

# Constructing and Evaluating Summary Data Sets for the Atmospheric Infrared Sounder

Amy Braverman and Eric Fetzer  
Jet Propulsion Laboratory  
Mail Stop 169-237  
4800 Oak Grove Drive  
Pasadena, CA 91109-8099  
Amy.Braverman@jpl.nasa.gov

## 1. Synopsis

The Atmospheric Infrared Sounder (AIRS) was launched into polar orbit on May 4, 2002 aboard NASA's EOS satellite, Aqua. Every day the instrument yields about 28.8 GB of data organized in 240 files: 120 ascending the daylight side of Earth, and 120 descending on the night side. Files contains 135 lines (along-track) of 90 footprints (cross-track) each. In each footprint AIRS measures radiance in 2378 spectral bands. A prime objective of the EOS program is to compile global, long-term data sets for climate change studies by the research community. All but the most well equipped users will find accessing and manipulating 28.8 GB per day collected over a period of years to be unmanagable. One traditional solution to this problem is to produce global maps of means and standard deviations for each parameter of interest. Typically, this is done on a one degree spatial grid, and data are summarized over a period of a day or month. Unfortunately, this discards a substantial amount of potentially important information about the data distribution. Means and standard deviations only fully characterize a distribution if it is normal, and capture no information about joint relationships among parameters. Covariances or correlations are sometimes reported to provide this information, but their numbers become large as the square of the number of parameters. Finally, summaries based on moments contain no information about outliers, which may be among the most interesting data for science analysis.

We propose a summary product that describes data belonging to each one degree grid cell by a set of multivariate representatives and weights and errors associated with them. For example, in the exercise below we summarize three days worth of AIRS radiances at eleven of the 2378 channels. Whereas traditional summary products might provide eleven maps of means, eleven maps of standard deviations, and some of the  $11(11-1)/2 = 55$  covariances or correlations, we provide a data product having  $K$  representative 11-vectors for each of the  $(180 \times 360 =)$  64,800 one degree grid cells. Each representative stands in for some number of the original 11-tuples of radiances acquired in the grid cell. That number is given by counts associated with the representatives.  $K$  may vary from grid cell to grid cell depending on how many representatives are required to adequately describe their data. Users may perform computations appropriately weighted by count, to estimate arbitrary functions of the data from this much smaller proxy data set. We also report the average squared euclidian distance between the representatives and the observations they stand for as a measure of error incurred using the representaives in place of the original data.

How useful this data product is for science analysis depends on whether conclusions drawn from the summary product are reasonably close to those that would have been drawn from the original data. Below we demonstrate that this is indeed the case for a simple analysis of AIRS radiance data.

## 2. AIRS Test Data

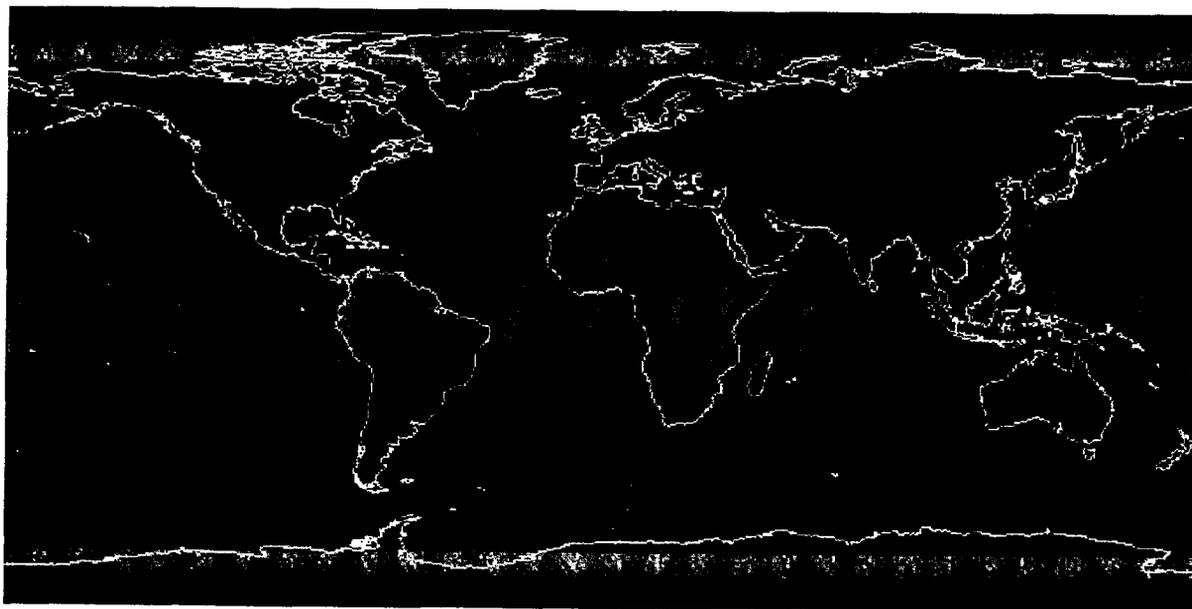
- AIRS brightness temperature in eleven channels, acquired from descending (nighttime) granules, July 20-22, 2002.
- One datum:  $x_{lat,lon} = (x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})'$ .

