



# A study on the influence of shape in classifying small spectral data sets



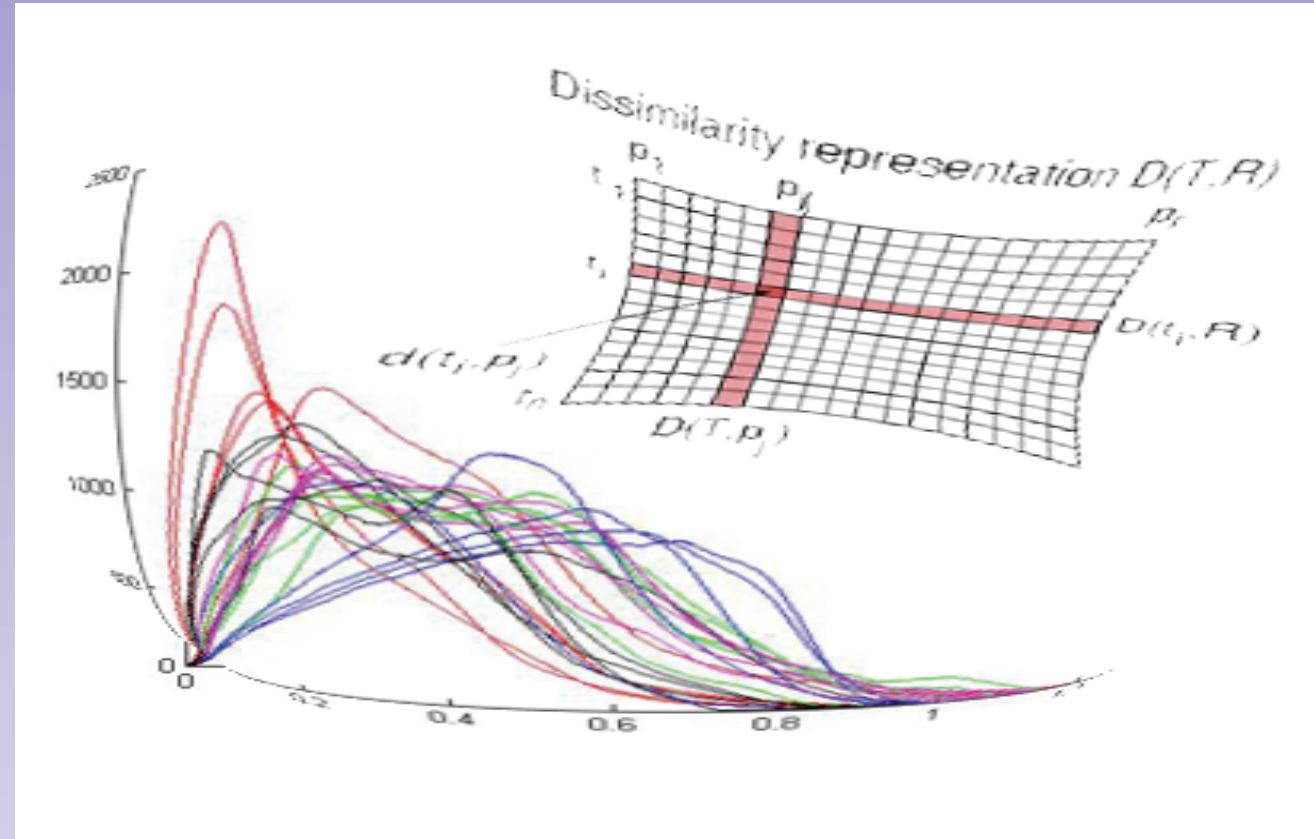
Diana Porro<sup>1,2</sup>, Robert W. Duin<sup>2</sup>, Isneri Talavera<sup>1</sup>, Mauricio Orozco<sup>3</sup>

<sup>1</sup>Advanced Technologies Application Centre, Havana, Cuba. <sup>2</sup>Pattern Recognition Lab, TU Delft, The Netherlands,

<sup>3</sup>Universidad Nacional de Colombia Sede Manizales, Colombia

\*{dporro,italavera}@cenatav.co.cu, r.duin@ieee.org, morozcoa@bt.unal.edu.co

## Introduction



### Feature representation

- ⊗ Considered as set of individual observations
- ⊗ Functional nature ignored i.e. connectivity between points, shape
- ⊗ Few objects in high dimensional space

