algorithms

MDPI

Article

A Linearly Involved Generalized Moreau Enhancement of $\ell_{2,1}$-Norm with Application to Weighted Group Sparse Classification

Yang Chen, Masao Yamagishi and Isao Yamada *

Department of Information and Communications Engineering, Tokyo Institute of Technology, 2-12-1 Okayama, Meguro-ku, Tokyo 152-8552, Japan; chen@sp.ict.e.titech.ac.jp (Y.C.); myamagi@ict.e.titech.ac.jp (M.Y.)
* Correspondence: isao@ict.e.titech.ac.jp

Abstract: This paper proposes a new group-sparsity-inducing regularizer to approximate $\ell_{2,0}$ pseudo-norm. The regularizer is nonconvex, which can be seen as a linearly involved generalized Moreau enhancement of $\ell_{2,1}$-norm. Moreover, the overall convexity of the corresponding group-sparsity-regularized least squares problem can be achieved. The model can handle general group configurations such as weighted group sparse problems, and can be solved through a proximal splitting algorithm. Among the applications, considering that the bias of convex regularizer may lead to incorrect classification results especially for unbalanced training sets, we apply the proposed model to the (weighted) group sparse classification problem. The proposed classifier can use the label, similarity and locality information of samples. It also suppresses the bias of convex regularizer-based classifiers. Experimental results demonstrate that the proposed classifier improves the performance of convex $\ell_{2,1}$ regularizer-based methods, especially when the training data set is unbalanced. This paper enhances the potential applicability and effectiveness of using nonconvex regularizers in the frame of convex optimization.

Keywords: convex optimization; proximal splitting algorithm; generalized Moreau enhancement; group sparsity; weighted $\ell_{2,1}$-norm; sparse representation-based classification

Citation: Chen, Y.; Yamagishi, M.; Yamada, I. A Linearly Involved Generalized Moreau Enhancement of $\ell_{2,1}$-Norm with Application to Weighted Group Sparse Classification. *Algorithms* **2021**, *14*, 312. https://doi.org/10.3390/a14110312

Academic Editor: Sorin-Mihai Grad

Received: 10 September 2021
Accepted: 26 October 2021
Published: 27 October 2021

1. Introduction

In recent decades, sparse reconstruction has become an active topic in many areas, such as in fields of signal processing, statistics, and machine learning [1]. By reconstructing a sparse solution from a linear measurement, we can obtain a certain expression of high-dimensional data as a vector with only a small number of nonzero entries. In practical applications, the data of interest can often be assumed to have a certain special structure. For example, in microarray analysis of gene expression [2,3], hyperspectral image unmixing [4–6], force identification in industrial applications [7], classification problems [8–13], etc., the solution of interest often possesses group-sparsity structure, namely the solution has a natural grouping of its coefficients and nonzero entries only occur in few groups.

This paper focuses on the estimation of group sparse solution, which is related to the Group LASSO (least absolute shrinkage and selection operator) [14]. Suppose $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_g^\top]^\top \in \mathbb{R}^n$ is a group sparse signal, where $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$, $\sum_{i=1}^g n_i = n$ and g is the number of groups. Just as with the use of ℓ_0 pseudo-norm for evaluation of the sparsity, the group sparsity of $\boldsymbol{x}$ can be evaluated with the $\ell_{2,0}$ pseudo-norm, i.e., $\|\boldsymbol{x}\|_{2,0} = \big\|\big(\|\boldsymbol{x}_1\|_2, \|\boldsymbol{x}_2\|_2, \cdots, \|\boldsymbol{x}_g\|_2\big)\big\|_0$, where $\|\cdot\|_2$ is the Euclidean norm, and $\|\cdot\|_0$ is the ℓ_0 pseudo-norm which counts the number of nonzero entries in the vector in $\mathbb{R}^g$.

The group sparse regularized least squares problem can be modeled as

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_{2,0}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ are known, and $\lambda > 0$ is the regularization parameter. However, the employment of the pseudo-norm $\ell_{2,0}$ makes (1) NP-hard [15]. Most studies in the application replace the nonconvex regularizer $\ell_{2,0}$ with its tightest convex envelope $\ell_{2,1}$ [16] (or its weighted variants), and the following regularized least squares problem has been proposed known as the Group LASSO [14],

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \sum_{i=1}^{g} w_i \|\boldsymbol{x}_i\|_2, \tag{2}$$

where $w_i > 0$ $(i = 1, \cdots, g)$ in the regularization term

$$\|\boldsymbol{x}\|_{w,2,1} := \sum_{i=1}^{g} w_i \|\boldsymbol{x}_i\|_2 \tag{3}$$

(this is $\ell_{w,2,1}$-norm of $\boldsymbol{x}$, i.e., a separable weighted version [17] of $\ell_{2,1}$-norm $\|\boldsymbol{x}\|_{2,1} = \sum_{i=1}^{g} \|\boldsymbol{x}_i\|_2$) are used to adjust for group sizes with $w_i = \sqrt{n_i}$ in [14,18]. We give a simple but clear explanation in Appendix A, to show the bias of $\ell_{2,1}$-norm caused by group size in the application of group sparse classification (GSC).

Although the convex optimization problem (2) has been used as a standard model for group sparse estimation applications, the convex regularizer $\ell_{w,2,1}$ does not necessarily promote group sparsity sufficiently, mainly due to the fact that $\ell_{w,2,1}$-norm is just an approximation of $\ell_{2,0}$ pseudo-norm within the severe restriction of the convexity. To promote the group sparsity more effectively than convex regularizers, nonconvex regularizers such as group SCAD (smoothly clipped absolute deviation) [3], group MCP (minimax concave penalty) [18,19], $\ell_{p,q}$ regularization ($\|\boldsymbol{x}\|_{p,q} := \left(\sum_{i=1}^{g} \|\boldsymbol{x}_i\|_p^q\right)^{1/q}$, $0 < q < 1 \leq p$) [20], iterative weighted group minimization [21], and $\ell_{2,0}$ [22] have been used for group sparse estimation problems. However, they lose the overall convexity (In [23], a nonconvex regularizer which can preserve the overall convexity was proposed, but the fidelity term of the optimization model is $\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2$ (limited to the case of $\boldsymbol{A} = \boldsymbol{I}_n$, where $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix), which cannot be applied to (1) for general $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.) of the optimization problems, which results in their algorithms of no guarantee of convergence to global minimizers of the overall cost functions.

In this paper, we propose a generalized weighted group sparse estimation model based on the linearly involved generalized-Moreau-enhanced (LiGME) approach [24] that uses nonconvex regularizer while maintaining the overall convexity of the optimization problem. Our contributions can be summarized as follows:

- We show in Proposition 2 that the generalized Moreau enhancement (GME) of $\ell_{w,2,1}$, i.e., $(\|\cdot\|_{w,2,1})_{\boldsymbol{B}}$ (see (11)), can bridge the gap between $\ell_{w,2,1}$ and $\ell_{2,0}$. For the nonseparable weighted $\ell_{2,1}$, i.e., $\|\boldsymbol{W}\cdot\|_{2,1}$, its GME can be expressed as LiGME of $\ell_{2,1}$ in the case of weight matrix $\boldsymbol{W}$ has full row-rank.
- We present a convex regularized least squares model with a nonconvex group sparsity promoting regularizer based on LiGME. It can be served as a unified model of many types of group sparsity related applications.
- We illustrate the unfairness of $\ell_{2,1}$ regularizer in unbalanced classification and then apply the proposed model to reduce the unfairness of it in GSC and weighted GSC (WGSC) [11].

The remainder of this paper is organized as follows. In Section 2, we give a brief review of LiGME model and WGSC method. In Section 3, we present our group sparse enhanced representation model and its mathematical properties. In Section 4, we apply the proposed model to group-sparsity-based classification problems. The conclusion is given in Section 5.

A preliminary short version of this paper was presented at a conference [25].

2. Preliminaries

2.1. Review of Linearly Involved Generalized-Moreau-Enhanced (LiGME) Model

We first give a brief review of linearly involved generalized-Moreau-enhanced (LiGME) models, which is closely related to our method. Although the convex function ℓ_1-norm (or nuclear norm) is the most frequently adopted regularizer for sparsity (or low-rank) pursuing problems, it tends to yield underestimation for high-amplitude value (or large singular value) [26,27]. The convexity-preserving nonconvex regularizers have been widely explored in [24,28–33], which promote sparsity (or low-rank) more effectively than convex regularizers without losing the overall convexity. Among them, the generalized minimax concave (GMC) function in [31] does not rely on certain strong assumptions in the least squares term and has great potential for dealing with nonconvex variations of $\|\cdot\|_1$. Motivated by GMC function, the LiGME model [24] provides a general framework for constructing linearly involved nonconvex regularizers for sparsity (or low-rank) regularized linear least squares while maintaining the overall convexity of the cost function.

