LAB10

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 $\{e_1, \dots, e_n\}$:

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

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

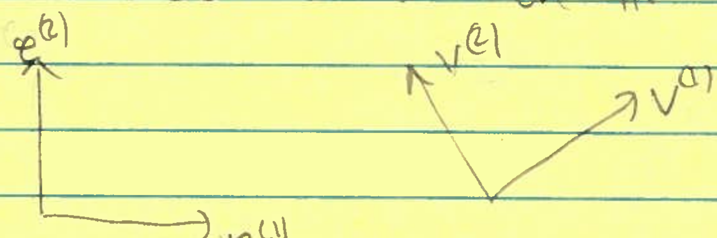
$$X = (X_{ij})_{i=1\dots m;\, j=1\dots n}$$

Now perform SVD:

SVD

$$\boxed{X = U \Sigma V^T}$$

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

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

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

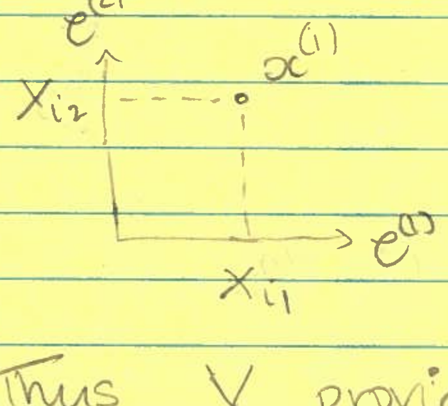
Thus $(*) \Rightarrow$

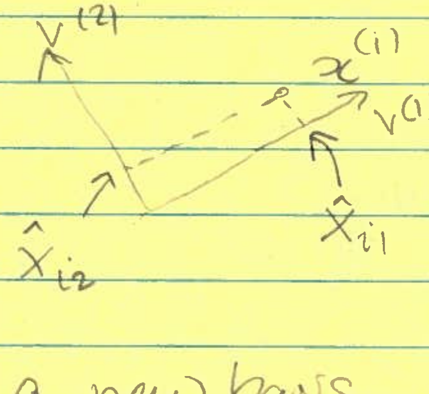
$$x_j^{(i)} = \sum_k \hat{X}_{ik} \; v_j^{(k)} \Rightarrow \boxed{\begin{array}{l} x^{(i)} = \sum_k \hat{X}_{ik} v^{(k)} \\ i = 1, \dots, n. \end{array}}$$

where:

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

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





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:

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

Now,

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

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

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

Then $(**) = 0$ and the covariance collapses to

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

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

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

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

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 $\{e^{(i)}\}$:

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

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

$$= \Sigma^T \Sigma$$

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

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

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

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

$$= \sigma_i^2 .$$

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