Similarity Estimation using Bayes Ensembles

Marisa Thoma

joint work with

Franz Graf, Tobias Emrich, Hans-Peter Kriegel and Matthias Schubert

Ludwig-Maximilians-Universität München
Database Group

Scientific and Statistical Database Management
22nd International Conference, SSDBM 2010
Heidelberg, Germany, June/July 2010

Outline

- Similarity measures
  - Requirements and pitfalls
- Bayes Ensemble Distance (BED)
  - Bayes Estimates (BE)
  - Relevance weights
  - Feature space optimization
- Experiments
- Summary and Outlook

Similarity Measures

Pair-wise object comparison

Objects $x_1, x_2 \in \mathbb{R}^d$

$s(x_1, x_2)$: Similarity of x_1 and x_2

$s(x_1, x_2) > s(x_1, x_3) \Rightarrow x_1$ more similar to x_2 than to x_3

Applications:

- Similarity / retrieval queries (ranking, range queries)
- Clustering
- Model Training

Distance Measures

Pair-wise object comparison

Distance Measures: $s(x_1, x_2) \sim 1 / d(x_1, x_2)$

e.g. L_p-norm: $L_p(x_1, x_2) = \left(\sum_{i=1}^{d} \left| x_{1,i} - x_{2,i} \right|^p \right)^{1/p}$

Problems:

- Target similarity is ignored
- Correlated Dimensions
- Irrelevant Dimensions
- 1 Dimension has a large influence on small distances
- Distances can grow arbitrarily large

Distance Measures

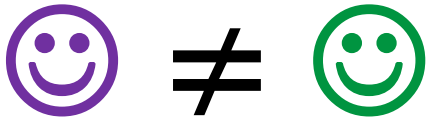
Pair-wise object comparison

Distance Measures: $s(x_1, x_2) \sim 1 / d(x_1, x_2)$

e.g. L_p-norm: $L_p(x_1, x_2) = \left(\sum_{i=1}^{d} \left| x_{1,i} - x_{2,i} \right|^p \right)^{1/p}$

Problems:

- Target similarity is ignored
- Correlated Dimensions
- Irrelevant Dimensions
- 1 Dimension has a large influence on small distances



- Distances can grow arbitrarily large

Distance Measures

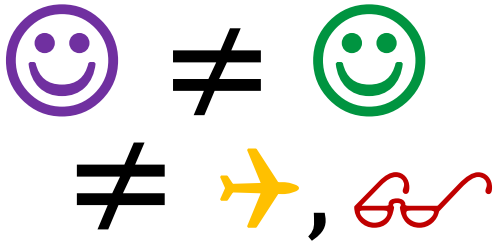
Pair-wise object comparison

Distance Measures: $s(x_1, x_2) \sim 1 / d(x_1, x_2)$

e.g. L_p-norm: $L_p(x_1, x_2) = \left(\sum_{i=1}^{d} \left| x_{1,i} - x_{2,i} \right|^p \right)^{1/p}$

Problems:

- Target similarity is ignored
- Correlated Dimensions
- Irrelevant Dimensions
- 1 Dimension has a large influence on small distances
- Distances can grow arbitrarily large



Distance Measures

Pair-wise object comparison

Distance Measures: $s(x_1, x_2) \sim 1 / d(x_1, x_2)$

e.g. L_p-norm: $L_p(x_1, x_2) = \left(\sum_{i=1}^{d} \left| x_{1,i} - x_{2,i} \right|^p \right)^{1/p}$

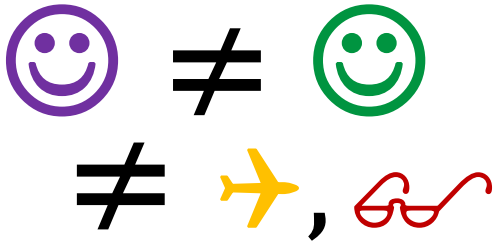
Problems:

