Principal Component Analysis

compiled by Alvin Wan from Professor Benjamin Recht's lecture, Professor Jonathan Shewchuck's notes, and Elements of Statistical Learning

1 Overview

The outline for our algorithm is the following: We take in an input X, which is $d \times n$ in dimension r.

1. Take $X_c = [x_1 - \mu_x, x_2 - \mu_x, \dots x_n - \mu_x]$, $\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$.
2. Compute SVD of $X_c = USV^T$
3. Return $\hat{X} = S_r V_r^T, U_r, \mu_x$

We will explore 3 different views of PCA, showing how each approach gives us the same result: the maximum eigenvalue and its corresponding eigenvector. In the last section, we will then explore an application of PCA to Latent Semantic Indexing (or Latent Factor Analysis).

View 1 : Maximizing Variance

The first view is finding maximum variance. How do we find the direction of maximum variance? Our goal is to find a $u \in \mathbb{R}^n$ such that the sample variance of $\{u^T x_1, u^T x_2 \dots u^T x_n\}$ is maximized. Assuming that our data is de-meaned, our objective function is the following

$$\begin{aligned}
&\textsc{Maximize}_{u:\|u\|_2=1}\textsc{var}(u^T x_i) \\
&= \textsc{Maximize}_{u:\|u\|_2=1}\frac{1}{n}\sum_{i=1}^{n}(u^T x_i)^2 \\
&= \textsc{Maximize}_{u:\|u\|_2=1}\frac{1}{n}\sum_{i=1}^{n} u^T x_i x_i^T u \\
&= \textsc{Maximize}_{u:\|u\|_2=1} u^T(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T)u \\
&= \textsc{Maximize}_{u:\|u\|_2=1}\frac{1}{n} u^T X X^T u
\end{aligned}$$

Why is u normalized? When u is normalized, $u^T x_i$ corresponds to the projection of x onto u with direction u. Why is X de-meaned? Only then is XX^T the covariance matrix Σ. Recall that the entries along the diagonal of Σ take the form $\text{cov}(x_i, x_i) = \text{var}(x_i)$. Additionally, if we do not de-mean, the first eigenvector will point to the mean.

Take $S = XX^T$. Since S is positive semi-definite, there exists an eigenvalue decomposition $S = PDP^T$, where P is orthonormal and D is a diagonal matrix with eigenvalues listed in descending order. Since P is orthonormal, Pu for any $u \in \mathbb{R}^d$ with conforming shape can be replaced with a uniquely determined $w \in \mathbb{R}^d$.

$$u^T XX^T u = u^T Su = u^T (PDP^T)u = w^T Dw = \sum_{i=1}^{d} \lambda_i w_i^2$$

Note that the constraint $\|u\|_2 = 1$ translates into $\|w\|_2 = 1$ since

$$\|w\|_2 = \|P^T u\|_2 = u^T PP^T u = u^T u = \|u\|_2 = 1$$

Since λ_is are listed in descending order, and $\|w\|_2 = 1$, then $\sum_{i=1}^{d} \lambda_i w_i^2$ is maximized when $w_1 = 1$ and all other $w_i = 0$. As a result $w^* = e_1$ and

$$u^* = Pw^* = Pe_1 = u_1$$

where u_1 is the first eigenvector. Our direction of maximum variance is u_1. How do we find subsequent principal components? Take $\hat{X}_i = X_i - (u_1^T X_i)u_1$.

$$u_1^T \hat{X}_i = u_1^T X_i - (u_1^T X_i)u_1^T u_1 = 0$$

Note that $\hat{X}$ is centered and orthogonal to u_1. To find the direction of maximum variance, we need the maximum singular value $\sigma_1(\hat{X})$. Note that this is the second principal component for our original X, $\sigma_2(X)$, in the direction u_2.

View 2 : Maximizing Likelihood

The second view is fitting a Gaussian model to our data. We will assume the following.

$$x_i = \alpha_i z_i + \omega_i$$

In the above, z_i, α_i are unknown but not random. However, $\omega_i \sim N(0, \sigma^2 I_d)$, $\alpha \in \mathbb{R}^d, z \in \mathbb{R}^d$. Our goal is to find z_i, and our objective is, formally, the following.

$$\begin{aligned}
&\textsc{Maximize}_{\alpha,z} \log p(X; \alpha, z) \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \|x_i - \alpha_i z_i\|^2 + \frac{d}{s} \log 2\pi\sigma^2 \\
&= \sigma_1 u_1 v_1^T
\end{aligned}$$

Note that this is the equivalent of the following, where we chain all of the x_i together. We wish to find the best rank-one matrix that models our data.

$$\textsc{Minimize}_{\alpha,z} \|X - z\alpha^T\|_F^2 = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

Does this uniquely define a z and an α? It doesn't, because we can scale α by k and z by k^{-1}. Even if z was constrained to have norm 1, z, α are not unique.

