Bi-Sparse Unsupervised Feature Selection

Xianchao Xiu

Department of Automation

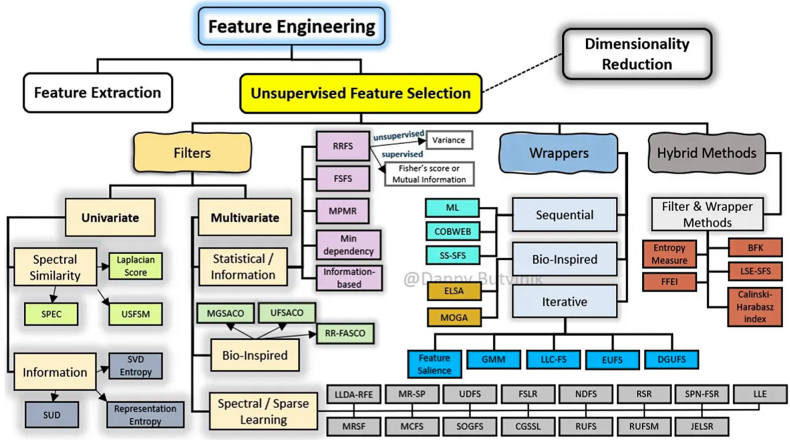
Operations Research Society of China, October 18-21, 2024

Joint work with Chenyi Huang (SHU), Pan Shang (CAS) and Wanquan Liu (SYSU)

Outline

UFS

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



PCA

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

$$\begin{aligned} \min_{W \in \mathbb{R}^{d \times p}} \quad & \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & W^\top W = I_p \end{aligned}$$

$$\Updownarrow$$

$$\begin{aligned} \min_{W \in \mathbb{R}^{d \times p}} \quad & -\operatorname{Tr}(W^\top X X^\top W) \\ \text{s.t.} \quad & W^\top W = I_p \end{aligned}$$

- Unsupervised feature selection by sparse PCA

$$\begin{aligned} \min_{W \in \mathbb{R}^{d \times p}} \quad & -\operatorname{Tr}(W^\top X X^\top W) \\ \text{s.t.} \quad & W^\top W = I_p,\ \|W\|_{2,0} \le s \end{aligned}$$

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

SOTA

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

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

- Li-Sun-Zhang, Unsupervised Feature Selection via Nonnegative Orthogonal Constrained Regularized Minimization, arXiv:2403.16966

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

- 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

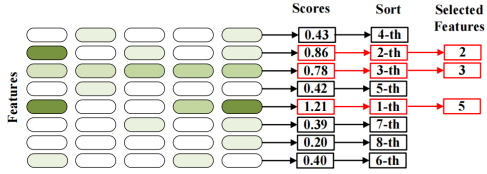
New Formulation

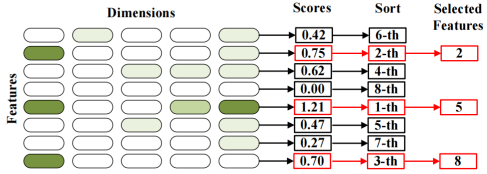
- Xiu-Huang-Shan-Liu, Bi-Sparse Unsupervised Feature Selection, IEEE TIP, 2025

$$\min_{W} \; -\mathrm{Tr}(W^\top S W) \; + \lambda \|W\|_{2,p}^{p} \; (0 < p < 1)$$
$$\text{s.t.} \;\; W^\top W = I$$

$$\Downarrow$$

$$\min_{W} \; -\mathrm{Tr}(W^\top S W) \; + \lambda_1 \|W\|_{2,p}^{p} + \lambda_2 \|W\|_{q}^{q} \; (0 \le p, q < 1)$$
$$\text{s.t.} \;\; W^\top W = I$$





