LIGO

MULTI-SCALE SMOOTHING IN NONPARAMETRIC CLASSIFICATION

PROBAL CHAUDHURI

probal@isical.ac.in

THEORETICAL STATISTICS AND MATHEMATICS UNIT

INDIAN STATISTICAL INSTITUTE, KOLKATA.

LIGO

Contents

Contents of this talk are based on

- Ghosh, A. K. and Chaudhuri, P. (2004) Optimal smoothing in kernel discriminant analysis. **Statistica Sinica, 14, 457-483.**
- Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2005) Classification using kernel density estimates : multi-scale analysis and visualization. **Technometrics (In Press).**
- Ghosh, A. K., Chaudhuri, P. and Murthy, C. A. (2005) On visualization and aggregation of nearest neighbor classifiers. **IEEE Trans. PAMI (In Press).**

PDF versions of these articles are available at

http://www.geocities.com/ghosh_anilk

LIGO

Some examples

- **Vowel recognition data**: A number of speakers spoke some words formed by ‘h’ followed by a vowel and then followed by ‘d’. Format frequencies of a speaker’s vocal tract were noted for different vowels.
  - **Vowel data-1** : 11 classes in 10 dimensions.
  - **Vowel data-2** : 10 classes in 2 dimensions.

LIGO

Some examples

- **Vowel recognition data**: A number of speakers spoke some words formed by ‘h’ followed by a vowel and then followed by ‘d’. Format frequencies of a speaker’s vocal tract were noted for different vowels.
  - **Vowel data-1** : 11 classes in 10 dimensions.
  - **Vowel data-2** : 10 classes in 2 dimensions.
- **Sonar data**: Patterns were obtained by bouncing sonar signals (frequency-modulated chirp) off a metal cylinder or from cylindrical rocks. Each of 60 observations on a pattern represents the energy within a particular frequency band, integrated over a certain period of time. These observations were averaged in a band of three to reduce the number of variables to 20.

LIGO

Some examples (contd..)

- **Image segmentation data**: 19 different measurements are taken on an image of a region consisting of 9 pixels. Images are taken from one of the following classes : brickface, sky, foliage, cement, window, path, grass.

LIGO

Some examples (contd..)

- **Image segmentation data**: 19 different measurements are taken on an image of a region consisting of 9 pixels. Images are taken from one of the following classes : brickface, sky, foliage, cement, window, path, grass.

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

LIGO

Some examples (contd..)

- **Image segmentation data**: 19 different measurements are taken on an image of a region consisting of 9 pixels. Images are taken from one of the following classes : brickface, sky, foliage, cement, window, path, grass.

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

- **Synthetic data (Ripley, 1994)**: Each of the two classes is an equal mixture of two bivariate normal distributions, which have the same dispersion matrix but different location parameters.

LIGO

Mathematical formulation

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

LIGO

Mathematical formulation

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

LIGO

Mathematical formulation

- **Data** : $(\mathbf{x}_n, c_n), n = 1, 2, \ldots, N$.
  Vector of measurement variables : $\mathbf{x}_n \in R^d$,
  Class labels :$c_n \in \{1, 2, \ldots, J\}$.
- **Decision rule** : $\mathrm{d}(\mathbf{x}) \ : \ R^d \rightarrow \{1, 2, \ldots, J.\}$
- **Bayes rule** : $\mathrm{d}_B(\mathbf{x}) = arg \max_j p(j \mid \mathbf{x}) = arg \max_j \pi_j f_j(\mathbf{x})$

  $f_j(\mathbf{x})$: density functions, π_j : prior probabilities.

LIGO

Mathematical formulation

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

Bayes classifier can be used only when the class densities and prior probabilities are known.

LIGO

Estimation of class densities

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

LIGO

Estimation of class densities

- **Parametric** : Assumes specific functional form for f and uses the training data to estimate its parameters.
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)
- **Nonparametric** : No specific assumption about the functional form of f.
  - Kernel Discriminant Analysis
  - Nearest Neighbor Classification
  - Classification and Regression Trees (CART)
  - Neural Networks
  - Support Vector Machines.

LIGO

Kernel discriminant analysis

- Kernel density estimate :

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

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

LIGO

Kernel discriminant analysis

- Kernel density estimate :

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

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

- Classification rule :

$$\mathrm{d}_K(\mathbf{x}) = \arg\max_j \pi_j \hat{f}_{jh_j}(\mathbf{x})$$

LIGO

Kernel discriminant analysis

- Kernel density estimate :

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

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

- Classification rule :

$$\mathrm{d}_K(\mathbf{x}) = \arg\max_j \pi_j \hat{f}_{jh_j}(\mathbf{x})$$

How to choose the optimum bandwidths?

LIGO

Standard methods for bandwidth selection

- Minimization of Mean Integrated Square Error (MISE) of density estimates

$$\begin{aligned} \text{MISE}(h) &= E\left[\int \left(f(\mathbf{x}) - \hat{f}_h(\mathbf{x})\right)^2 d\mathbf{x}\right] \\ &= \int \left\{Bias^2\left(\hat{f}_h(\mathbf{x})\right) + Var\left(\hat{f}_h(\mathbf{x})\right)\right\} d\mathbf{x} \end{aligned}$$

LIGO

Standard methods for bandwidth selection

- Minimization of Mean Integrated Square Error (MISE) of density estimates

$$\begin{aligned} \text{MISE}(h) &= E\left[\int \left(f(\mathbf{x}) - \hat{f}_h(\mathbf{x})\right)^2 d\mathbf{x}\right] \\ &= \int \left\{Bias^2\left(\hat{f}_h(\mathbf{x})\right) + Var\left(\hat{f}_h(\mathbf{x})\right)\right\} d\mathbf{x} \end{aligned}$$

- Minimization of cross-validated (CV) estimates of misclassification probabilities

LIGO

Standard methods for bandwidth selection

- Minimization of Mean Integrated Square Error (MISE) of density estimates

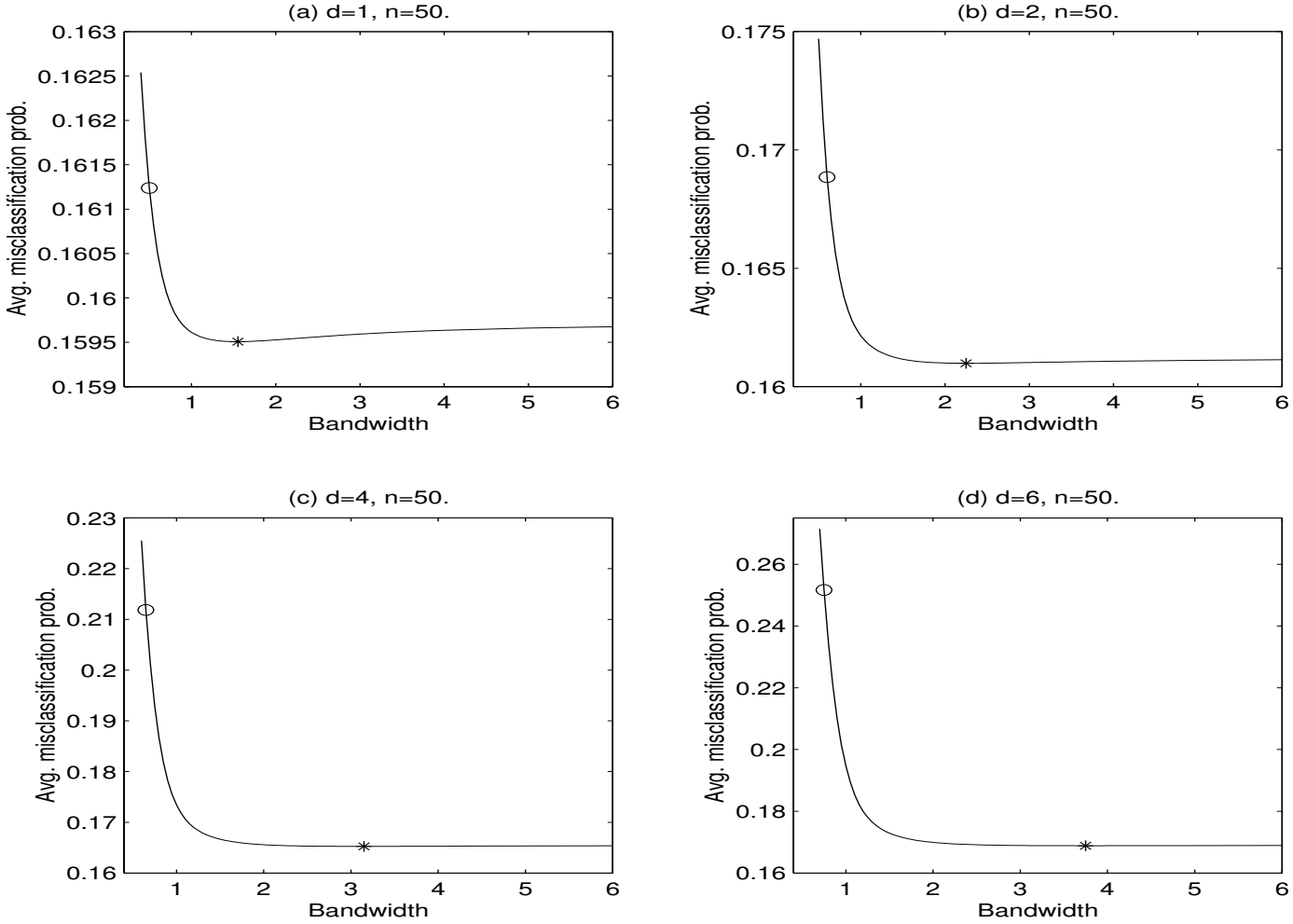
$$\begin{aligned} \text{MISE}(h) &= E\left[\int \left(f(\mathbf{x}) - \hat{f}_h(\mathbf{x})\right)^2 d\mathbf{x}\right] \\ &= \int \left\{Bias^2\left(\hat{f}_h(\mathbf{x})\right) + Var\left(\hat{f}_h(\mathbf{x})\right)\right\} d\mathbf{x} \end{aligned}$$

- Minimization of cross-validated (CV) estimates of misclassification probabilities

How good are these bandwidth selectors ?

LIGO How good is MISE bandwidth ($h_\circ$) ?

