EM Algorithm (Cont'd)

- Recap
- Generic EM Algorithm
- Apply to GMMs (Message: GMM algo is EM)
- Start Factor Analysis

We examined GMMs

Given: $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^d$ and $k > 0$

Do: find $P(z^{(i)} = j)$ for $i = 1 \dots n$, $j = 1 \dots k$

"A soft assignment"

We called $z^{(i)}$ a latent variable not observed directly

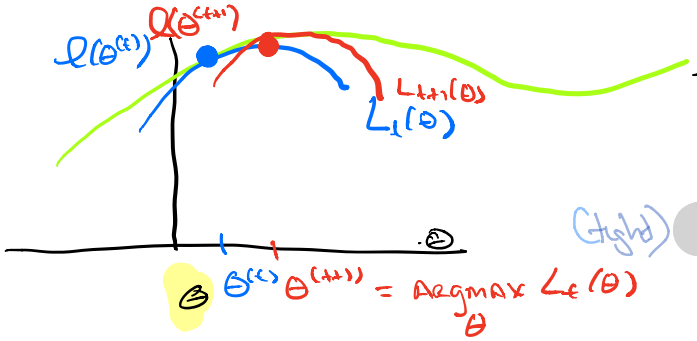
Today: Generalize model, relate it to MLE, EM Algorithm

$$\ell(\theta) = \sum_{i=1}^{n} \log P(x^{(i)}; \theta)$$

(θ: parameters)

$$= \sum_{i=1}^{n} \log \sum_{z} P(x^{(i)}, z; \theta)$$

Picture of Generic Algorithm



1. $L_t(\theta) \le \ell(\theta)$ (lowerbound)
2. $L_t(\theta^{(t)}) = \ell(\theta^{(t)})$ (tight)

Hope: $L_t(\theta)$ easy to optimize

Rough Algorithm

(E-step) 1. Find $L_t(\theta)$ given $\theta^{(t)}$

(M-STEP) 2. $\theta^{(t+1)} = \arg\max_\theta L_t(\theta)$

WE go from $\log \sum_z P(x^{(i)}, z; \theta)$ from fix a particular i

$$\log \sum_z P(x^{(i)}, z; \theta) = \log \sum_z \frac{Q^{(i)}(z)}{Q^{(i)}(z)} P(x^{(i)}, z; \theta)$$

Pick $Q^{(i)}(z)$ s.t. $\sum_z Q^{(i)}(z) = 1$ & $Q^{(i)}(z) \geq 0 \ \forall z$. (*)

$$= \log \mathbb{E}_{z \sim Q^{(i)}} \left[\frac{P(x^{(i)}, z; \theta)}{Q^{(i)}(z)} \right]$$

RECALL: $\log(\mathbb{E}(x)) \geq \mathbb{E}[\log(x)]$. GREAT!

$$\geq \mathbb{E}_{z \sim Q^{(i)}} \left[\log \frac{P(x^{(i)}, z; \theta)}{Q^{(i)}(z)} \right]$$

$$= \sum_z Q^{(i)}(z) \log \frac{P(x^{(i)}, z; \theta)}{Q^{(i)}(z)}$$

NB: this holds for ANY CHOICE $Q^{(i)}$ that SATISFIES (*)

Property 1

Pick $Q^{(i)}(z)$ SATISFY Property 2?

$$\ell(\theta^{(t)}) = L_t(\theta^{(t)})$$ "tight"

WANT $\log \sum_z P(x, z; \theta) = \sum_z Q^{(i)}(z) \log \frac{P(x^{(i)}, z; \theta)}{Q^{(i)}(z)}$ (†)

TAKE: $Q^{(i)}(z) = P(z \mid x^{(i)}; \theta^{(t)})$

- See piazza post 822 for more discussion.

† $P(x^{(i)}, z; \theta^{(t)}) = P(x^{(i)}; \theta^{(t)})$ does not depend

$$P(z \mid x^{(i)}; \theta^{(t)})$$

$$= \log P(x^{(i)}; \theta^{(t)}) \sum_z Q_i^{(t)}(z) \to 1$$

on z!

NOTE $Q^{(i)}(z)$ VARIES for every point.

