Spatial Range Querying for Gaussian-Based Imprecise Query Objects

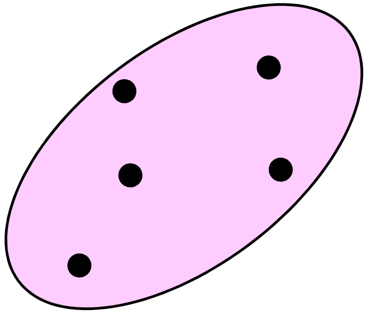
Yoshiharu Ishikawa, Yuichi Iijima
Nagoya University
Jeffrey Xu Yu
The Chinese University of Hong Kong

Outline

- Background and Problem Formulation
- Related Work
- Query Processing Strategies
- Experimental Results
- Conclusions

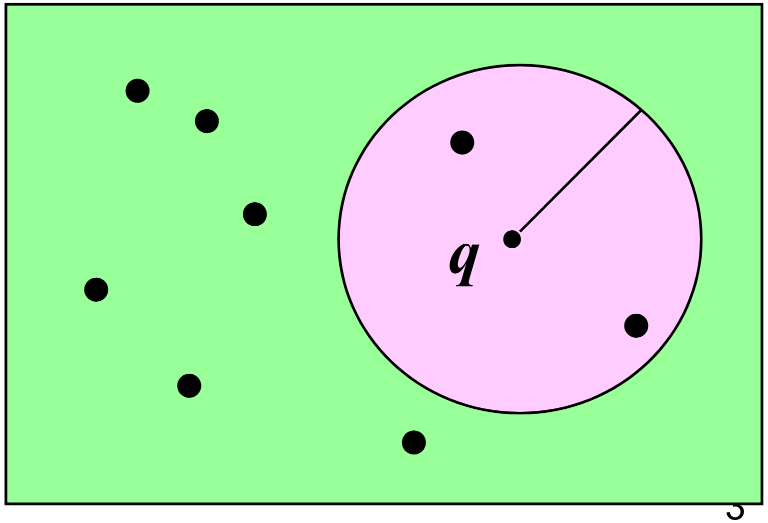
Imprecise Location Information

- Sensor Environments
  - Frequent updates may not be possible
    - GPS-based positioning consumes batteries
- Robotics
  - Localization using sensing and movement histories
  - Probabilistic approach has vagueness
- Privacy
  - Location Anonymity



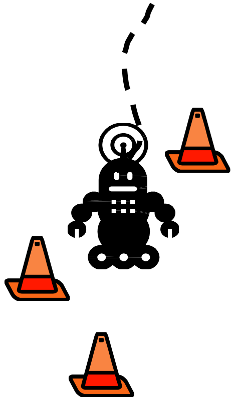
Location-based Range Queries

- Location-based Range Queries
  - Example: Find hotels located within 2 km from Yuyuan Garden
  - Traditional problem in spatial databases
    - Efficient query processing using spatial indices
    - Extensible to multi-dimensional cases (e.g., image retrieval)
- What happen if the location of query object is uncertain?



Probabilistic Range Query (PRQ) (1)

- Assumptions
  - Location of query object q is specified as a Gaussian distribution
  - Target data: static points



- Gaussian Distribution

$$p_q(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{x}-\boldsymbol{q})^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{q}) \right]$$

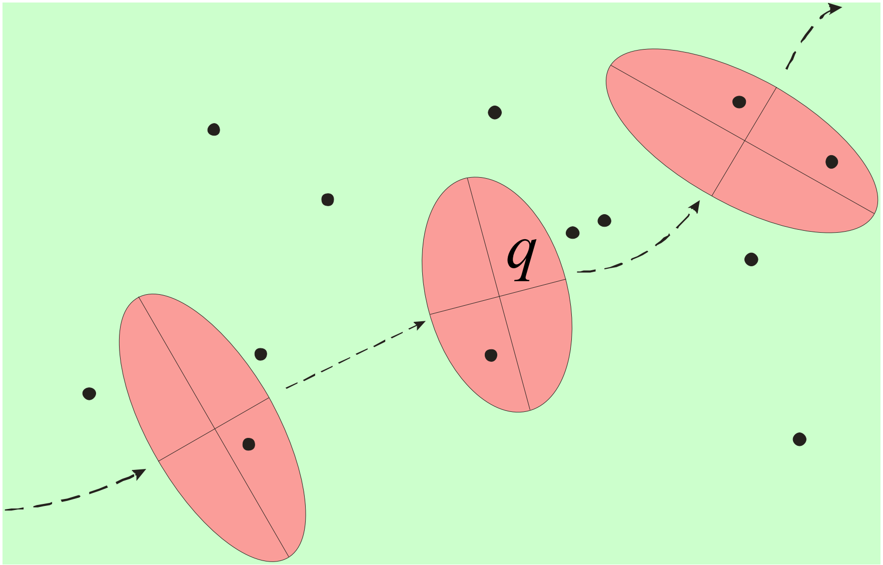
  - $\boldsymbol{\Sigma}$: Covariance matrix

Probabilistic Range Query (PRQ) (2)

- Probabilistic Range Query (PRQ)

$$PRQ(q,\ \delta,\ \theta) = \{o \mid o \in O,\ \Pr(\|\boldsymbol{x} - \boldsymbol{o}\|^2 \leq \delta^2) \geq \theta\}$$

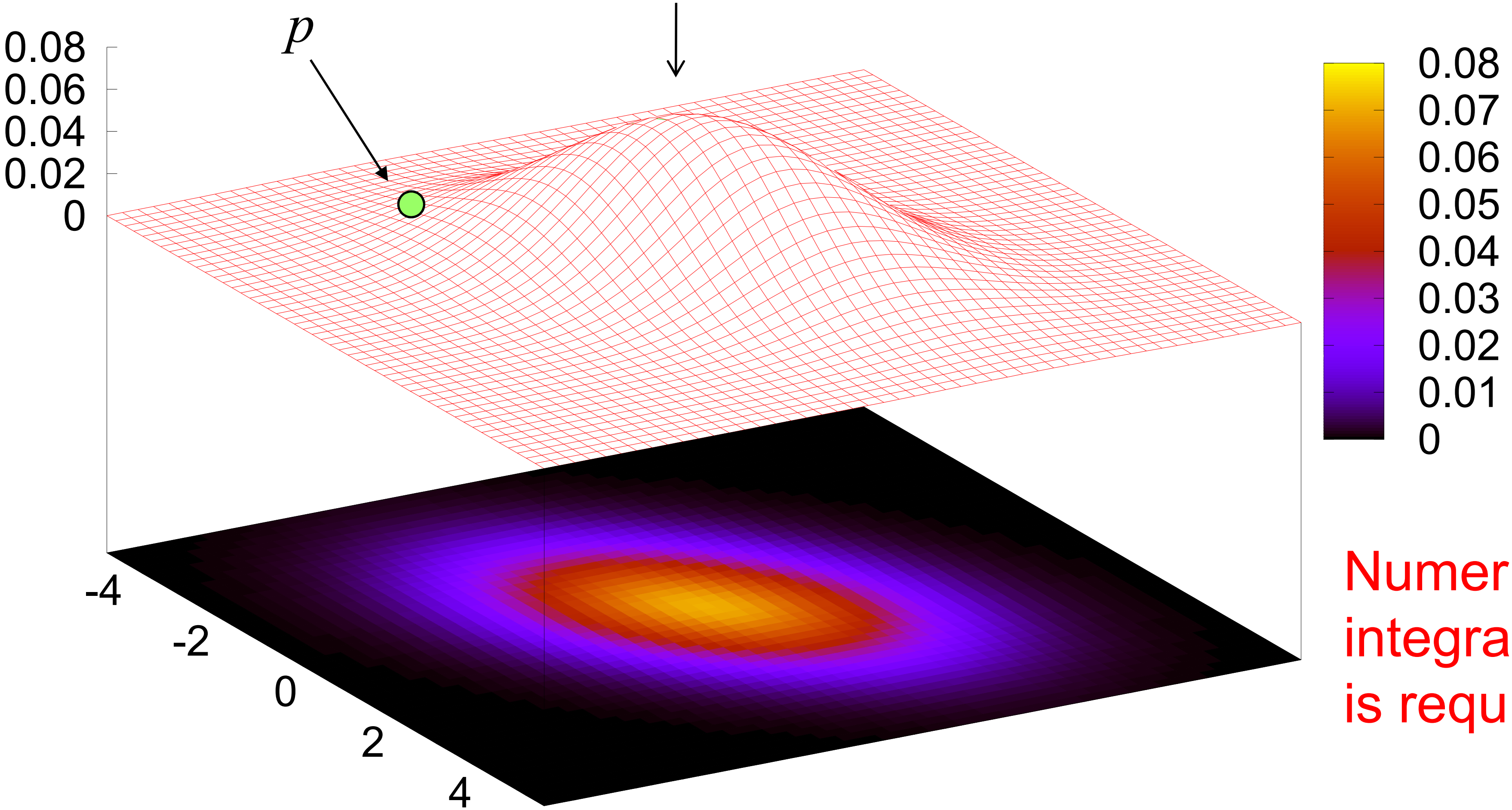
- Find objects such that the probabilities that their distances from q are less than δ are greater than θ



Probabilistic Range Query (PRQ) (3)

- Is distance between q and p within δ?

