Supervised Classification in High-Dimensional Space: Geometrical, Statistical, and Asymptotical Properties of Multivariate Data

Ana I. González Acuña

by *Luis O. Jimenez, and David A. Landgrebe*

in *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 28, NO. 1, FEBRUARY* ***1998***

09/03/2012

Contents

Introduction

- Actual remote-sensing systems enables the measurement of radiation in **many more spectral intervals** than previously possible.
- The increased dimensionality of such hyperspectral data greatly enhances the data information content, but provides a challenge to the current techniques for analyzing such data.
- Human experience in three-dimensional (3-D) space tends to mislead our intuition of geometrical and statistical properties in high-dimensional space.
- High-dimensional space properties are investigated and their implication for high-dimensional data and its analysis is studied.

Geometrical, statistical, and asymptotical properties

Present some unusual or unexpected hyperspace characteristics to show that higher dimensional space is quite different from the 3-D space.
As dimensionality increases:

- A. The Volume of a Hypercube Concentrates in the Corners
- B. The Volume of a Hypersphere Concentrates in an Outside Shell
- C. The Volume of a Hyperellipsoid Concentrates in an Outside Shell
- D. The Diagonals Are Nearly Orthogonal to All Coordinate Axes
- E. The Required Number of Labeled Samples for Supervised Classification Increases as a Function of Dimensionality
- F. For Most High-Dimensional Data Sets, Low Linear Projections Have the Tendency to be Normal, or a Combination of Normal Distributions, as the Dimension Increases

The Volume of a Hypercube Concentrates in the Corners

It has been shown [9] that the volume of the hypersphere of radius r and dimension d is given by

$$V_s(r) = \text{ volume of a hypersphere } = \frac{2r^d}{d}\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \quad (1)$$

and that the volume of a hypercube in $[-r, r]^d$ is given by

$$V_c(r) = \text{ volume of a hypercube } = (2r)^d. \quad (2)$$

The fraction of the volume of a hypersphere inscribed in a hypercube is

$$f_{d1} = \frac{V_s(r)}{V_c(r)} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \quad (3)$$

where d is the number of dimensions. We see in Fig. 1 how f_{d1} decreases as the dimensionality increases.

Note that $\lim_{d\to\infty} f_{d1} = 0$, which implies that the volume of the hypercube is increasingly concentrated in the corners as d increases.

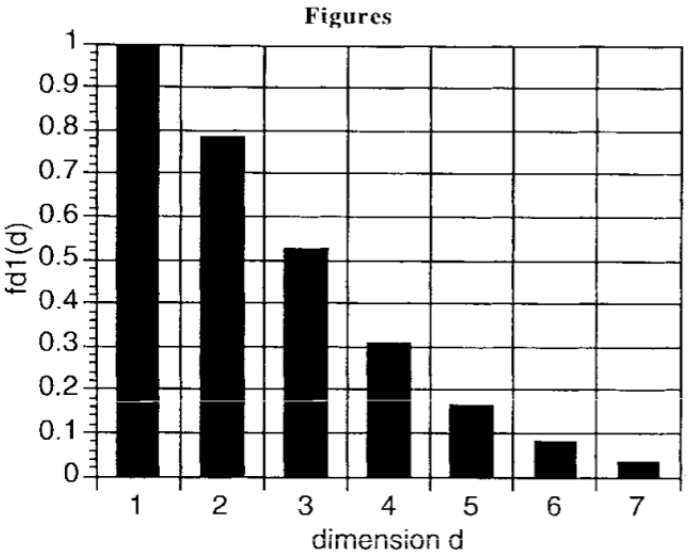


Fig. 1. Fractional volume of a hypersphere inscribed in a hypercube as a function of dimensionality.

The Volume of a Hypersphere Concentrates in an Outside Shell

The fraction of the volume in a shell defined by a sphere of radius $r - \varepsilon$ inscribed inside a sphere of radius r is

$$f_{d2} = \frac{V_d(r) - V_d(r-\varepsilon)}{V_d(r)} = \frac{r^d - (r-\varepsilon)^d}{r^d} = 1 - \left(1 - \frac{\varepsilon}{r}\right)^d.$$

In Fig. 2 we can observe, for the case $\varepsilon = r/5$, how as the dimension increases, the volume concentrates in the outside shell.

Note that $\lim_{d\to\infty} f_{d2} = 1, \forall\, \varepsilon > 0$, implying that most of the volume of a hypersphere is concentrated in an outside shell.

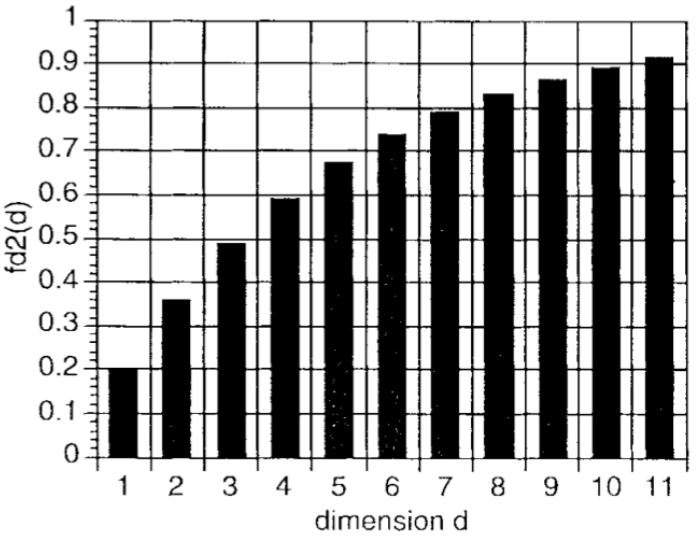


Fig. 2. Volume of a hypersphere contained in the outside shell as a function of dimensionality for $\varepsilon = r/5$.

