GSPL: A Succinct Kernel Model for Group-Sparse Projections Learning of Multiview Data

Danyang Wu[1,4] , **Jin Xu**[2*] , **Xia Dong**[1,4] , **Meng Liao**[2] , **Rong Wang**[3,4] ,
Feiping Nie[1,4*] and **Xuelong Li**[1,4]
[1]School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN)
[2]Data Quality Team, WeChat, Tencent Inc., Guangdong, P. R. China
[3]School of Cybersecurity and School of Artificial Intelligence, Optics and Electronics (iOPEN)
[4]Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
danyangwu41x@mail.nwpu.edu.cn, {jinxxu, maricoliao}@tencent.com, {xiadongpgh, feipingnie}@gmail.com, wangrong07@tsinghua.org.cn, li@nwpu.edu.cn

Abstract

This paper explores a succinct kernel model for Group-Sparse Projections Learning (GSPL), to handle multiview feature selection task completely. Compared to previous works, our model has the following useful properties: ***1) Strictness***: GSPL innovatively learns group-sparse projections strictly on multiview data via $\ell_{2,0}$-norm constraint, which is different with previous works that encourage group-sparse projections softly. ***2) Adaptivity***: In GSPL model, when the total number of selected features is given, the numbers of selected features of different views can be determined adaptively, which avoids artificial settings. Besides, GSPL can capture the differences among multiple views adaptively, which handles the inconsistent problem among different views. ***3) Succinctness***: Except for the intrinsic parameters of projection-based feature selection task, GSPL does not bring extra parameters, which guarantees the applicability in practice. To solve the optimization problem involved in GSPL, a novel iterative algorithm is proposed with rigorously theoretical guarantees. Experimental results demonstrate the superb performance of GSPL on synthetic and real datasets.

1 Introduction

Recently, a large amount of multiview data have emerged in real scenarios, in which the features are characterized from different sources, extractors and external environments. For example, a person can be described via iris, fingerprint and face features, a document can be narrated via multiple languages, an image can be described by different feature extractors, *etc* [Zhang *et al.*, 2016; Nie *et al.*, 2018; Wu *et al.*, 2020]. Considering high dimensionality of multiview data and expensive cost of label acquisition, how to select discriminative features on multiview data with unsupervised paradigm has attracted more attentions in recent years.

Related works. The previous works can be simply divided into two categories, including serial models and parallel models. 1) The serial models directly concatenate multiview feature vectors into a long vector, and explore a whole sparse Projection Matrix (PM) via employing single view unsupervised feature selection models (SOCFS [Han and Kim, 2015], SOGFS [Nie *et al.*, 2016], URAFS [Li *et al.*, 2018] models are frequently used), in which the discriminative features can be selected from sparse PM. Except for directly utilizing single-view models, [Dong *et al.*, 2019] proposed a ACSL model which learns both a unified projection and a collaborative graph to mine more discriminative information. 2) The parallel models explore a sparse PM for each view and integrate the selected features of different views artificially. The most exemplary works contain: ASVW [Hou *et al.*, 2017], which collectively learns sparse PM for each view via $\ell_{2,p}$-norm regularization and a common graph from multiple views; MVUFS [Qian and Zhai, 2014], which performs joint local learning regularized orthogonal nonnegative matrix factorization and $\ell_{2,1}$-norm minimization; RMFS [Liu *et al.*, 2016], which employs K-means to efficiently obtain labels for guiding $\ell_{2,1}$-norm regularized feature selection; CGMV-UFS [Tang *et al.*, 2018], which incorporates feature selection with $\ell_{2,1}$-norm constraint into non-negative matrix factorization (NMF) based clustering model.

Confronting problems. However, previous works suffer from several open flaws. In serial models, the inconsistency of multiple views cannot be captured completely, especially when concatenating multiview feature vectors. In parallel models, the number of selected features for each view depends on artificial settings, which is hard to apply in practice. Furthermore, both series and parallel works conventionally utilize $\ell_{2,1}$ or $\ell_{2,p}$-norm regularization to encourage the sparsity of PM, which cannot guarantee the group-sparsity of PM and leads to mismatched problem between the group-sparsity of PM and the number of selected features. More importantly, previous works contain several hyper-parameters for exploring the adaptive combination of different regularizations, including $\ell_{2,1}$ or $\ell_{2,p}$-norm regularization, which affects the applicability of model. Through the above analysis, we can conclude that a complete multiview feature selection model

[*]Corresponding Author

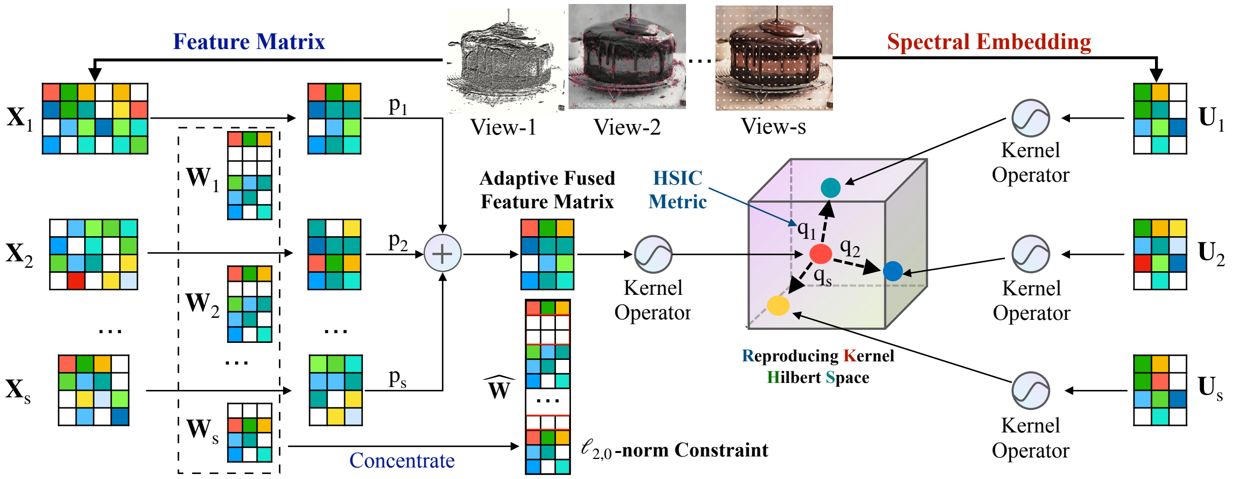


Figure 1: The workflow of GSPL Model.

should handle the following 4 problems:

- ***IC Problem*** **:** How to capture the ***InConsistency*** among multiple views?
- ***SF Problem*****:** How to allocate the number of ***Selected Features*** for multiple views based on a total number?
- ***GS Problem*****:** How to guarantee the ***Group-Sparsity*** of learnt projection matrix (matrices)?
- ***SC Problem*****:** How to guarantee the ***SuCcinctness*** of model?