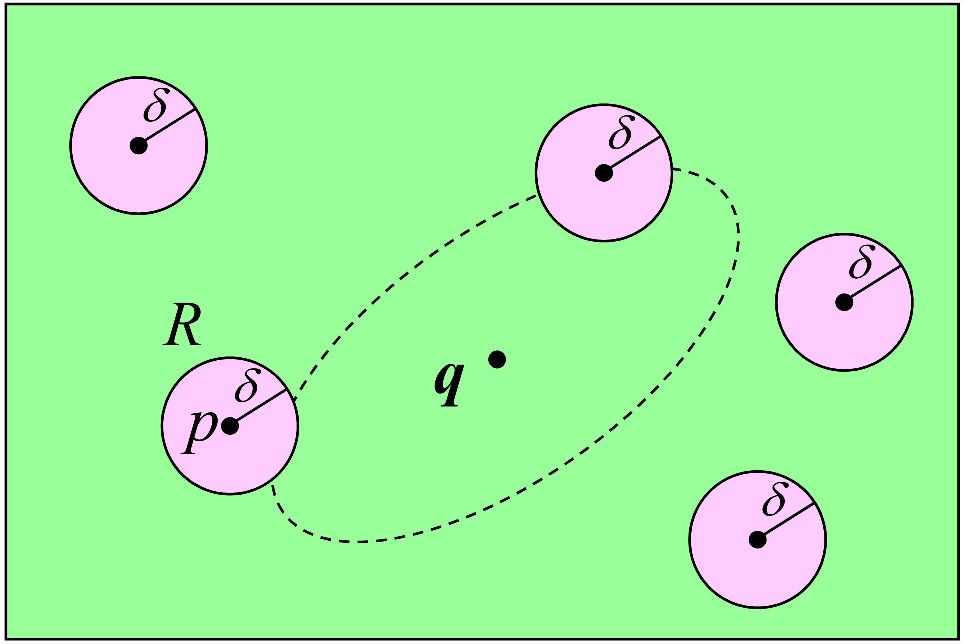
pdf of q (Gaussian distribution)



Numerical integraiton is required

Naïve Approach for Query Processing

- Exchanging roles
  - Pr[p is within δ from q] = Pr[q is within δ from p]
- Naïve approach
  - For each object p, integrate pdf for sphere region R
  - R : sphere with center p and radius δ
  - If the result $\geq \theta$, it is qualified
- Quite costly!



Outline

- Background and Problem Formulation
- Related Work
- Query Processing Strategies
- Experimental Results
- Conclusions

Related Work

- Query processing methods for uncertain (location) data
  - Cheng, Prabhakar, et al. (SIGMOD’03, VLDB’04, …)
  - Tao et al. (VLDB’05, TODS’07)
  - Parker, Subrahmanian, et al. (TKDE’07, ‘09)
  - Consider arbitrary PDFs or uniform PDFs
  - Target objects may be uncertain
- Research related to Gaussian distribution
  - Gauss-tree [Böhm et al., ICDE’06]
  - Target objects are based on Gaussian distributions

Outline

- Background and Problem Formulation
- Related Work
- Query Processing Strategies
- Experimental Results
- Conclusions

Outline of Query Processing

- Generic query processing strategy consists of three phases
  1. Index-Based Search: Retrieve all candidate objects using spatial index (R-tree)
  2. Filtering: Using several conditions, some candidates are pruned
  3. Probability Computation: Perform numerical integration (Monte Carlo method) to evaluate exact probability
- Phase 3 dominates processing cost
  – Filtering (phase 2) is important for efficiency

Query Processing Strategies

- Three strategies
  1. Rectilinear-Region-Based Approach (RR)
  2. Oblique-Region-Based Approach (OR)
  3. Bounding-Function-Based Approach (BF)
- Combination of strategies is also possible

Rectilinear-Region-Based (RR) (1)

- Use the concept of θ-region
  - Similar concepts are used in query processing for uncertain spatial databases
- θ-region: Ellipsoidal region for which the result of the integration becomes $1-2\theta$:

$$\int_{(\boldsymbol{x}-\boldsymbol{q})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{q}) \le r_\theta^2} p_q(\boldsymbol{x}) d\boldsymbol{x} = 1 - 2\theta$$

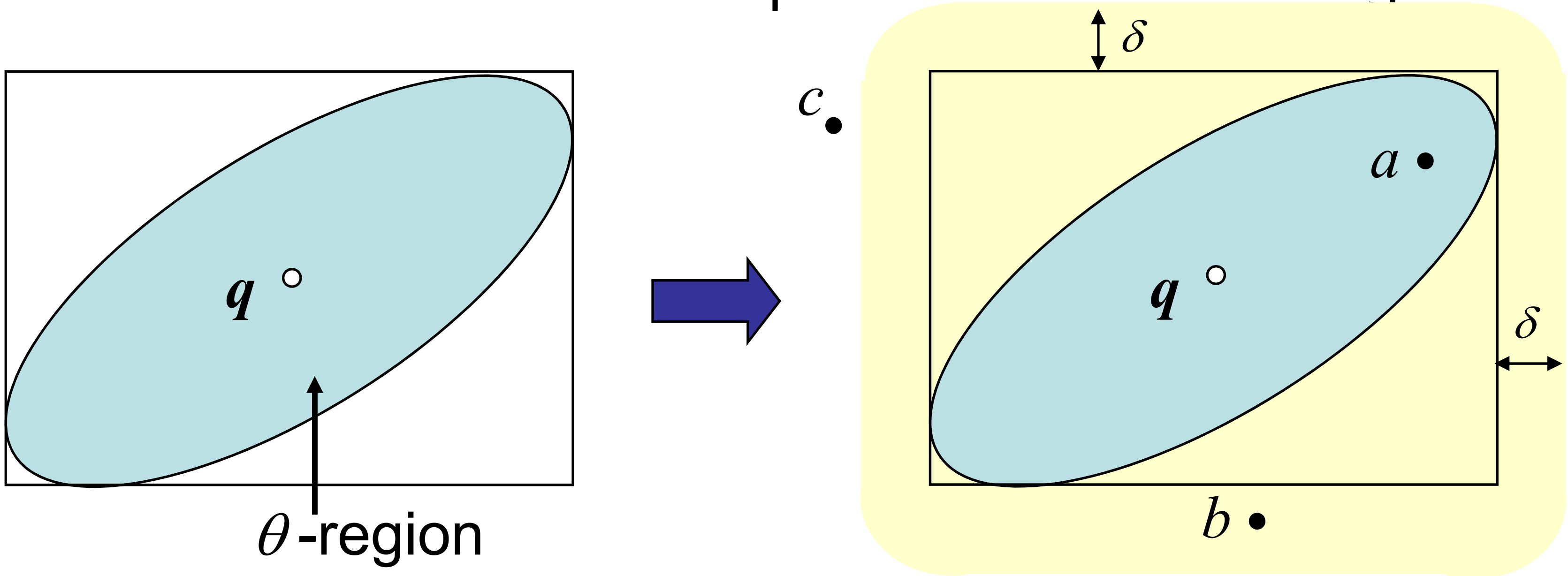
- The ellipsoidal region

$$(\boldsymbol{x}-\boldsymbol{q})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{q}) \le r_\theta^2$$

is the θ-region

Rectilinear-Region-Based (RR) (2)

- Query processing
  - Given a query, θ-region is computed: it is suffice if we have r_θ-table for "normal" Gaussian pdf
    - "Normal" Gaussian: $\boldsymbol{\Sigma} = \mathbf{I}, \boldsymbol{q} = \mathbf{0}$
    - Given θ, it returns appropriate r_θ
  - Derive MBR for θ-region and perform Minkowski Sum
  - Retrieve candidates then perform numerical integration



