Dimensionality Reduction

Jonathan Rosenblatt
Ben Gurion University

December 30, 2016

Example 1 (BMI)**.** Consider the heights and weights of a sample of individuals. The data may seemingly reside in 2 dimensions but given the height, we have a pretty good guess of a person's weight, and vice versa. We can thus state that heights and weights are not really two dimensional, but roughly lay on a 1 dimensional subspace of $\mathbb{R}^2$.

Example 2 (IQ)**.** Consider the correctness of the answers to a questionnaire with p questions. The data may seemingly reside in a p dimensional space, but assuming there is such a thing as "skill", then given the correctness of a person's reply to a subset of questions, we have a good idea how he scores on the rest. Put differently, we don't really need a 200 question questionnaire–100 is more than enough. If skill is indeed a one dimensional quality, then the questionnaire data should organize around a single line in the p dimensional cube.

Example 3 (Blind signal separation)**.** Consider n microphones recording an individual. The digitized recording consists of p samples. Are the recordings really a shapeless cloud of n points in $\mathbb{R}^p$? Since they all record the same sound, one would expect them to arrange around a single

1 Enter the King: Principal Component Analysis

Principal Component Analysis (PCA) is such a basic technique, it has been rediscovered and renamed independently in many fields. It can be found under the names of *Discrete Karhunen–Loève Transform; Hotteling Transform; Proper Orthogonal Decomposition (POD); Eckart–Young Theorem; Schmidt–Mirsky Theorem; Empirical Orthogonal Functions; Empirical Eigenfunction Decomposition; Empirical Component Analysis; Quasi-Harmonic Modes; Spectral Decomposition; Empirical Modal Analysis*, and possibly more[1]. The many names are quite interesting as they offer an insight into the different problems that led to PCA's (re)discovery.

Return to the BMI problem in Exampl 1. Assume you now wish to give each individual a "size score", that is a **linear** combination of height and weight: PCA does just that. It returns the linear combination that has the largest variability, i.e., the combination which best distinguishes between individuals.

The variance maximizing motivation above was the one that guided Hotelling Hotelling [1933]. But 30 years before him, Pearson [1901] derived the same procedure with a different motivation in mind. Pearson was also trying to give each individual a score. He did not care about variance maximization, however. He simply wanted a small set of coordinates in some (linear) space that approximates the original data well. As it turns out, the best linear-space approximation of X is also the variance maximizing one. More precisely: the *sequence* of $1, \ldots, p$ dimensional linear spaces that best approximate X, is exactly the sequence of $1, \ldots, p$ dimensional scores, that best separate between the n samples. Pearson and Hotelling (among others) thus arrived to the exact same solution, with different motivations.

[1] `http://en.wikipedia.org/wiki/Principal_component_analysis`

1.0.1 Bi Plot

The *Bi-Plot* shows the two first scores of the original data points. These scores are known as the *Principal Componets* (PCs). The contribution of each original variable to each PC, is called the *Loadings*. The plot also shows the contribution of each of the original variables to each of the scores. See example in Figure 1.

