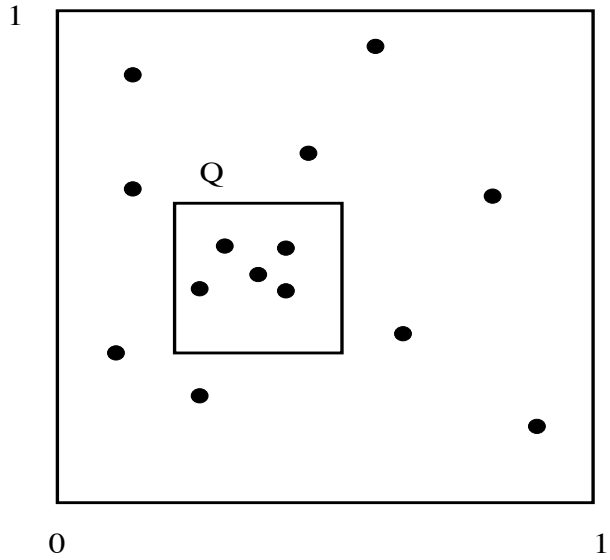
Using Kernels to Approximate Multidimensional Range Queries Over Real Attributes

Carlotta Domeniconi - UC Riverside
Dimitrios Gunopulos - UC Riverside
George Kollios - Boston Univ.
Vassilis Tsotras - UC Riverside

Query Selectivity Estimation

- Given a multidimensional point dataset, estimate the result of a range query.
- We assume that dimensions are numerical continuous attributes.
- Features may be correlated.
- We want to find efficiently an approximate answer.



Motivation

- Computing approximate answers to multidimensional range queries is a problem that arises in many settings:
  - **Database query optimization**: evaluate different execution plans, optimize top-k queries.
  - **Data exploration**: it allows efficient range and OLAP-type aggregate query approximation.

Density Estimators

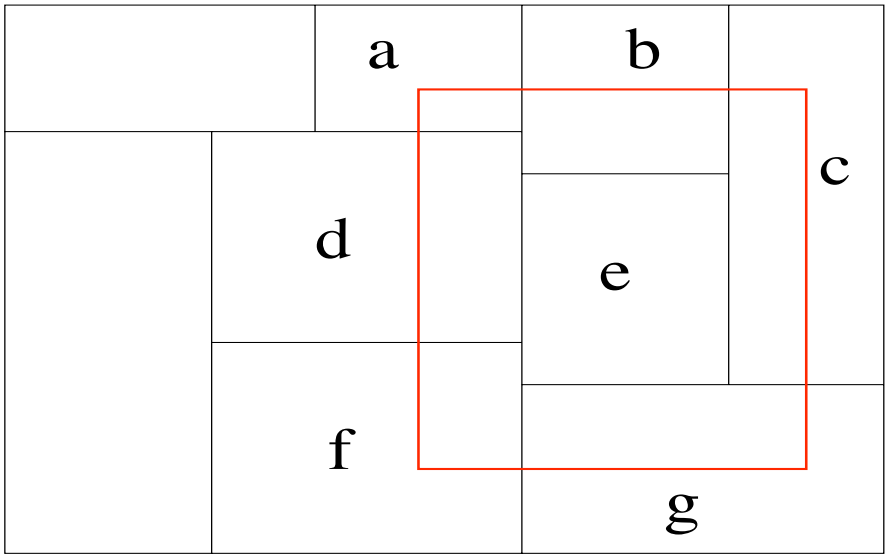
- A general approach is to use a non-parametric technique to approximate the data distribution (density function) using limited space.
- Density function:

$$\int_{[0,1]^d} f(x_1, \ldots, x_d) dx_1 \ldots dx_d = 1$$

- Selectivity estimation:

$$sel(f, Q) = N \int_Q f(x_1, \ldots, x_d) dx_1 \ldots dx_d$$

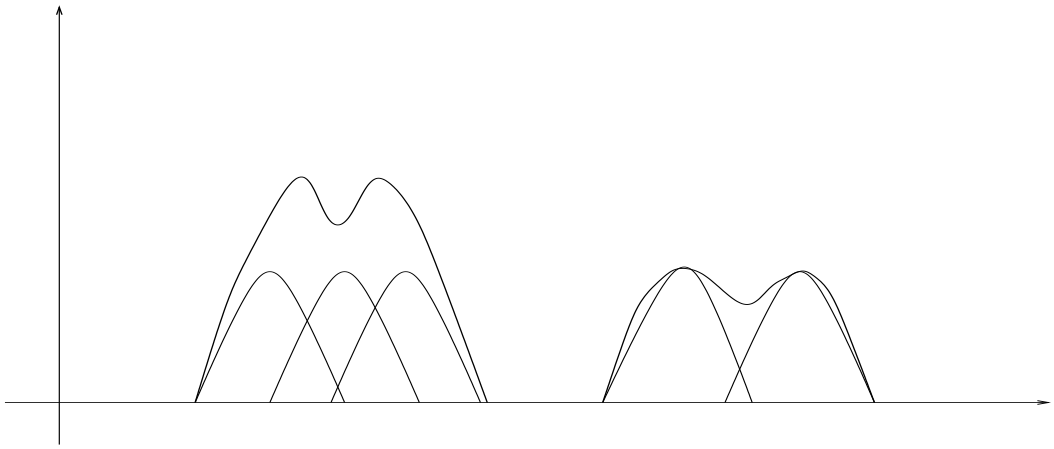
Multi-dimensional Histograms



- Particularly suited when each attribute has a finite discrete domain.
- **Problem**: Efficient construction of accurate histograms. Real valued attributes and high dimensionality $\rightarrow$ histogram methods are inefficient.

Kernel Density Estimators

- A generalization of sampling. Each sample distributes its weight over the space around it.
- Kernel function: describes the form of the weight distribution.
- The estimator is the sum of the kernel functions.



