Wavelet Synopses with Error Guarantees

Minos Garofalakis
Intel Research Berkeley

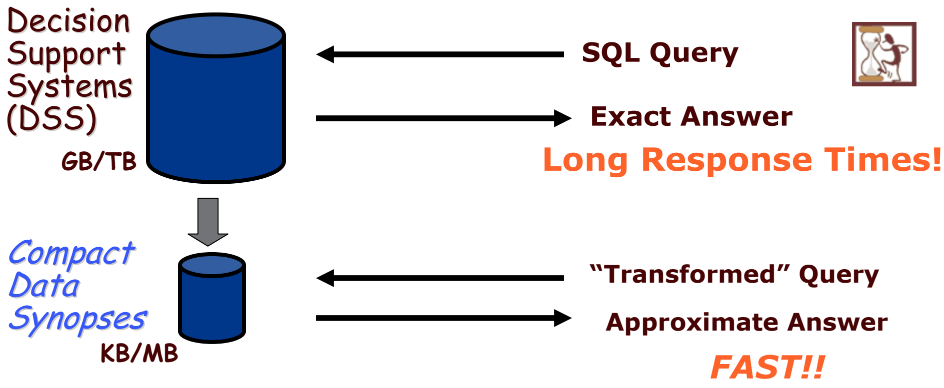
minos.garofalakis@intel.com
http://www2.berkeley.intel-research.net/~minos/

Joint work with ***Phil Gibbons*** *[ACM SIGMOD'02, ACM TODS'04]*
and ***Amit Kumar*** *[ACM PODS'04, ACM TODS'05]*

Outline

- Preliminaries & Motivation
  - Approximate query processing
  - Haar wavelet decomposition, conventional wavelet synopses
  - The problem
- A First solution: *Probabilistic Wavelet Synopses*
  - The general approach: Randomized Selection and Rounding
  - Optimization Algorithms for Tuning our Synopses
- More Direct Approach: Effective *Deterministic* Solution
- Extensions to Multi-dimensional Haar Wavelets
- Experimental Study
  - Results with synthetic & real-life data sets
- Conclusions

Approximate Query Processing



- Exact answers **NOT** always required
  - DSS applications usually *exploratory:* early feedback to help identify "interesting" regions
  - *Aggregate queries:* precision to "last decimal" not needed
    - e.g., "What percentage of the US sales are in NJ?"
- Construct effective *data synopses* ??

Haar Wavelet Decomposition

- Wavelets: mathematical tool for hierarchical decomposition of functions/signals
- Haar wavelets: simplest wavelet basis, easy to understand and implement
  - *Recursive pairwise averaging and differencing* at different resolutions

Resolution	Averages	Detail Coefficients
3	D = [2, 2, 0, 2, 3, 5, 4, 4]	----
2	[2, 1, 4, 4]	[0, -1, -1, 0]
1	[1.5, 4]	[0.5, 0]
0	[2.75]	[-1.25]

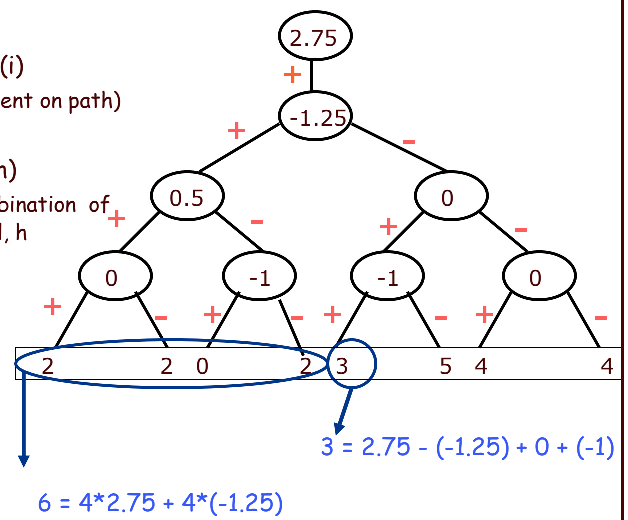
Haar wavelet decomposition: [2.75, -1.25, 0.5, 0, 0, -1, -1, 0]

- Construction extends naturally to multiple dimensions

Haar Wavelet Coefficients

- Hierarchical decomposition structure (a.k.a. ***Error Tree***)
  - Conceptual tool to "visualize" *coefficient supports* & *data reconstruction*
- Reconstruct data values d(i)
  - d(i) = $\sum$(+/-1) * (coefficient on path)
- Range sum calculation d(l:h)
  - d(l:h) = simple linear combination of coefficients on paths to l, h
- Only O(logN) terms

Original data



Wavelet Data Synopses

- Compute Haar wavelet decomposition of D
- *Coefficient thresholding* : only B<<|D| coefficients can be kept
  - B is determined by the available synopsis space
- Approximate query engine can do all its processing over such compact coefficient synopses (joins, aggregates, selections, etc.)
  - Matias, Vitter, Wang [SIGMOD'98]; Vitter, Wang [SIGMOD'99]; Chakrabarti, Garofalakis, Rastogi, Shim [VLDB'00]
- Conventional thresholding: Take B largest coefficients in *absolute normalized value*
  - Normalized Haar basis: divide coefficients at resolution j by $\sqrt{2^j}$
  - All other coefficients are ignored (assumed to be zero)
  - *Provably optimal* in terms of the overall Sum-Squared (L2) Error
- *Unfortunately*, no meaningful approximation-quality guarantees for
  - Individual reconstructed data values or range-sum query results

Problems with Conventional Synopses

intel.

- An example data vector and wavelet synopsis (|D|=16, B=8 largest coefficients retained)

Over 2,000% relative error!

Always accurate!

Original Data Values	127	71	87	31	59	3	43	99	100	42	0	58	30	88	72	130
Wavelet Answers	65	65	65	65	65	65	65	65	100	42	0	58	30	88	72	130

Estimate = 195, actual values: d(0:2)=285, d(3:5)=93!