Let $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}}, \|\cdot\|_{\mathcal{X}})$, $(\mathcal{Y}, \langle\cdot,\cdot\rangle_{\mathcal{Y}}, \|\cdot\|_{\mathcal{Y}})$, $(\mathcal{Z}, \langle\cdot,\cdot\rangle_{\mathcal{Z}}, \|\cdot\|_{\mathcal{Z}})$, and $(\tilde{\mathcal{Z}}, \langle\cdot,\cdot\rangle_{\tilde{\mathcal{Z}}}, \|\cdot\|_{\tilde{\mathcal{Z}}})$ be finite-dimensional real Hilbert spaces. Let a function $\Psi \in \Gamma_0(\mathcal{Z})$ be coercive with $\mathrm{dom}\Psi = \mathcal{Z}$. Here $\Gamma_0(\mathcal{Z})$ is the set of proper (i.e., $\mathrm{dom}\Psi := \{z \in \mathcal{Z} | \Psi(z) < \infty\} \neq \varnothing$) lower semicontinuous (i.e., $\mathrm{lev}_{\leq a}\Psi := \{z \in \mathcal{Z} | \Psi(z) \leq a\}$ is closed for $\forall a \in \mathbb{R}$) convex function (i.e., $\Psi(\theta z_1 + (1-\theta)z_2) \leq \theta\Psi(z_1) + (1-\theta)\Psi(z_2)$) for $\forall z_1, z_2 \in \mathrm{dom}\Psi, 0 \leq \theta \leq 1$) from $\mathcal{Z}$ to $(-\infty, \infty]$; a function $\Psi \in \Gamma_0(\mathcal{Z})$ is called coercive if $\|z\|_2 \to \infty \Rightarrow \Psi(z) \to \infty$. For $\Psi \in \Gamma_0(\mathcal{Z})$, the proximity operator of Ψ is defined by $\mathrm{Prox}_{\Psi} : \mathcal{Z} \to \mathcal{Z} : z \mapsto \arg\min_{v \in \mathcal{Z}} \Psi(v) + \frac{1}{2}\|v - z\|_{\mathcal{Z}}^2$.

The generalized Moreau enhancement (GME) of Ψ with $\boldsymbol{B} \in \mathcal{B}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is defined as

$$\Psi_{\boldsymbol{B}}(\cdot) := \Psi(\cdot) - \min_{v \in \mathcal{Z}} \left[\Psi(v) + \frac{1}{2}\|\boldsymbol{B}(\cdot - v)\|_{\tilde{\mathcal{Z}}}^2 \right], \tag{4}$$

where $\boldsymbol{B}$ is a tuning matrix for the enhancement. Then the LiGME model defined as the minimization of

$$J_{\Psi_{\boldsymbol{B}} \circ \mathfrak{L}} : \mathcal{X} \to \mathbb{R} : \boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\mathcal{Y}}^2 + \lambda \Psi_{\boldsymbol{B}} \circ \mathfrak{L}(\boldsymbol{x}), \tag{5}$$

where $(\boldsymbol{A}, \mathfrak{L}, \lambda) \in \mathcal{B}(\mathcal{X}, \mathcal{Y}) \times \mathcal{B}(\mathcal{X}, \mathcal{Z}) \times \mathbb{R}_+$.

Please note that GMC [31] can be seen as a special case of (5) with $\Psi = \|\cdot\|_1$ and $\mathfrak{L} = \mathrm{Id}$, where Id is the identity operator. Model (5) can also be seen as an extension of [32,33].

Although the GME function Ψ_B in (4) is not convex in general for $\boldsymbol{B} \neq \boldsymbol{O}_{\mathcal{B}(\mathcal{Z}, \tilde{\mathcal{Z}})}$, where $\boldsymbol{O}_{\mathcal{B}(\mathcal{Z}, \tilde{\mathcal{Z}})} \in \mathcal{B}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is the zero operator, the overall convexity of the cost function (5) can be achieved with $\boldsymbol{B}$ designed to satisfy the following convexity condition.

Proposition 1 ([24], Proposition 1)**.** *The cost function $J_{\Psi_{\boldsymbol{B}} \circ \mathfrak{L}}$ in (5) belongs to $\Gamma_0(\mathcal{X})$ for any $\boldsymbol{y} \in \mathcal{Y}$, if the GME regularizer $\Psi_{\boldsymbol{B}}$ in (4) satisfies that*

$$\boldsymbol{A}^*\boldsymbol{A} - \lambda \mathfrak{L}^* \boldsymbol{B}^* \boldsymbol{B} \mathfrak{L} \succeq O_{\mathcal{X}}, \tag{6}$$

where $\boldsymbol{A}^*$ denotes the adjoint of $\boldsymbol{A}$ and $\boldsymbol{O}_{\mathcal{X}} \in \mathcal{B}(\mathcal{X}, \mathcal{X})$ is the zero operator. In particular, when Ψ is a certain norm over the vector space $\mathcal{Z}$, $J_{\Psi_{\boldsymbol{B}} \circ \mathfrak{L}} \in \Gamma_0(\mathcal{X})$ if and only if (6) is satisfied.

A method of designing $\boldsymbol{B}$ satisfying (6) for $\mathcal{X} = \mathbb{R}^n$ is provided in [24]; see Proposition A1 in Appendix B. For any $\Psi \in \Gamma_0(\mathcal{Z})$ that is coercive, even symmetry and prox-friendly (Even symmetry means $\Psi \circ (-\mathrm{Id}) = \Psi$; prox-friendly means $\mathrm{Prox}_{\gamma\Psi}$ is computable ($\forall \gamma \in \mathbb{R}_{++}$).) with $\mathrm{dom}\Psi = \mathcal{Z}$, [24] provides a proximal splitting algorithm (see Proposition A2 in Appendix B) of guaranteed convergence to a globally optimal solution of model (5) under the overall-convexity condition (6).

2.2. Basic Idea of Weighted Group Sparse Classification (WGSC)

As a relatively simple but typical scenario for the application of the proposed idea in this paper, we introduce the main idea of weighted group sparse classification (WGSC). Classification is one of fundamental tasks in the field of the signal and image processing and pattern recognition. For a classification problem with g classes of subjects, the training samples can formulate a dictionary matrix $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_g] \in \mathbb{R}^{m\times n}$, where $\boldsymbol{A}_i = [\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2}, \cdots, \boldsymbol{a}_{in_i}] \in \mathbb{R}^{m\times n_i}$ is the subset of the training samples from subject i, $\boldsymbol{a}_{ij}$ is the j-th training sample from the i-th class, n_i is the number of training samples from class i, and $n = \sum_{i=1}^{g} n_i$ is the number of total training samples. The aim is to correctly determine which class the input test sample $\boldsymbol{y} \in \mathbb{R}^m$ belongs to. Although deep learning is very popular and powerful for classification tasks, it requires a very large-scale training set and computation resources for numerous parameters training with complicated back-propagation.

Wright et al. proposed the sparse representation-based classification (SRC) [34] for face recognition. With the assumption that samples of a specific subject lie in a linear subspace, a valid test sample $\boldsymbol{y}$ is expected to be approximated well by a linear combination of the training samples from the same class, which leads to a sparse representation coefficient over all training samples. Specifically, the test sample $\boldsymbol{y}$ is approximated by the linear combination of the dictionary items, i.e., $\boldsymbol{y} \approx \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}$ is the coefficient vector. A simple minimization model with sparse representation can be $\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_0$. In most SRC-based approaches, ℓ_0 regularizer is relaxed to ℓ_1, and the model becomes the well-known LASSO model [35] in statistics.

The label information of the dictionary atoms is not used in the simple model of SRC, hence the regression is based solely on the structure of each sample. When the subspaces spanned by different classes are not independent, SRC may lead the test image to be represented by training samples from multiple different classes. Considering ideal situation where the test image should only be approximated well by the training samples from one class corresponding to the correct one, in [8–10], the authors divided training samples into groups by prior label information and used group-sparsity regularizers. Naturally, the coefficient vector $\boldsymbol{x}$ has group structure $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_g^\top]^\top \in \mathbb{R}^n$, where $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in_i}]^\top \in \mathbb{R}^{n_i}$ $(i = 1, 2, \cdots, g)$. This kind of group sparse classification (GSC) approach aims to represent the test image using the minimum number of groups, and thus an ideal model is (1) which is NP-hard. As stated in Section 1, a convex approximation of $\ell_{2,0}$, i.e., $\ell_{2,1}$-norm, has been used widely as a best convex regularizer to incorporate the class labels.

