THE MATH BEHIND PCA

DATA

Suppose we have m data points in $\mathbb{R}^n$, ie. each data point has n coordinates (usually called features in machine learning), with respect to an arbitrary vector basis $\{e_1, \ldots, e_n\}$:

$$x^{(i)} = \sum_{j=1}^{n} X_{ij} e^{(j)} \qquad i = 1, \ldots, m.$$

Where $x^{(i)}, e^{(j)} \in \mathbb{R}^n$. Together, the data pts and the basis $e^{(i)}$ define a $m \times n$ matrix

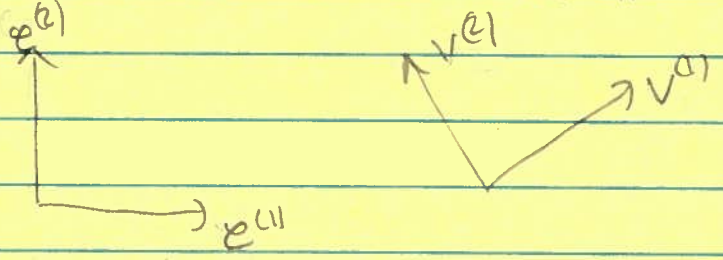
$$X = (X_{ij})_{i=1\cdots m;\, j=1\cdots n}$$

Now perform SVD:

SVD

$$X = U \Sigma V^T$$

The columns of V, denoted $v^{(i)}$, define a new basis in $\mathbb{R}^n$.



$$V = \begin{bmatrix} | & | & & | \\ v^{(1)} & v^{(2)} & \cdots & v^{(n)} \\ | & | & & | \end{bmatrix}$$

CHANGE OF BASIS

In component form, the SVD says:

$$X_{ij} = \sum_k (U\Sigma)_{ik} (V^T)_{kj} \qquad (*)$$

But

$$X_{ij} = x_j^{(i)}$$

$$(V^T)_{kj} = V_{jk} = v_j^{(k)}$$

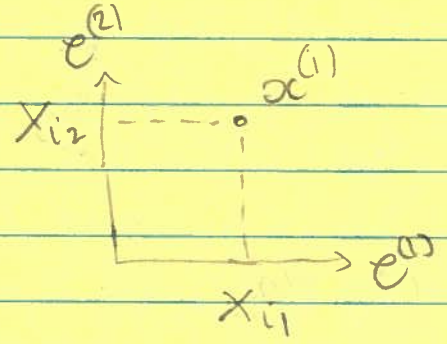
Thus $(*) \Rightarrow$

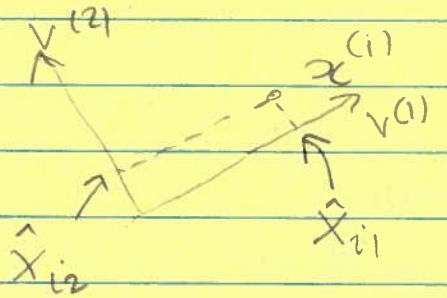
$$x_j^{(i)} = \sum_k \hat{X}_{ik} \, v_j^{(k)} \Rightarrow \boxed{x^{(i)} = \sum_k \hat{X}_{ik} v^{(k)}, \quad i = 1, \dots, n.}$$

where:

$\hat{X}_{ik} = k^{th}$ component of $x^{(i)}$ wrt basis $v^{(1)}, \dots, v^{(n)}$.

$= (U\Sigma)_{ik}$, i.e. $\boxed{\hat{X} = U\Sigma}$





Thus V provides a new basis and $U\Sigma$ provides the coordinates w.r.t. that new basis.

COVARIANCE OF FEATURES

Let X_i = random variable, realizations of which lie in the i^{th} column of X, ie. the n samples of X_i are $\{X_{1i}, \ldots, X_{mi}\}$.

