Dimension Reduction for Distance-Based Indexing

Rui Mao (Shenzhen University)
Willard L. Miranker (Yale University)
Daniel P. Miranker (Univ. of Texas at Austin)

Motivation

A theoretical framework for metric space indexing.

Outline

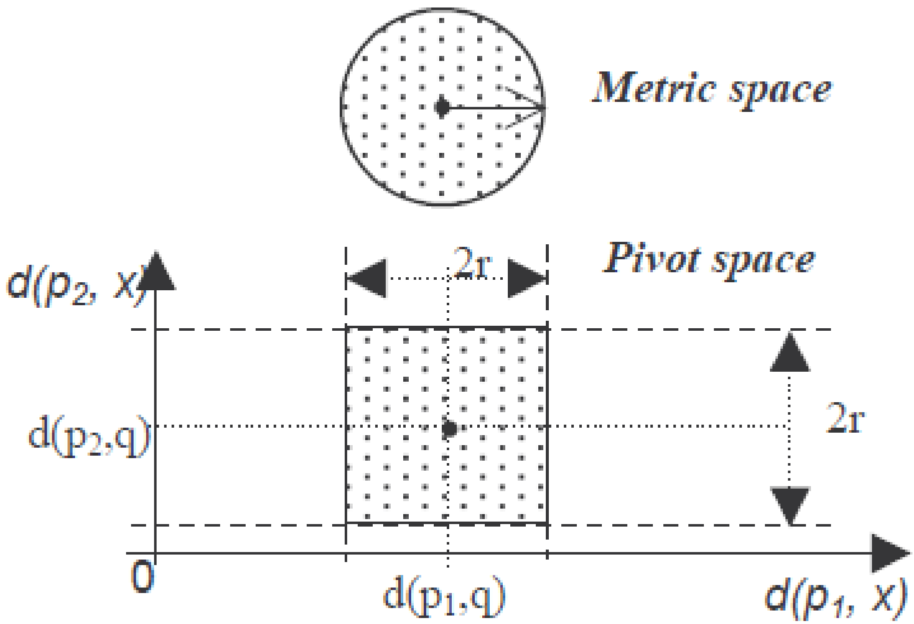
- Pivot space model
- Dimension reduction for distance-based indexing
- PCA for distance-based indexing
- Empirical results
- Conclusions and future work

Outline

- Pivot space model
  - General steps of distance-based indexing
  - Pivot space model
- Dimension reduction for distance-based indexing
- PCA for distance-based indexing
- Empirical results
- Conclusions and future work

General steps of Distance-based Indexing

1. metric space → R^k
2. multi-dimensional indexing → query cube
3. direct evaluation of cube



Pivot space model

- Pivot space F(S, P, d):
  - For data set S, pivot set P, and distance d:

$$F_{P,d}(S) = \{x_p \mid x_p = F_{P,d}(x) = (d(x,p_1), \ldots, d(x,p_k)), x \in S\}.$$

- Complete pivot space: P = S
- Theorem 1:

$$F(S, P, d) = F(\, F(S, \hat{P}, d), F(P, \hat{P}, d), L^{\infty})$$

  - Metric space → R^n

Outline

- Pivot space model
- Dimension reduction for distance-based indexing
  1. answer queries directly in the complete pivot space?
  2. dimension reduction for the complete pivot space?
  3. why is pivot selection important?
  4. how to select pivots?
- PCA in distance-based indexing
- Empirical results
- Conclusions and future work

1. Answer queries directly in the complete pivot space?

Theorem 2: Evaluation of similarity queries in the complete pivot space degrades the query performance to linear scan.

- Dimension reduction is inevitable

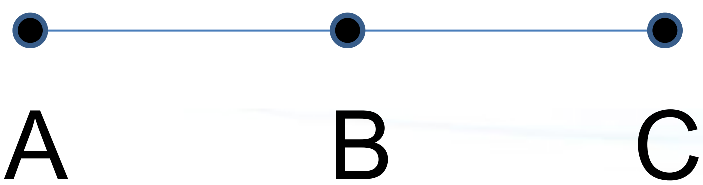
2. Dimension reduction for the complete pivot space?

Theorem 3: If a dimension reduction technique creates new dimensions based on all existing dimensions, evaluation of similarity queries degrades to a linear scan

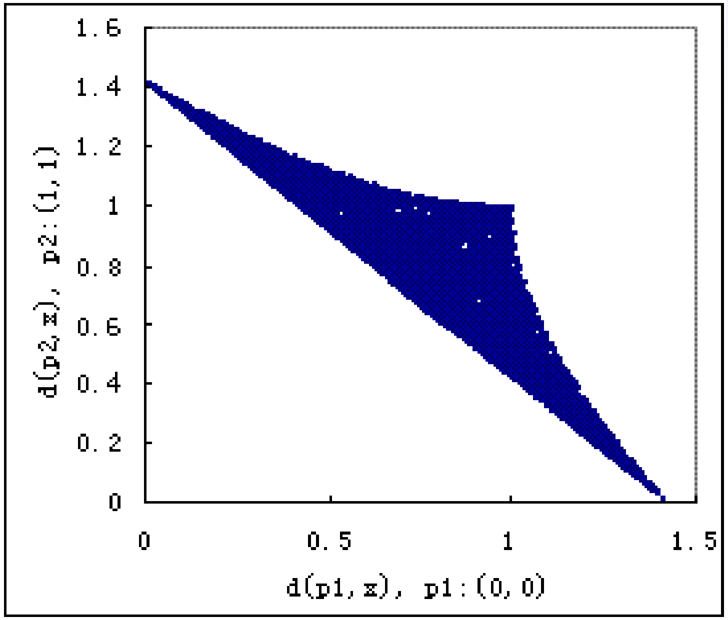
- Pivot selection: select only existing dimensions
- Metric space indexing vs. high dimensional indexing