WE CALL $ELBO(x, Q; \theta) = \sum_z Q(z) \log \frac{P(x, z; \theta)}{Q(z)}$

WE'VE SHOWN $\ell(\theta) \geq \sum_{i=1}^{n} ELBO(x^{(i)}, Q^{(i)}; \theta)$

$$\ell(\theta^{(t)}) = \sum_{i=1}^{n} ELBO(x^{(i)}, Q^{(i)}; \theta^{(t)})$$

↑ with our selection of $Q^{(i)}$

WARM UP: Mixture of Gaussians. EM RECOVERS AD HOC Algorithm

$$P(x^{(i)}, z^{(i)}) = P(x^{(i)} \mid z^{(i)}) P(z^{(i)})$$

"IN clusters" $z^{(i)} \sim \text{Multinomial}(\phi)$ $\phi_i \geq \sum \phi_i = 1$

"cluster means" $x^{(i)} \mid z^{(i)} = j \sim N(\mu_j, \Sigma_j^2)$

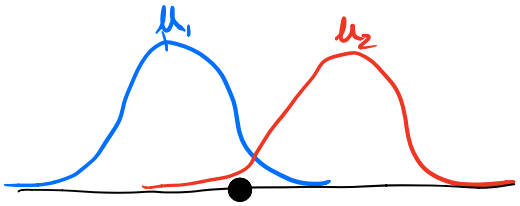
$z^{(i)}$ LATENT VARIABLE.

WHAT IS EM HERE?

$$Q^{(i)}(j) = P(z^{(i)} = j \mid x^{(i)}; \theta)$$

"Guess" IN TERMS of $\sim P(x^{(i)} / z^{(i)} = j; \theta)$

VIA BAYES Rule



M-STEP

MAX ϕ, μ, Σ

$$\sum_{i=1}^{n} \underbrace{\sum_{z^{(i)}} Q^{(i)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}}_{f_i(\theta)}$$

$$w_j^{(i)} = Q^{(i)}(j)$$

$$f_i(\theta) = \sum_j w_j^{(i)} \log \frac{\frac{1}{2\pi |\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right\} \phi_j}{w_j^{(i)}}$$

(MEANS, COVARIANCE)

$$\nabla_{\mu_j} f_i(\theta) = \sum_j \nabla_{\mu_j} w_j^{(i)} \log \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right\}$$

derivative $= \sum_j -\frac{1}{2} w_j^{(i)} \Sigma_j^{-1} (x^{(i)} - \mu_j)$

Assume Σ_j^{-1} exists (full rank)

Setting to 0 $\quad \mu_j = \frac{\sum w_j^{(i)} x^{(i)}}{\sum w_j^{(i)}}$ — old GMM update

$$f_i(\theta) = \sum_j w_j^{(i)} \log \frac{\frac{1}{2\pi |\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right\} \phi_j}{w_j^{(i)}}$$

(MEANS, COVARIANCE)

$$\nabla_{\phi_j} f_i(\theta) = \sum_j \nabla_{\phi_j} w_j^{(i)} \log \phi_j \qquad \phi_j \text{ is constrained}$$

$\sum \phi_j = 1$. constraint

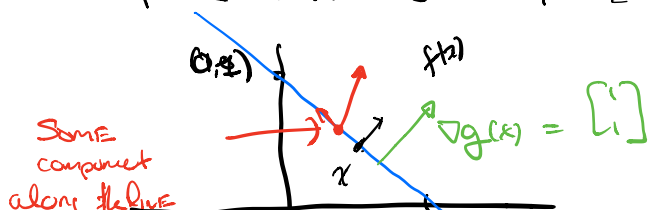
Lagrangian

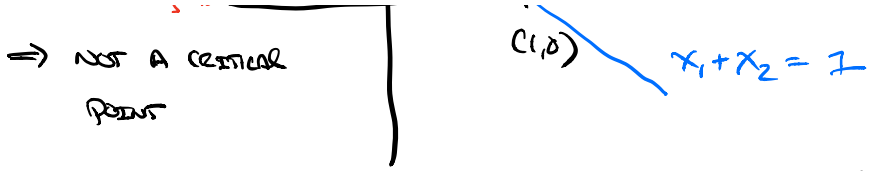
$$\Rightarrow \nabla_{\phi_j} f_i(\theta) = \sum_j \nabla_{\phi_j} \left(w_j^{(i)} \log \phi_j + \lambda \left(\sum_{m=1}^{k} \phi_j - 1 \right) \right)$$

