U\bar{\sigma}_X^2\bar{\sigma}^{-2}} \\ &= \frac{(t+q)(t+q_U)(1+q^{-1}t)}{tq^{-1}q_U\text{SNR}^{-1}}. \end{aligned} \quad (\text{C61})$$

The equation for the \mathcal{T} -transform is now

$$(\mathcal{T}_{\widehat{\mathbf{M}}^2} + q)(\mathcal{T}_{\widehat{\mathbf{M}}^2} + q_U)(1+q^{-1}\mathcal{T}_{\widehat{\mathbf{M}}^2}) = \frac{zq_U\mathcal{T}_{\widehat{\mathbf{M}}^2}}{q\text{SNR}}. \quad (\text{C62})$$

Using $\mathcal{T} = zg - 1$ we write down the cubic polynomial equation for g ,

$$ag_{\widehat{\mathbf{M}}^2}^3 + bg_{\widehat{\mathbf{M}}^2}^2 + cg_{\widehat{\mathbf{M}}^2} + d = 0, \quad (\text{C63})$$

with coefficients,

$$a = z^3, \quad (\text{C64})$$

$$b = 2qz^2 + q_Uz^2 - 3z^2, \quad (\text{C65})$$

$$c = q^2z + 2qq_Uz - 4qz - 2q_Uz + 3z - \frac{q_Uz^2}{\text{SNR}}, \quad (\text{C66})$$

$$d = q^2q_U - q^2 - 2qq_U + 2q + q_U - 1 + \frac{q_Uz}{\text{SNR}}. \quad (\text{C67})$$

The eigenvalue density is nonzero for complex solutions of the equation. The equation admits complex solutions when the discriminant Δ is negative,

$$\Delta = 4P^3 + 27Q^2, \quad (\text{C68})$$

where

$$P = \frac{3ac - b^2}{3a^2}, \quad (\text{C69})$$

$$Q = \frac{2b^2 - 9abc + 27a^2d}{27a^3}. \quad (\text{C70})$$

Written out explicitly, the determinant takes the form

$$\Delta = \frac{4(3ac - b^2)^3 + (2b^3 - 9abc + 27a^2d)^2}{27a^6}. \quad (\text{C71})$$

The equation $\Delta = 0$ yields a quadratic equation in z , giving the bounds on the eigenvalue density of $\widehat{\mathbf{M}}^2$,

$$z_{\pm} = \frac{8q^2 + 20qq_U - q_U^2 \pm \sqrt{q_U(8q + q_U)^3}}{8q_U\text{SNR}^{-1}}. \quad (\text{C72})$$

From the definitions of $q = N/T$ and $q_U = m/T$, we see that these bounds scale as

$$z_{\pm} \sim T^{-1}. \quad (\text{C73})$$

Since the singular values of \mathbf{M} are equal to the square root of the eigenvalues of $\widehat{\mathbf{M}}^2$, the eigenvalue bounds of the signal-noise cross terms, thus, scale as

$$\lambda_{\pm}^{\text{signal-noise}} \sim T^{-1/2}. \quad (\text{C74})$$

Thus, the contribution of the cross terms can be neglected in the classical statistics limit.

4. Adding the noise contribution $\mathbf{C}_{\bar{\sigma}\mathbf{R}}$

In Appendix C 2 we computed the spectrum of the pure signal contribution $\mathbf{C}_{\widehat{\mathbf{U}}\widehat{\mathbf{V}}}$ to the NECM in the classical statistics and the intensive limits. Now we will add the pure noise contribution $\mathbf{C}_{\bar{\sigma}\mathbf{R}}$ to obtain the spectrum of the approximate NECM. Since the noise and the signal contributions are free matrices with respect to each other, the spectrum of their sum can be computed using the \mathcal{R} -transform. The \mathcal{R} -transform of a random matrix \mathbf{A} is

$$\mathcal{R}_{\mathbf{A}}(z) = \mathcal{B}_{\mathbf{A}}(z) - 1/z, \quad (\text{C75})$$

where the \mathcal{B} transform is the functional inverse of the Stieltjes transform,

$$\mathcal{B}_{\mathbf{A}}[\mathfrak{g}_{\mathbf{A}}] = z. \quad (\text{C76})$$

The \mathcal{R} -transform is additive for free matrices,

$$\mathcal{R}_{\mathbf{A}+\mathbf{B}}(z) = \mathcal{R}_{\mathbf{A}}(z) + \mathcal{R}_{\mathbf{B}}(z). \quad (\text{C77})$$

