PCA meets RG

Serena Bradde[a] and William Bialek[a,b]
[a] *Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, 365 Fifth Ave., New York, New York 10016*
[b] *Joseph Henry Laboratories of Physics, and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton NJ 08544*
(Dated: March 3, 2016)

Come back and write an abstract at the end.

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 these systems more explicitly, describing the state of the 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 standard approach to the analysis of these high dimensional data is to look for a linear projection onto a lower dimensional space that captures most of the (interesting) variations. 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). 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 more continuous spectrum, so that any sharp division between important and unimportant dimensions would be arbitrary.

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; the evolution through the space of possible models is described by the renormalization group (RG). A central result of the RG 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 many macroscopic behaviors are insensitive to microscopic details (irrelevant operators).

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 simplified. Indeed, in a system where the many degrees of freedom live on a lattice, with translation invariant interactions, the eigenmodes of the covariance matrix (principal components) are Fourier modes, and typically we find that the variance of each mode decreases monotonically but smoothly with the wavelength of these modes. In the momentum space implementation of RG, we put a cutoff on the wavelength, and ask what happens to the joint distribution of the remaining variables as we move this cutoff to progressively longer wavelengths, averaging over the low variance 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 make this connection between PCA and RG more clear, and more general, so that we can construct RG approaches to the analysis of more complex high dimensional systems.

Familiar formulations of the RG make very explicit use of translation invariance (in the momentum space version) or the locality of interactions (in the real space version), but in more complex systems we cannot lean on these simplifications. We argue that the notion of relevant and irrelevant operators can be recast, at least perturbatively, in terms of the eigenvalue spectrum of a matrix that becomes the covariance matrix in the limit of nearly Gaussian distributions. This argument suggests conditions under which we can simplify our description of complex, high dimensional systems, well beyond the class of problems described by equilibrium statistical physics.

Let us imagine that the system we are studying is described by a set of variables $\phi_1, \phi_2, \cdots, \phi_N \equiv \{\phi_i\}$, 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

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

where the coefficients g_i allow us to describe weak kurtosis of each individual variable, and in the limit $g_i = 0$ the matrix K_{ij} is the inverse of the covariance matrix

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

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 kurtosis of each individual variable; in this sense it is a minimal model. We start in the Gaussian approximation and explore perturbations around this case.

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

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

with the usual normalization and orthogonality conditions,

$$\sum_i u_i(\mu)u_i(\mu') = \delta_{\mu\mu'} \tag{4}$$

$$\sum_\mu u_i(\mu)u_j(\mu) = \delta_{ij}, \tag{5}$$

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

$$\phi_i = \sum_\mu u_i(\mu)\tilde\phi_\mu; \tag{6}$$

if $g_i = 0$ then these modes are exactly the principal components. The Gaussian term in the probability distribution becomes

$$\frac{1}{2}\sum_{i,j}\phi_i K_{ij}\phi_j = \frac{1}{2}\sum_\mu \lambda_\mu\tilde\phi_\mu^2, \tag{7}$$

and hence the variance of each mode is given by $\langle\tilde\phi_\mu^2\rangle = 1/\lambda_\mu$. Importantly this means that the average variance of the individual variables is

$$\frac{1}{N}\sum_i\langle\phi_i^2\rangle = \frac{1}{N}\sum_\mu\frac{1}{\lambda_\mu} \to \int_0^\Lambda d\lambda\,\rho(\lambda)\frac{1}{\lambda}, \tag{8}$$

where in the last step we introduce the probability distribution of eigenvalues,

$$\rho(\lambda) = \frac{1}{N}\sum_\mu\delta(\lambda - \lambda_\mu), \tag{9}$$

which becomes smooth in the limit of infinite dimensionality, and we note explicitly that there is a largest eigenvalue Λ.

The essential idea is that we would like 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 (8). 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:

$$0 = \frac{d}{d\Lambda}\left[\frac{1}{N}\sum_i\langle(z_\Lambda\phi_i)^2\rangle\right] \tag{10}$$

$$= \frac{d}{d\Lambda}\left[\int_0^\Lambda d\lambda\,\rho(\lambda)\frac{(z_\Lambda)^2}{\lambda}\right] \tag{11}$$

$$\Rightarrow \frac{d\ln z_\Lambda}{d\ln\Lambda} = -\frac{1}{2}\rho(\Lambda)\left[\int_0^\Lambda d\lambda\,\rho(\lambda)\frac{1}{\lambda}\right]^{-1}. \tag{12}$$

Thus, as we change the scale of the smallest variance mode that we include in our analysis, we also have to change the scale of the variables $\{\phi_i\}$ themselves, and this scaling is determined by the probability distribution of eigenvalues of the (inverse) covariance matrix.