Optimization Algorithm

- First-order algorithm: PAM (Proximal Alternating Method)
- Model reformulation

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

$$\Downarrow$$

$$\begin{aligned} \min_{W,U,V} \quad & -\mathrm{Tr}(W^\top SW) + \lambda_1 \|V\|_{2,p}^p + \lambda_2 \|U\|_q^q \\ \text{s.t.} \quad & W^\top W = I, V = W, U = W \end{aligned}$$

$$\Downarrow$$

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

Optimization Algorithm

- Input: X, λ_1, λ_2, β_1, β_2, p,q, τ_1, τ_2, τ_3
- Initialize: W^0, U^0, V^0
- While not converged do
  - Update W^{k+1} by

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

  - Update U^{k+1} by

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

  - Update V^{k+1} by

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

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

Update W

- Riemannian gradient

$$\begin{aligned} \min_{W} \quad & -\operatorname{Tr}(W^\top S W) + \frac{\beta_1}{2}\|W - U^k\|_{\mathrm{F}}^2 + \frac{\beta_2}{2}\|W - V^k\|_{\mathrm{F}}^2 + \frac{\tau_1}{2}\|W - W^k\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & W^\top W = I \end{aligned}$$

$$\Downarrow$$

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

$$\Downarrow$$

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

Update W

- Riemannian Hessian

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

$$\Downarrow$$

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

$$\Downarrow$$

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

Update W

- Input: S, U^k, V^k, β_1, β_2, τ_1, ε, $\Delta' > 0$, $\rho' \in [0, \frac{1}{4})$
- While not converged do
  - Obtain η_i by solving trust domain subproblem

$$\min_{\eta \in \mathrm{T}_W \mathrm{St}(d,m)} \quad m_W(\eta) = g(W) + \mathrm{Tr}(\eta^\top \operatorname{grad} g(W)) + \frac{1}{2} \operatorname{vec}(\eta)^\top \operatorname{Hess} g(W) \operatorname{vec}(\eta)$$
$$\text{s.t.} \qquad \mathrm{Tr}(\eta^\top W \eta^\top) \leq \Delta^2$$

  - 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

- Closed-form solution

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

$$\Downarrow$$

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

$$\Downarrow$$

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

$$\Downarrow$$

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

Lemma

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

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

where

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

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

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

Update V

- Closed-form solution

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

$$\Downarrow$$

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

$$\Downarrow$$

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

$$\Downarrow$$

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

Outline

Experimental Details

- 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 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:

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

Experimental Details

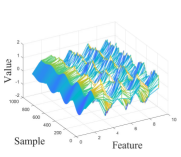
- Selected datasets

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

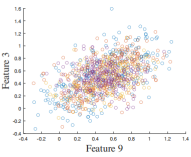
- Evaluation metrics
  - ACC: Accuracy
  - NMI: Normalized mutual information