- Large variation in answer quality
  - Within the same data set, when synopsis is *large,* when data values are about the same, when actual answers are about the same
  - Heavily-biased approximate answers!
- Root causes
  - Thresholding for aggregate L2 error metric
  - Independent, greedy thresholding (⇨ large regions without any coefficient!)
  - Heavy bias from dropping coefficients without compensating for loss

Approach: Optimize for Maximum-Error Metrics

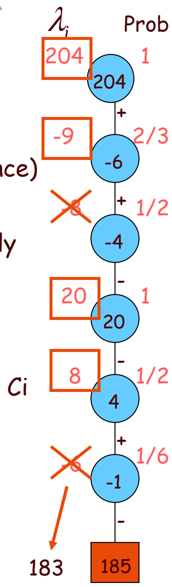
intel.

- Key metric for effective approximate answers: *Relative error with sanity bound* $\frac{|\hat{d}_i - d_i|}{\max\{|d_i|, s\}}$
  - Sanity bound "s" to avoid domination by small data values
- To provide tight error guarantees for *all* reconstructed data values

$$\text{Minimize} \quad \max_i \{\frac{|\hat{d}_i - d_i|}{\max\{|d_i|, s\}}\}$$

  - Minimize *maximum relative error* in the data reconstruction
- Another option: Minimize *maximum absolute error* $\max_i\{|\hat{d}_i - d_i|\}$
- Algorithms can be extended to general *"distributive" metrics* (e.g., average relative error)

A Solution: *Probabilistic Wavelet Synopses*



- Novel, *probabilistic thresholding scheme* for Haar coefficients
  - Ideas based on Randomized Rounding
- In a nutshell
  - Assign coefficient probability of retention (based on importance)
  - Flip biased coins to select the synopsis coefficients
  - Deterministically retain most important coefficients, randomly rounding others either up to a larger value or down to zero
  - ***Key:*** Each coefficient is *correct on expectation*
- Basic technique
  - For each non-zero Haar coefficient ci, define random variable Ci

$$C_i = \begin{cases} \lambda_i & \text{with probability } \frac{c_i}{\lambda_i} \in (0,1] \\ 0 & \text{with probability } 1 - \frac{c_i}{\lambda_i} \end{cases}$$

  - Round each ci independently to λ_i or zero by flipping a coin with success probability $\frac{c_i}{\lambda_i}$ (zeros are *discarded*)

Probabilistic Wavelet Synopses *(cont.)*

- Each Ci is correct on expectation, i.e., E[Ci] = ci
  - Our synopsis guarantees ***unbiased estimators*** for data values and range sums (by Linearity of Expectation)
- Holds for any λ_i's , BUT choice of λ_i's is crucial to quality of approximation and synopsis size
  - Variance of Ci: Var[Ci] = $(\lambda_i - c_i) \cdot c_i$
  - By independent rounding, Variance[*reconstructed di*] = $\sum_{path(di)} (\lambda_i - c_i) \cdot c_i$
    - *Better approximation/error guarantees for smaller* λ_i *(closer to ci)*
  - Expected size of the final synopsis E[size] = $\sum \frac{c_i}{\lambda_i}$
    - *Smaller synopsis size for larger* λ_i
- Novel optimization problems for "tuning" our synopses
  - Choose λ_i 's to ensure tight approximation guarantees (i.e., small reconstruction variance), while E[synopsis size] $\leq$ B
  - Alternative probabilistic scheme
    - Retain *exact* coefficient with probabilities chosen to *minimize bias*

MinRelVar: Minimizing Max. Relative Error

intel.

- *Relative error metric* $\dfrac{|\hat{d}_i - d_i|}{\max\{|d_i|, s\}}$

- Since estimate $\hat{d}_i$ is a random variable, we want to ensure a tight bound for our relative error metric with high probability
  - By Chebyshev's inequality

$$\Pr\left[\frac{|\hat{d}_i - d_i|}{\max\{|d_i|, s\}} < \alpha \cdot \frac{\sqrt{Var[\hat{d}_i]}}{\max\{|d_i|, s\}}\right] > 1 - \frac{1}{\alpha^2}$$

Normalized Standard Error (NSE) of reconstructed value

- To provide tight error guarantees for all data values
  - Minimize the *Maximum NSE* among all reconstructed values $\hat{d}_i$

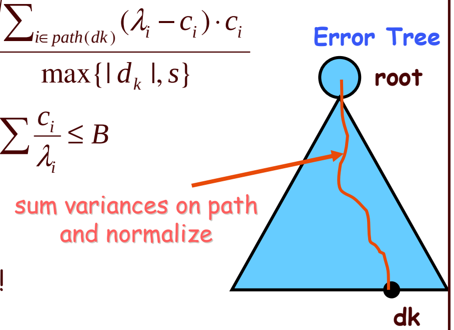
Minimizing Maximum Relative Error *(cont.)*

intel.

- *Problem:* Find rounding values λ_i to minimize the maximum NSE

$$\max_{path(dk) \in PATHS} \frac{\sqrt{\sum_{i \in path(dk)} (\lambda_i - c_i) \cdot c_i}}{\max\{|d_k|, s\}}$$

subject to $c_i / \lambda_i \in (0,1]$ and $\sum \dfrac{c_i}{\lambda_i} \le B$



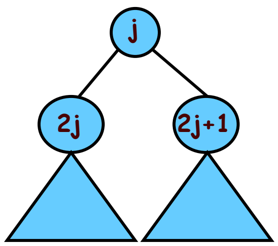
- Hard non-linear optimization problem!

- Propose solution based on a *Dynamic-Programming (DP) formulation*
  - Key technical ideas
    - Exploit the hierarchical structure of the problem (Haar error tree)
    - Exploit properties of the optimal solution
    - Quantizing the solution space

Minimizing Maximum Relative Error *(cont.)*

- Let $y_i = c_i / \lambda_i$ = the probability of retaining ci
  - yi = "fractional space" allotted to coefficient ci ($\sum yi \leq$ B)
- M[j,b] = optimal value of the *(squared)* maximum NSE for the subtree rooted at coefficient cj for a space allotment of b

$$M[j,b] = \min_{y \in (0, \min\{1,b\}], b_L \in [0, b-y]} \max\{ \frac{Var[j,y]}{Norm_{2j}} + M[2j, b_L], \frac{Var[j,y]}{Norm_{2j+1}} + M[2j+1, b-y-b_L] \}$$



- Normalization factors "Norm" depend only on the minimum data value in each subtree
  - *See paper for full details...*

