Informational Rescaling of PCA Maps with Application to Genetic Distance

Nassim Nicholas Taleb*, Pierre Zalloua†§, Khaled Elbassioni†, Andreas Henschel† and Daniel E. Platt‡
* Tandon School, New York University (Corresponding author, nnt1@nyu.edu)† Khalifa University ‡ Harvard University § IBM

March 2022 (First version, Dec 2019)

arXiv:2303.12654v1 [cs.IT] 14 Mar 2023

***Abstract*—We discuss the inadequacy of covariances/correlations and other measures in L-2 as relative distance metrics. We propose a computationally simple heuristic to transform a map based on standard principal component analysis (PCA) (when the variables are asymptotically Gaussian) into an entropy-based map where distances are based on mutual information (MI). Rescaling PCA distances using MI allows a representation of relative correlations. This entropy rescaled PCA, while preserving order relationships, changes the relative distances to make them linear to information.**

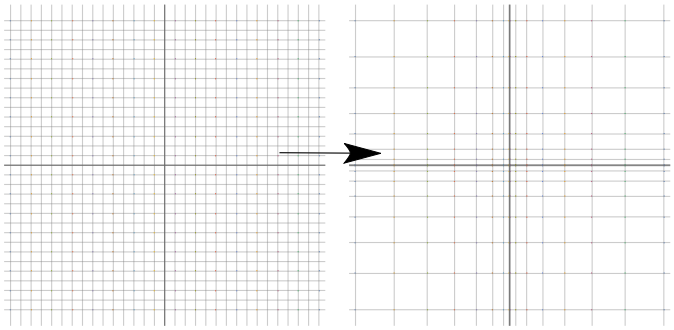

Fig. 1. Transformation of PCA maps to accommodate informational distances

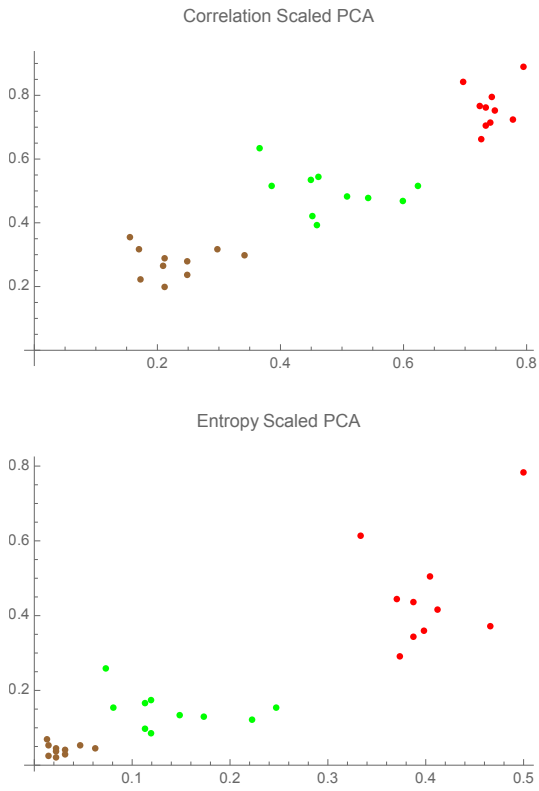

Fig. 2. Entropy rescaled principal component analysis changes the relative distances to make them linear to information. This is made possible thanks to the information-theoretic optimality of the PCAs under thin-tailed distributions.

I. Introduction: The problem of correlation

Correlation between two variables X and Y, even if we assumed that both variables are normally distributed (or in the class of rapid convergence to the normal, or "thin tailed" [1]), does not adequately reflect the information distance between them. This distortion becomes acute with Principal Component Analysis, PCA, and the genetic two-dimensional maps where there is a built-in correlation component.

For instance, if we are correlating 2 vectors X_1 and X_2 against Y (assuming it is the basis) the information does not scale linearly (even though correlation reflects a measure of the noise in a linear dependence). There must be some scaling of the correlation metric. A .5 correlation is vastly inferior to, say, .7.

A. Information and correlation

It has been shown that experts can be fooled by their own metrics under nonlinearity hence the need to "linearize" whatever metric is used. For cognitive limitations by experts are compounded by the nonlinearity of the measure; Soyer et al [2] showed how great many econometricians, while knowing their statistical equations down pat, don't get the real inferential and practical implications –all interpretation errors go in one direction, the *fooled by randomness* one (i.e. underestimation of noise). That 70 pct. of econometricians misinterpreted their own results is quite telling. The corresponding author has documented a version of the effect in [3]: professionals and graduate students failed to realize that they interpreted mean deviation as standard deviation, therefore underestimating volatility, especially under fat tails.