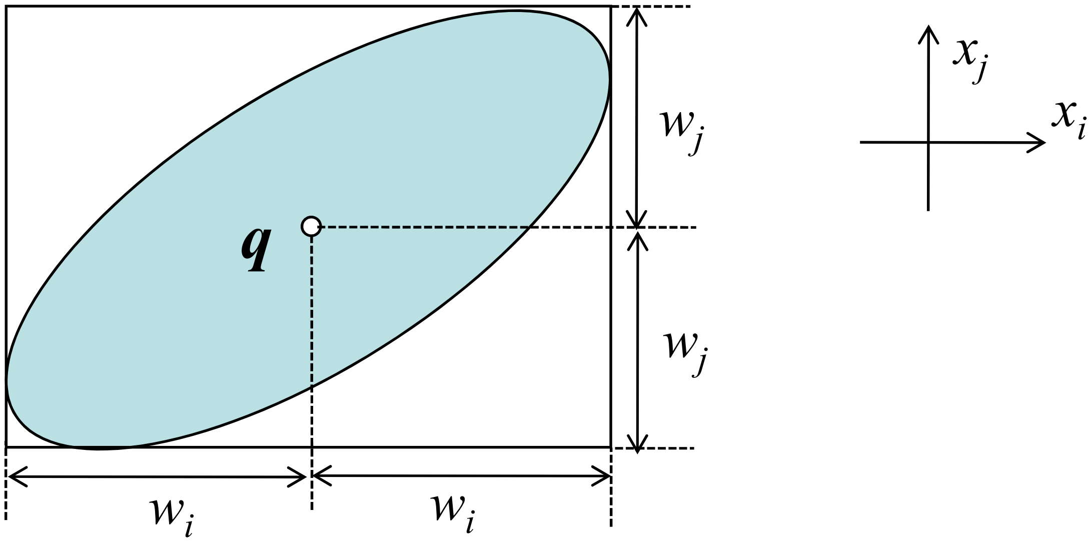
Rectilinear-Region-Based (RR) (3)

- Geometry of bounding box

$$w_i = \sigma_i r_\theta$$

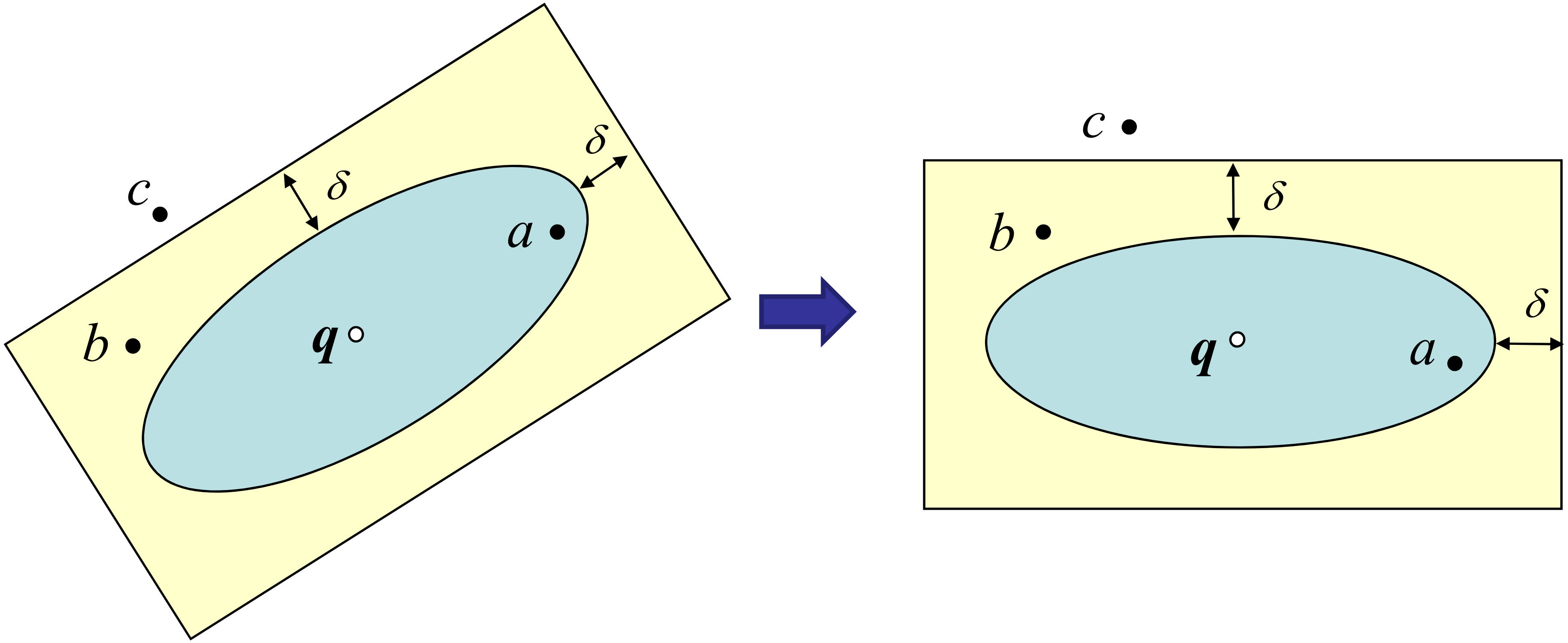
$$\sigma_i = \sqrt{(\boldsymbol{\Sigma})_{ii}}$$

where $(\boldsymbol{\Sigma})_{ii}$ is the (i, i) entry of $\boldsymbol{\Sigma}$



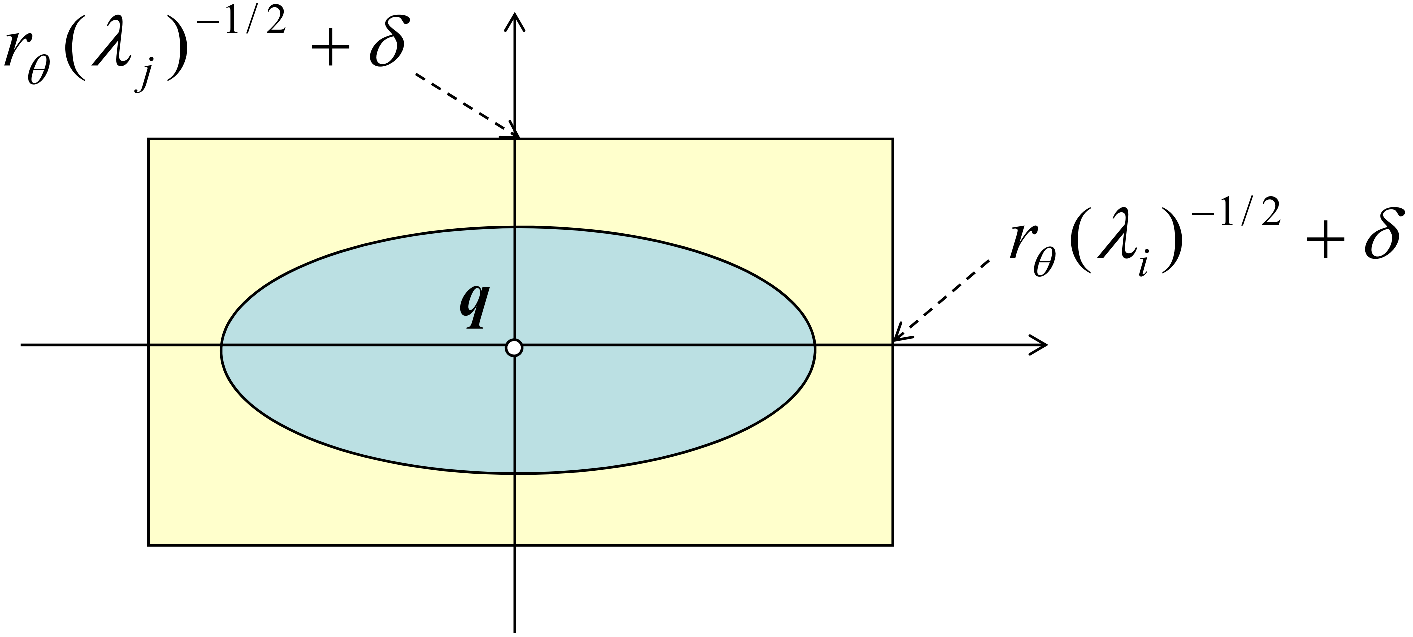
Oblique-Region-Based (OR) (1)

- Use of oblique rectangle
  - Query processing based on axis transformation
  - Not effective for phase 1 (index-based search): Only used for filtering (phase 2)



Oblique-Region-Based (OR) (2)

