

Rethinking Sparse Optimization Through Deep Learning

Xianchao Xiu

Department of Automation



Mathematical Optimization Society, May 16-19, 2025

Joint work with [Chenyi Huang](#) (SHU), [Long Chen](#) (SHU), and others

Outline

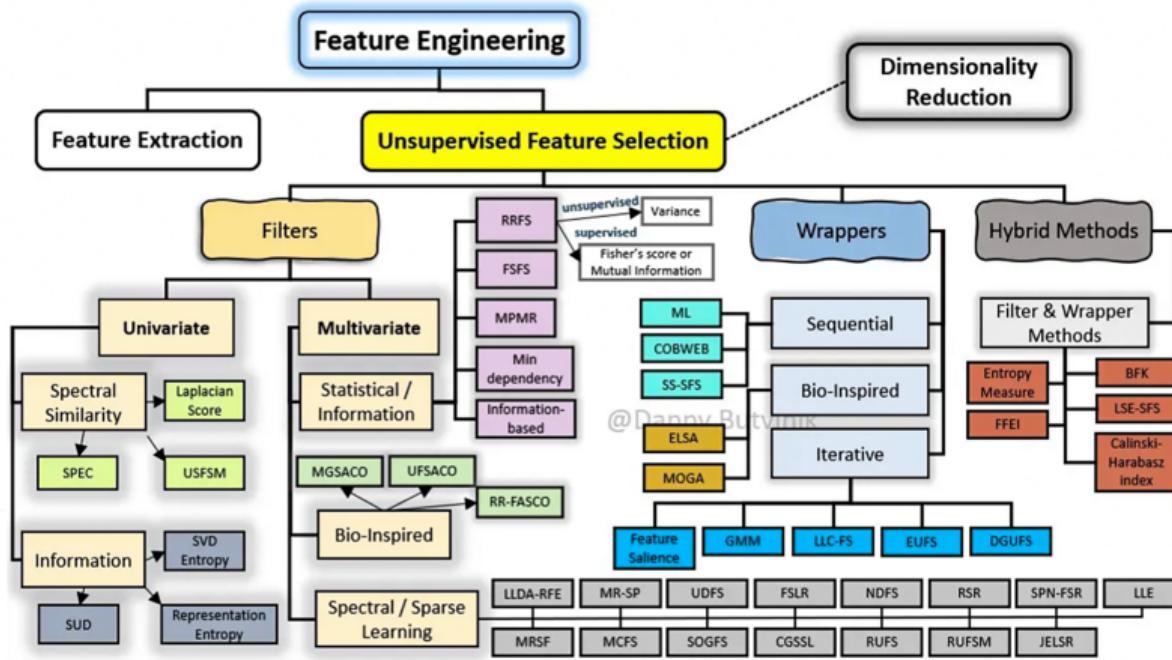
Introduction

Bi-Sparse Feature Selection

Deep Unfolding Feature Selection

Future Work

- ▶ Unsupervised feature selection *vs.* Feature extraction
- ▶ Select a subset of input features without labels



PCA

- Given $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$, principal component analysis (PCA) is

$$\min_{W \in \mathbb{R}^{d \times p}} \frac{1}{2} \|X - WW^\top X\|_F^2$$

$$\text{s.t. } W^\top W = I_p$$

\Updownarrow

$$\min_{W \in \mathbb{R}^{d \times p}} -\text{Tr}(W^\top X X^\top W)$$

$$\text{s.t. } W^\top W = I_p$$

- Unsupervised feature selection by sparse PCA

$$\min_{W \in \mathbb{R}^{d \times p}} -\text{Tr}(W^\top X X^\top W)$$

$$\text{s.t. } W^\top W = I_p, \|W\|_{2,0} \leq s$$

- The i -th feature can be measured by $\|\mathbf{w}^i\|$ since $\mathbf{z}_i = (\mathbf{w}^{1\top}, \mathbf{w}^{2\top}, \dots, \mathbf{w}^{d\top})\mathbf{x}_i$
- The dimension number is often omitted when it does not cause ambiguity

Motivation

- ▶ Li-Nie-Bian et al, Sparse PCA via $\ell_{2,p}$ -Norm Regularization for Unsupervised Feature Selection, [IEEE TPAMI](#), 2023

$$\begin{aligned} \min_W \quad & -\text{Tr}(W^\top X X^\top W) + \lambda \|W\|_{2,p}^p \quad (0 < p < 1) \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$

- ▶ Li-Sun-Zhang, Unsupervised Feature Selection via Nonnegative Orthogonal Constrained Regularized Minimization, [arXiv](#), 2024

$$\begin{aligned} \min_{W,Y} \quad & \text{Tr}(Y^\top LY) + \alpha \|Y - X^\top W\|_{2,1} + \beta \|W\|_{2,1} + \gamma \|W\|_F^2 \\ \text{s.t.} \quad & Y^\top Y = I, \quad Y \geq 0 \end{aligned}$$

- ▶ Hu-Wang-Zhang et al, Bi-Level Spectral Feature Selection, [IEEE TNNLS](#), 2025
- ▶ Jiao-Xue-Zhang, Sparse Learning-Based Feature Selection in Classification: A Multi-Objective Perspective, [IEEE TETCI](#), 2025
- ▶ Li-Yu-Yang et al, Exploring Feature Selection With Limited Labels: A Comprehensive Survey of Semi-Supervised and Unsupervised Approaches, [IEEE TKDE](#), 2024

Outline

Introduction

Bi-Sparse Feature Selection

Deep Unfolding Feature Selection

Future Work

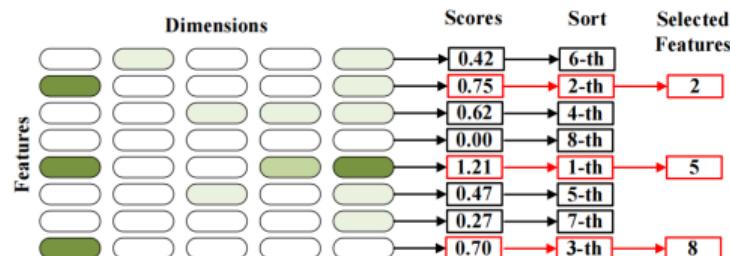
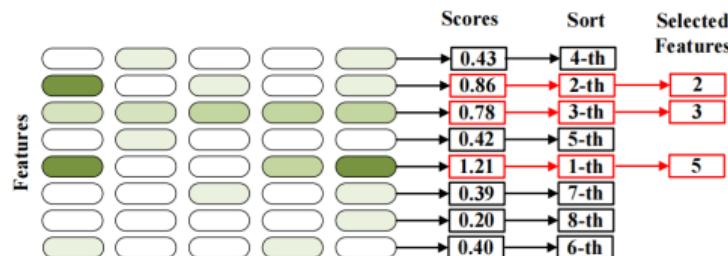
Model

- Xiu-Huang-Shang et al, Bi-Sparse Unsupervised Feature Selection, IEEE TIP, 2025

$$\begin{aligned} \min_W \quad & -\text{Tr}(W^\top SW) + \lambda \|W\|_{2,p}^p \quad (0 < p < 1) \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$

↓

$$\begin{aligned} \min_W \quad & -\text{Tr}(W^\top SW) + \lambda_1 \|W\|_{2,p}^p + \lambda_2 \|W\|_q^q \quad (0 \leq p, q < 1) \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$



Optimization

- ▶ Proximal alternating minimization (PAM)
- ▶ Model reformulation

$$\min_W -\text{Tr}(W^\top SW) + \lambda_1 \|W\|_{2,p}^p + \lambda_2 \|W\|_q^q$$

$$\text{s.t. } W^\top W = I$$

\Downarrow

$$\min_{W,U,V} -\text{Tr}(W^\top SW) + \lambda_1 \|V\|_{2,p}^p + \lambda_2 \|U\|_q^q$$

$$\text{s.t. } W^\top W = I, V = W, U = W$$

\Downarrow

$$\begin{aligned} \min_{W,U,V} & -\text{Tr}(W^\top SW) + \lambda_1 \|V\|_{2,p}^p + \lambda_2 \|U\|_q^q \\ & + \frac{\beta_1}{2} \|W - U\|_F^2 + \frac{\beta_2}{2} \|W - V\|_F^2 + \Phi(W) \end{aligned}$$

