Deterministic projection by growing cell structure networks for visualization of high-dimensionality datasets

Jason W.H. Wong & Hugh M. Cartwright

Journal of Biomedical Informatics 38 (2005) 322–330

Introduction

- Dimensionality reduction becoming more important as more complex datasets are becoming available.
- Data visualization allows us to use our most specialized sense (vision) to comprehend the data (clusters, patterns, outliers).

Projection Methods

- PCA (principle component analysis): linear method
  - sometimes doesn't reveal clusters
  - must be recalculated as data points are added
- SOM (self-organizing semantic map) >
  GCS (growing cell structure map)
  - distance matrix technique (U-method)
  - visualization restricted by the number of neurons
  - susceptible to over-fitting of data
  - computational limitations as the network grows

Random projection

- Projects a given point *u* from k-dimensional space down to a *t* dimensional subspace (t x k matrix).

- Proven by showing that the squared length of a random vector is heavily concentrated about its mean when projected onto a random *t*-dimensional subspace, and is not distorted by more than $(1 \pm \epsilon)$ with a probability $O(1/n^2)$, where *n* is the number of points.

- Unstable, performance affected by choice of the random orthonormal matrix, R.

Deterministic projection

- Combination of random projection with a GCS
- Unlike random projection, as data is added, a best-matching adjustment unit and vector is computed, enhancing the matrix.
- GCS networks adopt fractal growth behavior, so a small network can suffice to map the input data.
- “A trained GCS network provides *c* neurons that are laid out across a two-dimensional map in such a way that neurons whose vectors are similar, as defined by Euclidean distance, are closer to each other on the map,” and vice versa.

The algorithm

I. Initialize a two-dimensional GCS network, A, and train it by following the GCS algorithm using dataset, D.
II. From the GCS neuron positions generate and neuron vectors to matrices generate $P_{[c\times 2]}$ and $V_{[c\times k]}$, respectively.
III. Perform mean centering on each matrix.

$$P'_{[c\times 2]} = P_{[i\times j]} - \overline{P_j} \quad \forall i = 1,\ldots,c \text{ and } \forall j = 1,2,$$

$$V'_{[c\times k]} = V_{[i\times j]} - \overline{v_j} \quad \forall i = 1,\ldots,c \text{ and } \forall j = 1,\ldots,k.$$

IV. Compute the product of the mean centered matrices as described in Eq. (6).
V. Perform mean centering on matrix S.

$$S'_{[2\times k]} = S_{[i\times j]} - \overline{s_j} \quad \forall i = 1,2 \text{ and } \forall j = 1,\ldots,k.$$

VI. Apply the Gram–Schmidt orthogonalization algorithm [18] to S.
VII. Compute the product of S and data points in D as in Eq. (7).