It scales according to

$$\mathcal{R}_{a\mathbf{A}}(z) = a\mathcal{R}_{\mathbf{A}}(az), \quad (\text{C78})$$

where a is a real number. For a white Wishart matrix Eq. (C10), the \mathcal{R} -transform is known to be [39]

$$\mathcal{R}_{\mathbf{W}_Y}(z) = \frac{1}{1 - qz}. \quad (\text{C79})$$

For the pure noise contribution to the NECM, $\mathbf{C}_{\bar{\sigma}\mathbf{R}} \equiv \bar{\sigma}^2\mathbf{R}^T\mathbf{R}/T$, this results in

$$\mathcal{R}_{\mathbf{C}_{\bar{\sigma}\mathbf{R}}}(z) = \frac{\bar{\sigma}^2}{1 - qz\bar{\sigma}^2}. \quad (\text{C80})$$

Our goal is to first obtain the \mathcal{R} -transform of the sum of the signal and the noise contributions,

$$\mathcal{R}_{\mathbf{C}}(z) = \mathcal{R}_{\mathbf{C}_{\widehat{\mathbf{U}}\widehat{\mathbf{V}}}}(z) + \mathcal{R}_{\mathbf{C}_{\bar{\sigma}\mathbf{R}}}(z), \quad (\text{C81})$$

and from this to compute the Stieltjes transform to extract the eigenvalue density. Computing the \mathcal{R} -transform of the pure signal contribution $\mathbf{C}_{\widehat{\mathbf{U}}\widehat{\mathbf{V}}}$ in the classical statistics and the intensive limit requires additional steps.

a. Classical statistics limit

First, we compute the \mathcal{R} -transform of $\mathfrak{g}_{\mathbf{C}_{\widehat{\mathbf{U}}\widehat{\mathbf{V}}}}$. This is performed by solving the functional inverse

equation $g_{C_{\bar{v}v}}[\mathcal{B}_{C_{\bar{v}v}}] = z$, which gives us the \mathcal{B} transform from which we compute the \mathcal{R} -transform. From Eq. (C30), we see that \mathcal{B} transform satisfies

$$\mathcal{B}_{C_{\bar{v}v}}[\mathcal{B}_{C_{\bar{v}v}}z(-q_{V^T}\bar{\sigma}_X^2 + z) + q_{V^T}\bar{\sigma}_X^2 + q_{V^T}z - z] = 0. \quad (\text{C82})$$

A nontrivial solution of this equation is

$$\mathcal{B}_{C_{\bar{v}v}} = \frac{q_{V^T}\bar{\sigma}_X^2 + q_{V^T}z - z}{z(q_{V^T}\bar{\sigma}_X^2 - z)}. \quad (\text{C83})$$

Using Eq. (C75), this gives us the \mathcal{R} -transform $\mathcal{R}_{C_{\bar{v}v}}$. To it we add the \mathcal{R} -transform of the noise Eq. (C80) and subtract $-1/z$ to get the \mathcal{B} transform of the approximate NECM,

$$\mathcal{B}_C = \frac{\tilde{\sigma}^2z(q_{V^T}\bar{\sigma}_X^2 - z) + (-q\tilde{\sigma}^2z + 1)(q_{V^T}\bar{\sigma}_X^2 + q_{V^T}z - z)}{z(q_{V^T}\bar{\sigma}_X^2 - z)(-q\tilde{\sigma}^2z + 1)}. \quad (\text{C84})$$

The final step is to write down and solve the inverse function equation $\mathcal{B}_C[g_C] = z$. This is now equivalent to solving the third order polynomial equation,

$$ag_C^3 + bg_C^2 + cg_C + d = 0, \quad (\text{C85})$$

with

$$a = qz\tilde{\sigma}^2, \quad (\text{C86})$$

$$b = -qq_{V^T}z\bar{\sigma}_X^2\tilde{\sigma}^2 + [(q_{V^T} - 1)q + 1]\tilde{\sigma}^2 - z, \quad (\text{C87})$$

$$c = (q - 1)q_{V^T}\bar{\sigma}_X^2\tilde{\sigma}^2 + q_{V^T}z\bar{\sigma}_X^2 - q_{V^T} + 1, \quad (\text{C88})$$

$$d = -q_{V^T}\bar{\sigma}_X^2. \quad (\text{C89})$$

Written in terms of the SNR the coefficients take the form

$$a = \frac{qz}{1 + \text{SNR}}, \quad (\text{C90})$$

$$b = -\frac{qq_{V^T}z}{\text{SNR}} + \frac{(q_{V^T} - 1)q + 1}{1 + \text{SNR}} - z, \quad (\text{C91})$$

$$c = \frac{(q - 1)q_{V^T}}{\text{SNR}} + q_{V^T}z(1 + \text{SNR}^{-1}) - q_{V^T} + 1, \quad (\text{C92})$$

$$d = -q_{V^T}(1 + \text{SNR}^{-1}). \quad (\text{C93})$$

It is possible to solve this cubic equation analytically. However, the expressions become lengthy and provide little insight. Therefore, we rely on the numerical solution of the equation as shown in Fig. 2 as well as on the following analyses in the limits of small and large noise.

