Robust Subspace Clustering with Independent and Piecewise Identically Distributed Noise Modeling

Yuanman Li, Jiantao Zhou, Xianwei Zheng, Jinyu Tian, Yuan Yan Tang
Department of Computer and Information Science, University of Macau, Macau, China
{yuanmanli, jtzhou, yb47430, yb77405, yytang}@um.edu.mo

Abstract

Most of the existing subspace clustering (SC) frameworks assume that the noise contaminating the data is generated by an independent and identically distributed (i.i.d.) source, where the Gaussianity is often imposed. Though these assumptions greatly simplify the underlying problems, they do not hold in many real-world applications. For instance, in face clustering, the noise is usually caused by random occlusions, local variations and unconstrained illuminations, which is essentially structural and hence satisfies neither the i.i.d. property nor the Gaussianity. In this work, we propose an independent and piecewise identically distributed (i.p.i.d.) noise model, where the i.i.d. property only holds locally. We demonstrate that the i.p.i.d. model better characterizes the noise encountered in practical scenarios, and accommodates the traditional i.i.d. model as a special case. Assisted by this generalized noise model, we design an information theoretic learning (ITL) framework for robust SC through a novel minimum weighted error entropy (MWEE) criterion. Extensive experimental results show that our proposed SC scheme significantly outperforms the state-of-the-art competing algorithms.

1. Introduction

Many practical high-dimensional data usually lie in a low-dimensional structure, rather than being uniformly distributed over the ambient space [30, 36, 3, 5, 44]. Some representative examples include feature trajectories of rigidly moving objects in a video [30], face images of one subject [36], and the spectra of one instance in a hyperspectral image [3]. As a result, a collection of data from multiple categories can be regarded as the ones lying in a *union* of low-dimensional subspaces [5]. Subspace clustering (SC) refers to the problem of separating the data points according to their underlying subspaces, and has found numerous applications in motion segmentation [22, 43], image clustering [20, 21], data representation [14], etc.

There are many different types of SC approach proposed, e.g., the algebraic [17], the statistical [10], the iterative [41], and the spectral clustering based [5, 20] algorithms. In this work, we focus on the SC approaches based on the spectral clustering [5, 20], due to their state-of-the-art performance provided. Within the framework of spectral clustering based methods, an affinity matrix indicating the similarity between pairs of the data points is first built, and then the data points are separated by applying the spectral clustering [25] on this affinity matrix. The primary difference of various spectral clustering-based algorithms lies in how to learn a robust subspace representation (SR) of each data point, which seriously affects the clustering performance. Typically, the task of learning a robust SR is cast into a certain optimization problem, usually consisting of two terms: the fidelity term as well as the regularization term. The majority of the previous efforts along this line focused on designing the regularization functions with desirable properties, such as sparsity [5, 19], low-rankness [20, 33], manifold structures [28], or a combination of them [37, 42].

On the other hand, the studies on the fidelity term essentially accounting for the noise effect to robust SR are relatively limited. For the analytical tractability and the low complexity, most SC approaches simply adopted mean square error (MSE) criterion, which provides the optimality only when the noise is *i.i.d.* Gaussian [2]. Because of this limitation, MSE-based frameworks are very sensitive to the non-Gaussian noise [8, 29, 39]. Besides, MSE criterion only considers the second order statistics and may fail to capture sufficient statistical information of the noise signal. To remedy these drawbacks, information theoretic learning (ITL) [15, 29, 23, 7, 8] has been recently suggested to handle non-Gaussian noise, and successfully applied to image recognition [12, 34]. Specifically, ITL aims to find the solution that produces the coding residual with the minimal information [7, 34]. To this end, ITL replaces the MSE criterion with the one based on information theoretic measures, e.g., correntropy [23] and Rényi's entropy [6, 40]. Compared with MSE, ITL does not make Gaussianity assumption, and can exploit higher orders of statistical information of the signal [15]. Despite these desirable properties, *all* the

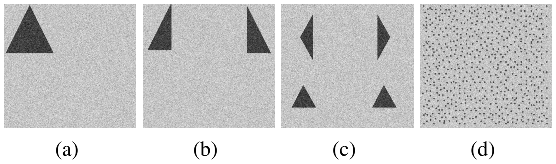

Figure 1. Four images corrupted with the additive white Gaussian noise (standard deviation 50). They have the same backgrounds and the same number of black pixels. The black pixels in (a)-(c) exhibit structural patterns, while the ones in (d) is purely random.

existing ITL-based algorithms still imposed *i.i.d.* assumption on the noise, namely, all the noise samples are generated from the same underlying distribution and no correlation exists among them. Unfortunately, such *i.i.d.* assumption often does not hold in reality. In many practical settings, different portions of the noise could have different statistical behaviors, exhibiting certain structures. A persuasive example is shown in Fig. 1. If we naively model the signals in Figs. 1(a)-(d) with an *i.i.d.* source, then all these four signals would have the same amount of information in terms of the traditional entropy [40]. Apparently, this *i.i.d.* noise model leads to inaccurate information estimations as the signal in Fig. 1(d) should have much larger amount of information from the view of information theory [11]. The aforementioned phenomenon calls for a more generic noise model that can better characterize the statistical behavior of the noise encountered in various practical scenarios.

In this work, we present a new robust SC algorithm through a more generic noise model called independent and piecewise identically distributed (*i.p.i.d.*) model, where we use a union of distributions, rather than a single one, to characterize the statistical behavior of the underlying noise. To the best of our knowledge, this is probably the first SC approach explicitly built upon a generic non-*i.i.d.* noise modeling. The major contributions of our work are as follows:

1. Our framework makes neither the *i.i.d.* nor Gaussianity assumptions on the noise, leading to the essential difference from the existing SC approaches.

2. We develop a novel minimum weighted error entropy (MWEE) criterion for the robust SC, through an *i.p.i.d.* noise model. We demonstrate its effectiveness in exploiting the inherent statistical information of the noise (including structural and purely random ones).

3. We design a relaxation technique to solve the optimization problem for the robust SC under the MWEE criterion, and an efficient implementation can be achieved.

4. The proposed MWEE criterion could be regarded as a general technique and readily incorporated into many existing learning systems to improve the robustness against various types of practical noise.

The rest of the paper is organized as follows. Section 2 reviews the spectral clustering-based SC. Section 3 presents the *i.p.i.d.* noise model. Section 4 introduces the MWEE-based SC algorithm and its optimization. Experimental results are given in Section 5 and Section 6 concludes.