Illustrative example

cube dataset

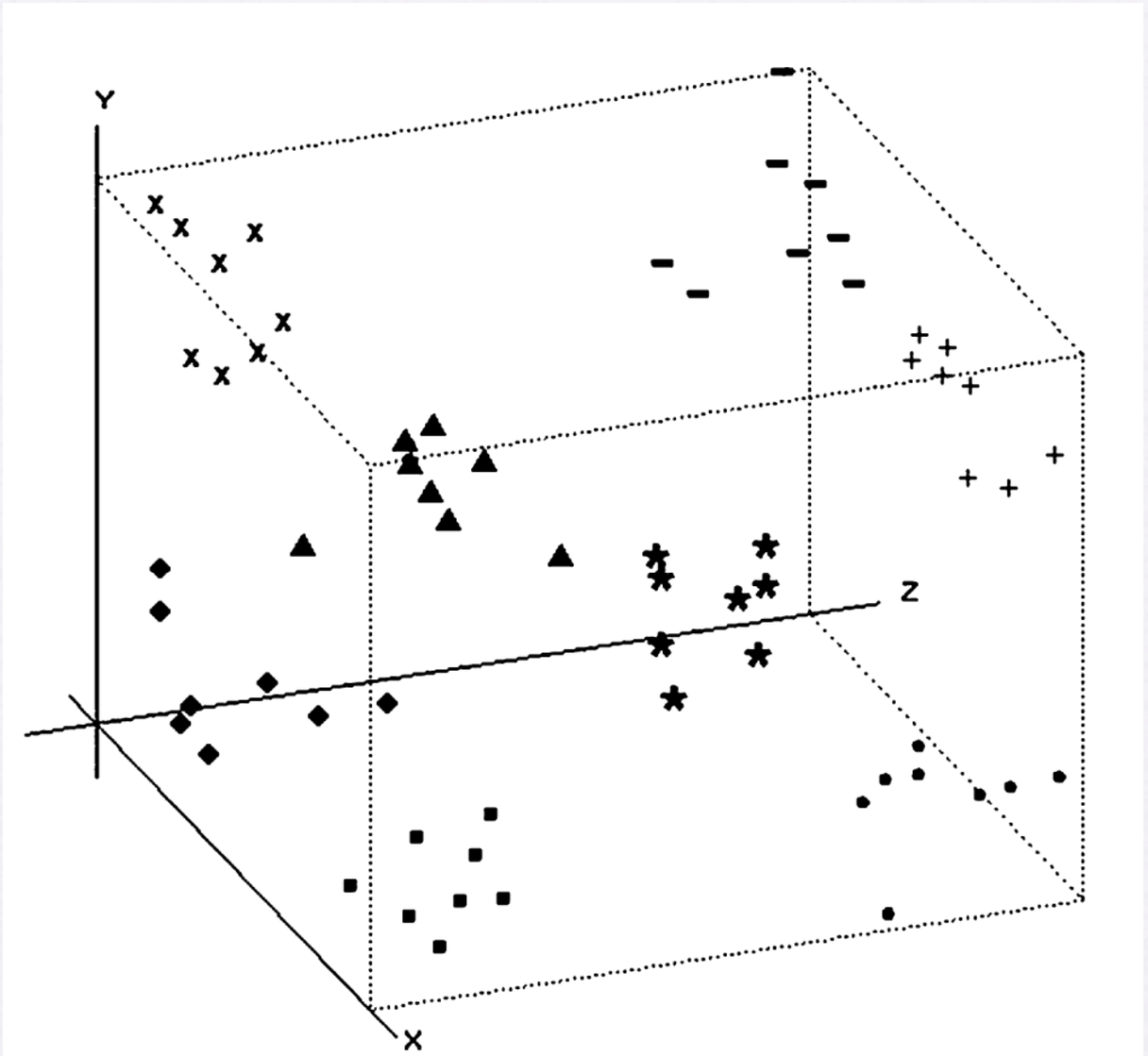
