Adapted Feature Extraction and Its Applications

Naoki Saito
Department of Mathematics
University of California
Davis, CA 95616
email: saito@math.ucdavis.edu
URL: http://www.math.ucdavis.edu/∼saito/

Acknowledgment

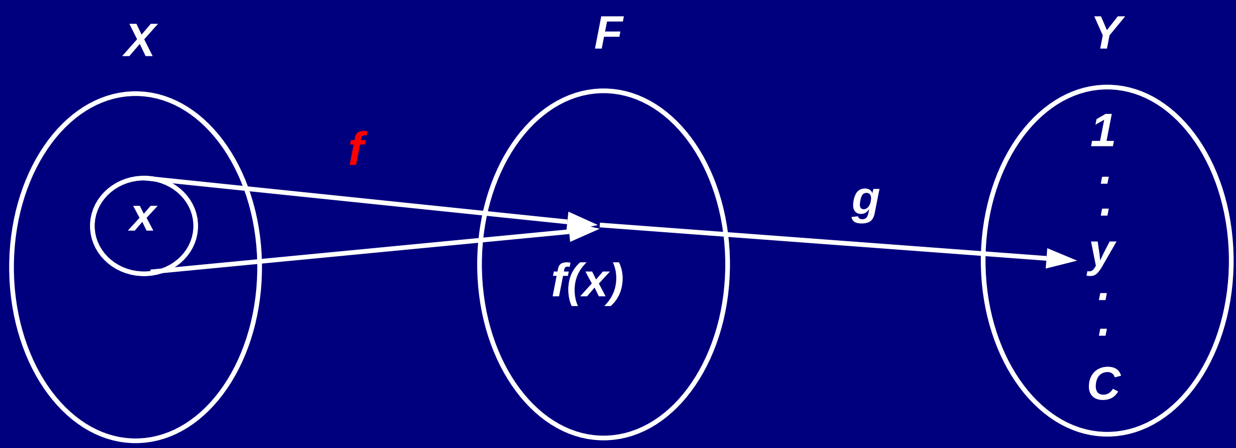
- Ronald R. Coifman (Yale/FMAH)
- Fred Warner (Yale/FMAH)
- Frank Geshwind (Yale/FMAH)

Outline

- Problem Formulation
- A Dictionary/Library of Orthonormal Bases
- Local Discriminant Basis (LDB)
- Improved LDB with Empirical Probability Density Estimation
- Example 1: Synthetic “Waveform” Classification
- Example 2: Geophysical Acoustic Waveform Classification
- Conclusion
- Future Directions

A Strategy for Pattern Recognition

- Normalization of input patterns – scaling, translation, rotation
- Feature extraction and selection
- Classification – LDA, CART, k-NN, Neural Networks, . . .



Problem Formulation

Learning (or Training): Given a set of data,

$$\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y},$$

where $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} = \{1, \ldots, K\}$ (class labels), find a *map* $d : \mathcal{X} \to \mathcal{Y}$ such that

$$R(d) = \frac{1}{N} \sum_{i=1}^{N} I(d(\boldsymbol{x}_i) \neq y_i) \to \text{small.}$$

Prediction: Apply the map d obtained during the training to a new dataset. Then interpret the result and evaluate d.

Problem Formulation …

- Let $(\boldsymbol{X}, Y) \in (\mathcal{X}, \mathcal{Y})$ be a random sample from

$$P(\boldsymbol{X} \in A, Y = y) = P(\boldsymbol{X} \in A \mid Y = y)\, P(Y = y),$$

where $P(Y = y) = \pi_y = N_y/N$ in practice.

- Let us assume the density $p(\boldsymbol{x} \mid y)$ exists.

- The Bayes classifier (or rule) d_B is:

$$d_B(\boldsymbol{x}) = kI(\boldsymbol{x} \in A_k),$$

where $A_k = \{\boldsymbol{x} \in \mathcal{X} : p(\boldsymbol{x} \mid k)\pi_k = \max_{y \in \mathcal{Y}} p(\boldsymbol{x} \mid y)\pi_y\}$.
In other words, "assign $\boldsymbol{x}$ to class k if $\boldsymbol{x} \in A_k$".
Note $\bigcup_{y \in \mathcal{Y}} A_y = \mathcal{X}$.

Problem Formulation …

Difficulties of the Bayes classifier:

- Don't know $p(\boldsymbol{x} \mid k)$ or
- Too difficult to estimate $p(\boldsymbol{x} \mid k)$ computationally because of the high dimensionality of the problem (*curse of dimensionality*).

$\Longrightarrow$ Need dimensionality reduction without loosing important information for classification.

Problem Formulation . . .

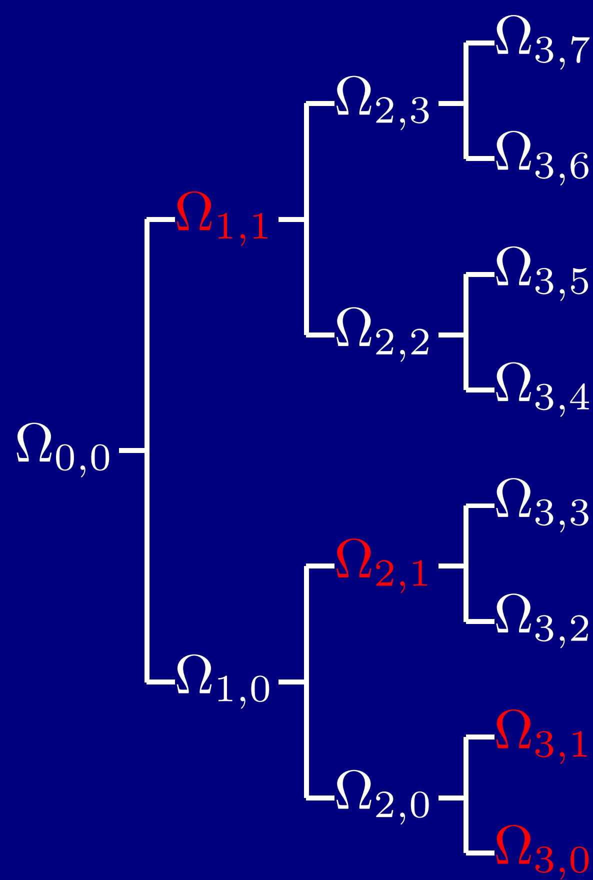
Therefore, we restrict our attention to the map $d : \mathcal{X} \to \mathcal{Y}$ of the following form:

$$d = g \circ f = g \circ \Theta_m \circ \Psi^T,$$

- $f : \mathcal{X} \to \mathcal{F} \subset \mathbb{R}^m$ is called a feature extractor consisting of:
  - Ψ: an n-dimensional orthogonal matrix selected from a dictionary or library of orthonormal bases.
  - Θ_m: a selection rule: it selects the most important $m\ (\leq n)$ coordinates from n-dimensional coordinates.
- $g : \mathcal{F} \to \mathcal{Y}$ is a conventional classifier, e.g., LDA, CART, ANN etc.

