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 General Terminology

Data A.k.a. *sample*, *observations*, depending on your community. Will typically consist of n, p dimensional vectors. We typically denote the data as an $n \times p$ 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 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 $1,\dots,p$ dimensional linear spaces that best approximate X, is exactly the sequence of $1,\dots,p$ dimensional scores, that best separate between the n samples. Pearson and Hotelling (among others) thus arrived to the exact same solution, with different motivations.

2.0.1 Scree Plot

2.0.2 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

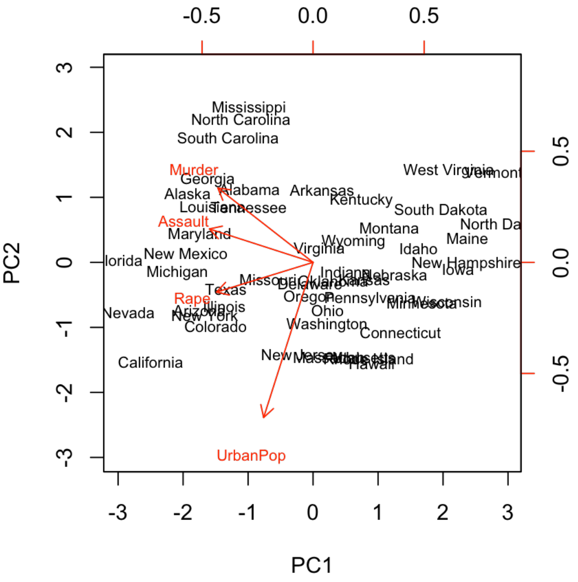

Figure 1: **BiPlot**. Arrest type data for USA states. Data includes urban population size, number of rape related arrests, assault related, and murder related ($p = 4$). Each city is presented against its two first PCs. Arrows encode the loadings. They show that PC1 encodes a general crime level, as 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] `http://en.wikipedia.org/wiki/Principal_component_analysis`

2.0.3 Mathematics of PCA

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

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 $\Sigma := \mathbf{Cov}\left[\mathbf{x}\right]$, so that for a fixed p-vector v: $\mathbf{Cov}\left[v'\mathbf{x}\right] = v'\Sigma v$. Finding a linear combination of $\mathbf{x}$ that best separates individuals, means maximizing $\mathbf{Cov}\left[v'x\right]$ w.r.t. to v. Clearly, $\mathbf{Cov}\left[v'x\right]$ 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:

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

So the P solutions for v are the eigen-vectors of Σ. 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.

Readers familiar with matrix norms will recognize that this is simply the derivation of the operator norm of Σ.

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 (scaled) empirical covariance $X'X$. □

The Linear-Space Embedding View We now seek to find a sequence of p approximations to X that lay in $1, \dots, p$ dimensional linear subspaces, with respect to a least squares loss. For simplicity of exposition, we will assume that X has been mean centred. The q'th problem to solve is thus

$$argmin_{f_q}\{\|X - f_q(X)\|_{Frob}\}. \tag{4}$$

Since f_rank is a map from $\mathbb{R}^p$ to some rank-q linear subspace, it must have the form $f_q(X) = H_qX$ where H_q is a $n \times n$ matrix of rank q. Since Eq.(4) minimizes sums of (squared) Euclidean distances, H_q has to be an orthogonal projection, thus symmetric. As such it can decomposed into an outer product $H_q = V_qV_q'$ where V_q is full rank $n \times q$ matrix [**?**, Eq.(5.13.4)]. Under the q-space constraint, and squared error, Eq.(4) collapses to

$$argmin_{V_q}\{\|X - V_qV_q'(X)\|_{Frob}\}. \tag{5}$$

Using some algebraic identities [**?**, Eq.(23.3)] Eq.(5) is equivalent to

$$argmax_{V_q}\{\mathrm{Tr}(V_q'XX'V_q)\}. \tag{6}$$

At this point we should note that the linear-space embedding problem has collapsed to the variance maximization problem! If you do not see this, just set $q = 1$ and compare to Eq.(3), recalling that $X'X$ estimates the features' covariance Σ.

2.0.4 Intuition

Notice we have currently offered two motivations for PCA: (i) Find linear combinations that best distinguish between observations, i.e., maximize variance. (ii) Find the linear subspace the bets approximates the data. The reason these two problems are equivalent, is due to the use of the squares error. Informally speaking, the data has some total variance. This variance can be decomposed into the part captured in $\mathcal{M}$, and the part not captured[2]. Since the variance in the data consists of sums of squares, minimizing the distance from X to $\mathcal{M}$, is the same as maximizing the variance of $X \hookrightarrow \mathcal{M}$, since their sum is fixed.

