Weighted Error Entropy based Information Theoretic Learning for Robust Subspace Representation

Yuanman Li, *Member, IEEE,* Jiantao Zhou, *Senior Member, IEEE,* Jinyu Tian, *Student Member, IEEE,* Xianwei Zheng, *Member, IEEE,* and Yuan Yan Tang, *Life Fellow, IEEE*

***Abstract*—In most of the existing representation learning frameworks, the noise contaminating the data points is often assumed to be independent and identically distributed (*i.i.d.*), where the Gaussian distribution is often imposed. This assumption, though greatly simplifies the resulting representation problems, may not hold in many practical scenarios. For example, the noise in face representation is usually attributable to local variation, random occlusion and unconstrained illumination, which is essentially structural and hence does not satisfy the *i.i.d.* property or the Gaussianity. In this paper, we devise a generic noise model, referred to as independent and piecewise identically distributed (*i.p.i.d.*) model for robust presentation learning, where the statistical behavior of the underlying noise is characterized using a union of distributions. We demonstrate that our proposed *i.p.i.d.* model can better describe the complex noise encountered in practical scenarios, and accommodate the traditional *i.i.d.* one as a special case. Assisted by the proposed noise model, we then develop a new information-theoretic learning (ITL) framework for robust subspace representation (SR) through a novel minimum weighted error entropy criterion. Thanks to the superior modeling capability of the *i.p.i.d.* model, our proposed learning method achieves superior robustness against various types of noise. When applying our scheme to the subspace clustering and image recognition problems, we observe significant performance gains over the existing approaches.**

***Index Terms*—Information-theoretic learning, subspace representation, weighted Parzen window, independent and piecewise identically distributed**

I. Introduction

HIGH-DIMENSIONAL data including images, videos, documents etc. are ubiquitous in many real-world problems. The high dimensionality brings many challenges in the data processing chain, not only increasing the computational complexity, but also leading to inferior performance due to the "curse of dimensionality" [1]–[3]. Fortunately, it has been observed that many high-dimensional data are not uniformly distributed in their ambient space, but usually lie in a latent subspace of low dimensionality. Some representative examples of such high-dimensional data include face images of the same subject [4], [5], multiple instances of a hand-written digit with different writing styles [3], and trajectories of the same moving object in videos [6], [7]. As a result, we can naturally assume that data from several classes are distributed in a collection of subspaces of low-dimensionality. It is therefore of paramount importance to reveal the low-dimensional structures of data embedded in the space of high-dimensionality, which could eventually bring performance gains for various machine learning and pattern recognition tasks.

Subspace representation (SR), as one of the most important learning techniques, aims to represent data drawn from the high-dimensional space as low-dimensional structures [8]. SR has been widely used in image clustering [9]–[11], motion segmentation [12]–[14], image classification [15], [16] etc. In reality, the observed data points are often contaminated with various types of noise, which could severely affect the performance of the resulting algorithms [17]. For simplicity, most learning algorithms for SR assumed that the noise is *i.i.d.*, i.e., all the noise points are independently generated from the same underlying distribution and have no correlations. Apparently, this assumption is too strong in many practical scenarios, leading to unsatisfactory performance. For instance, facial images captured in real environments are usually corrupted by severe contiguous occlusions, local variations and unconstrained illuminations, where the noise signals (w.r.t. clearly visible faces) are highly structural and hence violate both the *i.i.d.* assumption and the Gaussianity. Under this circumstance, how to design a model better characterizing true behaviors of different types of noise becomes an urgent and challenging problem for the robust SR.

A. Problem Statement

Let $\{\mathcal{S}_k\}_{k=1}^K$ represent a collection of K linear subspaces, and $\mathbf{x}_i$ be a noise-free data point. Define

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

where $\mathbf{\Gamma}$ is a permutation matrix, and $\mathbf{X}_i \in \mathbb{R}^{N\times n_k}$ contains the n_k data distributed in the subspace $\mathcal{S}_i$ of dimension $\mathbb{R}^N$. In our paper, we consider frameworks based on the self-expressiveness [6], where a generic point $\mathbf{y}$ from $\{\mathcal{S}_k\}_{k=1}^K$ could be written as a linear combination of the data points

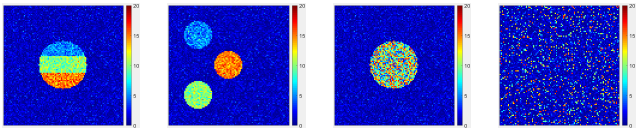
This work was supported by Natural Science Foundation of China under 62001304, 61901116 and 61971476, by Guangdong Basic and Applied Basic Research Foundation under 2019A1515110410 and 2019A1515010789, by Macau Science and Technology Development Fund under SKL-IOTSC-2018-2020, 077/2018/A2, and 0060/2019/A1, by Research Committee at University of Macau under MYRG2018-00029-FST and MYRG2019-00023-FST. (Corresponding author: Jiantao Zhou)

Y. Li is with Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: yuanmanli@szu.edu.cn).

J. Zhou and J. Tian are with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, and also with the Department of Computer and Information Science, University of Macau, Macau 999078. (e-mail: {jtzhou, yb77405}@um.edu.mo)

X. Zheng is with the School of Mathematics and Big Data, Foshan University, Guangdong 528000, China. (e-mail: alex.w.zheng@hotmail.com).

Y. Tang is with the Zhuhai UM Science and Technology Research Center, University of Macau, Macau 999078, and the Faculty of Science and Technology, UOW College Hong Kong, Hong Kong 999077.



(a) (b) (c) (d)

Fig. 1. Images contaminated with the Gaussian noise. They have the same noise elements (hence the same value in terms of MSE). The difference among them is that noise elements in (d) are randomly distributed, while the ones in (a)-(c) present structural patterns.

in $\mathbf{X}$, namely, $\mathbf{y} = \mathbf{X}\mathbf{z}$. Here $\mathbf{z}$ serves as the representation vector and $\mathbf{X}$ is called the dictionary.

Due to the fact that the dimension of a subspace is often smaller than the number of points in it, i.e., $rank(\mathbf{X}_k) < n_k$, the representation of $\mathbf{y}$ over $\mathbf{X}$ is generally not unique. To address the above issue, we can solve the following regularized optimization problem by incorporating some prior knowledge

$$\underset{\mathbf{z}}{\text{argmin}}\ \Phi(\mathbf{z}), \quad \text{s.t.} \quad \mathbf{y} = \mathbf{X}\mathbf{z}, \tag{2}$$

where $\Phi(\cdot)$ represents a certain regularization function. Practically, $\mathbf{y}$ could be corrupted with many types of noise. In the presence of noise, the following optimization problem is usually considered

$$\underset{\mathbf{z}}{\text{argmin}}\ \Phi(\mathbf{z}), \quad \text{s.t.} \quad \mathcal{L}(\mathbf{y} - \mathbf{X}\mathbf{z}) < \epsilon, \tag{3}$$

where $\mathcal{L}(\cdot)$ serves as a certain fidelity function.

In some cases, it is required to jointly calculate the representations of a collection of data points, represented by $\mathbf{Y}$, over the dictionary $\mathbf{X}$ [11]. We can then extend (3) to a matrix form as

$$\underset{\mathbf{Z}}{\text{argmin}}\ \Phi(\mathbf{Z}), \quad \text{s.t.} \quad \mathcal{L}(\mathbf{Y} - \mathbf{X}\mathbf{Z}) < \epsilon. \tag{4}$$

How to design the fidelity function $\mathcal{L}(\cdot)$ and the regularization function $\Phi(\cdot)$ plays an essential role in the robust representation problem. Typically, $\mathcal{L}(\cdot)$ is devised according to the noise behavior. For example, the widely used fidelity functions designed with the mean square error (MSE) criterion assume that the noise is *i.i.d.* Gaussian. On the other hand, the regularization function $\Phi(\cdot)$ can be designed in many forms to recover the subspaces, such as ℓ_1-norm based regularization functions [5], [6], and nuclear norm based regularization functions [11].

B. Related Works

Within the framework for robust SR, the majority of the previous efforts have been made to design an appropriate regularization function $\Phi(\cdot)$. By imposing different priors on the subspace representations, $\Phi(\cdot)$ could be devised from different perspectives. Along this line, sparsity has been extensively studied and employed, leading to the huge success in a wide range of applications [18], [19]. For the tractability, the regularization function is usually set as ℓ_1-norm to enforce sparsity. Wright *et al*. [5] pioneered the works on using a sparse representation-based classifier (SRC) for robust face recognition. Elhamifar and Vidal [6] devised a method named sparse subspace clustering (SSC), offering superior performance in motion segmentation and image clustering tasks. As a powerful variant of the traditional sparsity prior, the structured sparsity has also been incorporated in many works [18]. Low-rankness is another regularization technique investigated extensively, seeking for the lowest rank representation among all the potential candidates. The low-rank representation (LRR) is conducted by setting the regularization function to be the nuclear norm, and has achieved great success in a wide spectrum of tasks, e.g., recovering low-dimensional structures for the subspace clustering [11]. Besides the aforementioned techniques, there are many other regularization strategies, including the graph regularized learning [20], [21], and the combination of multiple regularizations [22]–[24].

Compared with the works on designing the regularization function $\Phi(\cdot)$, the studies on the fidelity function $\mathcal{L}(\cdot)$ are relatively limited. Considering the low complexity and the analytical tractability, most SR methods simply utilized the MSE criterion. As a well-known fact, the MSE criterion is optimal only when the noise obeys the *i.i.d.* Gaussian distribution [25]. Consequently, MSE-based SR frameworks were reported to be fragile to the non-Gaussian noise, especially when the noise distribution is asymmetric and of non-zero mean [26], [27]. Nevertheless, the noise encountered in many practical problems usually does not satisfy the *i.i.d.* assumption nor the Gaussianity. Besides, only the second-order statistics are considered by the MSE criterion, making it fail to exploit sufficient high-order information of the noise signal. As a result, it was observed that MSE-based approaches often cannot offer satisfactory performance in many practical scenarios [28], [29]. To remedy these drawbacks, there are some attempts to replace the traditional ℓ_2 measure with information-theoretic measures, such as Rényi's entropy [30] and correntropy [28], [31]. The resulting frameworks are called information-theoretic learning (ITL), as advocated by [26], [28], [29], [32]. Essentially, the ITL frameworks search for the solution producing the coding residual with minimal information [32], [33]. Different from MSE, ITL does not impose Gaussianity assumption on the noise signal, and can capture its higher-order statistical information [29]. It was shown that ITL-based approaches achieved promising performance when handling the non-Gaussian noise [19], [29], [30], [33]. Additional details regarding the ITL can be found in [26].

