

Hidden Data Structures and Statistics of Natural Images

Philipp Fleig ^{*}

Department of Cellular Biophysics, *Max Planck Institute for Medical Research*,
Jahnstraße 29, 69120 Heidelberg, Germany

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 13 June 2024; accepted 14 May 2025; published 12 June 2025)

Complex, high-dimensional data are often analyzed by grouping their constituent units into components, sometimes referred to as latent features, which afford physical or biological interpretation. However, *a priori* many different types of latent features and data decompositions can be defined, and one typically uses a trial-and-error approach to determine a decomposition that is natural to the system and its data. A principled understanding of which decomposition is appropriate for a given dataset is needed. In this work, we take a step in this direction and argue that sample-sample correlations carry important information to this effect. We construct a generative random latent feature matrix model of high-dimensional data based on linear mixing of latent features. As a key ingredient, we allow for a statistical dependence between the mixing coefficients. We argue that, with only two parameters (latent dimensionality and the structure of the statistical dependence), the resulting model captures characteristic properties found in many types of natural data. Specifically, the model can produce (overlapping) clusters, sparse mixing, and constrained (non-negative) mixing. We describe typical correlations and eigenvalue distributions of each pattern. Finally, we fit the model on correlation patterns in a benchmark natural images dataset, and we find a good match between the data and the sparse mixing regime of our model. This is in line with the well-known sparse coding structure in natural images, arguing that our approach identifies a correct latent decomposition of these data. We believe that our work will deliver similar insights for other diverse datasets.

DOI: [10.1103/8xd4-6hsr](https://doi.org/10.1103/8xd4-6hsr)

I. INTRODUCTION

Complex physical, biological, and social systems, characterized by numerous interacting units, often exhibit emergent collective properties [1]. To understand such systems, one typically relies on high-throughput experimental data. In such data, many variables (a.k.a., channels or degrees of freedom, such as neural activities, mutations, or pixel values) are measured in parallel. They are recorded across multiple samples, which may correspond to different time points, experimental trials, or distinct stimuli [2–5]. Considerable success in modeling such data has been achieved by the detailed study of correlations between the variables of the system [6–11]. In contrast, correlations among samples within these datasets have not been as extensively studied, and it remains less clear how they can unveil the hidden structures within the data. Nevertheless, several studies have argued that sample-

sample covariance can reveal crucial information about the data [12,13]. Similarly, statistical correlations among samples may determine the ability to learn from the sample [14].

Here, our primary focus is on analysis and modeling of sample-sample correlations within complex datasets. To this end, we develop a generative random latent feature model to capture the statistical properties of high-throughput datasets. Remarkably, with only a handful of parameters, our model qualitatively describes a wide variety of data patterns typically seen in experiments. Within our model, these patterns emerge from different ways of mixing latent features. Thus, we argue that the sample-sample correlations encode information on the nature of feature mixing and, as a result, about the suitable method for decomposing the data into features. As an example, we show how the model quantitatively describes statistical properties of image patches extracted from natural scenes, a classic dataset [15,16].

We consider data with N variables recorded over T samples. The data are high-dimensional, meaning $T, N \gg 1$. Generally, $T \sim N$, though modern experiments often push us to $N \gg T \gg 1$. Surprisingly, such high-dimensional data coming from natural systems are often simpler than they could have been in that they reveal an intrinsically lower-dimensional, latent structure. The number of *latent features* (a.k.a., collective degrees of freedom, which may be

^{*}Contact author: philipp.fleig@mr.mpg.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

externally imposed or emergent) present in each sample is often much smaller than N [9,17–28]. Importantly, we note that even if the total number m of features in the entire dataset is $m \sim N$ or even larger, individual data samples may still be low-dimensional, or sparse, in that only a few of the features may be combined to form each specific sample [16].

Apart from dimensionality, the structure of data is determined by *how* latent features are combined to form each datum. Here, we only explore *linear* mixing of features. Nonetheless, even for linear mixing, many different types of data can be observed, such as clusters [9], overlapping clusters [29], low-dimensional manifolds [30], and sparse data [16]. The richness of low-dimensional latent feature models is further increased since some types of data have constrained values: e.g., activities are often non-negative [31]. Ideally, one generative model should be able to reproduce as many of these data structures as possible under different values of model parameters.

A realistic way to build a generative model is to rely on a random latent feature model realized as $T \times N$ data matrices with elements taken from different probability distributions. Crucially, random matrix theory (RMT) can be used as a powerful tool to calculate analytical properties of such data [32]. A number of such random latent feature models have been described in a variety of contexts [30,32–36]. However, to the best of our knowledge, no model exists that captures the diversity of patterns between samples in the data described above.

It is therefore an open question how the distribution of observable quantities like the correlations between variables and between samples, or the eigenvalues of the correlation matrix (nonzero values of which are the same for variable-variable and sample-sample correlations) [32], depend on the properties of the model, and, in particular, on its latent dimensionality and the type of feature mixing. For instance, a gap in the eigenvalue spectrum often indicates a hidden low-dimensional structure due to the low-rank nature of the data. However, as pointed out above, for the case of sparse mixing of features, overall data need not have a low-rank structure, yet they still may possess a notion of low-dimensionality. With a generative random latent feature model at hand, precisely how observable distributions depend on model properties can be studied numerically or analytically.

Here, we present a random latent feature matrix model able to generate diverse types of mixing of latent features. The diversity is achieved by introducing statistically dependent mixing, which gives rise to distinct data patterns. We discuss the distribution of sample-sample correlations and the eigenvalue density in different regimes of the model. We show that, in the regime of sparse mixing of features, the model is in excellent agreement with the statistics of natural image patches. We explain how this fact points to the appropriate decomposition, or coding scheme, for the data.

II. THE MODEL

We consider data matrices $\mathbf{X}$ with N variables (columns) recorded in T samples, or observations (rows). We focus on models in which $\mathbf{X}$ is the product of three random matrices of the following form (note that $\odot$ denotes elementwise multi-

plication):

$$\mathbf{X} = (\mathbf{S} \odot \mathbf{U}) \mathbf{V} \text{ or, for the matrix elements,} \quad (1)$$

$$x_{tn} = \sum_{\mu=1}^m s_{t\mu} u_{t\mu} v_{\mu n}. \quad (2)$$

We refer to this as the SUV model and denote matrix elements by lower case letters throughout this article. We call $\mathbf{V}$ the *latent feature matrix* (dimensions $m \times N$). Each row of the matrix describes how one of m latent features, each specified as a row vector across the N variables, manifests itself in the data space. Further, $\mathbf{U}$ is the *coefficient matrix* (dimensions $T \times m$), which measure how “active” each of the m latent features is in each of the T samples. Finally, $\mathbf{S}$ (dimensions $T \times m$) is the *modulation matrix*, which can change the contribution of a feature to a sample by amplifying, zeroing it out, changing its sign, etc. Overall, the product $\mathbf{S} \odot \mathbf{U}$ is the *mixing matrix*, which describes how each of the T samples is obtained as a mixture of m latent features. The reason for representing the mixing matrix as a product of two factors will become clear later. Briefly, this is because products of two factors, each with simple statistical properties, can produce data with complex, more natural features.

As examples of data, which we hope to represent by this model, consider first a dataset of images. Here, the dataset may contain T images, each with N pixels. The latent features are m “basis” images, which are linearly combined to construct an image in the sample. Another example of data that we would like to represent is neural population activity. Here the variables are the N recorded neurons, with T samples of activity of each. The latent features represent m modes of collective neural activity [19–21]. We will assume that $T, N \gg 1$, which is often the case for such data.

In [34], we considered possibly the simplest model of the SUV structure, namely $\mathbf{X} = \mathbf{U}\mathbf{V}$ (with additional additive noise), for the case when the elements of $\mathbf{U}$ and $\mathbf{V}$ were i.i.d. Gaussian random variables, and the elements of $\mathbf{S}$ were all 1’s. We summarize the details of the model in Appendix A. In this model, each sample is a linear combination of m features, and the coefficients are independent Gaussian random variables themselves. Thus sample correlations are introduced by statistically independent mixing of the m latent features. Going forward, we will refer to this model as the Gaussian-Gaussian (GG) model. Here, we consider more complicated SUV models by dropping the condition of statistical independence of the mixing matrix components. Specifically, we seek to introduce mixing matrices that can model sparsity, often seen in biological data [37], so that only a few of many features contribute to a particular sample. We propose to model this by viewing the elements of $\mathbf{U}$ as taken from a Dirichlet distribution [38] (see also [39,40] for related models):

$$\text{Dir}(\{u_\mu\}; \beta_U) = \frac{1}{B(\beta_U)} \prod_{\mu=1}^m u_\mu^{\beta_U-1}, \quad \sum_{\mu} u_\mu = 1, \quad (3)$$

where $\beta_U \in \mathbb{R}_{>0}$ is a hyperparameter, $B(\beta_U)$ is

$$B(\beta_U) = \frac{\Gamma^m(\beta_U)}{\Gamma(m\beta_U)}, \quad (4)$$

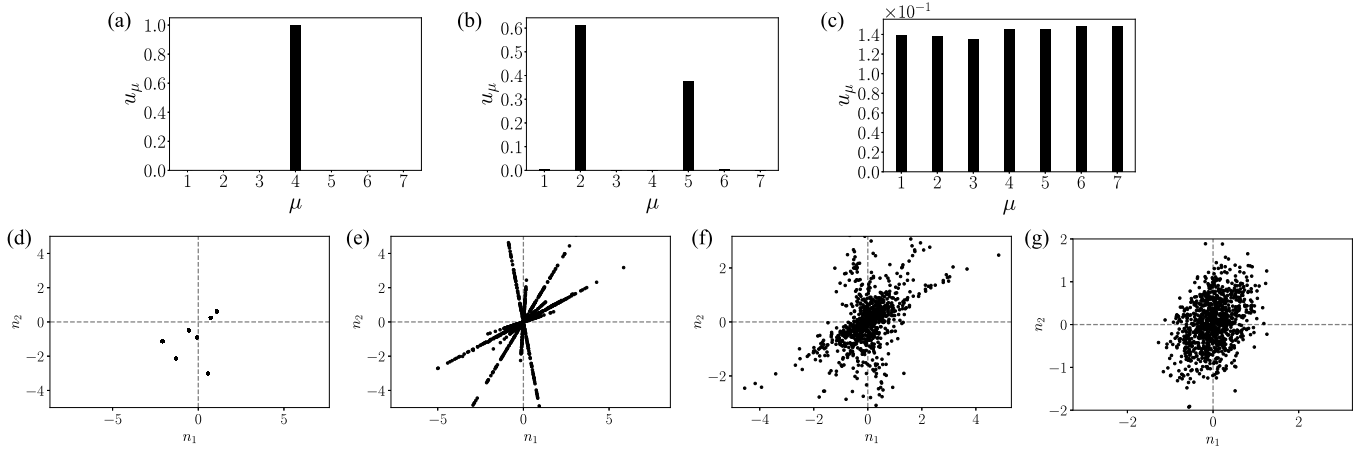


FIG. 1. In different parameter regimes, the SUV model admits locally low-dimensional and not low-dimensional data. We show the data scatter of the SUV model for different choices of $\mathbf{S}$ and β_U . The data have a characteristic aster structure. The matrix $\mathbf{V}$ with random Gaussian features is the same for all three plots, and $N = 2$, $T = 1000$, and $m = 7$. (a)–(c) Typical samples of the Dirichlet distribution for $\beta_U = 1 \times 10^{-8}$, 0.2, and 1×10^3 , respectively. (d) $s_{t\mu} = 1$ (trivial modulation; $p = 1$) and $\beta_U = 1 \times 10^{-8}$. Each sample is given by one of the $m = 7$ feature vectors (visible dots), creating well-defined clusters (dots), which can be further smeared by the measurement noise. (e) $s_{t\mu} \sim \mathcal{N}(0, 1)$ and $\beta_U = 1 \times 10^{-8}$. Each sample is still associated with a single feature vector, but the activation strength of the features follows the Gaussian distribution. This creates data that are locally low-dimensional, even though the overall dimensionality of the data cloud is not low. (f) $s_{t\mu} \sim \mathcal{N}(0, 1)$ and $\beta_U = 0.2$. Each sample is given by a mixture of multiple feature vectors (on average $m\beta_U = 1.4$ features), and the locally low-dimensional structure, while still visible, is not as prominent. (g) $s_{t\mu} \sim \mathcal{N}(0, 1)$ and $\beta_U = 1000$. On average, each sample is given by a mixture of all feature vectors, and the local low-dimensional structure has disappeared.

