Population Study of Session Traces

Discussion led by Steve Marron

8/4/00

Goals: Analyze "population of traces":

1. Summarize by low dim'al (10-20 d?) "feature vectors"
2. Analyze resulting population (e.g. clustering)
3. Use as basis to study change

Simple Summaries

a. Size Summaries:

Show UncSessionData\SessionData3p41d1s.ps

i. log10(Total Time)

ii. log10(Sign On Time)

iii. log10(Sign Off Time)

iv. log10(Total Packet Size)

v. log10(Total # of Packets)

vi. log10(# of Big Packets)

Recall: logs allow representing “ratios” and “proportions” as linear functions, e.g. % sign on time, Avg. Packet size

Simple Summaries

b. Shape Summaries:

Show UncSessionData\CombineSessionData1p51.pdf

For each linear piece:

i. log10(width)

- log10(time between knots)
- allows reconstruction

ii. log10(slope)

- log scale naturally includes ratios
- also allows reconstruction

“Shape” Summaries (cont.)

iii. Area, absolute Residuals

- Records “how far off”
  (trace is from linear fit)

- On “visual scale”

iv. Area, Residuals

- shows “direction” of deviation

- reflects “bias” component of error

- “difference” not “ratio”, so no log

Data Set:

Raw Data: 1,364 FTP Traces

Summarized Data: 829 “feature vec’s”

Rest gave summarization errors?

(seem to be too “small”,
need to check)

Data Analysis

Problem: how to “look” at the data?

Approach 1: “Parallel coordinates”

Show TraceChar1p2s1.ps

1. Size Variables could be Gaussian

2. Shape variables not Gaussian

3. Careful about differing orders of magnitude

Approach 2

Marginal distributions

Size variables:

Show TraceChar1p3s2.ps

- non-Gaussian distributions
- “clusters and “spikes”
- packet #’s both strongly skewed
- max sign on-off time $\approx 10^2$????

Marginal Dist'ns II

Shape Variables

Show TraceChar1p3s3.pdf

- all have many 0's
- except knot 0 widths
- other widths have "2 clusters"
- residual areas ~ skewed dist'ns
- should "separate out 0's"???

Approach 3

Visualize population mean

Idea: “Graphical view” of feature vector

Mean Size:

Show TraceChar1p11s2.ps

Mean Shape:

Show TraceChar1p11s3.ps

Combined Mean Size and Shape:

Show TraceChar1p11s1.ps

Approach 4

Visualize variability about mean

Principal Component Analysis:

Show ComplexPopn\CorneaRobust\SimplePCAeg.ps

(find directions of greatest variability)

Toy Example: family of curves

Show ComplexPopn\CurvDat\ParabsCurvDat.ps and ParabsUpDnCurvDat.ps

- separates out “dominant components of variability”
- “% explained” $\Rightarrow$ usefulness
- might find “clusters”

1st Principal Component

Show TraceChar1p12s1PC1.mpg

- dominated by “size”?
- big – small: # byte, # packets, % big packets, total time
- “big” ~ “slow start” shape
- “small” ~ long sign on-off, some shape???
- “really small” not a valid trace
- bytes / sec. rate not correlated
- explains 41% of variation

2nd Principal Component

Show TraceChar1p12s1PC2.mpg

- dominated by “sign on-off” time?
- Correlated with: size, time, rate…
- More traces at “larger end”???
- Largest part has “no traces”???
- Both ends ~ “Slow start” shape?
- Explains 14% of total variation (thus 55% total)

3rd Principal Component:

Show TraceChar1p12s1PC3.mpg

- dominated by total time?
- Nearly uncorrelated with size, and sign on-off time
- Shape: “later flat” (short time) vs. “earlier flat” (longer time)
- Explains 11% of total variation (thus 66% total)

4th Principal Component

Show TraceChar1p12s1PC4.mpg

- “big & fast” vs. “small & slow”
- “big & fast” ~ “slow start shape” and “long sign on”
- “small & slow” ~ “steppy shape” and “long sign off”
- strong correlation with bytes/sec rate, % big packets, bytes/packet
- Gaussian looking projections
- Explains 7% of total variation (thus 73% total)

PCA on Size variables only

Show TraceChar1p12s2PC1.mpg, TraceChar1p12s2PC2.mpg, TraceChar1p12s2PC3.mpg, TraceChar1p12s2PC4.mpg

- PC1's look very similar
- PC2 ~ earlier PC4 (since that PC4 more "size oriented???)
- PC3 similar to PC2, except opposite "sign off" relationship
- PC4 again similar, but now driven by sign on time
- Proj'n dist'ns more symmetric, with outliers
- Less total variability explained (only 43% total)

PCA on shape variables only:

Show TraceChar1p12s3PC1.mpg, TraceChar1p12s3PC2.mpg, TraceChar1p12s3PC3.mpg, TraceChar1p12s3PC4.mpg

PC1: “fast + step” vs. “slow start”

PC2: differing step locations

PC3 and PC4: more different steps

Conclusion: “step variation” is hard to summarize this way

To do next?

1. Simulate from PCA distribution?
2. Find problem with summaries?
3. Separate on knot numbers?
4. Clustering?
5. Look at “chosen direct’n” projections
6. Other types of traces?