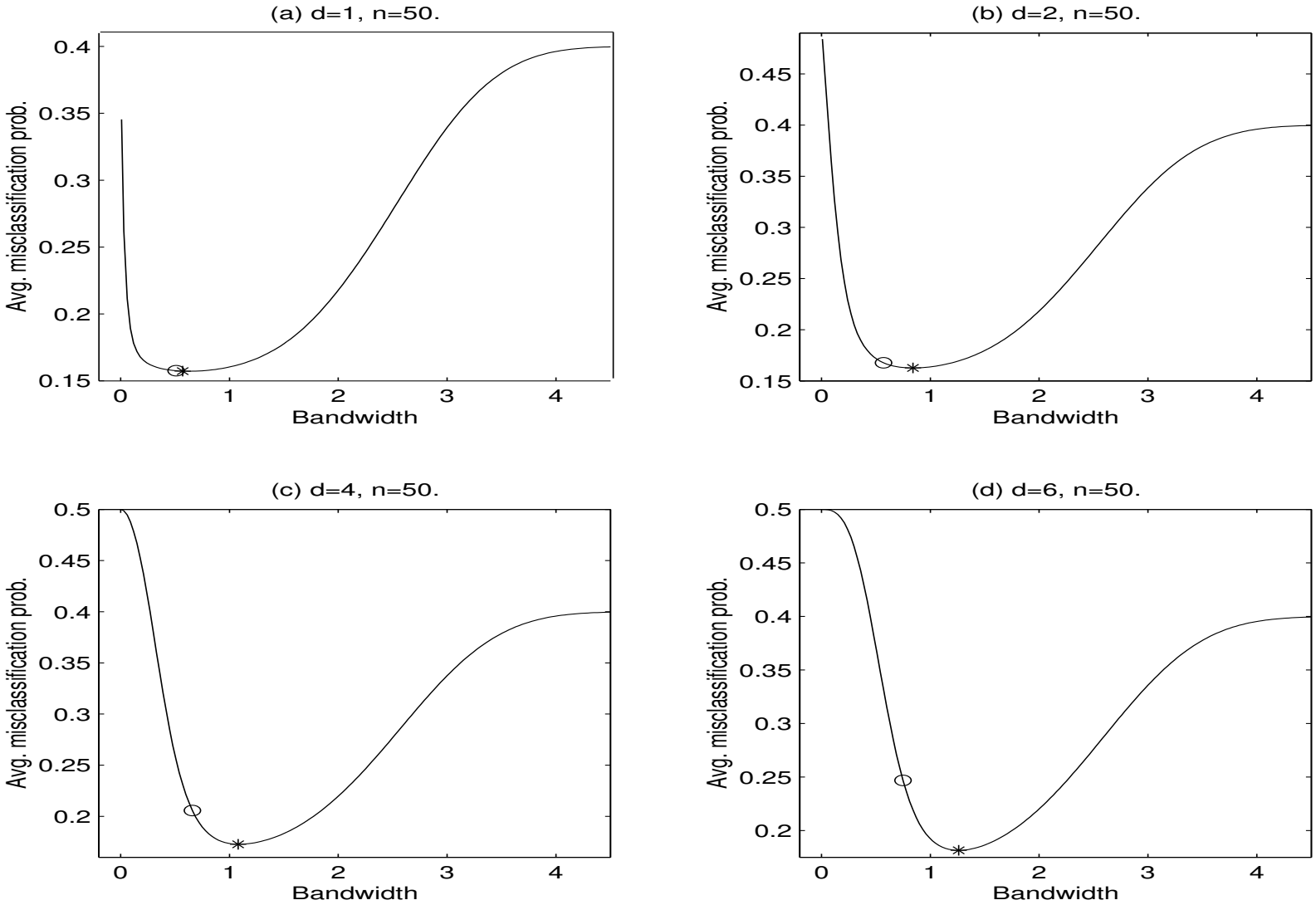
$$\pi_1 = \pi_2 = 0.5 \quad f_1 = N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \; f_2 = N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu}_1 = (2, 0, \ldots, 0)', \; \boldsymbol{\mu}_2 = (0, 0, \ldots, 0)', \boldsymbol{\Sigma} = \mathbf{I_d}$$



LIGO

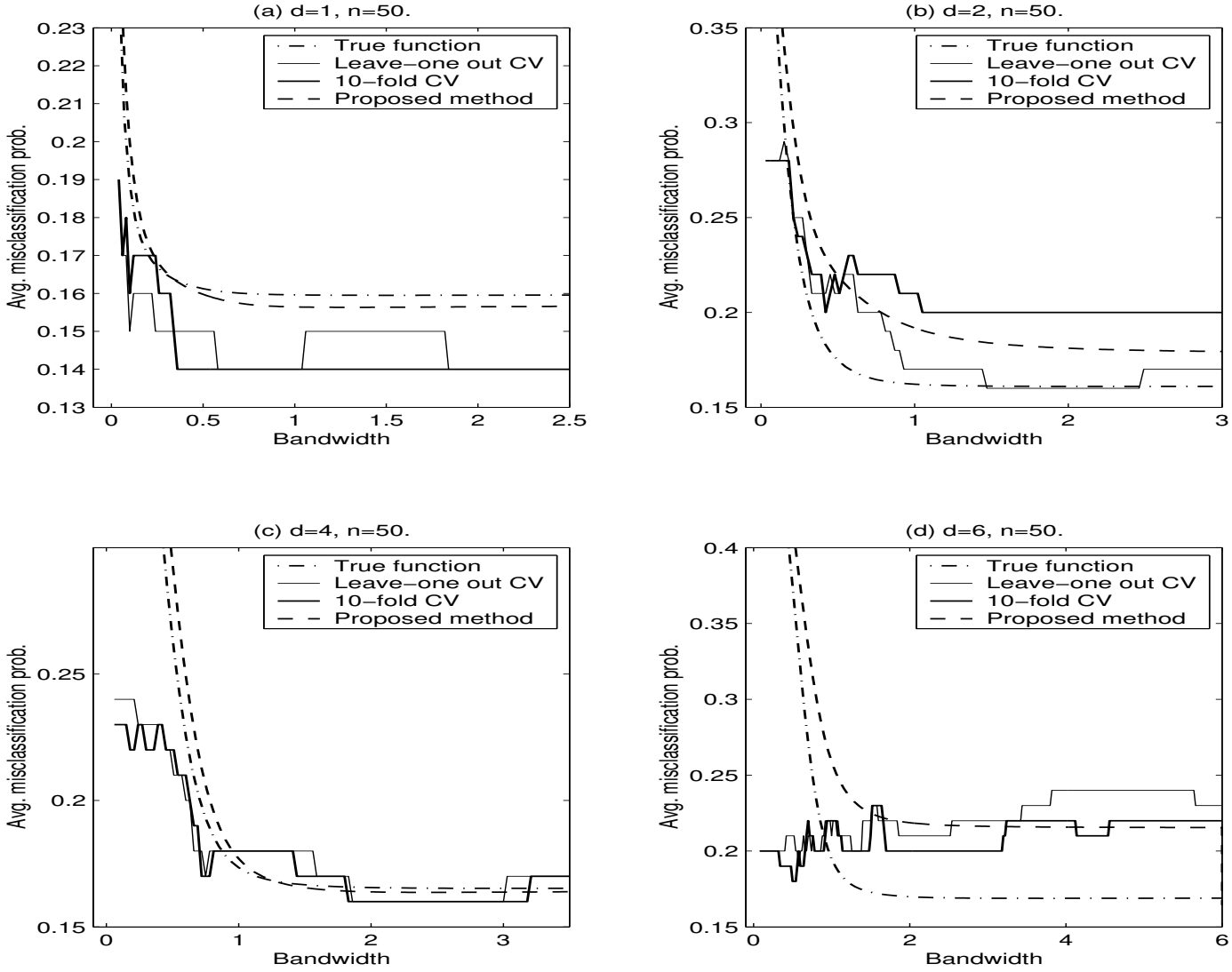
How good is $h_\circ$? (contd..)

$$\pi_1 = 0.6 \quad \pi_2 = 0.4$$



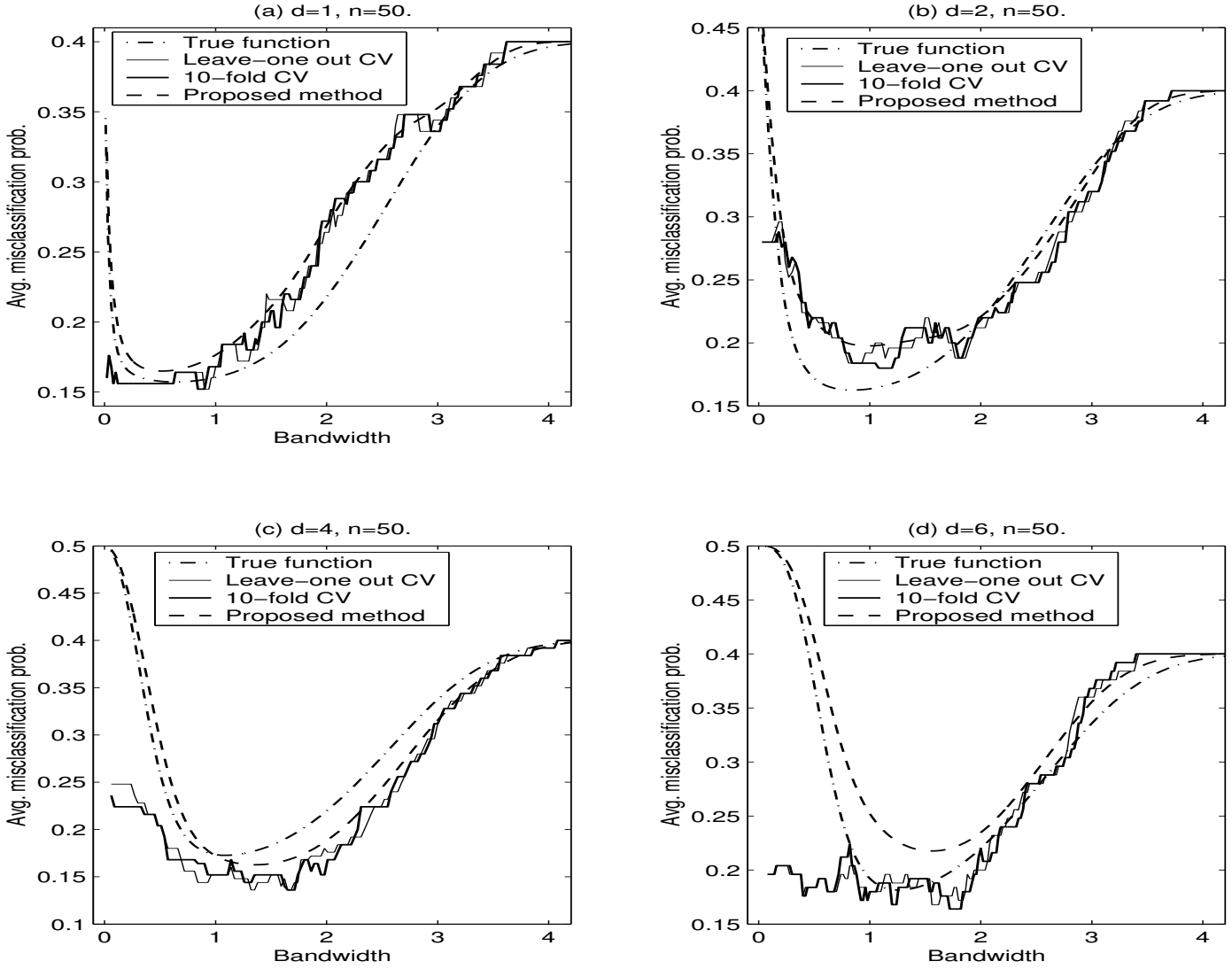
How good are CV techniques ?

