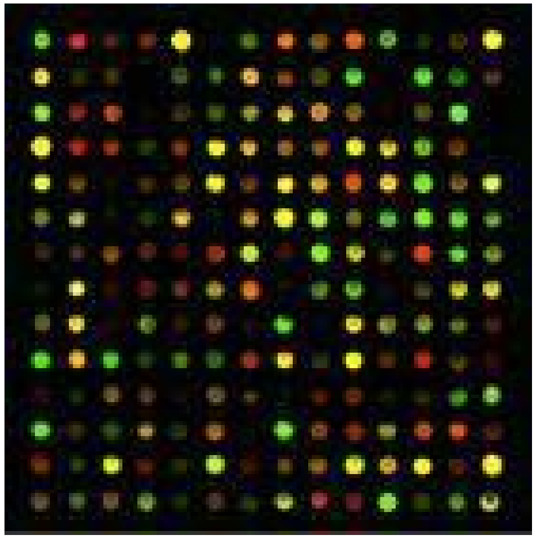
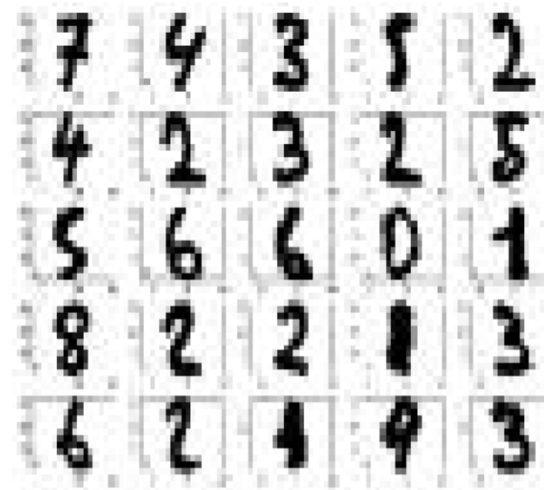
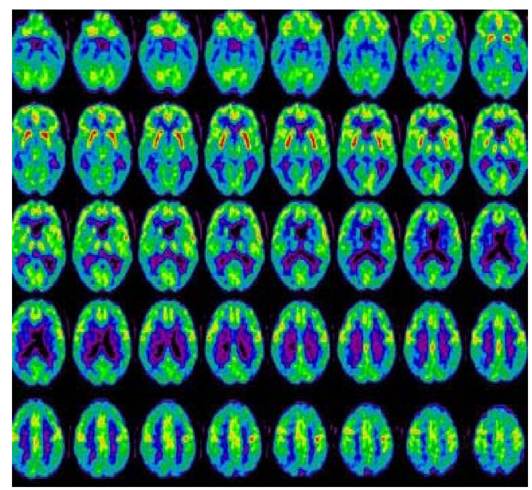
A Scalable Two-Stage Approach for a Class of Dimensionality Reduction Techniques

Liang Sun, Betul Ceran and Jieping Ye
Computer Science & Engineering
The Biodesign Institute
Arizona State University

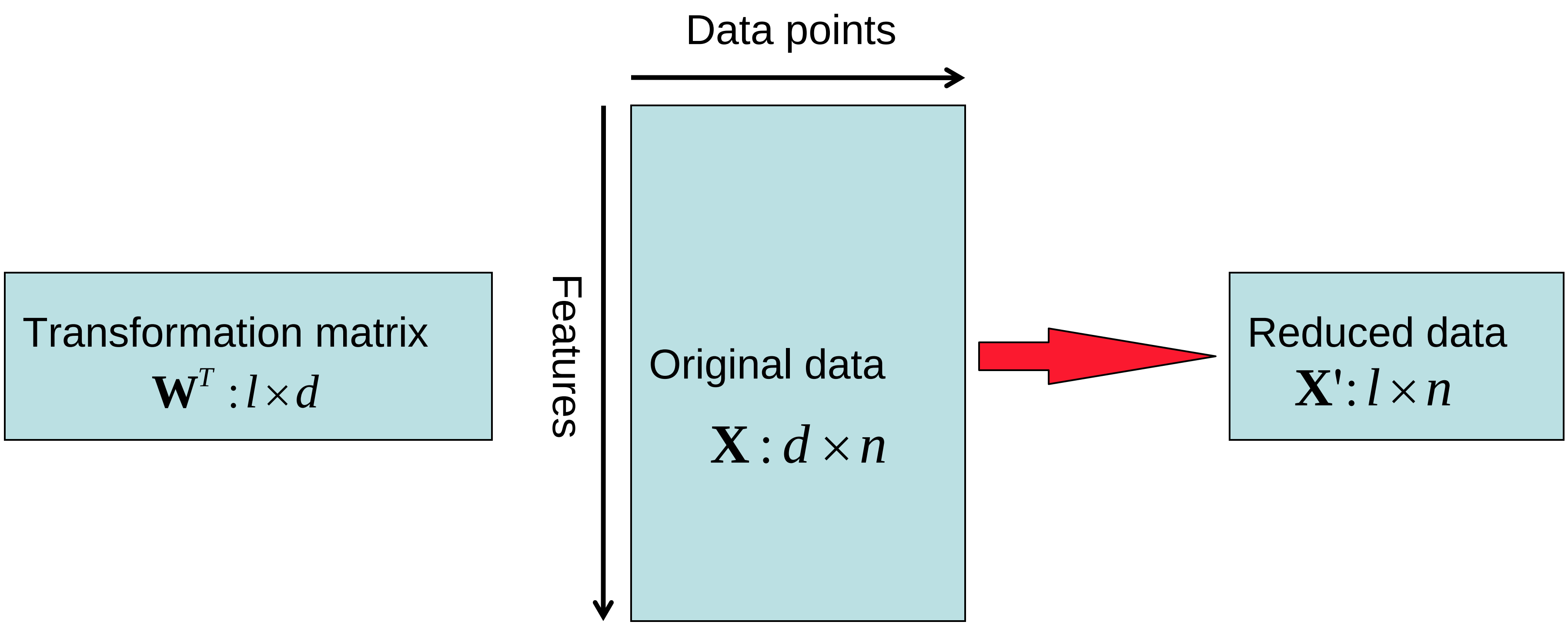
High-Dimensional Data is Ubiquitous







Linear Dimensionality Reduction



d: dimensionality
n: number of data points

W can be computed by optimizing a certain criterion

$$\mathbf{X}' = \mathbf{W}^{\mathrm{T}}\mathbf{X}$$

Why Dimensionality Reduction?