The only cure –visibly the level of statistical education doesn't help – is to avoid presenting nonlinear measures and linearizing whatever is presented to the specialist before the scientific implication.

Entropy methods being additive (unlike correlation) solve the problem.

Not all fall for correlation as a relatively uninformational metric. Machine learning loss functions rely on cross-entropy methods [4]. Since DNA is, well, information, an information-theoretic metric would be most certainly preferable to what is in current standard use.

Since mutual information maps to "how much can I gamble on X knowing Y", its information-theoretic quality is most applicable to genetic distance. Further, in addition to PCA analysis, entropy methods are helpful to properly scale runs of homozygosity (ROH) (that is, contiguous lengths of homozygous genotypes that are present in an individual due to parents transmitting identical haplotypes to their offspring).

B. Correlation and additivity

It has been shown in [5] that correlation is not additive across subsections of the domain under consideration –even when the variables are Gaussian.

C. This discussion

In the rest of this discussion we propose a new way to map PCs using mutual information. Conveniently, because PCA vectors for Gaussian variables are orthogonal both for correlation and mutual information, we can apply a simple heuristic for the translation.

II. PCA

We observe that conventional principal component analysis propose distances between groups and variables based on representation on maps built as follows.

Let $(X_1, \ldots, X_n)$ be the original vectors (in $\mathbb{R}^m$), and $(\pi_1, \ldots, \pi_n)$ the orthogonal principal components ordered by decreasing variance. Two–dimensional principal component representation typically maps X_i in Cartesian coordinates according to a metric μ such that the coordinates become

$$d_i = (\mu(X_i, \pi_j), \mu(X_i, \pi_{j'}))$$

where typically $j' = j + 1$. The same logic applies to three dimensions.

The function $\mu(.)$ in common use is expressed by the dot product $< X_i, \pi_j >$ scaled by $\frac{1}{m-1}$, or its decomposition via the scaled correlation

$$\mu(X_i, \pi_j) = \rho_{X_i,\pi_j} \sigma_{X_i} \sigma_{\pi_j} \tag{1}$$

and when the X are normalized,

$$\mu(X_i, \pi_j) = \rho_{X_i,\pi_j} \sqrt{\lambda_j} \tag{2}$$

where λ_j is the eigenvalue associated with the principal component π_j.

We will revisit with a matrix notation expressing the suggested transformations.

A. Mutual Information

We define $I_{X,Y}$ the mutual information between r.v.s X and Y.

$$I_{X,Y} = \int_{\mathcal{D}_X} \int_{\mathcal{D}_Y} f(x,y) \log \left(\frac{f(x,y)}{f(x)f(y)} \right) \mathrm{d}x \, \mathrm{d}y \tag{3}$$

and of course

$$\log \frac{f(x,y)}{f(x)f(y)} = \log \frac{f(x|y)}{f(x)} = \log \frac{f(y|x)}{f(y)}.$$

In effect mutual information is the Kullback-Leibler divergence between two distributions: the joint distribution $f(x, y)$ and the product $f(x)f(y)$ evaluated with respect to the joint distribution, [6].

We note some difficulties translating direct frequencies into continuous functions but in our case the problem is solved via the identity further down, allowing us to transfer from the pairwise correlation.[1]

Proposition 1

Under normality, the orthogonal principal components satisfy, for $i, j \leq m$

$$I_{\pi_i, \pi_{j \neq i}} = 0.$$

Proof. For bivariate normal distributions [7], [8] (though not all distributions in the elliptical class), uncorrelated means independence. Let Σ be the covariance matrix for $X, Y \sim \mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$ where M is a vector of means and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

Assume $M = (0, 0)$ to simplify. The PDFs are $f(x) = \frac{e^{-\frac{x^2}{2\sigma_1^2}}}{\sqrt{2\pi}\sigma_1}$; the joint PDF:

$$f(x,y) = \frac{\exp\left(-\frac{\sigma_2^2 x^2 - 2\rho\sigma_2\sigma_1 xy + \sigma_1^2 y^2}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\right)}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}}$$

The parametrization $\rho = 0$ implies the identity $f(x, y) = f(x)f(y)$, namely that lack of correlation implies independence, hence absence of mutual information between X and Y, that is, $I_{X,Y} = 0$.