Our proposals. To this goal, we propose a succinct kernel model for Group-Sparse Projections Learning (GSPL), to handle ***IC***, ***SF***, ***GS*** and ***SC*** problems simultaneously. The workflow is plotted into Fig. 1. Specifically, GSPL first learns a fused feature matrix, mapped and fused by sparse projection matrices and adaptive weights, and then encloses it to mutiview spectral embeddings based on Hilbert-Schmidt Independence Criterion (HSIC) on Reproducing Kernel Hilbert Spaces (RKHSs) adaptively. As aforementioned in Abstract, the proposed GSPL contains three useful functionalities, including ***Adaptivity***, ***Strictness*** and ***Succinctness***, wherein ***Adaptivity*** solves problems ***IC*** and ***SF*** via learning $\mathbf{p}$ and $\mathbf{z}$ in Fig. 1. ***Strictness*** solves problem ***GS*** via $\ell_{2,0}$-norm constraint of $\widehat{\mathbf{W}}$ in Fig. 1, ***Succinctness*** means that that our model does not add any hyper-parameter except for the number of selected features and the number of reduced dimensionality brought by feature selection task, which handles the ***SC*** problem effectively. Except for the above three functionalities, the HSIC metric on RKHSs helps the learnt fused feature matrix to capture the high-order information from multiview spectral embeddings. To solve the optimization problem involved in GSPL, we propose a novel iterative algorithm with theoretical guarantees. Eventually, the effectiveness of GSPL and the convergent speed of proposed algorithm are evaluated. The experimental results demonstrate that the proposed GSPL model achieves SOTA performance and the proposed solver can converge rapidly. In a word, we concisely summarize the main contributions of this paper as follows:

- We propose a GSPL model to solve the ***IC***, ***SF***, ***GS*** and ***SC*** vital problems simultaneously and effectively.
- We propose a novel iterative algorithm to solve the optimization problem involved in GSPL.
- The proposed GSPL model has ***no extra hyper-parameters***, but achieves ***SOTA*** performance in several real datasets compared to previous works.

2 Preliminaries

Hilbert-Schmidt Independence Criterion (HSIC). Given two Reproducing Kernel Hilbert Spaces (RKHSs) $\mathcal{F}$ and $\mathcal{G}$, and a joint measure p with $(\mathcal{A}\times\mathcal{B},\mathcal{P}\times\mathcal{J})$, where $\mathcal{A}$, $\mathcal{B}$ are separable spaces, $\mathcal{P}$ and $\mathcal{J}$ are Borel sets on $\mathcal{A}$ and $\mathcal{B}$ separately, HSIC is defined as the squared Hilbert-Schmidt norm with the cross-covariance. If there are n observations $\mathcal{Z}=\{(\mathbf{a}_1,\mathbf{b}_1),...,(\mathbf{a}_n,\mathbf{b}_n)\}\subseteq\mathcal{A}\times\mathcal{B}$ independently drawn from p, and the corresponding data matrices are $\mathbf{A}\in\mathbb{R}^{\mathrm{n}\times\mathrm{t}_1}$ and $\mathbf{B}\in\mathbb{R}^{\mathrm{n}\times\mathrm{t}_2}$, an empirical HSIC [Gretton *et al.*, 2005; Zhang *et al.*, 2018] can be written as

$$\mathcal{Q}(\mathcal{Z},\mathcal{F},\mathcal{G})=(\mathrm{n}-1)^{-2}\mathbf{Tr}(\mathbf{HK_A HK_B}),\tag{1}$$

where $\mathbf{K_A}\in\mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ and $\mathbf{K_B}\in\mathbb{R}^{\mathrm{n}\times\mathrm{n}}$ are kernel matrices and $\mathbf{H}=\mathbf{I}_{\mathrm{n}\times\mathrm{n}}-\frac{1}{\mathrm{n}}\mathbf{1}_\mathrm{n}\mathbf{1}_\mathrm{n}^\mathrm{T}$ is the centralized matrix. In theory, the larger the HSIC, the larger the dependence between $\mathbf{A}$ and $\mathbf{B}$.

Majorize-Maximization (MM) framework. For a general problem $\max_\mathrm{x} f(\mathrm{x})$, MM algorithm [Sun *et al.*, 2016; Nie *et al.*, 2020; Ham *et al.*, 2018] aims at seeking a surrogate problem $\max_\mathrm{x}\varphi(\mathrm{x})$ to help optimize the raw problem iteratively. Suppose x_0 is a current solution, $\max_\mathrm{x}\varphi(\mathrm{x}|\mathrm{x}_0)$ should satisfy the following conditions:

$$\forall x,\ f(\mathrm{x})\geqslant\varphi(\mathrm{x}|\mathrm{x}_0),\ f(\mathrm{x}_0)=\varphi(\mathrm{x}_0|\mathrm{x}_0).\tag{2}$$

Obviously, when $\tilde{\mathrm{x}}=\arg\max\varphi(\mathrm{x})$, we have

$$f(\tilde{\mathrm{x}})\geqslant\varphi(\tilde{\mathrm{x}}|\mathrm{x}_0)\geqslant\varphi(\mathrm{x}_0|\mathrm{x}_0)=f(\mathrm{x}_0),\tag{3}$$

which guarantees the ascent property.

