

# MULTISCALE SMOOTHING IN SUPERVISED STATISTICAL LEARNING

**ANIL K. GHOSH**

`akghosh@isical.ac.in`

**THEORETICAL STATISTICS AND MATHEMATICS UNIT  
INDIAN STATISTICAL INSTITUTE, KOLKATA**

- **Diabetes data** : Measurements are taken on fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight from chemical diabetic patients, overt diabetic patients and normal people.

- **Diabetes data** : Measurements are taken on fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight from chemical diabetic patients, overt diabetic patients and normal people.
- **Vowel recognition data**: A number of speakers spoke some words formed by 'h' followed by a vowel and then followed by 'd'. Resonant frequencies of a speaker's vocal tract are noted for different vowels.

- **Diabetes data** : Measurements are taken on fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight from chemical diabetic patients, overt diabetic patients and normal people.
- **Vowel recognition data**: A number of speakers spoke some words formed by 'h' followed by a vowel and then followed by 'd'. Resonant frequencies of a speaker's vocal tract are noted for different vowels.
- **Image segmentation data**: Images are taken from one of the following classes : brickface, sky, foliage, cement, window, path, grass. 19 measurements are taken on an image of a region consisting of 9 pixels.

- **Data** :  $(\mathbf{x}_n, c_n), n = 1, 2, \dots, N$ .  
Vector of measurement variables :  $\mathbf{x}_n \in R^d$ ,  
Class labels :  $c_n \in \{1, 2, \dots, J\}$ .

- **Data** :  $(\mathbf{x}_n, c_n), n = 1, 2, \dots, N$ .  
Vector of measurement variables :  $\mathbf{x}_n \in R^d$ ,  
Class labels :  $c_n \in \{1, 2, \dots, J\}$ .
- **Decision rule** :  $d(\mathbf{x}) : R^d \rightarrow \{1, 2, \dots, J\}$

- **Data** :  $(\mathbf{x}_n, c_n), n = 1, 2, \dots, N$ .  
Vector of measurement variables :  $\mathbf{x}_n \in R^d$ ,  
Class labels :  $c_n \in \{1, 2, \dots, J\}$ .
- **Decision rule** :  $d(\mathbf{x}) : R^d \rightarrow \{1, 2, \dots, J\}$
- **Bayes rule** :  $d_B(\mathbf{x}) = \arg \max_j \mathcal{P}(j | \mathbf{x}) = \arg \max_j \pi_j f_j(\mathbf{x})$   
 $f_j(\mathbf{x})$  : density functions,  $\pi_j$  : prior probabilities.

# Parametric and nonparametric classification

---

- **Parametric** : Assumes specific parametric form for  $f$  and uses the training data to estimate its parameters.
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)

# Parametric and nonparametric classification

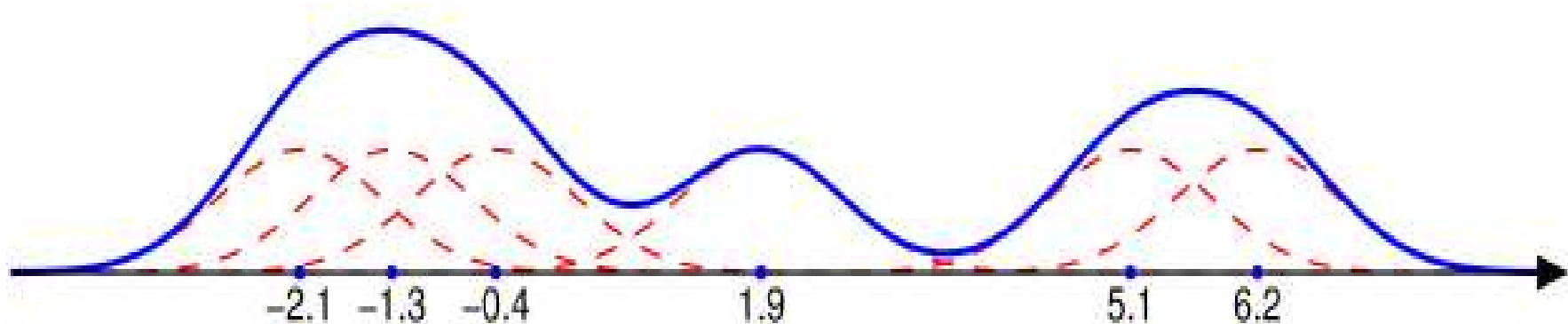
---

- **Parametric** : Assumes specific parametric form for  $f$  and uses the training data to estimate its parameters.
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)
- **Nonparametric** : No specific parametric assumption about the functional form of  $f$ .
  - Kernel Discriminant Analysis (KDA)
  - Nearest Neighbor Classification (NN)

