Ludwig Maximilians Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

Skript zur Vorlesung

Datenbanksysteme II

Sommersemester 2006

Kapitel 11: Indexstrukturen für Wahrscheinlichkeits-Verteilungen

Vorlesung: Christian Böhm
Übungen: Elke Achtert, Peter Kunath, Alexey Pryakhin

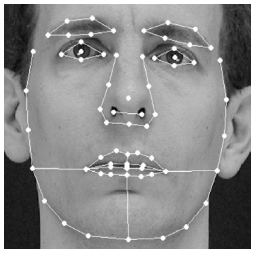
http://www.dbs.informatik.uni-muenchen.de/Lehre/DBSII

Datenbanksysteme II
Kapitel 11: Indexstrukturen für Wahrscheinlichkeits-Verteilungen

Inhalt

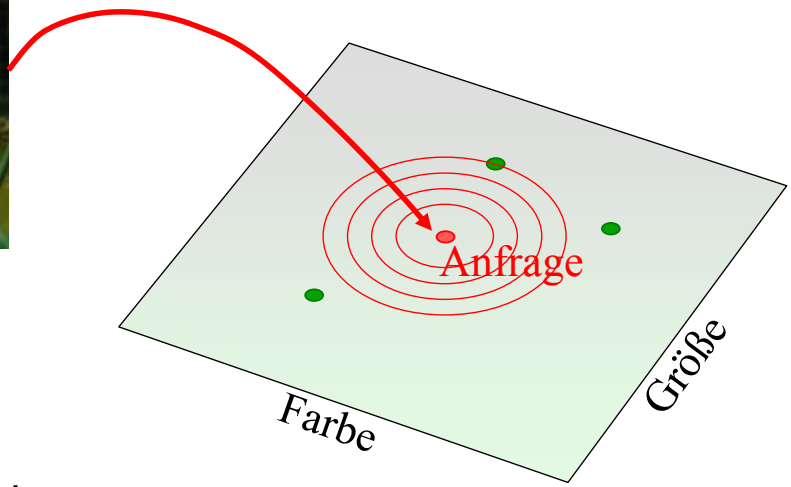
Effiziente Objekt-Identifikation

Phänotyp-Deskription → Botanische Klassifikation



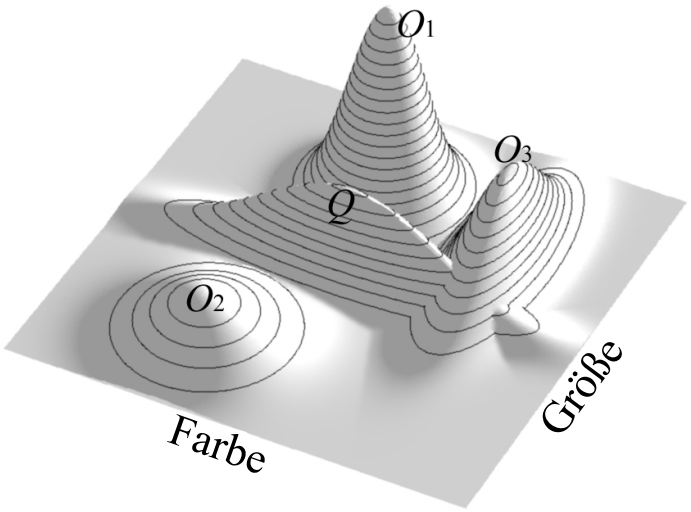
biometrische Merkmale → Personen-Identifikation

Von Vektoren zu Verteilungen (1)



Nächste-Nachbar-Anfragen erlauben nur globale Feature-Gewichtung

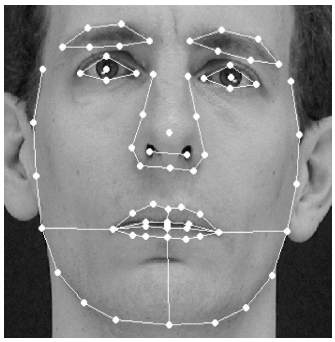
Von Vektoren zu Verteilungen (2)

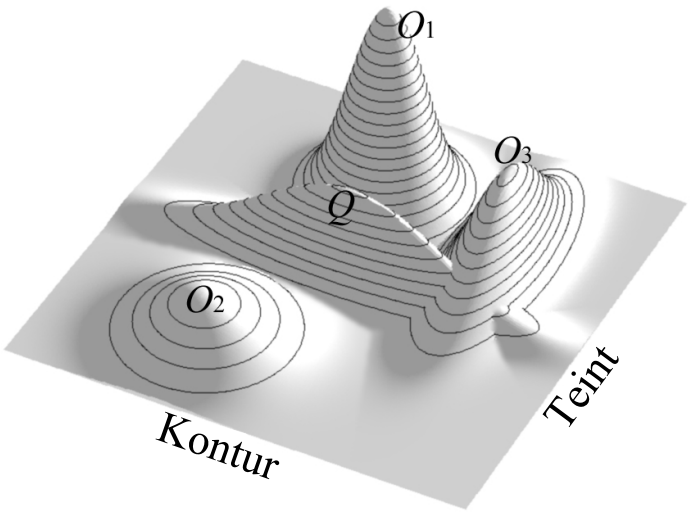


Unschärfe:

Die Spezies O_3
weist eine hohe Varianz
beim Größen-Feature auf.