3. Why is pivot selection important?

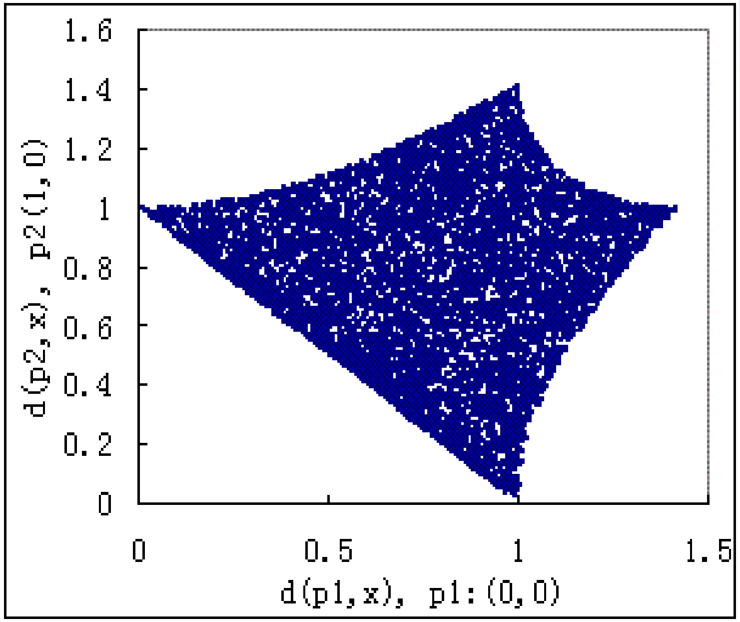
- Building index tree: a process of information loss
  - Information available to data partition is determined by pivot selection



Example: 2-d pivot space



Pivots: opposite corners (0,0) and (1,1)



Pivots: neighboring corners (0,0) and (1,0)

4. How to select pivots?

- **Heuristic**: for each new dimension, select the point with the largest projection on that new dimension in the pivot space.
  - Using of mathematical tools in R^n
  - Yet what is a good objective function for pivot selection?

Outline

- Pivot space model
- Dimension reduction for distance-based indexing
- PCA for distance-based indexing
  - Pivot selection
  - Estimate the intrinsic dimension
- Empirical results
- Conclusions and future work

PCA for pivot selection

- PCA for the complete pivot space.
- Apply the heuristic: for each PC, find the most similar dimension(point) in the complete pivot space
- Start with corners (farthest first traversal) as candidates

Estimate the intrinsic dimension

1. Pair wise distances

$$\rho = \mu^2/2\sigma^2$$

2. $|\text{Range}(q,r)| \sim r^d$
   - Linear regression: $\log(|\text{Range}(q,r)|)$ and $\log(r)$