It should be emphasized that, despite the desirable properties, *all* the existing ITL-based approaches still relied on the *i.i.d.* assumption on the noise. Though the *i.i.d.* assumption could significantly simplify the resulting problem, it may not hold practically. In reality, the noise usually exhibits certain structures, and hence different areas could present totally different statistical behaviors. In other words, a single distribution may not be sufficient to fully characterize the noise behavior, making the resultant information measures inaccurate and consequently degrading the performance of the developed algorithms. Fig. 1 gives a persuasive example, where four images are contaminated with the same noise elements of totally different arrangements. If we arbitrarily model the noise signals with an *i.i.d.* distribution, then the noises in Figs. 1(a)-(d) would obey the same distribution,

thus have the same amount of entropy [30]. Obviously, such an *i.i.d.* model results in inaccurate information estimations. According to the information theory [34], the noise in Fig. 1(d) is of the highest randomness, thus should have a much larger amount of information than the others. Based on the aforementioned phenomenon, to improve the robustness of SR, it requires to design more generic noise models with more powerful capabilities in describing the complex behavior of the underlying noise in many practical scenarios.

C. Our Contributions

In this paper, we devise a novel ITL framework for robust SR via a generic noise model, i.e., independent and piecewise identically distributed (*i.p.i.d.*) model. Different from traditional algorithms that assume the noise originates from a single $i.i.d.$ source, our framework describes the underlying noise using multiple distributions. We summarize our major contributions as below:

1) An *i.p.i.d.* noise model is proposed, assisted by which, we devise a novel minimum weighted error entropy (MWEE) criterion for the robust SR. We show that our MWEE criterion is effective to exploit the inherent statistical information of both the structural noise and the purely random one.
2) A robust SR framework is developed based on the MWEE criterion, which does not impose the *i.i.d.* assumption or the Gaussianity on the underlying noise. This enables the proposed framework to satisfactorily handle various kinds of noise in practical scenarios.
3) We establish the connections of the MWEE criterion with the existing ITL-based ones (i.e., criteria based on Rényi's entropy and correntropy), further showing the potential advantages of the MWEE-based frameworks.
4) As a general technique, the MWEE criterion proposed in this work could be readily extended to other learning frameworks, thus potentially improving their robustness.

This paper is the extension of our earlier conference version [12]. In comparison to the conference paper, both the technical and experimental parts have been substantially refined. We summarize the primary improvements as follows. First, [12] is tailored for the subspace clustering only, while in this work, we extend our framework to the problem of subspace representation, and study its applications to both the subspace clustering (Section IV.A) and the image recognition (Section IV.B). Second, a new Section III.B is inserted to discuss the interpretability of our MWEE criterion, and establish its connections with the traditional ITL criteria, further showing its generalization capability. Third, we redefine the entropy of our proposed *i.p.i.d.* source in Definition 2 based on the weighted average *information potential* (IP) [26], which greatly simplifies the related minimization problem. Fourth, we also provide more properties and theoretical proofs, such as Theorem 2 and Theorem 3, which make our model more technically solid. Last but not least, we insert a new Section VI, where extensive additional experiments are conducted to demonstrate the superiority of the proposed technique in the image recognition tasks.

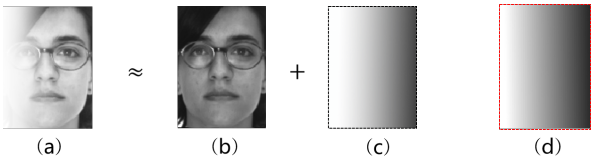


Fig. 2. An example of the illumination noise. (a)-(d) are the observed face, latent face, illumination noise, and synthetic noise generated by an *i.p.i.d.* source, respectively.

The remaining of this paper is organized as follows. We introduce the *i.p.i.d.* noise model in Section II, and then present the MWEE-based framework for robust SR, together with its optimization in Section III. In Section IV, we discuss how to apply our representation framework to deal with two real-world problems. Section V and Section VI conduct extensive experiments to validate the superiority of our scheme. Finally, Section VII concludes.

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

As a core part of this work, we focus on designing a new minimization criterion for robust representation. Motivated by the promising results of ITL-based approaches to deal with non-Gaussian noise, we propose a novel ITL-type fidelity function via a generic noise model. Specifically, the ITL-based approaches focus on finding a representation producing the coding residual with the minimum information [19], [33]. Under the ITL criterion, the representation problem defined in (3) can be specialized as

$$\underset{\mathbf{z}}{\operatorname{argmin}}\ \Phi(\mathbf{z}),\quad \text{s.t.}\ \ H(\mathbf{e}) \le \epsilon,\ \mathbf{y}-\mathbf{X}\mathbf{z}=\mathbf{e}, \tag{5}$$

where $H(\cdot)$ is a potential information measure, e.g., correntropy [19] or Rényi's entropy [30], [33]. Though ITL has many desirable properties, *all* the previous ITL-based approaches relied on the *i.i.d.* noise assumption. Unfortunately, in reality, the noise is generally signal-dependent, and hence does not obey the *i.i.d.* property. In our work, we propose an *i.p.i.d.* source for the noise modeling. It shows that our proposed *i.p.i.d.* source has superior advantages in characterizing the inherent statistical behavior of both the structural noise and the purely random one.

A. Definition of the i.p.i.d. Source and Its Properties

The 1-D *i.p.i.d.* source is defined as below.

Definition 1: Denote $\mathbf{x}=[x_1,...,x_N]\in\mathbb{R}^N$ as a sequence of N samples, and $\{\mathcal{P}_i\}_{i=1}^L$ as a non-overlapping, sequential partition of the index vector $[1,2,\cdots,N]$, which is given by

$$\mathcal{P}_i=[p_{i-1}+1,p_{i-1}+2,\cdots,p_i],\ i\in\{1,\cdots,L\}, \tag{6}$$

with $p_0=0$, $p_i<p_{i+1}$, and $p_L=N$. We say that $\mathbf{x}$ is generated by an *i.p.i.d.* source, if there exists a collection of probability density functions $\{f_i\}_{i=1}^L$, such that the subsequence of samples determined by $\mathcal{P}_i$ is independently generated according to f_i. Namely, for each $i\in\{1,2,\cdots,L\}$, we have

$$x_{p_{i-1}+1},x_{p_{i-1}+2},\cdots,x_{p_i}\overset{i.i.d.}{\sim} f_i. \tag{7}$$

We can readily extend the above definition of 1-D source to higher dimensional signals, such as images and videos. It is worth noting that a slightly similar definition was also adopted for coding the binary source [35]. We summarize some properties of the *i.p.i.d.* source as follows:

- **Locality**: By resorting to the non-overlapping, sequential partition, the local behavior of a certain signal can be fully exploited by the *i.p.i.d.* source. This is totally different from the traditional *i.i.d.* one, where the structural information is completely lost.
- **Fine-description**: Rather than using a single distribution, the *i.p.i.d.* source employs multiple density functions to characterize one signal. This makes it more powerful to characterize complex signals.
- **Generalization**: The *i.p.i.d.* source remains the ability to describe purely random signals, by considering the fact that the traditional *i.i.d.* source can be regarded as a special case with $L = 1$.

The above desirable properties endow the *i.p.i.d.* source with superior descriptive capability to characterize both purely random signals (e.g., Fig. 1(d)) and structural ones (e.g., Figs. 1(a)-(c)). As an example, we can satisfactorily model the noise in Fig. 1(b) using an *i.p.i.d.* source, with the blue, red and green circular areas, and the background originating from four different distributions. In Fig. 2, we give a more illustrative example with the unconstrained illumination noise. Apparently, such noise is highly structural, and hence one cannot appropriately model it using any $i.i.d.$ sources (e.g., Gaussian, Laplacian or MoG), since it is almost impossible to generate a similar noise using a certain $i.i.d.$ distribution. However, assisted by its desirable properties, the *i.p.i.d.* source is able to model the illumination noise satisfactorily. For clarification, Fig. 2(d) presents a synthetic signal generated by an *i.p.i.d.* source, where the elements in each disjoint 8×8 region obey an *i.i.d.* Gaussian distribution, whose mean is gradually decreased by a factor 0.5 from left to right. It can be readily observed that the synthetic signal well reflects the noise behavior in Fig. 2(c). Besides 2-D signals, 1-D signals (e.g., noise in a voice) and multi-D signals (e.g., noise in a video) can also be described by *i.p.i.d.* sources in a similar way.

B. Rényi's Entropy of the i.p.i.d. Source

Estimating the information of a signal is crucial for our proposed ITL-based SR framework. Without loss of generality, Rényi's entropy is employed in this work. Denote E as a random variable. Its Rényi's entropy of order α ($\alpha > 0, \alpha \neq 1$) takes the following form

$$H_\alpha(E) = \frac{1}{1-\alpha}\log\Big(\int (f_E(e))^\alpha de\Big). \tag{8}$$

The argument of the logarithm, $\int (f_E(e))^\alpha de$, is called the order-α *information potential* (IP) [26]. In many practical scenarios, the probability density function (PDF) $f_E(e)$ is unknown and only finite samples $\{e_i\}_{i=1}^N$ are available. Then we can use the Parzen window method to approximate $f_E(e)$, which is given by

$$\hat{f}_E(e) = \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma(e - e_i), \tag{9}$$

where $\kappa_\sigma(\cdot)$ is a Gaussian kernel with the parameter σ

$$\kappa_\sigma(e - e_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(e-e_i)^2}{2\sigma^2}\Big). \tag{10}$$

It should be emphasized that a fundamental assumption of the Parzen window method is that the samples $\{e_i\}_{i=1}^N$ satisfy the *i.i.d.* property.

In this work, we suggest a new entropy estimator for the *i.p.i.d.* source. To distinguish from the traditional Rényi's entropy estimator $H_\alpha(E)$ with the *i.i.d.* constraint, we refer to the new estimator as the *piecewise Rényi's entropy* (PRE). Note that the traditional entropy reflects the required minimum average number of bits to encode a sequence of *i.i.d.* symbols. With the same rule, we define the PRE as below:

Definition 2: Assume that the sequence $\mathbf{e} = [e_1, e_2, .., e_N]$ is generated by an *i.p.i.d.* source with the index partition $\{\mathcal{P}_q\}_{q=1}^L$. Then we define the PRE of $\mathbf{e}$ as

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

where $f_{E_q}(e)$ is the PDF estimated using the samples indexed by $\mathcal{P}_q$, and we have

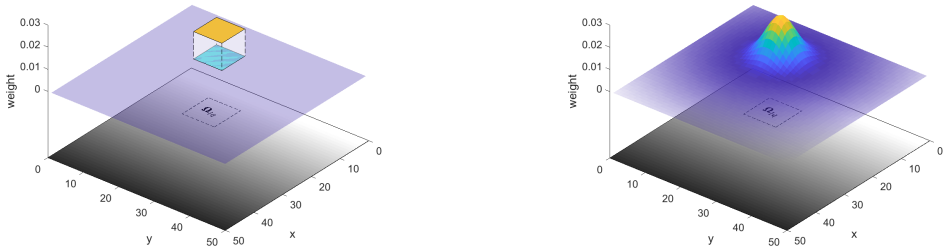
$$f_{E_q}(e) = \frac{1}{|\mathcal{P}_q|}\sum_{i\in\mathcal{P}_q} \kappa_\sigma(e - e_i). \tag{12}$$

Compared with the traditional entropy estimator $H_\alpha(\mathbf{e})$, PRE adopts a weighted average IP over different partitions. Since the samples indexed by $\mathcal{P}_q$ are *i.i.d.*, we approximate $f_{E_q}(e)$ using Parzen window estimation as shown in (12). Apparently, with the *Generalization* property of the *i.p.i.d.* source, PRE reduces to the traditional Rényi's entropy $H_\alpha(\mathbf{e})$ if the sequence $\mathbf{e}$ is actually *i.i.d.*; otherwise if it is *i.p.i.d.*, PRE can more precisely exploit its information. For simplicity, we set $\alpha = 2$, corresponding to the second-order statistics of the Rényi's entropy. The resulting estimator $\hat{H}_2(\mathbf{e})$ is then referred to as the second-order PRE.