- Most data mining algorithms may not be effective for high-dimensional data.
  - Curse of Dimensionality.
- The intrinsic dimension may be small.
  - For example, the number of genes responsible for a certain type of disease may be small.
- Visualization of the data

Dimensionality Reduction Algorithms

- Unsupervised
  - Latent Semantic Indexing (LSI)
  - Principal Component Analysis (PCA)
  - Manifold learning algorithms
- Supervised
  - Canonical Correlation Analysis (CCA)
  - Partial Least Squares (PLS)
  - Linear Discriminant Analysis (LDA)
  - Hypergraph Spectral Learning (HSL)
- Semi-supervised

Dimensionality Reduction Algorithms

- Many DR algorithms reduce to solving a generalized eigenvalue problem (GEP).
- We focus on algorithms in the form of the following GEP:

$$\mathbf{X}\mathbf{S}\mathbf{X}^{\mathrm{T}}\mathbf{w} = \lambda \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{w}$$

Example dimensionality reduction algorithms:
- ❑Canonical Correlation Analysis (CCA)
- ❑Orthonormalized Partial Least Squares (OPLS)
- ❑Hypergraph Spectral Learning (HSL)
- ❑Linear Discriminant Analysis (LDA)

- GEP
- In su m the label information.

Key Challenge: How to Solve the GEP Efficiently?

- Existing algorithms do not scale to large-size problems.
  - Algorithms solving the GEP in numerical linear algebra is generally computationally expensive.
- An equivalent least squares formulation for this class of GEP was proposed [Sun *et al*. ICML 09]
  - The equivalence is established under a strong assumption.
  - The equivalence only holds for the unregularized case.

Main Contributions

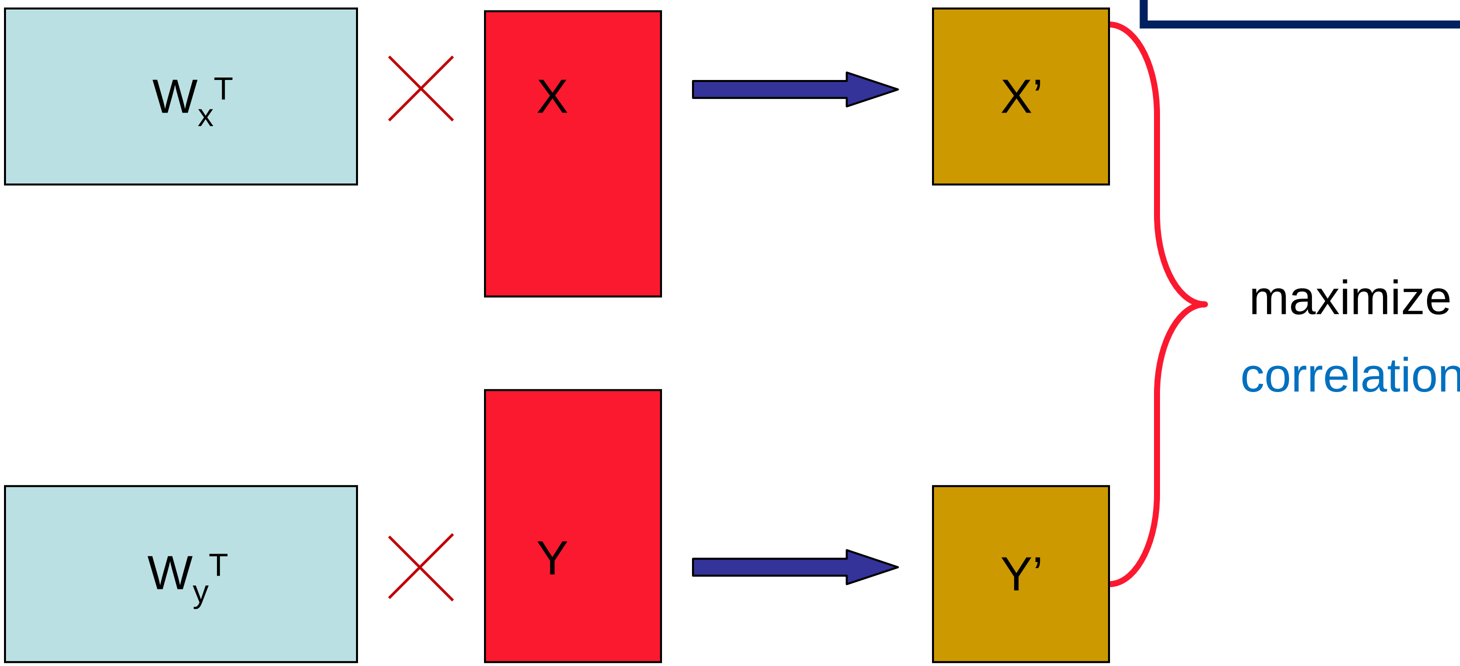
- We proposed a two-stage approach for a class of dimensionality reduction techniques including CCA, OPLS, HSL and LDA.
  - **No assumption** is required for establishing the equivalence relationship.
  - The equivalence relationship can be extended to the **regularization setting**.
  - The two-stage approach scales to large-size problems.