First, in the limit of the pure signal $\text{SNR} \rightarrow \infty$, we recover Eq. (20). Similarly, by truncating the polynomial coefficients in the pure noise limit $\text{SNR} \rightarrow 0$ at order $O(\text{SNR}^{-1})$, the MP density is recovered.

We also derive an approximate analytic expression for the bounds of the eigenvalue density $\lambda_{\pm}^{\text{SNR}}$, which is valid around both noise limits $\text{SNR} \rightarrow 0$ and $\text{SNR} \rightarrow \infty$. For this, we approximate the coefficients of the polynomial, noting that the smallest contribution to the coefficients common to both limits comes from terms of order $O[q/(1 + \text{SNR})]$ (recall that $q \rightarrow 0$ in the classical limit). Neglecting these terms leads to a quadratic polynomial equation for the Stieltjes transform,

$$rg_C^2 + sg_C + t \approx 0, \quad (\text{C94})$$

with

$$r = -\frac{qq_{V^T}z}{\text{SNR}} + \frac{1}{1 + \text{SNR}} - z, \quad (\text{C95})$$

$$s = \frac{(q - 1)q_{V^T}}{\text{SNR}} + q_{V^T}z(1 + \text{SNR}^{-1}) - q_{V^T} + 1, \quad (\text{C96})$$

$$t = -q_{V^T}(1 + \text{SNR}^{-1}). \quad (\text{C97})$$

The approximate bounds of the eigenvalue density then are given by the roots of the discriminant $\Delta g_C \approx s^2 - 4rt$, which gives the following bounds for the nonzero range of the eigenvalue density:

$$\begin{aligned} \lambda_{\pm}^{\text{SNR}} &\approx \frac{1 + q_{V^T}^{-1}}{1 + \text{SNR}^{-1}} + \frac{1 + q}{1 + \text{SNR}} \pm 2\sqrt{\frac{q_{V^T}^{-1}}{(1 + \text{SNR}^{-1})^2} + \frac{q}{(1 + \text{SNR})^2} + \frac{q}{(\sqrt{\text{SNR}} + \sqrt{\text{SNR}^{-1}})^2}} \\ &\approx \frac{1 + q_{V^T}^{-1}}{1 + \text{SNR}^{-1}} + \frac{1 + q}{1 + \text{SNR}} \pm 2\sqrt{\frac{q_{V^T}^{-1}}{(1 + \text{SNR}^{-1})^2} + \frac{q}{(1 + \text{SNR})^2}} \\ &= \frac{1 + q_{V^T}^{-1}}{1 + \text{SNR}^{-1}} + \frac{1 + q}{1 + \text{SNR}} \pm 2\sqrt{\left[\sqrt{\frac{q_{V^T}^{-1}}{(1 + \text{SNR}^{-1})^2}} + \sqrt{\frac{q}{(1 + \text{SNR})^2}}\right]^2 - \frac{2\sqrt{qq_{V^T}^{-1}}}{(1 + \text{SNR}^{-1})(1 + \text{SNR})}} \\ &\approx \frac{1 + q_{V^T}^{-1}}{1 + \text{SNR}^{-1}} + \frac{1 + q}{1 + \text{SNR}} \pm 2\left[\sqrt{\frac{q_{V^T}^{-1}}{(1 + \text{SNR}^{-1})^2}} + \sqrt{\frac{q}{(1 + \text{SNR})^2}}\right] \\ &= \frac{1}{1 + \text{SNR}^{-1}}\lambda_{\pm}^{\infty} + \frac{1}{1 + \text{SNR}}\lambda_{\pm}^{\text{MP}}. \end{aligned} \quad (\text{C98})$$

In the second line, we have dropped the third term under the square root since it is small in either of the two noise limits. In the third line, we have used $(\sqrt{a} + \sqrt{b})^2 = a + b + 2\sqrt{ab}$, and in the fourth line, we dropped the last term under the square root since it is also small in either of the two noise limits. In the final line we recognize that the terms form the weighted average of λ_{\pm}^{∞} Eq. (21) and the Marčenko-Pastur bounds $\lambda_{\pm}^{\text{MP}}$.