<i>Variable</i>	<i>Frequency</i>	<i>Measures</i>
$x_0$	724.742	temperature, high altitude
$x_1$	735.607	temperature, mid-atlitude
$x_2$	755.237	temperature, low-altitude
$x_3$	917.209	total window
$x_4$	1231.19	window
$x_5$	1285.323	total water vapor/methane
$x_6$	1345.174	total water vapor
$x_7$	2412.562	window
$x_8$	2450.02	window
$x_9$	2500.313	window
$x_{10}$	2616.095	window

Table 1: AIRS test data channels.

- Number of 11-dimensional data points per 1° grid cell:

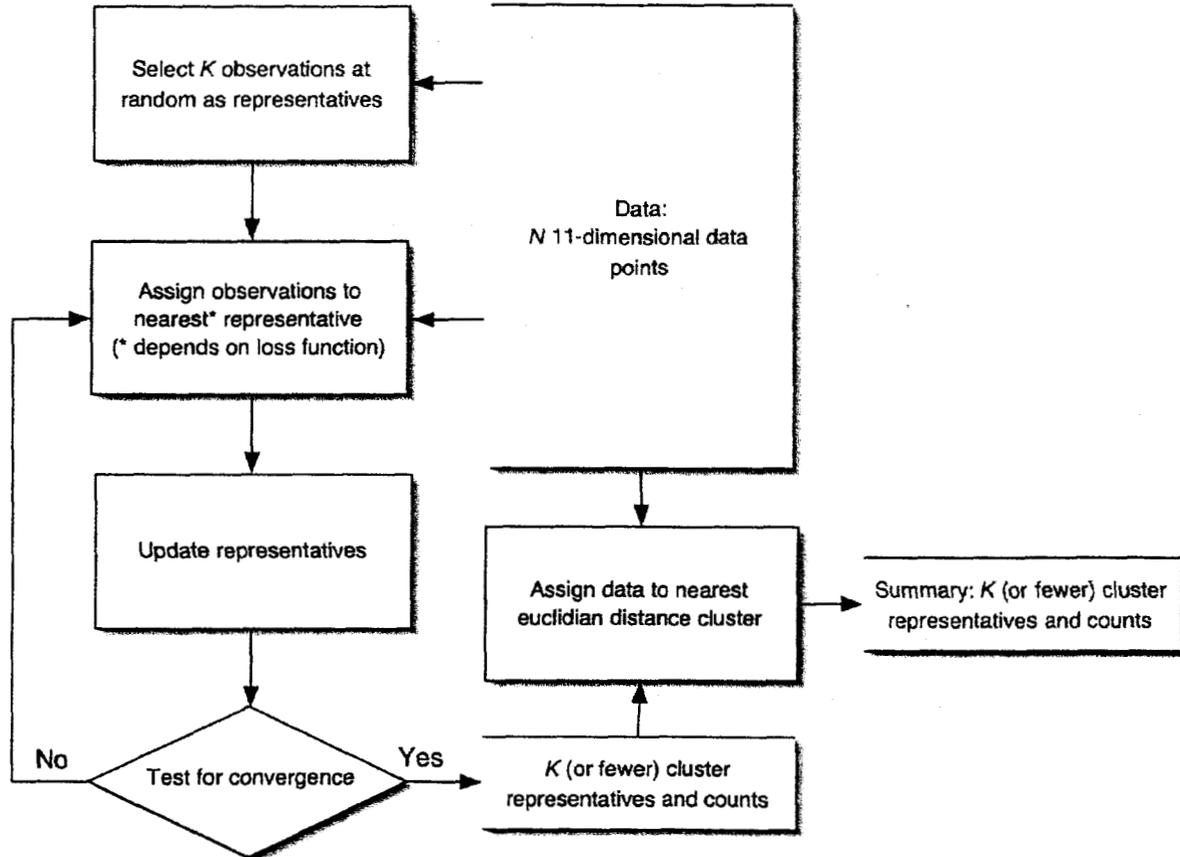
N\_20-22\_JUL2002



1.00000  117.000

### 3. Summarization Algorithm

1. Partition data points according to membership in  $1^\circ$  grid cell.
2. For each  $1^\circ$  subset, apply a clustering algorithm such as  $K$ -means (Macqueen, 1967) or Entropy-constrained Vector Quantization (ECVQ; Chou, Lookabaugh and Gray, 1989).



