Smooth Non-negative Sparse Representation for Face and Handwritten Recognition

Aboozar Ghaffari†, Mahdi Kafaee‡, and Vahid Abolghasemi⋆

†Biomedical Engineering Department, School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran, (Email: aboozar_ghaffari@iust.ac.ir)
‡Faculty of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran, (Email: kafaee@shahroodut.ac.ir)
⋆School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ, UK (Email: v.abolghasemi@essex.ac.uk)

June 21, 2021

Abstract

In sparse representation problem, there is always interest to reduce the solution space by introducing additional constraints. This can lead to efficient application-specific algorithms. Despite known advantages of sparsity and non-negativity for image data representation, limited studies have addressed these characteristics simultaneously, due to the challenges involved. In this paper, we propose a novel inexpensive sparse non-negative reconstruction method. We utilise a non-negativity penalty term within a convex function while imposing sparsity at the same time. Our method, termed as SnSA (smooth non-negative sparse approximation) applies a novel thresholding strategy on the sparse coefficients during the minimisation of the proposed convex function. The main advantage of SnSA algorithm is that hard zeroing the negative samples which leads to unstable and non-optimal sparse solution is avoided. Instead, a differentiable smoothing function is proposed that allows gradual suppression of negative samples leading to a sparse non-negative solution. This way, the algorithm is driven toward a solution with a balance in maximising the sparsity and minimising the reconstruction error. Our numerical and experimental results on both synthetic signals and well-established face and handwritten image databases, indicate higher classification performance of the proposed method compared to the state-of-the-art techniques.

Keywords— Non-negative sparse representation, Gradient descent, Smoothing function, face recognition, handwritten recognition.

1 Introduction

Sparse representation problem is one of the most attractive and demanding topics in signal processing, image processing, computer vision and pattern classification research **[1, 2, 3]**. It is now explicitly observed that one can represent variety of signals/images/patterns with only few non-zero samples using an overcomplete matrix, the so-called dictionary. In fact, the input data, e.g. face images, can be represented as a linear combination of few (sparse) coefficients with respect to a predefined or learned dictionary. This image representation scheme can then be used for various purposes from image denoising, to image classification and object tracking. There are many different data types in the world with underlying sparse structure which make the sparse analysis meaningful.

Original sparse recovery problem can be defined as follows

$$\min \|\mathbf{s}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{A}\mathbf{s} \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^n$ is sparse coefficients vector having at most k non-zero elements ($k \ll n$), $\mathbf{A} \in \mathbb{R}^{m \times n}$ is called dictionary, and $\mathbf{y} \in \mathbb{R}^m$ is the corresponding non-sparse-domain vector which can be regarded as input data sample, e.g. face image in vectorised form. The dictionary is normally chosen to be overcomplete, i.e. $m < n$. The columns of the dictionary are called atoms. In addition, the term $\|\mathbf{s}\|_0 = \sum s_i^0$ is called ℓ_0-norm and counts the number of non-zero elements in $\mathbf{s}$. It also worth noting that (1) has a unique and exact solution under specific conditions on k, m, and structure of dictionary. Depending on the application and data of interest, it might be required to impose additional constraint(s) on the sparse recovery problem for obtaining desired results. This is when it becomes very important to decide what family of methods to choose in order to mitigate the computational and analytical burden of adding new constraint(s) **as**

well as maintaining reconstruction quality. In general, solving (1), which is a non-convex problem, is NP-hard. Hence, various approaches have been proposed to convert it to a feasible problem. Most traditional techniques attempt to convexify (1) by replacing ℓ_0-norm with ℓ_1-norm. The reason is that ℓ_1-norm is a differentiable function and thus there exist many typical techniques to tackle it. **However, it normally requires expensive optimisation tools.** One of the important constraints, widely used in many applications, is non-negativity which is of particular interest in applications dealing with non-negative data [4, 5]. In fact, since the image pixels are naturally non-negative quantities, they can be used for parts-based description of the object of interest in the image. For instance, parts of a face image (e.g. eyes, eyebrows, lips) can be represented only by applying *addition* operator on a selection of pixels and hence the non-negativity condition is preserved.

In this paper, we propose a novel approach to solve sparse recovery problem (1) with additive non-negative penalty. **Motivated by the effectiveness of non-negativity constraint in learning parts of objects, particularly in applications like face and handwritten recognition [6], we derive and embed a mathematical smoothing function to simultaneously exploit sparsity and non-negativity. We consider direct minimisation of ℓ_0-norm, instead of ℓ_1-norm, to avoid encountering complex optimisation issues. To do this, a novel auxiliary function with tunable parameters to control smoothness and non-negativity is proposed. The main advantage of this function is that it is differentiable and can be directly embedded in the optimisation problem.** Our proposed approach can find a stable solution that avoids rigid weighting function such as those reported in previous works. **Our sparse reconstruction regime starts by allowing** existence of negative coefficients **but at a high cost.** These negative sparse coefficients are gradually suppressed **by appropriate weight functions to ultimately turn them into non-negative (and sparse) components while the reconstruction error is minimised simultaneously**. In other words, we do not blindly zero-out all negative values (unlike traditional techniques), but leave the algorithm to automatically adjust the reconstructed signal to a non-negative solution. This innovative dynamic suppression technique makes a great impact on the reconstructed coefficients compared to previous works. The mathematical tool we propose for this purpose is a smooth differentiable function that forms the proposed cost function. Then, a solution based on gradient descent minimisation is proposed. Finally, **the theoretical contributions achieved in this study are supported** by presenting a non-negative sparse representation classification utilised in face and handwritten image recognition applications.

The rest of the paper is organised as follows. In section 2, related works and state-of-the-art are reviewed. The proposed method and its associated mathematical formulations are described in section 3. Section 4 is devoted to represent the numerical experiments and the results. Finally, the paper is concluded in section 5.

2 Related works

One of the well-known sparse recovery methods is called basis pursuit (BP) [7]. In BP, the minimisation problem (1) is reformulated to be solved using linear programming. This family of approaches is precise and stable but too complex and heavy-run. There has been also reported a family of greedy techniques such as orthogonal matching pursuit (OMP) [8] to solve (1). The main advantages of these techniques are simplicity and fast implementation, despite less accuracy compared to BP. An alternative family of inexpensive sparse recovery methods, called iterative shrinkage techniques, has also been proposed in the literature [9, 10]. These methods fundamentally use an iterative scheme comprising a multiplication by dictionary and its adjoint, and a simple scalar shrinkage step. The shrinkage operation, which is a kind of sparsification, sets to zero those elements that fall below a threshold and leaves the remaining elements untouched. **Among other existing methods, Orthogonal Least-Squares (OLS) [11] has drawn attention in recent years in several applications. OLS has been proposed for recovery of sparse vectors in both noisy and noiseless scenarios. Unlike OMP which performs few linear inversions, OLS performs as many inversions and therefore it is relatively expensive. However, it has shown superior performance than OMP as a consequence. Relevance vector machine (RVM), as a statistical sparse coding technique, uses Bayesian model to obtain the parsimonious solutions for regression and probabilistic classification [12]. It is also called probabilistic sparse Kernel version of support vector machine (SVM) which can be used for sparse representation problems and classification.**

Sparsity and non-negativity have been used in areas such as pattern classification [13], particularly for image super-resolution [14], unsupervised feature selection [15], spectral clustering [16], and graph matching [17]. Sparse non-negative image representation has shown effectiveness in reducing the reconstruction error for local features and mitigating the computational cost of sparse coding-based image features [18]. There are many applications where *transform coefficients* are encountered to be sparse non-negative, e.g. in spectroscopy, hyperspectral imaging, tomography, DNA microarrays, and network monitoring [19, 20, 21]. This is of significant practical interest in X-ray computed tomography (CT), single photon emission computed tomography (SPECT), positron emission tomography (PET), and magnetic resonance imaging (MRI). For instance, an accelerated proximal-gradient technique for reconstructing non-negative signals being sparse in a transform domain from underdetermined measurements has proposed in [22]. The authors applied ℓ_1-norm and non-negativity constraint on the signal and its transform coefficients and reported a greater reconstruction performance compared to existing works [22]. Given the non-negative nature of sound, automatic music transcription using a non-negative sparse algorithm was proposed [23]. Similarly, a voice activity detection approach for noisy scenarios has been proposed in [24] under the non-negative sparse coding regime.

