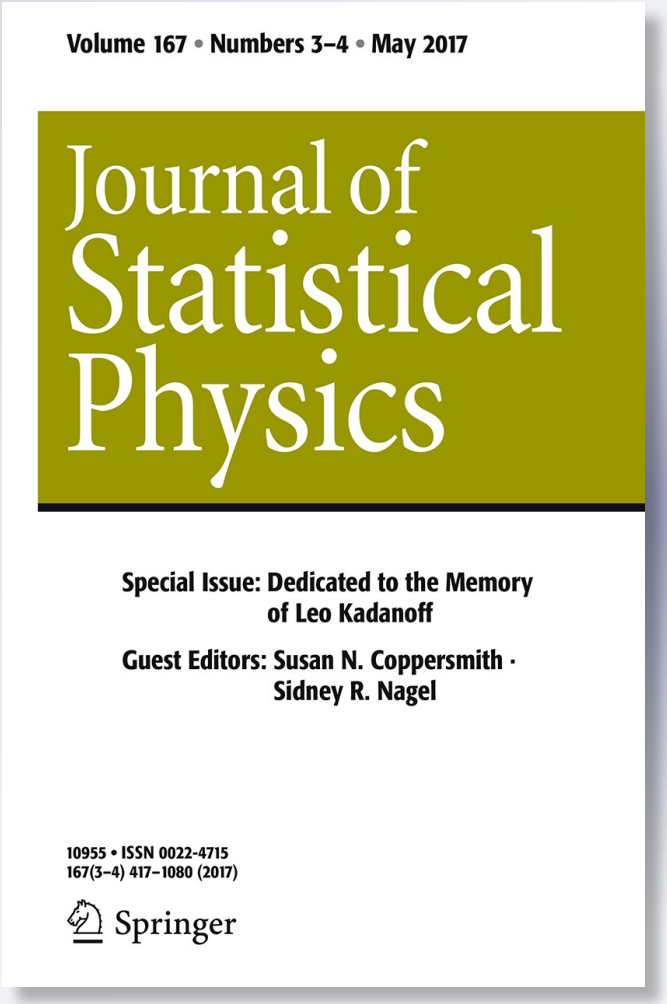
PCA Meets RG

Serena Bradde & William Bialek

Journal of Statistical Physics
1

ISSN 0022-4715
Volume 167
Combined 3-4

J Stat Phys (2017) 167:462-475
DOI 10.1007/s10955-017-1770-6



Springer

J Stat Phys (2017) 167:462–475
DOI 10.1007/s10955-017-1770-6

PCA Meets RG

Serena Bradde[1] · **William Bialek**[1,2]

Received: 29 October 2016 / Accepted: 17 March 2017 / Published online: 27 March 2017

Abstract A system with many degrees of freedom can be characterized by a covariance matrix; principal components analysis focuses on the eigenvalues of this matrix, hoping to find a lower dimensional description. But when the spectrum is nearly continuous, any distinction between components that we keep and those that we ignore becomes arbitrary; it then is natural to ask what happens as we vary this arbitrary cutoff. We argue that this problem is analogous to the momentum shell renormalization group. Following this analogy, we can define relevant and irrelevant operators, where the role of dimensionality is played by properties of the eigenvalue density. These results also suggest an approach to the analysis of real data. As an example, we study neural activity in the vertebrate retina as it responds to naturalistic movies, and find evidence of behavior controlled by a nontrivial fixed point. Applied to financial data, our analysis separates modes dominated by sampling noise from a smaller but still macroscopic number of modes described by a non-Gaussian distribution.

Keywords Renormalization group · Neural networks · Financial markets

1 Introduction

Many of the most interesting phenomena in the world around us emerge from interactions among many degrees of freedom. In the era of "big data," we are encouraged to think about this more explicitly, describing the state of a system as a point in a space with many dimensions: the state of a cell is defined by the expression level of many genes, the state of a financial market is defined by the prices of many stocks, and so on. One approach to the analysis of such high dimensional data is to look for a linear projection onto a lower dimensional

✉ William Bialek
wbialek@Princeton.EDU

[1] Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, 365 Fifth Ave., New York, NY 10016, USA

[2] Joseph Henry Laboratories of Physics, and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

space. Quantitatively, the best projection is found by diagonalizing the covariance matrix, which decomposes the variations into modes that are independent at second order; these are the principal components (PCs), and the method is called principal components analysis (PCA).[1] In favorable cases, very few modes will capture most of the variance, but it is much more common to find that the eigenvalues of the covariance matrix form a nearly continuous spectrum, so that any sharp division between important and unimportant dimensions would be arbitrary.

In the standard applications of PCA, the discussion focuses almost exclusively on the covariance matrix and its spectrum. In this case the crucial question is whether the structure that we see in the spectrum could have arisen at random, given that in analyzing real data we have access only to a finite number of samples, and this leads us to problems in random matrix theory. But behind the covariance matrix stands the entire (joint) probability distribution over the many variables in our system. In many cases it is the structure of this distribution that really interests us. In a network of neurons, for example, we might be searching for a distribution with many well resolved peaks, corresponding to the memories stored in a network with Hopfield-like dynamics [3]. But precisely because the system has many degrees of freedom, a direct visualization of this distribution is impossible. On the other hand, if PCA "works" and we can project onto just a few degrees of freedom, then we can visualize the distribution in this lower dimensional space. Is there any hope of getting at the structure of the full distribution without this collapse of dimensionality?

For physical systems in thermal equilibrium, it again is the case that the most interesting phenomena emerge from interactions among many degrees of freedom, but here we have a quantitative language for describing this emergence. In the classical view, we make precise models of the interactions on a microscopic scale, and then statistical mechanics is about calculating the implications of these interactions for the macroscopic behavior of matter. In the modern view, we admit that our microscopic description itself is approximate, incorporating "effective interactions" mediated by degrees of freedom that we might not want to describe explicitly, and that the distance scale at which we draw the boundary between explicit and implicit description also is arbitrary. Attention shifts from the precise form of our model to the way in which this model evolves as we move the boundary between degrees of freedom that we describe and those that we ignore [4]; the evolution through the space of possible models is described by the renormalization group (RG).

