On Wavelet Compression and Cardinality Estimation of Enterprise Data

Lakshminarayan Choudur, Umeshwar Dayal, Chetan Gupta, Ram Swaminathan

HP Laboratories
HPL-2010-132

Keyword(s):

compression, wavelets, thresholding, energy

Abstract:

Storing and analyzing large volume of structured or unstructured data at the scale of petabytes in applications such as business intelligence of an enterprise, is a daunting task. It is therefore desirable to store data in a compressed form. Compression often is performed using transform methods that permit capture of details in the data while at the same time representing it efficiently; wavelet transform is one such compression technique. Viewing the data as a signal, wavelet transform involves computing coefficients and selecting a small subset judiciously to approximate the signal. Because we pick a subset of the coefficients and not all of them to synthesize the data, wavelet based compression is inherently lossy. Compression in the wavelet domain is traditionally done using hard and soft thresholding techniques and variants thereof. Based on the notion of "total energy" of a signal, in this paper, we introduce two new thresholding methods called, level-independent energy and level-dependent energy thresholding. In particular, our energybased thresholding techniques exploit the relationship between total energy of the signal and wavelet coefficients to obtain compression to meet pre-specified error tolerances. In addition, level-dependent energy thresholding aims at determining wavelet coefficients that carry most "information" at various resolutions of the wavelet decomposition of the data distribution. Detailed studies and experiments over noise contaminated synthetic and TPCH benchmark data sets indicate that energy based thresholding methods yield approximately 100:1 compression with high reconstruction accuracy. As an application of our compression techniques, we show that energy-based thresholding methods improve accuracy of cardinality estimation in databases substantially over the popular methods based on equi-height and max-diff histograms as well as over hard, soft and probabilistic thresholding techniques from the statistics and database literature.

External Posting Date: October 6, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: October 6, 2010 [Fulltext]

On Wavelet Compression and Cardinality Estimation of Enterprise Data

Lakshminarayan Choudur
HP Labs
Austin, TX 78728
lakshminarayan.choudur@hp.com

Umeshwar Dayal
HP Labs
Palo Alto, CA 94304
umesh.dayal@hp.com

Chetan Gupta
HP Labs
Austin, TX 78728
chetan.gupta@hp.com

Ram Swaminathan
HP Labs
Palo Alto, CA 94304
ram.swaminathan@hp.com

***Abstract*— Storing and analyzing large volume of structured or unstructured data at the scale of petabytes in applications such as business intelligence of an enterprise, is a daunting task. It is therefore desirable to store data in a compressed form. Compression often is performed using transform methods that permit capture of details in the data while at the same time representing it efficiently; wavelet transform is one such compression technique. Viewing the data as a signal, wavelet transform involves computing coefficients and selecting a small subset judiciously to approximate the signal. Because we pick a subset of the coefficients and not all of them to synthesize the data, wavelet based compression is inherently lossy. Compression in the wavelet domain is traditionally done using *hard* and *soft* thresholding techniques and variants thereof. Based on the notion of "total energy" of a signal, in this paper, we introduce two new thresholding methods called, *level-independent energy* and *level-dependent energy* thresholding. In particular, our energy-based thresholding techniques exploit the relationship between total energy of the signal and wavelet coefficients to obtain compression to meet pre-specified error tolerances. In addition, level-dependent energy thresholding aims at determining wavelet coefficients that carry most "information" at various resolutions of the wavelet decomposition of the data distribution. Detailed studies and experiments over noise contaminated synthetic and TPCH benchmark data sets indicate that energy based thresholding methods yield approximately $100{:}1$ compression with high reconstruction accuracy. As an application of our compression techniques, we show that energy-based thresholding methods improve accuracy of cardinality estimation in databases substantially over the popular methods based on equi-height and max-diff histograms as well as over hard, soft and probabilistic thresholding techniques from the statistics and database literature.**

I. Introduction

Storing and analyzing large volume of structured or unstructured data at the scale of petabytes in applications such as business intelligence of an enterprise, is a challenging task due to computational, storage and network bandwidth limitations. One way to tackle this problem is to compress the data while preserving the properties of the data distribution. The compressed data then can be used as synopsis, for example, during query optimization in order to generate optimal query plans and reduce decision latency in real-time business intelligence applications.

Viewing data as a signal, in this paper, we consider wavelet based analysis and compression of large volume of enterprise data. Carefully performed wavelet analysis can model data distributions accurately, while providing the flexibility to retain only relevant information to enable compression. Intuitively, a *wavelet* is simply a versatile histogram whose shape and resolution can deftly be manipulated to obtain good approximations to data via a series of averaging and differencing operations. Mathematically, a wavelet is nothing but a *basis* vector. The elements of basis vector form an orthonormal system, and typical real-world data can be expressed as a linear combination of basis elements. The weights used in the linear combination are known as *wavelet coefficients* which capture the degree of association between the basis elements and the data. The wavelet basis elements are manipulated via *dilation* and *translation* operations and overlaid with data such that broad global features and minor local details are captured. Dilation operations produce a hierarchy of *levels*. Coefficients computed at each level of the hierarchy are indexed by their corresponding level.

Our approach to compression begins with the understanding that the "total energy" of a signal is directly related to its wavelet coefficients given by the Parseval's identity [14]. Under some conditions Parseval theorem states that the sum of squares of a set of data measured at a some discrete points of a signal, is equal to the sum of squares of its wavelet coefficients. Since all data is not created equal, some of the observations are noisy and so, can be eliminated. This process of elimination is done in wavelet-based compression via deleting noisy and small coefficients. Thus, in general only a small number of wavelet coefficients carry "information" while a preponderant majority hold very little information [24]. This sparse behavior of wavelet transform coding is exploited to compress data while preserving its basic properties. Because we pick a proper subset of the coefficients and not all of them to re-synthesize the data, the wavelet based compression is inherently lossy. Therefore, the challenge is to pick a small set of wavelet coefficients to achieve desired level of accuracy, and this is traditionally done using *hard*, *soft* and *probabilistic* thresholding techniques.

First, we use the notion of *energy*, the square of a wavelet coefficient, *cumulative energy*, the sum of the energies of a subset of coefficients, and *total energy*, the sum of the energies of all the coefficients, to propose new thresholding methods called, *level-independent energy* and *level-dependent energy* thresholding. Level-independent energy thresholding

picks coefficients after we order them based on their energies across all levels whereas level-dependent thresholding picks coefficients in each level separately and then splices them together. In particular, our energy-based thresholding techniques exploit the relationship between total energy of the signal and its corresponding wavelet coefficients to obtain data compression to meet pre-specified error tolerances, and in addition, the level-dependent energy thresholding aims at determining wavelet coefficients that carry most "information" at each resolution of the data distribution. These techniques are intuitive, reliable and flexible with good signal reconstruction properties. Our experiments with real-world enterprise data shows that a compression as good as 100:1 can be achieved using energy-based thresholding techniques.