PAM

- ▶ Input: $X, \lambda_1, \lambda_2, \beta_1, \beta_2, p, q, \tau_1, \tau_2, \tau_3$
- ▶ Initialize: W^0, U^0, V^0
- ▶ While not converged do
 - ▶ Update W^{k+1} by

$$\begin{aligned} \min_W \quad & -\text{Tr}(W^\top SW) + \frac{\beta_1}{2} \|W - U^k\|_F^2 + \frac{\beta_2}{2} \|W - V^k\|_F^2 + \frac{\tau_1}{2} \|W - W^k\|_F^2 \\ \text{s.t. } \quad & W^\top W = I \end{aligned}$$

- ▶ Update U^{k+1} by

$$\min_U \lambda_2 \|U\|_q^q + \frac{\beta_1}{2} \|W^{k+1} - U\|_F^2 + \frac{\tau_2}{2} \|U - U^k\|_F^2$$

- ▶ Update V^{k+1} by

$$\min_V \lambda_1 \|V\|_{2,p}^p + \frac{\beta_2}{2} \|W^{k+1} - V\|_F^2 + \frac{\tau_3}{2} \|V - V^k\|_F^2$$

- ▶ Output: $W^{k+1}, U^{k+1}, V^{k+1}$

- ▶ Update W -block: Riemannian gradient

$$\begin{aligned} \min_W \quad & -\text{Tr}(W^\top SW) + \frac{\beta_1}{2} \|W - U^k\|_F^2 + \frac{\beta_2}{2} \|W - V^k\|_F^2 + \frac{\tau_1}{2} \|W - W^k\|_F^2 \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$

 \Downarrow

$$\nabla g(W) = -2SW + \beta_1(W - U^k) + \beta_2(W - V^k) + \tau_1(W - W^k)$$

 \Downarrow

$$\begin{aligned} \text{grad } g(W) &= \mathcal{P}_W(\nabla g(W)) \\ &= \nabla g(W) - W \text{sym}(W^\top \nabla g(W)) \end{aligned}$$

- ▶ $\text{sym}(X) = (X + X^\top)/2$ extracts the symmetric part of X

- ▶ Update W -block: Riemannian Hessian

$$\nabla^2 g(W) = -2I \otimes S + (\beta_1 + \beta_2 + \tau_1)I$$

\Downarrow

$$\begin{aligned}\text{Hess } g(W) &= \mathcal{P}_W(\nabla^2 g(W)) \\ &= \nabla^2 g(W) - W \text{sym}(W^\top \nabla^2 g(W))\end{aligned}$$

\Downarrow

$$\text{Hess } g(W) \approx \frac{\text{grad } g(W + \varepsilon I) - \text{grad } g(W)}{\varepsilon}$$

PAM

- ▶ Input: $S, U^k, V^k, \beta_1, \beta_2, \tau_1, \varepsilon, \Delta' > 0, \rho' \in [0, \frac{1}{4})$
- ▶ While not converged do
 - ▶ Obtain η_i by solving trust domain subproblem

$$\begin{aligned} \min_{\eta \in T_W \text{St}(d,m)} m_W(\eta) &= g(W) + \text{Tr}(\eta^\top \text{grad } g(W)) + \frac{1}{2} \text{vec}(\eta)^\top \text{Hess } g(W) \text{vec}(\eta) \\ \text{s.t.} \quad \text{Tr}(\eta^\top W \eta^\top) &\leq \Delta^2 \end{aligned}$$

- ▶ Compute the trust ratio ρ_i
- ▶ if $\rho_i < \frac{1}{4}$ then
$$\Delta_{i+1} = \frac{1}{4}\Delta_i$$
- ▶ else if $\rho_i > \frac{3}{4}$ and $\|\eta_i\| = \Delta_i$ then
$$\Delta_{i+1} = \min(2\Delta_i, \Delta')$$
- ▶ else
$$\Delta_{i+1} = \Delta_i$$
- ▶ if $\rho_i > \rho'$ then
$$W_{i+1}^k = R_W(\eta_i)$$
- ▶ else
$$W_{i+1}^k = W_i^k$$
- ▶ Output: W

► Update U -block

$$\min_U \lambda_2 \|U\|_q^q + \frac{\beta_1}{2} \|W^{k+1} - U\|_{\text{F}}^2 + \frac{\tau_2}{2} \|U - U^k\|_{\text{F}}^2$$

↓

$$\min_U \lambda_2 \|U\|_q^q + \frac{\beta_1 + \tau_2}{2} \|U - \frac{\beta_1}{\beta_1 + \tau_2} W^{k+1} + \frac{\tau_2}{\beta_1 + \tau_2} U^k\|_{\text{F}}^2$$

↓

$$\min_{u_{ij}} \lambda_2 |u_{ij}|^q + \frac{\beta_1 + \tau_2}{2} (u_{ij} - y_{ij})^2$$

↓

$$u_{ij} = \text{Prox}(y_{ij}, \lambda_2 / (\beta_1 + \tau_2))$$

Lemma

- Revisiting ℓ_q ($0 \leq q < 1$) Norm Regularized Optimization, arXiv:2306.14394

$$\begin{aligned}\text{Prox}(a, \lambda) &= \operatorname{argmin}_x \frac{1}{2}(x - a)^2 + \lambda|x|^q \quad (0 \leq q < 1) \\ &= \begin{cases} \{0\}, & |a| < \kappa(\lambda, q) \\ \{0, \operatorname{sgn}(a)c(\lambda, q)\}, & |a| = \kappa(\lambda, q) \\ \{\operatorname{sgn}(a)\varpi_q(|a|)\}, & |a| > \kappa(\lambda, q) \end{cases}\end{aligned}$$

where

$$c(\lambda, q) = (2\lambda(1-q))^{\frac{1}{2-q}} > 0$$

$$\kappa(\lambda, q) = (2-q)\lambda^{\frac{1}{2-q}}(2(1-q))^{\frac{q+1}{q-2}}$$

$$\varpi_q(a) \in \{x : x - a + \lambda q \operatorname{sgn}(x)x^{q-1} = 0, x > 0\}$$

PAM

- ▶ Update V -block

$$\min_V \lambda_1 \|V\|_{2,p}^p + \frac{\beta_2}{2} \|W^{k+1} - V\|_F^2 + \frac{\tau_3}{2} \|V - V^k\|_F^2$$

↓

$$\min_V \lambda_1 \|V\|_{2,p}^p + \frac{\beta_2 + \tau_3}{2} \|V - \frac{\beta_2}{\beta_2 + \tau_3} W^{k+1} + \frac{\tau_3}{\beta_2 + \tau_3} V^k\|_F^2$$

↓

$$\min_{\mathbf{v}^i} \lambda_1 \sum_{i=1}^d \|\mathbf{v}^i\|_2^p + \frac{\beta_2 + \tau_3}{2} \|\mathbf{v}^i - \mathbf{z}^i\|_2^2$$

↓

$$\mathbf{v}^i = \text{Prox}(\|\mathbf{z}^i\|_2, \lambda_1 / (\beta_2 + \tau_3)) \cdot \frac{\mathbf{z}^i}{\|\mathbf{z}^i\|_2}$$

Experiments

- ▶ Compared methods
 - ▶ LapScore: He-Cai-Niyogi, NIPS, 2005
 - ▶ UDFS: Yang-Shen-Ma et al, IJCAI, 2011
 - ▶ SOGFS: Nie-Zhu-Li, IEEE TKDE, 2021
 - ▶ RNE: Liu-Ye-Li-Wang et al, KBS, 2020
 - ▶ FSPCA: Tian-Nie-Wang-Li et al, NIPS, 2020
 - ▶ SPCAFS: Li-Nie-Bian et al, IEEE TPAMI, 2023
 - ▶ SPCA-PSD: Zheng-Zhang-Liu et al, arXiv, 2023
 - ▶ FEN-PCAFS: Gao-Wu-Xu et al, IEEE TFS, 2024
- ▶ Implementation setups
 - ▶ Initialization: QR decomposition
 - ▶ Stopping criteria:

$$\frac{|f(W^{k+1}, U^{k+1}, V^{k+1}) - f(W^k, U^k, V^k)|}{\max\{1, |f(W^k, U^k, V^k)|\}} \leq 10^{-4}$$