A central result of the renormalization group is that many detailed features of models on a microscopic scale disappear as we coarse grain our description out to the macroscopic scale, and that in many cases we are left with only a few terms in our models, the "relevant operators." Thus, some of the success of simple models in describing the world comes not from an inherent simplicity, but rather from the fact that macroscopic behaviors are insensitive to most microscopic details (irrelevant operators). This result usually is phrased in terms of coupling constants in an effective Hamiltonian, but what we actually manipulate in the course of an RG analysis is the probability distribution over all the degrees of freedom in the system. Thus, the existence of a small number of relevant operators is the statement that these distributions become simple as we average over short distance details. Our emphasis on following the RG flow of distributions, rather than Hamiltonians, is in the spirit of a very early discussion by Jona-Lasinio [5].

The RG approach to statistical physics suggests that systems in which PCA fails to yield a clean separation between high variance and low variance modes may nonetheless be simpli-

[1] The idea of PCA goes back at least to the start of the twentieth century [1]. For a brief modern summary, see Ref. [2].

fied. Indeed, in a system where the many degrees of freedom live on a lattice, with translation invariant interactions, the principal components are Fourier modes, and typically we find that the variance of each mode decreases monotonically but smoothly with decreasing wavelength. In the momentum shell implementation of RG [6], we put a cutoff on the wavelength, and ask what happens to the joint distribution of the remaining variables as we move this cutoff, averaging over the short wavelength modes. In this language, the RG is about what happens as we vary the arbitrary distinction between high variance PCs that we keep, and low variance PCs that we ignore. The goal of this paper is to clarify this connection between PCA and RG, so that we can construct RG approaches to more complex, high dimensional systems.

2 Historical Note

This paper has been written for a volume dedicated to the memory of Leo Kadanoff. As has been described many times, the modern development of the RG began with Leo's intuitive construction of "block spins," in which he made explicit the idea of averaging over the fluctuations that occur on short wavelengths [7]. Later in his life, Kadanoff worked on more complex problems, from the dynamics of cities [8,9] to patterns [10], chaos [11], and singularities [12,13] in fluid flows, and more [14,15]. Beyond his own work, he was a persistent advocate for the physics community's exploration of complex systems, including biological systems. We have benefited, directly and indirectly, from his enthusiasm, as well as being inspired by his example.

Among Leo's last papers are a series of historical pieces reflecting on his role in the development of the RG, and on statistical physics more generally [16–19]. Although much can be said about these papers, surely one message is that the physicist's persistent search for simplification has been rewarded, time and again. We offer this paper in that spirit, as we try to carry Kadanoff's intuition about thinning out microscopic degrees of freedom away from its origins in systems with local interactions.

3 Correlation Spectra and Effective Dimensions

Let us imagine that the system we are studying is described by a set of variables $\phi_1, \phi_2, \ldots, \phi_N \equiv \{\phi_i\}$, where the dimensionality N is large. For the purposes of this discussion, "describing the system" means writing down the joint probability of all N variables, $P(\{\phi_i\})$. For simplicity we define these variables so that they have zero mean, and we'll assume that positive and negative fluctuations are equally likely (though this is not essential). We start with the guess that the fluctuations are nearly Gaussian, so we can write

$$P(\{\phi_i\}) = \frac{1}{Z} \exp\left[-\frac{1}{2} \sum_{i,j} \phi_i K_{ij} \phi_j - \frac{1}{4!} g \sum_i \phi_i^4 + \cdots \right], \tag{1}$$

where the coefficient g allow us to describe weak kurtosis of the random variables. It may be useful to note that the probability distribution in Eq. (1) is the maximum entropy, and hence least structured, model consistent with the full covariance matrix and the mean kurtosis of all the variables; in this sense it is a minimal model. Much of what we will say here can generalized to the case where each variable has a different kurtosis, so there is a distinct g_i associated with each term ϕ_i^4.

If $g = 0$, we are describing a system in which fluctuations are Gaussian, and in this limit the matrix K_{ij} is the inverse of the covariance matrix

$$C_{ij} = \langle \phi_i \phi_j \rangle. \tag{2}$$

In the conventional application of renormalization group ideas, we can classify non-Gaussian terms as relevant or irrelevant: as we coarse grain our description from microscopic to macroscopic scales, do departures from a Gaussian distribution become more or less important? Our first goal is to show how we can export this idea to the more general setting, where the kernel K_{ij} does not have any symmetries such as translation invariance. To do this, we start near $g = 0$, and work in perturbation theory.

It is useful to write the eigenvalues λ_μ and eigenvectors $\{u_i(\mu)\}$ of the matrix K,

$$\sum_j K_{ij} u_j(\mu) = \lambda_\mu u_i(\mu), \tag{3}$$

so that the variations in $\{\phi_i\}$ can be decomposed into modes $\{\tilde{\phi}_\mu\}$,

$$\phi_i = \sum_\mu u_i(\mu) \tilde{\phi}_\mu; \tag{4}$$

if $g = 0$ then these modes are exactly the principal components. The Gaussian term becomes

$$\frac{1}{2} \sum_{i,j} \phi_i K_{ij} \phi_j = \frac{1}{2} \sum_\mu \lambda_\mu \tilde{\phi}_\mu^2, \tag{5}$$

and hence at $g \to 0$ the variance of each mode is given by $\langle \tilde{\phi}_\mu^2 \rangle = 1/\lambda_\mu$. The average variance of the individual variables is

$$\frac{1}{N} \sum_i \langle \phi_i^2 \rangle = \frac{1}{N} \sum_\mu \frac{1}{\lambda_\mu} \to \int_0^\Lambda d\lambda\, \rho(\lambda) \frac{1}{\lambda}, \tag{6}$$

where in the last step we introduce the distribution

$$\rho(\lambda) = \frac{1}{N} \sum_\mu \delta(\lambda - \lambda_\mu), \tag{7}$$

which becomes smooth in the limit of large N, and we note explicitly that there is a largest eigenvalue Λ.[2]

