DENSITY ESTIMATION

Estimating an Unknown **P**robability **D**ensity **F**unction

Th&Ko §2.5 / DHS §3.1-3.4

Density Estimation

- Parametric techniques
  - Maximum Likelihood
  - Maximum A Posteriori
  - Bayesian Inference
  - **G**aussian **M**ixture **M**odels (GMM)
    - EM-Algorithm
- Non-parametric techniques
  - Histogram
  - Parzen Windows
  - k-nearest-neighbor rule

Estimation of an unknown PDF

Task:

Estimate the parameters $\underline{\theta}$ of a pdf with known structure from a set of data X.

(*e.g.* $p(\underline{x}) \equiv p(\underline{x};\underline{\theta})$ is known to be Gaussian $N(\underline{\mu},\Sigma)$, with unknown $\underline{\theta} = [\mu_1,...,\mu_l,\sigma_{1,1},....,\sigma_{l,l}]^T$)

Formal:

Let $\underline{x}_1$, $\underline{x}_2$,...., $\underline{x}_N$ known and independent (i.i.d.) $X = \{\underline{x}_1, \underline{x}_2, ...\underline{x}_N\}$

Let $p(\underline{x})$ be known within a vector parameter $\underline{\theta}$: $p(\underline{x}) \equiv p(\underline{x};\underline{\theta})$

$p(X;\underline{\theta}) \equiv p(\underline{x}_1, \underline{x}_2, ...\underline{x}_N;\underline{\theta})$

$= \prod_{k=1}^{N} p(\underline{x}_k;\underline{\theta})$ which is known as the likelihood of $\underline{\theta}$ *w.r.to* X

Estimation of an unknown PDF

- **Maximum Likelihood Method (ML)**
  - Search for the parameter Θ_{ML} that maximizes

$$\underline{\hat{\theta}}_{\mathrm{ML}} := \arg\max_{\underline{\theta}} \prod_{k=1}^{N} p(\underline{x}_k;\underline{\theta})$$

  - a necessary condition $\dfrac{\partial \prod_{k=1}^{N} p(\underline{x}_k;\underline{\theta})}{\partial(\underline{\theta})} = \underline{0}$
  - since ln is monotonic we can write the log-likelihood

$$L(\underline{\theta}) \equiv \ln p(X;\underline{\theta}) = \sum_{k=1}^{N} \ln p(\underline{x}_k;\underline{\theta})$$

$$\underline{\hat{\theta}}_{ML} : \frac{\partial L(\underline{\theta})}{\partial(\underline{\theta})} = \sum_{k=1}^{N} \frac{1}{p(\underline{x}_k;\underline{\theta})} \frac{\partial p(\underline{x}_k;\underline{\theta})}{\partial(\underline{\theta})} = \underline{0}$$

Properties of Maximum Likelihood Method

If there is a true value Θ_0, the ML estimate Θ_{ML} has the following properties (no proofs):

a) Θ_{ML} is *asymptotically unbiased* and *converges in the mean*.

$$p(\underline{x}) = p(\underline{x};\underline{\theta}_0), \text{ then } \quad \lim_{N\to\infty} E[\hat{\underline{\theta}}_{ML}] = \underline{\theta}_0$$

b) Θ_{ML} is *asymptotically consistent* and *converges in probability*.

$$\lim_{N\to\infty} \text{prob}\left\{\left\|\hat{\underline{\theta}}_{ML} - \underline{\theta}_0\right\| \le \varepsilon\right\} = 1$$

c) Θ_{ML} is *asymptotically consistent* and *converges in mean square*.

$$\lim_{N\to\infty} E\left[\left\|\hat{\underline{\theta}}_{ML} - \underline{\theta}_0\right\|^2\right] = 0$$

ML Example 1:

$p(\underline{x})$: $N(\underline{\mu},\Sigma)$: $\underline{\mu}$ unknown

$\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N$

$$p(\underline{x}_k) \equiv p(\underline{x}_k;\underline{\mu}) = \frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\underline{x}_k - \underline{\mu})^T \Sigma^{-1}(\underline{x}_k - \underline{\mu})\right)$$

$$L(\underline{\mu}) = \ln \prod_{k=1}^{N} p(\underline{x}_k;\underline{\mu}) = \sum_{k=1}^{N}\left(C - \frac{1}{2}(\underline{x}_k - \underline{\mu})^T \Sigma^{-1}(\underline{x}_k - \underline{\mu})\right)$$

$$\frac{\partial L(\underline{\mu})}{\partial(\underline{\mu})} \equiv \begin{bmatrix} \frac{\partial L}{\partial \mu_1} \\ \frac{\partial L}{\partial \mu_2} \\ . \\ . \\ \frac{\partial L}{\partial \mu_l} \end{bmatrix} = \sum_{k=1}^{N} \Sigma^{-1}(\underline{x}_k - \underline{\mu}) = \underline{0}$$

$$\Rightarrow \quad \underline{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{N} \underline{x}_k$$

Remember : if $A = A^T$ $\Rightarrow$ $\frac{\partial(\underline{\alpha}^T A \underline{\alpha})}{\partial \underline{\alpha}} = 2A\underline{\alpha}$

ML Example 2 :

$$p(\underline{x}):\ N(\underline{\mu},\Sigma):\ \underline{\mu}\ ,\ \Sigma=\sigma^2 I \quad \text{unknown}$$

$$\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_N \in \mathbb{R}^l$$

$$p(\underline{x}_k) \equiv p(\underline{x}_k;\underline{\mu},\sigma^2) = \frac{1}{(2\pi^l\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}(\underline{x}_k-\underline{\mu})^2\right)$$

$$L(\underline{\mu},\sigma^2) = \ln\prod_{k=1}^{N} p(\underline{x}_k;\underline{\mu},\sigma^2) = \sum_{k=1}^{N} -\frac{1}{2}\ln(2\pi^l\sigma^2) \quad - \quad \sum_{k=1}^{N}\frac{1}{2\sigma^2}(\underline{x}_k-\underline{\mu})^2$$

$$\frac{\partial L(\underline{\theta})}{\partial(\underline{\theta})} \equiv \begin{bmatrix} \frac{\partial L}{\partial \underline{\mu}} \\ \frac{\partial L}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{N}\frac{1}{\sigma^2}(\underline{x}_k-\underline{\mu}) \\ -\sum_{k=1}^{N}\frac{1}{2\sigma^2}+\sum_{k=1}^{N}\frac{1}{2\sigma^4}(\underline{x}_k-\underline{\mu})^2 \end{bmatrix} = \underline{0}$$

$$\Rightarrow \quad \underline{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{N}\underline{x}_k$$

$$\Rightarrow \quad \sigma^2_{ML} = \frac{1}{N}\sum_{k=1}^{N}(\underline{x}_k-\underline{\mu})^2$$

However, the true $\underline{\mu}$ is unknown, therefore we have to use $\underline{\mu}_{ML}$

Maximum Likelihood estimates are only asymptotically unbiased, so N should be large enough !

ML Example (3) :

$$p(\underline{x}):\ N(\underline{\mu},\Sigma):\ \underline{\mu}\ ,\ \Sigma=\sigma^2 I \quad \text{unknown}$$

$$\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_N \in \mathbb{R}^l$$

$$p(\underline{x}_k) \equiv p(\underline{x}_k;\underline{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{l}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\underline{x}_k-\underline{\mu})^T\Sigma^{-1}(\underline{x}_k-\underline{\mu})\right)$$

$$L(\underline{\mu},\sigma^2) = \ln\prod_{k=1}^{N} p(\underline{x}_k;\underline{\mu},\sigma^2) = -\frac{N}{2}\ln(2\pi^l|\Sigma|) - \frac{1}{2}\sum_{k=1}^{N}(\underline{x}_k-\underline{\mu})^T\Sigma^{-1}(\underline{x}_k-\underline{\mu})$$

$$\frac{\partial L(\underline{\theta})}{\partial(\underline{\theta})} \equiv \begin{bmatrix} \frac{\partial L}{\partial \underline{\mu}} \\ \frac{\partial L}{\partial \Sigma_{11}} \\ . \\ \\ \frac{\partial L}{\partial \Sigma_{ll}} \end{bmatrix} = \ldots\ldots\ldots \ldots = \underline{0}$$

$$\Rightarrow \quad \underline{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{N}\underline{x}_k$$

$$\Rightarrow \quad \Sigma_{ML} = \frac{1}{N}\sum_{k=1}^{N}(\underline{x}_k-\underline{\mu}_{ML})(\underline{x}_k-\underline{\mu}_{ML})^T$$

Maximum Likelihood estimates are only asymptotic unbiased, so we need a large N !