Outline

- Overview of Dimensionality Reduction Algorithms
  - Canonical Correlation Analysis (CCA)
  - Orthonormalized Partial Least Squares (OPLS)
  - Hypergraph Spectral Learning (HSL)
  - Linear Discriminant Analysis (LDA)
- The Proposed Two-Stage Approach
  - The main procedure
  - Equivalence relationship
  - Time complexity analysis
- Empirical Evaluation
- Conclusions

Canonical Correlation Analysis

$$\boxed{\mathbf{XSX^T w} = \lambda \mathbf{XX^T w}}$$

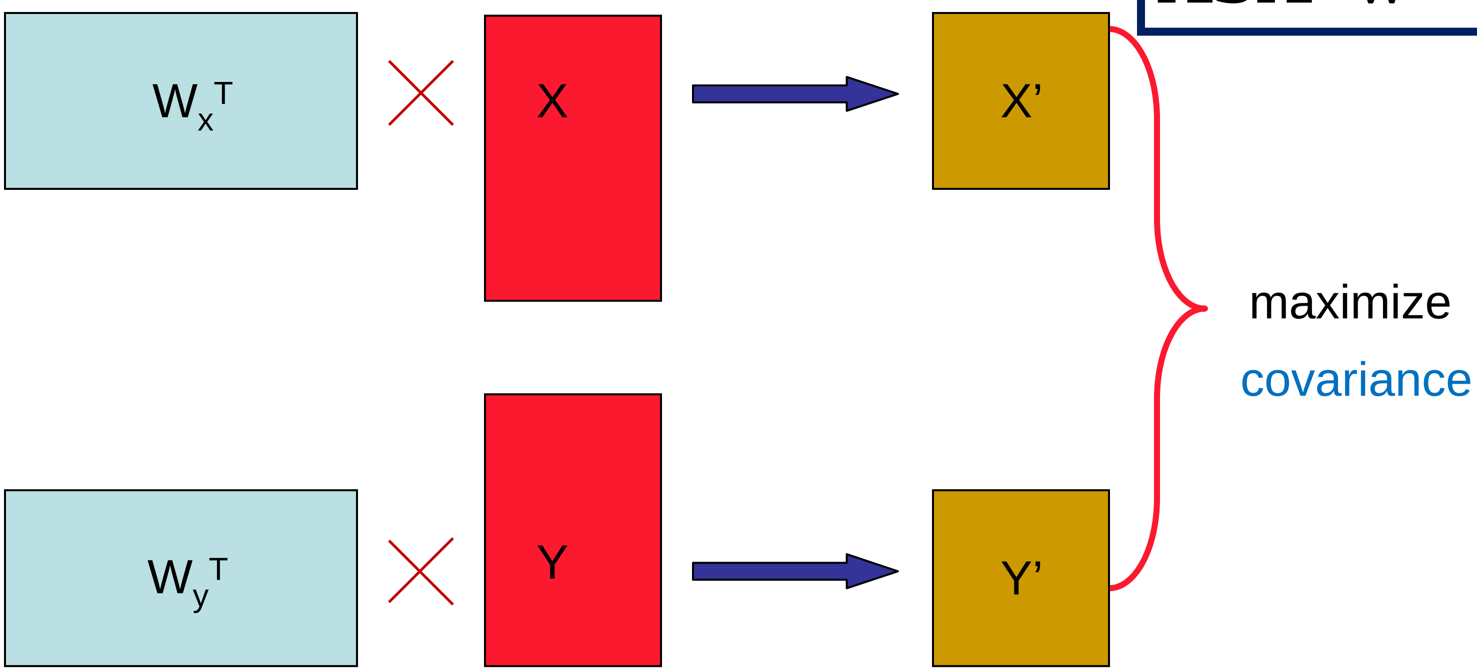


$$\mathbf{XY^T(YY^T)^{-1}YX^T w_x} = \lambda \mathbf{XX^T w_x}$$

$$\mathbf{S} = \mathbf{Y^T(YY^T)^{-1}Y} = \mathbf{HH^T}, \mathbf{H} = \mathbf{Y^T(YY^T)^{-1/2}}$$

Orthonormalized PLS

$$\boxed{\mathbf{XSX^T w} = \lambda \mathbf{XX^T w}}$$



$$\mathbf{XY^T YX^T w_x} = \lambda \mathbf{XX^T w_x},$$

$$\mathbf{S} = \mathbf{Y^T Y} = \mathbf{HH^T}, \mathbf{H} = \mathbf{Y^T}$$

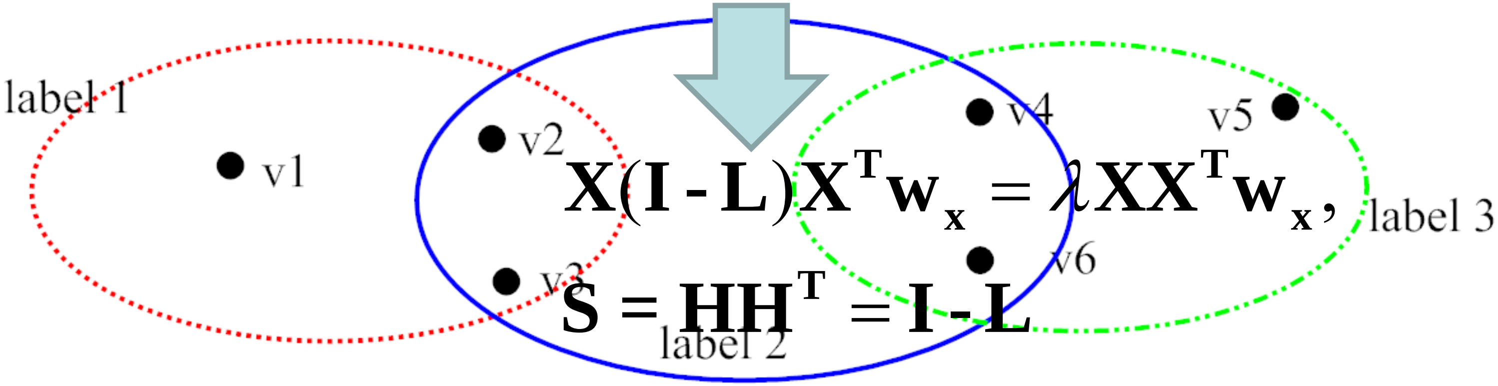
Hypergraph Spectral Learning

- HSL is a dimensionality reduction technique for multi-label classification.
- By capturing the correlation among different labels using hypergraph, HSL learns a low-dimensional embedding through a linear transformation **W**:

$$\min_{W} \quad \mathrm{Tr}(\mathbf{W}^{T}\mathbf{X}\mathbf{L}\mathbf{X}^{T}\mathbf{W})$$

$$\text{s.t.} \quad \mathbf{W}^{T}\mathbf{X}\mathbf{X}^{T}\mathbf{W} = \mathbf{I}_{\mathbf{k}}$$