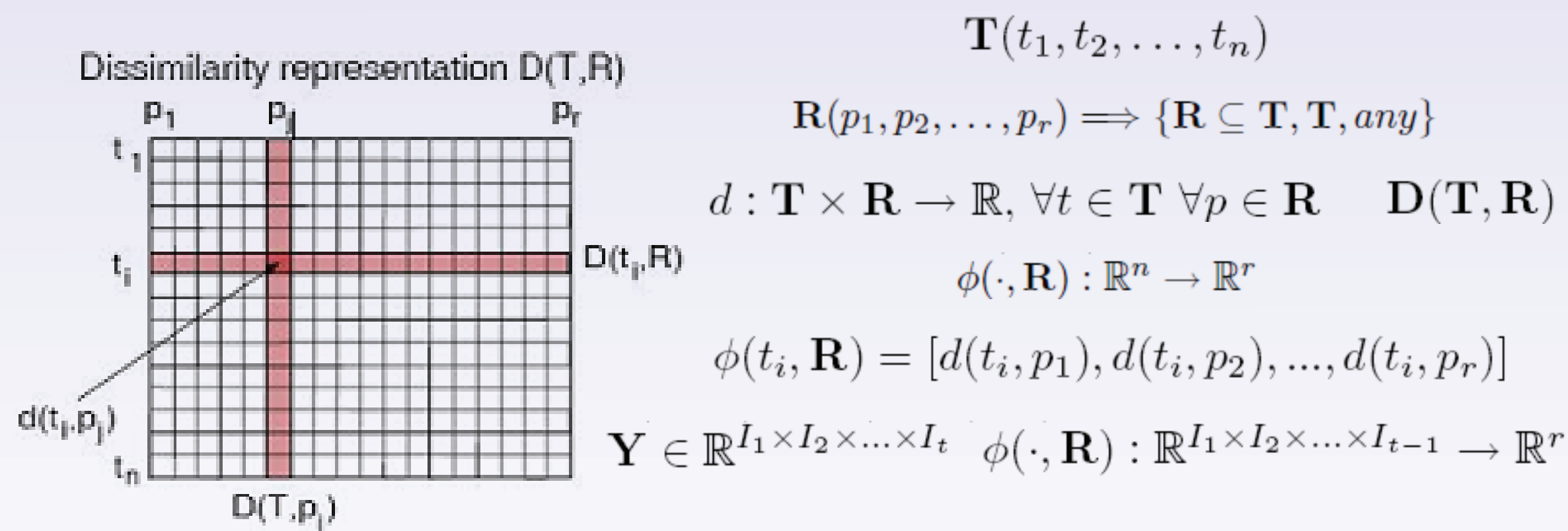
### Dissimilarity representation

- ☺ Allows including shape information in representation of spectra
- ☺ Dimensionality depends on the number of objects at most
- ☺ Curse of dimensionality can be tackled
- ☺ Any traditional classifier that works on feature spaces can be also used

The goal of this paper was to make a study on the benefit of using a dissimilarity measure which takes shape of spectra into account. We show that the dissimilarity representation of spectra with a shape-based measure leads to better classification results, and a certain number of objects is enough to achieve it.

## Dissimilarity Representation

Proposes to work on the space defined by the proximities between the objects instead of the space defined by their characteristics (features), based on the fact that classes are conformed by objects that have similar characteristics.



## Dissimilarity measures for spectral data

**Shape measure**  $d(t_a, t_b) = \sum_{j=1}^m |t_{aj}^\sigma - t_{bj}^\sigma|, \quad t^\sigma = \frac{d}{d_j} G(j, \sigma) * t$

### 2DShape measure

1. Compute the matrix  $D^1$

$$D_{a,b}^1 = \left( \sum_{k=1}^l \left( \sum_{j=1}^m (y_{a,j,k}^{\sigma_1} - y_{b,j,k}^{\sigma_1})^2 \right)^{p_1/2} \right)^{1/p_1}, \quad y_{i,j,k}^{\sigma_1} = \frac{d}{d_j} G(j, \sigma_1) * y_{i,j,k}$$