The essential idea is to eliminate the modes that have small variance. This corresponds to restricting our attention only to modes with λ *less* than some cutoff. Equivalently, it corresponds to decreasing the limit Λ on the integral over eigenvalues, e.g. in Eq. (6). This reduces the total variance, but it is natural to choose units in which the variance is fixed, and this implies that as we change the cutoff Λ we have to rescale the values of ϕ_i. So we replace $\phi_i \to z_\Lambda \phi_i$, and we can determine this scale factor by insisting that the mean variance stay fixed. Again, we are working at small g, so we do this calculation at $g = 0$:

$$0 = \frac{d}{d\Lambda} \left[\frac{1}{N} \sum_i \langle (z_\Lambda \phi_i)^2 \rangle \right] \tag{8}$$

[2] An alternative formulation treats the smallest eigenvalue separately, as with a mass term in field theory, measuring all eigenvalues by their distance from this minimum. Then $\rho(\lambda)$ would always have, as $N \to \infty$, support near $\lambda = 0$.

$$= \frac{d}{d\Lambda}\left[z_\Lambda^2 \int_0^\Lambda d\lambda\, \frac{\rho(\lambda)}{\lambda}\right] \tag{9}$$

$$\Rightarrow \frac{d\ln z_\Lambda}{d\ln\Lambda} = -\frac{1}{2}\rho(\Lambda)\left[\int_0^\Lambda d\lambda\, \rho(\lambda)\frac{1}{\lambda}\right]^{-1}. \tag{10}$$

When we reduce the cutoff, we also reduce the number of degrees of freedom in the system. The average of the quadratic term in the (log) probability distribution is automatically proportional to this effective number of degrees of freedom,

$$N_{\rm eff} = N\int_0^\Lambda d\lambda\, \rho(\lambda), \tag{11}$$

and this insures, for example, that the entropy of the probability distribution will be proportional to $N_{\rm eff}$ (extensivity). To be sure that this works also for the quartic terms, we write

$$Ng\left[\frac{1}{N}\sum_{\rm i}\phi_{\rm i}^4\right] = N_{\rm eff}\tilde{g}\left[\frac{1}{N}\sum_{\rm i}(z_\Lambda\phi_{\rm i})^4\right], \tag{12}$$

which defines the effective coupling constant

$$\tilde{g} = z_\Lambda^{-4} g\frac{N}{N_{\rm eff}}. \tag{13}$$

Now the scaling of the coefficient $\tilde{g}$ is given by

$$\frac{d\ln\tilde{g}}{d\ln\Lambda} = \rho(\Lambda)\left[\frac{2}{\int_0^\Lambda d\lambda\,\rho(\lambda)\frac{1}{\lambda}} - \frac{\Lambda}{\int_0^\Lambda d\lambda\,\rho(\lambda)}\right]. \tag{14}$$

Since this is the difference between two positive terms, we can find either sign for the result.

If the scaling function $d\ln\tilde{g}/d\ln\Lambda$ is positive, then as we decrease the cutoff Λ any small quartic term $\tilde{g}$ will become still smaller, and hence the distribution approaches a Gaussian. This seems to make sense, since decreasing Λ corresponds to averaging over more and more of the low variance modes, which means that each of the variables that remains is a weighted sum of many of the original variables; under these conditions we might expect the central limit theorem to enforce approximate Gaussianity of the resulting distribution. But if the scaling function $d\ln\tilde{g}/d\ln\Lambda < 0$, then as we average over more and more of the lower variance modes, the quartic term becomes more and more important to the structure of the distribution. To use the language of the RG, under these conditions the quartic term is a relevant operator.

If we consider the case where the density of eigenvalues is a power law, $\rho = B\lambda^{\alpha-1}$, we find

$$\frac{d\ln\tilde{g}}{d\ln\Lambda} = \alpha - 2. \tag{15}$$

Thus, the spectral density of eigenvalues determines the relevance of non-Gaussian terms in the distribution.

In the conventional field theoretic examples, where the variables ϕ live at positions $\mathbf{x}$ in a D dimensional Euclidean space, the correlations come from a "kinetic energy" term that enforces similarity among neighbors,

$$\frac{1}{2}\sum_{\rm i,j}\phi_{\rm i}K_{\rm ij}\phi_{\rm j} \to \frac{1}{2}\int d^Dx\,[\nabla\phi(\mathbf{x})]^2\,. \tag{16}$$

The eigenvectors of K are Fourier modes, indexed by a wave vector $\mathbf{k}$, with eigenvalues $\lambda = |\mathbf{k}|^2$. If the original variables are on a lattice with linear spacing a, then there is a maximum eigenvalue $\Lambda \sim (\pi/a)^2$, and the density

$$\rho(\lambda) \propto \int d^D k \, \delta\left(\lambda - |\mathbf{k}|^2\right) \propto \lambda^{D/2-1}, \tag{17}$$

corresponding to $\alpha = D/2$. From Eq (15) we find

$$\frac{d \ln \tilde{g}}{d \ln \Lambda} = \frac{D}{2} - 2 = \frac{1}{2}(D-4). \tag{18}$$

The quartic term is relevant if $d \ln \tilde{g}/d \ln \Lambda < 0$, which corresponds to $D < 4$, as is well known from the conventional RG analysis [20,21]; the extra factor of $1/2$ arises because Λ is a cutoff on the eigenvalue, which is the square of the wavevector.

Everything that we have said here can be carried over to the case where each variable is associated with a different coupling g_i. This corresponds to a maximum entropy description that captures the pairwise correlations among the variables and the kurtosis of each individual variable. The relevance or irrelevance of each term is controlled, in the same way, by the eigenvalue spectrum. If we include terms $\sim \phi_i^n$, allowing us to match higher moments of the marginal distributions for each ϕ_i, then as usual these terms are less relevant at larger n. These results suggest that the renormalization group may provide, as we hoped, a path to controlling the complexity of models, even outside the usual context of statistical field theory with local interactions.