and $\Gamma(\beta_U)$ is the usual Gamma function of its argument. The Dirichlet distribution generates random vectors of m weights $\{u_\mu : \mu = 1, \dots, m\}$, with the weights summing to unity; see Figs. 1(a)–1(c) for typical samples of Dirichlet random vectors with $m = 7$ for different values of β_U . If the rows of the coefficient matrix are Dirichlet-distributed, then (positive) contributions of different latent features to the observations are statistically dependent due to the row normalization condition. While in principle we can consider feature matrices $\mathbf{V}$ with more complex statistical structure (which may be important, for instance, for generating data that are required to be strictly non-negative), for most of this article we continue to model the features themselves as Gaussian-distributed:

$$v_{\mu n} \sim \mathcal{N}(\mu_V, \sigma_V^2). \quad (5)$$

For convenience, we set $\mu_V = 0$ and $\sigma_V^2 = 1$ in the following, unless explicitly stated otherwise.

To develop some intuition for what follows, we note that the mean, the variance, and the covariance of weights in the random vector sampled from the Dirichlet distribution are easy to calculate [41],

$$\mu_u \equiv \bar{u}_\mu = \frac{1}{m}, \quad (6)$$

$$\sigma_u^2 \equiv \text{var } u_\mu = \frac{1 - m^{-1}}{m(1 + m\beta)}, \quad (7)$$

$$\text{cov}(u_\mu, u_\nu) = \frac{\delta_{\mu\nu} - m^{-1}}{m(1 + m\beta)}. \quad (8)$$

Entries of $\mathbf{U}$ are limited to $[0, 1]$, which is too restrictive to allow agreement with real data. Thus, we use $\mathbf{S}$, which modulates the coefficient matrix by multiplying each of its entries by a random number. We take the elements of $\mathbf{S}$ to be statisti-

cally independent, so that all correlations between the entries of the mixing matrix are in the coefficients, and not in their modulation. Note that the sign and magnitude of the elements of $\mathbf{S}$ allow us to model qualitatively different aspects of the data structure. Having both positive and negative modulation coefficients captures anticorrelations between samples. On the other hand, the varying magnitude of elements of $\mathbf{S}$ affects the distribution in the activation strength of the latent features, allowing, for example, long-tailed (outlier) contributions or no contributions with very small activations. Modeling different types of data would require different choices for the distribution of elements of $\mathbf{S}$, and we will not investigate these choices exhaustively here.

Figure 1 illustrates typical data scatter of the SUV model for different parameter values and $m = 7$. In Fig. 1(d), we set all elements $s_{t\mu} = 1$, such that the modulation is trivial. For small β_U (e.g., $\beta = 1 \times 10^{-8}$), only one latent feature contributes significantly to each sample, and we can thus expect to see exactly m points in the scatter plot. In contrast, in Fig. 1(e), the components of $\mathbf{S}$ are i.i.d. random variables drawn from a Gaussian distribution. For small β_U , still only a single feature is associated with each sample, but the activation strength of features varies, producing a characteristic aster scatter of samples, reminiscent of some neural data [42]. These data are locally low-dimensional, but globally they are not. That is, in each single petal of the aster in Fig. 1(e), the data are locally one-dimensional, while the whole aster has the dimensionality of 2. In contrast, in Fig. 1(f), $\beta_U = 0.2$, and multiple features contribute at different activation strengths, so that the one-dimensional structure of the aster is broken, though remnants of it are still visible. Each petal now has the same dimensionality as the whole aster. Finally, in Fig. 1(g), $\beta_U = 1000$, so that all features contribute to each sample significantly, and no locally low-dimensional structure is observed.

For most of this paper, we are going to focus on the case of binary modulation, $s_{t\mu} \in \{\pm 1\}$, such that $\mathbf{S}$ only influences the distribution of correlation signs among samples. For this, we draw elements of the modulation matrix from the Bernoulli distribution:

$$s_{t\mu} \sim P(k) = \begin{cases} p, & k = 1, \\ 1 - p, & k = -1. \end{cases} \quad (9)$$

That is, each entry of $\mathbf{S}$ is an i.i.d. binary random number, equal to $+1$ with a probability of p , and -1 with a probability of $1 - p$. This modulation flips the sign of each entry of the coefficient matrix $\mathbf{U}$ with probability $1 - p$, allowing for negative mixing weights $\mathbf{S} \odot \mathbf{U}$. In Appendix B, we compute the distribution of mixing weights in this model.

Data analysis usually focuses on correlations among the measured variables (columns of the data matrix $\mathbf{X}$). Since such variables may have widely different units, one often uses empirically standardized data $\mathbf{X}'$, where each variable (a column) is independently centered to zero mean and scaled to unit variance. One then makes inferences about the data based on the empirical variable-variable (denoted by subscript v) correlation matrix,

$$\mathbf{C}_v = \frac{1}{T} \mathbf{X}'^T \mathbf{X}'. \quad (10)$$

In contrast, in this work we focus on the correlations between samples. We also work with empirically standardized data $\tilde{\mathbf{X}}$, where now each sample (row of the data matrix $\mathbf{X}$) is independently centered to zero mean and scaled to unit variance. The empirical sample-sample correlation matrix (denoted by subscript s) is

$$\mathbf{C}_s = \frac{1}{N} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T, \quad (11)$$

and we will simply refer to this matrix as the ECM. The empirically standardized data approximate the *theoretically* standardized random variables

$$\tilde{x} \equiv \frac{x - \mu_x}{\sigma_x}, \quad (12)$$

where the mean μ_x and the standard deviation σ_x of the pdf of each element x of $\mathbf{X}$ (also denoted as *theoretical* values) are obtained from Eq. (1):

$$\mu_x = \mathbb{E}[x] = \sum_{\mu} \mathbb{E}[s_{\mu} u_{\mu} v_{\mu}] = \sum_{\mu} \mathbb{E}[s_{\mu}] \mathbb{E}[u_{\mu}] \mathbb{E}[v_{\mu}], \quad (13)$$

$$\begin{aligned} \sigma_x^2 &= \mathbb{E}[(x - \mu_x)^2] = \sum_{\mu} \text{var}(s_{\mu} u_{\mu} v_{\mu}) \\ &+ \sum_{\mu \neq \nu} \text{cov}(s_{\mu} u_{\mu} v_{\mu}, s_{\nu} u_{\nu} v_{\nu}). \end{aligned} \quad (14)$$

Above we used the fact that s_{μ} , u_{μ} , and v_{μ} are statistically independent from each other.

We will study the eigenvalues of the sample-sample correlation matrix $\mathbf{C}_s$, denoted λ_s . However, we note that there is a well-known relation between the eigenvalues of $\mathbf{C}_s$ and eigenvalues of $\mathbf{C}_v$, denoted λ_v : nonzero eigenvalues of both matrices differ only by a multiplicative constant [32].

Specifically,

$$\lambda_v \approx \frac{\bar{\sigma}_s^2 N}{\bar{\sigma}_v^2 T} \lambda_s, \quad (15)$$

where $\bar{\sigma}_s$ is the empirical standard deviation along the dimension of variables (the bar denotes that we take the average value of all samples), and $\bar{\sigma}_v$ is the averaged empirical standard deviation along the dimension of samples (averaged over variables) of $\mathbf{X}$. For $T, N \gg 1$, $\bar{\sigma}_s$ and $\bar{\sigma}_v$ both converge to the theoretical value σ_x and $\bar{\sigma}_s/\bar{\sigma}_v \rightarrow 1$. In what follows, we focus on $N, T \gg 1$, so that the row and the column sample variances are well approximated by their theoretical values.

III. DATA PATTERNS AND DISTRIBUTIONS OF OBSERVABLES

To explore the effects of statistically dependent mixing of latent features on the structure of the data, we first choose $p = 1$ in Eq. (9). This choice implies a trivial modulation matrix $\mathbf{S}$ with all its entries $s_{t\mu} = 1$. As a result, the mixing matrix $\mathbf{S} \odot \mathbf{U}$ is identical to the coefficient matrix $\mathbf{U}$. We will refer to the model in this parameter regime as the Dirichlet-Gaussian (DG) model.

A. Data patterns in the DG model

We give a description of the data patterns obtained in different regimes of the parameter β_U . We consider both the case when $m \sim O(1)$ and when $m \sim O(N)$. In Figs. 2(a) and 3(b), we visualize data patterns for different values of β_U by presenting the pairwise correlation of the samples. We plot the correlation matrices with samples ordered using the nearest-neighbor clustering [43], which groups the samples hierarchically based on their correlations. The dendrograms above the correlation matrices illustrate the obtained hierarchical clustering. Below we describe typical patterns observed in the data.

Clusters, Fig. 2(a) (left column), emerge for $m \sim O(1)$ and $\beta_U \ll 1$. Here, each column of the coefficient matrix $\mathbf{U}$ is concentrated on a single feature, and each feature defines a separate cluster. The clusters are clearly visible in the correlation matrix. The dendrogram shows that the clusters are separated, and each cluster includes multiple samples.

Overlapping clusters, Fig. 2(a) (middle), are visible for $m \sim O(1)$ and $m\beta_U \sim O(1)$. Here, $O(1)$ features contribute to each sample. The correlation structure between samples becomes fuzzy, and samples may belong to more than one cluster. The dendrogram reveals groups of samples of different sizes and varying cluster distances.

Uniform mixing, Fig. 2(a) (right), describes the regime $m \sim O(1)$ and $\beta_U \gg 1$. All of the m latent features contribute approximately uniformly to each sample, and all the samples are therefore highly correlated. The dendrogram confirms that there is effectively a single cluster of samples (note that the range of cluster correlation distances is 10^{-3} in this subplot).