Next, we study wavelet analysis with energy-based thresholding techniques in database applications. Database researchers have demonstrated the capability of wavelets [5], [6], [15], [17], [16], and one such application of wavelets is in query optimization. Query optimizers use summaries because it is untenable to store or scan the entire table during optimization. Moreover, query optimizers use histograms as compression and analysis tool to get quick-albeit approximate-cardinality estimates of certain columns to generate plans based on a cost model. A histogram, which is simple to construct, is completely specified by number of bins, right boundary values and bin frequency. Few bins result in an out of focus, low resolution image of the true distribution of data, while many bins result in a grainy and choppy image of the distribution. Finding the best fitting histogram is tantamount to density estimation of the data, which is computationally expensive and impractical. In commercial systems, the equi-height histogram is constructed with an arbitrary choice of number bins, and the number of bins is often based on space limitations which is a major short coming. Also, the max-diff histogram is used where the bin boundaries are selected based on arbitrary choice of top k largest differences in the frequencies of the observations in the attribute-value distributions. We establish in this paper that a wavelet is merely a generalized histogram and that contrary to common myth, it is easy to implement. Furthermore, we argue that the gain in estimation accuracy of cardinalities outweigh the additional computational overhead. **[This para needs to be modified to reflect final experimental results.]**

Via detailed and comprehensive experiments over simulated and TPCH benchmark datasets, we demonstrate that the proposed wavelet-based thresholding methods perform at least as well as existing lossy compression methods in the mathematics and the databases literature. In particular, our experimental results show that the energy-based thresholding method outperforms hard thresholding, soft thresholding, and probabilistic thresholding relative to mean square error and minimum relative error with respect to approximating the original data and compression ratios. We tested the performance of the wavelet based methods to histograms and the max-diff histograms [22] for cardinality estimation in query optimization. Our experiments show that the cardinality estimated due to energy-based thresholding produce an accuracy in terms of relative error up to 30 percent for range queries and 25 percent for point queries. Moreover, for the same accuracy, energy-based thresholding achieves three times better compression than hard, soft and probabilistic thresholding. Level-dependent thresholding approach produces improved reconstruction accuracy and better compression ratios, relative to hard, soft, and energy-based thresholding. It is a promising technique which can conceivably be applied to data that follows a natural data hierarchy of levels.

Wavelet based modeling and synopsis in common data and streaming settings have been well studied and chronicled [10], [6], [7], [18], [16], [17]. The work in [6], [18] deal with advancing wavelets as a substitute for the equi-height, max-diff histograms used as synopsis tools for approximate query optimization. Most work has been limited to the application wavelets made up of square waves for modeling the distribution of the data. The data is condensed to a synopsis by the thresholding method of selecting the top (largest in magnitude) few coefficients. The wavelet as a synopsis has been advanced for analyzing multi-dimensional data and in streaming data situations [17], [10] as well.

The rest of the paper is organized as follows. In Section 2, we cover wavelet basics, and in Section 3, we discuss existing wavelet coefficient thresholding methods and related work in cardinality estimation in database applications. In Section 4, we discuss energy-based thresholding and level-dependent thresholding, and provide a mathematical justification and an algorithm to determine the number of wavelet coefficients needed to achieve specified reconstruction accuracy. In Section 5, we discuss experimental results comparing hard, soft, probabilistic thresholding and energy-based methods especially in the context of query optimization in databases. Finally, we conclude with Section 6.

II. Preliminaries

In this section, we begin by introducing wavelets and its basic basic building blocks. Next we discuss the *Haar wavelets*, and illustrate with an example, its construction, how it is used to transform the data, how to compute the wavelet coefficients, and finally, how to use them to reproduce the original data vector by the inverse transform. Finally, we outline *Daubechies wavelets*, another well known wavelet family that generalizes Haar wavelets.

A. Haar and Daubechies Wavelets

Wavelet analysis of data begins with data n-vector $\{y_i\}_{i=1}^n$. Wavelets are merely functions with some special properties that yield the representation

$$y_n(x) = \alpha\varphi(x) + \sum_{j=1}^{m}\sum_{k=0}^{2^{j-1}} \beta_{jk}\psi_{jk}(x),$$

where $m = \log_2 n$ and $\{\alpha, \beta_{j,k}\}$ represent the set of wavelet coefficients, denoted by C. The basic building blocks of wavelets consist of the *father* and the *mother* wavelets denoted

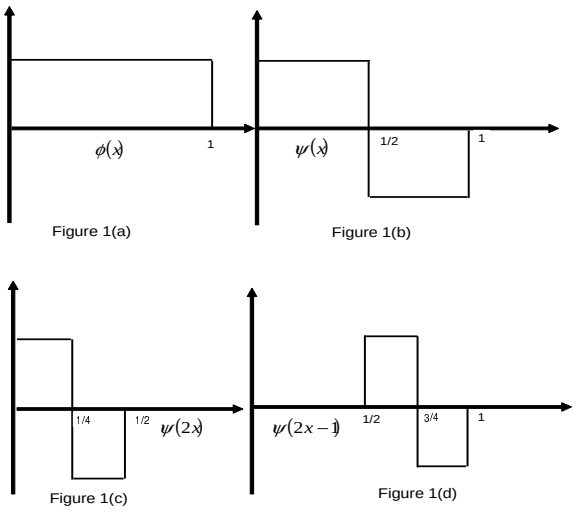

Fig. 1. The father wavelet $\varphi(x)$ is constant over the interval $[0, 1]$ is shown in 1(a). The mother wavelet $\psi(x)$ is given in 1(b). Dilation and translation of $\psi(x)$ are shown in 1(c) and 1(d), respectively. 1(c) is obtained by dividing $\psi(x) \in [0, 1/2)$ into two pieces and flipping the sign of the second piece. 1(d) is obtained by translating 1(c) by 1/2 unit.

by $\varphi(x)$ and $\psi(x)$, respectively. The father wavelet, a constant function, captures the broad global features in the data and the mother wavelet, through a series of *dilation* and *translation* operations, captures the details. The dilation operation recursively divides the interval over which $\psi(x)$ is defined, into segments of lengths 1/2, 1/4, 1/8, and so on. The translation operation shifts $\psi(x)$ by one unit of the interval. (See Figure 1 for an example.) Thus, the wavelet is a new tool to identify global and local features.

Mathematically, $\varphi(x)$ and $\psi(x)$ for Haar Wavelets are given by:

$$\varphi(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x \notin (0, 1) \end{cases} \tag{1}$$

$$\psi(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1/2 \\ -1 & \text{if } 1/2 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In general, the mother wavelet $\psi(x)$ subscripted by dilation and translation operations is written as;

$$\psi_{j,k} = 2^{j/2}\ (2^{-j}x - k),\ j \in \mathbb{Z},\ k \in \{0, 1, 2, \cdots, 2^{j-1}\},$$

where the indices j and k track the dilation and translation parameters, respectively. **[What is H matrix in general? Is there a characteristic function?]** Observe that any two columns of H are mutually orthogonal and that orthogonality property ensures that H is made up of minimum number of columns to compute the wavelet transform and its inverse.

As an example, consider the data array consisting four observations (8, 2, 1, 0). We will represent the data as a linear combination of Haar wavelet elements and the unknown coefficient vector C. The matrix formulation of linear combination is written as $Y_{(4\times1)}$=$H_{(4\times4)}C_{(4\times1)}$, where Y denotes the data vector, H is the wavelet matrix and C is the vector of unknown coefficients to be determined. Using C, obtained by solving the linear system, we will show how to reconstruct the data vector Y using the inverse of the Haar wavelet matrix. Explicitly, the linear system is given by:

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}}_{H} \underbrace{\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix}}_{C}$$

The solution to the linear system is $C = H^{-1}Y$, and the resulting coefficient vector $C = (2.75, 2.25, 3.0, 0.50)]$. To reconstruct the original vector, we pre-multiply vector C with H matrix, i.e., $Y = HC$.

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 2.75 \\ 2.25 \\ 3.00 \\ 0.50 \end{pmatrix} = \begin{pmatrix} 8 \\ 2 \\ 1 \\ 0 \end{pmatrix}$$

In general, approximation by a Haar wavelet to complex data distributions requires several dilations and translations, which in turn leads to computing large number of wavelet coefficients. However, the Haar wavelet is suitable for modeling data which is characterized by a slow changes over time. It has been shown that the top few coefficients minimize the error relative to the $\|L_2\|$ norm [21], [20] for Haar wavelets.