A Library of Orthonormal Bases

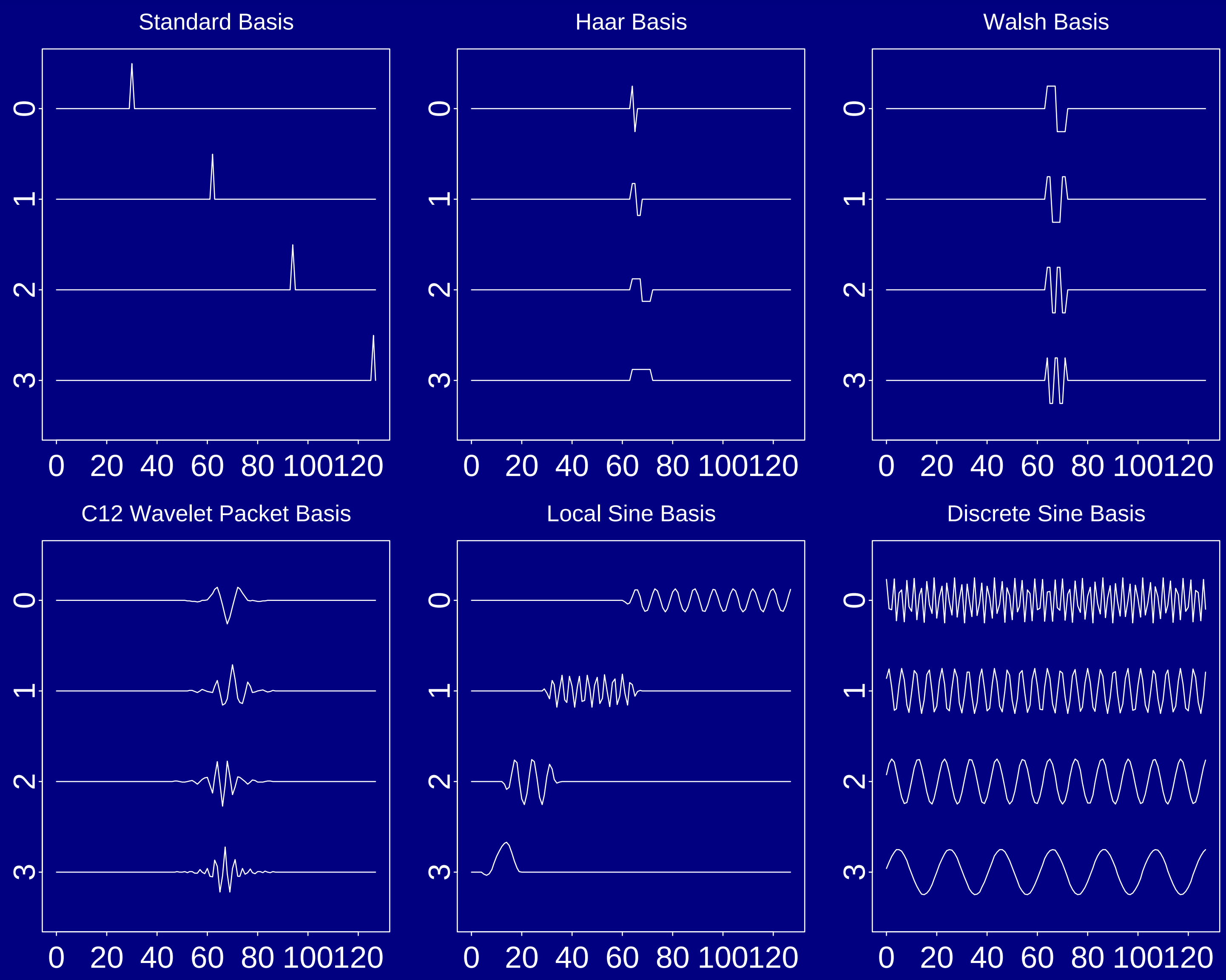
consists of dictionaries of orthonormal bases: each dictionary is a binary tree whose nodes are subspaces of $\Omega_{0,0} = \mathbb{R}^n$ with different time-frequency localization characteristics.



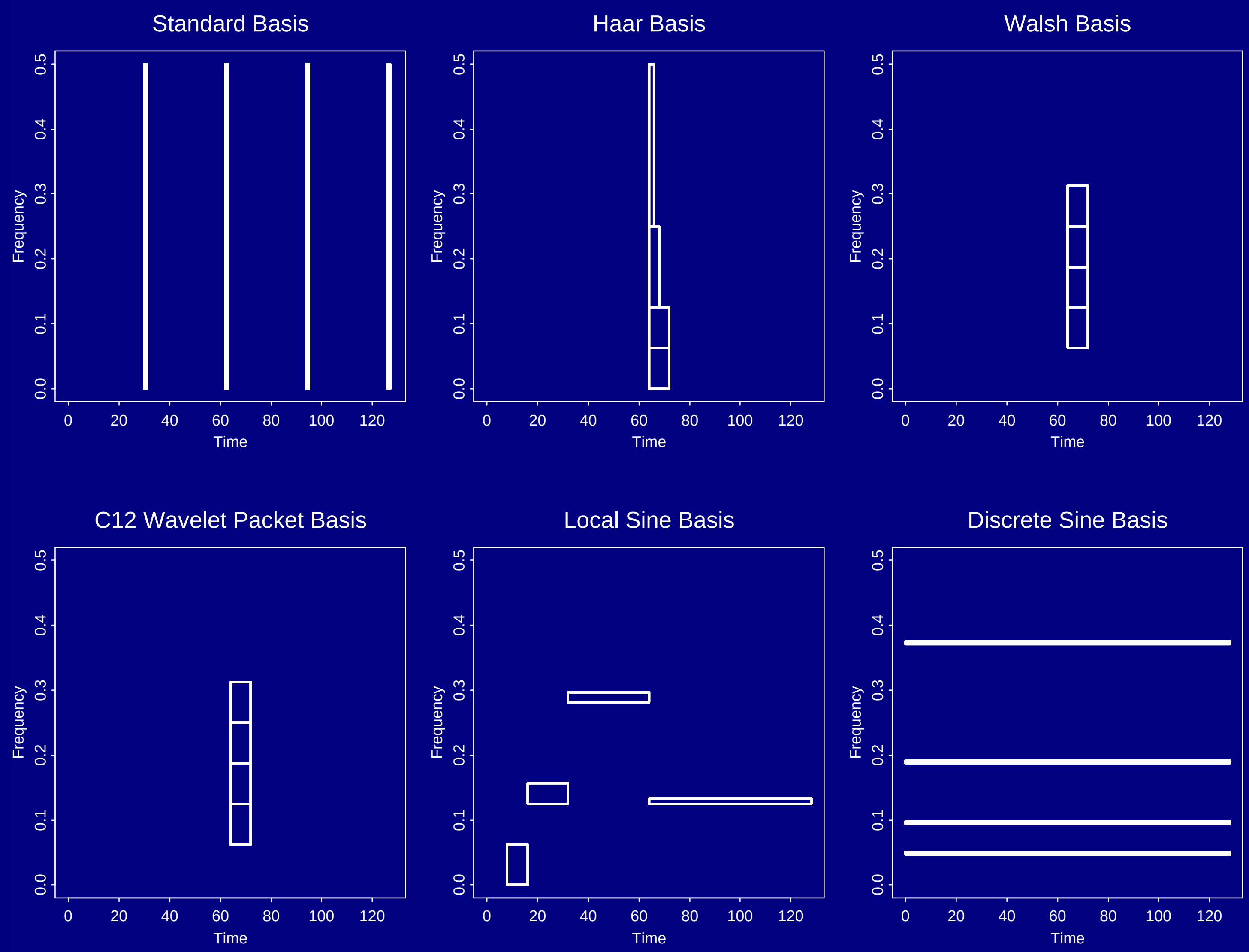
A Library of Orthonormal Bases . . .