Experiments

- ▶ Selected datasets

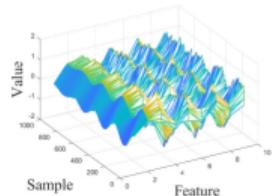
Type	Datasets	Features	Samples	Classes
Synthetic datasets	Dartboard1	9	1000	4
	Diamond9	9	3000	9
Real-world datasets	COIL20	1024	1440	20
	USPS	256	1000	10
	LUNG	325	73	7
	GLIOMA	4434	50	4
	UMIST	644	575	20
	pie	1024	1166	53
	Isolet	617	1560	26
	MSTAR	1024	2425	10

- ▶ Evaluation metrics

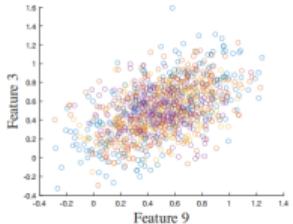
- ▶ ACC: Accuracy
- ▶ NMI: Normalized mutual information

Synthetic Experiments

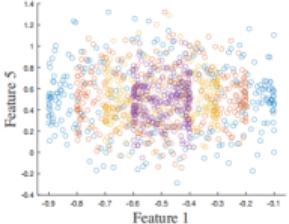
► Dartboard1



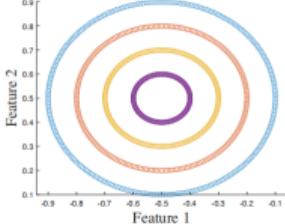
(a) Dartboard1



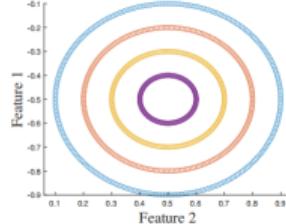
(b) LapScore



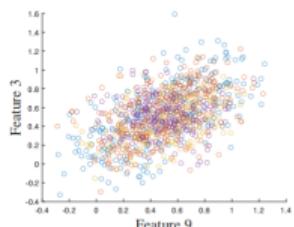
(c) SOGFS



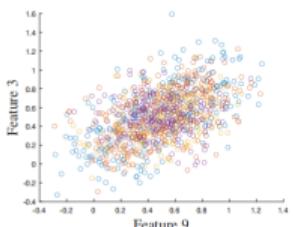
(d) RNE



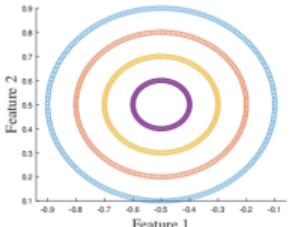
(e) UDFS



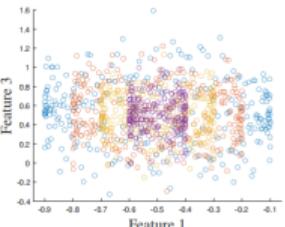
(f) SPCAFS



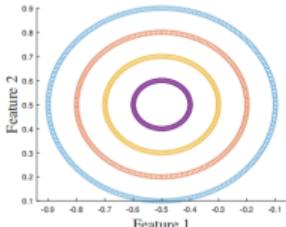
(g) FSPCA



(h) SPCA-PSD



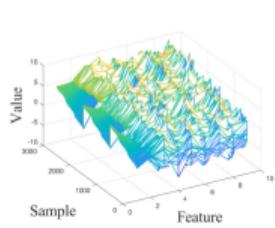
(i) FEN-PCAFS



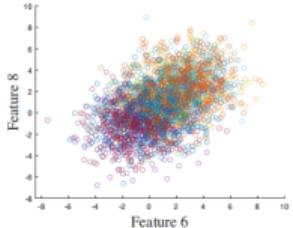
(j) BSUFS

Synthetic Experiments

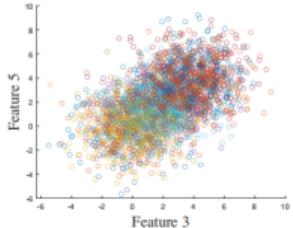
► Diamond9



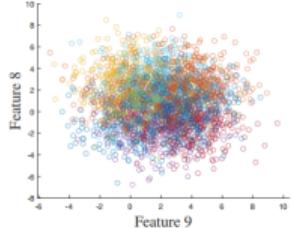
(a) Diamond9



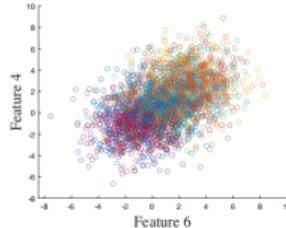
(b) LapScore



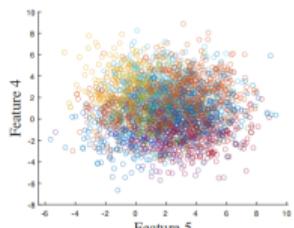
(c) SOGFS



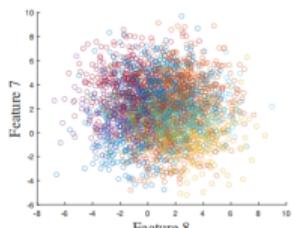
(d) RNE



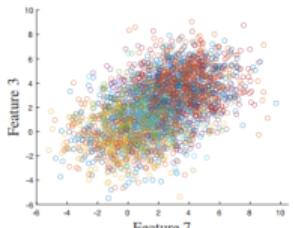
(e) UDFS



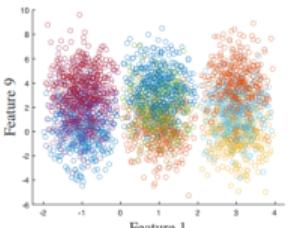
(f) SPCAFS



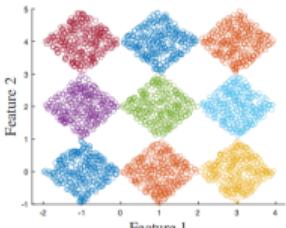
(g) FSPCA



(h) SPCA-PSD



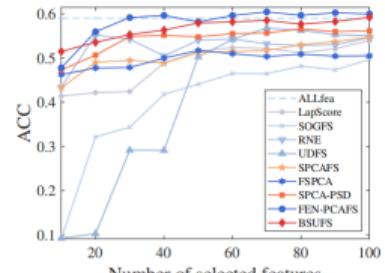
(i) FEN-PCAFS



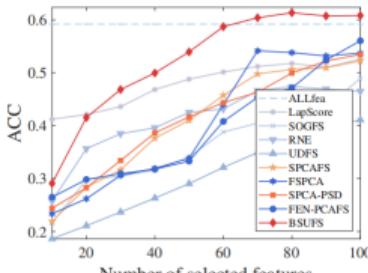
(j) BSUFS

Real Experiments

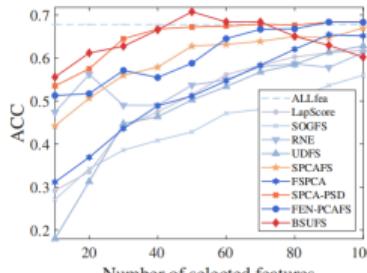
► Accuracy (ACC) ↑



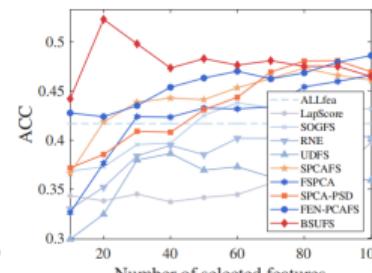
(a) COIL20



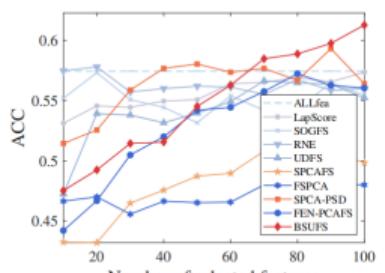
(b) Isolet



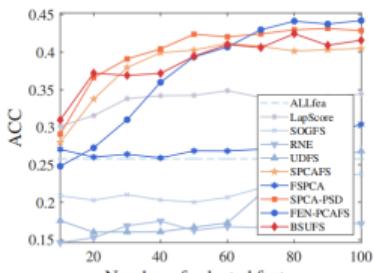
(c) USPS



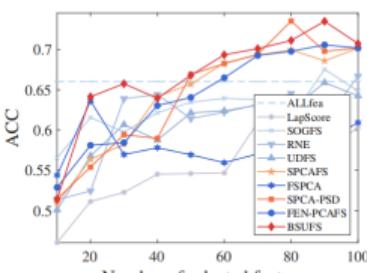
(d) umnist



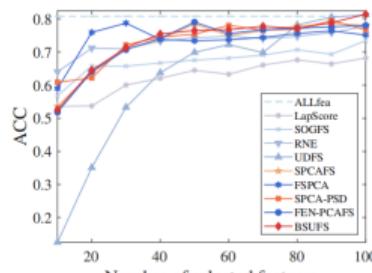
(e) GLIOMA



(f) pie



(g) LUNG



(h) MSTAR

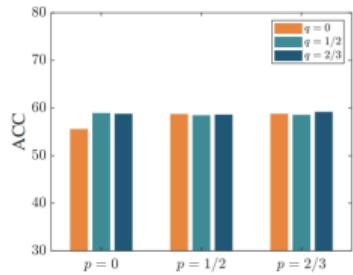
Real Experiments

► Accuracy (ACC) ↑

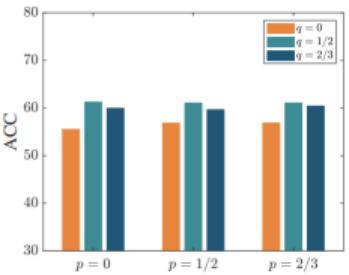