Synthetic Experiments

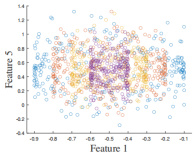
- Dartboard1



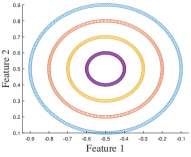
(a) Dartboard1



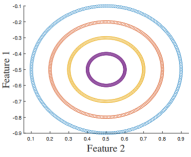
(b) LapScore



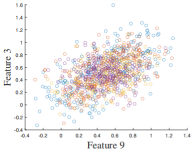
(c) SOGFS



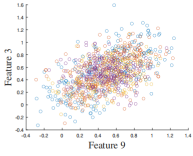
(d) RNE



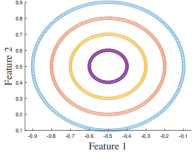
(e) UDFS



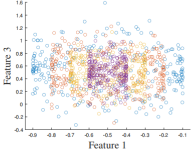
(f) SPCAFS



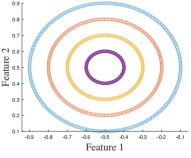
(g) FSPCA



(h) SPCA-PSD



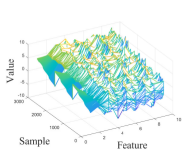
(i) FEN-PCAFS



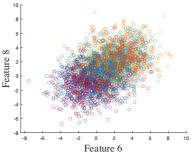
(j) BSUFS

Synthetic Experiments

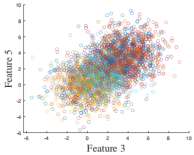
- Diamond9



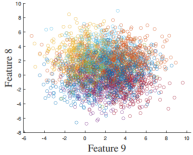
(a) Diamond9



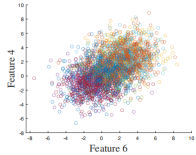
(b) LapScore



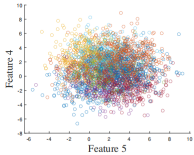
(c) SOGFS



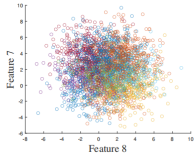
(d) RNE



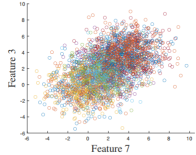
(e) UDFS



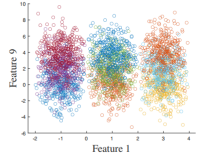
(f) SPCAFS



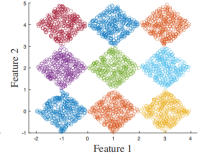
(g) FSPCA



(h) SPCA-PSD



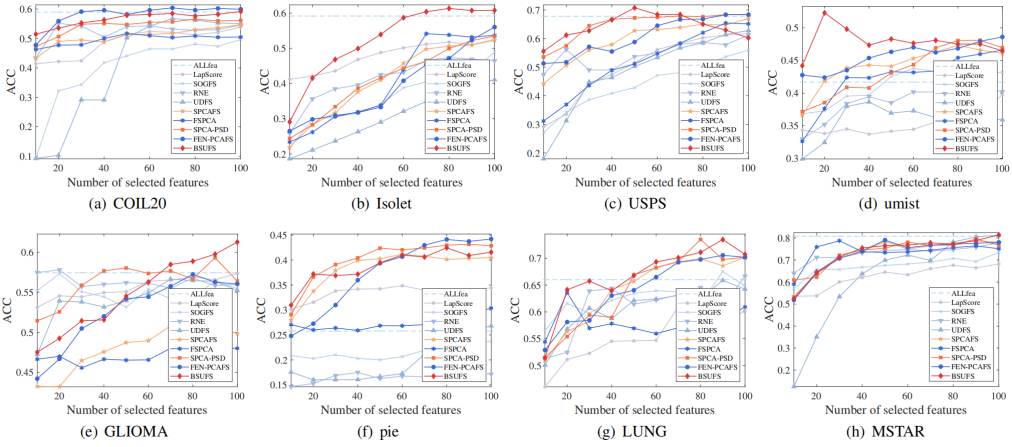
(i) FEN-PCAFS



(j) BSUFS

Real Experiments

- ACC comparisons