4 Can We Find Fixed Points?

Thus far our analysis has been confined to an analog of "power counting" in the conventional applications of the renormalization group. The next step is to integrate out the low variance degrees of freedom and compute corrections to the coupling constants that are beyond those generated from the spectrum of eigenvalues itself. Here we give a sketch of this calculation, following the conventional arguments as closely as possible.

We have a formulation in terms of discrete modes, so we can write $\phi_i \to \phi_i + u_i \psi$, where ψ is the variable describing fluctuations in the "last mode" that we have kept in our description. Our task is to average over the fluctuations in this last mode, and see how this changes the distribution of the remaining variables:

$$\exp\left[-\frac{\tilde{g}}{4!}\frac{N_{\text{eff}}}{N}\sum_i (z_\Lambda \phi_i)^4\right] \to \left\langle \exp\left[-\frac{\tilde{g}}{4!}\frac{N_{\text{eff}}}{N}\sum_i z_\Lambda^4 (\phi_i + u_i\psi)^4\right]\right\rangle. \tag{19}$$

In the limit of small g, ψ is Gaussian with $\langle \psi^2 \rangle = 1/\Lambda$, and we find

$$\left\langle \exp\left[-\frac{\tilde{g}}{4!}\frac{N_{\text{eff}}}{N}\sum_i z_\Lambda^4 (\phi_i + u_i\psi)^4\right]\right\rangle \tag{20}$$

$$= \exp\left[-\frac{\tilde{g}}{2}\frac{N_{\text{eff}}}{N}\sum_i z_\Lambda^4 \frac{u_i^2}{\Lambda}\phi_i^2 - \frac{\tilde{g}}{4!}\frac{N_{\text{eff}}}{N}\sum_i (z_\Lambda \phi_i)^4 + \frac{1}{2}\left(\frac{\tilde{g} N_{\text{eff}}}{4! N}\right)^2 \sum_{i,j} \frac{z_\Lambda^8}{\Lambda^2}\left(72\phi_i^2\phi_j^2 u_i^2 u_j^2 + 96\phi_i^3\phi_j u_i u_j^3\right) + \cdots\right]. \tag{21}$$

The first term is a correction to the matrix K, analogous to a mass renormalization. In the general case we not only get corrections to the coefficient of ϕ_i^4, we also generate terms $\sim\phi_i^2\phi_j^2$ and $\sim\phi_i^3\phi_j$. As in the standard discussion, we will assume that the fields ϕ_i are "slowly varying" functions of their index. More precisely, we will expand the correction terms around the point where $\phi_i = \phi_j$, and for now we drop the gradient-like terms $\sim(\phi_i - \phi_j)$, $\sim(\phi_i - \phi_j)^2$, …. In this approximation, we have

$$\sum_{i,j}\phi_i^2\phi_j^2u_i^2u_j^2 \approx \sum_i\phi_i^4u_i^2\sum_j u_j^2 \tag{22}$$

$$= \sum_i\phi_i^4u_i^2 \approx \frac{1}{N}\sum_i\phi_i^4, \tag{23}$$

where in the last step we again use the slow variation of ϕ_i to replace u_i^2 with its average. In the same approximation, the term $\sim u_iu_j^3$ vanishes. The net result is that

$$\tilde{g} \to \tilde{g} - \frac{3}{2}\tilde{g}^2\frac{N_{\rm eff}}{N}\frac{z_\Lambda^4}{N}\frac{1}{\Lambda^2}. \tag{24}$$

This is the change in coupling associated with integrating out one mode, which corresponds to a change in the cutoff such that $-d\Lambda\rho(\Lambda)N = 1$, so we can rewrite Eq (24) as

$$\frac{d\ln\tilde{g}}{d\ln\Lambda} = \frac{3}{2}\tilde{g}\frac{N_{\rm eff}}{N}z_\Lambda^4\frac{\rho(\Lambda)}{\Lambda}. \tag{25}$$

Combining with the scaling behavior in Eq (14), we find

$$\frac{d\ln\tilde{g}}{d\ln\Lambda} = \rho(\Lambda)\left[\frac{2}{\int_0^\Lambda d\lambda\,\rho(\lambda)\frac{\lambda}{\Lambda}} - \frac{\Lambda}{\int_0^\Lambda d\lambda\,\rho(\lambda)} + \frac{3}{2}\frac{\tilde{g}}{\Lambda}\right]. \tag{26}$$

In the case where $\rho(\lambda) \propto \lambda^{\alpha-1}$, this generates a fixed point $\tilde{g}_* \propto 2 - \alpha$, which is analogous to the Wilson-Fisher fixed point $\tilde{g}_* \propto 4 - D$ [20].

The calculation we have done here is aimed at showing that the conventional analysis of fixed points can be carried over to this different setting, away from equilibrium statistical physics with local interactions. We assume that this more complex setting allows for a richer variety of fixed points, which need to be explored.

5 An Approach to Data Analysis?

These arguments suggest that, at least in perturbation theory, much of the apparatus of the renormalization group for translation invariant systems with local interactions can be carried over to more complex systems. We can define relevant and irrelevant operators, and there is a path to identifying fixed points. The crucial role played by the dimensionality in systems with local interactions is played instead by the spectrum of the matrix K. Perhaps most important is that we can carry over the *concept* of renormalization.

