A Novel Stability based Feature Selection Framework for *k*-means Clustering

Dimitrios Mavroeidis and Elena Marchiori
Radboud University Nijmegen, The Netherlands

Presentation outline

- Main novelty of proposed framework
- Preliminary notions
  - *k*-means and PCA
  - stability of PCA
  - feature selection and sparse PCA
- Proposed framework
- Empirical results and further work

What's new

- Various conceptually different approaches for f.s.
- Most based on the notion of "relevant" features
- In the context of this work we adopt a bias-variance perspective:
  - Feature contribution to cluster separation vs. contribution to variance
  - Achieved through stability maximizing Sparse PCA
  - Novel greedy algorithm that optimizes a lower bound of the objective

k-means and PCA

- *k*-means objective

$$J_K = \sum_{k=1}^{K} \sum_{1 \in C_k} ||x_i - \mu_k||^2$$

- popular heuristic is Lloyd's algorithm
  - EM-style
  - iteratively updates cluster centers and assigns objects to closest centers
- alternative approach is PCA-based approximation
  - start with discrete cluster assignment problem
  - relax discrete problem to continuous
  - continuous *k*-means solution is derived by the eigenvectors of the Covariance matrix: PCA

Feature selection and Sparse PCA

- "Baseline" feature selection for *k*-means
  - select subset of features that "approximates" k-means objective
- "Baseline" feature selection for continuous PCA-based k-means
  - Select subset of features that "approximates" k-means continuous objective
- In PCA based k-means
  - objective function = eigenvalues of covariance matrix
  - features = rows and columns of covariance matrix
- Feature selection = select rows and columns of covariance matrix such that the eigenvalues are best approximated
  - Sparse PCA!

Stability of PCA

- Stability of the eigenvector solution is measured through the size of the relevant eigengap
- Stability of *k-1* dominant eigenvectors depends on the size of the eigengap

$$\lambda_{k-1} - \lambda_k$$

- Feature selection that maximizes stability?
- What are the semantics?

Stability maximizing sparse PCA

- Stability based f.s. is equivalent to cluster separation vs. Variance tradeoff

$$Cov = \mathbf{diag}(u) X_{fc} X_{fc}^T \mathbf{diag}(u)$$

$$\begin{aligned} Obj &= \max_{u \in \{0,1\}^m} \left(\tfrac{1}{n} \textstyle\sum_{i=1}^{n} (\lambda_1(Cov) - \lambda_i(Cov)) \right) \\ &= \max_{u \in \{0,1\}^m} \left(\tfrac{n-1}{n} \lambda_1(Cov) - \tfrac{1}{n} \textstyle\sum_{i=2}^{n} \lambda_i(Cov) \right) \\ &= \max_{u \in \{0,1\}^m} \left(\lambda_1(Cov) - \tfrac{1}{n} \mathbf{Trace}(Cov) \right) \end{aligned}$$

$$\max_{u \in \{0,1\}^m} \frac{n_1 n_2}{n} \left[2 \frac{d^{(u)}(c_1, c_2)}{n_1 n_2} - \frac{d^{(u)}(c_1, c_1)}{n_1^2} - \frac{d^{(u)}(c_2, c_2)}{n_2^2} \right] - \sum_{i=1}^{m} u_i \cdot \mathrm{var}(f_i)$$

$$d^{(u)}(c_i, c_j) = \sum_{k \in c_i} \sum_{l \in c_j} (x_k^{(u)} - x_l^{(u)})^2$$

Algorithmic approach

- We employ greedy forward search that optimizes lower bound of objective.
- Lower bound requires only 1 eigenvector computation per greedy step

$$Obj(I \cup \{m\}) \geq Obj(I) + B$$

where

$$\begin{aligned} B = (1 - \tfrac{1}{\boldsymbol{card}(I)+1})[(v^T x_{fc}(m)^2 & - \tfrac{1}{n} x_{fc}(m)^T x_{fc}(m)] \\ - \tfrac{2}{\boldsymbol{card}(I)+1}[(\textstyle\sum_{i \in I} v^T x_{fc}(i)) v^T x_{fc}(m) & - \tfrac{1}{n} (\textstyle\sum_{i \in I} x_{fc}(i))^T x_{fc}(m)] \\ + \tfrac{1}{\boldsymbol{card}(I)(\boldsymbol{card}(I)+1)}[(v^T \textstyle\sum_{i \in I} x_{fc}(i))^2 & - \tfrac{1}{n} (\textstyle\sum_{i \in I} x_{fc}(i))^T (\textstyle\sum_{i \in I} x_{fc}(i))] \end{aligned}$$

Algorithm

Algorithm 1 (X, p)

1: Initialize with index I_{k_0} where $i_0 = argmax_{j \in I} O_1\{j\}$.
2: **repeat**
3: Compute $i_k = argmax_{i \in I_k^c} B(i, I_k)$
4: Set $I_{k+1} = I_k \cup \{i_k\}$.
5: **until card**$(I_{k+1}) = p$.