We should note that directly computing $\hat{H}_2(\mathbf{e})$ is not straightforward, since the partition $\{\mathcal{P}_q\}_{q=1}^L$ is generally unknown in reality; and obtaining the partition is believed to be considerably difficult, if possible. This is typically true for those signals with complex or gradually varied patterns, such as the noise signal shown in Fig. 2(c). Fortunately, with the locality property of the *i.p.i.d.* source, it is reasonable to assume that samples in a sufficiently small local area are *i.i.d.* signals. Even without explicitly knowing $\{\mathcal{P}_q\}_{q=1}^L$, such a property permits us to first estimate the PDF for each small local region using the Parzen window method, and then approximate the PRE $\hat{H}_2(\mathbf{e})$ by averaging the results over all the local areas according to Definition 2.

Denote I_q as the location of e_q in the original data space [1]. For each location I_q, we construct a local region $\mathbf{\Omega}_{Iq}$, and the density function for $\mathbf{\Omega}_{Iq}$ then can be estimated as

[1] I_q is a scalar x for the 1-D data (e.g., voice), while it is a 2-D location (x, y) for the 2-D data (e.g., image).



(a) (b)

Fig. 3. An illustration of two types of weighting functions. (a) the weights of $f_{Iq}(e)$, which are zeros outside $\mathbf{\Omega}_{Iq}$, and (b) the weights of $\hat{f}_{I_q}(e)$, which are all non-zeros and decrease w.r.t. the distance to I_q.

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

We can see that (12) and (13) are equivalent if $\mathbf{\Omega}_{Iq}$ happens to coincide with $\mathcal{P}_q$. For simplicity, we write $f_{E_{Iq}}(e)$ as $f_{I_q}(e)$ in the sequel.

However, directly using (13) may lead to inaccurate estimation. This is because that in a small region, the number of samples is often insufficient for the density estimation. To address this issue, we introduce a weighting function, which allows us to use all the samples in $\mathbf{e}$ for estimating $f_{I_q}(e)$. A similar strategy was also proposed in [36] to approximate the density by using a few samples. When estimating $f_{I_q}(e)$, the weight of each sample is assigned based on its location distance from I_q, to preserve the locality property. Define $Dis(\cdot)$ as a certain distance function (e.g., ℓ_2-norm), to measure the distance of the location I_i from I_q in the data space. We write

$$D_{q,i} = Dis(I_q, I_i). \tag{14}$$

The PDF for $\mathbf{\Omega}_{Iq}$ is then approximated by

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

where $c(\cdot)$ is a weighting function related to the location distance between two samples. In our work, we refer to (15) as the weighted Parzen window (WPW) estimation.

As a weighting function, $c(\cdot)$ needs to satisfy the following conditions:

1) $c(\cdot) \geq 0$;
2) $\sum_{i=1}^{N} c(D_{q,i}) = 1$ for any I_q;
3) $c(x)$ is an even function and decreases w.r.t. $|x|$.

The first property eliminates the negative contribution of samples for the density estimation. The second property ensures that $\hat{f}_{I_q}(e)$ defined in (15) is a distribution. The third property indicates that the samples closer to I_q contribute more to the estimation of $\hat{f}_{I_q}(e)$. Note that WPW defined in (15) is a generalization of (12) and (13). Typically, (15) reduces to (13) when

$$c(D_{q,i}) = \begin{cases} \frac{1}{|\mathbf{\Omega}_{Iq}|} & \text{if } \ I_i \in \mathbf{\Omega}_{Iq}, \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, $c(\cdot)$ is chosen as a Gaussian-type function. Specifically,

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

where Q is a normalization term such that $\sum_{I_i} c(D_{q,i}) = 1$, and σ_w^2 is empirically set as $\frac{N}{1000}$. Fig. 3(a) and Fig. 3(b) illustrate the weights assigned to samples when calculating $f_{Iq}(e)$ and $\hat{f}_{Iq}(e)$, respectively.

Upon obtaining $\hat{f}_{Iq}(e)$ for each $\mathbf{\Omega}_{Iq}$, we can approximate the PRE by taking the average over all the locations. Denote the estimator as $\bar{H}_2(\mathbf{e})$. Mathematically,

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

Note that the integral of the product of two Gaussians equals the value of the Gaussian computed at the difference of the arguments, whose variance is the sum of the original two variances [26]. We have

$$\int \kappa_\sigma(e - e_i) \kappa_\sigma(e - e_j) de = \kappa_{\sqrt{2}\sigma}(e_i - e_j).$$

Finally, we obtain

$$\bar{H}_2(\mathbf{e}) = -\log \frac{1}{N} \sum_{I_q} \sum_{i,j=1}^{N} c(D_{q,i}) c(D_{q,j}) \kappa_{\sqrt{2}\sigma}(e_i - e_j). \tag{17}$$

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

Different from $\hat{H}_2(\mathbf{e})$ given in (11), the estimator $\bar{H}_2(\mathbf{e})$ does not require the partition $\{\mathcal{P}_q\}_{q=1}^{L}$ explicitly. In reality, the data could also be corrupted by *i.i.d.* noises, as will be demonstrated in Section VI-B. It is therefore important to show that $\bar{H}_2(\mathbf{e})$ is still effective in the *i.i.d.* case. Fortunately, with

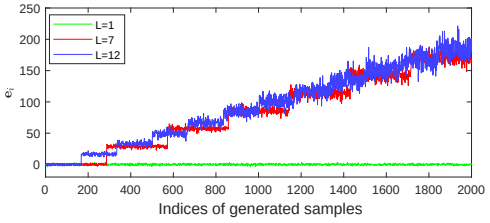


Fig. 4. The *i.p.i.d.* sequences with $L = 1$, $L = 7$ and $L = 12$.

the conditions of the weighting function $c(\cdot)$, the following theorem proves that the PRE estimator $\bar{H}_2(\mathbf{e})$ and the traditional Rényi's entropy $H_2(\mathbf{e})$ are equivalent under the *i.i.d.* case.

Theorem 1: Denote $\mathbf{e} = [e_1, e_2, .., e_N]$ as a sequence of symbols independently sampling from the same distribution $f(e)$. The approximation of the PRE estimator $\bar{H}_2(\mathbf{e})$ given in (17) and the traditional second-order Rényi's entropy $H_2(\mathbf{e})$ are equivalent.

We give the proof of Theorem 1 in the supplementary material. Apparently, the above theorem provides a fundamental theoretical basis for the capability of $\bar{H}_2(\mathbf{e})$ to characterize the *i.i.d.* noise.

In this following, we present a toy example to visualize the relationship among these three entropy estimators. Specifically, we first generate a sequence $\mathbf{e} = [e_1, e_2, ..., e_N]$ $(N = 2000)$ by an *i.p.i.d.* source with L piecewise partitions, where samples in the q-th partition are independently generated by the following Gaussian distribution

$$f_{E_q}(e) = \frac{1}{\sqrt{2\pi q^2}} \exp\left(-\frac{(e-\mu)^2}{2q^2}\right). \tag{18}$$

We set $\mu = \frac{200}{L}(q-1)$, and set the number of samples in each partition as $\lfloor N/L \rfloor$. Obviously, $L = 1$ corresponds to the case that the sequence $\mathbf{e}$ is $i.i.d.$, and the generated samples obey the standard normal distribution; while in the case of $L > 1$, the sequence becomes a 1-D structural non-$i.i.d.$ signal. Fig. 4 shows three examples, where the sequences are generated when $L = 1$ (green), $L = 7$ (red) and $L = 12$ (blue).

Then, we use three different estimators to estimate the information of generated sequences with various number of piecewise partitions, and the results are plotted in Fig. 5. It can be readily observed that $H_2(\mathbf{e})$, $\hat{H}_2(\mathbf{e})$ and $\bar{H}_2(\mathbf{e})$ are equivalent in the *i.i.d.* case $(L = 1)$, which coincides with Theorem 1. However, in the case of $L > 1$, $H_2(\mathbf{e})$ cannot well estimate the inherent information of the signal, and it becomes much larger than the other estimators. This phenomenon is unsurprising since $H_2(\mathbf{e})$ naively treats $\mathbf{e}$ as an *i.i.d.* signal, thus totally ignoring its structural information. Even without explicitly knowing the partition, we can observe from Fig. 5 that $\bar{H}_2(\mathbf{e})$ can still well approximate $\hat{H}_2(\mathbf{e})$. This demonstrates that $\bar{H}_2(\mathbf{e})$ is effective to estimate the inherent information of both the *i.i.d.* signals and the non-*i.i.d.* ones.