Von Vektoren zu Verteilungen (3)



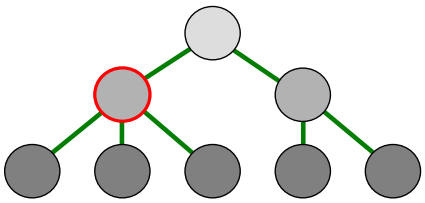


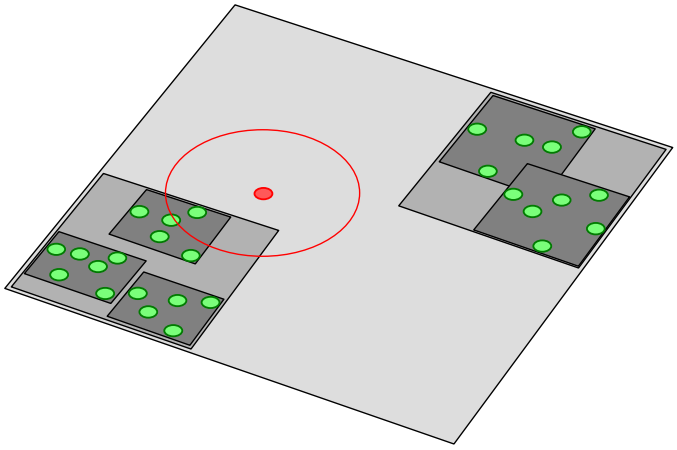
Unsicherheit:

Die Person Q
wurde nicht ganz
von vorne aufgenommen

Von Vektoren zu Verteilungen (4)

Das R-Tree-Prinzip:





Konservative Approximation
durch Minimum Bounding Rectangle (MBR)

Inhalt

Probabilistic Feature Vectors (PFV)

- Jedes beobachtete Objekt wird durch einen Probabilistic Feature Vector (PFV) v modelliert: $v = (\mu_i, \sigma_i)$, $1 \le i \le d$
  - *Merkmalsparameter* μ_i
  - *Unsicherheitsparameter* σ_i
- Normalverteilungsannahme für Messfehler/Unschärfe:

$$N_{\mu_i,\sigma_i}(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \cdot e^{\frac{-(x_i-\mu_i)^2}{2\cdot\sigma_i^2}}, i = 1,\ldots, d$$

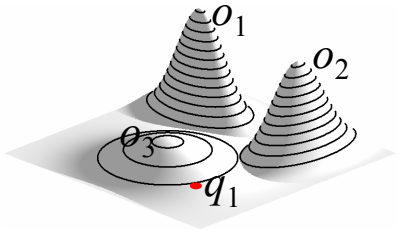
- Unabhängige multivariate Normalverteilung für Vektor v:

$$N_{\vec{\mu},\vec{\sigma}}(\vec{x}) = \prod_{1 \le i \le d} N_{\mu_i,\sigma_i}(x_i)$$

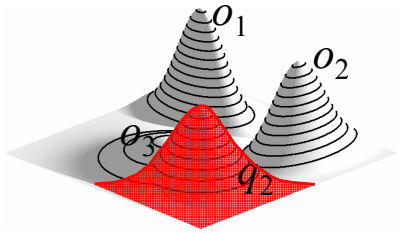
Anfragetypen (1)

- Unterscheidung nach Anfrage-Objekt:

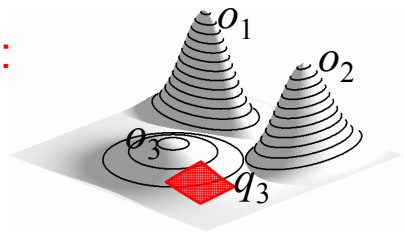
Punkt-Anfrage:



Gauss-Anfrage:



Intervall-Anfrage:



Anfragetypen (2)

- Unterscheidung nach Kardinalitäts-Spezifikation:

Gegeben: Punkt-, Gauss-, oder Intervall-Anfrage

- Threshold Identification Query (TIQ):
  Gegeben sei ein Grenzwert P_Θ mit $0 \leq P_\Theta \leq 1$:

$$TIQ(q, P_\Theta) = \{v \in DB \mid P(v \mid q) \geq P_\Theta\}$$

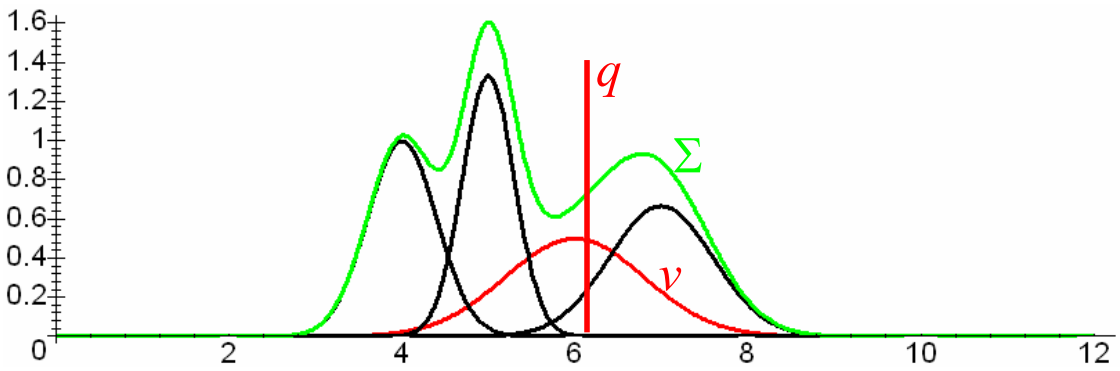
- k-Most Likely Identification Query (k-MLIQ):

$$v \in DB, k \in N, \forall v \in MLIQ_k(q), \forall w \in DB \setminus MLIQ_k(q): P(v \mid q) > P(w \mid q)$$