Deflation for multiple eigenvectors

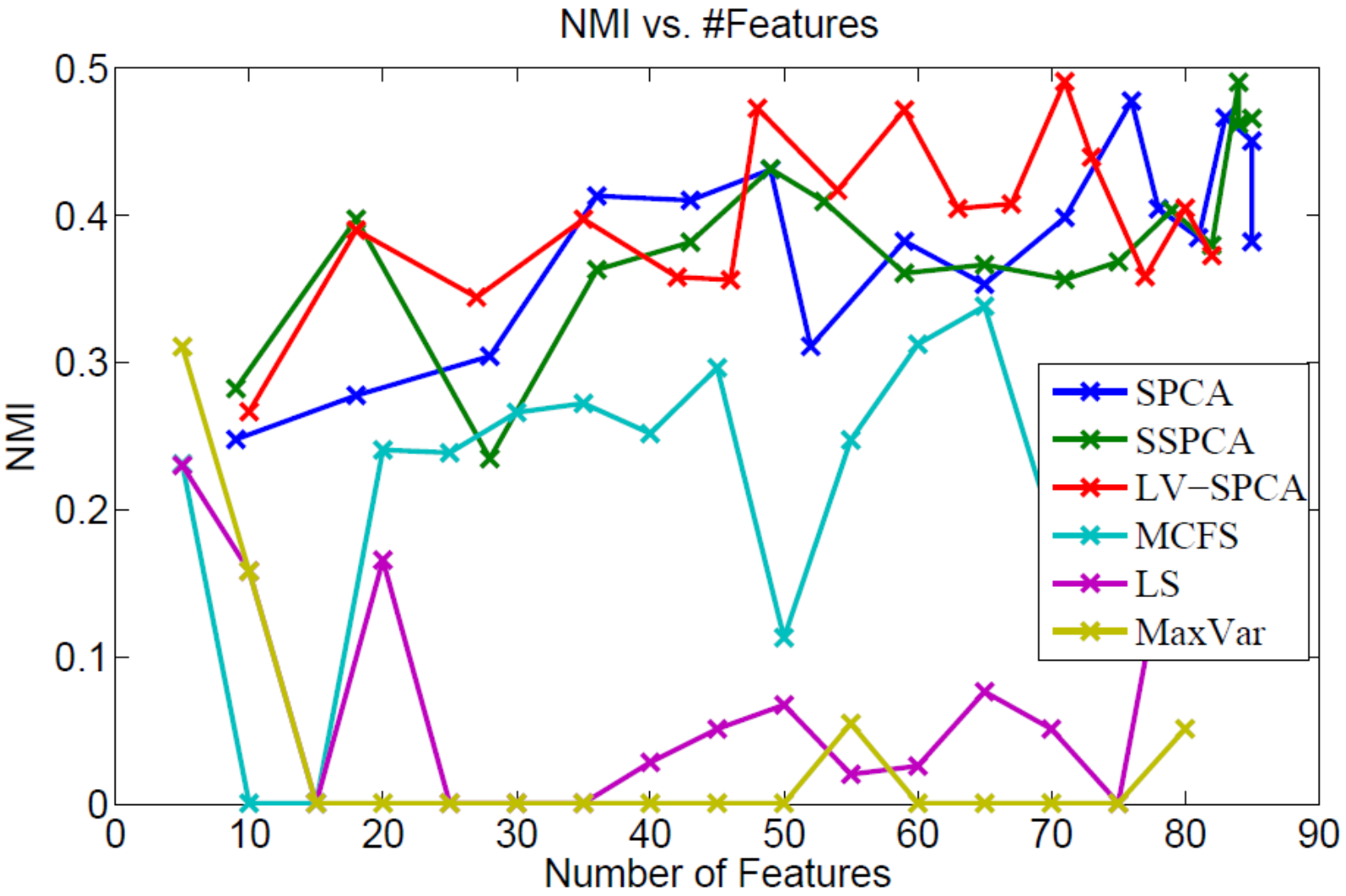
- For computing multiple sparse eigenvectors deflation is required
- In this paper we propose an efficient approach that is shown to be equivalent to Schur complement deflation