2. Compute the matrix  $D^2$

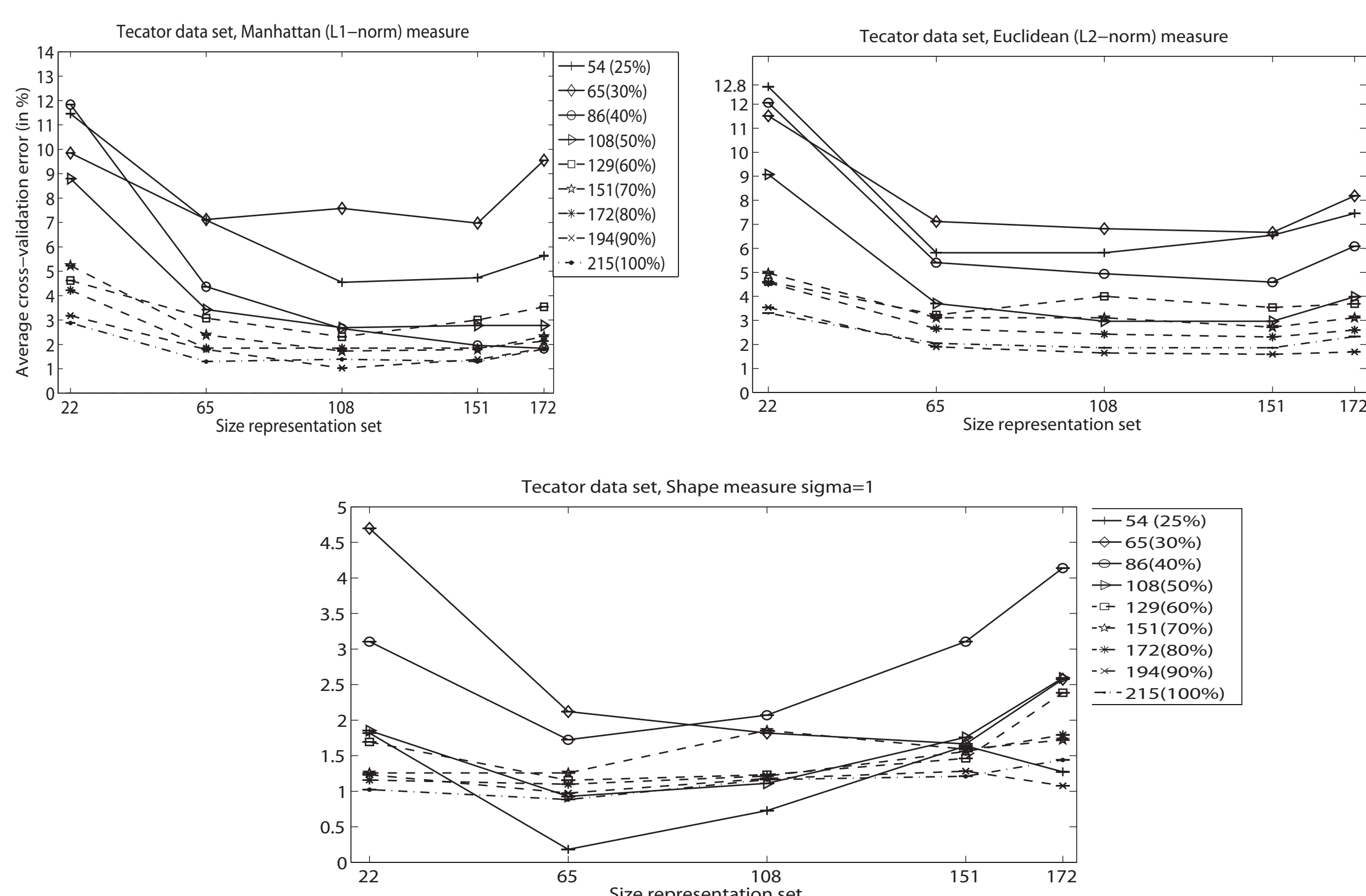
$$D_{a,b}^2 = \left( \sum_{k=1}^m \left( \sum_{j=1}^l (y_{a,j,k}^{\sigma_2} - y_{b,j,k}^{\sigma_2})^2 \right)^{p_2/2} \right)^{1/p_2}, \quad y_{i,j,k}^{\sigma_2} = \frac{d}{d_k} G(k, \sigma_2) * y_{i,j,k}$$