Kernel density estimate/ Parzen window estimate :

$$\hat{f}_{jh_j}(\mathbf{x}) = \frac{1}{n_j} \sum_{k=1}^{n_j} \left[ h_j^{-d} K \left\{ h_j^{-1} (\mathbf{x}_{jk} - \mathbf{x}) \right\} \right]$$

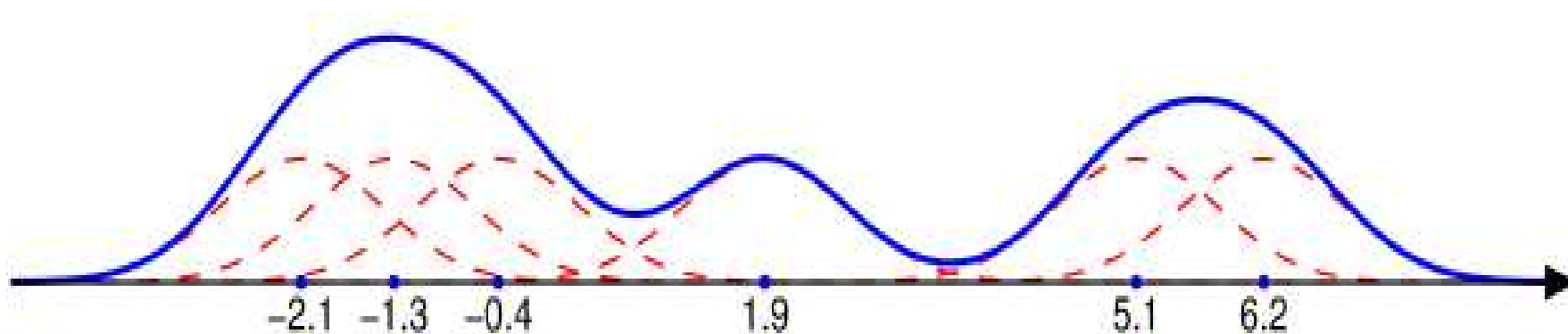
$K$ =kernel function  $h_j$ =bandwidth  $n_j$ =sample size  
 $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \dots, \mathbf{x}_{jn_j}$ =training sample observations.



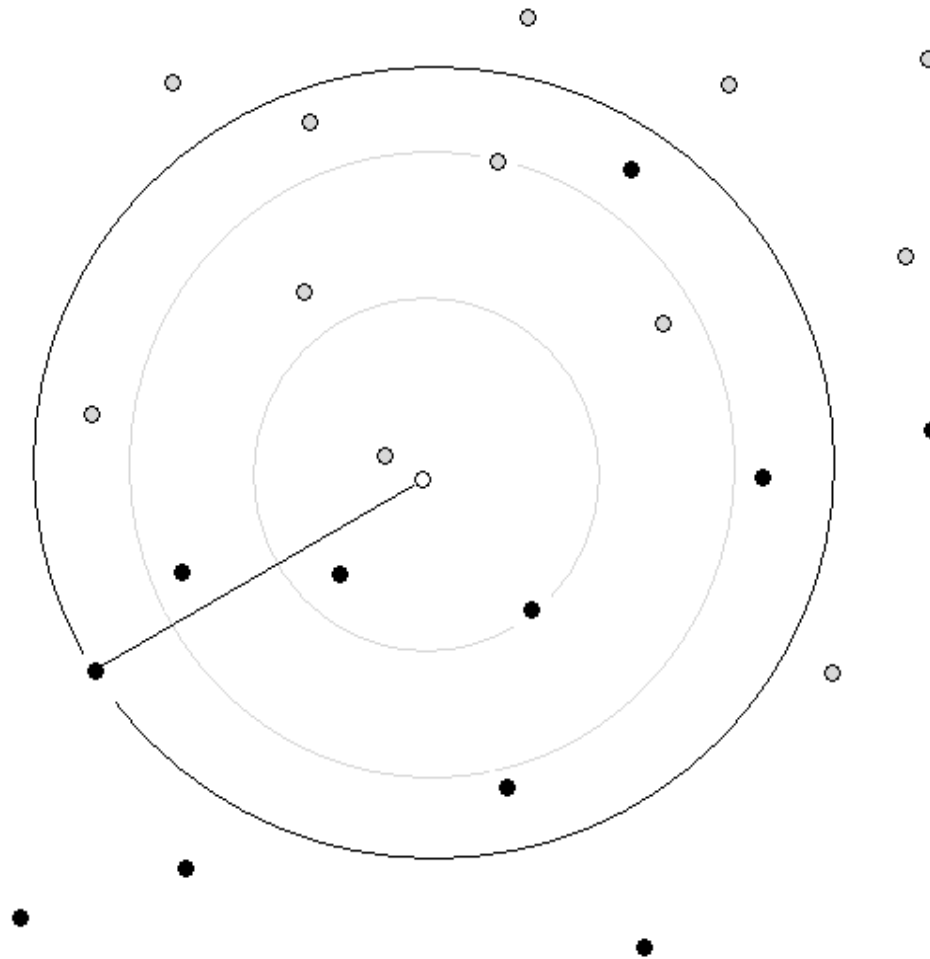
Kernel density estimate/ Parzen window estimate :

$$\hat{f}_{jh_j}(\mathbf{x}) = \frac{1}{n_j} \sum_{k=1}^{n_j} \left[ h_j^{-d} K \left\{ h_j^{-1} (\mathbf{x}_{jk} - \mathbf{x}) \right\} \right]$$

$K$ =kernel function  $h_j$ =bandwidth  $n_j$ =sample size  
 $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \dots, \mathbf{x}_{jn_j}$ =training sample observations.



