Understanding Local Structure Globally in Earth Science Remote Sensing Data Sets

Amy Braverman    Eric Fetzer

Jet Propulsion Laboratory
California Institute of Technology

August 1, 2007
Introduction

AIRS Level 2 Observations

The Idea

Algorithm

Analysis

Conclusions
Empirical probability distributions derived from the data are the signatures of physical processes generating the data.

Distributions defined on different space-time windows can be compared; differences or changes attributed to physical processes.

Approach:
- partition data on coarse, spatio-temporal grid (e.g. seasonal or monthly $5^\circ \times 5^\circ$);
- summarize each grid cell by a multivariate distribution estimate;
- use a “distance” between distributions as a basis for data mining.
Each L2 footprint represented by a vector of length 35:

<table>
<thead>
<tr>
<th>Indices</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-11</td>
<td>atmospheric temperature (tair) at 11 levels</td>
</tr>
<tr>
<td>12-22</td>
<td>water vapor (h2ommr) at 11 levels</td>
</tr>
<tr>
<td>23-32</td>
<td>cloud fraction (cldfrc) at 10 levels; excludes surface</td>
</tr>
<tr>
<td>33</td>
<td>land/water indicator</td>
</tr>
<tr>
<td>34</td>
<td>day/night indicator</td>
</tr>
<tr>
<td>35</td>
<td>quality indicator</td>
</tr>
</tbody>
</table>

- 240 files per day, each $45 \times 30$ footprints, since September 2002.
- Monthly data volume: $240 \times 45 \times 30 \times 35 \times 8 \times 30 = 2.72$ GB/month.
The Idea

- How to reduce data in a way that preserves information?

- Signal processing paradigm: rate-distortion theory.
The Idea

\[ \delta_k = \frac{1}{N_k} \sum_{i=1}^{N} \| x_i - y_k \|^2 1[\alpha(x_i) = k] \]

\[ N_k = \sum_{i=1}^{N} 1[\alpha(x_i) = k], \quad K \ll N \]

\[ \mathcal{I} = \{1, 2, \ldots, K\} \]

\[ \alpha : x \rightarrow \mathcal{I} \]

\[ y_k = \beta(k) = \frac{1}{N_k} \sum_{i=1}^{N} x_i 1[\alpha(x_i) = k] \]
\[ \delta_k = \frac{1}{N_k} \sum_{i=1}^{N} \| x_i - y_k \|^2 1[\alpha(x_i) = k] \]

\[ N_k = \sum_{i=1}^{N} 1[\alpha(x_i) = k], \quad K \ll N \]

\[ \mathcal{I} = \{1, 2, \ldots, K\} \]

\[ \alpha : \mathbf{x} \rightarrow \mathcal{I} \]

\[ y_k = \beta(k) = \frac{1}{N_k} \sum_{i=1}^{N} x_i 1[\alpha(x_i) = k] \]
The Idea

\[ \delta_k = \frac{1}{N_k} \sum_{i=1}^{N} \| x_i - y_k \|^2 1[\alpha(x_i) = k] \]

\[ N_k = \sum_{i=1}^{N} 1[\alpha(x_i) = k], \quad K \ll N \]

\[ I = \{1, 2, \ldots, K\} \]

\[ \alpha : x \to I \]

\[ y_k = \beta(k) = \frac{1}{N_k} \sum_{i=1}^{N} x_i 1[\alpha(x_i) = k] \]

\[ \Delta = E\| X - \beta[\alpha(X)] \|^2 \]

\[ H = E(|\gamma(\alpha(X))|) \]

\[ \gamma(k) = -\log_2 \left( \frac{N_k}{N} \right) \]

Compressed distribution

\[ P(Y = y) \]

\[ P(X = x) \]

Source encoder \( \alpha(x) \)

Channel encoder \( \gamma(k) \)

Channel (noiseless)

Source decoder \( \beta[\alpha(x)] \)

\[ \gamma^{-1}(k) \]