- Quantize choices for y to {1/q, 2/q, ..., 1}
  - q = input integer parameter, "knob" for run-time vs. solution accuracy
  - $O(Nq^2 B \log(qB))$ time, $O(qB \log N)$ memory

But, still…

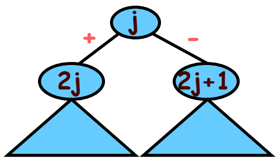
- Potential concerns for probabilistic wavelet synopses
  - Pitfalls of randomized techniques
    - Possibility of a "bad" sequence of coin flips resulting in a poor synopsis
  - Dependence on a quantization parameter/knob q
    - Effect on optimality of final solution is not entirely clear

- *"Indirect" Solution:* try to *probabilistically control* maximum relative error through appropriate probabilistic metrics
    - E.g., minimizing maximum NSE

- ***Natural Question***
  - Can we design an efficient *deterministic* thresholding scheme for minimizing non-L2 error metrics, such as maximum relative error?
    - Completely avoid pitfalls of randomization
    - *Guarantee* error-optimal synopsis for a given space budget B

Do our Earlier Ideas Apply?

- Unfortunately, probabilistic DP formulations rely on
  - Ability to assign *fractional storage* $y_i \in (0,1]$ to each coefficient ci
  - Optimization metrics (maximum NSE) with *monotonic/additive structure* over the error tree

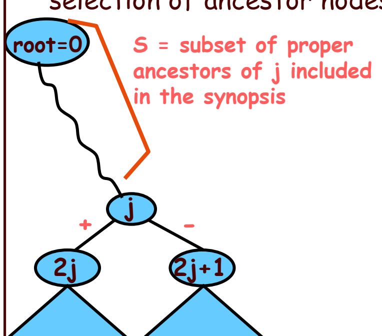


- M[j,b] = optimal NSE for subtree T(j) with space b
- ***Principle of Optimality***
  - Can compute M[j,*] from M[2j,*] and M[2j+1,*]

- When *directly* optimizing for maximum relative (or, absolute) error with storage $\in \{0,1\}$, principle of optimality fails!
  - Assume that M[j,b] = optimal value for $\max_{T(j)}\{\frac{|\hat{d}_i - d_i|}{\max\{|d_i|, s\}}\}$ with at most b coefficients selected in T(j)
  - Optimal solution at j may *not* comprise optimal solutions for its children
    - Remember that $\hat{d} = \sum$ (+/-)* SelectedCoefficient, where coefficient values can be positive or negative
- *BUT, it can be done!!*

Our Approach: Deterministic Wavelet Thresholding for Maximum Error

- ***Key Idea:*** Dynamic-Programming formulation that *conditions the optimal solution on the error that "enters" the subtree* (through the selection of ancestor nodes)



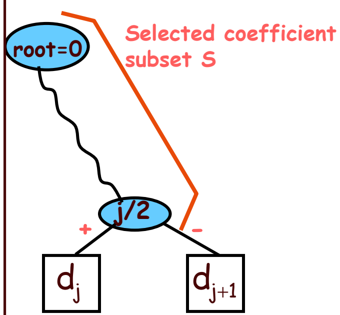
- Our DP table:

  M[j, b, S] = optimal maximum relative (or, absolute) error in T(j) with space budget of b coefficients (chosen in T(j)), *assuming subset S of j's proper ancestors have already been selected for the synopsis*
  - Clearly, $|S| \leq \min\{B-b, \log N+1\}$
  - Want to compute $M[0, B, \phi]$

- ***Basic Observation:*** Depth of the error tree is only logN+1 ⇨ we can explore and tabulate all S-subsets for a given node at a space/time cost of only O(N) !

Base Case for DP Recurrence: Leaf (Data) Nodes

- Base case in the bottom-up DP computation: Leaf (i.e., data) node d_j
  - Assume for simplicity that data values are numbered N, …, 2N-1



- M[j, b, S] is not defined for b>0
  - Never allocate space to leaves
- For b=0

$$M[j,0,S] = \frac{|d_j - \sum_{c \in S} sign(c,d_j) \cdot c|}{\max\{|d_j|, s\}}$$

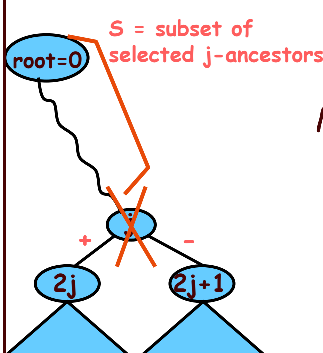
for each coefficient subset $S \subseteq path(d_j)$ with $|S| \leq \min\{B, \log N+1\}$
  - *Similarly for absolute error*

- Again, time/space complexity per leaf node is only O(N)

DP Recurrence: Internal (Coefficient) Nodes

- Two basic cases when examining node/coefficient j for inclusion in the synopsis: (1) Drop j; (2) Keep j

Case (1): Drop Coefficient j



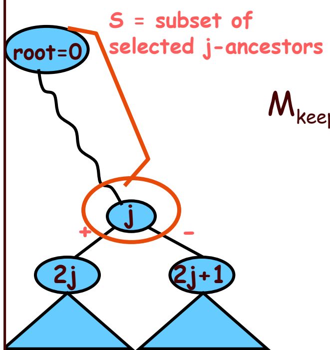
- In this case, the minimum possible maximum relative error in T(j) is

$$M_{drop}[j,b,S] = \min_{0 \leq b' \leq b} \max\{ M[2j,b',S], M[2j+1,b-b',S] \}$$

  - Optimally distribute space b between j's two child subtrees
- Note that the RHS of the recurrence is well-defined
  - Ancestors of j are obviously ancestors of 2j and 2j+1

DP Recurrence: Internal (Coefficient) Nodes *(cont.)*

Case (2): Keep Coefficient j



- In this case, the minimum possible maximum relative error in T(j) is

$$M_{keep}[j,b,S] = \min_{0 \le b' \le b-1} \max\{ M[2j,b',S \cup \{c_j\}], M[2j+1,b-b'-1,S \cup \{c_j\}] \}$$