3. If the loss is  $l(x, y) = \|x - y\|^2$ , where  $y$  is representative to which  $x$  is assigned, the algorithm is  $K$ -means. If the loss is  $l(x, y) = \|x - y\|^2 + \lambda(-\log p)$ , where  $\lambda$  is a lagrange multiplier and  $p$  is the proportion of grid cell data points assigned to  $y$ , the algorithm is ECVQ. In either case, the update step is averaging.

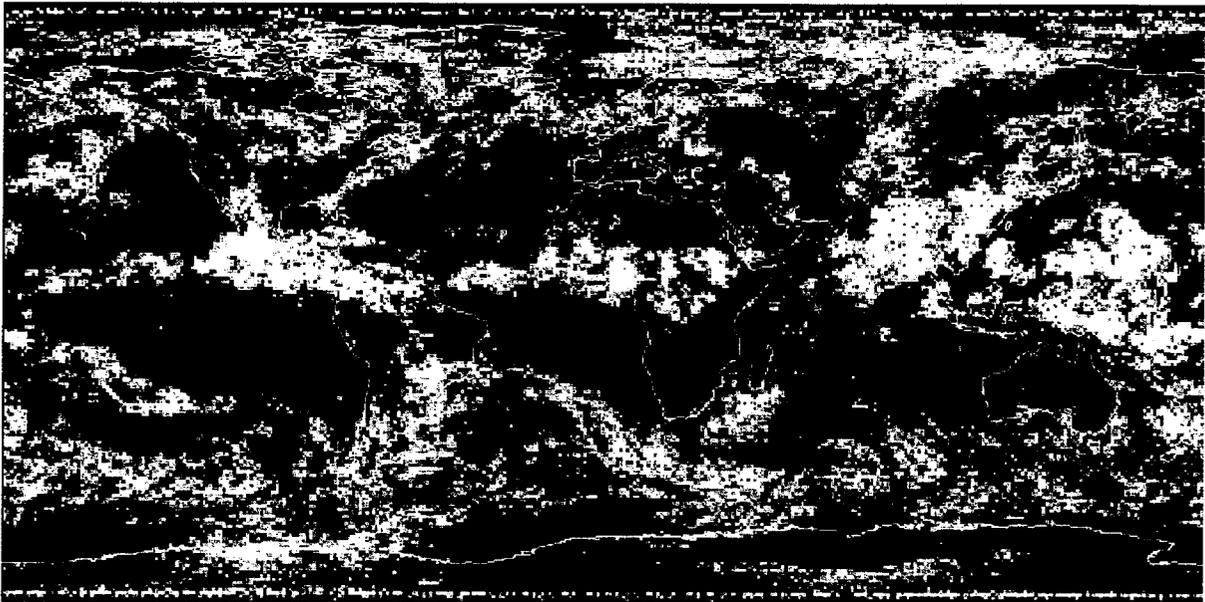
4. Result:  $K \ll N$

Original		Summarized			
Data Value	Cluster Assignment	Cluster Index	Cluster Representative	Cluster Count	Cluster Error
$x_0$	1	0	$y_0 = \text{avg}(x_{N-1})$	$N_1 = 1$	$D_0 = 0$
$x_1$	$K - 1$	1	$y_1 = \text{avg}(x_0, x_2)$	$N_2 = 2$	$D_1 = \text{avg}(\ x_0 - y_1\ ^2, \ x_2 - y_1\ ^2)$
$x_2$	1	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$K - 1$	$y_{K-1} = \text{avg}(x_1)$	$N_{K-1} = 1$	$D_{K-1} = 0$
$x_{N-1}$	0				