Datasets	ALLfea	LapScore	SOGFS	RNE	UDFS	SPCAFS	FSPCA	SPCA-PSD	FEN-PCAFS	BSUFS
COIL20	58.97±4.99 (10)	53.91±3.61 (100)	56.77±3.09 (70)	49.66±3.63 (100)	55.16±3.35 (20)	51.71±3.05 (50)	54.63±3.64 (100)	56.57±4.08 (80)	60.41±4.41 (70)	59.18±3.49 (100)
Isolet	59.18±3.19 (10)	52.55±2.83 (100)	41.11±1.71 (100)	48.93±2.69 (100)	47.39±2.91 (80)	54.15±2.69 (70)	52.26±2.81 (100)	53.45±2.82 (100)	56.04±3.50 (100)	61.34±3.33 (80)
USPS	67.79±4.96 (10)	61.76±4.52 (100)	62.83±3.79 (100)	56.00±3.48 (100)	61.28±3.46 (100)	65.43±4.90 (90)	66.98±3.92 (100)	68.38±3.85 (100)	68.36±4.62 (90)	70.77±3.73 (50)
umist	41.68±2.46 (10)	39.71±3.28 (100)	38.64±1.61 (40)	43.81±2.98 (60)	41.01±2.25 (90)	46.58±2.34 (100)	47.32±3.48 (80)	48.08±3.06 (90)	48.61±3.23 (100)	52.29±3.61 (20)
GLIOMA	57.44±6.40 (10)	57.36±3.60 (100)	56.64±6.47 (70)	57.32±6.47 (20)	57.80±2.98 (20)	48.04±5.26 (90)	52.08±3.64 (80)	59.32±6.27 (90)	57.24±8.16 (80)	61.28±9.01 (100)
pie	25.79±1.39 (10)	34.86±1.43 (60)	26.82±1.32 (100)	23.78±1.19 (100)	17.49±0.76 (40)	30.39±1.43 (100)	41.16±2.46 (60)	43.16±2.38 (90)	44.21±2.03 (100)	42.45±1.74 (80)
LUNG	66.03±7.23 (10)	60.93±8.02 (70)	65.89±7.43 (90)	67.53±7.73 (90)	66.68±8.32 (100)	63.62±5.45 (20)	70.16±7.71 (100)	73.53±8.91 (80)	70.58±6.88 (90)	73.51±6.80 (90)
MSTAR	80.81±8.76 (10)	68.21±4.57 (100)	81.25±7.48 (100)	73.46±5.61 (100)	77.82±6.16 (100)	78.74±5.20 (30)	78.63±8.68 (90)	79.53±6.75 (90)	79.03±6.02 (50)	81.43±6.89 (100)
Average	57.21±4.92	53.66±3.98	53.74±4.11	52.56±4.22	53.08±3.77	54.83±3.79	57.90±4.54	60.25±4.76	60.56±4.86	62.78±4.83

Real Experiments

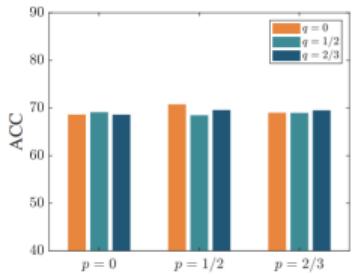
► Effects of p and q



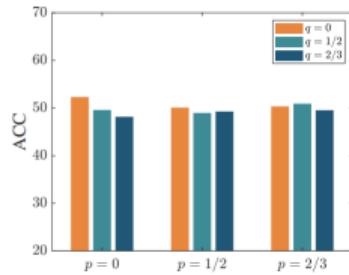
(a) COIL20



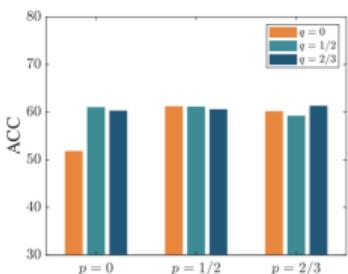
(b) Isolet



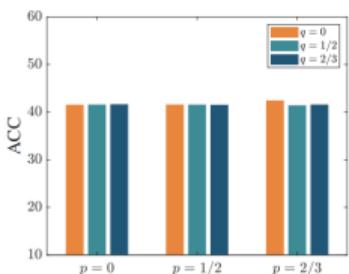
(c) USPS



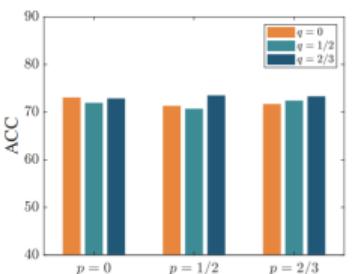
(d) umist



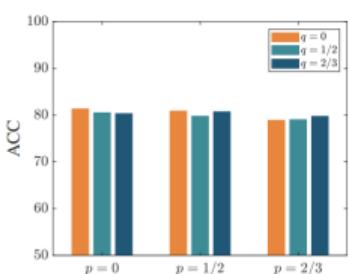
(e) GLIOMA



(f) pie



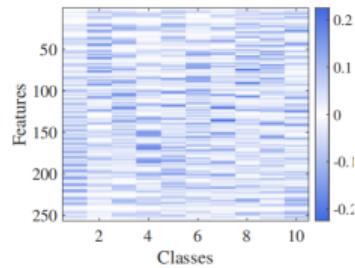
(g) LUNG



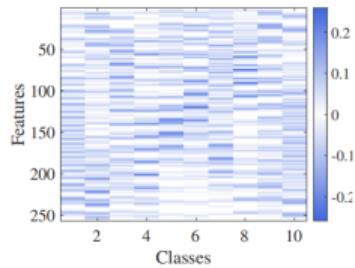
(h) MSTAR

Real Experiments

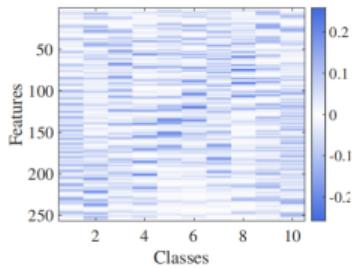
► Ablations on transformation matrices



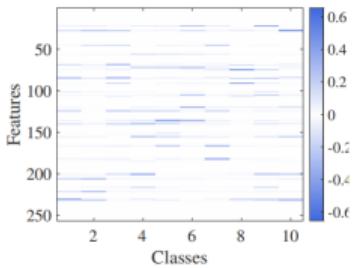
(a) USPS (Case I)



