T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A Study of Designs and Applications of Generalized Moreau Enhancement Matrix for Sparsity Aware LiGME Models
著者(和文)	CHEN YANG
Author(English)	Yang Chen
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12393号, 授与年月日:2023年3月26日, 学位の種別:課程博士, 審査員:山田 功,植松 友彦,府川 和彦,SLAVAKIS KONSTANTINOS,實松 豊
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12393号, Conferred date:2023/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Powered by T2R2 (Science Tokyo Research Repository)

Tokyo Institute of Technology
Department of Information and Communications Engineering

A Study of Designs and Applications of Generalized Moreau Enhancement Matrix for Sparsity Aware LiGME Models

Yang Chen

Supervisor Prof. Isao Yamada

Department of Communications and Computer Engineering,
School of Engineering

January 2023

Abstract

The Linearly involved Generalized-Moreau-Enhanced (LiGME) model has been established as a convex analytic framework to utilize linearly involved nonconvex regularizers for sparsity aware least squares estimation. In the LiGME model, the design of generalized Moreau enhancement matrix (GME matrix) is a key to guarantee the overall convexity of the optimization model. In this thesis, we study certain fundamental properties of the LiGME model and novel design strategies of the GME matrix. The proposed GME matrix design is applicable to general linear operators to be involved in the LiGME regularizer and does not require any eigendecomposition or iteration. By using the proposed GME matrix design, we also present applications to the group sparsity aware least squares estimation problems, including signal recovery and image classification. Numerical experiments demonstrate the effectiveness of the proposed GME matrix in the LiGME model.

Contents

Chapter 1

Introduction

Many tasks in data science, including signal processing and machine learning are formulated as the estimation of an unknown vector $x^\star \in \mathcal{X}$ from the observed vector $y \in \mathcal{Y}$, which follows the linear regression model

$$y = Ax^\star + \varepsilon, \tag{1.1}$$

where $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}}, \|\cdot\|_{\mathcal{X}})$ and $(\mathcal{Y}, \langle\cdot,\cdot\rangle_{\mathcal{Y}}, \|\cdot\|_{\mathcal{Y}})$ are finite dimensional real Hilbert spaces, $A : \mathcal{X} \to \mathcal{Y}$ is a known bounded linear operator and ε is an unknown noise vector [1, 2]. To tackle such estimations, a widely used approach is to minimize the regularized linear least squares cost function as follows

$$\underset{x\in\mathcal{X}}{\text{minimize}}\ \frac{1}{2}\|y - Ax\|_{\mathcal{Y}}^2 + \mu\Psi \circ \mathfrak{L}(x),\ \mu > 0, \tag{1.2}$$

where $\mathfrak{L} : \mathcal{X} \to \mathcal{Z}$ is a certain bounded linear operator, $\Psi : \mathcal{Z} \to (-\infty, +\infty]$ is a certain function, $\circ$ stands for composition and $(\mathcal{Z}, \langle\cdot,\cdot\rangle_{\mathcal{Z}}, \|\cdot\|_{\mathcal{Z}})$ is a finite dimensional Hilbert space. In problem (1.2), the least squares $\frac{1}{2}\|y - Ax\|_{\mathcal{Y}}^2$ is called fidelity term, $\Psi \circ \mathfrak{L}(x)$ is called regularizer (or penalty) which is designed based on a priori knowledge regarding $x^\star$, and the regularization parameter μ balances the two terms [1, 2]. The tuple $(\Psi, \mathfrak{L})$ is designed not only depending on each application but also based on mathematical tractability of the optimization model (1.2); see, for example, [3–6].

To design $(\Psi, \mathfrak{L})$ in (1.2), one of the most crucial properties to be handled in modern signal processing is the *sparsity* of vectors or the low-rankness of matrices. The sparsity and the low-rankness have been used for many scenarios in data science, such as compressed sensing [7, 8], sparsity aware applications [9–13], and machine learning [2, 14]. By reconstructing 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 non-zero entries.

Let $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Z} = \mathbb{R}^l$. Considering a prior information that $\mathfrak{L}(\boldsymbol{x}^\star)$ is sparse, i.e., the target signal $\boldsymbol{x}^\star$ becomes sparse via a linear transform $\mathfrak{L} : \mathbb{R}^n \to \mathbb{R}^l$, $\Psi(\cdot) = \|\cdot\|_0$, the ℓ_0 pseudo-norm which counts the number of nonzero entries in a vector, is a natural choice for Ψ. However, this choice makes the optimization problem (1.2) NP hard. Alternatively, $\Psi(\cdot) = \|\cdot\|_1$, the ℓ_1 norm defined as the sum of absolute value of entries, has been used widely because it is the largest convex minorant of $\|\cdot\|_0$ on $[-1,1]^l$ ($\subset \mathbb{R}^l$) [15–18]. For example, problem (1.2) reproduces the famous *Lasso (least absolute shrinkage and selection operator)* [19] with $(\Psi, \mathfrak{L}) := (\|\cdot\|_1, \mathrm{Id})$, reproduces the so-called Total Variation (TV) regularizer [20, 21] with $(\Psi, \mathfrak{L}) := (\|\cdot\|_1, D)$, the wavelet-based regularizer [21, 22] with $(\Psi, \mathfrak{L}) := (\|\cdot\|_1, \mathcal{W})$ respectively, where Id is the identity operator, D is the first order differential operator and $\mathcal{W}$ is a wavelet transform matrix.

To promote sparsity more effectively than ℓ_1 norm, utilization of nonconvex functions as Ψ has been considered, such as $(\|\boldsymbol{z}\|_q)^q = \sum_{i=1}^l |z_i|^q$ $(0 < q < 1)$ for $\boldsymbol{z} = [z_1, \cdots, z_l]^\top \in \mathbb{R}^l$ [23,24], smoothed ℓ_0 function [25], smoothed ℓ_q function [26], and the SCAD (smothly clipped absolute deviation) [27], to name but a few. These regularizers lead to nonconvexity of the cost function in (1.2), making the global minimizer of (1.2) hardly obtained. On the other hand, a great deal of effort has also been devoted to utilizing nonconvex function as Ψ while maintaining the overall convexity of the optimization problem (1.2) at the same time. The so-called convexity-preserving nonconvex penalties were introduced, in late 80's by Blake and Zisserman [28], and followed for example by Nikolova [29–31]. For recent developments of the convexity-preserving nonconvex penalties, see [32–40] and the references therein. Most of the convexity-preserving nonconvex penalties rely on a special assumption, i.e., the strong convexity of the least squares term in (1.2), equivalently the nonsingularity of A^*A, where A^* is the adjoint operator of A (see Section 2.1).

The *generalized minimax concave (GMC) model* [41] and the *Linearly involved Generalized-Moreau-Enhanced (LiGME) model* [5, 42] are exceptional examples which are free from the assumption of A^*A. The LiGME model (see Definition 2.13) is formulated as

$$\underset{x\in\mathcal{X}}{\text{minimize}}\ J_{\Psi_B\circ\mathfrak{L}} := \frac{1}{2}\|y - Ax\|_{\mathcal{Y}}^2 + \mu\Psi_B \circ \mathfrak{L}(x),\ \mu > 0, \tag{1.3}$$

where $\Psi_B(\cdot) := \Psi(\cdot) - \min_{v\in\mathcal{Z}} \left[\Psi(v) + \frac{1}{2}\|B(\cdot - v)\|_{\widetilde{\mathcal{Z}}}^2\right]$, $(\widetilde{\mathcal{Z}}, \langle\cdot,\cdot\rangle_{\widetilde{\mathcal{Z}}}, \|\cdot\|_{\widetilde{\mathcal{Z}}})$ is a finite dimensional Hilbert space and the linear operator B can be viewed as a tuning parameter. We call *B Generalized Moreau Enhancement matrix (GME matrix)*. It has been shown that Ψ_B in (1.3) can parametrically bridge the gap between ℓ_0 pseudo-norm and ℓ_1 norm [5, Ex-

ample 2(b)], and it can also parametrically bridge the gap between the rank function and the nuclear norm (sum of singular values of a matrix) [5, Example 2(c)]. Although the regularizer $\Psi_B \circ \mathfrak{L}$ in (1.3) is nonconvex in general, the *overall convexity* of $J_{\Psi_B \circ \mathfrak{L}}$ can be achieved if $A^*A - \mu \mathfrak{L}^* B^* B \mathfrak{L}$ is positive semidefinite (see Proposition 2.2). Since the LiGME model establishes a framework for constructing a class of nonconvex regularizers without losing overall convexity of the regularized least squares problem (1.3), it is applicable to a board range of sparsity and low rankness aware estimations [5, 6, 43, 44].

The design of GME matrix B is a key to guarantee the overall convexity of $J_{\Psi_B \circ \mathfrak{L}}$ in (1.3), which enables us to find a global minimizer of the LiGME model. To the best of the authors' knowledge, so far, the algebraic design methods (i.e., finite step algorithms) of GME matrices have been limited to certain special cases for linear operator $\mathfrak{L}$ of full row rank [5, 41]. However, there are many applications of LiGME models where $\mathfrak{L}$ does not have full row rank (see, e.g., Chapter 4 for group sparsity aware signal recovery). Very recently, to obtain B^*B satisfying the overall convexity condition for general $\mathfrak{L}$, a non-algebraic iterative algorithm, based on the so-called CQ method [45], has been proposed in [44]. However, at each update of the iterative algorithm in [44], the computation of the projection onto the set of all positive semidefinite matrices is required via eigendecomposition of a symmetric matrix.

The goal of this study is to establish an algebraic and scalable design of the GME matrix B for general linear operator $\mathfrak{L}$ not necessarily of full row rank, which would enhance the applicability of the LiGME model. Actually, in sparsity aware signal processing problems, we have many applications where a rank deficient linear operator $\mathfrak{L}$ plays important roles. For example, in a scenario of group sparsity aware signal recovery, the *grouping operator* $\mathscr{G}$ (see Definition 4.1) becomes rank deficient in general cases (see Example 4.1) [46, 47].

In this study, we propose a new algebraic design essentially based on a simple *LDU (lower-diagonal-upper) decomposition* [48–50] of the linear operator $\mathfrak{L}$. The proposed design is applicable to any $\mathfrak{L}$, which can be seen as a generalization of the existing algebraic designs in [5, 41]. Moreover, it does not require any eigendecomposition or any iterative computation, and thus more scalable than the existing designs including [44]. As an application of the proposed algebraic GME matrix design, we present a new LiGME model for group sparsity aware estimation with general group configurations which allows overlapping and incomplete group structures. In this application, we verified the applicability and effectiveness of the proposed algebraic design of the GME matrix through numerical experiments. Finally, we also present another application of the LiGME model

to group sparse classification (GSC) [51–53] and verified that the GSC based on the proposed LiGME model outperforms the standard GSC based on the Group Lasso model especially in cases of using unbalanced training sets.

The rest of this thesis is organized as follows. After the general introduction in this chapter, in Chapter 2, starting with the necessary mathematical notations and related definitions, we first present useful tools in the modern convex optimization and the fixed point theory of nonexpansive operators. Then we present a brief review on the existing strategies regarding the overall convexity preserving nonconvex regularizers, which includes the LiGME models. We also present a brief review of the prior arts for group sparsity aware signal recovery and group sparse classification.

In Chapter 3, after summarizing selected fundamental properties of the LiGME model, we propose a new algebraic design of the GME matrix based on the LDU decomposition. The proposed algebraic design is applicable to general linear operator $\mathfrak{L}$ in the LiGME model. Moreover, the proposed algebraic design can be seen as a generalization of the existing algebraic designs. We also present another algebraic design of the GME matrix, which is based on the singular value decomposition (SVD).

In Chapter 4, after introducing the grouping operator for general group configurations which allows overlapping and incomplete cover of group structures, we propose a new LiGME model for group sparsity aware signal recovery, where the proposed model can cope with the cases of unknown group partition. Since the grouping operator therein does not have full row rank, the proposed GME matrix design plays an essential role in this application. The numerical experiments show the effectiveness of the proposed LiGME model based on the proposed GME matrix design.

In Chapter 5, we consider another application of the LiGME model to the group sparse classification. After illustrating the unfairness of the $\ell_{2,1}$ regularizer in the appearance of different group sizes, we propose to use the LiGME model for group sparse classification of unbalanced training set. The numerical experiments demonstrate the effectiveness of the proposed LiGME model compared with the standard Group Lasso model, in this scenario of the group sparse classification as well.

Finally, in Chapter 6, we conclude this thesis with few additional remarks.

Chapter 2

Preliminaries

2.1 Mathematical Notation and Tools in Convex Optimization and Fixed Point Theory of Nonexpansive Operators

We use calligraphic font to denote Hilbert spaces and standard font letters to denote elements in the spaces. For readers who are not familiar with the notion of the Hilbert space, we encourage to consult, e.g., [54, 55]. For discussion in Euclidean space, we use boldface letters to express vectors and general font letters to represent scalars. We list the notations in Table 2.1 and Table 2.2.

Then we introduce some tools in convex optimization and fixed point theory of nonexpansive operators.

Definition 2.1. A function $f : \mathcal{H} \to (-\infty, \infty]$ is *convex* if it satisfies

$$(\forall x, y \in \mathcal{H}, \forall 0 \in (0,1)) \quad f(\theta x + (1-\theta)y) \leqslant \theta f(x) + (1-\theta) f(y).$$

f is *lower semicontinuous* if its lower level set $\mathrm{lev}_{\leqslant a} f := \{x \in \mathcal{H} | f(x) \leqslant a\}$ is closed for $\forall a \in \mathbb{R}$. f is said to be *proper* if $\mathrm{dom}\, f \neq \varnothing$.

The set of proper lower semicontinuous convex functions $f : \mathcal{H} \to (-\infty, \infty]$ is denoted by $\Gamma_0(\mathcal{H})$.

Definition 2.2. (Coercive) Let $f : \mathcal{H} \to [-\infty, \infty]$. Then f is *coercive* if

$$\lim_{\|x\|_{\mathcal{H}} \to \infty} f(x) = \infty.$$

Table 2.1: Notations related to Hilbert space.

Symbol	Definition
$\mathcal{H}, \widetilde{\mathcal{H}}$	Finite dimensional real Hilbert spaces
$\langle \cdot \mid \cdot \rangle_{\mathcal{H}}$	Scalar product in $\mathcal{H}$
$\|\cdot\|_{\mathcal{H}}$	Norm in $\mathcal{H}$
$\mathcal{B}(\mathcal{H}, \widetilde{\mathcal{H}})$	The set of all linear operators from $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}, \|\cdot\|_{\mathcal{H}})$ to $(\widetilde{\mathcal{H}}, \langle \cdot, \cdot \rangle_{\widetilde{\mathcal{H}}}, \|\cdot\|_{\widetilde{\mathcal{H}}})$
Id	The identity operator of general Hilbert spaces
$O_{\mathcal{B}(\mathcal{H},\widetilde{\mathcal{H}})}$	The zero operator from $\mathcal{H}$ to $\widetilde{\mathcal{H}}$, $O_{\mathcal{B}(\mathcal{H},\widetilde{\mathcal{H}})} \in \mathcal{B}(\mathcal{H}, \widetilde{\mathcal{H}})$
$O_{\mathcal{H}}$	The zero operator from $\mathcal{H}$ to $\mathcal{H}$, $O_{\mathcal{H}} \in \mathcal{B}(\mathcal{H}, \mathcal{H})$
$\|\mathfrak{L}\|_{\mathrm{op}}$	The operator norm of $\mathfrak{L} \in \mathcal{B}(\mathcal{H}, \widetilde{\mathcal{H}})$
	(i.e., $\|\mathfrak{L}\|_{\mathrm{op}} := \sup_{x \in \mathcal{H}, \|x\|_{\mathcal{H}} \leqslant 1} \|\mathfrak{L}x\|_{\widetilde{\mathcal{H}}}$)
$\mathfrak{L}^*$	The adjoint operator of $\mathfrak{L}$,
	(i.e., $\forall x \in \mathcal{H}, \forall y \in \widetilde{\mathcal{H}}, \langle \mathfrak{L}x, y \rangle_{\widetilde{\mathcal{H}}} = \langle x, \mathfrak{L}^* y \rangle_{\mathcal{H}}$, $\mathfrak{L}^* \in \mathcal{B}(\widetilde{\mathcal{H}}, \mathcal{H})$)
$\mathfrak{L} \succ O_{\mathcal{H}}$	$\mathfrak{L} \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ is positive definite
$\mathfrak{L} \succeq O_{\mathcal{H}}$	$\mathfrak{L} \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ is positive semidefinite
$\mathrm{ran}(\mathfrak{L})$	The range space of $\mathfrak{L} \in \mathcal{B}(\mathcal{H}, \widetilde{\mathcal{H}})$
	(i.e., $\mathrm{ran}(\mathfrak{L}) := \{y \in \widetilde{\mathcal{H}} \mid (\exists x \in \mathcal{H})\ \mathfrak{L}x = y\}$)
$\mathrm{null}(\mathfrak{L})$	The null space of $\mathfrak{L} \in \mathcal{B}(\mathcal{H}, \widetilde{\mathcal{H}})$
	(i.e., $\mathrm{null}(\mathfrak{L}) := \{x \in \mathcal{H} \mid \mathfrak{L}x = 0\}$)
$S^{\perp}$	The orthogonal complement subspace of a nonempty subset $S \subset \mathcal{H}$
	(i.e., $S^{\perp} := \{x \in \mathcal{H} \mid \langle x, y \rangle_{\mathcal{H}} = 0, \forall y \in S\}$)
$\mathrm{dom}\ f$	The domain of a function f
	(i.e., $\mathrm{dom} f = \{x \in \mathcal{H} \mid f(x) < \infty\}$)
P_C	The projection onto a nonempty closed convex set $C \subset \mathcal{H}$
	(i.e., $P_C : \mathcal{H} \to \mathcal{H} : x \mapsto \mathrm{argmin}_{y \in C} \|x - y\|_{\mathcal{H}}$)
$\Gamma_0(\mathcal{H})$	The set of proper lower semicontinuous convex functions (see Definition 2.1)
∂f	Subdifferential of a function f (see Definition 2.3)
f^*	Conjugate of a function f (see Definition 2.4)
$f \,\square\, g$	Infimal convolution of the functions f and g (see Definition 2.7)
$f \,\boxdot\, g$	Exact infimal convolution of the functions f and g (see Definition 2.7)
${}^{\gamma}f$	The Moreau envelope of a function f of parameter γ (see Definition 2.8)
Prox_f	Proximity operator of a function f (see Definition 2.9)

By convention, f is coercive if $\mathcal{H} = \{0\}$.

Definition 2.3. (Subdifferential) For a function $f \in \Gamma_0(\mathcal{H})$, the *subdifferential* of f is

Table 2.2: Notations related to Euclidean space.

Symbol	Definition
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_+$	The set of nonnegative real numbers
$\mathbb{R}_{++}$	The set of strictly positive real numbers
$\mathbb{N}$	The set of nonnegative intergers
$\boldsymbol{A}^\top$	The transpose of a matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	The inverse of a invertible matrix $\boldsymbol{A}$
$\boldsymbol{A}^\dagger$	The Moore-Penrose pseudoinverse of a matrix $\boldsymbol{A}$
$\boldsymbol{I}_n$	The identity matrix in $\mathbb{R}^{n\times n}$
$\boldsymbol{O}_{m\times n}$	The zero matrix in $\mathbb{R}^{m\times n}$
$\mathbf{0}_n$	The zero vector in $\mathbb{R}^n$
$\|\cdot\|_0$	The ℓ_0 pseudo-norm counting the number of nonzero entries in a vector
$\|\cdot\|_1$	The ℓ_1 norm calculating the sum of absolute values of all entries in a vector
$\|\cdot\|_2$	The ℓ_2 norm (also called Euclidean norm) calculating square root of the inner product of a vector with itself
$\|\cdot\|_{\text{spec}}$	The spectral norm calculating the largest singular value of a matrix

defined as the set valued operator

$$\partial f:\ \mathcal{H}\to 2^{\mathcal{H}}:\ x\mapsto\{u\in\mathcal{H}\mid\langle y-x,u\rangle_{\mathcal{H}}+f(x)\leqslant f(y),\ \forall y\in\mathcal{H}\}.$$

Definition 2.4. (Legendre-Fenchel conjugate) For any $f\in\Gamma_0(\mathcal{H})$, the *conjugate* (or *Legendre-Fenchel conjugate*) of f is defined by

$$f^*:\ \mathcal{H}\to(-\infty,\infty]:\ y\mapsto\sup_{x\in\mathcal{H}}\{\langle x,y\rangle_{\mathcal{H}}-f(x)\},$$

which satisfies $f^*\in\Gamma_0(\mathcal{H})$. Let $f\in\Gamma_0(\mathcal{H})$. Then,

$$(\forall(x,u)\in\mathcal{H}\times\mathcal{H})\quad u\in\partial f(x)\Leftrightarrow x\in\partial f^*(u).$$

Definition 2.5. (Nonexpensive operator) An operator $T:\mathcal{H}\to\mathcal{H}$ is called *nonexpansive* if

$$(\forall x,y\in\mathcal{H})\quad\|T(x)-T(y)\|_{\mathcal{H}}\leqslant\|x-y\|_{\mathcal{H}}.$$

In particular, a nonexpansive operator T is called α-*averaged* with $\alpha\in(0,1)$ if there exists a nonexpansive operator $\hat{T}:\mathcal{H}\to\mathcal{H}$ such that

$$T=(1-\alpha)\mathrm{Id}+\alpha\hat{T},$$

i.e., T is a convex combination of the identity operator and some nonexpansive operator $\hat{T}$.