Kernel Density Estimators

- Used extensively in statistics for approximating a probability distribution given a set of examples.
- One dimensional Kernel Estimators for query approximation [Blohsfeld et al. '99].

Kernel Density Estimator Parameters

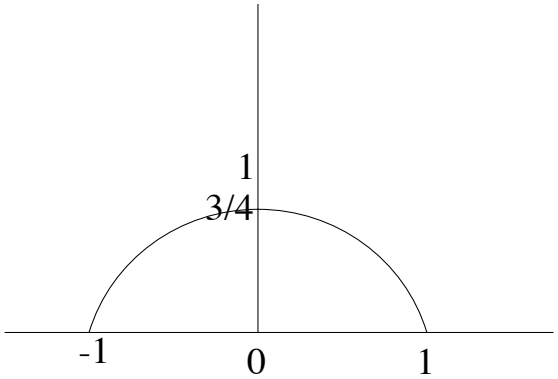
- What is the shape of the kernel function.
- What is the width of the kernel function.
- Where to place the kernel functions.
- How many kernel functions to use.

Epanechnikov Kernel Functions

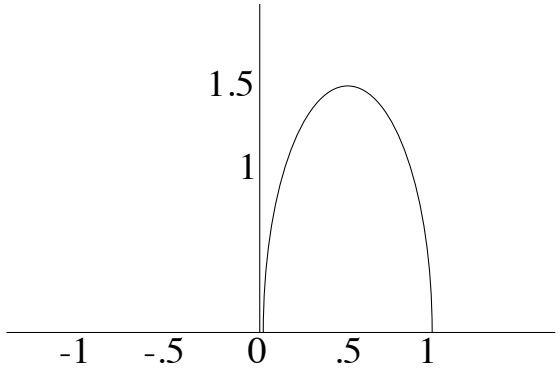
- d-dimensional Epanechnikov kernels:

$$k(x_1, \ldots, x_d) = (\frac{3}{4})^d \frac{1}{B_1 B_2 \ldots B_d} \prod_{1 \leq i \leq d} (1 - (\frac{x_i}{B_i})^2)$$

  if, for all i, $|\frac{x_i}{B_i}| < 1$, and 0 otherwise.



f(x) = 3/4 (1-x^2)



f(x) = 3/2 (1 - ((x-.5)/2)^2)

- Scott's rule to compute the bandwidth: $B_i = \sqrt{5}\ s_i\ |S|^{-\frac{1}{d+4}}$ where s_i is the standard deviation for the i-th dimension.

Computing the Kernel Estimator

- Scan the dataset and draw a random sample of size n $S = \{\bar{x_1}, ..., \bar{x_n}\}$.
- Compute the standard deviation for each dimension and using the Scott's rule set the bandwidths B_i for $i = 1, ..., d$.
- The approximation of the density function is:

$$F(\bar{x}) = \frac{1}{n} \sum_{\bar{x_i} \in S} k(\bar{x} - \bar{x_i})$$

Estimating the Selectivity of a Range Query

- Computing the selectivity of a range query is simple:

$$sel(F, [a_1, b_1] \times \ldots \times [a_d, b_d]) =$$

$$\frac{1}{n}(\frac{3}{4})^d \frac{1}{B_1 B_2 \ldots B_d} \sum_{1 \leq i \leq n} \int_{[a_1,b_1]} (1 - (\frac{x_1 - X_{i1}}{B_1})^2) \ldots$$

$$\int_{[a_d,b_d]} (1 - (\frac{x_d - X_{id}}{B_d})^2)\ dx_d \ldots dx_1$$

Kernel Estimator Properties

- **Efficient one pass construction**:

  The standard deviation for each dimension can be estimated in the sampling pass.

- **Efficient approximation**:

  Linear to the size of the estimator, and the dimensionality.

- **Accurate**.