$$X_{fc}^{(t)} = X_{fc}^{(t-1)}(I - v_t v_t^T)$$

Empirical results

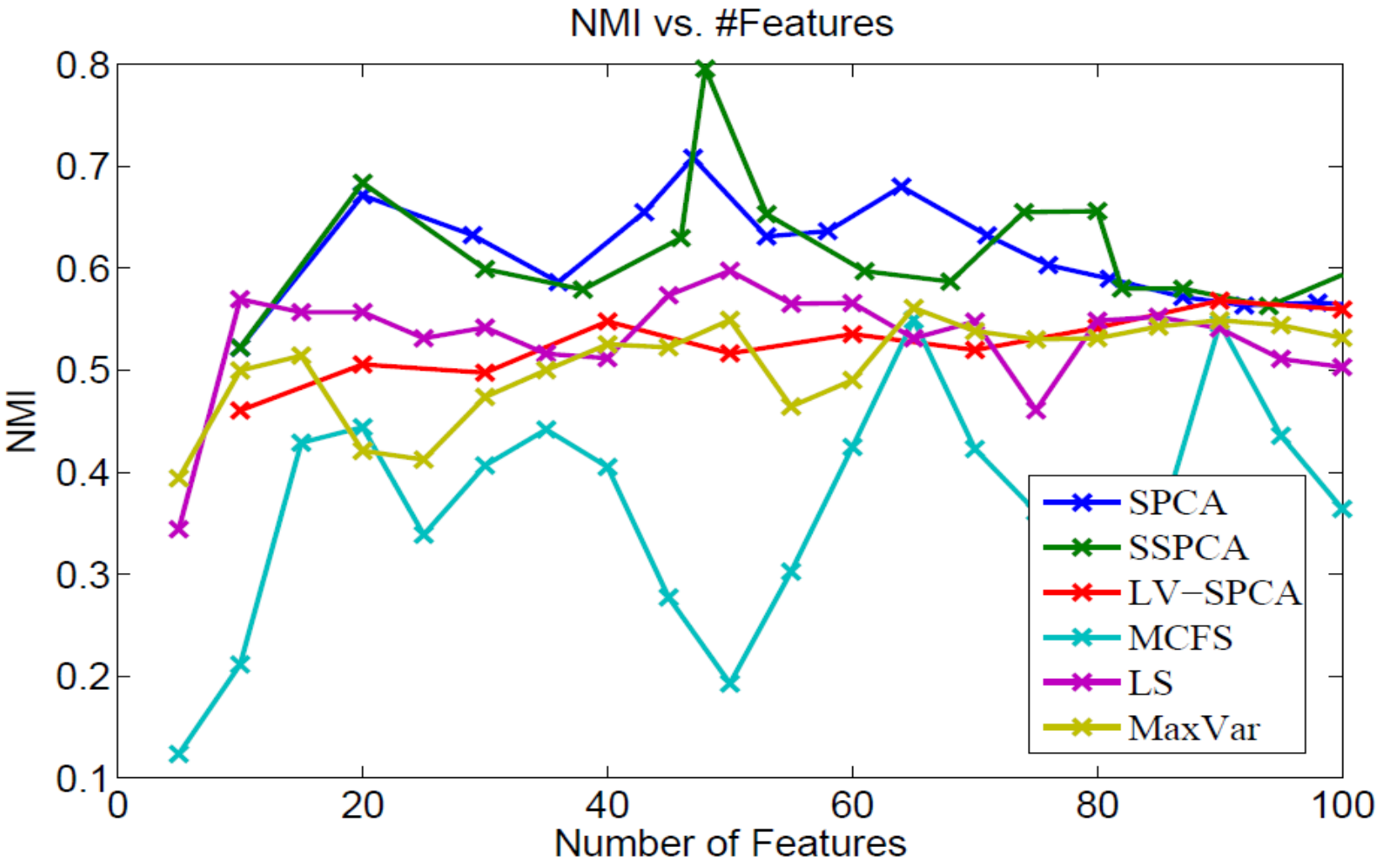
- 4 cancer research datasets
- 3 methods
  - SPCA: $\lambda_1(Cov)$
  - SSPCA: $\lambda_1(Cov) - \frac{1}{n}\mathbf{Trace}(Cov)$
  - LV-SPCA: $\lambda_1(Cov) - \gamma\mathbf{Trace}(Cov)$ $\gamma = \frac{\lambda_1(Cov_{full})}{\mathbf{Trace}(Cov_{full})}$
- Quantitative evaluation
  - clustering performance
- Qualitative evaluation
  - relevance of selected genes