#### 4. Summarization Results

- ECVQ algorithm applied to AIRS test data using a maximum of  $K = 15$  clusters per grid cell.
- Processing time: approximately 8 hours on four, 400 MHz Sun processors.
- Input data volume:  $\approx 550$  MB. Output:  $\approx 65$  MB.
- Number of clusters representing the data by grid cell:

K\_20-22\_JUL\_2002



1.00000 15.0000

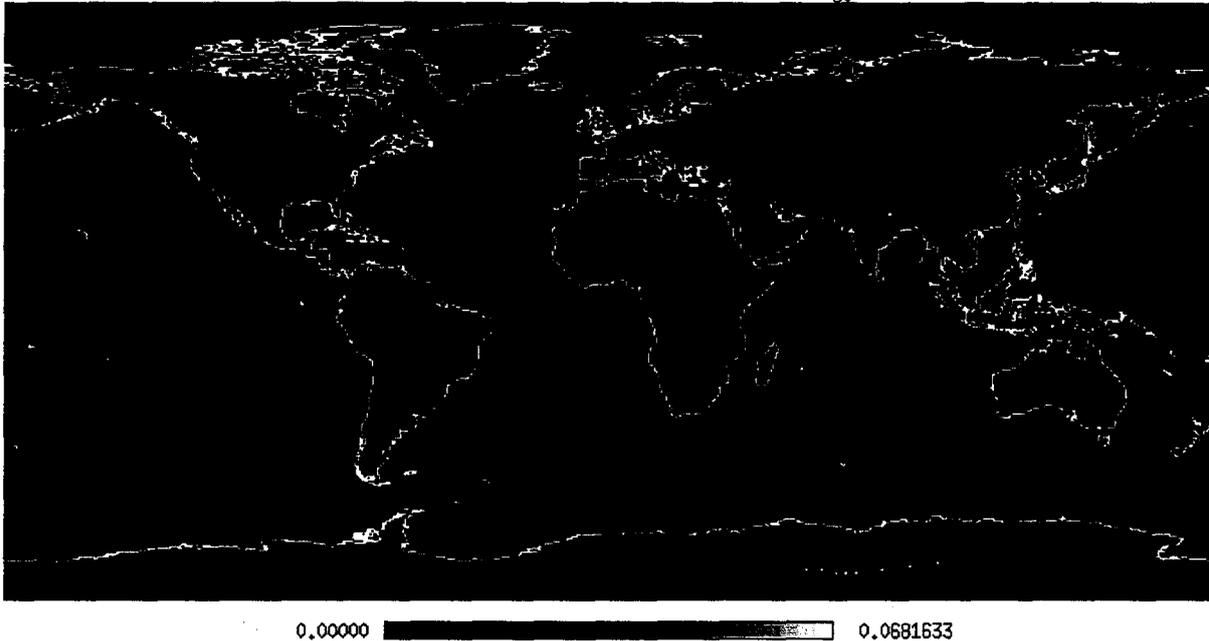
- Compare to the image in Section 2, and note that grid cells with large numbers of data points are not necessarily the same ones with lots of clusters.
- ECVQ allocates clusters according to information-theoretic data complexity, not numbers of data points.

## 5. Summarization Quality

- Average error by grid cell,  $D = N^{-1} \sum_{k=0}^{K-1} (N_k D_k)$ , relative to estimated average squared data vector norm,  $N^{-1} \sum_{k=0}^{K-1} (N_k \|y_k\|^2)$ :

Dx\_20-22\_Jul\_2002

90



- Relative errors are small; expect computations based on summaries to be close to computations based on raw data.
- How close depends on the nature of the computation.
- We think of the summaries as the "model" and the raw data as the "truth". Hence, this is a model testing problem with a twist: model quality depends on the use to which it is put.