Sparse mixing, Fig. 3, is observed for $m \gg 1$, and $m\beta_U \ll m$ [$m\beta_U$ can be $O(1)$, or $\gg 1$]. In this regime, the number of features available is much larger than the number of features that contribute to any given sample, see Fig. 3(a). In Appendix Fig. 5(a), we show the distribution of elements

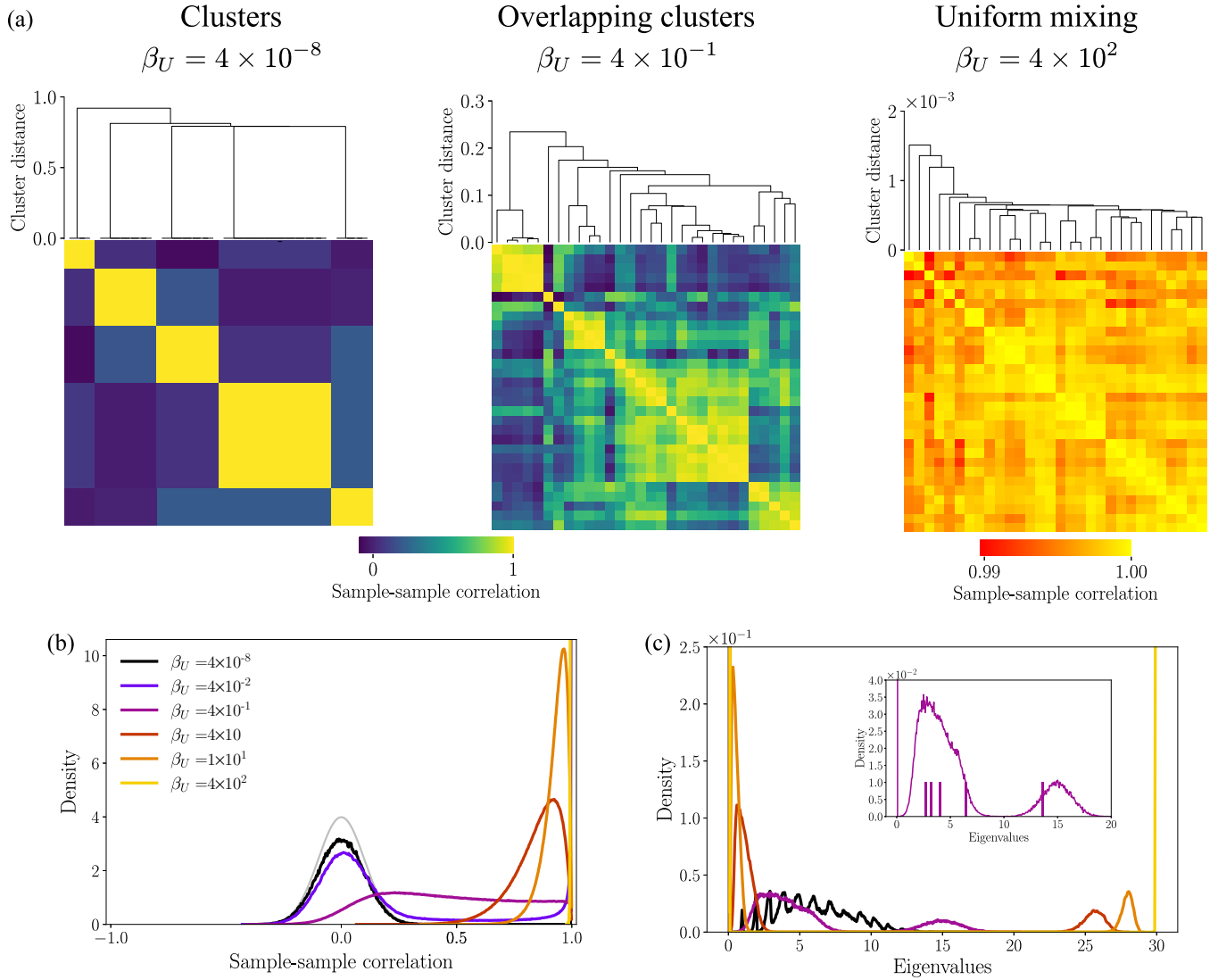


FIG. 2. Qualitatively different correlation patterns emerge in the Dirichlet-Gaussian model in different regimes of the parameter β_U : We illustrate this for $T = 30$, $N = 100$, and $m = 5$, and $\mathbf{V}$ has i.i.d. Gaussian entries with $\sigma_V^2 = 1$. (a) Sample-sample correlation matrices and dendrograms obtained from hierarchical clustering of the ECM for different β_U . (b) Probability densities of correlation coefficients for a range of β_U values. For reference, the gray curve shows the Gaussian distribution with the variance $1/N$, as expected for correlations of pure Gaussian noise data. (c) Eigenvalue densities for the same β_U values. Inset: zoom-in for $\beta_U = 4 \times 10^{-1}$, which is in the overlapping clusters regime. The vertical bars show the positions of eigenvalues for the single realization of the model; the height of the bars is not meaningful. Note a single large eigenvalue separated from the rest by a gap. Averages are computed from 2×10^4 independent realizations of the model.

of the simulated mixing matrix $\mathbf{S} \odot \mathbf{U}$ (gray) and an analytic curve (red) given by Eq. (B5). The plot shows a large number of small entries and a small number of large entries, as expected for sparse mixing. From the dendrogram, Fig. 3(b), we see that there are no clear clusters of more than two samples.

B. Distributions of second-order observables

Next, we qualitatively discuss the distribution of correlation coefficients and the spectral density of the ECM, Eq. (11). In the simulations, the pairwise correlations between variables are computed from empirically standardized data. These approximate the theoretically standardized random variables in Eq. (12), whose means and variances we now compute.

The mean of a product of a zero mean Gaussian and a Dirichlet random variable vanishes. Hence, by Eq. (13), the mean of elements x of the data matrix also vanishes,

$$\mu_x = \sum_{\mu} \mathbb{E}[v_{\mu}] \mathbb{E}[u_{\mu}] = 0. \quad (16)$$

In Appendix C we evaluate Eq. (14) to find the variance:

$$\sigma_x^2 = m(\sigma_u^2 + m^{-2})\sigma_V^2 = \left(\frac{1 - m^{-1}}{1 + m\beta_U} + \frac{1}{m} \right) \sigma_V^2. \quad (17)$$

We verify this expression numerically in Appendix Fig. 6, and we note the following limits: when $m \gg 1$, while $m\beta_U$ is finite, the variance approaches $\sigma_x^2 \approx \sigma_V^2/(1 + m\beta_U)$; when β_U

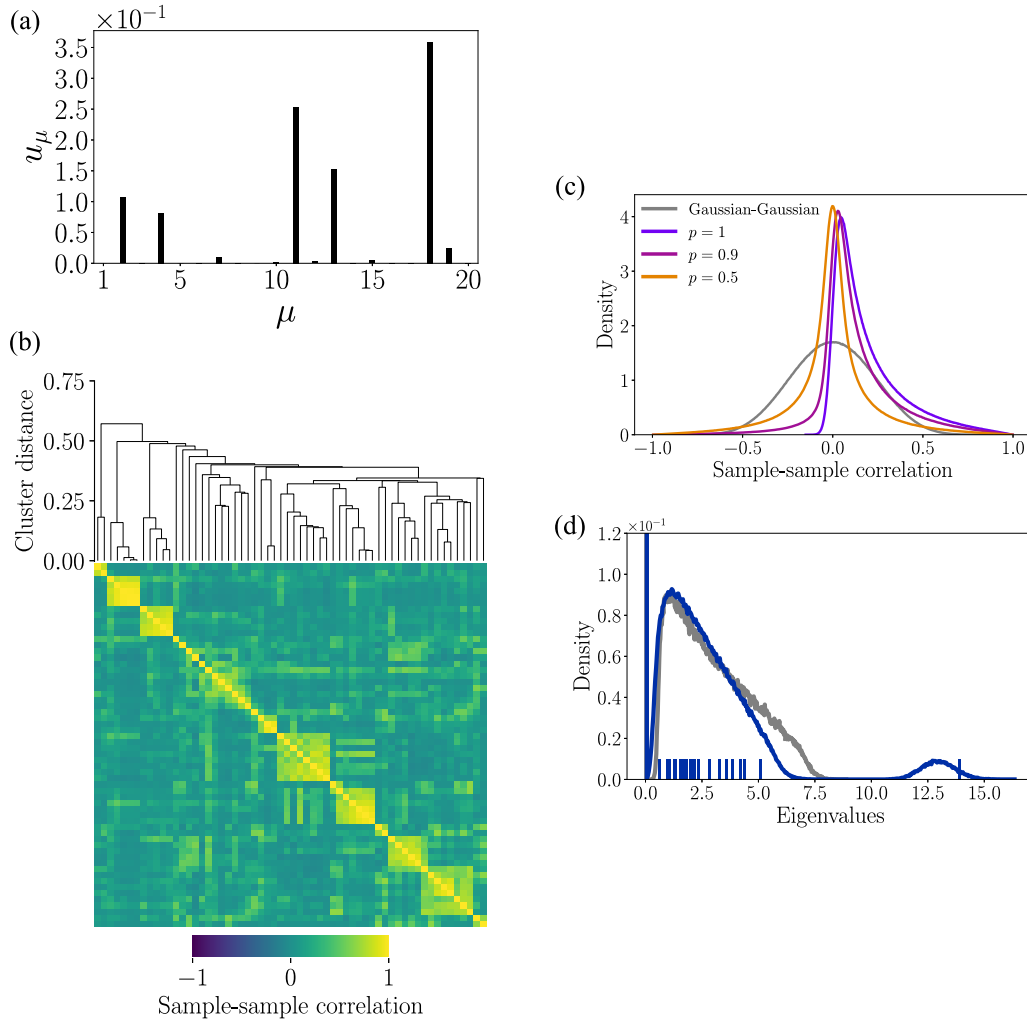


FIG. 3. The SUV model admits a sparse mixing regime. We use $T = 60$, $N = 1000$, $m = 20$, and $\sigma_V^2 = 1$. (a) Typical samples of the Dirichlet distribution for $\beta_U = 0.15$. (b) Sample-sample correlation matrix and dendrogram obtained from hierarchical clustering of the ECM. (c) Density of sample-sample correlations of the Dirichlet-Gaussian model for $p = 1$, 0.9, and 0.5. The observable density of the Gaussian-Gaussian model (gray) is shown for reference. (d) Eigenvalue densities of the same models. Vertical bars show positions of eigenvalues for the single realization of the Dirichlet-Gaussian model; the height of the bars is not meaningful. Note a single large eigenvalue separated from the rest by a gap. Averages are computed from 2×10^4 independent model realizations.

approaches zero, $\sigma_x^2 \approx \sigma_V^2$; and when β_U approaches infinity, $\sigma_x^2 \approx \sigma_V^2/m$.

1. Correlations

The family of correlation distributions in Fig. 2(b) reflects the transition between the two extreme limits (from pure clusters to uniform mixing). The first observation is that the density is not symmetric around zero. In the clusters limit (black curve), the density of correlations has a nearly δ -function peak at 1, corresponding to correlations between samples that are built from the same features. The density also has an approximately Gaussian peak around zero, which emerges from correlations between samples that include different features, and hence from feature-feature correlations. In the opposite limit of uniform mixing (yellow curve), all small correlations have disappeared and the density is approaching a δ function concentrated at 1 because all variables

approximately behave as a single cluster. For the regime of overlapping clusters between the two extremes, we observe intermediate correlation values between zero and one, signifying the existence of variables that can be attributed to more than one cluster.

In the sparse mixing regime, the correlation density [purple curve in Fig. 3(c)] is again nonsymmetric around zero. It has a large peak around small, positive correlations and few large correlation values. This is due to the sparse selection from a large number of features, so that there are now samples with different numbers of the same features, smearing the high correlation peak of the clusters limit. The peak near zero still comes from correlations between samples that do not have the same features, and hence from feature-feature correlations. For reference, we show the correlation density of the Gaussian-Gaussian random latent feature model. We also show the correlation density for $p = 0.9$ (magenta) and $p = 0.5$ (orange).