The rescaling of variables determines whether terms such as the quartic $\sim g_i$ in Eq (1) will become more or less important as we move the cutoff Λ. To begin, we want to write everything in terms of the rescaled variables. But, in addition, when we reduce the cutoff, we 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,

$$N_{\rm eff} = N\int_0^\Lambda d\lambda\,\rho(\lambda), \tag{13}$$

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

$$\sum_i g_i\phi_i^4 = N_{\rm eff}\sum_i\tilde g_i(z_\Lambda\phi_i)^4 \tag{14}$$

$$\tilde g_i = z_\Lambda^{-4}g_i/N_{\rm eff}. \tag{15}$$

Now the scaling of the coefficient $\tilde g_i$ is given by

$$\frac{d\ln\tilde g_i}{d\ln\Lambda} = \rho(\Lambda)\left[2\frac{1}{\int_0^\Lambda d\lambda\,\rho(\lambda)\frac{1}{\lambda}} - \frac{\Lambda}{\int_0^\Lambda d\lambda\,\rho(\lambda)}\right]. \tag{16}$$

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

If the scaling function $d\ln\tilde{g}_i/d\ln\Lambda$ is positive, then as we decrease the cutoff Λ and thus average over more and more of the low variance modes, any small quartic term $\tilde{g}_i$ will become still smaller, and hence the distribution approaches a Gaussian. This seems to make sense, since when we project onto a (much) lower dimensional space, each of the variables that remains is a weighted sum of many variables, and we might expect the central limit theorem to enforce approximate Gaussianity of the resulting distribution.

But if the scaling function $d\ln\tilde{g}_i/d\ln\Lambda < 0$, 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.

To see that our results for the scaling function and the definition of a relevant operator match that in the conventional RG analysis, let's think about a system in which the variables ϕ live at positions $\mathbf{x}$ in a D dimensional Euclidean space. Then the correlations come from a "kinetic energy" term that enforces similarity among neighbors,

$$\frac{1}{2}\sum_{i,j}\phi_i K_{ij}\phi_j \to \frac{1}{2}\int d^Dx\,[\nabla\phi(\mathbf{x})]^2\,. \tag{17}$$

The eigenvectors of K are then Fourier modes, as noted above, indexed by a wave vector $\mathbf{k}$, and the associated eigenvalue is $\lambda = |\mathbf{k}|^2$. If the original variables were on a lattice with linear spacing a, so that there are only a finite number of variables in total, then there is a maximum eigenvalue $\Lambda \sim (\pi/a)^2$. The probability density of eigenvalues is given by

$$\rho(\lambda) \propto \int d^Dk\,\delta\left(\lambda - |\mathbf{k}|^2\right) \propto \lambda^{D/2-1}. \tag{18}$$

Substituting into Eq (16), we find

$$\frac{d\ln\tilde{g}_i}{d\ln\Lambda} = \frac{D}{2} - 2 = \frac{1}{2}(D-4)\,. \tag{19}$$

The quartic term is relevant if $d\ln\tilde{g}_i/d\ln\Lambda < 0$, which corresponds to $D < 4$, as is well known from the conventional RG analysis; the extra factor of $1/2$ arises because Λ is a cutoff on the eigenvalue, which is the square of the wavevector. Thus the standard RG notions of relevant and irrelevant operators, which hinges on locality and translation invariance, can be recovered without using these ideas. The role of dimensionality in the RG analysis of local interactions is played, instead, by the eigenvalue spectrum of the matrix K_{ij}.

If we remember that the probability distribution $\rho(\lambda)$ is normalized, and choose units where the mean variance of the individual variables is equal to one, then Eq (16) simplifies to give $d\ln\tilde{g}_i/d\ln\Lambda \propto 2-\Lambda$. As a sanity check, we note that if $\rho(\lambda) = A\lambda^{D/2-1}$, then to enforce these two normalization conditions we must have $A = D/(2\Lambda^{D/2})$ and $\Lambda = D/(D-2)$, so that $2-\Lambda \propto D-4$. This normalization will be useful in what follows.