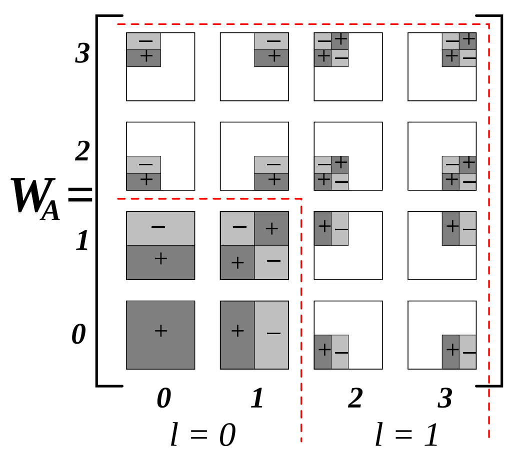
  - Take 1 unit of space for coefficient j, and optimally distribute remaining space
  - Selected subsets in RHS change, since we choose to retain j
- *Again, the recurrence RHS is well-defined*

- Finally, define $M[j,b,S] = \min\{ M_{drop}[j,b,S], M_{keep}[j,b,S] \}$
- Overall complexity: $O(N^2)$ time, $O(N \min\{B, \log N\})$ space

Multi-dimensional Haar Wavelets

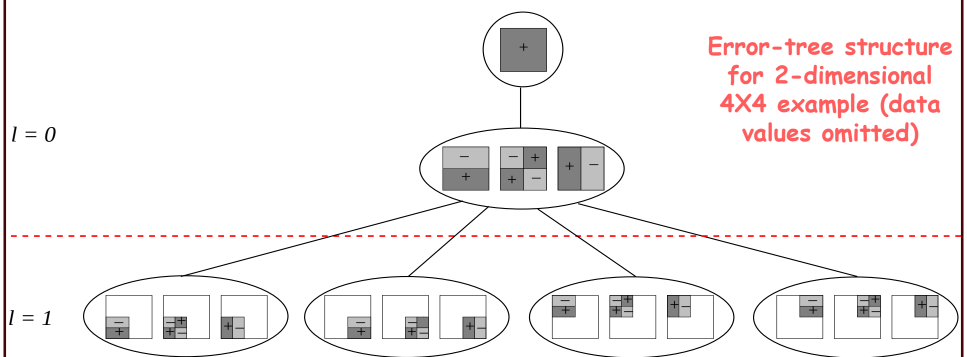
- Haar decomposition in d dimensions = d-dimensional array of wavelet coefficients
  - Coefficient *support region* = d-dimensional rectangle of cells in the original data array
  - *Sign* of coefficient's contribution can vary along the quadrants of its support

Support regions & signs for the 16 nonstandard 2-dimensional Haar coefficients of a 4X4 data array A

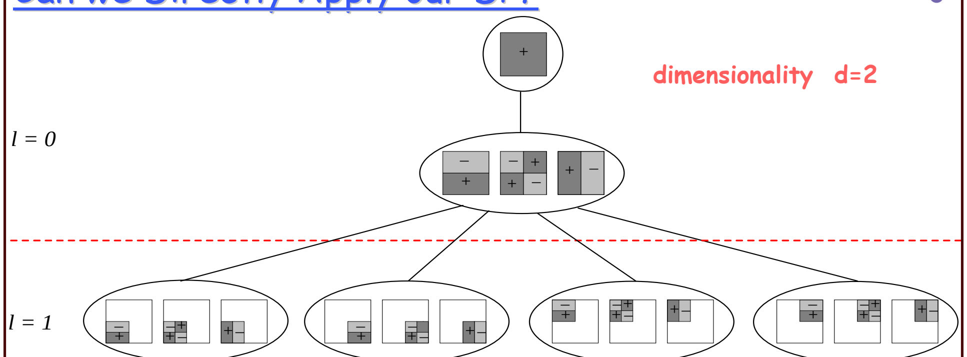


Multi-dimensional Haar Error Trees

- Conceptual tool for data reconstruction – more complex structure than in the 1-dimensional case
  - Internal node = *Set* of (up to) $2^d - 1$ coefficients (identical support regions, different quadrant signs)
  - Each internal node can have (up to) 2^d children (corresponding to the quadrants of the node's support)
- Maintains *linearity* of reconstruction for data values/range sums



Can we Directly Apply our DP?



- ***Problem:*** Even though depth is still O(logN), each node now comprises up to $2^d - 1$ coefficients, *all of which* contribute to every child
  - Data-value reconstruction involves up to $O((2^d - 1)\log N)$ coefficients
  - Number of potential ancestor subsets (S) explodes with dimensionality

    Up to $O(N^{2^d - 1})$ *ancestor subsets per node!*
  - Space/time requirements of our DP formulation quickly become infeasible (even for d=3,4)
- ***Our Solution:*** ε -approximation schemes for multi-d thresholding

Approximate Maximum-Error Thresholding in Multiple Dimensions

intel.

- *Time/space efficient approximation schemes* for deterministic multi-dimensional wavelet thresholding for maximum error metrics
- Propose two different approximation schemes
  - Both are based on *approximate dynamic programs*
  - Explore a much smaller number of options while offering ε-approximation gurantees for the final solution
- ***Scheme #1:*** *Sparse DP formulation* that rounds off possible values for subtree-entering errors to powers of $(1+\varepsilon)$
  - $O(\frac{\log R}{\varepsilon} N B \log N \log B)$ time
  - *Additive* ε-error guarantees for maximum relative/absolute error
- ***Scheme #2:*** Use *scaling & rounding of coefficient values* to convert a pseudo-polynomial solution to an efficient approximation scheme
  - $O(\frac{\log R}{\varepsilon} N B \log^2 N \log B)$ time
  - $(1+\varepsilon)$-approximation algorithm for maximum absolute error

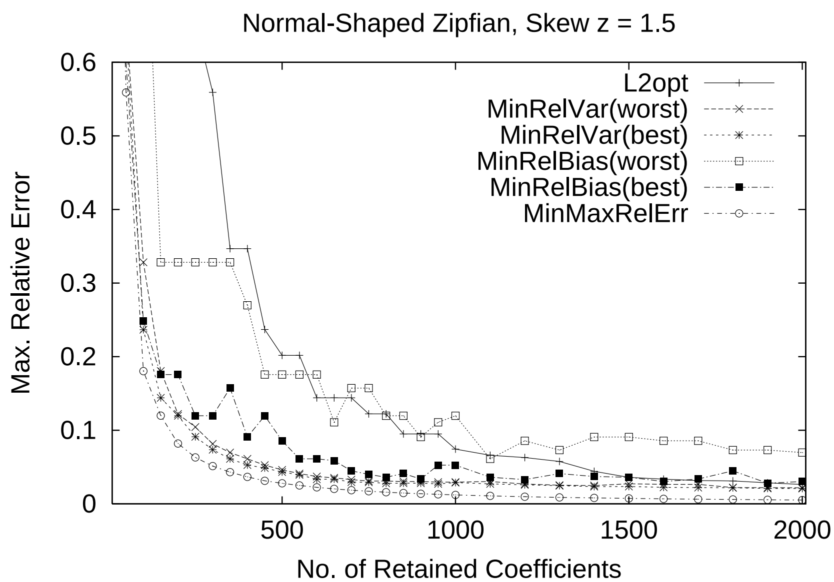
Experimental Study

intel.