III. Robust SR under the MWEE Criterion

In light of ITL [26], [28], [32], we propose to design the fidelity function $\mathcal{L}(\cdot)$ such that the solution produces the

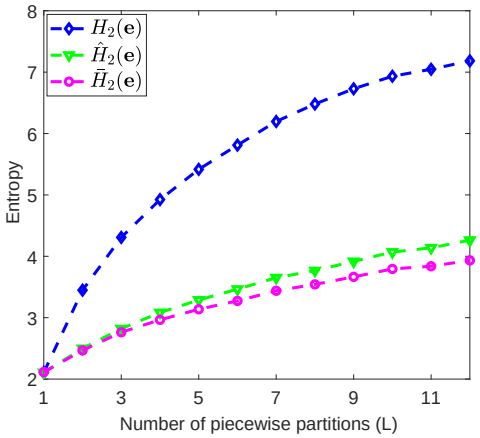


Fig. 5. Entropies of three estimators with different number of piecewise partitions (L). The signal is $i.i.d.$ when $L = 1$, while non-$i.i.d.$ when $L > 1$. $H_2(\mathbf{e})$, $\hat{H}_2(\mathbf{e})$ and $\bar{H}_2(\mathbf{e})$ are defined in (8) (with $\alpha = 2$), (11) and (17), respectively.

coding residual with the minimal information measured by the PRE $\bar{H}_2(\mathbf{e})$. Then the framework for robust SR defined in (5) can be specialized as

$$\underset{\mathbf{z}}{\text{argmin}} \ \ \Phi(\mathbf{z}), \quad \text{s.t.} \ \ \bar{H}_2(\mathbf{e}) \le \epsilon, \ \mathbf{y} - \mathbf{X}\mathbf{z} = \mathbf{e}. \tag{19}$$

Due to the weighted nature of the WPW estimation, we name the criterion of finding the solution with minimum $\bar{H}_2(\mathbf{e})$ as the minimum weighted error entropy (MWEE). It is worth noting that the proposed MWEE criterion designed with the *i.p.i.d.* model does not impose the *i.i.d.* assumption or the Gaussianity assumption on the noise, and our minimization target PRE could fully exploit the inherent information of both the *i.i.d.* noise and the non-*i.i.d.* one. This makes it fundamentally different from the traditional ITL and MSE criteria. As will be shown in the experimental stage, methods based on the MWEE criterion exhibit superior advantages to handle different kinds of noise.

A. Algorithm of the MWEE-based SR

Define

$$c^q_{i,j} = c(D_{q,i})c(D_{q,j}), \tag{20}$$

where $D_{q,i}$ is the distance measure defined in (14). For simplicity, we replace the notation $\sqrt{2}\sigma$ by σ in the definition of $\bar{H}_2(\mathbf{e})$ hereafter. We can rewrite $\bar{H}_2(\mathbf{e})$ as

$$\begin{aligned} \bar{H}_2(\mathbf{e}) &= -\log \frac{1}{N} \sum_{I_q} \sum_{i,j=1}^{N} c^q_{i,j} \kappa_\sigma(e_i - e_j) \\ &= -\log \sum_{i,j=1}^{N} w_{i,j} \kappa_\sigma(e_i - e_j) + \log N \\ &= -\log S(\mathbf{e}) + \log N, \end{aligned} \tag{21}$$

where

$$S(\mathbf{e}) = \sum_{i,j=1}^{N} w_{i,j} \kappa_\sigma(e_i - e_j), \tag{22}$$

and

$$w_{i,j} = \sum_{I_q} c_{i,j}^q. \tag{23}$$

Note that $w_{i,j}$ in (23) only depends on the locations of e_i and e_j in the data space, while not on their specific values.

Due to the fact that $\tilde{H}_2(\mathbf{e})$ is a monotonic decreasing function with respect to $S(\mathbf{e})$, we can minimize $-S(\mathbf{e})$ as a replacement of minimizing $\tilde{H}_2(\mathbf{e})$. Then the optimization problem (19) can be rewritten as

$$\underset{\mathbf{z}}{\operatorname{argmin}} - S(\mathbf{y} - \mathbf{X}\mathbf{z}) + \lambda\Phi(\mathbf{z}), \tag{24}$$

where λ is the Lagrange multiplier. Furthermore, let y_i represent the i-th entry of $\mathbf{y}$ and $\mathbf{a}_i$ denote the i-th row of $\mathbf{X}$, respectively. Then

$$\begin{aligned} S(\mathbf{y} - \mathbf{X}\mathbf{z}) &= \sum_{i,j=1}^{N} w_{i,j}\kappa_\sigma(y_i - y_j - (\mathbf{a}_i\mathbf{z} - \mathbf{a}_j\mathbf{z})) \\ &= \sum_{h=1}^{N^2} w_h\kappa_\sigma(\hat{y}_h - \hat{\mathbf{a}}_h\mathbf{z}), \end{aligned} \tag{25}$$

where $\hat{y}_h = y_i - y_j$, $\hat{\mathbf{a}}_h = \mathbf{a}_i - \mathbf{a}_j$ and $h = (i-1)N + j$. Let $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{N^2}]^T$, whose dimension is the square of the dimension of $\mathbf{y}$. The high dimension expansion makes it difficult to solve the problem (24). To tackle this issue, we approximate $S(\mathbf{e})$ in a similar way suggested in [33], which can be written as

$$\tilde{S}(\mathbf{e}) = \sum_{i=1}^{N} \kappa_\sigma \left(\sum_{j=1}^{N} w_{i,j}(e_i - e_j) \right). \tag{26}$$

We can observe that $S(\mathbf{e})$ and $\tilde{S}(\mathbf{e})$ have identical minimizers $\mathbf{e} = c\mathbf{1}$ for some constant c. Furthermore, Theorem 2 given below shows that $\tilde{S}(\mathbf{e})$ approaches $S(\mathbf{e})$ as σ gets large.

Theorem 2: For a signal $\mathbf{e} = [e_1, e_2, .., e_N]$ generated from the distribution $f(e)$, the gap between $\tilde{S}(\mathbf{e})$ and $S(\mathbf{e})$ is upper bounded by $M\sigma^{-3}$, where M is a constant.

The proof of Theorem 2 is given in the supplementary material. By replacing $S(\mathbf{y} - \mathbf{X}\mathbf{z})$ with $\tilde{S}(\mathbf{y} - \mathbf{X}\mathbf{z})$, we relax the problem (24) as the following MWEE-based problem

$$\underset{\mathbf{z}}{\operatorname{argmin}} - \tilde{S}(\mathbf{y} - \mathbf{X}\mathbf{z}) + \lambda\Phi(\mathbf{z}). \tag{27}$$

We further write

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

where the second equation holds due to the fact that $\sum_j w_{i,j} = 1$ (proof to be given in Theorem 3). Finally, we can rewrite (27) as

$$\underset{\mathbf{z}}{\operatorname{argmin}} - \sum_{i=1}^{N} \kappa_\sigma(\tilde{y}_i - \tilde{\mathbf{x}}_i\mathbf{z}) + \lambda\Phi(\mathbf{z}), \tag{28}$$

Algorithm 1 Half-quadratic optimization for MWEE-based SR

Input: The data matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$, a sample $\mathbf{y}$, the parameter λ, and $t = 0$.

1: Calculate $\tilde{\mathbf{y}} = [\tilde{y}_1, ..., \tilde{y}_N]^T$ and $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1^T, ..., \tilde{\mathbf{x}}_N^T]^T$ according to (29).
2: **Do** until convergence
3: $\quad u_i^{t+1} = \frac{1}{\sigma^2}\kappa_\sigma(\tilde{y}_i - \tilde{\mathbf{x}}_i\mathbf{z}^t),\ i = 1, 2, \cdots, N$
4: $\quad \mathbf{z}^{t+1} = \underset{\mathbf{z}}{\operatorname{argmin}}\ (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{z})^T \operatorname{diag}(\mathbf{u}^{t+1})(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{z}) + \lambda\Phi(\mathbf{z})$
5: $\quad t = t + 1$
6: **end**

Output: $\mathbf{z}$.

where

$$\tilde{y}_i = y_i - \sum_{j=1}^{N} w_{i,j}y_j,\ \tilde{\mathbf{x}}_i = \mathbf{a}_i - \sum_{j=1}^{N} w_{i,j}\mathbf{a}_j. \tag{29}$$

We can solve the above problem (28) by using the half-quadratic optimization theory [37], which is widely used for the ITL-based optimization problems [38]. Algorithm 1 summarizes the optimization procedures of the MWEE-based SR (28), where the detailed derivations are given in the supplementary material.

Remark: As shown in the problem (4), in some cases, we need to compute the representations of a collection of data points jointly. The fidelity function $\mathcal{L}(\cdot)$ can then be similarly defined as

$$\mathcal{L}(\mathbf{Y} - \mathbf{X}\mathbf{Z}) = -\sum_{i=1}^{n} \tilde{S}(\mathbf{y}_i - \mathbf{X}\mathbf{z}_i), \tag{30}$$

where $\mathbf{y}_i$ and $\mathbf{z}_i$ represent the i-th columns of $\mathbf{Y}$ and $\mathbf{Z}$, respectively. Then we can solve the resulting problem in a similar way as Algorithm 1.

B. Connection with the Traditional ITL Criterion

We first show that the MWEE criterion to minimize $\tilde{H}_2(\mathbf{e})$ is a generalization of the ITL criterion to minimize the traditional Rényi's entropy $H_2(\mathbf{e})$ [30], [32], [33]. More specifically, with some simple derivations, we can reformulate $H_2(\mathbf{e})$ of an *i.i.d.* source defined by (8) and (9) as

$$\begin{aligned} H_2(\mathbf{e}) &= -\log\frac{1}{N^2}\sum_{i,j=1}^{N} \kappa_{\sqrt{2}\sigma}(e_i - e_j) \\ &= -\log\sum_{i,j=1}^{N} \frac{1}{N}\kappa_{\sqrt{2}\sigma}(e_i - e_j) + \log N. \end{aligned} \tag{31}$$

Compared (31) with our minimization target $\tilde{H}_2(\mathbf{e})$ defined in (21), we can observe the difference is that $H_2(\mathbf{e})$ adopts a uniform weight $\frac{1}{N}$ for each term $e_i - e_j$, while $\tilde{H}_2(\mathbf{e})$ uses an adaptive weight $w_{i,j}$ instead. As demonstrated previously, they are equivalent in the *i.i.d.* scenarios. We show that $w_{i,j}$ has many desirable properties, which makes our MWEE criterion highly interpretable.

Theorem 3: With the conditions of the weighting function $c(x)$ listed in Section II-B, $w_{i,j}$ defined in (23) holds similar properties as

1) $w_{i,j} \geq 0$;
2) $\sum_i w_{i,j} = 1$ for any j;
3) $w_{i,j} = w_{j,i}$, and $w_{i,j}$ is decreasing with respect to the distance between I_i and I_j.

The proofs of these properties are provided in the supplementary material. Apparently, $w_{i,j}$ inherits the first two properties of the weighting function $c(\cdot)$ listed in Section II-B. Besides, We surprisingly find that $w_{i,j}$ holds a similar property as $c(x)$, in the sense that $w_{i,j}$ is still symmetric, and its value gets larger with the decreasing distance between I_i and I_j. As a result, the term $e_i - e_j$ would contribute more for $\tilde{H}_2(\mathbf{e})$ when their location distance is smaller. This property exhibits the capability of $\tilde{H}_2(\mathbf{e})$ to exploit the local behavior of signals. With these properties regarding $w_{i,j}$, the minimization target $\tilde{H}_2(\mathbf{e})$ is of high interpretability to characterize the *i.p.i.d.* sources.

As a special case, we prove that when $c(\cdot)$ is selected to be a Gaussian function as shown in (16), then $w_{i,j}$ is still Gaussian-type.

