# 2.5 PCA

Monday, January 13, 2020 5:25 PM

intoli.com/blog/pca-and-svd

We claimed earlier that SVD preserves variance. This is not quite true because we only preserved squared error unless we first center the data. How should we do that?

**Lemma 3.14** The $k$-dim affine subspace which minimizes the sum of squared perpendicular distances to the data points must pass through the centroid.

**proof sketch** Proof by contradiction. i.e. center at 0. Suppose not for a line. Then the same line slope but going through the origin is better, and can show via direct computation.

Implication: centering data does what we want.

## Principal Component Analysis

Goals: project onto few dimensions to retain as much "variability" in the data as possible.

The way we defined SVD, it's pretty clear it is related, but let's derive this again in a goal-oriented way.

Consider a dataset of $n$ $d$-dim vectors $\vec{a}_1, \dots, \vec{a}_n \in \mathbb{R}^d$, where each of the $d$-dimensions corresponds to a feature of the data point.

e.g. A medical dataset may have $n$ patients, and $d = 3$ features age, weight, height.

The data matrix $A = \begin{bmatrix} \vec{a}_1^T \\ \vdots \\ \vec{a}_n^T \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nd} \end{bmatrix}$.

The covariance matrix $S \in \mathbb{R}^{d \times d}$ is the symmetric array of numbers

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (a_{ij} - \bar{a}_j)(a_{ik} - \bar{a}_k), \text{ where } \bar{a}_j = \frac{1}{n} \sum_{i=1}^{n} a_{ij} \quad \text{(mean of var } j\text{)}$$

Covariance matrix gives all the pairwise covariances between variables.

Let's center $A$ by defining $A_c = A - \vec{1}_n \vec{\bar{a}}$, where $\vec{\bar{a}} = (\bar{a}_1, \dots, \bar{a}_d)$.

Then $S = \frac{1}{n-1} A_c^T A_c$ (exercise for reader)

$$= \frac{1}{n-1} \begin{bmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & & \vdots \\ a_{1d} & \cdots & a_{nd} \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & & \\ a_{n1} & \cdots & a_{nd} \end{bmatrix} = \frac{1}{n-1} \begin{bmatrix} \sum_{i=1}^{n} a_{i1}^2 & \cdots & \sum a_{i1} a_{id} \\ \vdots & & \vdots \\ \sum a_{id} a_{i1} & \cdots & \sum a_{id}^2 \end{bmatrix}$$

Note that the direction / combination of variables which maximizes variance is given by $\underset{|\vec{v}|=1}{\operatorname{argmax}}\ \vec{v}^T S \vec{v} = \underset{|\vec{v}|=1}{\operatorname{argmax}}\ \frac{\vec{v}^T A_c^T A_c \vec{v}}{n-1} = \underset{|\vec{v}|=1}{\operatorname{argmax}}\ \frac{|A_c \vec{v}|^2}{n-1}$,

which of course gives singular vectors of $A_c$.

i.e. choosing singular vectors of the centered data matrix immediately gives directions of high variance.

Furthermore, singular values tell us how much variance is explained in a particular direction.

For the first "principal component" $\vec{v}_1$ of the dataset, the variance along that component is $\vec{v}_1^T S \vec{v} = \frac{1}{n-1}|A_c \vec{v}|^2 = \frac{\sigma^2}{n-1}$.

But we have a problem. What if our original features were not comparable?

Suppose we have a dataset of 100 patients with age, weight, height.

| Patient | Age | Weight | Height |
|---|---|---|---|
| Alice | 25 years | 40 kg | 170 cm |
| Bob | 23 years | 50 kg | 180 cm |
| Carol | 24 years | 45 kg | 175 cm |
| Dan | 25 years | 70 kg | 200 cm |
| Erin | 23 years | 50 kg | 176 cm |
| ⋮ | | | |

Apply PCA. What will PCA capture first?

Weight + Height

What if we measure Age in minutes instead?

525,600 minutes in a year, so we suddenly scaled up the variance in age

Solution: first normalize all variables to have unit variance.

Relationship between PCA & SVD

SVD is an orthogonal matrix decomposition preserving squared distances.

PCA is a data analysis technique that tries to explain variability, applying SVD to various transformed data matrices.

Some people will disagree on choice of normalization or transform

e.g. unit variance if different units

log-transforms if log-normal, etc.

---

## Clustering a mixture of spherical Gaussians

# Clustering a mixture of spherical Gaussians

Simplest Gaussian mixture model.

Task: partition a set of points into $k$ subsets or clusters where each consists of nearby points.

**Define:** A mixture model is a probability density or distribution that is a weighted sum of simple component probability distributions.

e.g. of the form $f = w_1 p_1 + \cdots + w_k p_k$, where

$p_i$'s are basis probability

$w_i \in \mathbb{R}$, $\sum w_i = 1$. (mixture weights)

**Define:** The model fitting problem is to fit a mixture of $k$ basic densities to $n$ i.i.d. samples drawn from some $f$.

We know the class of basic density, but not all parameters & weights.

**Note:** First "machine learning" example, as our model "learns" weights & parameters from our data.

**Proposed solution:**

1. Cluster our data
2. Fit a Gaussian to each cluster. (spherical)
   ↳ easy because it turns out that the best fit Gaussian is just the one with the empirical mean & standard deviation.

difficult to cluster when there is lots of overlap

Easier when far apart.

e.g. $|\mu_1 - \mu_2| > 6 \max(\sigma_1, \sigma_2)$

implies $< 0.03$ overlap.