Clustering (1)



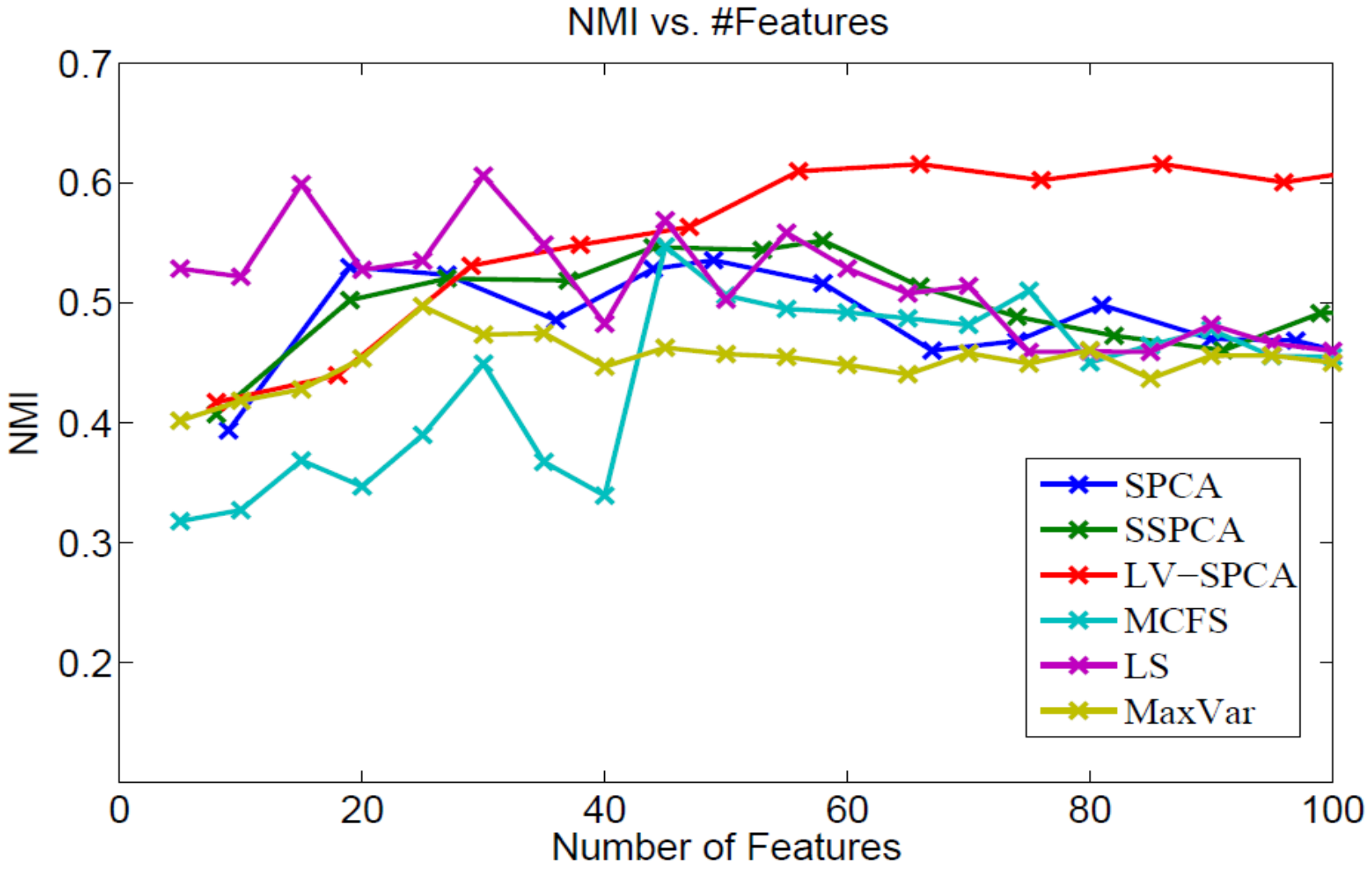
(a) Chen-2002

Clustering (2)