2. Review of the Spectral Clustering-based SC

Let $\{\mathcal{S}_k\}_{k=1}^K$ be a union of K linear subspaces of $\mathbb{R}^N$, and $\{\mathbf{x}_i\}_{i=1}^n$ be a collection of n observed data. Define

$$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] = [\mathbf{X}_1, ..., \mathbf{X}_K]\mathbf{P} \tag{1}$$

where $\mathbf{P}$ is a permutation matrix, and $\mathbf{X}_i \in \mathbb{R}^{N\times n_k}$ contains the n_k data points lying in the subspace $\mathcal{S}_i$. Given the data matrix $\mathbf{X}$, the goal of SC is to correctly separate the data points $\{\mathbf{x}_i\}_{i=1}^n$ into their underlying subspaces. In this work, we focus on the methods based on the spectral clustering [5, 20], which generally comprise two steps: i) learning an affinity matrix indicating the similarity between pairs of the data; and ii) obtaining the clustering results by applying the spectral clustering to the learned affinity matrix. The crucial difference among various SC algorithms lies in the techniques on how to learn the affinity matrix.

The state-of-the-art methods for learning the affinity matrix are based on the robust SR. According to the subspace learning, each data point can be effectively represented as a linear combination of the other points in $\mathbf{X}$, i.e.,

$$\mathbf{X} = \mathbf{X}\mathbf{Z}, \ \ \text{diag}(\mathbf{Z}) = \mathbf{0}, \tag{2}$$

where $\mathbf{X}$ is the self-expressive dictionary and $\mathbf{Z}$ serves as the representation coefficient matrix. Generally, the solution of $\mathbf{Z}$ is not unique, due to the fact that $rank(\mathbf{X}_k) < n_k$. To tackle this challenge, a commonly used technique is to incorporate the prior-domain knowledge, and solve the following regularized optimization problem

$$\min_{\mathbf{Z}} \mathcal{R}(\mathbf{Z}), \ \ \text{s.t.} \ \ \mathbf{X} = \mathbf{X}\mathbf{Z}, \ \text{diag}(\mathbf{Z}) = \mathbf{0}, \tag{3}$$

where $\mathcal{R}(\cdot)$ is a certain regularization function. In practice, $\mathbf{X}$ is often observed with various kinds of noise, i.e.,

$$\mathbf{X} = \mathbf{X}_o + \mathbf{E}_o, \tag{4}$$

where $\mathbf{X}_o$ is the noise-free data matrix and $\mathbf{E}_o$ denotes the noise term. Then (2) can be rewritten as

$$\mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \ \ \text{diag}(\mathbf{Z}) = \mathbf{0}, \tag{5}$$

where $\mathbf{E} = \mathbf{E}_o - \mathbf{E}_o\mathbf{Z}$. In the presence of noise, we usually consider the following problem

$$\min_{\mathbf{Z}} \mathcal{R}(\mathbf{Z}), \ \ \text{s.t.} \ \ \mathcal{L}(\mathbf{X} - \mathbf{X}\mathbf{Z}) < \epsilon, \ \text{diag}(\mathbf{Z}) = \mathbf{0}. \tag{6}$$

Here $\mathcal{L}(\cdot)$ is the fidelity function designed according to the noise behavior, and $\mathcal{R}(\cdot)$ represents the regularization function, which has been devised in many forms with different priors on the subspace structures. For example, ℓ_1-norm leads to the subspace-sparse representation [36, 5], and nuclear norm results in the subspace-low-rank representation [20]. Once obtaining $\mathbf{Z}$, the affinity matrix $\mathbf{M}$ can be induced from $\mathbf{Z}$, e.g., $\mathbf{M} = |\mathbf{Z}| + |\mathbf{Z}^T|$.

At step ii), we apply the spectral clustering [25] to $\mathbf{M}$, and eventually obtain the clustering results.

3. Construction of the *i.p.i.d.* Noise Model

Compared with $\mathcal{R}(\cdot)$, the studies on the fidelity function $\mathcal{L}(\cdot)$ to characterize the noise behavior are relatively limited. For simplicity, most of the existing SC approaches adopted MSE for the fidelity term, naively modeling the noise with *i.i.d.* Gaussian distribution. Though *i.i.d.* Gaussianity assumption greatly simplifies the underlying problems, it does not hold in many real-world scenarios. In this section, we present a generic noise model *i.p.i.d.*, where the *i.i.d.* property is satisfied only in a piecewise fashion. Later, we will show that the proposed *i.p.i.d.* noise model leads to a new design of $\mathcal{L}(\cdot)$ under the framework of SC given in (6). As neither *i.i.d.* nor Gaussianity assumptions are imposed, the resultant SC scheme exhibits superior robustness against various types of noise encountered in practice.

3.1. Definition of the *i.p.i.d.* source and its properties

We define the 1-D *i.p.i.d.* source as follows.

Definition 3.1 *Suppose that* $\mathbf{x} = [x_1, ..., x_N] \in \mathbb{R}^N$ *is a sequence of* N *independent samples. Let* $\{\mathcal{P}_i\}_{i=1}^L$ *be a non-overlapping, sequential partition of the index vector* $[1, 2, \cdots, N]$*, i.e.,*

$$\mathcal{P}_i = \{n_{i-1}+1, n_{i-1}+2, \cdots, n_i\}, i \in \{1, \cdots, L\}, \quad (7)$$

where $n_0 = 0$*,* $n_i < n_{i+1}$*, and* $n_L = N$*. The sequence* $\mathbf{x}$ *is said to be generated by an i.p.i.d. source, if there exists a union of probability density functions* $\{f_i\}_{i=1}^L$*, such that*

$$x_{n_{i-1}+1}, x_{n_{i-1}+2}, \cdots, x_{n_i} \overset{\text{i.i.d.}}{\sim} f_i, \; i \in \{1, \cdots, L\}. \quad (8)$$

The above definition can be readily extended to signals in higher dimensional space, e.g., images and videos. A somewhat similar definition was also given in [35] for the binary source coding. With the Definition 3.1, the *i.p.i.d.* source has the following properties. **Locality**: the *i.p.i.d.* source can well exploit the local behavior of a signal, which is different from a purely *i.i.d.* source; **Fine-description**, the *i.p.i.d.* source characterizes a signal using a union of density functions rather than a single one, providing it more powerful descriptive capability to describe a complex signal. **Generalization**: the traditional *i.i.d.* source is a special case of the *i.p.i.d.* source with $L = 1$.

Owning to these desirable properties, an *i.p.i.d.* source can describe both structural signals (as shown in Figs. 1(a)-(c)) and purely random ones (as shown in Fig. 1(d)). For example, Fig. 1(a) can be satisfactorily modeled by an *i.p.i.d.* source with $L = 2$, where the dark region and the background originate from two different distributions. A more illustrative example is given in Figs. 2(a)-(c), where the noise is dominated by the unconstrained illumination. Obviously, such noise cannot be appropriately modeled with any *i.i.d.* source; but it can be well characterized by the *i.p.i.d.* model. To clarify this point, we show a synthetic image generated by an *i.p.i.d.* source in Fig. 2(d). This synthetic image is produced in a way that each disjoint 8×8 patch is generated by a Gaussian distribution, with the mean gradually decreased by a constant 0.5 from left to right. We can observe that the synthetic image can well approximate the behavior of the illumination noise shown in Fig. 2(c).