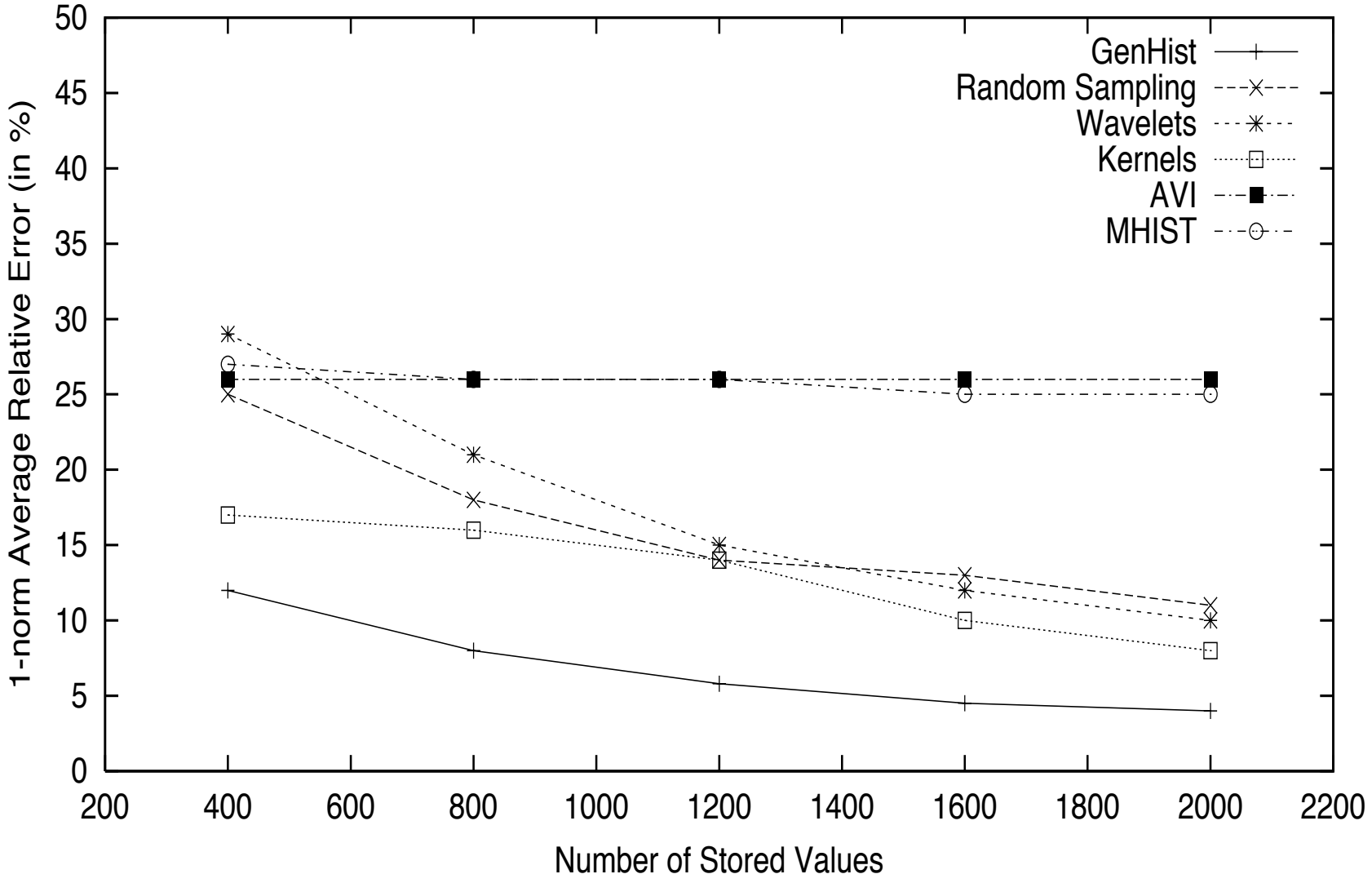
Previous Work

- Attribute Value Independence assumption.
- Random Sampling [Thompson '92].
- Multidimensional Histograms (MHIST) [Poosala and Ioannidis '97].
- Decomposition techniques (SVD, Wavelets, DCT) [Vitter and Wang '99, Lee et al. '99].
- One dimensional Kernel Estimators [Blohsfeld et al. '99].
- Density Estimation using clustering (EM) [Shanmugasundaram et al. '99].

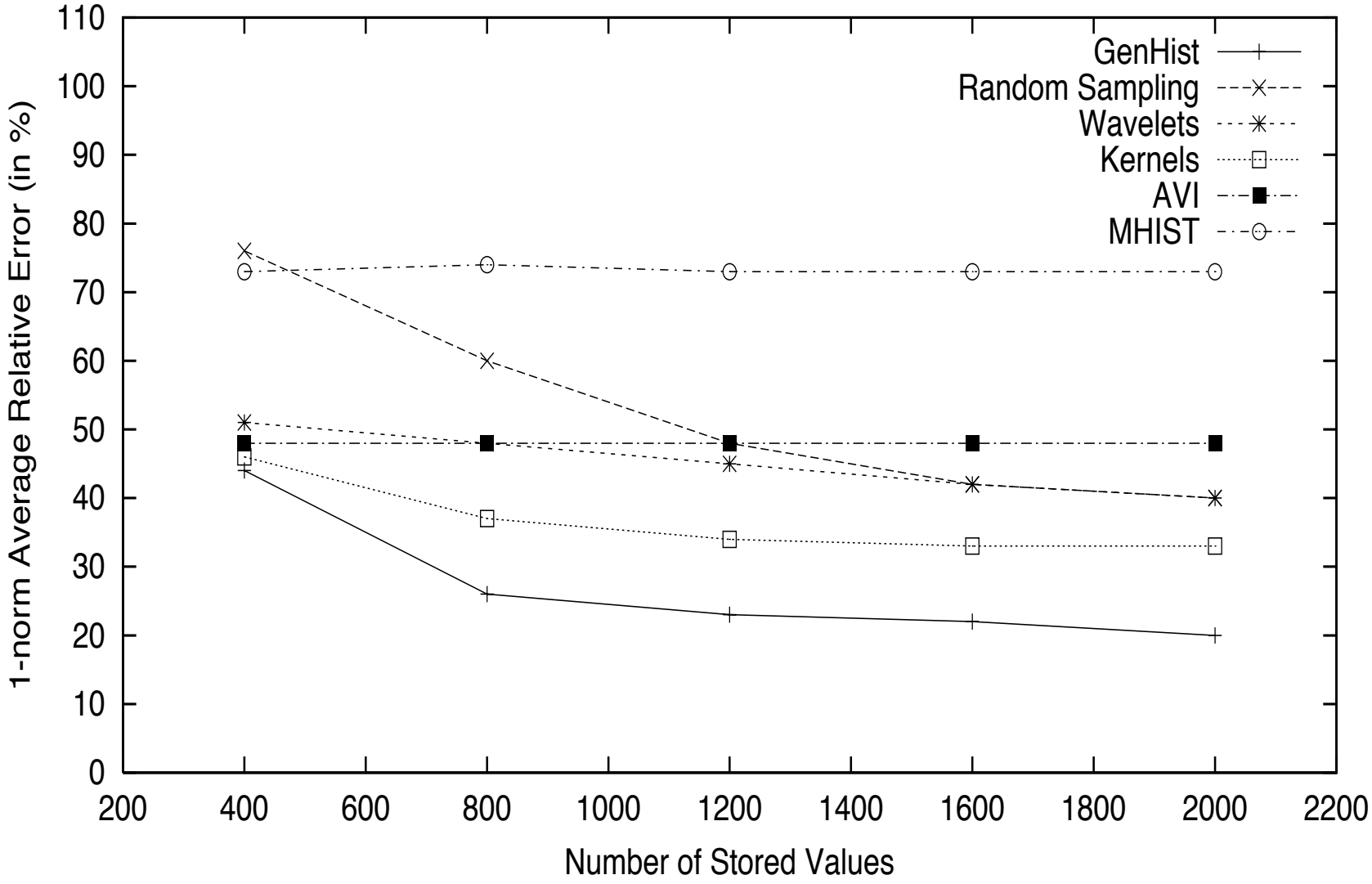
Experiments: Testbed

- **Synthetic Datasets**: TPC-D, DS1.
- **Real Datasets**: Forest Cover (FC), Multimedia.
- **Query types**: range queries with cardinality 1% and 10% and anchored queries.
- **Techniques**: AVI, GENHIST, Kernels, MHIST-2, Random Sampling, Wavelets.