$$\pi_1 = \pi_2 = 0.5$$



How good are CV techniques ? (contd..)

$$\pi_1 = 0.6 \quad \pi_2 = 0.4$$



LIGO

A smooth estimate for average misclassification probability

- $\Delta(h_1, h_2, \ldots, h_J)$

$$= 1 - \sum_{j=1}^{J} \pi_j \int P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) > \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for all } i \neq j\} \; f_j(\mathbf{x}) d\mathbf{x}$$

$$= 1 - \sum_{j=1}^{J} \pi_j \int \left[\int \prod_{i \neq j} P\{\pi_i \hat{f}_{ih_i}(\mathbf{x}) < u\} \; g_{jh_j}(u) \; du \right] \; f_j(\mathbf{x}) d\mathbf{x},$$

where $g_{jh_j}(\cdot)$ is the p.d.f. of $\pi_j \hat{f}_{jh_j}(\mathbf{x})$.

LIGO

A smooth estimate for average misclassification probability

- $\Delta(h_1, h_2, \ldots, h_J)$

$$= 1 - \sum_{j=1}^{J} \pi_j \int P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) > \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for all } i \neq j\} \; f_j(\mathbf{x}) d\mathbf{x}$$

$$= 1 - \sum_{j=1}^{J} \pi_j \int \left[\int \prod_{i \neq j} P\{\pi_i \hat{f}_{ih_i}(\mathbf{x}) < u\} \; g_{jh_j}(u) \; du \right] \; f_j(\mathbf{x}) d\mathbf{x},$$

where $g_{jh_j}(\cdot)$ is the p.d.f. of $\pi_j \hat{f}_{jh_j}(\mathbf{x})$.

- Need to find a data based estimate of $\Delta(h_1, h_2, \ldots, h_J)$ and then to minimize it.

LIGO

A smooth estimate for average misclassification probability

- $\Delta(h_1, h_2, \ldots, h_J)$

$$= 1 - \sum_{j=1}^{J} \pi_j \int P\{\pi_j \hat{f}_{jh_j}(\mathbf{x}) > \pi_i \hat{f}_{ih_i}(\mathbf{x}) \text{ for all } i \neq j\} \; f_j(\mathbf{x}) d\mathbf{x}$$

$$= 1 - \sum_{j=1}^{J} \pi_j \int \left[\int \prod_{i \neq j} P\{\pi_i \hat{f}_{ih_i}(\mathbf{x}) < u\} \; g_{jh_j}(u) \; du \right] \; f_j(\mathbf{x}) d\mathbf{x},$$

where $g_{jh_j}(\cdot)$ is the p.d.f. of $\pi_j \hat{f}_{jh_j}(\mathbf{x})$.

- Need to find a data based estimate of $\Delta(h_1, h_2, \ldots, h_J)$ and then to minimize it.
- A smooth estimate of $\Delta(h_1, h_2, \ldots, h_J)$ can be obtained using normal approximation of the distribution of kernel density estimates (Ghosh & Chaudhuri, 2004).

LIGO

Some problems in traditional approach

- It is difficult to optimize the estimated misclassification rate $\Delta(h_1, h_2, \ldots, h_J)$ when $J \geq 3$

LIGO

Some problems in traditional approach

- It is difficult to optimize the estimated misclassification rate $\Delta(h_1, h_2, \ldots, h_J)$ when $J \geq 3$
- It allows only one single bandwidth for each population density estimate

LIGO

Some problems in traditional approach

- It is difficult to optimize the estimated misclassification rate $\Delta(h_1, h_2, \ldots, h_J)$ when $J \geq 3$
- It allows only one single bandwidth for each population density estimate
- Optimum bandwidth is chosen based on the entire training sample. It is not flexible for the specific observation to be classified.

LIGO

Multi-scale analysis

- No bandwidth selection. Simultaneous study of discrimination measures for a wide range of bandwidths.

LIGO

Multi-scale analysis

- No bandwidth selection. Simultaneous study of discrimination measures for a wide range of bandwidths.
- In the presence of several competing classes, pairwise classification is followed by majority voting (Friedman, 1996) or coupling (Hastie & Tibshirani, 1998).

LIGO

Discrimination measures

- Posterior probability

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

LIGO

Discrimination measures

- Posterior probability

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

Instead of choosing the optimal values of the bandwidths, we look at $\mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x})$ for some appropriate range of values for h_1 and h_2.

LIGO

Discrimination measures (contd..)

- A p-value type measure

$$\mathbf{H_0} : \pi_1 E\{\hat{f}_{1h_1}(\mathbf{x})\} \leq \pi_2 E\{\hat{f}_{2h_2}(\mathbf{x})\}$$
$$\mathbf{H_a} : \pi_1 E\{\hat{f}_{1h_1}(\mathbf{x})\} > \pi_2 E\{\hat{f}_{2h_2}(\mathbf{x})\}$$

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

LIGO

Discrimination measures (contd..)

- A p-value type measure

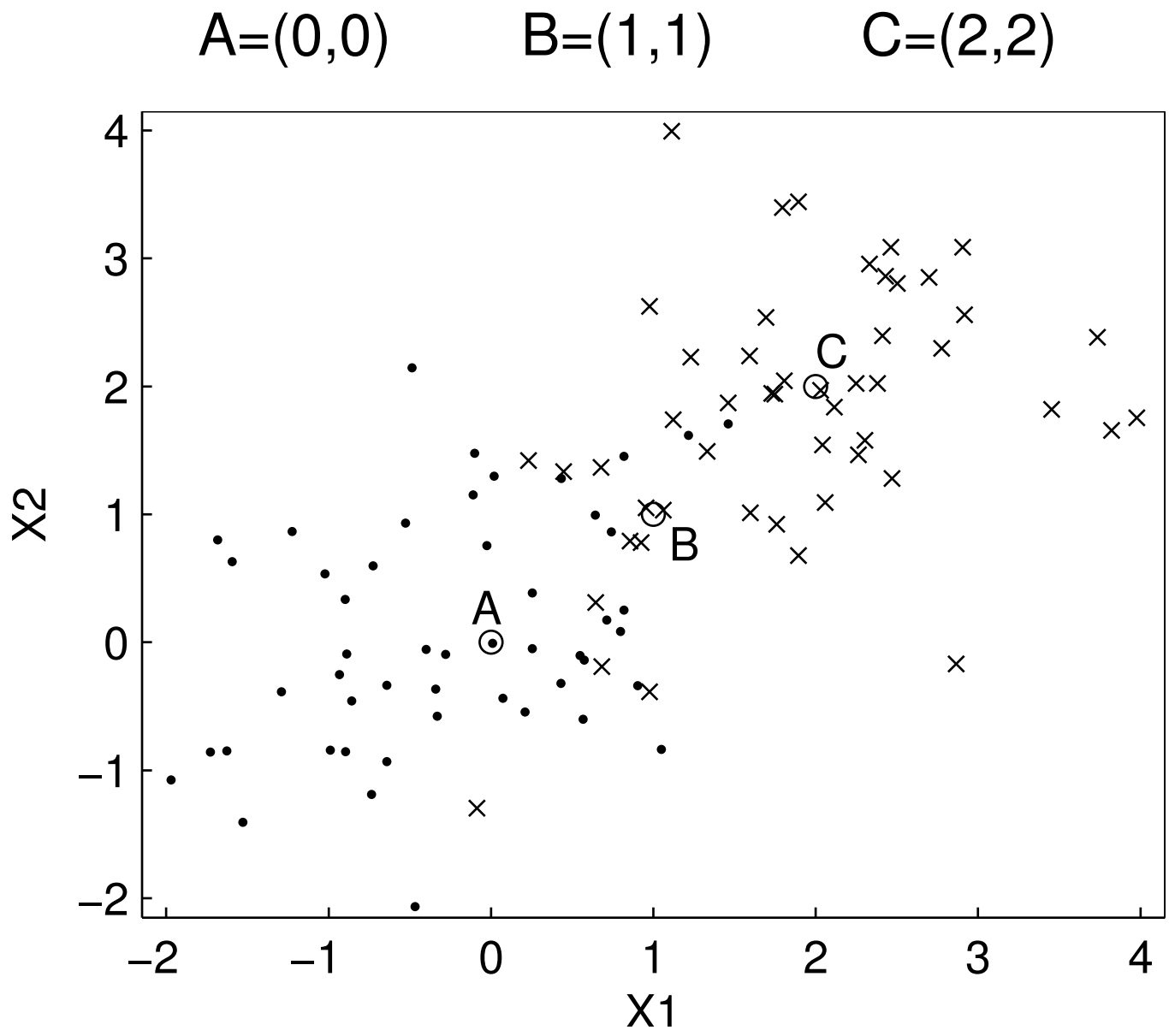
$$\mathbf{H_0} : \pi_1 E\{\hat{f}_{1h_1}(\mathbf{x})\} \leq \pi_2 E\{\hat{f}_{2h_2}(\mathbf{x})\}$$
$$\mathbf{H_a} : \pi_1 E\{\hat{f}_{1h_1}(\mathbf{x})\} > \pi_2 E\{\hat{f}_{2h_2}(\mathbf{x})\}$$