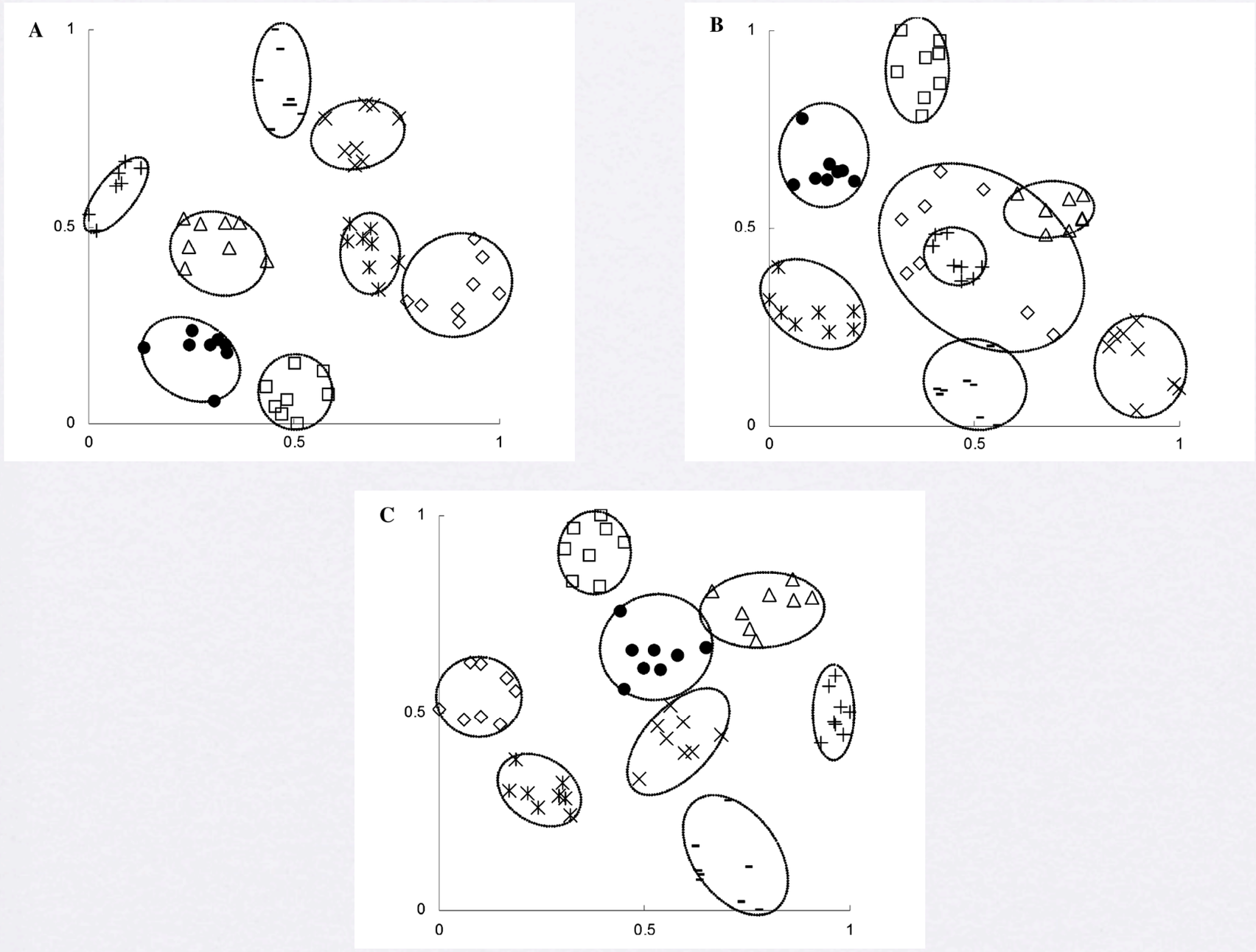