Utilising sparsity penalty into the non-negative matrix factorisation (NMF) problem has also been extensively studied with many applications from face recognition, [6, 25, 26] to biomedical engineering [5] and community detection [27]. In NMF, the aim is to extract meaningful features from input data matrix by factorising (approximating) it into two non-negative matrices. The main issue in NMF is that it cannot always guarantee sparse and parts-based representation of non-negative data. Therefore, enforcing sparsity to the objective function seems necessary but challenging. Meanwhile, there are some methods that add extra constraints to improve the convergence and speed of NMF [28, 29]. While ℓ_0-norm induces a natural sparsity measure, most works apply ℓ_1-norm constraint due to its well-posedness. However, we found one work that applies ℓ_0-norm constraint for approximate NMF by following an alternating least squares scheme [30, 31]. Since NMF has not been basically designed for classification problem, it cannot be directly suited for this purpose. However, it is encouraging to study how to exploit non-negativity and sparsity for classification of non-negative data, e.g. images. This idea, which has been rarely explored so far, will be addressed in this paper.

Sparse representation classification (SRC) techniques are among those that take advantages of sparsity for classification purposes [32]. Several extensions of this family of methods have been presented by adding specific constraints. For instance, Yuan et al. proposed a non-negative dictionary based on SRC for ear recognition [33]. They attempt to model partial occlusion and design a dictionary using Gabor features extracted from ear images. A label orthogonal regularised NMF was proposed in [34] for image classification. They combine label consistency, non-negativity and orthogonality for learning dictionary atoms that are discriminative. They evaluate the performance of this technique on digit and face databases. In microwave image classification, a method called aspect-aided dynamic non-negative sparse representation was proposed by Zhang et al. [35]. The authors attempt to classify active and inactive atoms via establishing a dynamic dictionary. Then, they use ℓ_1-regularised non-negative sparse representation for final sparse recovery and classification. Several other applications of sparse representations for classification include hyperspectral image classification [36], traffic sign classification [37] and plant recognition [38].

Although direct enforcing of ℓ_0-norm into the reconstruction problem is challenging, several researchers attempted to find innovative alternatives [30, 39, 40]. One of the interesting methods of this kind is called smoothed ℓ_0 (SL0) where ℓ_0-norm of a vector is approximated by an exponential smoothing function [39]. While there are several methods that apply sparsity and smoothness in general reconstruction problems [41], very few works have reported its efficacy for non-negative problems. Amongst few, Mohammadi et al. added non-negativity penalty to SL0, and proposed a method called constrained smoothed L0 (CSL0) [42]. In this method, the negative sparse coefficients are severely suppressed by introducing some weights against positive ones. The weights are static and cannot change with respect to the algorithm progress. In another work, a modification has been proposed to make orthogonal matching pursuit (OMP) non-negative [43], which was later improved in terms of computational complexity [44]. A robust non-negative sparse recovery method was proposed in [45] where the authors address recovery of non-negative vectors from non-negative compressive measurements. Random Bernoulli matrix (with 0/1 values) is considered for this purpose to preserve the non-negativity property.

3 Proposed method

As stated in previous section, a generic sparse recovery problem can be expressed by (1). Here, we add non-negativity penalty to (1) which forms the new cost function as follows:

$$\min \|\mathbf{s}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{A}\mathbf{s}, \ \mathbf{s} \geq 0 \tag{2}$$

Since ℓ_0-norm is not differentiable, minimisation problem (2) cannot be directly solved. One traditional solution is to replace ℓ_0-norm with ℓ_1-norm so that optimisation-based techniques, e.g. those based on linear programming, could be used. However, as mentioned in previous section, these techniques are computationally expensive and researchers are looking for alternatives. **Our approach in this paper is inspired by SL0 method [39] where a smoothing function was proposed to directly minimises the ℓ_0-norm in a coarse to fine approach. Their proposed function, which symmetrically affects both negative and non-negative values, is defined as:**

$$f_\sigma(s) = 1 - \exp\left(\frac{-s^2}{2\sigma^2}\right) \tag{3}$$

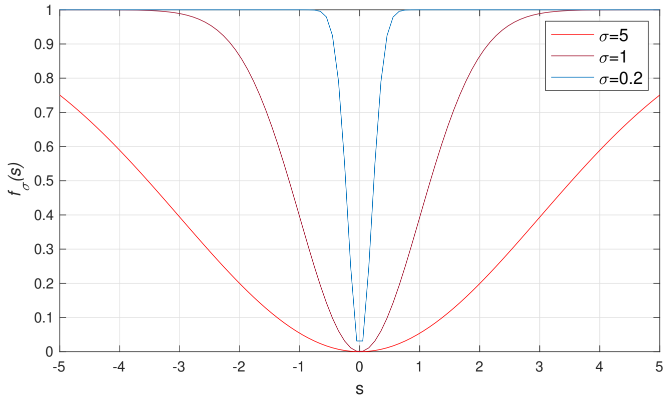


Figure 1: Sketch of smoothing function $f_\sigma(s)$ with three controller parameters. This function was used in [39] to convert ℓ_0-norm into a differentiable form.

where σ is a scalar parameter to control the degree of smoothness. Fig. 1 illustrates the shape of this function for three different σ's. According to this figure, as σ decreases the smoothness decreases, and the function becomes closer to exact ℓ_0-norm. In other words, $f_{\sigma=0}$ is equivalent to ℓ_0-norm problem (1), which is non-convex, and cannot be solved directly. The concept of embedding such a smoothing function into the original minimisation problem (1) is to relax this dilemma. Hence, taking (3) into account, the ℓ_0-norm minimisation problem (1) is approximated to:

$$\min \sum_{i=1}^{n} f_\sigma(s_i) \approx \|\mathbf{s}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{A}\mathbf{s} \tag{4}$$

which is convex and computationally inexpensive to solve (please refer to [39] for details of the minimisation process). While $f_\sigma(s)$ has shown to be very effective for solving ℓ_0-norm problem, it is not suitable for non-negative problems as it does not enforce any non-negative penalty (as can be observed from Fig. 1). Here, we design a different function to simultaneously apply smoothness and non-negativity, utilisable in (2). We aim to propose a differentiable function giving great flexibility to optimise the cost function as well as enforcing non-negativity. We start by modifying Fig. 1 so that $f(\cdot)$ be boosted for $s \leq 0$ while it remains unchanged for $s > 0$. In other words, our desire is to mathematically derive a function that can generate proposed curves in Fig. 2. As seen from Fig. 2, not only the proposed function incurs a large penalty to

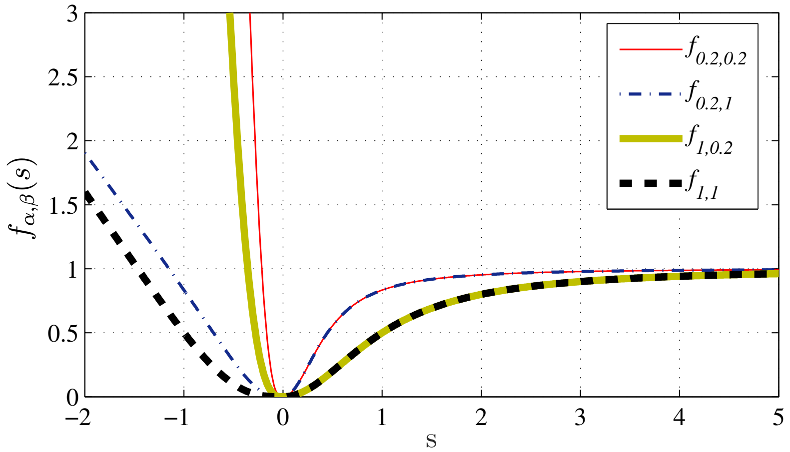


Figure 2: Function $f_{\alpha,\beta}(s)$ behaviour versus different values of s.