- Deterministic vs. Probabilistic (vs. Conventional L2)
- Synthetic and real-life data sets
  - Zipfian data distributions
    - Various permutations, skew z = 0.3 - 2.0
  - Weather, Corel Images (UCI), …
- Relative error metrics
  - Sanity bound = 10-percentile value in data
  - Maximum and average relative error in approximation
    - Deterministic optimization algorithms extend to any *"distributive" error metric*

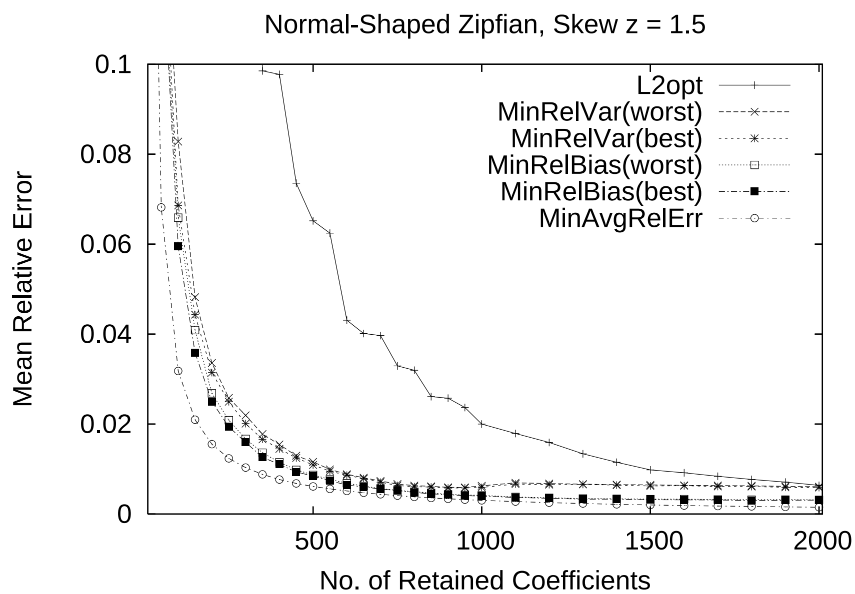
Synthetic Data – Max. Rel. Error

intel.

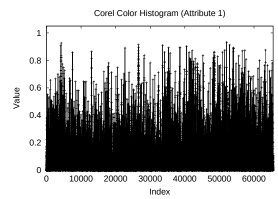


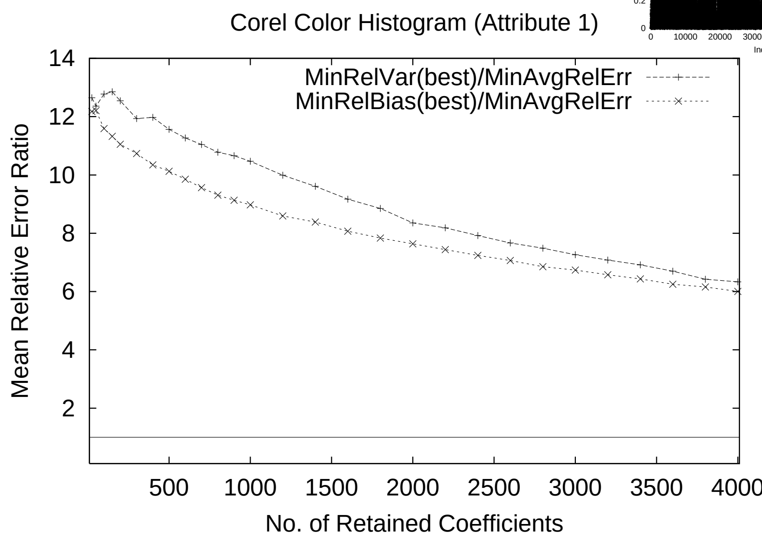
Synthetic Data – Avg. Rel. Error

intel.