- Target similarity is ignored
- Correlated Dimensions
- Irrelevant Dimensions

*Can be solved via **Mahalanobis distance**:*

$$D_{\text{Mah}}(x_1, x_2) = \left((x_1 - x_2)^T A (x_1 - x_2) \right)^{1/2}$$

- 1 Dimension has a large influence on small distances
- Distances can grow arbitrarily large



Distance Measures

Pair-wise object comparison

Distance Measures: $s(x_1, x_2) \sim 1 / d(x_1, x_2)$

e.g. L_p-norm: $L_p(x_1, x_2) = \left(\sum_{i=1}^{d} \left| x_{1,i} - x_{2,i} \right|^p \right)^{1/p}$

Problems:

- Target similarity is ignored
- Correlated Dimensions
- Irrelevant Dimensions

*Can be solved via **Mahalanobis distance**:*

$$D_{\text{Mah}}(x_1, x_2) = \left((x_1 - x_2)^T A (x_1 - x_2) \right)^{1/2}$$

- 1 Dimension has a large influence on small distances
- Distances can grow arbitrarily large

Bayes Ensemble Distance

Dissimilarity Probabilities of object pairs:

Bayes classifier for a feature i:

x_1, x_2 dissimilar (DIS): $P(\text{DIS} \mid (x_{1,i} - x_{2,i}))$

Bayes Estimate:

Priors for similar (p_{SIM}) and dissimilar objects (p_{DIS}):

$$\text{BE}_i(x_1, x_2) = \frac{p_{\text{DIS}} \cdot P\left((x_{1,i} - x_{2,i}) \mid \text{DIS}\right)}{p_{\text{DIS}} \cdot P\left((x_{1,i} - x_{2,i}) \mid \text{DIS}\right) + p_{\text{SIM}} \cdot P\left((x_{1,i} - x_{2,i}) \mid \text{SIM}\right)}$$

Bayes Ensemble Distance:

$$\text{BED}(x_1, x_2) = \frac{1}{d} \cdot \sum_{i=1}^{d} \text{BE}_i(x_1, x_2)$$

Why the Average?

Bayes Ensemble Distance:

$$\mathrm{BED}(x_1, x_2) = \frac{1}{d} \cdot \sum_{i=1}^{d} \mathrm{BE}_i(x_1, x_2)$$

Naïve Bayes Classifier:

$$\mathrm{NB}(x_1, x_2) = \frac{1}{\text{scale}} \cdot \prod_{i=1}^{d} \mathrm{BE}_i(x_1, x_2)$$

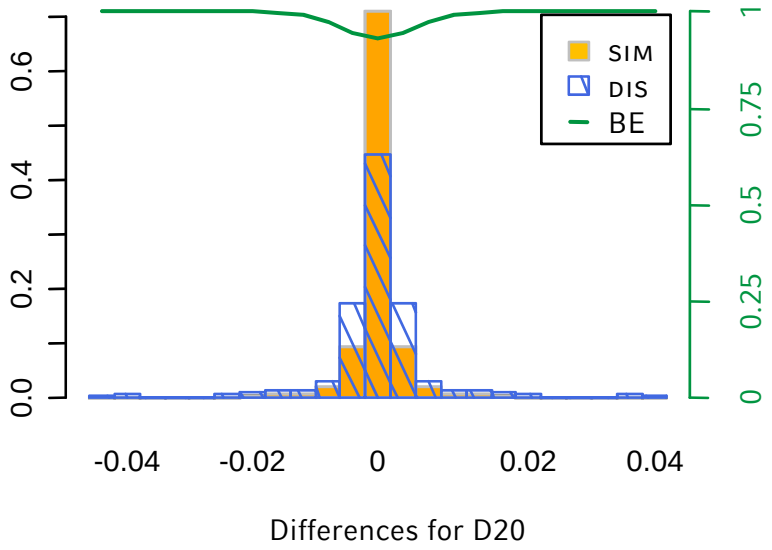
- BED also results in scores in [0,1]
- BED is more stable against outlier dimensions