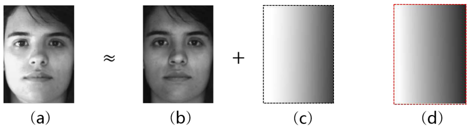

Figure 2. An illustration of the illumination noise. (a) observed image, (b) latent image, (c) noise, (d) an example of the synthetic noise generated by an *i.p.i.d.* source.

3.2. Rényi's entropy of an *i.p.i.d.* Source

We now discuss how to estimate the information of a signal under the *i.p.i.d.* model, which is crucial for the proposed robust SC scheme. We first review the traditional Rényi's entropy of an *i.i.d.* source. Let E be a random variable, and its Rényi's entropy with the order α ($\alpha > 0$ and $\alpha \neq 1$) is defined as

$$H_\alpha(E) = \frac{1}{1-\alpha} \log\Big(\int (f_E(e))^\alpha de \Big). \quad (9)$$

In practice, the probability density function $f_E(e)$ is generally unknown. Parzen window estimation [27] is a commonly adopted algorithm to approximate $f_E(e)$ using finite samples $\{e_i\}_{i=1}^N$, which is given by

$$\hat{f}_E(e) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(e - e_i), \quad (10)$$

where $\kappa_\sigma(\cdot)$ is the Gaussian kernel and σ is the kernel size. Note that an important assumption of the Parzen window estimation is that the samples $\{e_i\}_{i=1}^N$ are *i.i.d.*

To differentiate from the above traditional Rényi's entropy estimator defined under the *i.i.d.* assumption, we call our estimator under the *i.p.i.d.* assumption as the *piecewise Rényi's entropy* (PRE). As a well-known fact, the traditional entropy reflects the smallest number of bits on average, to represent a symbol generated by an *i.i.d.* source. To be consistent with this rule for an *i.p.i.d.* source, we define PRE as follows:

Definition 3.2 *Suppose that* $\mathbf{e} = [e_1, e_2, .., e_N]$ *is a sequence generated by an i.p.i.d. source with the index partition* $\{\mathcal{P}_q\}_{q=1}^L$*. Its information estimator PRE is given by*

$$\hat{H}_\alpha(\mathbf{e}) = \frac{1}{1-\alpha} \sum_q \frac{|\mathcal{P}_q|}{N} \log \int \Big(f_{E_q}(e)\Big)^\alpha de, \qquad (11)$$

where $f_{E_q}(e)$ *is the probability density function estimated with the samples indexed by the partition* $\mathcal{P}_q$*, i.e.,*

$$f_{E_q}(e) = \frac{1}{|\mathcal{P}_q|} \sum_{i \in \mathcal{P}_q} \kappa_\sigma(e - e_i). \qquad (12)$$

It can be shown that the PRE reduces to the traditional entropy $H_\alpha(\mathbf{e})$ when the signal is actually *i.i.d.*; otherwise if the signal is *i.p.i.d.*, PRE can more precisely exploit the entropy information. Without loss of generality, we set $\alpha = 2$, and the resulting estimator is denoted by $\hat{H}_2(\mathbf{e})$.

It should be noted that calculating $\hat{H}_2(\mathbf{e})$ is not straightforward in reality, because the partitions $\{\mathcal{P}_q\}_{q=1}^L$ are generally unknown. In fact, we believe that it is rather challenging, if possible, to estimate the partitions from the given data. This is typically true when the signal is very complex or gradually varied (see e.g., Fig. 2). Fortunately, thanks to the properties of the *i.p.i.d.* source, we can still approximate $\hat{H}_2(\mathbf{e})$ without explicitly knowing the partitions $\{\mathcal{P}_q\}_{q=1}^L$. Specifically, for the data samples in a sufficiently small local region, they can reasonably be assumed as *i.i.d.*, according to the locality property of the *i.p.i.d.* source. Then we can estimate the probability density function for each small local region, and approximate $\hat{H}_2(\mathbf{e})$ by taking the average over all the local regions by resorting to Definition 3.2.

Specifically, let I_q be the location of e_q in the original data space [1]. For each location I_q, we first construct a local region, denoted by $\mathbf{\Omega}_{Iq}$, centered at I_q. Then we estimate the probability density function for $\mathbf{\Omega}_{Iq}$ as

$$f_{E_{I_q}}(e) = \frac{1}{|\mathbf{\Omega}_{Iq}|} \sum_{i \in \mathbf{\Omega}_{I_q}} \kappa_\sigma(e - e_i), \qquad (13)$$

Note that (12) and (13) are equivalent if $\mathbf{\Omega}_{Iq}$ happens to be the actual partition $\mathcal{P}_q$. In the sequel, we use the notation $f_{I_q}(e)$ instead of $f_{E_{I_q}}(e)$ for simplicity.

However, the number of samples in a small local region is often insufficient for the density estimation. Alternatively, we propose to estimate $f_{I_q}(e)$ by using all the samples in $\mathbf{e}$ by introducing a weighting function, potentially achieving more accurate estimation. A somewhat similar strategy was employed in [1] for density estimation with a few samples. To preserve the locality property of the $i.p.i.d.$ source, when estimating $f_{I_q}(e)$, we assign larger weights to the samples with smaller distances to I_q. Concretely, we define the distance of two locations I_i and I_j in the data space as

$$D_{i,j} = Dis(I_i, I_j), \qquad (14)$$

where $Dis(\cdot)$ is a certain distance function, e.g., $||\cdot||_2$. Then the probability density function for $\mathbf{\Omega}_{Iq}$ is estimated by

$$\hat{f}_{I_q}(e) = \sum_{i=1}^N c(D_{q,i}) \kappa_\sigma(e - e_i). \qquad (15)$$

Here $c(\cdot)$ is an appropriately designed weighting function. In our work, we simply choose $c(\cdot)$ as a Gaussian function

$$c(D_{q,i}) = \frac{1}{Q} e^{-\frac{(D_{q,i})^2}{\sigma_w^2}}, \qquad (16)$$

where Q is the normalizer such that $\sum_{I_i} c(D_{q,i}) = 1$, and σ_w^2 is empirically set as $\frac{N}{1000}$. We then call (15) the weighted Parzen window (WPW) estimation.