Faced with real data on a system with many degrees of freedom, we don't know the matrix K. We do know that, if the system is close to being Gaussian, then K is close to being the inverse of the covariance matrix C, which we can estimate from the data. In systems with translation invariance, the eigenvectors of C and K are the same, which means that coarse graining by eliminating the modes with large eigenvalues of K (momentum shells) is exactly the same as eliminating the modes with small eigenvalues of C. Although this can't be true in

general, we can nonetheless try to use the eigenvalues of C as a way of ordering the degrees of freedom along an axis from the "details" that we want to ignore out to the macroscopic, collective variables that we suspect are most important. We implement this in several steps:

(1) We examine the spectrum of the covariance matrix. If a small number of eigenvalues are separated from the bulk, and capture most of the variance, then the system is genuinely low dimensional and we are done (PCA works). More commonly, we find a near continuum of eigenvalues, with no natural separation.
(2) Power-law behavior in a rank-ordered plot of the eigenvalues is analogous to power-law correlation functions in the usual field theoretic or statistical physics examples. In practice, however, it may be difficult to verify power-law behavior over a very wide range of scales.
(3) We coarse grain our description by projecting out a fraction of the modes with the smallest eigenvalues of C. In effect this replaces each variable ϕ_i by an average over low variance details $\phi_i \to \psi_i$, in the spirit of the block spin construction.

    More concretely, we start with variables $\{\phi_i\}$, with $i = 1, 2, \ldots, N$, as before, and we take these to have zero mean. We construct the covariance matrix C_{ij}, and find its eigenvalues and eigenvectors,[3]

$$\sum_j C_{ij} u_j(\mu) = \lambda_\mu u_i(\mu). \tag{27}$$

    We put these eigenvalues in order from largest ($\mu = 1$) to smallest ($\mu = N$). Coarse graining is a projection onto the subsets of modes that make the largest contributions to the total variance,

$$\phi_i \to \psi_i = \sum_j \hat{P}_{ij} \phi_j, \tag{28}$$

    where the projection operator is

$$\hat{P}_{ij} = \sum_{\mu=1}^{K} u_i(\mu) u_j(\mu), \tag{29}$$

    and we can think of this either as function of the fraction of the modes that we keep ($f = K/N$) or of the cutoff on the eigenvalues of the inverse covariance matrix, which connects more closely to our discussion above ($\Lambda = 1/\lambda_K$).
(4) To follow the results of coarse graining, we can measure the moments of the local variables, $\langle \psi_i^n \rangle$, or even their full distribution, as was done long ago for Monte Carlo data by Binder [22].

Notice that from the traditional point of view in applications of PCA, "coarse graining" is entirely trivial: we just keep some of the principal components and discard the others. But what is happening to the joint probability distribution of the remaining variables in the system need not be so trivial; indeed, this nontrivial evolution of the joint distribution, or the effective Hamiltonian, is the whole point of the conventional RG analysis. While the joint distribution is impossible to sample reliably in any realistic experiment or simulation, the point of Binder's construction is that we can see reflections of its changing structure

[3] In the analytic discussion of model distributions, above, the natural quantities were the eigenvalues and eigenvectors of the matrix K_{ij}. As noted, we don't have access to this matrix when we are confronted with real data, so we analyze the matrix C_{ij} instead. To emphasize that what we are doing with the data is in the same spirit as the analysis of the models, we abuse notation slightly and recycle the symbols $\{\lambda_\mu, u_i(\mu)\}$.

by monitoring the distribution of the individual variables. In the classical case, the bimodal distribution of a raw Ising variable evolves into the distribution of local magnetizations, which may be Gaussian away from a critical point or non-Gaussian at the critical point. Importantly, we never try to estimate a distribution over multiple variables, and so we never suffer from the curse of dimensionality.

5.1 A Network of Neurons

As a first example, we have analyzed an experiment on the activity of 160 neurons in a small patch of the vertebrate retina as it responds to naturalistic movies [23]; a full description will be given elsewhere, but here we focus on our ability to detect a nontrivial renormalization group flow as we coarse grain this system. As in previous analyses of these data, we divide time into small bins (width $\Delta\tau = 20$ ms), and in each bin a single neuron either generates an action potential or remains silent, so that the natural local variables are binary before any coarse graining. We can then take a state of the entire system to be the 160-dimensional vector of these binary variables, but we can also consider T successive vectors in time, as in the "time delayed embedding" analysis of dynamical systems [24]: increasing T compensates for not observing directly all the relevant degrees of freedom in the system, and gives us access to a higher dimensional description. The experiment runs for roughly one and one half hours (5660 s), and we estimate that this generates $\sim 10^5$ independent samples [23]. Given this size of the data set, we can use $T = 8$ without creating problems of undersampling, and this gives us $N = 1280$ dimensions. Because the different neurons are different from one another, we normalize each variable ϕ_i to have zero mean and unit variance, so the covariance matrix is the matrix of correlation coefficients in the raw data.

As we can see at left in Fig. 1, the eigenvalues of the correlation matrix have an essentially continuous spectrum, perhaps even showing hints of scale invariance.[4] This spectral structure is well outside the range that would be generated by an equally large random sample from uncorrelated variables.

Because the raw variables of this system are binary, the normalized fourth moments $(\langle\psi_i^4\rangle/\langle\psi_i^2\rangle^2)$ can be large and vary substantially from neuron to neuron. As we coarse grain, eliminating modes corresponding to small eigenvalues of the covariance matrix, this variation is reduced, as shown at right in Fig. 1. More strikingly, the normalized fourth moments are hardly varying as we move our cutoff beyond the first $\sim$90% of the modes, and the median value is stabilizing well above the value of 3 expected for a Gaussian distribution. It is interesting that the range of scales (fraction of modes included) over which see an approximately fixed fourth moment is the same as the range over which see approximately power-law behavior of the eigenvalue spectrum. These results suggest that the joint distribution of activity in this neural network is close to a nontrivial fixed point of the renormalization group transformation. This is consistent with previous evidence that this system is close to a critical point in the thermodynamic sense [25,26], but the renormalization group analysis connects more fully to our understanding of criticality in equilibrium systems.

[4] In some contexts it would be more natural to look at the distribution of eigenvalues, searching for modes that emerge clearly from a "bulk" that might be ascribed to sampling noise. Plotting eigenvalues vs their rank, as we do here, provides a representation of the cumulative distribution of eigenvalues, and does not require us to make bins along the eigenvalue axis. Rather than plotting from smallest to largest, we plot from largest to smallest, so that the spectra are more directly comparable to a plot of the susceptibility or propagator $G(k)$ vs momentum k in the usual statistical physics examples.