Feature Differences

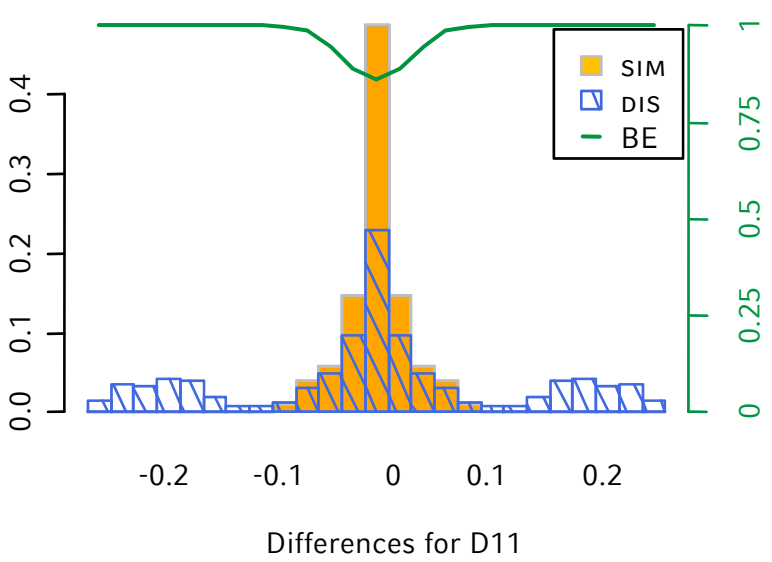
Examples:

Color Histogram Data

Suboptimal Dimension



Good Dimension



Feature Weighting

Introduce feature weights:

$$\mathrm{BED}(x_1, x_2) = \left(\sum_{i=1}^{d} q_i \right)^{-1} \cdot \sum_{i=1}^{d} q_i \cdot \mathrm{BE}_i(x_1, x_2)$$

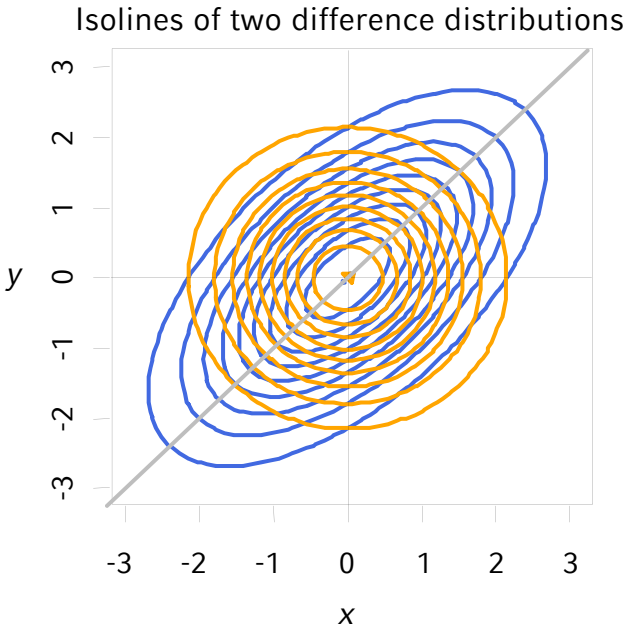
Assuming a Gaussian distribution:

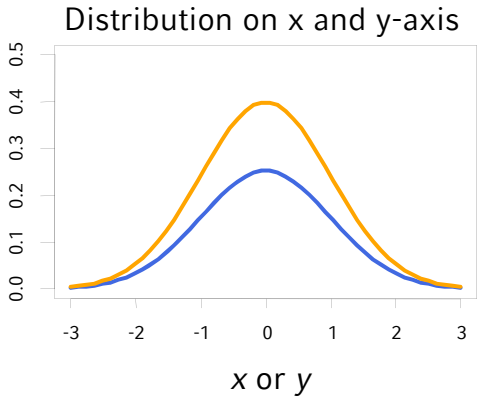
Use variance difference as quality measure