More generally, the non-separable weighted $\ell_{2,1}$-norm, i.e., $\|\boldsymbol{W}\cdot\|_{2,1}$ has also been used as the regularizer in GSC [11,36,37]. For example, Tang et al. [11] proposed a weighted GSC (WGSC) model as follows, by involving the information of the similarity between query sample and each class as well as the distance between query sample and each training sample,

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \sum_{i=1}^{g} w_i \|\boldsymbol{d}_i \odot \boldsymbol{x}_i\|_2, \tag{7}$$

where $\boldsymbol{d}_i = [d_{i1}, d_{i2}, \cdots, d_{in_i}] \in \mathbb{R}^{n_i}$ penalizes the distance between $\boldsymbol{y}$ and each training sample of i-th class, w_i is set to assess the relative importance of training samples from i-th class for representing the test sample, and here $\odot$ denotes element-wise multiplication. Specifically, the weights are computed by

$$d_{ij} = \exp\left(\frac{\|\boldsymbol{y} - \boldsymbol{a}_{ij}\|_2}{\sigma_1}\right) \quad \text{and} \quad w_i = \exp\left(\frac{r_i - r_{\min}}{\sigma_2}\right), \tag{8}$$

where σ_1 and σ_2 are bandwidth parameters, $\boldsymbol{x}_i^* = \arg\min_{\boldsymbol{x}_i} \|\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{x}_i\|_2^2$, $r_i = \|\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{x}_i^*\|$ computes the distance from $\boldsymbol{y}$ to the individual subspace generated by $\boldsymbol{A}_i$, and $r_{\min}$ denotes

the minimum reconstruction error of $\{r_i\}_{i=1}^g$. The regularizer in (7) can be written as a non-separable weighted $\ell_{2,1}$, i.e., $\|\boldsymbol{W}\boldsymbol{x}\|_{2,1}$, where

$$\boldsymbol{W} = \text{BlockDiag}(\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_g) \quad \text{and} \quad \boldsymbol{W}_i = \text{Diag}(w_i d_{i1}, w_i d_{i2}, \cdots, w_i d_{in_i}). \tag{9}$$

For the aforementioned methods, after obtaining the optimal solution (denoted by $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_1^\top, \hat{\boldsymbol{x}}_2^\top, \cdots, \hat{\boldsymbol{x}}_g^\top]^\top$), they assign $\boldsymbol{y}$ to the class that minimizes the class reconstruction residual defined by $\|\boldsymbol{y} - \boldsymbol{A}_i\hat{\boldsymbol{x}}_i\|_2$.

Although $\ell_{2,1}$ regularizer and its weighted variants are widely used in GSC and WGSC-based methods, they not only suppress the number of selected classes, but also suppress significant nonzero coefficients within classes. The later may lead to underestimation of high-amplitude elements and adversely affect the performance. The nonconvex regularizers such as $\ell_{2,p}$ $(0 < p < 1)$ [37] and group MCP [38] make the corresponding optimization problems nonconvex. Therefore, we hope to use a regularizer which can reduce the bias and approximate $\ell_{2,0}$ better than $\ell_{2,1}$ while ensuring the overall convexity of the problem.

3. LiGME Model for Group Sparse Estimation

3.1. GME of Weighted $\ell_{2,1}$-Norm and Its Properties

Although $\ell_{2,1}$-norm (or its weighted variants) acts as the favorable approach to approximate $\ell_{2,0}$ in the literature of group sparse estimation, it has large bias and does not promote group sparsity as effective as $\ell_{2,0}$. Since GME provides an approach to better approximate direct discrete measures (e.g., ℓ_0 for sparsity, matrix rank for low-rankness) than their convex envelopes, we propose to use it for designing group-sparsity pursuing regularizers.

More generally, let us consider the GME of $\|\cdot\|_{w,2,1}$ in (3). Clearly, $\|\cdot\|_{w,2,1} \in \Gamma_0(\mathbb{R}^n)$ is coercive, even symmetry and prox-friendly, whose proximity operator can be computed by

$$\text{Prox}_{\gamma\|\cdot\|_{w,2,1}} : \mathbb{R}^n \to \mathbb{R}^n : \boldsymbol{x} \mapsto \left\{ \left(1 - \frac{\gamma w_i}{\max\{\|\boldsymbol{x}_i\|_2, \gamma w_i\}} \right) \boldsymbol{x}_i \right\}_{i=1}^g, \tag{10}$$

where $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_g^\top]^\top \in \mathbb{R}^n$ is a signal with group structure, $\boldsymbol{x}_i \in \mathbb{R}^{n_i} (i = 1, 2, \cdots, g)$ and $\sum_{i=1}^g n_i = n$.

Actually, the GME of $\|\cdot\|_{w,2,1}$ with $\boldsymbol{B} \in \mathbb{R}^{b\times n}$ (see (4)):

$$(\|\cdot\|_{w,2,1})_{\boldsymbol{B}}(\boldsymbol{x}) = \sum_{i=1}^g w_i\|\boldsymbol{x}_i\|_2 - \min_{\boldsymbol{v}\in\mathbb{R}^n} \left\{ \sum_{i=1}^g w_i\|\boldsymbol{v}_i\|_2 + \frac{1}{2}\|\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{v})\|_2^2 \right\}, \tag{11}$$

where $\boldsymbol{v}_i \in \mathbb{R}^{n_i} (i = 1, 2, \cdots, g)$ and $\boldsymbol{v} = [\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top, \cdots, \boldsymbol{v}_g^\top]^\top \in \mathbb{R}^n$, can serve as a parametric bridge between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{w,2,1}$.

Proposition 2. *(GME of $\|\cdot\|_{w,2,1}$ can bridge the gap between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{w,2,1}$.) Let $\boldsymbol{B}_\gamma :=$ BlockDiag$(\frac{w_1}{\sqrt{\gamma}}\boldsymbol{I}_{n_1}, \frac{w_2}{\sqrt{\gamma}}\boldsymbol{I}_{n_2}, \cdots, \frac{w_g}{\sqrt{\gamma}}\boldsymbol{I}_{n_g})$ for $\gamma > 0$, where $w_i > 0$ is the weight in (3) for $i = 1, \cdots, g$. Then, for any $\boldsymbol{x} \in \mathbb{R}^n$,*

$$\lim_{\gamma\downarrow 0} \frac{2}{\gamma}(\|\cdot\|_{w,2,1})_{\boldsymbol{B}_\gamma}(\boldsymbol{x}) = \|\boldsymbol{x}\|_{2,0}. \tag{12}$$

Together with the fact that $(\|\cdot\|_{w,2,1})_{\boldsymbol{O}_{n\times n}}(\boldsymbol{x}) = \|\boldsymbol{x}\|_{w,2,1}$ where $\boldsymbol{O}_{n\times n} \in \mathbb{R}^{n\times n}$ is the zero matrix, the regularization term $\frac{2}{\gamma}(\|\cdot\|_{w,2,1})_{\boldsymbol{B}_\gamma}(\boldsymbol{x})$ can serve as a parametric bridge between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{w,2,1}$. As a special case, the GME of $\|\cdot\|_{2,1}$ can serve as a parametric bridge between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{2,1}$.

Proof. The regularization term $\frac{2}{\gamma}(\|\cdot\|_{w,2,1})_{\boldsymbol{B}_\gamma}(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R} : [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_g^\top]^\top \mapsto \sum_{i=1}^g \frac{2}{\gamma}\varphi_i(\boldsymbol{x}_i)$, where

$$\varphi_i(\boldsymbol{x}_i) := w_i\|\boldsymbol{x}_i\|_2 - \min_{\boldsymbol{v}_i \in \mathbb{R}^{n_i}} \left\{ w_i\|\boldsymbol{v}_i\|_2 + \frac{{w_i}^2}{2\gamma}\|\boldsymbol{x}_i - \boldsymbol{v}_i\|_2^2 \right\}$$

for $i = 1, \cdots, g$. By ([39], Example 24.20), we obtain

$$\frac{2}{\gamma}\varphi_i(\boldsymbol{x}_i) = \begin{cases} \dfrac{2w_i}{\gamma}\|\boldsymbol{x}_i\|_2 - \dfrac{{w_i}^2}{\gamma^2}\|\boldsymbol{x}_i\|_2^2, & \text{if } \|\boldsymbol{x}_i\|_2 \leq \dfrac{\gamma}{w_i} \\ 1, & \text{otherwise.} \end{cases} \tag{13}$$

Then, we obtain

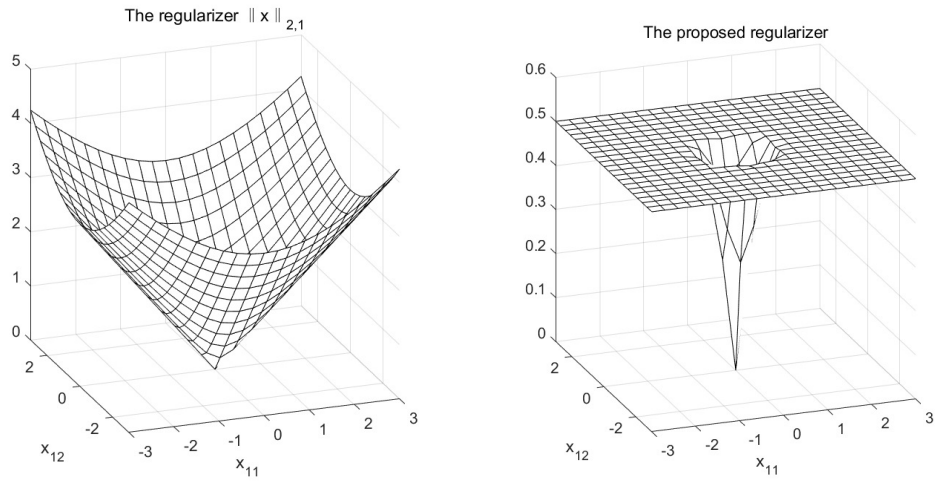
$$\lim_{\gamma \downarrow 0} \frac{2}{\gamma}\varphi_i(\boldsymbol{x}_i) = \begin{cases} 0, & \text{if } \|\boldsymbol{x}_i\|_2 = 0, \\ 1, & \text{otherwise,} \end{cases} \tag{14}$$

and

$$\lim_{\gamma \downarrow 0} \frac{2}{\gamma}(\|\cdot\|_{w,2,1})_{\boldsymbol{B}_\gamma} = \lim_{\gamma \downarrow 0} \sum_{i=1}^g \frac{2}{\gamma}\varphi_i(\boldsymbol{x}_i) = \|\boldsymbol{x}\|_{2,0}. \tag{15}$$

□

Figure 1 illustrates simple examples of $\|\boldsymbol{x}\|_{2,1}$ and $(\|\cdot\|_{2,1})_{\boldsymbol{B}}(\boldsymbol{x})$ when $g = 1$, $n = 2$ and $\boldsymbol{B} = \boldsymbol{I}_2$. As we can see, $(\|\cdot\|_{2,1})_{\boldsymbol{I}_2}(\boldsymbol{x})$ can approximate $\|\boldsymbol{x}\|_{2,0}$ better than $\|\boldsymbol{x}\|_{2,1}$.