Haar wavelet approximates data using square wave functions where as Daubchies approximates data using polynomial functions, and therefore, the latter is a generalization of the former. Unlike the Haar wavelet, the Daubechies wavelet has no formal functional representation, but begins with a vector of four numbers $\varphi = (a_0, a_1, a_2, a_3)$ and a mother wavelet specified by the vector $\psi = (a_3, -a_2, a_1, -a_0)$. Note that $\varphi \perp \psi$ because the inner product $<\varphi, \psi>$ is zero (to be clear, the inner product of coefficients of the elements of ϕ and ψ is zero). Furthermore to compress the constant vector $(1, 1, 1, 1)$ and the linear vector $(1, 2, 3, 4)$, Daubechies wavelet require two orthogonality conditions: $a_3 - a_2 + a_1 - a_0 = 1$ and $a_3 - 2a_2 + 3a_1 - 4a_0 = 1$. Finally to together with two more conditions $a_1a_3 + a_0a_2 = 1$ and $a_0 + a_1 + a_2 + a_3 = 1$, Daubechies wavelet defines a set of four equations, the solution of which produces a superior set of wavelet coefficients to that of Haar wavelet system.

The Daubechies wavelet begins with a *father wavelet* given by $\varphi(x)$ that is made up of smaller copies of itself. In other words,

$$\varphi(x) = \sum_{k=-\infty}^{\infty} a_k \varphi(2x - k). \tag{3}$$

The a_k's are called the wavelet coefficients. The corresponding *mother wavelet* is written as

$$\psi(x) = \sum_{k=-\infty}^{\infty} (-1)^k a_{1-k} \varphi(2x - k). \tag{4}$$

The mother wavelet coefficients are merely the coefficients of the father wavelet but appearing in reverse order, denoted by a_{1-k}. The father and mother wavelets, $\varphi(x)$ and $\psi(x)$, are chosen such that they satisfy three conditions: (1) Stability, (2) Convergence, and (3) Orthogonality. To satisfy the stability condition, the father wavelet is chosen such that $\int_{-\infty}^{\infty} \varphi(x)dx = 1$, and by solving this integral, we get $\sum a_k = 2$. Stability pertains to the finiteness and uniqueness of the father wavelet. The condition of convergence is needed to ensure that $\psi(x)$ is finite. The orthogonality condition is required to build the minimal set of wavelet basis elements for data transformation and reconstruction. When restricted to n-sized data vector, the above integrals reduce to finite sums between 0 and n. In summary, the four conditions are: **[How did we get (6), (7) and (8)? Can we explain?]**

$$\sum_{k=0}^{n/2-1} a_k = 2 \tag{5}$$

$$\sum_{k=0}^{n/2-1} (-1)^k k^m a_k = 0 \tag{6}$$

$$\sum_{k=0}^{n/2-1} a_k a_{k-2m} = 0 \tag{7}$$

$$\sum_{k=0}^{n/2-1} a_k^2 = 2 \tag{8}$$

where $m = 0, 1, 2, \cdots \frac{n}{2} - 1$ The solution to (1–4) gives the coefficient values, a_k. While, (5) guarantees stability of the father wavelet, it does not assure smoothness and convergence. In order to find a smooth solution, we apply Fourier transform to (3). To achieve a smooth solution, the Fourier transform $\frac{1}{2}\sum_{k=0}^{n/2-1} a_k e^{2\pi x}$ must be zero. Equivalently, the coefficients a_k must satisfy (6). To show othogonality, we must have, $\int(\sum a_k \varphi(2x-k))(\sum a_l \varphi(2x-2m-l))dx = \sum a_k a_{k-2m} \int \varphi_0^2(2x)dx = 0$, for $m \neq 0$. This condition reduces to (7) when the integral equals 1 which we can see is clearly true for the Haar wavelet.

Next, we will show that the Haar wavelet can be derived from the Daubechies wavelet. We will pick $n = 4$ and note that the same argument holds for any n. For $n = 4$, Equations 5 and 6 reduce to: $a_0 + a_1 = 2$, $a_0 - a_1 = 0$, $a_0^2 + a_1^2 = 2$. The unique solution to the system is given by $a_0 = a_1 = 1$. Using the values ofa_0 and a_1, the father wavelet in Equation 3 becomes, $\varphi(x) = \varphi(2x) + \varphi(2x-1)$ and the mother wavelet in Equation 4 becomes $\psi(x) = \varphi(2x) - \varphi(2x-1)$. It is easy to verify that $\varphi(x)$ satisfies Equation 1 and $\psi(x)$ satisfies Equation 2.

Define level precisely, discuss how to generate intermediate approximate coefficients corresponding to levels, and point out that matrix multiplication gives out the last level of coefficients directly.

B. Wavelet Thresholding

The wavelet coefficients are the nexus between the data and the wavelet basis, and they capture the degree of correlation between them. In many applications, the majority of wavelet coefficients are negligible in magnitude [2], [10], [11], [19] and do not contribute to describing the data by the wavelet transform. Therefore these coefficients can be discarded. The resulting set $\{c_i\}_{i=1}^{k}$, $k \ll n$ produces lossy compression of the original data. The process of discarding small magnitude coefficients is known as *thresholding*. Seminal work on thresholding is due to Donoho and Johnstone [11]. They proposed two methods known as *hard thresholding* and *soft thresholding*.

Hard Thresholding (HDT): Consider a dataset $\{y_i\}_{i=1}^{n}$. Hard thresholding selects coefficients based on a thresholding parameter λ. The parameter λ is computed based on the standard deviation σ of the wavelet coefficients at the highest resolution (first level of decomposition). An estimate of σ is given by $\hat{\sigma} = \sum_{i=1}^{n_1} |c_i - \text{med}(c_i)|/0.6745$[1], where med(.) is the median of the coefficients and n_1 denotes the number of coefficients at the first level of decomposition. The threshold is given by $\lambda = \hat{\sigma}\sqrt{(2\log n)/n}$; see [11] for more details. For every $i \leq n$, if $|c_i| \leq \lambda$, then set $c_i = 0$.

Soft Thresholding (SFT): Soft thresholding is based on the attractive idea of wavelet shrinkage [3]. It is similar to hard thresholding; it sets to zero all those coefficients whose absolute values are smaller than λ, and then shrinks coefficients that exceed λ in absolute value towards zero. More precisely, the shrinking is performed by $c_i^* = \text{sgn}(c_i)\ (c_i - \lambda)$, where sgn(.) is the standard signum function. Since λ is a sample estimate of the standard deviation of the noisy coefficients, $(c_i - \lambda)$ is the "de-noised" values of each of the remaining coefficients.

The error in reconstruction by HDT tends to be small but because the thresholding constant λ is usually small, compressibility is not that significant. Even though a strong theoretical argument underlies its construction, SFT fails both criteria of reconstruction error and compressibility. Both HDT and SFT procedures are conservative, and therefore admit many non-value added coefficients. Finally, in both thresholding methods, λ is fixed, which makes them less flexible.

While space-efficient and robust synopsis has been the holy-grail of wavelet based research, the application of wavelets to compressing real-world data has not advanced beyond the Haar wavelet system and basic thresholding methods. The Daubechies wavelet [2] has proven successful in modeling complex data distributions, and so is widely used in signal processing. For data synopsis, HDT and SFT methods have been used extensively [11]. Moreover, commercial software systems MATLAB and SAS use HDT and SFT methods.