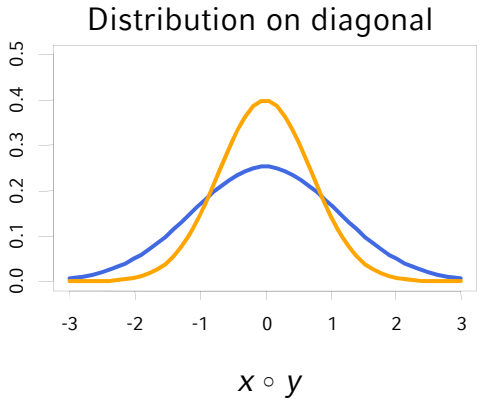
$$q_i = \sigma^2_{\mathrm{DIS}} - \sigma^2_{\mathrm{SIM}} = \underset{x_d \in \mathrm{DIS}}{\mathrm{avg}}\left(x^2_{d,i}\right) - \underset{x_s \in \mathrm{SIM}}{\mathrm{avg}}\left(x^2_{s,i}\right)$$

Feature Space Optimization

Exploit Feature Correlation



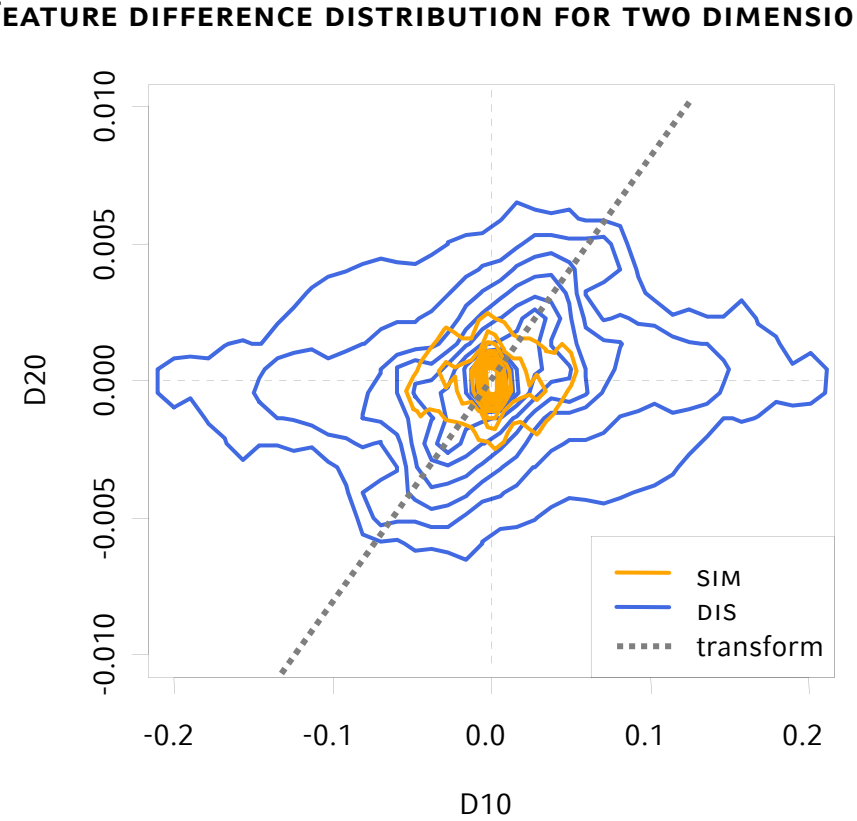




Feature Space Optimization

Example:

Color Histogram Data



Feature Space Optimization

Maximize variance difference

Transform to other space via $W = (w_1, \ldots, w_{d^*})$:

$$\begin{aligned} \max \quad & w_i^T \cdot \left(\Sigma_{\text{DIS}} - \Sigma_{\text{SIM}}\right) \cdot w_i \\ \text{s.t.} \quad & w_i \perp w_j \;\; \forall i, j \in \{1, \ldots, d^*\} \end{aligned}$$

$$\Sigma_{\text{SIM}} = \sum_{x_s \in \text{SIM}} x_s^T \cdot x_s \,, \quad \Sigma_{\text{DIS}} = \sum_{x_d \in \text{DIS}} x_d^T \cdot x_d$$