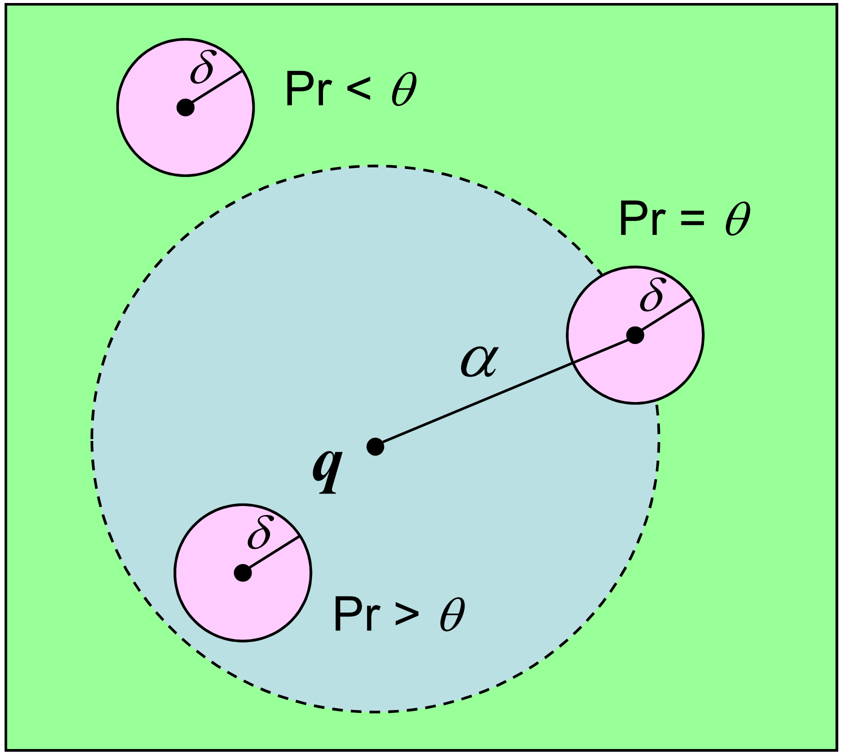
- Step 1: Rotate candidate objects
  - Based on the result of eigenvalue decomposition of $\mathbf{\Sigma}^{-1}$
- Step 2: Check whether each object is inside of the rectangle


$r_\theta(\lambda_j)^{-1/2} + \delta$
$r_\theta(\lambda_i)^{-1/2} + \delta$
$\boldsymbol{q}$

  - λ_i: Eigenvalue of $\mathbf{\Sigma}^{-1}$ for i-th dimension

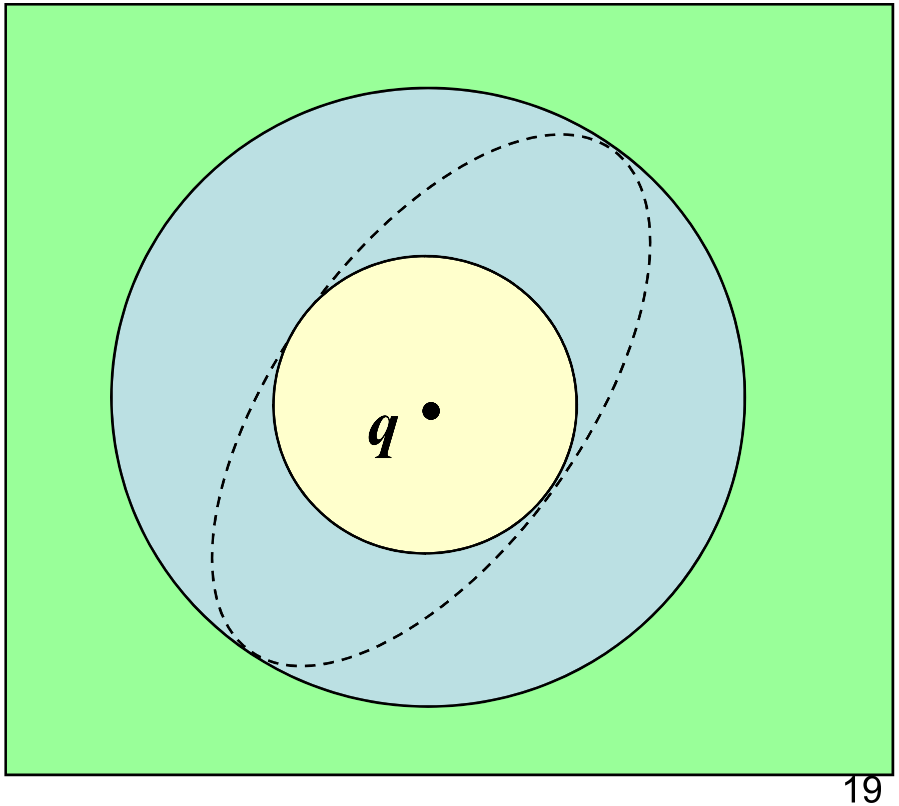
Bounding-Function-Based (BF) (1)

- Basic idea
  - Covariance matrix $\Sigma = \mathbf{I}$ ("normal" Gaussian pdf)
  - Isosurface of pdf has a spherical shape
- Approach
  - Let α be the radius for which the integration result is θ
  - If $\mathrm{dist}(q, p) \leq \alpha$ then p satisfies the condition
  - Construct a table that gives $(\delta, \theta) \rightarrow \alpha$ beforehand



Bounding-Function-Based (BF) (2)

- General case
  - isosurface has an ellipsoidal shape
- Approach
  - Use of upper- and lower-bounding functions for pdf
    - They have sphererical isosurfaces
    - Derived from covariance matrix



Bounding Functions

- Original Gaussian pdf

$$p_q(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{q})^t \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{q})\right]$$

- Upper- and lower-bounding functions

$$p_q^{\top}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left[-\frac{\lambda^{\top}}{2}\|\boldsymbol{x}-\boldsymbol{q}\|^2\right]$$

$$p_q^{\perp}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left[-\frac{\lambda^{\perp}}{2}\|\boldsymbol{x}-\boldsymbol{q}\|^2\right]$$

Isosurface has a spherical shape

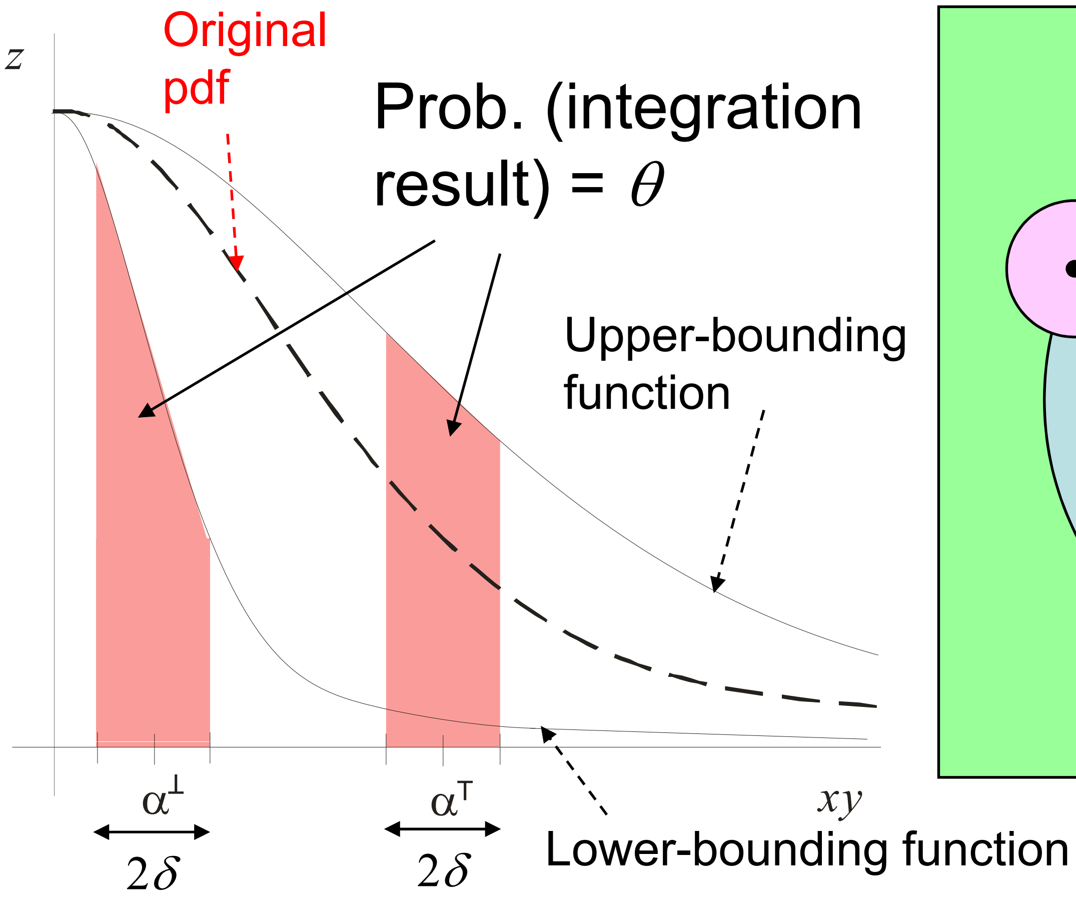
$p_q^{\perp}(\boldsymbol{x}) \le p_q(\boldsymbol{x}) \le p_q^{\top}(\boldsymbol{x})$ holds