(**a**) (**b**)

Figure 1. Simple examples of two group sparse regularizers (one group case): (**a**) The $\ell_{2,1}$ regularizer; (**b**) The regularizer $(\|\cdot\|_{2,1})_{\boldsymbol{I}_n}$.

Of course, as reviewed in Section 2.1, we can minimize $J_{(\|\cdot\|_{w,2,1})_{\boldsymbol{B}} \circ \mathrm{Id}}$ (see (5)) with the algorithm in (A3) in Proposition A2, under the overall-convexity condition $\boldsymbol{A}^\top\boldsymbol{A} - \lambda\boldsymbol{B}^\top\boldsymbol{B} \succeq \boldsymbol{O}_{n\times n}$.

In the following, we consider the GME of non-separable weighted $\ell_{2,1}$-norm $\|\boldsymbol{W}\cdot\|_{2,1}$, where $\boldsymbol{W} \in \mathbb{R}^{l\times n}$ is not necessarily a diagonal matrix. This is because in some applications, such as classification problems [11,36,37] stated in Section 2.2, and also heterogeneous feature selection [40], weights are introduced inside groups as well (i.e., the weight of every entry can be different) to improve the estimation accuracy. The GME of $\|\boldsymbol{W}\cdot\|_{2,1}$ with $\tilde{\boldsymbol{B}} \in \mathbb{R}^{b\times n}$ is well-defined (The lack of coercivity requires slight modification from min to inf.) as

$$(\|\boldsymbol{W}\cdot\|_{2,1})_{\tilde{\boldsymbol{B}}}(\boldsymbol{x}) = \|\boldsymbol{W}\boldsymbol{x}\|_{2,1} - \inf_{\boldsymbol{v}\in\mathbb{R}^n} \left\{ \|\boldsymbol{W}\boldsymbol{v}\|_{2,1} + \frac{1}{2}\|\tilde{\boldsymbol{B}}(\boldsymbol{x} - \boldsymbol{v})\|_2^2 \right\}, \tag{16}$$

and therefore we can formulate

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ J_{(\|\boldsymbol{W}\cdot\|_{2,1})_{\tilde{\boldsymbol{B}}}\circ\mathrm{Id}} = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda(\|\boldsymbol{W}\cdot\|_{2,1})_{\tilde{\boldsymbol{B}}}(\boldsymbol{x}). \tag{17}$$

However, we should remark that $\|\boldsymbol{W}\cdot\|_{2,1} \in \Gamma_0(\mathbb{R}^n)$ is even symmetric but not necessarily coercive or prox-friendly (As found in ([39], Proposition 24.14), it is known that for $\Psi \in \Gamma_0(\mathcal{Z})$ and $\mathfrak{L} \in \mathcal{B}(\mathcal{X},\mathcal{Z})$ satisfying $\mathfrak{L}\mathfrak{L}^* = \mu\mathrm{Id}$ with some $\mu \in \mathbb{R}_{++}$, we have $\mathrm{Prox}_{\Psi\circ\mathfrak{L}}(\boldsymbol{x}) = \boldsymbol{x} + \mu^{-1}\mathfrak{L}^*(\mathrm{Prox}_{\mu\Psi}(\mathfrak{L}\boldsymbol{x}) - \mathfrak{L}\boldsymbol{x})$ for $\boldsymbol{x}\in\mathcal{X}$. In such a special case, if Ψ is prox-friendly, $\Psi\circ\mathfrak{L}$ is also prox-friendly. However, for general $\mathfrak{L}\in\mathcal{B}(\mathcal{X},\mathcal{Z})$ not necessarily satisfying such standard conditions, we have to discuss the prox-friendliness of $\Psi\circ\mathfrak{L}$ case by case.). Fortunately, by Proposition 3 below, if $\mathrm{rank}(\boldsymbol{W}) = l$ and $\tilde{\boldsymbol{B}}$ can be expressed as $\tilde{\boldsymbol{B}} = \boldsymbol{B}\boldsymbol{W}$ for some $\boldsymbol{B}\in\mathbb{R}^{b\times l}$, we can show the useful relation

$$(\|\boldsymbol{W}\cdot\|_{2,1})_{\tilde{\boldsymbol{B}}}(\boldsymbol{x}) = (\|\cdot\|_{2,1})_{\boldsymbol{B}}\circ\boldsymbol{W}(\boldsymbol{x}), \tag{18}$$

which implies that the GME $(\|\boldsymbol{W}\cdot\|_{2,1})_{\tilde{\boldsymbol{B}}}$ of $\|\boldsymbol{W}\cdot\|_{2,1}$ can be handled as the LiGME $(\|\cdot\|_{2,1})_{\boldsymbol{B}}\circ\boldsymbol{W}$ of $\|\cdot\|_{2,1}$.

Proposition 3. *For $\Psi\in\Gamma_0(\mathcal{Z})$ which is coercive and $\boldsymbol{B}\in\mathcal{B}(\mathcal{Z},\tilde{\mathcal{Z}})$, assume $\mathfrak{L}\in\mathcal{B}(\mathcal{X},\mathcal{Z})$ has full row-rank. Then for any $\boldsymbol{x}\in\mathcal{X}$,*

$$(\Psi\circ\mathfrak{L})_{\boldsymbol{B}\circ\mathfrak{L}}(\boldsymbol{x}) = \Psi_{\boldsymbol{B}}\circ\mathfrak{L}(\boldsymbol{x}), \tag{19}$$

where $(\Psi\circ\mathfrak{L})_{\boldsymbol{B}\circ\mathfrak{L}}(\cdot) := \Psi(\mathfrak{L}\cdot) - \inf_{v\in\mathcal{X}}\left\{\Psi(\mathfrak{L}v) + \frac{1}{2}\|\boldsymbol{B}\mathfrak{L}(\cdot - v)\|_2^2\right\}$ and $\Psi_{\boldsymbol{B}}(\cdot) := \Psi(\cdot) - \min_{v\in\mathcal{Z}}\left\{\Psi(v) + \frac{1}{2}\|\boldsymbol{B}(\cdot - v)\|_2^2\right\}$.

Proof. On one hand, by the definition of GME, we have

$$\begin{aligned}(\Psi\circ\mathfrak{L})_{\boldsymbol{B}\circ\mathfrak{L}}(\boldsymbol{x}) &= \Psi(\mathfrak{L}\boldsymbol{x}) - \inf_{v\in\mathcal{X}}\left\{\Psi(\mathfrak{L}v) + \frac{1}{2}\|\boldsymbol{B}\mathfrak{L}(\boldsymbol{x}-v)\|_2^2\right\}\\ &= \Psi(\mathfrak{L}\boldsymbol{x}) - h(\boldsymbol{B}\mathfrak{L}\boldsymbol{x}),\end{aligned}$$

where $h(\boldsymbol{z}):\tilde{\mathcal{Z}}\to\mathbb{R}$ is given by

$$\begin{aligned}h(\boldsymbol{z}) &= \inf_{v\in\mathcal{X}}\left\{\Psi(\mathfrak{L}v) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{B}\mathfrak{L}v\|_2^2\right\}\\ &= \inf_{\boldsymbol{u}\in(\mathrm{null}\ \mathfrak{L})^\perp}\ \inf_{\hat{\boldsymbol{u}}\in\mathrm{null}\ \mathfrak{L}}\left\{\Psi(\mathfrak{L}(\boldsymbol{u}+\hat{\boldsymbol{u}})) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{B}\mathfrak{L}(\boldsymbol{u}+\hat{\boldsymbol{u}})\|_2^2\right\}\\ &= \inf_{\boldsymbol{u}\in(\mathrm{null}\ \mathfrak{L})^\perp}\left\{\Psi(\mathfrak{L}\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{B}\mathfrak{L}\boldsymbol{u}\|_2^2\right\}\\ &= \inf_{\boldsymbol{u}\in\mathrm{range}\ \mathfrak{L}^*}\left\{\Psi(\mathfrak{L}\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{B}\mathfrak{L}\boldsymbol{u}\|_2^2\right\}\\ &= \inf_{v\in\mathcal{Z}}\left\{\Psi(\mathfrak{L}\mathfrak{L}^*v) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{B}\mathfrak{L}\mathfrak{L}^*v\|_2^2\right\}\\ &= \inf_{v\in\mathcal{Z}}\left\{\Psi(\mathfrak{L}\mathfrak{L}^*(\mathfrak{L}\mathfrak{L}^*)^{-1}v) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{B}\mathfrak{L}\mathfrak{L}^*(\mathfrak{L}\mathfrak{L}^*)^{-1}v\|_2^2\right\}\\ &= \inf_{v\in\mathcal{Z}}\left\{\Psi(v) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{B}v\|_2^2\right\}.\end{aligned}$$

Therefore, $(\Psi\circ\mathfrak{L})_{\boldsymbol{B}\circ\mathfrak{L}}(\boldsymbol{x}) = \Psi(\mathfrak{L}\boldsymbol{x}) - \inf_{v\in\mathcal{Z}}\left\{\Psi(v) + \frac{1}{2}\|\boldsymbol{B}\mathfrak{L}\boldsymbol{x}-\boldsymbol{B}v\|_2^2\right\}$.