negative coefficients but the differentiability should be preserved. To do this, we start by reformulating non-negative penalty in (2) using the Lagrange method:

$$\min \sum_i (|s_i| + s_i)^0 + \lambda(|s_i| - s_i) \quad s.t. \quad \mathbf{y} = \mathbf{A}\mathbf{s}, \tag{5}$$

In order to provide a more precise description of the proposed cost function we rewrite it in a different form as follows:

$$f_{\alpha,\beta}(s) = \begin{cases} \frac{s^2}{s^2+\alpha} & s > 0 \\ 0 & s = 0 \\ \frac{|s|(\frac{|s|}{\beta})^{p+1}}{s^2+\alpha} & s < 0 \end{cases} \tag{6}$$

where s_i refers to i-th coefficient of vector $\mathbf{s}$, and the scalar λ is the Lagrange multiplier and defines the contribution of negative coefficients penalty to the whole cost function. For those coefficients in vector $\mathbf{s}$ in (5) that are negative (i.e. $s_i < 0$), the term $\lambda(|s_i| - s_i)$ turns into $2\lambda|s_i|$. This means that negative coefficients are imposed by a large penalty equal to 2λ. In contrast, if $s_i \geq 0$, then, $|s_i| - s_i = 0$, and therefore, no suppression is applied to the positive coefficients. This is desirable, as we aim not to impose any penalty rather than sparsity on positive coefficients to allow their natural evolution during the reconstruction procedure. However, the main challenge is to design a penalty function to simultaneously enforce sparsity as well as non-negativity on all coefficients. The term $(|s_i| + s_i)^0$ in (5) has been proposed for this purpose. It merely controls sparsness of positive coefficients and does not interfere the non-negativity penalty. If one defines $\lambda = \infty$ in (5), it turns into the non-negative problem (2). However, $(|s_i| + s_i)^0$ is not differentiable, and we cannot use this term directly as a plausible penalty. **Instead, we propose to add some new terms in form of numerators and a normalisation denominator, leading to the following function, which is differentaible and can generate our desired penalty function (as sketched in Fig. 2):**

$$f_{\alpha,\beta}(s) = \frac{1}{2}\frac{(|s| + s)s + (|s| - s)(\frac{|s|}{\beta})^{p+1}}{s^2 + \alpha} \tag{7}$$

where α, β, and p are positive scalars **to control the shape and smoothness of this function**. Notably, equation (7) presents working principle of the proposed penalty and it should be applied to all coefficients $s_i \in \{\mathbf{s}\}$. Fig. 2, illustrates several shapes of $f_{\alpha,\beta}(s)$ for selected values of α and β. As seen from this figure, the proposed function can provide a great flexibility in the amount of penalty that can be imposed on negative coefficients, while it does not have any significant impact on the positive coefficients.

As seen in (6), parameter α accounts for defining the sparsity degree. In other words, $\frac{s^2}{s^2+\alpha}$ is a smoothed version of ℓ_0-norm. Moreover, β is equivalent to λ in (5). If α tends to zero, then we will have:

$$\lim_{\alpha \to 0} f_{\alpha,\beta}(s) = \begin{cases} 1 & s > 0 \\ 0 & s = 0 \\ \frac{|s|^p}{\beta^{p+1}} & s < 0 \end{cases} \tag{8}$$

It is clear from the above equation that if α tends to zero, $f_{\alpha,\beta}(s)$ would be equivalent to ℓ_0-norm for positive values. In addition, when β tends to zero, a large amount of penalty is applied for negative values. It is important to note that parameter p controls the growing rate of the penalty imposing to negative values.

Now, we apply the defined function $f_{\alpha,\beta}(s)$ to the vector $\mathbf{s}$ and modify the optimisation problem (2) to:

$$\begin{gathered} \min F_{\alpha,\beta}(\mathbf{s}) = \min \textstyle\sum_i f_{\alpha,\beta}(s_i) = \\ \min \textstyle\sum_i \frac{1}{2}\frac{(|s_i|+s_i)s_i+(|s_i|-s_i)(|s_i|/\beta)^{p+1}}{s_i^2+\alpha} \;\; s.t. \; \mathbf{y} = \mathbf{A}\mathbf{s}. \end{gathered} \tag{9}$$

In order to solve the above optimisation problem we use the following steps:

1. Gradient descent algorithm (moving toward opposite direction of $\nabla F_{\alpha,\beta}(\mathbf{s})$)

2. Projection onto the constraints; non-negative-sparsity, and feasible set $\mathbf{y} = \mathbf{A}\mathbf{s}$.

These two steps start initially with large values for α and β, and then their values are gradually decreased. The initial solution of each step is taken from the result of the previous step. This process avoids the procedure to be trapped in local minima. On the other hand, small values of α and β in (8) is corresponding to (2) and (5). It is important to note that projection onto the three spaces, i.e. non-negativity, sparsity and $\mathbf{y} = \mathbf{A}\mathbf{s}$ is performed as follows. Values smaller than β in the non-negative and sparse domain are set to zero and then the result is projected onto the linear domain $\mathbf{y} = \mathbf{A}\mathbf{s}$. In practice, exact equality $\mathbf{y} = \mathbf{A}\mathbf{s}$ cannot be reachable, instead $\|\mathbf{y} - \mathbf{A}\mathbf{s}\|_2^2 \leq \epsilon$ is used. In order to impose this condition into the proposed cost function, inspired by SL0 method, the projection onto the linear space is performed when $\|\mathbf{y} - \mathbf{A}\mathbf{s}\|_2^2 \leq \epsilon$ does not meet [46]. The gradient of $F_{\alpha,\beta}(\mathbf{s})$ can be also computed as:

$$\nabla_s F_{\alpha,\beta}(\mathbf{s}) = [f'_{\alpha,\beta}(s_i)] \in \mathbb{R}^m \tag{10}$$

where f' is obtained via (11):

$$\begin{aligned} f'_{\alpha,\beta}(s) = 0.5((1 + sign(s)s + (s + |s|) + (sign(s) - 1)(\tfrac{|s|}{\beta})^{p+1} \\ + \tfrac{(p+1)sign(s)}{\beta}(|s| - s)(\tfrac{|s|}{\beta})^{p})(s^2 + \alpha) - 2s((|s| + s)s + (|s| - s)(\tfrac{|s|}{\beta})^{p+1}))(s^2 + \alpha)^{-2} \end{aligned} \tag{11}$$

Table 1 shows the summary of notations and symbols used in this paper. The pseudo-code of the proposed method (SnSA) is given in Algorithm

Table 1: **Summary of notations and symbols along with typical selected values.**

$\mathbf{s} \in \mathbb{R}^n$	sparse coefficients vector	k	number of non-zero coefficients
$\mathbf{A} \in \mathbb{R}^{m \times n}$	dictionary matrix	n	number of sparse coefficients
$\mathbf{y} \in \mathbb{R}^m$	raw input data vector	m	number of input samples
$\lambda > 0$	Lagrange multiplier	$\alpha > 10^{-9}$	smoothness controller scalar
$0 < \beta < 10$	penalty controller scalar	$p = 1$	penalty growing rate controller
ρ $(0.8 \sim 1)$	decreasing factor for α	$\gamma = 0.1$	non-negative penalty constant
$\mu = 0.001$	Gradient descent step size	$L = 5$	number of iterations
$\theta = 0.25$	estimator's threshold	ϵ	reconstruction error

Algorithm 1 Pseudo-code of the proposed SnSA.