(99.7% within $3\sigma$)

One naive approach: If $\vec{x}, \vec{y} \sim N(\mu, \sigma)$, $\vec{x}, \vec{y} \in \mathbb{R}^d$.

Then $|\vec{x} - \vec{y}|^2 \approx 2(\sqrt{d} \pm O(1))^2 \sigma^2$.

↑ "radius of Gaussian"

($\sigma$ same in every direction for spherical)

then $|\vec{x}-\vec{y}| \approx \sqrt{(d \pm O(1))}\sigma$.

↑ orthogonality of points. ↑ "radius of Gaussian"

Spherical

If $\vec{x} \sim N(\mu_1, \sigma)$, $\vec{y} \sim (\mu_2, \sigma)$, $|\mu_1 - \mu_2| = \Delta$,

then $|\vec{x}-\vec{y}|^2 \approx 2(\sqrt{d} \pm O(1))^2\sigma^2 + \Delta^2$

all directions "orthogonal"

So we need $2(\sqrt{d} \pm O(1))^2\sigma^2 + \Delta^2 > 2(\sqrt{d} \pm O(1))^2\sigma^2$

$(2d \pm O(1)\sqrt{d})\sigma^2 + \Delta^2 > (2d \pm O(1)\sqrt{d})\sigma^2$

$\Delta^2 > O(1)\sqrt{d}\,\sigma^2$

$\Delta > O(1)\sigma d^{\frac{1}{4}} = c\sigma d^{\frac{1}{4}}$, for some constant $c$.

↑ unfortunately separation needed is dependent on dimension, rather than just standard deviation.

Want to get rid of dependence on $d$.

**Better approach** Use SVD to project down to lower dimension, with less dependence on $d$.

Note, projecting a spherical Gaussian of standard dev $\sigma$ remains a " " " " " "

**Lemma 3.15** Suppose $p$ is a $d$-dim spherical subspace with center $\mu$ and standard dev $\sigma$. The density of $p$ projected onto a $k$-dim subspace $V$ is spherical with same stdev $\sigma$.

**proof sketch** Check densities.

**Def. 3.1** If $p$ is a prob. density in $d$-space, the best-fit line for $p$ is the line in the $\vec{v}_1$ direction where

$$\vec{v}_1 = \underset{|v|=1}{\arg\max}\ \mathbb{E}_{x\sim p}\left[(\vec{v}^T\vec{x})^2\right]$$

For a spherical Gaussian centered at the origin, any line passing through the center is a best-fit line.

**Lemma 3.16** Let the prob density $p$ be a spherical Gaussian with center $\mu \neq 0$.

**Lemma 3.16** Let the prob. density $p$ be a spherical Gaussian with center $\mu \neq 0$. The unique best-fit 1D subspace is the line passing through $\mu$ & the origin. If $\mu = 0$, any line through the origin is a best-fit line.

**proof.** For a randomly chosen $\vec{x}$ (according to $p$) and a fixed unit length $\vec{v}$,

$$\mathop{\mathbb{E}}_{\vec{x}\sim p}\left[(\vec{v}^T\vec{x})^2\right] = \mathop{\mathbb{E}}_{\vec{x}\sim p}\left[(\vec{v}^T(\vec{x}-\vec{\mu})+\vec{v}^T\vec{\mu})^2\right]$$

$$= \underbrace{\mathop{\mathbb{E}}_{\vec{x}\sim p}\left[(\vec{v}^T(\vec{x}-\vec{\mu}))^2\right]}_{\sigma^2} + 2(\vec{v}^T\vec{\mu})\underbrace{\mathop{\mathbb{E}}_{\vec{x}\sim p}\left[\vec{v}(\vec{x}-\vec{\mu})\right]}_{0 \text{ because } \mathbb{E}\vec{x}=\vec{\mu}} + (\vec{v}^T\vec{\mu})^2$$

$$= \sigma^2 + (\vec{v}^T\vec{\mu})^2$$

$\Rightarrow$ best-fit line maximizes $(\vec{v}^T\vec{\mu})^2$, so must be the line in the direction $\vec{\mu}$ from the origin. □

**Def 3.2** If $p$ is a prob. density in $d$-space, then the best-fit $k$-dim subsp. $V_k$ is

$$V_k = \mathop{\arg\max}_{\substack{V \\ \dim(V)=k}} \mathop{\mathbb{E}}_{\vec{x}\sim p}\left(|\text{proj}(\vec{x}, V)|^2\right),$$

where $\text{proj}(\vec{x}, V)$ is the orthogonal proj. of $\vec{x}$ onto $V$.

**Lemma 3.17** For a spherical Gaussian with center $\mu$, a $k$-dim subsp. is a best-fit iff it contains $\mu$.

**proof.** If $\vec{\mu} = 0$, by symmetry anything works.

If $\vec{\mu} \neq 0$, proj. perpendicular to $\vec{\mu}$ first as by prev. lemma must pass through $\vec{\mu}$. Then have case $\vec{\mu} = 0$. □

**Thm 3.18** If $p$ is a mixture of $k$ spherical Gaussians, the the best-fit $k$-dim subsp. contains the centers.

Thus, recall $\Delta > cd^{\frac{1}{4}}$. Here, we can now replace $d$ with $k$.

$\Delta > ck^{\frac{1}{4}}$.

So if $k = O(1)$, we only need $O(1)$ distance, as in