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

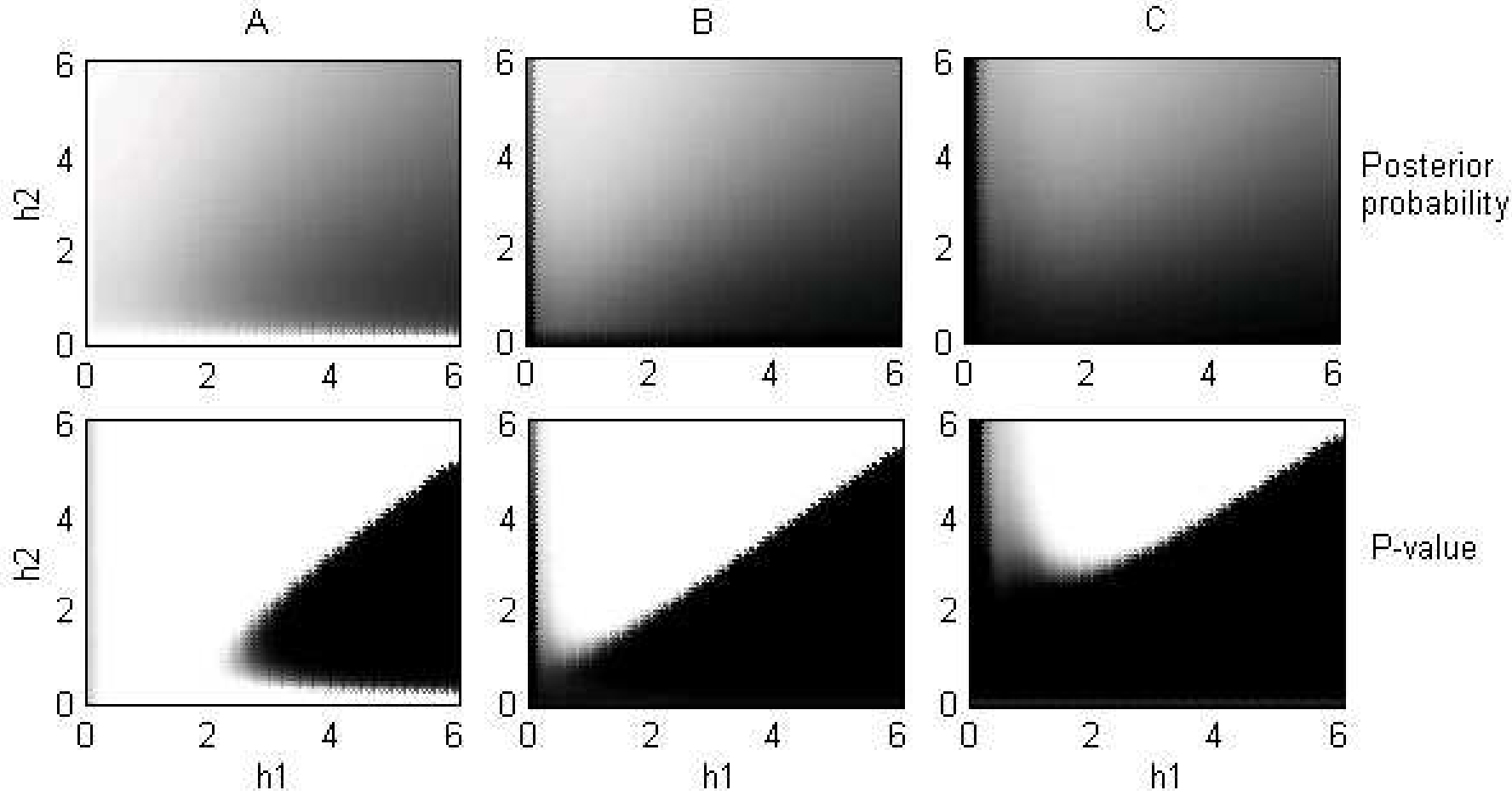
Instead of choosing the optimal values of the bandwidths, we look at $P_{h_1,h_2}(\mathbf{x})$ for some appropriate range of values for h_1 and h_2.

LIGO

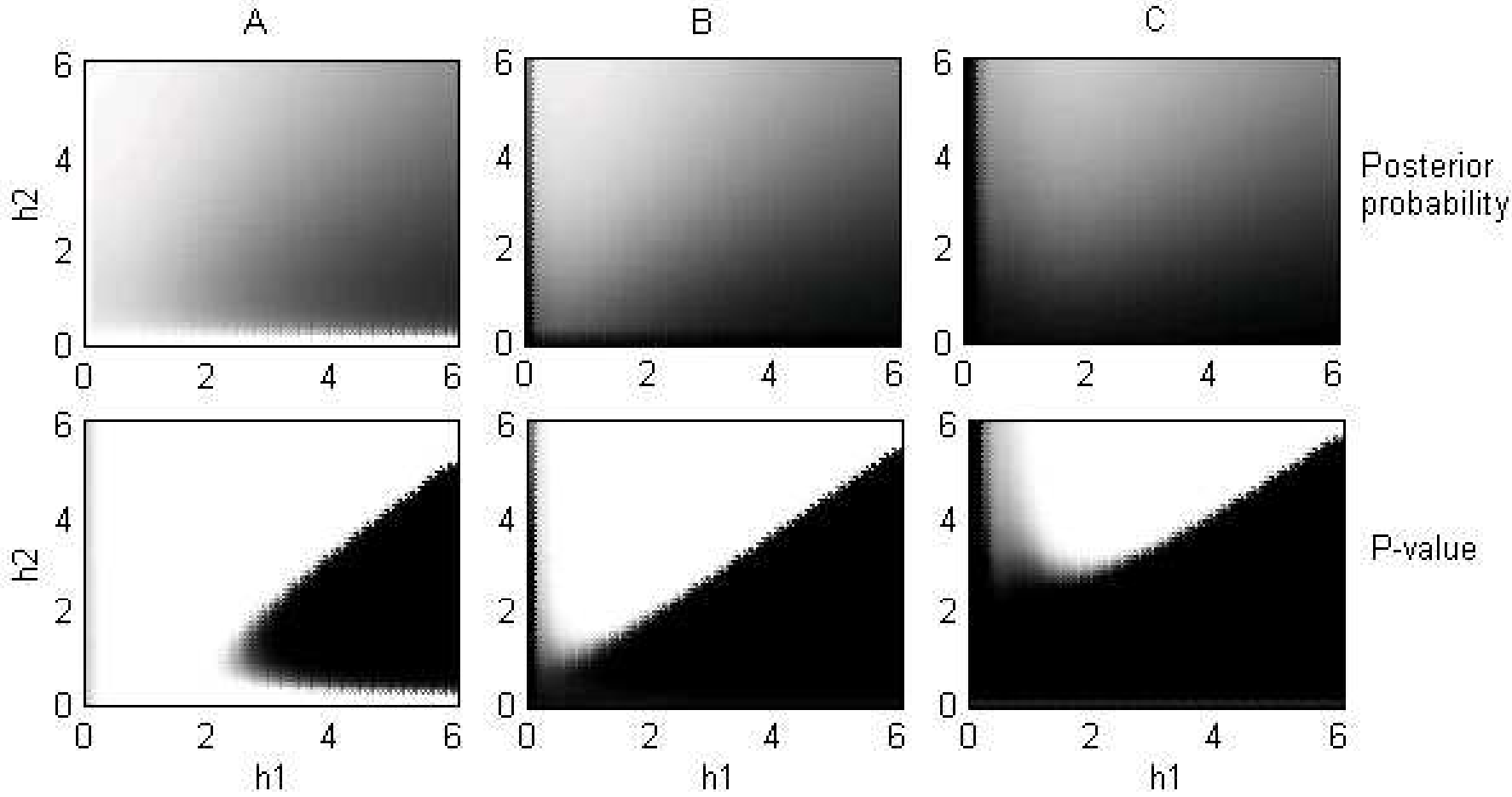
A simulated data set
N(0,0,1,1,0) Vs. N(2,2,1,1,0)



Multi-scale analysis and visualization



LIGO Multi-scale analysis and visualization



P-value sharpens the black and white contrast in the plot and hence facilitates the visualization of classification results

LIGO

Large sample properties

Scale space function : $\mathcal{S}_{jh_j}(\mathbf{x}) = E\{\hat{f}_{jh_j}(\mathbf{x})\} = f_j \circ K_{h_j}(\mathbf{x})$

LIGO

Large sample properties

Scale space function : $\mathcal{S}_{jh_j}(\mathbf{x}) = E\{\hat{f}_{jh_j}(\mathbf{x})\} = f_j \circ K_{h_j}(\mathbf{x})$

Result : Under appropriate regularity conditions, for any fixed h_1, h_2 and **x**,

- $\mid \mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x}) - \frac{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x})}{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) + \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})} \mid = O_P(N^{-1/2})$
- $\mid P_{h_1,h_2}(\mathbf{x}) - I\{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) > \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})\} \mid = O_P(N^{-1/2} e^{-CN})$ for some positive constant C

LIGO

Large sample properties

Scale space function : $\mathcal{S}_{jh_j}(\mathbf{x}) = E\{\hat{f}_{jh_j}(\mathbf{x})\} = f_j \circ K_{h_j}(\mathbf{x})$

Result : Under appropriate regularity conditions, for any fixed h_1, h_2 and **x**,

- $\mid \mathcal{P}_{h_1,h_2}(1 \mid \mathbf{x}) - \frac{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x})}{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) + \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})} \mid = O_P(N^{-1/2})$
- $\mid P_{h_1,h_2}(\mathbf{x}) - I\{\pi_1 \mathcal{S}_{1h_1}(\mathbf{x}) > \pi_2 \mathcal{S}_{2h_2}(\mathbf{x})\} \mid = O_P(N^{-1/2} e^{-CN})$ for some positive constant C

Because of fast convergence of P-values to indicator functions, the black and white contrast in the plot gets sharper.

LIGO

Aggregation of results

- **Final classification :** Assign an observation to class having the largest weighted posterior.

$$\mathrm{d}_{K^*}(\mathbf{x}) = \arg\max_j \sum_{h_1} \sum_{h_2} W_{\mathbf{x}}(h_1, h_2) \mathcal{P}_{h_1,h_2}(j \mid \mathbf{x}) \qquad j = 1, 2,$$

where $W_{\mathbf{x}}(h_1, h_2)$'s are weights for the kernel classifiers with different values of bandwidths.