□

We note that for other elliptical distributions, say the multivariate Student T or Cauchy, $\rho = 0$ does not mean independence [1]. For instance, for $X, Y \sim$ Multivariate Student T (α, ρ), the mutual information $I_\alpha(X, Y)$:

$$I_\alpha(X,Y) = -\frac{1}{2}\log\left(1-\rho^2\right) + \lambda_\alpha \tag{4}$$

where $\lambda_\alpha = -\frac{2}{\alpha} + \log(\alpha) + 2\pi(\alpha+1)\csc(\pi\alpha) + 2\log\left(B\left(\frac{\alpha}{2}, \frac{1}{2}\right)\right) - (\alpha+1)H_{-\frac{\alpha}{2}} + (\alpha+1)H_{-\frac{\alpha}{2}-\frac{1}{2}} - 1 - \log(2\pi)$,

[1] Common practice consists is smoothing the kernel distribution then computing the mutual information.

where $\csc(.)$ is the cosecant of the argument, $B(.,.)$ is the beta function and $H(.)^{(r)}$ is the harmonic number $H_n^r = \sum_{i=1}^{n} \frac{1}{i^r}$ with $H_n = H_n^{(1)}$. We note that $\lambda_\alpha \underset{\alpha\to\infty}{\to} 0$, the limit that corresponds to the Gaussian case.

This makes the proposed heuristic more straightforward than alternatives to PCA such as the t-distributed stochastic neighbor embedding (t-SNE) method [2].

We also note Linsker's result [9] showing that the conventional PCA provides an information-theoretic optimality so long as the noise is Gaussian.

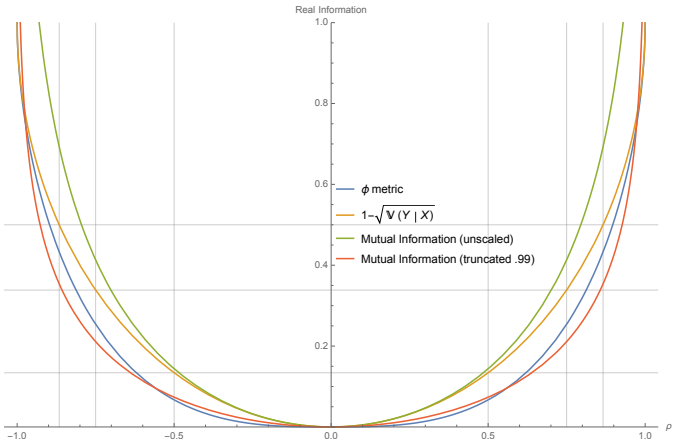

Fig. 3. Various rescaling methods, linerarizing information and putting correlation in perspective. The function ϕ is the "proportion of normalized similarity" for Y given X. The function $1 - \mathbb{V}(X|Y) = 1 - \mathbb{E}\left((X - \mathbb{E}(X|Y))^2 \middle| Y\right)$ represents the reduction in uncertainty of X knowing Y. While for large values there is no significant differences, these measures suffer from problems in additivity; only mutual information escapes such distortions.

Theorem 1

$I_{X,Y}$ is additive across partitions of $\mathcal{D}_X$ and $\mathcal{D}_Y$.

Proof. The result is immediate. We have:
$I_{X,Y} = \mathbb{E}(\log f(x,y)) - \mathbb{E}(\log f(x)) - \mathbb{E}(\log f(y))$. Consider the additivity of measures on subintervals. □

B. Re-scaling PCA distances using Mutual Information

We note that regardless of parametrization of X and Y, when the distributions are jointly Gaussian with $\rho_{X,Y}$, $I_{X,Y} = -(\rho_{X,Y})\frac{1}{2}\log\left(1-\rho^2\right)$.

$I_{X,Y}$ the mutual information between r.v.s X and Y and joint PDF $f(.,.)$, because of its additive properties, allows a representation of relative correlations, via the re-scaling function

$$r_{X,Y} = -\text{sgn}(\rho_{X,Y})\frac{1}{2}\log\left(1-\rho^2_{X,Y}\right); \qquad (5)$$

[2]We also note that the (standard) original stochastic neighbor embedding technique does not reflect information-theoretic distances; its aim is to reduce dimensionality.

see Figure 3 for details. Hence Eq. 2 can be modified for rescaling (marked as μ')

$$\mu(X_i,\pi_j)' = -\text{sgn}(\rho_{X_i,\pi_j})\frac{1}{2}\log\left(1-\rho^2_{X_i,\pi_j}\right)\sqrt{\lambda_i}, \qquad (6)$$

as shown in Figs. 1 and 2.