Input: $\mathbf{A}$ and $\mathbf{y}$

Initialisation:

1. α_{min}, ρ (decreasing factor), μ, β_0, γ, L, $t = 1$.
2. $\hat{\mathbf{s}} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y}$
3. $\alpha = 2 \max |\hat{\mathbf{s}}|$
4. $\beta = \beta_0$

Output: $\hat{\mathbf{s}}$

repeat

for $i = 1$ to L **do**

(a) Gradient descent: $\hat{\mathbf{s}} \leftarrow \hat{\mathbf{s}} - \mu \nabla_{\mathbf{s}} F_{\alpha,\beta}(\hat{\mathbf{s}})$

(b) Projection:

- if $\hat{s}_i < \beta$ $(i = 1, ...m)$ then $\hat{s}_i = 0$
- if $\|\mathbf{y} - \mathbf{A}\mathbf{s}\|_2^2 > \epsilon$ then
  $\hat{\mathbf{s}} \leftarrow \hat{\mathbf{s}} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\mathbf{s}} - \mathbf{y})$

end for

$\alpha = \rho\alpha$

$\beta = \beta_0 \exp(-\gamma t)$

$t = t + 1$

until $\alpha > \alpha_{min}$

1. **During execution of SnSA, β acts as a suppressor of negative s_i coefficients. This can be graphically and mathematically observed by referring to Fig. 2 and equation (7), where as β decreases, the shape of $f(\cdot)$ is become closer to ℓ_0-norm, while preserving only non-negative coefficients. We cannot simply zero out negative s_i coefficients as the fidelity approximation, i.e. y $\approx$ As, would not be met. Instead, we aim to gradually reduce β in an iterative manner so that the algorithm smoothly converges. To implement this, we vary β using $\beta = \beta_0 \exp(-\gamma t)$ in Algorithm 1 to monotonically control the non-negative penalty contribution. Using this exponential function, β will be large at the initial iterations of the algorithm (i.e. small t), but once the iterations pro-**

ceed, it decreases to ultimately gets close to zero. Conceptually, this way, the amount of penalty on negative coefficients is increased as the iterations grow.

4 Experimental results

In this section, the proposed algorithm is numerically compared with two common methods BP [7] and SL0 [39], and their corresponding extended versions, i.e. non-negative BP (NNBP) [47] and constrained SL0 (CSL0) [42]. In addition, non-negative orthogonal matching pursuit (NNOMP) [43] is included as a greedy sparse recovery technique for comparison. **Further, two more relevant methods, i.e. orthogonal least square (OLS) [11] and Bayesian sparse coding known as relevance vector machine (RVM) [12], were included in our experiments.** Two sets of experiments are conducted in this section. First, synthetic signals are generated and extensive simulations have been carried out to study the performance of the proposed method. Furthermore, two real scenarios, i.e., face recognition and handwritten digits recognition, are examined by applying the proposed method and related techniques using several well-established databases. **Finally, a comprehensive comparison and performance evaluation between the proposed method and several deep learning models is provided.** All experiments were carried out under the same environmental conditions in MATLAB software on a Core(TM)i7-2.6GHz machine with 12GB of memory. The parameters for SnSA are empirically selected as follows: $\beta_0 = 10$, $\rho = 0.9$ $\gamma = 0.1$, $L = 5$, $\alpha_{min} = 10^{-9}$, $\mu = 0.001$. Moreover, we set $p = 1$ in our simulations unless specified otherwise.

4.1 Synthetic data

In the first experiment, we generated random dictionary ensembles $\mathbf{A}$ of size 50×150, and applied different reconstruction methods for recovery of sparse vector $\mathbf{s}$ with k non-zero samples. The experiment was repeated 1000 times (each time with a random $\mathbf{A}$ and $\mathbf{s}$) for k varying from 1 to 50. The average signal-to-noise-ratio (SNR) against k has been illustrated in Fig. 3 **with SnSA for** $p = 1$ **and** $p = 5$**, as well as other related methods**. It is observed that SnSA outperforms other methods especially for severe conditions, i.e. $15 \preceq k \preceq 30$. Robustness of SnSA against different selection of p is evident from this figure. **The second best performance belongs to CSL0 yet slightly weaker than SnSA.**

Next, the phase-transition diagrams are evaluated as a very important and well-established performance measure for sparse recovery techniques [48, 49]. These diagrams are generated for 500 trials for signal length $n = 128$ while varying

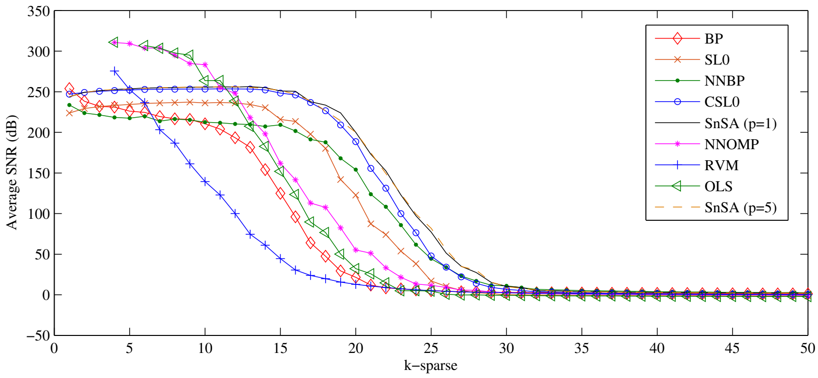


Figure 3: Reconstruction performance of different methods with random dictionary of size 50×100 for SnSA with both $p = 1$ and $p = 5$ and other relevant methods. Graphs with markers are associated to relevant methods.

measurement number m from 1 to $n/2$ and sparsity level k from 1 to $n/4$. The success rate was computed by giving a credit to the trials leading to reconstruction error less than 10^{-5}. The average success rates of all 500 independent trials for each point on the grid are sketched in Fig. 4. Darker areas correspond to higher success score and vice versa. The overlaid curves show the estimate at which the reconstruction is successful with probability $1 - \theta$. θ is the estimator's threshold set to $\theta = 0.25$ according to [50]. Fig. 5 illustrates the reconstruction performance among various relevant methods. It is seen from this figure that the performance of NNBP, BP, CSL0 and SL0 is comparable with that of SnSA when m and k are small. However, SnSA introduces higher success rate among all other techniques for larger m and k. This shows greater robustness of the proposed method.

Another aspect of advantage of SnSA is revealed by considering its performance against number of iterations. In this experiment, we conducted 100 trials of random ensembles with $\mathbf{A}$ of size 50×150 and $k = 10$. The reconstruction errors were then recorded against evolution of iterations. These results are plotted in Fig. 6 for three methods, i.e. SL0, CSL0, and SnSA, where all have iterative nature. It is seen from this figure that SnSA reaches to the minimum faster than other methods. Moreover, MSE of SnSA at iteration number 40 is about 0.00086 which is much less than that for SL0 and CSL0. It means that SnSA has a better convergence rate compared to other techniques.