The Volume of a Hyperellipsoid Concentrates in an Outside Shell

Generalization to a hyperellipsoid

The volume of a hyperellipsoid defined by

$$\frac{X_1^2}{(\lambda_1-\delta_1)^2}+\frac{X_2^2}{(\lambda_2-\delta_2)^2}+\cdots+\frac{X_d^2}{(\lambda_d-\delta_d)^2}=1$$

where $0 \le \delta_i < \lambda_i, \forall i$ is calculated by

$$V_e(\lambda_i-\delta_i)=\frac{2\prod_{i=1}^{d}(\lambda_i-\delta_i)}{d}\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}.$$

The fraction of the volume of $V_e(\lambda_i-\delta_i)$, inscribed in the volume $V_e(\lambda_i)$, is

$$f_{d3}=\frac{\prod_{i=1}^{d}(\lambda_i-\delta_i)}{\prod_{i=1}^{d}\lambda_i}=\prod_{i=1}^{d}\left(1-\frac{\delta_i}{\lambda_i}\right).$$

Let $\lambda_{\min}=\min(\delta_i/\lambda_i)$, then

$$f_{d3}=\prod_{i=1}^{d}\left(1-\frac{\delta_i}{\lambda_i}\right)\le\prod_{i=1}^{d}(1-\lambda_{\min})=(1-\lambda_{\min})^d.$$

Using the fact that $f_{d3}\ge 0$, it is concluded that $\lim_{d\to\infty} f_{d3}=0$

Consequences for high-dimensional data

1. **High-dimensional space is mostly empty**:
   Desirable to project to a lower dimensional subspace without losing significant information in terms of separability among the different statistical classes.
2. **Density estimation more difficult**:
   Normally distributed data will have a tendency to concentrate in the tails
   Uniformly distributed data will be more likely to be collected in the corners

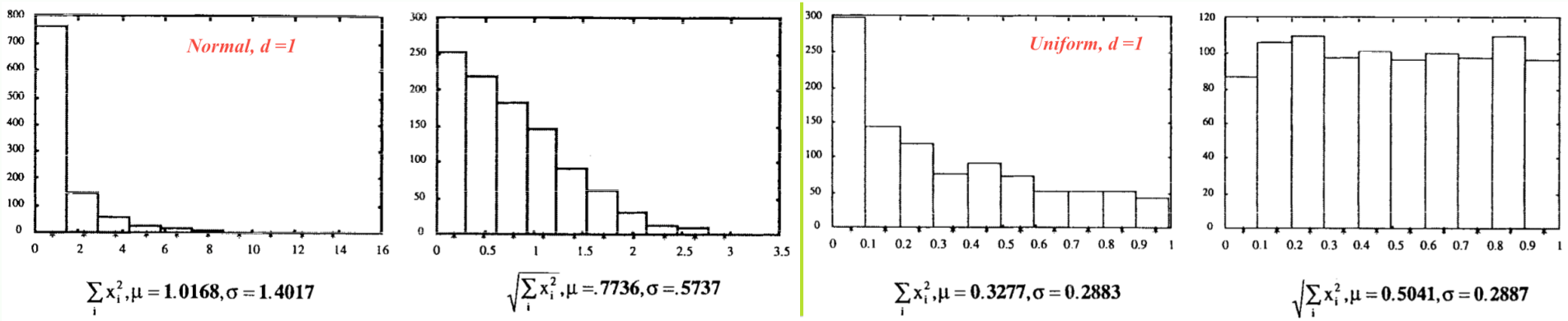
Statistical behavior of normally and uniformly distributed multivariate data at high dimensionality

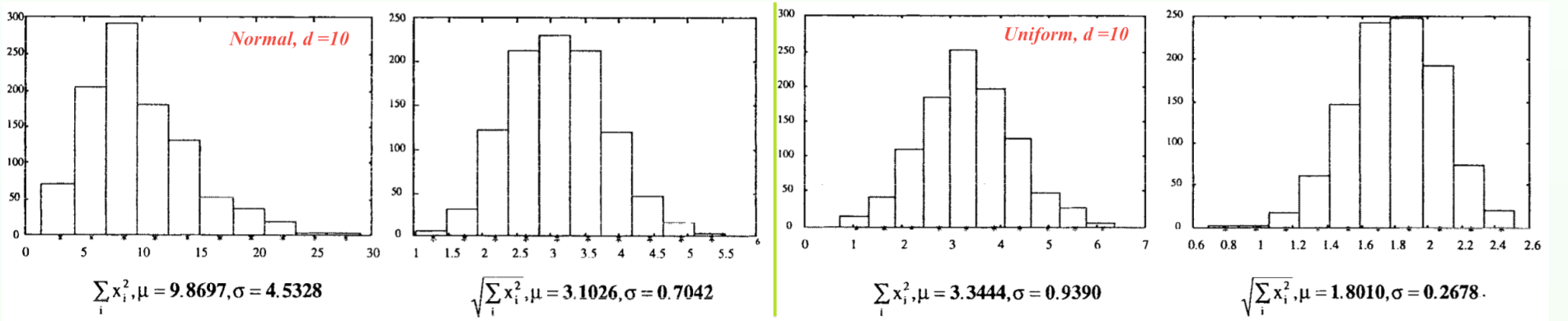
Experiment:

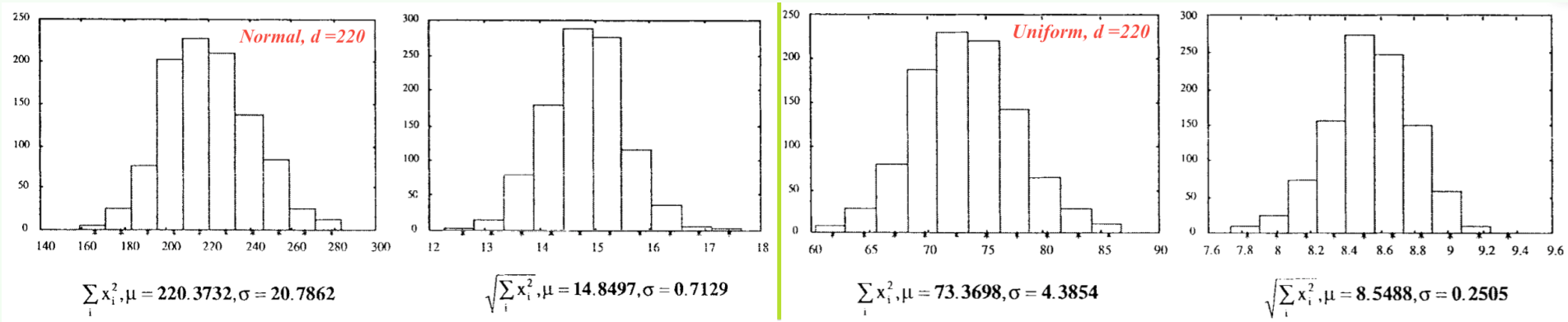
- Normal and uniform variables are i.i.d. samples from the distributions $N(0,1)$ and $U(-1,1)$, respectively.
- Two random variables: the distance from the zero coordinate and its square: $R = \sum_{i=1}^{d} x_i^2$, $r = \sqrt{\sum_{i=1}^{d} x_i^2}$

Results:

- The means and the standard deviations are functions of d.
- As d increases:
  the data will concentrate in an outside shell, and
  that shell will increase its distance from the origin.