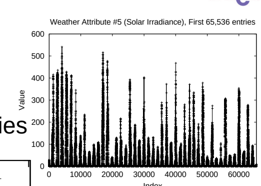


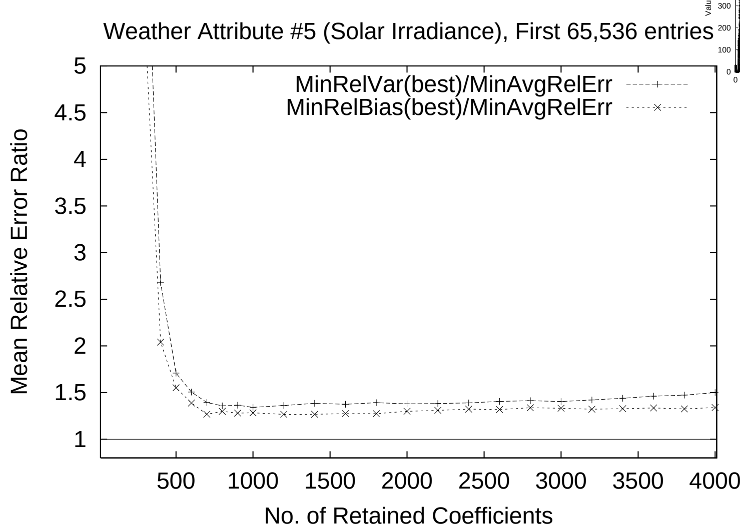
Real Data -- Corel





Real Data -- Weather

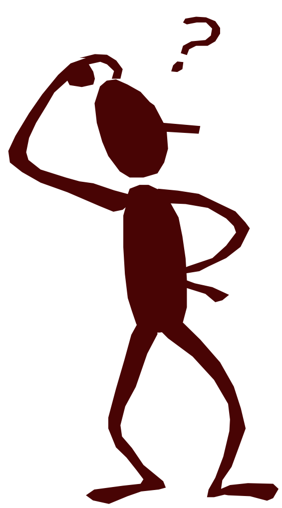




Conclusions & Future Work

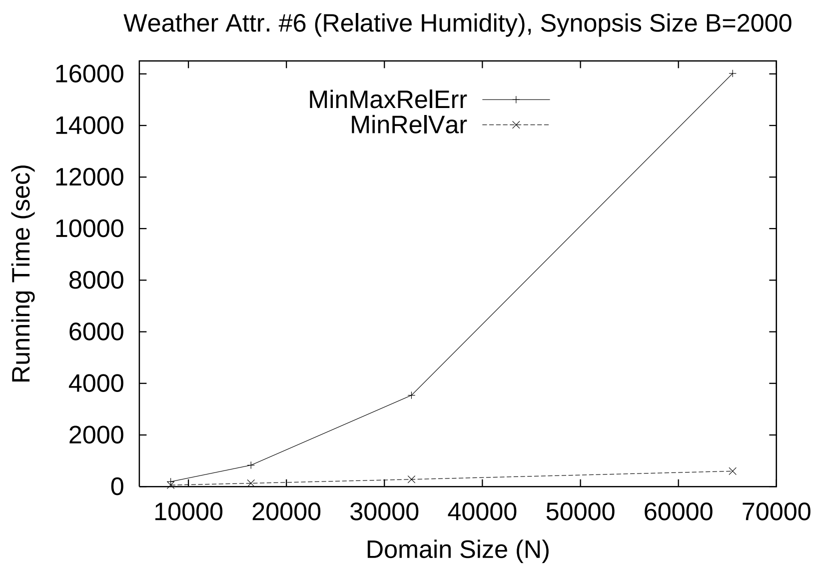
- Introduced the *first* efficient schemes for wavelet thresholding for maximum-error metrics
  - *Probabilistic* and *Deterministic*
  - Based on novel DP formulations
  - Deterministic avoids pitfalls of probabilistic solutions *and* extends naturally to *general error metrics*
- Extensions to multi-dimensional Haar wavelets
  - Complexity of exact solution becomes prohibitive
  - Efficient polynomial-time approximation schemes based on approximate DPs
- ***Future Research Directions***
  - Streaming computation/incremental maintenance of max-error wavelet synopses : *Heuristic solution* proposed recently (VLDB'05)
  - Extend methodology and max-error guarantees for more complex queries (joins??)
  - Suitability of Haar wavelets, e.g., for relative error? Other bases??

Thank you!

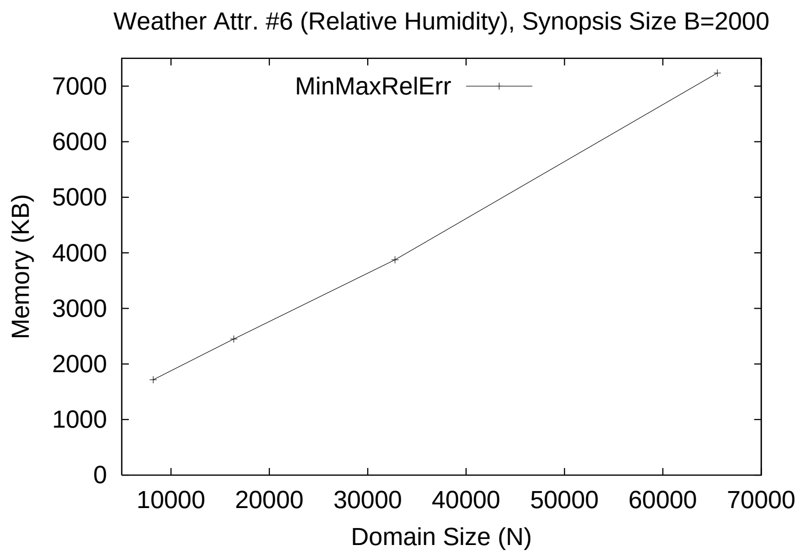


minos.garofalakis@intel.com
http://www2.berkeley.intel-research.net/~minos/

Runtimes



Memory Requirements



MinRelBias: Minimizing Normalized Bias

- *Scheme:* Retain the exact coefficient ci with probability yi and discard with probability (1-yi) -- no randomized rounding
  - Our Ci random variables are no longer unbiased estimators for ci
    - Bias[Ci] = | E[Ci] - ci | = |ci|*(1-yi)
- Choose yi's to minimize an upper bound on the *normalized reconstruction bias* for each data value; that is, minimize

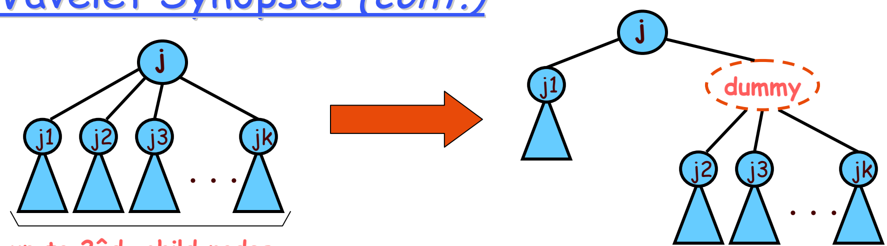
$$\max_{path(dk) \in PATHS} \frac{\sum_{i \in path(dk)} |c_i| \cdot (1 - y_i)}{\max\{|d_k|, s\}} \quad \text{subject to} \quad y_i \in (0,1] \quad \text{and} \quad \sum y_i \le B$$

- Same *dynamic-programming solution* as MinRelVar works!
- Avoids pitfalls of conventional thresholding due to
  - Randomized, non-greedy selection
  - Choice of optimization metric (minimize maximum resulting bias)