b. Intensive limit

First we compute the \mathcal{R} -transform of $\mathbf{g}_{\mathcal{C}_{\overline{\text{IV}}}}$. This is obtained by solving the functional inverse equation $\mathbf{g}_{\mathcal{C}_{\overline{\text{IV}}}}[\mathcal{B}_{\mathcal{C}_{\overline{\text{IV}}}}] = z$, which gives us the \mathcal{B} transform, from which we compute the \mathcal{R} -transform. To do this, we employed the symbolic algebra Python library SymPy v1.6.2. The \mathcal{B} transform satisfies the quadratic equation,

$$\begin{aligned} & \mathcal{B}_{\mathcal{C}_{\overline{\text{IV}}}}^2 (q^2 z - qq_{V^T} \bar{\sigma}_X^2 + 2qz - q_{V^T} \bar{\sigma}_X^2 + z) \\ & + \mathcal{B}_{\mathcal{C}_{\overline{\text{IV}}}} (q^2 q_{V^T} z - 2q^2 z + qq_{V^T} \bar{\sigma}_X^2 + 2qq_{V^T} z - 3qz + q_{V^T} \bar{\sigma}_X^2 + q_{V^T} z - z) \\ & - q^2 q_{V^T} + q^2 - qq_{V^T} + q = 0, \end{aligned} \quad (\text{C99})$$

for which we find the solution,

$$\begin{aligned} \mathcal{B}_{\mathcal{C}_{\overline{\text{IV}}}} = & \left(-qq_{V^T} z + 2qz - q_{V^T} \bar{\sigma}_X^2 - q_{V^T} z + z \right. \\ & \left. - \sqrt{q^2 q_{V^T}^2 z^2 - 2qq_{V^T}^2 \bar{\sigma}_X^2 z + 2qq_{V^T}^2 z^2 - 2qq_{V^T} z^2 + q_{V^T}^2 \bar{\sigma}_X^4 + 2q_{V^T}^2 \bar{\sigma}_X^2 z + q_{V^T}^2 z^2 - 2q_{V^T} \bar{\sigma}_X^2 z - 2q_{V^T} z^2 + z^2} \right) / \\ & [2z(qz - q_{V^T} \bar{\sigma}_X^2 + z)]. \end{aligned} \quad (\text{C100})$$

Using Eq. (C75), this gives us $\mathcal{R}_{\mathcal{C}_{\overline{\text{IV}}}}$ to which we add the \mathcal{R} -transform of the noise Eq. (C80) to obtain the \mathcal{R} -transform $\mathcal{R}_{\mathcal{C}}$ of the NECM. Subtracting $-1/z$, gives us the corresponding form of the \mathcal{B} transform,

$$\begin{aligned} \mathcal{B}_{\mathcal{C}} = & [2\bar{\sigma}^2 z(qz - q_{V^T} \bar{\sigma}_X^2 + z) + (-q\bar{\sigma}^2 z + 1)(-qq_{V^T} z + 2qz - q_{V^T} \bar{\sigma}_X^2 - q_{V^T} z + z \\ & - \sqrt{q^2 q_{V^T}^2 z^2 - 2qq_{V^T}^2 \bar{\sigma}_X^2 z + 2qq_{V^T}^2 z^2 - 2qq_{V^T} z^2 + q_{V^T}^2 \bar{\sigma}_X^4 + 2q_{V^T}^2 \bar{\sigma}_X^2 z + q_{V^T}^2 z^2 - 2q_{V^T} \bar{\sigma}_X^2 z - 2q_{V^T} z^2 + z^2})] / \\ & [2z(-q\bar{\sigma}^2 z + 1)(qz - q_{V^T} \bar{\sigma}_X^2 + z)]. \end{aligned} \quad (\text{C101})$$

The final step is to write down and solve the inverse functional equation $\mathcal{B}_{\mathcal{C}}[\mathbf{g}_{\mathcal{C}}] = z$. The sixth order polynomial equation that we need to solve is of the form

$$a\mathbf{g}_{\mathcal{C}}^6 + b\mathbf{g}_{\mathcal{C}}^5 + c\mathbf{g}_{\mathcal{C}}^4 + d\mathbf{g}_{\mathcal{C}}^3 + e\mathbf{g}_{\mathcal{C}}^2 + f\mathbf{g}_{\mathcal{C}} + g = 0, \quad (\text{C102})$$

with coefficients,

$$a = q^5 \bar{\sigma}^6 z^2 + 2q^4 \bar{\sigma}^6 z^2 + q^3 \bar{\sigma}^6 z^2, \quad (\text{C103})$$