3 Methodology

3.1 Proposed GSPL Model

Given a multi-view data $\boldsymbol{\mathcal{X}} = \{\mathbf{X}_1, \ldots, \mathbf{X}_s\}$, wherein $\mathbf{X}_v \in \mathbb{R}^{n \times d_v}$ is the feature matrix of v-th view, n and d_v are the numbers of samples and dimensions of $\mathbf{X}_v$. Besides, we denote the v-th spectral embedding as $\mathbf{U}_v \in \mathbb{R}^{n \times c}$ generated via Laplacian eigenmap [Belkin and Niyogi, 2001], *i.e.*, $\min_{\mathbf{U}_v^T\mathbf{U}_v = \mathbf{I}_{c \times c}} \mathbf{Tr}(\mathbf{U}_v^T \mathbf{L}_\mathbf{S} \mathbf{U}_v)$, on the similarity matrix $\mathbf{S}_v \in \mathbb{R}^{n \times n}$ constructed from $\mathbf{X}_v$, where c is the number of clusters and $\mathbf{L}_\mathbf{S} \in \mathbb{R}^{n \times n}$ is the Laplacian matrix of $\mathbf{S}$. Now we start the derivation of proposed GSPL model as Fig. 1 shows. At first, we map $\mathbf{X}_v \in \mathbb{R}^{n \times d_v}$ as $\mathbf{X}_v\mathbf{W}_v \in \mathbb{R}^{n \times m}$, where m is the reduced dimensionality and $\mathbf{W}_v \in \mathbb{R}^{d_v \times m}$ is the projection matrix. Then to integrate the information among multiple views, we explore a center $\sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v$ of multiple $\mathbf{X}_v\mathbf{W}_v \in \mathbb{R}^{n \times m}$, where p_v is the flip and scale operator variable of v-th view. Afterwards, we map $\sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v$ and multiview spectral embeddings $\{\mathbf{U}_g\}_{g=1}^s$ into Reproducing Kernel Hilbert Spaces (RKHSs), and enclose them with HSIC measure as follows:

$$\max_{\mathbf{p},\mathbf{z},\{\mathbf{W}_v\}_{v=1}^s} \sum_{g=1}^s z_g \cdot \mathbf{Tr}\left(\mathbf{H}\mathbf{K}_{\bar{\mathbf{X}}}\mathbf{H}\mathbf{K}_{\mathbf{U}_g}\right)$$
$$\text{s.t. } \mathbf{W}_v \in \mathbb{M}_v,\ \bar{\mathbf{X}} = \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v,$$
$$\|\mathbf{p}\|_2 = 1,\ \|\mathbf{z}\|_2 = 1,\ \mathbf{z} \geqslant \mathbf{0}, \tag{4}$$

where $\mathbb{M}_v$ is the sparse constraint of $\mathbf{W}_v$, $\mathbf{K}_{\bar{\mathbf{X}}}$ and $\mathbf{K}_{\mathbf{U}_g}$ are kernel matrices, z_g is the weight for the HSIC measure between $\bar{\mathbf{X}} = \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v$ and $\mathbf{U}_g$, and $\|\mathbf{z}\|_2 = 1$, $\mathbf{z} \geqslant 0$ is the normalized constraint. In this paper, we set the kernel as the inner product, *i.e.*, $\mathbf{K}_{\bar{\mathbf{X}}} = \bar{\mathbf{X}}\bar{\mathbf{X}}^T = (\sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v)(\sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v)^T$ and $\mathbf{K}_{\mathbf{U}_g} = \mathbf{U}_g \mathbf{U}_g^T$, then problem (4) can be written as

$$\max_{\mathbf{p},\mathbf{z},\{\mathbf{W}_v\}_{v=1}^s} \sum_{g=1}^s z_g \cdot \|\mathbf{U}_g^T \mathbf{H} \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v\|_F^2$$
$$\text{s.t. } \mathbf{W}_v \in \mathbb{M}_v,\ \|\mathbf{p}\|_2 = 1,\ \|\mathbf{z}\|_2 = 1,\ \mathbf{z} \geqslant \mathbf{0}. \tag{5}$$

In problem (5), $\mathbf{p}$ considers the flip and scale transformation when integrating multiple $\mathbf{X}_v\mathbf{W}_v$, and $\mathbf{z}$ considers the differences when approximating $\sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v$ to multiple spectral embeddings. These two variables capture the differences among multiple views comprehensively, which handles the ***IC*** problem effectively. Now we consider designing the sparse constraint $\mathbf{W}_v \in \mathbb{M}_v$. At first, we concentrate all projection matrices into $\widehat{\mathbf{W}} = [\mathbf{W}_1; \ldots; \mathbf{W}_s]^T \in \mathbb{R}^{\hat{d} \times m}$, where $\hat{d} = \sum_{v=1}^s d_v$, then constrain $\widehat{\mathbf{W}}$ via $\ell_{2,0}$-norm for group-sparsity. By this way, problem (5) can be improved as

$$\max_{\mathbf{p},\mathbf{z},\{\mathbf{W}_v\}_{v=1}^s} \sum_{g=1}^s z_g \cdot \|\mathbf{U}_g^T \mathbf{H} \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v\|_F^2$$
$$s.t.\ \widehat{\mathbf{W}} = [\mathbf{W}_1; \ldots; \mathbf{W}_s]^T,\ \widehat{\mathbf{W}}^T\widehat{\mathbf{W}} = \mathbf{I}_{m \times m},$$
$$\|\widehat{\mathbf{W}}\|_{2,0} = k,\ \|\mathbf{p}\|_2 = 1,\ \|\mathbf{z}\|_2 = 1,\ \mathbf{z} \geqslant \mathbf{0}, \tag{6}$$

where [1]$k \geqslant m$ is the number of selected features. In problem (6), according to $\|\widehat{\mathbf{W}}\|_{2,0} = k$, the group-sparsity of each $\mathbf{W}_v$ is also guaranteed and the number of 0 rows is determined adaptively from the perspective of the whole $\widehat{\mathbf{W}}$, which handles the ***SF*** and ***GS*** problems effectively. So far, the final version of GSPL model is obtained as problem (6), which handles the ***IC***, ***SF*** and ***GS*** problems effectively. More importantly, it is obvious to observe that except for m and k brought by multiview feature selection task, the proposed GSPL model does not bring any extra hyper-parameter, then the ***SC*** problem is also handled thoroughly.

3.2 Optimization for GSPL Model

Since problem (6) contains three variables, including $\mathbf{p}$, $\mathbf{z}$ and $\mathbf{W}$, we consider optimizing them alternatively.

Update p and fix others. When $\mathbf{W}$ and $\mathbf{z}$ are fixed, problem (6) becomes

$$\max_{\|\mathbf{p}\|_2=1} \mathbf{Tr}\left(\widehat{\mathbf{U}}(\sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v)(\sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v)^T\right), \tag{7}$$

where $\widehat{\mathbf{U}} = \sum_{g=1}^s z_g \mathbf{H}^T \mathbf{U}_g \mathbf{U}_g^T \mathbf{H}$. Let $\mathbf{A}_v = \widehat{\mathbf{U}}\mathbf{X}_v\mathbf{W}_v$, $\mathbf{B}_v = \mathbf{X}_v\mathbf{W}_v$, then problem (7) becomes

$$\max_{\|\mathbf{p}\|_2=1} \mathbf{Tr}\left((\sum_{v=1}^s p_v \mathbf{A}_v)(\sum_{v=1}^s p_v \mathbf{B}_v)^T\right). \tag{8}$$

For convenience, we denote $\widehat{\mathbf{A}} = [\mathbf{Vec}(\mathbf{A}_1), \ldots, \mathbf{Vec}(\mathbf{A}_s)]$, $\widehat{\mathbf{B}} = [\mathbf{Vec}(\mathbf{B}_1), \ldots, \mathbf{Vec}(\mathbf{B}_s)] \in \mathbb{R}^{nm \times s}$, where $\mathbf{Vec}(\cdot)$ is the column-based matrix vectorization operator. Then problem (8) can be rewritten as the following form:

$$\max_{\|\mathbf{p}\|_2=1} \mathbf{Tr}\left(\mathbf{p}^T \widehat{\mathbf{B}}^T \widehat{\mathbf{A}} \mathbf{p}\right). \tag{9}$$

The solution $\mathbf{p}$ of problem (9) can be formed as the eigenvectors corresponding to the s largest eigenvalues of $\widehat{\mathbf{B}}^T\widehat{\mathbf{A}}$.