This is equivalent to solving the EVD:

$$\lambda\, w = \left(\Sigma_{\text{DIS}} - \Sigma_{\text{SIM}}\right) \cdot w$$

BED Algorithm

Train_BED (X with $x_i \in \mathbb{R}^d$, sim, dis, d^*):

1. Derive Σ_{SIM} and Σ_{DIS}
2. Compute feature transformation $W \in \mathbb{R}^{d,d^*}$
3. Get weights q_i ($i \in \{1, \ldots, d^*\}$) for new feature space $W^T X$ using the features' variance differences

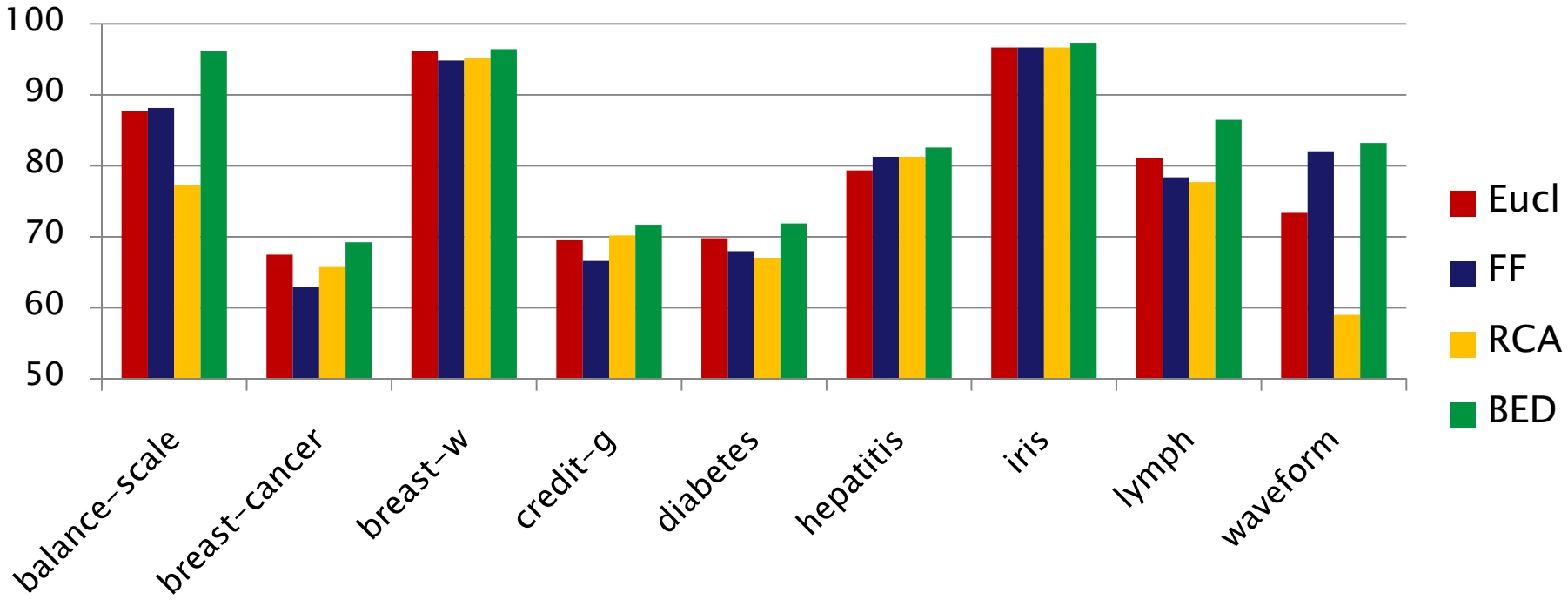
Output:

$$\mathrm{BED}(x_1, x_2) = \left(\sum_{i=1}^{d^*} q_i \right)^{-1} \cdot \sum_{i=1}^{d^*} q_i \cdot \mathrm{BE}_i(W^T x_1, W^T x_2)$$