2. Eigenvalues

The shape of the eigenvalue density depends on the model parameters T , N , m , and β_U . From RMT, we know that the eigenvalue densities of random matrices often converge to their limiting form when T , N , etc. tend to infinity, while ratios, such as N/T or the ratio of the number of latent features to the number of observables m/N , remain fixed. This is known as the thermodynamic or large- N limit, in which the fixed ratios control the shape of the density [32,34]. We expect these ratios to control the shape of the eigenvalue distributions of the Dirichlet-Gaussian model as well. However, here we focus on the qualitative description of eigenvalue densities, and we leave a detailed analytic description for the future.

When $m \ll N, T$, but we still consider it to be large, the rank of the ECM satisfies $\text{rank}(\mathbf{C}) < N, T$, implying that there are trivial zero eigenvalues. Beyond these, analogous to the correlations, the eigenvalue densities in Fig. 2(c) reflect the transition between the two limits of the model.

First, in the clusters limit (black curve), the eigenvalue density has a broad, “spiky” bump. The spikes result from the interaction of (on average) m dominant eigenvalues of a similar magnitude. In Appendix Fig. 7, we show how the resolution of the δ spikes increases at better sampling (smaller T/N ratio). In the opposite limit of uniform mixing (yellow curve), the spectrum becomes degenerate with two δ peaks. One peak is for a single eigenvalue at $\lambda = N$, corresponding to all variables being in one cluster, and the other peak has $N - 1$ eigenvalues at zero.

For the regime of overlapping clusters, between the two limits, the density has two bumps [inset in Fig. 2(c)]. The bumps move in opposite directions as β_U increases, and they develop into δ peaks in the extreme limit. To understand the origin of the two bumps, we show the position of eigenvalues of a particular realization of the overlapping clusters model from Fig. 2(a) by vertical bars in Fig. 2(c), inset. The right density bump is due to a single, top-ranked eigenvalue, and the left bump is due to the remaining $m - 1$ nontrivial eigenvalues. In the particular case shown, the left bump is the distribution of four eigenvalues. We verified numerically our expectation that the left bump carries approximately four times as much mass as the right bump. The top-ranked eigenvalue is due to a baseline correlation across all variables that comes from the nonzero, positive mean of the correlation density. Its origin can be further understood from analyzing the eigenmode associated with the top-ranked eigenvalue. For this, we decompose the correlation matrix into two contributions, as expressed in Appendix Fig. 8(a). The first contribution is the projection of the correlation matrix onto the top-ranked eigenmode, and the second contribution is the projection on all other eigenmodes. The decomposition is shown in Appendix Fig. 8(b), and the distributions of matrix entries of each contribution is shown on the right. The mean values of the correlation density and the top-ranked contribution are the same, and the mean of the remaining contribution vanishes. Thus the top-ranked mode can be considered as a baseline contribution, similar to the “market mode” found in the analysis of financial data, which captures the collective up-and-down of the stock market [44]. As β_U grows, more and more of the total variance goes into this mode, until all variables are strongly correlated. The remaining $m - 1$ modes

capture correlations between specific subgroups of variables. As β_U is increased, less and less of the total variance belongs to these modes.

Finally, in Fig. 3(d) we show the eigenvalue density in the *sparse* mixing regime (blue curve). For reference, we also show the eigenvalue density of the Gaussian-Gaussian model (gray curve) [34]. The density of the Gaussian-Gaussian model has a single peak with upper and lower eigenvalue bounds given by $\lambda_{\pm} = (1 \pm \sqrt{N/m})^2 \approx 0.54, 7.46$, respectively, as was derived in [34]. In contrast, the eigenvalue density of the sparse mixing model has two peaks. The qualitative understanding of the origin of the two peaks is analogous to the one just described for the overlapping clusters regime. The eigenvalues of a specific realization of the sparse Dirichlet-Gaussian model are shown by the vertical blue bars. The right and the left bumps are due to a single top-ranked and the $m - 1$ remaining nontrivial eigenvalues, respectively. The decomposition of the sparse Dirichlet-Gaussian correlation matrix into the two contributions is shown in Appendix Fig. 8(c), together with the contributions’ densities.

For the computation of the sparse model, we have selected parameters such that the density is well sampled. This is the case when both T and N are large. In particular, N has to be large enough to ensure that sampling of the different ways of sparsely selecting features is sufficient. In other sampling limits, the form of the eigenvalue density is more complicated, and a detailed analysis will be needed to understand its structure.

Finally, in Appendix Fig. 9, we show the influence that feature structure has on the correlation and eigenvalue observable distributions for the sparse mixing regime. As examples, we consider features that are Dirichlet random vectors, features with entries that are i.i.d. random variables drawn from an exponential distribution, and features with entries that are i.i.d. binary random variables taking values $\{\pm 1\}$. We note that the first two examples lead to non-negative data, and the SUV model is a form of non-negative matrix factorization [31] in these cases. We observe that feature structure only has a small influence on the correlation and eigenvalue distributions.

IV. SAMPLE STATISTICS OF NATURAL IMAGES

We now turn to the analysis of the statistics of natural scene images. Significant success has been achieved in studying the correlation statistics between different pixels of natural images [8,11]. Here, we study the correlations between samples (patches from natural scene images), fit SUV random latent feature models on the correlations, and we also make predictions for the eigenvalues of the sample-sample correlation matrix.

Our data come from the classic *Sparsenet* dataset [15] used by Olshausen *et al.* to train a model of receptive fields in the mammalian visual system [16]. The dataset consists of ten natural scene images (of North American landscape) in grayscale, and each image has been whitened. Specifically, we construct a $T \times N$ data matrix $\mathbf{X}$, containing T flattened, square image patches with N pixels. These relatively small image patches are randomly sampled from across the ten much larger images of the dataset.

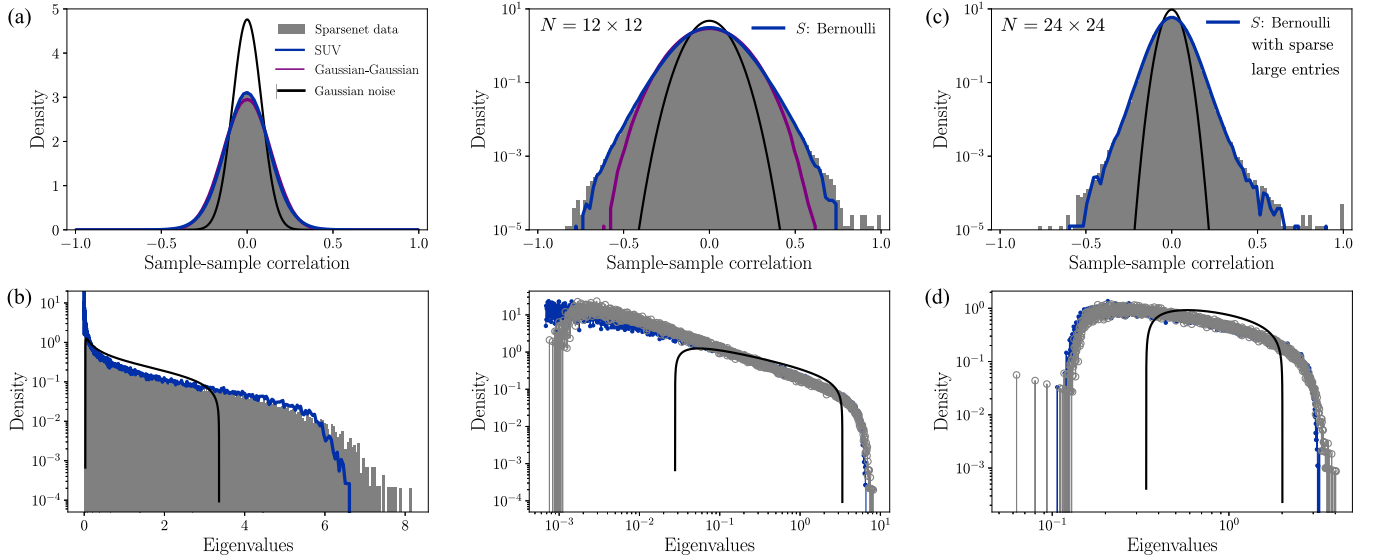


FIG. 4. Sample statistics of natural scene images is well-described by the SUV model. We compare the sample-sample correlation and eigenvalue distributions of the *Sparsenet* dataset with predictions of the maximum-likelihood (ML) SUV model. Our *Sparsenet* data matrix consists of $T = 100$ square image patches each of size N . To obtain the observable distributions (gray), we sample the data matrix multiple times, stated as the number of trials below. (a) Image patch size $N = 12 \times 12 = 144$: The elements of the modulation matrix $\mathbf{S}$ are i.i.d. binary random variables taking values $\{\pm 1\}$, each with probability $p = 0.5$. We show the comparison of sample-sample correlations of the *Sparsenet* data (gray) and the ML SUV model with $m_{\text{ML}} = 92 \pm 2$ and $\beta_{\text{ML}} = 0.30 \pm 0.05$ (blue), on linear-linear (left) and log-linear scale (right). For reference, we also show the distribution of correlations expected for pure noise data (black), and the ML Gaussian-Gaussian latent feature model with $m_{\text{ML}} = 91 \pm 2$ (purple). (b) Comparison of the eigenvalue density of the *Sparsenet* sample-sample correlation matrix (gray) and the model prediction (blue), on log-linear scale (left) and log-log scale (right). For reference, we show the Marčenko-Pastur distribution (black), with sampling ratio T/N . (c) Image patch size $N = 24 \times 24 = 576$: The elements of the modulation matrix $\mathbf{S}$ are i.i.d. binary random variables as above, but with a small number of large entries randomly introduced into the matrix ($z = 5$, $p_{\text{replace}} = 5 \times 10^{-3}$; see the main text). We show the comparison of sample-sample correlations of the *Sparsenet* data (gray) and the ML SUV model with $m_{\text{ML}} = 310 \pm 2$ and $\beta_{\text{ML}} = 0.25 \pm 0.05$ (blue). (d) Comparison of the eigenvalue density of the *Sparsenet* sample-sample correlation matrix (gray), the model prediction (blue), and the Marčenko-Pastur distribution (black). All curves from model simulations and *Sparsenet* data were computed with 800 trials.

We standardize the rows of the data matrix and compute its sample-sample correlation matrix Eq. (11). The distribution of the correlations is shown in gray in Fig. 4(a). There are, in principle, two ways of arranging the samples. The first is to view them as a single data matrix with a large number T of image patches. The second is to arrange them into multiple data matrices with a smaller T , and average the observables over these matrices. This results in fewer correlation coefficients or eigenvalues, but averaging determines their distributions better. We adopt the latter approach, choose $T = 100$, and average over 800 repetitions, which we will refer to as trials in the following. We compare the two approaches in Appendix Figs. 10(a) and 10(c), showing that they lead to very similar correlation distributions.

To first demonstrate that the sample-sample correlation distribution captures the nontrivial structure of the data, we compare empirical correlations to the distribution correlations for i.i.d. Gaussian noise data (black) in all panels of Fig. 4. Such correlations follow a Beta distribution [45], whose expression we provide in Eq. (D1) of Appendix D. For large enough N , the curve approaches a Gaussian normal distribution with variance $1/N$. The distribution of such pure noise correlations substantially differs from the correlations in the *Sparsenet* data, as seen in Figs. 4(a)–4(c). Next, we model the *Sparsenet* correlations with a variant of our SUV model.