Probabilistic Thresholding: Deterministic thresholding, a commonly used technique, entails selecting the top k coefficients for some chosen value of k. The top k coefficients are chosen after iteratively adding and dropping coefficients until no significant error in reconstructing the original data is observed. It has been argued in [23] that optimality of deterministic

[1]This constant is chosen carefully for unbiased estimate of σ.

thresholding under the mean square error metric does not guarantee good reconstruction of the individual data values. So probabilistic thresholding under maximum relative error was suggested as an alternative. This procedure involves rounding up or down of individual data points, and by flipping a coin, it assigns a probability to each coefficient based on its importance to the reconstruction of individual values to build a compressed coefficient vector. Probabilistic thresholding guarantees unbiased estimates of the individual coefficients but as we show later in Section IV, for real-world data, the improvement in reconstruction accuracy is inferior to that of energy-based thresholding under mean square error and minimum relative error criteria.

III. Our Contribution

[Fix this after the Introduction is fixed.] Both HRT and SFT weed out coefficients based on a fixed threshold λ. While it is an attractive argument to use a simple threshold, it turns out that it keeps more coefficients than necessary, which lowers the ability to compress the data. We will leverage the cumulative energy of the data captured by the coefficients to determine the number of coefficients for achieving data smoothing and better compression ratio. The Parseval's identity [14] states that the sum of squares of wavelet coefficients equals total energy in the data. Since coefficients which are small in magnitude have little energy, they provide less information, and so can be discarded. By plotting a graph of cumulative energy versus the number of coefficients, we can select a small set of significant coefficients which leads to substantial compression. This method of thresholding is motivated by principal components analysis where the choice of principal components is based on the idea of cumulative sum of Eigen values.

A. Energy-Based Thresholding

Our approach involves cumulative energy of wavelet coefficients to capture information in the n-vector data. The graphing parameters are given by the set $\{i, \sum_{k=1}^{i} \ c_k^2\}$, where $i \leq n$, and $\sum_{k=1}^{i} \ c_k^2$ is the cumulative energy up to the ith coefficient. As an example, for a data vector whose sample size is 64, approximately 25 coefficients contribute 95% of the total energy. The advantage of the method is that it gives the flexibility to compromise between accuracy and space constraints. In addition, it is not dependent on a rigid thresholding constant. The user can adaptively choose the set of coefficients that meet dual criteria of space and accuracy. In contrast, hard and soft thresholding methods cut off of coefficients based on a threshold determined *a priori*. We now state the algorithm for energy-based thresholding.

Algorithm *Energy-Based-Thresholding*

1) For a given n-vector data, Select a wavelet basis.
2) Obtain the wavelet coefficient set $\{c_i\}_{i=1}^n$ by using the n-vector data and the chosen wavelet basis.
3) Arrange the wavelet coefficients computed in the descending order of magnitude. Denote the ordered set also by $\{c_{(i)}\}_{i=1}^n$.
4) Plot the number coefficients as a function of cumulative energy given by the sum of ordered wavelet coefficients $\{c_{(i)}\}_{i=1}^n$.
5) Find the point on the y-axis of the graph where the incremental cumulative energy is less than a pre-specified value.
6) Select the corresponding point on the x-axis. The point on the x-axis corresponds to number of coefficients whose cumulative energy contribution is approximately maximal for rebuilding the original data.

Energy based thresholding is anchored on the idea that cumulative energies converge to the total energy of a function and that the consecutive differences in cumulative sums converge to zero. In other words, a decreasing sequence of positive numbers bounded below converges to its infimum, that is zero. This assertion validates discarding coefficients beyond some number. Theorem 1 below guarantees the convergence of the differences of cumulative energies (which are always positive) to its infimum, that is clearly zero.

[Fix the theorem and proof.]

Theorem 1: For $1 < j \leq n$, let $\Delta E_j = \sum_{j=1}^{i} c_{(j)}^2 - \sum_{j=1}^{i-1} c_{(j)}^2$, $i \in \{2, \cdots, n\}$, Then sequence $\{\Delta E_j\}_{j=1}^n$ converges to its infimum, $\inf_j(E_j)$.

Proof: We prove that if the decreasing sequence $\{\Delta E_i\}$ is bounded below, then it is convergent and the limit is $\inf_i(\Delta E_i)$. Since $\{\Delta E_i\}$ is decreasing, positive and it is bounded below, by the greatest lower bound property of real numbers $L = \inf_i(\Delta E_i)$ exists and is finite. For every $\epsilon > 0$, there exists a $\Delta E_N < L + \epsilon$, otherwise $L + \epsilon$ is the infimum of $\{\Delta E_i\}_{i=1}^n$ which contradicts that $L = inf_i(\Delta E_i)$. Since $\{\Delta E_i)$ is a decreasing sequence, by Cauchy's convergence criterion, for all $n > N(\epsilon)$, $\Delta E_n - L \leq \Delta E_N - L < \epsilon$. Hence the limit of the sequence $\{\Delta E_i\}_{i=1}^n$ is given by $L = inf_i(\Delta E_i)$. ■

The theorem justifies the use of energy based thresholding. When a wavelet transform is applied to data with finite energy, the corresponding wavelet coefficients set tends to zero.

B. Thresholding Based on Pre-specified accuracy

In many situations, the user wants to know the number of coefficients needed to meet a pre-specified level of accuracy (quality of reconstruction). This is helpful to estimate tradeoffs between storage space needed and accuracy of reconstruction. This is the converse of the problem we looked at previously, where we picked the number of coefficients that provided a certain amount of information (cf. 3.0). Towards this end, we use the Cauchy's convergence principle and the monotone convergence theorem from mathematical analysis [14] to determine number of coefficients required to meet a specified level of accuracy. Preliminary to outlining the algorithm to determine coefficients required to attaining certain accuracy, we outline the mathematical machinery required to verify our assertions.

Theorem 2: Let $\{\phi_0, \phi_1, \ldots\}$ be orthonormal on I. Assume that $f \in L^2(I)$ and suppose that $f(x) = \sum_{n=0}^{\infty} c_n \phi_n(x)$.

Then:

1) The series $\sum |c_n|^2$ converges and satisfies the inequality $\sum_{n=0}^{\infty} |c_n|^2 \leq ||f^2||$. The inequality is known as the Bessel's inequality.
2) The equation $\sum_{n=0}^{\infty} |c_n|^2 = ||f^2||$ is known as the Parseval's formula.

Holds iff $Lim_{n\to\infty}||f - s_n|| = 0$, where $\{s_n\}$ is the sequence of partial sums defined by $s_n(x) = \sum_{k=0}^{n} c_k \phi_k(x)$.

Theorem 3: Let $\{E_i\}_{i=1}^{\infty}$ be an increasing series of partial sums ($E_i = \sum_{j=1}^{i} c_j^2$) which is real, positive and bounded by $\| (f) \|^2$=M, where $\| (f) \|^2$ is the total energy of $f \in L^2(I)$ and is a real number. The number of $\{c_i\}$ required to obtain an accuracy is given by the solution n to the inequality $|\sum_{i=1}^{n} E_i - M| \leq \epsilon$.

Proof: Consider a sequence of partial sums $\{E_i\}_{i=1}^{n}$. Since $E_n \leq M$ by Bessel's inequality in Theorem 2, we have each element in the sequence $\{E_i\}_{k=1}^{\infty}$ is bounded by M. Since the series converges to M, this means that for $n \geq N$ we have $|\sum_{i=1}^{n} c_i^2 - M| \leq \epsilon$ by the Cauchy convergence criterion. Since the difference $|\sum_{i=1}^{n} c_i^2 - M|$ measures the error in energy between the finite set of coefficients $\{c_i\}_{i=1}^{n}$ and M, an error within ϵ is obtained by finding the value of n, which satisfies the inequality $|\sum_{i=1}^{n} c_i^2 - M| \leq \epsilon$. ■

Table 7 shows the dependency of degree of compression (number of coefficients required) versus accuracy. The box below describes algorithmically, the steps required to determine the number of coefficients needed to achieve an accuracy of ϵ.