$$\begin{aligned} b = & q^5 q_{V^T} \bar{\sigma}^6 z - 2q^5 \bar{\sigma}^6 z - 2q^4 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^6 z^2 + 2q^4 q_{V^T} \bar{\sigma}^6 z - q^4 \bar{\sigma}^6 z - 3q^4 \bar{\sigma}^4 z^2 \\ & - 2q^3 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^6 z^2 + q^3 q_{V^T} \bar{\sigma}^6 z + 3q^3 \bar{\sigma}^6 z - 6q^3 \bar{\sigma}^4 z^2 + 2q^2 \bar{\sigma}^6 z - 3q^2 \bar{\sigma}^4 z^2, \end{aligned} \quad (\text{C104})$$

$$\begin{aligned} c = & -q^5 q_{V^T} \bar{\sigma}^6 + q^5 \bar{\sigma}^6 - q^4 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^6 z + 3q^4 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^6 z - 3q^4 q_{V^T} \bar{\sigma}^4 z \\ & - q^4 \bar{\sigma}^6 + 6q^4 \bar{\sigma}^4 z + q^3 q_{V^T}^2 \bar{\sigma}_X^4 \bar{\sigma}^6 z^2 \\ & - q^3 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^6 z - 2q^3 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^6 z + 6q^3 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^4 z^2 + 2q^3 q_{V^T} \bar{\sigma}^6 - 6q^3 q_{V^T} \bar{\sigma}^4 z \\ & - 2q^3 \bar{\sigma}^6 + 5q^3 \bar{\sigma}^4 z + 3q^3 \bar{\sigma}^2 z^2 \\ & - 4q^2 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^6 z + 6q^2 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^4 z^2 + q^2 q_{V^T} \bar{\sigma}^6 - 3q^2 q_{V^T} \bar{\sigma}^4 z + q^2 \bar{\sigma}^6 \\ & - 5q^2 \bar{\sigma}^4 z + 6q^2 \bar{\sigma}^2 z^2 + q\bar{\sigma}^6 - 4q\bar{\sigma}^4 z + 3q\bar{\sigma}^2 z^2, \end{aligned} \quad (\text{C105})$$

$$\begin{aligned} d = & q^4 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^6 - q^4 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^6 + 3q^4 q_{V^T} \bar{\sigma}^4 - 3q^4 \bar{\sigma}^4 - q^3 q_{V^T}^2 \bar{\sigma}_X^4 \bar{\sigma}^6 z - q^3 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^6 \\ & + 3q^3 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^4 z + 3q^3 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^6 - 9q^3 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^4 z + q^3 q_{V^T} \bar{\sigma}^4 + 3q^3 q_{V^T} \bar{\sigma}^2 z + q^3 \bar{\sigma}^4 \\ & - 6q^3 \bar{\sigma}^2 z + 2q^2 q_{V^T}^2 \bar{\sigma}_X^4 \bar{\sigma}^6 z - 3q^2 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^4 z^2 - q^2 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^6 + 3q^2 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^4 z \\ & + 2q^2 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^4 z - 6q^2 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^2 z^2 - 4q^2 q_{V^T} \bar{\sigma}^4 \\ & + 6q^2 q_{V^T} \bar{\sigma}^2 z + 5q^2 \bar{\sigma}^4 - 7q^2 \bar{\sigma}^2 z - q^2 z^2 - 2qq_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^6 + 8qq_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^4 z \\ & - 6qq_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^2 z^2 - 2qq_{V^T} \bar{\sigma}^4 \\ & + 3qq_{V^T} \bar{\sigma}^2 z + q\bar{\sigma}^2 z - 2qz^2 - \bar{\sigma}^4 + 2\bar{\sigma}^2 z - z^2, \end{aligned} \quad (\text{C106})$$

$$\begin{aligned} e = & -3q^3 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^4 + 3q^3 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^4 - 3q^3 q_{V^T} \bar{\sigma}^2 + 3q^3 \bar{\sigma}^2 - q^2 q_{V^T}^2 \bar{\sigma}_X^4 \bar{\sigma}^6 + 3q^2 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^4 z \\ & + 2q^2 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^4 - 3q^2 q_{V^T}^2 \bar{\sigma}_X^2 \bar{\sigma}^2 z - 6q^2 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^4 + 9q^2 q_{V^T} \bar{\sigma}_X^2 \bar{\sigma}^2 z \end{aligned}$$

