A Comparison of Dimensionality Reduction Techniques in Scientific Applications

Ya Ju Fan & Chandrika Kamath

Lawrence Livermore National Laboratory

May 23, 2012

LLNL-PRES-543451. This work was funded by the SciDAC-e project, "MINDES: Data Mining for Inverse Design. This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Dimension Reduction

Feature Subset Selection

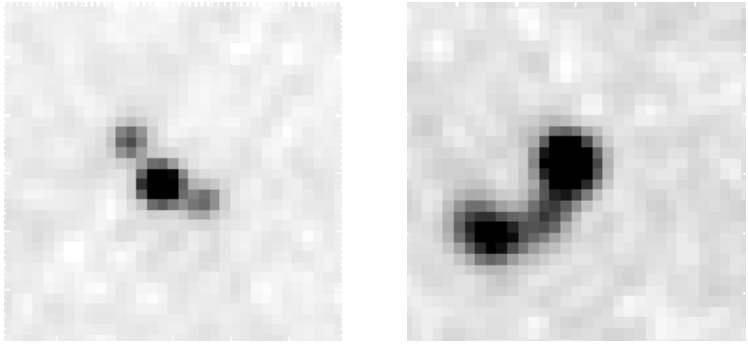
- Stump Filter
- Distance Filter
- Chi-squared Filter
- ReliefF

Data Transformation

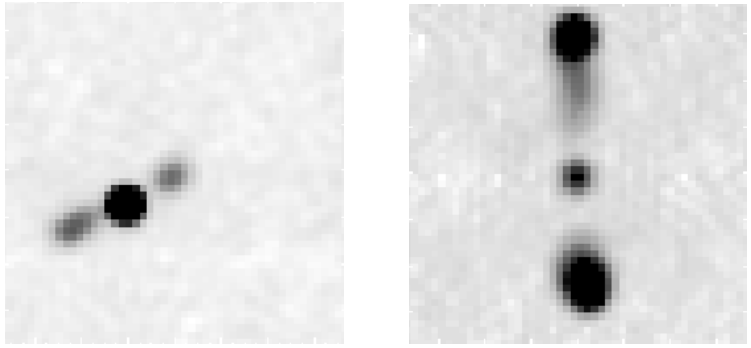
- Principal Component Analysis (PCA)
- Isomap (Tenenbaum et al. 2000)
- Locally Linear Embedding (LLE) (Roweis & Saul 2000)
- Laplacian Eigenmaps (Belkin & Niyogi 2002)
- Local Tangent Space Alignment (LTSA) (Zhang & Zha 2002)