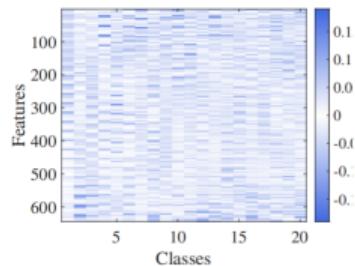
(b) USPS (Case II)



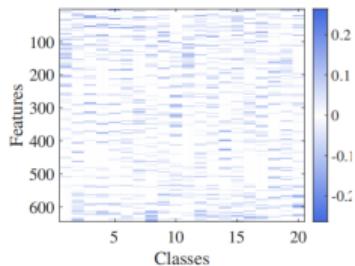
(c) USPS (Case III)



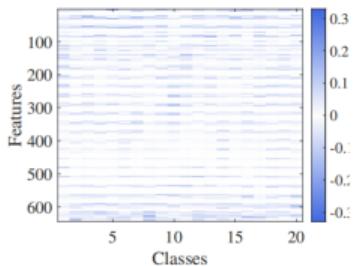
(d) USPS (Case IV)



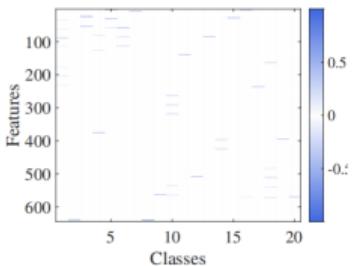
(e) umist (Case I)



(f) umist (Case II)



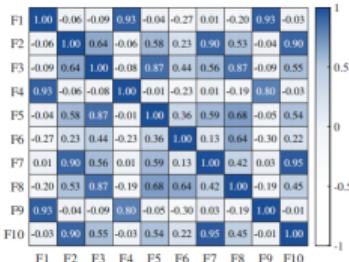
(g) umist (Case III)



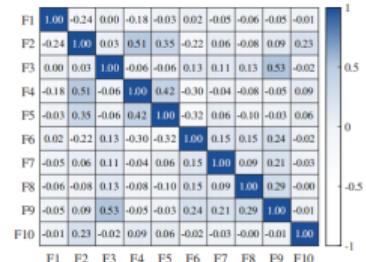
(h) umist (Case IV)

Real Experiments

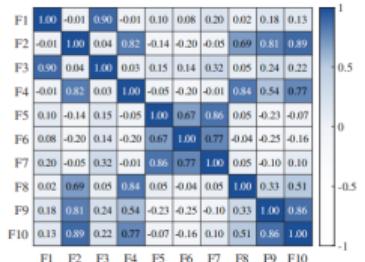
► Heatmaps of feature correlations



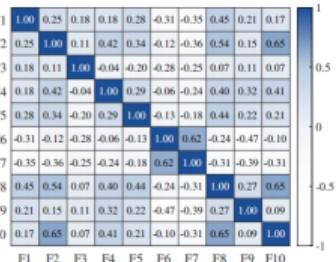
(a) COIL20 (SPCAFS)



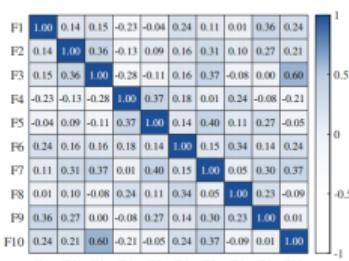
(b) Isolet (SPCAFS)



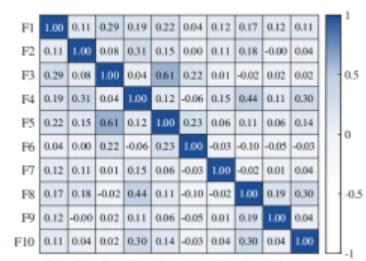
(c) USPS (SPCAFS)



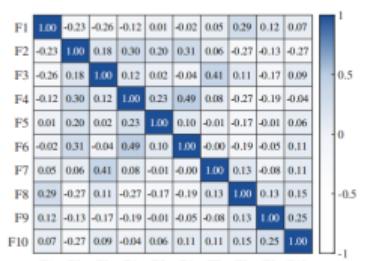
(d) LUNG (SPCAFS)



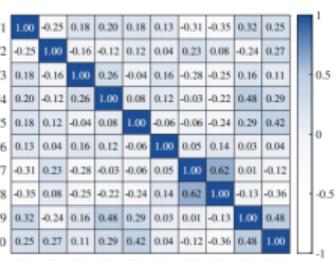
(e) COIL20 (BSUFS)



(f) Isolet (BSUFS)



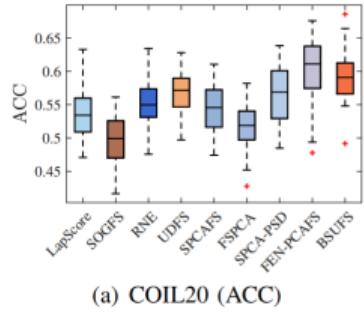
(g) USPS (BSUFS)



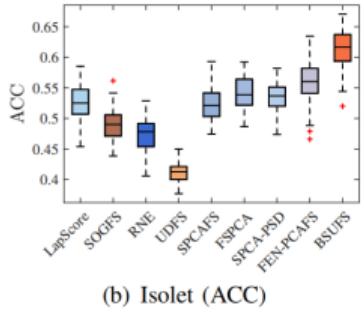
(h) LUNG (BSUFS)

Real Experiments

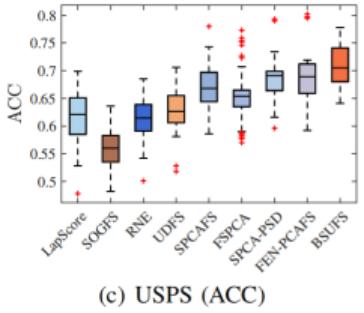
► Model stability analysis



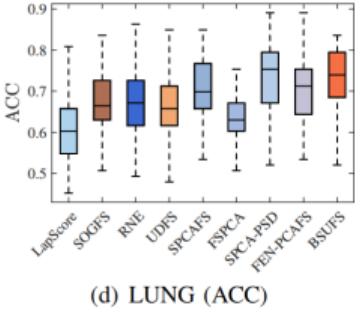
(a) COIL20 (ACC)



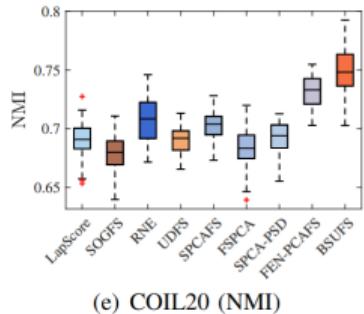
(b) Isolet (ACC)



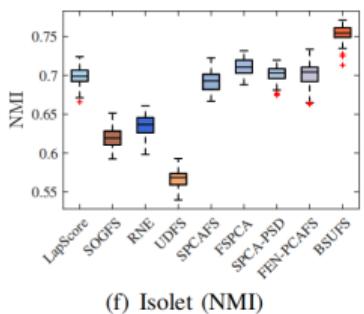
(c) USPS (ACC)



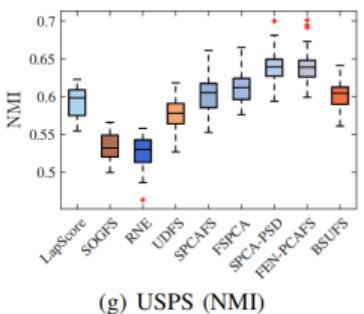
(d) LUNG (ACC)



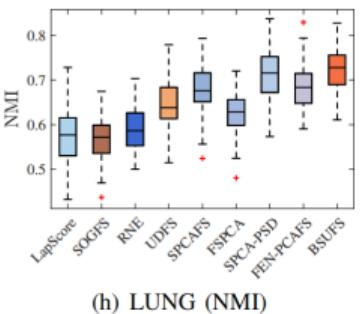
(e) COIL20 (NMI)



(f) Isolet (NMI)



(g) USPS (NMI)



(h) LUNG (NMI)