Classification rule :  $d(\mathbf{x}) = \arg \max_j \pi_j \hat{f}_{jh_j}(\mathbf{x})$



The decision rule depends on the value of  $k$

# Standard methods for smoothing parameter selection

---

- Maximization of the likelihood function (likelihood cross-validation).

# Standard methods for smoothing parameter selection

---

- Maximization of the likelihood function (likelihood cross-validation).
- Minimization of a data based estimate (e.g., bootstrap or cross-validation estimate) of misclassification probability.

# Standard methods for smoothing parameter selection

---

- Maximization of the likelihood function (likelihood cross-validation).
- Minimization of a data based estimate (e.g., bootstrap or cross-validation estimate) of misclassification probability.
- Minimization of Mean Integrated Square Error (MISE) of density estimates (in case of kernel discriminant analysis).

# LIGO Some problems in traditional approach based on a fixed scale of smoothing

---

- Optimum level of smoothing is chosen based on the entire training sample, while a good choice of smoothing parameter may also depend on the observation to be classified.

# LIGO Some problems in traditional approach based on a fixed scale of smoothing

---

- Optimum level of smoothing is chosen based on the entire training sample, while a good choice of smoothing parameter may also depend on the observation to be classified.
- One may like to assess the strength of evidence in favor of different competing class at different scale of smoothing.

# LIGO Some problems in traditional approach based on a fixed scale of smoothing

---

- Optimum level of smoothing is chosen based on the entire training sample, while a good choice of smoothing parameter may also depend on the observation to be classified.
- One may like to assess the strength of evidence in favor of different competing class at different scale of smoothing.
- It allows only one single bandwidth for each population density estimate irrespective of the competing class density.

- 
- No selection of smoothing parameter.

- 
- No selection of smoothing parameter.
  - Simultaneous study of discrimination measures for a wide range of smoothing parameters.

# Discrimination measures (kernel discriminant analysis)

---

- Posterior probability :

$$\mathcal{P}_{h_1, h_2}(1 \mid \mathbf{x}) = \frac{\pi_1 \hat{f}_{1h_1}(\mathbf{x})}{\pi_1 \hat{f}_{1h_1}(\mathbf{x}) + \pi_2 \hat{f}_{2h_2}(\mathbf{x})}$$

# Discrimination measures (kernel discriminant analysis)

- Posterior probability :