Detour: want to find critical points of

$f: \mathbb{R}^2 \to \mathbb{R}$ s.t. $x_1 + x_2 = 1$ $\quad (g(x) = 1)$

$g(x) = x_1 + x_2$





If z is a critical point then $\nabla f(z)$ is parallel to $\nabla g(z)$.

$$\nabla f(z) = -\lambda \nabla g(z)$$

(λ exists)

$$L(x, \lambda) = f(x) + \lambda(g(x) - 1)$$

$$\nabla_x L = \nabla f(x) + \lambda \nabla g(x) = 0 \rightarrow \text{Parallel Condition}$$

$$\nabla_\lambda L = g(x) - 1 = 0 \rightarrow \text{Constraint is satisfied}$$

$$\nabla_{\phi_j} \left(\sum_{i=1}^{n} w_j^{(i)} \log \phi_j + \lambda \left(\sum_{m=1}^{k} \phi_m - 1 \right) \right)$$

$$= \sum_{i=1}^{n} \frac{w_j^{(i)}}{\phi_j} + \lambda = 0 \Rightarrow \phi_j = -\frac{1}{\lambda} \sum_{i=1}^{n} w_j^{(i)}$$

Since $\sum_j \phi_j = 1$ $\Rightarrow 1 = \sum_j \phi_j = -\frac{1}{\lambda} \sum_i \sum_j w_j^{(i)} = -\frac{n}{\lambda}$ (the inner sum equals 1)

$$\Rightarrow \phi_j = \frac{1}{n} \sum_{i=1}^{n} w_j^{(i)}$$

MESSAGE: EM Recovers Gauss Algorithm

NB: z is discrete Here, can Replace Sums w/ Integrals

Factor Analysis

Many fewer points than dimensions "$n \ll d$"

$n \gg d$ Gauss

How does this happen?

Place sensors all over canada Record temperature

@ 1000s of locations on 1000s

But only record for 30 days ($N < d$)

Want to fit a density but seems hopeless...

Key idea Assume there is some latent r.v. that is not complex and explains behavior

1st lets see the problems w/ fitting a single Gaussian...

Given: $x^{(1)} \cdots x^{(n)} \in \mathbb{R}^d$ (n: smaller, d: huge)

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \rightarrow \text{seems ok...}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

$\text{rank}(\hat{\Sigma}) \leq n < d$ not full rank

$|\hat{\Sigma}| = 0$ (not defined)

$$P(x;\mu,\Sigma) = \frac{1}{2\pi |\hat{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \hat{\Sigma}^{-1}(x-\mu)\right\}$$

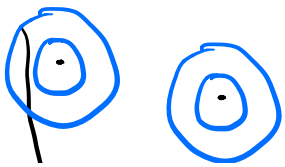
($|\hat{\Sigma}|$: zero; $\hat{\Sigma}^{-1}$: not defined)

We will fix issues by examining these simpler models

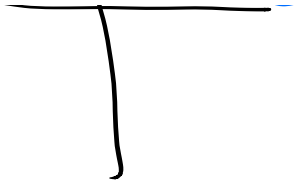
$\Rightarrow$ Spoiler: Combine all these as final model

Building Block 1

Suppose independent & identical r.v.



Covariance are circles



$\Sigma_i = \sigma^2 I$

↑ scalar.

$\Sigma_i \in \mathbb{R}^{d \times d}$

$= \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix}$

$$\min_{\mu, \Sigma} \sum_{i=1}^{n} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu) + \log |\Sigma|$$

(Assumption $\Sigma_i = \sigma^2 I$)

$$= \sigma^{-2} \sum_{i=1}^{n} (x^{(i)} - \mu)^T (x^{(i)} - \mu) + \log \sigma^{2d}$$

$z = \sigma^2$

$$= z^{-1} \underbrace{\left(\sum_{i=1}^{n} \|x^{(i)} - \mu\|^2 \right)}_{c} + d \log z$$

$$\nabla_z = -z^{-2} c + n \frac{d}{z} = 0 \qquad -c = dz$$

$$z = \frac{c}{nd}$$

$$\sigma^2 = \frac{c}{nd}$$

$$|\Sigma| = \left| \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix} \right| = \sigma^{2d}$$