- Examples of dictionaries include wavelet packet bases, local trigonometric bases, and local Fourier bases.
- It costs $O(n[\log n]^p)$ to generate a dictionary for a signal of length n ($p = 1$ for wavelet packets, $p = 2$ for LTB/LFB).
- Each dictionary may contain up to $n(1 + \log_2 n)$ basis vectors and more than 2^n possible orthonormal bases.
- How to select the best possible basis for the problem at hand is a key issue.

Example of Local Basis Functions



Time-Frequency Characteristics



Local Discriminant Basis

- Let $\mathcal{D} = \{\boldsymbol{w}_i\}_{i=1}^{N_w}$ be a time-frequency dictionary
- $\mathcal{D} = \{B_j\}_{j=1}^{N_B}$ as a list of all possible orthonormal bases where $B_j = (\boldsymbol{w}_{j_1}, \ldots, \boldsymbol{w}_{j_n})$
- Let $\mathcal{M}^+(B_j)$ be a measure of efficacy of B_j for discriminantion
- Then the local discriminant basis (LDB) can be written as

$$\Psi = \arg\max_{B_j \in \mathcal{D}} \mathcal{M}^+(B_j).$$

- **Proposition:** There exists a fast algorithm (divide-and-conquer, $O(n)$) to find Ψ from each dictionary $\mathcal{D}$ if $\mathcal{M}^+$ is additive.

$$\mathcal{M}^+(0) = 0, \quad \mathcal{M}^+(\{\boldsymbol{w}_{j_1}, \ldots, \boldsymbol{w}_{j_n}\}) = \sum_{i=1}^{n} \mathcal{M}^+(\boldsymbol{w}_{j_i}).$$

Discriminant Measures

- For $\boldsymbol{w}_i \in \mathcal{D}$, consider the projection $Z_i \triangleq \boldsymbol{w}_i \cdot \boldsymbol{X}$ of an input random signal $\boldsymbol{X} \in \mathcal{X}$
- Let δ_i be the efficacy for discrimination (or discriminant power) of $\boldsymbol{w}_i$
- Some possibilities of measuring the efficacy of the basis B_j:
- $\mathcal{M}^+(B_j) = \sum_{i=1}^n \delta_{j_i}$
- $\mathcal{M}^+(B_j) = \sum_{i=1}^k \delta_{(j_i)}$, where $\{\delta_{(j_i)}\}$ is the decreasing rearrangement of $\{\delta_{j_i}\}$ and $k < n$.
- $\mathcal{M}^+(B_j) = \sum_{i=1}^n \varepsilon_{j_i}\delta_{j_i}$, where $\varepsilon_{j_i} = 1$ if $E[Z_i^2] > \theta$, $= 0$ otherwise.

Discriminant Measures . . .

- What are the basic quantities to use for 1D efficacy?
- Differences in normalized energies of Z_i among classes:

$$V_i^{(y)} \triangleq \frac{E[Z_i^2 \,|\, Y = y]}{\sum_{i=1}^{n} E[Z_i^2 \,|\, Y = y]} \to \hat{V}_i^{(y)} = \frac{\sum_{k=1}^{N_y} |\boldsymbol{w}_i \cdot \boldsymbol{x}_k^{(y)}|^2}{\sum_{k=1}^{N_y} \|\boldsymbol{x}_k^{(y)}\|^2}$$

- Differences in probability density functions (pdfs) of Z_i:

$$q_i^{(y)}(z) \triangleq \int_{\boldsymbol{w}_i \cdot \boldsymbol{x} = z} p(\boldsymbol{x} \,|\, y)\, d\boldsymbol{x} \to \hat{q}_i^{(y)}(z) \quad \text{via e.g., ASH}$$

Discriminant Measures ...

- Some possibilites of "discrepancy" measures between two pdfs $p(x)$, $q(x)$:
- Relative entropy [Kullback-Leibler divergence]:

$$D_{KL}(p, q) = \int_{-\infty}^{\infty} p(x) \log_2 \frac{p(x)}{q(x)} \, dx$$

- Hellinger distance:

$$D_H(p, q) = \int_{-\infty}^{\infty} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \, dx$$

- L^2 distance:

$$D_2(p, q) = \int_{-\infty}^{\infty} \left(p(x) - q(x) \right)^2 \, dx$$

Discriminant Measures …

- Using normalized energy:

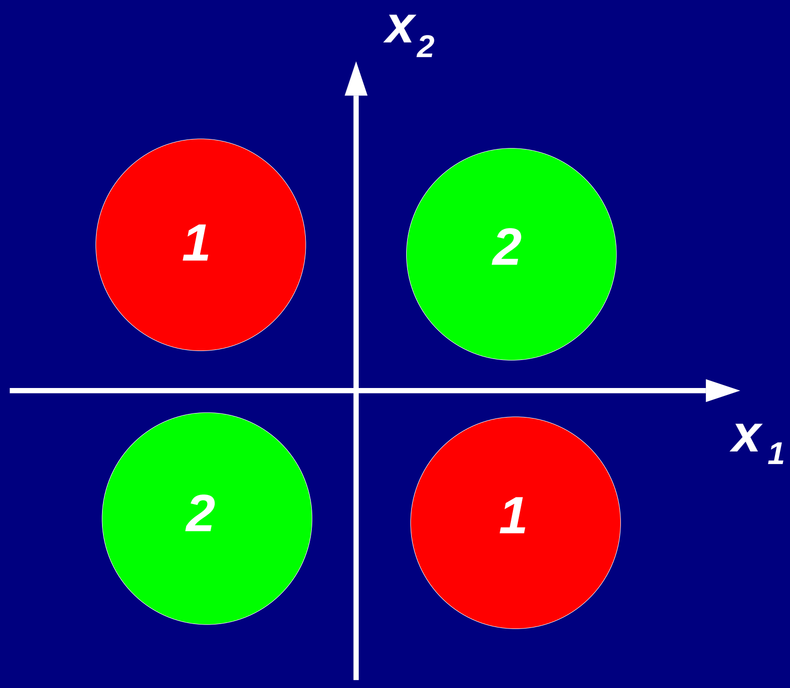
$$\delta_i = \hat{V}_i^{(1)} \log_2 \frac{\hat{V}_i^{(1)}}{\hat{V}_i^{(2)}}, \left(\sqrt{\hat{V}_i^{(1)}} - \sqrt{\hat{V}_i^{(2)}}\right)^2, \text{or} \left(\hat{V}_i^{(1)} - \hat{V}_i^{(2)}\right)^2$$

- Using empirical pdfs:

$$\delta_i = D_{KL}(\hat{q}_i^{(1)}, \hat{q}_i^{(2)}), D_H(\hat{q}_i^{(1)}, \hat{q}_i^{(2)}), \text{or} D_2(\hat{q}_i^{(1)}, \hat{q}_i^{(2)})$$

Discriminant Measures. . .