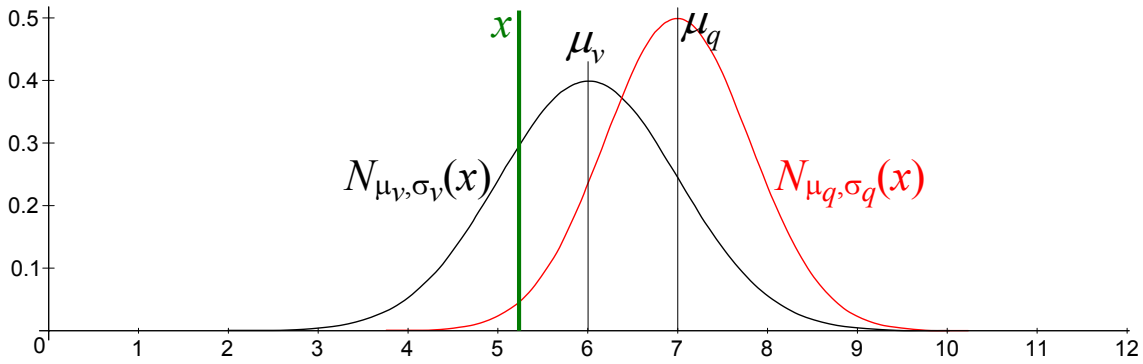
Trefferwahrscheinlichkeit

- Gegeben: Nicht-probabilistischer Anfrage-Vektor $q = [q_1, q_2, \ldots]$
  Ein gespeichertes Objekt (PFV) $v = [\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots]$
- Trefferwahrscheinlichkeit nach dem Satz von Bayes:

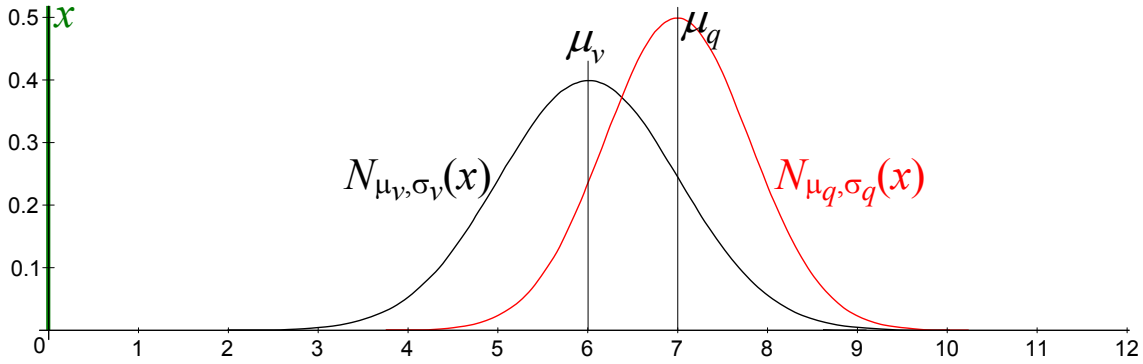
$$P(v \mid q) = \frac{P(v) \cdot p(q \mid v)}{\sum_{w \in DB}(P(w) \cdot p(q \mid w))}$$



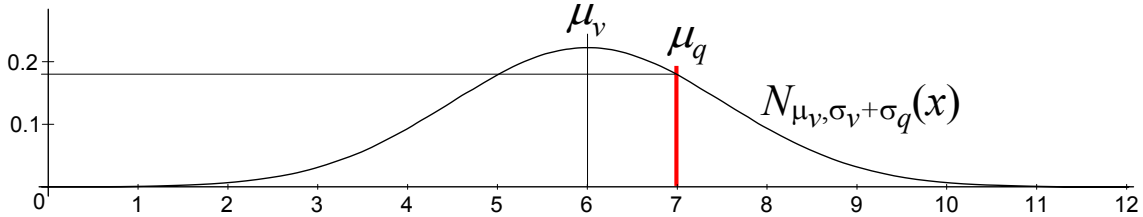
Transformation der Anfrage-PFVs (1)



Transformation der Anfrage-PFVs (2)

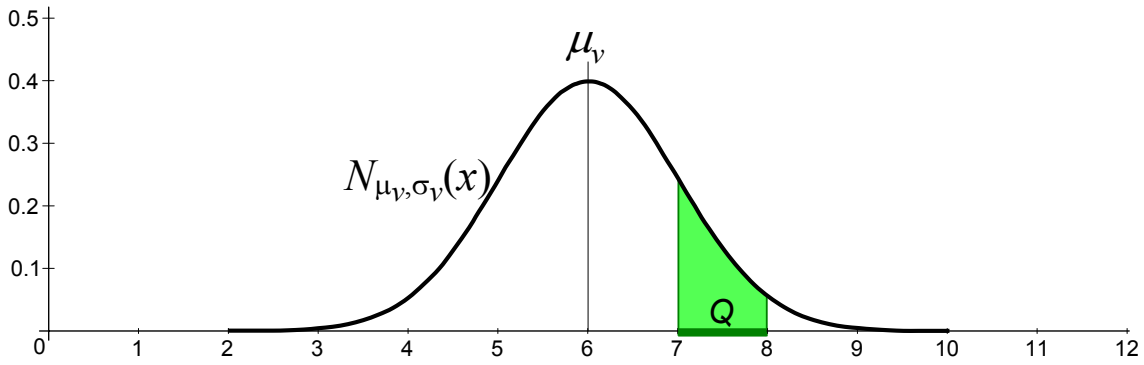


$$p(q \mid v) = \int_{-\infty}^{+\infty} N_{\mu_v,\sigma_v}(x) \cdot N_{\mu_q,\sigma_q}(x)\,dx = N_{\mu_v,\sigma_v+\sigma_q}(\mu_q)$$