The first step is to fit the correlation matrix $\mathbf{S}$. We start with relatively small patches, $N = 12 \times 12 = 144$ pixels. As before, we choose the elements of the modulation matrix to be binary $\{\pm 1\}$, obeying the Bernoulli distribution in Eq. (9). We note that the correlation distribution in image patches, Fig. 4(a), is symmetric around zero, taking on both positive and negative correlation values. This requires setting the Bernoulli distribution parameter to $p = 1/2$. The second step is to fit the model parameters β_U and m to the empirical distribution of the correlations [gray in Fig. 4(a)] using maximum-likelihood (ML) estimation. We find as ML values $m_{\text{ML}} = 92 \pm 2$ and $\beta_{\text{ML}} = 0.30 \pm 0.05$, and the fit is shown in blue in Fig. 4(a). We find a near-perfect match between the *Sparsenet* correlations and our model, both in the bulk (near zero) and in the tails of the distribution, where the density is five orders of magnitude smaller [cf. Fig. 4(a)]. To verify that the ML value $m_{\text{ML}} \approx 92$ is not controlled by the sample size, $T = 100$, in Appendix Fig. 10(b) we use $T = 200$ and again find $m_{\text{ML}} \approx 93$.

We next investigate the quality of fits as a function of (i) the patch size N , and (ii) different choices for the distribution of $\mathbf{S}$ matrix elements. First, in Fig. 4(c) and Appendix Figs. 11(a)–11(c), we compare different patch sizes with $N = 12 \times 12$, 18×18 , and 24×24 pixels. With the increasing patch size, the tails of the sample-sample correlation distribution data

(gray) become more pronounced, especially for the positive correlation values. We wondered if we could find a simple way to model these tails. This would require a small number of strongly positively correlated samples. To achieve this, we start as above, with an $\mathbf{S}$ with $\{\pm 1\}$ drawn from a Bernoulli distribution with $p = 1/2$. However, now, with a small probability $p_{\text{replace}} = 5 \times 10^{-3}$, we randomly replace some elements by a large, positive value $z = 5$. In Appendix Fig. 5(c), we present a simulation, showing the effect that this change has on the distribution of elements of the mixing matrix $\mathbf{S} \odot \mathbf{U}$. This ad hoc modeling leads to an improved fit, even of the right tail, shown in blue in Fig. 4(c). It is likely possible to make this fit even better via a systematic search over the parameters p_{replace} and z . However, we choose not to do this since the fit is already great and since the main point here is not a perfect fit, but a demonstration that adjustments leading to quantitative improvements of the fit can be incorporated without changing the general structure of the SUV model.

Next, we investigate the effect of using a distribution of $\mathbf{S}$ matrix elements different from Bernoulli. Specifically, we consider the case when the elements of the modulation matrix $\mathbf{S}$ are drawn from a Gaussian distribution with zero mean and unit variance. We show the corresponding ML fit with $m_{\text{ML}} = 93$ and $\beta_{\text{ML}} = 0.8$ in Appendix Fig. 12(b). We find that the fit is in very good agreement with the data. However, we note that β_{ML} is significantly larger than for Bernoulli-distributed $\mathbf{S}$ elements ($\beta_{\text{ML}} \approx 0.3$). This shows that for the continuous Gaussian case, on average more modes are required to represent any given sample (image patch), presumably because many modulation values are near zero.

Now, we ask what the general agreement between the SUV model and the data tells us about the underlying structure of the data. First, we note that the ML parameter values $m_{\text{ML}} \approx 92$ and $\beta_{\text{ML}} \approx 0.3$ imply that, for each sample (image patch), only a relatively small number of the total available features is used for mixing, and the ML model is thus of the sparse mixing type. Note that *sparse* in the name of the original dataset refers to the activity of the neural network used to analyze it. It is thus encouraging that the data themselves can be modeled as sparse.

Second, we point out that, with the inferred parameters, the overall dataset is not particularly low-dimensional (even if each patch is), since the number of features, ~ 92 , is similar to the number of variables, 144. Thus we check whether the data could also be fit by a simpler, low-rank model that does not model correlations between features. For this, we use the Gaussian-Gaussian model from [34], summarized in Appendix A, and we fit its single parameter m on the sample-sample correlation data. We find an ML value of $m_{\text{ML}}^{\text{GG}} = 91 \pm 2$ closely matching the ML value determined for the SUV model. Yet, as seen in Fig. 4(a), the distribution of sample-sample correlations produced by the best-fit Gaussian-Gaussian model is dramatically different from the data. In other words, these data are not purely low-rank in the traditional sense, but their low dimensionality emerges through sparse mixing of features. We suggest that this is the reason why data decomposition methods like independent component analysis and sparse coding applied to these natural images yield latent features that are interpretable as receptive fields [16,46].

Finally, to make a *prediction* instead of a fit, we use the SUV model with parameters inferred from the distribution of correlations and compare the eigenvalue distribution of its correlation matrix to the *Sparsenet* data. We show the comparison in Figs. 4(b) and 4(d). We find a good match between the data (gray) and the model prediction (blue). For reference, we also show the distribution of eigenvalues expected for pure noise data (black curve), given by the well-known Marčenko-Pastur distribution [47], which we state in Appendix D in Eq. (D5). In Appendix Fig. 13, we show the prediction versus data for eigenvalue distributions for larger T . While the predictions are good for modest T (≈ 100), the prediction quality for small eigenvalues becomes bad as T is increased. In other words, as a larger number T of samples per trial is available, the smaller eigenvalues of the sample-sample correlation matrix are sampled better, and the SUV model cannot capture these small details. This behavior is different from the sample-sample correlation values themselves, whose distribution does not significantly depend on T . We leave a detailed investigation of this phenomenon to future work.

V. DISCUSSION

We have presented a random latent feature model for structured data based on linear mixing of latent features, Eq. (1). We call this the SUV model. A key ingredient in our model is the statistical dependence of the mixing coefficients. In different parameter regimes, this dependence gives rise to different patterns, commonly found in natural datasets, from clusters to sparse models. As an example, we showed that the SUV model is an *excellent* model for the sample-sample statistics of natural images, requiring fitting of just two hyperparameters: the size of the feature set and the average number of features used to generate any particular sample, though the agreement becomes quantitatively worse at larger patch sizes. We are not aware of similar simple, analytical generative models that can approximate such statistics in this or other similar natural datasets—usually fitting of many parameters, such as synaptic weights of generative neural networks, is required for a similar accuracy [6,21,48–51].

A central point of our analysis is that the qualitative structure of the distribution of various observables can inform us about a class of models best suited for representing the data. For example, in the case of *Sparsenet* data, we observed that quantitatively small differences in the tails of the distribution of sample-sample correlations suggest that sparse mixing of features is more appropriate to describe this dataset than more traditional low-rank models. For data analysis, this means that decomposition schemes like independent component analysis [52,53], non-negative matrix factorization [31], and similar sparse coding models are more natural to apply to the data—and are more likely to yield physically and biologically interpretable latent features—than other decomposition methods, such as principal component analysis [54]. Similarly, a characteristic feature of our model is the aster structure of the data scatters, cf. Figs. 1(e) and 1(f). These data are locally low dimensional, but globally they are not. Since such structures, indeed, occur in data and models of neural activity [42,55], one should question whether models based on

low-dimensional neural activity manifolds are the best model for such datasets.

In this article, we have focused mostly on sample-sample correlations. This is because, in our opinion, these statistics do not receive enough attention in the literature, though the kernel representation of data covariance is used routinely. However, nonzero parts of spectra of sample-sample and variable-variable correlations are the same up to a normalization constant, and the SUV model can equally be applied to the variable-variable statistics, making the analysis more general.

The SUV model, Eq. (1), is very simple, and it is unclear *a priori* why it should be so effective in modeling *Sparsenet* data. We believe that there are two reasons for this. First, note that we assumed that the latent features are Gaussian i.i.d. vectors, which is certainly a bad model for receptive fields in [16]. However, the structure of features becomes practically irrelevant to sample-sample correlations as long as the individual features are uncorrelated either due to (a) the central limit theorem at large enough N , or (b) features forming a sparse, overcomplete code, as in [16]. Second, in our model, sparsity emerges as a consequence of the Dirichlet distribution in a particular parameter regime. This is analytically tractable, but is also unlikely to be a precise model of the world. Yet Olshausen *et al.* [16] showed that different ways of imposing sparsity on their model resulted in quantitatively similar conclusions. Thus, one may expect that the details of the sparseness mechanism in the generative model would not be especially important either. The insensitivity of our model to the feature and the sparsity structures points to a notion of universality—a yet undiscovered class of models with quantitatively similar sample-sample correlations. Analysis of this class would need to be done using RMT approaches [32], and would parallel other universality classes discovered within the theory. Note, however, that traditional universal results in RMT emerge from matrices with i.i.d. entries, and tools for analysis of statistically dependent mixing coefficients are not well developed. We hope that the need to analyze the SUV or similar models will lead to advances in the RMT methods as well.

ACKNOWLEDGMENT

I.N. was supported, in part, by the Simons Foundation Investigator award, NSF Grant No. PHY/201052, and NIH Grants No. 1R01NS099375 and No. 2R01NS084844.

APPENDIX A: GAUSSIAN-GAUSSIAN MODEL

We give a brief summary of the Gaussian-Gaussian model, which has been analyzed and discussed at length in [34]. This model can be used to describe pure low-rank structure in data. While in Ref. [34] the model included additive noise, we focus here only on the signal part of the model. The model's data matrix is given by a product of two random matrices:

$$\mathbf{X} = \mathbf{U}\mathbf{V}, \quad (\text{A1})$$

with matrix entries given by i.i.d. zero mean Gaussian random numbers:

$$u_{t\mu} \sim \mathcal{N}(0, \sigma_U^2), \quad v_{\mu n} \sim \mathcal{N}(0, \sigma_V^2), \quad (\text{A2})$$

$$t = 1, \dots, T, \quad \mu = 1, \dots, m, \quad n = 1, \dots, N. \quad (\text{A3})$$

In general, the random variables have variances σ_U^2 and σ_V^2 , but throughout this article we set the variances to unity $\sigma_U^2 = \sigma_V^2 = 1$. The dimensions of the factor matrices $\mathbf{U}$ and $\mathbf{V}$ are $T \times m$ and $m \times N$, respectively, and the latent dimensionality is m .