Update z and fix others. When $\mathbf{W}$ and $\mathbf{p}$ are fixed, problem (6) becomes

$$\max_{\|\mathbf{z}\|_2=1, \mathbf{z} \geqslant 0} \sum_{g=1}^s z_g \cdot \|\mathbf{U}_g^T \mathbf{H} \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v\|_F^2. \tag{10}$$

Considering the term $\|\mathbf{U}_g^T \mathbf{H} \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v\|_F^2 \geq 0$, we have the following derivations according to Cauchy-Schwarz inequality [Steele, 2004]:

$$\sum_{g=1}^s z_g \cdot \|\mathbf{U}_g^T \mathbf{H} \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v\|_F^2$$
$$\overset{(a)}{\leq} \sqrt{\left(\sum_{g=1}^s \|\mathbf{U}_g^T \mathbf{H} \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v\|_F^4\right)\left(\sum_{g=1}^s z_g^2\right)}$$
$$= \sqrt{\sum_{g=1}^s \|\mathbf{U}_g^T \mathbf{H} \sum_{v=1}^s p_v \mathbf{X}_v \mathbf{W}_v\|_F^4}. \tag{11}$$

[1] k and m are two intrinsic parameters brought by projection-based feature selection task. In our model, we set $m \leqslant k$ for the correction of the $\ell_{2,0}$-norm optimization.

Considering $\|\mathbf{z}\|_2 = 1$, the equality in (a) holds when

$$\mathrm{z_g} = \frac{\|\mathbf{U}_\mathrm{g}^\mathrm{T}\mathbf{H}\sum_{\mathrm{v}=1}^{\mathrm{s}} \mathrm{p_v}\mathbf{X}_\mathrm{v}\mathbf{W}_\mathrm{v}\|_\mathrm{F}^2}{\sqrt{\sum_{\mathrm{g}=1}^{\mathrm{s}}\|\mathbf{U}_\mathrm{g}^\mathrm{T}\mathbf{H}\sum_{\mathrm{v}=1}^{\mathrm{s}} \mathrm{p_v}\mathbf{X}_\mathrm{v}\mathbf{W}_\mathrm{v}\|_\mathrm{F}^4}}, \tag{12}$$

which is the closet solution of problem (10).

Update $\{\mathbf{W}_\mathrm{v}\}_{\mathrm{v}=1}^{\mathrm{s}}$ and fix others. When $\mathbf{p}$ and $\mathbf{q}$ are fixed, problem (6) can be written as

$$\max_{\widehat{\mathbf{W}}^T\widehat{\mathbf{W}}=\mathbf{I}_{\mathrm{m}\times\mathrm{m}},\|\widehat{\mathbf{W}}\|_{2,0}=\mathrm{k}} \sum_{\mathrm{g}=1}^{\mathrm{s}} \mathrm{z_g}\cdot\|\mathbf{U}_\mathrm{g}^\mathrm{T}\mathbf{H}\widehat{\mathbf{X}}\tilde{\mathbf{P}}\widehat{\mathbf{W}}\|_\mathrm{F}^2, \tag{13}$$

where $\widehat{\mathbf{X}} = [\mathbf{X}_1, \ldots \mathbf{X}_\mathrm{s}] \in \mathbb{R}^{\mathrm{n}\times\hat{\mathrm{d}}}$, $\tilde{\mathbf{P}} = \mathbf{diag}(\tilde{\mathbf{p}}) \in \mathbb{R}^{\hat{\mathrm{d}}\times\hat{\mathrm{d}}}$ is a diagonal matrix and $\tilde{\mathbf{p}} = [\mathrm{p}_1\mathbf{1}_{\mathrm{d}_1}; \ldots; \mathrm{p_s}\mathbf{1}_{\mathrm{d_s}}] \in \mathbb{R}^{\hat{\mathrm{d}}}$ is the diagonal vector of it, $\mathbf{diag}(\cdot)$ is the vector-diagonal operator. Denote $\mathbf{E}_\mathrm{g} = \mathbf{U}_\mathrm{g}^\mathrm{T}\mathbf{H}\widehat{\mathbf{X}}\tilde{\mathbf{P}}$, problem (13) can be written as

$$\max_{\widehat{\mathbf{W}}^\mathrm{T}\widehat{\mathbf{W}}=\mathbf{I}_{\mathrm{m}\times\mathrm{m}},\ \|\widehat{\mathbf{W}}\|_{2,0}=\mathrm{k}} \sum_{\mathrm{g}=1}^{\mathrm{s}} \mathrm{z_g}\cdot\mathbf{Tr}\left(\widehat{\mathbf{W}}^\mathrm{T}\mathbf{E}_\mathrm{g}^\mathrm{T}\mathbf{E}_\mathrm{g}\widehat{\mathbf{W}}\right). \tag{14}$$