The distance from zero coordinate increases

The Diagonals Are Nearly Orthogonal to All Coordinate Axes

The cosine of the angle between any diagonal vector and a Euclidean coordinate axis is

$$\cos(\theta_d) = \pm\frac{1}{\sqrt{d}}.$$

Fig. 5 illustrates how the angle between the diagonal and the coordinates θ_d approaches 90° with increases in dimensionality.

Note that $\lim_{d\to\infty} \cos(\theta_d) = 0$, which implies that in high-dimensional space, the diagonals have a tendency to become orthogonal to the Euclidean coordinates.

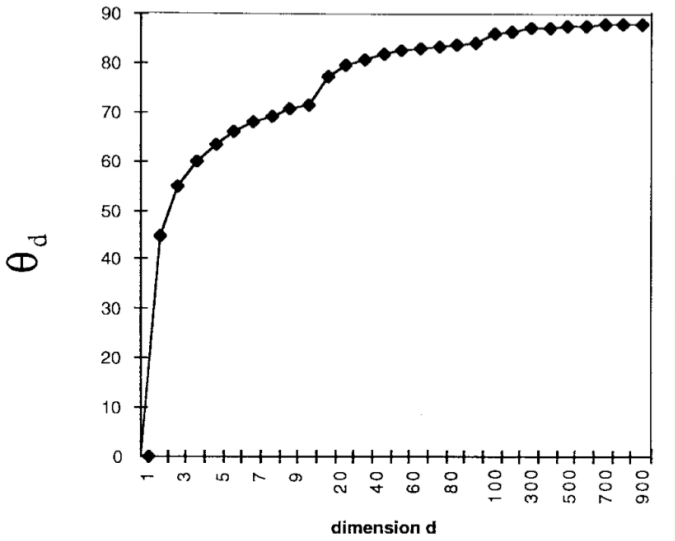


Fig. 5. Angle (in degrees) between a diagonal and a Euclidean coordinate versus dimensionality.

Consequence: the projection of any cluster onto any diagonal (e.g., by averaging features) could destroy information contained in multispectral data

The Required Number of Labeled Samples for Supervised Classification Increases as a Function of Dimensionality

- The required number of training samples is linearly related to the dimensionality for a linear classifier and to the square of the dimensionality for a quadratic classifier
- With a limited number of training samples, there is a penalty in classification accuracy as the number of features increases beyond some point.

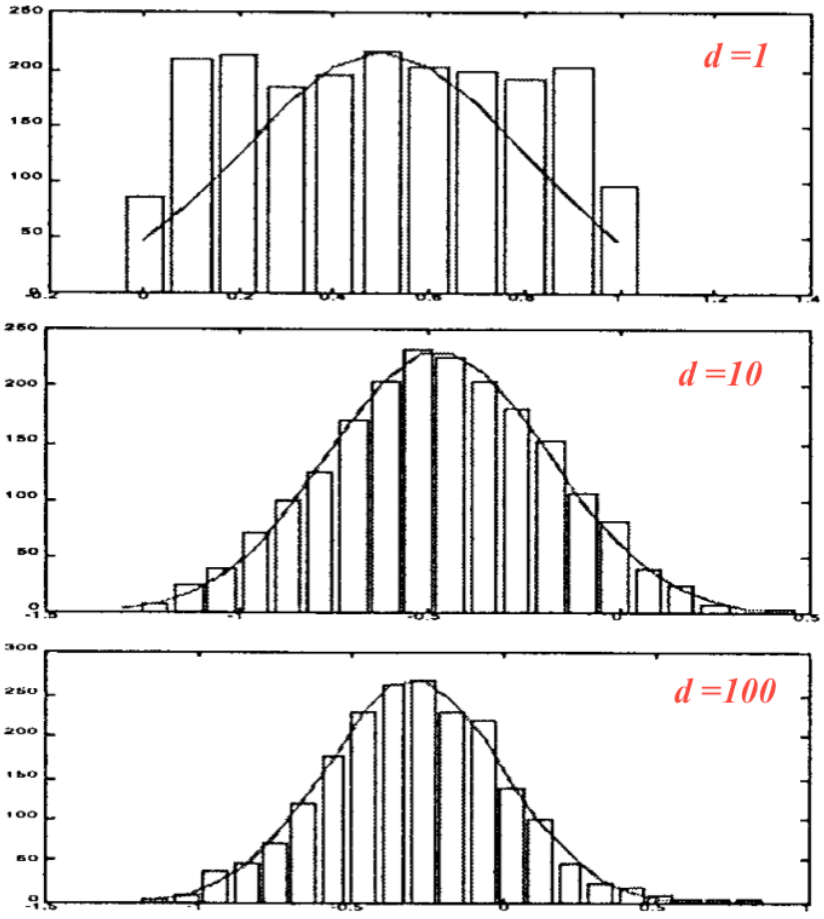
Low Linear Projections

For most high-dimensional data sets, low linear projections have the tendency to be normal, or a combination of normal distributions (normality), as the dimension increases.

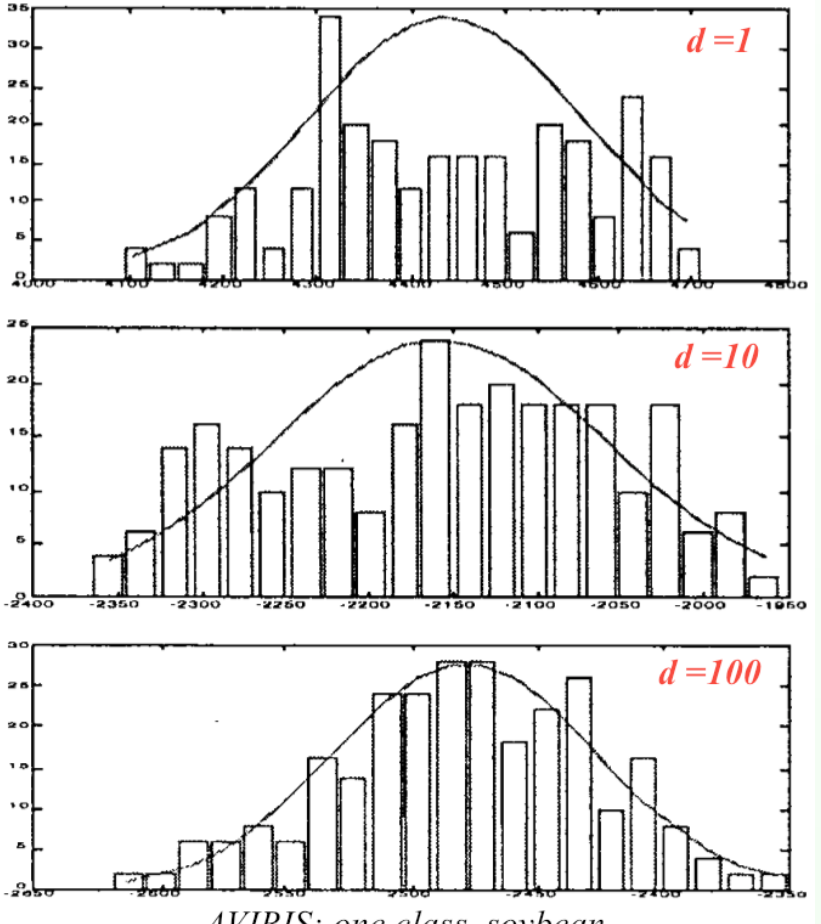
Experiments:

- Project the data from a high-dimensional space to a one-dimensional (1-D) subspace.
- In the original high-dimensional space $d = \{1, 10, 100\}$
- Method of data projection: multiply data with a normal vector with random angles from the coordinates.

One class

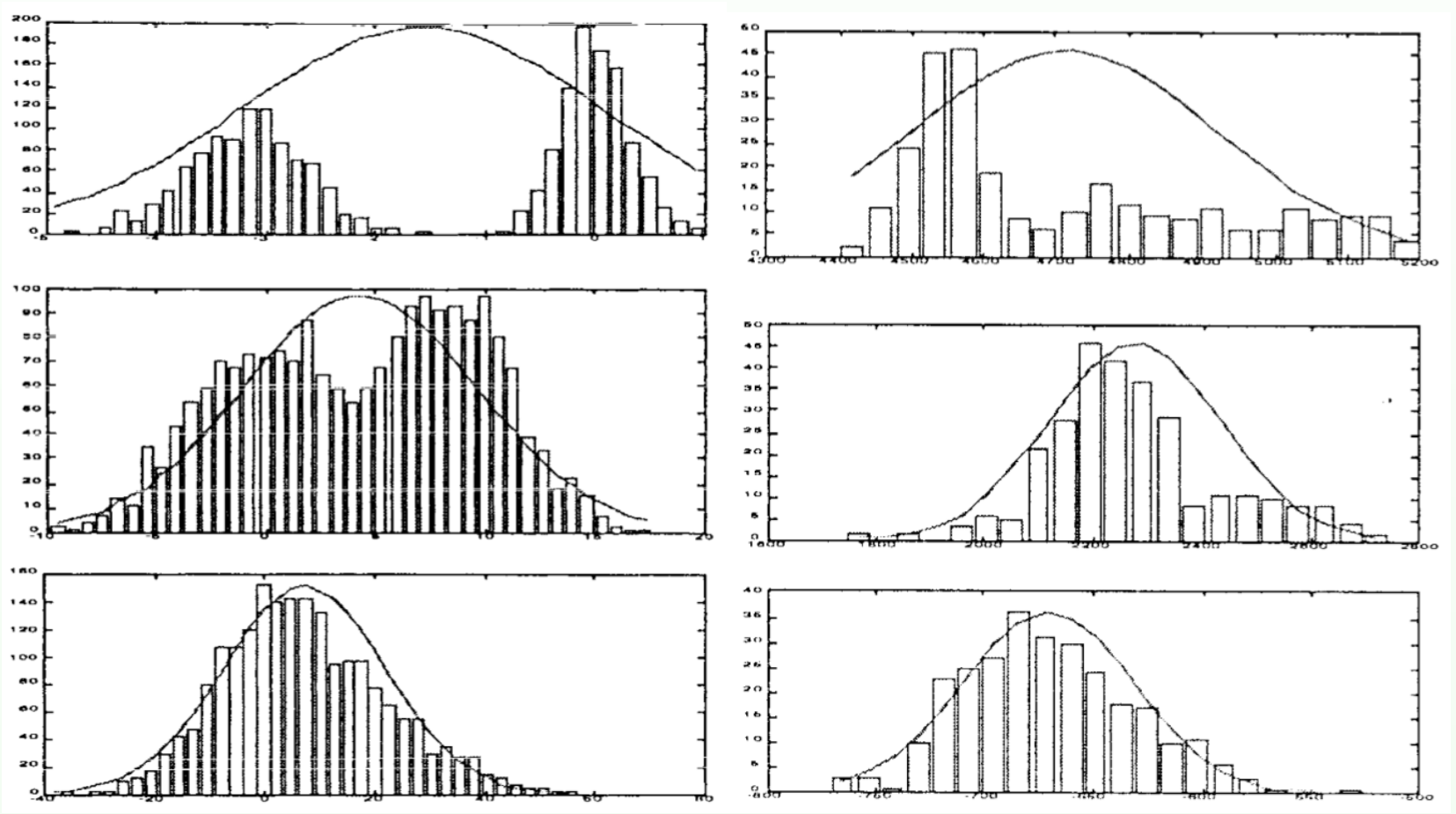


one class with uniform distribution



AVIRIS: one class, soybean

Two class



Low Linear Projections

The first results tempt us to expect that the **data** can be assumed to be **a combination of normal distributions** in the projected subspace without any problem.

The second ones show the risk of damaging data projecting it into one normal distribution, losing separability and information.

We can see the advantage of developing an algorithm that will **estimate the projection directions** that separate the explicitly defined classes by **doing the computations in a lower dimensional space**.

Asymptotical first- and second-order statistics properties

The conditions required for the predominance of either first- or second-order statistics in the discrimination among the statistical classes in high-dimensional space.

- As the number of features increases, the potential information content in multispectral data increases.
- In supervised classification, that increment of information is translated to the number of classes and their separability.
- We will use Bhattacharyya distance here as the measure of separability.
- It provides a bound of classification accuracy, taking into account first- and second-order statistics.

Bhattacharyya distance

Bhattacharyya distance under the assumption of normality:

The Bhattacharyya distance is the sum of the contribution of the difference of the means and the difference of the covariances. $\mu = \mu_M + \mu_C$, where

$$\mu_M = \frac{1}{8}(M_2 - M_1)^T \overline{\Sigma}^{-1} (M_2 - M_1), \overline{\Sigma} = \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]$$

and

$$\mu_C = \frac{1}{2} \ln \left(\frac{|\overline{\Sigma}|}{\sqrt{|\Sigma_1||\Sigma_2|}} \right).$$

Experiments

¿How Bhattacharyya distance and its mean and covariance components can aid in the understanding of the role of first- and second-order statistics?

