Random Projection Background:

Idea: Points in a sufficiently high dim vector space can be projected onto a lower dim space while maintaining distances.

Setup: Have $X \in \mathcal{M}^{d \times N}$, matrix of N d dim. data points. Want to project onto k ($k \ll d$) space.

Method: Let $R \in \mathcal{M}^{k \times d}$ be a random matrix with columns of unit length:

$$R X \in \mathcal{M}^{k \times N} \text{ is our projection}$$

$$\begin{pmatrix} \uparrow & & \uparrow \\ \vec{r}_1 & \cdots & \vec{r}_d \\ \downarrow & & \downarrow \end{pmatrix} \begin{pmatrix} \uparrow & & \uparrow \\ x_1 & \cdots & x_n \\ \downarrow & & \downarrow \end{pmatrix}$$

Expectation Maximization (EM) Notes

side note: generative model for X to Y is a joint dist on X & Y.

BIH

Setup: Have some data with K clusters, and want to make a generative (Gaussian) model for each cluster. We need EM to get the parameters of these models.

ex. $x_1, \ldots, x_n$ 1-d observations with K=2 clusters.
Want to get μ, σ^2 for both clusters.
Note this is trivial if we know the clusters since we can directly compute mean & variance.
If we don't know the clusters but know the parameters, we could assign each point to whichever dist was most likely to produce it.

EM Alg:

1. Start w/ 2 randomly placed gaussians $(\mu_a, \sigma_a^2), (\mu_b, \sigma_b^2)$
2. For each point, assign to a group
3. Adjust parameters based on points in each group.
4. Repeat from 2 till stable.

Multivariate case:

Data w/ d attributes from K sources.

$d \times d$

1. Randomly initialize $(\mu_1, \Sigma_1), \dots, (\mu_K, \Sigma_K)$
2. Group all points based on the normals.
3. Readjust parameters.
4. Repeat from 2 till convergence.

2 is the expectation step

3 is the maximization step.

Agglomerative Clustering:

Start w/ n clusters $c_1, \dots, c_n$ each containing 1 point.

While there are $> K$ clusters, merge the most similiar clusters based on the similiarity matrix.

$$\text{sim}(c_i, c_j) = \min_{x_i \in c_i,\, x_j \in c_j} P_{ij}$$

ie $\text{sim}(c_i, c_j)$ = the similiarity of the 2 most dissimiliar points from clusters i & j. Might be better to do:

$$\text{sim}(c_i, c_j) = \sum_{x_i \in c_i} \sum_{x_j \in c_j} P_{ij}$$