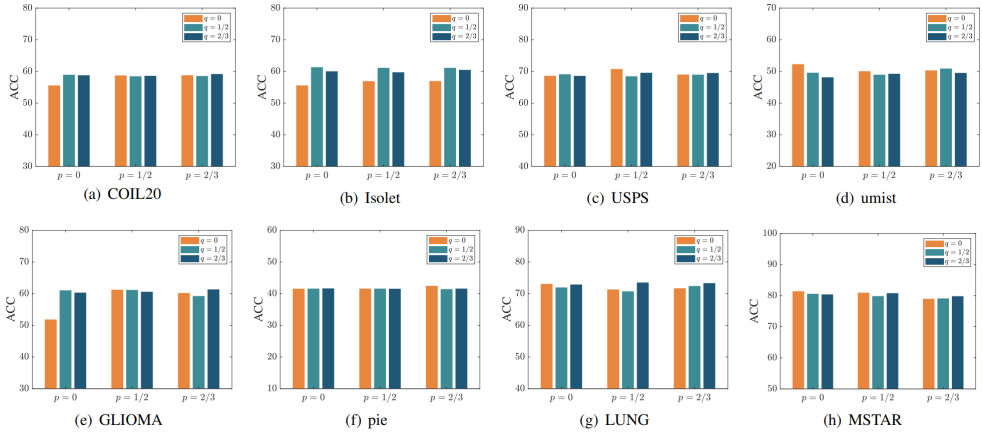
Real Experiments

- ACC comparisons

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**

Effects of p and q

- ACC comparisons

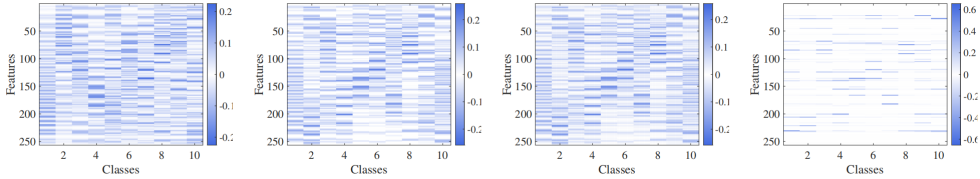


(a) COIL20 (b) Isolet (c) USPS (d) umist

(e) GLIOMA (f) pie (g) LUNG (h) MSTAR

Ablation Experiments

- Transformation visualizations

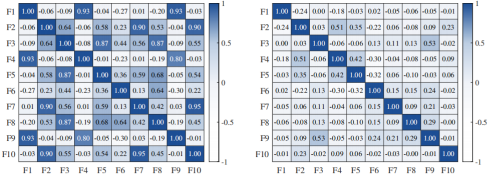


(a) USPS (Case I) (b) USPS (Case II) (c) USPS (Case III) (d) USPS (Case IV)

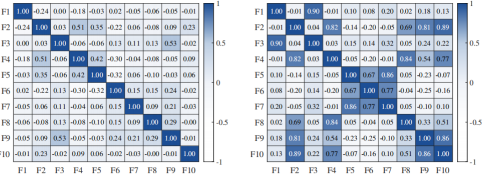
(e) umist (Case I) (f) umist (Case II) (g) umist (Case III) (h) umist (Case IV)

Feature Correlation

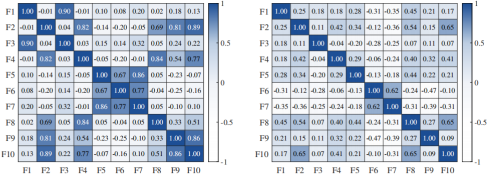
- Heatmap visualizations



(a) COIL20 (SPCAFS)

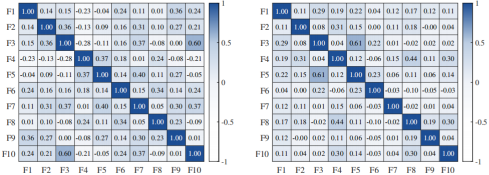


(b) Isolet (SPCAFS)

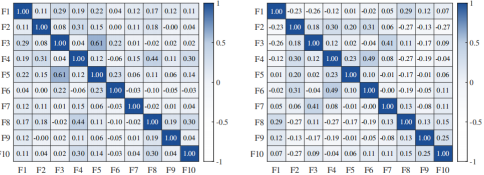


(c) USPS (SPCAFS)

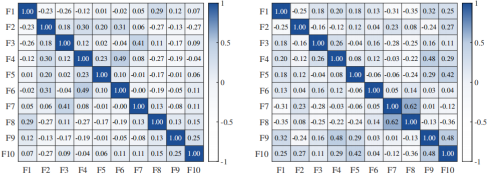
(d) LUNG (SPCAFS)



(e) COIL20 (BSUFS)



(f) Isolet (BSUFS)

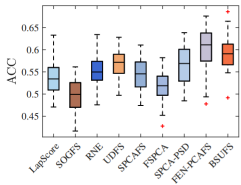


(g) USPS (BSUFS)

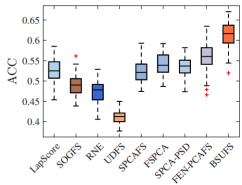
(h) LUNG (BSUFS)

Model Stability

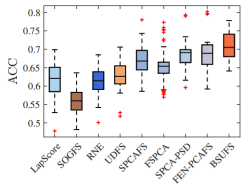
- Box-plots visualizations



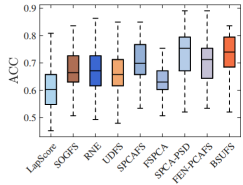
(a) COIL20 (ACC)