Further, we denote $\mathbf{S_E} = \sum_{\mathrm{g}=1}^{\mathrm{s}} \mathrm{z_g}\mathbf{E}_\mathrm{g}^\mathrm{T}\mathbf{E}_\mathrm{g} + \lambda\mathbf{I}_{\mathrm{m}\times\mathrm{m}}$, where λ guarantees that $\mathbf{S_E}$ is positive semi-definite, then problem (14) can be written as

$$\max_{\widehat{\mathbf{W}}^\mathrm{T}\widehat{\mathbf{W}}=\mathbf{I}_{\mathrm{m}\times\mathrm{m}},\ \|\widehat{\mathbf{W}}\|_{2,0}=\mathrm{k}} \mathbf{Tr}\left(\widehat{\mathbf{W}}^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}\right), \tag{15}$$

which is NP-hard, then we consider to solve it into two cases. At first, we consider the case $\mathbf{rank}(\mathbf{S_E}) \leqslant \mathrm{m}$. Since $\|\widehat{\mathbf{W}}\|_{2,0} = \mathrm{k}$, suppose $^2\mathbf{q} \in \mathbf{Ind}(\mathrm{k}, \hat{\mathrm{d}})$ is the $\mathbb{R}^\mathrm{k}$ indicator vector of non-sparse rows of $\widehat{\mathbf{W}}$, then $\widehat{\mathbf{W}}$ can be decomposed into $\widehat{\mathbf{W}} = \mathbf{BD}$. Wherein, $\mathbf{B} = \mathbf{\Pi}(\mathbf{q}) \in \{\mathbf{0}, \mathbf{1}\}_{\hat{\mathrm{d}}\times\mathrm{k}}$, whose $\langle\mathrm{u}, \mathrm{v}\rangle$-th element $\mathrm{b_{uv}} = 1$ only if $\mathrm{u} = \mathrm{q_v}$, and $\mathbf{\Pi}(\cdot)$ is the mapping function $\mathbf{Ind}(\mathrm{k}, \hat{\mathrm{d}}) \to \{\mathbf{0}, \mathbf{1}\}_{\hat{\mathrm{d}}\times\mathrm{k}}$; $\mathbf{D} \in \mathbb{R}^{\mathrm{k}\times\mathrm{m}}$ and the u-th row of $\mathbf{D}$ is the $\mathrm{q_u}$-th row of $\widehat{\mathbf{W}}$, and naturally $\mathbf{D}^\mathrm{T}\mathbf{D} = \mathbf{I}_{\mathrm{m}\times\mathrm{m}}$. Then problem (15) can be written as the following problem $w.r.t.$ $\mathbf{B}$ and $\mathbf{D}$:

$$\max_{\mathbf{B}=\mathbf{\Pi}(\mathbf{q}),\mathbf{q}\in\mathbf{Ind}(\mathrm{k},\hat{\mathrm{d}}),\mathbf{D}^\mathrm{T}\mathbf{D}=\mathbf{I}_{\mathrm{m}\times\mathrm{m}}} \mathbf{Tr}\left(\mathbf{D}^\mathrm{T}\mathbf{B}^\mathrm{T}\mathbf{S_E}\mathbf{BD}\right), \tag{16}$$

where the optimal $\mathbf{B} = \mathbf{\Pi}(\tilde{\mathbf{q}})$, where $\tilde{\mathbf{q}} \in \mathbf{Ind}(\mathrm{k}, \hat{\mathrm{d}})$ is the indicate vector of the first k largest values of the diagonal vector of $\mathbf{S_E}$. Then in problem (16), the solution $\mathbf{D}$ can be formed by the eigenvectors corresponding to first-m largest eigenvalues of $\mathbf{B}^\mathrm{T}\mathbf{S_E}\mathbf{B}$. Finally, $\widehat{\mathbf{W}}$ can be calculated as $\mathbf{BD}$. More details about this solution can be referred to supplementary material. Then we consider the case $\mathbf{rank}(\mathbf{S_E}) > \mathrm{m}$. In this case, we consider utilizing the famous Majorize-Minimization (MM) framework. Suppose the current solution is $\widehat{\mathbf{W}}_0$, via observing that $\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0 = \widehat{\mathbf{W}}_0^\mathrm{T}\left(\mathbf{S_E}\widehat{\mathbf{W}}_0(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)^\dagger\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\right)\widehat{\mathbf{W}}_0$, where $(\cdot)^\dagger$ is the Moore-Penrose inverse operator, and considering the conditions of MM framework, we guess the following surrogate problem of problem (15) based on MM framework:

Algorithm 1: The Algorithm for Solving problem (6)

```
Input: $\mathcal{X} \in \mathbb{R}^{\mathrm{n}\times\hat{\mathrm{d}}}$, $\{\mathbf{U}_\mathrm{g}\}_{\mathrm{g}=1}^{\mathrm{s}}$, $\hat{\mathrm{d}}$, k, m
Initialization: $\mathbf{p}$, $\mathbf{q}$, $\widehat{\mathbf{W}} = [\mathbf{W}_1; \ldots; \mathbf{W}_\mathrm{s}]^\mathrm{T}$.
while not converge do
    Update $\mathbf{p}$ via solving problem (9).
    Update $\mathbf{z}$ via Eq. (12).
    if $\mathbf{rank}(\mathbf{S_E}) \leqslant \mathrm{m}$ then
        Update $\mathbf{B}$ and $\mathbf{D}$ via solving problem (16).
        Update $\widehat{\mathbf{W}}$ via $\widehat{\mathbf{W}} = \mathbf{BD}$.
    else
        while not converge do
            Set $\widehat{\mathbf{W}}_0$ by current $\widehat{\mathbf{W}}$.
            Update $\widehat{\mathbf{W}}$ via solving problem (17).
Output: $\mathbf{p}$, $\mathbf{z}$, $\{\mathbf{W}_1, \ldots, \widehat{\mathbf{W}}_\mathrm{s}\}$.
```

$$\max_{\widehat{\mathbf{W}}} \mathbf{Tr}\left(\widehat{\mathbf{W}}^\mathrm{T}\left(\mathbf{S_E}\widehat{\mathbf{W}}_0(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)^\dagger\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\right)\widehat{\mathbf{W}}\right)$$
$$\text{s.t. } \widehat{\mathbf{W}}^T\widehat{\mathbf{W}} = \mathbf{I}_{\mathrm{m}\times\mathrm{m}}, \|\widehat{\mathbf{W}}\|_{2,0} = \mathrm{k}, \tag{17}$$

with the vital property as the following remark:

Remark 1. *Suppose that the objective functions of problem* (15) *and problem* (17) *are* $\mathcal{J}_R(\widehat{\mathbf{W}})$ *and* $\mathcal{J}_S(\widehat{\mathbf{W}})$ *respectively, at any point* $\widehat{\mathbf{W}}^T\widehat{\mathbf{W}} = \mathbf{I}_{\mathrm{m}\times\mathrm{m}}, \|\widehat{\mathbf{W}}\|_{2,0} = \mathrm{k}$, *we have* $\mathcal{J}_R(\widehat{\mathbf{W}}) \geqslant \mathcal{J}_S(\widehat{\mathbf{W}})$ *and* $\mathcal{J}_R(\widehat{\mathbf{W}}_0) = \mathcal{J}_S(\widehat{\mathbf{W}}_0)$.