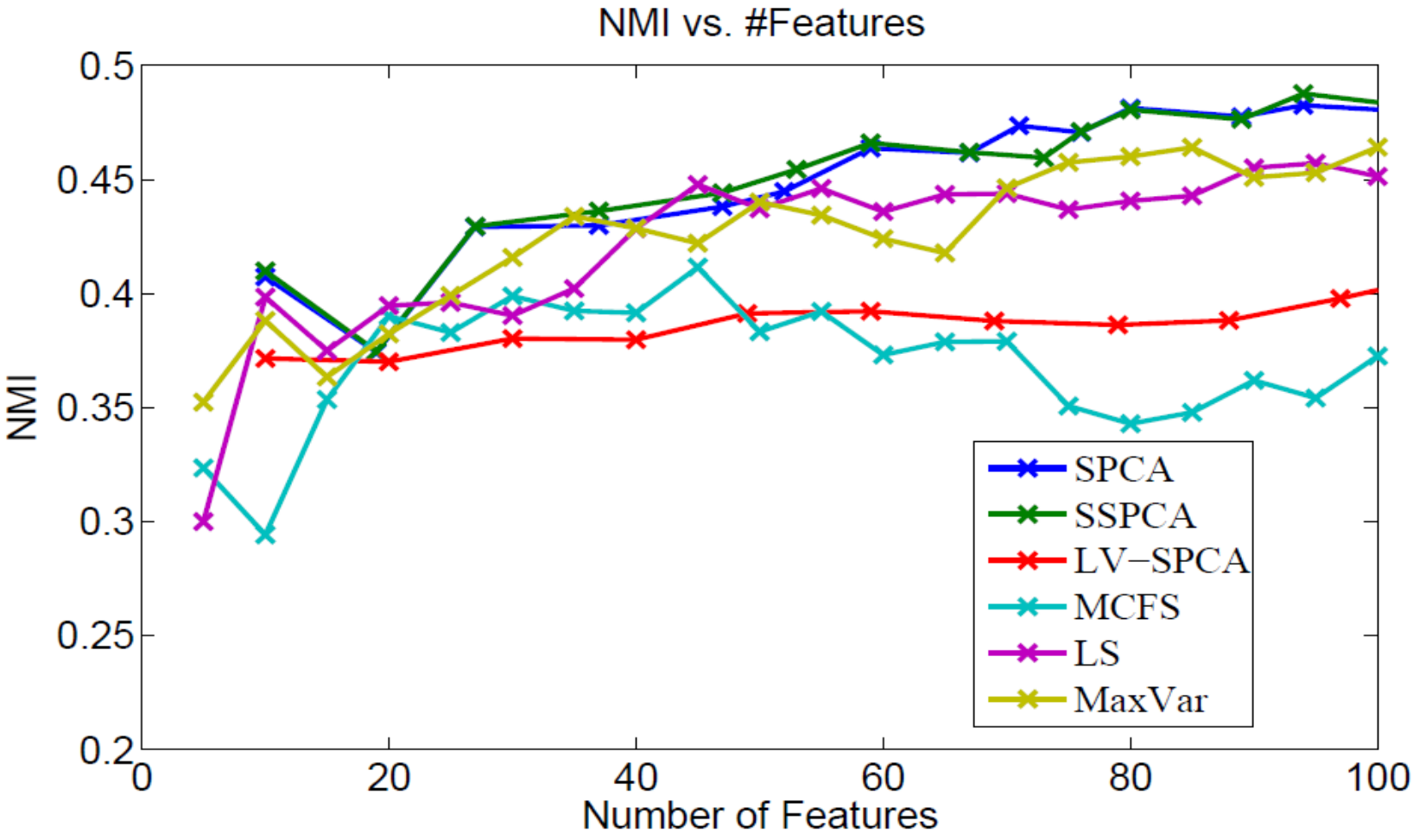
(b) Golub-1999-v2

Clustering (3)



(c) Pomeroy-2002-v2

Clustering (4)



(d) Ramaswamy-2001

Qualitative evaluation

- Evaluated relevance of selected features in the biology literature for Golub dataset
- Proposed framework identified relevant genes that were missed by competitive methods
- Results highlight the viability of considering stability based f.s. algorithms

Further work

- Alternate optimization approaches
- Kernel k-means
- Spectral Clustering
- Parameter tuning for separation vs variance tradeoff