Outline

Introduction

Bi-Sparse Feature Selection

Deep Unfolding Feature Selection

Future Work

Model

- ▶ Chen-Xiu, Tuning-Free Structured Sparse PCA via Deep Unfolding Networks, CCC, 2025

$$\begin{aligned} \min_W \quad & \frac{1}{2} \|X - WW^\top X\|_F^2 + \lambda \|W\|_{2,1} + \mu \|W\|_1 \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$

- ▶ Alternating direction method of multipliers (ADMM)

$$\begin{aligned} \min_W \quad & \frac{1}{2} \|X - WW^\top X\|_F^2 + \lambda \|Y\|_{2,1} + \mu \|Z\|_1 \\ \text{s.t.} \quad & W^\top W = I, \quad W = Y, \quad W = Z \end{aligned}$$

↓

$$\begin{aligned} \mathcal{L}(W, Y, Z, \Lambda, \Pi) = & \frac{1}{2} \|X - WW^\top X\|_F^2 + \lambda \|Y\|_{2,1} + \mu \|Z\|_1 \\ & + \langle \Lambda, W - Y \rangle + \frac{\alpha}{2} \|W - Y\|_F^2 + \langle \Pi, W - Z \rangle + \frac{\beta}{2} \|W - Z\|_F^2 \end{aligned}$$

SPCA-Net

- ▶ Update W -block

$$\min_W \quad f(W) := \frac{1}{2} \|X - WW^\top X\|_F^2 + \frac{\alpha}{2} \|W - Y^k + \Lambda^k/\alpha\|_F^2 + \frac{\beta}{2} \|W - Z^k + \Pi^k/\beta\|_F^2$$

$$\text{s.t. } W^\top W = I$$

↓

$$\min_W \quad f(W^k) + \langle \nabla f(W^k), W - W^k \rangle + \frac{1}{2\eta} \|W - W^k\|_F^2$$

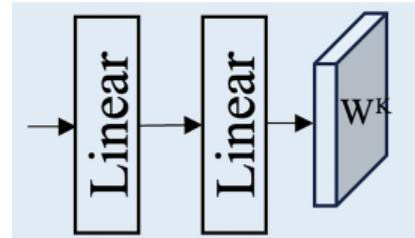
$$\text{s.t. } W^\top W = I$$

↓

$$W^{k+1} = UV^\top$$

↓

$$W^{k+1} = \text{LargNet}(U, V^\top)$$



SPCA-Net

- ▶ Update Y -block

$$\min_Y \lambda \|Y\|_{2,1} + \frac{\alpha}{2} \|X^{k+1} - Y + \Lambda^k/\alpha\|_F^2$$

\Downarrow

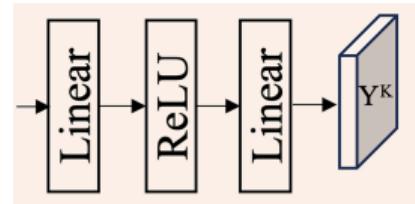
$$Y^{k+1} = \text{sign}(\|X^{k+1} + \Lambda^k/\alpha\|_2) \circ \max(\|X^{k+1} + \Lambda^k/\alpha\|_2 - \lambda/\alpha, 0)$$

\Downarrow

$$Y^{k+1} = \frac{X^{k+1} + \Lambda^k/\alpha}{\|X^{k+1} + \Lambda^k/\alpha\|_2} \text{ReLU}(\|X^{k+1} + \Lambda^k/\alpha\|_2 - \lambda/\alpha)$$

\Downarrow

$$Y^{k+1} = \text{GSoftNet}(X^{k+1} + \Lambda^k/\alpha, \lambda/\alpha)$$



SPCA-Net

- ▶ Update Z -block

$$\min_Z \mu \|Z\|_1 + \frac{\beta}{2} \|X^{k+1} - Z + \Pi^k / \beta\|_F^2$$

↓

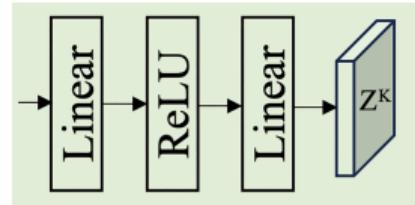
$$Z^{k+1} = \text{sign}(X^{k+1} + \Pi^k / \beta) \circ \max(|X^{k+1} + \Pi^k / \beta| - \mu / \beta, 0)$$

↓

$$Z^{k+1} = \frac{X^{k+1} + \Pi^k / \beta}{|X^{k+1} + \Pi^k / \beta|} \text{ReLU}(|X^{k+1} + \Pi^k / \beta| - \mu / \beta)$$

↓

$$Z^{k+1} = \text{SoftNet}(X^{k+1} + \Pi^k / \beta, \mu / \beta)$$



SPCA-Net

- ▶ **Input:** $X, \lambda, \mu, \alpha, \beta$
- ▶ **Initialize:** $(W^0, Y^0, Z^0, \Lambda^0, \Pi^0)$
- ▶ **While** $k = 1, \dots, K$ **do**
 - ▶ Update W^{k+1} by

$$W^{k+1} = \text{LargNet}(U, V^\top)$$

- ▶ Update Y^{k+1} by

$$Y^{k+1} = \text{GSoftNet}(X^{k+1} + \Lambda^k / \alpha, \lambda / \alpha)$$

- ▶ Update Z^{k+1} by

$$Z^{k+1} = \text{SoftNet}(X^{k+1} + \Pi^k / \beta, \mu / \beta)$$

- ▶ Update Λ^{k+1}, Π^{k+1} by

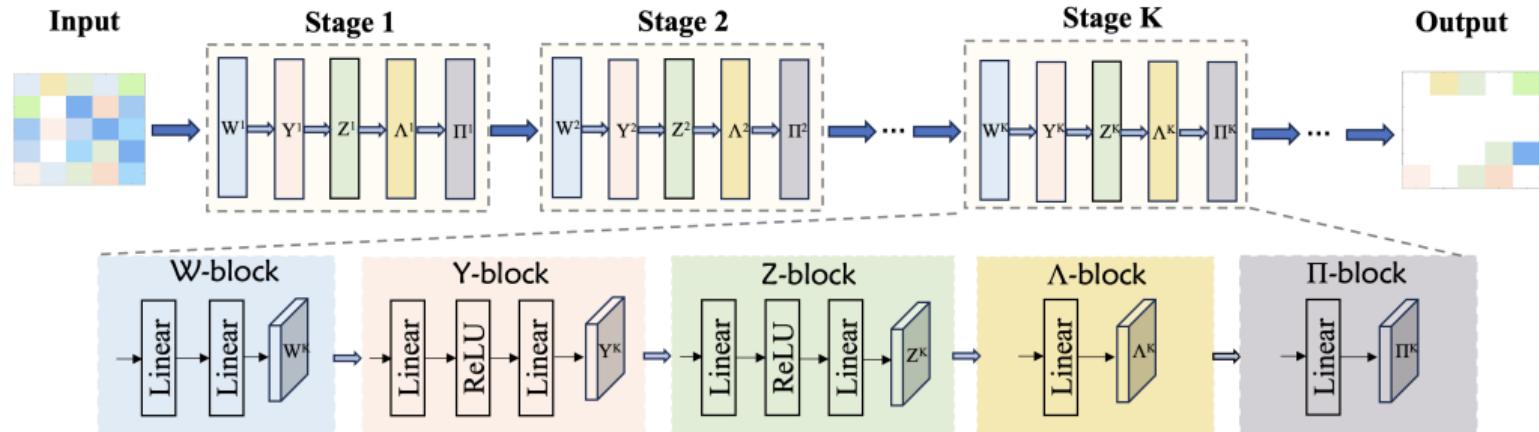
$$\Lambda^{k+1} = \text{Linear}(W^{k+1}, Y^{k+1}, \Lambda^k, \alpha), \quad \Pi^{k+1} = \text{Linear}(W^{k+1}, Z^{k+1}, \Pi^k, \beta)$$

- ▶ **Output:** Trained W

SPCA-Net

- All parameters $(\lambda, \mu, \alpha, \beta)$ are trained in an end-to-end manner
- The loss is defined as