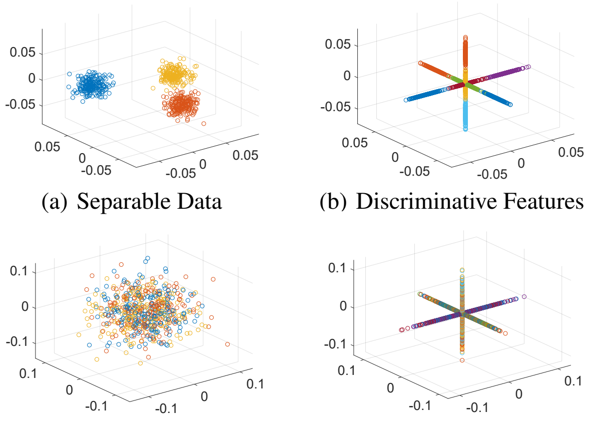
Then we can conclude that problem (17) can be a surrogate problem for problem (15). The proof of Remark 1 is provided into supplementary material. Afterwards, we focus on the surrogate problem, $i.e.$, problem (17), and propose the following two observations: ***1)*** Since $\mathbf{rank}(\widehat{\mathbf{W}}_0) \leqslant \mathrm{m}$, we have $\mathbf{rank}\left(\mathbf{S_E}\widehat{\mathbf{W}}_0(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)^\dagger\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\right) \leqslant \mathrm{m}$. ***2)*** Considering $\mathbf{S_E}$ is positive semi-definite so that $\mathbf{S_E} = \mathbf{GG}^\mathrm{T}$, and the fact that $\mathbf{A}^\dagger = \mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger$, we have

$$\mathbf{S_E}\widehat{\mathbf{W}}_0(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)^\dagger\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E} = \mathbf{S_E}\widehat{\mathbf{W}}_0(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)^\dagger$$
$$(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)^\dagger\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E} = \widehat{\mathbf{G}}\widehat{\mathbf{G}}^\mathrm{T}. \tag{18}$$

where $\widehat{\mathbf{G}} = \mathbf{S_E}\widehat{\mathbf{W}}_0(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)^\dagger\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{G}$, then we have $\mathbf{S_E}\widehat{\mathbf{W}}_0(\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}\widehat{\mathbf{W}}_0)^\dagger\widehat{\mathbf{W}}_0^\mathrm{T}\mathbf{S_E}$ is positive semi-definite. Then problem (17) can be solved as the same way of problem (15) in the case of $\mathbf{rank}(\mathbf{S_E}) \leqslant \mathrm{m}$. For simplicity of expression, we do not provide the detailed optimization procedures. The algorithm to solve problem (6) is summarized in Algorithm 1.

Convergence guarantee. Algorithm 1 optimizes $\mathbf{p}$, $\mathbf{z}$ and $\widehat{\mathbf{W}}$ via solving the subproblems iteratively, wherein $\mathbf{p}$-subproblem and $\mathbf{z}$-problem contain closed solutions. For $\widehat{\mathbf{W}}$, when $\mathbf{rank}(\mathbf{S_E}) \leqslant \mathrm{m}$, $\widehat{\mathbf{W}}$ has closed solution; when $\mathbf{rank}(\mathbf{S_E}) > \mathrm{m}$, the convergence of the subproblem (15) is guaranteed by MM framework. Therefore, the objective value

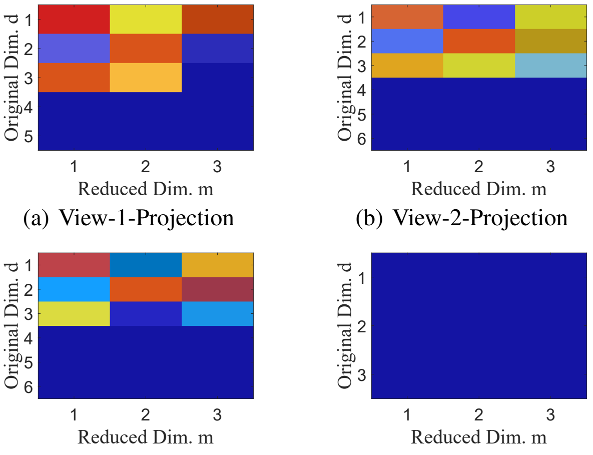
[2] $\mathbf{Ind}(\mathrm{k}, \hat{\mathrm{d}})$ is a $\mathbb{R}^\mathrm{k}$ indicator vector, which selectes k elements from $\{1, \ldots, \hat{\mathrm{d}}\}$ as ascending order and the elements are not duplicated. For example, $\mathbf{Ind}(2, 3)$ can be $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$.



(a) Separable Data (b) Discriminative Features

(c) Noise Data (d) Noisy Features

Figure 2: Illustration of the discriminative and noisy features.



(a) View-1-Projection (b) View-2-Projection

(c) View-3-Projection (d) View-4-Projection

Figure 3: Visualization of the learnt projection matrices on synthetic dataset. Dark blue means the element is 0.

of problem (6) will increase when optimizing any one of the subproblems, which means the objective value of problem (6) will monotonically increase per iteration until convergence.

Computational cost. Update $\mathbf{p}$ and $\mathbf{z}$: $\mathcal{O}(n^2(m+c)s + nkms)$; Update $\widehat{\mathbf{W}}$: $\mathcal{O}((n^2 + n\widehat{d} + \widehat{d}^2)cs + (k^3 + \widehat{d}km)t_1)$; The whole computational cost can be calculated by: $\mathcal{O}(((n^2c + \widehat{d}^2c + n^2m + nkm)s + (\widehat{d}km + k^3)t_1)t_2)$ where t_1, t_2 are the iterations of the inner and outer loops separately.

4 Experiments

Evaluations on synthetic dataset. This part aims to answer the following questions intuitively: *Can GSPL model select the discriminative features on the multiview data with noise features and noise views?* To this end, we carefully design a synthetic dataset which contains 4 views, each of which has 600 samples with 3 clusters. As Table 1 shows, in view 1-3, the 1-3 features are discriminative and the other features are noisy. View 4 is chaotic in which all the features are noisy. Fig. 2 intuitively shows the discriminative and noisy features, where (b) and (d) are respectively the projections of the data in (a) and (c). In Fig. 2-(a) and -(b), all the three features are discriminative, and in Fig. 2-(c) and -(d), all the three features are not discriminative. On the designed synthetic dataset, we run GSPL algorithm to select 9 features from 20 features via setting m as 3, and the visualization of the the learnt projections $\{\mathbf{W}_v\}_{v=1}^s$ are plotted into Fig. 3. From the results, we can find that the non-sparse rows of each $\mathbf{W}_v$ correspond to the discriminative features of this view. Particularly in chaotic view 4, GSPL ignores all the features, which means GSPL can resist noisy views. Besides, we record the learnt $\mathbf{z}$, *i.e.*, $z_1 = \mathbf{0.5065}$, $z_2 = \mathbf{0.6165}$, $z_3 = \mathbf{0.6041}$, $z_4 = \mathbf{0.0013}$, which satisfies the designed criterion that views 1-3 are normal and view 4 is chaotic. Thus we can conclude that the variable $\mathbf{z}$ has the ability to resist noisy views.