- Important to consider two-dimensional projection



- Applicable to the complex coefficients of local Fourier bases

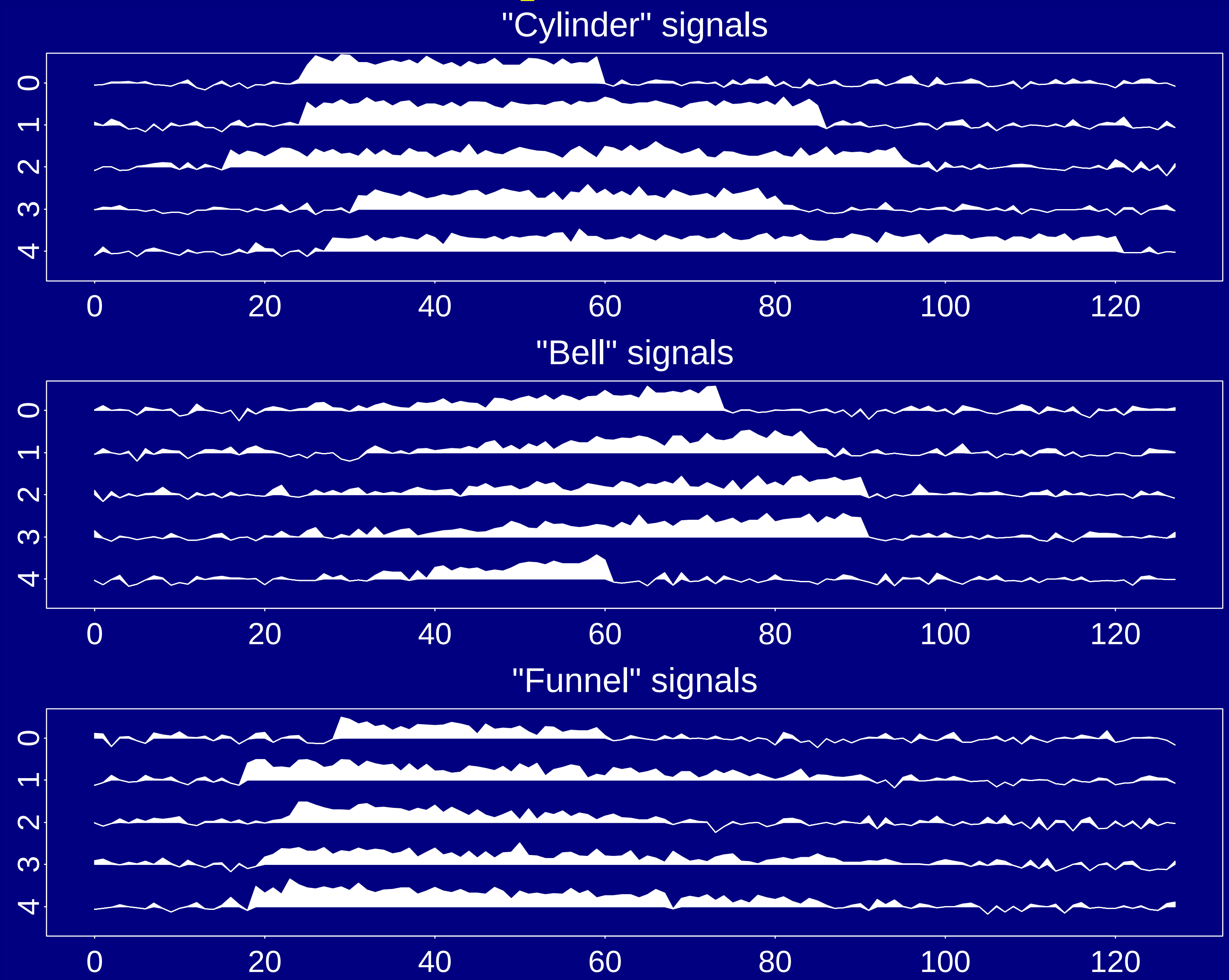
Example 1: Signal Shape Classification

Objective: Classify synthetic signals of length 128 to three possible classes,

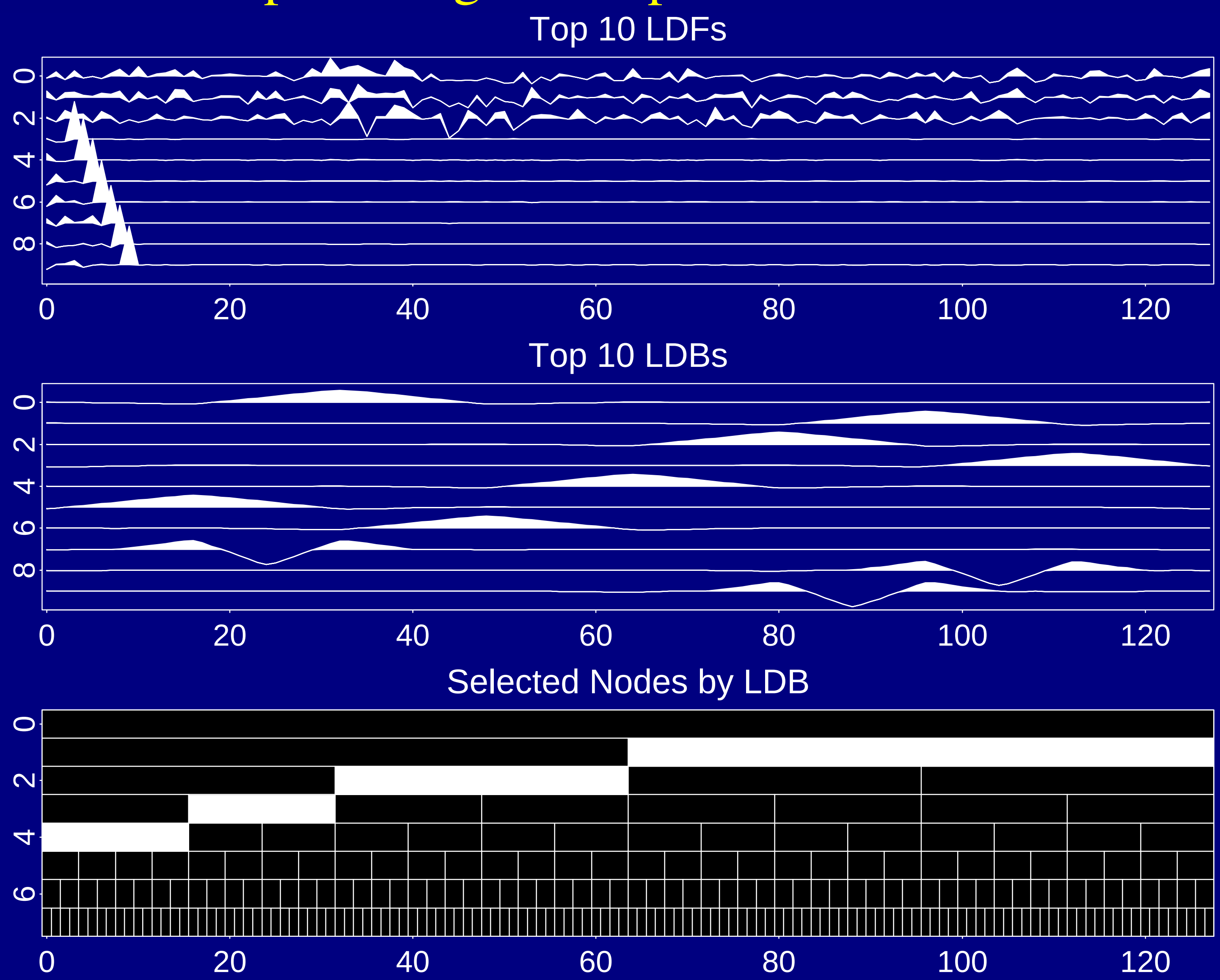
$$
\begin{aligned}
c(i) &= (6+\eta)\cdot\chi_{[a,b]}(i) + \epsilon(i) && \text{for "cylinder,"}\\
b(i) &= (6+\eta)\cdot\chi_{[a,b]}(i)\cdot(i-a)/(b-a) + \epsilon(i) && \text{for "bell,"}\\
f(i) &= (6+\eta)\cdot\chi_{[a,b]}(i)\cdot(b-i)/(b-a) + \epsilon(i) && \text{for "funnel."}
\end{aligned}
$$

where $a = \mathrm{unif}[16, 32],\ b - a = \mathrm{unif}[32, 96]$,
$\eta \sim \mathcal{N}(0, 1),\ \epsilon(i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

Example Waveforms



Example 1: Signal Shape Classification . . .