LIGO

Aggregation of results

- **Final classification :** Assign an observation to class having the largest weighted posterior.

  $$\mathrm{d}_{K^*}(\mathbf{x}) = \arg\max_j \sum_{h_1} \sum_{h_2} W_{\mathbf{x}}(h_1, h_2) \mathcal{P}_{h_1,h_2}(j \mid \mathbf{x}) \qquad j = 1, 2,$$

  where $W_{\mathbf{x}}(h_1, h_2)$'s are weights for the kernel classifiers with different values of bandwidths.

- Weights depend on the misclassification probabilities $\Delta(h_1, h_2)$ and also on the specific observation $\mathbf{x}$.

LIGO

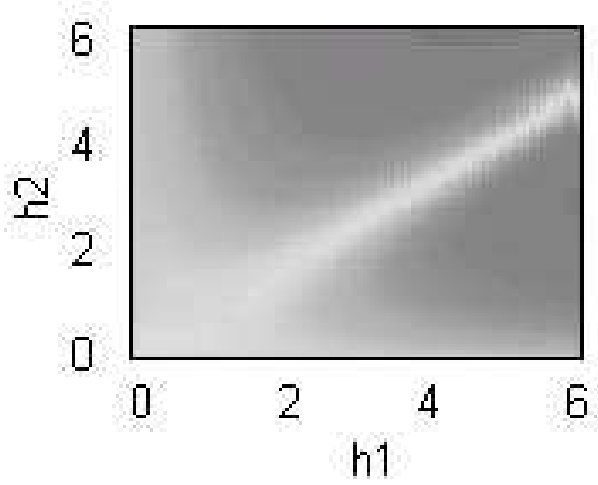
Aggregation of results

- **Final classification :** Assign an observation to class having the largest weighted posterior.

$$\mathrm{d}_{K^*}(\mathbf{x}) = \arg\max_j \sum_{h_1} \sum_{h_2} W_{\mathbf{x}}(h_1, h_2) \mathcal{P}_{h_1,h_2}(j \mid \mathbf{x}) \qquad j = 1, 2,$$

where $W_{\mathbf{x}}(h_1, h_2)$'s are weights for the kernel classifiers with different values of bandwidths.

- Weights depend on the misclassification probabilities $\Delta(h_1, h_2)$ and also on the specific observation **x**.



Grey-scale values for probability of correct classification

LIGO

Aggregation of results (contd..)

- Weight function :

$$w(h_1,h_2) = \begin{cases} \exp\left\{-\frac{1}{2}\frac{(\widehat{\Delta}(h_1,h_2)-\widehat{\Delta}_0)^2}{\widehat{\Delta}_0(1-\widehat{\Delta}_0)/N}\right\} & \text{if } \frac{\widehat{\Delta}(h_1,h_2)-\widehat{\Delta}_0}{[\widehat{\Delta}_0(1-\widehat{\Delta}_0)/N]^{1/2}} \leq \tau \text{ and } \\ & \widehat{\Delta}(h_1,h_2) < \min\{\pi_1,\pi_2\} \\ 0 & \text{otherwise,} \end{cases}$$

where $\widehat{\Delta}_0 = \min \widehat{\Delta}_{h_1,h_2}$

LIGO

Aggregation of results (contd..)

- Weight function :

$$w(h_1,h_2) = \begin{cases} \exp\left\{-\frac{1}{2}\frac{(\widehat{\Delta}(h_1,h_2)-\widehat{\Delta}_0)^2}{\widehat{\Delta}_0(1-\widehat{\Delta}_0)/N}\right\} & \text{if } \frac{\widehat{\Delta}(h_1,h_2)-\widehat{\Delta}_0}{[\widehat{\Delta}_0(1-\widehat{\Delta}_0)/N]^{1/2}} \leq \tau \text{ and} \\ & \widehat{\Delta}(h_1,h_2) < \min\{\pi_1,\pi_2\} \\ 0 & \text{otherwise,} \end{cases}$$

where $\widehat{\Delta}_0 = \min \widehat{\Delta}_{h_1,h_2}$

- Choice of τ :
  - $\tau = \mathbf{0}$: Assign weights only on those bandwidth pairs (h_1, h_2) for which $\widehat{\Delta}(h_1, h_2)$ is minimum.
  - $\tau = \mathbf{3}$: Considers all possible choices of bandwidths within some appropriate range of values for h_1 and h_2.

LIGO

Aggregation of results (contd..)

- Adjusted regional weight

$$W_{\mathbf{x}}(h_1, h_2) = w(h_1, h_2) \left| P_{h1,h2}(\mathbf{x}) - 0.5 \right|$$

LIGO

Aggregation of results (contd..)

- Adjusted regional weight

$$W_{\mathbf{x}}(h_1, h_2) = w(h_1, h_2) \left| P_{h1,h2}(\mathbf{x}) - 0.5 \right|$$

- Final classification results for A, B & C

	Weighted posteriors		
	A	B	C
τ= 0	0.521	0.497	0.464
τ =3	0.665	0.484	0.299

LIGO

Aggregation of results (contd..)

- Adjusted regional weight

$$W_{\mathbf{x}}(h_1, h_2) = w(h_1, h_2) \left| P_{h1,h2}(\mathbf{x}) - 0.5 \right|$$

- Final classification results for A, B & C

	Weighted posteriors		
	A	B	C
τ= 0	0.521	0.497	0.464
τ =3	0.665	0.484	0.299

- Multi-class problem : Pairwise classification is followed by majority voting or coupling.

LIGO

Results on benchmark data sets

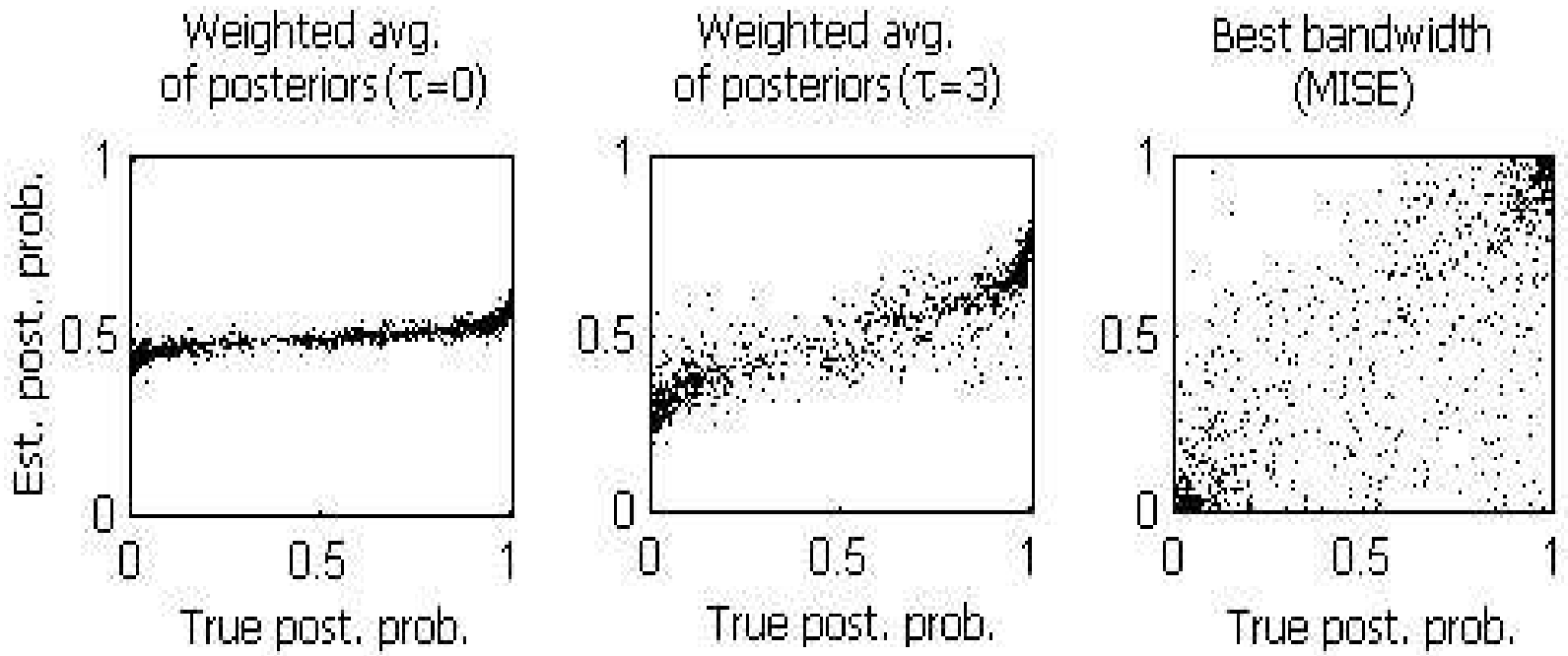
Classifiers	Diabetes	Image	Vowel
LDA	11.0	11.4	55.6
QDA	9.7	14.6	52.8
CART	—	12.6	56.4
Neural Net	—	12.1	50.9
Nearest Neighbor	9.0	18.2	43.7
Kernel (with $\mathbf{h}_0$)	12.4	15.7	62.1
Kernel (Wt. avg.)			
Voting ($\tau = 0$)	6.2	10.5	50.6
Coupling ($\tau = 0$)	15.2	11.7	47.2
Voting ($\tau = 3$)	6.2	11.0	51.9
Coupling ($\tau = 3$)	8.3	10.6	48.9

LIGO

Results on benchmark data sets (contd..)

Classifiers	Synthetic	Vowel-2	Sonar
LDA	10.8	25.2	20.2
QDA	10.2	19.8	15.4
CART	10.1	23.7	20.2
Neural Net	9.4	18.6	19.2
FDA-MARS	9.3	20.7	22.1
(degree 2)	9.6	19.8	19.2
Kernel (with $\mathbf{h}_0$)	9.3	18.9	14.4
Kernel (Wt. avg.)			
Voting ($\tau = 0$)	9.0	19.8	14.4
Coupling ($\tau = 0$)	9.0	36.6	14.4
Voting ($\tau = 3$)	9.2	17.4	12.5
Coupling ($\tau = 3$)	9.2	22.8	12.5

Estimates of posterior probability



LIGO

Nearest neighbor (NN) classification

Bayes rule : $d_B(\mathbf{x}) = arg\max_j p(j \mid \mathbf{x}) = arg\max_j \pi_j f_j(\mathbf{x})$

$f_j(\mathbf{x})$: density functions, π_j : prior probabilities.