APPENDIX B: ANALYTICAL DISTRIBUTION OF THE MIXING MATRIX ELEMENTS FOR BINARY MODULATION

We derive an analytical expression for the distribution of mixing coefficients $(\mathbf{S} \odot \mathbf{U})_{t\mu}$ for the case when the rows of the coefficient matrix $\mathbf{U}$ are given by Dirichlet random vectors as defined in Eq. (3), and each element of $\mathbf{S}$ is an i.i.d. binary random variable $\{\pm 1\}$, drawn according to Eq. (9). To obtain the distribution of individual entries of the mixing matrix, we marginalize the Dirichlet distribution Eq. (3) over all, but a single component. It is well-known that this results in the Beta distribution

$$\int du_1 \dots du_{m-1} \delta\left(1 - \sum_{\mu} u_{\mu}\right) \text{Dir}(\{u_{\mu}\}; \beta_U) \quad (\text{B1})$$

$$= \text{Beta}(u_m; \beta_U, (m-1)\beta_U). \quad (\text{B2})$$

Since the hyperparameter β_U of the Dirichlet distribution in Eq. (3) is the same for all variables u_{μ} , the result of marginalizing is independent of which of the $m-1$ directions we choose to marginalize over. All elements of the coefficient matrix thus follow the same Beta distribution:

$$u_{t\mu} \sim \text{Beta}(u_{t\mu}; \beta_U, (m-1)\beta_U). \quad (\text{B3})$$

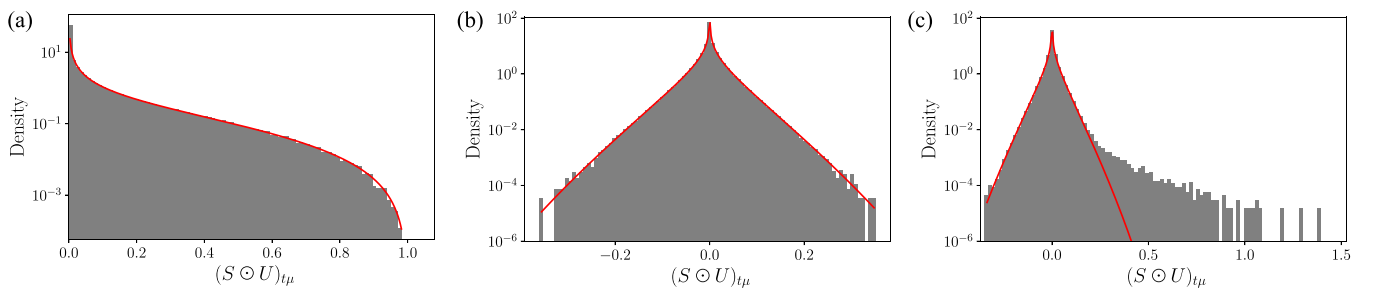


FIG. 5. Distribution of mixing matrix elements. (a) $p = 1$, $m = 20$, and $\beta_U = 0.15$. (b) $p = 0.5$, $m = 90$, and $\beta_U = 0.3$. (c) Same parameters as (b) but with some elements of the $\mathbf{S}$ matrix randomly replaced with a probability of $p_{\text{replace}} = 5 \times 10^{-3}$ by $z = 5$. For all simulations, $T = 100$ with 400 independent trials. The red line shows the distribution without disorder for reference.

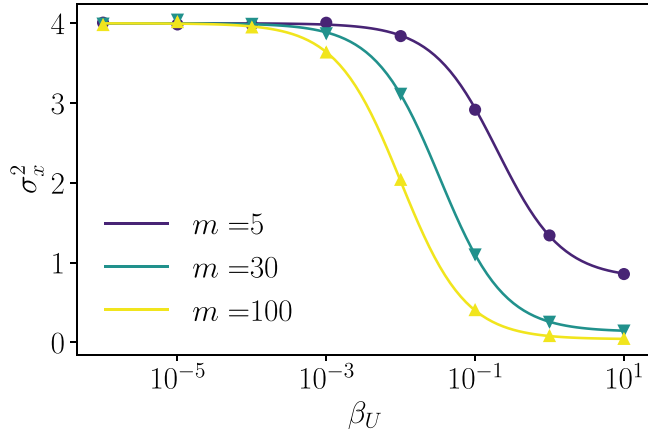


FIG. 6. Comparison of the analytic (solid lines) and simulated (markers) values of the data variance. Dirichlet-Gaussian model with a trivial modulation matrix, $\sigma_v^2 = 4$ for $m = 5, 30$, and 100 over a range of β_U values. Simulated values are averages over 3.6×10^4 independent model realizations.

To obtain the elements of the mixing matrix, the coefficients are modulated by $\mathbf{S}$, flipping their sign with a probability of $1 - p$. To account for the modulation, we have to weight the distribution of positive and negative elements of the mixing matrix by p and $1 - p$, respectively. The distribution of mixing matrix elements

$$x \equiv (\mathbf{S} \odot \mathbf{U})_{t\mu} \quad (\text{B4})$$

is therefore given by

$$\text{pdf}(x) = [pH(x) + (1 - p)H(-x)]\text{Beta}(|x|; \beta, (m - 1)\beta), \quad (\text{B5})$$

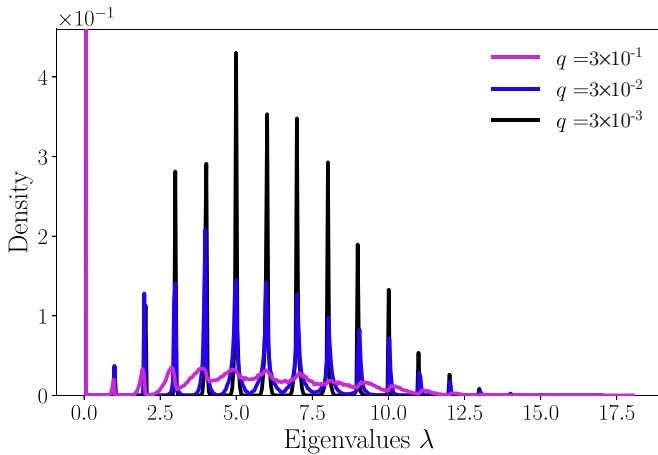


FIG. 7. Eigenvalue density spikes in the clusters regime of the Dirichlet-Gaussian model are resolved by small sampling ratios: We show this for $T = 30$, $m = 5$, $\beta_U = 4 \times 10^{-8}$, and $\sigma_v^2 = 1$. The resolution of the spikes increases as the sampling ratio $q = T/N$ decreases. Averages are computed from 2×10^4 independent realizations.

where H is the Heaviside step function. Figure 5 shows a comparisons between the analytics and simulations.

APPENDIX C: DATA MEAN AND VARIANCE IN THE DIRICHLET-GAUSSIAN MODEL

We compute the theoretical mean and variance of data from the Dirichlet-Gaussian model with a trivial modulation matrix ($s_{t\mu} = 1$), and i.i.d. Gaussian latent features. Starting from the model definition, Eq. (1), we see that the mean, Eq. (13), and variance, Eq. (14), are

$$\mu_x = \sum_{\mu} \mathbb{E}[u_{\mu} v_{\mu}] = \sum_{\mu} \mathbb{E}[u_{\mu}] \mathbb{E}[v_{\mu}], \quad (\text{C1})$$

$$\sigma_x^2 = \sum_{\mu} \text{var}(u_{\mu} v_{\mu}) + \sum_{\mu \neq v} \text{cov}(u_{\mu} v_{\mu}, u_v v_v), \quad (\text{C2})$$

where we imposed trivial modulation. We note that the expression for the variance of the product of two independent random variables is

$$\text{var}(u_{\mu} v_{\mu}) = (\sigma_u^2 + \mu_u^2)(\sigma_v^2 + \mu_v^2) - (\mu_u \mu_v)^2. \quad (\text{C3})$$

To compute the first term on the right-hand side of Eq. (C2), we use Eqs. (7), (5), and (C3) to find

$$\begin{aligned} \text{var}(u_{\mu} v_{\mu}) &= (\sigma_u^2 + \mu_u^2)(\sigma_v^2 + \mu_v^2) - (\mu_u \mu_v)^2 \\ &= (\sigma_u^2 + \mu_u^2)\sigma_v^2 = \left(\frac{1 - m^{-1}}{m(1 + m\beta_U)} + \frac{1}{m^2} \right) \sigma_v^2. \end{aligned} \quad (\text{C4})$$

The second term on the right-hand side of Eq. (C2) vanishes for $\mu \neq v$, since

$$\begin{aligned} \text{cov}(u_{\mu} v_{\mu}, u_v v_v) &= \mathbb{E}[(u_{\mu} v_{\mu} - \mathbb{E}(u_{\mu} v_{\mu}))(u_v v_v - \mathbb{E}(u_v v_v))] \\ &= \mathbb{E}[u_{\mu} v_{\mu} u_v v_v] - \mathbb{E}[u_{\mu} u_v] \mathbb{E}[v_{\mu} v_v] = 0, \end{aligned} \quad (\text{C5})$$

where we have used statistical independence of u and v , and the fact that v has zero mean. Therefore, the expression for the variance is

$$\sigma_x^2 = m \text{var}(u_{\mu} v_{\mu}) = \left(\frac{1 - m^{-1}}{1 + m\beta_U} + \frac{1}{m} \right) \sigma_v^2. \quad (\text{C6})$$

We verify this expression numerically in Fig. 6.

APPENDIX D: OBSERVABLE DENSITIES OF I.I.D. GAUSSIAN NOISE DATA

In this Appendix, we collect some known results on the distribution of correlations and eigenvalues of the correlation matrix for i.i.d. Gaussian noise data. Specifically, we consider a $T \times N$ matrix with i.i.d. Gaussian random variables as entries.

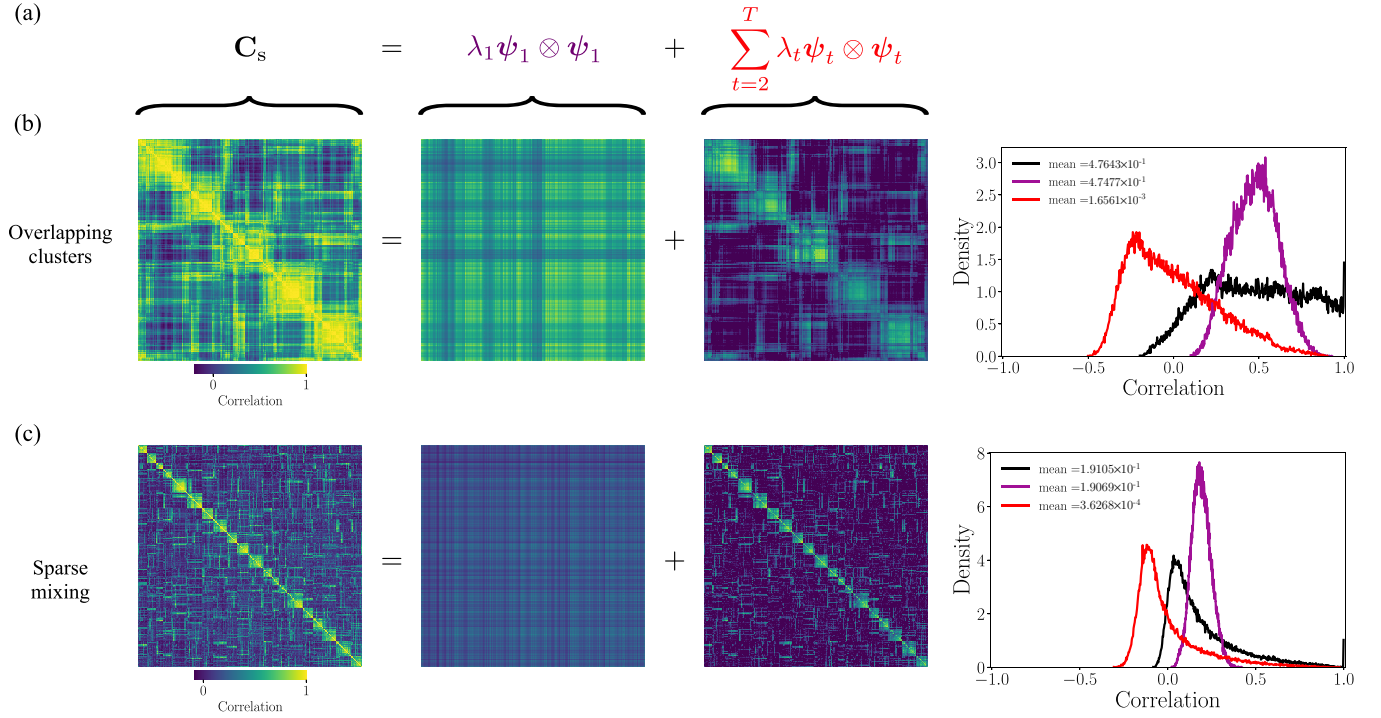


FIG. 8. Decomposition of the correlation matrix of the Dirichlet-Gaussian model. We show this for a single realization with $T = 360$. (a) General decomposition of the correlation matrix (black) into contributions of the top-ranked eigenmode ψ_1 (purple) and the sum of the remaining eigenmodes $\psi_2, \dots, \psi_T$ (red). The contributions are given by the outer product $\otimes$ of eigenmodes, weighted by the eigenvalues. (b) The overlapping clusters model with $N = 100$, $m = 5$, and $\beta_U = 0.4$. The distributions of matrix elements of the different contributions are shown on the right in their respective color. The numerical mean values are also stated. (c) The sparse mixing model with $N = 1000$, $m = 20$, and $\beta_U = 0.15$. To the extent that the mean of the matrix elements of the first mode is essentially indistinguishable from the means of the overall ECM, and the means of all other eigenvalue contributions are nearly zero, this shows that the mean correlation in the ECM is captured by the top-ranked eigenmode.