Principal Components

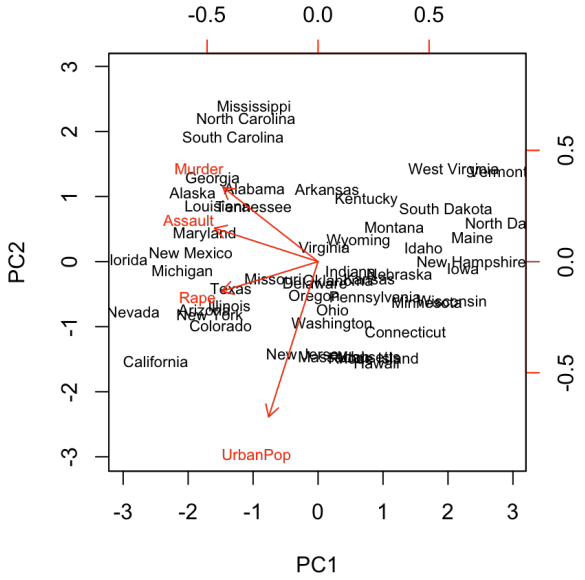

Figure 1: **BiPlot**. Arrest type data for USA states. Data includes urban population size, number of rape related arrests, assault related, and murder related ($p = 4$). Each city is presented against its two first PCs. Arrows encode the loadings. They show that PC1 encodes a general crime level, as it is the average of all type of crimes. PC2 measures the level of urbanization, as it is dominated by the UrbanPopulation variable.
Source: `https://goo.gl/85qtKv`

1.0.2 Scree Plot

[TODO]

1.1 Mathematics of PCA

We now present the derivation of PCA from the two different motivations.

1.1.1 Variance Maximizing View of PCA

Proof. The sketch of the proof is the following: We will first show that the weight vector that maximizes the variance is the eigenvector that corresponds to the first principal component. We will do so for the *population* covariance, Σ, and wrap up by plugging its empirical counterpart, $X'X$ (assuming a centered X).

Starting with the first principal component. For a random p-vector, $\mathbf{x}$ denote $\Sigma := \mathbf{Cov}[\mathbf{x}]$, so that for a fixed p-vector v: $\mathbf{Cov}[v'\mathbf{x}] = v'\Sigma v$. Finding a linear combination of $\mathbf{x}$ that best separates individuals, means maximizing $\mathbf{Cov}[v'x]$ w.r.t. to v. Clearly, $\mathbf{Cov}[v'x]$ may explode if any v is allowed. It is most convenient, mathematically, to constrain the l_2 norm: $\|v\|_2^2 = 1$. Maximizing under a constraint, using Lagrange-Multipliers:

$$argmax_v\{v'\Sigma v - \lambda(\|v\|_2^2 - 1)\}. \tag{1}$$

Differentiating w.r.t v and equating zero:

$$(\Sigma - \lambda I)v = 0 \tag{2}$$

We thus see that any of the p eigenvalue-eigenvector pairs of Σ is a local extremum. Which of them to pick? To find a *global* maximum we return to the original problem, as plug our result:

$$argmax_{v:\|v\|_2^2=1}\{v'\Sigma v\} = argmax_\lambda\{v'\lambda v\} \tag{3}$$

so that the global maximum is obtained with the largest eigen-value λ. Put differently, the weight vector that returns the score that best separates individuals, is the eigenvector of Σ with the largest eigenvalue.

The second principal component can be found by solving the same problem, with the additional constraint of v_2 orthogonal to v_1.

The last missing ingredient is that instead of the true covariance between the features, Σ, we use the (centered) empirical covariance $X'X$. □

Remark 1. Readers familiar with matrix norms will recognize that the above is exactly the derivation of the spectral norm of Σ.

1.1.2 Linear-Space approximation view

In here, we try to find a series of $\mathcal{M}_q; q = 1, \ldots, p$, such that $\mathcal{M}_q$ is a *linear* subspace of dimension q which well approximates X in some (matrix) norm. For the deatils, see for instance Shalev-Shwartz and Ben-David [2014].

1.1.3 Why did Hotelling and Pearson arrive to the same solution?

We have currently offered two motivations for PCA: (i) Find linear combinations $v_1, \ldots, v_p$ that best distinguish between observations, i.e., maximize variance. (ii) Find the linear subspaces $\mathcal{M}_1, \ldots, \mathcal{M}_p$ that best approximates the data. The reason these two problems are equivalent, is due to the use of the squares-error/Euclidean norms.

Informally speaking, the data has some total variance. In analogy to the $SST = SSR + SSE$ decomposition in linear regression, the total variance of X can be decomposed into the part in $\mathcal{M}_q$, and the part orthogonal. The orthogonal part is the distance of X from $\mathcal{M}_q$. Maximizing the variance in $\mathcal{M}_q$ is thus the same as minimizing the distance from X to $\mathcal{M}_q$.