Experiments

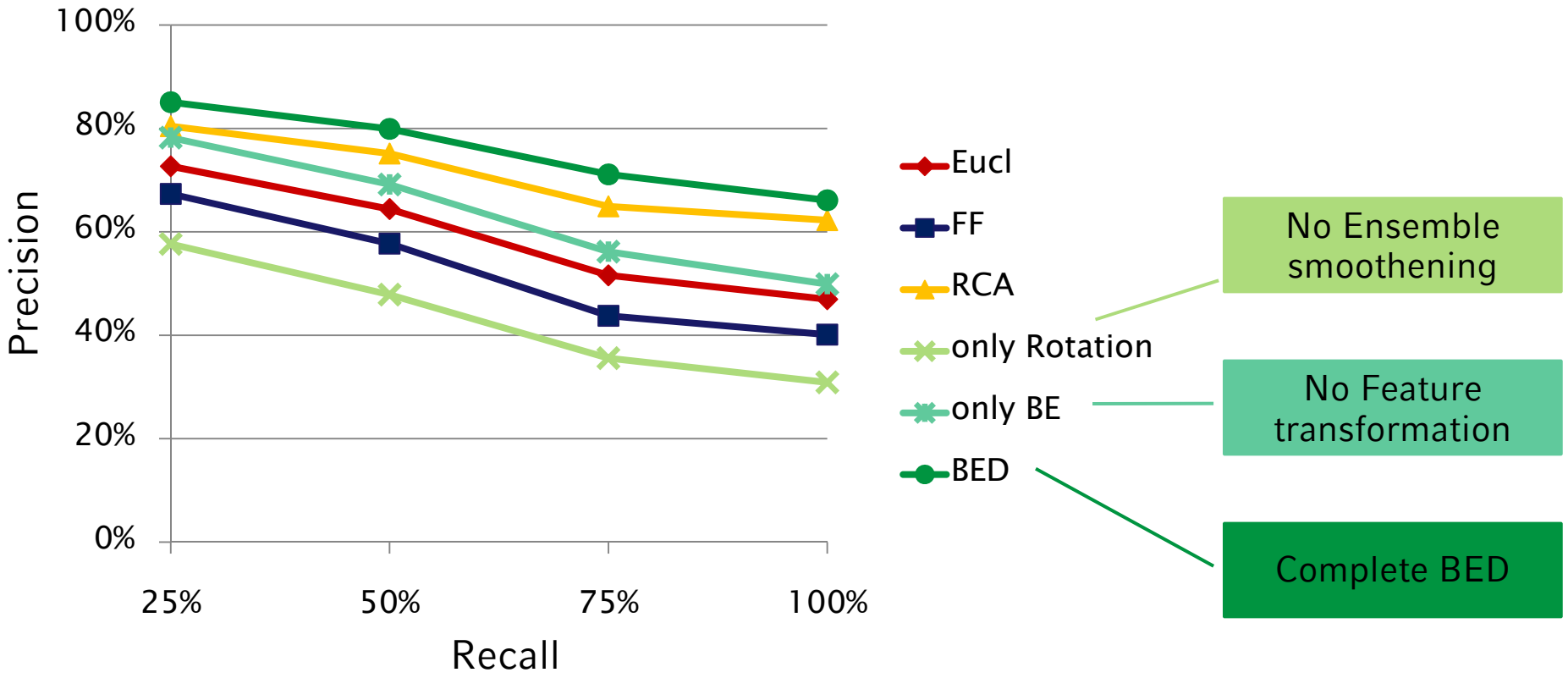
UCI classification datasets:

Accuracy Comparison to Euclidian Distance (Eucl), Fisher Faces (FF) and Relevant Component Analysis (RCA)



Experiments

Image Retrieval Dataset (Precision Recall Curves)



Conclusion and Outlook

BEDs now:

- Balanced, adaptive distance measure
- Easily interpretable
- Flexible w.r.t. data input

Improvements:

- Runtime / Indexing ideas
- Local adaptivity
- Alternative weighting scheme

Thank you.