3. Combine both dissimilarities matrices  $D = \alpha_1 D^1 + \alpha_2 D^2$

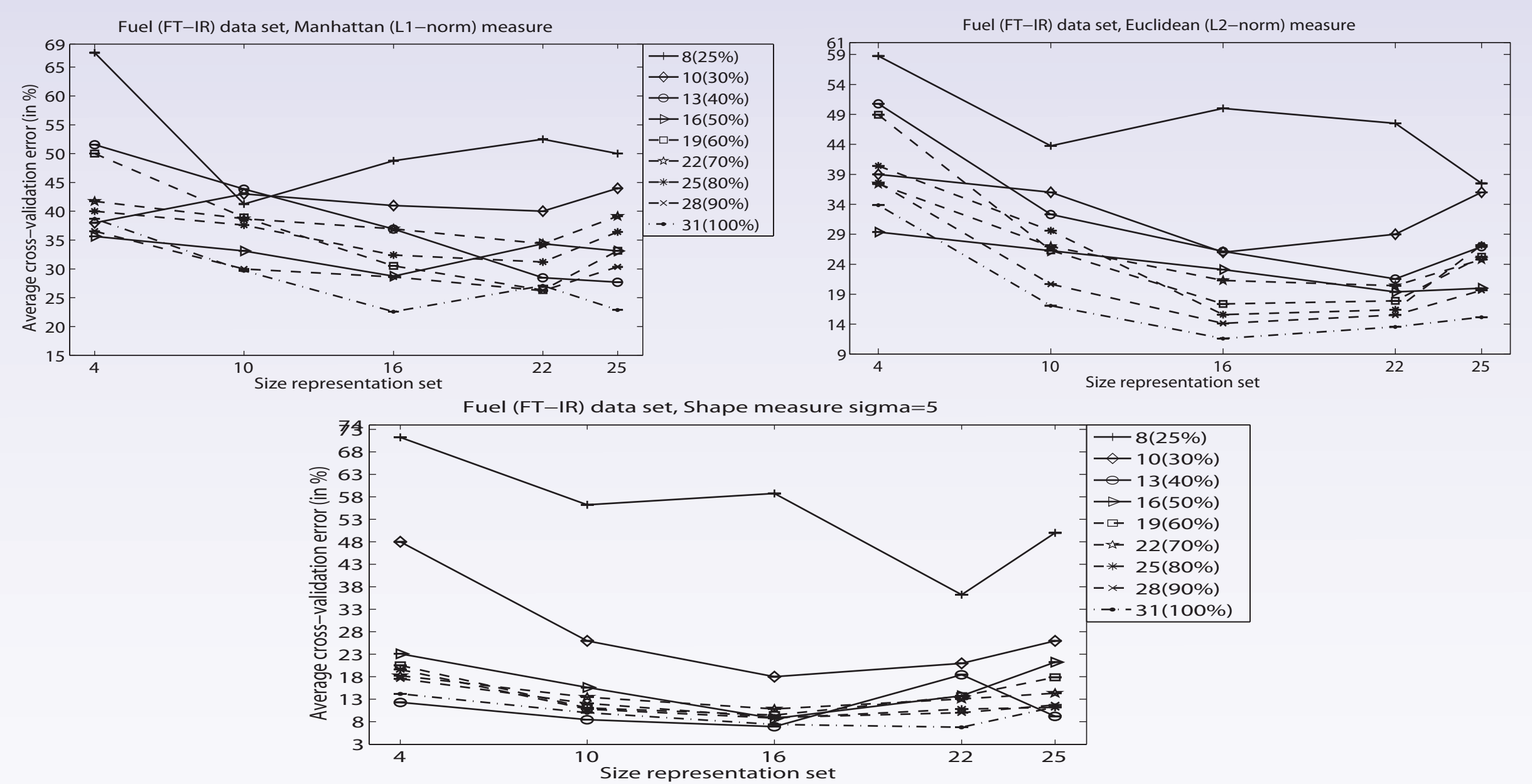
## Experiments

**Figures.** Average 10 times k-fold cross-validation error (in %) for all data sets with (top-left) Manhattan, (top-right) Euclidean and (bottom) Shape measures. Learning curves are shown for various sizes of training and representation sets. The Fisher classifier was used.