Algorithm *Pre-Specified Accuracy*

1) Select a wavelet basis for modeling and compressing $\{y_i\}_{i=1}^{n}$
2) Obtain the wavelet coefficient set $\{c_i\}_{i=1}^{n}$ by applying the desired wavelet transform
3) Arrange the wavelet coefficients computed at all resolutions in the descending order of magnitude. Denote the ordered set by $\{c_{(i)}\}_{i=1}^{n}$.
4) Compute the total energy of the data vector $(y_1, y_2, \ldots, y_n)$ given by $\sum_{i=1}^{n} y_i^2 = M$
5) Compute the cumulative energies $\sum_{i=1}^{k} c_{(i)}^2$
6) Choose desired precision/accuracy (ϵ)
7) Find $k(\epsilon) \exists |\sum_{i=1}^{k} c_i^2 - M| \leq \epsilon$

The value $k(\epsilon)$ is the number of coefficients required to achieve an accuracy of ϵ, i.e., one needs to find the value of k which depends on ϵ such that the inequality is satisfied.

Table 7 is a summary of how wavelet energies can be used to determine a subset of wavelet coefficients that guarantee a desired level of accuracy(ϵ). For the illustration, data is generated according to the Doppler function. For a sample of size 16, a subset of 9 coefficients (largest in magnitude) is needed to achieve a reconstruction error of 5%. If the desired error level is 10%, fewer coefficients equal to 7 are required. Similarly, for other sample sizes, as precision increases (error decreases) the number coefficients increases. The same pattern holds for other members of the Daubechies (DbN) family wavelets.

C. Level Dependent Thresholding

Thresholding methods in previous sections were based on computing a threshold after computing all the coefficients at every resolution. The energy based thresholding too deleted coefficients based on all the computed coefficients. It has been observed that even though the original data is highly correlated, the wavelet coefficients however, exhibit much less dependence [25]. If the noise in the data is correlated and the process is stationary (stationarity means, the data moves around a constant mean), the variance of the coefficients will depend on the resolution level, but will be a constant within the resolution level. This phenomenon argues for a level dependent thresholding approach. In the level dependent thresholding (LDT) approach, we compute within level total energy and select coefficients within that level which capture a large proportion of the information. In the end, we will combine the surviving coefficients at each level into a single compressed vector and use that for decompression. Table 7 shows the performance of level dependent energy based thresholding (LDT) applied to the noise contaminated Doppler, Bumps, Quadchirp, and Mishmash functions using the Daubechies (Db4) basis analyzed at a sample size of 16384 discrete points. Examining the RC statistic in the last column, the gain in compression is as high as 2.25 times and accuracy (SSE) is between 40%-80% due to LDT. Examining the results for the "Bumps" function, while we sacrifice compression i.e, RC statistic is (0.40), the gain inaccuracy is 80%! If an application requires reconstructing vectors at various levels in a hierarchy: as an example; consider the configuration of a data center. Topologically, a server has several inlet/outlet sensors, multiple servers belong to a rack, racks belong to a zone and multiple zones constitute a data center. Within that topological architecture, it may be desirable to compress data at the level in the hierarchy. In that scenario, LDT may be handy. We are conducting further research in this area, to exploit the property of de-correlated coefficients across levels and the practical utility of hierarchical dependent analysis.

IV. Experimental Results

In this section, we discuss the performance of the many thresholding methods we introduced, proposed and discussed. The methods include: hard thresholding, soft thresholding, energy based thresholding, probabilistic thresholding, and level dependent thresholding, respectively denoted by HRT, SFT and EBT, GAR, and LDT. We will evaluate the techniques relative to compression ratio (CR) mean square error (MSE) and/or minimum relative error (MRE) criteria. The evaluation environment includes synthetic data generated from six different underlying functions contaminated by Gaussian Noise [11]. They are respectively, 'blocks', 'bumps', 'heavy sine', 'doppler', 'quadchirp', and 'mishmash' as well as data from the TPCH benchmark tables used in databases studies. We will demonstrate the utility of the techniques in cardinality estimation by comparing against the standard compression techniques such as equi-height and max-diff histograms. Please note that since typical wavelet analysis applies to data whose

size is a power of two such as 4, 16, 64, 128, and so on, our experimental data consist of vectors whose lengths are powers of two.

The statistic CR is the compression ratio used to measure degree of compression is given as,

$$CR = \frac{n}{\#\text{of coefficients used}}$$

, where *n* is the total number of coefficients.

To measure the quality of reconstruction of the original signal, we use the mean sum of squares(MSE) criterion given by $\frac{1}{n}(Y_i - Y_i^*)^T(Y_i - Y_i^*)$ where Y_i and Y_i^* are respectively the original and the reconstructed data vectors. The symbol *T* denotes the transpose of a vector.

The minimum relative error (MRE) statistic also is used to measure quality of reconstruction of the original signal, is given as,

$\sum_{i=1}^{n} \frac{|\hat{y_i}-y|}{(|y_i|,S)}$, where S is a sanity bound, typically chosen to be the 10^{th} percentile of the distribution of the data.

The measures MSE and MRE give us an idea about the accuracy of reconstruction. An estimate of the accuracy of reconstruction is obtained by computing the difference between the original data (Y) and a version of it obtained after applying a thresholding technique. We also measure, "Energy" measured in % units is given by

$$PE = \frac{\sum_{i=1}^{k} c_i^2}{\sum_{i=1}^{n} c_i^2} \times 100, k < n$$

.

In the following, we will discuss the results of the experiments in detail. Figure 2 consists of graphs corresponding to the six Gaussian noise contaminated distributions.

Method	n	# of Coeffs.	% Energy Used	MSE	CR
HRT	1024	62	100.0	0.0000	16.52
SFT	1024	62	100.0	0.0000	16.52
EBT	1024	54	99.90	0.0030	18.96
HRT	16384	103	100.0	0.0000	159.1
SFT	16384	103	100.0	0.0000	159.1
EBT	16384	76	99.90	0.0020	215.6
HRT	32768	112	100.0	0.0000	292.6
SFT	32768	112	100.0	0.0000	292.6
EBT	32768	76	99.9	0.0030	431.2
HRT	65536	117	100.0	0.0000	560.1
SFT	65536	117	100.0	0.0000	560.1
EBT	65536	76	99.90	0.0000	862.3

Fig. 3. Comparison of the thresholding methods using the Haar wavelet analyzing blocks distribution.

The columns in Figures 3-5 are the following: column 1 is the thresholding method applied, Column 2 is the sample size, column 3 is the number of coefficients retained after applying thresholding, column 4 is the % energy used, Column 6 is the mean square error (MSE) and Column 7 is the compression ratio (CR). In Figures 3-4 we summarize the

Method	n	# of Coeffs.	% Energy Used	MSE	CR
HRT	1024	420	99.97	0.0003	2.44
SFT	1024	420	97.35	0.0020	2.44
EBT	1024	305	99.90	0.0010	3.36
HRT	16384	5253	100.0	0.0000	3.12
SFT	16384	5253	99.91	0.0000	3.12
EBT	16384	436	99.90	0.0010	37.58
HRT	32768	10096	100.0	0.0000	3.25
SFT	32768	10096	100.0	0.0000	3.25
EBT	32768	437	96.90	0.0010	74.98
HRT	65536	19415	100.0	0.0000	3.38
SFT	65536	19415	100.0	0.0000	3.38
EBT	65536	436	96.89	0.0010	150.31

Fig. 4. Comparison of the thresholding methods using the Haar wavelet analyzing doppler distribution.