Example 1: Signal Shape Classification …

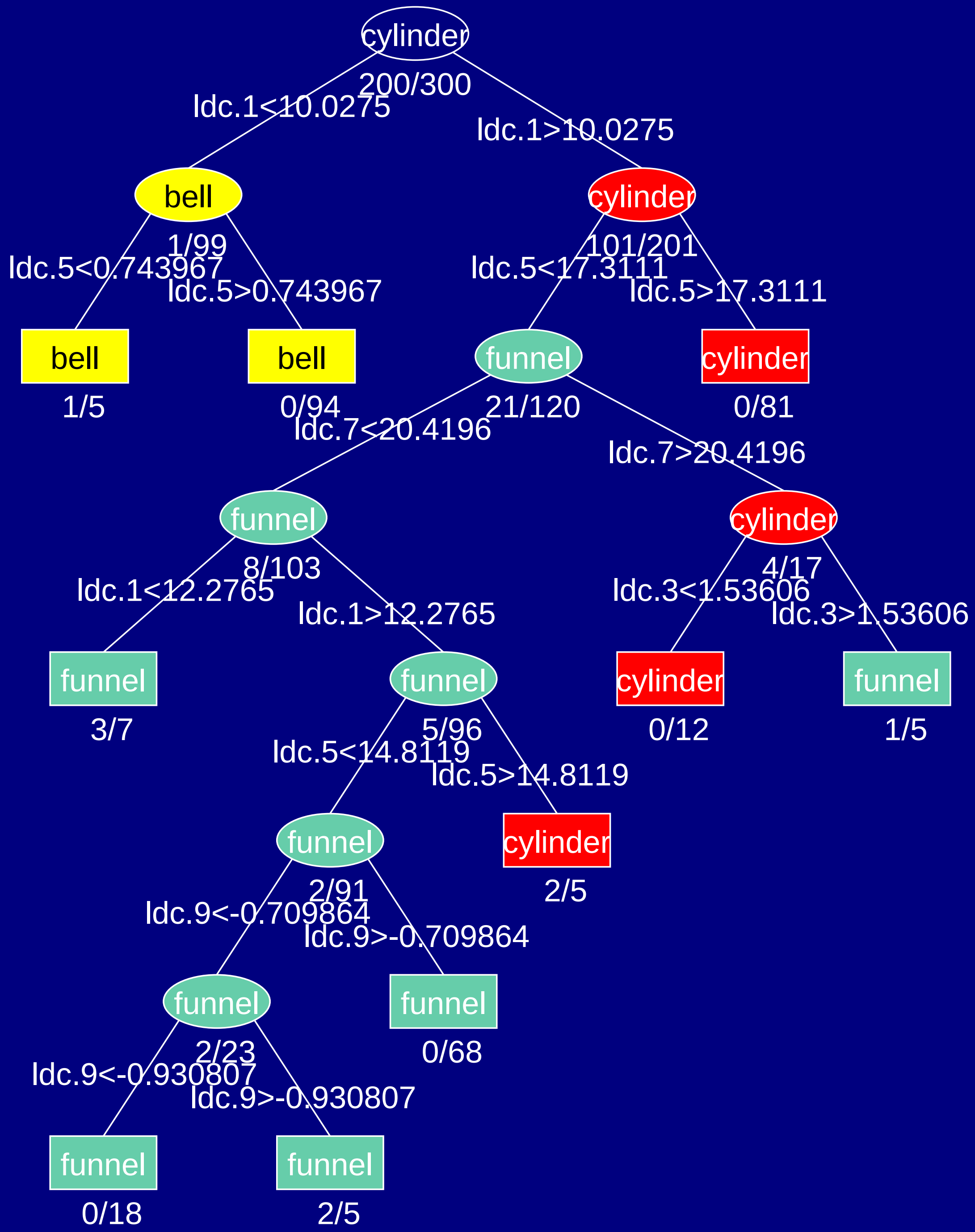
Table of Misclassifications (average over 10 simulations)

Method (Coordinates)	Training Data	Test Data
LDA on STD	0.83 %	12.31 %
CT on STD	2.83 %	11.28 %
LDA on Top 10 LDB	7.00 %	8.37 %
CT on Top 10 LDB	2.67 %	5.54 %

- 100 training signals and 1000 test signals were generated for each class per simulation
- 12-tap Coiflet filter was used for LDB

Example 1: Signal Shape Classification . . .

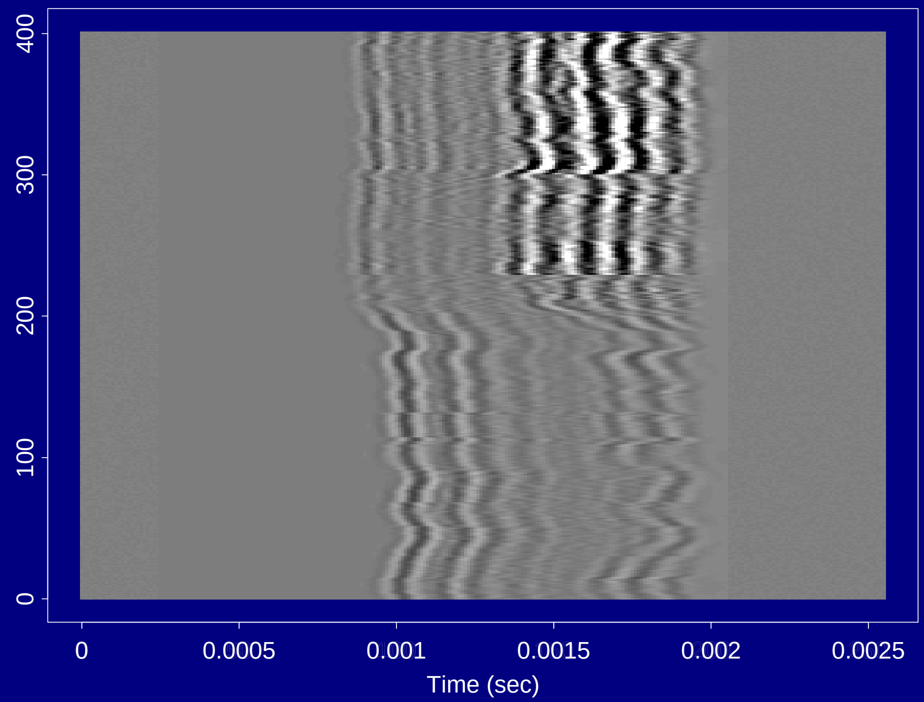
Full Classification Tree on Top 10 LDB Coordinates