Hypergraph Laplacian

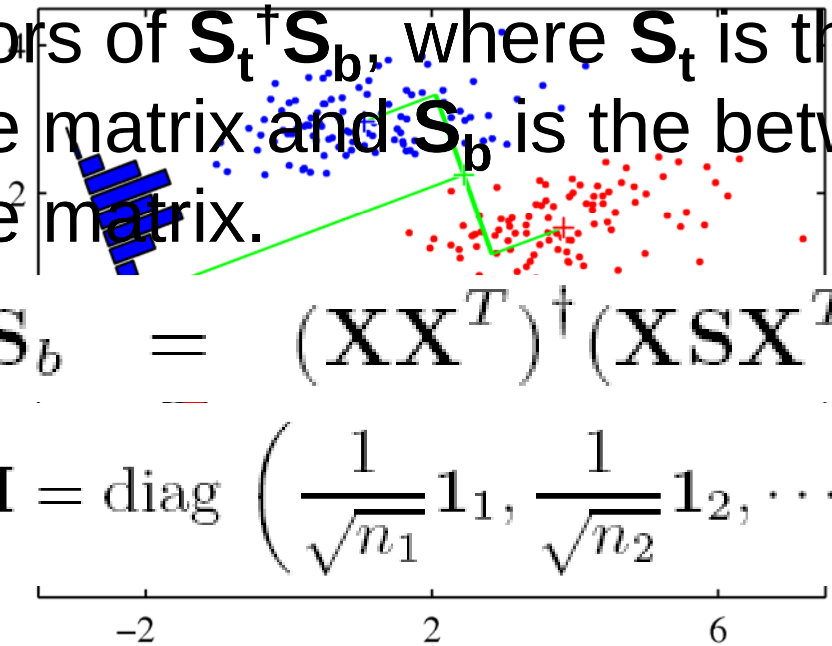


$$\mathbf{X}(\mathbf{I}-\mathbf{L})\mathbf{X}^{T}\mathbf{w}_{\mathbf{x}} = \lambda\mathbf{X}\mathbf{X}^{T}\mathbf{w}_{\mathbf{x}},$$

$$\mathbf{S} = \mathbf{H}\mathbf{H}^{T} = \mathbf{I}-\mathbf{L}$$

Linear Discriminant Analysis

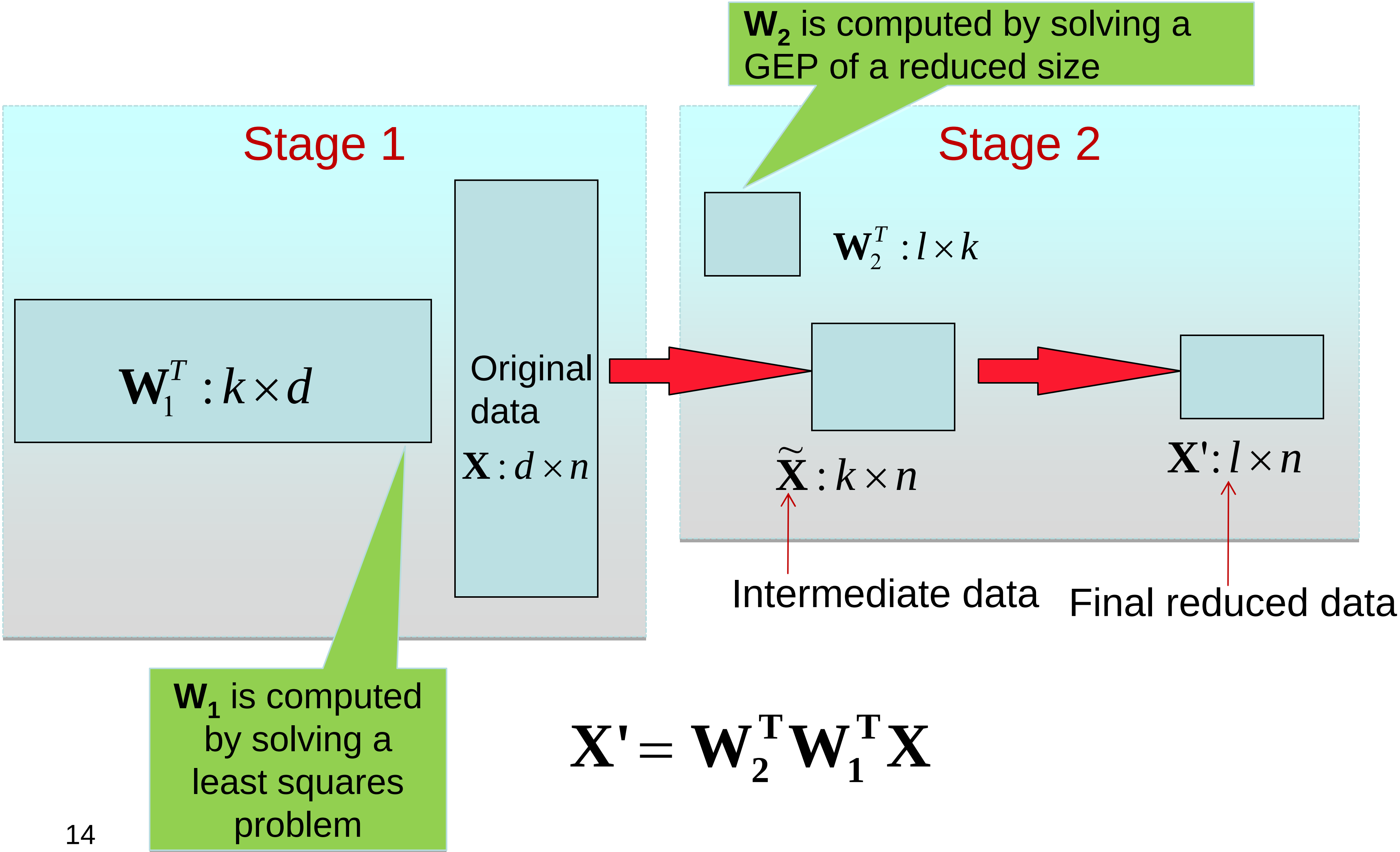
- LDA attempts to minimize the within-class variance while maximizing the between-class variance after the linear projection.
- The optimal linear projection consists of the top eigenvectors of $\mathbf{S_t}^\dagger\mathbf{S_b}$, where $\mathbf{S_t}$ is the total covariance matrix and $\mathbf{S_b}$ is the between-class covariance matrix.