On the other hand, $\Psi_{\boldsymbol{B}}\circ\mathfrak{L}(\boldsymbol{x}) = \Psi(\mathfrak{L}\boldsymbol{x}) - \min_{v\in\mathcal{Z}}\left\{\Psi(v) + \frac{1}{2}\|\boldsymbol{B}(\mathfrak{L}\boldsymbol{x}-v)\|_2^2\right\}$ by definition. Thus, we obtain the conclusion. □

In the rest of the paper, we focus on LiGME model of $\ell_{2,1}$-norm.

3.2. LiGME of $\ell_{2,1}$-Norm

For simplicity as well as for effectiveness in application to GSC and WGSC, we focus on the LiGME model of $\|\cdot\|_{2,1}$ with an invertible linear operator $\boldsymbol{W}$,

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ J_{(\|\cdot\|_{2,1})_{\boldsymbol{B}}\circ\boldsymbol{W}} = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda(\|\cdot\|_{2,1})_{\boldsymbol{B}}\circ\boldsymbol{W}(\boldsymbol{x}). \tag{20}$$

In this case, for achieving $J_{(\|\cdot\|_{2,1})_{\boldsymbol{B}}\circ\boldsymbol{W}} \in \Gamma_0(\mathbb{R}^n)$, we can simply design $\boldsymbol{B} \in \mathbb{R}^{m\times n}$, in a way similar to ([31], (48)), as in the next proposition.

Proposition 4. *For an invertible $\boldsymbol{W} \in \mathbb{R}^{n\times n}$, let*

$$\boldsymbol{B} = \sqrt{\theta/\lambda}\boldsymbol{A}\boldsymbol{W}^{-1},\ 0 \le \theta \le 1, \tag{21}$$

then for the LiGME model in (20), $J_{(\|\cdot\|_{2,1})_{\boldsymbol{B}}\circ\boldsymbol{W}} \in \Gamma_0(\mathbb{R}^n)$.

Proof. By $\boldsymbol{A}^\top\boldsymbol{A} - \lambda\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W} = \boldsymbol{A}^\top\boldsymbol{A} - \lambda\boldsymbol{W}^\top(\sqrt{\theta/\lambda}\boldsymbol{A}\boldsymbol{W}^{-1})^\top(\sqrt{\theta/\lambda}\boldsymbol{A}\boldsymbol{W}^{-1})\boldsymbol{W} = (1-\theta)\boldsymbol{A}^\top\boldsymbol{A} \succeq \boldsymbol{O}_{n\times n}$ and Proposition 1, $J_{(\|\cdot\|_{2,1})_{\boldsymbol{B}}\circ\boldsymbol{W}} \in \Gamma_0(\mathbb{R}^n)$ is ensured. □

Model (20) can be applied to many different applications that conform to group-sparsity structure.

4. Application to Classification Problems

4.1. Proposed Algorithm for Group-Sparsity Based Classification

Since $\ell_{2,1}$ regularizer in GSC is unfair for classes of different sizes (see Appendix A) while $\ell_{2,0}$-regularizer is not, our purpose is to use a better approximation of $\ell_{2,0}$ as the regularizer. Therefore, we apply model (20) to group-sparsity-based classification. Following GSC, we can set $\boldsymbol{W} = \boldsymbol{I}_n$ in (20).

Inspired by WGSC [11] which well designs weights to enforce locality and similarity information of samples, we can also set the weight matrix $\boldsymbol{W}$ according to (9). The classification algorithm is summarized in Algorithm 1.

The $\ell_{2,1}$-norm regularized least squares problem in WGSC can be solved by a proximal gradient method [41]. Compared with it, the step 2 in Algorithm 1 for solving (20) requires at each update only one additional computation for $\mathrm{Prox}_{\gamma\|\cdot\|_{2,1}}$ (see (10) with $w_i = 1$).

4.2. Experiments

First, by setting $\boldsymbol{W} = \boldsymbol{I}_n$, we conduct the experiments on a relatively simple dataset to investigate the influence by bias of $\ell_{2,1}$ regularizer on the classification problem (especially when training set is unbalanced), and verify the performance improvement using $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$ as the regularizer by conducting the experiments on a relatively simple dataset. The USPS handwritten digit database [42] has 11,000 samples of digits "0" through "9" (1100 samples per class). The dimension of each sample is 16×16. In our classification experiments, we vectorized them to 256-D vectors. The number of training samples for each class is not necessarily equal, which varies from 5 to 50 (the size of test set is fixed to 50 images per class).

We set $\boldsymbol{W} = \boldsymbol{I}_n$ (the initialization of $\boldsymbol{W}$ should be modified in Algorithm 1) for the proposed model (20) and compared it with GSC (with $\ell_{2,1}$ regularizer) [10]. We set $\boldsymbol{B} = \sqrt{\theta/\lambda}\boldsymbol{A}$ and fix $\theta = 0.9$ to achieve the overall convexity of proposed method, and set $\kappa = 1.1$, $\iota = \|(\kappa/2)\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{I}_n\|_{\mathrm{spec}} + (\kappa-1)$, $\tau = (\kappa/2 + 2/\kappa)\lambda\|\boldsymbol{B}\|_{\mathrm{spec}}^2 + (\kappa-1)$. The initial estimate is set as $(\boldsymbol{x}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{w}^{(0)}) = (\boldsymbol{O}_{n\times 1}, \boldsymbol{O}_{n\times 1}, \boldsymbol{O}_{n\times 1})$, and the stopping criterion is set to either $\|(\boldsymbol{x}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{w}^{(k)}) - (\boldsymbol{x}^{(k+1)}, \boldsymbol{u}^{(k+1)}, \boldsymbol{w}^{(k+1)})\|_2 < 10^{-4}$ or steps reaching 10,000.

Figure 2 shows an example of unbalanced training set (digits "0" through "4" have 5 samples per class and "5" through "9" have 25 samples per class). The input (an image of

digit "0") was misclassified (into digital "6") by GSC while classified correctly by proposed method. The obtained coefficient vectors by GSC and proposed method (both with $\lambda = 4$) are illustrated respectively, and some samples corresponding to nonzero coefficients are also displayed in Figure 2. It can be seen that the samples from digit "6" made the greatest contribution to the representation in GSC, and samples from "5" and "0" also made small contribution. In our method, samples from the correct class "0" made the biggest contribution and led to correct result. It is reasonable, because our method did not suppress the high value coefficients too much whereas $\ell_{2,1}$ did. The big suppression of $\ell_{2,1}$ made the coefficients of the correct class cannot be large enough, and thus easily led to misclassification.

Algorithm 1: The proposed group-sparsity enhanced classification algorithm

Input: A matrix of training samples $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_G] \in \mathbb{R}^{m\times n}$ grouped by class information, a test sample vector $\boldsymbol{y} \in \mathbb{R}^m$, parameters λ, σ_1 and σ_2.

1. Initialization: Let $(\boldsymbol{x}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$.
Compute the weight matrix $\boldsymbol{W}$ by (9).
Choose $\boldsymbol{B}$ satisfying $\boldsymbol{A}^\top\boldsymbol{A} - \lambda\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W} \succeq \boldsymbol{O}_{n\times n}$.
Choose $(\iota, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, +\infty)$ satisfying

$$\iota\boldsymbol{I}_n - \frac{\kappa}{2}\boldsymbol{A}^\top\boldsymbol{A} - \lambda\boldsymbol{W}^\top\boldsymbol{W} \succeq \boldsymbol{O}_{n\times n} \quad \text{and} \quad \tau \geq \Big(\frac{\kappa}{2} + \frac{2}{\kappa}\Big)\lambda\|\boldsymbol{B}\|_{\text{spec}}^2. \tag{22}$$

2. For $k = 0, 1, 2, \cdots$, compute

$$\begin{aligned}
\boldsymbol{x}^{(k+1)} =& \Big[\boldsymbol{I}_n - \frac{1}{\iota}(\boldsymbol{A}^\top\boldsymbol{A} - \lambda\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W})\Big]\boldsymbol{x}^{(k)} - \frac{\lambda}{\iota}\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{u}^{(k)} - \frac{\lambda}{\iota}\boldsymbol{W}^\top\boldsymbol{v}^{(k)} + \frac{1}{\iota}\boldsymbol{A}^\top\boldsymbol{y},\\
\boldsymbol{u}^{(k+1)} =& \text{Prox}_{\frac{\lambda}{\tau}\|\cdot\|_{2,1}}\Big[\frac{2\lambda}{\tau}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}\boldsymbol{x}^{(k+1)} - \frac{\lambda}{\tau}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}\boldsymbol{x}^{(k)} + (\boldsymbol{I}_n - \frac{\lambda}{\tau}\boldsymbol{B}^\top\boldsymbol{B})\boldsymbol{u}^{(k)}\Big],\\
\boldsymbol{v}^{(k+1)} =& 2\boldsymbol{W}\boldsymbol{x}^{(k+1)} - \boldsymbol{W}\boldsymbol{x}^{(k)} + \boldsymbol{v}^{(k)} - \text{Prox}_{\|\cdot\|_{2,1}}\Big(2\boldsymbol{W}\boldsymbol{x}^{(k+1)} - \boldsymbol{W}\boldsymbol{x}^{(k)} + \boldsymbol{v}^{(k)}\Big)
\end{aligned}$$

until the stopping criterion is fulfilled.