Raw distribution

\[ y_1, y_2, \ldots, y_K \]

\[ x_1, x_2, \ldots, x_N \]
The Idea

\[ P(X = x) \]

\[ \Delta = E \| X - \beta[\alpha(X)] \|^2 \]

\[ H = E (|\gamma(\alpha(X))|) \]

\[ \gamma(k) = -\log_2 \left( \frac{N_k}{N} \right) \]

\[ \delta_k = \frac{1}{N_k} \sum_{i=1}^{N} \| x_i - y_k \|^2 1[\alpha(x_i) = k] \]

\[ N_k = \sum_{i=1}^{N} 1[\alpha(x_i) = k], \quad K << N \]

\[ \mathcal{I} = \{1, 2, \ldots, K\} \]

\[ \alpha : x \rightarrow \mathcal{I} \]

\[ y_k = \beta(k) = \frac{1}{N_k} \sum_{i=1}^{N} x_i 1[\alpha(x_i) = k] \]

Compressed distribution

\[ P(Y = y) \]

\[ P(X = x) \]

\[ \Delta \]

Source encoder \( \alpha(X) \)

Channel encoder \( \gamma(k) \)

Channel (noiseless)

Source decoder \( \beta[\alpha(X)] \)

Channel decoder \( \gamma^{-1}(k) \)
The Idea

- A data set is a stochastic information source: draw from it randomly and with replacement.

- Theoretical rate-distortion function is a property of the information source.

- Estimate the convex hull of the rate-distortion function empirically.

The Idea

- Also need to preserve local data structure within global context.

- Needed for comparative climatological studies and for data mining.
“Gridded" multivariate distribution estimates with similar fidelity to raw data implies comparable information.

Data mining: can we characterize the evolution of multivariate distributions across space, time, and resolution?
Algorithm

ECVQ module: repeat S times

ECVQ loss function:

\[ L_n = \frac{1}{N} \sum_{i=1}^{N} \left[ \| x_i - \beta(\alpha(x_i)) \|^2 + \lambda \left( -\log_2 \frac{N\alpha(x_i)}{N} \right) \right] \]

Randomly assign data points to K clusters

Compu cluster centroids and counts

Reassign data points to minimum loss clusters

Test for convergence

No

Yes

K or fewer representatives

S sets of representatives

K or fewer representatives

K or fewer representatives

Assign to closest representative, update

Assign to closest representative, update

Assign to closest representative, update

K or fewer representatives with estimated distortions

K or fewer representatives with estimated distortions

K or fewer representatives with estimated distortions

Input data for one grid cell

Best summary: K or fewer representatives

Choose minimum distortion representatives

Assign to closest representative, update

Convert back to data space

Best grid cell summary K or fewer representatives, counts and distortions

Standardize/project to PC space

Design Sample

S design samples

1 cross-validation sample

Standardize/project to PC space

Design Sample

Standardize/project to PC space

Design Sample

Standardize/project to PC space

Cross-validation Sample

Run in stages, partially at GDAAC and partially at JPL.

Average compressed data set size: 12.5 MB/month.

- Winter $200x = \text{Dec } 200x, \text{Jan } 200(x+1), \text{and Feb } 200(x+1)$.
- Constructed from constituent monthly summaries.

Winter 2002, grid cell [-10,-60] (Amazon)

- Look at relationships among grid cells (globally) for each winter (data mining).
Global Summaries

Winter 2002

Graphs showing data for h2ommr, cldfrc, and tair with various colored lines and km and tair scales.
Winter 2003
Global Summaries

Winter 2004

Graphs showing data for h2ommr, cldfrc, and tair.
Global Summaries

Winter 2005

[Graphs showing data for different parameters such as h2o, mmr, cldfrc, and tair over a range of km and degrees.]
Global Spatial Relationships