If each individual variable has unit variance, then the matrix C_{ij} is the matrix of correlation coefficients, and Λ is the inverse of the smallest eigenvalue of this matrix. The mean eigenvalue is one, so large Λ corresponds to a wide range of eigenvalues, which are generated by strong correlations. As an example, if the correlation coefficients are random numbers with standard deviation $\delta C = c/\sqrt{N}$, then for large N random matrix theory tells us that $\Lambda = 1/(1-2c)$. This suggests that, for weakly correlated systems, with $\delta C < 1/(4\sqrt{N})$, the Gaussian description can be self–consistent, in that small quartic corrections are irrelevant in the RG sense. On the other hand, for $\delta C > 1/(4\sqrt{N})$ these corrections are relevant, so that non–Gaussianity is always important for systems that reach this level of correlation.

Thus far our analysis has been confined to "power counting." The next step is to actually integrate out the the low variance degrees of freedom and compute corrections to the coupling constants that are beyond those generated from the spectrum of eigenvalues itself. Since we can think about discrete modes, 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, and we want to average over these fluctuations; in the limit of small g, ψ is Gaussian with $\langle\psi^2\rangle = 1/\Lambda$, and we find

$$\exp\left[-\frac{1}{4}N_{\rm eff}\sum_i \tilde{g}_i z_\Lambda^4\phi_i^4\right] \to \left\langle \exp\left[-\frac{1}{4}N_{\rm eff}\sum_i \tilde{g}_i z_\Lambda^4(\phi_i+u_i\psi)^4\right]\right\rangle \tag{20}$$

$$= \exp\left[-\frac{1}{2}N_{\rm eff}\sum_i \tilde{g}_i z_\Lambda^4\frac{u_i^2}{\Lambda}\phi_i^2 - \frac{1}{4}N_{\rm eff}\sum_i \tilde{g}_i z_\Lambda^4\phi_i^4 + \frac{1}{4}N_{\rm eff}^2\sum_{i,j}\tilde{g}_i\tilde{g}_j z_\Lambda^8\phi_i^2\phi_j^2\frac{u_i^2u_j^2}{\Lambda^2} + \cdots\right] \tag{21}$$

The first term is a correction to the matrix K, analogous to a mass renormalization. From the third term, we can absorb the $\mathrm{i} = \mathrm{j}$ contributions as a correction to $\tilde{g}_\mathrm{i}$,

$$\tilde{g}_\mathrm{i} \to \tilde{g}_\mathrm{i} - N_\mathrm{eff} \tilde{g}_\mathrm{i}^2 z_\Lambda^4 \frac{u_\mathrm{i}^4}{\Lambda^2}. \tag{22}$$

Notice that we have integrated out exactly one mode, which corresponds to a shift in the cutoff by an amount $d\Lambda$ such that $N_\mathrm{eff} \rho(\Lambda) d\Lambda = -1$, so we can write

$$d\tilde{g}_\mathrm{i} = \rho(\Lambda) z_\Lambda^4 \tilde{g}_\mathrm{i}^2 \left(N_\mathrm{eff} u_\mathrm{i}^2\right)^2 \frac{d\Lambda}{\Lambda^2}. \tag{23}$$

If we combine these perturbative effects with the dimensional effects from Eq (16), we have

$$\frac{d \ln \tilde{g}_\mathrm{i}}{d \ln \Lambda} = \rho(\Lambda) \left[2 - \Lambda + \tilde{g}_\mathrm{i} \left(N u_\mathrm{i}^2\right)^2 \frac{1}{\Lambda} \right], \tag{24}$$

where we use the normalization conditions above. We recall from Eq (4) that $\sum_\mathrm{i} u_\mathrm{i}^2 = 1$, so that $N u_\mathrm{i}^2$ is a number of order unity at large N. Since we are decreasing the cutoff Λ, this flow generate a stable fixed point at $\tilde{g}_\mathrm{i} = 0$ for $\Lambda < 2$, or at

$$\tilde{g}_\mathrm{i}^* = \frac{\Lambda(\Lambda - 2)}{\left(N u_\mathrm{i}^2\right)^2} \tag{25}$$

for $\Lambda > 2$.

Equation (25) for the fixed point of the flow may look a bit odd, since it has an explicit dependence on the cutoff, but remember that we have chosen units in which the average variance of the individual degrees of freedom is equal to one, so as we move the cutoff we have to rescale the eigenvalues themselves. As noted above, in the usual case of local interactions with $-\nabla^2$ in place of the matrix K_ij, this fixes $\Lambda = D/(D-2)$, and hence $\tilde{g}_\mathrm{i}^* \propto 4 - D$, as usual.

We thank [***]. 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.