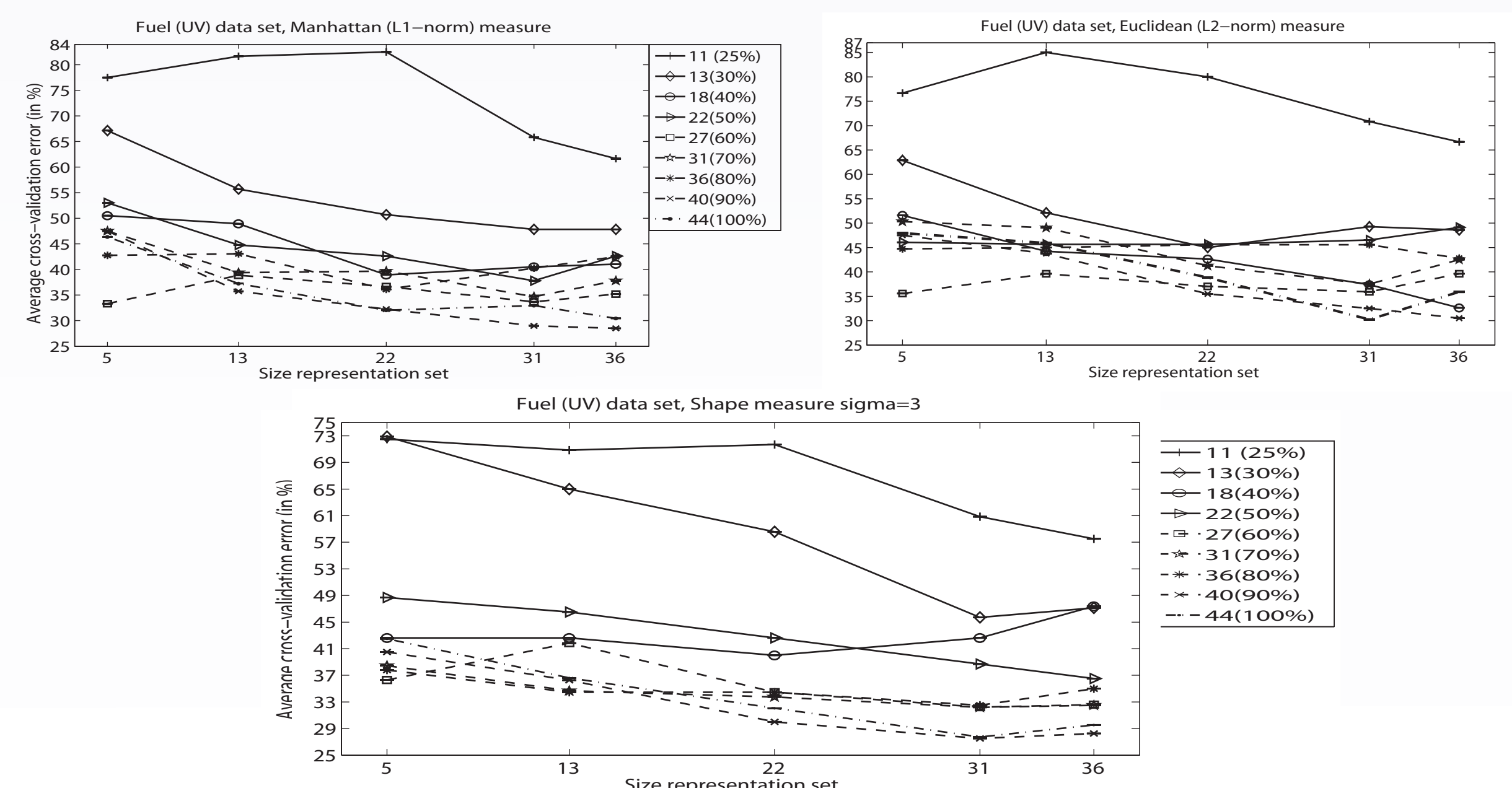
### Tecator (215 x 100) : High fat content (77) and Low fat content (138) NIR spectra