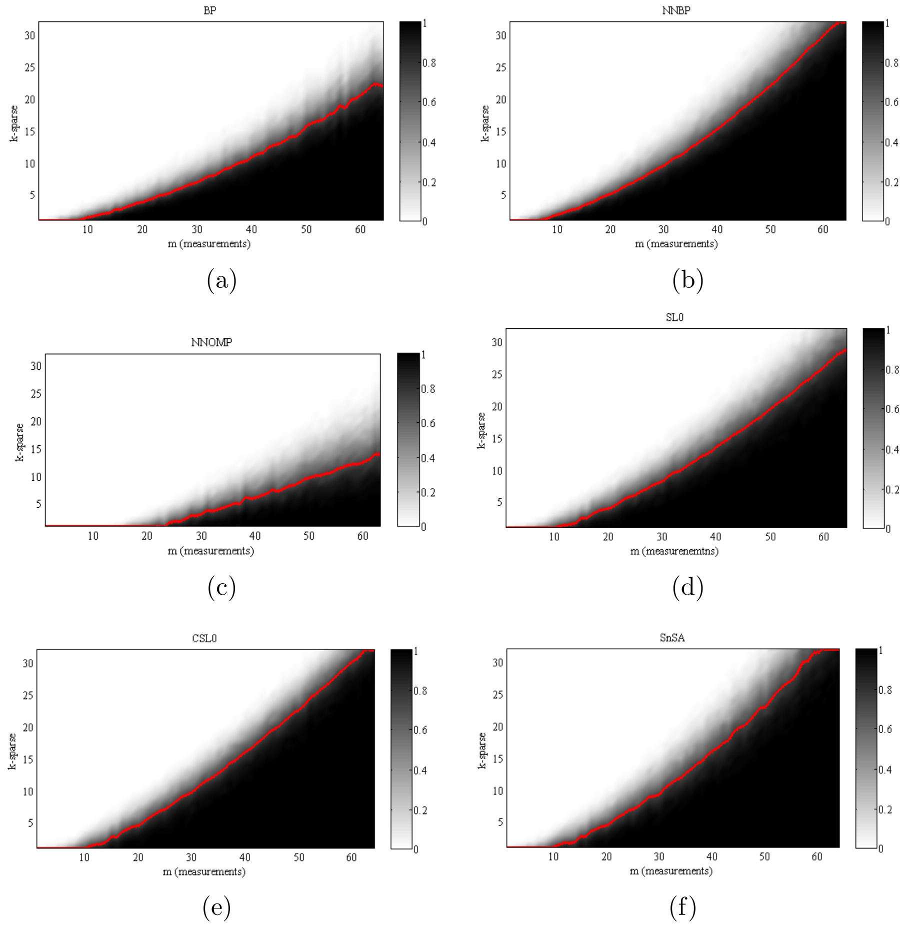


Figure 4: Phase transitions for (a) BP, (b) NNBP, (c) NNOMP, (d) SL0, (e) CSL0, and (f) SnSA. Darker areas correspond to higher success rate.

4.2 Real data

4.2.1 Face recognition

Four different face databases are considered here for evaluation of the proposed method in real scenarios. Some sample images of each database are given in Fig. 7. A brief description of these databases are:

- *Yale*: it contains 165 GIF images of 15 subjects of size 64×64. There are 11 images per subject, one for each of the following facial expressions or configurations: center-light, with glasses, happy, left-light, without glasses,

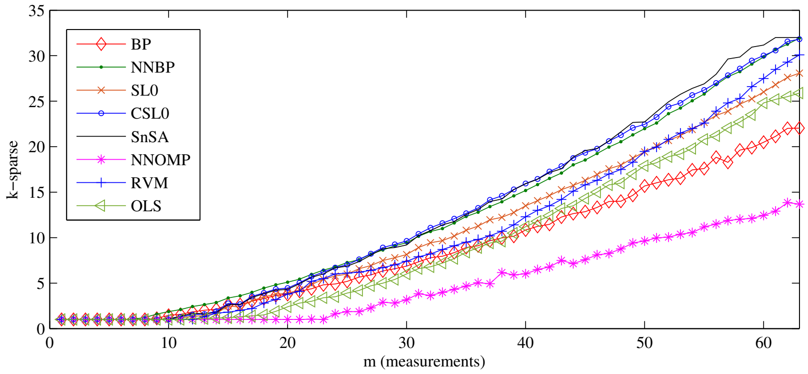


Figure 5: Comparison of different phase transitions.

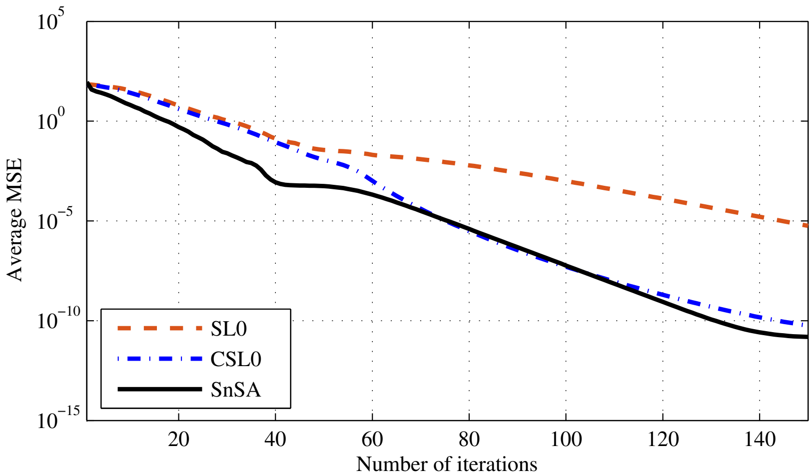


Figure 6: Average MSEs of different methods for 100 trials. (Dictionary Size: 50×100, $k = 10$, $p = 1$).

normal, right-light, sad, sleepy, surprised, and wink [51].

- *ORL*: it contains 400 images of size 48×48, 10 different images per person for 40 subjects. For some individuals, the images were acquired at different times. The facial expressions in these images are different, e.g. open or closed eyes and smiling or non-smiling. Other facial details such as glasses or no glasses also exist [52].

Table 2: Comparison of classification accuracy (%) for different methods using four face databases.

	YALE	CK+	AR	ORL
BP	85.32	84.76	87.10	94.37
NNBP	86.67	88.18	89.54	95.63
NNOMP	85.33	80.00	82.29	93.13
OLS	88.00	85.00	86.57	95.75
RVM	81.33	83.29	85.43	95.63
SL0	86.00	87.29	86.86	94.82
CSL0	86.67	93.33	89.71	95.75
SnSA	**91.33**	**96.67**	**92.00**	**96.88**

- *CK+*: it consists of 321 emotion sequences with labels (angry, contempt, disgust, fear, happiness, sadness, surprise). Images are of size 128×128 [53].
- *AR*: it consists of 4000 images corresponding to 126 people's faces (70 men and 56 women). The images size is 165×120. Images feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf) [54]. Here a subset of 50 males and 50 females are used.

For all four databases sparse representation classification (SRC) technique was used [32]. Following previous works, we assume for CK+ database that the information of neutral face is provided and subtract from all images both training and testing. Also, the preprocessings such as removing background have been applied to input images wherever needed prior applying the algorithms.

The average accuracies of classification of different facial expressions on four databases are given in Table 2. As seen from Table 2, SnSA outperforms with all databases. Inspection of this table confirms the overall improved performance achieved using the proposed method. In addition, non-negative-based methods generally give better results confirming the compatibility of these methods for non-negative data such as face images.

In the process of preparing the face images as input for the algorithms, there is a conventional stage of eigenface production. In this step, face images are projected onto a lower dimensional feature space, performed using principle comment analysis (PCA) technique [55]. This process greatly reduces the computational burden while preserving most important information of the images. However, selecting the dimension of lower space is challenging and could influence on the ultimate

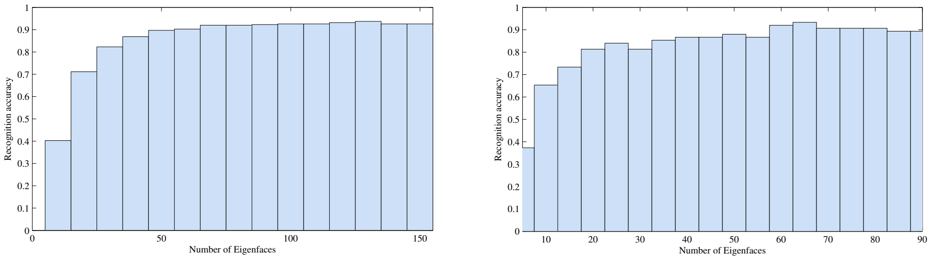
(a) Yale

(b) ORL

(c) CK+

(d) AR

Figure 7: Sample images from various databases.



