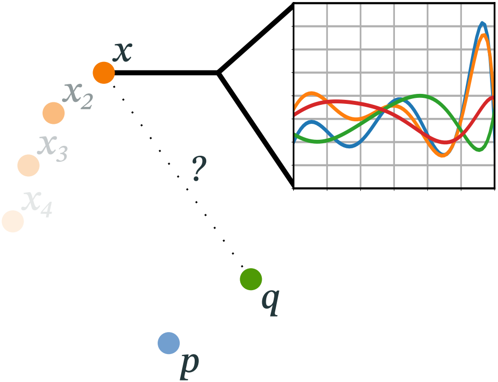
Towards Ptolemaic metric properties of the z-normalized Euclidean distance for MulTiS indexing

Max Pernklau, Christian Beecks

May 13, 2024

Department of Mathematics and Computer Science, University in Hagen, Germany

Motivation: Comparing Time Series



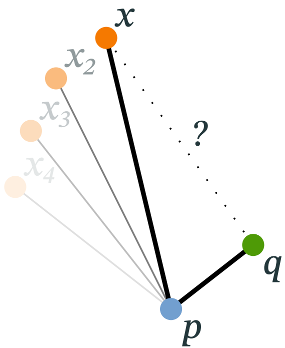
Compare time series using a (dis-)similarity measure d

Find time series similar to example q in database $\mathbb{D}$

$$d : \mathbb{D} \times \mathbb{D} \to \mathbb{R}^+$$

$$\{x \in \mathbb{D} \mid d(q, x) \leq r\}$$

Motivation: Comparing Time Series Efficiently

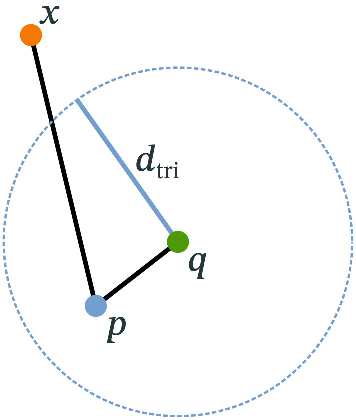


Cache distances to special pivot points $d(p, x_i)$ beforehand

Calculate $d(p, q)$ online

Determine $d(q, x) \leq r$

Motivation: Metric Index



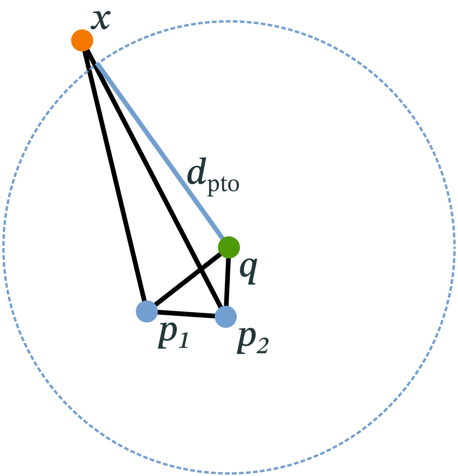
If $(\mathbb{D}, d)$ is a metric space:

Use the triangle inequality to access index

$$d(q,x) \geq d_{\text{tri}} = |d(p,x) - d(p,q)|$$

exclude if $d_{\text{tri}} > r$

Ptolemaic Index



If $(\mathbb{D}, d)$ is also Ptolemaic:

Even better approximation!

$$d(q,x) \geq d_{\text{pto}} = \frac{|\overline{xp_1} \cdot \overline{qp_2} - \overline{qp_1} \cdot \overline{xp_2}|}{\overline{p_1p_2}}$$

Objective

Find a distance function for MulTiS that is

metric and Ptolemaic to speed up range queries!

Structure

- Univariate z-normalized Euclidean distance
- Multivariate z-normalized Euclidean distance
- Indexability

Univariate z-normalized Euclidean Distance

aka Matrix Profile distance

$$d_z(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$$

$$d_z(x, y) = \sqrt{\sum_i^n (\nu(x_i) - \nu(y_i))^2}$$

$$\text{with} \quad \nu : x_i \mapsto \frac{x_i - \mathrm{E}(x)}{\sqrt{\mathrm{Var}(x)}}$$

Univariate z-normalized Euclidean Distance

aka Matrix Profile distance

$$d_z(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$$

$$d_z(x, y) = \sqrt{\sum_i^n (\nu(x_i) - \nu(y_i))^2}$$

$$\text{with} \quad \nu : x_i \mapsto \frac{x_i - \mathrm{E}(x)}{\sqrt{\mathrm{Var}(x)}}$$

$$\iff d_z(x, y) = \sqrt{2n}\sqrt{1 - \mathrm{Corr}(x, y)}$$

Homomorphic to Euclidean Metric

$$
\begin{aligned}
d_z(x, y) = \quad & d_{\text{eucl}}(\nu(x), \nu(y)) = \quad & (d_{\text{eucl}} \circ \nu)(x, y) \\
\Rightarrow (\mathbb{R}^n, d_z) = \quad & (\mathbb{R}^n, d_{\text{eucl}} \circ \nu) \sim \quad & (\nu(\mathbb{R}^n), d_{\text{eucl}})
\end{aligned}
$$

but $\nu(\mathbb{R}^n) \subset \mathbb{R}^n$, so $(\mathbb{R}^n, d_z) \subset (\mathbb{R}^n, d_{\text{eucl}})$

