CMU SCS

15-826: Multimedia and Data Mining

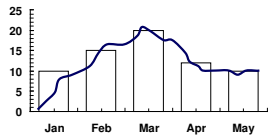
Information recovery

Christos Faloutsos

CMU SCS

1. INTRODUCTION

- 1.1 Problem:
  - given monthly sums of sales
  - find daily sales



CMU SCS

1.2 Motivation - Applications

- Qopt - selectivities
- data warehousing
- transaction recording systems (details: in tertiary store)
- statistical/scientific db
- data integration (partial info from many sources)

CMU SCS

OUTLINE

- Introduction
- Survey
- Proposed Method
- Experiments
- Extensions
- Conclusions

CMU SCS

2. SURVEY

- Qopt - selectivity estimation
  - uniformity assumption [Selinger et al, 79]
  - ... shown pessimistic [Christodoulakis 84]
  - histograms [Muralikrishna DeWitt 88], [Ioannidis et al 95, 96]
  - query feedbacks [Chen Roussopoulos 94]
  - spatial databases [Theodoridis Sellis 96]
- Statistical databases [Malvestuto 93], [Ng Ravishankar 93]

CMU SCS

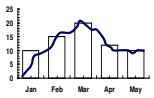
OUTLINE

- Introduction
- Survey
- Proposed Method
- Experiments
- Extensions
- Conclusions

CMU SCS

3. PROPOSED METHOD

- 3.1 Problem formulation



Estimate the vector

$\vec{x} = [x_i], i = 1, ..., N$

Under the constraints

$C_k(\vec{x}) = 0 \quad (k = 1, \ldots, n)$

CMU SCS

observations:

- Under-specified -> infinite # of solutions
- constraints: can be arbitrary
- trivial generalization for *n*-dim:

count11	count12	count13	count14
count21	count22	count23	count24
count31	count32	count33	count34
count41	count42	count43	count44

CMU SCS

special case

- 1-d - no overlap - no gaps (eg., monthly sums)
- Obvious solution: uniformity
- BUT
  - ignores continuity
  - overlapping sums?
  - impossible for missing values

CMU SCS

3.2 Solution

(Linear) Regularization

- Intuition: choose some solution, e.g., the smoothest:
- minimize a smoothness function *F()*
- under the given constraints

$$F(\vec{x}) = \sum_{i=1}^{N-1} (x_i - x_{i+1})^2$$

$$C_k(\vec{x}) = S_k - \sum_{i \in B_k} x_i = 0$$

CMU SCS

Observations

- problem is now well defined!
- any functional *F()* can be used
- Lemma: Entropy for *F()* -> uniformity assumption

CMU SCS

Lin-Reg

- =Phillips-Tomey = Tikhonov-Miller = ...
- Lagrange multipliers -> matrix inversion
- in *linear* time!
- Theory of ``Inverse Problems''
  - image restoration
  - tomography etc

CMU SCS

Theorem:

- Given partial sums of a *band-limited* signal with small enough batch-size: Perfect reconstruction!
- In practice, Lin-Reg => almost perfect reconstruction for ``smooth'' signals

CMU SCS

OUTLINE

- Introduction
- Survey
- Proposed Method
- Experiments
- Extensions
- Conclusions

CMU SCS

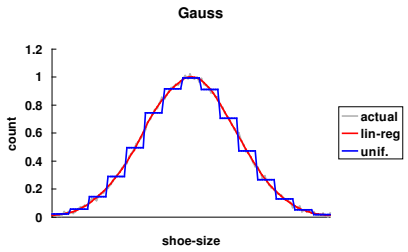
4. EXPERIMENTS

- synthetic datasets (GAUSS, SINE)
- real ones (IBM, LYNX)

Experiments, to see
- accuracy of reconstruction
- effect of batch-size

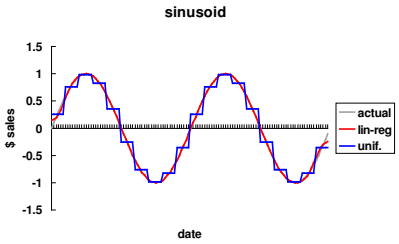
CMU SCS

4.1. Accuracy



CMU SCS

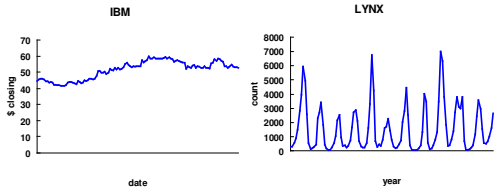
Synthetic - sinusoid



almost band-limited ->
almost perfect reconstruction!

CMU SCS

Real datasets

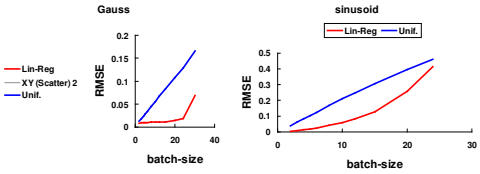


- IBM: ``fractal'' - fd=1.5
- LYNX: ~ periodic (9-10 years)

CMU SCS

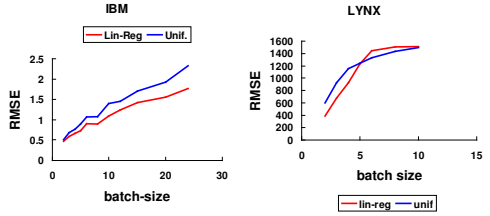
4.2 Effect of batch-size

Performance for synthetic datasets:
up to 89% savings



CMU SCS

Real datasets



up to 21% savings up to 35% savings

CMU SCS

5. EXTENSIONS

- 5.1 Merging histograms

age

40-50	3,000	??	??	??
30-40	5,000	??	??	??
20-30	2,000	??	??	??
	10,000	9,000	900	100
		0-10K	10K-50K	>50K

salary

CMU SCS

Solution:

- Lemma: Max Entropy -> independence assumption
- Lin-Reg should do better on smooth distributions

CMU SCS

5.2. Data Integration

- given (some) monthly sums
- (some) weekly sums and
- (some) daily amounts
- estimate all daily sales

CMU SCS

Specifically:

- minimize

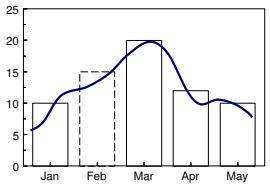
$$F(\vec{x}) = \sum_{i=1}^{N-1} (x_i - x_{i+1})^2$$

$$C_{monthly,i}(\vec{x}) = 0$$

$$C_{weekly,j}(\vec{x}) = 0$$

CMU SCS

5.3 Missing values



- Lin-Reg works, while uniformity/Max Entropy do **not**!

CMU SCS

6. CONCLUSIONS

- Lin-Reg -> good reconstruction
- up to 89% better RMSE
- Lemmas: Max Entropy -> Uniformity and Independence assumptions
- Theorem that guarantees perfect reconstruction
- Theory of Inverse Problems: seems valuable for DB.

CMU SCS

Reference

- Christos Faloutsos, H.V. Jagadish and Nikolaos D. Sidiropoulos *Recovering Information from Summary Data* VLDB, Athens, Greece, Aug. 1997.