WARNING: An unbiased estimator is also no guarantee for a correct result!!

Estimation of an unknown PDF

❖ **Maximum Aposteriori Probability estimation (MAP)**

- (In ML $\underline{\theta}$ was considered as a parameter)
  Here we shall look at $\underline{\theta}$ as a random vector described by a pdf $p(\underline{\theta})$, assumed to be known
- Given $\mathcal{X} = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$

  Compute the maximum of $p(\underline{\theta}|X)$
- From Bayes theorem

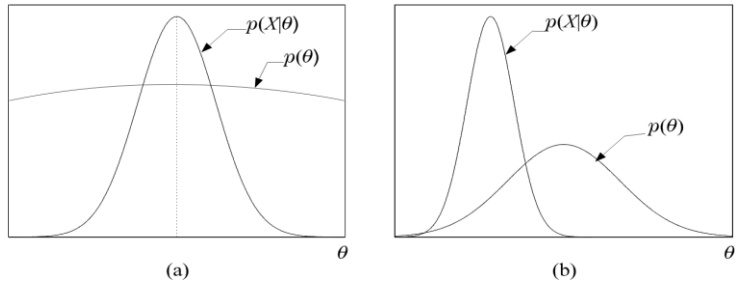
$$p(\underline{\theta})\,p(X|\underline{\theta}) = p(X)\,p(\underline{\theta}|X) \quad \text{or} \quad p(\underline{\theta}|X) = \frac{p(\underline{\theta})\,p(X|\underline{\theta})}{p(X)}$$

- The method:

$$\hat{\underline{\theta}}_{MAP} = \arg\max_{\underline{\theta}} p(\underline{\theta}|X) \quad \text{or}$$

$$\hat{\underline{\theta}}_{MAP} : \frac{\partial}{\partial \underline{\theta}}\left(P(\underline{\theta})\,p(X|\underline{\theta})\right) = 0$$

If $p(\underline{\theta})$ is uniform or broad enough $\hat{\underline{\theta}}_{MAP} \cong \underline{\theta}_{ML}$



❖ MAP Example:

$p(\underline{x}): N(\underline{\mu}, \Sigma = \sigma^2 I)$, $\underline{\mu}$ unknown

$X = \{\underline{x}_1, ..., \underline{x}_N\}$

$$p(\underline{\mu}) = \frac{1}{(2\pi)^{\frac{l}{2}} \sigma_\mu^l} \exp\left(-\frac{\left\|\underline{\mu} - \underline{\mu}_0\right\|^2}{2\sigma_\mu^2}\right)$$

$$\underline{\hat{\theta}}_{MAP}: \quad \frac{\partial}{\partial \underline{\mu}} \ln\left(\prod_{k=1}^{N} p(\underline{x}_k | \underline{\mu}) p(\underline{\mu})\right) = \underline{0} \quad \text{or} \quad \sum_{k=1}^{N} \frac{1}{\sigma^2}(\underline{x}_k - \underline{\hat{\mu}}) - \frac{1}{\sigma_\mu^2}(\underline{\hat{\mu}} - \underline{\mu}_0) = \underline{0}$$

$$\Rightarrow \quad \underline{\hat{\mu}}_{MAP} = \frac{\underline{\mu}_0 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{k=1}^{N} \underline{x}_k}{1 + \frac{\sigma_\mu^2}{\sigma^2} N}$$

For $\frac{\sigma_\mu^2}{\sigma^2} >> 1$, or for $N \to \infty$

$$\Rightarrow \quad \underline{\hat{\mu}}_{MAP} \cong \underline{\hat{\mu}}_{ML} = \frac{1}{N} \sum_{k=1}^{N} \underline{x}_k$$