$$\begin{aligned}
& -2q^2 q_{VT} \tilde{\sigma}^2 - q^2 q_{VT} z + q^2 \tilde{\sigma}^2 + 2q^2 z + qq_{VT}^2 \tilde{\sigma}_X^4 \tilde{\sigma}^6 \\
& -4qq_{VT}^2 \tilde{\sigma}_X^4 \tilde{\sigma}^4 z + 3qq_{VT}^2 \tilde{\sigma}_X^4 \tilde{\sigma}^2 z^2 + 2qq_{VT}^2 \tilde{\sigma}_X^2 \tilde{\sigma}^4 - 3qq_{VT}^2 \tilde{\sigma}_X^2 \tilde{\sigma}^2 z - 2qq_{VT} \tilde{\sigma}_X^2 \tilde{\sigma}^4 + 2qq_{VT} \tilde{\sigma}_X^2 \tilde{\sigma}^2 z \\
& + 2qq_{VT} \tilde{\sigma}_X^2 z^2 + 2qq_{VT} \tilde{\sigma}^2 - 2qq_{VT} z - 3q\tilde{\sigma}^2 + 3qz + 2q_{VT} \tilde{\sigma}_X^2 \tilde{\sigma}^4 - 4q_{VT} \tilde{\sigma}_X^2 \tilde{\sigma}^2 z \\
& + 2q_{VT} \tilde{\sigma}_X^2 z^2 + q_{VT} \tilde{\sigma}^2 - q_{VT} z - \tilde{\sigma}^2 + z,
\end{aligned} \tag{C107}$$

$$\begin{aligned}
f &= 3q^2 q_{VT}^2 \tilde{\sigma}_X^2 \tilde{\sigma}^2 - 3q^2 q_{VT} \tilde{\sigma}_X^2 \tilde{\sigma}^2 + q^2 q_{VT} - q^2 + 2qq_{VT}^2 \tilde{\sigma}_X^4 \tilde{\sigma}^4 - 3qq_{VT}^2 \tilde{\sigma}_X^4 \tilde{\sigma}^2 z - qq_{VT}^2 \tilde{\sigma}_X^2 \tilde{\sigma}^2 + qq_{VT}^2 \tilde{\sigma}_X^2 z \\
& + 3qq_{VT} \tilde{\sigma}_X^2 \tilde{\sigma}^2 - 3qq_{VT} \tilde{\sigma}_X^2 z + qq_{VT} - q - q_{VT}^2 \tilde{\sigma}_X^4 \tilde{\sigma}^4 + 2q_{VT}^2 \tilde{\sigma}_X^2 \tilde{\sigma}^2 z - q_{VT}^2 \tilde{\sigma}_X^4 z^2 \\
& - q_{VT}^2 \tilde{\sigma}_X^2 \tilde{\sigma}^2 + q_{VT}^2 \tilde{\sigma}_X^2 z + 2q_{VT} \tilde{\sigma}_X^2 \tilde{\sigma}^2 - 2q_{VT} \tilde{\sigma}_X^2 z,
\end{aligned} \tag{C108}$$

$$g = -qq_{VT}^2 \tilde{\sigma}_X^2 + qq_{VT} \tilde{\sigma}_X^2 - q_{VT}^2 \tilde{\sigma}_X^4 \tilde{\sigma}^2 + q_{VT}^2 \tilde{\sigma}_X^4 z. \tag{C109}$$

We solve this polynomial equation numerically, looking for complex roots which yield nonzero values of the eigenvalue density. For large signal-to-noise ratio, we encounter numerical instabilities when trying to determine the eigenvalue density bounds. We run into these instabilities in the determination of the true density bounds in Fig. 2(b) bottom plot. To fix this, we start at the peak of the density and determine the values of λ for which the density hits zero for the first time to either side of the peak. All other zero density crossings are assumed to be due to numerical instabilities.

The ranges of nonzero density are shown in Fig. 2(b) bottom plot. We see that as the signal-to-noise ratio increases, there is a bifurcation point where the density splits into two bumps. The left bump is associated with the noise, and the

right bump is associated with the pure latent feature signal. From our approximate expression for the eigenvalue bounds Eq. (22), we can estimate the value of the SNR at which the splitting occurs. The defining equation for this is given by the intersection between the right boundary of the noise region and the left boundary of the signal part of the density in Eq. (22), resulting in the condition,

$$(1 + \text{SNR}^{-1})\lambda_+^{\text{MP}} = \lambda_-^{\text{SNR}}. \tag{C110}$$

Solving for SNR, we obtain the following estimation for the splitting point:

$$\text{SNR}_{\text{split}} \approx \frac{\lambda_+^{\text{MP}} - \lambda_-^{\text{MP}}}{\lambda_-^{\infty}}. \tag{C111}$$