Corollary 1: Let $c(D_{q,i}) = \frac{1}{Q} e^{-\frac{(D_{q,i})^2}{\sigma_w^2}}$, where Q is a normalization term such that $\sum_{I_i} c(D_{q,i}) = 1$. Then $w_{i,j}$ is given by

$$w_{i,j} = \frac{e^{-\frac{(D_{i,j})^2}{2\sigma_w^2}}}{\sum_i e^{-\frac{(D_{i,j})^2}{2\sigma_w^2}}}. \tag{32}$$

The proof is given in the supplementary material. Assisted by Corollary 1, we can set $w_{i,j}$ according to (32) directly without the tedious computation of (23).

Furthermore, the MWEE criterion also has a strong connection with the Correntropy Induced Metric (*CIM*) [28], which has been widely adopted to handle non-Gaussian noises with large outliers. Given any two vectors $\mathbf{a} = (a_1, ..., a_N)$, and $\mathbf{b} = (b_1, ..., b_N)$, *CIM* aiming to measure their similarity is defined as

$$CIM(\mathbf{a}, \mathbf{b}) = \Big(\kappa_\sigma(0) - \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma(e_i)\Big)^{\frac{1}{2}}, \tag{33}$$

where $\mathbf{e} = (e_1, .., e_N)$ with $e_i = a_i - b_i$. It can be observed that when e_i's are very large, *CIM* becomes close to 1, which is much smaller than the mean absolute error $\frac{1}{N}||\mathbf{a}-\mathbf{b}||_1$ and the mean squared error $\frac{1}{N}||\mathbf{a}-\mathbf{b}||_2^2$. This endows the capability of *CIM* to suppress the effect of large outliers. Essentially, *CIM* is a local metric, which behaves like the ℓ_2-norm when e_i is small, exhibits similar effects to the ℓ_1-norm when e_i becomes larger, and eventually approaches the ℓ_0-norm when e_i is very large [28]. In fact, our framework can be regarded as an extension of *CIM* induced frameworks. By setting all $w_{i,j}$'s in (29) to 0, our framework (28) would become those *CIM* based frameworks such as [39], [40]. However, compared with *CIM*, which treats the entries of $\mathbf{e}$ independently and totally ignores their correlations, the MWEE criterion designed with our generic noise model is more powerful to characterize the behaviors of those non-*i.i.d.* signals.

Algorithm 2 MWEE-based Subspace Clustering (MWEE-S)

Input: A matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ to be clustered, the parameter λ and the number of subspaces K.
1: Normalize $\mathbf{X}$ so that each column has unit l_2 norm.
2: Solve the problem (35) to deal with the linear subspace, or the problem (39) to address the affine subspace.
3: Compute the similarity matrix $\mathbf{M} = |\mathbf{Z}| + |\mathbf{Z}|^T$.
4: Apply the spectral clustering algorithm [41] to the similarity matrix $\mathbf{M}$.
Output: K clusters.

Algorithm 3 MWEE-based Classifier (MWEE-C)

Input: A matrix of training samples $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$, the parameter λ, and a test sample $\mathbf{y}$.
1: Normalize $\mathbf{X}$ so that each column has unit l_2 norm.
2: Solve the MWEE-based problem (40) according to Algorithm 1
3: Compute the reconstruction residuals (41).
4: Predict Identity($\mathbf{y}$) $= \underset{k}{\operatorname{argmin}}\ R_k(\mathbf{y})$.
Output: Identity($\mathbf{y}$).

IV. Applications to High-level Vision Tasks

We now discuss how to apply our proposed SR framework designed under the MWEE criterion to real applications. Specifically, two fundamental problems in computer vision, i.e., subspace clustering and image recognition are considered in this work.

A. MWEE-based Subspace Clustering

Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ be a set of points drawn from a collection of K linear subspaces $\{\mathcal{S}_i\}_{i=1}^K$. Subspace clustering aims to correctly separate the points according to their underlying subspaces. Specifically, we only consider the spectral clustering based approaches [6], [11] in this work. Such methods divide the subspace clustering into two phases: i) learn an affinity matrix from the data; and ii) separate the data by applying the spectral clustering [41] to this affinity matrix.

The sparse subspace clustering (SSC) designed with the MSE criterion is a popular subspace clustering technique using the self-expressiveness property, which suggested to compute the representation coefficients by solving the following optimization problem

$$\underset{\mathbf{Z}}{\operatorname{argmin}}\ ||\mathbf{X} - \mathbf{X}\mathbf{Z}||_F^2 + \lambda||\mathbf{Z}||_1, \quad \text{s.t.}\ \ \operatorname{diag}(\mathbf{Z}) = \mathbf{0}. \tag{34}$$

The constraint $\operatorname{diag}(\mathbf{Z}) = \mathbf{0}$ can eliminate the trivial solution to represent a point by itself. According to our proposed SR framework, we redesign the problem (34) under the MWEE criterion. As a result, we have

$$\underset{\mathbf{Z}}{\operatorname{argmin}}\ \mathbf{\Psi}(\mathbf{X} - \mathbf{X}\mathbf{Z}) + \lambda||\mathbf{Z}||_1, \quad \text{s.t.}\ \ \operatorname{diag}(\mathbf{Z}) = \mathbf{0}, \tag{35}$$

where

$$\mathbf{\Psi}(\mathbf{X} - \mathbf{X}\mathbf{Z}) \triangleq -\sum_{i=1}^{n} \tilde{S}(\mathbf{x}_i - \mathbf{X}\mathbf{z}_i). \tag{36}$$

Fig. 6. Illustration of face clustering, which aims to cluster multiple images according to their subjects.

Here $\tilde{S}(\cdot)$ is defined in (26) reflecting the MWEE criterion. We can further decompose the problem (35) into n independent subproblems, where the i-th subproblem is

$$\underset{\mathbf{z}_i \in \mathbb{R}^n}{\text{argmin}} \; -\tilde{S}(\mathbf{x}_i - \mathbf{X}\mathbf{z}_i) + \lambda ||\mathbf{z}_i||_1, \quad \text{s.t.} \;\; \mathbf{z}_{i,i} = 0, \tag{37}$$

where $\mathbf{z}_i$ corresponds to the i-th column of $\mathbf{Z}$. To deal with the above problem, we first solve the problem

$$\underset{\mathbf{z}'_i \in \mathbb{R}^{n-1}}{\text{argmin}} \; -\tilde{S}(\mathbf{x}_i - \hat{\mathbf{X}}\mathbf{z}'_i) + \lambda ||\mathbf{z}'_i||_1. \tag{38}$$

Here $\hat{\mathbf{X}} = [\mathbf{x}_1, .., \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, .., \mathbf{x}_n]$. Note that (38) has the same form as the problem (27), which can be directly solved using Algorithm 1 presented in Section III-A. Upon obtaining $\mathbf{z}'_i$, we compute $\mathbf{z}_i$ by

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

Note that rather than the linear subspaces, data could lie in a union of affine in some real-world problems. For example, the motion segmentation task aims to cluster the data distributed in a collection of 3D affine subspaces [6]. In our work, we consider the framework proposed in [6] to address affine subspaces, which can be written as

$$\begin{aligned} &\underset{\mathbf{Z}}{\text{argmin}} \; \mathbf{\Psi}(\mathbf{X} - \mathbf{X}\mathbf{Z}) + \lambda ||\mathbf{Z}||_1, \\ &\text{s.t.} \;\; \mathbf{Z}^T\mathbf{1} = \mathbf{1}, \text{diag}(\mathbf{Z}) = \mathbf{0}. \end{aligned} \tag{39}$$

Compared with the problem (35), the above optimization problem for the case of affine subspaces introduces additional linear equality constraints. We can solve (39) using the alternating direction method of multipliers (ADMM) [6], incorporating with a similar optimization strategy in Algorithm 1.

We summarize our subspace clustering algorithm in Algorithm 2, where we first solve the problem (35) or (39) to calculate the similarity matrix $\mathbf{M} = |\mathbf{Z}| + |\mathbf{Z}|^T$, and then obtain the clustering results by applying the spectral clustering [41] to $\mathbf{M}$. In our work, we simply name the proposed MWEE-based subspace clustering method as MWEE-S.

B. MWEE-based Image Recognition

We now design a classifier based on the MWEE criterion for image recognition. In this task, objects from the same class are regarded as the ones lying in the same linear subspace. Assume that there are K classes, which correspond to a collection of K linear subspaces $\{\mathcal{S}_k\}_{k=1}^K$. With notations defined in Section I-A, let $\mathbf{X} \in \mathbb{R}^{N \times n}$ record n training samples, where each sample is a vector of dimension N. $\mathbf{X}_k$ contains n_k samples from the k-th class. Define $\mathbf{y} \in \mathbb{R}^N$ as a test sample which could be observed with severe noise. The fundamental task for image recognition is to assign $\mathbf{y}$ a correct class label by using the training samples in $\mathbf{X}$ [42].

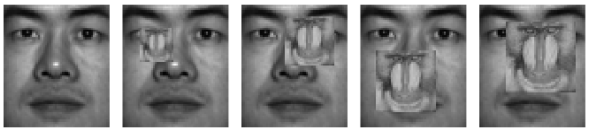


Fig. 7. Some examples of simulated images. From left to right: 0%, 10%, 20%, 30% and 40% regions are randomly selected and occluded.

Specifically, we first compute the representation of $\mathbf{y}$ through the proposed MWEE-based SR problem

$$\underset{\mathbf{z}}{\text{argmin}} \; -\tilde{S}(\mathbf{y} - \mathbf{X}\mathbf{z}) + \lambda \Phi(\mathbf{z}). \tag{40}$$

Then the reconstruction residual is calculated over each class by

$$R_k(\mathbf{y}) = -\tilde{S}(\mathbf{y} - \mathbf{X}_k T_k(\mathbf{z})), \;\; k \in \{1, ..., K\}. \tag{41}$$

Here $T_k(\mathbf{z})$ is a truncated operator extracting only the elements in $\mathbf{z}$ indexed by $\mathcal{I}_k$, which is a set consisting of the indices of the training samples from the k-th class. Eventually, we assign $\mathbf{y}$ to the class with the minimum reconstruction residual. Algorithm 3 summarizes the complete procedure. For simplicity, in this work, we set $\Phi(\cdot)$ as the ℓ_1 norm for the MWEE-based classifier. In the sequel, we name the MWEE-based classifier as MWEE-C.

V. Experimental Results for the Subspace Clustering

We evaluate the effectiveness of our proposed subspace clustering method MWEE-S on two subspace clustering tasks, i.e., face clustering and motion segmentation. Our code is available at https://github.com/YuanmanLi/github-MWEE.