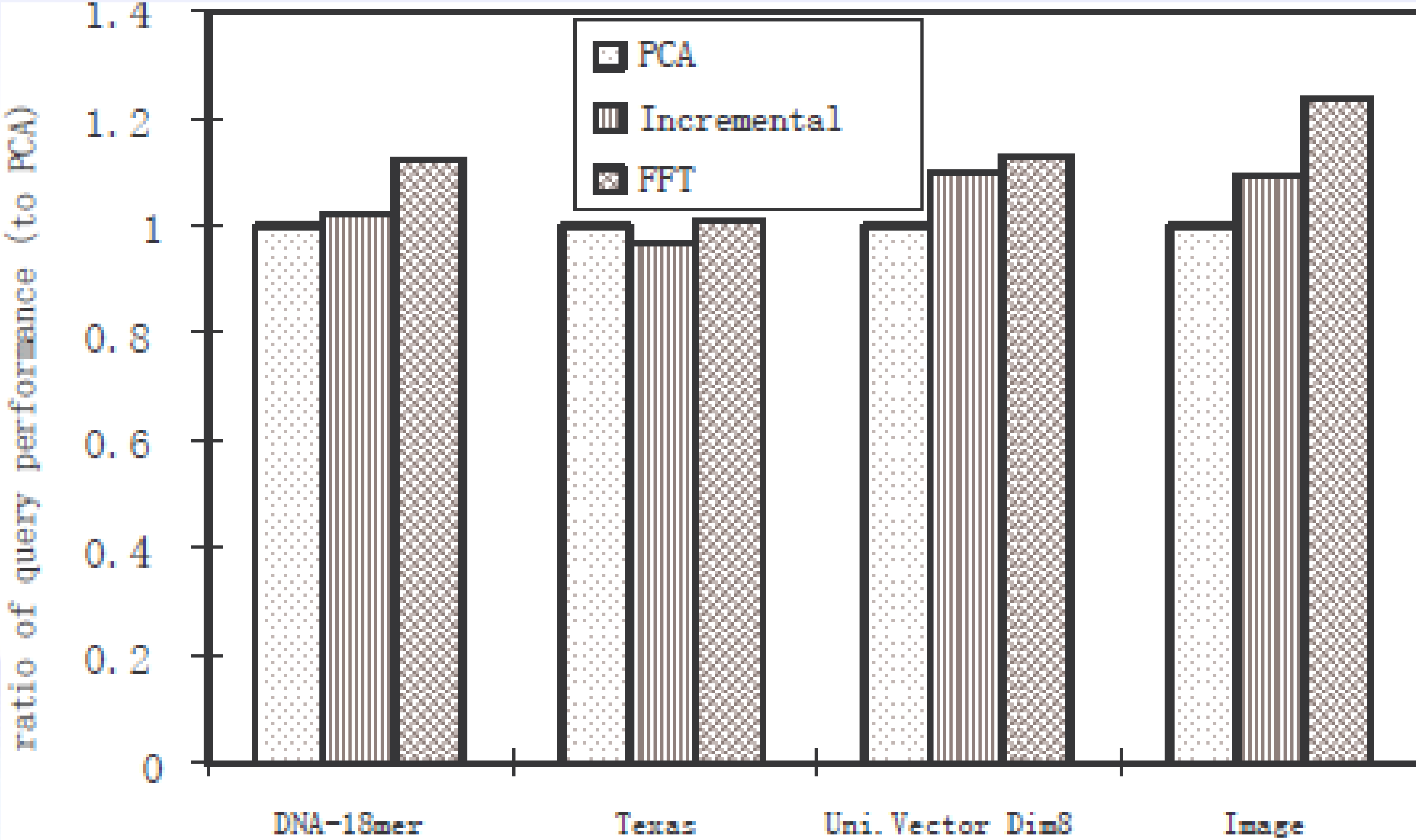
3. Where eigenvalue changes the most:
   - $\text{argmax}_i (\lambda_i / \lambda_{i+1})$, $0.015 \le \lambda_{i+1} \le 0.035$, $\sum_{j=1}^{i} \lambda_j > 0.6$

- Yet how to define the intrinsic dimension?

Outline

- Pivot space model
- Dimension reduction for distance-based indexing
- PCA for distance-based indexing
- Empirical results
  - Query performance
  - intrinsic dimension
- Conclusions and future work

Query performance



Intrinsic dimension

Workload	Domain dimension	Distance oracle	Intrinsic dimension		
			$\mu^2/2\sigma^2$	regression	$\text{argmax}_i(\lambda_i/\lambda_{i+1})$
Vector (uniform)	D=1-20	L^∞	1.72d - 1.81	0.73d + 0.88	d+1 (d≠3,4), 4, 7 (d=3, 4)
		L^1	d	0.75d + 0.84	d+1
		L^2	1.41d - 0.71	0.78d - 0.72	d+1
Vector (exponential)	D = 1-10	L^∞	0.244d + 0.446	0.676d + 0.62	d+1
		L^1	0.499d - 0.0006	0.737d + 0.482	d+1
		L^2	0.427d + 0.113	0.72d + 0.534	d+1
Vector (normal)	D = 1-10	L^∞	0.644d + 0.559	0.858d + 0.325	d+1
		L^1	0.875d + 0.002	0.863d + 0.32	d+1
		L^2	0.989d - 0.145	0.872d + 0.305	d+1
Texas	2	L^∞ / L^1 / L^2	1.29 / 1.42 / 0.87	1.54 / 1.54 / 1.51	3
Hawaii	2	L^∞ / L^1 / L^2	0.31 / 0.26 / 0.36	1.47 / 1.45 / 1.44	2
Protein q-gram	q = 6-18	Weighted edit distance	2.46q + 2.32	-0.08q + 4.16	q+1 (q<18), 17 (q=18)
DNA q-gram	q = 9-18	Hamming distance	1.27q + 0.37	0.14q + 2.52	q+1 (q<18), 21 (q=18)
Mass-spectra	40,000	Fuzzy cosine distance	0.62	1.23	2
Image	66	Linear combination of L-norms	5.26	4.85	5

Outline

- Pivot space model
- Dimension reduction for distance-based indexing
- PCA for distance-based indexing
- Empirical results
- Conclusions and future work

Conclusions and future work

- Established a parallel between metric space indexing and high dimensional indexing
- More mathematical tools for pivot selection?
- Objective function of pivot selection?
- Pivot space model for data partition?
- Intrinsic dimension?
- Optimal num of pivots vs. intrinsic dimension?

Thank you!

Acknowledgement:

- Gonzalo Navarro
- Glen Nuckolls
- Piotr Indyk