FirstTriples Data Set

Bents

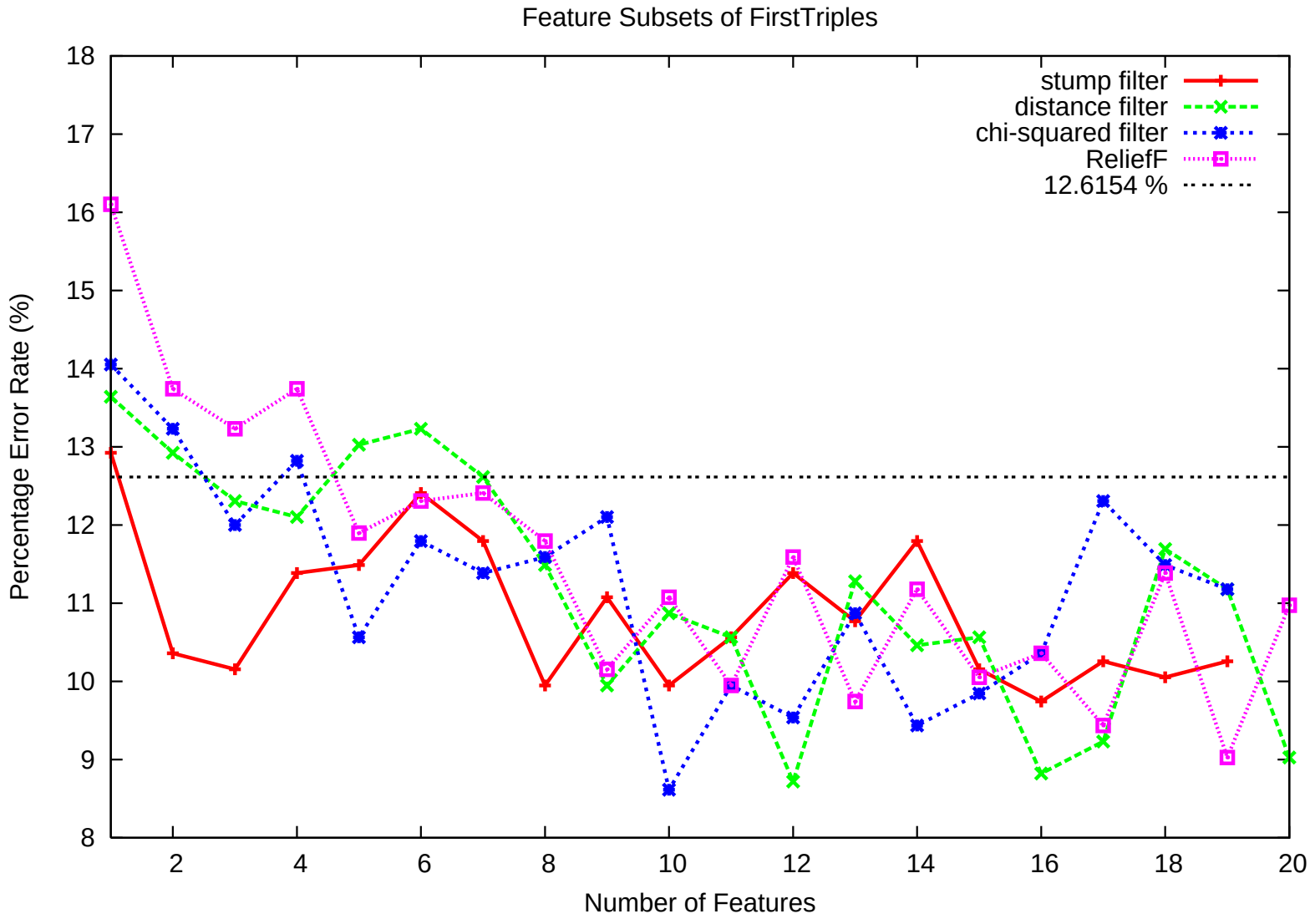


Non-bents

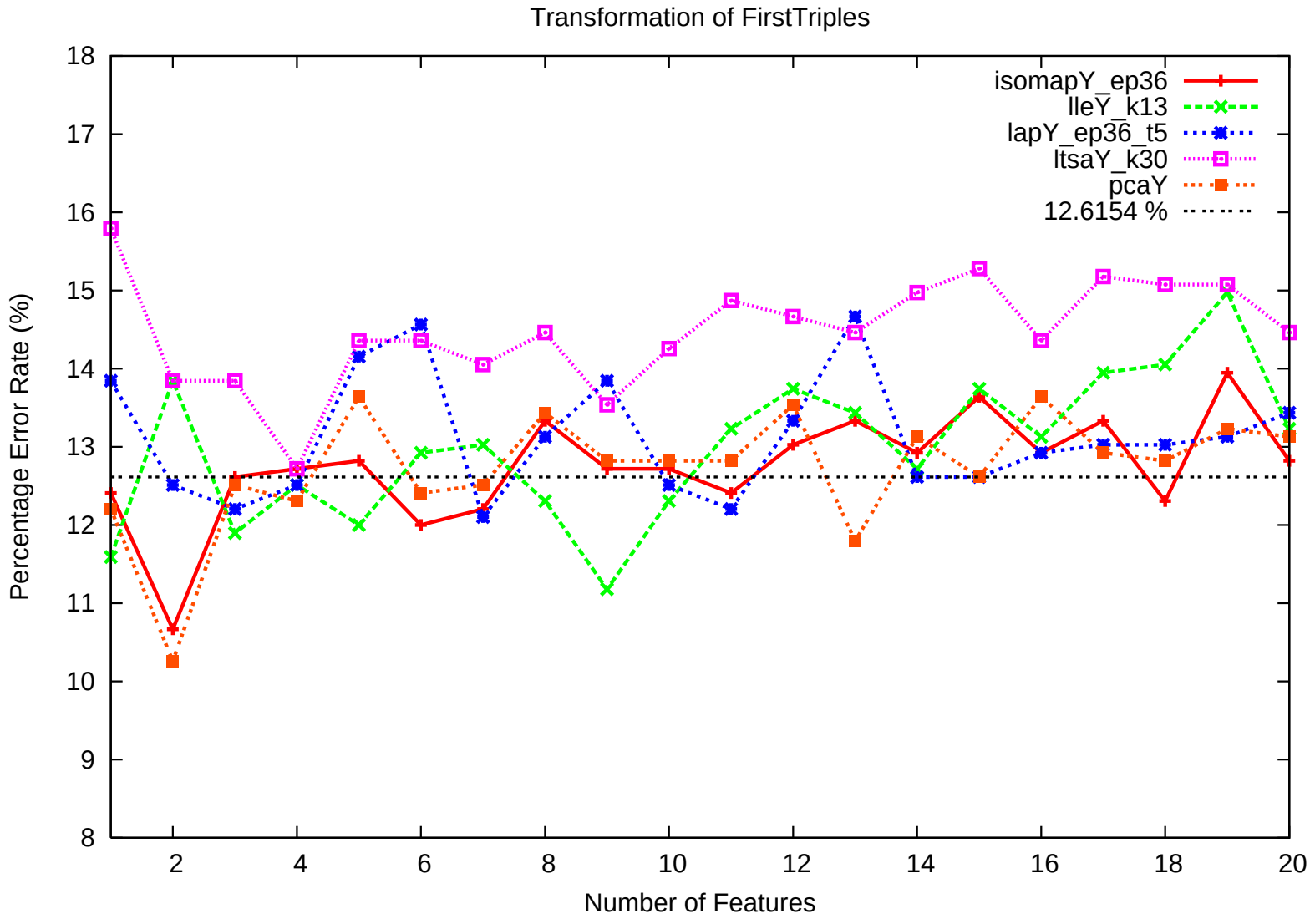


- Classify radio-omitting galaxies into two classes:
  - Bent - one with a bent-double morphology;
  - Non-bent - one without a bent.
- Raw image data are processed to create a catalog by fitting two-dimensional, elliptic Gaussians to each galaxy.
- Each entry in the catalog corresponds to a Gaussian and includes information such as
  - the location and size of the Gaussian
  - the major and minor axes of the ellipse
  - the peak flux