$$\mathbf{S}_t^\dagger \mathbf{S}_b = (\mathbf{X}\mathbf{X}^T)^\dagger (\mathbf{X}\mathbf{S}\mathbf{X}^T)$$

$$\mathbf{S} = \mathbf{H}\mathbf{H}^T, \quad \mathbf{H} = \mathrm{diag}\left(\frac{1}{\sqrt{n_1}}\mathbf{1}_1, \frac{1}{\sqrt{n_2}}\mathbf{1}_2, \cdots, \frac{1}{\sqrt{n_k}}\mathbf{1}_k\right) \in \mathbb{R}^{n\times k}$$

Overview of the Two-Stage Approach



$$\mathbf{X}' = \mathbf{W}_2^{\mathbf{T}} \mathbf{W}_1^{\mathbf{T}} \mathbf{X}$$

The Two-Stage Approach without Regularization

In the second stage, we replace X in the original GEP with the intermediate data, and solve a GEP (or optimization problem) of a reduced size.

Algorithm 1 The Two-Stage Appr
ization

Input: $\mathbf{X}$, $\mathbf{H}$

Output: $\mathbf{W}$

Stage 1: Solve the following least sq problem:

$$\min_{\mathbf{W}_1} \|\mathbf{W}_1^T \mathbf{X} - \mathbf{H}^T\|_F^2. \tag{10}$$

Stage 2: Compute $\tilde{\mathbf{X}} = \mathbf{W}_1^T \mathbf{X}$, and so e the following optimization problem:

$$\begin{aligned} \max_{\mathbf{W}_2} \quad & \mathrm{Tr}(\mathbf{W}_2^T \tilde{\mathbf{X}} \mathbf{H} \mathbf{H}^T \tilde{\mathbf{X}}^T \mathbf{W}_2) \quad (11) \\ \text{s. t.} \quad & \mathbf{W}_2^T \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \mathbf{W}_2 = \mathbf{I}_\ell. \end{aligned}$$

Compute $\mathbf{W} = \mathbf{W}_1 \mathbf{W}_2$ as the final solution.

The Two-Stage Approach without Regularization

- In the first stage, $\mathbf{H}^T$ can be considered as the "latent target" encoded by the label information in **Y**.
- Advantages of using LSQR to solve least squares in the first stage:
  - Good scalability.
  - Reliable for even ill-conditioned problems.
- In the second stage, we project the data matrix **X** onto a subspace, and solve the resulting generalized eigenvalue problem of a reduced size.

Time Complexity Analysis

Algorithm 1 The Tw… ization

Input: $\mathbf{X}$, $\mathbf{H}$

Output: $\mathbf{W}$

Stage 1: Solve the fol… …ing least squares problem:

$$\min_{\mathbf{W}_1} \|\mathbf{W}_1^T\mathbf{X} - \mathbf{H}^T \ldots$$

Stage 2: Compute $\tilde{\mathbf{X}} = \mathbf{W}_1^T\mathbf{X}$, and … the following optimi… problem:

(11)

s. t. $\mathbf{W}_2^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{W}_2 = \ldots$

Compute $\mathbf{W} = \mathbf{W}_1\mathbf{W}_2$ as the final solution.

The total computational cost of the first stage is O($Nk(3n+5d+2z)$) using LSQR when **X** is sparse, where N is the total number of iterations, z is the number of nonzero entries in **X**.

The cost of the second stage is O($kz+nk^2$).

The total computational cost is O($Nk(3n + 5d + 2z) + kz + nk^2 + dk^2$) when X is sparse.

The cost of combining the results in two stages is O(dk^2)

Equivalence without Regularization

THEOREM 1. *The top ℓ ($\ell \leq \mathrm{rank}(\mathbf{A})$) projection vectors computed by Eq. (11) are given by*

$$\mathbf{W}_2 = (\mathbf{U}_A \mathbf{\Sigma}_A^{-1})_\ell, \tag{22}$$

where $(\mathbf{U}_A \mathbf{\Sigma}_A^{-1})_\ell$ consists of the first ℓ columns of $(\mathbf{U}_A \mathbf{\Sigma}_A^{-1})$. Thus, the projection vectors computed by the two-stage approach are

$$\boxed{\mathbf{W} = \mathbf{W}_1 \mathbf{W}_2 = \mathbf{U}_1 \mathbf{\Sigma}_1^{-1} \mathbf{V}_{A\ell}} \tag{23}$$

THEOREM 2. *The eigenvectors corresponding to the top ℓ ($\ell \leq \mathrm{rank}(\mathbf{A})$) eigenvalues of $(\mathbf{X}\mathbf{X}^T)^\dagger(\mathbf{X}\mathbf{H}\mathbf{H}^T\mathbf{X}^T)$ are*

$$\boxed{\mathbf{W}_0 = \mathbf{U}_1 \mathbf{\Sigma}_1^{-1} \mathbf{V}_{A\ell},} \tag{26}$$

where $\mathbf{V}_{A\ell}$ consists of the first ℓ columns of $\mathbf{V}_A$. Thus, the two-stage approach produces the same solution as the direct approach which solves the original generalized eigenvalue problem directly.

The Two-Stage Approach with Regularization

- The two-stage approach can be extended to the regularization setting
  - A **penalized** least squares problem using the same target is solved in the first stage.
  - The equivalence relationship can also be rigorously established

A significant improvement of existing work

- The computational cost of the two-stage approach in the regularization setting is the same as the unregularized one.