2.0.5 PCA as a Graph Method

Starting from the maximal variance motivation, it is perhaps not surprising that PCA depends only on the similarities between features, as measured by their empirical covariance. The linearity of the target manifold was there by assumption.

Following the linear-space embedding motivation, it is was surprising that the solutions depend only on the empirical covariances. This fact can be attributed to the use of squared error loss, which implied we were trying to decompose the total variance into the part in $\mathcal{M}_q$ and the orthogonal part.

From both motivations we see that the values of X are of no importance given $X'X$, which can be informally thought of as a sufficient statistic[3].

In-turn, $X'X$ depends only on the empirical covariances between *individuals* ($\mathfrak{S} = XX'$), or on the Euclidean distances between individuals ($\mathfrak{D} = (\|x_i - x_j\|)$).

The building blocks of all these graph-based dimensionality reduction methods are:

1. Compute some similarity graph $\mathfrak{S}$ (or dissimilarity graph $\mathfrak{D}$) from the raw features.

2. Call upon graph embedding theory to map the data points into the target manifold $\mathcal{M}$.

The fact that the linear-space embedding of the data depends only some similarity graph has laid a bridge between feature embedding, such as PCA, and *graph embedding* methods such as MDS (§**??**). Moreover, it has opened the door for replacing the covariance similarity, with many other similarity measures. Classic MDS (§**??**) is simply PCA when starting from $\mathfrak{S}$, thus viewed as a graph embedding problem. kPCA (§**??**) plugs kernel similarities (§**??**) instead of covariance similarities. Isomap (§**??**), LocalMDS (§**??**), and LLE (§**??**) follow a similar motivation using *local* measures of similarity. Spectral Clustering (§**??**) does some linear-space embedding à-la PCA, then wrapping up with a clustering algorithm in $\mathcal{M}$ à-la K-means.

We now prove that the PCA solution can be cast in terms of the covariance between individuals ($\mathfrak{S} = XX'$) or the Euclidean distances ($\mathfrak{D} = \|x_i - x_j\|$). In particular, we show that all the information on the location (mean) of X, needed for the PCA reconstruction, is actually encoded in $\mathfrak{S}$ (or $\mathfrak{D}$).

The following exposition takes from [**?**, Section 18.5.2]

PCA with the Covariance Similarity Graph To begin, we need to cast the solution to the PCA problem in Eq.(6) using the Singular Value Decomposition (SVD). SVD

Definition 1 (SVD)**.** Any $n \times p$ matrix X, can be decomposed into $X = UDV'$ where U is an $n \times p$ orthogonal matrix ($U'U = I_p$); D is a $p \times p$ diagonal matrix with diagonal elements $d_1 \geq d_2 \geq \cdots \geq d_p$; V is a $p \times p$ orthogonal matrix ($V'V = I_p$).

For mean centered X, the series of embeddings $f_q(X)$ for $q = 1, \dots,$ resulting from Eq.(6) is

[2] Analogous to $SST = SSR + SSE$ in linear regression.

[3] It is not a proper sufficient statistic as no generative model has been assumed.

given by $f_q(X) = U_q D_q$, where U_q D_q are the q leading columns of U and D respectively. UD is thus the sequence of all solutions.

Now denoting $\mathfrak{S} = XX'$ and calling SVD: $\mathfrak{S} = UD^2U'$. We thus see that by decomposing $\mathfrak{S}$ we can recover U, D, and thus $f_q(X)$.

If X is not mean centred, the relation still holds, but we skip the presentation.

PCA with the Euclidean Distance Dissimilarity Graph Can we convert Euclidean distances to empirical covariances? Yes!

Denote the matrix of distances of a non-centred X: $\mathfrak{D}^2 = (\|x_i - x_j\|^2)$.

$$\begin{aligned}
\mathfrak{D}^2_{i,j} =& \|x_i - x_j\|^2 && (7)\\
=& \|x_i - \bar{x}\|_2 + \|x_j - \bar{x}\|_2 - 2\left\langle x_i - \bar{x}, x_j - \bar{x}\right\rangle && (8)\\
=& \|x_i - \bar{x}\|_2 + \|x_j - \bar{x}\|_2 - 2\mathfrak{S}_{i,j} && (9)
\end{aligned}$$

where $\mathfrak{S}_{i,j}$ is the empirical covariance between individual i and j. We thus have

$$\mathfrak{S} = -(I - M)\frac{\mathfrak{D}^2}{2}(I - M) \qquad (10)$$

where M is the centring matrix: $M := \frac{1}{n}\mathbf{1}\mathbf{1}'$, and $\mathbf{1}$ an n vector of 1's.

3 FA

4 ICA

References

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

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.