3. Compute the class label $i^\star$ of $\boldsymbol{y}$ by

$$i^\star = \arg\min_i \|\boldsymbol{y} - \boldsymbol{A}_i\boldsymbol{x}_i^{(k+1)}\|_2. \tag{23}$$

Output: The class label $i^\star$ corresponding to $\boldsymbol{y}$.

(For example, any $\kappa > 1$, $\iota = \|(\kappa/2)\boldsymbol{A}^\top\boldsymbol{A} + \lambda\boldsymbol{W}^\top\boldsymbol{W}\|_{\text{spec}} + (\kappa - 1)$ and $\tau = (\kappa/2 + 2/\kappa)\lambda\|\boldsymbol{B}\|_{\text{spec}}^2 + (\kappa - 1)$ can satisfy (22).)

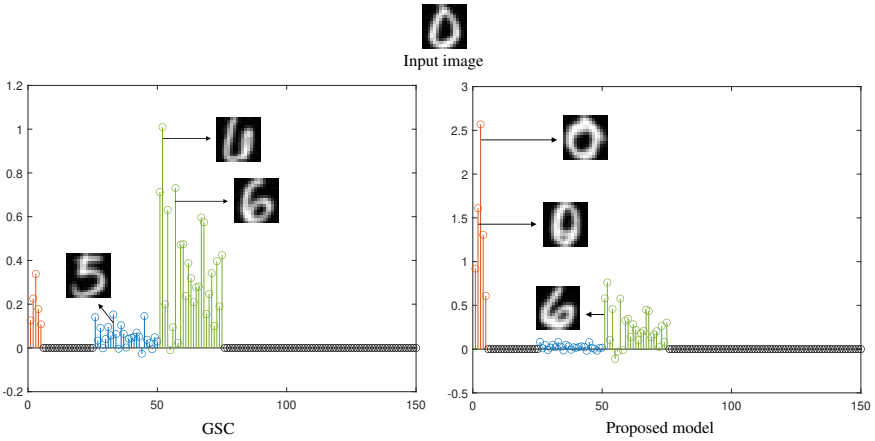


Figure 2. Estimated sparse coefficients $\hat{\boldsymbol{x}}$ by GSC and proposed method respectively.

Table 1 summarizes the recognition accuracy of GSC and the proposed method with $\boldsymbol{W} = \boldsymbol{I}_n$. The training set includes digits "0" through "4" β samples per class and "5" through "9" α samples per class. Through numerical experiments, we found that GSC with $\lambda = \lambda_{\mathrm{GSC}} = 1.5$ and the proposed method with $\lambda = \lambda_{\mathrm{prop}} = 3$ perform well on this dataset. We also experimented the proposed method of using λ_{GSC}, which did not degrade too much compared with using λ_{prop}. We see that the GSC model degrades when the training set is unbalanced, and the proposed method outperforms GSC especially in such case.

Table 1. Recognition results on the USPS database.

Method	**Training Set Size ($\alpha = \max_i\{n_i\}$, $\beta = \min_i\{n_i\}$)**					
	α	**10**		**25**		**50**
	β	**5**	**10**	**5**	**25**	**25**	**50**
GSC(with $\lambda = 1.5$)		81.4%	86.6%	73.6%	91.4%	88.4%	93.2%
Proposed ($\lambda = 1.5$)		82.0%	87.2%	79.0%	92.2%	89.4%	93.0%
Proposed ($\lambda = 3$)		82.6%	87.8%	80.8%	92.2%	90.6%	93.4%

Next, we conduct the experiments on a classic face dataset to verify the validity of the proposed linearly involved model by setting the weight matrix $\boldsymbol{W}$ according to (9). The ORL Database of Faces [43] contains 400 images from 40 distinct subjects (10 images per subject) with variations in lighting, facial expressions (open or closed eyes, smiling or not smiling) and facial details (glasses or no glasses). In our experiments, following [44], all images were downsampled from 112×92 to 16×16 and then formed 256-D vectors. The number of training samples for each class is not necessarily equal, which varies from 4 to 8 (test set is fixed to 2 images per class).

We compared the proposed model (20) ($\boldsymbol{W} = \boldsymbol{I}_n$ and $\boldsymbol{W}$ by (9) respectively) with GSC [10] and WGSC [11]. In order to achieve the overall convexity, we set $\boldsymbol{B} = \sqrt{\theta/\lambda}\boldsymbol{A}\boldsymbol{W}^{-1}$, $0 \leq \theta \leq 1$ and fix $\theta = 0.9$ for proposed method. Settings of (ι, τ, κ), initial estimate and stopping criterion are the same as those in the previous experiment. When the parameter λ is assigned too small, the obtained coefficient vector is not group sparse; when the parameter σ_1 or σ_2 is assigned too small, the information of locality or similarity plays a decisive role. We found that $\lambda = 0.05$ for $\ell_{2,1}$ regularizer-based methods (i.e., GSC and WGSC), $\lambda = 0.2$ for the proposed method and $\sigma_1 \in [2, 4], \sigma_2 \in [0.5, 2]$ for weights involved methods (i.e., WGSC and proposed method with $\boldsymbol{W}$ by (9)) work well on this dataset.

Figure 3 shows a classification result of WGSC and proposed method ($\boldsymbol{W}$ by (9)) (both with $\sigma_1 = 4, \sigma_2 = 2$)) when training set is unbalanced (20 subjects have 8 samples per class and the others have 6 samples per class). The input is an image of subject 10 which was misclassified into subject 8 by WGSC while classified correctly by proposed method.

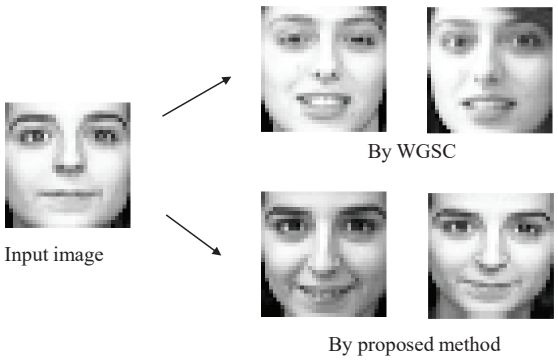


Figure 3. An example of results by WGSC and proposed method.

Table 2 summarizes the recognition accuracy of GSC, the proposed method with $\boldsymbol{W} = \boldsymbol{I}_n$, WGSC and the proposed method with $\boldsymbol{W}$ computed by (9). Training set setting is that 20 subjects have β samples per class and the others have α samples per class. With the strategically designed matrix (9), WGSC achieves a significant improvement over GSC. By using the proposed method with $\boldsymbol{W}$ computed by (9), the performance can be further improved, especially when the training set is unbalanced.

Table 2. Recognition results on the ORL database.

Method	Training Set Size ($\alpha = \max_i\{n_i\}$, $\beta = \min_i\{n_i\}$)					
	α	4	6		8	
	β	4	4	6	4	6	8
GSC		86.3%	85.0%	91.3%	85.0%	92.5%	93.8%
Proposed ($\boldsymbol{W} = \boldsymbol{I}_n$)		88.8%	86.3%	93.8%	86.3%	93.8%	95.0%
WGSC		90.6%	87.5%	95.0%	88.8%	93.8%	96.3%
Proposed ($\boldsymbol{W}$ by (9))		91.3%	89.4%	95.6%	91.9%	94.4%	96.3%

5. Conclusions

In this paper, the potential applicability and effectiveness of using nonconvex regularizers in convex optimization framework was explored. We proposed a generalized Moreau enhancement (GME) of weighted $\ell_{2,1}$ function and analyzed its relationship with the linearly involved GME of $\ell_{2,1}$-norm. The proposed regularizer is nonconvex and promotes group sparsity more effectively than $\ell_{2,1}$ while maintaining the overall convexity of the regression model at the same time. The model can be used in many applications and we applied it to classification problems. Our model makes use of the grouping structure by class information and suppresses the tendency of underestimation of high-amplitude coefficients. Experimental results showed that the proposed method is effective for image classification.

Author Contributions: Conceptualization, M.Y. and I.Y.; methodology, Y.C., M.Y. and I.Y.; software, Y.C.; writing-original draft, Y.C., writing-review and editing, M.Y. and I.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS Grants-in-Aid grant number 18K19804 and by JST SICORP grant number JPMJSC20C6.