C. In Matrix Notation

Using matrix notation, we can express the problem in the following way. Centering and scaling in the correct order yields a matrix suitable for computing correlations. We start by defining a matrix $\mathbf{G} = (g_{ij})$ features indexed by $i \in \mathbb{Z}_m$ samples, and $j \in \mathbb{Z}_n$. [3].

Theorem 2

Define

$$\mu_i = \frac{1}{n-1}\sum_{j\in\mathbb{Z}_n} g_{ij}, \qquad (7)$$

$$\sigma_{ii'} = \frac{1}{n-1}\sum_{j\in\mathbb{Z}_n}(g_{ij}-\mu_i)(g_{i'j}-\mu_{i'}), \qquad (8)$$

$$\sigma_i^2 = \sigma_{ii} \qquad (9)$$

$$\mathbf{Z} = \left(\frac{g_{ij}-\mu_i}{\sigma_i}\right) \qquad (10)$$

$$\rho_{ii'} = \frac{\sigma_{ii'}}{\sigma_i\sigma_{i'}}. \qquad (11)$$

Then $\mathbf{ZZ^T} = (n-1)(\rho_{ii'})$.

Proof. $\mathbf{ZZ^T} = \left(\sum_{j\in\mathbb{Z}_n}\frac{(g_{ij}-\mu_i)(g_{i'j}-\mu_{i'})}{\sigma_i\sigma_{i'}}\right) = (n-1)\left(\frac{\sigma_{ii'}}{\sigma_i\sigma_{i'}}\right) = (n-1)(\rho_{ii'})$. □

Therefore the correlation matrix $\mathbf{C}$ may be represented [4] by [5]

$$\mathbf{C} = (\rho_{ii'}) = \frac{1}{n-1}\mathbf{ZZ^T} = \text{cov}(\mathbf{Z},\mathbf{Z^T}). \qquad (12)$$

Theorem 3

$\mathbf{C}$ is symmetric and positive definite.

Proof. Since, for any vector $\mathbf{w}$, the expression $\mathbf{w^T C w} = \frac{1}{n-1}(\mathbf{Z^T w})^T(\mathbf{Z^T w}) \geq 0$, it follows $\mathbf{C}$ is positive definite. Also, $\mathbf{C^T} = \frac{1}{n-1}(\mathbf{ZZ^T})^\mathbf{T} = \frac{1}{n-1}\mathbf{ZZ^T} = \mathbf{C}$, and so is symmetric. □

[3]These features could be biallelic diploid SNPs coded in $\mathbb{Z}_2$

[4]The transpose aims at ascertaining that in some software programs such as Mathematica, the eigenvectors are presented as the columns of the matrix.

[5]The centering by rows for genotypic analysis differs from Patterson[10], but connforms with Price et al.[11]; the "smartpca" app computes the appropriate correlations with "altnormstyle: NO".

The diagonalization of $\mathbf{C}$ provides a decomposition of the feature vectors into an orthogonal set that spans the subspace containing the samples.

Theorem 4

the $\mathbf{U^T Z}$ *rows are orthogonal, and the covariance diagonal.*

Proof. Given that $\mathbf{C}$ is positive definite and symmetric, $\mathbf{C}$ is diagonalized by an orthonormal matrix $\mathbf{U}$ of the normalized orthogonal eigenvectors to yield a diagonal matrix $\mathbf{D}$, so that $\mathbf{CU} = \mathbf{UD}$. $\mathbf{S}^2 = (n-1)\mathbf{D}$ is in common usage so that $(\mathbf{ZZ^T})\mathbf{U} = \mathbf{US^2}$. Therefore $\mathbf{D} = \mathbf{U^T C U} = \text{cov}((\mathbf{U^T Z}), (\mathbf{U^T Z})^\mathbf{T}) = \frac{1}{n-1}\mathbf{U^T Z Z^T U}$. Since $\mathbf{D}$ is diagonal, the $\mathbf{U^T Z}$ rows are orthogonal, and the covariance $\mathbf{D}$ in that basis is diagonal. □

