PCA

- interpretation of data can be complicated by correlation amongst features and noise in the data
- PCA can completely get rid of correlation, and eliminates the ^(most noisy) dims.

Assumptions

i.e. an elliptical cloud of data

① principle directions that correspond to actual signal are ⊥ to each other (necessary to diagonalize Σ)

② data behave linearly

③ directions w/ higher variance correspond to actual signal

①  not = (ICA would be better)

②  (not linear) kernel PCA

Theory

- let $X \in \mathbb{R}^{m \times n}$, and let all cols. be centered by the mean (so we can do linear algebra)

- the sample Cov-Matrix is

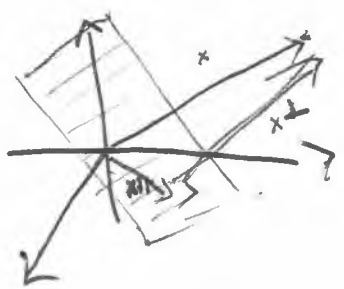
$$\underset{n \times n}{\Sigma} = \frac{1}{m-1} \underset{n \times m}{X^T} \underset{m \times n}{X}$$; which is a real, symmetric matrix

- it can be thus diagonalized in a basis of its eigenvectors (which are orthonormal by spectral thm)

$$S = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_n & - \end{bmatrix} \Sigma \begin{bmatrix} | & & | \\ v_1 & \dots & v_n \\ | & & | \end{bmatrix}$$; where $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_n^2$ by construction

(eliminates all correlation)

- to reduce dimensionality, we project data onto-subspace defined by directions w/ highest signal ($i \leq k$ principle directions), the projection can be expressed as:



$$x = x^{\parallel} + x^{\perp} = \alpha_1 v_1 + \dots \alpha_k v_k + \dots + \alpha_n v_n$$

$$= \langle x, v_1 \rangle v_1 + \dots \langle x, v_k \rangle v_k + \dots \langle x, v_n \rangle v_n$$

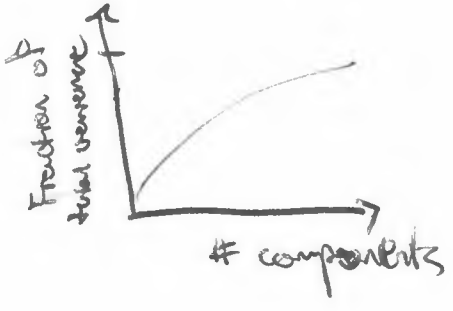
$$\Rightarrow \underset{m \times k}{X_{PCA}} = \underset{m \times n}{X} \underset{n \times k}{\begin{bmatrix} | & & | \\ v_1 & \dots & v_k \\ | & & | \end{bmatrix}}$$

notes

- usually use truncated SVD to obtain eigenvalues / eigenvectors of X^TX since SVD is more numerically stable and ~~taking~~ computing eigenvectors of an $n \times n$ matrix might be daunting

- usually need to center (since we are looking for a subspace that goes through the data) and scale to omit variance (since 1 dimension could have high variance purely due to units).

- can get a handle on # components to use by looking at



uses

- visualization in 2D
- reducing # predictors for a supervised method (or even in a clustering method)
- data compression
- can eliminate noisy dimensions w/ no meaningful signal.

other ways of interpreting PCA

- a plane that goes through data w/ minimal sum of distances to plane.

→

- a low rank matrix that is as close as possible to X:

$$X_{pca} = \underset{\substack{\tilde{X} \in \mathbb{R}^{m \times n} \\ \text{rank}(\tilde{X}) = k}}{\operatorname{argmin}} \|\tilde{X} - X\|_F^2$$

- 1^{st} PC = direction w/ max variance
  2^{nd} PC = " " " orthogonal to 1^{st}
  3^{rd} PC = " " " " " previous
  ⋮

(Follows from SVD)