-
- [1] C. Killer, T. Bockwoldt, S. Schütt, M. Himpel, A. Melzer, and A. Piel, Phase Separation of Binary Charged Particle Systems with Small Size Disparities using a Dusty Plasma, *Phys. Rev. Lett.* **116**, 115002 (2016).
- [2] M. T. Valentine, P. D. Kaplan, D. Thota, J. C. Crocker, T. Gisler, R. K. Prud'homme, M. Beck, and D. A. Weitz, Investigating the microenvironments of inhomogeneous soft materials with multiple particle tracking, *Phys. Rev. E* **64**, 061506 (2001).
- [3] M. Sinhuber, K. Van Der Vaart, R. Ni, J. G. Puckett, D. H. Kelley, and N. T. Ouellette, Three-dimensional time-resolved trajectories from laboratory insect swarms, *Sci. Data* **6**, 190036 (2019).
- [4] B. Lusch, J. N. Kutz, and S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, *Nat. Commun.* **9**, 4950 (2018).
- [5] D. Schanz, S. Gesemann, and A. Schröder, Shake-the-box: Lagrangian particle tracking at high particle image densities, *Exp. Fluids* **57**, 70 (2016).
- [6] NOAA Physical Sciences Laboratory, Gridded climate data, <https://psl.noaa.gov/data/gridded/>, accessed: 2021-06-30.
- [7] J. L. Natale, D. Hofmann, D. G. Hernández, and I. Nemenman, Reverse-engineering biological networks from large data sets, in *Quantitative Biology: Theory, Computational Methods and Examples of Models*, edited by B. Munsky, L. Tsimring, and W. S. Hlavacek (MIT Press, Cambridge, MA, 2018).
- [8] L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, Collective behavior of place and non-place neurons in the hippocampal network, *Neuron* **96**, 1178 (2017).
- [9] G. J. Stephens, B. Johnson-Kerner, W. Bialek, and W. S. Ryu, Dimensionality and dynamics in the behavior of *c. elegans*, *PLoS Comput Biol* **4**, e1000028 (2008).
- [10] G. J. Berman, D. M. Choi, W. Bialek, and J. W. Shaevitz, Mapping the stereotyped behaviour of freely moving fruit flies, *J. R. Soc. Interface* **11**, 20140672 (2014).
- [11] W. Weisser, C. Roscher, S. Meyer, A. Ebeling, G. Luo, E. Allan, H. Bessler, R. Barnard, N. Buchmann, F. Buscot *et al.*, Biodiversity effects on ecosystem functioning in a 15-year grassland experiment: Patterns, mechanisms, and open questions, *Basic Appl. Ecol.* **23**, 1 (2017).
- [12] A. I. Dell, J. A. Bender, K. Branson, I. D. Couzin, G. G. de Polavieja, L. P. Noldus, A. Pérez-Escudero, P. Perona, A. D. Straw, M. Wikelski, and U. Brose, Automated image-based tracking and its application in ecology, *Trends in Ecology & Evolution* **29**, 417 (2014).
- [13] C. Pandarinath, D. J. O'Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg *et al.*, Inferring single-trial neural population dynamics using sequential auto-encoders, *Nat. Methods* **15**, 805 (2018).
- [14] D. J. Schwab, I. Nemenman, and P. Mehta, Zipf's Law and Criticality in Multivariate Data without Fine-Tuning, *Phys. Rev. Lett.* **113**, 068102 (2014).

