Principal Component Analysis

Cornell CS 4/5780 — Lecture 5 — Spring 2022

Whereas k-means clustering sought to partition the data into homogeneous subgroups, principal component analysis (PCA) will seek to find, if it exists, low-dimensional structure in the data set $\{x\}_{i=1}^n$ (as before, $x_i \in \mathbb{R}^d$). This problem can be recast in several equivalent ways and we will see a few perspectives in these notes. Accordingly, PCA has many uses including data compression (analogous to building concise summaries of data points), item classification, data visualization, and more.

First, we will consider a simple mathematical model for data that directly motivates the PCA problem. Assume there exists some unit vector $u \in \mathbb{R}^d$ such that $x_i \approx \alpha_i u$ for some scalar α_i.[12] While x_i is high dimensional (assuming d is large), there is a sense in which it could be well approximated by a much smaller number of "features." In fact, given u (which is the same for all x_i) we could well approximate our data using only n numbers—the α_i. More concisely, we say that the x_i approximately lie in a subspace of dimension 1. Moreover, assuming the variability in α_i is larger than whatever is hiding in the approximately equals above, just knowing the coefficients α_i would also explain most of the variability in the data—if we want to understand how different various x_i are we could simply compare α_i. This is illustrated in fig. 9, where we see two dimensional data that approximately lies in a one dimensional subspace.

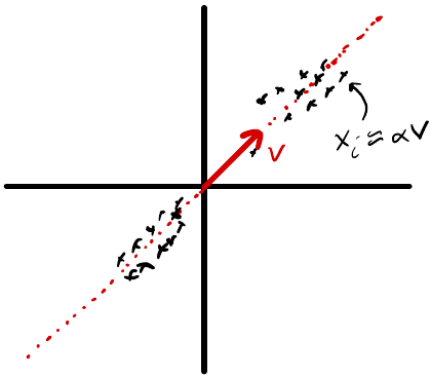

Figure 9: An example where two dimensional data approximately lies in a one dimensional subspace.

More generally, we will describe PCA from two perspectives. First we will view PCA as finding a low-dimensional representation of the data that captures most of the interesting behavior. Here, "interesting" will be defined as variability. This is analogous to computing "composite" features (i.e., linear combinations of entries in each x_i) that explain most of the variability in the data. Second, we will see how PCA can be thought of as providing low-dimensional approximation to the data that is the best possible given the dimension (i.e., if we only allow ourselves k dimensions to represent d dimensional data we will do so in the best possible way).

Centering the data. Before proceeding with a mathematical formulation of PCA there is one important pre-processing step we need to perform on the data. Typically, in unsupervised learning we are interested in understanding relationships between data points and not necessarily bulk properties of the data.[13] Taking the best approximation view of PCA this highlights a key problem of working with raw data points. In particular if the data has a sufficiently large mean, i.e., $\mu = \frac{1}{n}\sum_i x_i$ is sufficiently far from zero, the best approximation of each data point is roughly μ. An example of this is seen in fig. 10.

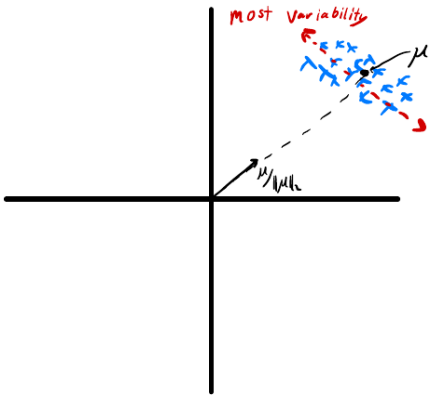

Figure 10: For data with a non-zero mean the best approximation is achieved using a vector similar to the mean; in contrast, most of the interesting behavior in the data may occur in completely different directions.