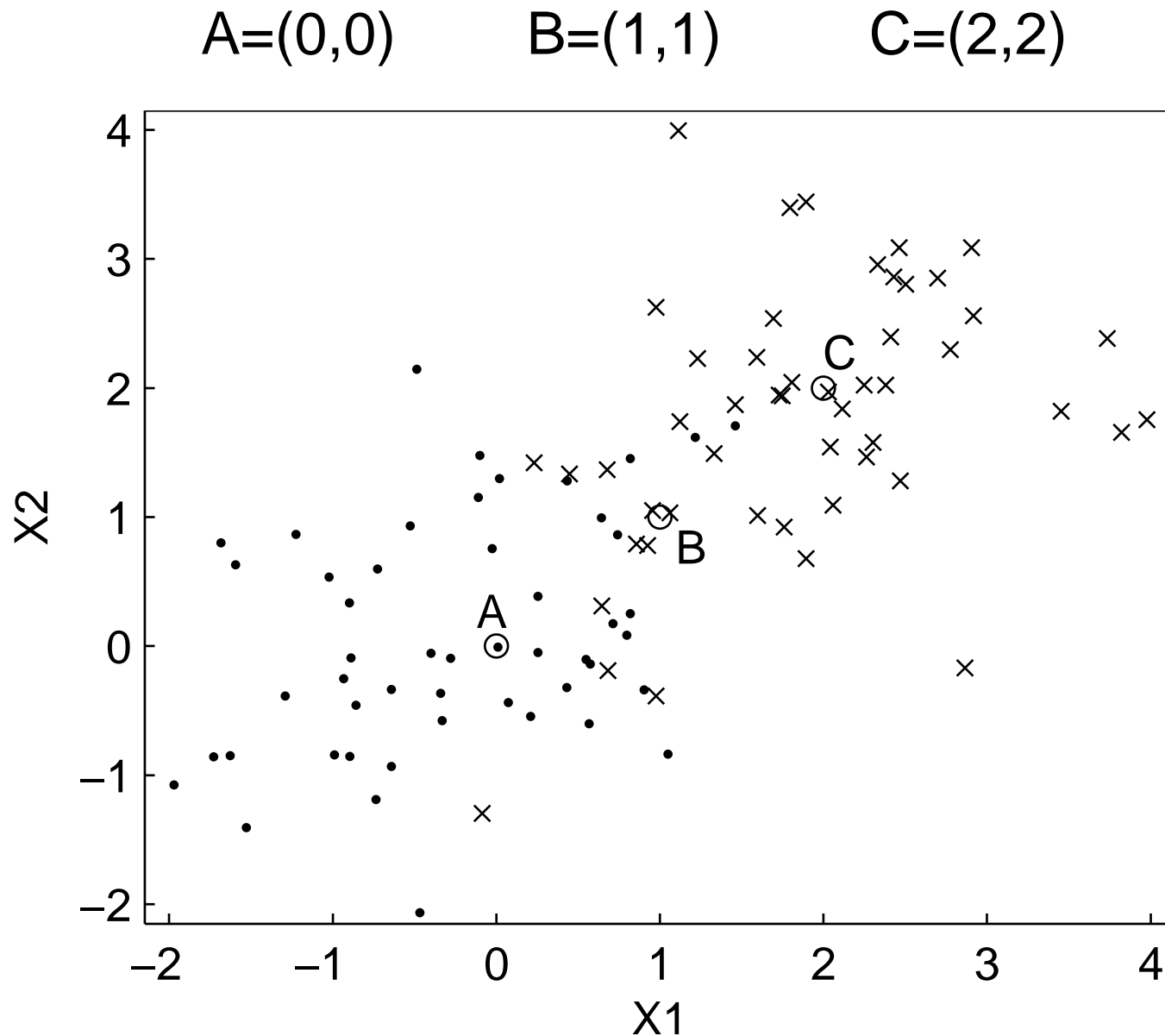
$$\mathcal{P}_{h_1, h_2}(1 \mid \mathbf{x}) = \frac{\pi_1 \hat{f}_{1h_1}(\mathbf{x})}{\pi_1 \hat{f}_{1h_1}(\mathbf{x}) + \pi_2 \hat{f}_{2h_2}(\mathbf{x})}$$

- A statistical measure of evidence (p-value)  
[Technometrics, 2006] :

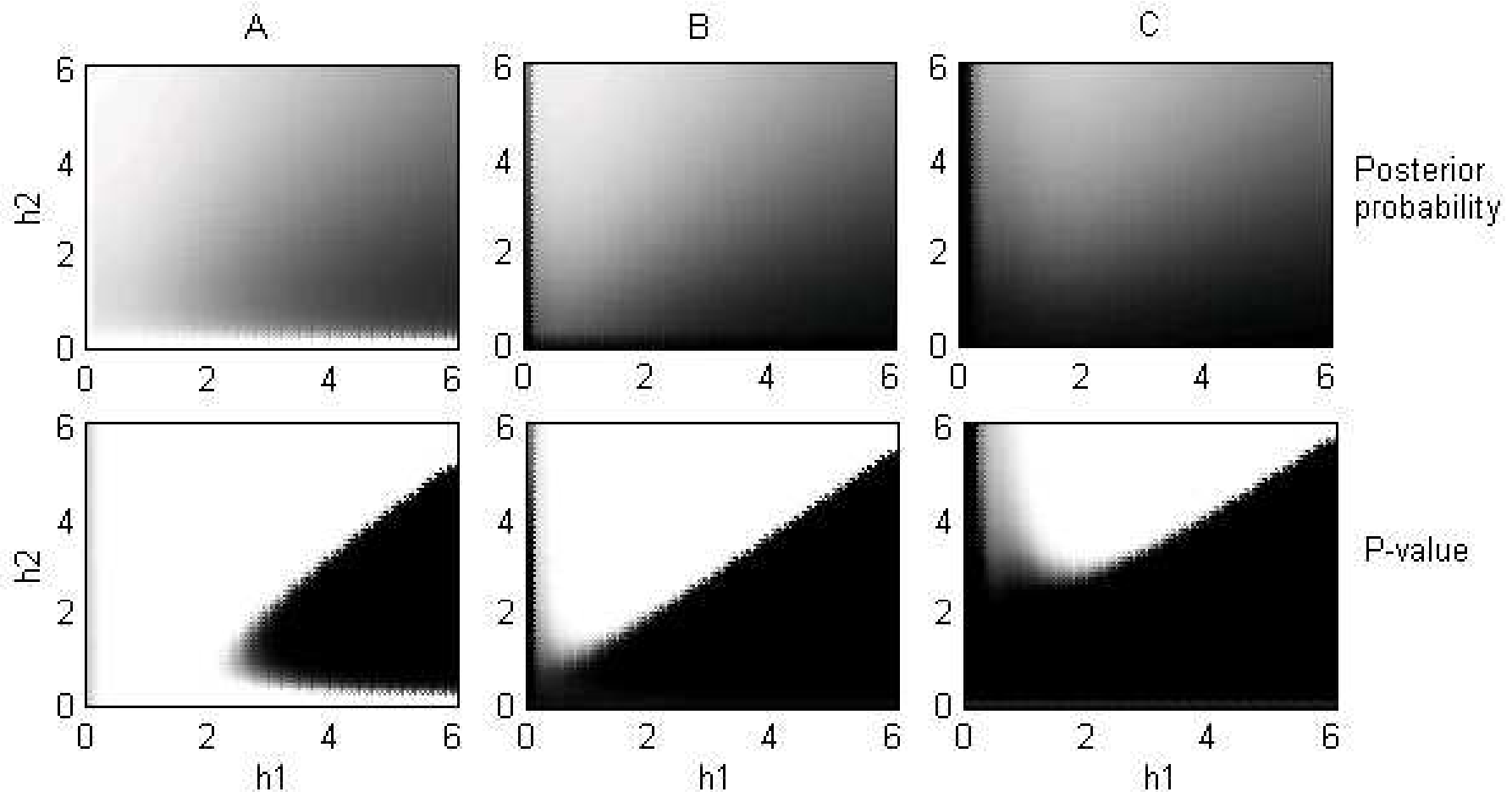
$$\begin{aligned} P_{h_1, h_2}(\mathbf{x}) &= P\{\pi_1 \hat{f}_{1h_1}(\mathbf{x}) > \pi_2 \hat{f}_{2h_2}(\mathbf{x})\} \\ &\simeq \Phi \left( \frac{\pi_1 \hat{f}_{1h_1}(\mathbf{x}) - \pi_2 \hat{f}_{2h_2}(\mathbf{x})}{\sqrt{\frac{\pi_1^2 s_{1h_1}^2(\mathbf{x})}{n_1} + \frac{\pi_2^2 s_{2h_2}^2(\mathbf{x})}{n_2}}} \right) \end{aligned}$$