$\Rightarrow (\mathbb{R}^n, d_z)$ is a Ptolemaic pseudometric

Multivariate Extension

$$d_m(X,Y) : (\mathbb{R}^{n\times m}) \times (\mathbb{R}^{n\times m}) \to \mathbb{R}^+$$

$$d_m(X,Y) = \sqrt{\sum_i^k (d_z(X_i, Y_i))^2} = ||\vec{d}_z(X,Y)||$$

Properties of the Multivariate Extension

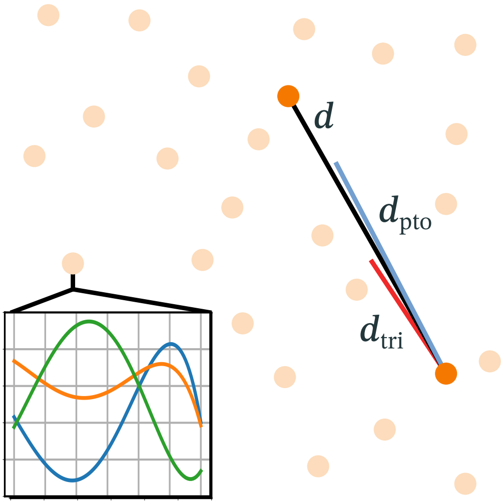
is a pseudometric

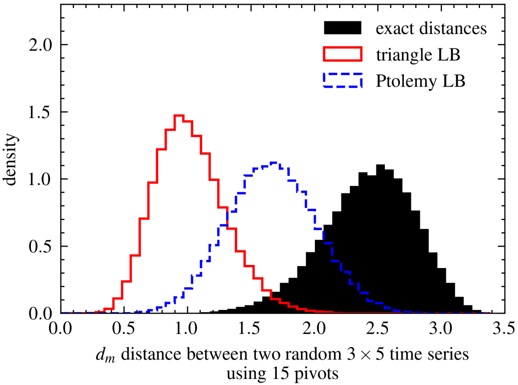
Use norm subadditivity ($||X+Y|| \leq ||X|| + ||Y||$)

is probably Ptolemaic

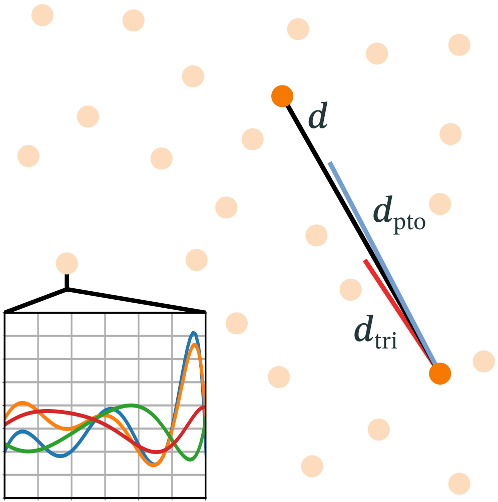
only numerical evidence yet

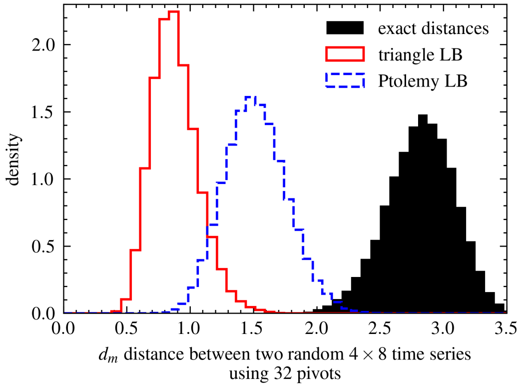
Indexing Improvements: Low Dimensionality





Indexing Improvements: High Dimensionality





Conclusion

Summary

- The multivariate z-normalized Euclidean distance is a Ptolemaic pseudometric
- Ptolemy's inequality improves indexability of such metric spaces

Future Work

- Analytical proof
- Integrate into indexing structures, test on real data
- Pivot selection

Backup Slides

Experiment: Counterexample Search

- generate $5 \cdot 10^7$ MulTiS quadruplets
- with $X \in \mathbb{R}^{4\times5}$ (4 observables, 5 time slices)
- draw each element of the matrix from a uniform distribution
    1. $\{X_{ij} \in \mathbb{Z} \mid -20 \leq X_{ij} \leq 20\}$
    2. $\{X_{ij} \in \mathbb{R} \mid -20 \leq X_{ij} \leq 20\}$

Experiment: Indexability

- generate random points $\{X_{ij} \in \mathbb{R} \mid -20 \leq X_{ij} \leq 20\}$
- for each measured distance:
  - choose 15 (32) random pivots
  - only use the pivots that generate the highest lower bounds
  - calculate actual distance, best triangle lower bound, best Ptolemaic lower bounds

Metric Space

metric space (M, d) with time series $x, y, z \in M$

$d(x, y) = 0 \iff x = y$ (identity of indiscernibles)

$d(x, y) = d(y, x)$ (symmetry)

$d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

The “fourth” metric axiom

Do we need $d(x, y) > 0$?

$$0 = d(x, x) \leq d(x, y) + d(y, x) \qquad \text{(triangle inequality)}$$

$$0 = d(x, x) \leq 2d(x, y) \qquad \text{(symmetry)}$$

Ptolemaic Spaces

quadratic form distance $\sqrt{(x-y)^T A(x-y)}$

Jensen–Shannon distance

$$\sqrt{H(P/2+Q/2)-H(P)/2-H(Q)/2}$$

triangular distance $\sqrt{\sum_i \frac{(P_i-Q_i)^2}{P_i+Q_i}}$

cosine distance $\sqrt{1-\cos\measuredangle XY}$

$(M,\sqrt{d})$ for any metric space (M,d)