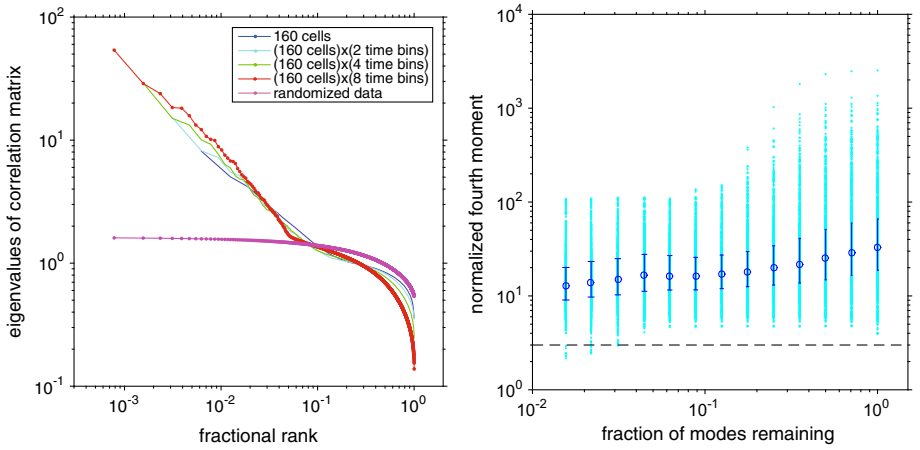


Fig. 1 Analysis of neural activity in the retina. States are defined by the patterns of spiking and silence in successive time bins from 160 neurons, as described in the text. *Left* The spectrum of eigenvalues of the correlation matrix, with states constructed from different numbers of time bins, plotted versus fractional mode number in descending order; results from randomized data shown for comparison. *Right* Normalized fourth moments for each of the 8×160 variables (*cyan dots*), as a function of the fraction of modes remaining after the coarse graining procedure; *blue circles* show medians $\pm$ one quartile. More precisely, what we plot is $\langle \psi_i^4 \rangle / \langle \psi_i^2 \rangle^2$, with the coarse grained variables ψ_i defined by Eqs. (28, 29). The dashed line is the value of this normalized moment for Gaussian random variables. The plot suggests that the fourth moments flow to a non-trivial fixed value, well above the Gaussian prediction (Color figure online)

5.2 Daily Returns on the NYSE

As a second example, we consider a set of 4000 assets traded on the New York Stock Exchange [27,28]. In the data, spanning nearly ten years from 1 Jan 1990 through 30 Apr 1999, 2445 assets appear for more than 2300 days out of the total of 2356 trading days, and we focus on a random subset of $N = 2048$ from this group. On each day t an asset i opens at price $p_i^{\text{open}}(t)$ and closes at price $p_i^{\text{close}}(t)$; we define the state of the system on day t by the vector of daily returns $\{r_i(t)\}$, with $r_i(t) = \ln[p_i^{\text{close}}(t)/p_i^{\text{open}}(t)]$. At left in Fig. 2 we see the spectrum of eigenvalues of the correlation matrix for these variables. In contrast to the neural data, the number of samples here is comparable to the dimensionality of the system, so we expect that the spectrum will be substantially affected by random sampling. Indeed, if we project out the ten percent of modes with the largest variance (opposite to our RG procedure), the resulting spectrum is very close to the predictions of the Marchenko–Pastur distribution for covariance matrices constructed from samples of uncorrelated variables [29].

There is a large literature on the analysis of correlation matrices for financial data, with more recent work emphasizing the use of random matrix theory as a null model or perhaps even a tool for "cleaning" the inferences that can be drawn from finite data sets. As emphasized at the outset, what stands behind the correlation matrix is the full joint distribution over the many degrees of freedom in the system, and it is this distribution that we would like to get at using RG ideas. The comparison of the eigenvalue spectrum with the predictions of random matrix theory does suggest, however, that with the resolution available in these data, integrating out many of the low variance degrees of freedom will correspond simply to removing random noise, and only once this noise is removed will we see features of the market itself.

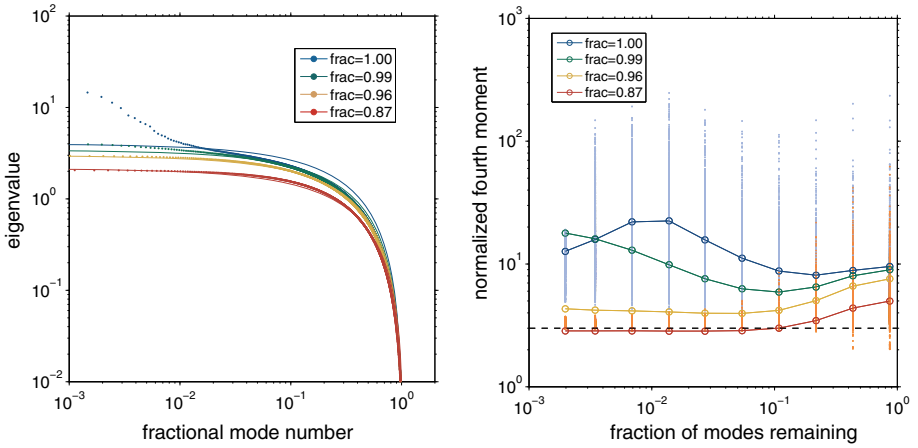


Fig. 2 Analysis of daily returns on $N = 2048$ assets in the NYSE for a period of $T = 2356$ days. *Left* Eigenvalues of the correlation matrix, in descending order, as a function of fractional mode number; results are shown for all the data (*blue*), as well as cases in which we remove the largest 1% (*green*), 4% (*yellow*), or 13% (*red*) of the eigenvalues. *Solid lines* are theoretical expectations from the Marchenko–Pastur distribution [29]. *Right* The flow of the normalized fourth moments for each of the $N = 2048$ variables ψ_i when we integrate out a fraction of high eigenmode. The normalization procedure after eingenmode integration is described in the main text. *Colors* code different cases where we track all the modes, or first remove different fractions of the large variance modes, as in the analysis on the left. The *dashed line* is the prediction for Gaussian random variables. We see that the fourth moments have a nonmonotonic behavior when all eigenvalues are included, while they flow rapidly to the Gaussian fixed point when the top 10% of eigenvalues are removed (Color figure online)