Method	n	# of Coeffs.	% Energy Used	MSE	CR
HRT	1024	543	100.00	0.0000	1.89
SFT	1024	543	99.36	0.0000	1.89
EBT	1024	398	99.42	0.0400	2.57
HRT	16384	559	100.00	0.0000	29.31
SFT	16384	559	99.65	0.0000	29.31
EBT	16384	398	98.96	0.0000	41.17
HRT	32768	565	100.00	0.0000	58.00
SFT	32768	565	99.74	0.0000	58.00
EBT	32768	398	98.96	0.0000	82.33
HRT	65536	570	100.00	0.0000	114.96
SFT	65536	570	99.81	0.0000	114.98
EBT	65536	570	98.82	0.0000	164.66

Fig. 5. Comparison of the thresholding methods using the Db4 wavelet analyzing doppler distribution.

relative performance of the HRT, SFT and EBT procedures in conjunction with the Haar wavelet over the "blocks," and "doppler" distributions. An examination of Figure 3 reveals that the energy based thresholding produces compression ratios as high as 800:1, while its mean square error is as low as those of HRT and SFT. From Figure 2, we notice that the "blocks," signal is relatively slowly changing. The Haar wavelet, which approximates the data by a set of square-waves(boxes) is suitable and hence the compression ratio and MSE are excellent. Figure 4 are the results due to the application of the Haar wavelet to the "doppler" data. Approximation by Haar, results in satisfactory compression ratio and MSE, but we will show that the fit can be improved by a suitably chosen higer order Daubechies wavelet. This is because the signal is rapidly changing and approximation by a higher order (polynomial) wavelet is more appropriate. Figure 5 summarizes the results of applying Daubechies (Db4) to the "doppler signal. Comparing Figures 4 and 5, It is quite evident that the Db4 wavelet provides a better fit than Haar both in

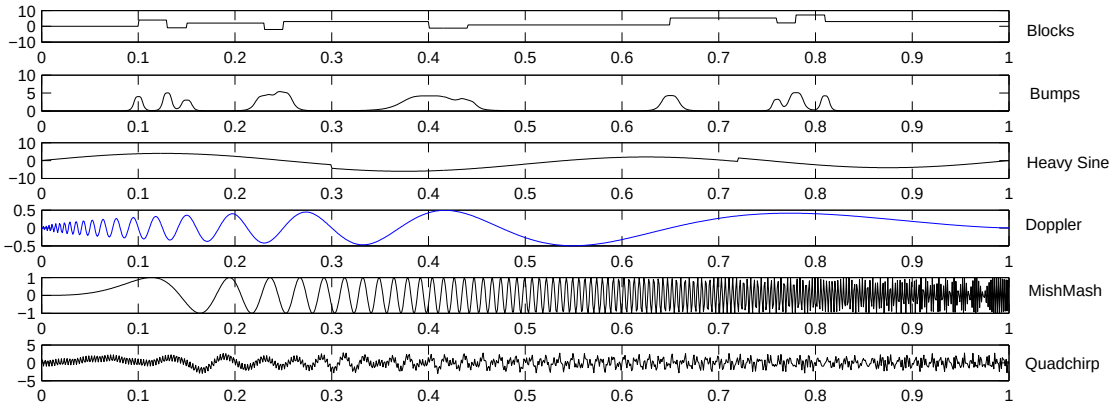

Fig. 2. The six functions contaminated by Gaussian noise.

terms of compression ratio as well as MSE. The illustration suggests the importance of proper selection of a wavelet basis in fitting to data distributions. Finally, the highlight of our experimental study is that EBT outperforms HRT and SFT thresholding substantially without sacrificing reconstruction accuracy. This is a step forward as EBT enables higher levels of compression, while maintaining fidelity to the original data.

In the following, we compare hard, soft and energy based thresholding against the probabilistic thresholding (GAR). For the comparative analysis, we have chosen the TPCH benchmark data. We used the attribute (line item part key) for analysis. The summary columns in Figure 6 are: column 1 is the sample size, columns 2-5 are the coefficients retained after thresholding, columns (6-9) capture the respective mean square errors. In Figure 7, we tabulated the mean relative errors incolumns (6-9), and columns 1-5 being the same as those in Figure 6.

Figure 6 compares the procedures in terms of the MSE criterion. Figure 7 compares the estimates in terms of the MRE criterion. The data analyzed over different sample sizes of the attribute values, demonstrates superior performance of HRT, SFT, and EBT procedures compared to GAR. The HRT, SFT, and EBT produce mean squared errors comparable to one other, while the MSE due to GAR is worse by a magnitude of at least 2x. Since, it was argued that the MSE crterion as an overall metric does not estimate the individual data values and MRE is a suitable alternative measure [**?**], we compared the four procedures relative to the MRE statistic. Figure 7 is summarizes the results. Again, HRT, SFT, and EBT outperform GAR. While HRT, SFT, and EBt are comparable, probabilistic thresholding (GAR) on the average is worse by a magnitude of at least 2x. The compression ratios relative to the four methods are not reported for space considerations. But a quick look at columns 1 and (2-5) and (performing necessary ratio calculations) shows that HRT, SFT, and EBT result in higher compression rates compared to GAR. Comparisons among HRT, SFT, and EBT reveals that HRT and SFT produce

n	Data	% Accuracy	# of Coeffs	MSE	MRE	CR
4096	TPCH	1%	2826	26.41	1.26	1.45
4096	TPCH	2%	2526	26.51	1.30	1.62
4096	TPCH	4%	2176	26.72	1.29	1.88
4096	TPCH	6%	1946	26.90	1.28	2.10
4096	TPCH	8%	1767	27.09	1.27	2.32
4096	TPCH	10%	1622	27.25	1.26	2.53
32768	TPCH	1%	22709	26.81	1.32	1.44
32768	TPCH	2%	20332	26.92	1.31	1.61
32768	TPCH	4%	17482	27.12	1.30	1.87
32768	TPCH	6%	15593	27.30	1.29	2.10
32768	TPCH	8%	14152	27.48	1.28	2.32
32768	TPCH	10%	12894	27.64	1.27	2.54
131072	TPCH	1%	90846	26.97	1.32	1.44
131072	TPCH	2%	81428	27.08	1.31	1.61
131072	TPCH	4%	70076	27.27	1.30	1.87
131072	TPCH	6%	62542	27.45	1.29	2.10
131072	TPCH	8%	56788	27.62	1.28	2.31
131072	TPCH	10%	52096	27.79	1.27	2.52

Fig. 8. Relationship between number of coefficients required and accuracy of reconstruction using "Line item Part key" column from TPCH.

lower mean square errors then EBT. Since we chose to retain coefficients that account for 90% of the energy for compression due to EBT, it is less accurate. We can increase the accuracy due to EBT by adding coefficients that account for more energy. This enables more accurate reconstruction, but there will be a proportional reduction in compressibility. In the following discussion, we will show the relationship between accuracy and compressibility, summarized succinctly in Figure 8.