While finding μ this tells us something about a data set, we know how to compute it and that is not the goal of PCA. Therefore, to actually understand the relationship between features we would like to omit this complication. Fortunately, this can be accomplished by centering our data before applying PCA. Specifically, we let $\hat{x}_i = x_i - \mu$, where $\mu = \frac{1}{n}\sum_i x_i$. We now simply work with the centered feature vectors $\{\hat{x}_i\}_{i=1}^n$ and will do so throughout the remainder of these notes. Note that in some settings it may also be important to scale the entries of $\hat{x}_i$ (e.g., if they correspond to measurements that have vastly different scales).

Maximizing the variance. The first goal we will use to describe PCA is finding a small set of composite features that capture most of the variability in the data. To illustrate this point, we will first consider finding a single composite feature that captures as much of the variability in the data set as possible. Mathematically, this corresponds to finding a vector $\phi \in \mathbb{R}^d$ such that the sample variance of the scalars $z_i = \phi^T \hat{x}_i$ is as large as possible.[14] By convention we require that $\|\phi\|_2 = 1$, otherwise we could artificially inflate the variance by simply increasing the magnitude of the entries in ϕ. Using the definition of sample variance we can now formally define the first principal component of a data set.

> **First principal component:** The first principal component of a data set $\{x_i\}_{i=1}^n$ is the vector $\phi \in \mathbb{R}^d$ that solves
>
> $$\max_{\|\phi\|_2=1} \frac{1}{n} \sum_{i=1}^{n} (\phi^T \hat{x}_i)^2. \tag{4}$$

When we discussed k-means above we framed it as an optimization problem and then discussed how actually solving that problem was hard. In this case we do not have such a complication—the problem in eq. 4 has a known solution. It is useful to introduce a bit more notation to state the solution. Specifically, we will consider the data matrix

$$\widehat{X} = \begin{bmatrix} | & & | \\ \hat{x}_1 & \cdots & \hat{x}_n \\ | & & | \end{bmatrix}.$$

This allows us to rephrase eq. 4 as

$$\max_{\|\phi\|_2=1} \|\hat{X}^T \phi\|_2 = \sqrt{\max_{\|\phi\|_2=1} \phi^T \hat{X} \hat{X}^T \phi} = \sqrt{\max_{\|\phi\|_2=1} \phi^T \Sigma \phi},$$

where $\Sigma = \hat{X}\hat{X}^T$. In other words, ϕ is the unit vector that the matrix $\widehat{X}^T$ makes as large as possible.

How can we solve this using an eigendecomposition?

Solving via singular value decomposition. Doing PCA with an eigendecomposition is easier to understand At this point we need to (re)introduce one of the most powerful matrix decompositions—the singular value decomposition (SVD).[15] To simplify this presentation we make the reasonable assumption that $n \geq d$.[16] In particular, we can always decompose the matrix $\widehat{X}$ as $\widehat{X} = U \Sigma V^T$ where U is an $d \times d$ matrix with orthonormal columns, V is an $n \times d$ matrix with orthonormal columns, Σ is a $d \times d$ diagonal matrix with $\Sigma_{ii} = \sigma_i \geq 0$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$. Letting

$$U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_d \\ | & & | \end{bmatrix}, \quad V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_d \\ | & & | \end{bmatrix}, \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \end{bmatrix}$$

we call u_i the left singular vectors, v_i the right singular vectors, and σ_i the singular values. While methods for computing the SVD are beyond the scope of this class, efficient algorithms exist that cost $\mathcal{O}(nd^2)$ and, in cases where n and d are large it is possible to efficiently compute only the first few singular values and vectors (e.g., for $i = 1, \ldots, k$).

Under the mild assumption that $\sigma_1 > \sigma_2$ the SVD the solution to eq. 4 becomes apparent: $\phi = u_1$.[17] All ϕ can be written as $\phi = \sum_{i=1}^{d} a_i u_i$ where $\sum_i a_i^2 = 1$ (because we want $\|\phi\|_2 = 1$). We now observe that

$$\|\widehat{X}^T\phi\|_2^2 = \|V\Sigma U^T\phi\|_2^2 = \left\|\sum_{i=1}^{d}(\sigma_i a_i)v_i\right\|_2^2 = \sum_{i=1}^{d}(\sigma_i a_i)^2,$$

so the quantity is maximized by setting $a_1 = 1$ and $a_i = 0$ for $i \neq 1$.