If we apply our RG procedure to the raw data, we see a non-monotonic trajectory of the fourth moments, first moving toward the Gaussian fixed point and then away (Fig. 2, right). The turning point is roughly when we have integrated out all but the last ten percent of high variance modes, which is consistent with the lower $\sim 90\%$ of the eigenvalue spectrum being well described by the Marchenko–Pastur distribution. Indeed, if we first exclude the top ten percent of high variance modes, the fourth moments flow very quickly to the Gaussian fixed point and the third moments flow to zero (not shown). The ten percent of high variance modes clearly are not just noise, however, although it is not clear from the data whether their distribution is described by a fixed point of the RG. We emphasize that the boundary between noise-like and non-noise modes is a property not of the system, but of the finite sample of data; it is possible that the RG analysis we propose here could be combined with denoising [30,31] to give more insight.

6 Not Quite Conclusions

The idea that the RG might be useful in more complex systems is a widely held intuition. In particular, there have been efforts to move from regular lattices to graphs [32], as well as to construct a real space renormalization group for spin glasses [33–35]. What is new here, we think, is that, at least in perturbation theory, we can free ourselves completely from assumptions of locality, which seem so crucial to the usual notions of relevant and irrelevant operators. Perhaps more importantly, connecting RG and PCA allows us to look at data in a new way, with interesting results in two very different complex systems.

PCA is a search for simplification. The hope is that a system with variables that live in a high dimensional space can be captured by a projection of these variables into a low dimensional space. Although the RG involves (repeated) projections onto lower dimensional spaces, this dimensionality reduction is *not* the source of simplification. Indeed, when we study models, the full renormalization group transformation involves expanding the system back to restore the original number of degrees of freedom; admittedly, this is difficult to do with real data. RG is a search for simplification, not in the space of the system variables but in the space of models.

An interesting connection is to the question of how well different terms in a model are determined by experimental data. Starting with models for biological signaling networks [36], Sethna and colleagues have argued that models for complex systems typically have a wide range of parameter sensitivities, so that some directions in parameter space have coordinates that are easily determined by data while other directions are almost never determined. This pattern is quantified by the spectrum of eigenvalues in the Fisher information matrix (FIM), and in many cases this spectrum is nearly uniform on a logarithmic scale [37,38], a property termed "sloppiness." Many of these models can be written so that parameter spaces are compact, and simplification then is achievable by moving along the ill-determined directions until reaching the edge of the space, leaving a model with one less parameter [39]. Recent work has shown that conventional statistical physics models do not exhibit sloppiness if experiments involve measurements on the microscopic scale, but that this pattern develops when measurements are restricted to coarse grained variables [40]. The spreading of the FIM eigenvalues is controlled by the RG scaling of the different operators out of which the model is constructed, suggesting that the notions of simplification that are inherent to the RG are equivalent to a more data-driven simplification in which we keep only model components that are well determined by experiment. It is possible that there are even more direct connections between the renormalization group and the learning of probabilistic models [41].

In the conventional implementations of the renormalization group, we put variables in order by their length scale, with small length scales at one end and long length scales at the other. The intuition is that, when interactions are local, smaller scales are less important, or at least less interesting, and so we average over scales shorter than some distance ℓ. The RG then is the exploration of what happens as we change ℓ. In more complex systems, simplification requires us to find a natural coordinate system in state space, and then put these coordinates in order of their likely importance, with fine-grained details at one end and crucial collective degrees of freedom at the other. The spectrum of the covariance matrix gives us one possible answer to these questions, which we have explored here, but surely there are other possibilities, even in the two examples discussed above. The more significant idea is that once we have identified an axis along which coarse graining seems to make sense, rather than looking for the right place to put the boundary between what we include and what we ignore, we should use the RG as inspiration to explore the evolution of our description as we move this boundary.

Acknowledgements We thank D Amodei, MJ Berry II, and O Marre for making available the data of Ref. [23] and M Marsili for the data of Ref. [27]. We are especially grateful to G Biroli, J–P Bouchaud, MP Brenner, CG Callan, A Cavagna, I Giardina, MO Magnasco, A Nicolis, SE Palmer, G Parisi, and DJ Schwab for helpful discussions and comments on the manuscript. Work at CUNY was supported in part by the Swartz Foundation. Work at Princeton was supported in part by Grants from the National Science Foundation (PHY-1305525, PHY-1451171, and CCF-0939370) and the Simons Foundation.