Figure 8, shows the relationship between compressibility and accuracy and the ability of EBT to let the experimenter control accuracy and compressibility. Accuracy again pertains to the gap between the original data and the data vector reconstructed after compression. Column 1 consists of the sample size of the data vector, column 2 refers to the source of the attribute values, column 3 is the accuracy desired, column

n	# of Coeffs(HRT)	# of Coeffs(SFT)	# of Coeffs(EBT)	# of Coeffs(GAR)	MSE(HRT)	MSE(SFT)	MSE(EBT)	MSE(GAR)
4096	3890	3890	2052	4012	26.28	26.47	26.81	53.21
8192	7893	7893	4110	8090	26.52	26.66	27.06	58.73
16384	15905	15905	8184	16110	26.51	26.60	27.04	63.22
32768	32103	32103	16467	32748	26.68	26.75	27.21	68.10
65536	64594	64594	33032	65536	26.77	21.86	27.29	72.72
131072	129623	129623	66013	130011	26.84	26.88	27.37	77.49

Fig. 6. Comparison of the thresholding methods using the Db4 wavelet analyzing Doppler distribution.

n	# of Coeffs(HRT)	# of Coeffs(SFT)	# of Coeffs(EBT)	# of Coeffs(GAR)	MRE(HRT)	MRE(SFT)	MRE(EBT)	MRE(GAR)
4096	3890	3890	2052	4012	1.31	1.31	1.33	2.08
8192	7893	7893	4110	8090	1.32	1.34	1.36	2.16
16384	15905	15905	8184	16110	1.33	1.35	1.36	2.25
32768	32103	32103	16467	32748	1.33	1.35	1.37	2.32
65536	64594	64594	33032	65536	1.33	1.36	1.37	2.38
131072	129623	129623	66013	130011	1.33	1.33	1.39	2.43

Fig. 7. Comparison if the thresholding methods using the Db4 wavelet analyzing Doppler distribution.

4 is the number of coefficients required to meet the specified accuracy (column 3), Columns 5 and 6 are the MSE and MRE respectively, and column 7 is the CR statistic. It is clear from the data that as the gap between the true vector and reconstructed vector widens, accuracy (col 3) decreases, and compressibility (CR)given in column 6 increases. The main point of the table is to demonstrate that EBT gives the user the flexibility to evaluate trade-off between compressibility (storage) and accuracy . If space requirements is a constraint, one can see its effect on accuracy and chose that combination of accuracy and compresibility that meets requirements. Notice that as CR increases, the statistics MSE and MRE increase as well.

In the following subsection, section, we will apply the disparate techniques to the cardinality estimation problem. A typical cardinality estimation scenario consists of estimating the number of tuples (observations) from a random sample of tuples drawn from a column of a database table. Cardinality is simply the count of the number of tuples in column of a database table. Estimating occurrences of specific tuple is known as a point query, and estimating the number of tuples occurring within an interval is known as a range query.

A. Wavelets vs Histograms for Cardinality Estimation

In the previous sections, we presented ways to compress data and saw EBT as a viable method alternative to hard thresholding,soft thresholding, and probabilistic thresholding. In database applications, especially in query optimization, synopses of some specific columns of the tables are stored. These synopses are used by the optimizer at compile-time to generate optimal query plans. The query plan is based on a cost model relative to run-time execution. Needless to say, an accurate synopsis of the attributes of interest is imperative. Common synopsis methods are histograms. A variety of histogram techniques are in vogue. In commercial database systems, the equi-height and max-diff histograms are used. While the equi-height and max-diff histograms are simple to implement, they do not produce satisfactory cardinality estimates when the distributions are complex, which lead to poor query plans.

We compared the performance of wavelet thresholding to equi-height histograms using data drawn from the six distributions(Figure 2). We begin with a set of range queries $a \leq X \leq b$, where X is a sample from one of the six distributions (Figure 2). We compare the true cardinality (the actual occurrence of tuples between the two limits) and the estimated cardinalities from the four methods. The performance was evaluated relative to the absolute relative error (RE) criterion. The relative error is defined as: $\frac{|C-\hat{C}|}{C} \times 100$, where C and $\hat{C}$ are the true and estimated cardinalities respectively. Figure 9 captures the comparison of the wavelet based compression to the equi-height histogram. Column 1 is the distribution of data used, column 2 is the sample size of the data vector, column 3 is the actual cardinality (actual number of tuples within the chosen range), columns 4-8 are the estimated cardinalities by the equi-height, and the wavelet methods, columns 9-12 are the relative errors. Column 5 (wavelet) is the estimated cardinality when no thresholding is applied.

Examining the table, it is clear that wavelets based cardinality estimation is in general better at estimating the cardinalities than the equi-height histogram. Observe that the error in estimation due to equi-height histogram is as high as 75%. The three wavelet methods perform well except the soft thresholding method in some instances. We may be reminded that the performance of EBT can be adjusted to improve estimation accuracy by decreasing the the amount of thresholding. We did not report the degree of compression due to space considerations, but the degree of compression achieved by EBT is significantly larger, in excess of 100% compared to the competing procedures. For this analysis, the number of bins for the histogram was set to $\sqrt{n}$, where n is the sample size. Note that we are not reporting the

Dist	n	Actual Cardinality	EQH	Wavelet	HRT	SFT	EBT	RE(EQH)	RE(HRT)	RE(SFT)	RE(EBT)
Blocks	16384	1310	1024	1310	1268	1241	1308	21.83%	3.21%	5.27%	0.15%
Bumps	16384	171	211	171	168	164	168	23.39%	1.75%	4.09%	1.75%
Heavy-Sine	16384	277	486	277	288	256	278	75.45%	3.97%	7.58%	0.36%
Doppler	16384	455	470	455	456	448	456	3.30%	0.22%	1.54%	0.22%
Quadchirp	16384	572	577	572	579	709	573	0.87%	1.22%	23.95%	0.17%
Mishmash	16384	563	586	563	566	544	564	4.09%	0.53%	3.37%	0.18%
Blocks	32768	32113	32128	32818	32113	32113	32110	0.05%	0.00%	0.00%	0.01%
Bumps	32768	32188	32256	32188	32183	32183	32186	0.21%	0.02%	0.02%	0.01%
Heavy-Sine	32768	32116	32111	31116	32116	32116	32118	0.02%	0.00%	0.00%	0.01%
Doppler	32768	32117	32128	32117	32120	32122	32123	0.03%	0.01%	0.02%	0.02%
Quadchirp	32768	32111	32237	32111	32114	32118	32122	0.39%	0.01%	0.02%	0.03%
Mishmash	32768	32274	32263	32274	32275	32355	32274	0.03%	0.00%	0.25%	0.00%

Fig. 9. Cardinality estimate from the four procedures for large samples.

relative error corresponding to the wavelet procedure with no thresholding in the table as it produces perfect reconstruction.

Continuing, we evaluated the performance of HRT, SFT, EBT, GAR, Equi-height histogram (EQH), and Max-Diff histogram (MDF) using the TPCH benchmark data. The attribute value is the column (line item part key). The results are summarized in Table 10.

Table above summarizes cardinality estimates in response to a point query. Clearly, the performance of EBT is much superior to all the competing procedures in terms of the relative error (RE) metric. The performance of probabilistic thresholding (GAR) produces inaccurate estimates of the true cardinality. The max-diff and equi-height histograms compete well with the wavelet based thresholding, but EBT results in more accurate estimates. Based on the results, we can argue in favor of wavelets in conjunction with hard, soft, and energy based thresholding as an alternative to histogram based synopsis.

In the next section, we will summarize some results pertaining to the idea of level dependent thresholding we advanced in section 3.1.

B. Level Dependent Thresholding

Level dependent thresholding is rooted in the idea that wavelet coefficients obtained at each level of decomposition are independent of their cousins at other levels. This assertion argues for thresholding at every level of decomposition. Figure 11 compares HRT and SFT thresholding procedures against energy based level dependent method denoted by LDT. Columns in Figure 11, capture the method applied, the sample size of the data vector (n), the data distribution, the number of coefficients used, the statistics MSE, and CR respectively.