Fact 1. (Krasnoselskii-Mann iterations for finding a fixed point of averaged non-expansive operator [56, Section 5.2]) Let $\mathrm{Fix}(T) := \{x \in \mathcal{H} \mid T(x) = x\}$. For a nonexpansive operator $T : \mathcal{H} \to \mathcal{H}$ with $\mathrm{Fix}(T) \neq \varnothing$ and any initial point $x^{(0)} \in \mathcal{H}$, if $(\alpha^{(k)})_{k\in\mathbb{N}} \subset [0,1]$ satisfies

$$\sum_{k\in\mathbb{N}} \alpha^{(k)}(1-\alpha^{(k)}) = \infty,$$

then the sequence $(x^{(k)})_{k\in\mathbb{N}} \subset \mathcal{H}$ generated by

$$x^{(k+1)} = \left[(1-\alpha^{(k)})\mathrm{Id} + \alpha^{(k)}T\right](x^{(k)})$$

convergence weakly to a point in $\mathrm{Fix}(T)$. In particular, if T is α-averaged with some $\alpha \in (0,1)$, then

$$x^{(k+1)} = T(x^{(k)})$$

convergence weakly to a point in $\mathrm{Fix}(T)$.

Definition 2.6. (Even symmetry) A function $f : \mathcal{H} \to (-\infty, \infty]$ is said to be *even symmetry* if

$$f(-x) = f(x).$$

Definition 2.7. (Infimal convolution) Let f and g be functions from $\mathcal{H}$ to $(-\infty, \infty]$. Then *infimal convolution* of f and g is defined by

$$f \,\square\, g : \mathcal{H} \to [-\infty, \infty] : x \mapsto \inf_{y\in\mathcal{H}} \left(f(y) + g(x-y)\right).$$

It is exact at a point $x \in \mathcal{H}$ if

$$(\exists y \in \mathcal{H}) \quad (f \,\square\, g)(x) = f(y) + g(x-y) \in (-\infty, \infty].$$

If $f \,\square\, g$ is exact at every point of its domain, it is denoted by $f \,\boxdot\, g$.

Definition 2.8. (Moreau envelope) Let $f : \mathcal{H} \to (-\infty, \infty]$ and $\gamma \in \mathbb{R}_{++}$. The *Moreau envelope* of f of parameter γ is defined by

$${}^{\gamma}f = f \,\square\, \left(\frac{1}{2\gamma}\|\cdot\|_{\mathcal{H}}^2\right).$$

Definition 2.9. (Proximity operator) The *proximity operator* of $f \in \Gamma_0(\mathcal{H})$ is defined by

$$\mathrm{Prox}_f : \mathcal{H} \to \mathcal{H} : x \mapsto \operatorname*{argmin}_{y\in\mathcal{H}} \left[f(y) + \frac{1}{2}\|x-y\|_{\mathcal{H}}^2\right].$$

The proximity operator Prox_f is well-defined because the minimizer of $f(\cdot) + \frac{1}{2}\|\cdot - x\|_{\mathcal{H}}^2$ exists uniquely in $\mathcal{H}$ for every $f \in \Gamma_0(\mathcal{H})$ and $x \in \mathcal{H}$. A function $f \in \Gamma_0(\mathcal{H})$ is said to be *prox-friendly* if $\mathrm{Prox}_{\gamma f}$ is available as a computational tool for any $\gamma \in \mathbb{R}_{++}$.

2.2 Linearly involved Generalized-Moreau-Enhanced Model

Before introducing the Linearly involved Generalized-Moreau-Enhanced (LiGME) model, we briefly review the minimax-concave (MC) penalty function and the generalized MC (GMC) penalty function.

Definition 2.10. (Huber function) The *Huber function* $s : \mathbb{R} \to \mathbb{R}$ is defined as

$$s(x) := \begin{cases} \frac{1}{2}x^2, & |x| \leqslant 1 \\ |x| - \frac{1}{2}, & |x| \geqslant 1. \end{cases}$$

The Huber function is illustrated in Fig. 2.1(d). It is a example of the Moreau envelope, which can be written as

$$s = |\cdot| \,\square\, \frac{1}{2}(\,\cdot\,)^2 = {}^{1}|\cdot|.$$

The scaled versions of Huber function $s_b : \mathbb{R} \to \mathbb{R}$ is given by

$$s_b = |\cdot| \,\square\, \frac{1}{2}b^2(\,\cdot\,)^2 = {}^{1/b^2}|\cdot|. \tag{2.1}$$

Definition 2.11. (Minimax-concave penalty) The *minimax-concave (MC) penalty* function $\phi : \mathbb{R} \to \mathbb{R}$ is defined as

$$\phi(x) := \begin{cases} |x| - \frac{1}{2}x^2, & |x| \leqslant 1 \\ \frac{1}{2}, & |x| \geqslant 1. \end{cases}$$

The MC penalty is illustrated in Fig. 2.1(b). It can be written as

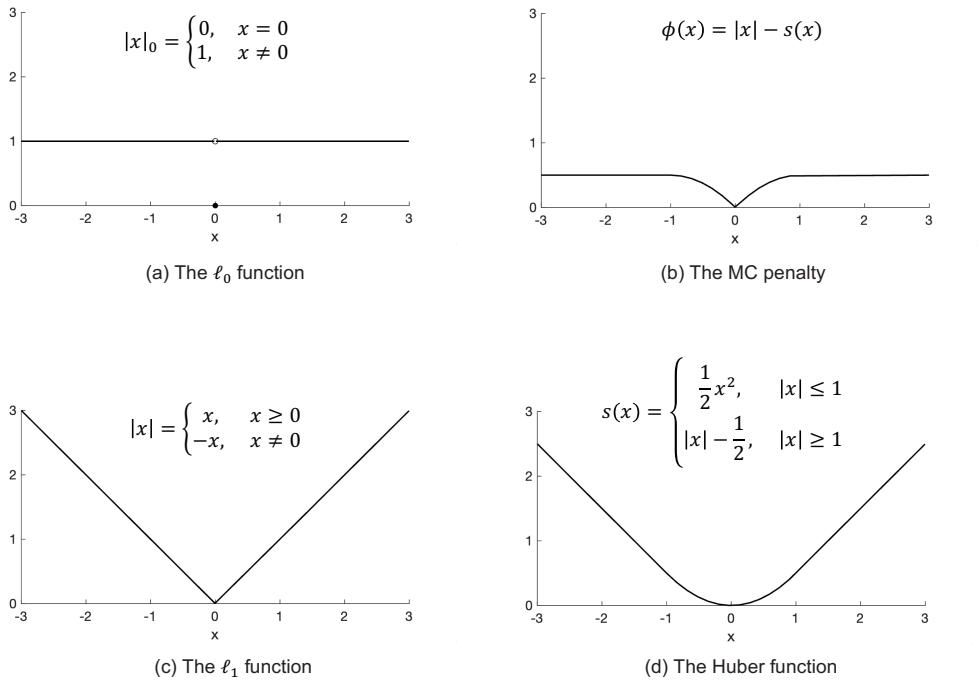
$$\phi(x) = |x| - s(x).$$

The scaled version of MC penalty $\phi_b : \mathbb{R} \to \mathbb{R}$ is given by

$$\phi_b(x) = |x| - s_b(x), \tag{2.2}$$

where s_b is the scaled Huber function in (2.1).

Some examples of one-dimensional sparsity-pursuing regularizers are illustrated in Fig. 2.1. As can be observed, the nonconvex MC penalty in Fig. 2.1(b) approximates ℓ_0 much better than convex ℓ_1 in Fig. 2.1(c). The introduction of parameter b into the scaled version in (2.2) makes MC penalty more convenient and flexible.



(a) The ℓ_0 function (b) The MC penalty

(c) The ℓ_1 function (d) The Huber function

Fig. 2.1: The ℓ_0 function, MC penalty, ℓ_1 function and Huber function of one-dimensional case.

Proposition 2.1. [41, Proposition 3] Let $\mu > 0$ and $a \in \mathbb{R}$. Define $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \frac{1}{2}(y - ax)^2 + \mu\phi_b(x), \tag{2.3}$$

where ϕ_b is the scaled MC penalty in (2.2). The function f is convex if $b^2 \leqslant a^2/\mu$.

Although the regularizer ϕ_b in (2.3) itself is nonconvex in general, the cost function f can be convex under certain conditions. Regularizers with this kind of property are called *convexity preserving regularizers*.

For multidimensional cases,

$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \mu \sum_{i=1}^{n} \phi_{b_i}(x_i) \tag{2.4}$$

is a direct generalization of (2.3), where $\boldsymbol{x} = [x_1, x_2 \cdots, x_n] \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. However, this kind of separable penalty are fundamentally limited. If $\boldsymbol{A}^\top \boldsymbol{A}$ is singular,

then a separable penalty constrained to maintain cost function convexity can only improve on the ℓ_1 norm to a very limited extent [57].

A generalized non-separable version of scaled MC penalty has been proposed in [41].

Definition 2.12. (Generalized MC penalty) The *generalized MC (GMC)* penalty function $\psi_{\boldsymbol{B}} : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\psi_{\boldsymbol{B}}(\boldsymbol{x}) := \|\boldsymbol{x}\|_1 - S_{\boldsymbol{B}}(\boldsymbol{x}), \tag{2.5}$$

where $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ and $S_{\boldsymbol{B}} : \mathbb{R}^n \to \mathbb{R}$ is the generalized Huber function defined as

$$S_{\boldsymbol{B}} = \|\cdot\| \,\square\, \frac{1}{2}\|\boldsymbol{B}\cdot\|_2^2. \tag{2.6}$$

The generalized Huber function $S_{\boldsymbol{B}}$ is a proper lower semicontinuous convex function and the infimal convolution in (2.6) is exact [41, Proposition 5]. Thus the GMC penalty in (2.5) can be written as

$$\psi_{\boldsymbol{B}}(\boldsymbol{x}) = \|\boldsymbol{x}\| - \min_{\boldsymbol{v}\in\mathbb{R}^n}\left\{\|\boldsymbol{v}\|_1 + \frac{1}{2}\|\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{v})\|_2^2\right\}.$$

The GMC penalty has great potential for dealing with many non-convex variations of $\|\cdot\|_1$ under modern convex analysis.

In order to maximize the applicability of the excellent ideas of the MC penalty and GMC penalty, the *Linearly involved Generalized Moreau Enhanced (LiGME)* model has been proposed in [5].

Definition 2.13. (LiGME model [5, Definition 1]) 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 $(\widetilde{\mathcal{Z}}, \langle\cdot,\cdot\rangle_{\widetilde{\mathcal{Z}}}, \|\cdot\|_{\widetilde{\mathcal{Z}}})$ be finite dimensional real Hilbert spaces, $\Psi \in \Gamma_0(\mathcal{Z})$ coercive with $\mathrm{dom}\Psi = \mathcal{Z}$, $B \in \mathcal{B}(\mathcal{Z}, \widetilde{\mathcal{Z}})$ and $(A, \mathfrak{L}, \mu) \in \mathcal{B}(\mathcal{X}, \mathcal{Y}) \times \mathcal{B}(\mathcal{X}, \mathcal{Z}) \times \mathbb{R}_+$. Then:

(a) Generalized Moreau enhanced penalty function Ψ_B is defined as

$$\Psi_B(\cdot) := \Psi(\cdot) - \min_{v\in\mathcal{Z}}\left[\Psi(v) + \frac{1}{2}\|B(\cdot - v)\|_{\widetilde{\mathcal{Z}}}^2\right], \tag{2.7}$$

where B can be viewed as a tuning parameter and we called it *Generalized Moreau Enhancement matrix (GME matrix)* in this thesis.

(b) Linearly involved Generalized-Moreau-Enhanced (LiGME) penalty is defined as $\Psi_B \circ \mathfrak{L} : \mathcal{X} \to (-\infty, \infty]$.

(c) LiGME mode is defined as the minimization of

$$J_{\Psi_B\circ\mathfrak{L}} : \mathcal{X} \to \mathbb{R} : x \mapsto \frac{1}{2}\|y - Ax\|_{\mathcal{Y}}^2 + \mu\Psi_B \circ \mathfrak{L}(x), \tag{2.8}$$

where we assume that $J_{\Psi_B\circ\mathfrak{L}}$ has a minimizer in $\mathcal{X}$.

The regularizer term of (2.8) can be write as

$$\Psi_B \circ \mathfrak{L}(x) = \Psi(\mathfrak{L}x) - \left(\Psi \,\square\, \frac{1}{2}\|B \cdot\|_{\widetilde{\mathcal{Z}}}^2\right)(\mathfrak{L}x). \tag{2.9}$$

For $B = O_{\mathcal{B}(\mathcal{Z},\widetilde{\mathcal{Z}})}$ in (2.7), Ψ_B reproduces Ψ. Otherwise Ψ_B is not necessarily convex in general. If B satisfies a certain condition (see Proposition 2.2), the cost function $J_{\Psi_B\circ\mathfrak{L}}$ in (2.8) becomes convex, which means that the overall convexity is achieved in (2.8). We remark that the GMC model [41] is a special case of (2.8) with $(\mathcal{X},\mathcal{Y},\mathcal{Z},\widetilde{\mathcal{Z}}) = (\mathbb{R}^n,\mathbb{R}^m,\mathbb{R}^n,\mathbb{R}^m)$, $\Psi = \|\cdot\|_1$ and $\mathfrak{L} = \mathrm{Id}$. The relation between MCP [58], GMC penalty [41] and LiGME penalty [5] is illustrated in Fig. 2.2.

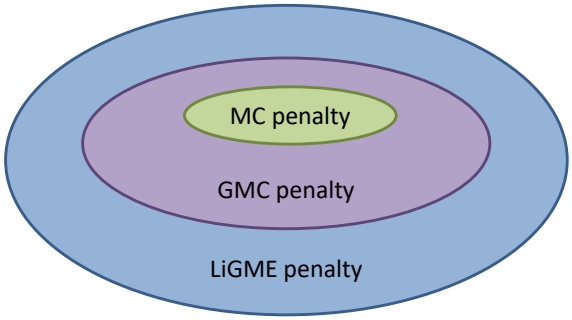


Fig. 2.2: The relation between MCP [58], GMC penalty [41] and LiGME penalty [5].

Proposition 2.2. (Overall convexity condition for the LiGME model [5, Proposition 1]) Let $(A,\mathfrak{L},\mu) \in \mathcal{B}(\mathcal{X},\mathcal{Y}) \times \mathcal{B}(\mathcal{X},\mathcal{Z}) \times \mathbb{R}_{++}$. The GME penalty function Ψ_B in Definition 2.13 has the following property: for the three conditions

(C1) $A^*A - \mu\mathfrak{L}^*B^*B\mathfrak{L} \succeq O_{\mathcal{X}}$,

(C2) $J_{\Psi_B\circ\mathfrak{L}} \in \Gamma_0(\mathcal{X})$ for any $y \in \mathcal{Y}$,

(C3) $J^{(0)}_{\Psi_B\circ\mathfrak{L}} := \frac{1}{2}\|A\cdot\|_{\mathcal{Y}}^2 + \mu\Psi_B \circ \mathfrak{L} \in \Gamma_0(\mathcal{X})$,

the relation (C1)⇒(C2)⇔(C3) holds.

In particular, if $\Psi \in \Gamma_0(\mathcal{Z})$ satisfies the condition as a norm of the vector space $\mathcal{Z}$, the equivalence (C1)⇔(C2)⇔(C3) is achieved.

Proposition 2.3. (Iterative approximation of a global minimizer of $J_{\Psi_B \circ \mathfrak{L}}$ [5, Theorem 1]) Let $\Psi \in \Gamma_0(\mathcal{Z})$ in Definition 1 be even symmetry and prox-friendly. Assume the condition (C1) in Proposition 2.2 and the existence of a minimizer, of $J_{\Psi_B \circ \mathfrak{L}}$, in $\mathcal{X}$. Let $\mathcal{H} := \mathcal{X} \times \mathcal{Z} \times \mathcal{Z}$. Define $T_{\mathrm{LiGME}} : \mathcal{H} \to \mathcal{H} : (x, v, w) \mapsto (\xi, \zeta, \eta)$ by

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

where $(\sigma, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, +\infty)$ is chosen to satisfy

$$\begin{cases}
\sigma \mathrm{Id} - \dfrac{\kappa}{2} A^* A - \mu \mathfrak{L}^* \mathfrak{L} \succ O_{\mathcal{X}}, \\
\tau \geqslant \left(\dfrac{\kappa}{2} + \dfrac{2}{\kappa}\right) \mu \|B\|_{\mathrm{op}}^2.
\end{cases} \tag{2.10}$$

Then, T_{LiGME} is a $\frac{\kappa}{2\kappa - 1}$-averaged nonexpensive operator in the Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathfrak{P}}, \| \cdot \|_{\mathfrak{P}})$, where

$$\mathfrak{P} := \begin{bmatrix}
\sigma \mathrm{Id} & -\mu \mathfrak{L}^* B^* B & -\mu \mathfrak{L}^* \\
-\mu B^* B \mathfrak{L} & \tau \mathrm{Id} & O_{\mathcal{Z}} \\
-\mu \mathfrak{L} & O_{\mathcal{Z}} & \mu \mathrm{Id}
\end{bmatrix} \succ O_{\mathcal{H}}. \tag{2.11}$$

For any initial point $(x^{(0)}, v^{(0)}, w^{(0)}) \in \mathcal{H}$, the sequence $(x^{(k)}, v^{(k)}, w^{(k)})_{k \in \mathbb{N}} \subset \mathcal{H}$ generated by

$$(x^{(k+1)}, v^{(k+1)}, w^{(k+1)}) = T_{\mathrm{LiGME}}(x^{(k)}, v^{(k)}, w^{(k)}) \tag{2.12}$$

converges weakly to point $(x^\star, v^\star, w^\star) \in \mathrm{Fix}(T_{\mathrm{LiGME}})$ and

$$\lim_{k \to \infty} x^{(k)} = x^\star,$$

where $x^\star$ is a global minimizer of $J_{\Psi_B \circ \mathfrak{L}}$.

Remark 2.1. An example of $(\sigma, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, +\infty)$ satisfying (2.10) is

$$\begin{cases}
\forall\, \kappa > 1, \\
\sigma = \left\| \dfrac{\kappa}{2} A^* A + \mu \mathfrak{L}^* \mathfrak{L} \right\|_{\mathrm{op}} + (\kappa - 1), \\
\tau = \left(\dfrac{\kappa}{2} + \dfrac{2}{\kappa}\right) \mu \|B\|_{\mathrm{op}}^2 + (\kappa - 1).
\end{cases}$$

Remark 2.2. The derivation of the algorithm in (2.12) is inspired by Condat ' s primal-dual algorithm [59] and is based on the so-called forward-backward splitting method; see [5, Remark 4].

From Proposition 2.3, we can see that design of GME matrix B is a key to guarantee the overall convexity of the LiGME model, and the overall convexity is indispensable for minimizing $J_{\Psi_B \circ \mathfrak{Q}}$ in (2.8). To our knowledge, so far, the algebraic designs of GME matrix B have been limited to a special case where the linear operator $\mathfrak{Q}$ has full row rank [5,41]. A main goal of this thesis is to establish a unified algebraic design of the GME matrix B (see Section 3.2).

2.3 Group Sparsity Aware Signal Recovery and Classification

2.3.1 Group Sparsity Aware Signal Recovery

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 [60, 61], MRI reconstruction [62], hyperspectral image unmixing [63–65], impact force identification in industrial applications [66], classification problems [51–53, 67–69], 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.

Suppose $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_t^\top]^\top \in \mathbb{R}^n$ is a group sparse signal, where $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$, $\sum_{i=1}^t n_i = n$ and t is the number of groups. Just like 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., $\|\boldsymbol{x}\|_{2,0} = \|(\|\boldsymbol{x}_1\|_2, \|\boldsymbol{x}_2\|_2, \cdots, \|\boldsymbol{x}_t\|_2)\|_0$.

The group sparse regularized least squares problem can be modeled as

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu\|\boldsymbol{x}\|_{2,0}, \tag{2.13}$$

where $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ are known, and $\mu > 0$ is the regularization parameter. However, the employment of the pseudo-norm $\ell_{2,0}$ makes (2.13) NP-hard [70]. Most studies in the application replace the nonconvex regularizer $\ell_{2,0}$ with its tightest convex

envelope $\ell_{2,1}$ [71] (or its weighted variants), and the following regularized least squares problem has been proposed named as the *Group Lasso* [72],

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\mu\sum_{i=1}^{t} w_i\|\boldsymbol{x}_i\|_2, \tag{2.14}$$

where $w_i > 0$ $(i = 1, \cdots, t)$ in the regularization term are used to adjust for group sizes with $w_i = \sqrt{n_i}$ in [72, 73]. The Lasso model [19]

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\mu\|\boldsymbol{x}\|_1 \tag{2.15}$$

is a specialization of Group Lasso model (2.14) with $t = n$ and $n_i = 1$ $(i = 1, 2, \cdots, t)$.

Although the convex optimization problem (2.14) has been used as a standard model for group sparse estimation applications, the convex regularizer $\sum_{i=1}^{t} w_i\|\boldsymbol{x}_i\|_2$ does not necessarily promote group sparsity sufficiently, mainly due to the fact that it 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) [61], group MCP (minimax concave penalty) [73, 74], $\ell_{p,q}$ regularization ($\|\boldsymbol{x}\|_{p,q} := \left(\sum_{i=1}^{t}\|\boldsymbol{x}_i\|_p^q\right)^{1/q}$, $0 < q < 1 \leqslant p$) [75], iterative weighted group minimization [76], and $\ell_{2,0}$ [77] have been used, for promoting the group sparsity more effectively than convex regularizers. However, they lose the overall convexity of such optimization problems, which results in their algorithms of no guarantee of convergence to a global minimizer of the overall cost functions. In [78], a non-convex 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 $\boldsymbol{A} = \boldsymbol{I}_n$), which can not be applied to group sparsity aware least squares model of general $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.