Features	View-1	View-2	View-3	View-4
Discriminative	1-3	1-3	1-3	—
Noise	4-5	4-6	4-6	1-3

Table 1: The descriptions of designed synthetic datasets.

Datasets	ORL	Caltech101-7	Reuters	COIL20
View-1	GIST(512)	Gabor(48)	French(2000)	INTE(1024)
View-2	LBP(59)	WM(40)	German(2000)	LBP(3304)
View-3	GIST(512)	CENT(254)	Spanish(2000)	GIST(6750)
View-4	HOG(864)	HOG(1984)	Italian(2000)	—
View-5	CENT(254)	GIST(512)	English(2000)	—
View-6	—	LBP(928)	—	—
Total Dim	1689	3766	10000	11078
Class	40	7	6	20
Sample	400	1474	1200	1440
Type	Image	Image	Text	Image

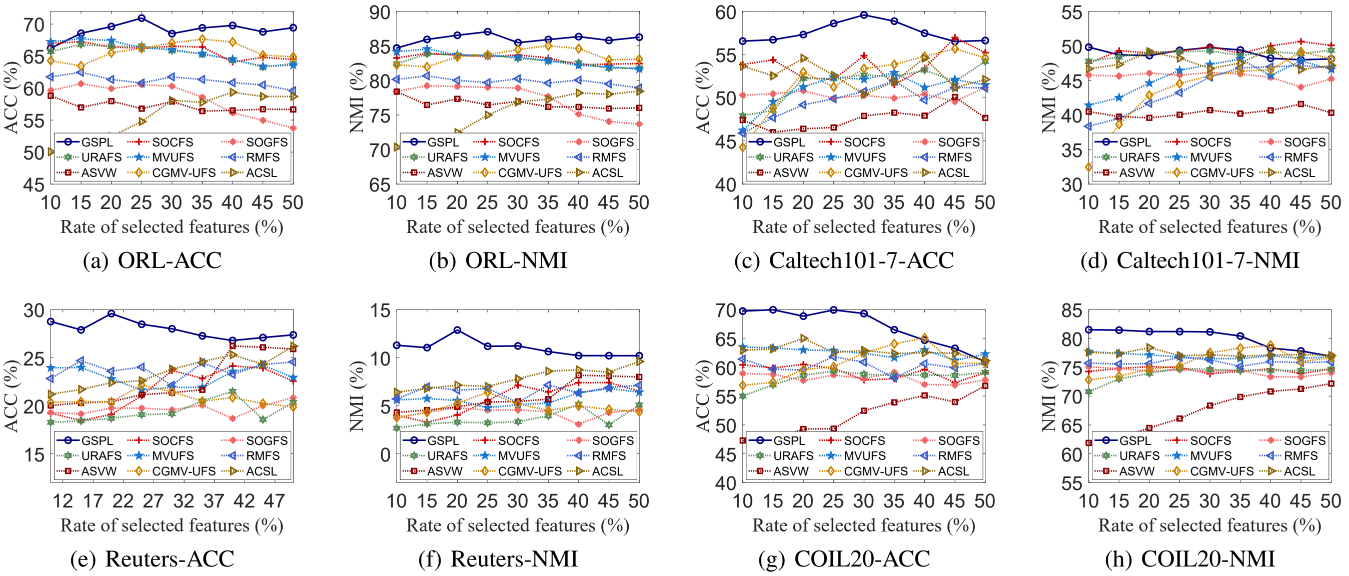
Table 2: The descriptions of real datasets.

Evaluations on real datasets. In this part, we evaluate the clustering performance of proposed method GSPL on 4 real multi-view datasets, including **ORL**[3], **Caltech101-7** [Li *et al.*, 2004], **Reuters**[4], **COIL20**[5]. The details are available in Table 2. Moreover, We employ 8 SOTA models as competitors, including 3 series models (*i.e.*, **SOCFS** [Han and Kim, 2015], **SOGFS** [Nie *et al.*, 2016], **URAFS** [Li *et al.*, 2018]), and 5 parallel models (*i.e.*, **MVUFS** [Qian and Zhai, 2014], **RMFS** [Liu *et al.*, 2016], **ASVW** [Hou *et al.*, 2017], **CGMV-UFS** [Tang *et al.*, 2018], **ACSL** [Dong *et al.*, 2019]). We evaluate the clustering performance via K-means on the selected features of original data, and the performance is measured by ACCuracy (ACC) and Normalized Mutual Information (NMI) [Strehl and Ghosh, 2002]. For fair comparison, all the parameters are set according to the suggestions in original articles, and the total reduced dimensionality m in our GSPL is empirically selected around $2k/3$ to k, and the number of selected features is set to $\widehat{d} * r$, where r is the rate of the selected features, traversed from 0.1 to 0.5 with 0.05 as interval. To alleviate the random effect caused by K-means, we run it for 30 times and report the mean in Fig. 4.

[3] http://www.uk.research.att.com/facedatabase.html

[4] https://archive.ics.uci.edu/ml/datasets.html

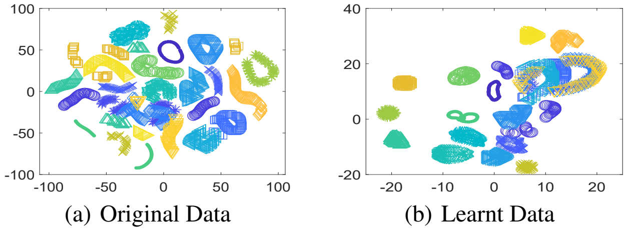
[5] http://www.cs.columbia.edu/CAVE/software/softlib/



(a) ORL-ACC (b) ORL-NMI (c) Caltech101-7-ACC (d) Caltech101-7-NMI

(e) Reuters-ACC (f) Reuters-NMI (g) COIL20-ACC (h) COIL20-NMI

Figure 4: Comparison of clustering performance (ACC and NMI) on 4 datasets with different rates of selected features.