The correlation coefficient r between samples is given by the product of two mutually independent random vectors of length N with i.i.d. Gaussian entries, and the distribution of the correlation coefficient was derived in [45]. See also [34] for an extensive discussion. The coefficients are distributed according to a symmetric Beta distribution

$$\text{pdf}(r) = \text{Beta}(r; \alpha, \alpha; \ell = -1, s = 2) \quad (\text{D1})$$

$$= \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(\alpha)} (1 - r^2)^{\alpha-1}, \quad (\text{D2})$$

where Γ is the Gamma function, and ℓ and s are the location and scale parameter, respectively. The parameters ℓ and s are set such that the density is defined on the correlation interval $r \in [-1, 1]$, and

$$\alpha = \frac{N-1}{2}. \quad (\text{D3})$$

The variance of the distribution with the scale $s = 2$ is given by

$$\text{var} = \frac{s^2}{4(2\alpha + 1)} = \frac{1}{2\alpha + 1} = \frac{1}{N}. \quad (\text{D4})$$

For large values N , the distribution approaches the normal distribution with variance $1/N$.

Next, we turn to the eigenvalues of the sample-sample correlation matrix. The eigenvalues follow the famous Marčenko-Pastur distribution:

$$\rho(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi q\lambda} \quad (\text{D5})$$

with

$$\lambda_{\pm} = (1 \pm \sqrt{q})^2 \quad (\text{D6})$$

and $q = T/N$. We note that, as we consider the eigenvalue spectrum of the sample-sample correlation matrix $\mathbf{C}_s$, the sampling ratio $q = T/N$ is the inverse of the ratio N/T , which would apply for the variable-variable correlation matrix $\mathbf{C}_v$, more commonly encountered in applications.

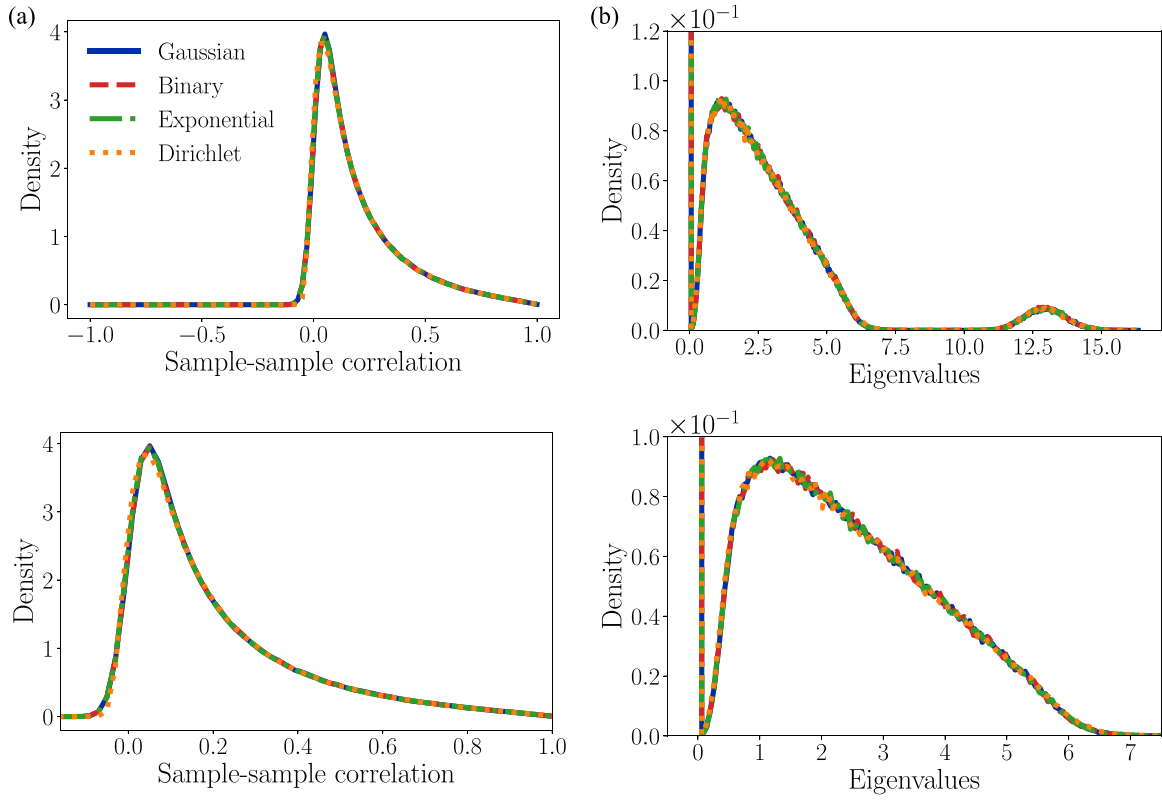


FIG. 9. Latent feature structure has almost no influence on the correlation and eigenvalue densities of the SUV model in the sparse mixing regime. We use $T = 60$, $N = 1000$, $m = 20$, $\beta_U = 0.15$, and $s_{t\mu} = 1$ for all entries. We test the following types of latent features $\mathbf{V}$: blue line, i.i.d. Gaussian entries for reference; dashed red, binary $\{\pm 1\}$ i.i.d. Dashed red, Bernoulli entries with $p = 0.5$; dash-dotted green, i.i.d. exponentially distributed entries with rate parameter $\lambda = 1$; dotted orange, Dirichlet random vectors with $\beta_V = 5 \times 10^{-2}$. (a) Correlation densities computed from 800 independent realizations of simulated data. Bottom plot is a zoom-in version of the top plot. (b) Eigenvalue densities computed from 2×10^4 independent realizations of simulated data. Bottom plot is a zoom-in version of the top plot.

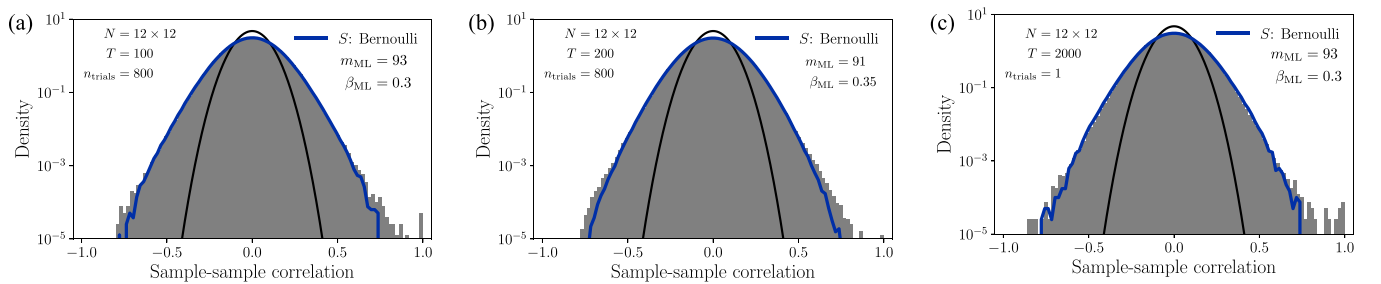


FIG. 10. Different approaches to sampling lead to similar sample-sample correlation distributions. Maximum-likelihood fits on the correlation distribution of image patches of size $N = 12 \times 12$ for different sampling approaches. (a) $T = 100$, $n_{\text{trials}} = 800$. (b) $T = 200$, $n_{\text{trials}} = 800$. (c) $T = 2000$, and a single sample, $n_{\text{trial}} = 1$.

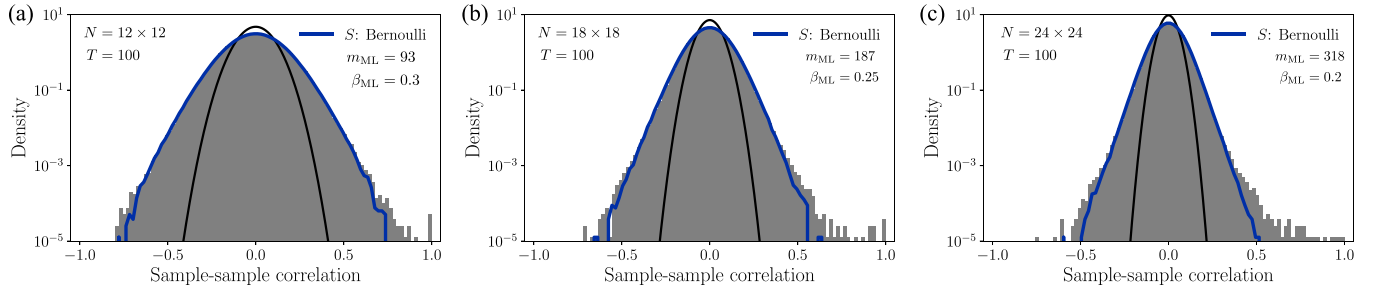


FIG. 11. The correlation distribution develops tails for increasing patch size. Maximum-likelihood fits on data of increasing patch size. (a) $N \times N = 12 \times 12$. (b) $N \times N = 18 \times 18$. (c) $N = 24 \times 24$. All simulations with $T = 100$ computed from 800 independent samples.