The only unresolved matter- is why the solution to the variance maximization problem is a *linear* subspace? This is simply because all the scores, are linear combinations of columns of X, thus span a linear subspace, as is sought in the linear-subspace approximation view.

1.2 How many PCs can you recover?

On the face of it, with p variables you can find p PCs. Things are not that simple however.

In the population version of the problem, i.e., when Σ is known, there may be as many non zero eigenvalues as the rank of Σ. Stating that Σ is full rank, is stating that the variables of $\mathbf{x}$ are not fully correlated.

In the empirical version of the problem, i.e., when $X'X$ is known, there may be as many non zero eigenvalues as the rank of $X'X$. Clearly, if $p > n$, variables of X have to be linearly dependent, so that $X'X$ cannot possibly be of full rank. To say that the kernel of X is of rank $p - n > 0$, is to say that there are $p - n$ scores that are identically zero, thus have no variance.

Problems do not end when $p < n$. This is because if $p < n$ but $p \sim n$ then we do not have many observations per estimated parameter. In the statistical literature, this is known as a *high dimensional* problem. In the engineering parlance, we say we have low *signal to noise*. For a rigorous treatment of the statistical properties of PCA, see Nadler [2008].

1.3 PCA as a Graph Method

It turns out that we may find the sequence of best approximating linear subspaces, i.e., the PCs, without the actual measurements X, but only with a dissimilarity graph. In particular, with the $n \times n$ graph $\mathfrak{D}$ of Euclidean distances between individuals: $\mathfrak{D}_{i,j} := \|x_i - x_j\|_2$.

It should come of no surprise that we done need the actual measurements, X, since the optimal loadings, v, only depend on the covariance Σ, or its empirical counterpart, $X'X$. It may, however, be quite surprising that given the distances between individuals, we may not recover the covariance between variables, $X'X$, but we can recover the PCs. Put differently, to find the low dimensional $\mathcal{M}$ that approximates the data, we don't need the whole data, but rather, only the graph of distances between data points.

For proof of the above statement, we refer the reader to [Friedman et al., 2001, Sec.18.5.2]

This observation will later be very useful for other dimensionality reduction algorithms, which operate not on the original data points, but rather, on dissimilarity graphs.

2 Preliminaries

2.1 Terminology

Variable A.k.a. *dimension*, or *feature* in the machine learning literature, or *column* for reasons that will be obvious in the next item.

Data A.k.a. *sample*, *observations*, depending on your community. Will typically consist of n, p dimensional vectors, i.e., with p variables in each. We typically denote the data as a $n \times p$ matrix X.

Manifold A space which is regular enough so that it is *locally* has all the properties of a linear space. We will denote an arbitrary manifold by $\mathcal{M}$.

Embedding Informally speaking: a "shape preserving" mapping of a space into another.

Linear Embedding An embedding done via a linear operation (thus representable by a matrix).

Generative Model Known to statisticians as the *sampling distribution*. The assumed stochastic process that generated the observed data.

2.2 Motivations

Scoring Give each observation a score (Hotelling motivation).

Latent structure Recover unobservables from indirect measurements. E.g: Blind signal reconstruction, CT scan, cryo-electron microscopy, etc.

SNR Denoise measurements before further processing like clustering, supervised learning, etc.

Compression Save on RAM ,CPU, and communication when operating on a lower dimensional representation of the data.

2.3 Taxonomy

Generative vs. algorithmic Refers to the motivation of the approach. Is it stated as an algorithm, or stated via some generative probabilistic model. PCA is purely algorithmic.

Linear $\mathcal{M}$ vs. non-linear $\mathcal{M}$. Is the target manifold linear or not? In PCA, $\mathcal{M}$ is linear.

Linear embedding vs. non-linear embedding . Is the embedding into $\mathcal{M}$ a linear operation? In PCA, the embedding is linear, and indeed, represented by a matrix.