Empirical Evaluation

- Goals:
  - To verify the equivalence relationship between the direct approach and the two-stage approach.
  - To demonstrate the scalability of the two-stage approach.
- Setup
  - All experiments are performed on a PC with Intel Core 2 Duo T9500 2.6G CPU and 4GB RAM.
  - Synthetic data are generated using the Gaussian distribution.
  - To verify the equivalence relationship, we compare $||W_0W_0^T - WW^T||_2$ under different values of the regularization parameter γ.

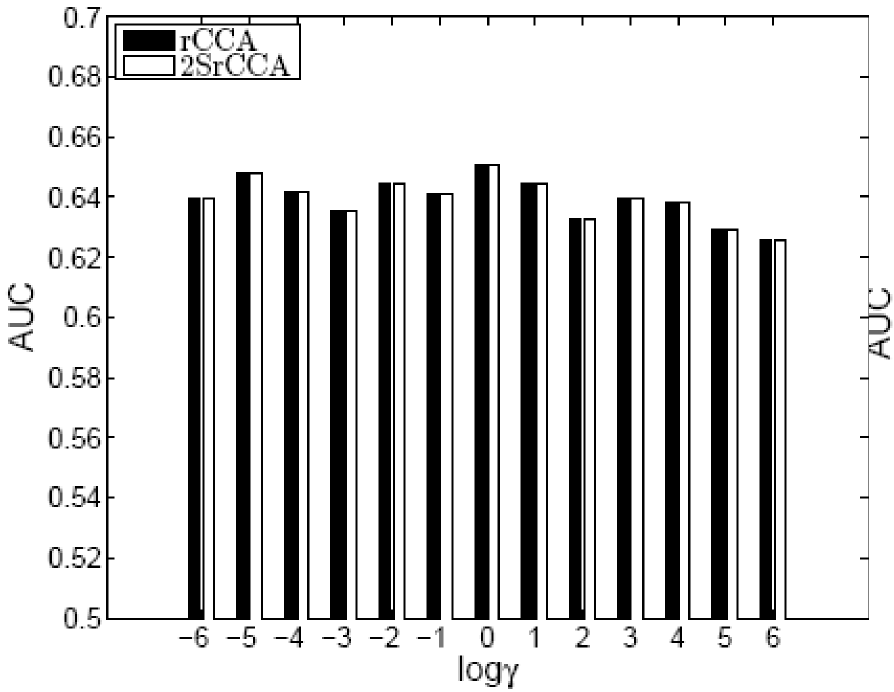
Data Description

Table 1: Statistics of the data sets: n is the number of samples, d is the data dimensionality, and k is the number of labels (classes).

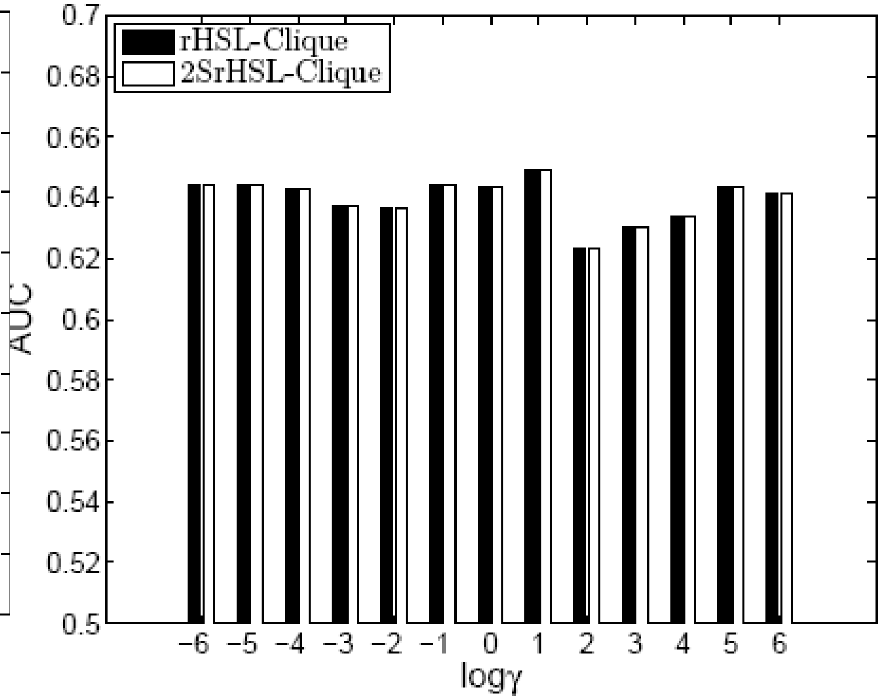
Data Set	Type	n	d	k
Syn1	Multi-class	1000	100	5
Syn2	Multi-class	1000	5000	5
Syn3	Multi-label	1000	100	5
Syn4	Multi-label	1000	5000	5
Ionosphere	Multi-class	351	34	2
Optical digits	Multi-class	5620	64	10
Satimage	Multi-class	6435	36	6
USPS	Multi-class	9298	256	10
Wine	Multi-class	178	13	3
Scene	Multi-label	2407	294	6
Yeast	Multi-label	2417	103	14
news20	Multi-class	15935	62061	20
rcv1v2	Multi-label	3000	47236	101

AUC Comparison on the Yeast Data Set

- Sample size n = 2417, dimensionality d=103, number of labels k=14.
- Regularization parameter γ= 1e-6 ~ 1e6.