When the *group structure* (location and size of each group $\boldsymbol{x}_i$ $(i = 1, 2, \cdots, t)$, see Section 4.1) of target signal is not known a prior, it is reported that fully overlapping group setting in the OGS (overlapping group shrinkage) model [46] is more robust against unknown grouping structure than nonoverlapping but possibly erroneous group setting. As an alternative approach, Obozinski, Jacob and Vert have proposed the latent group lasso [79], for selection of relevant blocks from predefined overlapping groups. Recently, by applying the proximity operator of the perspective function [80], Kuroda and Kitahara propose in [81] a further alternative approach to nonoverlapping group sparse signal recovery problem by estimating the optimal partition.

2.3.2 Group Sparse Classification

Classification is one of fundamental tasks in the field of the signal and image processing and pattern recognition. For a classification problem with t classes of subjects, the training samples can formulate a dictionary matrix $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_t] \in \mathbb{R}^{m\times n}$, 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 $n = \sum_{i=1}^{t} n_i$ 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) [82] 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 $\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu\|\boldsymbol{x}\|_0$. In most SRC based approaches, ℓ_0 regularizer is relaxed to ℓ_1, and the model becomes the well-known Lasso model (2.15) [19].

The label information of the dictionary atoms is not utilized 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 [51–53], 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 $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_t^\top]^\top \in \mathbb{R}^n$, where $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in_i}]^\top \in \mathbb{R}^{n_i}$ $(i = 1, 2, \cdots, t)$. 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 (2.13) which is NP-hard. 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. We give a simple but clear explanation in Section 5.1, to show the bias of $\ell_{2,1}$-norm caused by group size in the application of group sparse classification.

More generally, weighted $\ell_{2,1}$-norm has also been used as the regularizer [67, 83, 84]. For example, Tang, et al. [67] 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,

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\mu\|\boldsymbol{W}\boldsymbol{x}\|_{2,1}, \tag{2.16}$$

where the weight matrix $\boldsymbol{W}$ is obtained by

$$\boldsymbol{W}=\text{BlockDiag}(\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_t) \quad \text{and} \quad \boldsymbol{W}_i=\text{Diag}(w_i d_{i1}, w_i d_{i2},\cdots,w_i d_{in_i}), \tag{2.17}$$

$$d_{ij}=\exp\left(\frac{\|\boldsymbol{y}-\boldsymbol{a}_{ij}\|_2}{\sigma_1}\right) \quad \text{and} \quad w_i=\exp\left(\frac{r_i-r_{\min}}{\sigma_2}\right), \tag{2.18}$$

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 $\{r_i\}_{i=1}^t$.

For the aforementioned methods, after obtaining the optimal solution (denoted by $\hat{\boldsymbol{x}}=[\hat{\boldsymbol{x}}_1^\top,\hat{\boldsymbol{x}}_2^\top,\cdots,\hat{\boldsymbol{x}}_t^\top]^\top$), 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 $\ell_{2,p}$ ($0<p<1$) [84] and group MCP [85] make the corresponding optimization problems nonconvex.

Chapter 3

Properties and GME Matrix Designs for LiGME Models

In this chapter, we first analysis some basic properties of the LiGME model. After that, we propose a new algebraic design of GME matrix. The design is unified but simple, which is applicable to general linear operator and does not require any iteration.

3.1 Fundamental Properties of LiGME Model

In this section, we analysis some basic properties of the LiGME model. Firstly, we show that the regularizer term of the LiGME model is nonnegative.

Proposition 3.1. The regularizer $\Psi_B \circ \mathfrak{L}(\cdot)$ in LiGME model (2.8) satisfies

$$\Psi_B \circ \mathfrak{L}(x) \geqslant 0 \quad \text{for all } x \in \mathcal{X}. \tag{3.1}$$

Proof.

$$\begin{aligned} &\Psi_B \circ \mathfrak{L}(x) \\ =&\Psi \circ \mathfrak{L}(x) - \min_{v \in \mathcal{Z}} \left[\Psi(v) + \frac{1}{2} \|B(\mathfrak{L}(x) - v)\|_{\widetilde{\mathcal{Z}}}^2 \right] \\ \geqslant&\Psi \circ \mathfrak{L}(x) - \Psi \circ \mathfrak{L}(x) \\ =&0 \end{aligned} \tag{3.2}$$

for $\forall x \in \mathcal{X}$, where the inequality is achieved by letting $v = \mathfrak{L}(x)$. □

Then, we show a relation between the null space of A and $B \circ \mathfrak{L}$.

Proposition 3.2. Assume that $\mu \in \mathbb{R}_{++}$ in (2.8) and GME matrix B satisfies the overall convexity condition (C1) in Proposition 2.2, then

$$\mathrm{null}(A) \subset \mathrm{null}(B\mathfrak{L}). \tag{3.3}$$

Proof. Let $x \in \mathrm{null}(A) \subset \mathcal{X}$. Then by (C1), we have

$$\begin{aligned} 0 \leqslant & \langle (A^*A - \mu \mathfrak{L}^* B^* B \mathfrak{L}) x, x \rangle \\ = & \langle A^*Ax, x \rangle - \mu \langle \mathfrak{L}^* B^* B \mathfrak{L} x, x \rangle \\ = & \langle Ax, Ax \rangle - \mu \langle B\mathfrak{L}x, B\mathfrak{L}x \rangle \\ = & -\mu \langle B\mathfrak{L}x, B\mathfrak{L}x \rangle \leqslant 0, \end{aligned}$$

which implies $B\mathfrak{L}x = 0$, and $x \in \mathrm{null}(B\mathfrak{L})$. Therefore, $\mathrm{null}(A) \subset \mathrm{null}(B\mathfrak{L})$. □

In Definition 2.13(c) and Proposition 2.3, we assume the existence of a minimizer of the LiGME model. In fact, under the overall convexity condition (C2) in Proposition 2.2, the existence of a minimizer of $J_{\Psi_B \circ \mathfrak{L}}$ is guaranteed in many cases, since the convex function $J_{\Psi_B \circ \mathfrak{L}}$ in (2.8) is guaranteed to be continuous over $\mathcal{X}$ [86, Corollary 8.40]. Example 3.1 lists some cases (not limited to these special cases).

Example 3.1. (a) If A^*A is nonsingular, the function $J_{\Psi_B \circ \mathfrak{L}}$ is coercive, which ensures the existence of its minimizer in $\mathcal{X}$ [86, Proposition 11.15]. In this case, the coercivity of $J_{\Psi_B \circ \mathfrak{L}}$ is verified by

(i) the coercivity of the least squares fidelity term of (2.8),

(ii) $(\forall x \in \mathcal{X})\ \Psi_B \circ \mathfrak{L}(x) \geqslant 0$ by Proposition 3.1.

(b) Even for general $A \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$, $J_{\Psi_B \circ \mathfrak{L}}$ is guaranteed to have a minimizer over any nonempty compact subset $S \subset \mathcal{X}$ [86, Theorem 1.29], by imposing a constraint $x \in S$ if necessary; see, e.g., (4.12) in Section 4.2, in a scenario of group sparsity aware estimation. In particular, a recently proposed constrained LiGME (cLiGME) model [6] admits certain *multiple-sets split feasibility type convex constraints* [87] (see also [43] for a single constraint case).

For a linearly involved regularizer $\Psi \circ \mathfrak{L}(x)$, there are two approaches to consider its enhancement by using the idea of Moreau envelope.

(a) Consider the Moreau enhancement of $\Psi \circ \mathfrak{L}$ straightly, like in [88]. It is given as

$$(\Psi \circ \mathfrak{L})_{\hat{B}}(\cdot) := \Psi(\mathfrak{L}\cdot) - \inf_{v \in \mathcal{X}} \left\{ \Psi(\mathfrak{L}v) + \frac{1}{2}\|\hat{B}(\cdot - v)\|_2^2 \right\}, \tag{3.4}$$

and variable is x.

(b) Considering the Moreau enhancement of Ψ as in LiGME model [5], which is defined as

$$\Psi_B(\cdot) := \Psi(\cdot) - \min_{v \in \mathcal{Z}} \left\{ \Psi(v) + \frac{1}{2}\|B(\cdot - v)\|_2^2 \right\}, \tag{3.5}$$

and then take $\mathfrak{L}(x)$ as the variable of (3.5).

We can show a useful relation between these two approaches.

Proposition 3.3. For $\Psi \in \Gamma_0(\mathcal{Z})$ which is coercive and $B \in \mathcal{B}(\mathcal{Z}, \tilde{\mathcal{Z}})$, assume $\mathfrak{L} \in \mathcal{B}(\mathcal{X}, \mathcal{Z})$ has full row-rank. Then for any $x \in \mathcal{X}$,

$$(\Psi \circ \mathfrak{L})_{B \circ \mathfrak{L}}(x) = \Psi_B \circ \mathfrak{L}(x), \tag{3.6}$$

Proof. See Appendix A.1. □

Remark 3.1. Applicability of (3.5) beyond (3.4).

(a) Even though Ψ is even symmetric, coercive and prox-friendly (easy to deal with least squares model with (3.5)), $\Psi \circ \mathfrak{L}$ is not necessarily coercive or prox-friendly, which makes the least squares model with (3.4) requires inner loop to obtain proximal operation of $\Psi \circ \mathfrak{L}$. As found in [86, 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 $\mathrm{Prox}_{\Psi \circ \mathfrak{L}}(x) = x + \mu^{-1}\mathfrak{L}^*(\mathrm{Prox}_{\mu\Psi}(\mathfrak{L}x) - \mathfrak{L}x)$ for $x \in \mathcal{X}$. 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.

(b) The Moreau enhancement of Ψ, i.e., (3.5), is proven to bridge the gap between Ψ and discrete measure functions. For example, $(\|\cdot\|_1)_B$ bridges the gap between $\|\cdot\|_0$ and $\|\cdot\|_1$ (see [5, Example 2(b)]), $(\|\cdot\|_{\mathrm{nuc}})_B$ bridges the gap between $\mathrm{rank}(\cdot)$ and $\|\cdot\|_{\mathrm{nuc}}$ (see [5, Example 2(c)]), $(\|\cdot\|_{2,1})_B$ bridges the gap between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{2,1}$ (see Proposition 4.1 in Section 4.1). However, the meaning of (3.4) has not been proven.

Therefore, in this thesis, we focus on (3.5).

3.2 Proposed Algebraic Design of GME Matrix Based on LDU Decomposition

In the LiGME model, since an algebraic design of GME matrix for rank deficient linear operator $\mathfrak{L}$ in (2.8) has not yet been reported, we propose a new algebraic design of GME matrix applicable to general linear operators, which can further enhances the applicability of the LiGME model.

Theorem 3.1. (Proposed design of GME matrix) In Definition 2.13, let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^l)$, $(\boldsymbol{A}, \mathfrak{L}, \mu) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{l\times n} \times \mathbb{R}_{++}$, and $r := \mathrm{rank}(\mathfrak{L}) > 0$. Choose nonsingular matrices $\boldsymbol{L} \in \mathbb{R}^{l\times l}$ and $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ satisfying

$$\boldsymbol{L}\mathfrak{L}\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{O}_{r\times(n-r)} \\ \boldsymbol{O}_{(l-r)\times r} & \boldsymbol{O}_{(l-r)\times(n-r)} \end{bmatrix}. \tag{3.7}$$

Let

$$\left[\begin{array}{c|c} \boldsymbol{A}_1 & \boldsymbol{A}_2 \end{array}\right] := \boldsymbol{A}\boldsymbol{R}, \tag{3.8}$$

where $\boldsymbol{A}_1 \in \mathbb{R}^{m\times r}$, $\boldsymbol{A}_2 \in \mathbb{R}^{m\times(n-r)}$, and choose $\boldsymbol{M} \in \mathbb{R}^{h\times r}$ to satisfy

$$\boldsymbol{M}^\top \boldsymbol{M} = {\boldsymbol{A}_1}^\top \boldsymbol{A}_1 - {\boldsymbol{A}_1}^\top \boldsymbol{A}_2 {\boldsymbol{A}_2}^\dagger \boldsymbol{A}_1 \in \mathbb{R}^{r\times r}. \tag{3.9}$$

Then

(a) For any $\theta \in [0, 1]$,

$$\boldsymbol{B}_\theta := \left[\begin{array}{cc} \sqrt{\theta/\mu}\boldsymbol{M} & \boldsymbol{O}_{h\times(l-r)} \end{array}\right] \boldsymbol{L} \in \mathbb{R}^{h\times l} \tag{3.10}$$

ensures $J_{\Psi_{\boldsymbol{B}_\theta}\circ\mathfrak{L}} \in \Gamma_0(\mathbb{R}^n)$.

(b) The existence of $\boldsymbol{M}$ satisfying (3.9) is guaranteed. For example,

$$\boldsymbol{M} = [P_{(\mathrm{ran}(\boldsymbol{A}_2))^\perp}]\boldsymbol{A}_1 \ (= \boldsymbol{A}_1 - [P_{\mathrm{ran}(\boldsymbol{A}_2)}]\boldsymbol{A}_1),$$

the projection of $\boldsymbol{A}_1$ onto the orthogonal complement subspace $(\mathrm{ran}(\boldsymbol{A}_2))^\perp$, satisfies (3.9). We denote the matrix expression of the projection operator $P_{\mathrm{ran}(\boldsymbol{A}_2)} : \mathbb{R}^m \to \mathbb{R}^m$ (see Table 2.1 in Section 2.1) by $[P_{\mathrm{ran}(\boldsymbol{A}_2)}] \in \mathbb{R}^{m\times m}$.

(c) For computation of the operator T_{LiGME} in Proposition 2(b), $\boldsymbol{B}_\theta^\top \boldsymbol{B}_\theta$ with $\theta \in [0, 1]$ is required, which is given by

$$\boldsymbol{B}_\theta^\top \boldsymbol{B}_\theta = \boldsymbol{L}^\top \begin{bmatrix} \frac{\theta}{\mu}\boldsymbol{M}^\top \boldsymbol{M} & \boldsymbol{O}_{r\times(l-r)} \\ \boldsymbol{O}_{(l-r)\times r} & \boldsymbol{O}_{(l-r)\times(l-r)} \end{bmatrix} \boldsymbol{L} \in \mathbb{R}^{l\times l}. \tag{3.11}$$

Proof. (a) For $\theta = 0$, we have $\boldsymbol{B}_\theta = \boldsymbol{O}_{h\times l}$ and $J_{\Psi_{\boldsymbol{O}_{h\times l}}\circ\mathfrak{Q}} \in \Gamma_0(\mathbb{R}^n)$ for all $\boldsymbol{y} \in \mathbb{R}^m$ by Proposition 2.2(a). For $\theta \in (0,1]$, we can show $\boldsymbol{A}^\top\boldsymbol{A} - \mu\mathfrak{Q}^\top\boldsymbol{B}_\theta^\top\boldsymbol{B}_\theta\mathfrak{Q} \succeq \boldsymbol{O}_{n\times n}$ as follows. By equation (3.7) and definition (3.10), we obtain

$$
\begin{aligned}
&\mathfrak{Q}^\top\boldsymbol{B}_\theta^\top\boldsymbol{B}_\theta\mathfrak{Q}\\
=&\boldsymbol{R}^{-\top}\begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{O}_{r\times(l-r)}\\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{O}_{(n-r)\times(l-r)}\end{bmatrix}\boldsymbol{L}^{-\top}\boldsymbol{B}_\theta^\top\boldsymbol{B}_\theta\boldsymbol{L}^{-1}\begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{O}_{r\times(n-r)}\\ \boldsymbol{O}_{(l-r)\times r} & \boldsymbol{O}_{(l-r)\times(n-r)}\end{bmatrix}\boldsymbol{R}^{-1}\\
=&\boldsymbol{R}^{-\top}\begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{O}_{r\times(l-r)}\\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{O}_{(n-r)\times(l-r)}\end{bmatrix}\begin{bmatrix}\frac{\theta}{\mu}\boldsymbol{M}^\top\boldsymbol{M} & \boldsymbol{O}_{r\times(l-r)}\\ \boldsymbol{O}_{(l-r)\times r} & \boldsymbol{O}_{(l-r)\times(l-r)}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{O}_{r\times(n-r)}\\ \boldsymbol{O}_{(l-r)\times r} & \boldsymbol{O}_{(l-r)\times(n-r)}\end{bmatrix}\boldsymbol{R}^{-1}\\
=&\boldsymbol{R}^{-\top}\begin{bmatrix}\frac{\theta}{\mu}\boldsymbol{M}^\top\boldsymbol{M} & \boldsymbol{O}_{r\times(n-r)}\\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{O}_{(n-r)\times(n-r)}\end{bmatrix}\boldsymbol{R}^{-1}. \qquad (3.12)
\end{aligned}
$$

According to (3.8) and (3.12), we have

$$
\begin{aligned}
&\boldsymbol{O}_{n\times n} \preceq \boldsymbol{A}^\top\boldsymbol{A} - \mu\mathfrak{Q}^\top\boldsymbol{B}_\theta^\top\boldsymbol{B}_\theta\mathfrak{Q}\\
\Leftrightarrow\quad &\boldsymbol{O}_{n\times n} \preceq (\boldsymbol{A}\boldsymbol{R})^\top(\boldsymbol{A}\boldsymbol{R}) - \mu\begin{bmatrix}\frac{\theta}{\mu}\boldsymbol{M}^\top\boldsymbol{M} & \boldsymbol{O}_{r\times(n-r)}\\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{O}_{(n-r)\times(n-r)}\end{bmatrix}\\
\Leftrightarrow\quad &\boldsymbol{O}_{n\times n} \preceq \begin{bmatrix}\boldsymbol{A}_1^\top\boldsymbol{A}_1 - \theta\boldsymbol{M}^\top\boldsymbol{M} & \boldsymbol{A}_1^\top\boldsymbol{A}_2\\ \boldsymbol{A}_2^\top\boldsymbol{A}_1 & \boldsymbol{A}_2^\top\boldsymbol{A}_2\end{bmatrix}\\
\Leftrightarrow\quad &\boldsymbol{A}_2^\top\boldsymbol{A}_2 \succeq \boldsymbol{O}_{(n-r)\times(n-r)},\\
&\boldsymbol{A}_2^\top\boldsymbol{A}_2(\boldsymbol{A}_2^\top\boldsymbol{A}_2)^\dagger\boldsymbol{A}_2^\top\boldsymbol{A}_1 = \boldsymbol{A}_2^\top\boldsymbol{A}_1, \qquad (3.13)\\
&\text{and } \boldsymbol{A}_1^\top\boldsymbol{A}_1 - \theta\boldsymbol{M}^\top\boldsymbol{M} - \boldsymbol{A}_1^\top\boldsymbol{A}_2(\boldsymbol{A}_2^\top\boldsymbol{A}_2)^\dagger\boldsymbol{A}_2^\top\boldsymbol{A}_1 \succeq \boldsymbol{O}_{r\times r},
\end{aligned}
$$

where the last equivalence is verified by [89, Theorem 1][1]. Let us verify the last three conditions in (3.13).

(i) $\boldsymbol{A}_2^\top\boldsymbol{A}_2 \succeq \boldsymbol{O}_{r\times r}$ holds obviously.

(ii) Since $\boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger$ is a symmetric idempotent matrix [90, (3.7.6)], i.e., $(\boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger)^\top = \boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger$ and $\boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger(\boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger) = \boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger$, we have

$$\boldsymbol{A}_2(\boldsymbol{A}_2^\top\boldsymbol{A}_2)^\dagger\boldsymbol{A}_2^\top = \boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger(\boldsymbol{A}_2^\top)^\dagger\boldsymbol{A}_2^\top = \boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger.$$

Therefore, $\boldsymbol{A}_2^\top\boldsymbol{A}_2(\boldsymbol{A}_2^\top\boldsymbol{A}_2)^\dagger\boldsymbol{A}_2^\top\boldsymbol{A}_1 = \boldsymbol{A}_2^\top\boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger\boldsymbol{A}_1 = \boldsymbol{A}_2^\top\boldsymbol{A}_1$, that is, the second condition holds.

[1] [89, Theorem 1] Let $\boldsymbol{S} = \begin{bmatrix}\boldsymbol{S}_{11} & \boldsymbol{S}_{12}\\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22}\end{bmatrix} \in \mathbb{R}^{n\times n}$ be a symmetric matrix, where $\boldsymbol{S}_{11} \in \mathbb{R}^{r\times r}$, $\boldsymbol{S}_{12} \in \mathbb{R}^{r\times(n-r)}$, $\boldsymbol{S}_{21} \in \mathbb{R}^{(n-r)\times r}$ and $\boldsymbol{S}_{22} \in R^{(n-r)\times(n-r)}$. Then $\boldsymbol{S} \succeq \boldsymbol{O}_{n\times n}$ if and only if the following three conditions hold simultaneously: (i) $\boldsymbol{S}_{22} \succeq \boldsymbol{O}_{(n-r)\times(n-r)}$, (ii) $\boldsymbol{S}_{21} = \boldsymbol{S}_{22}\boldsymbol{S}_{22}^\dagger\boldsymbol{S}_{21}$, (iii) $\boldsymbol{S}_{11} \succeq \boldsymbol{S}_{12}\boldsymbol{S}_{22}^\dagger\boldsymbol{S}_{21}$.

