Data Reduction Fundamentals :

Principal Component Analysis -

* Based on the Singular Value Decomposition
* Non-parametric approach towards denoising
* Usually leads to shrinkage in one dimension of the data

Variance $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}$

Q Why is statistical variance divided by "n-1" and not "n"?
That is, why not $S^2 = \frac{\sum_i (x_i - \mu)^2}{n}$? Observe that

$$S^2 = \frac{\sum_i [x_i^2 + \mu^2 - 2x_i\mu]}{n} = \frac{\sum_i x_i^2}{n} - \mu^2$$

Now, if we compute S^2 over many mini-batches of size "n"

$$\mathbb{E}[S^2] = \frac{1}{n}\sum_i \mathbb{E}[x_i^2] - \mathbb{E}[\mu^2]$$

If $x_i \sim N(\mu, \sigma)$ then $\mathbb{E}[x_i^2] = \sigma^2 + \mu^2$

and $\mathbb{E}[\mu^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_i x_i\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_i x_i^2 + 2\sum_i x_i x_j\right]$

$$= \frac{1}{n^2}\left[n(\sigma^2+\mu^2) + 2\sum_i \mathbb{E}[x_i x_j]\right]$$

$n(n-1)$ (Assumes x_i, x_j are i.i.d)

$$= \mu^2 + \frac{\sigma^2}{n}$$

$$\Rightarrow \mathbb{E}[S^2] = \sigma^2 + \mu^2 - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \sigma^2\left(\frac{n-1}{n}\right)$$

Thus, to match ensemble average of i.i.d variances with the theoretical ("generator") average, we define $S^2 = \frac{\sum_i (x_i - \mu)^2}{n-1}$ instead.

→ This argument is due to Bessel but there are a lot of non-technical reasons as well to prefer $(n-1)$ in the denominator

Consequently, define covariance as $\text{cov}(X,Y) = \frac{\sum (X_i - \mu_X)(Y_i - \mu_Y)}{n-1}$

If $\text{cov}(X,Y) > 0$ we say variables are positively correlated, $\text{cov}(X,Y) < 0$ neg. correlated, $\text{cov}(X,Y) = 0$ variables are uncorrelated (*)

(This is wrong! Independence $\Rightarrow \text{cov}(X,Y) = 0$

But dependent X, Y can be made to have $E(XY) = E(X)E(Y)$!)

Eg: Take RV X s.t. that $E(X) = 0$ and $E(X^3) = 0$

i.e., $X \sim N(0, \sigma^2)$ then $\text{cov}(X,Y) = E(X^3) = 0$ but Y, X are clearly dependent!

if $Y = X^2$

=

Linear Independence:

If vectors $X_1, X_2, \ldots, X_k$ are LI then if

$$\sum_1^k c_j X_j = 0 \implies c_1 = c_2 = \ldots c_k = 0.$$

be written

That is no vector can linearly as a combination of the rest. Basically linear dependence is a measure of redundancy among the vectors.

Span: The span of a set of vectors is the set of all vectors generated by linearly combining them.

Basis: If the spanning vectors are linearly independent and # of such vectors = dimension of vector space then the Span is the Basis for that vector space

Orthogonal / Orthonormal:

If the basis vectors furthermore have mutually 0 dot product and if they have each norm 1

Using the Gram-Schmidt method you can take any linearly independent set of vectors and produce an orthogonal set of vectors:

Start with $v_1, v_2, \ldots, v_n$. Define $\text{proj}_u^{\circ} v = \frac{\langle u, v \rangle}{\langle u, u \rangle} u$

Then

$$u_k = v_k - \sum_{j=1}^{k-1} \text{proj}_{u_j}^{\circ}(v_k)$$

The idea behind PCA:

Noise information cannot have high variance or else it would be something interesting i.e. a signal

consequently, if we can find the basis of representation for a dataset in which the axes are aligned towards directions of high variance (signal directions) we can read off qualitative information more easily

$$P_{m \times m} X_{m \times n} = Y_{m \times n} \qquad \underline{p}_i \in \mathbb{R}^m$$

x_i, y_i

$$\begin{bmatrix} \underline{p}_1 \\ \underline{p}_2 \\ \vdots \\ \underline{p}_m \end{bmatrix} \begin{bmatrix} \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_n \end{bmatrix} = \begin{bmatrix} \underline{y}_1 & \underline{y}_2 & \cdots & \underline{y}_n \end{bmatrix}$$

$$\underline{y}_1^T = \left(\langle \underline{p}_1, \underline{x}_1 \rangle, \langle \underline{p}_2, \underline{x}_1 \rangle, \ldots, \langle \underline{p}_m, \underline{x}_1 \rangle \right)$$

