The Application of Principal Component Analysis (PCA) to AIRS Data Compression

Lihang Zhou, Mitchell D. Goldberg, Walter W. Wolf, Larry Mcmillin, Chris Barnet
NOAA/NESDIS/ORA, Camp Springs, MD 20746;
*Q55 Group, Inc, Lanham, MD 20706

1. INTRODUCTION

Why Data Compression?

- Data compression is a topic of much importance to data archives, in regards to the access and distribution of the new generation of high spectral resolution infrared sounders.

- NOAA/NESDIS is processing and distributing AIRS data and products in near real-time. This offers us a great opportunity to use real AIRS observations for a test bed for data compression studies of the future operational hyper spectral sounding instruments.

How should we do it?

- Instead of using some general-purpose compression technique, we want to develop our application in a way that can exploit the special characteristics of hyper spectral data, such as:
  - Redundant information:
    Observations from hyper spectral instruments are not independent. Principal Components Analysis (PCA) is the most economical way to extract the independent information and reduce the dimension of the data.

- Small reconstruction residues:
  The differences between the PCA reconstructed brightness temperature and the original observation are mostly very small numbers.

- Stable distribution:
  The range and distribution of the residues vary little from granule to granule.

In this poster we present our studies of the applications of PCA to AIRS data compression, and our experiments for achieving the near loss-less compression.

Implementation of PCA

- The method to be used to generate and apply the eigenvectors is described in detail in Goldberg et al.
- For data compression, we use the dependent eigenvector to compute the PCS, to achieve higher compression ratios.
- Since timing is one of the biggest concerns when running operational algorithms, we divided the complete AIRS spectrum into three bands to speed up the computation, as show in table 2.

Table 2: Calculation Time for Three Spectral Bands:

<table>
<thead>
<tr>
<th>Band</th>
<th>Channel</th>
<th>Frequency</th>
<th>Calculation Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>688</td>
<td>650-920</td>
<td>2m56.87s</td>
</tr>
<tr>
<td>2</td>
<td>689</td>
<td>920-1400</td>
<td>2m59.19s</td>
</tr>
<tr>
<td>3</td>
<td>719</td>
<td>1400-2650</td>
<td>3m39.11s</td>
</tr>
</tbody>
</table>

2. APPROACH

3. RESULTS

Lossy Compression

- Figures in this sector display the RMS error and other statistical information, when using different numbers of PCS to reconstruct brightness temperatures, and then comparing them with the original observations.
- Such information can be served as statistical metadata for archiving.

Near Lossless Compression

- Figures above show the BIAS and RMS of reconstructed brightness temperature (Blue curve), and reconstructed plus the residues (0.1K-Green curve; 0.01K-red curve), compared with the observed data.
- For 0.1K accuracy, the RMS are mostly less than 0.05K, the Bias are less than 0.002K. For 0.01K, both RMS and Bias appear to be close to zero.

- Although Huffman code is appropriate, an even simpler and more efficient compression technique to use is Elias Gamma Coding, which does not need to store the tree for decoding.

- We used Huffman coding to code the residuals. Below is a chart that shows the average bit lengths needed for keeping the brightness temperature accuracy to 0.01K, and to 0.1K, respectively.

Encoding Residues

4. Summary

Where We Are

- Lossy compression: 40 PCS: ~11
- Near lossless: 40 PCS + residual:
  - Store brightness temperature to 0.01K: ~4
  - Store brightness temperature to 0.1K: ~8

- To develop a data compression system using the approach presented in this poster to be used for archiving the compressed AIRS level 1B data and the corresponding metadata.
- To develop web capability to make the compressed 1B data accessible by the interested users.
- Continued optimizing of the compression and residue encoding technique.

REFERENCES:


- Comments/feedback: please email to: Lihang.zhou@noaa.gov

Figure 1: Channels of sounders (from Chris Barnet)

Figure 2: Residuals are mostly very small numbers, and the distribution is stable.

Figure 3: Elias Gamma Coding, which is like a static Huffman Coding, seems to be a more efficient way to encode the residues.

Figure 4: Average Bit Length

Figure 5: RMS and BIAS of reconstructed brightness temperature (Blue curve), and reconstructed plus the residues (0.1K-Green curve; 0.01K-red curve), compared with the observed data.

Table 1: Residues

<table>
<thead>
<tr>
<th>Residue Range</th>
<th>Band 1</th>
<th>Band 2</th>
<th>Band 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.1</td>
<td>30.30%</td>
<td>67.50%</td>
<td>52.07%</td>
</tr>
<tr>
<td>&lt;0.4 &amp;&gt;0.1</td>
<td>50.30%</td>
<td>31.70%</td>
<td>42.18%</td>
</tr>
<tr>
<td>&lt;0.8 &amp;&gt;0.4</td>
<td>15.89%</td>
<td>0.79%</td>
<td>5.30%</td>
</tr>
<tr>
<td>&lt;1.6 &amp;&gt;0.8</td>
<td>3.16%</td>
<td>0.01%</td>
<td>0.45%</td>
</tr>
<tr>
<td>&gt;3.2 &amp;&gt;1.6</td>
<td>0.199%</td>
<td>0.0005%</td>
<td>0.0079%</td>
</tr>
<tr>
<td>&gt;=3.2 &amp;&gt;1.6</td>
<td>0.0001%</td>
<td>0.0000%</td>
<td>0.0000%</td>
</tr>
</tbody>
</table>

File Type | File Size (MB)
---|---
L1B | 121
Eigenvector | 1/1/0.7
PC Scores (40pcs) | 1.9/1.9/1.9
Encoded (0.1K) | 4.0/2.0/2.3
Encoded (0.01) | 9.7/5.9/5.6

Above table gives the file size for original and compressed files.