Experiments:

- The first one has conditions in which second- order statistics are more relevant in discriminating among the classes.
- The second experiment has conditions for the predominance of first-order statistics.

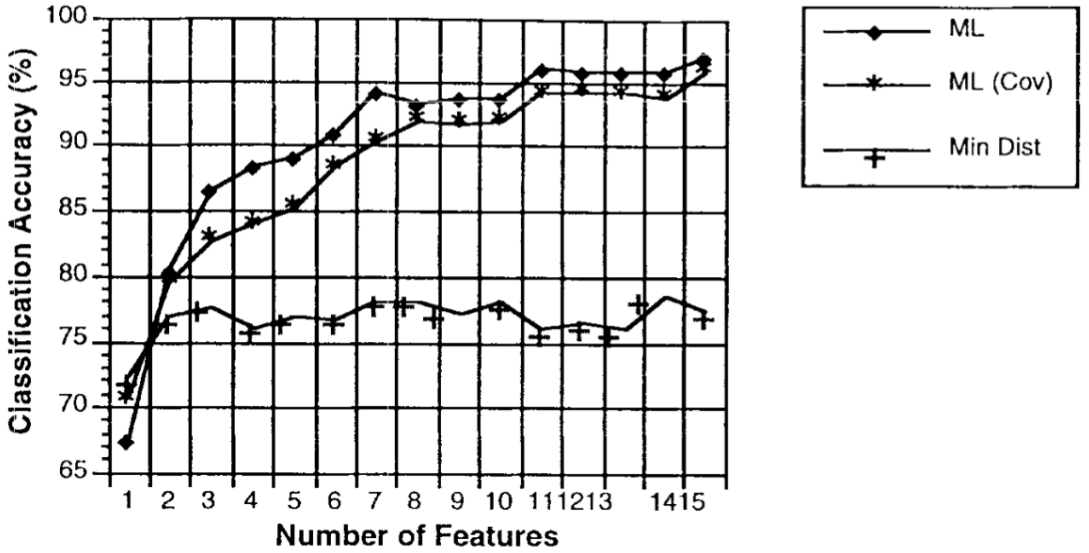
The mean (Bhatt Mean) and covariance (Bhatt Cov) components of Bhattacharyya distance and its sum were computed, also the ratio of Bhatt Mean/Bhatt Cov

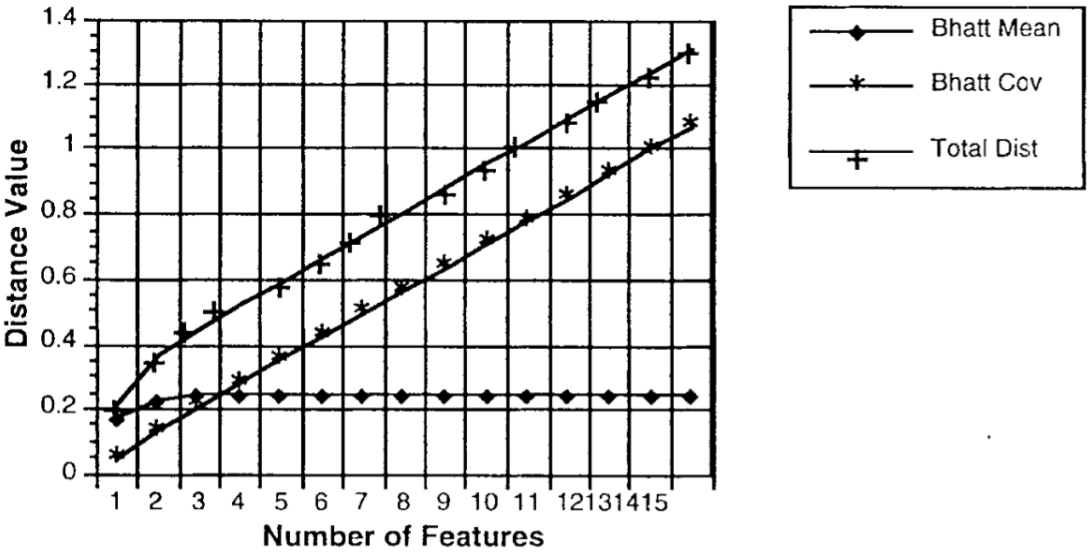
Experiment 1

- Data generated for two classes
- Both classes belong to normal distributions with different means and covariances.
- Each class has 500 points.
- Parameters:

$$M_1 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$
$$M_2 = [1.5 \quad 1 \quad 0.5 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$
$$\Sigma_1 = \begin{bmatrix} 1 & & & & & & & & & 0 \\ & 1 & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 1 & & & & & & \\ & & & & 1 & & & & & \\ & & & & & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & & 1 & & \\ & & & & & & & & 1 & \\ 0 & & & & & & & & & 1 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1.5 & & & & & & & & & 0 \\ & 1.9 & & & & & & & & \\ & & 3 & & & & & & & \\ & & & 3 & & & & & & \\ & & & & 3 & & & & & \\ & & & & & 3 & & & & \\ & & & & & & 3 & & & \\ & & & & & & & 3 & & \\ & & & & & & & & 3 & \\ 0 & & & & & & & & & 3 \end{bmatrix}.$$

- Classifiers: Gaussian ML, Gaussian ML with zero-mean data, and Minimum Distance classifier





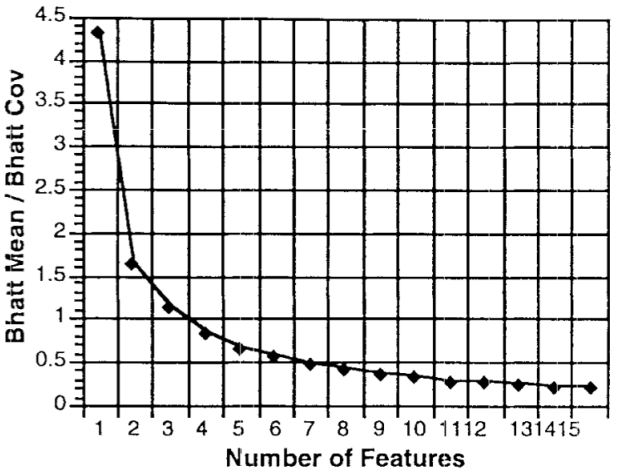


Fig. 13. Ratio of Bhattacharyya distance-mean component over the covariance component.