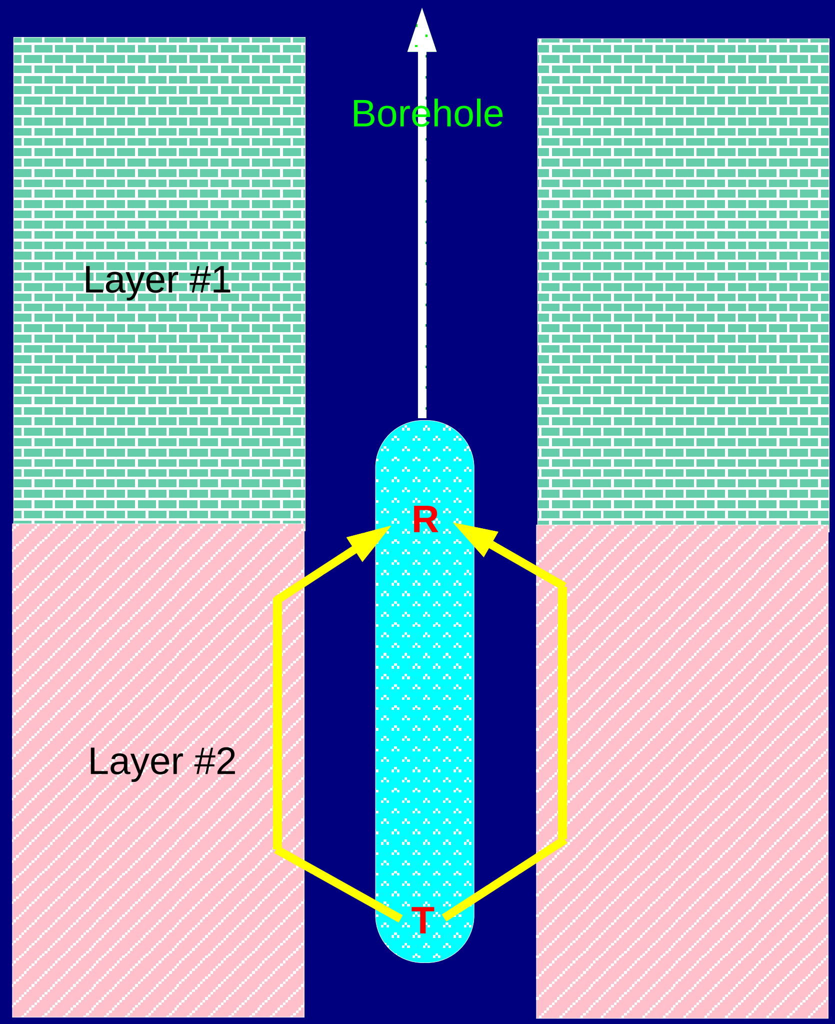
saito@math.ucdavis.edu

Example 2: Classification of Geophysical Acoustic Waveforms

- 402 acoustic waveforms (256 time samples) were recorded at a gas producing well.
- Region consists of sand or shale layers.
- Sand layers contain either water or gas.