## 6. Sample Analysis

- Can we distinguish between clear and cloudy scenes in the AIRS test data?
- Proposition: since different channels see different levels of the atmosphere, cloudy data points should have relatively homogeneous brightness temperatures across ten of the eleven channels (excluding  $x_0$ , which observes at very high altitude). If the view to the ground is obscured by a cloud, the channels all see the same thing and show similar brightness temperatures. If a data point represents a clear scene, the channels see different things and will be heterogeneous.
- For each data point,  $\mathbf{x} = (x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})'$ , we can compute the standard deviation of the last ten components:

$$w = \sqrt{\frac{1}{10} \sum_{j=1}^{10} (x_j - \bar{x})^2}, \quad \text{where } \bar{x} = \frac{1}{10} \sum_{j=1}^{10} x_j.$$

To account for differences of scale, we also examine the coefficient of variation,  $c = w/\bar{x}$ .

- In each grid cell we compute average values of  $w$  and  $c$ :

$$\bar{w} = \frac{1}{N} \sum_{n=1}^N w_n \quad \text{and} \quad \bar{c} = \frac{1}{N} \sum_{n=1}^N c_n,$$

where  $w_n$  and  $c_n$  are the standard deviation and coefficient of variation of the  $n$ th data point in a grid cell, and  $N$  is the number of data points in the grid cell.

- How well can  $\bar{w}$  and  $\bar{c}$  be estimated from the summary data?
- Estimates:

$$\hat{w} = \frac{1}{N} \sum_{k=1}^K N_k \sqrt{\frac{1}{10} \sum_{j=1}^{10} (y_{kj} - \bar{y}_k)^2}, \quad \text{where } \bar{y}_k = \frac{1}{10} \sum_{j=1}^{10} y_{kj},$$

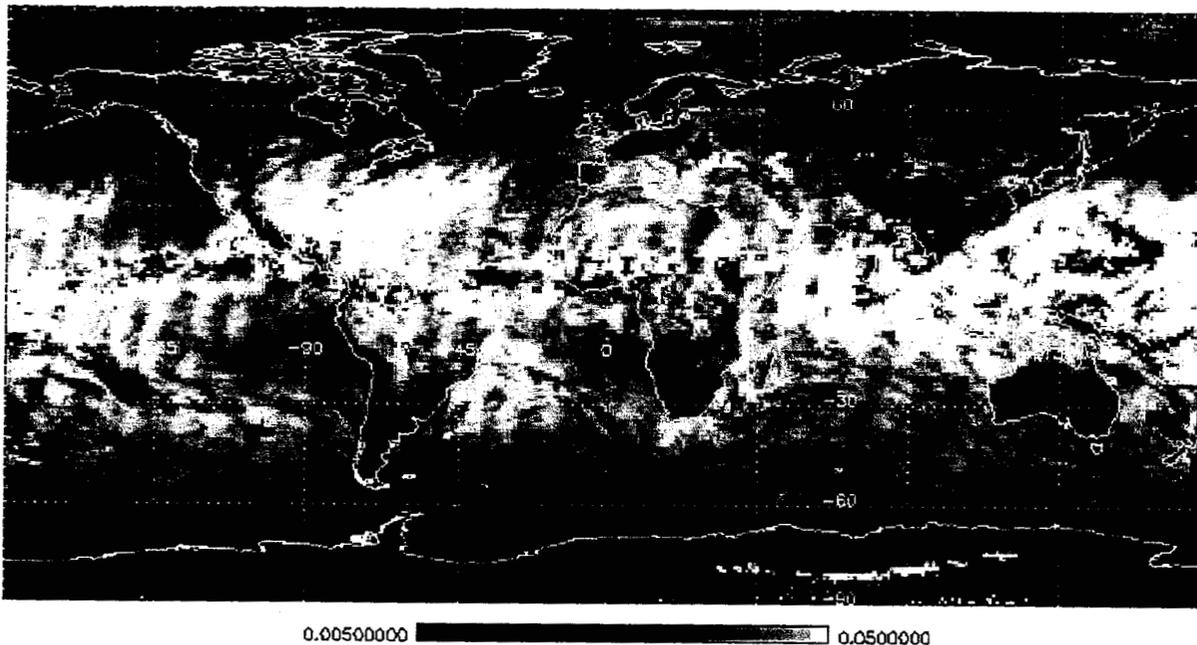
$$\hat{c} = \frac{\hat{w}}{\hat{x}}, \quad \text{where } \hat{x} = \frac{1}{N} \sum_{k=1}^K N_k \frac{1}{10} \sum_{j=1}^{10} y_{kj} = \bar{x},$$

and  $y_{kj}$  is the  $j$ th element of the  $k$ th representative for the grid cell.

