

$$\text{total var}_{(A)} = \text{Sum Squares}_{(A)} = \sum_{x \in A} \|x - \mu_A\|^2$$

ward  $d(A, B) = \text{increase in total var}$   
 by linking  $AB = \text{total var}(AB) - \text{total var}(A) - \text{total var}(B)$

$$= \sum_{x \in AB} \|x - \mu_{AB}\|^2 - \sum_{x \in A} \|x - \mu_A\|^2 - \sum_{x \in B} \|x - \mu_B\|^2$$

$$= \sum_{x \in A} \|x - \mu_{AB}\|^2 + \sum_{x \in B} \|x - \mu_{AB}\|^2 - \sum_{x \in A} \|x - \mu_A\|^2 - \sum_{x \in B} \|x - \mu_B\|^2$$

$$\sum_{x \in A} \|x - \mu_{AB}\|^2 = \sum_{x \in A} \|x - \mu_A + \frac{|B|}{|A+B|} (\mu_A - \mu_B)\|^2 =$$

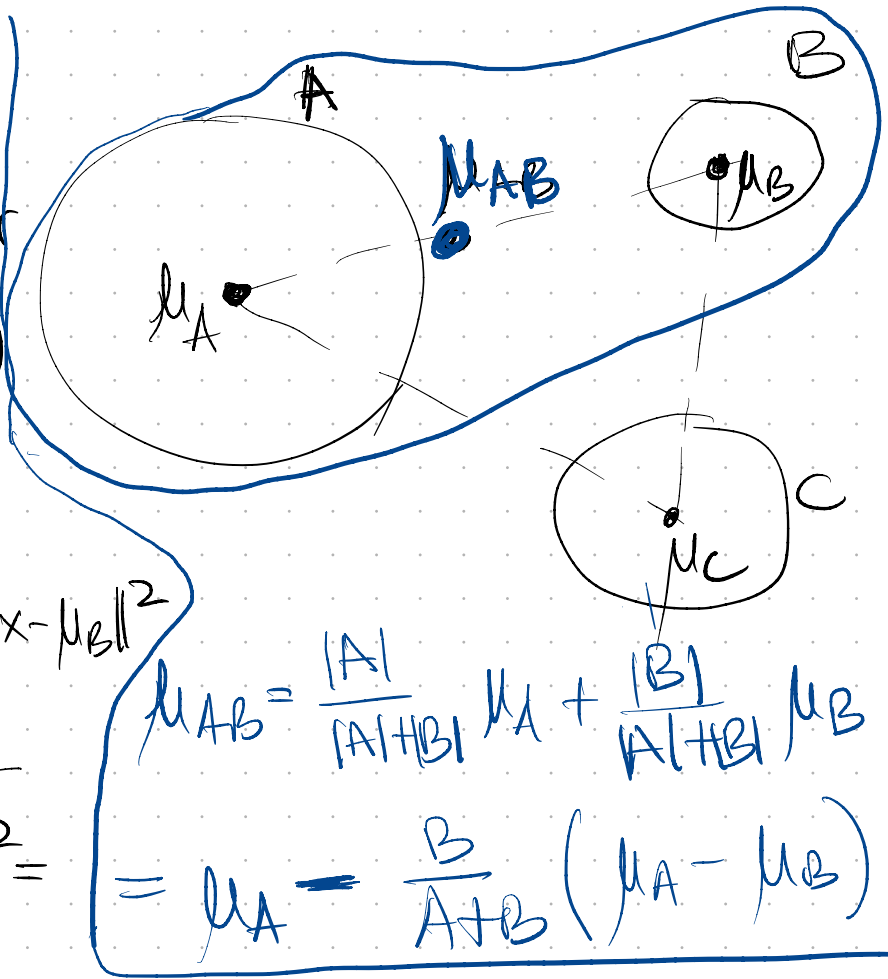
$$= \sum_{x \in A} \|x - \mu_A\|^2 + \frac{|B|^2}{(|A+B|)^2} \|\mu_A - \mu_B\|^2 \cdot |A| + 2 \frac{|B|}{|A+B|} \sum_{x \in A} (x - \mu_A) \cdot (\mu_A - \mu_B) \stackrel{=0}{\Rightarrow}$$

$$= \text{total var}(A) + \frac{|B|^2 |A|}{(|A+B|)^2} \|\mu_A - \mu_B\|^2$$

Similar:  $\sum_{x \in B} \|x - \mu_{AB}\|^2 = \text{total var}(B)$

$$+ \frac{|A|^2 |B|}{(|A+B|)^2} \|\mu_A - \mu_B\|^2$$

$$d(A, B) = \frac{|B|^2 |A|}{|A+B|} \|\mu_A - \mu_B\|^2 + \frac{|A|^2 |B|}{|A+B|} \|\mu_A - \mu_B\|^2 = \frac{|A| |B|}{|A+B|} \|\mu_A - \mu_B\|^2 = \frac{|A| |B|}{|A+B|} \|\mu_A - \mu_B\|^2$$



$(A+B) d(A, B) = AB \|\mu_A - \mu_B\|^2$  How to UPDATE dist when  $\text{join}(A, B) \Rightarrow AB$

$$(A+B+C) d(AB, C) = (A+B) \cdot C \|\mu_{AB} - \mu_C\|^2 = (A+B)C \left\| \frac{A}{A+B} \mu_A + \frac{B}{A+B} \mu_B - \mu_C \right\|^2$$

$$= \frac{C}{A+B} \|A(\mu_A - \mu_C) + B(\mu_B - \mu_C)\|^2$$

$$= \frac{C}{A+B} \cdot A^2 \|\mu_A - \mu_C\|^2 + \frac{C}{A+B} \cdot B^2 \|\mu_B - \mu_C\|^2 + 2 \frac{CAB}{A+B} (\mu_A - \mu_C)(\mu_B - \mu_C)$$

$$= \boxed{AC \|\mu_A - \mu_C\|^2} - AC \cdot \frac{B}{A+B} \|\mu_A - \mu_C\|^2 + \boxed{BC \|\mu_B - \mu_C\|^2} - BC \frac{A}{A+B} \|\mu_B - \mu_C\|^2$$

$$= (A+C) d(A, C) + (B+C) d(B, C) - \frac{CAB}{A+B} \left( \|\mu_A - \mu_C\|^2 + \|\mu_C - \mu_B\|^2 + 2(\mu_A - \mu_C)(\mu_C - \mu_B)^T \right)$$

$$= (A+C) d(A, C) + (B+C) d(B, C) - \frac{CAB}{A+B} \|\mu_A - \mu_C + \mu_C - \mu_B\|^2$$

$$= (A+C) d(A, C) + (B+C) d(B, C) - C \cdot d(A, B)$$

So update word dist on  $\text{join}(A, B) \Rightarrow AB$ :

$$d(AB, C) = \frac{A+C}{A+B+C} d(A, C) + \frac{B+C}{A+B+C} d(B, C) - \frac{C}{A+B+C} d(A, B)$$