(a) AR (b) YALE

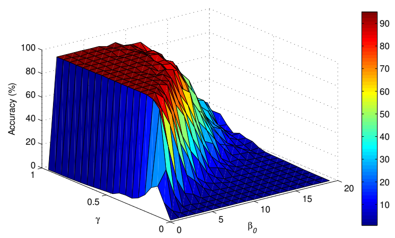
Figure 8: Average recognition rate of SnSA versus the number of selected eigenfaces. As seen, the accuracy becomes stable and maximised when the number of eigenfaces are more than 60.

results. We setup an experiment to illustrate how the reduced dimension was chosen. Based on observations, if the length of the feature vector to be higher than 50, the stable and optimal performance is guaranteed. These results are given in Fig. 8. We have chosen 80 for the number of eigenfaces in all experiments.

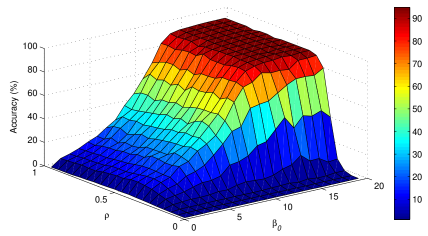
Next, we conduct an experiment to study the robustness of the proposed approach. We evaluated the influence of variation of key parameters, i.e. β_0, γ, ρ and L on the classification accuracy for AR database. In particular, we recorded the recognition accuracy while varying these parameters within a wide range and keeping other parameters fixed. The results of this experiment are depicted in Fig. 9. Following observations can be revealed by inspecting graphs in Fig. 9. SnSA is highly robust against variations of γ, β_0 and L, as observed from Fig. 9 (d), (e) and (f). Most sensitivity occurs where γ and ρ are changing while keeping other parameters fixed (Fig. 9 (c)). This is reasonable since γ is exponential index and ρ is the step-size of the outer loop (Algorithm 1). Hence, smaller values for ρ leads to a higher accuracy (Fig. 9 (c)). Also, inspecting Fig. 9 (a) and (b) implies that too small (too large) β_0 degrades the accuracy. Therefore, a moderate value for β_0 (e.g. $\beta_0 \approx 10$) would provide the best performance.

4.2.2 Handwritten Digits Recognition

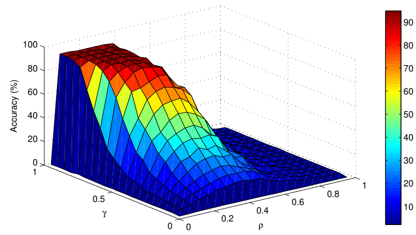
In this part, we investigate the effectiveness of SnSA and compare its recognition performance with related methods on a different data type, i.e., handwritten digits. We consider two databases for this purpose, i.e., MNIST and USPS. MNIST involves a training set of 60,000, and a test set of 10,000 grayscale image examples of digits ‘0’ through ‘9’. It is a subset of a larger set available from NIST. The digits have been size-normalised and centered in a fixed-size image [56]. USPS has



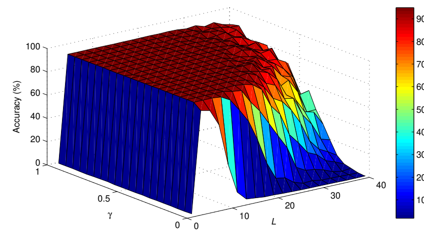
(a) γ and β_0



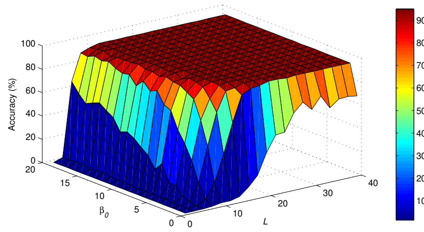
(b) ρ and β_0



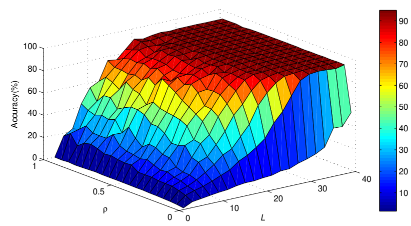
(c) γ and ρ



(d) γ and L

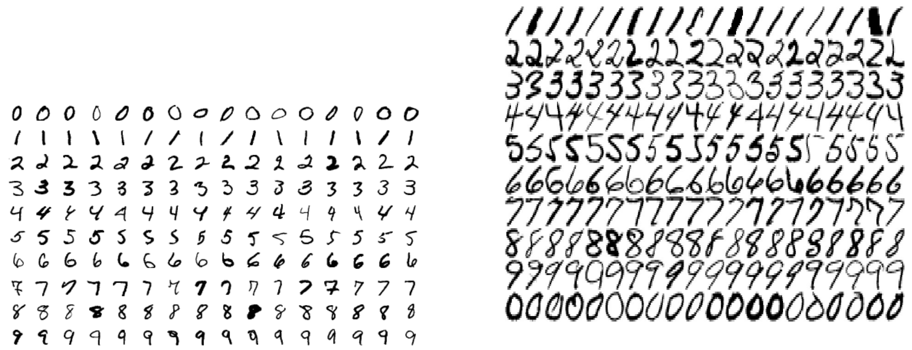


(e) β_0 and L



(f) ρ and L

Figure 9: The classification accuracy of SnSA versus variations of parameters β_0, γ, ρ, and L. We fixed $\mu = 0.001$ and $\alpha_{min} = 10^{-9}$ for all trials, and fixed $\gamma = 0.1$ $\beta_0 = 10$ $\rho = 1$ and $L = 5$ where needed at each specific sub-figure shown above.



(a) MNIST (b) USPS

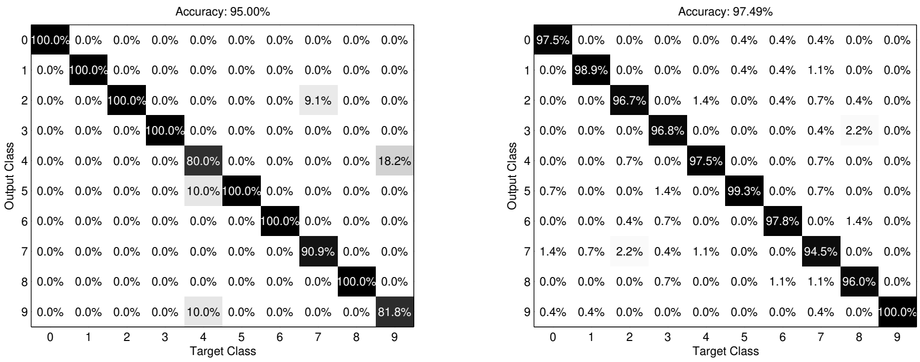
Figure 10: Sample grayscale images of handwritten digits. The images have made negative for ease of representation.

Table 3: Classification accuracy (%) and running time (ms) for different methods with MNIST and USPS handwritten digits database. The running time was calculated as the average reconstruction time per image.

	BP	NNBP	NNOMP	OLS	RVM	SL0	CSL0	SnSA
MNIST (%)	93.10	91.32	92.40	94.00	82.67	90.40	91.21	**94.52**
USPS (%)	95.28	93.11	94.87	95.30	96.50	94.68	95.98	**97.49**
Time (ms)	3126	1350	76.00	144.4	83.32	75.00	731.0	**52.00**

7291 train and 2007 test images of digits '0' through '9'. The images are 16-by-16 grayscale pixels [57]. Sample representations of these images for both databases are given in Fig. 10. Table 3 represents the classification results of applying several sparse recovery techniques within SRC for these databases. SnSA parameter settings were the same as those in the previous experiments. It can be observed from the results of Table 3 that the proposed method outperforms all other techniques. In particular, SnSA performs best among its non-negative competitors i.e. NNBP and NNOMP. Table 3 also reports the running times of different sparse recovery method per image. It is seen that SnSA is the fastest method among others. **Furthermore, the running time of RVM and SL0 are comparable with that of the proposed method.** As expected, BP achieved second highest accuracy in the table, however, it is the slowest by far among others due to its high computational complexity.