Upon estimating the density $\hat{f}_{Iq}(e)$ for each $\mathbf{\Omega}_{Iq}$, the PRE $\hat{H}_2(\mathbf{e})$ can then be approximated by $\bar{H}_2(\mathbf{e})$ through taking the average over all the locations, i.e.,

$$\begin{aligned} \bar{H}_2(\mathbf{e}) &= -\frac{1}{N} \sum_{I_q} \log \int \Big(\hat{f}_{I_q}(e)\Big)^2 de \\ &= -\frac{1}{N} \sum_{I_q} \log \sum_{i,j=1}^N c(D_{q,i}) c(D_{q,j}) \kappa_{\sqrt{2}\sigma}(e_i - e_j). \end{aligned} \qquad (17)$$

3.3. Relationship among $H_2(\mathbf{e})$, $\hat{H}_2(\mathbf{e})$ and $\bar{H}_2(\mathbf{e})$

Compared with $\hat{H}_2(\mathbf{e})$ defined in (11), $\bar{H}_2(\mathbf{e})$ does not need to know the partitions $\{\mathcal{P}_q\}_{q=1}^L$ explicitly. Furthermore, it can be proved below that $\bar{H}_2(\mathbf{e})$ derived under the $i.p.i.d.$ assumption is still equivalent to the traditional Rényi's entropy $H_2(\mathbf{e})$ under the $i.i.d.$ setting.

Theorem 3.1 *Suppose that the signal elements in* $\mathbf{e} = [e_1, e_2, .., e_N]$ *are independently sampled from the same distribution* $f(e)$*. Then the PRE estimator* $\bar{H}_2(\mathbf{e})$ *defined in (17) is equivalent to the traditional Rényi's entropy* $H_2(\mathbf{e})$ *given by (9) and (10).*

The proof is given in the supplementary file. It provides a fundamental theoretical basis for the capability of $\bar{H}_2(\mathbf{e})$ to characterize the $i.i.d.$ source.

To further show the relationship among the traditional Rényi's entropy $H_2(\mathbf{e})$, the PRE $\hat{H}_2(\mathbf{e})$ and its approximation $\bar{H}_2(\mathbf{e})$, we give a toy example here by generating an *i.p.i.d.* sequence $\mathbf{e} = [e_1, e_2, ..., e_N]$ $(N = 2000)$ with L partitions. For the q-th partition, the samples are independently generated by a Gaussian distribution

$$f_{E_q}(e) = \frac{1}{\sqrt{2\pi q^2}} \exp\Big(-\frac{(e-\mu)^2}{2q^2}\Big), \qquad (18)$$

where $\mu = \frac{200}{L}(q-1)$. For simplicity, each partition has the same number of samples, i.e., $\lfloor N/L \rfloor$. Obviously, $L = 1$

[1] For the 1-D signal (e.g., voice), I_q is a scalar. For the 2-D signal (e.g., image), I_q is a 2-D index.

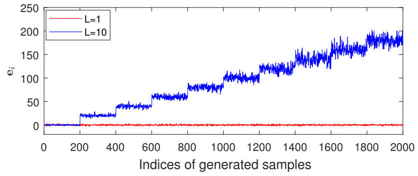

Figure 3. Two examples of the *i.p.i.d.* sequence.

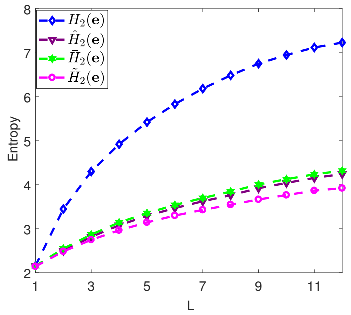

Figure 4. $H_2(\mathbf{e})$, $\hat{H}_2(\mathbf{e})$, $\bar{H}_2(\mathbf{e})$ and $\tilde{H}_2(\mathbf{e})$ w.r.t. L

implies that the sequence $\mathbf{e}$ is generated by a standard normal distribution; while when $L > 1$, $\mathbf{e}$ is a 1-D signal with structures. Fig. 3 shows two examples of the sequences when $L = 1$ (red) and $L = 10$ (blue). Fig. 4 plots the curves of different estimators with various number of partitions. We can observe that under the *i.i.d.* case, i.e., $L = 1$, $H_2(\mathbf{e})$, $\hat{H}_2(\mathbf{e})$ and $\bar{H}_2(\mathbf{e})$ are the same, which coincides with Theorem 3.1. When $L > 1$, $H_2(\mathbf{e})$ becomes much larger than the PRE $\hat{H}_2(\mathbf{e})$ and $\bar{H}_2(\mathbf{e})$. This is because $H_2(\mathbf{e})$ totally ignores the structural information of $\mathbf{e}$, and simply treats all the signal elements as *i.i.d.* From Fig. 4, we can also observe that $\bar{H}_2(\mathbf{e})$ can well approximate $\hat{H}_2(\mathbf{e})$, even without knowing the partitions.

4. Proposed SC Method and Its Optimization

Based on the proposed *i.p.i.d.* noise model, we now suggest a new fidelity function $\mathcal{L}(\cdot)$ in (6) for the robust SC.

4.1. MWEE-based sparse SC

In this work, we focus on the popular Sparse Subspace Clustering (SSC) method [5], which adopts ℓ_1-norm for $\mathcal{R}(\cdot)$ to achieve a subspace-sparse representation, while designing $\mathcal{L}(\cdot)$ under MSE criterion. Namely,

$$\underset{\mathbf{Z}}{\text{argmin}}\ ||\mathbf{Z}||_1,\ \text{ s.t. }\ ||\mathbf{X}-\mathbf{XZ}||_F^2 < \epsilon,\ \text{diag}(\mathbf{Z}) = \mathbf{0}. \quad (19)$$

As aforementioned, the MSE criterion has many serious limitations. Motivated by the great success of ITL to handle non-Gaussian noise, we suggest to design a new ITL-type fidelity function $\mathcal{L}(\cdot)$ through the proposed *i.p.i.d.* noise model. Specifically, given a data point and a dictionary, the ITL-based framework aims to find a representation producing the coding residual with minimal information [12, 34]. In light of this motivation, we replace the Frobenius norm in (19) with our proposed PRE, and we have

$$\underset{\mathbf{Z}}{\text{argmin}}\ ||\mathbf{Z}||_1,\ \text{ s.t. }\ \mathbf{\Phi}(\mathbf{X}-\mathbf{XZ}) < \epsilon,\ \text{diag}(\mathbf{Z}) = \mathbf{0}, \quad (20)$$

where

$$\mathbf{\Phi}(\mathbf{X}-\mathbf{XZ}) = \sum_{i=1}^{n} \bar{H}_2(\mathbf{x}_i - \mathbf{X}\mathbf{z}_i), \quad (21)$$

and $\mathbf{z}_i$ is the i-th column of $\mathbf{Z}$. In this work, we name the criterion of minimizing $\bar{H}_2(\mathbf{e})$ as Minimum Weighted Error Entropy (MWEE) criterion, due to the weighted nature of $\bar{H}_2(\mathbf{e})$. Different from MSE and *all* the existing ITL criteria, such as the ones based on correntropy [12] and Rényi's entropy [40, 34], MWEE built upon the *i.p.i.d.* model makes neither the *i.i.d.* nor Gaussianity assumptions. The minimization target PRE can better reflect the inherent information of the noise, no matter it is structural or purely random. As expected and will be shown experimentally, our algorithm is very robust against various kinds of noise.