Data Availability Statement: Publicly available data sets were analyzed in this study. These data can be found here: https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#usps; https://cam-orl.co.uk/facedatabase.html; and http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html (accessed on 25 October 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LASSO	Least Absolute Shrinkage and Selection Operator
SCAD	Smoothly Clipped Absolute Deviation)
MCP	Minimax Concave Penalty
GMC	Generalized Minimax Concave
GME	Generalized Moreau Enhancement
LiGME	Linearly involved Generalized-Moreau-Enhanced (or Enhancement)
SRC	Sparse Representation-based Classification
GSC	Group Sparse Classification
WGSC	Weighted Group Sparse Classification

Appendix A. The Bias of $\ell_{2,1}$ Regularizer in Group Sparse Classification

Using $\ell_{2,1}$ regularizer in classification problems not only minimizes the number of the selected classes, but also minimizes the ℓ_2-norm of coefficients within each class. The later may adversely affect the classification result, since the optimal representation of a test sample by training samples of the correct subject may contain large coefficients. Moreover, in many classification applications, the number of training samples from different classes is not the same. We argue that the bias of $\ell_{2,1}$ regularizer makes it unfair for classes of different sizes.

Example A1. *Suppose that a test sample $\boldsymbol{y} \in \mathbb{R}^m$ can be represented by a combination of all n_i samples from class i without error, i.e., $\boldsymbol{y} = \boldsymbol{A}_i \boldsymbol{x}_i$ and $\|\boldsymbol{x}_i\|_1 = 1$, where $\boldsymbol{A}_i \in \mathbb{R}^{m \times n_i}$ and $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$.*

(a) If the number of samples in this class is doubled by duplication, the training set of class i becomes $\tilde{\boldsymbol{A}}_i = [\boldsymbol{A}_i, \boldsymbol{A}_i] \in \mathbb{R}^{m \times 2n_i}$. Obviously, $\boldsymbol{y}$ can also be well represented by $\boldsymbol{y} = \tilde{\boldsymbol{A}}_i \tilde{\boldsymbol{x}}_i$, where $\tilde{\boldsymbol{x}}_i = [\eta \boldsymbol{x}_i^\top, (1-\eta)\boldsymbol{x}_i^\top]^\top \in \mathbb{R}^{2n_i}$ $(0 \le \eta \le 1)$ and $\|\tilde{\boldsymbol{x}}_i\|_1 = 1$. However, $\|\boldsymbol{x}_i\|_2^2 - \|\tilde{\boldsymbol{x}}_i\|_2^2 = 2\eta(1-\eta)\|\boldsymbol{x}_i\|_2^2 \ge 0$. That is, $\ell_{2,1}$ value of the first representation (before duplication) is greater than that of the second one (after duplication).

(b) If the number of samples in this class is increased to dn_i by copying $d-1$ times $(d > 1)$, the training set of class i becomes $\tilde{\boldsymbol{A}}_i = [\boldsymbol{A}_i, \cdots, \boldsymbol{A}_i] \in \mathbb{R}^{m \times dn_i}$. Obviously, $\boldsymbol{y} = \tilde{\boldsymbol{A}}_i \tilde{\boldsymbol{x}}_i$ is a representation of $\boldsymbol{y}$, where $\tilde{\boldsymbol{x}}_i = [\frac{1}{d}\boldsymbol{x}_i^\top, \cdots, \frac{1}{d}\boldsymbol{x}_i^\top]^\top \in \mathbb{R}^{dn_i}$ and $\|\tilde{\boldsymbol{x}}_i\|_1 = 1$. Then $\|\tilde{\boldsymbol{x}}_i\|_2 = \frac{1}{\sqrt{d}}\|\boldsymbol{x}_i\|_2 < \|\boldsymbol{x}_i\|_2$.

Example A1 tells us that the group size affects the value of $\ell_{2,1}$ regularizer. Even if the new training sample is only a copy of the original samples (without adding any new information), the value of $\ell_{2,1}$ regularizer will decrease. Therefore, $\ell_{2,1}$ regularizer is unfair for classes of different sizes. It has the tendency to refuse the class has relatively few samples, because the coefficient vector is more likely to have a large $\ell_{2,1}$ regularizer value. Please note that $\ell_{2,0}$-regularizer is independent of group size and it does not have such unfairness.

Appendix B. Parameter Tuning and Proximal Splitting Algorithm for LiGME Model

Proposition A1 ([24], Proposition 2)**.** *In (5), let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^l)$, $(\boldsymbol{A}, \mathfrak{L}, \lambda) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{l \times n} \times \mathbb{R}_{++}$, and $\mathrm{rank}(\mathfrak{L}) = l$. Choose a nonsingular $\tilde{\mathfrak{L}} \in \mathbb{R}^{n \times n}$ satisfying $[\boldsymbol{O}_{l \times (n-l)} \ \boldsymbol{I}_l]\tilde{\mathfrak{L}} = \mathfrak{L}$. Then $\boldsymbol{B}_\theta := \sqrt{\theta/\lambda}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top \in \mathbb{R}^{l \times l}$, $\theta \in [0,1]$, ensures $J_{\Psi_{\boldsymbol{B}_\theta} \circ \mathfrak{L}} \in \Gamma_0(\mathbb{R}^n)$, where $[\tilde{\boldsymbol{D}}_1 \ \tilde{\boldsymbol{D}}_2] := \boldsymbol{A}(\tilde{\mathfrak{L}})^{-1}$ and $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top := \tilde{\boldsymbol{D}}_2^\top \tilde{\boldsymbol{D}}_2 - \tilde{\boldsymbol{D}}_2^\top \tilde{\boldsymbol{D}}_1 (\tilde{\boldsymbol{D}}_1^\top \tilde{\boldsymbol{D}}_1)^\dagger \tilde{\boldsymbol{D}}_1^\top \tilde{\boldsymbol{D}}_2 \in \mathbb{R}^{l \times l}$ is an eigendecomposition.*

Proposition A2 ([24], Theorem 1)**.** *Consider minimization of $J_{\Psi_{\boldsymbol{B}} \circ \mathfrak{L}}$ in (5) under the overall-convexity condition (6). Let a real Hilbert space $(\mathcal{H} := \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \langle \cdot, \cdot \rangle_{\mathcal{H}}, \|\cdot\|_{\mathcal{H}})$ be a product space and define an operator $\mathcal{T}_{\mathrm{LiGME}} : \mathcal{H} \to \mathcal{H} : (\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \to (\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\eta})$ with parameters $(\iota, \tau) \in \mathbb{R}_{++} \times \mathbb{R}_{++}$, by*

$$
\begin{aligned}
\boldsymbol{\xi} &:= \left[\mathrm{Id} - \frac{1}{\sigma}(\boldsymbol{A}^*\boldsymbol{A} - \lambda \mathfrak{L}^* \boldsymbol{B}^* \boldsymbol{B} \mathfrak{L})\right]\boldsymbol{x} - \frac{\lambda}{\iota}\mathfrak{L}^*\boldsymbol{B}^*\boldsymbol{B}\boldsymbol{u} - \frac{\lambda}{\iota}\mathfrak{L}^*\boldsymbol{v} + \frac{1}{\iota}\boldsymbol{A}^*\boldsymbol{y}, \\
\boldsymbol{\zeta} &:= \mathrm{Prox}_{\frac{\lambda}{\tau}\Psi}\left[\frac{2\lambda}{\tau}\boldsymbol{B}^*\boldsymbol{B}\mathfrak{L}\boldsymbol{\xi} - \frac{\lambda}{\tau}\boldsymbol{B}^*\boldsymbol{B}\mathfrak{L}\boldsymbol{x} + \left(\mathrm{Id} - \frac{\lambda}{\tau}\boldsymbol{B}^*\boldsymbol{B}\right)\boldsymbol{u}\right], \\
\boldsymbol{\eta} &:= 2\mathfrak{L}\boldsymbol{\xi} - \mathfrak{L}\boldsymbol{x} + \boldsymbol{v} - \mathrm{Prox}_{\Psi}(2\mathfrak{L}\boldsymbol{\xi} - \mathfrak{L}\boldsymbol{x} + \boldsymbol{v}).
\end{aligned}
$$

Then the following holds:

1. $\arg\min_{\boldsymbol{x} \in \mathcal{X}} J_{\Psi_{\boldsymbol{B}} \circ \mathfrak{L}}(\boldsymbol{x}) = \{\boldsymbol{x}^\star \in \mathcal{H} \mid (\boldsymbol{x}^\star, \boldsymbol{u}^\star, \boldsymbol{v}^\star) \in \mathrm{Fix}(\mathcal{T}_{\mathrm{LiGME}})\}$, *where* $\mathrm{Fix}(\mathcal{T}_{\mathrm{LiGME}}) := \{(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \in \mathcal{H} \mid \mathcal{T}_{\mathrm{LiGME}}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})\}$.

2. *Choose* $(\iota, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, \infty)$ *satisfying*

$$\begin{aligned} &\iota \mathrm{Id} - \frac{\kappa}{2} \boldsymbol{A}^* \boldsymbol{A} - \lambda \mathfrak{L}^* \mathfrak{L} \succ \boldsymbol{O}_{\mathcal{X}}, \\ &\tau \geq \left(\frac{\kappa}{2} + \frac{2}{\kappa} \right) \lambda \|\boldsymbol{B}\|_{\mathrm{op}}^2, \end{aligned} \tag{A1}$$

where $\|\cdot\|_{\mathrm{op}}$ *is the operator norm. Then*

$$\mathfrak{P} := \begin{bmatrix} \iota \mathrm{Id} & -\lambda \mathfrak{L}^* \boldsymbol{B}^* \boldsymbol{B} & -\lambda \mathfrak{L}^* \\ -\lambda \boldsymbol{B}^* \boldsymbol{B} \mathfrak{L} & \tau \mathrm{Id} & \boldsymbol{O}_{\mathcal{Z}} \\ -\lambda \mathfrak{L} & \boldsymbol{O}_{\mathcal{Z}} & \lambda \mathrm{Id} \end{bmatrix} \succ \boldsymbol{O}_{\mathcal{H}} \tag{A2}$$

and $\mathcal{T}_{\mathrm{LiGME}}$ *is* $\frac{\kappa}{2\kappa-1}$*-averaged nonexpansive in the Hilbert space* $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathfrak{P}}, \|\cdot\|_{\mathfrak{P}})$.

3. *Assume the condition (A1) holds. Then, for any initial point* $(\boldsymbol{x}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)})$*, the sequence* $\{(\boldsymbol{x}^{(k)}, \boldsymbol{v}^{(k)}, \boldsymbol{u}^{(k)})\}_{k \in \mathbb{N}}$ *generated by*

$$(\boldsymbol{x}^{(k+1)}, \boldsymbol{u}^{(k+1)}, \boldsymbol{v}^{(k+1)}) = \mathcal{T}_{\mathrm{LiGME}}(\boldsymbol{x}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{v}^{(k)}) \tag{A3}$$

converges to a point $(\boldsymbol{x}^\star, \boldsymbol{u}^\star, \boldsymbol{v}^\star) \in \mathrm{Fix}(\mathcal{T}_{\mathrm{LiGME}})$ *and*

$$\lim_{k \to \infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^\star \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} J_{\Psi_{\boldsymbol{B}} \circ \mathfrak{L}}(\boldsymbol{x}).$$

References

1. Theodoridis, S. *Machine Learning: A Bayesian and Optimization Perspective*; Academic Press: Cambridge, MA, USA, 2015.
2. Ma, S.; Song, X.; Huang, J. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinform.* **2007**, *8*, 60. [CrossRef] [PubMed]
3. Wang, L.; Chen, G.; Li, H. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **2007**, *23*, 1486–1494. [CrossRef] [PubMed]
4. Wang, X.; Zhong, Y.; Zhang, L.; Xu, Y. Spatial group sparsity regularized nonnegative matrix factorization for hyperspectral unmixing. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6287–6304. [CrossRef]
5. Drumetz, L.; Meyer, T.R.; Chanussot, J.; Bertozzi, A.L.; Jutten, C. Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms. *IEEE Trans. Image Process.* **2019**, *28*, 3435–3450. [CrossRef] [PubMed]
6. Huang, J.; Huang, T.Z.; Zhao, X.L.; Deng, L.J. Nonlocal tensor-based sparse hyperspectral unmixing. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6854–6868. [CrossRef]
7. Qiao, B.; Mao, Z.; Liu, J.; Zhao, Z.; Chen, X. Group sparse regularization for impact force identification in time domain. *J. Sound Vib.* **2019**, *445*, 44–63. [CrossRef]
8. Majumdar, A.; Ward, R.K. Classification via group sparsity promoting regularization. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 861–864.
9. Elhamifar, E.; Vidal, R. Robust classification using structured sparse representation. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1873–1879.
10. Huang, J.; Nie, F.; Huang, H.; Ding, C. Supervised and projected sparse coding for image classification. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, WA, USA, 14–18 July 2013.
11. Tang, X.; Feng, G.; Cai, J. Weighted group sparse representation for undersampled face recognition. *Neurocomputing* **2014**, *145*, 402–415. [CrossRef]
12. Rao, N.; Nowak, R.; Cox, C.; Rogers, T. Classification with the sparse group lasso. *IEEE Trans. Signal Process.* **2015**, *64*, 448–463. [CrossRef]
13. Tan, S.; Sun, X.; Chan, W.; Qu, L.; Shao, L. Robust face recognition with kernelized locality-sensitive group sparsity representation. *IEEE Trans. Image Process.* **2017**, *26*, 4661–4668. [CrossRef]
14. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2006**, *68*, 49–67. [CrossRef]
15. Natarajan, B.K. Sparse approximate solutions to linear systems. *SIAM J. Comput.* **1995**, *24*, 227–234. [CrossRef]
16. Argyriou, A.; Foygel, R.; Srebro, N. Sparse prediction with the k-support norm. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Siem Reap, Cambodia, 13–16 December 2018; Volume 1, pp. 1457–1465.
17. Deng, W.; Yin, W.; Zhang, Y. Group sparse optimization by alternating direction method. In *Wavelets and Sparsity XV*; International Society for Optics and Photonics: Bellingham, WA, USA, 2013; Volume 8858, p. 88580R.

18. Huang, J.; Breheny, P.; Ma, S. A selective review of group selection in high-dimensional models. *Stat. Sci. A Rev. J. Inst. Math. Stat.* **2012**, *27*. [CrossRef]
19. Breheny, P.; Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat. Comput.* **2015**, *25*, 173–187. [CrossRef]
20. Hu, Y.; Li, C.; Meng, K.; Qin, J.; Yang, X. Group sparse optimization via lp, q regularization. *J. Mach. Learn. Res.* **2017**, *18*, 960–1011.
21. Jiang, L.; Zhu, W. Iterative Weighted Group Thresholding Method for Group Sparse Recovery. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 63–76. [CrossRef] [PubMed]
22. Jiao, Y.; Jin, B.; Lu, X. Group Sparse Recovery via the $\ell^0(\ell^2)$ Penalty: Theory and Algorithm. *IEEE Trans. Signal Process.* **2016**, *65*, 998–1012. [CrossRef]
23. Chen, P.Y.; Selesnick, I.W. Group-sparse signal denoising: Non-convex regularization, convex optimization. *IEEE Trans. Signal Process.* **2014**, *62*, 3464–3478. [CrossRef]
24. Abe, J.; Yamagishi, M.; Yamada, I. Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition. *Inverse Probl.* **2020**, *36*, 035012. [CrossRef]
25. Chen, Y.; Yamagishi, M.; Yamada, I. A Generalized Moreau Enhancement of $\ell_{2,1}$-norm and Its Application to Group Sparse Classification. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021.
26. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
27. Larsson, V.; Olsson, C. Convex low rank approximation. *Int. J. Comput. Vis.* **2016**, *120*, 194–214. [CrossRef]
28. Blake, A.; Zisserman, A. *Visual Reconstruction*; MIT Press: Cambridge, MA, USA, 1987.
29. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
30. Nikolova, M.; Ng, M.K.; Tam, C.P. Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. *IEEE Trans. Image Process.* **2010**, *19*, 3073–3088. [CrossRef] [PubMed]
31. Selesnick, I. Sparse regularization via convex analysis. *IEEE Trans. Signal Process.* **2017**, *65*, 4481–4494. [CrossRef]
32. Yin, L.; Parekh, A.; Selesnick, I. Stable principal component pursuit via convex analysis. *IEEE Trans. Signal Process.* **2019**, *67*, 2595–2607. [CrossRef]
33. Abe, J.; Yamagishi, M.; Yamada, I. Convexity-edge-preserving signal recovery with linearly involved generalized minimax concave penalty function. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 4918–4922.
34. Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Ma, Y. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 210–227. [CrossRef]
35. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
36. Xu, Y.; Sun, Y.; Quan, Y.; Luo, Y. Structured sparse coding for classification via reweighted $\ell_{2,1}$ minimization. In *CCF Chinese Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 189–199.
37. Zheng, J.; Yang, P.; Chen, S.; Shen, G.; Wang, W. Iterative re-constrained group sparse face recognition with adaptive weights learning. *IEEE Trans. Image Process.* **2017**, *26*, 2408–2423. [CrossRef]
38. Zhang, C.; Li, H.; Chen, C.; Qian, Y.; Zhou, X. Enhanced group sparse regularized nonconvex regression for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]
39. Bauschke, H.H.; Combettes, P.L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed.; Springer International Publishing: New York, NY, USA, 2017.
40. Zhao, L.; Hu, Q.; Wang, W. Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Trans. Multimed.* **2015**, *17*, 1936–1948. [CrossRef]
41. Qin, Z.; Scheinberg, K.; Goldfarb, D. Efficient block-coordinate descent algorithms for the group lasso. *Math. Program. Comput.* **2013**, *5*, 143–169. [CrossRef]
42. Hull, J.J. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 550–554. [CrossRef]
43. Samaria, F.S.; Harter, A.C. Parameterisation of a stochastic model for human face identification. In Proceedings of the1994 IEEE Workshop on Applications of Computer Vision, Sarasota, FL, USA, 5–7 December 1994; pp. 138–142.
44. Cai, D.; He, X.; Hu, Y.; Han, J.; Huang, T. Learning a spatially smooth subspace for face recognition. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–7.