LIGO Nearest neighbor (NN) classification

Bayes rule : $d_B(\mathbf{x}) = arg\max_j p(j \mid \mathbf{x}) = arg\max_j \pi_j f_j(\mathbf{x})$

$f_j(\mathbf{x})$: density functions, π_j : prior probabilities.

- Nearest neighbor density estimates :
  $\hat{f}_j^{(v)}(\mathbf{x}) = n_j^{(v)}/(n_j \times v)$

LIGO Nearest neighbor (NN) classification

Bayes rule : $d_B(\mathbf{x}) = arg\max_j p(j \mid \mathbf{x}) = arg\max_j \pi_j f_j(\mathbf{x})$

$f_j(\mathbf{x})$: density functions, π_j : prior probabilities.

- Nearest neighbor density estimates : $\hat{f}_j^{(v)}(\mathbf{x}) = n_j^{(v)}/(n_j \times v)$
- Estimates for priors : $\hat{\pi}_j = n_j/N$

LIGO

Nearest neighbor (NN) classification

Bayes rule : $d_B(\mathbf{x}) = arg\max_j p(j \mid \mathbf{x}) = arg\max_j \pi_j f_j(\mathbf{x})$

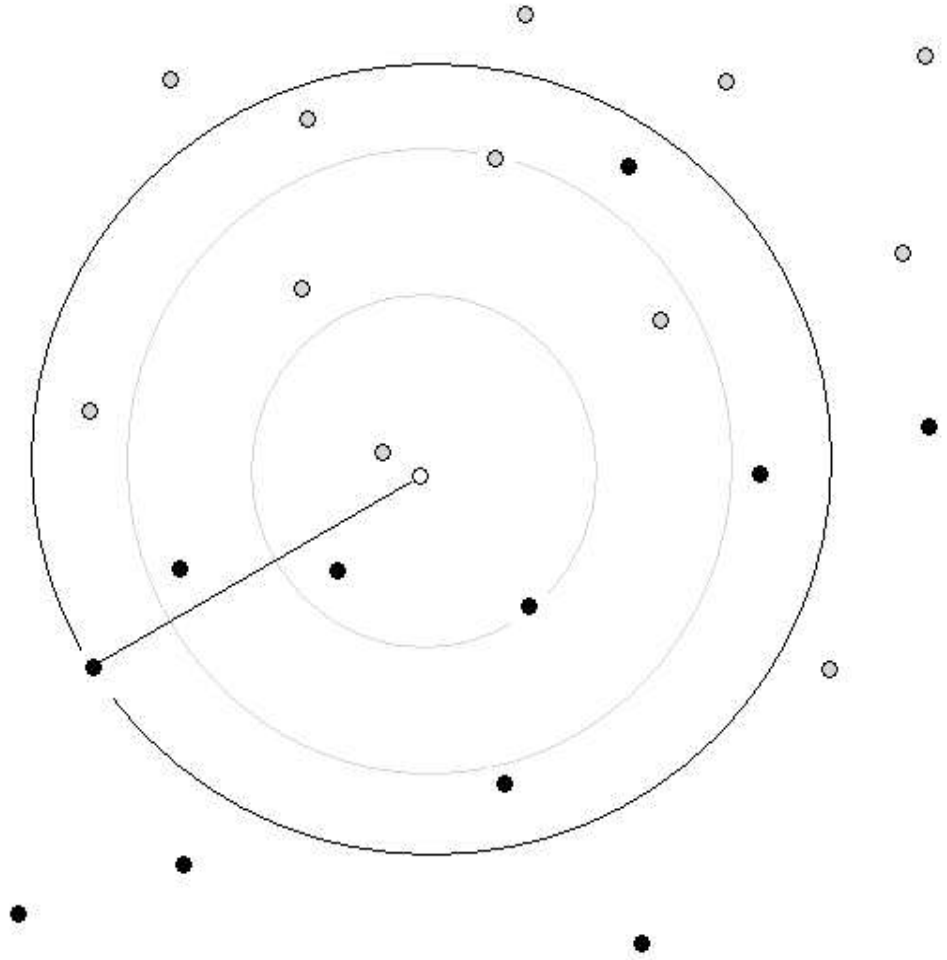
$f_j(\mathbf{x})$: density functions, π_j : prior probabilities.

- Nearest neighbor density estimates : $\hat{f}_j^{(v)}(\mathbf{x}) = n_j^{(v)}/(n_j \times v)$
- Estimates for priors : $\hat{\pi}_j = n_j/N$
- If same neighborhood is used for all populations, it leads to a classification in favor of the most frequent class in that neighborhood.

  $\mathrm{d}_{NN}(\mathbf{x})$: arg $\max_j \left\{ n_j^{(v)} \right\}$.

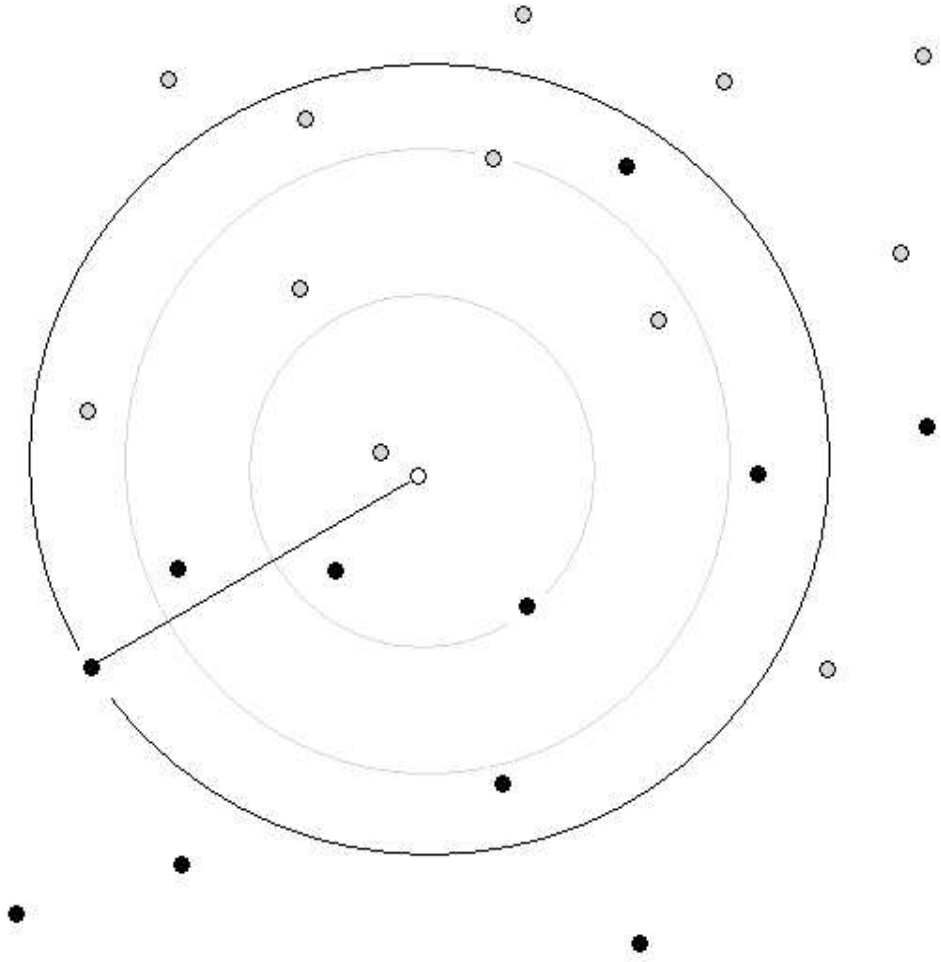
LIGO

NN Classification (contd..)



LIGO

NN Classification (contd..)



What is the optimum value of neighborhood parameter k ?

LIGO

NN Classification using multiple values of k

- No estimation for optimum neighborhood parameter

LIGO

NN Classification using multiple values of k

- No estimation for optimum neighborhood parameter
- Simultaneous study of classification results (discrimination measures) for all possible values of k

LIGO

NN Classification using multiple values of k

- No estimation for optimum neighborhood parameter
- Simultaneous study of classification results (discrimination measures) for all possible values of k
- Aggregation of these results for final classification

LIGO

Discrimination measures

Grey-scale values of discrimination measures are plotted for different classes and different values of k.