Let us now compute the covariance of X_i and X_j:

$$\text{cov}(X_i, X_j) = \mathbb{E}\left[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right]$$

Now,

$$\mathbb{E}[X_i] \simeq \frac{X_{1i} + \cdots X_{mi}}{m} \qquad (\star\star)$$

Let us suppose we have "mean normalized" the data, ie:

$$x^{(i)} \leftarrow x^{(i)} - \frac{x^{(1)} + \cdots + x^{(m)}}{m}$$

Then $(\star\star) = 0$ and the covariance collapses to

$$\text{cov}(X_i, X_j) = \mathbb{E}[X_i X_j]$$

$$\simeq \frac{1}{m} \sum_{k=1}^{m} X_{ki} X_{kj}$$

$$= \frac{1}{m} \sum_{k} (X^T)_{ik} X_{kj}$$

$$= \frac{1}{m}(X^T X)_{ij}$$

In general, all elements of $X^T X$ will be "non-zero", ie. all features are correlated w/ one another.

Contrast that with the covariance of the new features (coordinates) defined by the SVD basis $\{e^{(i)}\}$:

$$\hat{X}^T \hat{X} = (U\Sigma)^T (U\Sigma)$$

$$= \Sigma^T U^T U \Sigma$$

$$= \Sigma^T \Sigma$$

$$= \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ & & \sigma_r^2 \\ 0 & & & \ddots \\ & & & & 0 \end{bmatrix} \quad (= n \times n)$$

Thus in the SVD basis, the features are independent (uncorrelated), and their variance is:

$$\text{var}(\hat{X}_i) = \text{cov}(\hat{X}_i, X_i)$$

$$= (X^T X)_{ii}$$

$$= \sigma_i^2.$$

In summary, the spread of the data pts along $v^{(i)}$, as measured by $\text{var}(\hat{x}_i)$, is σ_i^2.

PCA / LSA (cf wiki/Latent_semantic_analysis #Derivation)

github:
Math 105A → lab_worksheets – Lab10.

Add this to p.6 of Lab10.pdf:

The eigenvectors & "covariance matrix" are $v^{(i)}$. Here's how:

$$X = \begin{matrix} \uparrow \\ m \\ \downarrow \end{matrix} \overset{\longleftarrow n \longrightarrow}{\begin{bmatrix} & \\ & \end{bmatrix}}$$

$$X = U \Sigma V^T$$

where $\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ & 0 & \end{bmatrix}$ $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$

$$V = \begin{bmatrix} | & & | \\ v^{(1)} & \cdots & v^{(n)} \\ | & & | \end{bmatrix}$$

X_i = values of a R.V.
= i^{th} column of X.

$$\text{cov}(X_i, X_j) = \frac{1}{m} (X^T X)_{ij}$$

$$X^T X v^{(i)}$$

$$= (U \Sigma V^T)^T (U \Sigma V^T) v^{(i)}$$

$$= V \Sigma \underbrace{U^T U}_{I} \Sigma V^T v^{(i)}$$

$$= V \underbrace{\Sigma \Sigma}_{\begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2 \end{bmatrix}} \underbrace{V^T v^{(i)}}_{\begin{bmatrix} - & v^{(1)} & - \\ & \vdots & \\ - & v^{(n)} & - \end{bmatrix} \begin{bmatrix} | \\ v^{(i)} \\ | \end{bmatrix}}$$

$$= \begin{bmatrix} 0 \\ \vdots \\ \sigma_i^2 \\ \vdots \\ 0 \end{bmatrix}$$

$$= V \begin{bmatrix} 0 \\ \vdots \\ \sigma_i^2 \\ \vdots \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} | & & | \\ v^{(1)} & \cdots & v^{(n)} \\ | & & | \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ \sigma_i^2 \\ \vdots \\ 0 \end{bmatrix}$$

$$= \sigma_i^2 \begin{bmatrix} v_1^{(i)} \\ \vdots \\ v_n^{(i)} \end{bmatrix} = \sigma_i^2 v^{(i)}$$

We have proved that

$$(X^T X)\, v^{(i)} = \sigma_i^2 v^{(i)}.$$

The proof that SVD of X or eigendecomposition of $X^T X$ is equivalent to an optimization problem where one successively finds eigenvectors that maximize variance of data along them can be found in "Details" section of Wikipedia article on PCA.

MathWorks / Lab10: probabilistic PCA paper by Bishop - Eqs (4) and (3) show how to generate observed data using a hidden variable s.t. MLE of parameters in the model yields the principal axes of the data, and the variance of the data along those axes!