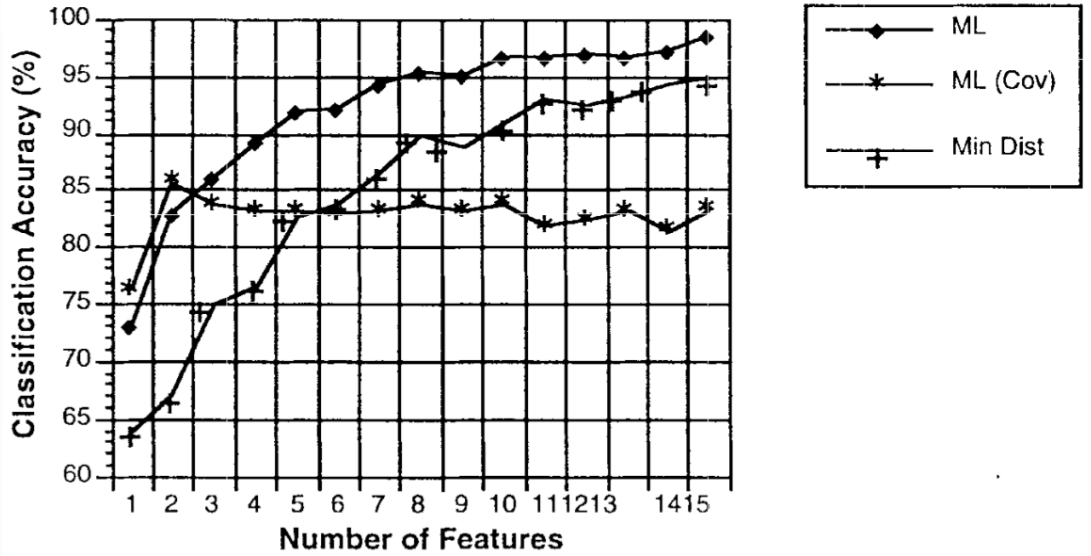
Experiment 2

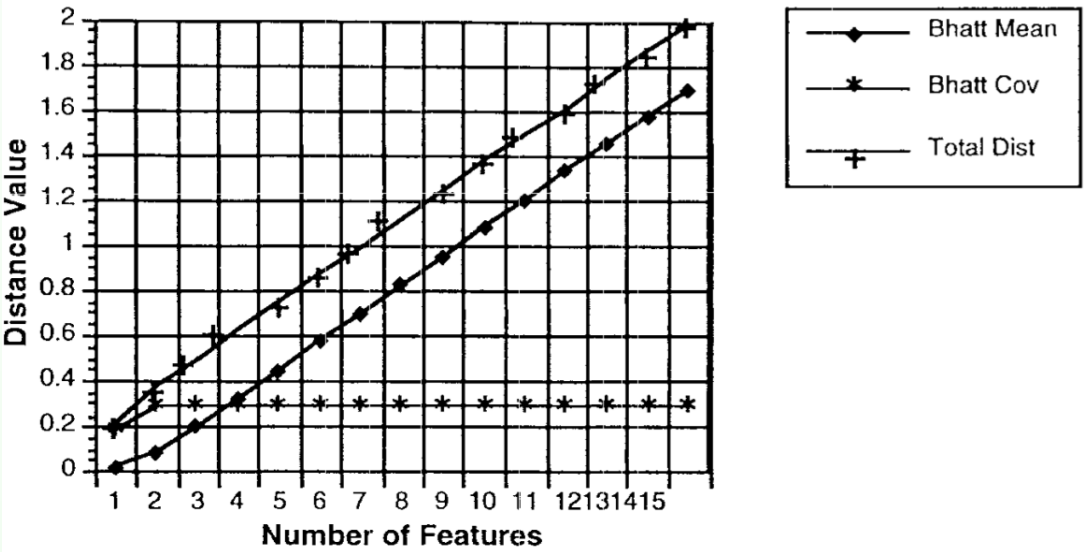
- Similar experiment1 but the first-order statistics are predominant in this case.
- Parameters:

$$M_1 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$
$$M_2 = [1.5 \quad 1 \quad 0.5 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & & & & & & & & & 0 \\ & 1 & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 1 & & & & & & \\ & & & & 1 & & & & & \\ & & & & & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & & 1 & & \\ & & & & & & & & 1 & \\ 0 & & & & & & & & & 1 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2.5 & & & & & & & & & 0 \\ & 2 & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 1 & & & & & & \\ & & & & 1 & & & & & \\ & & & & & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & & 1 & & \\ & & & & & & & & 1 & \\ 0 & & & & & & & & & 1 \end{bmatrix}.$$

- Classifiers: Gaussian ML, Gaussian ML with zero-mean data, and Minimum Distance classifier





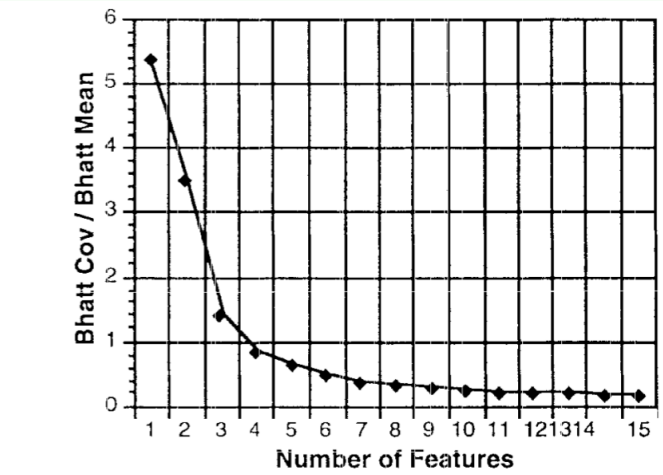


Fig. 16. Ratio of Bhattacharyya distance-covariance component over the mean component.

Upper bounds

¿When the mean difference plays a predominant role?

¿When the covariance difference became predominant?

Case 1

Case 1) Covariance Difference as the Dominant Role in Statistical Class Separability: Assume a two-class problem, where without loss of generality, the first- and second-order statistics are

$$\Sigma_1 = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} \alpha_1\sigma^2 & & 0 \\ & \ddots & \\ 0 & & \alpha_d\sigma^2 \end{bmatrix},$$

$$(M_2 - M_1) = [\varepsilon_1 \cdots \varepsilon_k \quad \hat{\varepsilon}_{k+1} \cdots \hat{\varepsilon}_d]^T.$$

Under the following conditions:

1) $\alpha_i \in (\alpha_{\min}, \alpha_{\max})$, where $\alpha_{\min} > 0$ and at least there exists α_i, such that $\alpha_i \neq 1$;
2) $\varepsilon_{\max} = \max_{\forall i \in (k+1,\, d)} (|\hat{\varepsilon}_i|)$, such that $\varepsilon_{\max} \approx 0$;
3) $k = f(d) \ni \lim_{d\to\infty}(k/d) = 0$ [as an example, $\forall \lambda > 0, d = k^{(1+\lambda)}$];
4) $\varepsilon_i^2 \in (E_{\min}, E_{\max})$, $\forall i \in (1, k)$, and $(E_{\max} < \infty)$ (to see the validity of this last assumption, see Appendix B).