So, the first left singular value of $\widehat{X}$ gives us the first principal component of the data. What about finding additional directions? A natural way to set up this problem is to look for the composite feature with the next most variability. Informally, we could look for ψ such that $y_i = \psi^T\hat{x}_i$ has maximal sample variance. However, as stated we would simply get the first principal component again. Therefore, we need to force the second principal component to be distinct from this first. This is accomplished by forcing them to be orthogonal, i.e., $\psi^T\phi = 0$. While this may seem like a complex constraint it is actually not here. In fact, the SVD still reveals the solution: the second principal component is $\psi = u_2$. Fig. 11 illustrates how the first two principal components look for a stylized data set. We see that they reveal directions in which the data varies significantly.

More generally, we may want to consider the top k principal components. In other words the k directions in which the data varies the most. We denote the principal component ℓ as ϕ_ℓ and to enforce that we find different directions we require that $\phi_\ell^T\phi_{\ell'} = 0$ for $\ell \neq \ell'$. We also order them by the sample variance of $\phi_\ell^T\hat{x}_i$, i.e., $\phi_\ell^T\hat{x}_i$ has greater variability than $\phi_{\ell'}^T\hat{x}_i$ for $\ell < \ell'$.

While we will not present a full proof, the SVD actually gives us all the principal components of the data set $\{x_i\}_{i=1}^n$.

> **Principal components:** Principal component ℓ of data set $\{x_i\}_{i=1}^n$ is denoted ϕ_ℓ and satisfies $\phi_\ell = u_\ell$, where $\widehat{X} = U\Sigma V^T$ is the SVD of $\widehat{X}$.

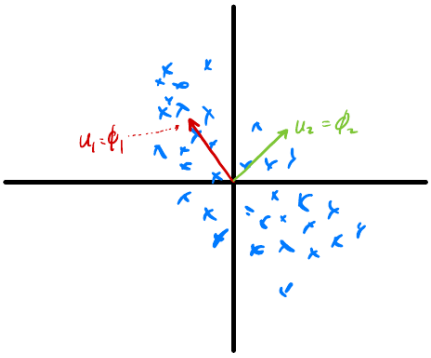

Figure 11: Two principal components for a simple data set.

Explaining variability in the data Our discussion of PCA started with considering variability in the data. Therefore, we would like to consider how the principal components explain variability in the data. First, by using the SVD of $\widehat{X}$ we have already computed the sample variance for each principal component. In fact, we can show that

$$\mathrm{Var}(\phi_\ell^T\hat{x}_i) = \sigma_\ell^2/n.$$

In other words, the singular values reveal the sample variance of $\phi_\ell^T\hat{x}_i$.

Since we may want to think of principal components as providing a low-dimensional representation of the data a reasonable question to ask is how many we should compute/use for downstream tasks. One way to address this question is to try and address a sufficient fraction of the variability in the data. In other words, we pick k such that $\phi_1, \ldots, \phi_k$ capture most of the variability in the data. To understand this we have to understand what the total variability of the data is. Fortunately, this is can be easily computed as

$$\sum_{j=1}^{d}\sum_{i=1}^{n}(\hat{x}_i(j))^2.$$

In other words we simply sum up the squares of all the entries in all the data points.