Note: $\lambda^{\top} = \min\{\lambda_i\}$

$\lambda^{\perp} = \max\{\lambda_i\}$

Bounding-Function-Based (BF) (3)

- $\alpha^{\top}$ ($\alpha^{\perp}$): Radius with which the integration result of upper- (lower-) bounding function is θ

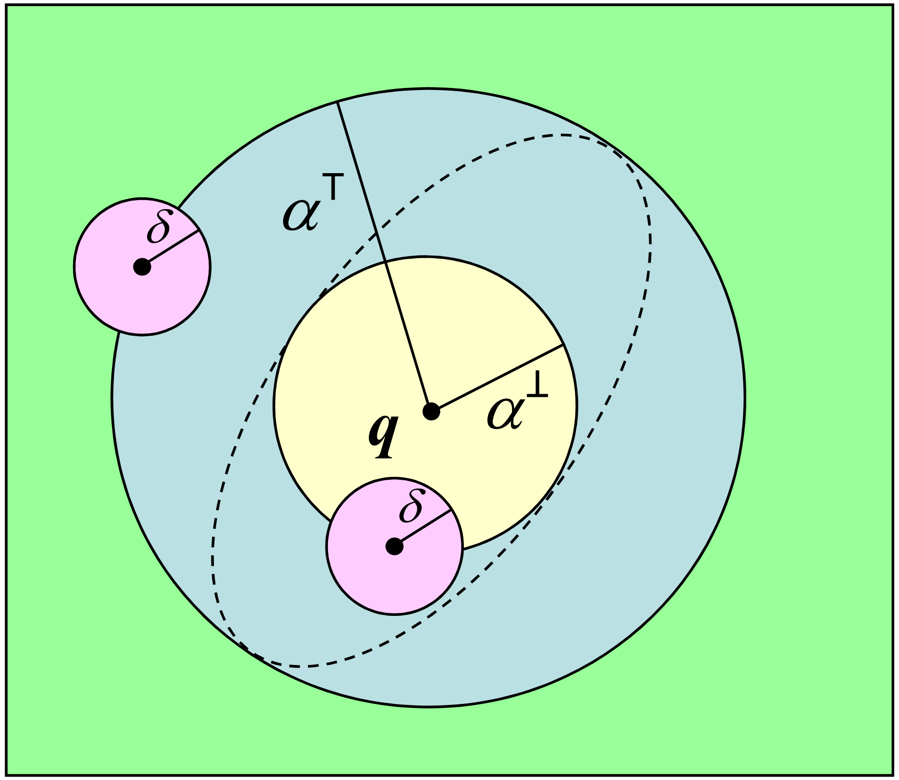


Original pdf

Prob. (integration result) = θ

Upper-bounding function

Lower-bounding function



Bounding-Function-Based (BF) (4)

- Theoretical result
  - Let S^{T} be a spherical region with radius $\sqrt{\lambda^{\mathsf{T}}}\,\delta$ and its center relative to the origin is β^{T}, and assume that S^{T} satisfies the following equation:

$$\int_{\boldsymbol{x}\in S^{\mathsf{T}}} p_{\text{norm}}(\boldsymbol{x})d\boldsymbol{x} = (\lambda^{\mathsf{T}})^{d/2}\,|\,\boldsymbol{\Sigma}\,|^{1/2}\ \theta$$

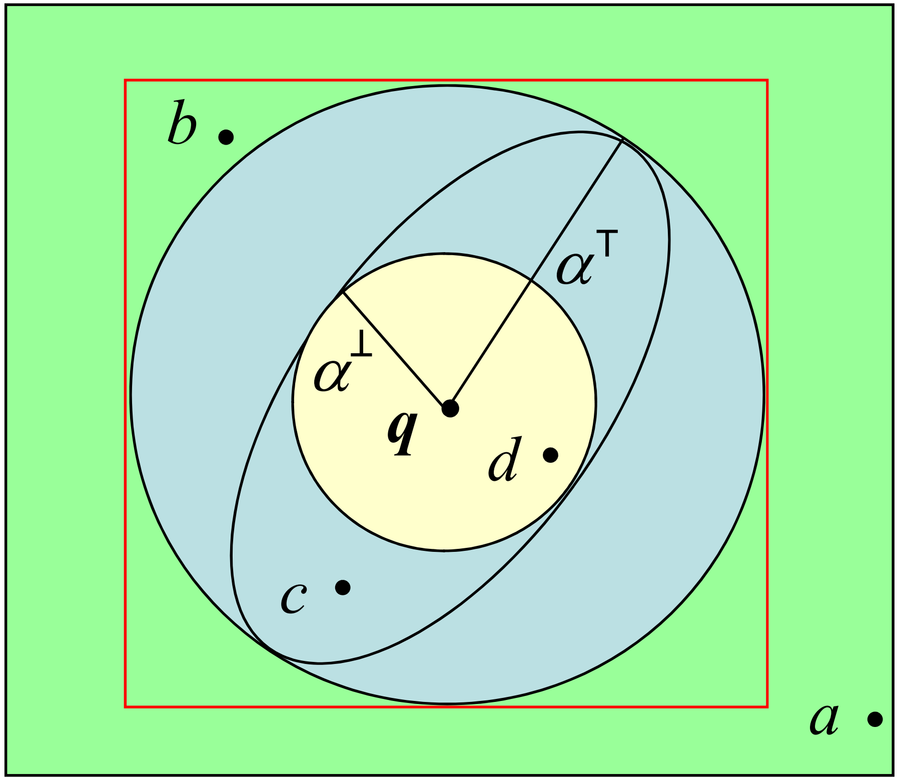
  - Using table that gives $(\delta, \theta) \rightarrow \alpha$, we can get β^{T}:

$$(\sqrt{\lambda^{\mathsf{T}}}\,\delta,\ (\lambda^{\mathsf{T}})^{d/2}\,|\,\Sigma\,|^{1/2}\ \theta) \rightarrow \beta^{\mathsf{T}}$$

  - Then we can get $\alpha^{\mathsf{T}} = \dfrac{\beta^{\mathsf{T}}}{\sqrt{\lambda^{\mathsf{T}}}}$

Bounding-Function-Based (BF) (5)

- Step 1: Use of R-tree
  - $\{b, c, d\}$ are retrieved as candidates
- Step 2: Filtering using $\alpha^{\top}$
  - b is deleted
- Step 2': Filtering using $\alpha^{\perp}$
  - We can determine d as an answer without numerical integration



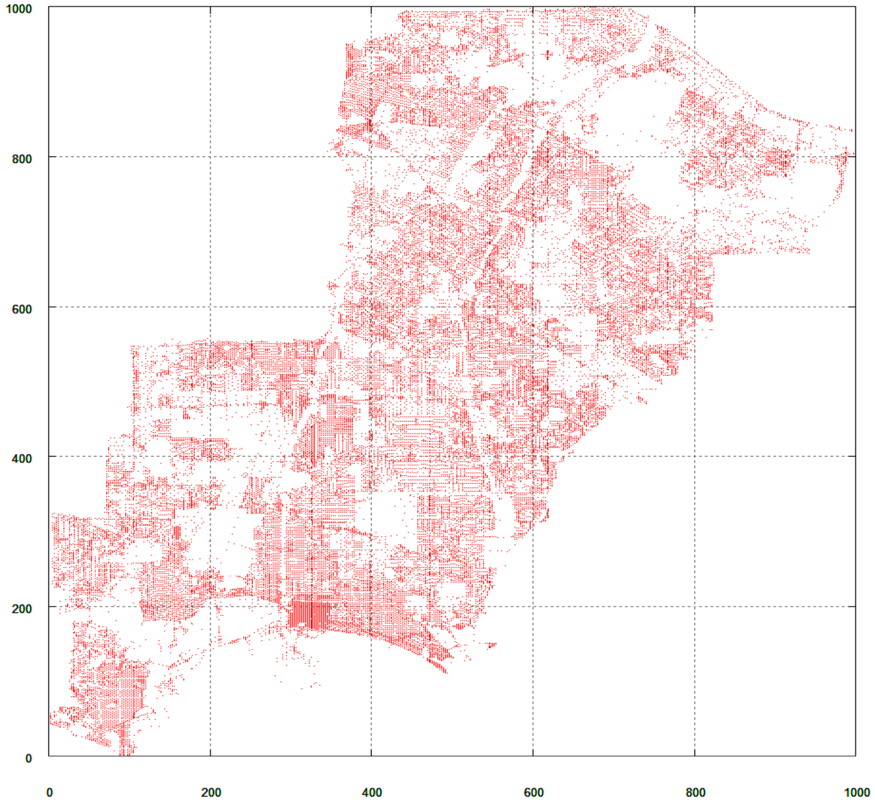
- Step 3: Numerical integration
  - Performed on $\{c\}$