We can identify the n columns, m rows, matrix of n feature-wise orthogonal principal components π_i as:

$$\mathbf{P} = \mathbf{U^T Z} \tag{13}$$

Note that, since the covariances of the $\mathbf{P}$ are $\text{cov}(\mathbf{P}, \mathbf{P^T}) = \mathbf{D}$ is diagonal, the rows are orthogonal, as noted previously. The matrix

$$\mathbf{V} = (n-1)^{-1/2}\mathbf{D^{-1/2}P} = \mathbf{S^{-1}U^T Z} \tag{14}$$

is normalized so that $\mathbf{VV^T} = \mathbf{I}$. $\mathbf{V}$ is half-orthornormal; the transposes are not: $\mathbf{V^T V} \neq \mathbf{I}$. The reason for this is that the number of individual vectors of SNPs for the individuals in $\mathbf{Z}$ does not span the space of SNP vectors since $m \ll n$. These are the familiar matrices in the singular value decomposition

$$\mathbf{Z} = \mathbf{USV^T}. \tag{15}$$

This decomposition also shows that the vectors in $\mathbf{V^T}$ represent an orthogonal basis in which $\mathbf{Z}$ can be represented, and so spans the subspace spanned by $\mathbf{Z}$.

Also $\text{cov}(\mathbf{S}, \mathbf{S^T}) = \mathbf{U^T}\text{cov}(\mathbf{Z}, \mathbf{Z^T})\mathbf{U}$ will be useful.

We define the correlation matrix

$$\mathbf{M} = \text{cor}(\mathbf{Z}, \mathbf{P^T}) \tag{16}$$

Then

Theorem 5

$$\mathbf{M} = \mathbf{U}. \tag{17}$$

Proof. $\mathbf{M} = [\text{cov}(\mathbf{Z}, \mathbf{Z^T})]^{-1/2}\text{cov}(\mathbf{Z}, \mathbf{P^T})[\text{cov}(\mathbf{P}, \mathbf{P^T})]^{-1/2}$. Noting that $\text{cov}(\mathbf{Z}, \mathbf{Z^T}) = \frac{1}{n-1}\mathbf{U^T S^2 U}$, $\text{cov}(\mathbf{P}, \mathbf{P^T}) = \frac{1}{n-1}\mathbf{S^2}$, and $\text{cov}(\mathbf{Z}, \mathbf{P^T}) = \frac{1}{n-1}\mathbf{ZZ^T U} = \frac{1}{n-1}\mathbf{US^2}$, then $\mathbf{M} = \mathbf{US^{-1}U^T US^2 S^{-1}} = \mathbf{U}$. □

This is therefore the standard principal component matrix that we expect, *and*, since this is a correlation, this may be re-scaled as a mutual information. The information re-scaled version $\mathbf{M}'$ becomes

$$\mathbf{M}' = \mathbf{R}(\mathbf{M}) = \mathbf{R}(\mathbf{U}),$$

et voilà.

III. Discussion

We showed how, under some conditions satisfied in population genetics, to efficiently and effectively convert a principal components based map to one representing information-based distance.

References

[1] N. N. Taleb, *The Statistical Consequences of Fat Tails*. STEM Academic Press, 2020.

[2] E. Soyer and R. M. Hogarth, "The illusion of predictability: How regression statistics mislead experts," *International Journal of Forecasting*, vol. 28, no. 3, pp. 695–711, 2012.

[3] D. Goldstein and N. Taleb, "We don't quite know what we are talking about when we talk about volatility," *Journal of Portfolio Management*, vol. 33, no. 4, 2007.

[4] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[5] N. N. Taleb, "Common misapplications and misinterpretations of correlation in social" science," *Preprint, Tandon School of Engineering, New York University*, 2020.

[6] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[7] I. M. Gel'fand and A. M. Yaglom, "Computation of the amount of information about a stochastic function contained in another such function," *Uspekhi Matematicheskikh Nauk*, vol. 12, no. 1, pp. 3–52, 1957.

[8] A. Gel'fand, I.M.; Yaglom, "Calculation of amount of information about a random function contained in another such function," in *Eleven Papers on Analysis, Probability and Topology*, American Mathematical Society, Dec. 1959, vol. 12, iSSN: 0065-9290, 2472-3193.

[9] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[10] N. Patterson, A. L. Price, and D. Reich, " Population Structure and Eigenanalysis," *PLOS Genetics*, vol. 2, no. 12, p. e190, Dec. 2006, publisher: Public Library of Science.

[11] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, " Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, Aug. 2006, number: 8