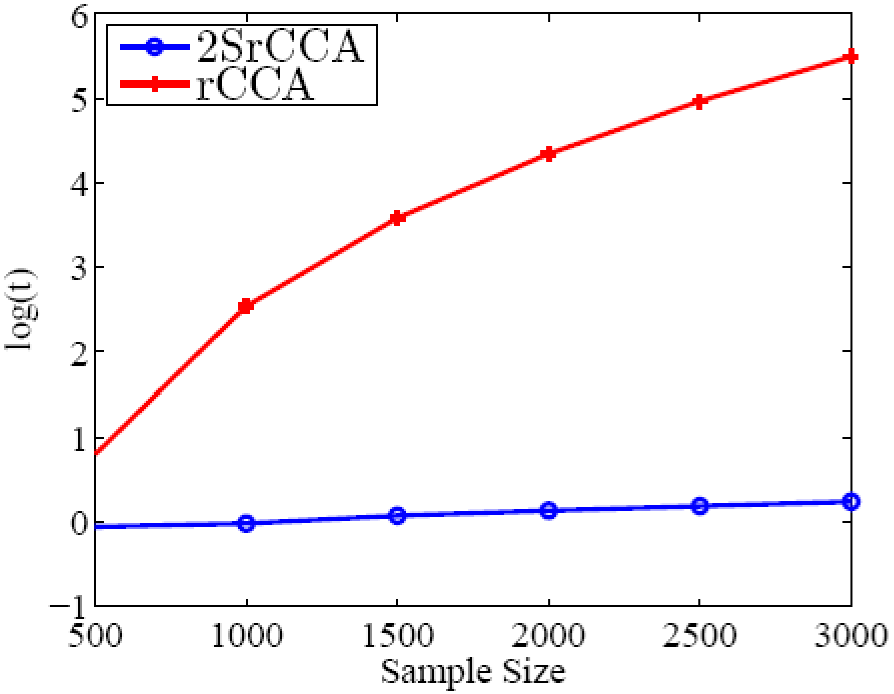
(A) CCA



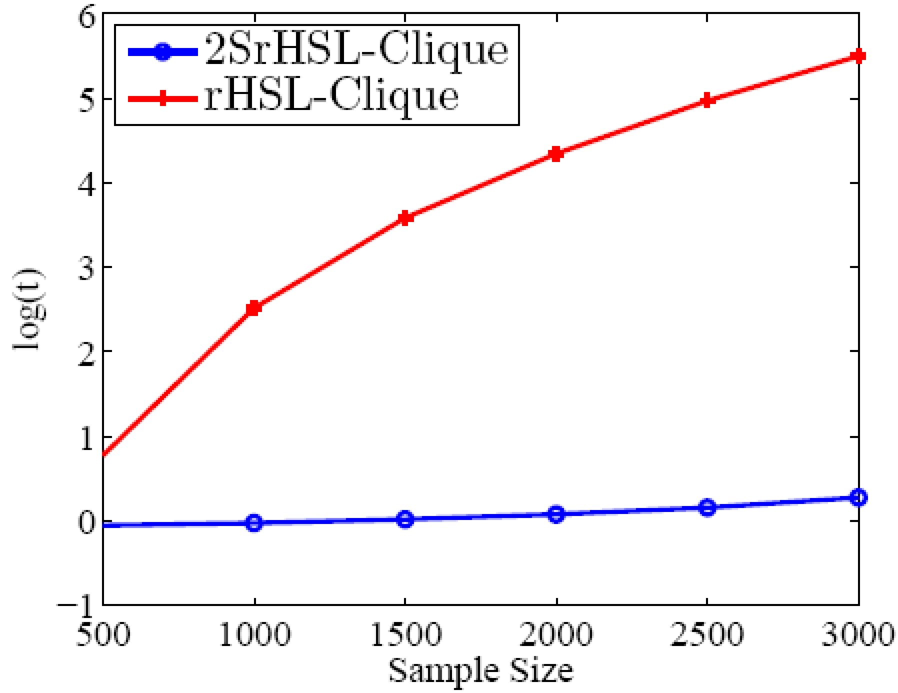
(C) HSL-Clique

Scalability Comparison on the rcv1v2 Data Set (1)

- Sample size n=500:500:3000.
- Dimensionality d=5000.
- Number of labels k=101.



