Reducing Size and Complexity of Remote Sensing Data Sets

Amy Braverman
Jet Propulsion Laboratory,
California Institute of Technology
Motivation

- Earth Observing System satellites (Terra, Aqua, Aura) to return vast quantities of data.

- New type of data provides a global view of local phenomena.

- Data to be made available to the public for analysis, but large segment of the user community can’t handle the volume.
Multi-angle Imaging SpectroRadiometer (MISR) aboard Terra produces about 2 TB per month of radiance and geophysical data.
Motivation

Strategy

- Partition data on a monthly, global, one degree by one degree grid.

- Replace original data in each cell with a set of representatives and associated weights called a summary.

- Discrete distribution so defined should be close to the original empirical distribution, but also parsimonious.

- Similar to data squashing (DuMouchel, 2001), quantization (Cover and Thomas, 1991).
\[ \alpha \text{ assigns } y\text{'s to clusters:} \]
\[ \alpha(y) = k. \]

Compressed Data, Quantized Data, Summary

\[ \beta(k) \text{ produces cluster means:} \]
\[ \beta(k) = \frac{1}{N(k)} \sum_{n=1}^{N} y_n 1[\alpha(y_n) = k]. \]

\[ N(k) \text{ produces cluster counts:} \]
\[ N(k) = \sum_{n=1}^{N} 1[\alpha(y_n) = k]. \]

\[ q(y) \text{ is the average of the cluster to which } y \text{ belongs:} \]
\[ q(y) = \beta[\alpha(y)]. \]

R.V. version: \( Q = q(Y) = E(Y|Q). \) \( \Rightarrow E(Q) = E(Y), \]
\[ Var(Q) \leq Var(Y), \]
\[ Cov(Q,Y) = Cov(Q). \]

Two figures of merit for \( q, \) or equivalently, \( \alpha: \)
\[ \Delta(q) = \frac{1}{N} \sum_{n=1}^{N} \| y_n - q(y_n) \|^2 = trCov(Y - Q). \]
\[ h(q) = - \sum_{k=1}^{K} \frac{N(k)}{N} \log \frac{N(k)}{N}. \]
Want fewer clusters when that will do ⟷ want distortions as similar as possible.
Algorithm

Assign to $K$ clusters, compute $\beta(k)$’s.

Reassign rows to minimize $d(y, k)$.

Update $\beta(k)$’s and $N(k)$’s. Delete empty clusters.

Test for convergence, Yes

Minimize $L_\lambda = \frac{1}{N} \sum_{n=1}^{N} d(y, k)$ by choice of $\alpha$:

$$d(y, k) = ||y - \beta(k)||^2 + \lambda \left[ -\log \frac{N(k)}{N} \right].$$

Original application: estimate distortion-rate functions of information sources.

$\Delta(q^*)$

$\lambda_1 \lambda_2 \lambda_3 \lambda_4 \lambda_5$

$h(q^*)$

Computationally intensive, depends on initial random assignment, not distortion-minimizing.
Algorithm

Assign to $K$ clusters, compute $\beta(k)$'s.

Reassign rows to minimize $d(y, k)$.

Update $\beta(k)$'s and $N(k)$'s. Delete empty clusters.

Test for convergence.

Yes

No

Samples

Summary

\[ \left\{ \tilde{\beta}(k), \tilde{N}(k) \right\}_{k=1}^{K} \]

Minimizes error.

Iteration and multiple scans on samples only.

Best preliminary summary is the one with smallest $\hat{\Delta}$.

$\hat{\Delta}$ is a goodness-of-fit measure.

Average of $\hat{\Delta}$'s, $\bar{\Delta}$, is a process performance measure which accounts for sampling variation.
MISR aerosol data over southern Africa, August-September 2000.

$y_n$ is a six-dimensional observation (row of data):
$y_n = (\tau_n, \chi_{n1}, \chi_{n2}, \chi_{n3}, \chi_{n4}, \chi_{n5})$ with a latitude and longitude, representing a $17.6 km^2$ region.

$\tau_n$ is optical depth. $\chi_{nj}$’s measures how close the vector of observed MISR radiances is from that predicted by aerosol model $j$.

Original data: 6,304,861 observations.
$\sum_{lat,lon} \tilde{K} = 9,322$ observations.
Is there a relationship between optical depth and heterogeneity of the model-fit χ’s?

Measure heterogeneity by \( w_n = \frac{1}{5} \sum_{j=1}^{5} (\chi_{nj} - \bar{\chi}_n)^2 \).

Compare: \( \rho(\tau, w) \) computed using original data to \( \hat{\rho}(\tau, w) \) computed from summaries:

\[
\rho(\tau, w) = \frac{\sum_{n=1}^{N} (\tau_n - \bar{\tau})(w_n - \bar{w})}{\sqrt{\sum_{n=1}^{N} (\tau_n - \bar{\tau})^2} \sqrt{\sum_{n=1}^{N} (w_n - \bar{w})^2}},
\]

\[
\hat{\rho}(\tau, w) = \frac{\sum_{k=1}^{K} N(k)(\hat{\tau}_k - \bar{\tau})(\hat{w}_k - \bar{w})}{\sqrt{\sum_{k=1}^{K} N(k)(\hat{\tau}_k - \bar{\tau})^2} \sqrt{\sum_{k=1}^{K} N(k)(\hat{w}_k - \bar{w})^2}}.
\]
Example

(a) $N$.  
(b) $\tilde{K}$.  
(c) Relative $\sqrt{\Delta}$.  

(d) $\rho(\tau, \omega)$.  
(e) $\hat{\rho}(\tau, \omega)$.  

Key:

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\geq 20,000$</td>
<td>$\geq 0.05$</td>
<td>$\geq 0.05$</td>
<td>40</td>
<td>$\geq 4.5$</td>
</tr>
</tbody>
</table>
Conclusion

- Effectiveness and computational efficiency depend on inherent clustering of the data.


- Parameter settings still a bit ad hoc.

- Provides nonparametric, descriptive summary of the data, not inferential statistics.