The problem (20) can be decomposed into n independent subproblems, with the i-th one expressed as

$$\underset{\mathbf{z}_i\in\mathbb{R}^n}{\text{argmin}}\ ||\mathbf{z}_i||_1,\ \text{ s.t. }\ \bar{H}_2(\mathbf{x}_i - \mathbf{X}\mathbf{z}_i) < \epsilon_i,\ \mathbf{z}_{i,i} = 0, \quad (22)$$

where $\epsilon = \sum_{i=1}^{n} \epsilon_i$. To handle the problem (22), we can first solve

$$\underset{\mathbf{z}_i'\in\mathbb{R}^{n-1}}{\text{argmin}}\ ||\mathbf{z}_i'||_1,\ \text{ s.t. }\ \bar{H}_2(\mathbf{x}_i - \hat{\mathbf{X}}\mathbf{z}_i') < \epsilon_i. \quad (23)$$

where $\hat{\mathbf{X}} = [\mathbf{x}_1, .., \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, .., \mathbf{x}_n]$. Then the solution $\mathbf{z}_i$ of (22) is eventually computed by

$$\mathbf{z}_i = [\mathbf{z}_{i,1}', ..., \mathbf{z}_{i,i-1}', 0, \mathbf{z}_{i,i}', ..., \mathbf{z}_{i,n-1}'].$$

Nevertheless, it is very difficult to solve the problem (23), since 1) $\bar{H}_2(\mathbf{e})$ defined in (17) is a summation of complex logarithmic functions; and 2) the kernel function $\kappa_{\sqrt{2}\sigma}(\cdot)$ is highly non-convex [12]. To tackle this challenge, we now propose a relaxation technique for $\bar{H}_2(\mathbf{e})$ such that the resulting problem can be efficiently solved.

4.2. Relaxation of $\bar{H}_2(\mathbf{e})$

Define

$$c_{i,j}^q = c(D_{q,i})c(D_{q,j}), \quad (24)$$

where $D_{q,i}$ is given in (14). Since $\sum_{I_q} \frac{1}{N} = 1$, by the convexity of the negative log function and the Jassen's inequality, we have

$$\bar{H}_2(\mathbf{e}) \geq -\text{log} \sum_{I_q} \frac{1}{N} \sum_{i,j=1}^{N} c_{i,j}^q \kappa_{\sqrt{2}\sigma}(e_i - e_j) \quad (25)$$

$$= -\text{log} \sum_{I_q} \sum_{i,j=1}^{N} c_{i,j}^q \kappa_{\sqrt{2}\sigma}(e_i - e_j) + \text{log}N.$$

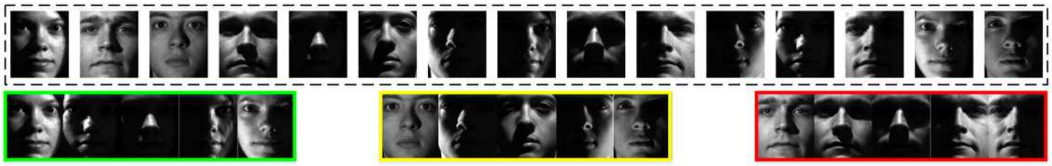

Figure 5. Illustration of the face clustering. Images are from the `Extended Yale B` database [18].

Algorithm 1 Half-quadratic algorithm for the problem (32)

```
Input: The data matrix $\mathbf{A} = [\mathbf{x}_1, .., \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, .., \mathbf{x}_n]$, a data point $\mathbf{y} = \mathbf{x}_i$, the parameter $\lambda$ and $t = 0$.
Calculate $\tilde{\mathbf{y}} = [\tilde{y}_1, ..., \tilde{y}_N]^T$ and $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1^T, ..., \tilde{\mathbf{x}}_N^T]^T$.
While 'not converged', do
    $u_i^{t+1} = \frac{1}{2\sigma^2}\kappa_{\sqrt{2}\sigma}(\tilde{y}_i - \tilde{\mathbf{x}}_i\mathbf{z}^t),\ i = 1, 2, \cdots, N$
    $\mathbf{z}^{t+1} = \underset{\mathbf{z}}{\text{argmin}}\ (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{z})^T \text{diag}(\mathbf{u}^{t+1})(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{z}) + \lambda||\mathbf{z}||_1$
    $t = t + 1$
end while
Output: The representation vector $\mathbf{z}$.
```

Algorithm 2 MWEE-based SC (MWEE-S)

```
Input: The data matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$, the number of subspaces $K$, the parameter $\lambda$.
Normalize the columns of $\mathbf{X}$ to have unit $l_2$ norm.
Compute the representation matrix $\mathbf{Z}$ by solving (20) to deal with linear subspaces , or solving (33) for affine subspaces.
Construct the similarity matrix $\mathbf{M} = |\mathbf{Z}| + |\mathbf{Z}|^T$.
Apply the spectral clustering algorithm [25] to $\mathbf{M}$.
Output: $K$ clusters.
```

Let

$$\tilde{H}_2(\mathbf{e}) = -\text{log}S(\mathbf{e}) + \text{log}N, \tag{26}$$

where

$$S(\mathbf{e}) = \sum_{i,j=1}^{N} w_{i,j}\kappa_{\sqrt{2}\sigma}(e_i - e_j),\quad w_{i,j} = \sum_{I_q} c_{i,j}^q. \tag{27}$$

From (25), we can see that $\tilde{H}_2(\mathbf{e})$ is a lower bound of $\bar{H}_2(\mathbf{e})$. It can be shown that when the signal elements in $\mathbf{e}$ are *i.i.d.*, the inequality in (25) holds with equality. Further, $\bar{H}_2(\mathbf{e})$ and $\tilde{H}_2(\mathbf{e})$ have the same minimizer $\mathbf{e} = c\mathbf{1}$ for some constant c. In Fig. 4, we also plot the curve of $\tilde{H}_2(\mathbf{e})$ with different number of partitions. Motivated by the work [16], we suggest to substitute $\bar{H}_2(\mathbf{e})$ with $\tilde{H}_2(\mathbf{e})$ in (23).