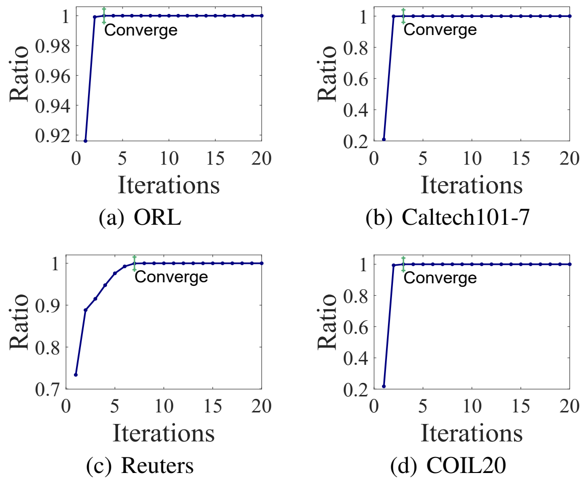
(a) Original Data (b) Learnt Data

Figure 5: T-SNE visualization on COIL20 dataset.

Performance evaluation. In Fig. 4: 1) GSPL achieves over 4.84%, 4.58%, 1.28% and 7.51% higher ACC than the second highest results on the 4 datasets, and obtains around 2.39%, 3.41% and 3.40% larger NMI than the best results of competitors on ORL, Reuters and COIL20 datasets; 2) GSPL outperforms others when the rate of the selected features is small ($i.e.$ 10% or 15%) in all the cases, which indicates that GSPL can achieve better performance than others with fewer selected features and select more discriminative features.

TSNE visualization. In this part, we visualize the original concentrated feature matrix $\widehat{\mathbf{X}} = [\mathbf{X}_1, \dots \mathbf{X}_{\mathrm{s}}]$ and learnt $\widetilde{\mathbf{X}} \in \mathbb{R}^{\mathrm{n}\times k}$ via T-SNE on them to obtain the low-dimensional representations of them. The results are plotted in Fig. 5. It is obvious that via group-sparse projection and adaptive fusion, the separability of data has superb improvement.

Convergence analysis. In this part, we empirically show the convergence behavior of Algorithm 1. For all the datasets, we set k and m to $0.5 * \widehat{\mathrm{d}}$ and $2\mathrm{k}/3$, respectively. The convergent curves are plotted in Fig. 6, from which we can see that Algorithm 1 can converge rapidly (within 10 iterations).



(a) ORL (b) Caltech101-7

(c) Reuters (d) COIL20

Figure 6: The convergent curves of GSPL model. Ratio is the proportion of each objective value in maximum objective value.

5 Conclusion

This paper proposed a GSPL model for multi-view feature selection. The three properties of our model, including Strictness, Adaptivity and Succinctness, effectively handle the four key problems of multiview feature selection task, including ***IC***, ***SF***, ***GS***, ***SC***. Moreover, an efficient algorithm with rigorous convergence guarantee is innovatively proposed to optimize GSPL model. Sufficient experimental results verify the promising performance of the proposed method.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0101902, in part by the Natural Science Basic Research Program of Shaanxi (Program No. 2021JM-071), in part by the National Natural Science Foundation of China under Grant 61936014 and Grant 61772427, and in part by the Fundamental Research Funds for the Central Universities under Grant G2019KY0501.

References

[Belkin and Niyogi, 2001] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, pages 585–591, 2001.

[Dong *et al.*, 2019] Xiao Dong, Lei Zhu, Xuemeng Song, Jingjing Li, and Zhiyong Cheng. Adaptive collaborative similarity learning for unsupervised multi-view feature selection. *arXiv preprint arXiv:1904.11228*, 2019.

[Gretton *et al.*, 2005] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

[Ham *et al.*, 2018] Bumsub Ham, Minsu Cho, and Jean Ponce. Robust guided image filtering using nonconvex potentials. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):192–207, 2018.

[Han and Kim, 2015] Dongyoon Han and Junmo Kim. Unsupervised simultaneous orthogonal basis clustering feature selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5016–5023. IEEE, 2015.

[Hou *et al.*, 2017] Chenping Hou, Feiping Nie, Hong Tao, and Dongyun Yi. Multi-view unsupervised feature selection with adaptive similarity and view weight. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1998–2011, 2017.

[Li *et al.*, 2004] Fei-Fei Li, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE, 2004.

[Li *et al.*, 2018] Xuelong Li, Han Zhang, Rui Zhang, Yun Liu, and Feiping Nie. Generalized uncorrelated regression with adaptive graph for unsupervised feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1587–1595, 2018.

[Liu *et al.*, 2016] Hongfu Liu, Haiyi Mao, and Yun Fu. Robust multi-view feature selection. In *Proceedings of the 16th IEEE International Conference on Data Mining*, pages 281–290. IEEE, 2016.

[Nie *et al.*, 2016] Feiping Nie, Wei Zhu, and Xuelong Li. Unsupervised feature selection with structured graph optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1302–1308, 2016.

[Nie *et al.*, 2018] Feiping Nie, Lai Tian, and Xuelong Li. Multiview clustering via adaptively weighted procrustes. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2022–2030, 2018.

[Nie *et al.*, 2020] Feiping Nie, Danyang Wu, Rong Wang, and Xuelong Li. Truncated robust principle component analysis with a general optimization framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Qian and Zhai, 2014] Mingjie Qian and Chengxiang Zhai. Unsupervised feature selection for multi-view clustering on text-image web news data. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pages 1963–1966, 2014.

[Steele, 2004] John Michael Steele. *The Cauchy-Schwarz master class: An introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

[Strehl and Ghosh, 2002] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(Dec):583–617, 2002.

[Sun *et al.*, 2016] Ying Sun, Prabhu Babu, and Daniel P Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2016.

[Tang *et al.*, 2018] Chang Tang, Jiajia Chen, Xinwang Liu, Miaomiao Li, Pichao Wang, Minhui Wang, and Peng Lu. Consensus learning guided multi-view unsupervised feature selection. *Knowledge-Based Systems*, 160:49–60, 2018.

[Wu *et al.*, 2020] Danyang Wu, Feiping Nie, Rong Wang, and Xuelong Li. Multi-view clustering via mixed embedding approximation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3977–3981. IEEE, 2020.

[Zhang *et al.*, 2016] Zhenyue Zhang, Zheng Zhai, and Limin Li. Uniform projection for multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1675–1689, 2016.

[Zhang *et al.*, 2018] Changqing Zhang, Yeqing Liu, Yue Liu, Qinghua Hu, Xinwang Liu, and Pengfei Zhu. Fish-mml: Fisher-hsic multi-view metric learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3054–3060, 2018.