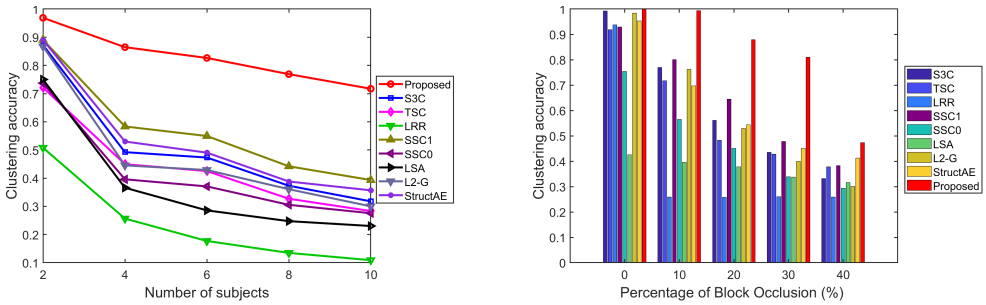
A. Face Clustering

As illustrated in Fig. 6, given a collection of face images of multiple subjects captured with various kinds of noise, face clustering aims to cluster them into their underlying subjects [43]. For this experiment, we adopt the `Extended Yale B` dataset [44], which consists of 38 individuals with 64 frontal face images for each subject. Images in this dataset are acquired under various illumination conditions. To reduce the complexity, all the images are downsampled to the resolution 96×84.

We employ the model MWEE-S presented in (35) for the problem of face clustering, and compare its performance with SSC0 [10], SSC1 [6], LRR [11], TSC [45], LSA [46], S^3C [47]

TABLE I
CLUSTERING ACCURACY (%) OVER THE EXTENDED YALE B DATASET. BEST RESULTS ARE MARKED IN BOLD.

Methods	LSA	SSC0	SSC1	LRR	TSC	L2-G	S^{3}C	StructAE	MWEE-S
2 subjects	71.09	99.22	96.89	97.66	97.66	98.44	99.22	98.44	**100.0**
4 subjects	42.58	75.39	92.97	93.75	91.80	98.44	99.22	95.31	**100.0**
6 subjects	45.05	85.94	94.01	96.62	93.49	98.44	95.83	95.05	**100.0**
8 subjects	33.98	60.35	93.75	75.59	90.43	97.66	94.92	95.51	**100.0**
10 subjects	32.50	53.75	87.19	76.56	86.41	96.56	94.69	94.38	**99.84**



(a) (b)

Fig. 8. (a) Average clustering accuracy against 25% contiguous occlusion. (b) Average clustering accuracy against various levels of occlusion.

and L2-G [21]. For our method MWEE-S, we empirically set the parameter λ as 10^{-4}. For LRR, we adopt the newly updated code [11]. For S^{3}C, we utilize the soft S^{3}C implementation. Besides the above model-based methods, we also compare MWEE-S with the deep learning based method StructAE [48]. There are five layers in StructAE, with 300-200-150-200-300 neurons. Similar to [48], we perform PCA on the data and reduce the dimension to 300 before feeding into the network.

As illustrated in Fig. 6, images in `Extended Yale B` are mainly degraded by the unconstrained illumination, where the noise obviously violates the *i.i.d.* assumption. However, such noise can be well described by our proposed *i.p.i.d.* source as we discussed previously. The clustering results of different algorithms for the first 2, 4, 6, 8 and 10 subjects are presented in Table I. It can be readily figured out that the clustering accuracy of those MSE-based methods (e.g., LLR and SSC0) drops rapidly as the number of subjects increases. Assisted by the proposed MWEE criterion, our MWEE-S obtains the best results for *all* the cases. Notably, when the number of subject is below 10, our method can correctly cluster all the images to their subjects, while in the case of 10 subjects, our method only miss-clusters one image. Note that L2-G [21], S^{3}C [47] and StructAE [48] also perform quite well on this dataset. However, it should be noted that L2-G applied a post-procedure to refine the representation coefficients, while S^{3}C employed a very complex regularization function. Furthermore, we in the following will show that they are not robust in the case of complex noise scenarios.

To further show the robustness of our approach, we simulate the contiguous occlusion by randomly selecting a local region in each face image, and then substituting it with an unrelated '`Baboon`' image. Some examples are shown in Fig. 7.

We should emphasize that, for the subspace clustering problem, all the images are occluded. While for the face recognition problem [5], the images in the training set are not corrupted. Clearly, such a difference makes the former problem more challenging. Note that the noise caused by the occlusion is a combination of the unconstrained illumination and the '`Baboon`' image, both of which are highly structural, and should be of low entropy. Such noise obviously dose not satisfy the *i.i.d.* property. Fig. 8(a) plots the curves of the clustering accuracy against 25% occlusion for different algorithms, where each result is the average of 10 runs. It can be observed from Fig. 8(a) that MWEE-S beats all the competing algorithms by a big margin, especially when the number of subjects gets larger. Even under 25% occlusion, we can see that the performance degradation of MWEE-S is graceful in comparison with the results listed in Table I. On the contrary, in the same case, the accuracy of all the other algorithms is severely degraded. The results further demonstrate the effectiveness of our MWEE criterion in modeling the practically encountered noise. We can also observe that the performance of the deep learning based method StructAE is highly degraded under complex noise. This is not surprising since for subspace clustering, the network is trained in an unsupervised fashion (without labels), where the representations learned by the deep network highly depend on the designed objective function. StructAE still adopted an MSE-based objective function, making it ineffective to model the behaviors of non-Gaussian noises.

To analyze how the occlusion level affects the clustering accuracy, we fix the number of subjects as 4, while varying the occlusion level from 0 to 40%. The number of subjects is fixed as 4. The results of different methods are plotted in Fig. 8(b). We can still see that MWEE-S obtains remarkable performance gains compared with the competing algorithms, especially for the occlusion level above 10%.

B. Motion Segmentation

Motion segmentation refers to the task of segmenting a video clip of several rigidly moving objects into multiple spatiotemporal regions, where each region corresponds to one motion in the scene. Illustrations are shown in Fig. 9.

(a) (b) (c) (d) (e) (f)

Fig. 9. Examples of motion segmentation. The first row shows the feature points of multiple moving objects tracked in a video, while the second row visualizes the segmented results. Here we only show two frames of each video.

TABLE II
CLUSTERING ERROR (%) OVER HOPKINS155 DATASET. BEST RESULTS ARE MARKED IN BOLD.

Methods		LSR	SSC0	SSC1	LRR	LRSC	L2-G	S^3C	StructAE	MWEE-S
2F	Avg.	2.98	6.97	2.18	1.60	3.42	5.54	2.20	4.70	**1.22**
	Med.	0.30	0.21	**0.00**	**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	7.89	**4.20**
4K	Avg.	3.21	7.05	2.42	2.35	3.35	5.81	2.33	6.05	**1.77**
	Med.	0.38	0.21	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	1.06	**0.00**
	Std.	7.79	12.82	7.51	7.30	8.76	11.59	6.98	8.26	**6.12**

Given a video sequence with F frames, the motion segmentation generally consists of two steps. The first step is extracting and tracking a collection of n feature points $\{\mathbf{x}_{f,i} \in \mathbb{R}^2\}_{i=1}^n$ through all the frames $f = 1, .., F$ of the video. The second step is to apply the clustering algorithms to these feature points and then obtain the segmentation results.

Specifically, each feature point, also called the feature trajectory, is a $2F$-dimensional vector formed by stacking its spatial positions of all frames, i.e.,

$$\mathbf{x}_i = [\mathbf{x}_{1,i}^T, \mathbf{x}_{2,i}^T, .., \mathbf{x}_{F,i}^T]^T \in \mathbb{R}^{2F}. \tag{42}$$

We use $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ to denote the data matrix containing all the feature trajectories of a video. According to [6], feature trajectories of a single motion lie in an affine subspace of $\mathbb{R}^{2F}$, and hence we can regard that the ones of K rigid motions lie in a union of K low-dimensional subspaces of $\mathbb{R}^{2F}$. For the performance evaluation of the motion segmentation, we use the `Hopkins155` dataset [49], which contains 156 video sequences, where each video has 2 or 3 motions. Some video frames are shown in Fig. 9. On average, in this dataset, each sequence of 2 motions has about 30 frames and 266 trajectories, and the one of 3 motions has about 29 frames and 398 trajectories.

We use MWEE-S (see 39) tailored for the affine subspace clustering for the motion segmentation task, and we empirically set the parameter $\lambda = 10^{-4}$. The considered competing methods include SSC0 [10], SSC1 [6], LSR [50], LRR [11], LRSC [51], L2-G [21], S^3C [47] and StructAE [48]. Different from SSC0, the method SSC1 adopts an additional affine constraint $\mathbf{Z}^T\mathbf{1} = \mathbf{1}$. For StructAE, since the dimension of feature points (i.e., $2F$) in a video sequence is less than 300, we set the architecture of the network as $2F$-100-F-100-$2F$. We have carefully tuned the parameters and the best results are reported. We first conduct the experiment using the original $2F$-dimensional feature trajectories, and similar to [6], we also perform another experiment by using PCA to project the trajectories into a $4K$-dimensional subspace (K denotes the number of motions). We summarize the average, median and standard deviation of the clustering errors of different algorithms in Table II. First, we can see that the clustering performance of different approaches generally slightly degrades when reducing the feature dimension to $4K$ because of the information loss. Second, it can also be observed that our method MWEE-S achieves much smaller clustering errors and standard deviations, than all the other competing methods in both cases. The results imply that our MWEE criterion designed with the *i.p.i.d.* noise model is indeed beneficial for the motion segmentation.

VI. EXPERIMENTAL RESULTS FOR IMAGE RECOGNITION

We now demonstrate the effectiveness of our method MWEE-C for robust face recognition, where the images are observed with different types of noise. The recognition rates are recorded for the performance evaluation of different algorithms.

A. Experimental Setting and Datasets

Three popular public face datasets are employed in our experiments, i.e., the `Extended Yale B` [44], the `AR` [52], and the `Yale` [53] datasets. Images from these datasets present large variations over occlusions, poses, expressions and illuminations. We described the `Extended Yale B` in Section V-A, while the descriptions of the remaining two datasets are given below:

`AR` dataset: This dataset is comprised of over 4000 facial images taken from 126 subjects (56 women and 70 men).

TABLE III
RECOGNITION RATES (%) UNDER DIFFERENT LEVELS OF RANDOM PIXEL CORRUPTION. BEST (SECOND BEST) RESULTS ARE MARKED IN RED (BLUE).