Note that this is equivalent to scaling + stretching of $\underline{x}_1$ over m axes

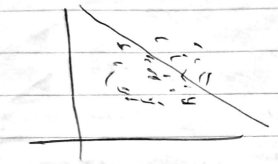
That is, we are changing the basis of $\underline{x}_k$ from $\mathbb{R}^m$ to P and the new representation is $\underline{y}_k \in \mathbb{R}^m$ spanned by P.

What is key here is that change of bases does not modify the data itself, it just provides a new viewpoint → It is the same as rotation/translation of axes in 2D, 3D.

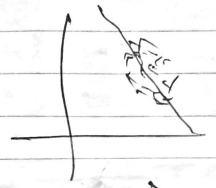
As described earlier, we want to represent our data in a basis where the noise variance is minimized. That is, if we call

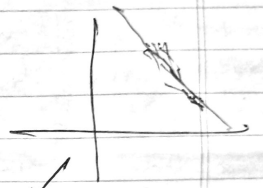
$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

we want to re-express X such that in Y, SNR is maximized.



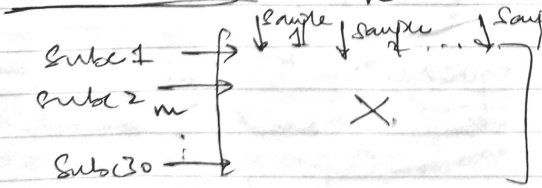
there there is no basis where SNR is nice because noise variance itself is high (cannot distinguish from signal)





there ~~exists~~ exists a rotation of axis where the highly "correlated" samples lie on. This is our (first) principal component

Data Matrix:



Basically row dimension is object dimension, column dimension is true dimension

We subtract the mean across row dimension (over time) to center our data.

Thus, we have $\hat{\Sigma}_X = \frac{1}{n-1} X X^T$

(Note that we are now calling X as row vectors i.e, $X \in \mathbb{R}^n$)

$$\hat{\Sigma}_{X_{ij}} = \frac{1}{n-1} \langle x_i, x_j \rangle$$

→ Covariance matrix is symmetric square
→ The diagonal is packed with variance of individual subcarriers
→ The off diagonal is filled with cross variance or correlation b/w different subcarriers.

Signal Energy is on diagonal, noise energy is off diagonal, can we find a change of basis transformation such that $\hat{\Sigma}_X \to \hat{\Sigma}_Y$ is diagonalized.

We want change of basis / PCA matrix to be orthonormal i.e, $\langle p_i, p_j \rangle = \delta_{ij}$ so that we can find directions of descending variance by iteratively eliminating previously exposed orthogonal directions.

$$\hat{\Sigma}_Y = \frac{1}{n-1} \hat{Y} \hat{Y}^T = \frac{1}{n-1} (P\hat{X})(P\hat{X})^T$$

$$= \frac{1}{n-1} P(\hat{X}\hat{X}^T) P^T$$

$$= P \hat{\Sigma}_X P^T.$$

So, we need to find P s.t for any $\hat{\Sigma}_X$, $\hat{\Sigma}_Y$ is diagonalized.

Recall, eigenvalue diagonalization → any square matrix $\hat{\Sigma}_x = \Lambda D \Lambda^T$ where D is a diag mtx composed of eigenvalues and Λ has corresponding eigenvectors.

(* The Karhunen Loeve expansion is a generalization of the e.v basis of Cov matrices for arbitrary random processes)

If we pick $P = \Lambda^T$ then $\hat{\Sigma}_y = D$ as

$P^{-1} = P^T = \Lambda^{-1}$.

In summary, the change of basis is the eigenbasis of the covariance matrix of the signal data structure is, the new representation in the Karhunen-Loeve basis yields a diagonal cov.

Therefore, if we sort the eigenvalues in a descending order, we have the most to-least significant principal components

If you drop the last k eigenvectors, you have a lower dimensional representation of your data.

$$\text{Proportion of Variance} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_r}{\lambda_1 + \lambda_2 + \cdots + \lambda_m}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$.

When you swap eigenvalue decomp with its parent the SVD you obtain LSA / Linear Discriminant Analysis via the Eckart-Young-Mirsky Theorem (Next Lecture)