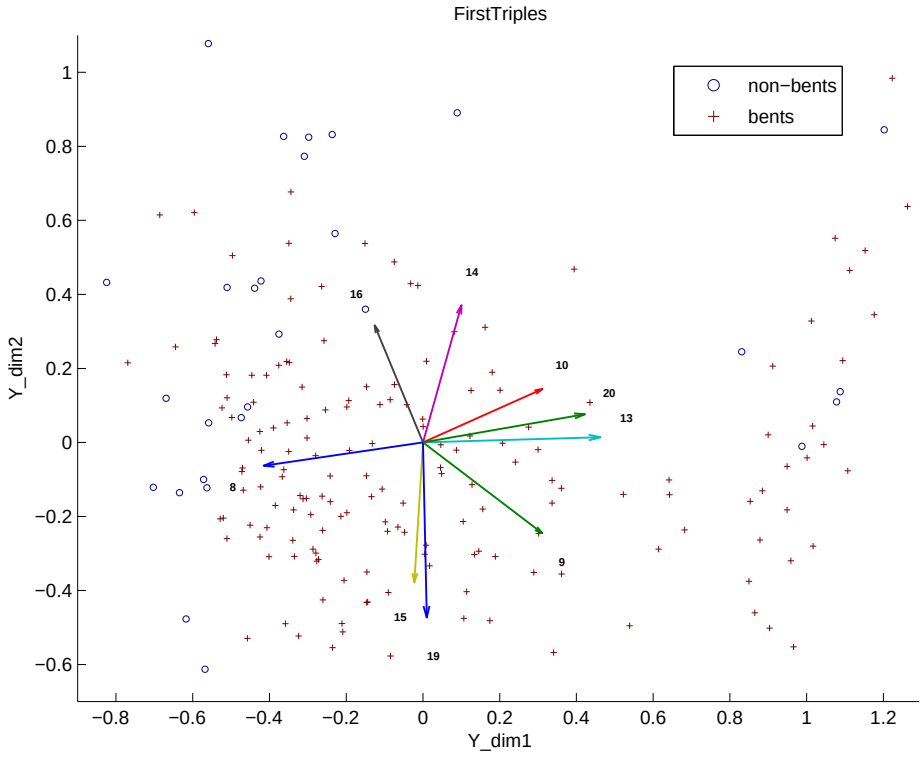
Feature Subsets of FirstTriples Data

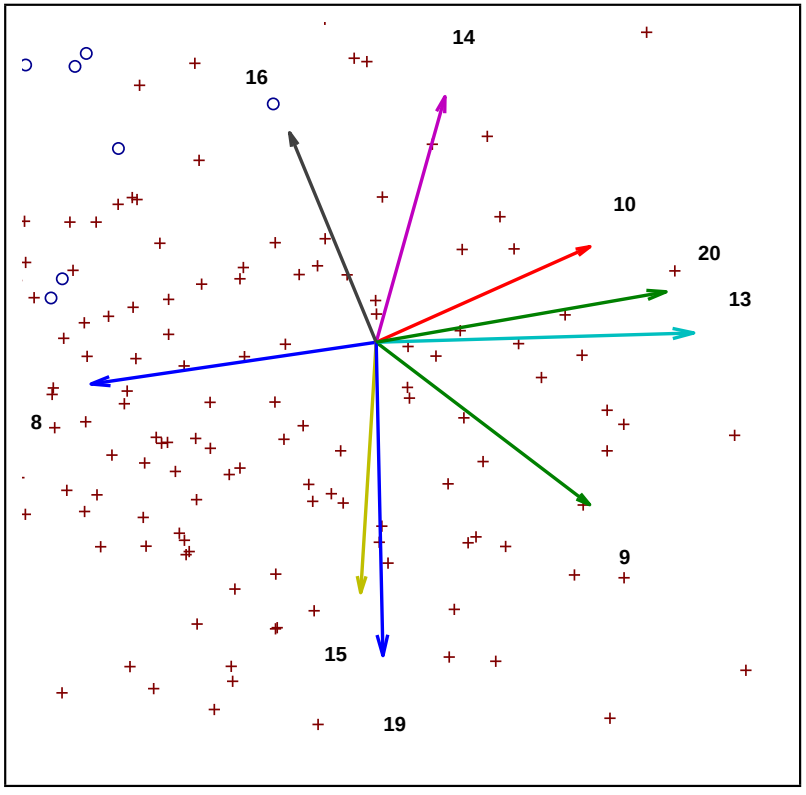


Data Transformation of FirstTriples Data

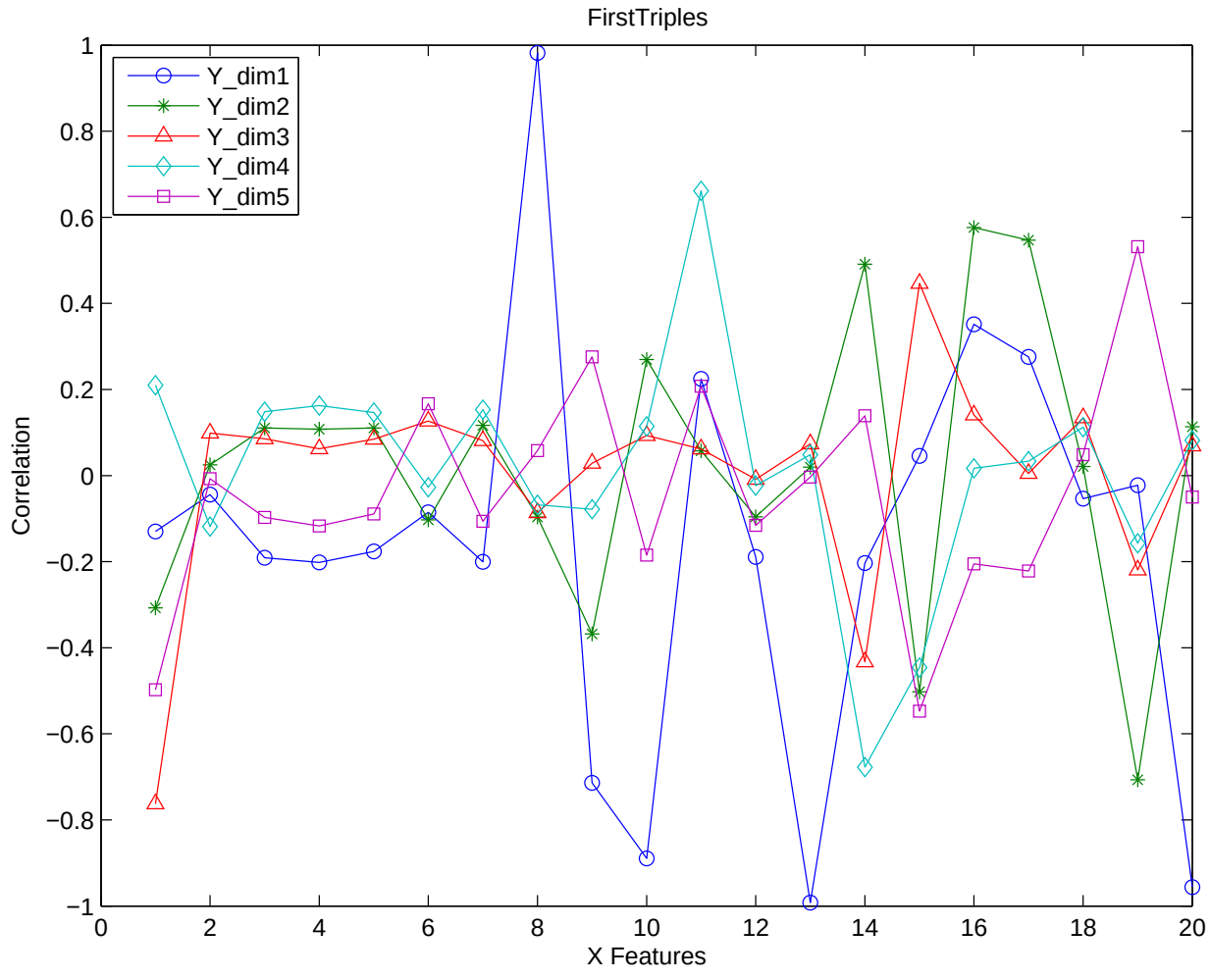


Left: PCA biplot of the *FirstTriples* dataset. Right: zoomed-in view





Correlation between top five Isomap reduced dimensions and all original features for FirstTriples Data



Significant Features

PCA:

- First coordinate:
  A negative correlation between feature CoreAngl (8) and angles (9, 10, 13 and 14) & symmetry (20)
- Non-bents fall around the extreme values of angle features.
- Second coordinate:
  symmetry (19) v.s. angle (14) & distance (16)

Isomap:

- First dimension:
  A negative correlation between feature CoreAngle(8) and angles (9,10 and 13) & symmetry (20)
- Second dimension:
  symmetry (19)

Feature subsets: symmetries and angles

Conclusion

Feature Subset Selection Techniques

- Supervised
- Consistently improve the classification
- The best: filter-based

Data Transformation Methods

- Non-supervised
- May be able to find properties related to class labels
- Those that employ the eigenvectors corresponding to the **largest** eigenvalues seem to perform better
- Best: PCA and Isomap
- Isomap may capture some nonlinear properties of the data that are useful