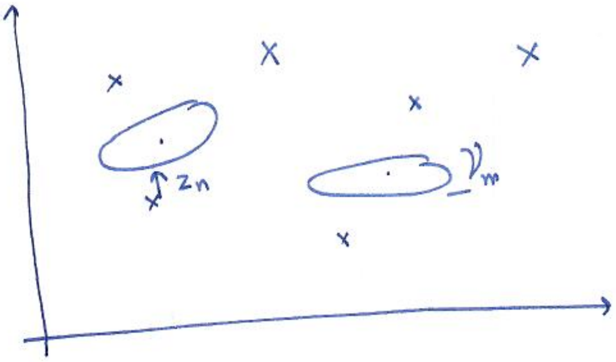
Lec 10: 03.06.2022



X - known $\underline{X}_n$

$\underline{\nu}$ - unknown parameters

$\underline{z}_n$ - unknown mode measurements

It is difficult to exactly or partially determined the X and $\underline{\nu}$ together i.e. coupled. However, if we decouple them then determining them becomes a relatively easy solution. This is where EM (Expectation Maximization) Algorithm comes into play.

4-35

4-36

We have bad/missing data $\underline{z}$ and that is to be eliminated.

In the eq.^{n}

$$\underline{Q}(\underline{\theta}) = \int \ln p(\underline{x}, \underline{z}, \underline{\theta})\, p(\underline{z} \mid \underline{x}, \underline{\theta})\, dz$$

we need to eliminate

Hence we do this step $p(\underline{z} / \underline{x}, \underline{\theta})$

4-37

Apply the general EM algorithm to special GMM model.

4-38, 4-39

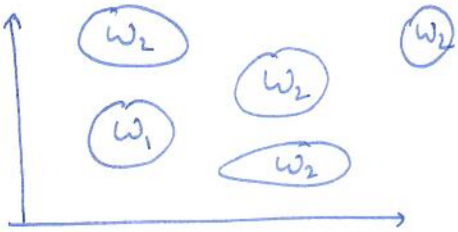
4-40

MATLAB : gm distribution

GMM classifier :

c GMM models

- one GMM $p(\underline{x} \mid \omega_j ; \underline{\vartheta}_j)$ per class ω_j $1 \leq j \leq c$



No of modes M_j per class ω_j may not be same for each class.

- run EM algorithm c times
- then have a better idea

4-41

4-42

GMM classifier contains

Gaussian classifier : $M_j = 1 \quad \forall j$

and naive " " : $M_j = 1$, $\underline{\underline{C}}_j$ diagonal $\forall j$

as special cases

Q How to choose the no of modes?

Model order estimation :

But model orders are M_j are usually unknown

→ need order estimation : Estimate

M_j (and $\underline{\vartheta}_j$) from data

↑ discrete valued ↑ continuous valued

2 popular criteria :

Akaike Information Criteria (AIC)

Bayesian " " (BIC)

Q How to know which one to use?

Intuition + perform over all possibilities and then find the one which gives the best result.

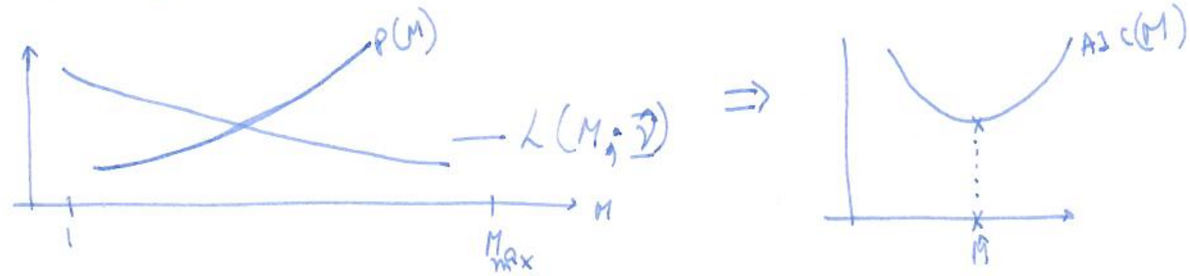
For one GMM:

- Try EM for $M = 1 \cdots M_{max}$
- Compute $AIC(M) = -\underset{\uparrow}{\mathcal{L}}(M; \hat{\underline{\theta}}) + \overset{\text{penalty term}}{\overset{\downarrow}{p}}(M)$

  log likelihood $\approx \ln p(\underline{X}, \underline{Z}; \underline{\theta}^{\circ})$ in EM

- $\hat{M} = \arg\min\limits_{M} (AIC(M))$

Q Why penalty term



Reason for decay - if we increase M, then we have more parameters $\rightarrow$ better fit of data to data $\rightarrow \mathcal{L}(M; \hat{\underline{\theta}}) \uparrow$ $\rightarrow$ overfitting

Hence we add $p(M)$ to avoid overestimation of M

Difference b/w AIC and BIC:

Choice of penalty term $p(M)$

4-43, 4-49

4-1 → Now we will talk about density estimation (Parzen window)

4.4 Density estimation

Idea

- Non-parametric estimate of $p(\underline{x} \mid \omega_j)$
- Apply Bayesian decision theory

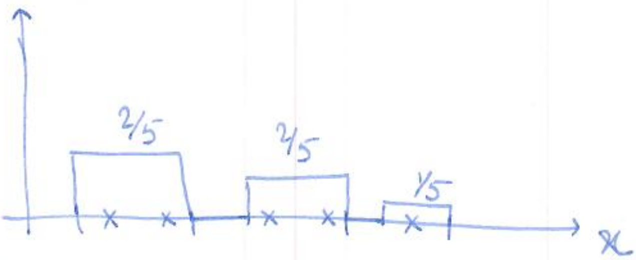
eg Parzen window method

For one class:

Given: N iid sample of $\underline{x}_n$ from an unknown $p(\underline{x})$

Desired: PDF estimate $\hat{p}(\underline{x})$

Simples: Normalized histogram



Total area under curve should be 1, however this is not smooth. To smooth this, we need kernels.

Better: Kernel based PDF

Kernel: in $\mathbb{R}^d$: a mathematical PDF-like function $\phi(x)$

- $\phi(\underline{x}) \geq 0 \;\; \forall \underline{x}$
- $\int \phi(\underline{x})\, d\underline{x} = 1$

Gaussian kernel : $N(\underline{0}, \underline{\underline{I}})$

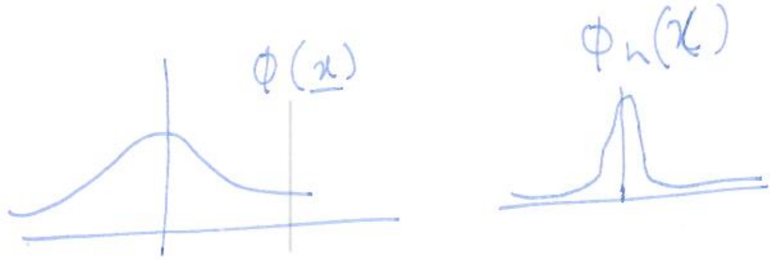
$$\phi(\underline{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} \|\underline{x}\|^2\right),$$

has a fixed width (SD)

Scaled kernel :

$$\phi_h(\underline{x}) = \frac{1}{h^d}\phi\left(\frac{1}{h}\underline{x}\right) \geq 0, \quad \int \phi_h(\underline{x})\, d\underline{x} = 1$$

$h \rightarrow$ bandwidth parameter $h > 0$



$h \downarrow$ = narrow kernel $\rightarrow$ detailed, however sometimes too much details in PDF

$h \uparrow$ = broad kernel $\rightarrow$ less detailes.

PDF estimate :

$$\hat{p}(\underline{x}) = \frac{1}{N} \sum_{n=1}^{N} \phi_h(\underline{x} - \underline{x}_n) \geq 0, \quad \int \hat{p}(\underline{x})\, d\underline{x} = 1$$

4-45

4-46