# A simulated data set

## $N(0,0,1,1,0)$ Vs. $N(2,2,1,1,0)$



# LIGO Multi-scale analysis and visualization



**P-value makes the plots sharper and facilitates the visualization of classification results**

# Discrimination measures (nearest neighbor classification)

---

- Posterior probability :

$$\mathcal{P}_k(j | \mathbf{x}) = k_j/k$$

Proportion of class  $j$  observations among the  $k$  nearest neighbors of  $\mathbf{x}$ .

# Discrimination measures (nearest neighbor classification)

- Posterior probability :

$$\mathcal{P}_k(j | \mathbf{x}) = k_j/k$$

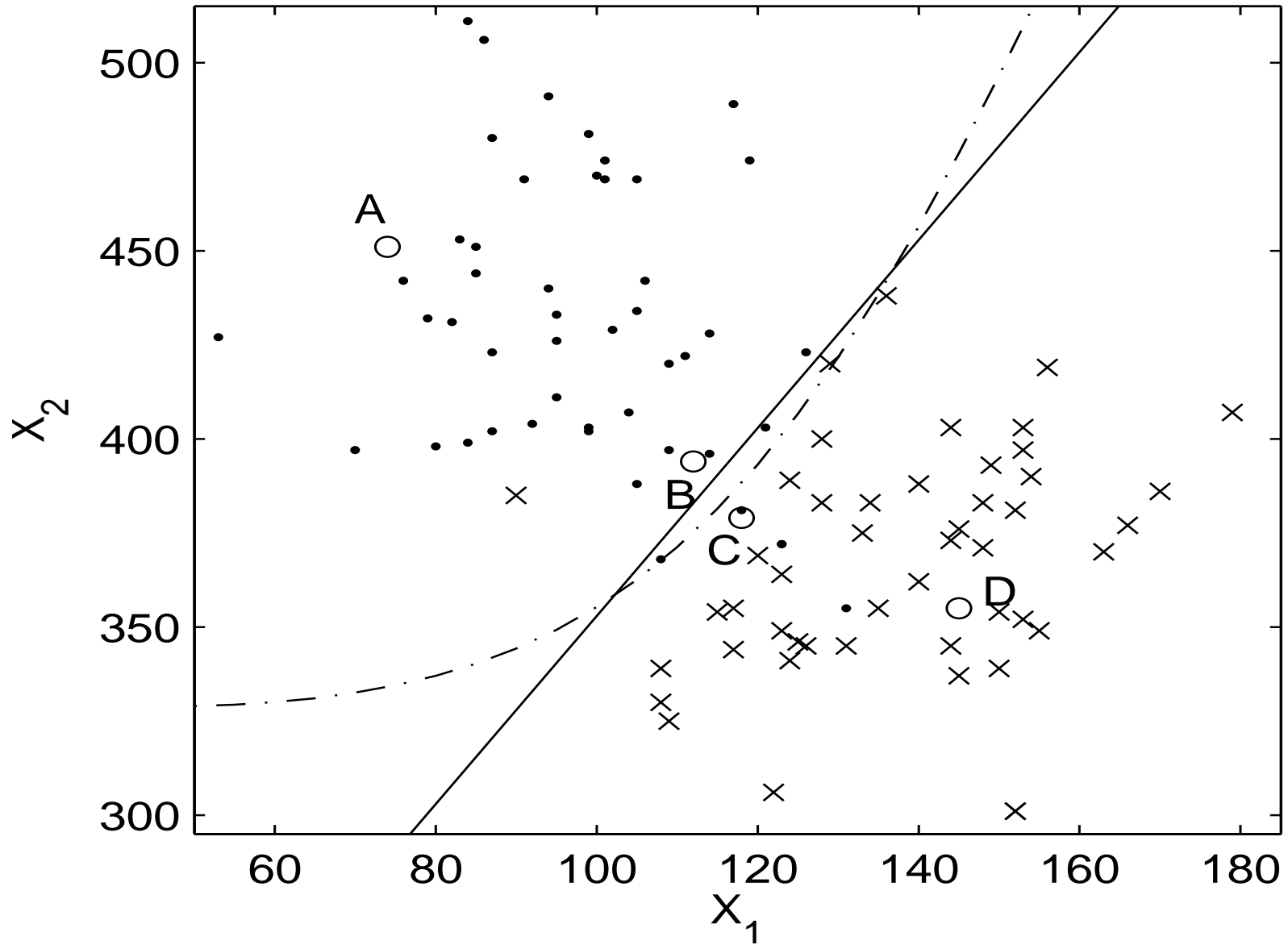
Proportion of class  $j$  observations among the  $k$  nearest neighbors of  $\mathbf{x}$ .