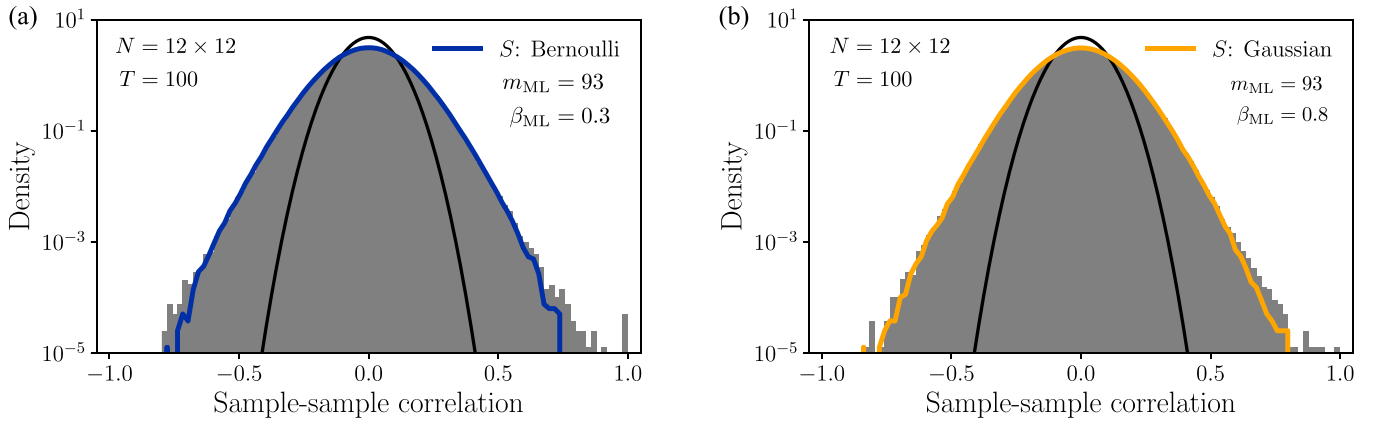


FIG. 12. The correlation distribution can be fit with $\mathbf{S}$ matrix elements following a Gaussian distribution. (a) $N \times N = 12 \times 12$, $T = 100$. (b) $N = 12 \times 12$ image patches with $T = 100$ and elements of $\mathbf{S}$ are drawn as i.i.d. Gaussian random variables. All simulations with $T = 200$ computed from 800 independent trials.

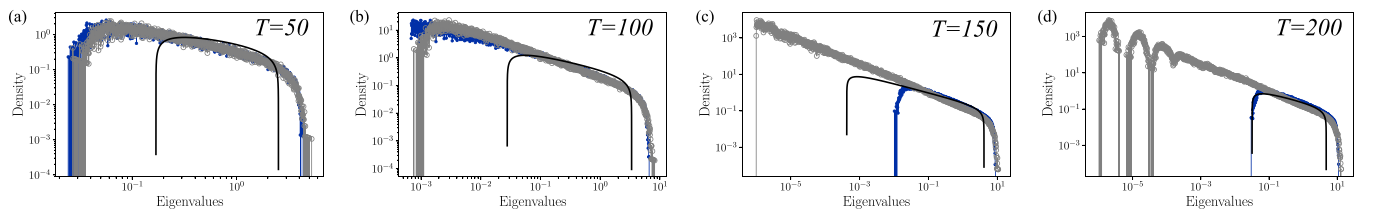


FIG. 13. The prediction quality of the eigenvalue spectrum is good for modest T but deteriorates for large T . Prediction of the eigenvalue spectrum (blue) made by maximum-likelihood fits on the correlation distribution for image patches of size $N = 12 \times 12$ for increasing numbers of observation: (a) $T = 50$ ($m_{\text{ML}} = 89$, $\beta_{\text{ML}} = 0.3$); (b) 100 ($m_{\text{ML}} = 92$, $\beta_{\text{ML}} = 0.3$); (c) 150 ($m_{\text{ML}} = 93$, $\beta_{\text{ML}} = 0.35$); and (d) 200 ($m_{\text{ML}} = 91$, $\beta_{\text{ML}} = 0.3$). The eigenvalue spectrum computed from data is shown in gray. The Marčenko-Pastur distribution (black) is shown for reference. For all simulations, we use 800 independent trials.

- [1] P. W. Anderson, More is different: Broken symmetry and the nature of the hierarchical structure of science, *Science* **177**, 393 (1972).
- [2] O. Marre, D. Amodei, N. Deshmukh, K. Sadeghi, F. Soo, T. E. Holy, and M. J. Berry, II, Mapping a complete neural population in the retina, *J. Neurosci.* **32**, 14859 (2012).
- [3] N. A. Steinmetz, C. Aydin, A. Lebedeva, M. Okun, M. Pachitariu, M. Bauza, M. Beau, J. Bhagat, C. Böhm, M. Broux, S. Chen, J. Colonell, R. J. Gardner, B. Karsh, F. Kloosterman, D. Kostadinov, C. Mora-Lopez, J. O'Callaghan, J. Park, J. Putzeys *et al.*, Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings, *Science* **372**, eabf4588 (2021).
- [4] M. Nagy, H. Naik, F. Kano, N. V. Carlson, J. C. Koblitz, M. Wikelski, and I. D. Couzin, Smart-barn: Scalable multimodal arena for real-time tracking behavior of animals in large numbers, *Sci. Adv.* **9**, eadf8068 (2023).
- [5] A. Cavagna, X. Feng, S. Melillo, L. Parisi, L. Postiglione, and P. Villegas, Como: A novel comoving 3d camera system, *IEEE Trans. Instrum. Meas.* **70**, 1 (2021).
- [6] E. Schneidman, M. J. Berry, II, R. Segev, and W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population, *Nature (London)* **440**, 1007 (2006).
- [7] 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).
- [8] D. L. Ruderman, Origins of scaling in natural images, *Vis. Res.* **37**, 3385 (1997).
- [9] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Protein sectors: Evolutionary units of three-dimensional structure, *Cell* **138**, 774 (2009).
- [10] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, Protein 3d structure computed from evolutionary sequence variation, *PLoS ONE* **6**, e28766 (2011).
- [11] D. L. Ruderman and W. Bialek, Statistics of natural images: Scaling in the woods, *Phys. Rev. Lett.* **73**, 814 (1994).
- [12] C. Qin and L. J. Colwell, Power law tails in phylogenetic systems, *Proc. Natl. Acad. Sci. USA* **115**, 690 (2018).
- [13] M. Nitzan and M. P. Brenner, Revealing lineage-related signals in single-cell gene expression using random matrix theory, *Proc. Natl. Acad. Sci. USA* **118**, e1913931118 (2021).
- [14] U. Tomasini and M. Wyart, How deep networks learn sparse and hierarchical data: The sparse random hierarchy model, *arXiv:2404.10727*.
- [15] <http://www.rctn.org/bruno/sparsenet/>.
- [16] B. A. Olshausen and D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature (London)* **381**, 607 (1996).
- [17] J. P. Cunningham and B. M. Yu, Dimensionality reduction for large-scale neural recordings, *Nat. Neurosci.* **17**, 1500 (2014).
- [18] P. Gao, E. Trautmann, B. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, A theory of multineuronal dimensionality, dynamics and measurement, *bioRxiv* (2017), doi: [10.1101/214262](https://doi.org/10.1101/214262).
- [19] J. A. Gallego, M. G. Perich, L. E. Miller, and S. A. Solla, Neural manifolds for the control of movement, *Neuron* **94**, 978 (2017).
- [20] S. M. Perkins, E. A. Amematsro, J. Cunningham, Q. Wang, and M. M. Churchland, An emerging view of neural geometry in motor cortex supports high-performance decoding, *eLife* **12**, RP89421 (2025).
- [21] 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).
- [22] 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).
- [23] 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).
- [24] 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).
- [25] J. Moran and M. Tikhonov, Defining coarse-grainability in a model of structured microbial ecosystems, *Phys. Rev. X* **12**, 021038 (2022).
- [26] J. Moran and M. Tikhonov, Emergent predictability in microbial ecosystems, *arXiv:2403.19372*.
- [27] D. Jordan, S. Kuehn, E. Katifori, and S. Leibler, Behavioral diversity in microbes and low-dimensional phenotypic spaces, *Proc. Natl. Acad. Sci. USA* **110**, 14018 (2013).
- [28] V. Ngampruetikorn, I. Nemenman, and D. J. Schwab, Extrinsic vs intrinsic criticality in systems with many components, *Phys. Rev. Res.* **7**, 013188 (2025).
- [29] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* **95**, 14863 (1998).
- [30] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the influence of data structure on learning in neural networks: The hidden manifold model, *Phys. Rev. X* **10**, 041044 (2020).
- [31] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature (London)* **401**, 788 (1999).
- [32] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, 2020).
- [33] F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, and L. Zdeborová, The role of regularization in classification of high-dimensional noisy gaussian mixture, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2020), pp. 6874–6883.
- [34] P. Fleig and I. Nemenman, Statistical properties of large data sets with linear latent features, *Phys. Rev. E* **106**, 014102 (2022).
- [35] R. B. Grosse, R. Salakhutdinov, W. T. Freeman, and J. B. Tenenbaum, Exploiting compositionality to explore a large space of model structures, in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI'12* (AUAI Press, Arlington, VA, 2012), pp. 306–315.
- [36] I. D. Landau, G. C. Mel, and S. Ganguli, Singular vectors of sums of rectangular random matrices and optimal estimation of high-rank signals: The extensive spike model, *Phys. Rev. E* **108**, 054129 (2023).
- [37] B. A. Olshausen and D. J. Field, Sparse coding with an over-complete basis set: A strategy employed by v1? *Vis. Res.* **37**, 3311 (1997).
- [38] I. Nemenman, F. Shafee, and W. Bialek, Entropy and inference, revisited, in *Advances in Neural Information Processing Sys*

- tems*, edited by T. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, Cambridge, MA, 2001), Vol. 14.
- [39] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, *J. Mach. Learn. Res.* **3**, 993 (2003).
 - [40] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen *et al.*, Toy models of superposition, [arXiv:2209.10652](https://arxiv.org/abs/2209.10652).
 - [41] https://en.wikipedia.org/wiki/Dirichlet_distribution.
 - [42] M. J. Berry and G. Tkačik, Clustering of neural activity: A design principle for population codes, *Front. Comput. Neurosci.* **14**, 20 (2020).
 - [43] <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html> (SciPy v1.7.1).
 - [44] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Noise dressing of financial correlation matrices, *Phys. Rev. Lett.* **83**, 1467 (1999).
 - [45] H. Hotelling, New light on the correlation coefficient and its transforms, *J. R. Stat. Soc. Ser. B (Methodol.)* **15**, 193 (1953).
 - [46] J. H. van Hateren and D. L. Ruderman, Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex, *Proc. R. Soc. London B* **265**, 2315 (1998).
 - [47] V. A. Marčenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, *Math. USSR-Sborn.* **1**, 457 (1967).
 - [48] S. Saremi and T. J. Sejnowski, Hierarchical model of natural images and the origin of scale invariance, *Proc. Natl. Acad. Sci. USA* **110**, 3071 (2013).
 - [49] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Statistical mechanics for natural flocks of birds, *Proc. Natl. Acad. Sci. USA* **109**, 4786 (2012).
 - [50] Y. Gao, E. W. Archer, L. Paninski, and J. P. Cunningham, Linear dynamical neural population models through nonlinear embeddings, in *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Red Hook, NY, 2016), Vol. 29.
 - [51] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2015), pp. 2256–2265.
 - [52] A. J. Bell and T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, *Neural Comput.* **7**, 1129 (1995).
 - [53] A. Hyvärinen and E. Oja, Independent component analysis: Algorithms and applications, *Neural Netw.* **13**, 411 (2000).
 - [54] J. Shlens, A tutorial on principal component analysis, [arXiv:1404.1100](https://arxiv.org/abs/1404.1100).
 - [55] W. Młynarski and G. Tkačik, Efficient coding theory of dynamic attentional modulation, *PLoS Biol.* **20**, e3001889 (2022).