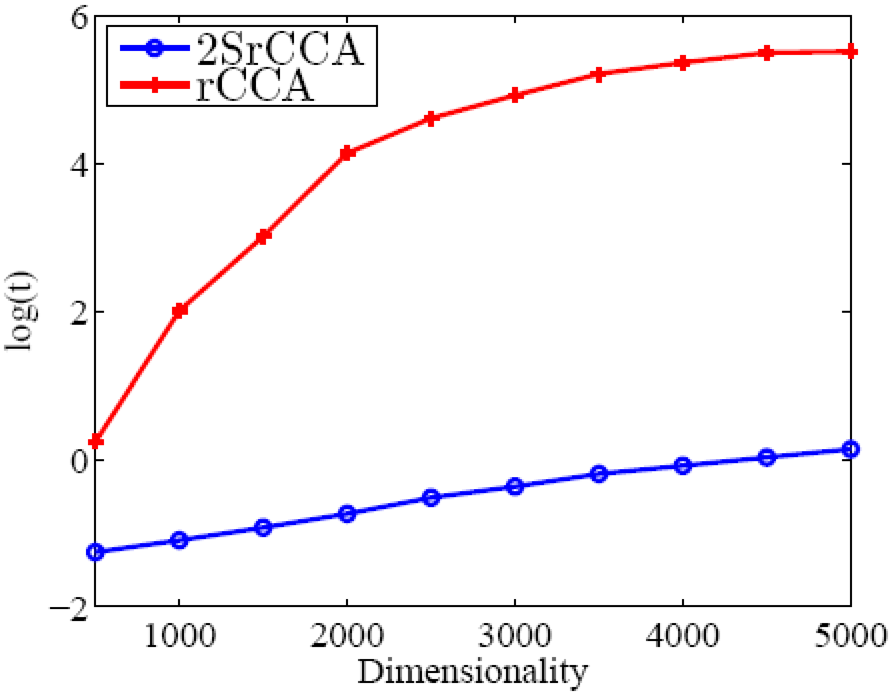
(A) CCA



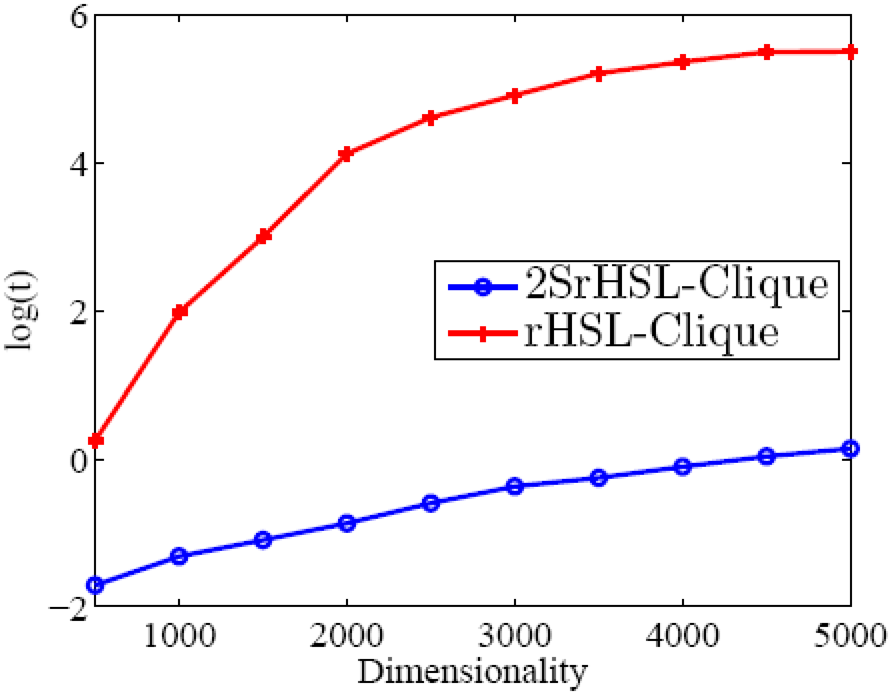
(C) HSL-Clique

Scalability Comparison on the rcv1v2 Data Set (2)

- Sample size n=3000.
- Dimensionality d=500:500:5000.
- Number of labels k=101.



(A) CCA



(C) HSL-Clique

Conclusions and Future Work

- Establish the two-stage approach for a class of dimensionality reduction techniques including CCA, OPLS, LDA, and HSL.
  - The equivalence relationship is rigorously proved.
  - Advantages of the two-stage approach:
    - **No assumption** is required.
    - It can be applied in the **regularization setting**.
    - Good **scalability**.

- Future Work
  - Extend the two-stage approach to other algorithms similar to the GEP formulation.
  - Online algorithm for the two-stage approach.

Thank you!

Equivalence Verification

Data	Technique	0	1.0e-006	1.0e-004	1.0e-002	1.0e+000	1.0e+002	1.0e+004	1.0e+006
Syn1	LDA	2.9e-018	3.6e-018	3.4e-018	3.1e-018	2.6e-018	2.5e-018	3.1e-019	3.0e-021
Syn2	LDA	5.8e-019	1.4e-018	1.2e-018	8.9e-019	1.2e-018	9.9e-019	2.3e-019	2.9e-021
Syn3	CCA	4.9e-018	8.4e-018	7.0e-018	6.5e-018	9.5e-018	6.0e-018	5.1e-019	7.2e-021
	OPLS	4.6e-018	5.0e-018	8.7e-018	5.0e-018	6.6e-018	6.1e-018	5.4e-019	5.0e-021
	HSL-Clique	1.0e-017	1.8e-017	1.2e-017	1.2e-017	1.5e-017	1.4e-017	2.9e-018	2.5e-020
	HSL-Star	1.4e-017	2.4e-017	9.3e-018	2.6e-017	2.1e-017	5.0e-017	9.8e-019	1.3e-020
Syn4	CCA	1.3e-018	5.2e-018	3.2e-018	1.8e-018	1.3e-018	1.8e-018	4.2e-019	5.9e-021
	OPLS	1.0e-018	1.1e-018	1.3e-018	1.5e-018	1.3e-018	1.3e-018	2.9e-019	5.9e-021
	HSL-Clique	2.7e-018	2.9e-018	2.7e-018	5.0e-018	3.2e-018	2.7e-018	8.9e-019	1.4e-020
	HSL-Star	2.5e-018	3.7e-018	2.9e-018	5.7e-018	4.1e-018	2.9e-018	1.1e-018	3.1e-020
Scene	CCA	2.4e-015	2.1e-015	6.1e-015	3.7e-015	1.2e-015	1.8e-016	6.0e-018	9.0e-020
	OPLS	2.0e-015	3.4e-015	3.8e-015	2.5e-015	1.1e-015	2.3e-016	1.1e-017	1.4e-019
	HSL-Clique	4.5e-015	9.1e-015	2.6e-014	1.2e-014	3.6e-015	1.3e-015	5.9e-017	1.0e-018
	HSL-Star	4.6e-015	3.3e-014	2.1e-014	7.7e-015	1.1e-014	2.5e-016	1.0e-016	6.5e-019
Yeast	CCA	1.6e-012	1.5e-011	1.2e-012	1.4e-015	6.9e-016	5.9e-017	1.7e-018	1.4e-020
	OPLS	4.1e-012	1.6e-011	3.7e-012	1.2e-014	1.5e-015	3.7e-016	3.2e-018	2.9e-020
	HSL-Clique	1.5e-012	1.4e-011	3.7e-012	3.9e-015	1.6e-015	2.7e-016	5.1e-018	2.5e-020
	HSL-Star	2.1e-012	1.0e-011	2.4e-012	1.1e-014	9.4e-015	1.1e-015	1.5e-017	4.4e-019
Wine	LDA	5.9e-017	2.1e-016	2.3e-016	2.1e-016	3.2e-017	2.2e-018	1.3e-020	2.0e-020
Satimage	LDA	4.6e-016	2.2e-015	8.4e-016	7.3e-016	7.7e-016	8.1e-017	3.9e-017	6.2e-019
Ionosphere	LDA	8.5e-018	1.0e-017	4.3e-018	2.1e-017	6.8e-018	6.6e-018	6.6e-020	1.1e-021
Optical digits	LDA	6.2e-018	7.2e-018	6.7e-018	5.7e-018	1.9e-018	1.5e-019	5.9e-020	5.6e-021
USPS	LDA	7.0e-015	3.0e-014	2.6e-014	6.6e-015	1.1e-016	3.0e-018	4.1e-019	6.6e-021