Multi-dimensional Probabilistic Wavelet Synopses

- *A First Issue:* Data density can increase dramatically due to recursive pairwise averaging/differencing (during decomposition)
  - Previous approaches suffer from *additional bias* due to ad-hoc construction-time thresholding
- *Our Solution:* "Adaptively threshold" coefficients probabilistically during decomposition <u>without introducing reconstruction bias</u>
- Once decomposition is complete, basic ideas/principles of probabilistic thresholding carry over directly to the d-dimensional case
  - *Linear* data/range-sum reconstruction
  - Hierarchical error-tree structure for coefficients
- Still, our algorithms need to deal with the added complexity of the d-dimensional error-tree…

Multi-dimensional Probabilistic Wavelet Synopses *(cont.)*



up to 2^d child nodes

- Computing M[j, B] = optimal max. NSE value at node j for space B, involves examining all possible allotments to j's children
- Naïve/brute-force solution would increase complexity by

$$O((qB)^{2^d-1})$$

- ***Idea:*** Generalize optimal DP formulation to effectively "order" the search
- M[<nodeList>, B] = optimal max. NSE for all subtrees with roots in <nodeList> and total space budget B
- M[<j>, B] only examines possible allotments between <j1> and <j2,…,jk>
- Only increases space/time complexity by 2^d (typically, d $\leq$ 4-5 dimensions)

- *Sets of coefficients* per error-tree node can also be effectively handled
- Details in the paper…

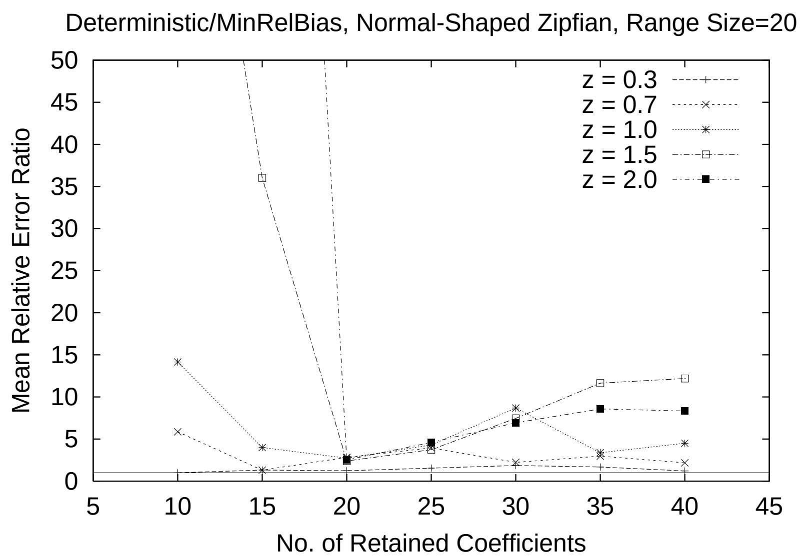
MinL2: Minimizing Expected L2 Error

- *Goal:* Compute rounding values λ_i to minimize *expected value* of overall L2 error
  - Expectation since L2 error is now a random variable
- *Problem:* Find λ_i that minimize $\sum \frac{(\lambda_i - c_i)\cdot c_i}{2^{level(ci)}}$, subject to the constraints

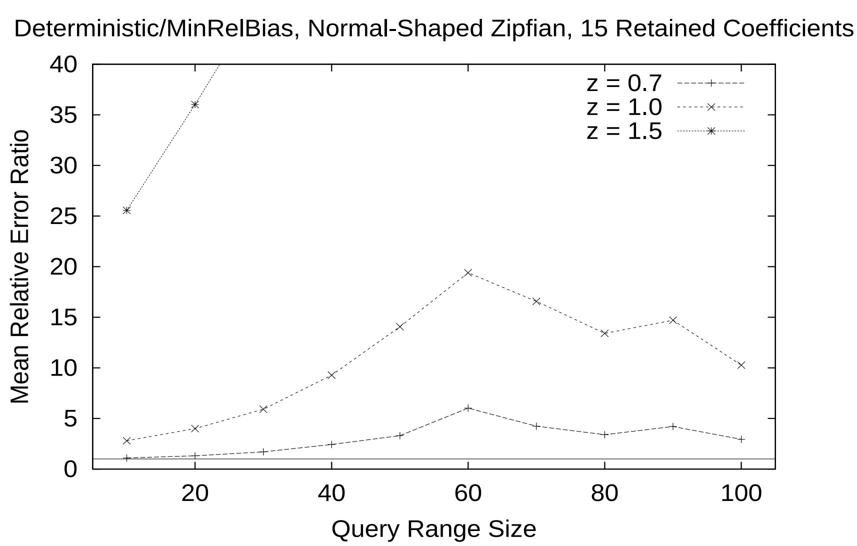
$$c_i/\lambda_i \in (0,1] \quad \text{and} \quad \sum \frac{c_i}{\lambda_i} \leq B$$

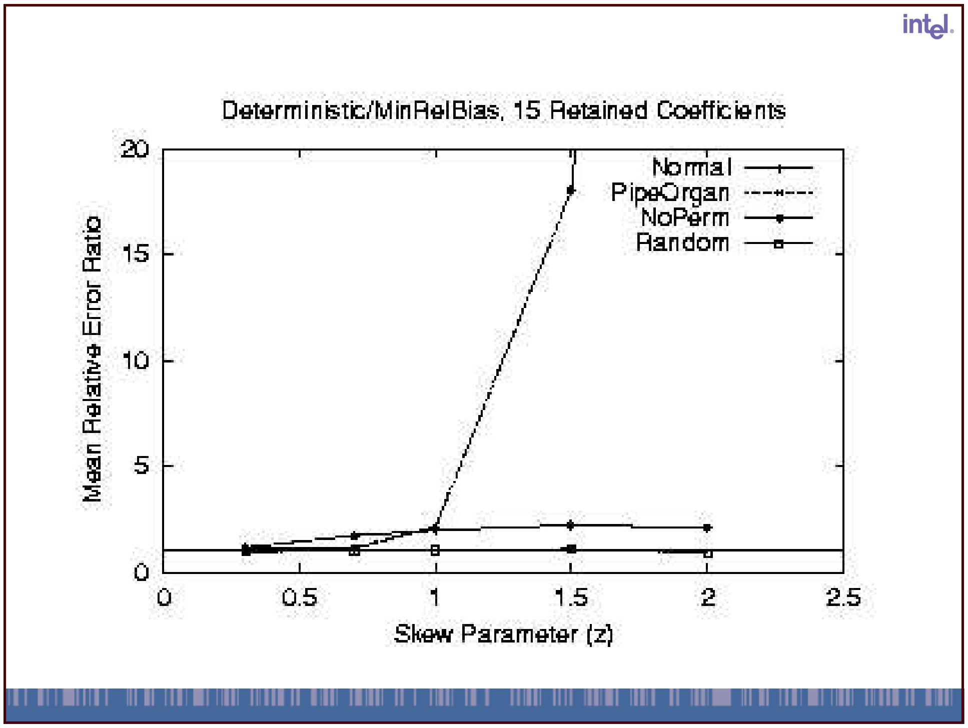
- Can be solved optimally: Simple iterative algorithm, O(N logN) time
- BUT, again, overall L2 error cannot offer error guarantees for individual approximate answers (data/range-sum values)