What is true, though less apparent, is that the total variability in the data is also encoded in the singular values of $\widehat{X}$ since

$$\sum_{j=1}^{d}\sum_{i=1}^{n}(\hat{x}_i(j))^2 = \sum_{i=1}^{d}\sigma_i^2.$$

Similarly we can actually encode the total variability of the data captured by the first k principal components as $\sum_{i=1}^{k} \sigma_i^2$. (This is a consequence of the orthogonality of the principal components.) Therefore, the proportion of total variance explained by the first k principal components is

$$\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{d} \sigma_i^2}.$$

So, we can simply pick k to explain whatever fraction of the variance we want. Or, similar to with k-means we can pick k by identifying when we have diminishing returns in explaining variance by adding more principal components, i.e., looking for a knee in the plot of singular values.

Best approximations. We now briefly touch on how PCA is also solving a best approximation problem for the centered data $\hat{x}_i$. Specifically, say we want to approximate every data point $\hat{x}_i$ by a point w_i in a fixed k dimensional subspace. Which subspace should we pick to minimize $\sum_{i=1}^{n} \|\hat{x}_i - w_i\|_2^2$? Formally, this can be stated as finding a $n \times k$ matrix W with orthonormal columns that solves

$$\min_{\substack{W \in \mathbb{R}^{n \times d} \\ W^T W = I}} \sum_{i=1}^{n} \|\hat{x}_i - W(W^T \hat{x}_i)\|_2^2. \tag{5}$$

This is because if we force $w_i = W z_i$ for some $z_i \in \mathbb{R}^k$ (i.e., w_i lies in the span of the columns of W) the choice of z_i that minimizes $\|\hat{x}_i - W z_i\|_2$ is $z_i = W^T \hat{x}$.

While this is starting to become a bit predictable, the SVD again yields the solution to this problem. In particular, the problem specified in eq. 5 is solved by setting the columns of W to be the first k left singular vectors of $\widehat{X}$ or, analogously, the first k principal components. In other words, if we want to project our data $\hat{x}_i$ onto a k dimensional subspace the best choice is the subspace defined by the first k principal components. Notably, this fact is tied to our definition of best as the subspace where the sum of squared distances between the data points and their projections are minimized. The geometry of this is illustrated in fig. 12.

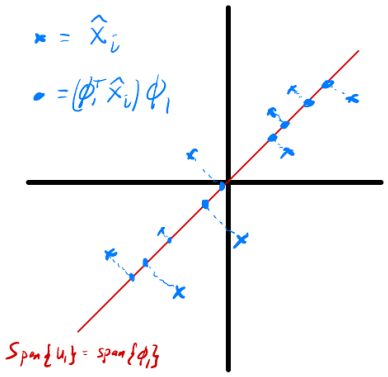

Figure 12: The best one dimensional approximation to a two dimensional data set.

PCA for visualization. A common use of PCA is for data visualization. In particular, if we have high dimensional data that is hard to visualize we can sometimes see key features of the data by plotting its projection onto a few (1, 2, or 3) principal components. For example, if $d = 2$ this corresponds to forming a scatter plot of $(\phi_1^T \hat{x}_i, \phi_2^T \hat{x}_i)$. It is important to consider that only a portion of the variability in the data is "included" in visualizations of this type and any key information that is orthogonal to the first few principal components will be completely missed. Nevertheless, this way of looking at data can prove quite useful—see the class demo.

Footnotes

12. Assume that the error in this approximation is much smaller than the variation in the coefficients α_i.

13. For example, k-means is shift invariant—if we add an arbitrary vector $w \in \mathbb{R}^d$ to every data point it does not change the outcome of our algorithm/analysis.

14. We call this a composite feature because z_i is a linear combination of the features (entries) in each data vector $\hat{x}_i$.

15. This is often taught using eigenvectors instead. The connection is through the sample covariance matrix $\widehat{\Sigma} = \widehat{X}\widehat{X}^T$. The singular vectors we will use are equivalent to the eigenvectors of $\widehat{\Sigma}$ the the eigenvalues of $\widehat{\Sigma}$ are the singular values of $\widehat{X}$ squared.

16. Everything works out fine if $n < d$ but we need to write $\min(n, d)$ in a bunch of places.

17. If $\sigma_1 = \sigma_2$ the first principal component is not well defined.