Noticing that $\tilde{H}_2(\mathbf{e})$ is monotonically decreasing w.r.t. $S(\mathbf{e})$, minimizing $\tilde{H}_2(\mathbf{e})$ is equivalent to minimizing $-S(\mathbf{e})$. By replacing $\bar{H}_2(\mathbf{e})$ with $-S(\mathbf{e})$, and introducing a Lagrange multiplier λ, we reformulate the problem (23) as

$$\underset{\mathbf{z}_i'}{\text{argmin}} - S(\mathbf{x}_i - \hat{\mathbf{X}}\mathbf{z}_i') + \lambda||\mathbf{z}_i'||_1. \tag{28}$$

For the notation simplicity, we further let $\mathbf{y} = \mathbf{x}_i$, $\mathbf{A} = \hat{\mathbf{X}}$ and $\mathbf{z} = \mathbf{z}_i'$. Then, the problem (28) becomes

$$\underset{\mathbf{z}}{\text{argmin}} - S(\mathbf{y} - \mathbf{A}\mathbf{z}) + \lambda||\mathbf{z}||_1. \tag{29}$$

By resorting to a similar strategy proposed in [34], we approximate $S(\mathbf{e})$ with

$$\tilde{S}(\mathbf{e}) = \sum_{i=1}^{N} \kappa_{\sqrt{2}\sigma}\left(\sum_{j=1}^{N} w_{i,j}(e_i - e_j)\right). \tag{30}$$

Denote the i-th entry of $\mathbf{y}$ by y_i and the i-th row of $\mathbf{A}$ by $\mathbf{a}_i$. Since $\sum_j w_{i,j} = 1$ (proof given in the supplementary file), we have

$$\begin{aligned}\tilde{S}(\mathbf{y} - \mathbf{A}\mathbf{z}) &= \sum_{i=1}^{N} \kappa_{\sqrt{2}\sigma}\left(\sum_{j=1}^{N} w_{i,j}\left((y_i - \mathbf{a}_i\mathbf{z}) - (y_j - \mathbf{a}_j\mathbf{z})\right)\right)\\ &= \sum_{i=1}^{N} \kappa_{\sqrt{2}\sigma}\left(y_i - \sum_{j=1}^{N} w_{i,j}y_j - \left(\mathbf{a}_i - \sum_{j=1}^{N} w_{i,j}\mathbf{a}_j\right)\mathbf{z}\right).\end{aligned} \tag{31}$$

Substituting $S(\mathbf{y} - \mathbf{A}\mathbf{z})$ by $\tilde{S}(\mathbf{y} - \mathbf{A}\mathbf{z})$, the problem (29) finally becomes

$$\underset{\mathbf{z}}{\text{argmin}}\ -\sum_{i=1}^{N} \kappa_{\sqrt{2}\sigma}(\tilde{y}_i - \tilde{\mathbf{x}}_i\mathbf{z}) + \lambda||\mathbf{z}||_1, \tag{32}$$

where $\tilde{y}_i = y_i - \sum_{j=1}^N w_{i,j}y_j$ and $\tilde{\mathbf{x}}_i = \mathbf{a}_i - \sum_{j=1}^N w_{i,j}\mathbf{a}_j$. (32) can be efficiently solved via the half-quadratic theory [26]. Algorithm 1 shows the optimization procedures, and the detailed derivation is given in the supplementary file.

Remark: Note that in many cases, data points lie in a union of *affine* subspaces rather than linear subspaces, as will be discussed in Section 5.2. To deal with affine subspaces, we adopt the strategy suggested in [5], by introducing additional linear equality constraints in (20), i.e.,

$$\begin{aligned}&\underset{\mathbf{Z}}{\text{argmin}}\ ||\mathbf{Z}||_1,\\ &\text{s.t.}\ \ \boldsymbol{\Phi}(\mathbf{X} - \mathbf{X}\mathbf{Z}) < \epsilon,\ \ \mathbf{Z}^T\mathbf{1} = \mathbf{1},\ \text{diag}(\mathbf{Z}) = \mathbf{0}.\end{aligned} \tag{33}$$

This problem can be efficiently solved by using a similar technique described in Algorithm 1, incorporating with the Alternating Direction Method of Multipliers (ADMM) method [5].

Upon having the representation matrix $\mathbf{Z}$, we then build the affinity matrix by $\mathbf{M} = |\mathbf{Z}| + |\mathbf{Z}|^T$. Finally, we apply spectral clustering [25] to $\mathbf{M}$, and obtain the clustering results. The whole SC algorithm is summarized in Algorithm 2. In this work, we refer the MWEE-based SC algorithm to as MWEE-S for short.

Table 1. Clustering accuracy (%) of different algorithms on the `Extended Yale B`.

Methods	LSA	SSC0	SSC1	LRR	TSC	L2-G	S^{3}C	MWEE-S
2 subjects	71.09	99.22	96.89	97.66	97.66	98.44	99.22	**100.0**
4 subjects	42.58	75.39	92.97	93.75	91.80	98.44	99.22	**100.0**
6 subjects	45.05	85.94	94.01	96.62	93.49	98.44	95.83	**100.0**
8 subjects	33.98	60.35	93.75	75.59	90.43	97.66	94.92	**100.0**
10 subjects	32.50	53.75	87.19	76.56	86.41	96.56	94.69	**99.84**

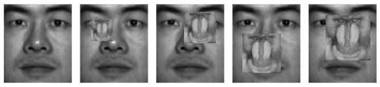

Figure 6. Some simulated examples. Images from left to right are randomly occluded by 0%, 10%, 20%, 30% and 40%, respectively.

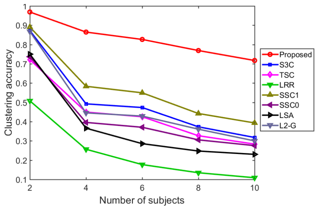

Figure 7. Average clustering accuracy of different algorithms on the `Extended Yale B` against 25% contiguous occlusion.

5. Experimental Results

We evaluate our proposed SC algorithm MWEE-S, in dealing with two practical problems: face clustering and motion segmentation. To embrace the concept of reproducible research [32], the code of our paper will be available upon the acceptance.

5.1. Face clustering

Given a set of facial images from multiple subjects, the goal of face clustering is to separate them according to their underlying subjects. An example is shown in Fig. 5. The `Extended Yale B` dataset [18] is adopted in this experiment, which contains 2432 frontal face images from 38 subjects, with 64 instances for each subject. Images in this dataset are captured under various lighting conditions. For efficiency, we resize all the images to 96×84. We compare the performance of our MWEE-S in (20) with SSC0 [4], SSC1 [5], LRR [20], TSC [13], LSA [38], S^{3}C [19] and L2-G [28]. We use the codes provided by their authors with the default parameter settings. More specifically, for S^{3}C, we adopt the soft S^{3}C implementation, since it leads to the best performance among all the variants [19]. For LRR, we use the code newly updated [20]. The difference between SSC0 and SSC1 is that SSC0 adopts $||\cdot||_F^2$ as the fidelity function while SSC1 uses $||\cdot||_1$. For our proposed MWEE-S, the parameter λ is consistently set as 10^{-4}.

As shown in Fig. 5, the noise in the `Extended Yale B` is mainly caused by the unconstrained illumination, which obviously satisfies neither the *i.i.d.* assumption nor the Gaussianity. Table 1 reports the clustering accuracy of different algorithms over the `Extended Yale B`, for the first 2, 4, 6, 8 and 10 subjects. We can see that when the number of subjects increases, the performance of the MSE-based algorithms (such as SSC0 and LRR) drops rapidly. This is because MSE criterion simply treats the noise as *i.i.d.* Gaussian. By resorting to the MWEE-based criterion derived under the *i.p.i.d.* noise model, our MWEE-S outperforms the competing methods for *all* the cases. Notably, MWEE-S obtains 100% clustering accuracy when the number of subjects is below 10, while only miss-clustering one image when the number of subjects is 10. It can also be seen that the recent works S^{3}C [19] and L2-G [28] also achieved rather good performance on this dataset. However, S^{3}C adopts more complex regularization functions, while L2-G applies a post-processing on the representation coefficients. Furthermore, as will be clear soon, they are not robust under more complex noise scenarios.