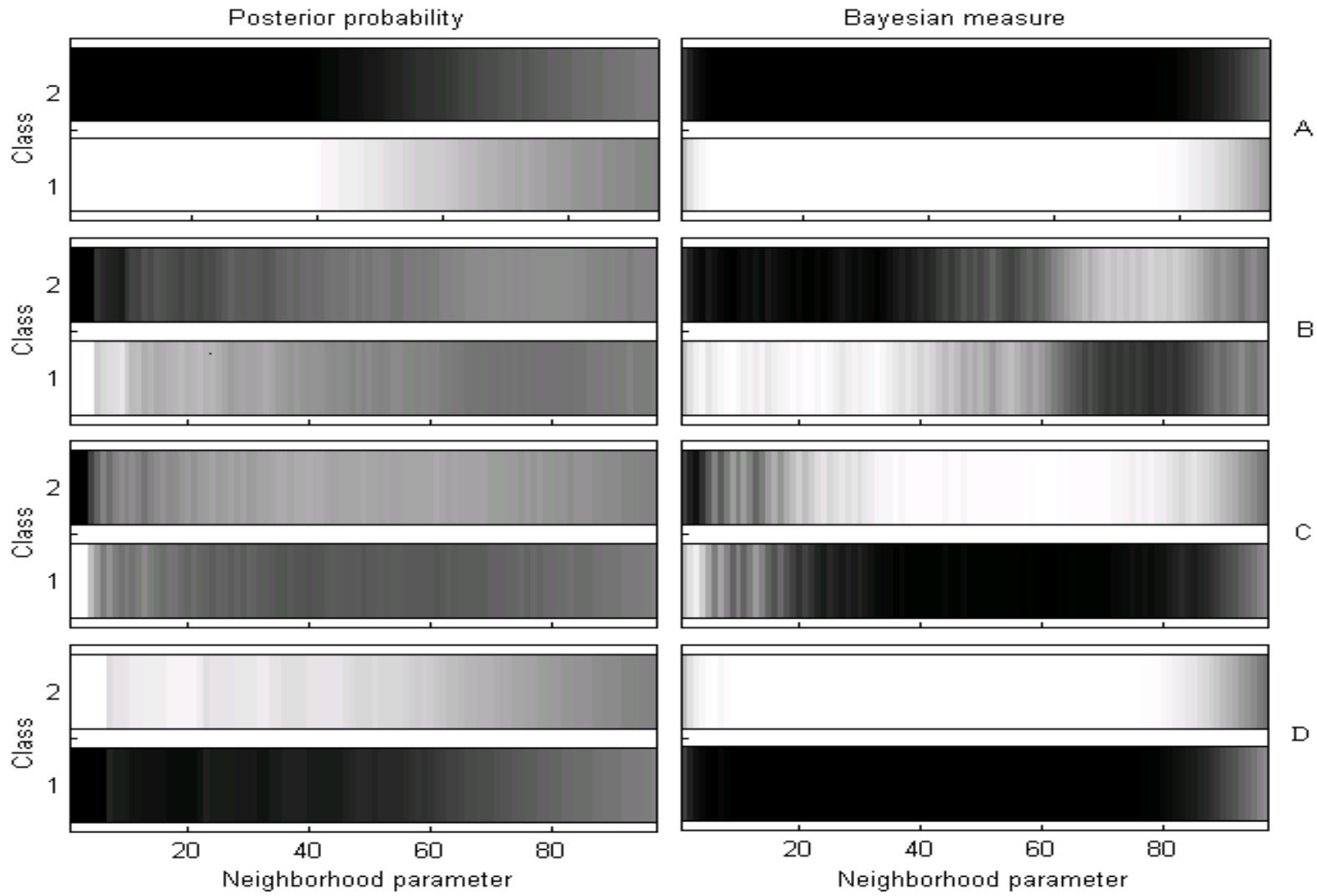
- Bayesian measure of strength [IEEE PAMI, 2005] :