References

1. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philos. Mag. **2**, 559–572 (1901)
2. Shlens, J.: A tutorial on principal components analysis. arXiv:1404.1100 [cs.LG] (2014)
3. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA **79**, 2554–2558 (1982)
4. Wilson, K.G.: Problems in physics with many scales of length. Sci. Am. **241**, 158–179 (1979)
5. Jona-Lasinio, G.: The renormalization group: a probabilistic view. Il Nuovo Cimento **26B**, 99–119 (1975)
6. Wilson, K.G., Kogut, J.: The renormalization group and the ϵ expansion. Phys. Rep. **12**, 75–200 (1974)
7. Kadanoff, L.P.: Scaling laws for Ising models near T_c. Physics **2**, 263–272 (1966)
8. Kadanoff, L.P.: From simulation model to public policy: an examination of Forrester's "Urban Dynamics". Simulation **16**, 261–268 (1971)
9. Kadanoff, L.P., Weinblatt, H.: Public policy conclusions from urban growth models. IEEE Trans. Syst. Man Cybern. **SMC–2**, 139–165 (1972)
10. Bensimon, D., Kadanoff, L.P., Liang, S., Shraiman, B.I., Tang, C.: Viscous flows in two dimensions. Rev. Mod. Phys. **58**, 977–999 (1986)
11. Halsey, T.C., Jensen, M.H., Kadanoff, L.P., Procaccia, I., Shraiman, B.I.: Fractal measures and their singularities: the characterization of strange sets. Phys. Rev. A**33,** 1141–1151 (1986); erratum **34,** 1601 (1986)
12. Constantin, P., Kadanoff, L.P.: Singularities in complex interfaces. Philos. Trans. R. Soc. Lond. Ser. A **333**, 379–389 (1990)
13. Bertozzi, A., Brenner, M., Dupont, T.F., Kadanoff, L.P.: Singularities and similarities in interface flows. In: Sirovich, L.P. (ed.) Trends and Perspectives in Applied Mathatematics. Springer Verlag Applied Math Series Vol. 100, pp. 155–208 (1994)
14. Coppersmith, S.N., Blank, R.D., Kadanoff, L.P.: Analysis of a population genetics model with mutation, selection, and pleitropy. J. Stat. Phys. **97**, 429–459 (1999)
15. Povinelli, M.L., Coppersmith, S.N., Kadanoff, L.P., Nagel, S.R., Venkataramani, S.C.: Noise stabilization of self-organized memories. Phys. Rev. E **59**, 4970–4982 (1999)
16. Kadanoff, L.P.: More is the same: mean field theory and phase transitions. J. Stat. Phys. **137**, 777–797 (2009)
17. Kadanoff, L.P.: Relating theories via renormalization. Stud. Hist. Philos. Sci. B **44**, 22–39 (2013)
18. Kadanoff, L.P.: Reflections on Gibbs: from statistical physics to the Amistad. J. Stat. Phys. **156**, 1–9 (2014)
19. Kadanoff, L.P.: Innovations in statistical physics. Annu. Rev. Cond. Matter Phys. **6**, 1–14 (2015)
20. Wilson, K.G., Fisher, M.E.: Critical exponents in 3.99 dimensions. Phys. Rev. Lett. **28**, 240–243 (1972)
21. Amit, D.J., Martin-Mayor, V.: Field Theory, the Renormalization Group, and Critical Phenomena. Graphs to Computers, 3rd edn. World Scientific, Singapore (2005)
22. Binder, K.: Finite size scaling analysis of Ising model block distribution functions. Z. Phys. B **43**, 119–140 (1981)
23. Tkačik, G., Marre, O., Amodei, D., Schneidman, E., Bialek, W., Berry II, M.J.: Searching for collective behavior in a large network of sensory neurons. PLoS Comput. Biol. **10**, e1003408 (2014)
24. Abarbanel, H.D.I., Brown, R., Sidorowich, J.J., Tsimring, L.S.: The analysis of observed chaotic data in physical systems. Rev. Mod. Phys. **65**, 1331–1392 (1993)
25. Mora, T., Bialek, W.: Are biological systems poised at criticality? J. Stat. Phys. **144**, 268–302 (2011)
26. Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S.E., Berry II, M.J., Bialek, W.: Thermodynamics and signatures of criticality in a network of neurons. Proc. Natl. Acad. Sci. USA **112**, 11508–11513 (2015)
27. Marsili, M.: Dissecting financial markets: sectors and states. Quant. Financ. **2**, 297–302 (2002)
28. Lillo, F., Mantegna, R.N.: Variety and volatility in financial markets. Phys. Rev. E **62**, 6126–6134 (2000)
29. Bouchaud, J.P., Potters, M: Financial applications. In: Akemann, G., Baik, J., Di Francesco, P. (eds.) The Oxford Handbook of Random Matrix Theory. Oxford University Press, Oxford (2011). arXiv:0910.1205 [q–fin.ST] (2009)
30. Bun, J., Allez, R., Bouchaud, J.P., Potters, M.: Rotational invariant estimator for general noisy matrices. arXiv:1502.06736 [cond–mat.stat–mech] (2015)
31. Bun, J., Bouchaud, J.-P., Potters, M.: Cleaning large correlation matrices: tools from random matrix theory. arXiv:1610.08104 [cond–mat.stat–mech] (2016)
32. Aygün, E., Erzan, A.: Spectral renormalization group theory on networks. J. Phys. Conf. Ser. **319**, 012007 (2011)

33. Castellana, M.: Real-space renormalization group analysis of a non-mean-field spin-glass. EPL **95**, 47014 (2011)
34. Angelini, M.C., Parisi, G., Ricci-Tersenghi, F.: Ensemble renormalization group for disordered systems. Phys. Rev. B **87**, 134201 (2013)
35. Angelini, M.C., Biroli, G.: Spin glass in a field: a new zero-temperature fixed point in finite dimensions. Phys. Rev. Lett. **114**, 095701 (2015)
36. Brown, K.S., Hill, C.C., Calero, G.A., Myers, C.R., Lee, K.H., Sethna, J.P., Cerione, R.A.: The statistical mechanics of complex signaling networks: Nerve growth factor aignaling. Phys. Biol. **1**, 184–195 (2004)
37. Waterfall, J.J., Casey, F.P., Gutenkunst, R.N., Brown, K.S., Myers, C.R., Brouwer, P.W., Elser, V., Sethna, J.P.: Sloppy model universality class and the Vandermonde matrix. Phys. Rev. Lett. **97**, 150601 (2006)
38. Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., Sethna, J.P.: Universally sloppy parameter sensitivities in systems biology. PLoS Comput. Biol. **3**, e189 (2007)
39. Transtrum, M.K., Machta, B.B., Sethna, J.P.: Geometry of nonlinear least squares with applications to sloppy models and optimization. Phys. Rev. E **83**, 036701 (2011)
40. Matcha, B.B., Chachra, R., Transtrum, M.K., Sethna, J.P.: Parameter space compression underlies emergent theories and predictive models. Science **342**, 604–607 (2013)
41. Mehta, P., Schwab, D.J.: An exact mapping between the variational renormalization group and deep learning. arXiv:1410.3831 [stat.ML] (2014)