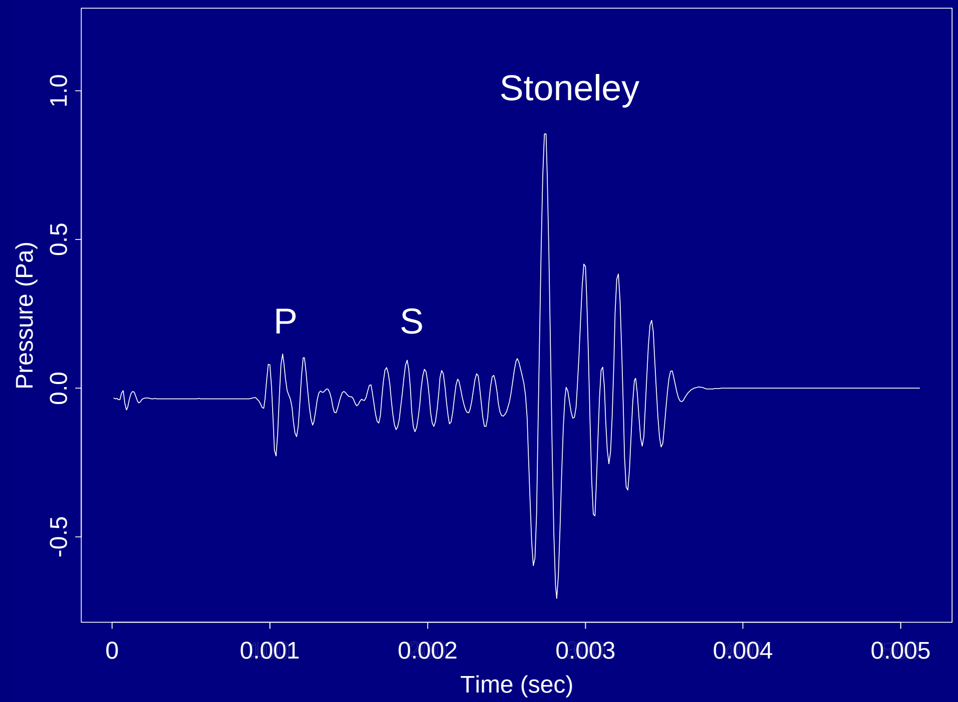


Sonic Waveform Measurement



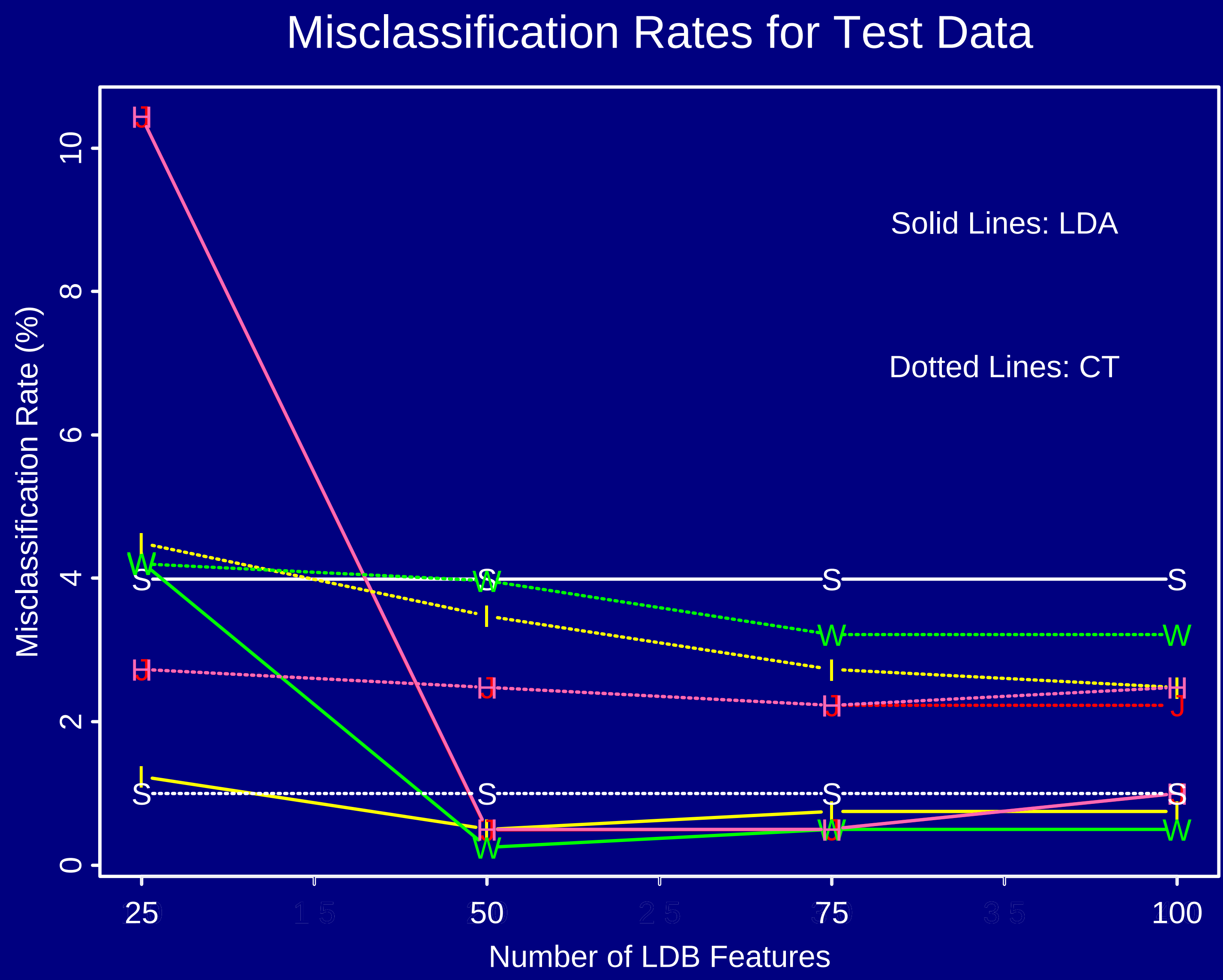
Objectives

- Can we classify acoustic waveforms in terms of mineral contents of layers?
- Can we automate the classification process?
- If so, what features (or wave components) are important?

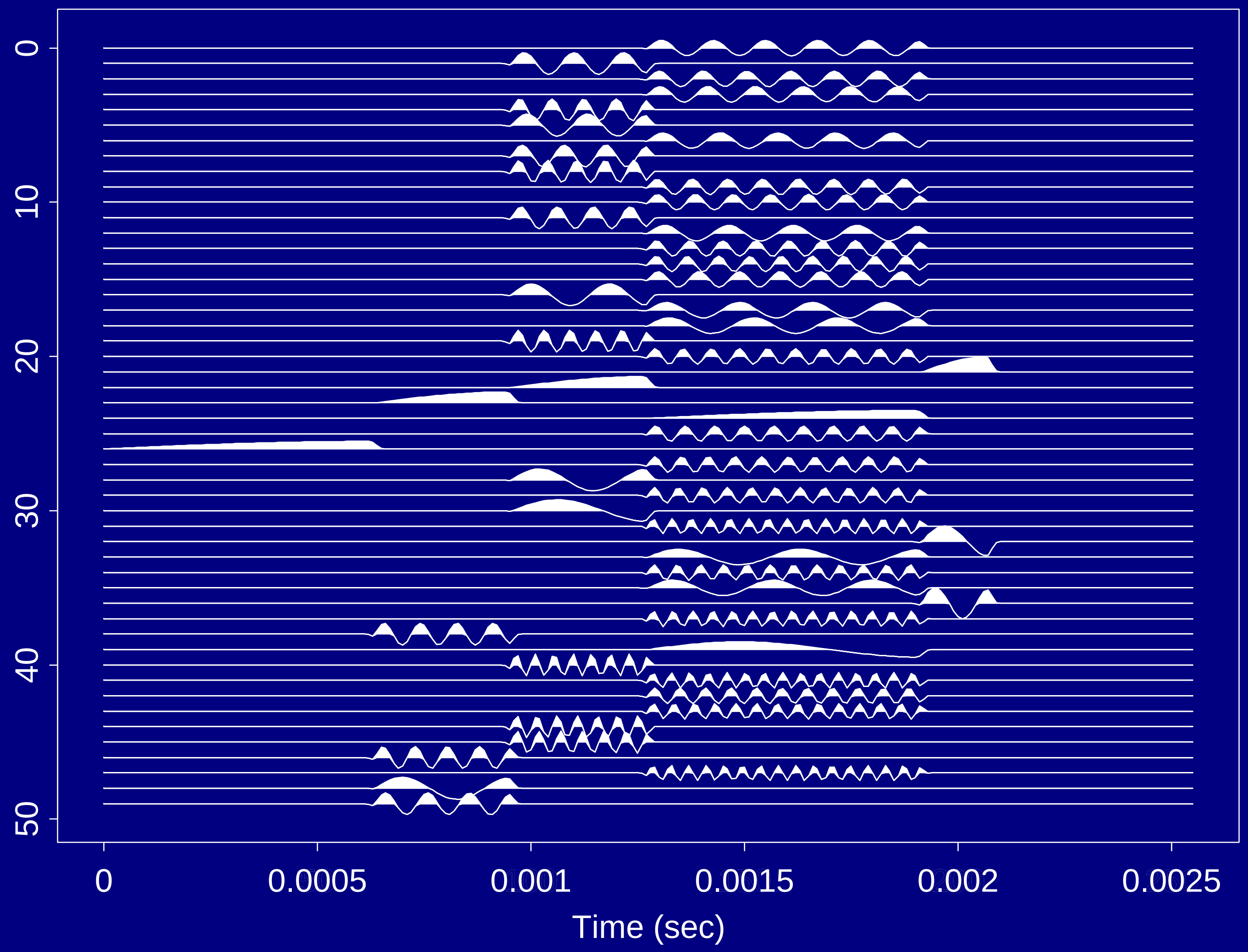


Classification Procedure

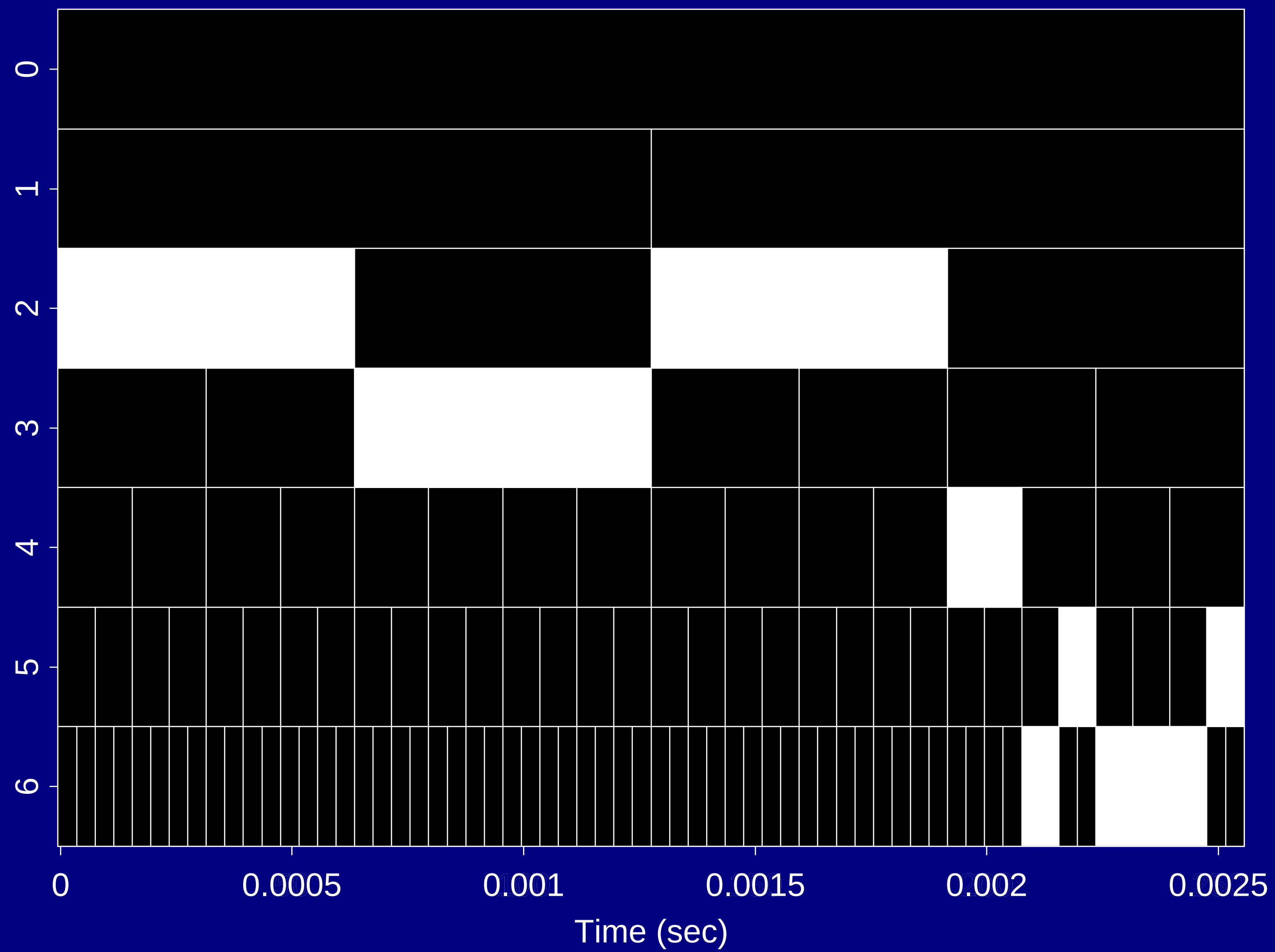
- Adopt 10-fold cross validation; split waveforms randomly into training and test datasets, and repeat the experiments.
- Decompose training waveforms into the local sine dictionary.
- Choose the LDB using various discrepancy measures.
- Choose 25, 50, 75, 100 most discriminant coordinates.
- Construct classifiers (LDA and Classification Tree) using these.
- Decompose test waveforms into the LDB and classify them.
- Compute the average misclassification rates.