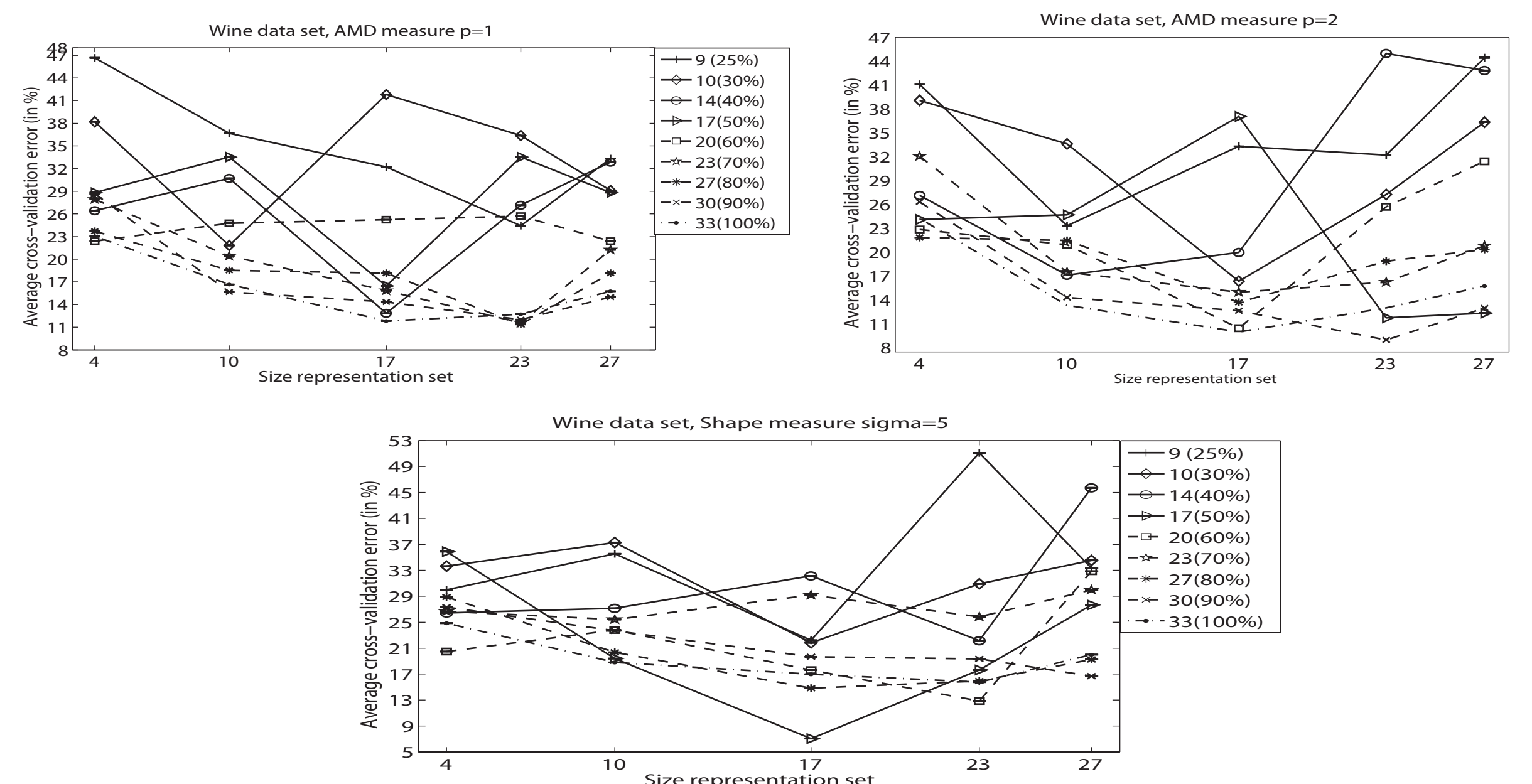
### Fuel (31 x 3528) : Regular gasoline (16) and Special gasoline (15) FT-IR spectra



### Fuel (44 x 127) : Regular gasoline (23) and Special gasoline (21) UV-VIS spectra



### Wine (33 x 2700 x 200) : South America (21) and Australia (12) GC-MS



## Conclusions

- DR seems to be a good solution to tackle the curse-of-dimensionality problem in most spectral data sets, if a suitable dissimilarity measure is used.
- It is important to take into account the functional nature of spectra in its representation.
- It is shown that if we use the shape information, a few samples are enough for classifiers to learn well.