Levels		Extended Yale B								Yale							
		LRC	CRC	SRC	HQ	L2-G	MEESRC	NMR	MWEE-C	LRC	CRC	SRC	HQ	L2-G	MEESRC	NMR	MWEE-C
0	Min	94.2	97.0	96.5	92.3	94.1	93.3	99.3	99.5	80.0	92.2	96.7	81.1	88.9	82.2	95.6	92.2
	Max	98.5	99.2	99.7	95.7	96.6	96.5	99.8	99.8	83.3	98.9	100	83.3	94.4	85.6	100	100
	Mean	97.0	98.6	98.6	93.8	95.1	94.7	99.7	99.6	81.4	96.1	98.6	82.5	91.7	83.6	98.3	94.4
10	Min	93.4	87.4	89.6	96.7	75.2	92.6	91.3	97.2	80.0	91.1	95.6	84.4	85.6	82.2	94.4	92.2
	Max	97.9	93.3	96.0	98.4	77.0	97.1	97.2	99.1	83.3	93.3	97.8	85.6	92.2	86.7	97.8	94.4
	Mean	96.4	91.7	94.0	97.3	75.9	95.1	95.5	98.6	82.2	92.2	96.7	84.7	90.0	84.4	95.8	93.9
30	Min	88.0	77.6	81.7	95.6	42.4	88.1	84.0	93.3	68.9	70.0	91.1	82.2	76.7	82.2	88.9	87.8
	Max	94.5	82.8	87.3	98.4	50.3	94.1	91.9	97.0	76.7	81.1	92.2	85.6	83.3	85.6	93.3	97.8
	Mean	92.6	80.4	85.0	97.2	46.2	91.7	89.0	95.8	73.6	78.3	91.9	84.7	81.1	84.4	90.0	94.2
50	Min	68.4	54.9	68.5	92.4	23.4	81.8	74.9	86.2	24.4	41.1	53.3	65.6	48.9	82.2	56.7	85.6
	Max	73.9	56.6	72.3	95.9	28.4	89.7	79.4	92.1	33.3	51.1	60.0	76.7	61.1	86.7	65.6	87.8
	Mean	72.0	55.3	69.6	94.4	25.0	87.0	76.9	90.1	30.6	46.4	55.8	70.0	55.3	84.4	61.1	86.7
70	Min	21.0	22.0	30.3	69.5	9.6	75.7	38.9	76.2	12.2	18.9	15.6	32.2	23.3	77.8	15.6	78.9
	Max	25.5	22.6	33.6	76.1	11.2	80.8	42.9	81.5	13.3	26.7	17.8	61.1	26.7	81.1	22.2	83.3
	Mean	23.4	22.2	32.4	72.8	10.4	78.1	40.8	78.9	12.8	20.8	16.7	44.2	24.2	79.7	19.4	81.4

There are 26 images for each subject with different expressions (anger, smile, scream and neutral), illumination conditions and occlusions (sunglasses and scarf). A subset [5] of this dataset containing 50 female subjects and 50 male subjects is used in our experiments. We resize the images to the resolution of 165×120.

`Yale` dataset: The `Yale` dataset includes 15 subjects with totally 165 facial images (11 images for each). These images are captured under different lighting conditions, and with varying facial expressions. Images in this dataset are resized to the resolution of 48×48.

To demonstrate the effectiveness of our algorithm MWEE-C, we compare the performance with the following approaches, including the benchmark classifiers: SRC [5], LRC [54], CRC [55], and those methods designed for robust face recognition: CESR [19], RSC [27], HQ [56], L2-G [21], MEESRC [33] and NMR [4]. Note that L2-G adopts k-NN as the classifier, while all the other methods classify images based on the reconstruction residual. For a fair comparison, the associated regularization parameters of the competing methods are carefully tuned from the set $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, and we report the best results. For our algorithm MWEE-C, we set the regularization parameter λ as 10^{-4} for all the experiments.

B. Simulated Random Pixel Corruption

In this subsection, we verify the effectiveness of MWEE-C against the *i.i.d.* noise. In Theorem 1, we have theoretically demonstrated the equivalence of $\bar{H}_2(\mathbf{e})$ and the traditional Rényi's entropy in the *i.i.d.* scenario, implying that MWEE-C can also handle the *i.i.d.* noise corruption satisfactorily. The `Extended Yale B` [44] and the `Yale` [53] datasets are employed in the following experiments. For efficiency, we resize the images in `Extended Yale B` to 96×84 resolution. For each subject of these two datasets, half of the images are randomly selected for training (32 images per subject for `Extended Yale B`, and 5 for `Yale`), while the remaining half for testing. To simulate the random pixel corruption, a set of pixels of each test image are randomly selected and replaced by the random values generated from a uniform distribution over $[0, 255]$. The percentage of replaced pixels is varied from 10% to 70% to have different levels of corruption. Fig. 10 shows some examples.

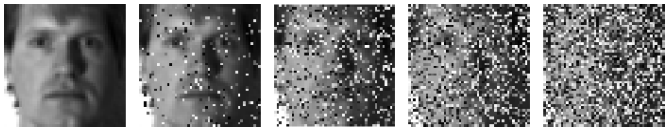
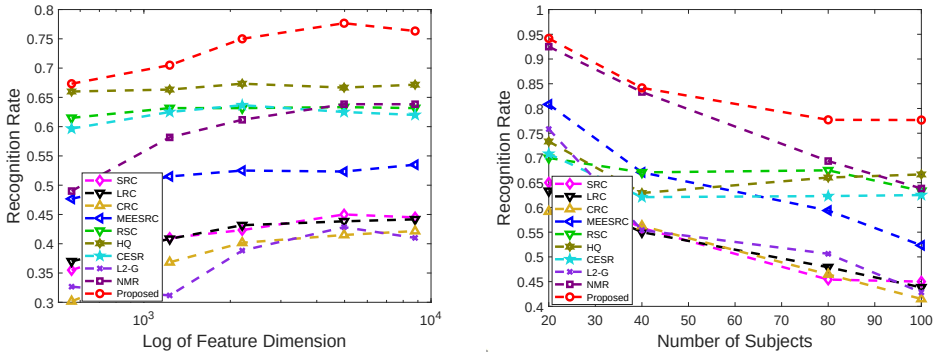


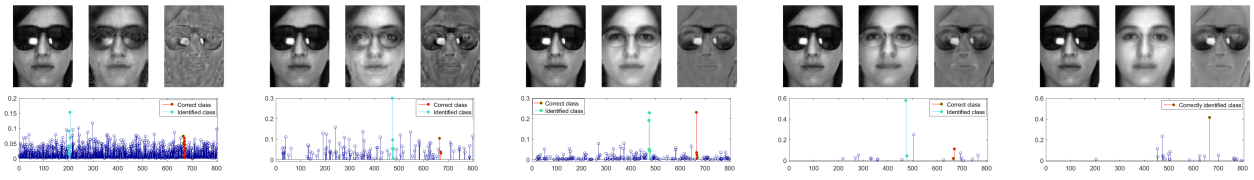
Fig. 10. Images from left to right are the original image, and the ones randomly corrupted by 10%, 30%, 50% and 70%, respectively.

Table III reports the performance of different algorithms under various levels of random pixel corruption. It can be seen that all the methods, including those MSE-based ones (i.e., LRC, CRC, L2-G and SRC), achieve good performance under low levels of corruption. Particularly, SRC performs the best on `Yale` dataset under 10% corruption. One potential reason is that the statistical behavior of the noise in a small image is not prominent under the low levels of corruption. On the other hand, it can be observed that only the ITL-based approaches, i.e., MWEE-C, HQ and MEESRC are robust against the high levels of corruption. For instance, the average recognition rate of MWEE-C is still 78.9% on the `Extended Yale B` dataset against 70% corruption. For those MSE-based methods, their performance degrades rapidly when the corruption level is above 30%. Specifically, when the corruption level reaches 70%, the mean recognition rates of all the MSE-based algorithms are below 25% on the `Yale` dataset; nevertheless, in the same case, the recognition rate of MWEE-C is still over 80%. This demonstrates that the MSE criterion is not robust against non-Gaussian noise, even if the noise satisfies the *i.i.d.* property. In addition, it can be seen that the performance gain of MWEE-C over MEESRC becomes smaller as the corruption level gets higher. Under low levels of corruption, the noise is mainly caused by the varying expressions and different lighting conditions, rather than the *i.i.d.* noise generated from the uniform distribution. In this case, the MWEE criterion derived from the *i.p.i.d.* model can more accurately characterize the true noise behavior. Meanwhile, for the high levels of corruption, the noise is dominated by the purely random signal. Then our model degrades to MEESRC based on the *i.i.d.* noise model. It should be noted that, in



(a) (b)

Fig. 11. (a) Recognition rates over different feature dimensions. (b) Recognition rates over the first K subjects with the feature dimension of 83×60.



(a) (b) (c) (d) (e)

Fig. 12. Representation coefficients and reconstructed results under sunglasses occlusion. For each subfigure, pictures in the top row are the occluded image, the reconstructed image and the residual map, respectively. The second row from left to right shows the magnitudes of coefficients computed by CRC, SRC, RSC, MEESRC and the proposed algorithm, respectively.

reality, the former cases are the most frequently occurring ones.

C. Real-World Occlusion

We now test the robustness of MWEE-C against the real disguise problem, i.e., sunglasses occlusion. The `AR` dataset is employed in our experiment. For training, we use the 800 non-occluded facial images with varying expressions (8 samples per subject except for the corrupted image `w-027-14.bmp`) in the `AR` dataset. For testing, we take the 600 images with sunglasses occlusion under different illumination conditions (6 samples per subject). Fig. 11(a) reports the performance with feature dimensions of 28×20, 41×30, 55×40, 83×60 and 110×80. We can notice that MWEE-C outperforms the other methods for all the feature dimensions consistently. Fig. 11(b) further shows the performance of different algorithms with various numbers of subjects. We can still observe significant gains of MWEE-C, compared with the other competitors especially when K (the number of subjects) is large. For instance, we can see from Fig. 11 that, with 83×60 feature dimension and $K = 100$, the performance gains of MWEE-C over SRC, CRC, L2-G, LRC, MEESRC, CESR, HQ, RSC and NMR are 33.2%, 36.2%, 34.84%, 33.8%, 25.3%, 20.2%, 11%, 14.5% and 13.9%, respectively, which are quite remarkable. This shows that our MWEE criterion designed with the generic noise model is more robust against the noise caused by the sunglasses occlusion.