LIGO

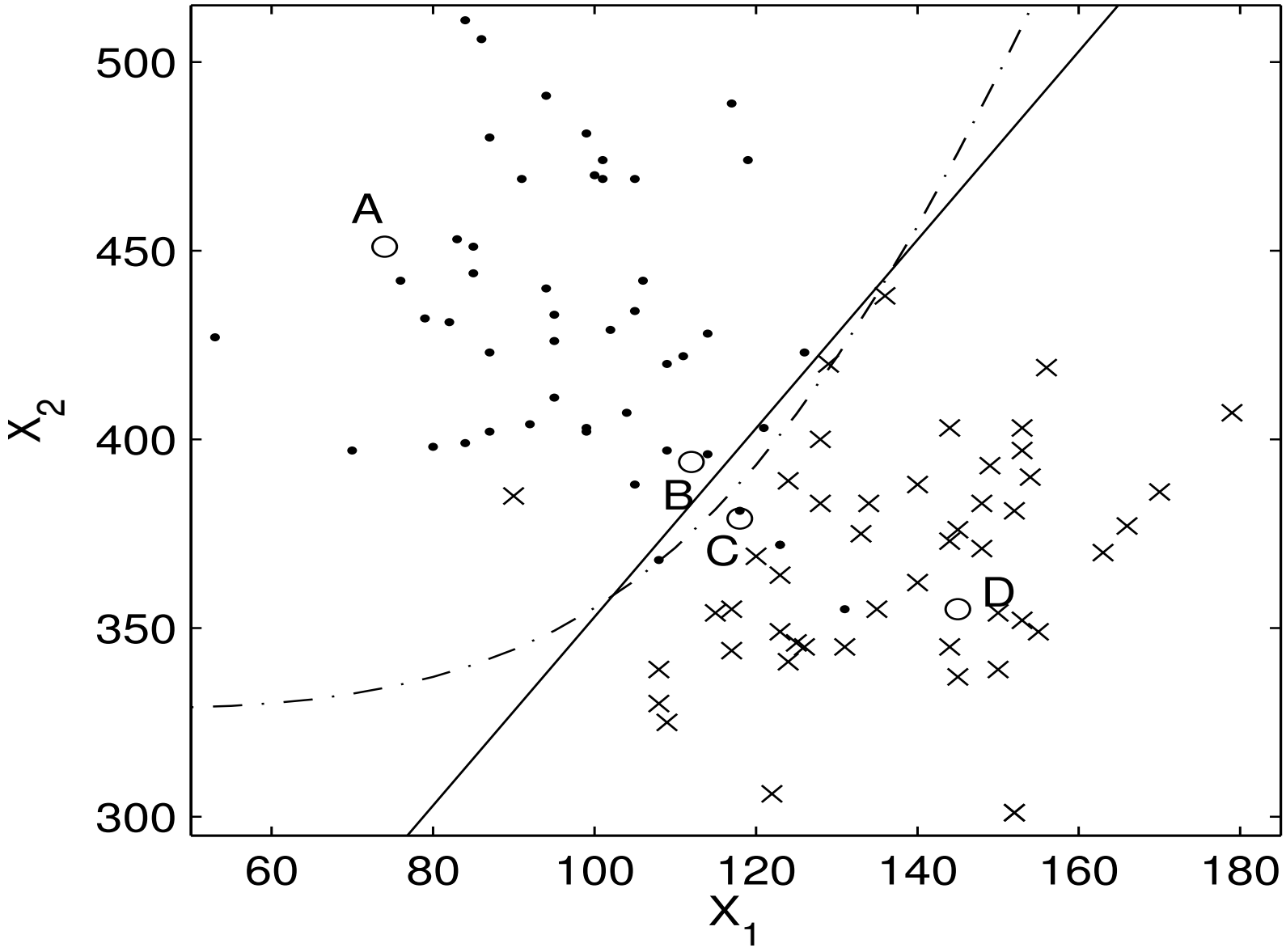
Discrimination measures

Grey-scale values of discrimination measures are plotted for different classes and different values of k.

- **Posterior probability :**
  $\hat{p}_k(j \mid \mathbf{x})$ = Proportion of class j observations among the k nearest neighbors of **x**.

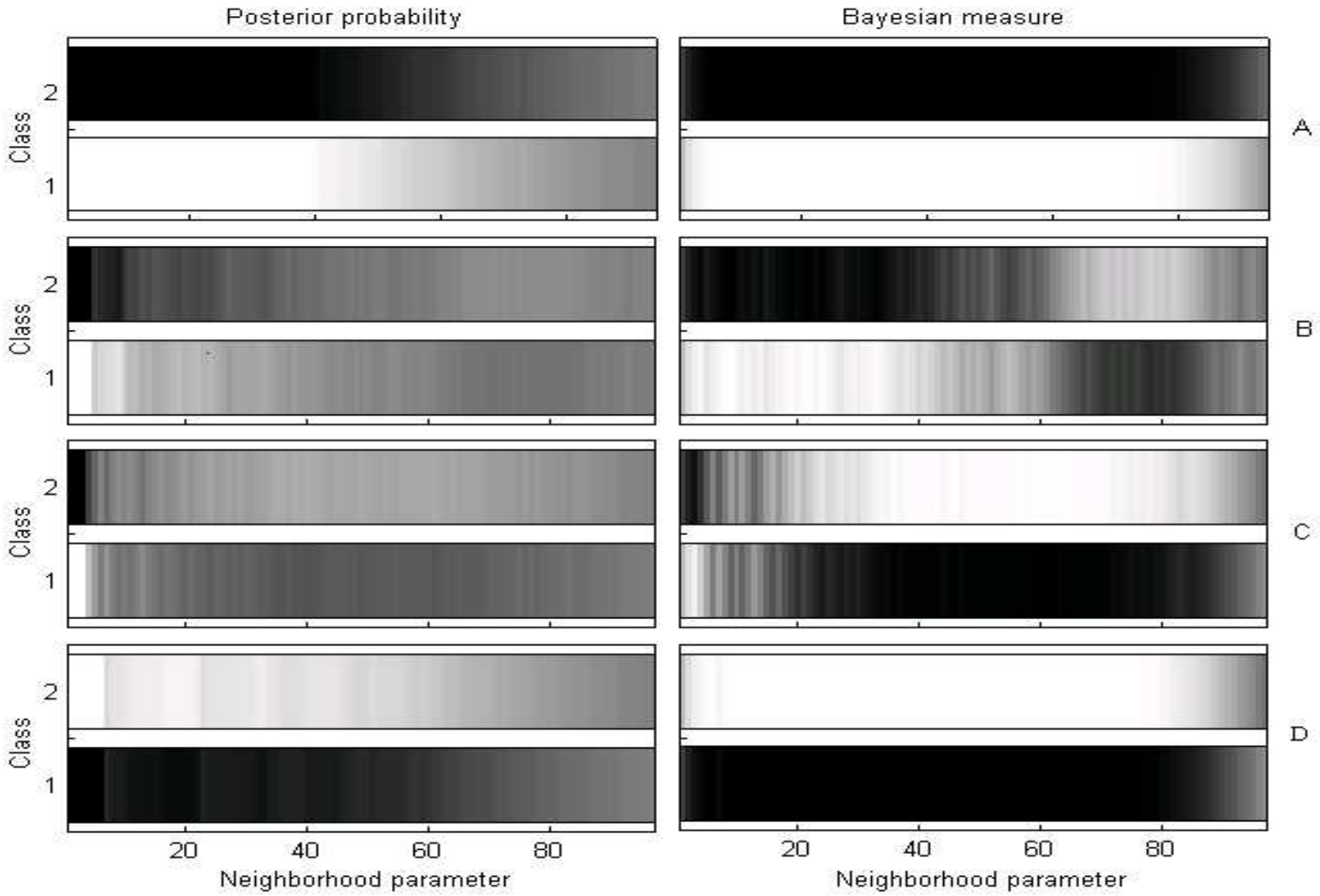
LIGO

Salmon data

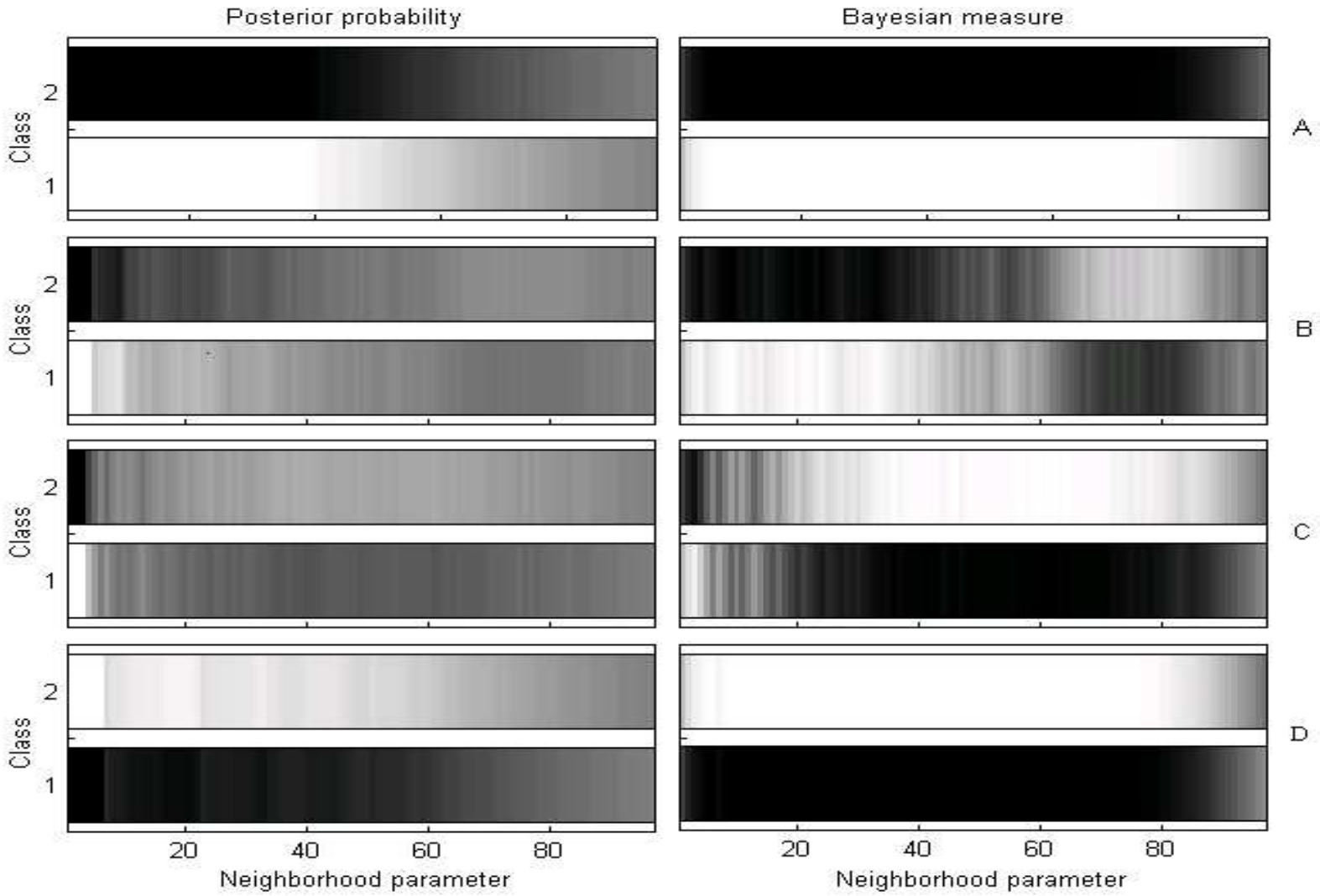


LIGO

Multi-scale visualization



Multi-scale visualization



Bayesian measure sharpens the plot without loss of information

LIGO

Bayesian Strength Function

- **Bayesian measure of strength :**

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

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

and $\varphi(\mathbf{t}_k \mid \mathbf{p}, k) = \dfrac{k!}{t_{1_k}!\ t_{2_k}!\ \ldots\ t_{J_k}!} \prod_{j=1}^{J} p_j^{t_{j_k}}.$

LIGO

Bayesian Strength Function

- **Bayesian measure of strength :**

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

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

and $\varphi(\mathbf{t}_k \mid \mathbf{p}, k) = \dfrac{k!}{t_{1_k}!\ t_{2_k}!\ \ldots\ t_{J_k}!} \prod_{j=1}^{J} p_j^{t_{j_k}}.$

Result : If $\pi(\mathbf{p})$ is symmetric in $p_1, p_2, \ldots, p_J$,
$\hat{p}_k(j \mid \mathbf{x}) \geq \hat{p}_k(i \mid \mathbf{x}) \Leftrightarrow S(j \mid k) \geq S(i \mid k)$.