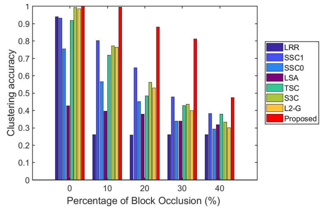

Figure 8. Average clustering accuracy of different algorithms against various levels of occlusion.

Effect of the contiguous block occlusion: We now test the effectiveness of the proposed MWEE-S in the presence of contiguous occlusions. For each facial image, we first randomly select a region, and then substitute it with an unrelated image patch. Specifically, the image 'Baboon' is used for the occlusion simulation, which was adopted in [36, 34] as well. Some examples are given in Fig. 6.

Note that the noise in this scenario can hardly be assumed to be *i.i.d.*, since it is the combination of the illumination and the unrelated image 'Baboon', both of which are highly structural. Fig. 7 plots the clustering accuracy of different methods against 25% occlusion, over the facial images of various number of subjects. To alleviate the impact of random occlusion positions, all the results are averaged over 10 random runs. We can observe that our MWEE-S achieves considerably better performance than the competitors, and the gain margin becomes larger when the number of subjects increases. Compared with the results in Table 1, it can be noticed that the performance drops severely, for all the competing methods against 25% occlusions. In contrast, the performance degradation of WMEE-S is much more graceful. Such phenomenon further demonstrates that the WMEE criterion designed under the *i.p.i.d.* can better characterize the noise behavior.

Effect of the occlusion level: To investigate the impact of different occlusion levels on the clustering performance, we vary the occlusion level from 0 to 40%, while fixing the number of subjects to be 4. Fig. 8 depicts the results of different algorithms. As can be observed, our method achieves the best clustering performance for all the occlusion level-

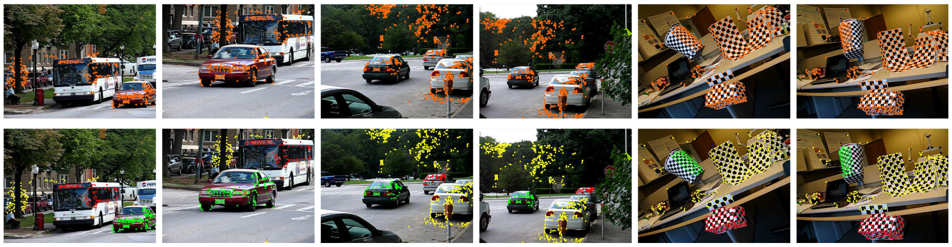

Figure 9. Example frames with tracked features from three videos in `Hopkins 155` [31]. Given feature points on multiple rigidly moving objects tracked in a video (top), motion segmentation aims to separate the feature trajectories according to moving objects (bottom).

Table 2. Clustering error (%) of different algorithms on the `Hopkins155` database.

Methods		LSR	SSC0	SSC1	LRR	LRSC	L2-G	S^{3}C	MWEE-S
2F	Avg.	2.98	6.97	2.18	1.60	3.42	5.54	2.20	**1.22**
	Med.	0.30	0.21	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Std.	7.48	12.69	7.24	4.66	8.83	11.18	6.89	**4.20**
4K	Avg.	3.21	7.05	2.42	2.35	3.35	5.81	2.33	**1.77**
	Med.	0.38	0.21	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Std.	7.79	12.82	7.51	7.30	8.76	11.59	6.98	**6.12**

s. The performance gaps of our scheme over the competing methods are quite remarkable, especially when the occlusion level is between 10% to 30%.

5.2. Motion segmentation

Motion segmentation aims to segment a video sequence with multiple rigidly moving objects into several spatiotemporal regions, each of which corresponds to one moving object in the scene. Fig. 9 shows some examples of three video sequences, where we only draw two frames with tracked points for each video. Let F be the number of frames in a video sequence. Generally, the motion segmentation problem can be solved by first tracking the spatial positions of n feature points $\mathbf{x}_{f,i} \in \mathbb{R}^2$ ($f = [1,..,F], i = [1,..,n]$) across the frames of the video, and then clustering the feature point trajectories according to their underlying motions [5, 19]. Specifically, the trajectory of the i-th feature point is formed by stacking its spatial positions in the video, namely

$$\mathbf{x}_i = [\mathbf{x}_{1,i}^T, \mathbf{x}_{2,i}^T, .., \mathbf{x}_{F,i}^T]^T \in \mathbb{R}^{2F}. \tag{34}$$

Then all the trajectories of a video can be represented by a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$. Since the trajectories of a rigid motion lie in an affine subspace of $\mathbb{R}^{2F}$ of dimension at most 3 [5], the feature trajectories of K rigid motions lie in a union of K subspaces of $\mathbb{R}^{2F}$.

We adopt the `Hopkins155` database [31] for this experiment, which provides an extensive benchmark for testing many subspace segmentation methods. The `Hopkins155` database consists of 156 video sequences (hence 156 subspace clustering tasks), with 2 or 3 motions in each video. The feature points are extracted and tracked across frames. On average, each video of 2 motions has 266 trajectories and 30 frames, while each video of 3 motions has 398 trajectories and 29 frames. Some example frames are given in Fig. 9, where the feature points from one moving object are marked in the same color (bottom). For the motion segmentation task, we adopt MWEE-S in (33) tailored for the affine subspace, and compare it with LSR[24], SSC0 [4], SSC1 [5], LRR [20], LRSC [9], L2-G [28], S^{3}C[19]. The difference between SSC0 and SSC1 is that SSC1 uses the affine constraint $\mathbf{Z}^T\mathbf{1} = \mathbf{1}$, while SSC0 does not. For LSR, we utilize the LSR1 implementation.

The first experiment is conducted on the original $2F$-dimensional data. Since the rank of each linear subspace is at most 4 [5], we also do the experiment on the $4K$-dimensional data, where K is the number of motions in each sequence. PCA is adopted for the dimension reduction. The clustering results are summarized in Table 2, where we report the average, median and standard deviation of the clustering errors. Generally, the clustering error of each method increases when the data dimension is reduced from $2F$ to $4K$, due to the information loss. However, in both cases, the proposed method *significantly* outperforms all the competing algorithms, in terms of both the average clustering error and standard deviation. This implies that modeling the noise of trajectories with an *i.p.i.d.* source, other than an *i.i.d.* Gaussian, indeed helps for motion segmentation.