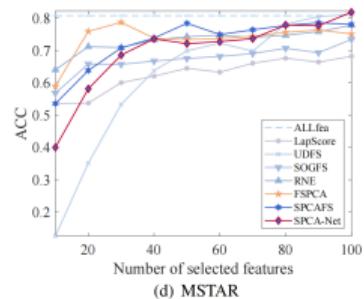
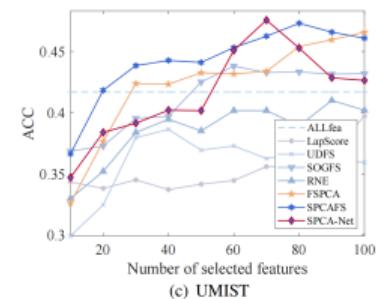
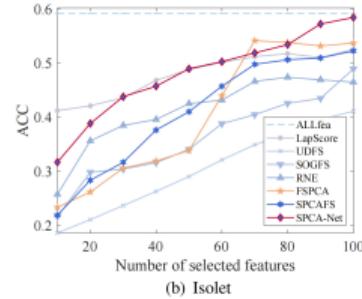
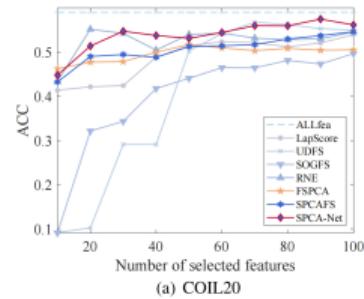
$$\text{Loss} = \frac{1}{2} \|X - \bar{W}\bar{W}^\top X\|_F^2 + \lambda\|\bar{W}\|_{2,1} + \mu\|\bar{W}\|_1$$



Real Experiments

► Accuracy (ACC) ↑

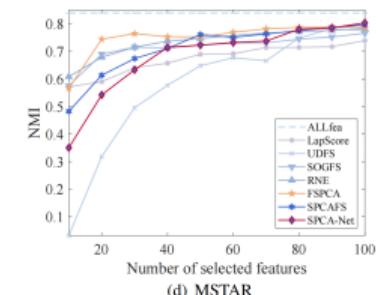
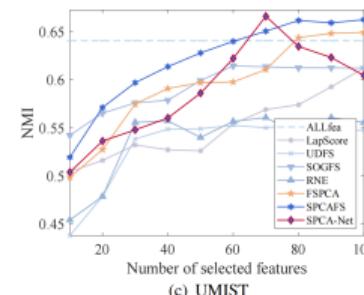
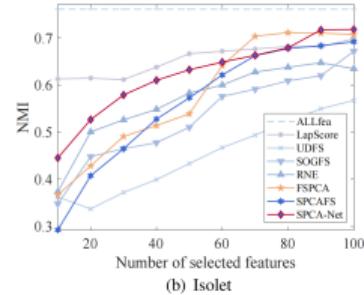
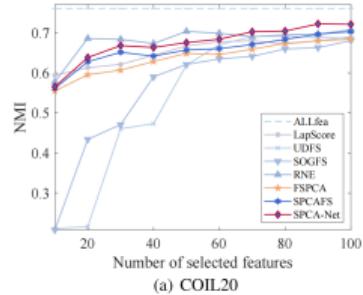
Datasets	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAFS	SPCA-Net
COIL20	58.97±4.99 (10)	53.91±3.61 (100)	56.70±3.09 (70)	49.66±3.63 (100)	55.16±3.35 (20)	51.71±3.05 (50)	54.63±3.64 (100)	57.46±2.76 (90)
Isolet	59.18±3.19 (10)	52.55±2.83 (100)	41.11±1.71 (100)	48.93±2.69 (100)	47.39±2.91 (80)	54.15±2.69 (70)	52.26±2.81 (100)	58.43±4.31 (100)
UMIST	41.68±2.46 (10)	39.71±3.28 (100)	38.64±1.61 (40)	43.81±2.98 (80)	41.01±2.25 (90)	46.58±2.34 (100)	47.32±3.48 (80)	47.58±4.97 (70)
MSTAR	80.81±8.76 (10)	68.21±4.57 (100)	81.25±7.48 (100)	73.46±5.61 (100)	77.82±6.16 (100)	78.74±5.20 (30)	78.63±8.68 (90)	81.90±6.87 (100)



Real Experiments

► Normalized mutual information (NMI) ↑

Datasets	ALLfea	LapScore	UDFS	SOGFS	RNE	FSPCA	SPCAF	SPCA-Net
COIL20	76.04±1.69 (10)	69.01±1.53 (100)	69.12±1.17 (80)	68.03±1.59 (100)	70.76±2.07 (100)	68.41±1.60 (100)	70.29±1.31 (100)	72.21±2.68 (90)
Isolet	76.09±1.77 (10)	69.86±1.26 (100)	56.73±1.05 (100)	67.15±1.45 (100)	64.74±1.28 (90)	71.12±1.11 (80)	69.18±1.33 (100)	71.80±1.59 (100)
UMIST	64.07±1.76 (10)	61.23±2.15 (100)	55.43±1.50 (80)	61.46±2.03 (70)	56.08±1.80 (60)	64.94±1.65 (100)	66.26±1.74 (100)	66.62±7.52 (70)
MSTAR	83.96±3.14 (10)	73.90±1.62 (100)	78.18±3.64 (90)	76.56±1.54 (100)	78.26±2.51 (100)	78.87±2.52 (90)	79.62±2.30 (100)	80.67±3.47 (90)



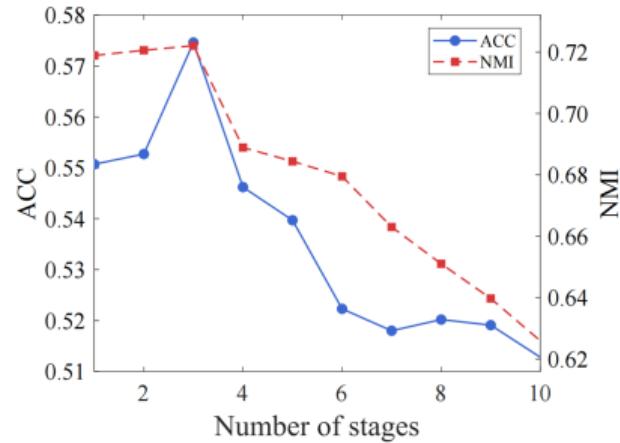
Real Experiments

► Ablation studies

Datasets	Network	ACC ↑	NMI ↑
COIL20	✗	55.12±2.67	70.44±1.37
	✓	57.46±2.76	72.21±2.68
Isolet	✗	51.84±2.82	67.02±1.43
	✓	58.43±4.31	71.80±1.59
UMIST	✗	40.65±2.29	55.88±1.62
	✓	47.58±4.97	66.62±7.52
MSTAR	✗	80.65±6.47	80.53±2.41
	✓	81.90±6.87	80.67±3.47

Datasets	Dynamic	ACC ↑	NMI ↑
COIL20	✗	56.71±3.83	71.49±3.67
	✓	57.46±2.76	72.21±2.68
Isolet	✗	52.06±3.71	68.91±2.36
	✓	58.43±4.31	71.80±1.59
UMIST	✗	42.63±2.78	60.12±1.69
	✓	47.58±4.97	66.62±7.52
MSTAR	✗	80.74±5.28	80.59±3.67
	✓	81.90±6.87	80.67±3.47

► Effect of deep unfolding stages



► More comparisons with BSUFS?