(iii) The third condition can be written as $(1-\theta)(\boldsymbol{A}_1^\top \boldsymbol{A}_1 - \boldsymbol{A}_1^\top \boldsymbol{A}_2 \boldsymbol{A}_2{}^\dagger \boldsymbol{A}_1) \succeq \boldsymbol{O}_{r\times r}$. Since the eigenvalues of $\boldsymbol{A}_2\boldsymbol{A}_2^\dagger$ are not larger than 1, $\boldsymbol{A}_1^\top \boldsymbol{A}_1 - \boldsymbol{A}_1^\top \boldsymbol{A}_2 \boldsymbol{A}_2{}^\dagger \boldsymbol{A}_1 = \boldsymbol{A}_1^\top (\boldsymbol{I}_m - \boldsymbol{A}_2 \boldsymbol{A}_2{}^\dagger)\boldsymbol{A}_1$ is positive semidefinite. Together with the fact that $1-\theta \geqslant 0$, the third condition also holds.

Thus $J_{\Psi_{\boldsymbol{B}_\theta}\circ \mathfrak{L}} \in \Gamma_0(\mathbb{R}^n)$ has been proven.

(b) By [86, Corollary 3.24, Proposition 3.30], $\boldsymbol{M} = [P_{(\mathrm{ran}(\boldsymbol{A}_2))^\perp}]\boldsymbol{A}_1 = (\boldsymbol{I}_m - \boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger)\boldsymbol{A}_1$. Then

$$\begin{aligned}\boldsymbol{M}^\top \boldsymbol{M} &= \boldsymbol{A}_1^\top (\boldsymbol{I}_m - \boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger)(\boldsymbol{I}_m - \boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger)\boldsymbol{A}_1 \\ &= \boldsymbol{A}_1^\top (\boldsymbol{I}_m - 2\boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger + \boldsymbol{A}_2\boldsymbol{A}_2{}^\dagger)\boldsymbol{A}_1 \\ &= \boldsymbol{A}_1{}^\top \boldsymbol{A}_1 - \boldsymbol{A}_1{}^\top \boldsymbol{A}_2 \boldsymbol{A}_2{}^\dagger \boldsymbol{A}_1.\end{aligned}$$

(c) It is clear by Proposition 2.3 and (3.10).

□

Remark 3.2. (Applicability and scalability of the proposed design in Theorem 3.1)

(a) The design in Theorem 3.1 is applicable to any linear operator $\mathfrak{L} \in \mathbb{R}^{l\times n}$, not necessarily of full row rank, which broadens the applicability of the LiGME model. For application to the case of rank deficient $\mathfrak{L}$, see Example 4.1(c) in group sparsity aware estimations.

(b) The existence of nonsingular matrices $\boldsymbol{L} \in \mathbb{R}^{l\times l}$ and $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ satisfying (3.7) is guaranteed [91, 0.4.6(f)]. One can use elementary row operations to attain RREF (reduced row-echelon form) of $\mathfrak{L} \in \mathbb{R}^{l\times n}$, that is, $\boldsymbol{L}\mathfrak{L}$, and attain RCEF (reduced column-echelon form) of $\boldsymbol{L}\mathfrak{L}$ by elementary column operations, where $\boldsymbol{L}$ and $\boldsymbol{R}$ are multiplication of the elementary row operations and the elementary column operations, respectively.

This computation is equivalent to the LU (lower-upper) decomposition or the LDU decomposition, whose complexity is at most $ln^2 - \frac{1}{3}n^3$ $(l \geqslant n)$ or $nl^2 - \frac{1}{3}l^3$ $(l \leqslant n)$ [50, Section 5.5.2], [48, Section 4.2]. For any matrix $\mathfrak{L} \in \mathbb{R}^{l\times n}$ with rank $r > 0$, the LU decomposition (general case for rectangular matrix) factors it as

$$(\text{if } l \geqslant n) \quad \mathfrak{L} = \boldsymbol{P}\begin{bmatrix}\boldsymbol{L}_1 & \boldsymbol{O}_{r\times(n-r)} \\ \boldsymbol{L}_2 & \boldsymbol{O}_{(l-r)\times(n-r)}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_1 & \boldsymbol{U}_2 \\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{I}_{n-r}\end{bmatrix}\boldsymbol{Q}, \tag{3.14}$$

$$(\text{if } l \leqslant n) \quad \mathfrak{L} = \boldsymbol{P}\begin{bmatrix}\boldsymbol{L}_1 & \boldsymbol{O}_{r\times(l-r)} \\ \boldsymbol{L}_2 & \boldsymbol{I}_{l-r}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_1 & \boldsymbol{U}_2 \\ \boldsymbol{O}_{(l-r)\times r} & \boldsymbol{O}_{(l-r)\times(n-r)}\end{bmatrix}\boldsymbol{Q}, \tag{3.15}$$

where $\boldsymbol{L}_1 \in \mathbb{R}^{r\times r}$ is a unit lower triangular matrix, $\boldsymbol{U}_1 \in \mathbb{R}^{r\times r}$ is an upper triangular matrix, $\boldsymbol{P} \in \mathbb{R}^{l\times l}$ and $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ are two permutation matrices, $\boldsymbol{L}_2 \in \mathbb{R}^{(l-r)\times r}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{r\times(n-r)}$. By LDU decomposition, $\mathfrak{Q}$ can be factored as

$$\mathfrak{Q} = \boldsymbol{P}\begin{bmatrix}\boldsymbol{L}_1 & \boldsymbol{O}_{r\times(l-r)}\\ \boldsymbol{L}_2 & \boldsymbol{I}_{l-r}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_r & \boldsymbol{O}_{r\times(n-r)}\\ \boldsymbol{O}_{(l-r)\times r} & \boldsymbol{O}_{(l-r)\times(n-r)}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_1 & \boldsymbol{U}_2\\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{I}_{n-r}\end{bmatrix}\boldsymbol{Q}. \tag{3.16}$$

(c) Since the computation $\boldsymbol{M} = [P_{(\mathrm{ran}(\boldsymbol{A}_2))^\perp}]\boldsymbol{A}_1$ in Theorem 3.1(b) does not require in principle any eigendecomposition, the proposed design of $\boldsymbol{B}_\theta$ with $\boldsymbol{M} = [P_{(\mathrm{ran}(\boldsymbol{A}_2))^\perp}]\boldsymbol{A}_1$ in Theorem 3.1 is more scalable than the existing design in [5, Proposition 2] (see Corollary 3.1(a) below). Moreover, the computation for $\boldsymbol{B}_\theta^\top \boldsymbol{B}_\theta$ does not need iterative computation (unlike the non-algebraic iterative algorithm in [44, Algorithm 1]).

Corollary 3.1. Theorem 3.1 is a generalization of the following existing algebraic design methods of GME matrix.

(a) ([5, Proposition 2]) In Definition 2.13, let $(\mathcal{X},\mathcal{Y},\mathcal{Z}) = (\mathbb{R}^n,\mathbb{R}^m,\mathbb{R}^l)$, $(\boldsymbol{A},\mathfrak{Q},\mu) \in \mathbb{R}^{m\times n}\times\mathbb{R}^{l\times n}\times\mathbb{R}_{++}$, and $\mathrm{rank}(\mathfrak{Q}) = l$. Choose a nonsingular $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ satisfying $\mathfrak{Q}\boldsymbol{R} = [\boldsymbol{I}_l\ \boldsymbol{O}_{l\times(n-l)}]$. Then

$$\boldsymbol{B}_\theta := \sqrt{\theta/\mu}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top \in \mathbb{R}^{l\times l},\ \theta\in[0,1], \tag{3.17}$$

ensures $J_{\Psi_{\boldsymbol{B}_\theta}\circ\mathfrak{Q}} \in \Gamma_0(\mathbb{R}^n)$, where $[\boldsymbol{A}_1 \mid \boldsymbol{A}_2] := \boldsymbol{A}\boldsymbol{R}$ and

$$\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top := \boldsymbol{A}_1^\top\boldsymbol{A}_1 - \boldsymbol{A}_1^\top\boldsymbol{A}_2(\boldsymbol{A}_2^\top\boldsymbol{A}_2)^\dagger\boldsymbol{A}_2^\top\boldsymbol{A}_1 \in \mathbb{R}^{l\times l}$$

is an eigendecomposition. (NOTE: the GME matrix design in (3.17) is a special case of (3.10) with $r = l$, $\boldsymbol{L} = \boldsymbol{I}_l$ and $\boldsymbol{M} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top$.)

(b) ([41]) In Definition 2.13, let $(\mathcal{X},\mathcal{Y},\mathcal{Z}) = (\mathbb{R}^n,\mathbb{R}^m,\mathbb{R}^n)$, $(\boldsymbol{A},\mathfrak{Q},\mu) \in \mathbb{R}^{m\times n}\times\mathbb{R}^{n\times n}\times\mathbb{R}_{++}$, and $\mathfrak{Q} = \boldsymbol{I}_n$. Then

$$\boldsymbol{B}_\theta := \sqrt{\theta/\mu}\boldsymbol{A} \in \mathbb{R}^{m\times n},\ \theta\in[0,1], \tag{3.18}$$

ensures $J_{\Psi_{\boldsymbol{B}_\theta}\circ\mathfrak{Q}} \in \Gamma_0(\mathbb{R}^n)$. (NOTE: The GME matrix design in (3.18) is a special case of (3.10) with $r = l = n$, $\boldsymbol{L} = \boldsymbol{R} = \boldsymbol{I}_n$ and $\boldsymbol{M} = \boldsymbol{A}$.)

Corollary 3.2. (Proposed design of GME matrix for multiple LiGME regularizers)
If a LiGME model has multiple regularizers, i.e., the LiGME model as follows,

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \sum_{i=1}^{\mathcal{M}}\mu_i(\Psi_i)_{\boldsymbol{B}_i}\circ\mathfrak{Q}_i(\boldsymbol{x}), \tag{3.19}$$

where $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\mu_i > 0$, $\Psi_i \in \Gamma_0(\mathbb{R}^{l_i})$ coercive, dom $\Psi_i = \mathbb{R}^{l_i}$, $\boldsymbol{B}_i \in \mathbb{R}^{h_i\times l_i}$ and $\mathfrak{L}_i \in \mathbb{R}^{l_i\times n}$, $(\Psi_i)_{\boldsymbol{B}_i}(\cdot) = \Psi_i(\cdot) - \min_{\boldsymbol{v}\in\mathbb{R}^{l_i}} \left[\Psi_i(\boldsymbol{v}) + \frac{1}{2}\|\boldsymbol{B}_i(\cdot - \boldsymbol{v})\|_2^2\right]$ $(i = 1, 2, \cdots, \mathcal{M})$.

Choose $\alpha_i \in \mathbb{R}_{++}$ $(i = 1, 2, \cdots, \mathcal{M})$ satisfying $\sum_{i=1}^{\mathcal{M}} \alpha_i = 1$ and apply Theorem 3.1 to $(\sqrt{\alpha_i}\boldsymbol{A}, \mathfrak{L}_i, \mu_i)$ to obtain $\boldsymbol{B}_i$ satisfying

$$\alpha_i \boldsymbol{A}^\top \boldsymbol{A} - \mu_i \mathfrak{L}_i^\top \boldsymbol{B}_i^\top \boldsymbol{B}_i \mathfrak{L}_i \succeq \boldsymbol{O}_{n\times n} \ (i = 1, 2, \cdots, \mathcal{M}), \tag{3.20}$$

then the overall convexity of (3.19) is achieved.

Proof. Let $\Psi := \mathbb{R}^l \to (-\infty, \infty] : \ \boldsymbol{z} = [\boldsymbol{z}_1^\top, \boldsymbol{z}_2^\top, \cdots, \boldsymbol{z}_{\mathcal{M}}^\top]^\top \mapsto \sum_{i=1}^{\mathcal{M}} \mu_i \Psi_i(\boldsymbol{z}_i)$ with $\boldsymbol{z}_i \in \mathbb{R}^{l_i}$ $(i = 1, 2, \cdots, \mathcal{M})$ and $l = \sum_{i=1}^{\mathcal{M}} l_i$, $\mathfrak{L} : \mathbb{R}^n \to \mathbb{R}^l : \ \boldsymbol{x} \mapsto [(\mathfrak{L}_1\boldsymbol{x})^\top, (\mathfrak{L}_2\boldsymbol{x})^\top, \cdots, (\mathfrak{L}_{\mathcal{M}}\boldsymbol{x})^\top]^\top$ and $\boldsymbol{B} : \mathbb{R}^l \to \mathbb{R}^h : \ [\boldsymbol{z}_1^\top, \boldsymbol{z}_2^\top, \cdots, \boldsymbol{z}_{\mathcal{M}}^\top]^\top \mapsto [(\sqrt{\mu_1}\boldsymbol{B}_1\boldsymbol{z}_1)^\top, (\sqrt{\mu_2}\boldsymbol{B}_2\boldsymbol{z}_2)^\top, \cdots, (\sqrt{\mu_{\mathcal{M}}}\boldsymbol{B}_{\mathcal{M}}\boldsymbol{z}_{\mathcal{M}})^\top]^\top$ with $h = \sum_{i=1}^{\mathcal{M}} h_i$. We have

$$\Psi_{\boldsymbol{B}} \circ \mathfrak{L} = \sum_{i=1}^{\mathcal{M}} \mu_i (\Psi_i)_{\boldsymbol{B}_i} \circ \mathfrak{L}_i. \tag{3.21}$$

Together with (3.20), we have

$$\begin{aligned}
&\boldsymbol{A}^\top \boldsymbol{A} - \mathfrak{L}^\top \boldsymbol{B}^\top \boldsymbol{B} \mathfrak{L} \\
=&\boldsymbol{A}^\top \boldsymbol{A} - \sum_{i=1}^{\mathcal{M}} \mu_i \mathfrak{L}_i^\top \boldsymbol{B}_i^\top \boldsymbol{B}_i \mathfrak{L}_i \\
=&\sum_{i=1}^{\mathcal{M}} \left(\alpha_i \boldsymbol{A}^\top \boldsymbol{A} - \mu_i \mathfrak{L}_i^\top \boldsymbol{B}_i^\top \boldsymbol{B}_i \mathfrak{L}_i\right) \succeq \boldsymbol{O}_{n\times n}.
\end{aligned}$$

Therefore, the overall convexity of (3.19) is achieved according to Proposition 2.2. □

Remark 3.3. In Corollary 3.2, the choice of α_i depends on how much we want to enhance the nonconvexity of the i-th regularizer. The larger α_i is, the stronger the nonconvexity of the i-th regularizer.

3.3 Proposed Algebraic Design of GME Matrix Based on SVD

We also present an alternative design of GME matrix in Proposition 3.4 when the singular value decomposition (SVD) of $\mathfrak{L}$ is available.

Proposition 3.4. In Definition 2.13, let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^l)$, $(\boldsymbol{A}, \mathfrak{L}, \mu) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{l\times n} \times \mathbb{R}_{++}$, and $r := \mathrm{rank}(\mathfrak{L})$. Let $\mathfrak{U}\boldsymbol{\Sigma}\mathfrak{V}^\top = \mathfrak{L}$ be the SVD of $\mathfrak{L}$, where $\mathfrak{U} \in \mathbb{R}^{l\times l}$ and $\mathfrak{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\boldsymbol{\Sigma} \in \mathbb{R}^{l\times n}$ is a rectangular diagonal matrix with non-negative real numbers on the diagonal. Let $[\ \widehat{\boldsymbol{A}}_1 \mid \widehat{\boldsymbol{A}}_2\ ] := \boldsymbol{A}\mathfrak{V}$, where $\widehat{\boldsymbol{A}}_1 \in \mathbb{R}^{m\times r}$, $\widehat{\boldsymbol{A}}_2 \in \mathbb{R}^{m\times(n-r)}$, and choose $h \in \mathbb{N}$, $\widehat{\boldsymbol{M}} \in \mathbb{R}^{h\times r}$ satisfying $\widehat{\boldsymbol{M}}^\top\widehat{\boldsymbol{M}} := \widehat{\boldsymbol{A}}_1^\top\widehat{\boldsymbol{A}}_1 - \widehat{\boldsymbol{A}}_1^\top\widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^\dagger\widehat{\boldsymbol{A}}_1 \in \mathbb{R}^{r\times r}$. Then for any $\theta \in [0,1]$,

$$\widehat{\boldsymbol{B}}_\theta := \left[\sqrt{\theta/\mu}\widehat{\boldsymbol{M}} \quad \boldsymbol{O}_{h\times(n-r)}\right]\boldsymbol{\Sigma}^\dagger\mathfrak{U}^\top \in \mathbb{R}^{h\times l} \tag{3.22}$$

ensures $J_{\Psi_{\widehat{\boldsymbol{B}}_\theta}\circ\mathfrak{L}} \in \Gamma_0(\mathbb{R}^n)$.

Proof. If $\theta = 0$, $\widehat{\boldsymbol{B}}_\theta = \boldsymbol{O}_{h\times l}$ and $J_{\Psi_{\boldsymbol{O}_{h\times l}}\circ\mathfrak{L}} \in \Gamma_0(\mathbb{R}^n)$ for all $\boldsymbol{y} \in \mathbb{R}^m$. Let $\theta \in (0,1]$. We need to show that $\boldsymbol{A}^\top\boldsymbol{A} - \mu\mathfrak{L}^\top\widehat{\boldsymbol{B}}_\theta^\top\widehat{\boldsymbol{B}}_\theta\mathfrak{L} \succeq \boldsymbol{O}_{n\times n}$. By $\mathfrak{U}\boldsymbol{\Sigma}\mathfrak{V}^\top = \mathfrak{L}$ and the definition (3.22), we have that

$$\begin{aligned}
&\mathfrak{L}^\top\widehat{\boldsymbol{B}}_\theta^\top\widehat{\boldsymbol{B}}_\theta\mathfrak{L}\\
=&\mathfrak{V}\boldsymbol{\Sigma}^\top\mathfrak{U}^\top\mathfrak{U}(\boldsymbol{\Sigma}^\dagger)^\top\begin{bmatrix}\sqrt{\theta/\mu}\widehat{\boldsymbol{M}}^\top\\ \boldsymbol{O}_{(n-r)\times h}\end{bmatrix}\left[\sqrt{\theta/\mu}\widehat{\boldsymbol{M}} \quad \boldsymbol{O}_{h\times(n-r)}\right]\boldsymbol{\Sigma}^\dagger\mathfrak{U}^\top\mathfrak{U}\boldsymbol{\Sigma}\mathfrak{V}^\top\\
=&\mathfrak{V}\begin{bmatrix}\theta/\mu\widehat{\boldsymbol{M}}^\top\widehat{\boldsymbol{M}} & \boldsymbol{O}_{r\times(n-r)}\\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{O}_{(n-r)\times(n-r)}\end{bmatrix}\mathfrak{V}^\top.
\end{aligned} \tag{3.23}$$

Therefore, by $[\ \widehat{\boldsymbol{A}}_1 \mid \widehat{\boldsymbol{A}}_2\ ] := \boldsymbol{A}\mathfrak{V}$ and (3.23), we have

$$\begin{aligned}
&\boldsymbol{O}_{n\times n} \preceq \boldsymbol{A}^\top\boldsymbol{A} - \mu\mathfrak{L}^\top\widehat{\boldsymbol{B}}_\theta^\top\widehat{\boldsymbol{B}}_\theta\mathfrak{L}\\
\Leftrightarrow\ &\boldsymbol{O}_{n\times n} \preceq (\boldsymbol{A}\mathfrak{V})^\top(\boldsymbol{A}\mathfrak{V}) - \mu\begin{bmatrix}\frac{\theta}{\mu}\widehat{\boldsymbol{M}}^\top\widehat{\boldsymbol{M}} & \boldsymbol{O}_{r\times(n-r)}\\ \boldsymbol{O}_{(n-r)\times r} & \boldsymbol{O}_{(n-r)\times(n-r)}\end{bmatrix}\\
\Leftrightarrow\ &\boldsymbol{O}_{n\times n} \preceq \begin{bmatrix}\widehat{\boldsymbol{A}}_1^\top\widehat{\boldsymbol{A}}_1 - \theta\widehat{\boldsymbol{M}}^\top\widehat{\boldsymbol{M}} & \widehat{\boldsymbol{A}}_1^\top\widehat{\boldsymbol{A}}_2\\ \widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_1 & \widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_2\end{bmatrix}\\
\Leftrightarrow\ &\widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_2 \succeq \boldsymbol{O}_{(n-r)\times(n-r)},\\
&\widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_2(\widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_2)^\dagger\widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_1 = \widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_1,\\
&\text{and } \widehat{\boldsymbol{A}}_1^\top\widehat{\boldsymbol{A}}_1 - \theta\widehat{\boldsymbol{M}}^\top\widehat{\boldsymbol{M}} - \widehat{\boldsymbol{A}}_1^\top\widehat{\boldsymbol{A}}_2(\widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_2)^\dagger\widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_1 \succeq \boldsymbol{O}_{r\times r},
\end{aligned} \tag{3.24}$$

where the last equivalence is confirmed by [89, Theorem 1]. We verify the last three conditions in (3.24) one by one. Firstly, $\widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_2 \succeq \boldsymbol{O}_{r\times r}$ holds obviously. Since $\widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^\dagger$ is a symmetric idempotent matrix [90, (3.7.6)], we have $\widehat{\boldsymbol{A}}_2(\widehat{\boldsymbol{A}}_2^\top\widehat{\boldsymbol{A}}_2)^\dagger\widehat{\boldsymbol{A}}_2^\top = \widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^\dagger(\widehat{\boldsymbol{A}}_2^\top)^\dagger\widehat{\boldsymbol{A}}_2^\top =$

$\widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^{\dagger}$. Therefore, $\widehat{\boldsymbol{A}}_2^{\top}\widehat{\boldsymbol{A}}_2(\widehat{\boldsymbol{A}}_2^{\top}\widehat{\boldsymbol{A}}_2)^{\dagger}\widehat{\boldsymbol{A}}_2^{\top}\widehat{\boldsymbol{A}}_1 = \widehat{\boldsymbol{A}}_2^{\top}\widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^{\dagger}\widehat{\boldsymbol{A}}_1 = \widehat{\boldsymbol{A}}_2^{\top}\widehat{\boldsymbol{A}}_1$, that is, the second condition holds. The third condition can be written as $(1-\theta)(\widehat{\boldsymbol{A}}_1^{\top}\widehat{\boldsymbol{A}}_1 - \widehat{\boldsymbol{A}}_1^{\top}\widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^{\dagger}\widehat{\boldsymbol{A}}_1) \succeq \boldsymbol{O}_{r\times r}$. Since the eigenvalues of $\widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^{\dagger}$ are not larger than 1, $\widehat{\boldsymbol{A}}_1^{\top}\widehat{\boldsymbol{A}}_1 - \widehat{\boldsymbol{A}}_1^{\top}\widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^{\dagger}\widehat{\boldsymbol{A}}_1 = \widehat{\boldsymbol{A}}_1^{\top}(\boldsymbol{I}_m - \widehat{\boldsymbol{A}}_2\widehat{\boldsymbol{A}}_2^{\dagger})\widehat{\boldsymbol{A}}_1$ is positive semidefinite. Together with the fact that $1-\theta \geqslant 0$, the third condition also holds. Thus $J_{\Psi_{\boldsymbol{B}_\theta}\circ\Omega} \in \Gamma_0(\mathbb{R}^n)$ has been proven. □

Remark 3.4. For the two designs in Theorem 3.1 and Proposition 3.4, although the obtained GME matrices $\boldsymbol{B}_\theta$ and $\widehat{\boldsymbol{B}}_\theta$ are different, we verified experimentally that they lead to the same estimation results of the LiGME model.