- The "distance" between two distributions (a measure of similarity).
- $36 \times 72 = 2592$, $5^\circ \times 5^\circ$ degree grid cells, each containing a distribution.
- Form a $2592 \times 2592$ symmetric distance matrix.
- Use multidimensional scaling (MDS) to analyze the distance matrix.
Distance Between Distributions

- Distance between two distributions, \( p \) and \( q \): \( \Delta(p, q) \)

- \( X, Y, p \)'s and \( q \)'s are given.
- Find \( \pi \)'s to maximize \( \text{Corr}(X, Y) \)

- \( p_i = P(X = x_i) \)
- \( X \) is a random draw from grid cell 1

<table>
<thead>
<tr>
<th></th>
<th>( q_1 )</th>
<th>( q_2 )</th>
<th>( q_3 )</th>
<th>( q_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 )</td>
<td>( \pi_{11} )</td>
<td>( \pi_{12} )</td>
<td>( \pi_{13} )</td>
<td>( \pi_{14} )</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>( \pi_{21} )</td>
<td>( \pi_{22} )</td>
<td>( \pi_{23} )</td>
<td>( \pi_{24} )</td>
</tr>
<tr>
<td>( p_3 )</td>
<td>( \pi_{31} )</td>
<td>( \pi_{32} )</td>
<td>( \pi_{33} )</td>
<td>( \pi_{34} )</td>
</tr>
</tbody>
</table>

\[
\Delta(p, q) = \min_{\pi: \sum_i \pi_{ij} = \sum_j \pi_{ij}} \frac{E\|X - Y\|^2}{p_i = \sum_j \pi_{ij}, \quad q_j = \sum_i \pi_{ij}}
\]
Distance Matrix

\[ \Delta(p, q) \]

2592 5° x 5° grid cells
For each object (e.g. grid cell) in a distance matrix, represent it as a point in a low dimensional (e.g. 2-dimensional) space.

Situate the points in the low dimensional space so that relative inter-point distances approximate those in the distance matrix.
Multidimensional Scaling

(Color indicates latitude zone.)
Multidimensional Scaling
Multidimensional Scaling
Multidimensional Scaling

Histogram of MDS1, Winter 2002

Histogram of MDS1, Winter 2003

Histogram of MDS1, Winter 2004

Histogram of MDS1, Winter 2005
Multidimensional Scaling

2002 MDS1 LT -6

[Graphs showing multidimensional scaling plots for different columns.]
Multidimensional Scaling

2002 MDS1 GT -6 and LT -5
Multidimensional Scaling

2002 MDS1 GT -5 and LT -4
Multidimensional Scaling

2002 MDS1 GT -4 and LT -3
2002 MDS1 GT -3 and LT -2
Multidimensional Scaling

2002 MDS1 GT -2 and LT -1
Multidimensional Scaling

2002 MDS1 GT -1 and LT 0
Multidimensional Scaling

2002 MDS1 GT 0 and LT 1
Multidimensional Scaling

2002 MDS1 GT 1 and LT 2
Multidimensional Scaling

2002 MDS1 GT 2 and LT 3

---

The diagrams show the relationship between different variables using Multidimensional Scaling (MDS) techniques. The axes represent different parameters, and the data points are plotted to illustrate the similarities and differences among them. The plots help in visualizing the multidimensional data in a two-dimensional space.
Multidimensional Scaling

2002 MDS1 GT 3 and LT 4
Multidimensional Scaling

2002 MDS1 GT 4 and LT 5
Multidimensional Scaling

2002 MDS1 GT 5 and LT 6
Multidimensional Scaling

2002 MDS1 GT 6 and LT 7
Multidimensional Scaling

2002 MDS1 GT 7
Conclusions

- MDS1 captures “tropicalness”.
- MDS2? Rainfall?
- Inter-winter differences shown in MDS plots are striking; represent changes in relative relationships among grid cells.
- 2003, 2005 similar; 2004 is different.
- What is the physical interpretation/explanation?

Comments and suggestions welcome! Send email to Amy.Braverman@jpl.nasa.gov