Range-SUM Queries: Relative Error Ratio vs. Space

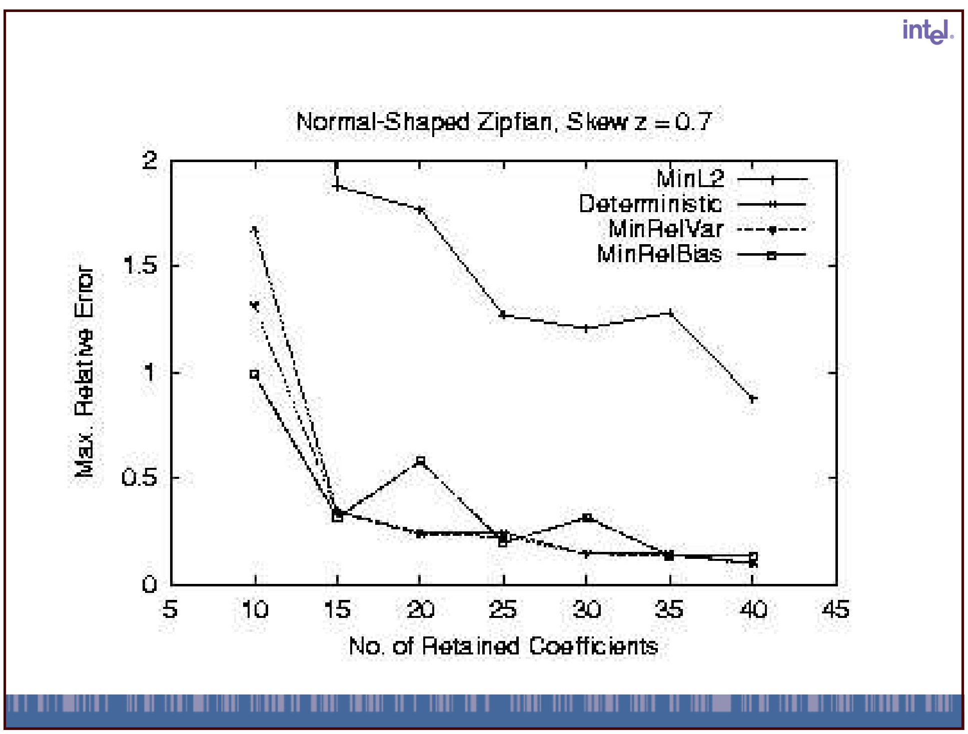


Range-SUM Queries: Relative Error Ratio vs. Range Size

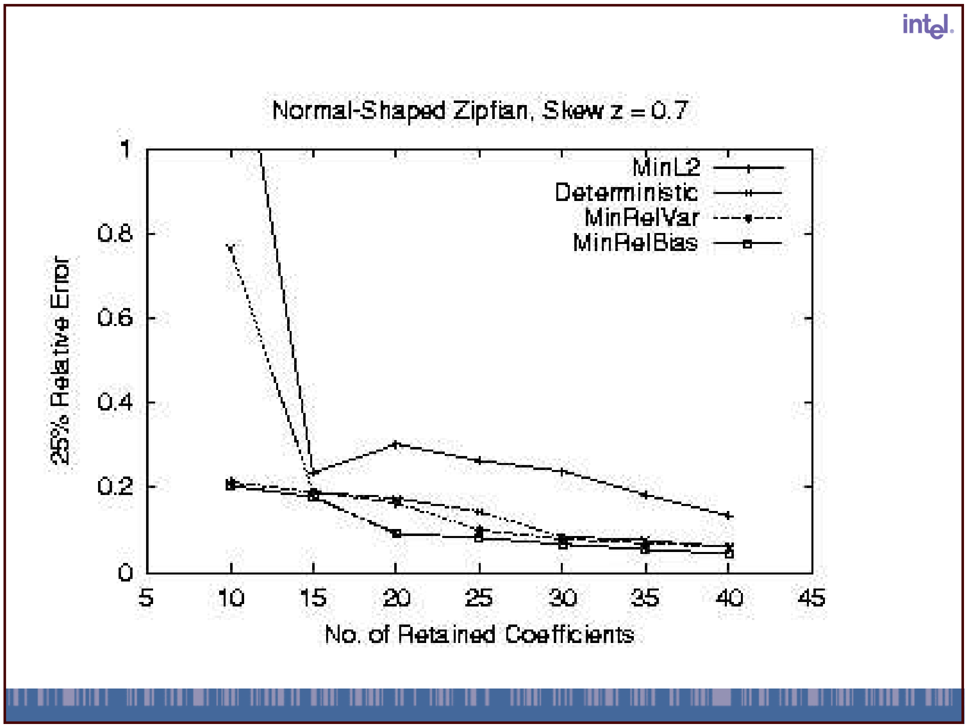




intel.
Deterministic/MinRelBias, 15 Retained Coefficients
Mean Relative Error Ratio
Skew Parameter (z)
Normal
PipeOrgan
NoPerm
Random
0
5
10
15
20
0
0.5
1
1.5
2
2.5



intel.
Normal-Shaped Zipfian, Skew z = 0.7
Max. Relative Error
No. of Retained Coefficients
MinL2
Deterministic
MinRelVar
MinRelBias
0
0.5
1
1.5
2
5
10
15
20
25
30
35
40
45



Normal-Shaped Zipfian, Skew z = 0.7
MinL2
Deterministic
MinRelVar
MinRelBias
25% Relative Error
No. of Retained Coefficients