Finally, we depict the confusion matrix as a result of applying SnSA to MNIST and USPS databases in Fig. 11. As seen from Fig. 11 (a), classification accuracy is more that 90% in most classes except for digits ‘4’ and ‘9’. Precise inspection through the shape of these digits (Fig. 10 (a)) reveals high similarity between them which explains the reason of misclassification in Fig. 11 (a). However, this is not the case for USPS database as the classification accuracy for all classes are very good according to Fig. 11 (b).



(a) MNIST

(b) USPS

Figure 11: Confusion matrix for handwritten digits classification using SnSA.

4.3 Comparison with Deep learning models

The fast pacing developments of deep learning techniques has led to an increased tendency to embedding them in numerous problems such as pattern classification. Since mathematical developments proposed in this paper was utilised in face and handwritten recognition as potential applications, here we provide a comparison with state-of-the-art deep learning methods. To this end, five different architectures have been employed in our experiments: three pure convolutional neural networks (CNNs) with 1, 2, and 3 convolutional layer(s) under ReLU activation function, one LeNet-5 [58] with Sigmoid activation function, and one well-established pre-trained deep network, i.e., ResNet [59]. LeNet-5 has a convolution and subsampling layer that are alternated twice. All the models except ResNet have been locally trained using the datasets of interest in this work. ResNet (with 152 layers) was pre-trained on the large well-known ImageNet database and is adopted here using transfer

Table 4: Classification accuracy (%) among various deep neural network architectures and the proposed method with face and handwritten datasets.

	CNN-1	CNN-2	CNN-3	LeNet-5	ResNet152	SnSA
YALE	80.74	84.63	85.21	91.32	82.68	**91.33**
AR	81.73	86.91	92.55	**97.88**	96.75	92.00
ORL	87.33	88.67	89.53	88.37	92.33	**96.88**
CK+	74.88	81.04	73.46	76.30	85.00	**96.67**
MNIST	97.45	98.33	**98.62**	97.13	97.86	94.52
USPS	89.78	89.57	89.69	71.10	95.51	**97.49**

learning technique to work with our datasets. Table 4 depicts the results of this experiment with all the face and handwritten datasets used in this paper. According to this table, the proposed method has achieved highest accuracy with all datasets except with AR and MNIST. We reasonably believe that this is mainly dependent on the scale of the dataset. In fact, deep learning methods naturally perform weaker on small datasets such as YALE, ORL, and CK+. Nevertheless, deep networks present greater performance with large-scale datasets such as AR and MNIST. Also, pre-trained network, i.e. ResNet152, has slightly

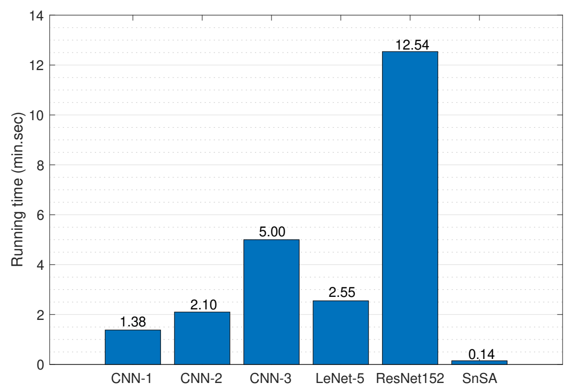


Figure 12: **Comparative analysis of the running time(s) elapsed to train various deep models and the proposed method with YALE dataset. Learning rate and number of epochs were 0.001 and 40, respectively, for deep neural network models.**

improved the performance with ORL, MNIST or USPS, but still performs weaker than SnSA.

Unlike deep network models which mainly require high power computers, our proposed method runs locally and fast on general-purpose computers. Figure 12 provides a comparative illustration of the processing time elapsed for training the model with YALE face dataset. As seen from this figure, as the depth of neural network increases the running time also increases dramatically. Figure 12 shows that complex networks like ResNet152 takes significantly longer to be trained even with datasets like YALE which includes only 165 images of small sizes 32×32. In contrast, Figure 12 shows that the proposed method is $\times 5$ and $\times 50$ faster than CNN-1 and ResNet152, respectively. Moreover, well-framed deep models require enormous number of parameters (e.g. ResNet with 25 million parameters), while the proposed method only requires 6 parameters to be fine-tuned. In summary, the proposed method is preferred when small datasets and less computing resources are available.

5 Conclusions

In this paper, a novel technique for non-negative sparse recovery problem was presented. A smooth non-negative function was proposed for this purpose. This convex function allows existence of negative coefficients at initial iterations which are gradually suppressed until a non-negative solution is achieved. The main advantages of proposed SnSA compared to CSL0 are as follows. The penalty term of non-negative coefficients in SnSA has the convex form and therefore is differentiable. The thresholding step is embedded into the optimisation. These properties result in better convergence and higher performance as explored through our extensive experiments. In addition, the superiority of the proposed method for real-world applications of face recognition and handwritten digits recognition with several well-established databases were verified. **It was observed that the proposed method outperforms deep learning methods on small-scale datasets, and performs competitively when large-scale datasets are available. We are interested and aim to further study how the proposed method can be utilised as a complementary algorithm, e.g. activation function, contributing as a layer within deep learning techniques. This will also provide further opportunity to investigate the utilisation of the proposed approach in deep dictionary learning framework.**

References

[1] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.* Springer, 2010.

[2] F. Mokhayeri and E. Granger, “A paired sparse representation model for robust face recognition from a single sample,” *Pattern Recognition*, vol. 100, p. 107129, 2020.

[3] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, “A survey of sparse representation: Algorithms and applications,” *IEEE Access*, vol. 3, pp. 490–530, 2015.

[4] M. A. Nazari Siahsar, S. Gholtashi, V. Abolghasemi, and Y. Chen, “Simultaneous denoising and interpolation of 2d seismic data using data-driven nonnegative dictionary learning,” *Signal Processing*, vol. 141, pp. 309–321, 2017.

[5] S. Ferdowsi, V. Abolghasemi, B. Makkiabadi, and S. Sanei, “A New Spatially Constrained NMF With Application To FMRI,” in *33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society – EMBC'11*, August 2011.

[6] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, October 1999.

[7] S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[8] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Transactions on Information Theory*, vol. 53, pp. 4655–4666, December 2007.

[9] T. Blumensath and M. E. Davies, “Iterative thresholding for sparse approximations,” *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 629–654–654, 2008.

[10] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communciation Pure Application*, vol. LVII, pp. 1413–1457, 2004.

[11] A. Hashemi and H. Vikalo, “Sparse recovery via orthogonal least-squares under presence of noise,” 2016.

[12] W. Wang, D. Sun, P. Shao, H. Kuang, and C. Sui, "Fully bayesian analysis of relevance vector machine classification with probit link function for imbalanced data problem," *IEEE Access*, vol. 9, pp. 77451–77463, 2021.

[13] J. Xu, W. An, L. Zhang, and D. Zhang, "Sparse, collaborative, or nonnegative representation: Which helps pattern classification?," *Pattern Recognition*, vol. 88, pp. 679 – 688, 2019.

[14] R. Abiantun, F. Juefei-Xu, U. Prabhu, and M. Savvides, "SSR2: Sparse signal recovery for single-image super-resolution on faces with extreme low resolutions," *Pattern Recognition*, vol. 90, pp. 308 – 324, 2019.

[15] H. Yuan, J. Li, L. L. Lai, and Y. Y. Tang, "Joint sparse matrix regression and nonnegative spectral analysis for two-dimensional unsupervised feature selection," *Pattern Recognition*, vol. 89, pp. 119 – 133, 2019.