Misclassification Rates for Test Data
Solid Lines: LDA
Dotted Lines: CT
Misclassification Rate (%)
0
2
4
6
8
10
25
50
75
100
Number of LDB Features

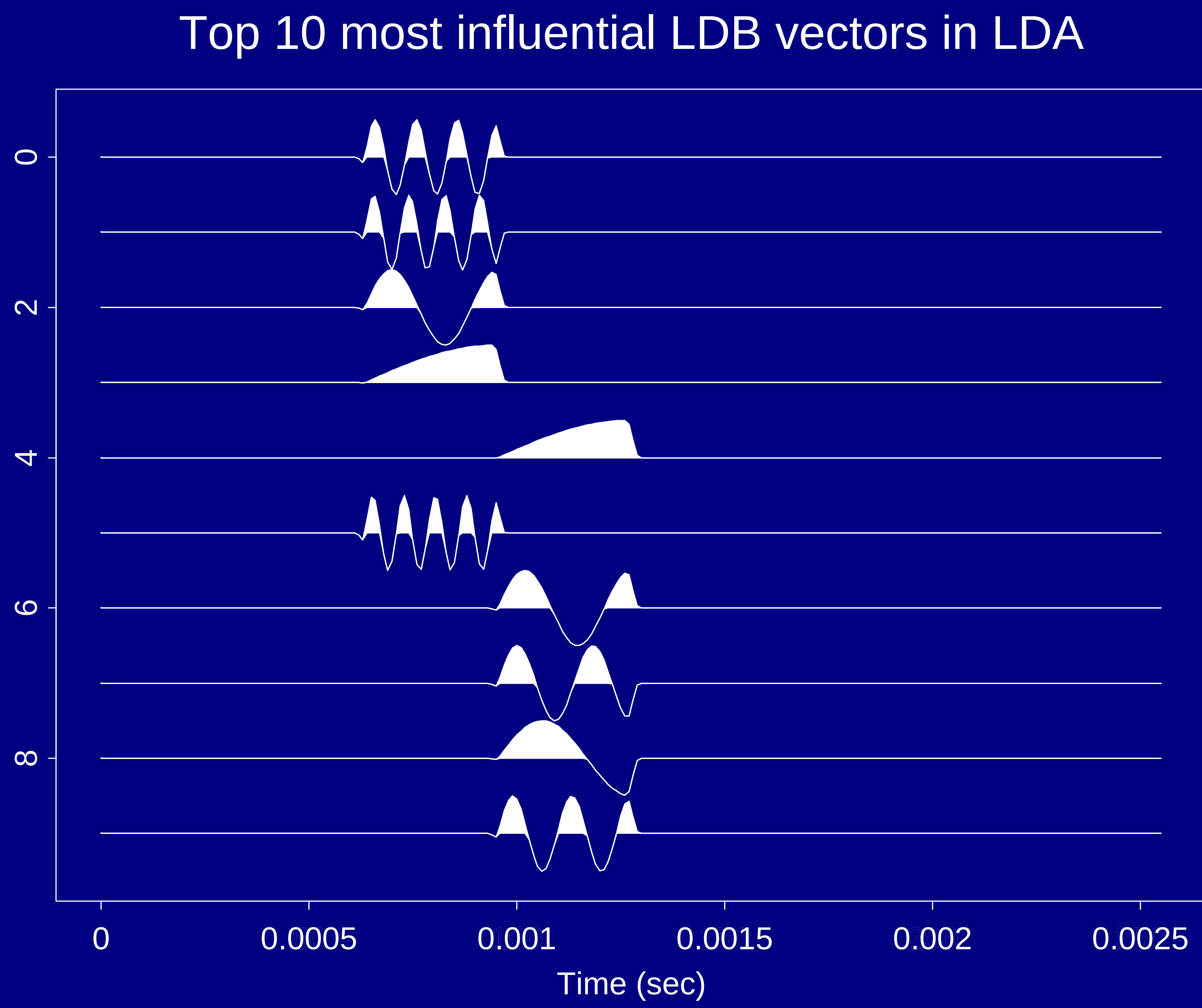
The Best LDB Vectors



The Best LDB Pattern



The Best LDB Vectors . . .



Observations

- LDA on LDB gave better results than CT on LDB $\Longrightarrow$ features are oblique.
- Top 25 LDB features were clearly not sufficient.
- Supplying too many features degraded the performance.
- Most important LDB vectors are clustered around P wave components.

Improved LDB with Empirical PD Estimation

- This can be viewed as a specific yet fast version of the Projection Pursuit algorithm (Friedman-Tukey, Huber).
- Compared to top-down strategy of PP, LDB is not greedy, i.e., bottom-up approach.
- This approach also works for complex-valued expansion coefficients (e.g., local Fourier bases).
- How to estimate the empirical pdf? $\longrightarrow$ histograms, ASH (averaged shifted histograms), nearest neighbor, kernel-based methods . . .

Conclusion

- LDB functions can capture relevant local features in data
- Interpretation of the results becomes easier and more intuitive
- Computational cost is at most $O(n[\log n]^2)$
- These methods enhance both traditional and modern statistical methods

Conclusion . . .

- Our algorithms are being applied and tested to:
  - Geophysical signal/image classification at Schlumberger
  - Noise reduction in hearing aids at Northwestern University
  - Diagnostics of mammography at University of Chicago Hospital
  - Radar target discrimination at Lockeed-Martin

Future Directions

- How about the good basis for clustering?
- Parameterization of low dimensional structures in high dimensional space
- Explore the relationship with the David-Jones-Semmes geometric analysis
- Develop two dimensional version for the complex coefficients provided by the local Fourier bases or brushlets