Outline

- Background and Problem Formulation
- Related Work
- Query Processing Strategies
- Experimental Results
- and Conclusions

Experiments on 2D Data (1)

- Map of Long Beach, CA
  - Normalized into $[0, 1000] \times [0, 1000]$



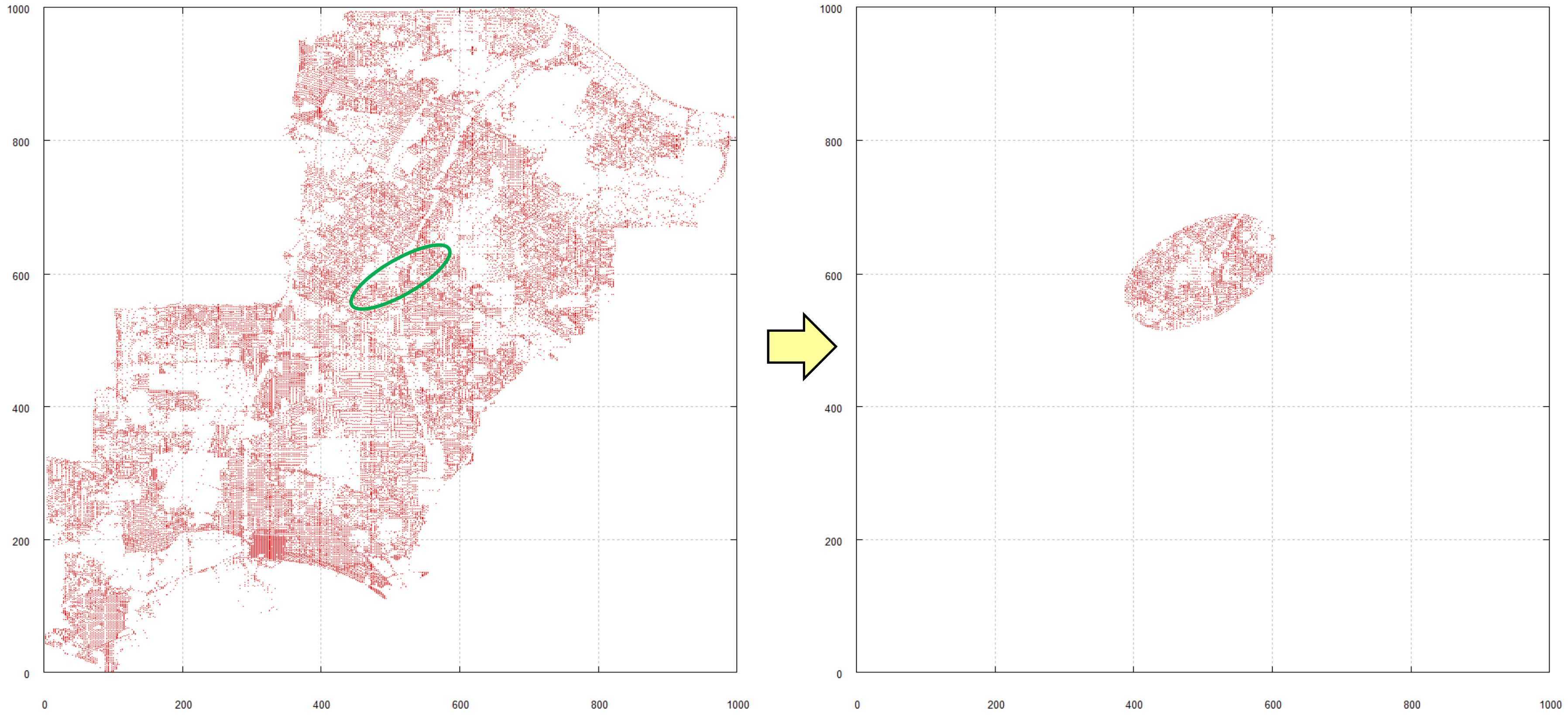
- 50,747 entries
- Indexed by R-tree
- Covariance matrix

$$\boldsymbol{\Sigma} = \gamma \begin{bmatrix} 7 & 2\sqrt{3} \\ 2\sqrt{3} & 7 \end{bmatrix}$$

- γ : Scaling parameter
  - Default: $\gamma = 10$

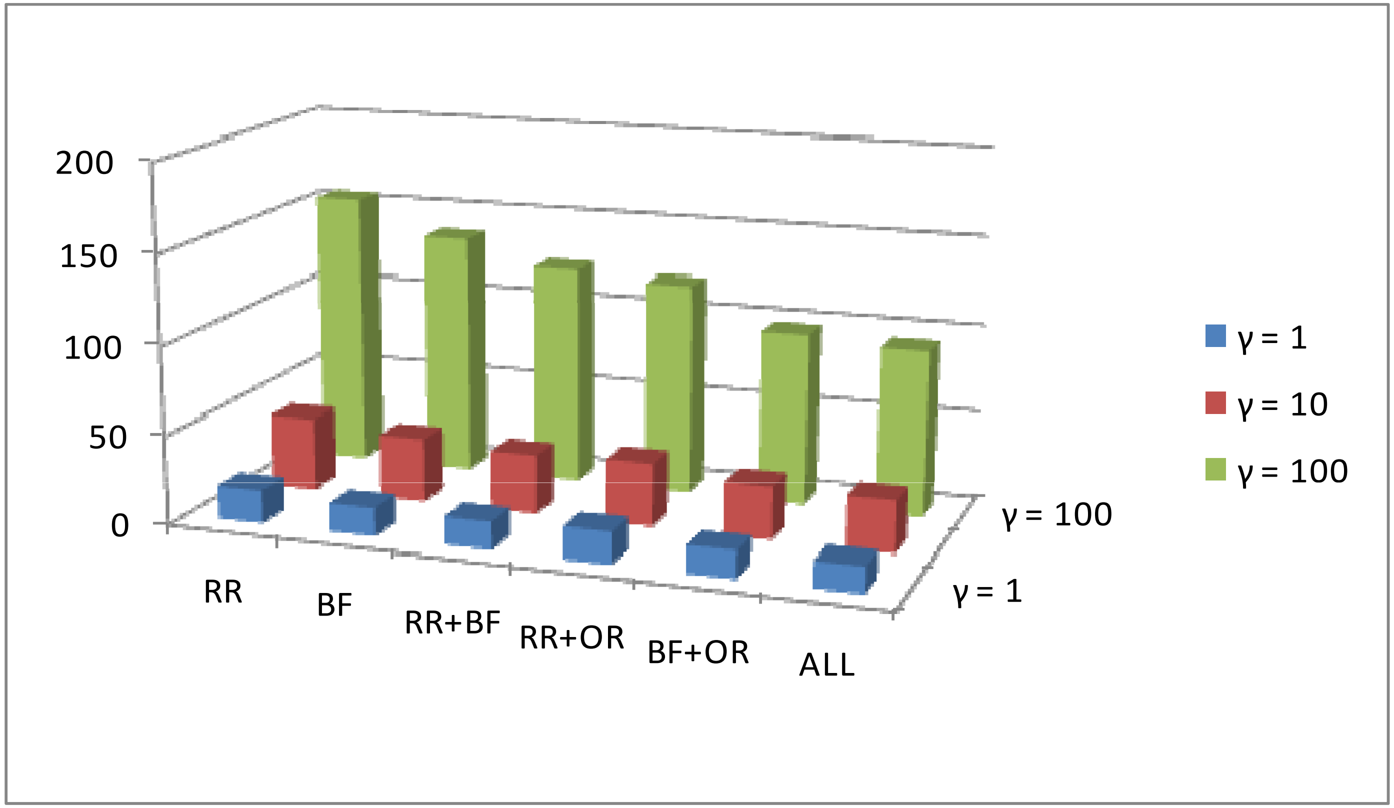
Example Query

- Find objects within distance $\delta = 50$ with probability threshold $\theta = 1\%$