6. Conclusions

This paper has presented a new robust SC approach via the *i.p.i.d.* noise modeling. Different from the traditional SC algorithms, our method makes neither the *i.i.d.* nor Gaussianity assumptions on the noise. Based on the proposed *i.p.i.d.* model, we have developed a novel optimization criterion MWEE, which can well characterize the inherent information of the noise, despite it is structural or purely random. Such desirable properties make the developed SC method very robust against various kinds of noise encountered in practice. In fact, the proposed MWEE criterion can be readily integrated into many learning systems, which could provide better performance than the state-of-the-art methods, as will be demonstrated in our future work.

Acknowledgments: This work was supported in part by the Macau Science and Technology Development Fund under Grants FDCT/022/2017/A1 and FDCT/077/2018/A2, and in part by the Research Committee at the University of Macau under Grant MYRG2016-00137-FST and MYRG2018-00029-FST.

References

[1] G. A. Babich and O. I. Camps. Weighted parzen windows for pattern classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(5):567–570, 1996. 4

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 1

[3] Y. Chen, N. M. Nasrabadi, and T. D. Tran. Hyperspectral image classification using dictionary-based sparse representation. *IEEE Trans. Geosci. Remote Sens.*, 49(10):3973–3985, 2011. 1

[4] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2790–2797, 2009. 7, 8

[5] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013. 1, 2, 3, 5, 6, 7, 8

[6] D. Erdogmus, K. E. Hild, J. C. Principe, M. Lazaro, and I. Santamaria. Adaptive blind deconvolution of linear channels using renyi's entropy with parzen window estimation. *IEEE Trans. Signal Process.*, 52(6):1489–1498, 2004. 1

[7] D. Erdogmus and J. C. Principe. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Trans. Signal Process.*, 50(7):1780–1786, 2002. 1

[8] D. Erdogmus and J. C. Principe. From linear adaptive filtering to nonlinear information processing - the design and analysis of information processing systems. *IEEE Signal Process. Mag.*, 23(6):14–33, 2006. 1

[9] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution to robust subspace estimation and clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 1801–1807, 2011. 8

[10] A. Gruber and Y. Weiss. Multibody factorization with uncertainty and missing data using the em algorithm. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, volume 1, pages 707–714. IEEE, 2004. 1

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2009. 2

[12] R. He, W. S. Zheng, and B. G. Hu. Maximum correntropy criterion for robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1561–1576, 2011. 1, 5

[13] R. Heckel and H. Bölcskei. Robust subspace clustering via thresholding. *IEEE Trans. Signal Process.*, 61(11):6320–6342, 2015. 7

[14] W. Hong, J. Wright, K. Huang, and Y. Ma. Multiscale hybrid linear models for lossy image representation. *IEEE Trans. Image Process.*, 15(12):3655–3671, 2006. 1

[15] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou. Learning theory approach to minimum error entropy criterion. *J. Mach. Learn. Res.*, 14:377–397, 2013. 1

[16] T. Huang, Y. Liu, H. Meng, and X. Wang. Adaptive compressed sensing via minimizing cramer?rao bound. *IEEE Sig. Process. Lett.*, 21(3):270–274, 2014. 6

[17] K.-i. Kanatani. Motion segmentation by subspace separation and model selection. In *Proc. IEEE Int. Conf. Comput. Vis.*, volume 2, pages 586–591. IEEE, 2001. 1

[18] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(5):684–698, 2005. 6, 7

[19] C. G. Li, C. You, and R. Vidal. Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework. *IEEE Trans. Image Process.*, 26(6):2988–3001, 2017. 1, 7, 8

[20] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):171–184, 2013. 1, 2, 3, 7, 8

[21] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *Proc.Int. Conf. Mach. Learn*, pages 663–670, 2010. 1

[22] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1615–1622. IEEE, 2011. 1

[23] W. Liu, P. P. Pokharel, and J. C. Principe. Correntropy: Properties and applications in non-gaussian signal processing. *IEEE Trans. Signal Process.*, 55(11):5286–5298, 2007. 1

[24] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan. Robust and efficient subspace segmentation via least squares regression. In *Proc. Eur. Conf. Comput. Vis.,*, pages 347–360. Springer, 2012. 8

[25] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proc. Neural Information Processing Systems,*, pages 849–856, 2002. 1, 3, 6

[26] M. Nikolova and M. K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27(3):937–966, 2005. 6

[27] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. 3

[28] X. Peng, Z. Yu, Z. Yi, and H. Tang. Constructing the l2-graph for robust subspace learning and subspace clustering. *IEEE Trans. on Cybern.*, 47(4):1053–1066, 2017. 1, 7, 8

[29] J. C. Principe. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer, New York, 1st edition, 2010. 1

[30] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. Comput. Vis.,*, 9(2):137–154, 1992. 1

[31] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 1–8, 2007. 8

[32] P. Vandewalle, J. Kovacevic, and M. Vetterli. Reproducible research in signal processing. *IEEE Signal Process. Mag.*, 26(3):37–47, 2009. 7

[33] R. Vidal and P. Favaro. Low rank subspace clustering (LRSC). *Pattern Recognit. Lett.*, 43:47 – 61, 2014. 1

[34] Y. Wang, Y. Y. Tang, and L. Li. Robust face recognition via minimum error entropy-based atomic representation. *IEEE Trans. Image Process.*, 24(12):5868–5878, 2015. 1, 5, 6, 7

[35] F. M. J. Willems. Coding for a binary independent piecewise-identically-distributed source. *IEEE Trans. Signal Process.*, 42(6):2210–2217, 1996. 3

[36] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009. 1, 3, 7

[37] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang. Discriminative transfer subspace learning via low-rank and sparse representation. *IEEE Trans. Image Process.*, 25(2):850–863, 2016. 1

[38] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Proc. European Conf. Computer Vision*, pages 94–106. Springer, 2006. 7

[39] M. Yang, L. Zhang, J. Yang, and D. Zhang. Regularized robust coding for face recognition. *IEEE Trans. Image Process.*, 22(5):1753–1766, 2013. 1

[40] X.-T. Yuan and B.-G. Hu. Robust feature extraction via information theoretic learning. In *Proc. Int. Conf. Mach. Learn.*, pages 1193–1200. ACM, 2009. 1, 2, 5

[41] T. Zhang, A. Szlam, and G. Lerman. Median k-flats for hybrid linear modeling with many outliers. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 234–241. IEEE, 2009. 1

[42] Y. Zhang, D. Shi, J. Gao, and D. Cheng. Low-rank-sparse subspace representation for robust regression. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 7445–7454, 2017. 1

[43] Y. Zhang, Z. Sun, R. He, and T. Tan. Robust subspace clustering via half-quadratic minimization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3096–3103, 2013. 1

[44] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang. Robust principal component analysis with complex noise. In *Proc. Int. Conf. Mach. Learn.*, pages 55–63. ACM, 2014. 1