View 3 : Minimizing Projection Error

This is reminiscent of regression: when we run linear regression, the error runs along the y-axis. For PCA, the error is the minimum distance from the point to the line, making the error perpendicular to our line. In other words, if our decision boundary is defined by

$$\{w^T z = 0\} = L(w)$$

We have that distance is defined is to be the following. Note that the projection of x_i onto w is $\text{proj}_w x_i = \frac{\langle w, x_i \rangle}{\langle w, w \rangle} = \frac{w^T x_i}{\|w\|^2} w$. To take the distance to our line, we then consider the magnitude of the difference between the projection with x_i.

$$\textsc{Dist}(x_i, L(w) = \|x_i - \frac{w^T x_i}{\|w\|^2} w\|^2$$

As a result, our objective is simply the sum of all these distances to the line. In the first step, we use $(a-b)^2 = a^2 - 2ab + b^2$. In the second, we use the fact that $w^T x_i$ is a scalar and that $x_i^T w = (w^T x)^T$ is the same scalar. In the fourth, we know that $\frac{w}{\|w\|}$ is a unit vector and that $\frac{w^T x_i}{\|w\|}$ is a scalar.

$$\begin{aligned}
&\textsc{Minimize}_w \sum_{i=1}^{n} \|x_i - \frac{w^T x_i}{\|w\|^2} w\|^2 \\
&= \sum_{i=1}^{n} \|x_i\|^2 - 2 x_i^T \frac{w^T x_i}{\|w\|^2} w + \|\frac{w^T x_i}{\|w\|^2} w\|^2 \\
&= \sum_{i=1}^{n} \|x_i\|^2 - 2 \frac{(w^T x_i)^2}{\|w\|^2} + \|\frac{w^T x_i}{\|w\|^2} w\|^2 \\
&= \sum_{i=1}^{n} \|x_i\|^2 - 2 (\frac{w^T x_i}{\|w\|})^2 + \|\frac{w^T x_i}{\|w\|} \frac{w}{\|w\|}\|^2 \\
&= \sum_{i=1}^{n} \|x_i\|^2 - 2 (\frac{w^T x_i}{\|w\|})^2 + (\frac{w^T x_i}{\|w\|})^2 \\
&= \sum_{i=1}^{n} \|x_i\|^2 - (\frac{w^T x_i}{\|w\|})^2
\end{aligned}$$

Since x_i are fixed, then $\|x_i\|_2^2$ are fixed. We can reformulate this as maximization problem, considering only the second term.

$$
\begin{aligned}
&\textsc{Maximize}_w \sum_i (\frac{w^T x_i}{\|w\|})^2 \\
&= \frac{\sum_i w^T x_i x_i^T w}{\|w\|^2} \\
&= \frac{w^T X X^T w}{\|w\|^2} \\
&= (\frac{w}{\|w\|})^T X X^T \frac{w}{\|w\|}
\end{aligned}
$$

$\frac{w}{\|w\|}$ are unit vectors, so we can consider $u \in \mathbb{R}^d$, where $\|u\|_2 = 1$.

$$\textsc{Maximize}_{u:\|u\|_2=1} u^T X X^T u$$

Note this objective function is precisely the formulation for PCA.

2 Latent Factor Analysis

In general, **factor analysis** encompasses variability in a set of observed variables using a smaller subset of unobserved variables. **Latent Factor Analysis** or **Latent Semantic Indexing** uses PCA to find a low-rank approximation. This allows us to recognize archetypes and user affinities for archetypes, or relationships between terms and concepts, by applying dimensionality reduction to a **term-document matrix**.

Each row is a document and each column is a word, so X_{ij} represents the number of occurrences of word j in document i. We see that this term-document matrix X is effectively a bag-of-words model, representing an unstructured piece of text.

2.1 Applications

1. **fuzzy search** The reduced-rank X' clusters synonyms, by SVD.
2. **denoising** Reducing dimensionality may improve classification, as it removes noise. $x_i = \alpha_i z + w_i$ as opposed to $\hat{x_i} = z^T x = \alpha + z^T \omega_i$. In other words, X may be a measurement of some low-rank matrix. Thus, the reduced-rank X' may be a better estimator.
3. **collaborative filtering** LFA may allow us to fill in values (i.e., matrix completion). Just as X' clusters synonyms, it groups users with similar tastes.

2.2 Relation to SVD

Recall that SVD gives us a decomposition of $X = U\Sigma V^T$, where this can be written entry-wise as

$$X = \sum_{i=1}^{d} \delta_i u_i v_i^T$$

, where δ_is are ordered from greatest to least. Consider each δ_i as a "genre". u_i then tells us which documents are associated with that genre, and v_i tells us which terms are associated with that genre.

This is a form of clustering, where we notice that similar genres or books see stronger cluster memberships. However, clusters can *overlap*, meaning that documents selected by u_1 are not necessarily disjoint from documents selected by u_2.

2.3 Relation to PCA

Formally, we're looking to factor X into AB^T. This yields the following objective function.

$$\textsc{Minimize}_{(A,B)} \|X - AB^T\|_F^2$$

As it turns out, the solution is $A = U_r S_r^{1/2}, B = V_r S_r^{1/2}$, where U_r is the r-rank approximation of U, S is the diagonal matrix with the first r singular values, and V is the r-rank approximation of V. In this case, we note that $\|A\|_F = \|B\|_F$. Consider the low-rank approximation of X, $X' \in \mathbb{R}^{r \times r}$. Per PCA, we select the r u_i, v_i with the largest singular values δ_i.

$$X' = \sum_{i=1}^{r} \delta_i u_i v_i^T$$

X' is the rank-r approximation that minimizes the squared Frobenius norm.

$$\|X - X'\|_F^2 = \sum_{i,j} (X_{ij} - X'_{ij})^2$$