Then, as d increases, the covariance contribution will dominate the Bhattacharyya distance.

The total covariances information plays a more important role in discriminating among the classes than the means information.

Case 2

Case 2) Mean Differences as Dominant in Statistical-Class Separability: Assume a two-class problem, where without loss of generality, the first- and second-order statistics are

$$\Sigma_1 = \begin{bmatrix} \sigma^2 & & & & 0 \\ & \ddots & & & \\ & & \sigma^2 & & \\ & & & \ddots & \\ 0 & & & & \sigma^2 \end{bmatrix}$$

and

$$\Sigma_2 = \begin{bmatrix} \alpha_1\sigma^2 & & & & 0 \\ & \ddots & & & \\ & & \alpha_k\sigma^2 & & \\ & & & \hat{\alpha}_{k+1}\sigma^2 & \\ & & & & \ddots \\ 0 & & & & \hat{\alpha}_d\sigma^2 \end{bmatrix},$$

$$(M_2 - M_1) = [\varepsilon_1 \cdots \varepsilon_d]^T.$$

Under the following assumptions:

1) $\alpha_i \in (\alpha_{\min}, \alpha_{\max})$, where $0 < \alpha_{\min} < \alpha_{\max} < \infty, \forall i \in (1, k)$;
2) $\hat{\alpha}_i \in (1-\delta, 1+\delta), \forall i \in (k+1, d)$, where $\delta \approx 0$;
3) $\varepsilon_i^2 \geq E_{\min} > 0, \forall i \in (1, d)$;
4) $\lim_{d\to\infty}(k/d) = 0$ [as an example, $\forall \lambda > 0, d = k^{(1+\lambda]}$).

As d increases, the means differences will dominate the Bhattacharyya distance.

The total mean differences will provide more information for classes discrimination than covariances differences.

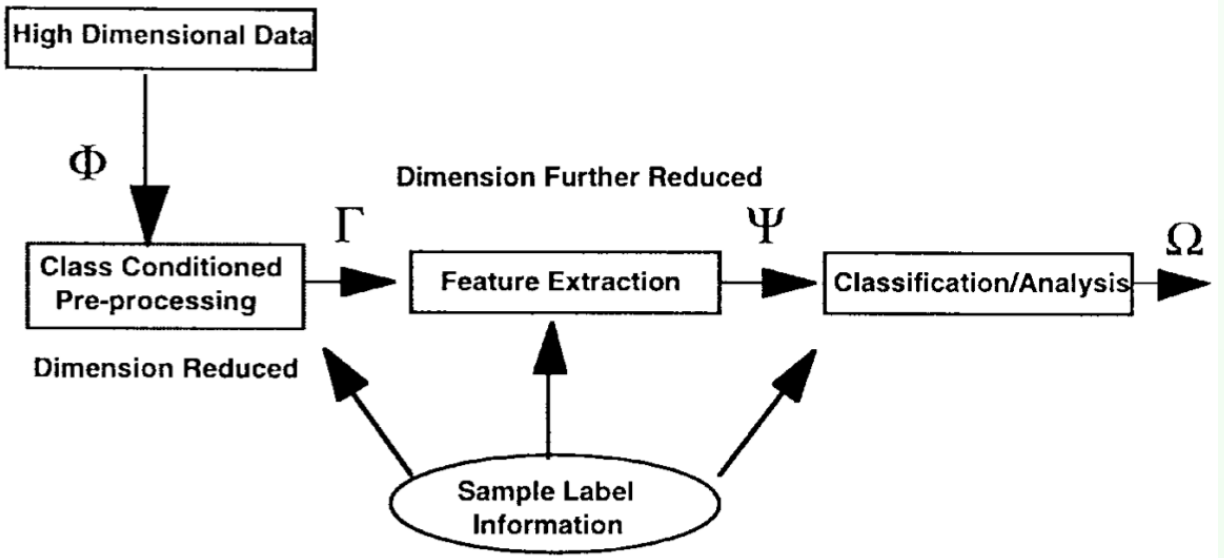
High-dimensional-characteristics implications for supervised classification

- Apparent high-dimensional space is mostly empty and multivariate data is usually in a lower dimensional structure.
  It is possible to reduce the dimensionality without losing significant information and separability.
- A parametric version of data-analysis algorithms may be expected to provide better performance doing the difficulties of density estimation in nonparametric approaches.
- The increased number of labeled samples required for supervised classification as the dimensionality increases presents a problem to current feature-extraction algorithms where computation is done at full dimensionality

A **new method** (¿?) is required that computes in a lower dimensional subspace: preprocessing method is called **parametric projection pursuit**

Parametric projection pursuit (PPP)

- Reduces the dimensionality of the data, maintaining as much information as possible by optimizing a projection index that is a measure of separability.
- The projection index: minimum Bhattacharyya distance among the classes, taking in consideration first- and second-order characteristics.
- The calculation is performed in the lower dimensional subspace where the data is to be projected.



Classification of high-dimensional data, including reprocessing of high-dimensional data.

Experiment 1

- A segment of AVIRIS data taken of northwest Indiana's Indian Pine test site. From the original 220 spectral channels, 200 were used, discarding the atmospheric absorption bands
- Classes = 8. Number of training samples = 1790, and number of test samples =1630.
- Four types of dimension reduction algorithms (200-22)
  - DB: decision-boundary feature extraction [from Φ directly to Ψ]
  - DA: discriminant analysis [from Φ directly to Ψ]
  - PPDB: PPP [from Φ directly to Γ] + decision boundary [from Γ to Ψ]
  - PPDA: PPP [from Φ directly to Γ] + discriminant analysis [from Γ to Ψ]
- Four types of classifiers [from Ψ to Ω] : ML classifier, ML with 2% threshold, a spectral-spatial classifier named ECHO and the fourth is ECHO with a 2% threshold.