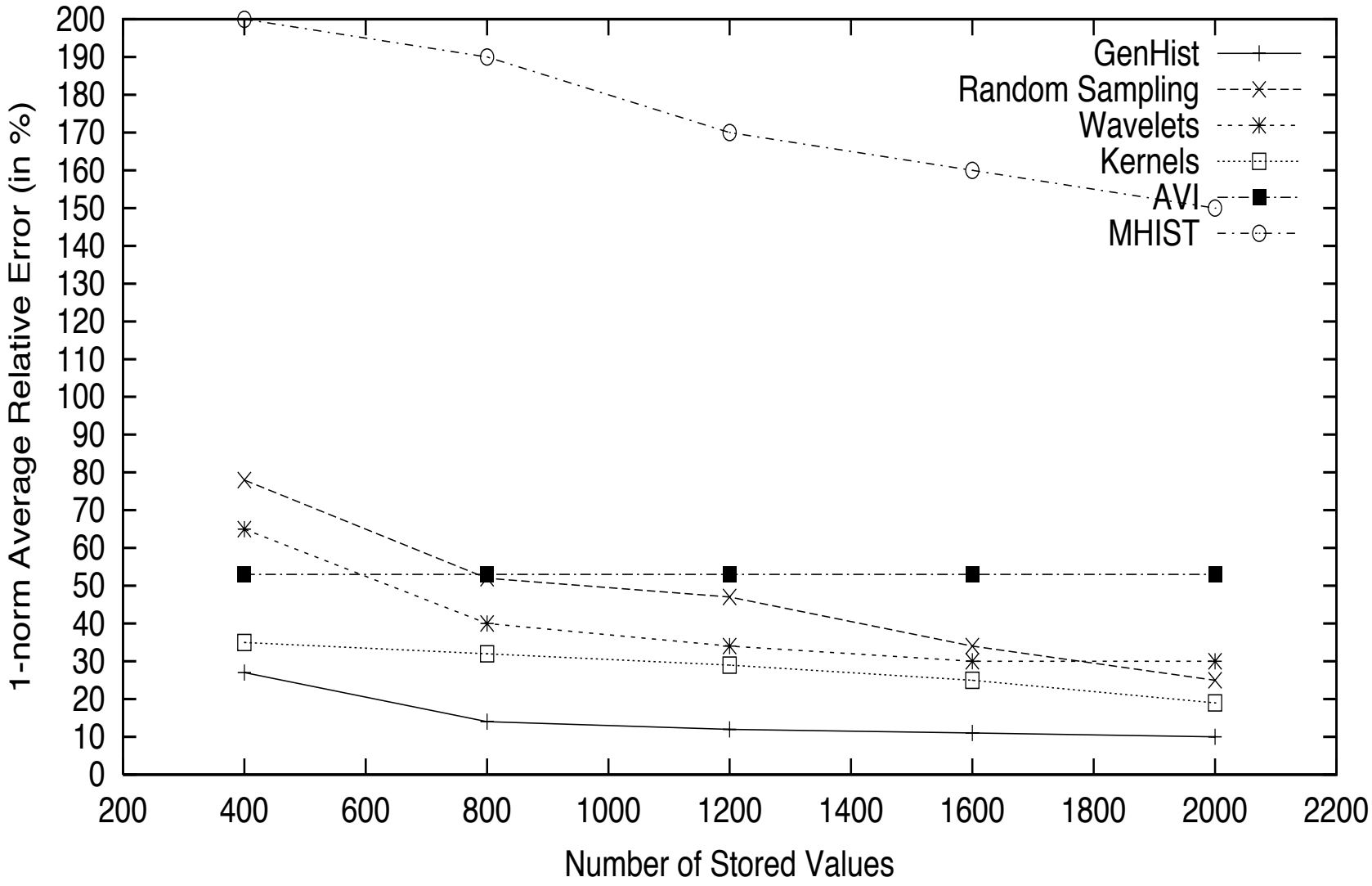
FC Dataset, 4-dim, 10% queries



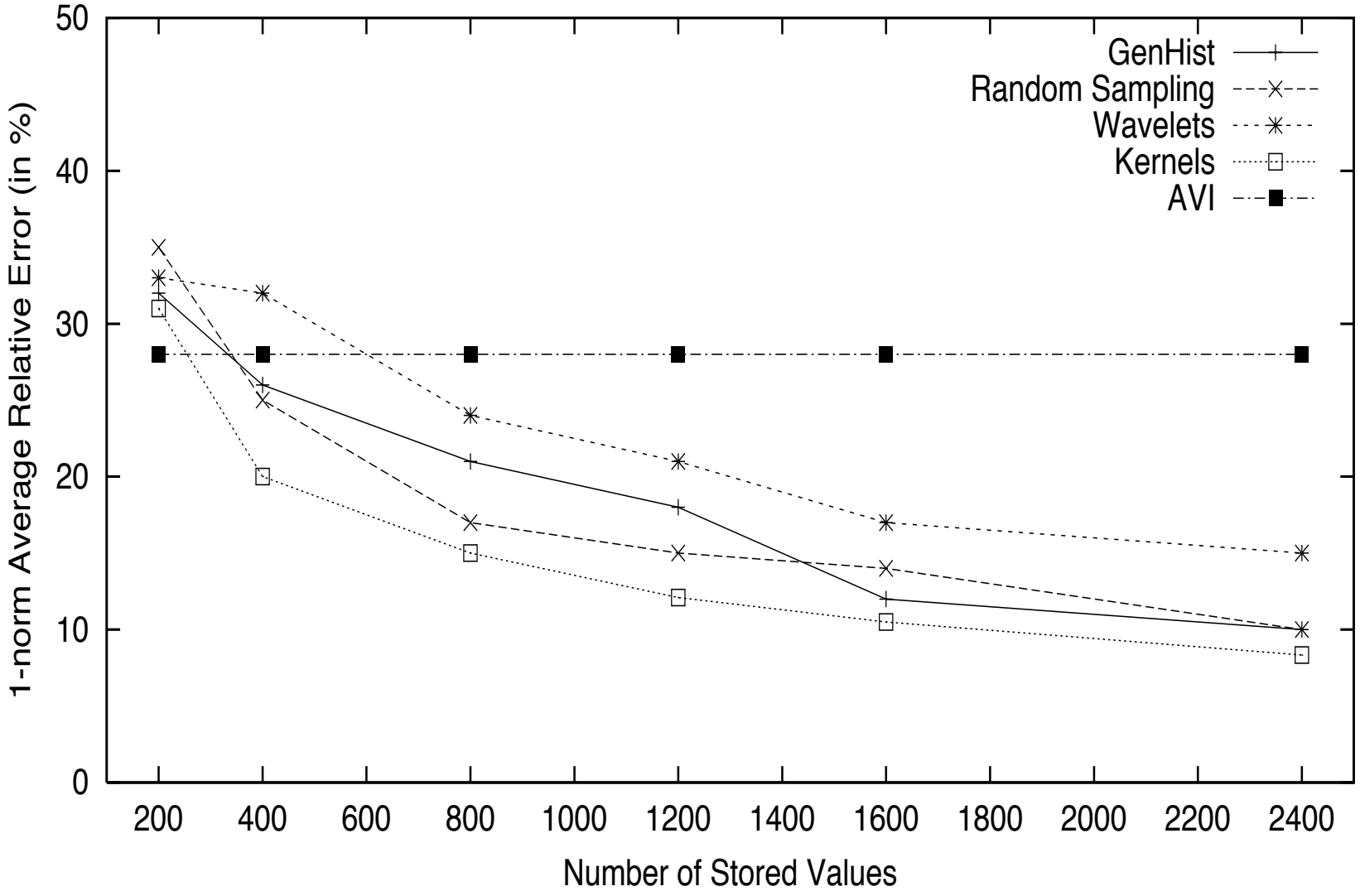
DS1 Dataset, 4-dim, anchored queries



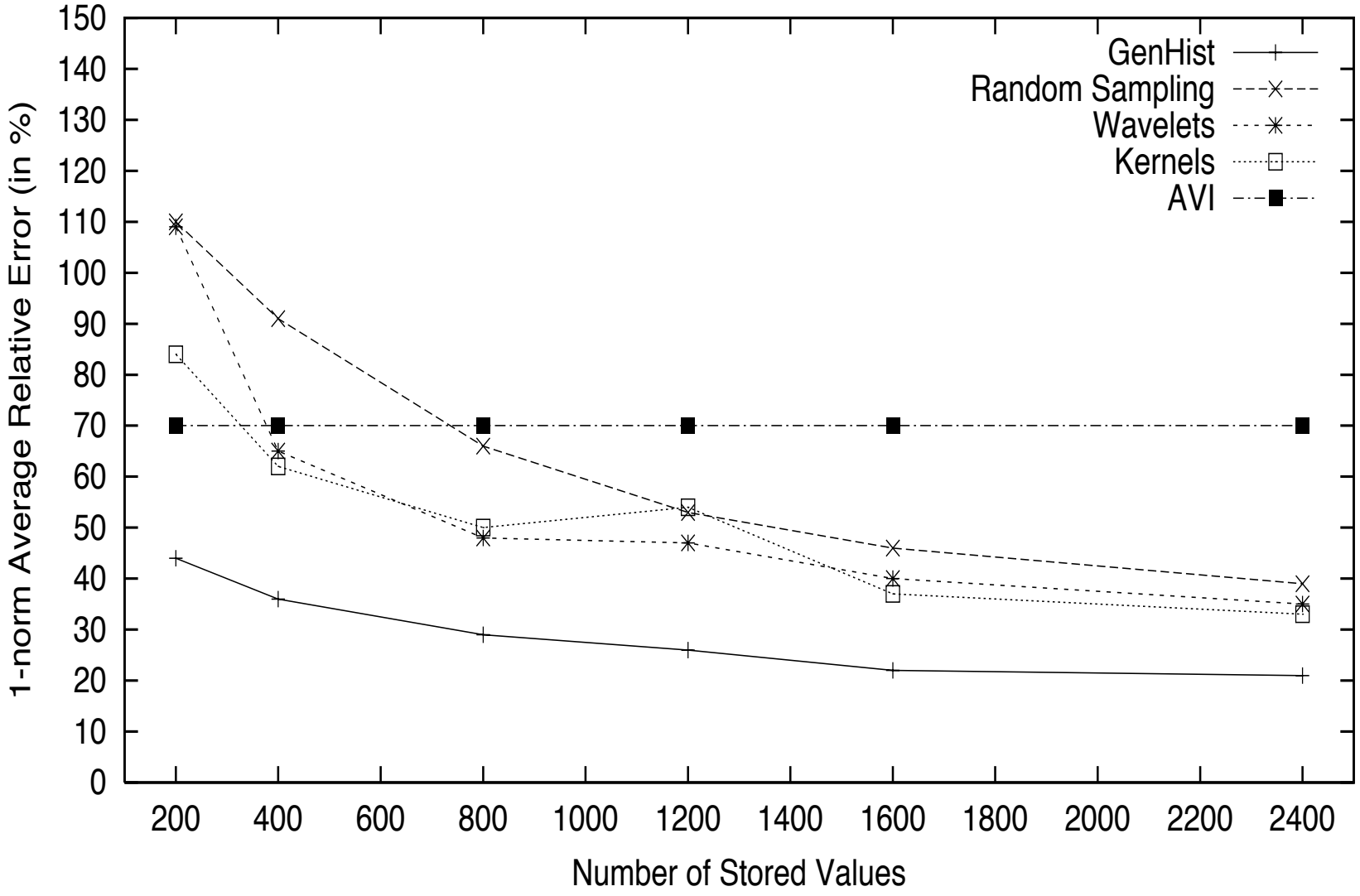
FC Dataset, 4-dim, anchored queries