Experiments on 2D Data (2)

- Numerical integration dominates the total cost
- R-tree-based search is negligible
- ALL is the most effective strategy

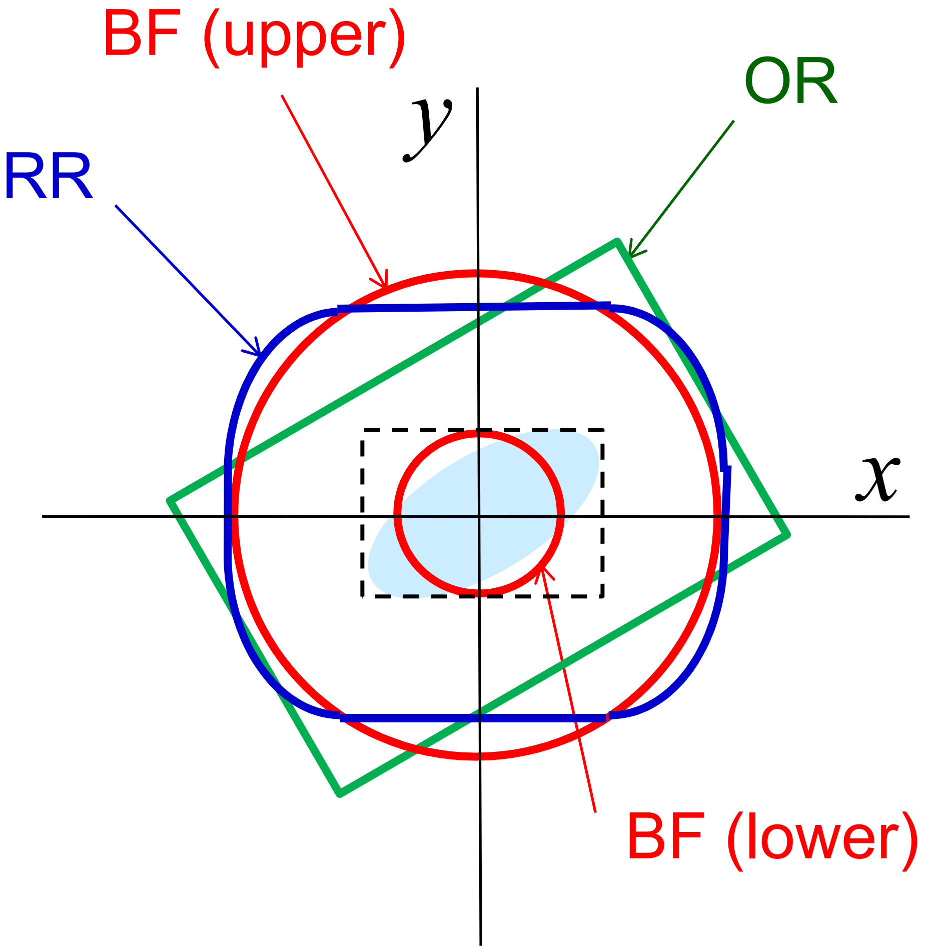


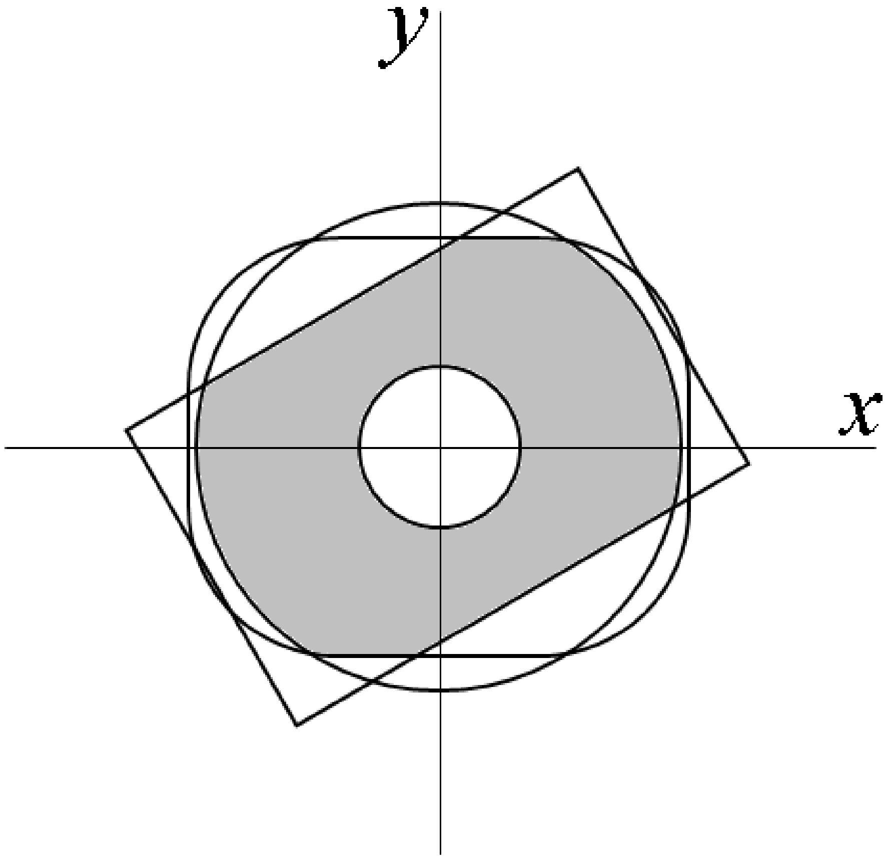
$\delta = 25$

$\theta = 0.01$

Experiments on 2D Data (3)

- Filtering regions ($\delta = 25$, $\theta = 0.01$, $\gamma = 10$)

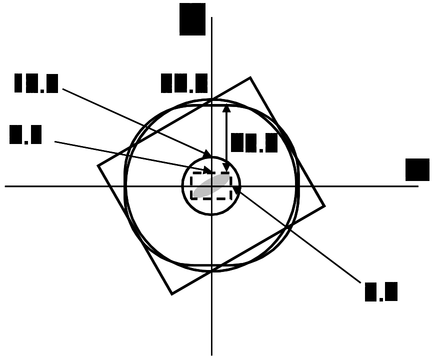




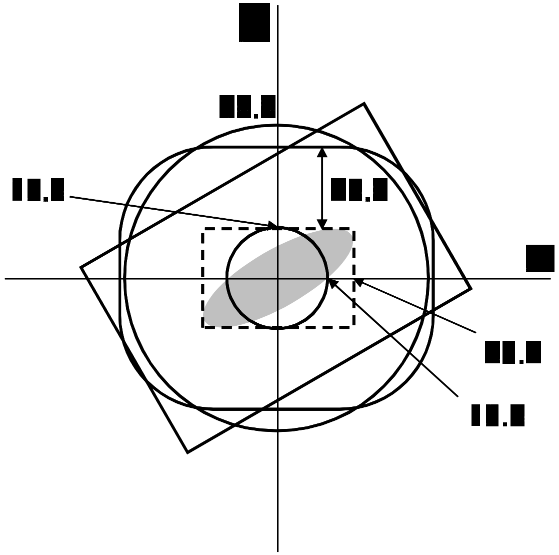
Integration region for ALL

Experiments on 2D Data (4)

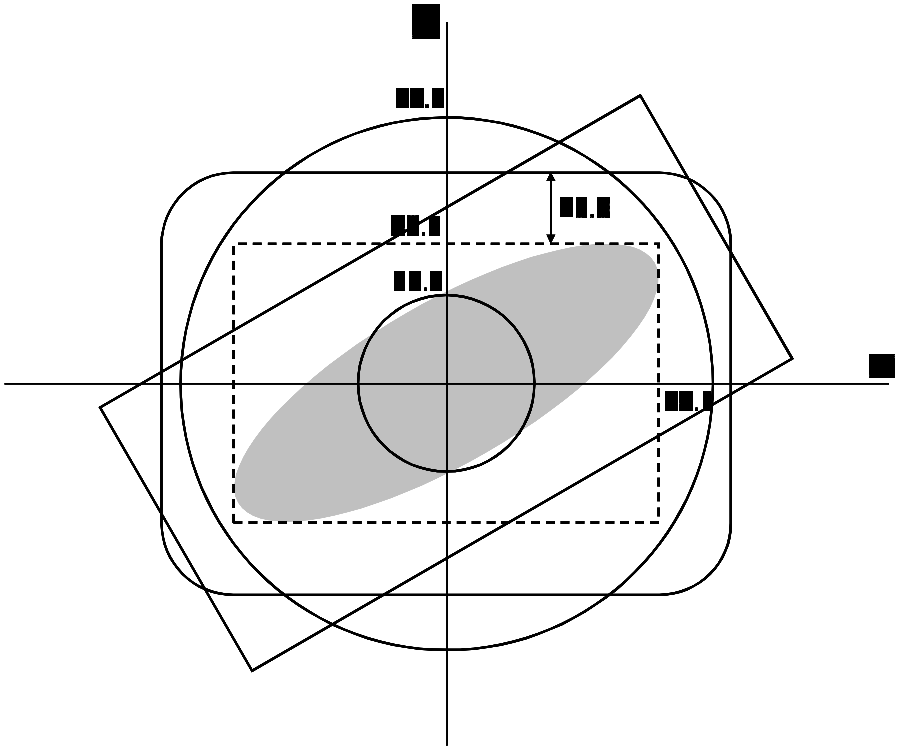
- Filtering regions for different uncertainty setting $(\delta = 25, \theta = 0.01)$