Fig. 12 provides some representative reconstruction results of SRC, CRC, RSC, MEESRC and MWEE-C. It is worth emphasizing that *only* our proposed method MWEE-C correctly classifies the test image in this case (the correct subject identity is 84 in the `AR` dataset). Since the 60-th and the 84-th subjects are quite similar, RSC and MEESRC misclassify the test image to the 60-th subject who wears glasses as shown in Figs. 12(c)-(d). This is because RSC and MEESRC rely on the *i.i.d.* assumption, leading to the insufficient robustness against the contiguous occlusion. For CRC and SRC, they simply treat the noise as *i.i.d.* Gaussian, and hence, in fact regard the sunglasses as one part of the face. In other words, sunglasses and the face components such as nose and mouth would have the same importance for face representation. This explains why the MSE criterion tends to search many unrelated faces to reconstruct the sunglasses as shown in Figs. 12(a)-(b). Due to the superior modeling capability of our generic noise model, the sunglasses, as the dominating degradation source, can still be satisfactorily characterized. From Fig. 12(e), we can observe that the reconstruction residual obtained by MWEE-C is of very limited entropy, despite its large error under the MSE criterion. Though the 60-th and 84-th subjects are quite similar, our proposed MWEE-C can still assign the image to the correct subject.

VII. Conclusions

This paper has presented a novel information-theoretic learning approach for robust SR. Different from the traditional representation schemes, the proposed framework does not make the *i.i.d.* assumption and the Gaussianity on the noise. To the best of our knowledge, this may be the first work explicitly using a union of distributions to characterize the noise in the context of robust data representation. With the designed *i.p.i.d.* noise model, we have proposed a new optimization criterion MWEE for robust SR, which exhibits superior advantages to exploit the inherent information of both the *i.i.d.* noise and the non-*i.i.d.* one. Such desirable properties make the resulting

frameworks very robust to handle various practical noises. We have demonstrated the usefulness of our developed approach through solving the subspace clustering and image recognition problems, and our method has achieved very promising results. As an interesting direction to be exploited in the future, the designed MWEE criterion can be utilized as a generic tool in the other learning systems, thus potentially improving their robustness against various kinds of practical noise.

REFERENCES

[1] R. E. Bellman, *Dynamic Programming*. Dover Publications, Inc., 2003.

[2] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Proc. of ACM Int. Conf. Multimedia*, 2007, pp. 301–304.

[3] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, "Deep clustering with sample-assignment invariance prior," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–12, 2019.

[4] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, 2017.

[5] 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.*, vol. 31, no. 2, pp. 210–227, 2009.

[6] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[7] C. Lu, J. Feng, Z. Lin, T. Mei, and S. Yan, "Subspace clustering by block diagonal representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 487–501, 2019.

[8] J. C. Lv, Z. Yi, and J. Zhou, *Subspace learning of neural networks*. Crc Press, 2018, vol. 42.

[9] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4833–4843, 2018.

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2009, pp. 2790–2797.

[11] 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.*, vol. 35, no. 1, pp. 171–184, Jan 2013.

[12] Y. Li, J. Zhou, X. Zheng, J. Tian, and Y. Y. Tang, "Robust subspace clustering with independent and piecewise identically distributed noise modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2019, pp. 8720–8729.

[13] C. You, C.-G. Li, D. P. Robinson, and R. Vidal, "Is an affine constraint needed for affine subspace clustering?" in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9915–9924.

[14] R. He, L. Wang, Z. Sun, Y. Zhang, and B. Li, "Information theoretic subspace clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2643–2655, 2016.

[15] Z. Yue, H. Yong, D. Meng, Q. Zhao, Y. Leung, and L. Zhang, "Robust multiview subspace learning with nonindependently and nonidentically distributed complex noise," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2019.

[16] K. Guo, L. Liu, X. Xu, D. Xu, and D. Tao, "Godec+: Fast and robust low-rank matrix decomposition based on maximum correntropy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2323–2336, 2018.

[17] J. C. Lv, K. K. Tan, Z. Yi, and S. Huang, "A family of fuzzy learning algorithms for robust principal component analysis neural networks," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 1, pp. 217–226, 2010.

[18] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4094–4107, 2012.

[19] R. He, W. S. Zheng, and B. G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, 2011.

[20] M. Yin, S. Xie, Z. Wu, Y. Zhang, and J. Gao, "Subspace clustering via learning an adaptive low-rank graph," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3716–3728, 2018.

[21] X. Peng, Z. Yu, Z. Yi, and H. Tang, "Constructing the l2-graph for robust subspace learning and subspace clustering," *IEEE Trans. on Cybern.*, vol. 47, no. 4, pp. 1053–1066, 2017.

[22] L. Xie, M. Yin, X. Yin, Y. Liu, and G. Yin, "Low-rank sparse preserving projections for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5261–5274, 2018.

[23] 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.*, 2017, pp. 7445–7454.

[24] 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.*, vol. 25, no. 2, pp. 850–863, 2016.

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

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

[27] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, 2013.

[28] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, 2007.

[29] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou, "Learning theory approach to minimum error entropy criterion," *J. Mach. Learn. Res.*, vol. 14, no. 2, pp. 377–397, 2013.

[30] X.-T. Yuan and B.-G. Hu, "Robust feature extraction via information theoretic learning," in *Proc. Int. Conf. Mach. Learn.* ACM, 2009, pp. 1193–1200.

[31] X. Zhang, B. Chen, H. Sun, Z. Liu, Z. Ren, and Y. Li, "Robust low-rank kernel subspace clustering based on the schatten p-norm and correntropy," *IEEE Trans. on Know. and Data Eng.*, pp. 1–1, 2019.

[32] D. Erdogmus and J. C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1780–1786, 2002.

[33] Y. Wang, Y. Y. Tang, and L. Li, "Robust face recognition via minimum error entropy-based atomic representation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5868–5878, 2015.

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

[35] F. M. J. Willems, "Coding for a binary independent piecewise-identically-distributed source," *IEEE Trans. Signal Process.*, vol. 42, no. 6, pp. 2210–2217, 1996.

[36] G. A. Babich and O. I. Camps, "Weighted parzen windows for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 5, pp. 567–570, 1996.

[37] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.

[38] R. He, B. Hu, X. Yuan, L. Wang *et al.*, *Robust recognition via information theoretic learning*. Amsterdam, The Netherlands: Springer, 2014.

[39] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin, "Correntropy induced l2 graph for robust subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1801–1808.

[40] T. Jin, R. Ji, Y. Gao, X. Sun, X. Zhao, and D. Tao, "Correntropy-induced robust low-rank hypergraph," *IEEE Trans. on Image Process.*, vol. 28, no. 6, pp. 2755–2769, 2019.

[41] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Neural Information Processing Systems,*, 2002, pp. 849–856.

[42] J. Feng, B. Ni, Q. Tian, and S. Yan, "Geometric ℓ_p-norm feature pooling for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2011, pp. 2609–2704.

[43] D. Chen, J. Lv, and Y. Zhang, "Unsupervised multi-manifold clustering by learning deep representation," in *Proc. Workshops AAAI Conf. Artif. Intell.*, 2017, pp. 385–391.

[44] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, 2005.

[45] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *IEEE Trans. Signal Process.*, vol. 61, no. 11, pp. 6320–6342, 2015.

[46] 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*. Springer, 2006, pp. 94–106.

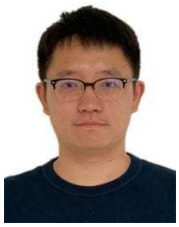
[47] C. G. Li, C. You, and R. Vidal, "Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2988–3001, 2017.

[48] X. Peng, J. Feng, S. Xiao, W. Yau, J. T. Zhou, and S. Yang, "Structured autoencoders for subspace clustering," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5076–5086, 2018.

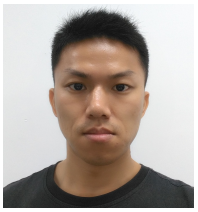
[49] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2007, pp. 1–8.
[50] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. Eur. Conf. Comput. Vis.,.* Springer, 2012, pp. 347–360.
[51] 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.* IEEE, 2011, pp. 1801–1807.
[52] A. Martinez and R. Benavente, "The ar face database," *Computer Vision Center, Tech. Rep. 24*, 1998.
[53] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.
[54] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, 2010.
[55] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. Int. Conf. Comput. Vis.* IEEE, 2011, pp. 471–478.
[56] R. He, W. S. Zheng, T. Tan, and Z. Sun, "Half-quadratic-based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 261–275, 2014.

Yuanman Li (Member, IEEE) received the B.Eng. degree in software engineering from Chongqing University, Chongqing, China, in 2012, and the Ph.D. degree in computer science from University of Macau, Macau, 2018. From 2018 to 2019, he was a Post-doctoral Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau. He is currently an Assistant Professor with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include data representation, multimedia security and forensics, computer vision and machine learning.

Jiantao Zhou (Senior Member, IEEE) received the B.Eng. degree from the Department of Electronic Engineering, Dalian University of Technology, in 2002, the M.Phil. degree from the Department of Radio Engineering, Southeast University, in 2005, and the Ph.D. degree from the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, in 2009. He held various research positions with the University of Illinois at Urbana-Champaign, The Hong Kong University of Science and Technology, and the McMaster University. He is currently an Associate Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau. His research interests include multimedia security and forensics, multimedia signal processing, artificial intelligence and big data. He holds four granted U.S. patents and two granted Chinese patents. He has coauthored two articles that received the Best Paper Award at the IEEE Pacific-Rim Conference on Multimedia in 2007 and the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2016. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.



Jinyu Tian (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in mathematics from Chongqing University, Chongqing, China. He is currently pursuing the Ph.D. degree with the Faculty of Science and Technology, University of Macau, Taipa, China. His current research interests include machine learning and pattern recognition.



Xianwei Zheng (Member, IEEE) received the B.Sc. degree in mathematics from Hanshan Normal University, Chaozhou, China, in 2009, the M.Sc. degree in applied mathematics from Shantou University, Shantou, China, in 2012, the Ph.D. degree in mathematics from Shantou University, China, and the Ph.D. degree in computer science from the University of Macau, China, in 2018, respectively. He joined the School of Mathematics and Big Data, Foshan University in 2018 as a young research fellow. His current research interests include graph signal processing, image processing, machine learning, wavelet analysis and frame theory.

Yuan Yan Tang (Life Fellow, IEEE) is currently a Chair Professor with the Faculty of Science and Technology, University of Macau, and a Professor/an Adjunct Professor/a Honorary Professor with several institutes, including Chongqing University, China; Concordia University, Canada; Hong Kong Baptist University, Hong Kong; and the Advanced Innovation Center for Big Data and Brain Computing, Beihang University. He has authored over 400 academic papers and has authored or co-authored over 25 monographs/books/book chapters. His current interests include wavelets, pattern recognition, image processing, and artificial intelligence. He is a fellow of the International Associate of Pattern Recognition (IAPR). He is the Founder and the Editor-in-Chief of the International Journal on Wavelets, Multiresolution, and Information Processing and an associate editor of several international journals. He is the Founder and the Chair of the Pattern Recognition Committee in the IEEE SMC. He served as the general chair, the program chair, and a committee member for many international conferences. He is the Founder and the General Chair of the series International Conferences on Wavelets Analysis and Pattern Recognition. He is the Founder and the Chair of the Macau Branch of the IAPR.