Outline

Introduction

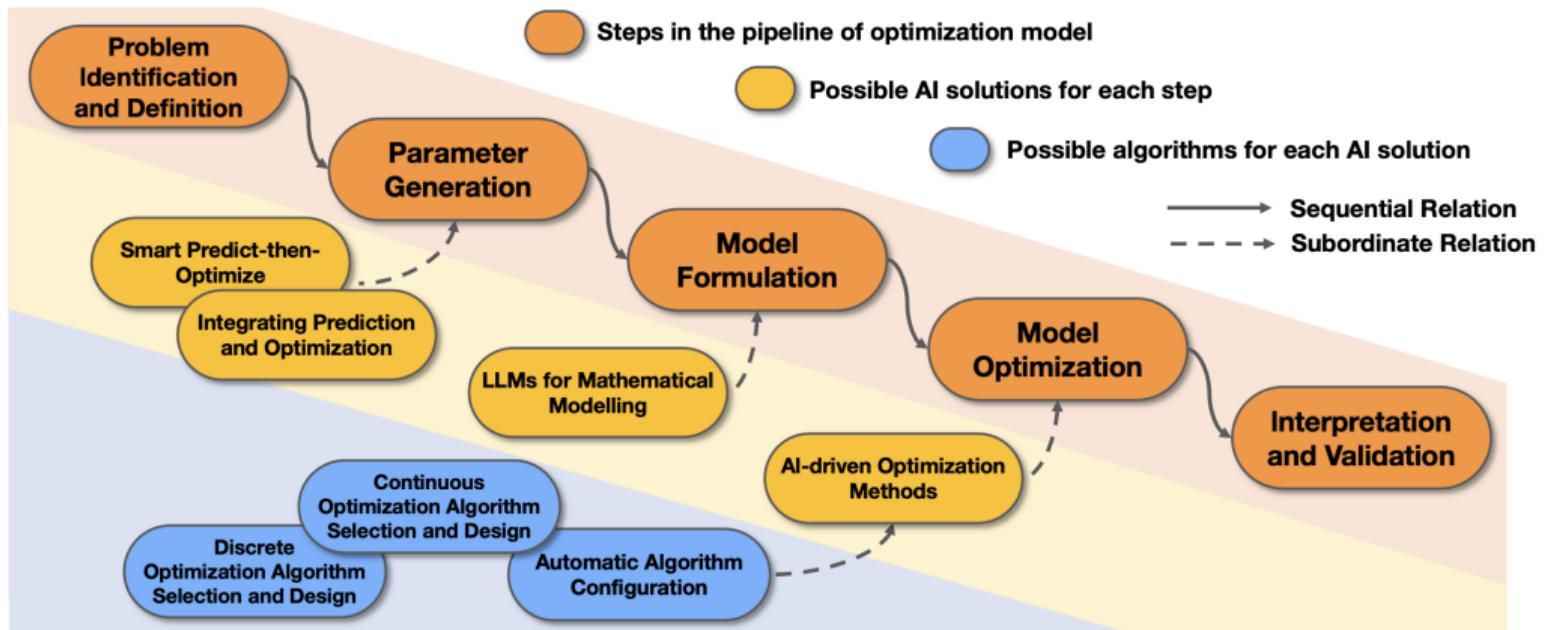
Bi-Sparse Feature Selection

Deep Unfolding Feature Selection

Future Work

Future Work

► Large Language Models for Optimization

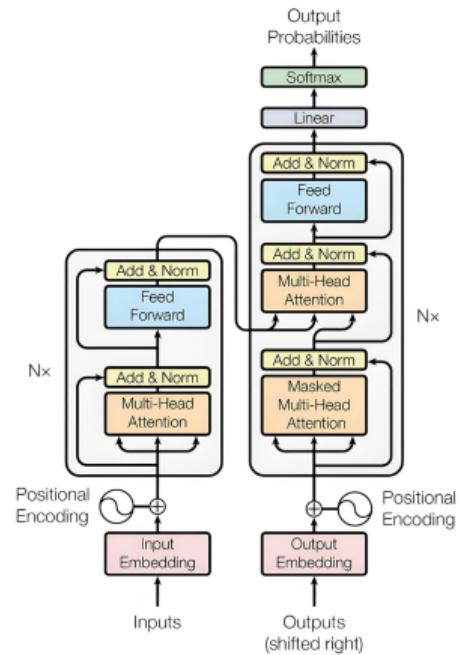
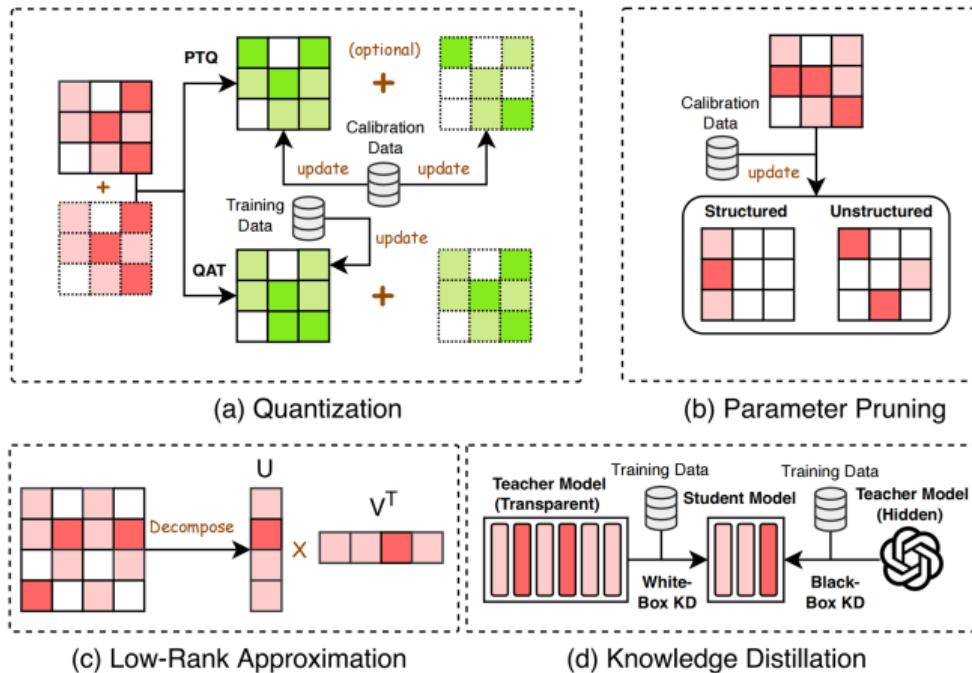


Future Work

- ▶ Ramamonjison-Yu-Li et al, NL4Opt Competition: Formulating Optimization Problems Based on Their Natural Language Descriptions, [NeurIPS](#), 2022
- ▶ Yang-Wang-Lu et al, Large Language Models as Optimizers, [ICLR](#), 2024
- ▶ AhmadiTeshnizi-Gao-Udell, OptiMUS: Scalable Optimization Modeling with (MI)LP Solvers and Large Language Models, [ICML](#), 2024
- ▶ Gao-Jiang-Cai et al, StrategyLLM: Large Language Models as Strategy Generators, Executors, Optimizers, and Evaluators for Problem Solving, [NeurIPS](#), 2024
- ▶ Romera-Paredes-Barekatain et al, Mathematical Discoveries from Program Search with Large Language Models, [Nature](#), 2024
- ▶ Jiang-Shu-Qian et al, LLMOPT: Learning to Define and Solve General Optimization Problems from Scratch, [ICLR](#), 2025
- ▶ Huang-Tang-Hu et al, ORLM: A Customizable Framework in Training Large Models for Automated Optimization Modeling, [Operations Research](#), 2025
- ▶ Chen-Flores-Mantri et al, OptiChat: Bridging Optimization Models and Practitioners with Large Language Models, [INFORMS Journal on Data Science](#), 2025

Future Work

► Optimization for Large Language Models



Future Work

- ▶ Frantar-Alistarh, SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot, [ICML](#), 2023
- ▶ Ma-Fang-Wang, LLM-Pruner: On the Structural Pruning of Large Language Models, [NeurIPS](#), 2023
- ▶ Sun-Liu-Bair et al, A Simple and Effective Pruning Approach for Large Language Models, [ICLR](#), 2024
- ▶ Deng-Jiao-Liu, et al, DRPruning: Efficient Large Language Model Pruning through Distributionally Robust Optimization, [ACL](#), 2025
- ▶ Zhao-Hu-Li, et al, FISTAPruner: Layer-wise Post-training Pruning for Large Language Models, [EMNLP](#), 2025
- ▶ Liu-Liu-Wang et al, ARMOR: High-Performance Semi-Structured Pruning via Adaptive Matrix Factorization, [arXiv](#), 2025

Thank you for your attention!

xcxiu@shu.edu.cn