Fig. 1. Cube dataset shown as a three-dimensional scatter plot.

Illustrative example

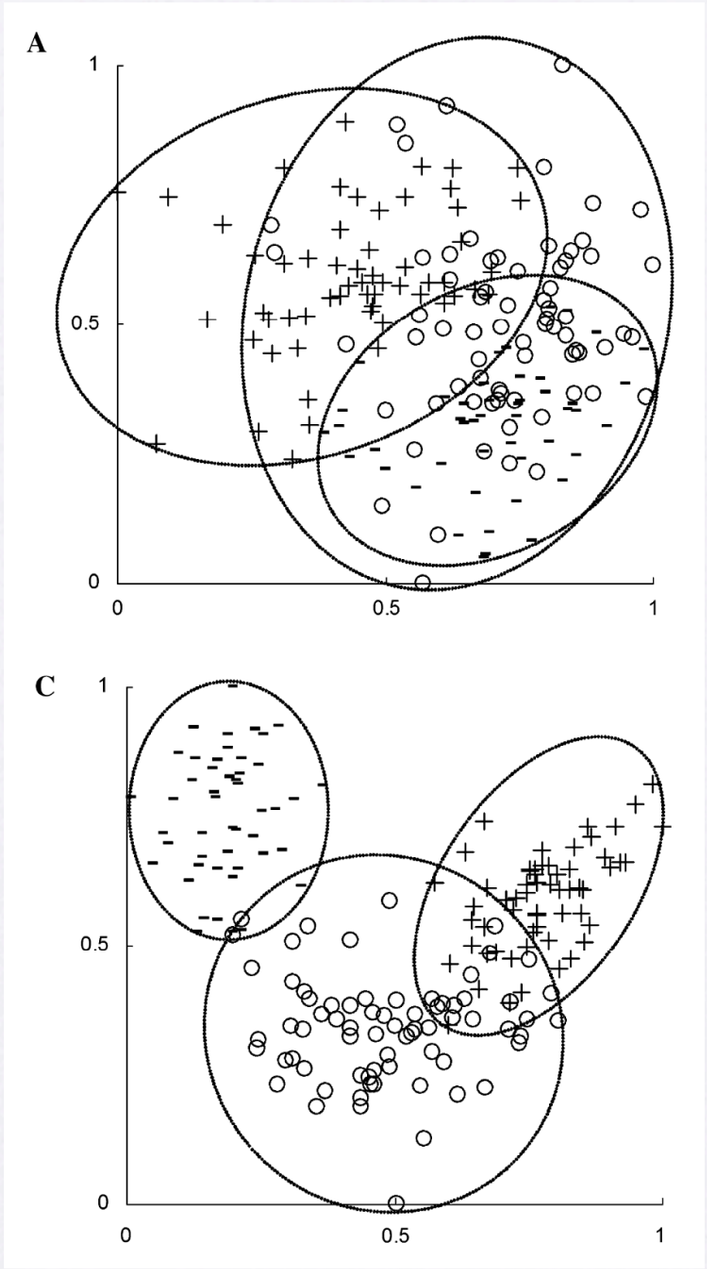
cube dataset

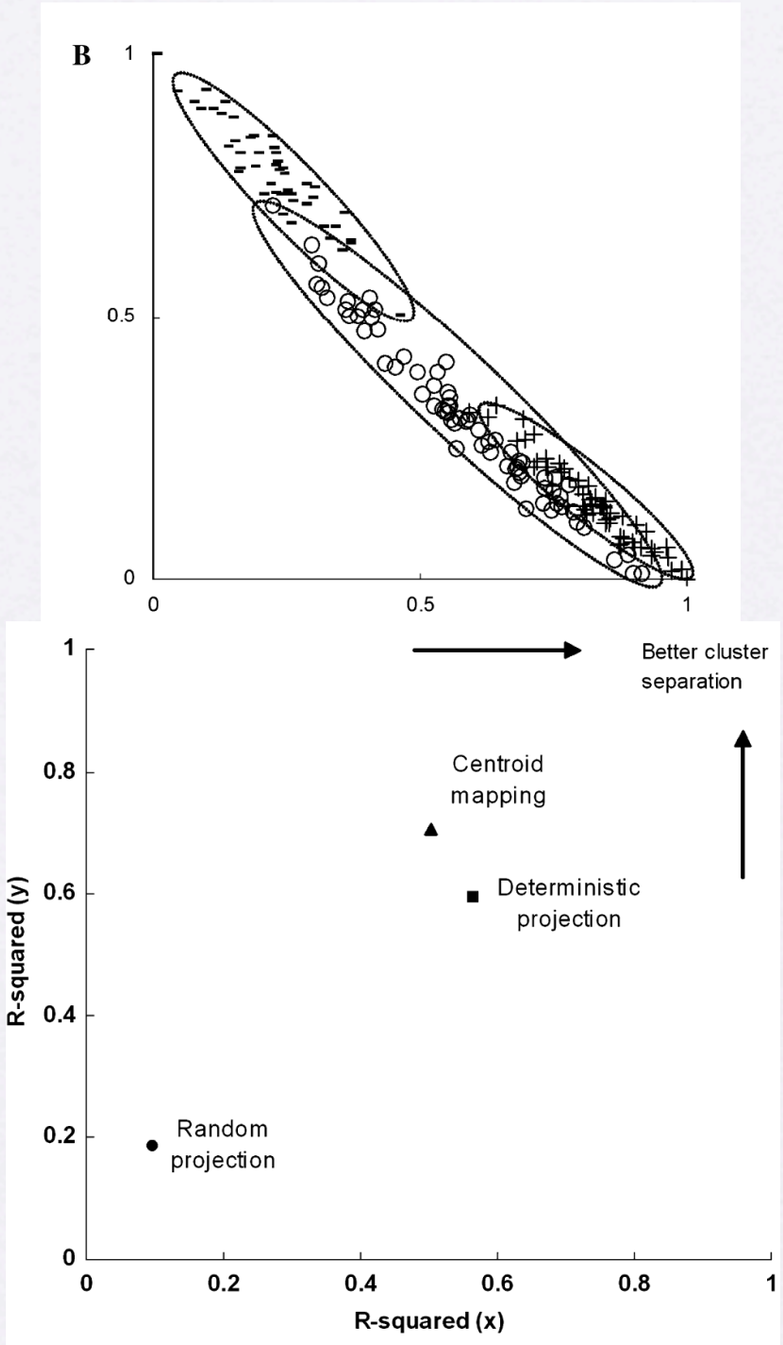


(A) Deterministic projection (B) Centroid mapping (C) Random projection

Illustrative example

wine dataset

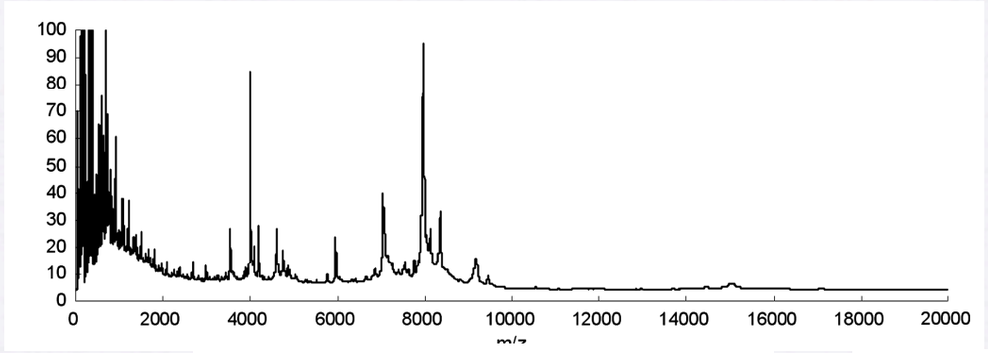




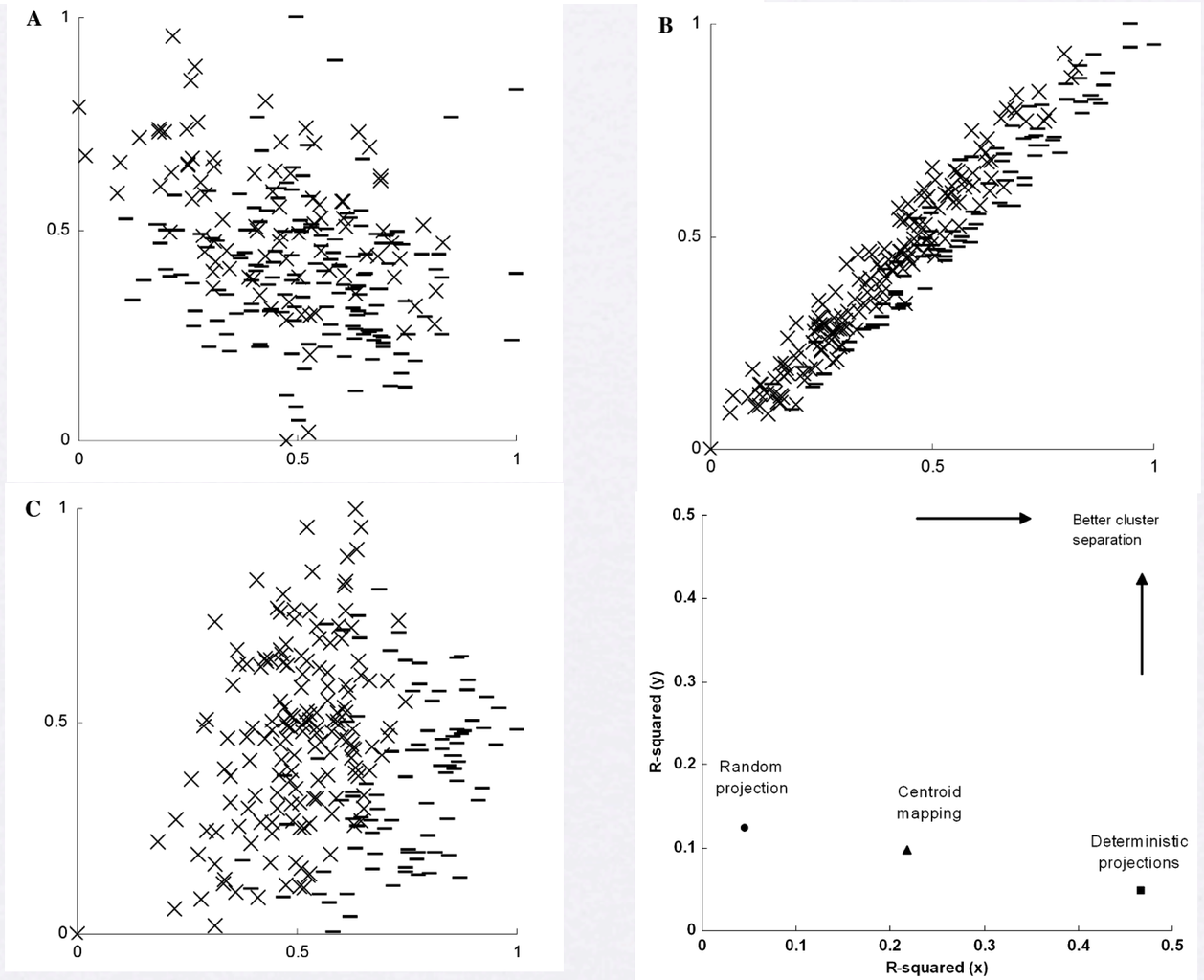
(A) Deterministic projection (B) Centroid mapping (C) Random projection

Illustrative example

clinical proteomics dataset



Typical mass spectrum of proteomics dataset.



(A) Deterministic projection (B) Centroid mapping (C) Random projection

Discussion

Table 1
Leave-one-out cross-validation of the wine dataset using different projection methods

Projection method	Correctly classified (%)[a]
Deterministic projection	90.8 ± 0.9
Centroid mapping	84.3 ± 1.5
Random projection	71.8 ± 4.0

[a] Correct classification of a data point is evaluated using the *k*-nearest neighbor method. The percentage is an average of 20 independent projections.

- Effective for mapping multi-dimensional datasets, even with blind test data (wine set).
- More stable than random projection, average performance higher, standard deviation lower (Table 1).
- Better than centroid mapping at depicting cluster separation.
- Deterministic projection can be applied rapidly without the need to pre-select features within the dataset before projection.
- Much faster (2 min vs 1 sec).
- Not limited to visualization in two dimensions. Possibly arbitrarily high dimensions.

Discussion

- Can be applied using genetic algorithms to discover features within a dataset that results in optimum visual separation between known classes.
- "This may prove to be a particularly relevant application in biomarker discovery from clinical proteomic mass spectrometry data. In rapidly growing research fields such as microarray analysis and biomarker discovery, it is of great interest to explore patterns within very high dimensionality data. The development of a deterministic projection method for data visualization should aid such data analyses by providing a way to deduce patterns in data using the power of our visual perception."