## 7. Empirical Comparison

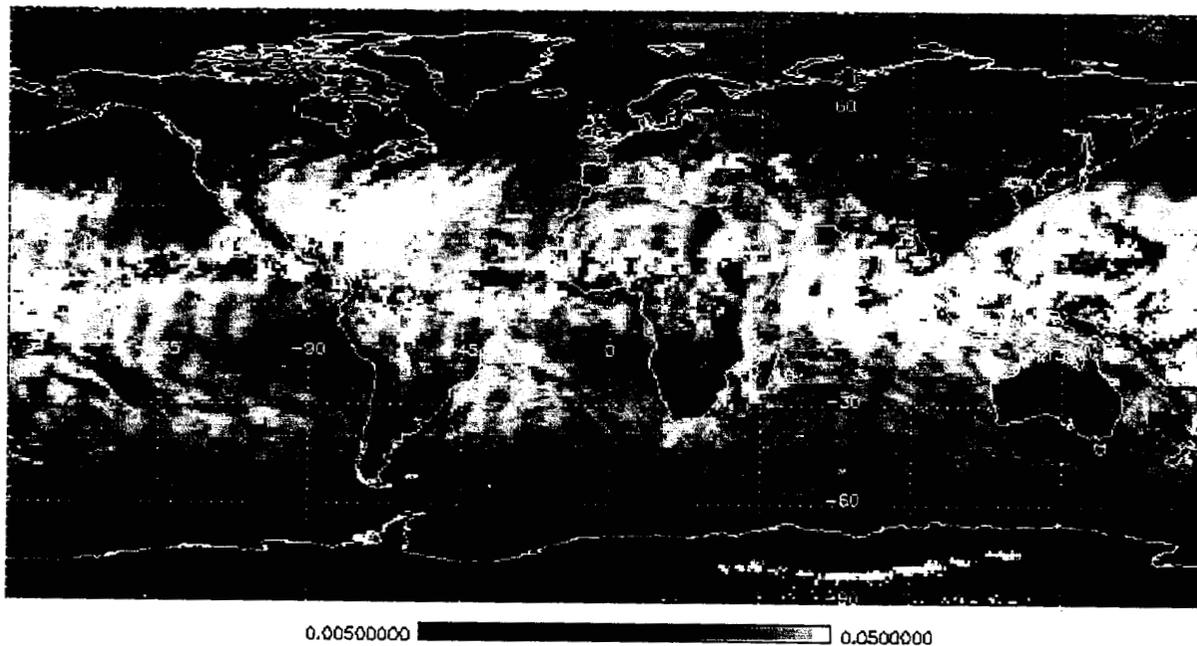
- Estimated average coefficient of variation of brightness temperatures by grid cell:

CV\_Brightness\_20-22\_Jul\_2002



- True average coefficient of variation of brightness temperatures by grid cell:

True\_CV\_Brightness\_20-22\_Jul\_2002



## 8. Conclusions

- For linear functions, calculations using summaries accurately reproduce the truth.
- For nonlinear, differentiable functions, we can approximate accuracy using a Taylor Expansion. For arbitrary functions we must look at test cases and try to understand where small changes in input data will result in large changes in output. Where large discrepancies between the data and the summaries exist, we must beware. "Large" depends on the analysis being conducted.
- In this example we have the luxury of being able to compute the truth. If that were true universally, there would be no need for this methodology to begin with. However, these results give us confidence that similarly summarized three day chunks of AIRS data can be used to find candidate clear scenes by examining brightness temperature heterogeneity.
- In those cases where we can't compute the truth globally, we suggest doing so for selected areas. The areas should be geographically stratified.

## References:

- Chou, P.A., Lookagaugh, T., and Gray, R. (1989). Entropy-constrained Vector Quantization, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 31-42. Macqueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-296.