$$S(j | k) = \int_{p_j = \max\{p_1, p_2, \dots, p_J\}} f(\mathbf{p} | k, \mathbf{t}) d\mathbf{p},$$

where  $f(\mathbf{p} | k, \mathbf{t}_k) = \pi(\mathbf{p}) \varphi(\mathbf{t}_k | \mathbf{p}, k) / \int \pi(\mathbf{p}) \varphi(\mathbf{t}_k | \mathbf{p}, k) d\mathbf{p}$

# Salmon data





**Bayesian measure sharpens the plot without loss of information**

# Aggregation of results : weighted averaging of posteriors

---

- Weight function  $w(s)$  should be a decreasing function of the misclassification rate.

# Aggregation of results : weighted averaging of posteriors

---

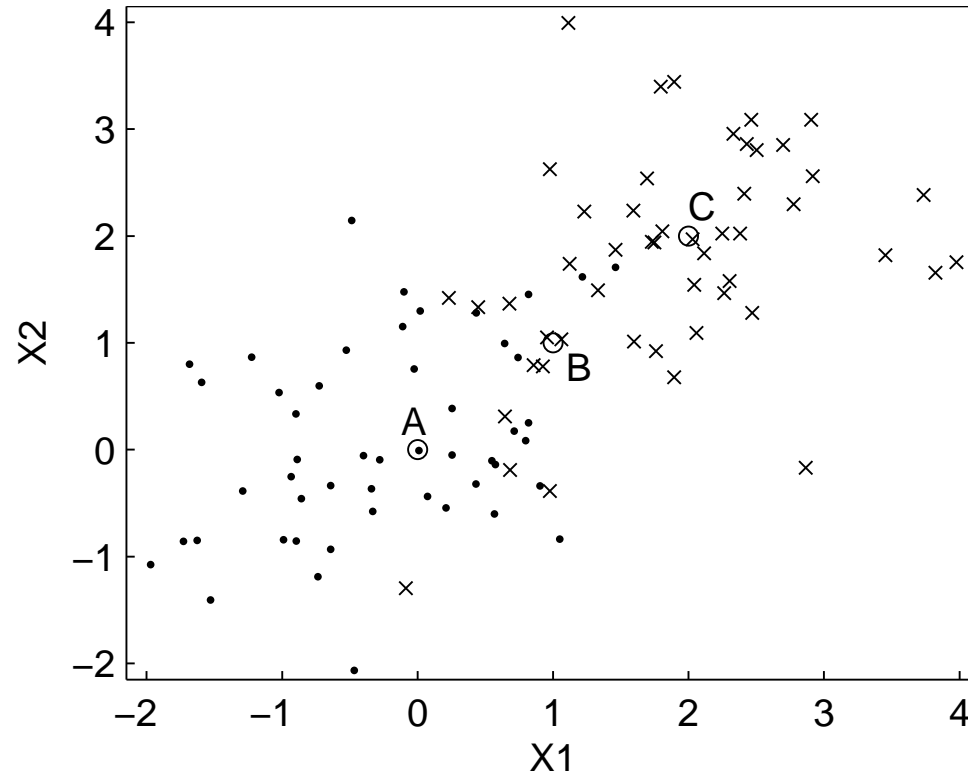
- Weight function  $w(s)$  should be a decreasing function of the misclassification rate.
- It is desirable to have some adjustment to  $w(s)$  depending on the observation to be classified. For example, in case of kernels, we can use  $w_{\mathbf{x}}(s) = w(s) |P_s(\mathbf{x}) - 0.5|$ .

# Aggregation of results : weighted averaging of posteriors

- Weight function  $w(s)$  should be a decreasing function of the misclassification rate.
- It is desirable to have some adjustment to  $w(s)$  depending on the observation to be classified. For example, in case of kernels, we can use  $w_{\mathbf{x}}(s) = w(s) |P_s(\mathbf{x}) - 0.5|$ .
- Assign an observation to class having the largest weighted posterior.