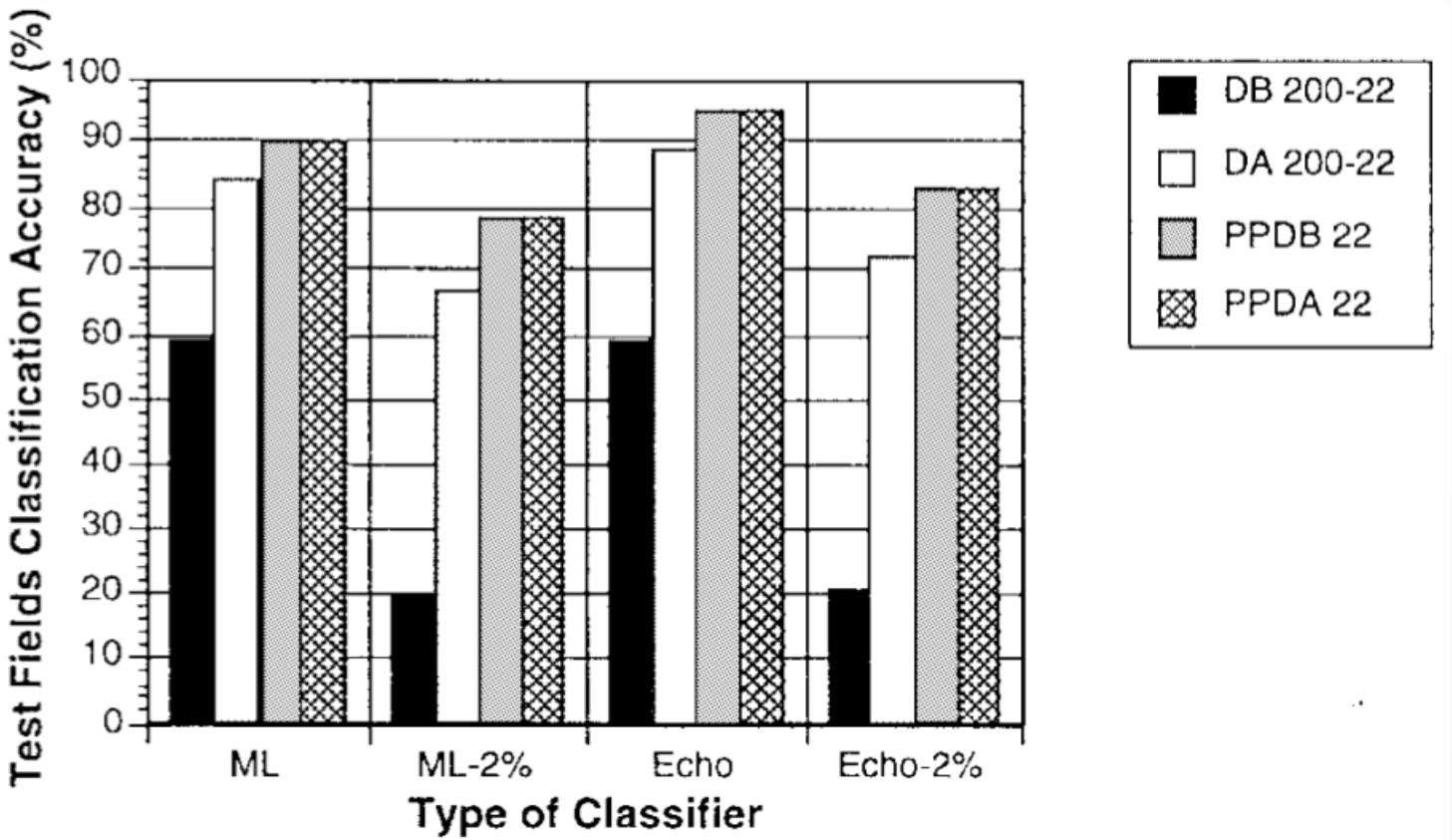


Fig. 18. Test fields classification accuracy for four feature-extraction methods and four classifiers.

Experiment 2

- Classes = 4. Number of training samples = 179, and number of test samples =3501.
- Two types of dimension reduction algorithms
  - DA (200-3): discriminant analysis [from Φ directly to Ψ]
  - PPDA: PPP (200-22) [from Φ directly to Γ] + discriminant analysis (22-3) [from Γ to Ψ]

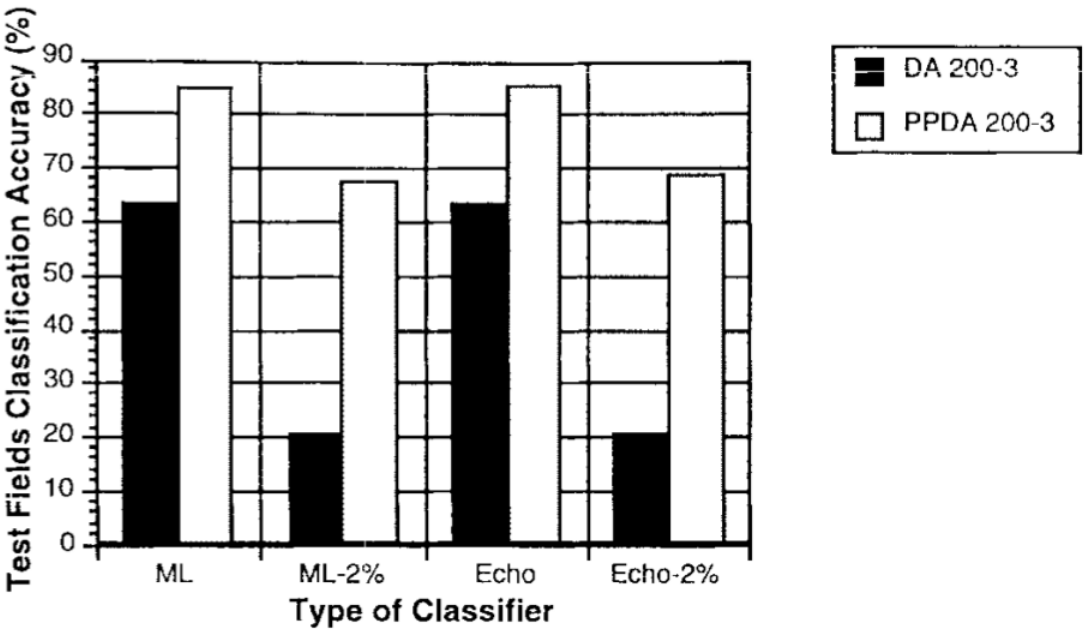


Fig. 19. Test fields classification accuracy for two feature-extraction methods and four classifiers.

Conclusions

- Characteristics of high-dimensional space are different from those of the 3-D space
- Implications in the context of supervised classification techniques
- A large number of samples are required to make estimation and grows as the dimensionality increases
- The goal is to reduce the dimensionality of the data to the right subspace without losing separability information.
- Describe a procedure to make the computations in a lower dimensional space

Application program: Multispec@