Learning an embedding vs. an embedding function? Will we need to apply the reduction to new data? If yes, we need to learn an *embedding function*. If no, and we merely want to low dimensional representation of existing data, we only need to learn an embedding.

3 Latent Variable Generative Approaches

All generative approaches to dimensionality reduction will specify some unobserved set of variables, which we can observe indirectly up to some measurement noise. The unobservable variables will typically have a lower dimension than the observables, thus, dimension is reduced. We start with the simplest case of linear Factor Analysis.

3.1 Factor Analysis (FA)

To fix ideas, we start by revisiting the IQ problem in Example 2:

Example 4 (g-factor[2])**.** Assume n respondents answer p quantitative questions: $x_i \in \mathbb{R}^p, i = 1, \ldots, n$. Also assume, their responses are some linear function $A \in \mathbb{R}^p$ of a single personality attribute, s_i. We can think of s_i as the subject's "intelligence". We thus have

$$x_i = As_i + \varepsilon_i \tag{4}$$

And in matrix notation:

$$X = A\mathbf{s} + \varepsilon \tag{5}$$

The problem is to recover the unobservable intelligence scores, $s_1, \ldots, s_n$, from the observed answers X.

Assuming a generative distribution on $\mathbf{s}$ and ε, we may try to estimate $A\mathbf{s}$ by assuming some distribution on $\mathbf{s}$ and ε and apply maximum likelihood. Under standard assumptions on the distribution of $\mathbf{s}$ and ε, recovering $\mathbf{s}$ from $\widehat{A\mathbf{s}}$ is still impossible as there are infinitely many such solutions. To see this, consider an orthogonal *rotation* matrix R ($R'R = I$). For each such R: $A\mathbf{s} = AR'R\mathbf{s} = A^*\mathbf{s}^*$. While mathematically equivalent, A and A^* may have very different interpretations. This is why many researchers find FA an unsatisfactory inference tool.

Remark 2 (Identifiability in PCA)**.** The non-uniqueness (non-identifiability) of the FA solution under variable rotation is never mentioned in the PCA context. Why is this? This is because the methods solve different problems. The reason the solution to PCA is well defined is that PCA does not seek a single $\mathbf{s}$ but rather a *sequence* of $\mathbf{s}$ with dimensions growing from 1 to n.

FA Terminology The FA terminology is slightly different than PCA:

- **Factors**: The unobserved attributes $\mathbf{s}$. Not to be confused with the *principal components* in the context of PCA.
- **Loadings**: The A matrix; the contribution of each attribute to the observed X.

[2] `https://en.wikipedia.org/wiki/G_factor_(psychometrics)`

- **Rotation**: An arbitrary orthogonal re-combination of the latent attributes $\mathbf{s}$ and loadings, which changes the interpretation of the result.

The FA literature does offer several heuristics to "fix" the solution of the FA. These are known as *rotations*:

- **Varimax**: By far the most popular rotation. Attempts to construct factors that are similar to the original variables, thus facilitating interpretation. This can be seen as a "soft" approach to sPCA (Sec.5.1).
- **Quartimax**: Seeks a minimal number of factors to explain each variable. May thus result factors that are uninterpretable, since they all rely on the same variables.
- **Equimax**: A compromise between Varimax and Quartimax.
- **Oblimin**: Relaxes the requirement of the factors to be uncorrelated, so that they may be similar to the original variables; even more so than in varimax. This facilitates the interpretability of the factors.
- **Promax**: A computationally efficient approximation of oblimin.

3.1.1 Bibliographic Notes

For a brief review of Factor Analysis see Friedman et al. [2001]. For an full exposition, and a discussion of the differences with PCA, see Jolliffe [2002].

3.2 Non Linear Factor Analysis

Classical FA deals with features, X, that are linear in the latent factors, $\mathbf{s}$. Like any other generative model approach, it can be easily extended to deal with non-linear functions of the latent factors: $X = g(S)$, provided that $g(.)$ is one-to-one.