Intervall-Anfragen

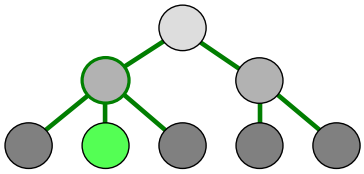
- Wahrscheinlichkeit ergibt sich durch Integration über den Anfrage-Bereich Q

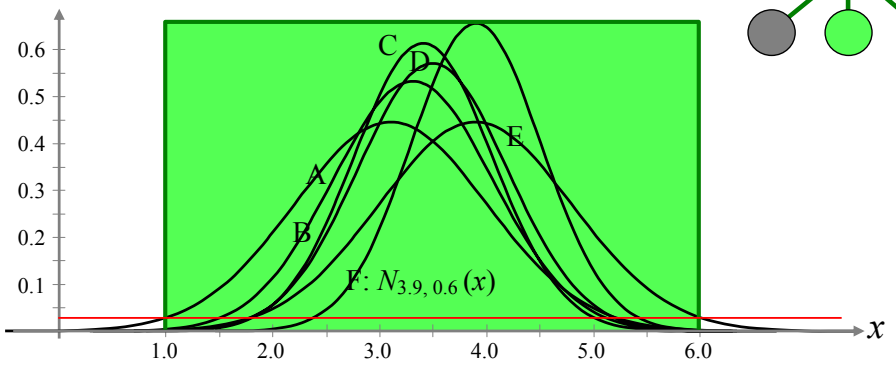


Inhalt

Approximation der Gausskurven

Schwellwert-basierte Approximation:





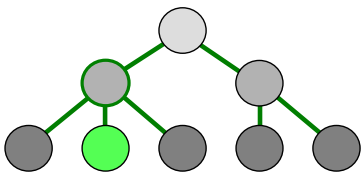
Probleme:

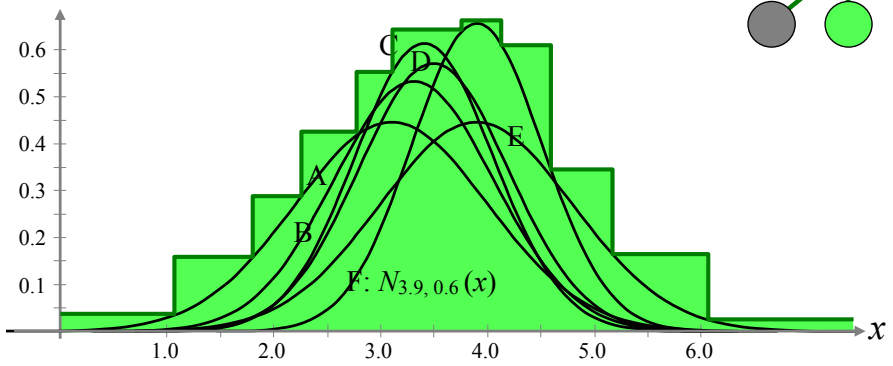
- Nicht konservativ: False Dismissals
- Geringe Approximationsgüte
- Wahl eines geeigneten Schwellwerts unklar

[VLDB 2004: Efficient Indexing Methods for Probabilistic Threshold Queries, Cheng, Xia, Prabhakar, Shah, Vitter]

Approximation der Gausskurven

Konservative Approximation:



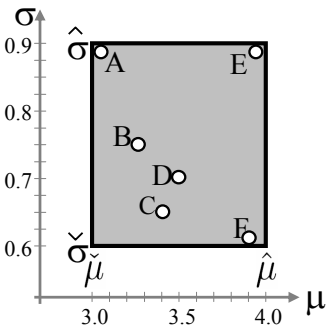


Problem:
Zu speicherintensiv, v.a. im multivariaten Fall bei hoher Dimensionalität

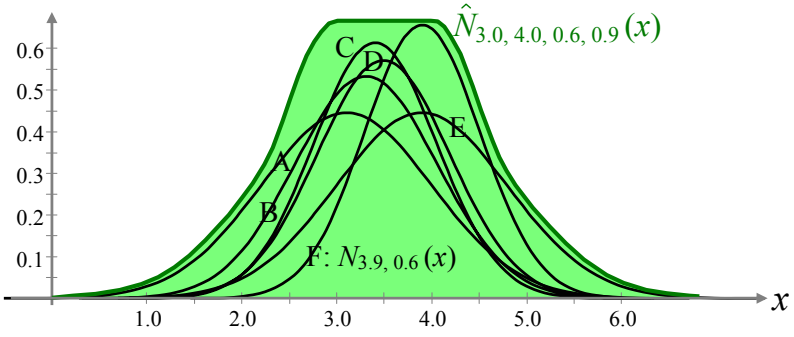
[VLDB 2005: Indexing Multidimensional Uncertain Data, Tao, Cheng, Xiao, Ngai, Kao, Prabhakar]

Indexierung des Parameterraums

μ-σ-Raum:



Feature-Wahrscheinlichkeitsdichte-Raum:



Konservative Approximation:

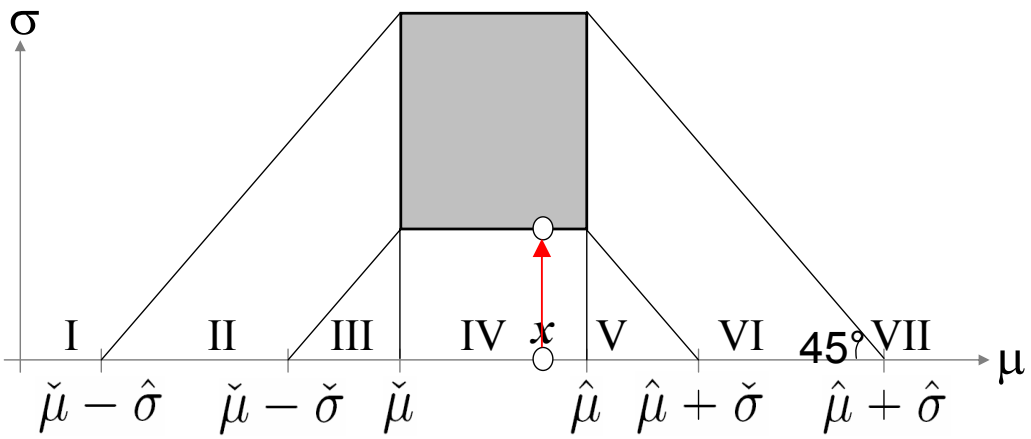
$$\hat{N}_{\breve{\mu},\hat{\mu},\breve{\sigma},\hat{\sigma}}(x) = \max_{\substack{\breve{\mu}\le\mu\le\hat{\mu}\\ \breve{\sigma}\le\sigma\le\hat{\sigma}}}\{N_{\mu,\sigma}(x)\}$$

[ICDE 2006: The Gauss-Tree: Efficient Object Identification of Probabilistic Feature Vectors
mit Alexey Pryakhin und Matthias Schubert

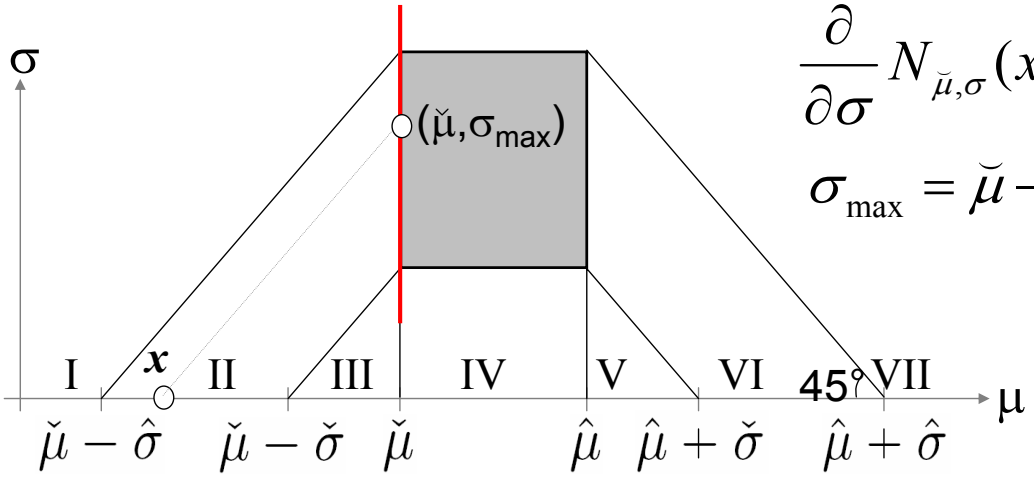
Analytische Lösung (1)

$$\hat{N}_{\breve{\mu},\hat{\mu},\breve{\sigma},\hat{\sigma}}(x) = \begin{cases} N_{\breve{\mu},\hat{\sigma}}(x) \;\; if \;\; x < \breve{\mu} - \hat{\sigma} & (I) \\ N_{\breve{\mu},\breve{\mu}-x}(x) \;\; if \;\; \breve{\mu} - \hat{\sigma} \le x < \breve{\mu} - \breve{\sigma} & (II) \\ N_{\breve{\mu},\breve{\sigma}}(x) \;\; if \;\; \breve{\mu} - \breve{\sigma} \le x < \breve{\mu} & (III) \\ N_{x,\breve{\sigma}}(x) \;\; if \;\; \breve{\mu} \le x < \hat{\mu} & (IV) \\ N_{\hat{\mu},\breve{\sigma}}(x) \;\; if \;\; \hat{\mu} \le x < \hat{\mu} + \breve{\sigma} & (V) \\ N_{\hat{\mu},x-\hat{\mu}}(x) \;\; if \;\; \hat{\mu} + \breve{\sigma} \le x < \hat{\mu} + \hat{\sigma} & (VI) \\ N_{\hat{\mu},\hat{\sigma}}(x) \;\; if \;\; \hat{\mu} + \hat{\sigma} \le x & (VII) \end{cases}$$

Analytische Lösung (2)