$\gamma = 1$:
Nearly exact



$\gamma = 10$:
Medium uncertainty



$\gamma = 100$:
Uncertain

Experiments on 9D Data (1)

- Motivating Scenario: Example-Based Image Retrieval
  - User specifies sample images
  - Image retrieval system estimates his interest as a Gaussian distribution

Experiments on 9D Data (2)

- Data set: Corel Image Features data set
  - – From UCI KDD Archive
  - – Color Moments data
  - – 68,040 9D vectors
  - – Euclidean-distance based similarity
- Experimental Scenario: Pseudo-Feedback
  - – Select a random query object, then retrieve *k*-NN query ($k = 20$) as sample images
  - – Derive the covariance matrix from samples

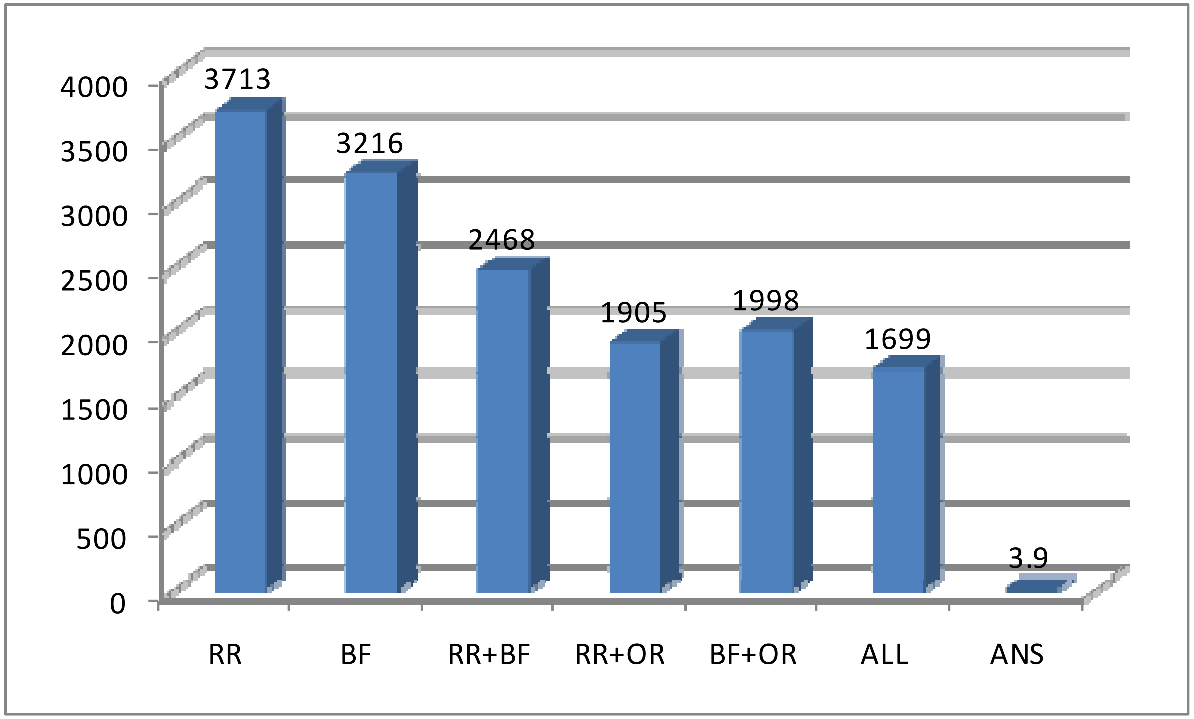
$$\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}} + \kappa \mathbf{I}$$

$\tilde{\boldsymbol{\Sigma}}$: Sample covariance matrix

κ : Normalization parameter

Experiments on 9D Data (3)

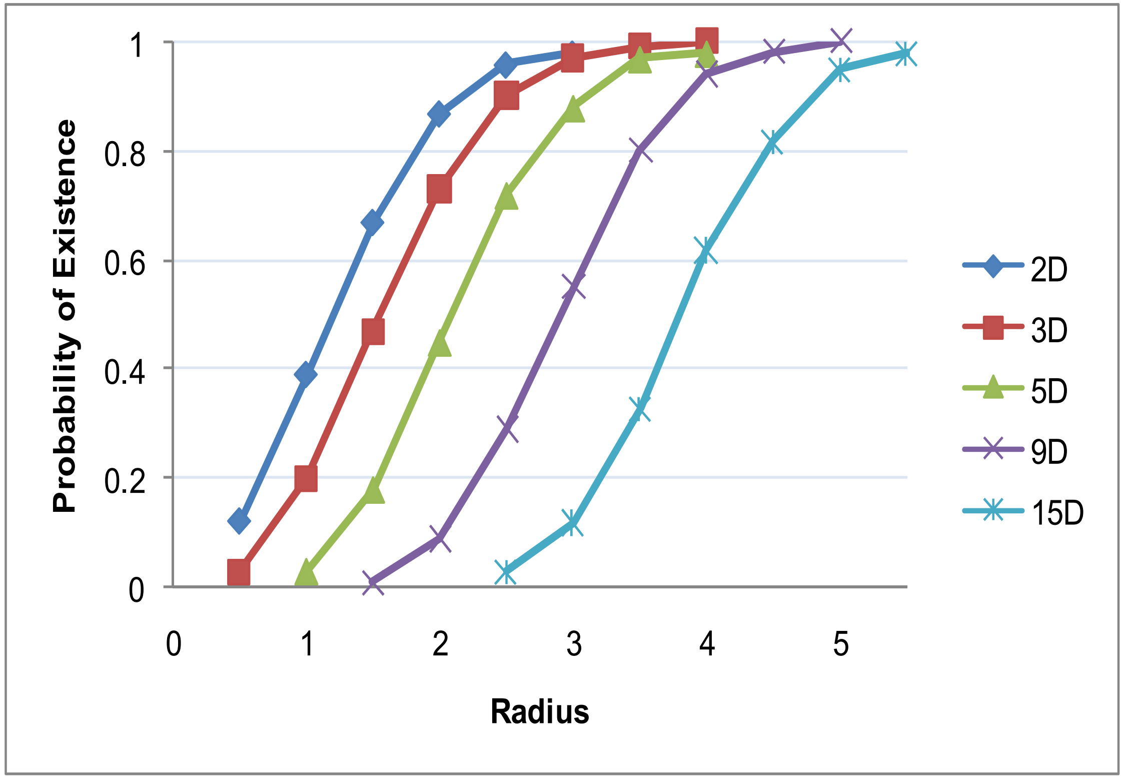
- Parameters
  - $\delta = 0.7$: For exact case, it retrieves 15.3 objects
  - $\theta = 40\%$
- Number of candidates (ANS: answer objs)



Too many candidates to retrieve only 3.9 objects!

Experiments on 9D Data (4)

- Reason: Curse of dimensionality
- Plot shows existence probability for p_{norm} for different radii and dimensions



Location of query object is too vague: In medium dimension, it is quite apart from its distribution center on average

Example: For 9D case, the probability that query object is within distance two is only 9%

Outline

- Background and Problem Formulation
- Related Work
- Query Processing Strategies
- Experimental Results
- Conclusions

Conclusions

- Spatial range query processing methods for imprecise query objects
  - Location of query object is represented by Gaussian distribution
  - Three strategies and their combinations
  - Reduction of numerical integration is important
  - Problem is difficult for medium- and high-dimensional data
- Our related work
  - Probabilistic Nearest Neighbor Queries (MDM’09)

Spatial Range Querying for Gaussian-Based Imprecise Query Objects

Yoshiharu Ishikawa, Yuichi Iijima

Nagoya University

Jeffrey Xu Yu

The Chinese University of Hong Kong