3.3 Independent Component Analysis (ICA)

ICA is a family of latent space models, thus, a *meta-method*. It assumes data is generated as some function of the latent variables $\mathbf{s}$. In many cases this function is assumed to be linear in $\mathbf{s}$ so that ICA is compared, if not confused, with PCA and even more so with FA. In its most popular form, X is assume to be a *linear* function of the latent independent components: $X = A\mathbf{s}$.

The fundamental idea of ICA is that $\mathbf{s}$ has a joint distribution of *non-Gaussian independent* variables. This independence assumption, solves the the non-uniquness of $\mathbf{s}$ in FA.

Being a generative model, estimation of $\mathbf{s}$ can then be done using maximum likelihood, or other estimation principles. A popular information theoretic estimation principle, replacing the maximum-likelihood principle, is known as *infomax*.

ICA is a popular technique in signal processing, where $\mathbf{s}$ is actually the signal, such as sound in Example 3. Recovering $\mathbf{s}$ is thus recovering the original signals mixing in the recorded X.

Remark 3 (ICA and FA)**.** The solutions to the (linear) ICA problem can ultimately be seen as a solution to the FA problem with a particular rotation R implied by the probabilistic assumptions on $\mathbf{s}$. Put differently, the formulation of the (linear) ICA problem, implies a unique rotation, which can be thought of as the rotation that returns components that are as far from Gaussian as possible.

Mathematics of ICA For ease of presentation we present a simple setup, which can be considerably generalized. In this setup, we will first analyze the population problem, i.e., in terms of random variables. We thus replace the data X, with the random vector $\mathbf{x}$, and afterwards consider implementation for finite samples.

- $\mathbf{x} = A\mathbf{s}$, implying that $\mathbf{x}$ is *linear* in the latent components, and the latent space is of dimension $q = p$. It follows that $s = A'X$.
- $\mathbf{x}$ has been pre-whitented, so that $\mathbf{Cov}[\mathbf{x}] = I$.
- Distance between distributions are measured using the Kullback-Leibler divergence (KL): $D_{KL}(\mathbf{x}\|\mathbf{s})$.

The optimization problem in this simple ICA is to find an orthogonal matrix A, for which: (i) the components of $A'\mathbf{x}$ are independent; (ii) $A'\mathbf{x}$ is a good approximation of $\mathbf{x}$. Formally:

$$argmin_{A \text{ orthogonal };A'\mathbf{x} \text{ independent}}\{D_{KL}(A'\mathbf{x}\|\mathbf{x})\}. \tag{6}$$

By enforcing the independence constraint in Eq.(6), and due to the properties of the KL divergence, Eq.(6) is equivalent to

$$argmin_{A \text{ orthogonal}}\{\sum_{j=1}^{q} H(A_j\mathbf{x}) - H(\mathbf{x})\} \tag{7}$$

where $H(\mathbf{x})$ denotes the Entropy of the random variable $\mathbf{x}$ (Definition **??**). Now, $H(\mathbf{x})$ is obviously fixed, so we need to minimize $H(A_j\mathbf{x})$. A classical result in information theory, is that the Gaussian distribution has the maximal entropy. Minimizing $H(A_j\mathbf{x})$ can thus be interpreted as finding a matrix A such that its columns return random variables, $A_j\mathbf{x}$, that are as *non-Gaussian* as possible.

This is where the population analysis ends. The insight we take from it, is that finding independent components, is actually finding non-Gaussian combinations of $\mathbf{x}$. The different implementations of ICA, indeed look for a matrix A which returns the most non-Gaussian combinations of the observed X.

3.3.1 Bibliographic notes

For a general discussion of ICA see Jolliffe [2002]. For a brief exposition of the linear ICA see Friedman et al. [2001]. For a detailed review of ICA see Hyvärinen and Oja [2000].

4 Purely Algorithmic Approaches

5 Dealing with the high-dimension

5.1 Sparse Principal Component Analysis

References

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

B. Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, pages 2791–2817, 2008.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.