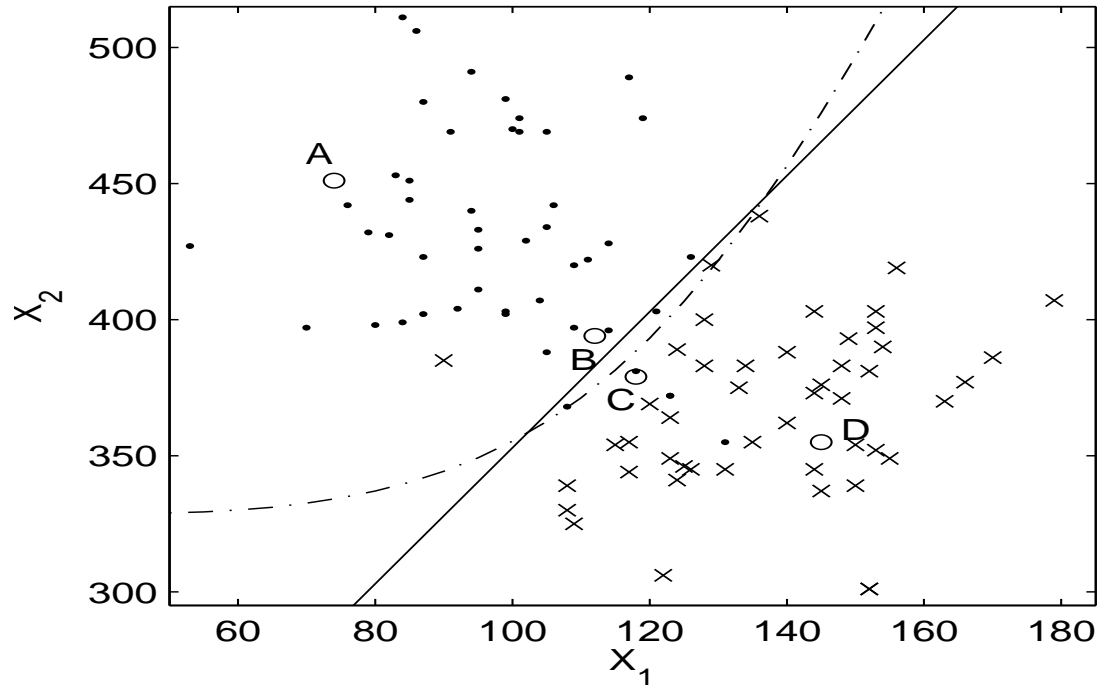
$$\mathcal{P}^{(w)}(j | \mathbf{x}) = \sum_{s \in S} w_{\mathbf{x}}(s) \mathcal{P}_s(j | \mathbf{x})$$

# Aggregation of classifiers



	Weighted posteriors		
	A	B	C
Cross-validation	0.521	0.497	0.464
Wt. posterior	0.665	0.484	0.299

# Aggregation of classifiers



	A	B	C	D
Cross-validation	1	6/7	4/7	1/7
Wt. posterior	0.8804	0.6261	0.4146	0.1445

Classifiers	Diabetes	Image	Vowel	Synthetic	Vowel-2	Sonar
LDA	11.0	11.4	55.6	10.8	25.2	20.2
QDA	9.7	14.6	52.8	10.2	19.8	15.4
CART	—	12.6	56.4	10.1	23.7	20.2
Neural Net	—	12.1	50.9	9.4	18.6	19.2
Kernel (MISE)	12.4	15.7	62.1	9.3	18.9	14.4
Wt. avg	6.2	11.0	51.9	9.2	17.4	12.5

# Results on benchmark data sets : comparison with cross validation method

---

Data sets	$k$ -NN (cross-valid.)	Weighted posterior
Salmon	9.38 (0.18)	8.30 (0.14)
Wine	1.05 (0.07)	0.45 (0.05)
Vowel-2	17.75 (2.09)	18.93 (2.15)
Diabetes	11.50 (0.16)	10.47 (0.15)
Biomedical	17.61 (0.18)	17.10 (0.16)

- Ghosh, A. K., Chaudhuri, P. and Murthy, C. A. (2005) On visualization and aggregation of nearest neighbor classifiers. *IEEE Trans. Pattern Anal. Machine Intell.*, **27**, 1592-1602.
- Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2006) Classification using kernel density estimates : multi-scale analysis and visualization. *Technometrics*, **48**, 120-132.
- Ghosh, A. K., Chaudhuri, P. and Murthy, C. A. (2006) Multiscale classification using nearest neighbor density estimates. *IEEE Trans. Sys. Man Cybern.*, **36**, 1139-1148.
- Chaudhuri, P., Ghosh, A. K. and Oja, H. (2008) Classification based on hybridization of parametric and nonparametric density estimates. *IEEE Trans. Pattern Anal. Machine Intell.*, To appear (preprint available on journal's website).