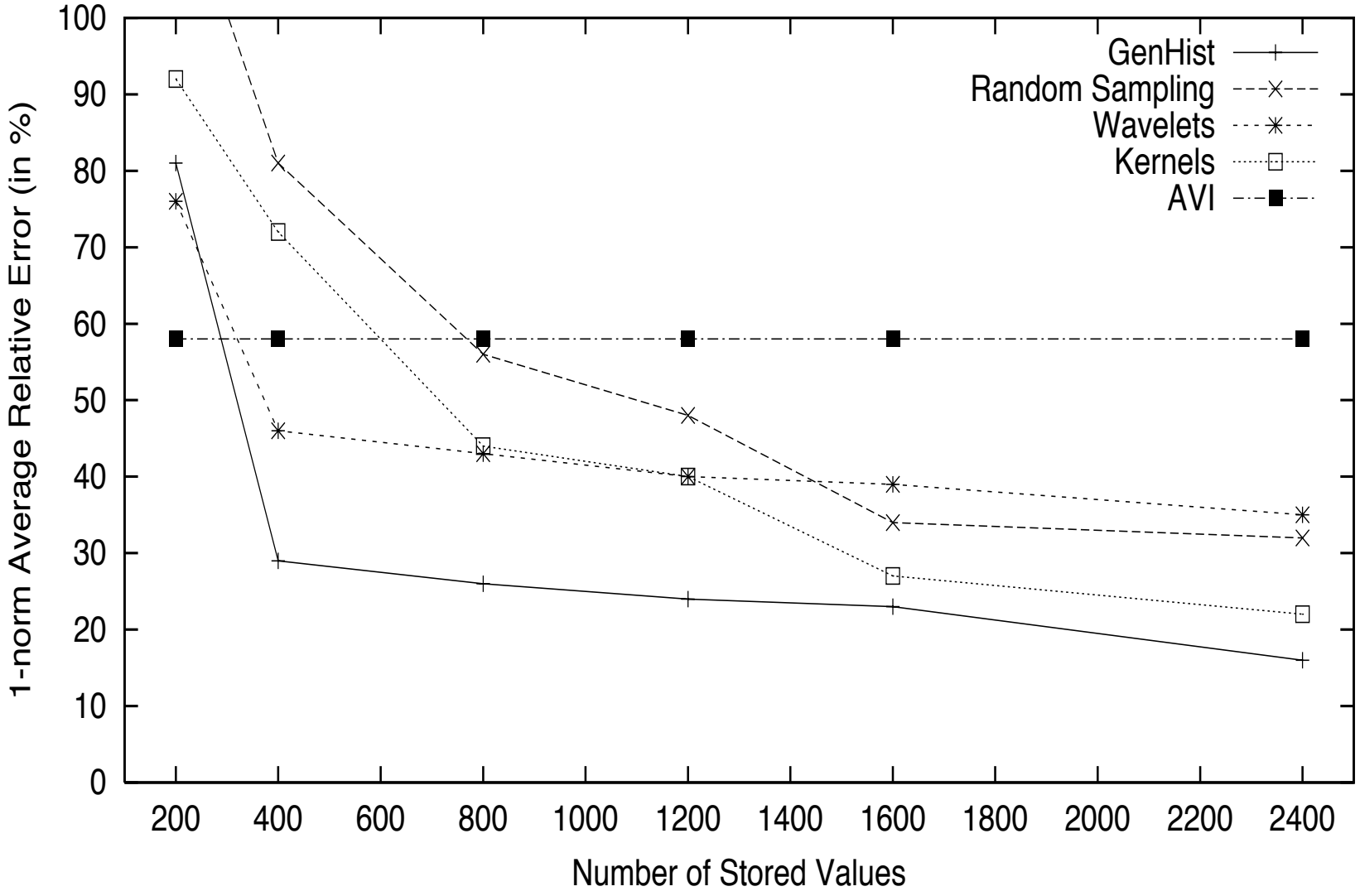
FC Dataset, 5-dim, 10% queries



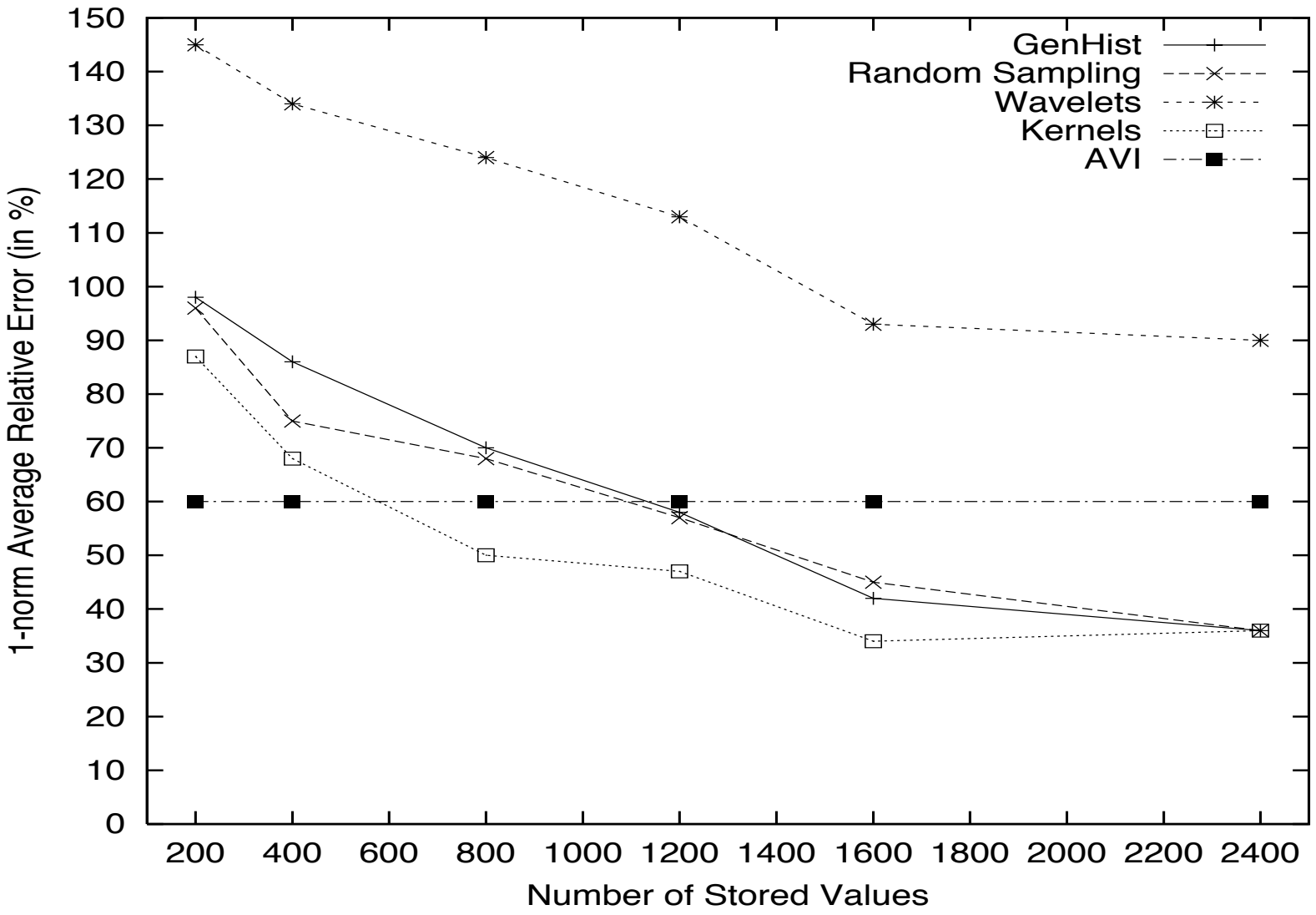
TPC-D Dataset, 5-dim, 1% queries



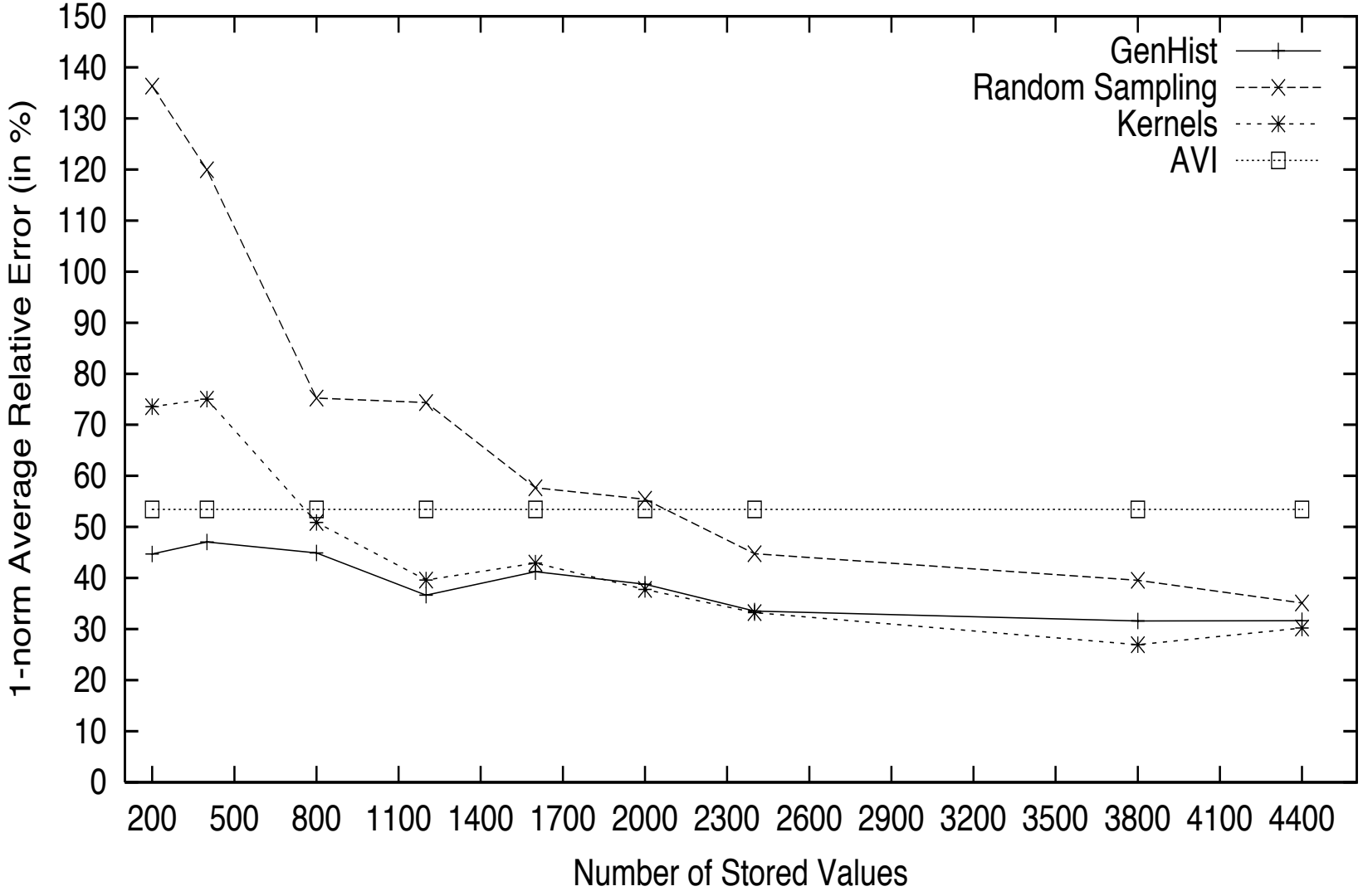
FC Dataset, 5-dim, 1% queries



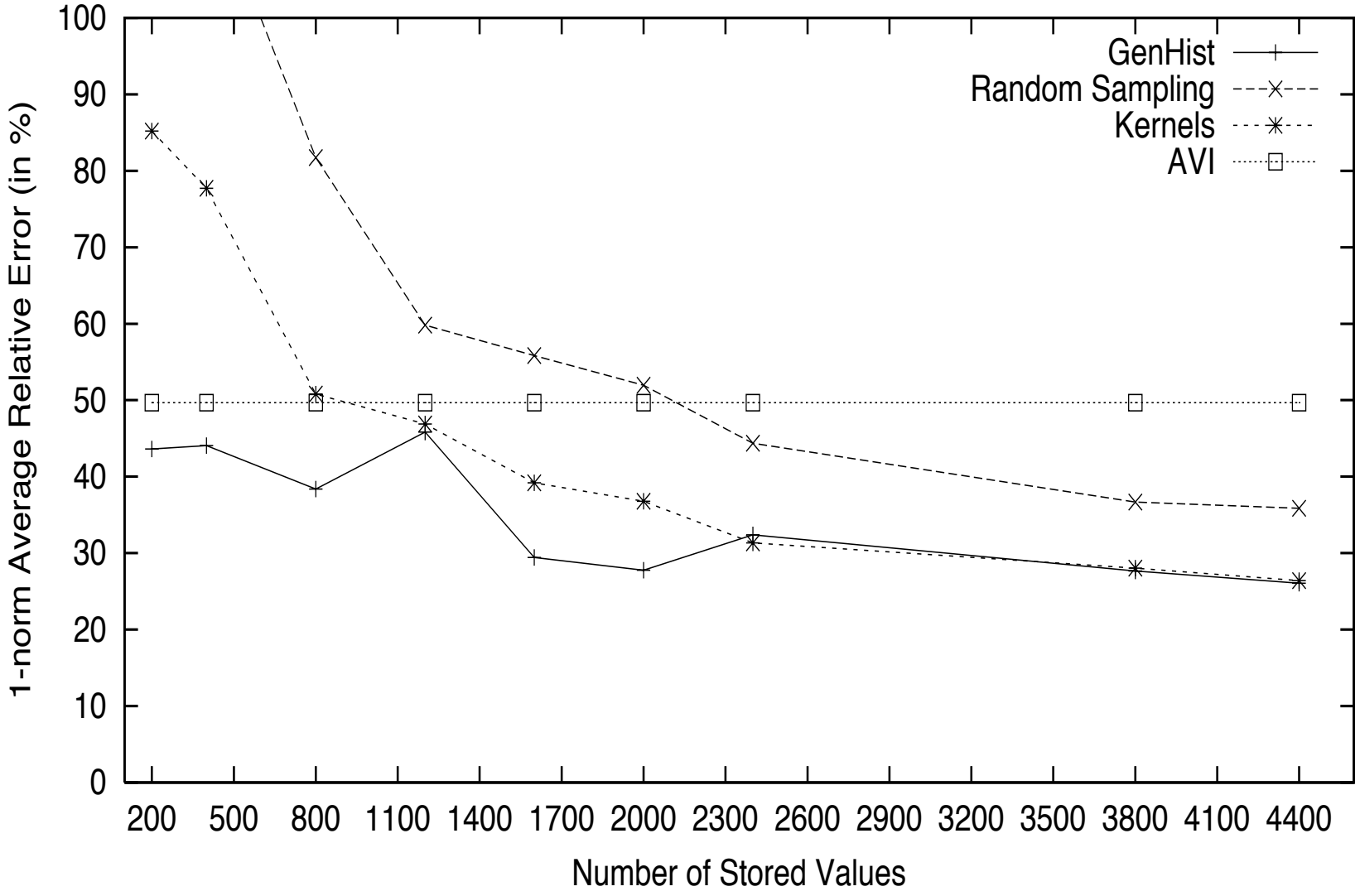
TPC-D Dataset, 5-dim, anchored queries



FC Dataset, 10-dim, 1% queries



FC Dataset, 10-dim, 1% queries in 8-dim



Experiments: Running Time

- 5-dim, 20K queries, 2K stored values, 1M points, running time in seconds.

Method	Construction Time	Estimation Time
Random Sampling	30	7
Kernels	31	35
GenHist	550	30
Wavelets	650	41

Remarks

- Kernel Estimators and GENHIST outperform the other techniques in most of the experiments.
- Kernel Estimators perform well (typically better than Random Sampling) and are inexpensive to compute.
- The accuracy of the techniques decreases quickly as the dimensionality increases.

Future Work

- Perform experiments in higher dimensions.
- Adaptive kernel estimators.