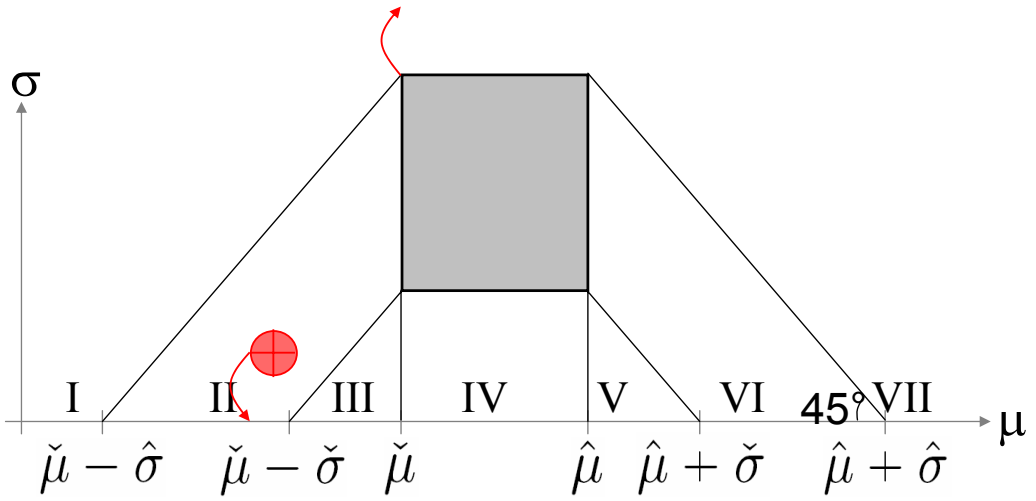
Analytische Lösung (3)



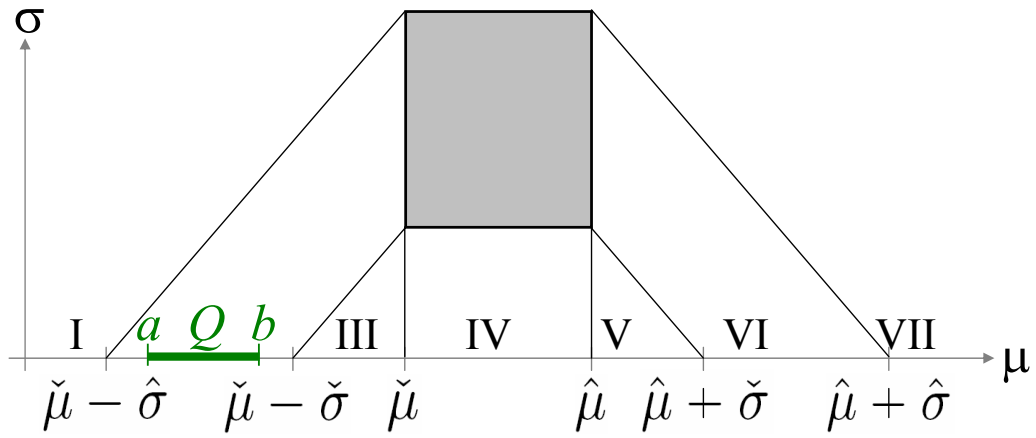
$$\frac{\partial}{\partial\sigma} N_{\check{\mu},\sigma}(x) = 0$$

$$\sigma_{\max} = \check{\mu} - x$$

Probabilistische Anfragen



Intervall-Anfragen



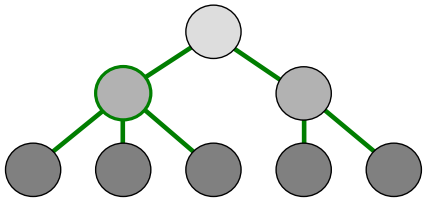
$$\mu_{max} = \max\{\breve{\mu}, \min\{1/2(a+b), \hat{\mu}\}\}$$

$$\sigma_{max} = \frac{\sqrt{2\ln\left(\frac{\mu-b}{\mu-a}\right)(a-b)(2\mu-a-b)}}{2\ln\left(\frac{\mu-b}{\mu-a}\right)}$$

falls Q ganz in I-III liegt
(hier nur ein Fall von mehreren)

[SSDBM 2006: Probabilistic Ranking Queries on Gaussians mit Alexey Pryakhin und Matthias Schubert]

Anfragebearbeitungs-Algorithmen



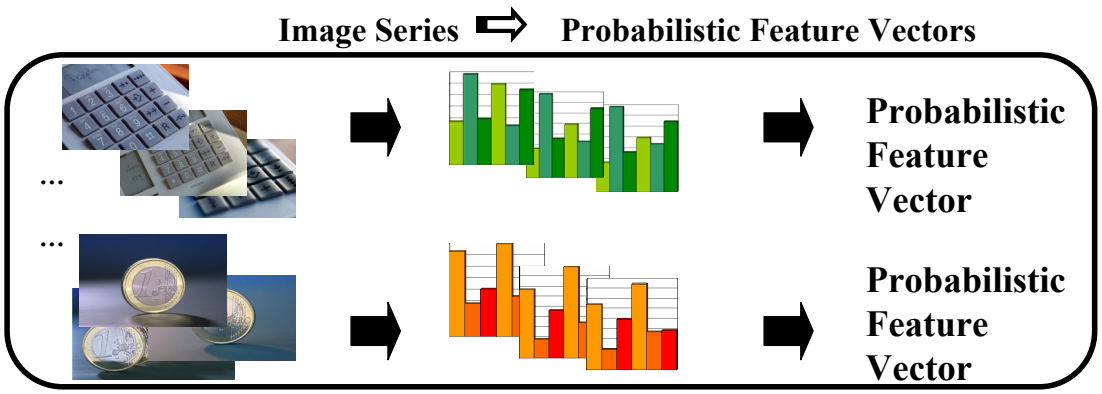
Grundschema:

- Beginne mit der Baum-Wurzel
- Expandiere so lange Knoten des Baumes, bis kein Knoten mehr existiert, der Treffer enthalten kann
  (Ausschluss wegen konservativer Approximation)
- Immer bei Blatt-Knoten:
  Ermittle Wahrscheinlichkeiten der gespeicherten Objekte
  - Wenn sichere Treffer: Ausgeben
  - Wenn noch nicht sicher: In Puffer merken
- Sonst: Sortiere Kind-Knoten in Warteschlange ein

Inhalt

Experimente und Ergebnisse

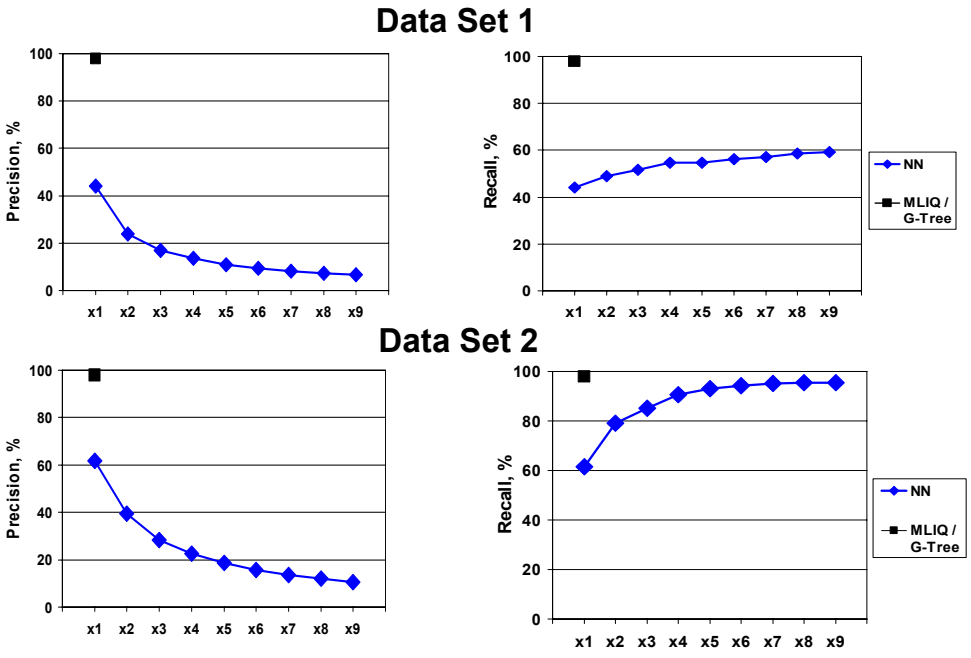
- Testumgebung



Datensatz	# Dimensionen	Größe	# Anfragen
Data Set 1	27	10,987	100
Data Set 2	10	100,000	500

Experimente und Ergebnisse

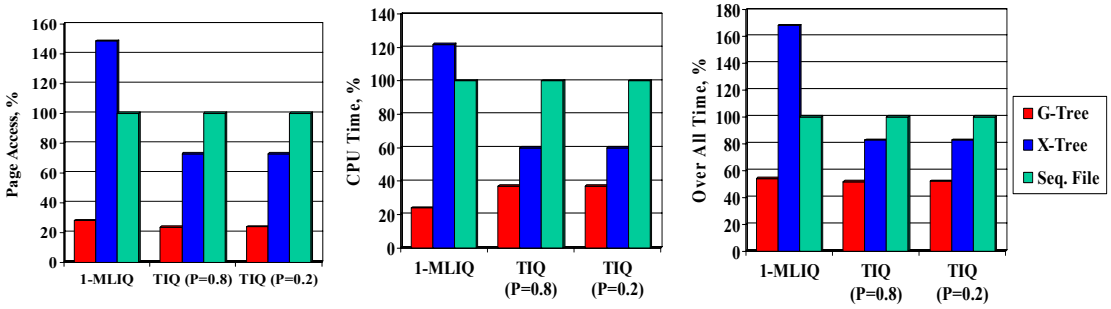
Vergleich Recall und Precision of 3MLIQ and 3NN



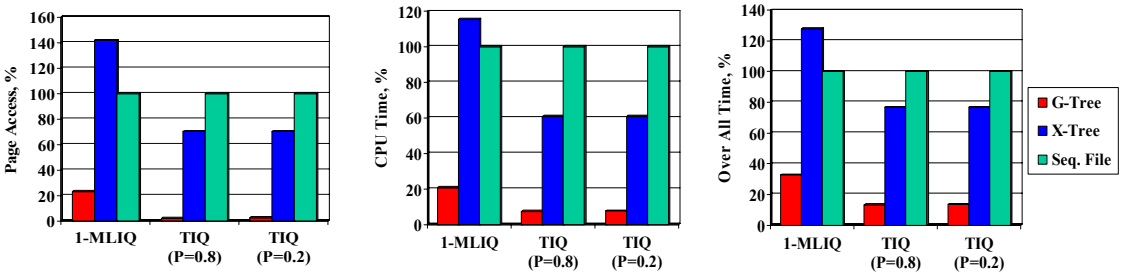
Experimente und Ergebnisse

Sequential Scan, X-Tree and Gauss-Tree bzgl. Effizienz

Data Set 1



Data Set 2

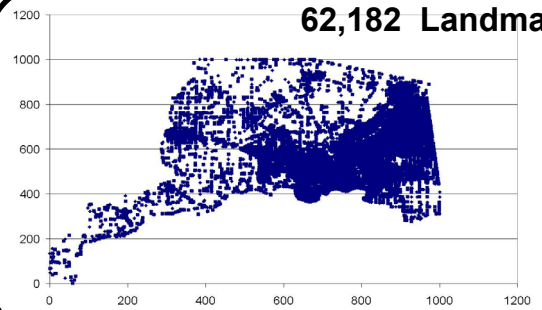


Experimente und Ergebnisse

Testumgebung

Datensatz	# Dimensionen	Größe	# Anfragen
Data Set 1	1	100,000	200
Data Set 2	2	62,182	200