- [15] M. C. Morrell, A. J. Sederberg, and I. Nemenman, Latent Dynamical Variables Produce Signatures of Spatiotemporal Criticality in Large Biological Systems, *Phys. Rev. Lett.* **126**, 118302 (2021).
- [16] J. A. Gallego, M. G. Perich, L. E. Miller, and S. A. Solla, Neural manifolds for the control of movement, *Neuron* **94**, 978 (2017).
- [17] J. Page, M. P. Brenner, and R. R. Kerswell, Revealing the state space of turbulence using machine learning, *Phys. Rev. Fluids* **6**, 034402 (2021).
- [18] E. H. Nieh, M. Schottdorf, N. W. Freeman, R. J. Low, S. Lewallen, S. A. Koay, L. Pinto, J. L. Gauthier, C. D. Brody, and D. W. Tank, Geometry of abstract learned knowledge in the hippocampus, *Nature (London)* **595**, 80 (2021).
- [19] J. Shlens, A tutorial on principal component analysis, [arXiv:1404.1100](https://arxiv.org/abs/1404.1100).
- [20] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, UK, 2020).
- [21] V. A. Marčenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, *Math. USSR Sb.* **1**, 457 (1967).
- [22] A. M. Sengupta and P. P. Mitra, Distributions of singular values for some random matrices, *Phys. Rev. E* **60**, 3389 (1999).
- [23] 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).
- [24] 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).
- [25] M. Capitaine and C. Donati-Martin, Spectrum of deformed random matrices and free probability, [arXiv:1607.05560](https://arxiv.org/abs/1607.05560).
- [26] R. Salakhutdinov and A. Mnih, Probabilistic matrix factorization, in *Proceedings of the 20th International Conference on Neural Information Processing Systems, NIPS'07, Vancouver, BC, Canada, 2007* (Curran Associates, Red Hook, NY, 2007), pp. 1257–1264.
- [27] R. Grosse, R. R. Salakhutdinov, W. T. Freeman, and J. B. Tenenbaum, Exploiting compositionality to explore a large space of model structures, in *28th Conference on Uncertainty in Artificial Intelligence, Catalina Island, 2012* (AUAI, Arlington, VA, 2012), pp. 306–315.
- [28] L. Meshulam, J. L. Gauthier, C. D. Brody, D. W. Tank, and W. Bialek, Coarse Graining, Fixed Points, and Scaling in a Large Population of Neurons, *Phys. Rev. Lett.* **123**, 178103 (2019).
- [29] H. Hotelling, New light on the correlation coefficient and its transforms, *J. R. Stat. Society: Ser. B: (Methodological)* **15**, 193 (1953).
- [30] H. (<https://math.stackexchange.com/users/6460/henry>), Sum of n i.i.d Beta-distributed variables, Mathematics Stack Exchange, <https://math.stackexchange.com/q/3096929> (version: 2019-02-02).
- [31] Marčenko-Pastur (MP) distribution, $\rho^{\text{MP}}(\lambda) = \sqrt{(\lambda - \lambda_-^{\text{MP}})(\lambda_+^{\text{MP}} - \lambda)} / (2\pi q\lambda)$ with $\lambda_{\pm}^{\text{MP}} = (1 \pm \sqrt{q})^2$, Ref. [20].
- [32] W. B. Johnson and J. Lindenstrauss, Extensions of lipschitz mappings into a hilbert space 26, *Contemporary mathematics* **26**, 28 (1984).
- [33] N. Halko, P.-G. Martinsson, and J. A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Rev.* **53**, 217 (2011).
- [34] R. Phillips, N. M. Belliveau, G. Chure, H. G. Garcia, M. Razo-Mejia, and C. Scholes, Figure 1 theory meets Figure 2 experiments in the study of gene expression, *Ann. Rev. Biophys.* **48**, 121 (2019).
- [35] M. D. Wang, M. Nicodemi, N. H. Dekker, T. Gregor, D. Holcman, A. M. V. Oijen, and S. Manley, Physics meets biology: The joining of two forces to further our understanding of cellular function, *Mol. Cell* **81**, 3033 (2021).
- [36] J. Wishart and M. S. Bartlett, The distribution of second order moment statistics in a normal system, *Math. Proc. Camb. Philos. Soc.* **28**, 455 (1932).
- [37] Wolfram Research, Inc., functions.wolfram.com, <http://functions.wolfram.com/03.04.21.0008.01>, accessed: 2021-06-30.
- [38] H. (<https://math.stackexchange.com/users/6460/henry>), Sum of n i.i.d Beta-distributed variables, Mathematics Stack Exchange, <https://math.stackexchange.com/q/3096929> (version: 2019-02-02).
- [39] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, UK, 2020).
- [40] D. V. Voiculescu, K. J. Dykema, and A. Nica, *Free Random Variables*, Vol. 1 (American Mathematical Soc., Providence, RI, 1992).
- [41] R. R. Müller, A random matrix model of communication via antenna arrays, *IEEE Trans. Inf. Theory* **48**, 2495 (2002).
- [42] Z. Burda, A. Jarosz, G. Livan, M. A. Nowak, and A. Swiech, Eigenvalues and singular values of products of rectangular gaussian random matrices, *Phys. Rev. E* **82**, 061114 (2010).
- [43] T. Dupic and I. P. Castillo, Spectral density of products of Wishart dilute random matrices. Part I: the dense case, [arXiv:1401.7802](https://arxiv.org/abs/1401.7802) [cond-mat.dis-nn].
- [44] G. Livan, M. Novaes, and P. Vivo, *Introduction to Random Matrices*, SpringerBriefs in Mathematical Physics, Vol. 26 (Springer, Cham, 2018).
- [45] S. Rabinowitz, How to find the square root of a complex number, *Mathematics and Informatics Quarterly* **3**, 54 (1993).
- [46] W. Cui, J. W. Rocks, and P. Mehta, The perturbative resolvent method: Spectral densities of random matrix ensembles via perturbation theory, [arXiv:2012.00663](https://arxiv.org/abs/2012.00663) [cond-mat.dis-nn].