Large sample properties

For a given **x**, define

$$P_j(\mathbf{x}) = \pi_j f_j(\mathbf{x}) / \sum_{i=1}^{J} \pi_i f_i(\mathbf{x}) \quad \text{for } j = 1, 2, \ldots, J$$

where π_j's are prior probabilities and f_j's are continuous density functions.

LIGO

Large sample properties

For a given **x**, define

$$P_j(\mathbf{x}) = \pi_j f_j(\mathbf{x}) / \sum_{i=1}^{J} \pi_i f_i(\mathbf{x}) \quad \text{for } j = 1, 2, \ldots, J$$

where π_j's are prior probabilities and f_j's are continuous density functions.

Assume that
(i) $P_i(\mathbf{x}) > P_j(\mathbf{x})$ for all $j \neq i$, and
(ii) $k \to \infty$ and $k/N \to 0$ as $N \to \infty$.

Large sample properties

For a given **x**, define

$$P_j(\mathbf{x}) = \pi_j f_j(\mathbf{x}) / \sum_{i=1}^{J} \pi_i f_i(\mathbf{x}) \quad \text{for } j = 1, 2, \ldots, J$$

where π_j's are prior probabilities and f_j's are continuous density functions.

Assume that
(i) $P_i(\mathbf{x}) > P_j(\mathbf{x})$ for all $j \neq i$, and
(ii) $k \to \infty$ and $k/N \to 0$ as $N \to \infty$.

Result : If $\pi(\mathbf{p})$ is symmetric in its arguments,

(i) $S(i \mid k) \xrightarrow{P} 1$ $\qquad$ (ii) $S(j \mid k) \xrightarrow{P} 0$ for all $j \neq i$.

LIGO

Aggregation of classifiers

- **Aggregated classification rule :**
  - $\mathrm{d}_{NN^*}(\mathbf{x}) : \underset{j}{\mathrm{argmax}} \sum_k w(k)\ \hat{p}_k(j \mid \mathbf{x})$
  - $\mathrm{d}_{NN^*}(\mathbf{x}) : \underset{j}{\mathrm{argmax}} \sum_k w(k)\ S(j \mid k)$,

    where $w(k)$ is the weight of the k-nearest neighbor classifier.

LIGO

Aggregation of classifiers

- **Aggregated classification rule :**
  - $\mathrm{d}_{NN^*}(\mathbf{x}) : \operatorname{argmax}_j \sum_k w(k)\ \hat{p}_k(j \mid \mathbf{x})$
  - $\mathrm{d}_{NN^*}(\mathbf{x}) : \operatorname{argmax}_j \sum_k w(k)\ S(j \mid k)$,

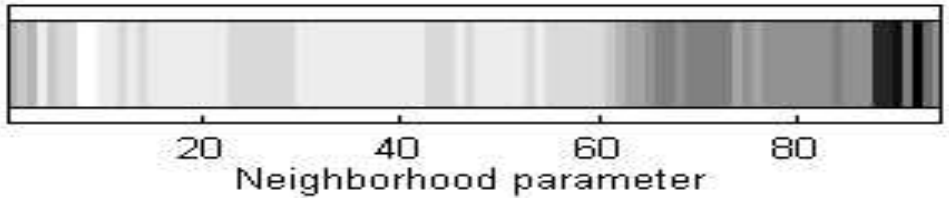
    where $w(k)$ is the weight of the k-nearest neighbor classifier.
- The weight function $w(k)$ is a decreasing function of the misclassification probability $\Delta(k)$.

Aggregation of classifiers

- **Aggregated classification rule :**
  - $\mathrm{d}_{NN^*}(\mathbf{x}) : \underset{j}{\operatorname{argmax}} \sum_k w(k)\ \hat{p}_k(j \mid \mathbf{x})$
  - $\mathrm{d}_{NN^*}(\mathbf{x}) : \underset{j}{\operatorname{argmax}} \sum_k w(k)\ S(j \mid k)$,

    where $w(k)$ is the weight of the k-nearest neighbor classifier.
- The weight function $w(k)$ is a decreasing function of the misclassification probability $\Delta(k)$.
- Leave-one-out cross-validation method is used to estimate $\Delta(k)$ for different values of k.

Plot for probability of correct classification



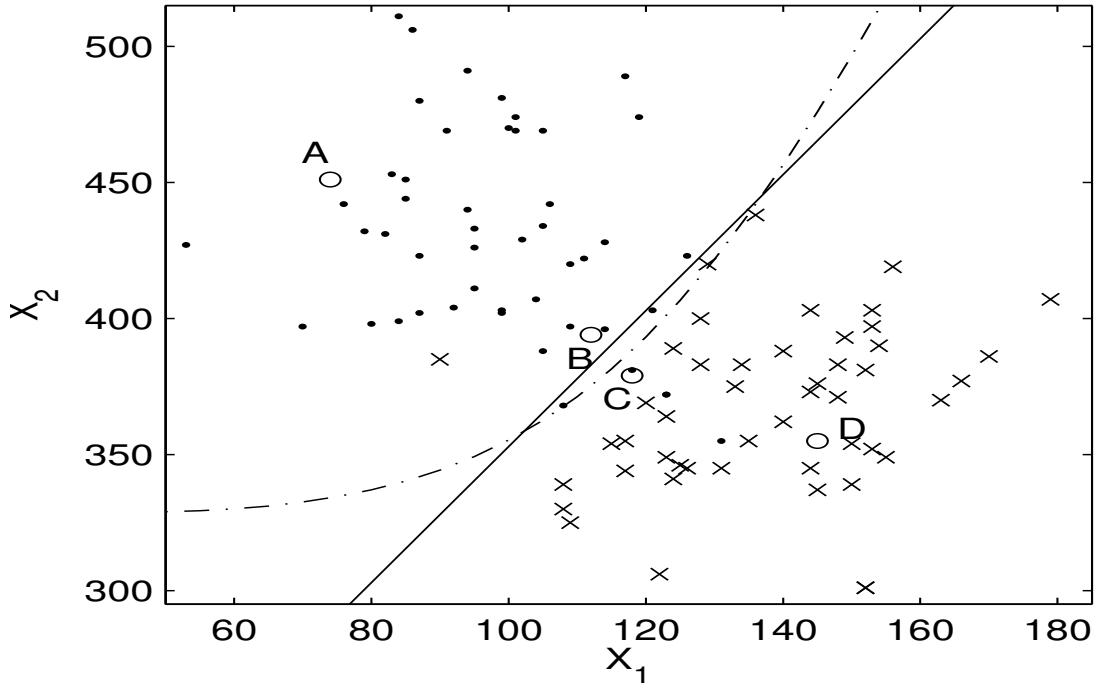
Aggregation of classifiers (contd..)

Weight function :

$$w(k) = \begin{cases} \exp\left\{-\frac{1}{2}\frac{\left(\widehat{\Delta}(k)-\Delta_0\right)^2}{\Delta_0(1-\Delta_0)/N}\right\} & \text{if } \frac{\widehat{\Delta}(k)-\Delta_0}{[\Delta_0(1-\Delta_0)/N]^{1/2}} \leq \tau \\ 0 & \text{otherwise,} \end{cases}$$

- $\tau = 0$: Assignment of whole weight on those k-nearest neighbor classifiers for which $\widehat{\Delta}(k) = \Delta_0$
- $\tau = 3$: Maximum value of τ to be used in practice

Aggregation of classifiers (contd..)



	A	B	C	D
Cross-validation	1	6/7	4/7	1/7
Wt. posterior	0.8804	0.6261	0.4146	0.1445
Wt. strength	0.9712	0.7777	0.2808	0.0134

LIGO

Results on benchmark data sets : comparison with cross validation method

Data sets	k-NN (cross-valid.)	Weighted posterior	Weighted strength
Salmon	9.38 (0.18)	8.30 (0.14)	8.30 (0.15)
Wine	1.05 (0.07)	0.45 (0.05)	0.41 (0.04)
Vowel-2	17.75 (2.09)	18.93 (2.15)	19.25 (2.16)
Diabetes	11.50 (0.16)	10.47 (0.15)	10.71 (0.15)
Biomedical	17.61 (0.18)	17.10 (0.16)	17.56 (0.17)

LIGO

Other aggregation techniques

- Prob. NN (Holmes and Adams, 2002; JRSS-B)
  - For fixed k and logistic parameter β, consider a logistic regression model for conditional probabilities of different classes.
  - Consider a prior distribution for k and β.
  - Use MCMC to generate samples from posterior distribution of (k, β) and hence to find aggregated posterior probability for different classes.

LIGO

Other aggregation techniques

- Prob. NN (Holmes and Adams, 2002; JRSS-B)
  - For fixed k and logistic parameter β, consider a logistic regression model for conditional probabilities of different classes.
  - Consider a prior distribution for k and β.
  - Use MCMC to generate samples from posterior distribution of (k, β) and hence to find aggregated posterior probability for different classes.
- Likelihood based approach (HA, 2003; Biometrika)
  - Consider a logistic model for posteriors that uses multiple values of k simultaneously.
  - Use iterative re-weighted least squares method to get MLE of logistic regression parameters and hence the estimate of posteriors.

LIGO

Comparison with the aggregation techniques of Holmes & Adams

Data sets	k-NN (cross valid.)	Likelihood (H.A.'03)	Prob. NN (H.A.'02)	Weighted posterior	Weighted strength
Synthetic	11.70 (1.02)	8.2	8.4	9.80 (0.94)	9.90 (0.94)
Vowel-1	46.75 (2.32)	49.3	—	46.75 (2.32)	46.75 (2.32)
Pima	25.27 (0.25)	23.9	24.7	24.48 (0.24)	24.64 (0.24)
Australian	13.20 (0.23)	13.3	14.7	13.16 (0.24)	12.97 (0.23)