Examining the table, clearly the LDT method outperforms both HRT, and SFT methods. The compression ratio (CR) and the mean square error (MSE) due to LDT are significantly better. As the sample size becomes large, compression due to LDT is less significant. The MSE due to LDT is low and is less than or equal to those of HRT and SFT. This is especially true for smaller sample sizes. Level dependent thresholding is a promising technique and we will study its applications in query optimization and other business intelligence applications and report findings in a future paper.

Method	n	Distibution	# of Coeffs.	MSE	CR
HRT	16384	Doppler	404	0.27	40.55
SFT	16384	Doppler	404	0.27	40.55
LDT	16384	Doppler	228	0.17	71.86
HRT	16384	Quadchirp	9143	1.20	1.79
SFT	16384	Quadchirp	9143	1.19	1.79
LDT	16384	Quadchirp	4085	0.50	4.01
HRT	16384	Bumps	609	7.85	26.90
SFT	16384	Bumps	609	7.85	26.90
LDT	16384	Bumps	1526	6.30	10.74
HRT	16384	Mishmash	15258	3.42	1.07
SFT	16384	Mishmash	15258	3.35	1.07
LDT	16384	Mishmash	11217	2.54	1.46
HRT	16384	TPCH	16031	0.00	1.07
SFT	16384	TPCH	16032	0.01	1.07
LDT	16384	TPCH	14399	0.00	1.46
HRT	32768	TPCH	32270	0.00	1.02
SFT	32768	TPCH	32270	0.01	1.02
LDT	32768	TPCH	21083	0.02	1.55
HRT	131072	TPCH	129988	0.00	1.01
SFT	131072	TPCH	129988	0.00	1.01
LDT	131072	TPCH	129759	0.00	1.09

Fig. 11. Comparison of Hard, Soft and Level Dependent using the Db4 wavelet over Doppler, Quadchirp and Mishmash and TPCH using Db4

V. Conclusion

In conclusion, the argument for using energy based wavelet based methods as a tool for compression as well as application in query optimization is compelling. We attempted to demonstrate how the technology of wavelets in itself can be improved by the introduction of the novel energy based thresholding, and level dependent thresholding. The introduction of higher order wavelets of Daubechies hold great promise for data characterization and subsequent compression. Their potential for application in cardinality estimation in the context of

Dist	n	Actual Cardinality	EQH	MDF	GAR	HRT	SFT	EBT	RE(EQH)	RE(MDF)	RE(GAR)	RE(HRT)	RE(SFT)	RE(EBT)
TPCH	1024	110	112	116	98	112	83	109	1.82%	5.45%	10.91%	1.82%	24.55%	0.92%
TPCH	16384	1670	1638	1631	1516	1634	1710	1632	1.92%	2.34%	9.22%	2.16%	2.40%	2.33%
TPCH	32768	3356	3277	3412	3703	3303	3454	3299	2.35%	1.67%	10.34%	1.58%	2.92%	1.73%
TPCH	65536	6660	6554	6563	8945	6495	6847	6643	1.59%	1.46%	34.31%	2.48%	2.81%	0.26%
TPCH	131072	13138	13107	13162	16121	13136	13513	13149	0.24%	0.18%	22.71%	0.02%	2.85%	0.08%

Fig. 10. Cardinality estimate from the six procedures for large samples.

approximate query processing is also promising. Cardinality estimation is key in BI intelligence environments because of the multitude of query mixed workloads. So efficient data management via compression and analysis is desirable to reduce run time execution of queries to improve query throughput.

ACKNOWLEDGMENT

The authors would like to thank various sponsors for supporting their research.

REFERENCES

[1] A. Cohen, I. Daubechies, B. Jawerth, and P. Vial, "Multiresolution Analysis, Wavelets, and Fast Algorithms on an Interval," *Comptes Rendus Academie des Sciences*, Paris (A), 316, pp. 417–421, 1993.

[2] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Communications in Pure and Applied Mathematics*, 41, pp. 909–996, 1988.

[3] I. Daubechies, "Ten Lectures on Wavelets," *CBMS-NSF Series in Applied Mathematics*, No. 61, Society for Industrial and Applied Mathematics, 1992.

[4] Y. Meyer, "Ondelettes et Operateurs; Ondelettes, Operateurs de Calderon-Zygmund, III," *Operateurs multilineaires*, (English translation published by Cambridge University Press), 1990.

[5] M.H.D. Kiem, "Wavelets and their application in Databases," *Tutorial Notes of ICDE*, 2001

[6] K. Chakrabarthi, M. Garofolakis, R. Rastogi, and K. Shim, "Approximate query processing using wavelets," *The VLDB Journal*, 10, pp. 199–223, 2001.

[7] M. Garofolakis, J. Gehkre, and R. Rastogi, "Querying and mining data streams: you only get one look" *ACM SIGMOD*, 2002.

[8] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M.J. Strauss "One–pass wavelet decompositions of data streams," *IEEE Transactions on Knowledge and Data Engineering*, 15(3), pp. 541–554, 2003.

[9] G. Strang, "Wavelets," *American Scientist*, pp. 250–255, 1994.

[10] A. Gilbert, Y. Kotidis, S. Muthukrishnan, M.J. Strauss, "Surfing Wavelets on Streams: One–Pass Summaries for Approximate Aggregate Queries," *VLDB*, 2001.

[11] D.L. Donoho, and I. Johnstone, "Spatial Adaptations by Wavelet Shrinkage," *Biometrika*, 81(3), pp. 425–455, 1994.

[12] S. Mallat, "A Theory of Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), pp. 674–693, 1989.

[13] I. Johnstone and B. Silverman, "Wavelet Threshold Estimators for Data Correlated Noise," *Journal of Royal Statistical Society*, 59(2), pp. 319–351, 1997.

[14] T. Apostol, *Mathematical Analysis*, Addison Wesley Longman, 1974.

[15] M. Garofalakis and P. Gibbons, "Wavelet Synopsis with Error Guarantees," *ACM SIGMOD*, 2002.

[16] Y. Matias, J.Vitter and M. Wang, "Dynamic Maintenance of Wavelet Based Histograms," *VLDB*, 2000.

[17] J. Vitter and M. Wang, "Approximate Computation of Multidimensional Aggregates of Sparse Data using Wavelets," *ACM SIGMOD*, 1999.

[18] Y. Matias, J. Vitter and M. Wang, "Wavelet-Based Histograms for Selectivity Estimation," *ACM SIGMOD*, 1998.

[19] W. Hardle, G. Kerkyacharian, D. Picard, and A. Tsybakov, Wavelet, *Approximation, and Statistical Applications*, Springer, 1998.

[20] S. Guha and B. Harb, "Wavelet synopsis for data streams: minimizing non-Euclidean error," *KDD*, 2005.

[21] M. Garofalokis and A. Kumar, "Determining wavelet thresholding for maximum error metric," *PODS*, 2004.

[22] P. Viswanath, P.J. Haas, Y.E. Ionnadis and E.J. Shekita, "Improved histograms for selectivity estimation of range predicates," *SIGMOD Record* 25(2), pp. 294–305, 1996.

[23] M. Garofalakis, P.B. Gibbons, "Probabilistic Wavelet Synopsis," *ACM Transactions on Database Systems* 29(1), pp. 43–90, 2004.

[24] E.J. Candés and M.B. Wakin, "An introdution to compressive sampling," *IEEE Signal Processing Magazine*, 2008.

[25] "Threshold Estimators for Data with Correlated Noise," *Journal of the Royal Statistical Society, Series B* 59(2), pp. 319–351, 1997.