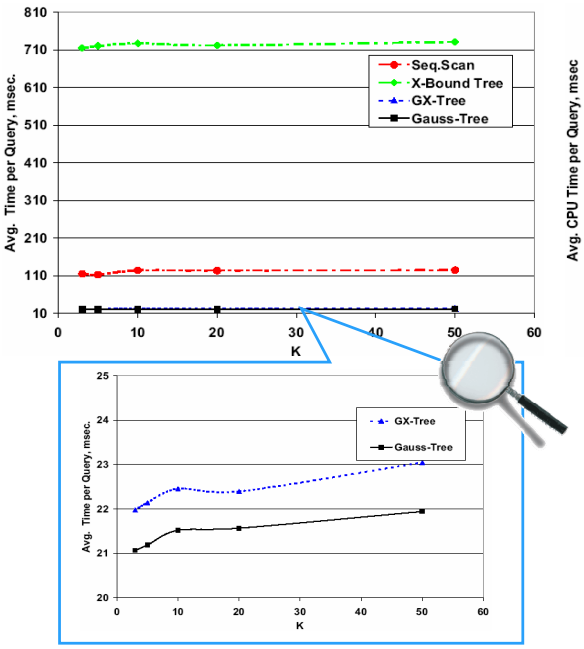
62,182 Landmarks of Sacramento

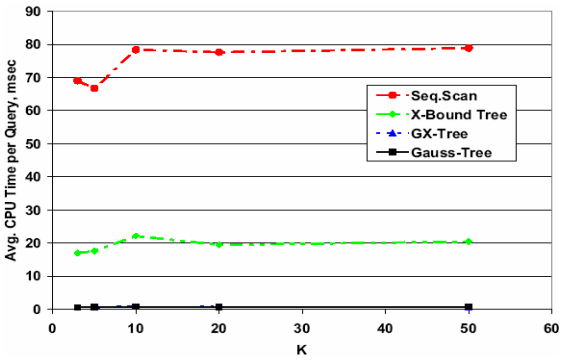


$$v_1 = (\mu_{11}, \sigma_{11}, \ldots\ldots, \mu_{1d}, \sigma_{1d})$$
$$\vdots$$
$$v_j = (\mu_{j1}, \sigma_{j1}, \ldots\ldots, \mu_{jd}, \sigma_{jd})$$
$$\vdots$$
$$v_N = (\mu_{N1}, \sigma_{N1}, \ldots\ldots, \mu_{Nd}, \sigma_{Nd})$$

Experimente und Ergebnisse

Komplette Laufzeit und CPU Zeit für PRQ (Data Set 1)

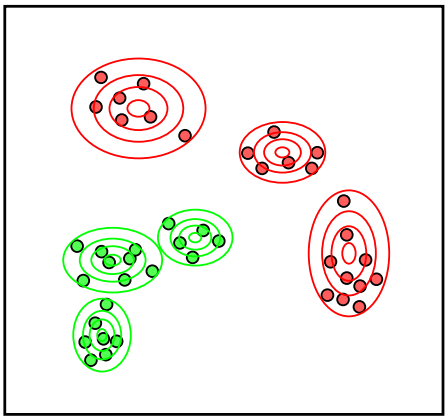




Inhalt

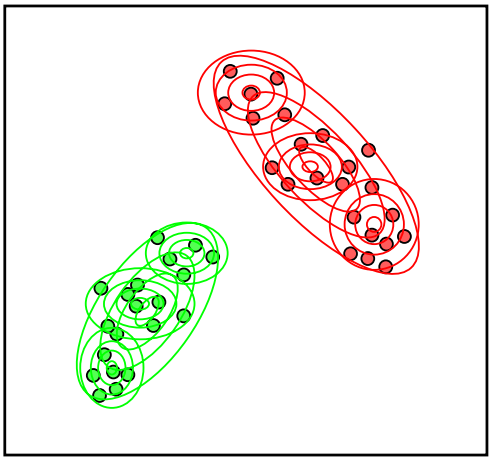
Multi-modale Verteilungen

- Vor allem verbreitet in Anwendungen wie z.B. Video-Suche
- Sequenz der Bilder formt nicht einzelne Gauss-Kurve, sondern eine beliebige Verteilung mit mehreren Peaks
- Kann häufig durch ein Mix mehrerer Gauss-Kurven gut approximiert werden



Verteilungen mit Kovarianz

- Nicht achsenparallel ausgerichtete Gausskurven
- Sondern mit beliebiger Rotation (durch Kovarianz-Matrix)
- Anzahl einzelner Instanzen oft deutlich verringerbar
- Erhöhte Speicherkomplexität pro Instanz: $O(d^2)$



Weiteres Future Work

- Weitere Verteilungsfunktionen:
  - Laplace-Verteilung
  - Multinomial-Verteilung (relevant für Text-Ähnlichkeit)
- Behandlung von Missing Values und Missing Objects
- Bereichsanfragen mit nicht-rechteckiger Gestalt
- Kostenmodelle und Optimierung
- Weitere Anwendungen:
  - Phenotype Description
  - Biometric Identification
  - Selectivity Estimation