Chapter 4

Application to Group Sparsity Aware Signal Recovery

In this chapter, we consider a specific problem, that is, group sparsity aware signal recovery. First, by defining a grouping operator $\mathscr{G}$, we consider a general expression of grouping, which allows overlapping and incomplete cover cases. Then we focus on the group sparsity pursuing regularizers, where the proposed nonconvex GME penalty $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$ can bridge the gap between $\|\cdot\|_{2,1}$ and $\|\cdot\|_{2,0}$. By combining the grouping operator $\mathscr{G}$ and the GME penalty $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$, we present a new LiGME formulation for group sparsity aware signal recovery problems. For general case, we encounter the rank deficiency of $\mathfrak{L}$. The overall convexity of the LiGME model is achieved by the proposed GME matrix in Theorem 3.1. We applied the model to group sparsity aware problem of unknown group structure. We verify the effectiveness of the proposed LiGME model with the GME matrix design by numerical experiments.

4.1 Grouping Operator and Proposed LiGME Regularizer

In many applications, the target signal has group sparsity property. In this study, we consider a general expression of grouping. Let $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^\top \in \mathbb{R}^n$ be a vector and $\boldsymbol{x}_{\mathfrak{g}_i} \in \mathbb{R}^{n_i}$ $(i = 1, 2, \cdots, t)$ be subvectors of $\boldsymbol{x}$, where t is the number of groups, $(\varnothing \neq)\ \mathfrak{g}_i = \{\varrho \in \{1, 2, \cdots, n\} \mid p_i \leqslant \varrho \leqslant q_i\}$ for some $p_i, q_i \in \{1, 2, \cdots, n\}$ (i.e. $p_i = \min \mathfrak{g}_i$ and $q_i = \max \mathfrak{g}_i$) is the index set corresponding to i-th subvector and $\mathfrak{g}_i \neq \mathfrak{g}_j$ $(1 \leqslant i < j \leqslant t)$. That is, $\boldsymbol{x}_{\mathfrak{g}_i} = (x_\varrho)_{\varrho \in \mathfrak{g}_i}$ and $\#\mathfrak{g}_i = q_i - p_i + 1 = n_i$. Then $[\boldsymbol{x}_{\mathfrak{g}_1}^\top, \boldsymbol{x}_{\mathfrak{g}_2}^\top, \cdots, \boldsymbol{x}_{\mathfrak{g}_t}^\top]^\top \in \mathbb{R}^{n'}$ with

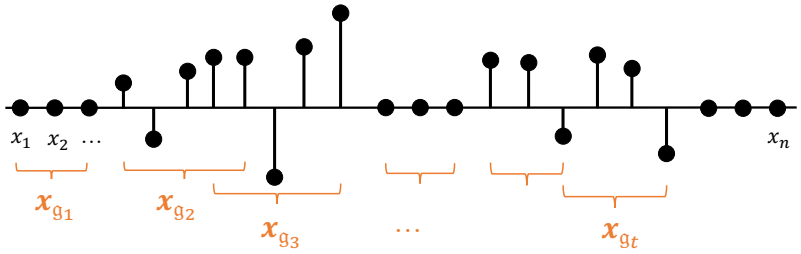


Fig. 4.1: Illustration of a group structure of a vector $\boldsymbol{x}$ (with overlapping and incomplete).

$\sum_{i=1}^{t} n_i = n'$ forms a new vector with group partition. The index sets $\mathfrak{g}_i$ $(i = 1, \cdots, t)$ express a *group structure* of $\boldsymbol{x} \in \mathbb{R}^n$.

Note that we consider general group configurations which allows overlapping and/or incomplete cover (see Fig. 4.1 for an illustration). For example, if $\mathfrak{g}_i \cap \mathfrak{g}_j \neq \varnothing$ $(i \neq j)$, the i-th group and j-th group overlap; if $\bigcup_{i=1}^{t} \mathfrak{g}_i \subsetneq \{1, 2, \cdots, n\}$, this case corresponds an incomplete cover, which means that some entries of $\boldsymbol{x}$ are not included in $[\boldsymbol{x}_{\mathfrak{g}_1}^\top, \boldsymbol{x}_{\mathfrak{g}_2}^\top, \cdots, \boldsymbol{x}_{\mathfrak{g}_t}^\top]^\top$.

Definition 4.1. (Grouping operator) We define a grouping operator

$$\mathscr{G} := [\mathscr{G}_1^\top, \mathscr{G}_2^\top, \cdots, \mathscr{G}_t^\top]^\top \in \mathbb{R}^{n' \times n}, \tag{4.1}$$

where $\mathscr{G}_i \in \mathbb{R}^{n_i \times n}$ extracts consecutive entries from $\boldsymbol{x} \in \mathbb{R}^n$ to form the i-th group $\boldsymbol{x}_{\mathfrak{g}_i} = \mathscr{G}_i \boldsymbol{x} \in \mathbb{R}^{n_i}$ $(i = 1, 2, \cdots, t)$, and $[\boldsymbol{x}_{\mathfrak{g}_1}^\top, \boldsymbol{x}_{\mathfrak{g}_2}^\top, \cdots, \boldsymbol{x}_{\mathfrak{g}_t}^\top]^\top = \mathscr{G}\boldsymbol{x} \in \mathbb{R}^{n'}$. Obviously, $\mathscr{G}_i$ is composed of the zero matrices and the identity matrix, i.e.,

$$\mathscr{G}_i = [\boldsymbol{O}_{n_i \times (p_i - 1)}, \boldsymbol{I}_{n_i}, \boldsymbol{O}_{n_i \times (n - q_i)}]. \tag{4.2}$$

We show an example in Fig. 4.2, where the grouping operator $\mathscr{G}$ transforms $\boldsymbol{x}$ (with overlaps) to $\mathscr{G}\boldsymbol{x}$ (without overlaps).

Remark 4.1. In Definition 4.1, a natural group sparsity measure of $\boldsymbol{x} \in \mathbb{R}^n$ with grouping operator $\mathscr{G} = [\mathscr{G}_1^\top, \mathscr{G}_2^\top, \cdots \mathscr{G}_t^\top]^\top \in \mathbb{R}^{n' \times n}$ is

$$\begin{aligned} \|\boldsymbol{x}\|_{2,0}^{(\mathscr{G}_i)_{i=1}^t} &:= \|(\|\mathscr{G}_1 \boldsymbol{x}\|_2, \|\mathscr{G}_2 \boldsymbol{x}\|_2, \cdots, \|\mathscr{G}_t \boldsymbol{x}\|_2)\|_0 \\ &= \|(\|\boldsymbol{x}_{\mathfrak{g}_1}\|_2, \|\boldsymbol{x}_{\mathfrak{g}_2}\|_2, \cdots, \|\boldsymbol{x}_{\mathfrak{g}_t}\|_2)\|_0 . \end{aligned} \tag{4.3}$$

As the standard convex approximation of (4.3), we introduce the sum (i.e. ℓ_1 norm) of the group-wise ℓ_2 norms as

$$\begin{aligned} \|\boldsymbol{x}\|_{2,1}^{(\mathscr{G}_i)_{i=1}^t} &:= \sum_{i=1}^{t} \|\mathscr{G}_i \boldsymbol{x}\|_2 = \sum_{i=1}^{t} \|\boldsymbol{x}_{\mathfrak{g}_i}\|_2 \\ &= \|(\|\boldsymbol{x}_{\mathfrak{g}_1}\|_2, \|\boldsymbol{x}_{\mathfrak{g}_2}\|_2, \cdots, \|\boldsymbol{x}_{\mathfrak{g}_t}\|_2)\|_1 , \end{aligned} \tag{4.4}$$

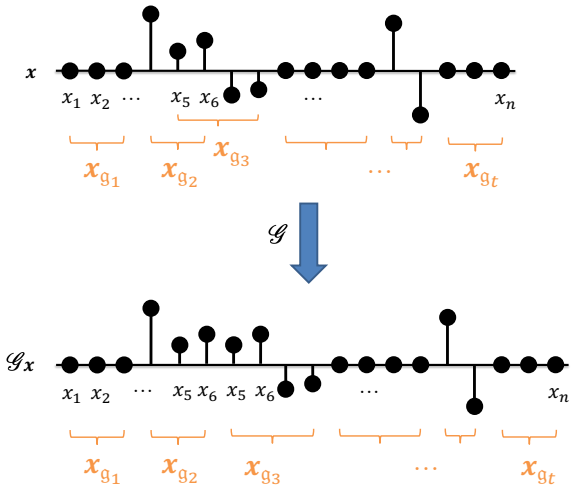


Fig. 4.2: Illustration of the performing of grouping operator $\mathscr{G}$ on a vector $\boldsymbol{x}$ (with overlapping).

because ℓ_1-norm is known as the largest convex minorant of $\|\cdot\|_0$ on $[-1,1]^l$ $(\subset \mathbb{R}^l)$ of ℓ_0 pseudo-norm. Clearly, by using the $\ell_{2,0}$ pseudo-norm (see, e.g., [77]),

$$\begin{aligned} \|\cdot\|_{2,0} : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_t} &\to \mathbb{R} \\ : [\boldsymbol{x}_{\mathfrak{g}_1}^\top, \boldsymbol{x}_{\mathfrak{g}_2}^\top, \cdots, \boldsymbol{x}_{\mathfrak{g}_t}^\top]^\top &\mapsto \|(\|\boldsymbol{x}_{\mathfrak{g}_1}\|_2, \|\boldsymbol{x}_{\mathfrak{g}_2}\|_2, \cdots, \|\boldsymbol{x}_{\mathfrak{g}_t}\|_2)\|_0, \end{aligned} \tag{4.5}$$

and $\ell_{2,1}$ norm (see, e.g., [47])

$$\begin{aligned} \|\cdot\|_{2,1} : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_t} &\to \mathbb{R} \\ : [\boldsymbol{x}_{\mathfrak{g}_1}^\top, \boldsymbol{x}_{\mathfrak{g}_2}^\top, \cdots, \boldsymbol{x}_{\mathfrak{g}_t}^\top]^\top &\mapsto \sum_{i=1}^{t} \|\boldsymbol{x}_{\mathfrak{g}_i}\|_2, \end{aligned} \tag{4.6}$$

we can express (4.3) and (4.4) as

$$(\forall \boldsymbol{x} \in \mathbb{R}^n) \qquad \|\boldsymbol{x}\|_{2,0}^{(\mathscr{G}_i)_{i=1}^t} = \|\cdot\|_{2,0} \circ \mathscr{G}(\boldsymbol{x}), \quad \text{and} \tag{4.7}$$

$$(\forall \boldsymbol{x} \in \mathbb{R}^n) \qquad \|\boldsymbol{x}\|_{2,1}^{(\mathscr{G}_i)_{i=1}^t} = \|\cdot\|_{2,1} \circ \mathscr{G}(\boldsymbol{x}) \tag{4.8}$$

respectively. Therefore, the function $\|\cdot\|_{2,1}^{(\mathscr{G}_i)_{i=1}^t} : \mathbb{R}^n \to \mathbb{R}$ is a norm of $\mathbb{R}^n$ if and only if $\operatorname{rank}(\mathscr{G}) = n$.

Example 4.1. (Typical grouping)

(a) (Simplest case without grouping) If $t = n$, and $\mathfrak{g}_i = \{i\}$ for all $i = 1, 2, \cdots, n$, each group has only one entry without duplication. Here $\mathscr{G} = \boldsymbol{I}_n$.

(b) (Standard partitioning) If $\mathfrak{g}_i \cap \mathfrak{g}_j = \varnothing$ for any $1 \leqslant i < j \leqslant t$ and $\bigcup_{i=1}^{t} \mathfrak{g}_i = \{1, 2, \cdots, n\}$, $\sum_{i=1}^{t} n_i = n$, the groups do not overlap and all entries are covered, which is the most widely considered group partitioning in the applications. Here $\mathscr{G}_i = [\boldsymbol{O}_{n_i \times \sum_{j=1}^{i-1} n_j}, \boldsymbol{I}_{n_i}, \boldsymbol{O}_{n_i \times \sum_{j=i+1}^{t} n_j}]$, and $\mathscr{G} = \boldsymbol{I}_n$.

(c) (Grouping operator with common group size and common overlapping degree) If $n_i = s$ for all $1 \leqslant i \leqslant t$ and $\#(\mathfrak{g}_i \cap \mathfrak{g}_{i+1}) = d$ $(0 \leqslant d \leqslant s-1)$ for all $1 \leqslant i \leqslant t-1$, all the groups have the common size s and common *overlapping degree* d. Here

$$\mathscr{G}_i = [\boldsymbol{O}_{s \times (i-1)(s-d)}, \boldsymbol{I}_s, \boldsymbol{O}_{s \times (n-is+id-d)}]. \tag{4.9}$$

When $d = 0$, the groups are nonoverlapping; when $d = s-1$, the groups are fully overlapping [46]. Note that when $d > 0$, $\mathscr{G}$ does not have full row rank.

For group sparsity aware estimation, we consider the GME penalty $(\|\cdot\|_{2,1})_{\boldsymbol{B}} : \mathbb{R}^n \to (-\infty, \infty]$ as follows,

$$(\|\cdot\|_{2,1})_{\boldsymbol{B}}(\boldsymbol{x}) = \|\boldsymbol{x}\|_{2,1} - \min_{\boldsymbol{v} \in \mathbb{R}^n} \left[\|\boldsymbol{v}\|_{2,1} + \frac{1}{2}\|\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{v})\|_2^2 \right], \tag{4.10}$$

which is specialization of Ψ_B in (2.7). First, we show that the GME penalty $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$ can bridge the gap between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{2,1}$.

Proposition 4.1. (GME of $\|\cdot\|_{2,1}$ can bridge the gap between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{2,1}$.) Let $\boldsymbol{B}_\gamma := \frac{1}{\sqrt{\gamma}}\boldsymbol{I}_n$ for $\gamma > 0$. Then, for any $\boldsymbol{x} \in \mathbb{R}^n$,

$$\lim_{\gamma \downarrow 0} \frac{2}{\gamma} (\|\cdot\|_{2,1})_{\boldsymbol{B}_\gamma}(\boldsymbol{x}) = \|\boldsymbol{x}\|_{2,0}. \tag{4.11}$$

Together with the fact that $(\|\cdot\|_{2,1})_{\boldsymbol{O}_{n\times n}}(\boldsymbol{x}) = \|\boldsymbol{x}\|_{2,1}$, the regularization term $\frac{2}{\gamma}(\|\cdot\|_{2,1})_{\boldsymbol{B}_\gamma}(\boldsymbol{x})$ can serve as a parametric bridge between $\|\cdot\|_{2,0}$ and $\|\cdot\|_{2,1}$. As a special case, the GME of $\|\cdot\|_1$ can serve as a parametric bridge between $\|\cdot\|_0$ and $\|\cdot\|_1$.

Proof. See Appendix A.2. □

Fig. 4.3 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 $\|\boldsymbol{x}\|_{2,1}$.

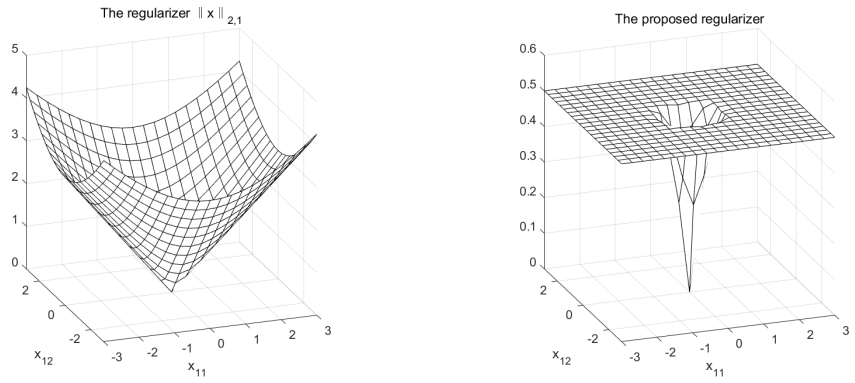


Fig. 4.3: Examples of of $\|\boldsymbol{x}\|_{2,1}$ and $(\|\cdot\|_{2,1})_{\boldsymbol{B}}(\boldsymbol{x})$ ($g = 1$, $n = 2$ and $\boldsymbol{B} = \boldsymbol{I}_2$).

4.2 Proposed Group Sparsity Aware Signal Recovery Model and Algorithm

By combining grouping operator $\mathscr{G}$ and GME penalty $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$, we propose a LiGME model for group sparsity aware estimation as follows,

$$\underset{\boldsymbol{x}\in\bar{B}(\boldsymbol{0},R)}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu(\|\cdot\|_{2,1})_{\boldsymbol{B}} \circ \boldsymbol{W} \circ \mathscr{G}(\boldsymbol{x}), \tag{4.12}$$

where $\boldsymbol{W} \in \mathbb{R}^{l\times n'}$ is a certain matirx and $\bar{B}(\boldsymbol{0},R) \subset \mathbb{R}^n$ is a closed ball centered at $\boldsymbol{0} \in \mathbb{R}^n$ with a sufficiently large radius $R > 0$. This constraint $\boldsymbol{x} \in \bar{B}(\boldsymbol{0},R)$ is newly introduced to ensure the existence of a minimizer of (4.12), see Example 3.1(b) in Section 3.1. The proposed model (4.12) admits any $\boldsymbol{W} \in \mathbb{R}^{l\times n'}$ and general grouping operator $\mathscr{G} \in \mathbb{R}^{n'\times n}$ in Definition 4.1, thus it can be used in the case of Example 4.1(c).

Remark 4.2. If $\boldsymbol{B} = \boldsymbol{O}_{n'\times n'}$, the regularizer in (4.12) reproduces $\|\cdot\|_{2,1} \circ \boldsymbol{W} \circ \mathscr{G}(\boldsymbol{x})$, which can be viewed as a generalization of $\|\boldsymbol{x}\|_{2,1}^{(\mathscr{G}_i)_{i=1}^t} = \|\cdot\|_{2,1} \circ \mathscr{G}(\boldsymbol{x})$ in Remark 4.1. The group weights in [72, 73] can be expressed in the form of a positive diagonal matrix $\boldsymbol{W}$.

If $\boldsymbol{B}$ is diagonal and $\mathscr{G} = \boldsymbol{W} = \boldsymbol{I}_n$, the regularizer $(\|\cdot\|_{2,1})_{\boldsymbol{B}} \circ \boldsymbol{W} \circ \mathscr{G}(\boldsymbol{x})$ in (4.12) reproduces the minimax concave penalty used in [92].

Remark 4.3. Model (4.12) is a generalization of some existing models as follows.

(a) For the special case in Example 4.1(a) with $\boldsymbol{W} = \boldsymbol{I}_n$, model (4.12) reproduces the GMC penalty model in [41].

(b) For the special $\mathscr{G}$ in Example 4.1(c), and $\boldsymbol{B} = \boldsymbol{O}_{n'\times n'}$, model (4.12) reproduces a convex model

$$\underset{\boldsymbol{x}\in\bar{B}(\boldsymbol{0},R)}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\mu\|\cdot\|_{2,1}\circ\boldsymbol{W}\circ\mathscr{G}(\boldsymbol{x}), \tag{4.13}$$

which is a slight generalization of the Group Lasso model. For simplicity, we call (4.13) *$\mathscr{G}$-Group Lasso* (or *g-Group Lasso* in Fig. 4.4(c) and Fig. 4.5(c)) in this study. More specifically, $\mathscr{G}$-Group Lasso in (4.13) reproduces the OGS model in [46] by letting $\boldsymbol{A}=\boldsymbol{I}_n$ and $\boldsymbol{W}=\boldsymbol{I}_n$.

For better understanding of the relation and differences between the proposed model (4.12), $\mathscr{G}$-Group Lasso (4.13) and some existing models, we summarize them in Table 4.1.

Table 4.1: Some existing models can be written as specializations of $\underset{\boldsymbol{x}\in S}{\text{minimize}}\ J_{\Psi_{\boldsymbol{B}}\circ\mathfrak{L}}(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\mu\Psi_{\boldsymbol{B}}\circ\mathfrak{L}(\boldsymbol{x})$.

Method	$\boldsymbol{A}$	Ψ	$\mathfrak{L}$	$\boldsymbol{B}$	S
Lasso [4]	general	$\|\cdot\|_1$	$\boldsymbol{I}_n$	$\boldsymbol{O}_n$	$\mathbb{R}^n$
Group Lasso [14]	general	$\|\cdot\|_{2,1}$	$\boldsymbol{I}_n$	$\boldsymbol{O}_n$	$\mathbb{R}^n$
OGS [15]	$\boldsymbol{I}_n$	$\|\cdot\|_{2,1}$	$\mathscr{G}$	$\boldsymbol{O}_{n'}$	$\mathbb{R}^n$
$\mathscr{G}$-Group Lasso (4.13)	general	$\|\cdot\|_{2,1}$	$\mathscr{G}$	$\boldsymbol{O}_{n'}$	$\bar{B}(\boldsymbol{0},R)$
Proposed model (4.12)	general	$\|\cdot\|_{2,1}$	$\mathscr{G}$	by Theorem 1	$\bar{B}(\boldsymbol{0},R)$

Clearly, $\|\cdot\|_{2,1}\in\Gamma_0(\mathbb{R}^l)$ is coercive and even symmetry. Moreover, it is well known that $\|\cdot\|_{2,1}$ is prox-friendly (see, e.g., [86, Example 24.20]) by

$$\begin{aligned}\mathrm{Prox}_{\gamma\|\cdot\|_{2,1}}:\ &\mathbb{R}^l\to\mathbb{R}^l\\ :\ &[\boldsymbol{x}_{\mathfrak{g}_1}^\top,\cdots,\boldsymbol{x}_{\mathfrak{g}_t}^\top]^\top\mapsto\left\{\left(1-\frac{\gamma}{\max\{\|\boldsymbol{x}_{\mathfrak{g}_i}\|_2,\gamma\}}\right)\boldsymbol{x}_{\mathfrak{g}_i}\right\}_{i=1}^t.\end{aligned} \tag{4.14}$$

In order to solve (4.12), we apply a simple but useful modification of the T_{LiGME} update in (2.12) in Proposition 2.2(b), essentially based on the idea found in [43, (16)]. We use in (4.17) in Algorithm 1 the metric projection onto $\bar{B}(\boldsymbol{0},R)$, i.e.,

$$\begin{aligned}P_{\bar{B}(\boldsymbol{0},R)}:\ &\mathbb{R}^n\to\bar{B}(\boldsymbol{0},R)\\ :\ &\boldsymbol{x}\mapsto\begin{cases}\boldsymbol{x}, & \boldsymbol{x}\in\bar{B}(\boldsymbol{0},R)\\ R\dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}, & \boldsymbol{x}\notin\bar{B}(\boldsymbol{0},R)\end{cases}\end{aligned} \tag{4.15}$$

Algorithm 1 The algorithm of LiGME model for group sparsity aware estimation

Input: $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\mathscr{G} \in \mathbb{R}^{n'\times n}$ and $\boldsymbol{W} \in \mathbb{R}^{l\times n'}$.

1. Initialization: $(\boldsymbol{x}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{w}^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l$.

Choose $\boldsymbol{B}$ satisfying $\boldsymbol{A}^\top\boldsymbol{A} - \mu\mathscr{G}^\top\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}\mathscr{G} \succeq \boldsymbol{O}_{n\times n}$ by Theorem 3.1.

Choose $(\sigma, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, +\infty)$ satisfying[2]

$$\begin{cases} \sigma\boldsymbol{I}_n - \dfrac{\kappa}{2}\boldsymbol{A}^\top\boldsymbol{A} - \mu\mathscr{G}^\top\boldsymbol{W}^\top\boldsymbol{W}\mathscr{G} \succ \boldsymbol{O}_{n\times n} \\ \tau \geqslant \left(\dfrac{\kappa}{2} + \dfrac{2}{\kappa}\right)\mu\|\boldsymbol{B}\|_{\mathrm{spec}}^2. \end{cases} \tag{4.16}$$

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

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= P_{\bar{B}(\boldsymbol{0},R)}\Big[\boldsymbol{x}^{(k)} - \frac{1}{\sigma}(\boldsymbol{A}^\top\boldsymbol{A} - \mu\mathscr{G}^\top\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}\mathscr{G})\boldsymbol{x}^{(k)} - \frac{\mu}{\sigma}\mathscr{G}^\top\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{v}^{(k)} \\ &\qquad - \frac{\mu}{\sigma}\mathscr{G}^\top\boldsymbol{W}^\top\boldsymbol{w}^{(k)} + \frac{1}{\sigma}\boldsymbol{A}^\top\boldsymbol{y}\Big], \\ \boldsymbol{v}^{(k+1)} &= \mathrm{Prox}_{\frac{\mu}{\tau}\|\cdot\|_{2,1}}\Big[\frac{2\mu}{\tau}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}\mathscr{G}\boldsymbol{x}^{(k+1)} - \frac{\mu}{\tau}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}\mathscr{G}\boldsymbol{x}^{(k)} + (\boldsymbol{I}_l - \frac{\mu}{\tau}\boldsymbol{B}^\top\boldsymbol{B})\boldsymbol{v}^{(k)}\Big], \\ \boldsymbol{w}^{(k+1)} &= (\mathrm{Id} - \mathrm{Prox}_{\|\cdot\|_{2,1}})\Big(2\boldsymbol{W}\mathscr{G}\boldsymbol{x}^{(k+1)} - \boldsymbol{W}\mathscr{G}\boldsymbol{x}^{(k)} + \boldsymbol{w}^{(k)}\Big) \end{aligned} \tag{4.17}$$

until the stopping criterion is fulfilled.

Output: $\tilde{\boldsymbol{x}} := \boldsymbol{x}^{(k+1)}$.

in place of the metric projection onto $\mathbb{R}_+^N$ in [43, (16)]. Algorithm 1 has a guarantee of convergence to a global minimizer of (4.12); see the discussion in [43, Section III-C].

As we clarified in Remark 4.3, the proposed model (4.12) is a generalization of the Group Lasso model [72]. Compared with proximal gradient method for Group Lasso model [93], Algorithm 1 requires at each update only one additional computation for $\mathrm{Prox}_{\gamma\|\cdot\|_{2,1}}$ in (4.14). If the proposed model (4.12) degenerates to Group Lasso, in each update (4.17) of Algorithm 1, there is no necessary to compute $\boldsymbol{v}^{(k+1)}$. Therefore, the computational cost is the same as the proximal gradient method for Group Lasso model.

[2]For example, any $\kappa > 1$, $\sigma = \|(\kappa/2)\boldsymbol{A}^\top\boldsymbol{A} + \mu\mathscr{G}^\top\boldsymbol{W}^\top\boldsymbol{W}\mathscr{G}\|_{\mathrm{spec}} + (\kappa - 1)$ and $\tau = (\kappa/2 + 2/\kappa)\mu\|\boldsymbol{B}\|_{\mathrm{spec}}^2 + (\kappa - 1)$ can satisfy (4.16).

4.3 Numerical Experiments

We consider group sparsity aware signal recovery where the true grouping structure of the target group sparse signal $\boldsymbol{x}^\star$ is unknown. In such case, inspired by [46], we tested the performance of the proposed LiGME model (4.12) with fully overlapping grouping operator (see $\mathscr{G}$ in Example 4.1(c) with $d = s-1$, $t = n-s+1$), where the GME matrix design in Theorem 3.1 is essential. The size of original signal $\boldsymbol{x}^\star$ is 100 ($n = 100$). By constructing a matrix with i.i.d. random entries obeying a zero-mean, unit-variance Gaussian distribution, the measurement matrix $\boldsymbol{A} \in \mathbb{R}^{64\times 100}$ is generated by normalizing it row by row ($m = 64$). The simulated noise $\boldsymbol{\varepsilon}$ is set to be independent white Gaussian noise with standard deviation σ_{noise}. The observation is generated by $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{\varepsilon} \in \mathbb{R}^{64}$.

We compared the performance of the proposed LiGME model (4.12) with the Lasso model (2.15) [19] and the $\mathscr{G}$-Group Lasso model (4.13). We set the regularization parameter to $\mu = 3\sigma_{\text{noise}}$ for Lasso, and $\mu = 0.8\sigma_{\text{noise}}$ for $\mathscr{G}$-Group Lasso and the LiGME model (4.12), which is a simple and effective procedure as well as maintains sparsity of the signal. The matrix $\boldsymbol{W}$ in both $\mathscr{G}$-Group Lasso and LiGME model are set to $\boldsymbol{I}_n$.

The Lasso model is implemented by iterative shrinkage thresholding algorithm (ISTA) [21] and $\mathscr{G}$-Group Lasso is implemented by Algorithm 1 with $\boldsymbol{B} = \boldsymbol{O}_{n'\times n'}$. We set $\kappa = 1.1$, $\sigma = \|(\kappa/2)\boldsymbol{A}^\top\boldsymbol{A} + \mu\boldsymbol{I}_n\|_{\text{spec}} + (\kappa-1)$, $\tau = (\kappa/2 + 2/\kappa)\mu\|\boldsymbol{B}\|_{\text{spec}}^2 + (\kappa-1)$. For the LiGME model, we design $\boldsymbol{B} = \boldsymbol{B}_\theta$ ($\theta = 0.9$) in Theorem 3.1, to achieve the overall convexity of the model. For the sufficiently large ball in (4.12) and (4.17), we set $R = 1000$. To obtain the final estimate $\tilde{\boldsymbol{x}} \in \mathbb{R}^{100}$, the initial estimate is set as $(\boldsymbol{x}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{w}^{(0)}) = (\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$, and the stopping criterion is set to either $\|(\boldsymbol{x}^{(k)}, \boldsymbol{v}^{(k)}, \boldsymbol{w}^{(k)}) - (\boldsymbol{x}^{(k+1)}, \boldsymbol{v}^{(k+1)}, \boldsymbol{w}^{(k+1)})\|_2 < 10^{-5}$ or $k = 10,000$.

In the following, we present two experiments for estimating target signals $\boldsymbol{x}^\star$ (in Fig.4.4(a) and Fig.4.5(a) respectively) of different group structures.

4.3.1 Estimation of $\boldsymbol{x}^\star$ of common group sizes

In the first experiment, the signal $\boldsymbol{x}^\star$ has a common group size 5 for each group (unknown information in the experiments). We set $\sigma_{\text{noise}} = 0.1$. The final estimate $\tilde{\boldsymbol{x}}$ by Lasso and $\mathscr{G}$-Group Lasso with fully overlapping setting in Example 4.1(c) with $s = 4$, $d = 3$ are illustrated in Fig. 4.4(b) and (c) respectively. Compared to Lasso in Fig. 4.4(b),

the $\mathscr{G}$-Group Lasso in Fig. 4.4(c) can recover small nonzero coefficient (see, e.g., the 45th entry), keep large coefficient better (see, e.g., the 83th entry) and achieves lower RMSE (root-mean-square error) value, where RMSE of $\tilde{\boldsymbol{x}}$ from $\boldsymbol{x}^\star$ is defined as

$$\text{RMSE} := \sqrt{\frac{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^\star\|^2}{n}}. \tag{4.18}$$

The performance of $\mathscr{G}$-Group Lasso shows the effectiveness of fully overlapping group setting for the case of general measurement matrix $\boldsymbol{A}$, which is consistent with the case of $\boldsymbol{A} = \boldsymbol{I}_n$ in [46]. Improving $\mathscr{G}$-Group Lasso further is challenging and we test whether the proposed LiGME model (4.12) can do so. Under the same setting of $\mathscr{G}$ ($s = 4$, $d = 3$), the result of (4.12) solved by Algorithm 1 is illustrated in Fig. 4.4(d). Compared to Lasso in Fig. 4.4(b) and $\mathscr{G}$-Group Lasso in Fig. 4.4(c), the LiGME model performs better and achieves the lowest RMSE value. For a clearer comparison, the entry-wise absolute values of estimation error $\tilde{\boldsymbol{x}} - \boldsymbol{x}^\star$ of the three methods are shown in Fig. 4.4(e). As can be observed, the error of LiGME model is smaller than the other two in general.

For $\mathscr{G}$-Group Lasso and the LiGME model (4.12), we also changed the group size s of grouping operator $\mathscr{G}$, and tested the performance of nonoverlapping setting ($d = 0$) and fully overlapping setting ($d = s - 1$). The results are list in Table 4.2. We can see that the LiGME model outperforms $\mathscr{G}$-Group Lasso in all cases. The lowest RMSE value is obtained by the LiGME model with $s = 5$, $d = 0$ (nonoverlapping). This is because 5 is the real group size of target signal $\boldsymbol{x}^\star$. The performance of nonoverlapping setting is very dependent on the real group structure. In addition to $s = 5$, overlapping setting performs better than nonoverlapping. The results demonstrate the effectiveness of LiGME model (4.12) with GME matrix in Theorem 3.1. We have found that the performance of overlapping setting degrades when $s > 5$, which is probably due to the decrease of group sparsity under the overlapping setting. Additionally, it has been well studied in [46] that partial overlapping, i.e., $0 < d < s - 1$, yields an inferior RMSE compared to both fully overlapping and nonoverlapping cases, as it leads to non-uniformly penalization of the entries.

4.3.2 Estimation of $\boldsymbol{x}^\star$ of different group sizes

In the second experiment, the sizes of different groups in original signal $\boldsymbol{x}^\star$ are different. Fig. 4.5 shows (a) $\boldsymbol{x}^\star$, (b)-(d) the final estimates $\tilde{\boldsymbol{x}}$ obtained for $\sigma_{\text{noise}} = 0.1$ by Lasso (2.15) with $t = n$, $n_i = 1$ for $i = 1, 2, \cdots, t$, $\mathscr{G}$-Group Lasso (4.13) and the LiGME model (4.12) respectively, and (e) entry-wise absolute values of estimation error $\tilde{\boldsymbol{x}} - \boldsymbol{x}^\star$, where

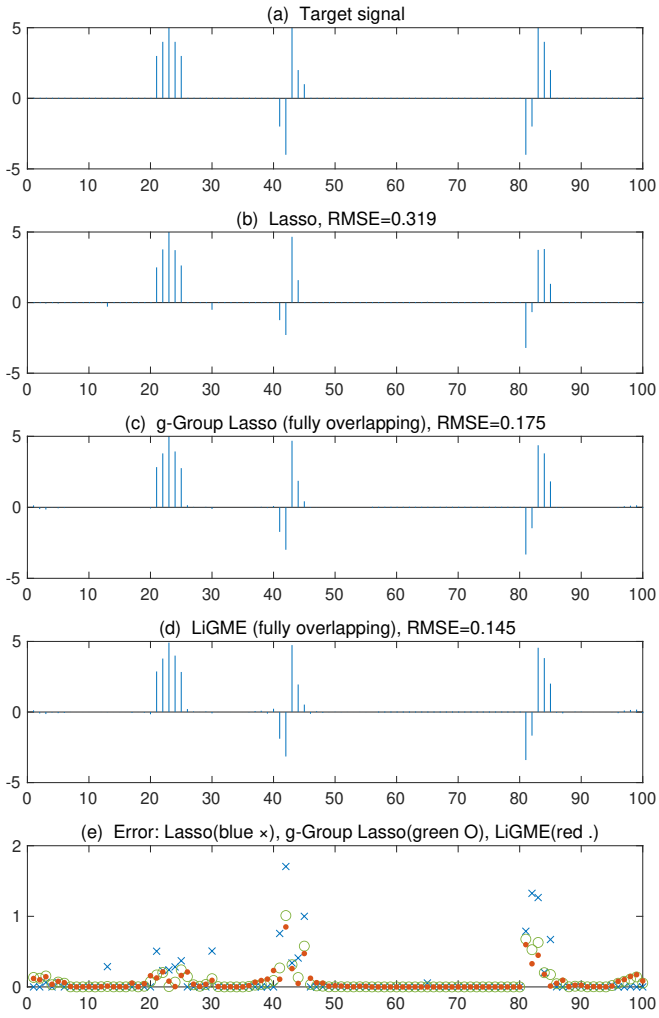


Fig. 4.4: Experiment 1: (a) target signal $\boldsymbol{x}^\star$, (b) final estimates $\tilde{\boldsymbol{x}}$ obtained by Lasso, (c) $\tilde{\boldsymbol{x}}$ obtained by $\mathscr{G}$-Group Lasso with fully overlapping setting ($s = 4$, $d = 3$), (c) $\tilde{\boldsymbol{x}}$ obtained by the LiGME model with fully overlapping setting ($s = 4$, $d = 3$), (e) entry-wise absolute values of estimation error $\tilde{\boldsymbol{x}} - \boldsymbol{x}^\star$.

we used fully overlapping grouping operator in Example 4.1(c) with $s = 4, d = 3$ for $\mathscr{G}$-Group Lasso and the LiGME model. Comparing Fig. 4.5(b) and Fig. 4.5(c), we can see that the $\mathscr{G}$-Group Lasso with fully overlapping setting recovers small nonzero coefficient between two large coefficients (see, e.g. the 59th entry) and keeps large coefficient better (see, e.g. the 84th entry). Although it is challenging to further improve the performance

Table 4.2: Experiment 1: RMSE of final estimate $\tilde{\boldsymbol{x}}$ by Lasso (i.e., $s = 1$, $d = 0$), $\mathscr{G}$-Group Lasso and the LiGME model with different grouping operator setting.

Group size s of grouping operator	Lasso and $\mathscr{G}$-Group Lasso		LiGME model	
	Nonoverlapping ($d = 0$)	Fully overlapping ($d = s - 1$)	Nonoverlapping ($d = 0$)	Fully overlapping ($d = s - 1$)
1	0.319	/	0.178	/
2	0.292	0.241	0.153	0.121
3[3]	0.260	0.186	0.177	0.123
4	0.241	0.175	0.149	0.145
5	0.124	0.216	0.086	0.175

of $\mathscr{G}$-Group Lasso, the proposed LiGME model achieves it. The result of the LiGME model in Fig. 4.5(d) performed better and achieved lower RMSE. For a more intuitive comparison, Fig. 4.4(e) illustrates the error between original signal and reconstructed results by the three methods. We can find that the error of the LiGME model is smaller than the other two in overall terms.

RMSE values with different grouping operator setting are summarized in Table 4.3. As visible, the fully overlapping settings offer an improved performance relative to nonoverlapping settings when s is smaller than 5. This is reasonable. Since the true division of groups is unknown, nonoverlapping may completely divide components of the same group into different groups while the fully overlapping operator fairly takes into account combinations of consecutive entries. The performance of fully overlapping setting decreases when s becomes larger, probably due to the reduced group sparsity under the fully overlapping setting. In all instances, the LiGME model outperforms $\mathscr{G}$-Group Lasso, since the GME penalty in the LiGME model is a nonconvex enhancement of $\|\cdot\|_{2,1}$ in $\mathscr{G}$-Group Lasso.

For the signal in Fig. 4.5(a), we compared the performance of Lasso, $\mathscr{G}$-Group Lasso with fully overlapping setting, and LiGME model with fully overlapping setting, at different noise levels. The results are summarized in Table 4.4, where the signal-to-noise ratio (SNR) values defined by

$$\text{SNR} := 10 \log_{10} \frac{\|\boldsymbol{x}^\star\|^2}{\|\boldsymbol{\varepsilon}\|^2} \qquad [\text{dB}] \tag{4.19}$$

of different noise levels are also listed. It can be seen that, at different noise levels, $\mathscr{G}$-

[3]When group size of grouping operator is set to $s = 3$ and there is no overlapping ($d = 0$), the grouping setting is: divide the first 99 entries into 33 groups in order and put the last entry into a separate group.

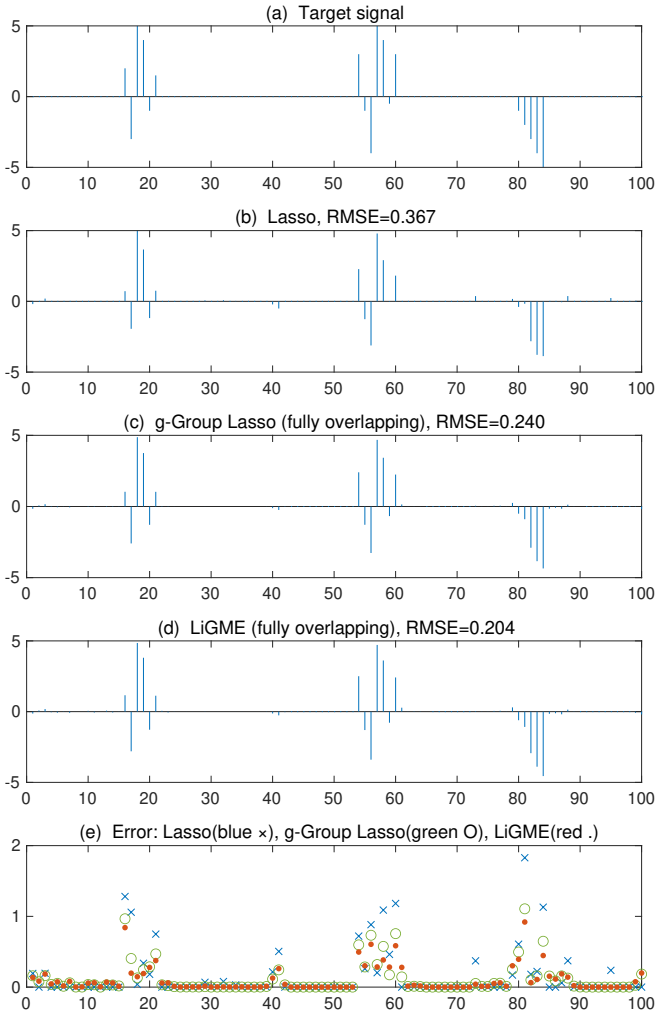


Fig. 4.5: Experiment 2: (a) target signal $\boldsymbol{x}^\star$, (b) final estimates $\tilde{\boldsymbol{x}}$ obtained by Lasso, (c) $\tilde{\boldsymbol{x}}$ obtained by $\mathscr{G}$-Group Lasso with fully overlapping setting ($s = 4$, $d = 3$), (c) $\tilde{\boldsymbol{x}}$ obtained by the LiGME model with fully overlapping setting ($s = 4$, $d = 3$), (e) entry-wise absolute values of estimation error $\tilde{\boldsymbol{x}} - \boldsymbol{x}^\star$.

Group Lasso with fully overlapping setting improves the performance of Lasso, and the LiGME model enhances the performance further. It indicates that the LiGME model based on overlapping group setting is robust to noise.

Table 4.3: Experiment 2: RMSE of final estimate $\tilde{\boldsymbol{x}}$ by Lasso (i.e., $s = 1$, $d = 0$), $\mathscr{G}$-Group Lasso and the LiGME model with different grouping operator setting.

Group size s of grouping operator	Lasso and $\mathscr{G}$-Group Lasso		LiGME model	
	Nonoverlapping ($d = 0$)	Fully overlapping ($d = s - 1$)	Nonoverlapping ($d = 0$)	Fully overlapping ($d = s - 1$)
1	0.367	/	0.336	/
2	0.356	0.282	0.289	0.201
3[4]	0.301	0.269	0.252	0.201
4	0.313	0.240	0.222	0.204
5	0.271	0.273	0.220	0.240

Table 4.4: Experiment 2: RMSE of final estimate $\tilde{\boldsymbol{x}}$ by Lasso, $\mathscr{G}$-Group Lasso with fully overlapping setting ($s = 4$, $d = 3$) and the LiGME model with fully overlapping setting ($s = 4$, $d = 3$) under different noise level.

Noise		PSNR		
σ_{noise}	SNR (dB)	Lasso	$\mathscr{G}$-Group Lasso	LiGME model
0.1	24.71	0.367	0.240	0.204
0.2	18.69	0.647	0.514	0.434
0.4	12.67	0.840	0.753	0.678

[4]When group size of grouping operator is set to $s = 3$ and there is no overlapping ($d = 0$), the grouping setting is: divide the first 99 entries into 33 groups in order and put the last entry into a separate group.

Chapter 5

Application to Group Sparse Classification

In this chapter, we apply the proposed LiGME model (4.12) to another application, that is, group sparse classification (GSC) [51–53]. We consider supervised classification, where the label information of the training data is known. More importantly, we consider unbalanced training set, which means that the number of training samples in different classes is different. We first analyze the bias of $\ell_{2,1}$ in unbalanced case and then consider to suppress the bias by using the proposed LiGME model. We verify the effectiveness of the proposed model to GSC by numerical experiments.

5.1 Bias of $\ell_{2,1}$ in Group Sparse Classification

We consider a supervised classification problem with t classes of subjects. Let $\boldsymbol{A}_i = [\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2}, \cdots, \boldsymbol{a}_{in_i}] \in \mathbb{R}^{m\times n_i}$ be the subset of the training samples from subject i, where $\boldsymbol{a}_{ij}$ is the j-th training sample from the i-th class and n_i is the number of training samples from class i. Then the dictionary matrix matrix $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_t] \in \mathbb{R}^{m\times n}$ is formulated by the training samples. $n = \sum_{i=1}^{t} n_i$ is the number of total training samples. Under the idea of GSC, the test sample $\boldsymbol{y}$ is approximated by the linear combination of the dictionary items $\boldsymbol{y} \approx \boldsymbol{A}\boldsymbol{x}$ and the coefficient vector $\boldsymbol{x}$ is group sparse with group structure $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_t^\top]^\top \in \mathbb{R}^n$, where $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in_i}]^\top \in \mathbb{R}^{n_i}$ $(i = 1, 2, \cdots, t)$.

In GSC methods, $\ell_{2,1}$-norm is the most frequently adopted regularizer to measure the group sparsity of $\boldsymbol{x}$. We firstly argue that its tendency to yield biased estimates for high-

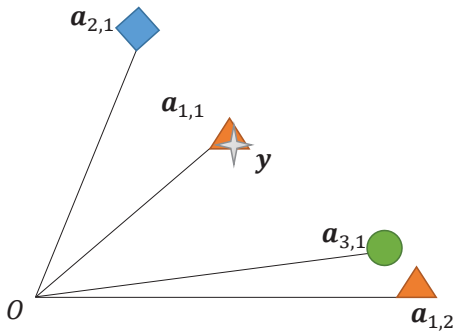


Fig. 5.1: A toy example of two different representations. A query image $\boldsymbol{y}$ (star) can be well represented by samples $\boldsymbol{a}_{1,1}$ and $\boldsymbol{a}_{1,2}$ of one class (triangle). It can also be well represented by samples $\boldsymbol{a}_{2,1}$ and $\boldsymbol{a}_{3,1}$ of another two classes (diamond and circle respectively).

Table 5.1: Penalty term values of representations in Fig. 5.1

Penalty	ℓ_0	ℓ_1	$\ell_{2,0}$	$\ell_{2,1}$	$\ell_{2,1/2}$	$(\|\cdot\|_{2,1})_{\boldsymbol{I}_2}$
First	2	0.99	**1**	0.951	**0.951**	**0.50**
Second	2	0.90	2	0.90	1.794	0.695

amplitude coefficients might lead to undesirable results. A toy example in Fig. 5.1 illustrates the potential risk of the $\ell_{2,1}$ regularizer, where four training samples (i.e., $\boldsymbol{a}_{1,1}$, $\boldsymbol{a}_{1,2}$, $\boldsymbol{a}_{2,1}$, $\boldsymbol{a}_{3,1}$) from three different classes (represented by triangle, diamond and circle respectively) and a query sample $\boldsymbol{y}$ (represented by star) are shown. The query sample can be well represented by the samples from triangle, i.e., $\boldsymbol{y} = 0.95\boldsymbol{a}_{1,1} + 0.04\boldsymbol{a}_{1,2}$. It can also be well represented by a combination of one sample from diamond and one sample from circle as $\boldsymbol{y} = 0.50\boldsymbol{a}_{2,1} + 0.40\boldsymbol{a}_{3,1}$.

Table 5.1 gives the value of different penalties (regularization parameter equals 1) by the above two representations. As we can see, ℓ_0 penalty cannot distinguish between the two representations in this case whereas $\ell_{2,0}$ penalty certainly chooses the first one, which demonstrates the advantage of GSC over SRC. However, $\ell_{2,1}$ penalty chooses the second one, the representation by samples from two classes instead of a single class, which is undesirable for GSC framework. This is due to that $\ell_{2,1}$ penalty refuses the large coefficient in the first representation, but accepts the two small coefficients in the second one.

Although $\ell_{2,0}$ and its approximation $\ell_{2,1/2}$ succeed in this example, they lead to non-convex optimization. Therefore, we can consider the proposed penalty in (4.12). For example, $(\|\cdot\|_{2,1})_{\frac{1}{\sqrt{\gamma}}\boldsymbol{I}_2}$ will prefer the first representation in Fig. 5.1 as long as $\gamma \leqslant 4.85$. This also reflects its role as a bridge between $\ell_{2,1}$ and $\ell_{2,0}$.

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. The influence caused by bias of $\ell_{2,1}$ penalty would be amplified and the bias makes $\ell_{2,1}$ unfair for classes of different sizes.

Example 5.1. 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 $\tilde{\boldsymbol{x}}_i = [\eta\boldsymbol{x}_i^\top, (1-\eta)\boldsymbol{x}_i^\top]^\top \in \mathbb{R}^{2n_i}$ $(0 \leqslant \eta \leqslant 1)$ and $\|\tilde{\boldsymbol{x}}_i\|_1 = 1$. However, $\|\boldsymbol{x}_i\|_2^2 - \|\tilde{\boldsymbol{x}}_i\|_2^2 = 2\eta(1-\eta)\|\boldsymbol{x}_i\|_2^2 \geqslant 0$. 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 5.1 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, when training set is unbalanced, $\ell_{2,1}$ 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. The bias can easily cause misclassification, especially when the correct class has relatively few samples. Note that $\ell_{2,0}$ regularizer is independent of group size and it does not have such unfairness.

Since $\ell_{2,1}$ regularizer in GSC is unfair for classes of different sizes 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 the GME penalty $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$ in (4.10) to GSC framework, which can bridge the gap between $\ell_{2,1}$ and $\ell_{2,0}$.

5.2 Proposed Group Sparse Classification Model and Algorithm

We expect to improve the performance of Group Lasso on unbalanced training sets by replacing the $\ell_{2,1}$ penalty with $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$. Since the label information of the training data is known, the group structure is known, and there is no overlapping or incomplete cover. As a specialization of (4.12) with $\mathscr{G} = \boldsymbol{I}_n$, we use the model as follows

$$\underset{\boldsymbol{x}\in\bar{B}(\boldsymbol{0},R)}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu(\|\cdot\|_{2,1})_{\boldsymbol{B}} \circ \boldsymbol{W}(\boldsymbol{x}) \tag{5.1}$$

for group sparse classification problems.

Following GSC, we can set $\boldsymbol{W} = \boldsymbol{I}_n$ in (5.1). We can also follow WGSC [67] to set $\boldsymbol{W}$ as in (2.17). Since $\boldsymbol{W}$ is invertible in general, we can use a specialization of the proposed GME matrix design in Theorem 3.1 as follows to ensure the overall convexity of (5.1).

Corollary 5.1. In Definition 2.13, let $(\mathcal{X},\mathcal{Y},\mathcal{Z}) = (\mathbb{R}^n,\mathbb{R}^m,\mathbb{R}^n)$, $(\boldsymbol{A},\mathfrak{Q},\mu) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}_{++}$. Let $\mathfrak{Q} = \boldsymbol{W}$, where $\boldsymbol{W}$ is invertible. Then

$$\boldsymbol{B}_\theta := \sqrt{\theta/\mu}\boldsymbol{A}\boldsymbol{W}^{-1} \in \mathbb{R}^{m\times n},\ \theta\in[0,1], \tag{5.2}$$

ensures the overall convexity of (5.1).

Proposed algorithm for group sparse classification is summarized in Algorithm 2. The sequence $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ generated by Algorithm 2 is guaranteed to converge to a globally optimal solution of problem (5.1).

5.3 Numerical Experiments

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

Algorithm 2 The proposed group sparsity enhanced classification algorithm

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

1. Initialization: Let $(\boldsymbol{x}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{w}^{(0)}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$;

Set $\boldsymbol{W} = \boldsymbol{I}_n$ or set $\boldsymbol{W}$ according to (2.17).

Choose $\boldsymbol{B}$ satisfying $\boldsymbol{A}^\top\boldsymbol{A} - \mu\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W} \succeq \boldsymbol{O}_{n\times n}$ by Corollary 5.1.

Choose $(\sigma, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, +\infty)$ satisfying

$$\begin{cases} \sigma\boldsymbol{I}_n - \dfrac{\kappa}{2}\boldsymbol{A}^\top\boldsymbol{A} - \mu\boldsymbol{W}^\top\boldsymbol{W} \succ \boldsymbol{O}_{n\times n} \\ \tau \geqslant \left(\dfrac{\kappa}{2} + \dfrac{2}{\kappa}\right)\mu\|\boldsymbol{B}\|_{\text{spec}}^2. \end{cases} \tag{5.3}$$

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

$$\begin{aligned} \boldsymbol{x}^{(k+1)} =& P_{\bar{B}(\boldsymbol{0},R)}\Big[\boldsymbol{x}^{(k)} - \frac{1}{\sigma}(\boldsymbol{A}^\top\boldsymbol{A} - \mu\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W})\boldsymbol{x}^{(k)} - \frac{\mu}{\sigma}\boldsymbol{W}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{v}^{(k)} \\ & - \frac{\mu}{\sigma}\boldsymbol{W}^\top\boldsymbol{w}^{(k)} + \frac{1}{\sigma}\boldsymbol{A}^\top\boldsymbol{y}\Big] \\ \boldsymbol{v}^{(k+1)} =& \text{Prox}_{\frac{\mu}{\tau}\|\cdot\|_{2,1}}\Big[\frac{2\mu}{\tau}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}\boldsymbol{x}^{(k+1)} - \frac{\mu}{\tau}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}\boldsymbol{x}^{(k)} + (\boldsymbol{I}_n - \frac{\mu}{\tau}\boldsymbol{B}^\top\boldsymbol{B})\boldsymbol{v}^{(k)}\Big], \\ \boldsymbol{w}^{(k+1)} =& \left(\text{Id} - \text{Prox}_{\|\cdot\|_{2,1}}\right)\left(2\boldsymbol{W}\boldsymbol{x}^{(k+1)} - \boldsymbol{W}\boldsymbol{x}^{(k)} + \boldsymbol{w}^{(k)}\right), \end{aligned}$$

until the stopping criterion is fulfilled and let $\hat{\boldsymbol{x}} := \boldsymbol{x}^{(k+1)}$.

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

$$i^\star = \arg\min_i \|\boldsymbol{y} - \boldsymbol{A}_i\hat{\boldsymbol{x}}_i\|_2.$$

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

training set is unbalanced), and verify the performance improvement by using $(\|\cdot\|_{2,1})_{\boldsymbol{B}}$ in (4.10) as the regularizer by conducting the experiments on a relatively simple dataset. The USPS handwritten digit database [94] has 11,000 samples of digits "0" through "9" (1,100 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 compared the performance of the proposed model (5.1) with Group Lasso model (2.14) [53]. We set $\boldsymbol{B} = \sqrt{\theta/\mu}\boldsymbol{A}$ and fix $\theta = 0.9$ to achieve the overall convexity of proposed model, and set $\kappa = 1.1$, $\iota = \|(\kappa/2)\boldsymbol{A}^\top\boldsymbol{A} + \mu\boldsymbol{I}_n\|_{\text{spec}} + (\kappa - 1)$, $\tau = (\kappa/2 + 2/\kappa)\mu\|\boldsymbol{B}\|_{\text{spec}}^2 + (\kappa - 1)$. The initial estimate is set as $(\boldsymbol{x}^{(0)}, \boldsymbol{u}^{(0)}, \boldsymbol{w}^{(0)}) = (\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$, 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. For the sufficiently large ball in (5.1), we set $R = 1000$.

Figure 5.2 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 Group Lasso model (2.14) while classified correctly by the proposed model (5.1). The obtained coefficient vectors by Group Lasso and the proposed LiGME model (both with $\mu = 4$) are illustrated respectively, and some samples corresponding to nonzero coefficients are also displayed in Figure 5.2. It can be seen that the samples from digit "6" made the greatest contribution to the representation in Group Lasso, and samples from "5" and "0" also made small contribution. In the figure of the proposed LiGME model, samples from the correct class "0" made the biggest contribution and led to correct result. It is reasonable, because the proposed model 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 lead to misclassification.

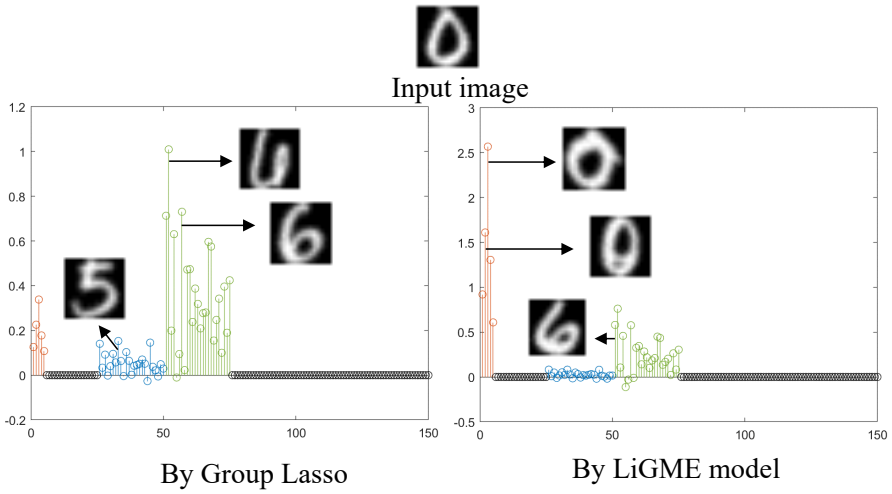


Fig. 5.2: Estimated sparse coefficients $\hat{\boldsymbol{x}}$ by Group Lasso and the proposed model respectively.

Table 5.2 summarizes the recognition accuracy of Group Lasso and the proposed model (5.1) 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 Group Lasso with $\mu = \mu_{\text{gl}} = 1.5$ and the proposed model with $\mu = \mu_{\text{prop}} = 3$ perform well on this dataset. We also experimented the proposed model of using μ_{gl}, which did not degrade too much compared with using μ_{prop}. We see that the Group Lasso model degrades when the training set is unbalanced, and the proposed LiGME model outperforms Group Lasso especially in such case.

Table 5.2: 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
Group Lasso ($\mu = 1.5$)		81.4%	86.6%	73.6%	91.4%	88.4%	93.2%
LiGME ($\mu = 1.5$)		82.0%	87.2%	79.0%	92.2%	89.4%	93.0%
LiGME ($\mu = 3$)		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 (2.17). The ORL Database of Faces [95] 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 [96], 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 performance of the proposed LiGME model (5.1) ($\boldsymbol{W} = \boldsymbol{I}_n$ and $\boldsymbol{W}$ by (2.17) respectively) with both Group Lasso model (2.14) [53] and weighted Group Lasso model (2.16) [67]. In order to achieve the overall convexity, we set $\boldsymbol{B} = \boldsymbol{B}_\theta$ with $\theta = 0.9$ by Corollary 5.1 for the proposed model (5.1). Settings of (ι, τ, κ), initial estimate, stopping criterion and R are the same as those in the previous experiment. When the parameter μ 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 $\mu = 0.05$ for $\ell_{2,1}$ regularizer based methods (i.e. Group Lasso and weighted Group Lasso), $\mu = 0.2$ for the proposed LiGME model and

$\sigma_1 \in [2,4], \sigma_2 \in [0.5,2]$ for weights involved methods (i.e. weighted Group Lasso and proposed LiGME model with $\boldsymbol{W}$ by (2.17)) work well on this dataset.

Figure 5.3 shows a classification result of weighted Group Lasso and the proposed LiGME model ($\boldsymbol{W}$ by (2.17) 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 weighted Group Lasso while classified correctly by the proposed LiGME model (5.1).

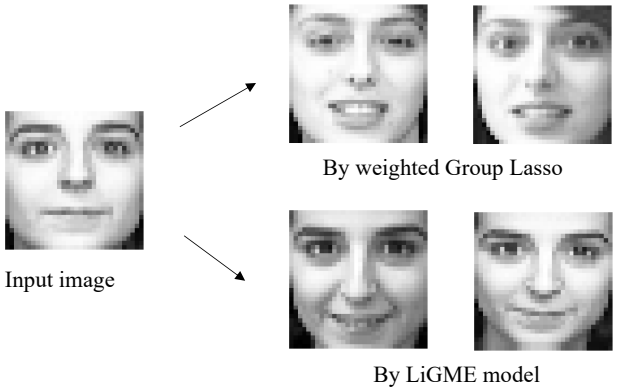


Fig. 5.3: An example of results by weighted Group Lasso and the proposed LiGME model.

Table 5.3: 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
Group Lasso		86.3%	85.0%	91.3%	85.0%	92.5%	93.8%
LiGME ($\boldsymbol{W} = \boldsymbol{I}_n$)		88.8%	86.3%	93.8%	86.3%	93.8%	95.0%
Weighted Group Lasso		90.6%	87.5%	95.0%	88.8%	93.8%	96.3%
LiGME ($\boldsymbol{W}$ by (2.17))		91.3%	89.4%	95.6%	91.9%	94.4%	96.3%

Table 5.3 summarizes the recognition accuracy of Group Lasso, the proposed LiGME model with $\boldsymbol{W} = \boldsymbol{I}_n$, weighted Group Lasso and the proposed model with $\boldsymbol{W}$ computed by (2.17). 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 (2.17), weighted Group Lasso achieves a significant improvement over Group Lasso. By using the proposed LiGME model (5.1) with $\boldsymbol{W}$ computed by (2.17), the performance can be further improved, especially when the training set is unbalanced. The proposed LiGME model and Algorithm 2 for GSC application is effective.

Chapter 6

Conclusion

In this study, we have studied the designs and applications of generalized Moreau enhancement (GME) matrix for sparsity aware linearly involved generalized-Moreau-enhanced (LiGME) models.

In Chapter 2, we have introduced the basic definitions in convex optimization and fixed point theory of nonexpansive operators. We have reviewed the LiGME model as well as the group sparsity aware estimation and classification.

In Chapter 3, we have studied some certain fundamental properties of the LiGME model and included a discussion on the existence of a minimizer of the LiGME model. We have proposed a unified algebraic design of the GME matrix, which is a key to guarantee the overall convexity of the LiGME model. The proposed design is a generalization of the existing algebraic designs. It can be used for general linear operators to be involved in the LiGME regularizer. Moreover, it does not require any eigendecomposition or iteration. Therefore, the applicability of the LiGME model is enhanced by using the proposed design. We have also provided a design method of GME matrix, for the case that the singular value decomposition of the linear operator is available.

In Chapter 4, by using the proposed GME matrix design, we have presented an application to group sparsity aware signal recovery. We have proposed a group sparsity aware LiGME model based on the grouping operator, and the model can be applied to general group configurations which allows overlapping and incomplete cover. The model is applied to group sparsity aware signal recovery of unknown group structure, where fully overlapping setting is chosen for the grouping operator. Numerical experiments have demonstrated the effectiveness of the proposed GME matrix in the LiGME model.

In Chapter 5, we have applied the proposed LiGME model to group sparse classification. We have analyzed the bias of $\ell_{2,1}$ norm in Group Lasso model in this application. Then, we have used the proposed LiGME model to group sparse classification of unbalanced training set. Numerical experiments demonstrate the effectiveness of the proposed LiGME model.

Finally, we would like to remark humbly that we are ready to encourage the data science community to apply the LiGME models to broader areas of research relying essentially on the sparsity promoting convex regularizers.

Acknowledgements

First and foremost I would like to express extremely grateful to my supervisor, Professor Isao Yamada, for his invaluable advice, continuous support, and patience. During the past years, I have learned a lot from him, about how to find a problem, how to do a research and how to solve problems. He has given me many instructions on how to write a paper and deliver a presentation well. In addition to the technical knowledge and skills, I also learned philosophy of life from him. When I encountered failure in research and fall into depression in life, he would give me great encouragement to boost my confidence and have a more positive outlook on life. He is so nice, so patient, and so supportive to me. Without his great support and continuous encouragement, I could not complete my study.

I am also grateful to Assistant Professor Masao Yamagishi for his constructive comments on my research, writing and presentation preparation. His plentiful knowledge and generously provided support have encouraged me in all the time of my academic research and daily life.

I would like to thank the thesis review committee of this dissertation, Professor Isao Yamada, Professor Tomohiko Uyematsu, Professor Kazuhiko Fukawa, Professor Konstantinos Slavakis and Associate Professor Yutaka Jitsumatsu, not only for their time and extreme patience, but also for their constructive comments helping me to improve my research.

I would like to thank the secretary, Ms. Hamasaki, for her help with various procedures and for her warm encouragement and support. I am also grateful to the members of the Yamada Laboratory, who have deepen my knowledge not only in research but also in various other fields through daily discussions and chats. It is their kind help and support that have made my study and life in Japan a wonderful time.

I would like to express my sincere thanks to the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan for giving me such a precious opportunity to study in Japan and financing my life.

How time flies. Although most of my time of studying abroad is under COVID-19, I have an unforgettable and cherished experience in Japan. I would like to thank all the people who gave me warmth in my daily life. Finally, I would like to express my gratitude to my parents. Thank them for their continuous understanding and support in the past years.

Appendix

A.1 Proof of Proposition 3.3

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

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

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

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

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

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

A.2 Proof of Proposition 4.1

The regularization term can be written as

$$\begin{aligned}\frac{2}{\gamma}\left(\|\cdot\|_{2,1}\right)_{\boldsymbol{B}_\gamma}(\boldsymbol{x}) : \mathbb{R}^n &\to \mathbb{R} \\ : [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_g^\top]^\top &\mapsto \sum_{i=1}^{g} \frac{2}{\gamma}\varphi_i(\boldsymbol{x}_i),\end{aligned}$$

where

$$\varphi_i(\boldsymbol{x}_i) := \|\boldsymbol{x}_i\|_2 - \min_{\boldsymbol{v}_i \in \mathbb{R}^{n_i}} \left\{ \|\boldsymbol{v}_i\|_2 + \frac{1^2}{2\gamma}\|\boldsymbol{x}_i - \boldsymbol{v}_i\|_2^2 \right\}$$

for $i = 1, \cdots, g$. By [86, Example 24.20], we obtain

$$\frac{2}{\gamma}\varphi_i(\boldsymbol{x}_i) = \begin{cases} \dfrac{2}{\gamma}\|\boldsymbol{x}_i\|_2 - \dfrac{1^2}{\gamma^2}\|\boldsymbol{x}_i\|_2^2, & \text{if } \|\boldsymbol{x}_i\|_2 \leqslant \dfrac{\gamma}{1} \\ 1, & \text{otherwise.} \end{cases} \tag{6.1}$$

Then, we get

$$\lim_{\gamma \downarrow 0} \frac{2}{\gamma}\varphi_i(\boldsymbol{x}_i) = \begin{cases} 0, & \text{if } \|\boldsymbol{x}_i\|_2 = 0, \\ 1, & \text{otherwise,} \end{cases} \tag{6.2}$$

and

$$\lim_{\gamma \downarrow 0} \frac{2}{\gamma}\left(\|\cdot\|_{2,1}\right)_{\boldsymbol{B}_\gamma} = \lim_{\gamma \downarrow 0} \sum_{i=1}^{g} \frac{2}{\gamma}\varphi_i(\boldsymbol{x}_i) = \|\boldsymbol{x}\|_{2,0}. \tag{6.3}$$

References

[1] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Vol. 2. Springer, 2009.

[2] Sergios Theodoridis. *Machine Learning: a Bayesian and Optimization Perspective*. Academic press, 2015.

[3] Adi Ben-Israel and Thomas NE Greville. *Generalized Inverses: Theory and Applications*, Vol. 15. Springer Science & Business Media, 2003.

[4] Mario Bertero, Patrizia Boccacci, and Christine De Mol. *Introduction to Inverse Problems in Imaging*. CRC press, 2021.

[5] Jiro Abe, Masao Yamagishi, and Isao Yamada. Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition. *Inverse Problems*, Vol. 36, No. 3, p. 035012, 2020.

[6] Wataru Yata, Masao Yamagishi, and Isao Yamada. A convexly constrained LiGME model and its proximal splitting algorithm. *Journal of Applied and Numerical Optimization*, Vol. 4, No. 2, pp. 245–271, 2022.

[7] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, Vol. 59, No. 8, pp. 1207–1223, 2006.

[8] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, Vol. 52, No. 4, pp. 1289–1306, 2006.

[9] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, Vol. 98, No. 6, pp. 1031–1044, 2010.

[10] Richard G Baraniuk, Emmanuel Candes, Michael Elad, and Yi Ma. Applications of sparse representation and compressive sensing. *Proceedings of the IEEE*, Vol. 98, No. 6, pp. 906–909, 2010.

[11] Zheng Zhang, Yong Xu, Jian Yang, Xuelong Li, and David Zhang. A survey of sparse representation: algorithms and applications. *IEEE Access*, Vol. 3, pp. 490–530, 2015.

[12] Qiang Zhang, Yi Liu, Rick S Blum, Jungong Han, and Dacheng Tao. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. *Information Fusion*, Vol. 40, pp. 57–75, 2018.

[13] Zhijin Qin, Jiancun Fan, Yuanwei Liu, Yue Gao, and Geoffrey Ye Li. Sparse representation for wireless communications: a compressive sensing approach. *IEEE Signal Processing Magazine*, Vol. 35, No. 3, pp. 40–58, 2018.

[14] Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Vol. 2. Springer, 2010.

[15] Rafael E Carrillo, Kenneth E Barner, and Tuncer C Aysal. Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 2, pp. 392–408, 2010.

[16] Junfeng Yang and Yin Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM Journal on Scientific Computing*, Vol. 33, No. 1, pp. 250–278, 2011.

[17] Yunhai Xiao, Hong Zhu, and Soon-Yi Wu. Primal and dual alternating direction algorithms for ℓ_1-ℓ_1-norm minimization problems in compressive sensing. *Computational Optimization and Applications*, Vol. 54, No. 2, pp. 441–459, 2013.

[18] Juan Marcos Ramirez and Jose Luis Paredes. Robust transforms based on the weighted median operator. *IEEE Signal Processing Letters*, Vol. 22, No. 1, pp. 120–124, 2015.

[19] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288, 1996.

[20] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, Vol. 60, No. 1-4, pp. 259–268, 1992.

[21] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, Vol. 57, No. 11, pp. 1413–1457, 2004.

[22] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis*. Cambridge university press, 2015.

[23] Simon Foucart and Ming-Jun Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leqslant 1$. *Applied and Computational Harmonic Analysis*, Vol. 26, No. 3, pp. 395–407, 2009.

[24] Jeevan K Pant, Wu-Sheng Lu, and Andreas Antoniou. New improved algorithms for compressive sensing based on ℓ_p norm. *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 61, No. 3, pp. 198–202, 2014.

[25] Hosein Mohimani, Massoud Babaie-Zadeh, and Christian Jutten. A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm. *IEEE Transactions on Signal Processing*, Vol. 57, No. 1, pp. 289–301, 2008.

[26] Ming-Jun Lai, Yangyang Xu, and Wotao Yin. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM Journal on Numerical Analysis*, Vol. 51, No. 2, pp. 927–957, 2013.

[27] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, Vol. 96, No. 456, pp. 1348–1360, 2001.

[28] Andrew Blake and Andrew Zisserman. *Visual Reconstruction*. MIT press, 1987.

[29] Mila Nikolova. Estimation of binary images by minimizing convex criteria. In *Proceedings 1998 International Conference on Image Processing*, Vol. 2, pp. 108–112. IEEE, 1998.

[30] Mila Nikolova. Markovian reconstruction using a gnc approach. *IEEE Transactions on Image Processing*, Vol. 8, No. 9, pp. 1204–1220, 1999.

[31] Mila Nikolova. Energy minimization methods, 2011.

[32] Yin Ding and Ivan W Selesnick. Artifact-free wavelet denoising: non-convex sparse regularization, convex optimization. *IEEE Signal Processing Letters*, Vol. 22, No. 9, pp. 1364–1368, 2015.

[33] Thomas Möllenhoff, Evgeny Strekalovskiy, Michael Moeller, and Daniel Cremers. The primal-dual hybrid gradient method for semiconvex splittings. *SIAM Journal on Imaging Sciences*, Vol. 8, No. 2, pp. 827–857, 2015.

[34] Ilker Bayram. On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Transactions on Signal Processing*, Vol. 64, No. 6, pp. 1597–1608, 2015.

[35] Emmanuel Soubies, Laure Blanc-Féraud, and Gilles Aubert. A continous exact l0 penalty (cel0) for least squares regularized problem. *SIAM Journal on Imaging Sciences*, Vol. 8, pp. 1574–1606, 2015.

[36] Marcus Carlsson. On convexification/optimization of functionals including an l2-misfit term. *arXiv preprint arXiv:1609.09378*, 2016.

[37] Mohammadreza Malek-Mohammadi, Cristian R Rojas, and Bo Wahlberg. A class of nonconvex penalties preserving overall convexity in optimization-based mean filtering. *IEEE Transactions on Signal Processing*, Vol. 64, No. 24, pp. 6650–6664, 2016.

[38] Alessandro Lanza, Serena Morigi, and Fiorella Sgallari. Convex image denoising via non-convex regularization with parameter selection. *Journal of Mathematical Imaging and Vision*, Vol. 56, No. 2, pp. 195–220, 2016.

[39] Alessandro Lanza, Serena Morigi, Ivan Selesnick, and Fiorella Sgallari. Nonconvex nonsmooth optimization via convex–nonconvex majorization–minimization. *Numerische Mathematik*, Vol. 136, No. 2, pp. 343–381, 2017.

[40] Ivan Selesnick and Masoud Farshchian. Sparse signal approximation via nonseparable regularization. *IEEE Transactions on Signal Processing*, Vol. 65, No. 10, pp. 2561–2575, 2017.

[41] Ivan Selesnick. Sparse regularization via convex analysis. *IEEE Transactions on Signal Processing*, Vol. 65, No. 17, pp. 4481–4494, 2017.

[42] Jiro Abe, Masao Yamagishi, and Isao Yamada. Convexity-edge-preserving signal recovery with linearly involved generalized minimax concave penalty function. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4918–4922. IEEE, 2019.

[43] Daichi Kitahara, Rikako Kato, Hiroki Kuroda, and Akira Hirabayashi. Multi-contrast CSMRI using common edge structures with LiGME model. In *European Signal Processing Conference (EUSIPCO)*, pp. 2119–2123. IEEE, 2021.

[44] Xiaoqian Liu and Eric C Chi. Revisiting convexity-preserving signal recovery with the linearly involved gmc penalty. *Pattern Recognition Letters*, Vol. 156, pp. 60–66, 2022.

[45] Charles Byrne. Iterative oblique projection onto convex sets and the split feasibility problem. *Inverse Problems*, Vol. 18, No. 2, p. 441, 2002.

[46] Po-Yu Chen and Ivan W Selesnick. Translation-invariant shrinkage/thresholding of group sparse signals. *Signal Processing*, Vol. 94, pp. 476–489, 2014.

[47] Wei Deng, Wotao Yin, and Yin Zhang. Group sparse optimization by alternating direction method. In *Wavelets and Sparsity XV*, Vol. 8858, p. 88580R. International Society for Optics and Photonics, 2013.

[48] Yasuhiko Ikebe, Yoshiko Ikebe, Nobuyoshi Asai, and Yoshinori Miyazaki. *Modern Linear Algebra Through Decomposition Theorems*. Kyoritsu-Shuppan, (in Japanese), 2009.

[49] Gene H Golub and Charles F Van Loan. *Matrix Computations*. JHU press, 2013.

[50] Robert van de Geijn and Margaret Myers. *Advanced Linear Algebra: Foundations to Frontiers*. Creative Commons NonCommercial (CC BY-NC), 2022.

[51] Angshul Majumdar and Rabab K Ward. Classification via group sparsity promoting regularization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 861–864. IEEE, 2009.

[52] Ehsan Elhamifar and René Vidal. Robust classification using structured sparse representation. In *CVPR 2011*, pp. 1873–1879. IEEE, 2011.

[53] Jin Huang, Feiping Nie, Heng Huang, and Chris Ding. Supervised and projected sparse coding for image classification. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[54] Erwin Kreyszig. *Introductory Functional Analysis with Applications*, Vol. 17. John Wiley & Sons, 1991.

[55] Isao Yamada. *Kougaku no Tameno Kansu Kaiseki (Functional Analysis for Engineering)*. Suurikougaku-Sha/Saiensu-Sha, 2009.

[56] CW Groetsch. A note on segmenting mann iterates. *Journal of Mathematical Analysis and Applications*, Vol. 40, No. 2, pp. 369–372, 1972.

[57] Ivan W Selesnick and Ilker Bayram. Enhanced sparsity by non-separable regularization. *IEEE Transactions on Signal Processing*, Vol. 64, No. 9, pp. 2298–2313, 2016.

[58] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, Vol. 38, No. 2, pp. 894–942, 2010.

[59] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, Vol. 158, No. 2, pp. 460–479, 2013.

[60] Shuangge Ma, Xiao Song, and Jian Huang. Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*, Vol. 8, No. 1, pp. 1–17, 2007.

[61] Lifeng Wang, Guang Chen, and Hongzhe Li. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, Vol. 23, No. 12, pp. 1486–1494, 2007.

[62] Jian Zou and Yuli Fu. Split Bregman algorithms for sparse group lasso with application to MRI reconstruction. *Multidimensional Systems and Signal Processing*, Vol. 26, No. 3, pp. 787–802, 2015.

[63] Xinyu Wang, Yanfei Zhong, Liangpei Zhang, and Yanyan Xu. Spatial group sparsity regularized nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55, No. 11, pp. 6287–6304, 2017.

[64] Lucas Drumetz, Travis R Meyer, Jocelyn Chanussot, Andrea L Bertozzi, and Christian Jutten. Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms. *IEEE Transactions on Image Processing*, Vol. 28, No. 7, pp. 3435–3450, 2019.

[65] Jie Huang, Ting-Zhu Huang, Xi-Le Zhao, and Liang-Jian Deng. Nonlocal tensor-based sparse hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[66] Baijie Qiao, Zhu Mao, Jinxin Liu, Zhibin Zhao, and Xuefeng Chen. Group sparse regularization for impact force identification in time domain. *Journal of Sound and Vibration*, Vol. 445, pp. 44–63, 2019.

[67] Xin Tang, Guocan Feng, and Jiaxin Cai. Weighted group sparse representation for undersampled face recognition. *Neurocomputing*, Vol. 145, pp. 402–415, 2014.

[68] Nikhil Rao, Robert Nowak, Christopher Cox, and Timothy Rogers. Classification with the sparse group lasso. *IEEE Transactions on Signal Processing*, Vol. 64, No. 2, pp. 448–463, 2015.

[69] Shoubiao Tan, Xi Sun, Wentao Chan, Lei Qu, and Ling Shao. Robust face recognition with kernelized locality-sensitive group sparsity representation. *IEEE Transactions on Image Processing*, Vol. 26, No. 10, pp. 4661–4668, 2017.

[70] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, Vol. 24, No. 2, pp. 227–234, 1995.

[71] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 1*, pp. 1457–1465, 2012.

[72] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 68, No. 1, pp. 49–67, 2006.

[73] Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical Science*, Vol. 27, No. 4, 2012.

[74] Patrick Breheny and Jian Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, Vol. 25, No. 2, pp. 173–187, 2015.

[75] Yaohua Hu, Chong Li, Kaiwen Meng, Jing Qin, and Xiaoqi Yang. Group sparse optimization via $\ell_{p,q}$ regularization. *The Journal of Machine Learning Research*, Vol. 18, No. 1, pp. 960–1011, 2017.

[76] Lanfan Jiang and Wenxing Zhu. Iterative weighted group thresholding method for group sparse recovery. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, No. 1, pp. 63–76, 2020.

[77] Yuling Jiao, Bangti Jin, and Xiliang Lu. Group sparse recovery via the $\ell^0(\ell^2)$ penalty: Theory and algorithm. *IEEE Transactions on Signal Processing*, Vol. 65, No. 4, pp. 998–1012, 2016.

[78] Po-Yu Chen and Ivan W Selesnick. Group-sparse signal denoising: non-convex regularization, convex optimization. *IEEE Transactions on Signal Processing*, Vol. 62, No. 13, pp. 3464–3478, 2014.

[79] Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

[80] Patrick L Combettes and Christian L Müller. Perspective functions: Proximal calculus and applications in high-dimensional statistics. *Journal of Mathematical Analysis and Applications*, Vol. 457, No. 2, pp. 1283–1306, 2018.

[81] Hiroki Kuroda and Daichi Kitahara. Block-sparse recovery with optimal block partition. *IEEE Transactions on Signal Processing*, Vol. 70, pp. 1506–1520, 2022.

[82] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 2, pp. 210–227, 2008.

[83] Yong Xu, Yuping Sun, Yuhui Quan, and Yu Luo. Structured sparse coding for classification via reweighted $\ell_{2,1}$ minimization. In *CCF Chinese Conference on Computer Vision*, pp. 189–199. Springer, 2015.

[84] Jianwei Zheng, Ping Yang, Shengyong Chen, Guojiang Shen, and Wanliang Wang. Iterative re-constrained group sparse face recognition with adaptive weights learning. *IEEE Transactions on Image Processing*, Vol. 26, No. 5, pp. 2408–2423, 2017.

[85] Chao Zhang, Huaxiong Li, Chunlin Chen, Yuhua Qian, and Xianzhong Zhou. Enhanced group sparse regularized nonconvex regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[86] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces (Second Edition)*. Springer International Publishing, 2017.

[87] Yair Censor, Tommy Elfving, Nirit Kopf, and Thomas Bortfeld. The multiple-sets split feasibility problem and its applications for inverse problems. *Inverse Problems*, Vol. 21, No. 6, p. 2071, 2005.

[88] Alessandro Lanza, Serena Morigi, Ivan W Selesnick, and Fiorella Sgallari. Sparsity-inducing nonconvex nonseparable regularization for convex image processing. *SIAM Journal on Imaging Sciences*, Vol. 12, No. 2, pp. 1099–1134, 2019.

[89] Paul A Bekker. The positive semidefiniteness of partitioned matrices. *Linear Algebra and its Applications*, Vol. 111, pp. 261–278, 1988.

[90] Arthur Albert. Regression and the Moore-Penrose pseudoinverse. *Mathematics in Science and Engineering*, Vol. 94, , 1972.

[91] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, second edition, 2012.

[92] Kyohei Suzuki and Masahiro Yukawa. Robust recovery of jointly-sparse signals using minimax concave loss function. *IEEE Transactions on Signal Processing*, Vol. 69, pp. 669–681, 2020.

[93] Zhiwei Qin, Katya Scheinberg, and Donald Goldfarb. Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation*, Vol. 5, No. 2, pp. 143–169, 2013.

[94] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 5, pp. 550–554, 1994.

[95] Ferdinando S Samaria and Andy 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. IEEE, 1994.

[96] Deng Cai, Xiaofei He, Yuxiao Hu, Jiawei Han, and Thomas Huang. Learning a spatially smooth subspace for face recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7. IEEE, 2007.

List of Publications

Journal Papers

Yang Chen, Masao Yamagishi and Isao Yamada. A Linearly involved Generalized-Moreau-Enhancement of $\ell_{2,1}$-norm with Application to Weighted Group Sparse Classification, *Algorithms*, special issue "Algorithms for Convex Optimization" (invited), Vol. 14, No. 11, pp.312-325, 2021.

Yang Chen, Masao Yamagishi and Isao Yamada. A Unified Design of Generalized Moreau Enhancement Matrix for Sparsity Aware LiGME Models, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, accepted, 2023.

Peer-Reviewed International Conference

Yang Chen, Masao Yamagishi and Isao Yamada. A Generalized Moreau Enhancement of $\ell_{2,1}$-norm and Its Application to Group Sparse Classification. In *IEEE International Conference on European Signal Processing Conference (EUSIPCO)*, Ireland, 2021.

Domestic Conference

Yang Chen, Masao Yamagishi and Isao Yamada. A Simple Design of Generalized Moreau Enhancement Matrix for LiGME Models. In *IEICE SIP Symposium*, Japan, 2022.

Other Publication

(not related to this thesis) Yang Chen and Chunlin Wu. Data-Driven tight frame construction for impulsive noise removal. *Journal of Computational Mathematics*, Vol. 40, No. 1, pp.89-107, 2022.