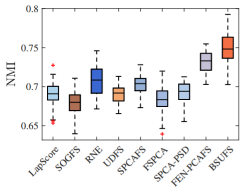
(b) Isolet (ACC)



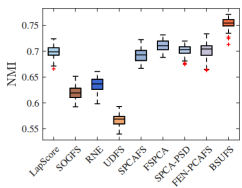
(c) USPS (ACC)



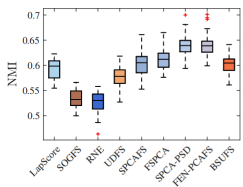
(d) LUNG (ACC)



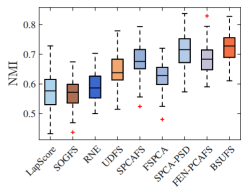
(e) COIL20 (NMI)



(f) Isolet (NMI)



(g) USPS (NMI)



(h) LUNG (NMI)

Outline

Conclusions and Future Work

- How to learn data distributions?

$$\begin{aligned} \min_{W} \ & \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 + \lambda_1 \|W\|_{2,p}^p + \lambda_2 \|W\|_q^q \\ \text{s.t.} \ & W^\top W = I \end{aligned}$$

  - Convex loss: ℓ_1-norm, $\ell_{2,1}$-norm, quantile, Huber
  - Nonconvex loss: ℓ_p-norm, $\ell_{2,p}$-norm, SCAD, MCP, capped ℓ_1

- Xiu-Yang-Li, Unsupervised feature selection via sparse and low-rank contrastive learning, Operations Research Transactions, 2025

A simple framework for contrastive learning of visual representations
T Chen, S Kornblith, M Norouzi... - ... on machine learning, 2020 - proceedings.mlr.press
... In our contrastive learning, as positive pairs are computed in the same device, the model can exploit the local information leakage to improve prediction accuracy without improving ...
☆ 保存 99 引用 被引用次数：22684 相关文章 所有 24 个版本 »

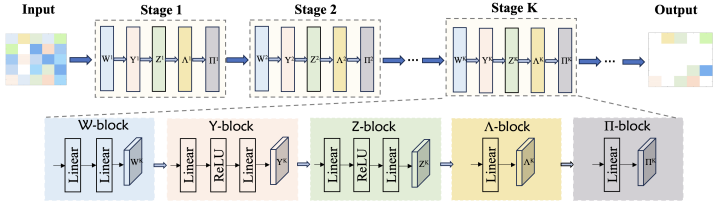
Conclusions and Future Work

- How to learn regularization parameters?

$$\min_{W} \ \frac{1}{2}\|X - WW^\top X\|_{\mathrm{F}}^2 + \lambda_1 \|W\|_{2,p}^p + \lambda_2 \|W\|_q^q$$
$$\text{s.t.} \ \ W^\top W = I$$

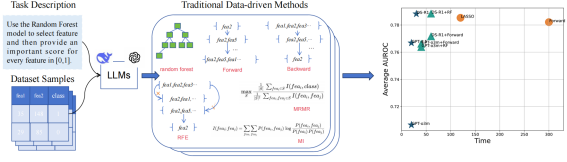
  - $\lambda_1, \lambda_2 \in \{10^{-6}, 10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}, 10^{6}\}$
  - β_1, β_2, p,q, τ_1, τ_2, τ_3

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



Conclusions and Future Work

- How to learn feature selection?
  - Cho-Cund-Srivastava et al, LMPriors: Pre-Trained Language Models as Task-Specific Priors, NeurIPS, 2022
  - Han-Yoon-Arik et al, Large Language Models Can Automatically Engineer Features for Few-Shot Tabular Learning, ICML, 2024
  - Li-Tan-Liu, Exploring Large Language Models for Feature Selection: A Data-centric Perspective, SIGKDD, 2025
- Li-Xiu, LLM4FS: Leveraging Large Language Models for Feature Selection, CAC, 2025



Thank you for your attention!

xcxiu@shu.edu.cn