[16] H. Lu, Z. Fu, and X. Shu, "Non-negative and sparse spectral clustering," *Pattern Recognition*, vol. 47, no. 1, pp. 418 – 426, 2014.

[17] B. Jiang, H. Zhao, J. Tang, and B. Luo, "A sparse nonnegative matrix factorization technique for graph matching problems," *Pattern Recognition*, vol. 47, no. 2, pp. 736 – 747, 2014.

[18] S. Zhang, J. Wang, W. Shi, Y. Gong, Y. Xia, and Y. Zhang, "Normalized nonnegative sparse encoder for fast image representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 7, pp. 1962–1972, 2019.

[19] M. A. Khajehnejad, A. G. Dimakis, W. Xu, and B. Hassibi, "Sparse recovery of nonnegative signals with minimal expansion," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 196–208, 2011.

[20] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "This is spiral-tap: Sparse poisson intensity reconstruction algorithms—theory and practice," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1084–1096, 2012.

[21] R. M. Willett, M. F. Duarte, M. A. Davenport, and R. G. Baraniuk, "Sparsity and structure in hyperspectral imaging : Sensing, reconstruction, and target detection," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 116–126, 2014.

[22] R. Gu and A. Dogandzic, "Reconstruction of nonnegative sparse signals using accelerated proximal-gradient algorithms," 2015.

[23] K. O'Hanlon, M. D. Plumbley, and H. Nagano, "Group Non-negative Basis Pursuit for Automatic Music Transcription," in *5th International Workshop on Machine Learning and Music (MML12)*, (Edinburgh, United Kingdom), June 2012.

[24] P. Teng and Y. Jia, "Voice activity detection via noise reducing using non-negative sparse coding," *IEEE Signal Processing Letters*, vol. 20, pp. 475–478, May 2013.

[25] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* John Wiley, 2009.

[26] M. Rajapakse and L. Wyse, "Face recognition with non-negative matrix factorization," vol. 5150, pp. 1838–1847, SPIE, 2003.

[27] M. Zhang and Z. Zhou, "Structural deep nonnegative matrix factorization for community detection," *Applied Soft Computing*, vol. 97, p. 106846, 2020.

[28] Z. Yang, Y. Zhang, W. Yan, Y. Xiang, and S. Xie, "A fast non-smooth nonnegative matrix factorization for learning sparse representation," *IEEE Access*, vol. 4, pp. 5161–5168, 2016.

[29] B. Sabzalian and V. Abolghasemi, "Iterative weighted non-smooth non-negative matrix factorization for face recognition," *International Journal of Engineering*, vol. 31, no. 10, pp. 1698–1707, 2018.

[30] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with ℓ_0-constraints," *Neurocomputing*, vol. 80, pp. 38 – 46, 2012. Special Issue on Machine Learning for Signal Processing 2010.

[31] N. Binesh and M. Rezghi, "Fuzzy clustering in community detection based on nonnegative matrix factorization with two novel evaluation criteria," *Applied Soft Computing*, vol. 69, pp. 689 – 703, 2018.

[32] 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, pp. 210–227, Feb. 2009.

[33] L. Yuan, W. Liu, and Y. Li, "Non-negative dictionary based sparse representation classification for ear recognition with occlusion," *Neurocomputing*, vol. 171, pp. 540 – 550, 2016.

[34] W. Zhu and Y. Yan, "Label and orthogonality regularized non-negative matrix factorization for image classification," *Signal Processing: Image Communication*, vol. 62, pp. 139 – 148, 2018.

[35] X. Zhang, Q. Yang, M. Liu, Y. Jia, S. Liu, and G. Li, "Aspect-aided dynamic non-negative sparse representation-based microwave image classification," in *Sensors*, 2016.

[36] R. Lan, Z. Li, Z. Liu, T. Gu, and X. Luo, "Hyperspectral image classification using k-sparse denoising autoencoder and spectral-restricted spatial characteristics," *Applied Soft Computing*, vol. 74, pp. 693 – 708, 2019.

[37] P. H. Kassani and A. B. J. Teoh, "A new sparse model for traffic sign classification using soft histogram of oriented gradients," *Applied Soft Computing*, vol. 52, pp. 231 – 246, 2017.

[38] S. Zhang, C. Zhang, Z. Wang, and W. Kong, "Combining sparse representation and singular value decomposition for plant recognition," *Applied Soft Computing*, vol. 67, pp. 164 – 171, 2018.

[39] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm," *IEEE Transactions on Signal Processing*, vol. 57, pp. 289–301, January 2009.

[40] A. Ghaffari, M. Babaie-Zadeh, and C. Jutten, "Sparse decomposition of two dimensional signals," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3157–3160, IEEE, 2009.

[41] A. Atamturk, A. Gomez, and S. Han, "Sparse and smooth signal estimation: Convexification of l0 formulations," 2018.

[42] M. Mohammadi, E. Fatemizadeh, and M. Mahoor, "Non-negative sparse decomposition based on constrained smoothed $\ell 0$ norm," *Signal Processing*, vol. 100, pp. 42 – 50, 2014.

[43] A. M. Bruckstein, M. Elad, and M. Zibulevsky, "Sparse non-negative solution of a linear system of equations is unique," in *2008 3rd International Symposium on Communications, Control and Signal Processing*, pp. 762–767, March 2008.

[44] M. Yaghoobi, D. Wu, and M. E. Davies, "Fast non-negative orthogonal matching pursuit," *IEEE Signal Processing Letters*, vol. 22, pp. 1229–1233, Sept 2015.

[45] R. Kung and P. Jung, “Robust nonnegative sparse recovery and 0/1-bernoulli measurements,” in *2016 IEEE Information Theory Workshop (ITW)*, pp. 260–264, Sept 2016.

[46] A. Eftekhari, M. Babaie-Zadeh, C. Jutten, and H. A. Moghaddam, “Robust-SL0 for stable sparse representation in noisy settings,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3433–3436, April 2009.

[47] D. L. Donoho and J. Tanner, “Sparse nonnegative solutions of underdetermined linear equations by linear programming,” in *Proceedings of the National Academy of Sciences*, pp. 9446–9451, 2005.

[48] D. Donoho and J. Tanner, “Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing,” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[49] D. L. Donoho and J. Tanner, “Precise undersampling theorems,” *Proceedings of the IEEE*, vol. 98, pp. 913–924, June 2010.

[50] D. L. Donoho, I. Drori, V. Stodden, and Y. Tsaig, “Sparselab,” December 2005.

[51] P. N. Belhumeur, J. a. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 711–720, July 1997.

[52] F. S. Samaria and A. C. Harter, “Parameterisation of a stochastic model for human face identification,” in *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pp. 138–142, Dec 1994.

[53] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, June 2010.

[54] A. Martinez and R. Benavente, “The ar face database,” in *CVC Technical Report No.24*, June 1998.

[55] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, Jan. 1991.

[56] “MNIST: Database of handwritten digits.” http://yann.lecun.com/exdb/mnist/. Accessed: 01.10.2020.

[57] “USPS: Database of handwritten digits.” https://cs.nyu.edu/ roweis/data.html. Accessed: 01.10.2020.

[58] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[59] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

- Nonnegative smoothed L0 (SL0) for sparse recovery is proposed.
- The proposed cost function has a convex form and hence differentiable.
- Numerical experiments demonstrate high performance and robustness.
- Successful performance on face and handwritten recognition has been verified.

Credit Author Statement:

Aboozar Ghaffari: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Writing- Reviewing and Editing **Mahdi Kafaee**: Conceptualization, Resources, Software, Writing- Reviewing and Editing, **Vahid Abolghasemi**: Methodology, Software, Validation, Visualization, Investigation, Data curation, Writing- Original draft, Project administration.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: