axioms

MDPI

Article

Group Forward–Backward Orthogonal Matching Pursuit for General Convex Smooth Functions

Zhongxing Peng, Gengzhong Zheng and Wei Huang *

School of Computer Information Engineering, Hanshan Normal University, Chaozhou 521041, China; zxpeng@hstc.edu.cn (Z.P.); zhenggz@hstc.edu.cn (G.Z.)
* Correspondence: hw@hstc.edu.cn

Abstract: This paper introduces the Group Forward–Backward Orthogonal Matching Pursuit (Group-FoBa-OMP) algorithm, a novel approach for sparse feature selection. The core innovations of this algorithm include (1) an integrated backward elimination process to correct earlier misidentified groups; (2) a versatile convex smooth model that generalizes previous research; (3) the strategic use of gradient information to expedite the group selection phase; and (4) a theoretical validation of its performance in terms of support set recovery, variable estimation accuracy, and objective function optimization. These advancements are supported by experimental evidence from both synthetic and real-world data, demonstrating the algorithm's effectiveness.

Keywords: compressed sensing; group orthogonal matching pursuit; group sparse; group restricted isometry property; instance optimality

MSC: 41A65; 94A15; 94A12; 68P30

check for updates

Citation: Peng, Z.; Zheng, G.; Huang, W. Group Forward–Backward Orthogonal Matching Pursuit for General Convex Smooth Functions. *Axioms* **2024**, *13*, 774. https://doi.org/10.3390/axioms13110774

Academic Editors: Cristian S. Calude and Fevrier Valdez

Received: 24 August 2024
Revised: 25 October 2024
Accepted: 7 November 2024
Published: 8 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The problem of feature selection has garnered significant interest from the disciplines of data mining and machine learning [1–4]. It aims to recover the explanatory variable $\beta \in \mathbb{R}^d$ from $y = X\beta$, where $y \in \mathbb{R}^n$ is an observed respond, and $X \in \mathbb{R}^{n\times d}$ is a data matrix with every row as a data point and every column as a feature. Usually, we are interested in the scenarios with many more features than the number of data points, i.e., $d \gg n$. In such a case, given y and X, the estimation of β may not be unique and leads to overfitting [5]. Additional prior knowledge is required to make this problem well posed. Owing to the inherent sparsity of pertinent features, the utilization of ℓ_0-norm regularization serves as a sound choice, as it affords explicit constraint over the degree of sparsity. Thus, the feature selection problem with ℓ_0 regularization can be formulated as follows:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^d}{\arg\min} \|y - X\beta\|_2^2, \quad \text{s.t.} \|\beta\|_0 \leq k, \tag{1}$$

where k is the sparsity number of β.

However, to solve the above problem with ℓ_0 regularization is NP-hard, because the ℓ_0 norm is nonconvex and nonsmooth, which requires intractable combinatorial search [6]. As a result, approximation is necessary to solve Equation (1). One popular approximation is the Least Absolute Shrinkage and Selection Operator (Lasso) [7], where the ℓ_1 norm is used to replace ℓ_0 norm. Since ℓ_1 norm is convex, the problem of Lasso can be solved efficiently. Despite its utility, this approach has two drawbacks: (1) The ℓ_1 norm indirectly influences the level of sparsity, necessitating additional strong assumptions to ensure a suitably sparse solution. (2) To achieve a sparse feature representation, a substantial regularization parameter is required, which can be problematic since it tends to suppress both irrelevant and pertinent features to zero. To address these shortcomings, a two-stage approach is

proposed in [8] for thresholding the resultant coefficients, but it needs additional tuning parameters and makes the procedure more complex and less robust [5].

Orthogonal Matching Pursuit (OMP), which is a greedy algorithm that selects one feature at a time, serves as a compelling alternative to Lasso in the approximation of the ℓ_0 norm [9–13]. In contrast to Lasso, OMP is simpler, more efficient, and explicitly maintains sparsity. Since the residual of OMP is always orthogonal to the features already chosen, the same feature cannot be selected twice, resulting in a sparsity of k after k steps. Notably, OMP is referred to as boosting in the field of machine learning [14]. However, standard OMP lacks a mechanism to eliminate incorrectly chosen features from earlier steps. To address this, backward steps have been introduced to remove erroneous features from OMP [5,15]. Furthermore, incorporating gradient information enhances the efficiency of OMP [11] and FoBa-OMP [15].

Nonetheless, all the aforementioned OMP-based algorithms neglect to incorporate a widely embraced prior knowledge concerning the group structure among features. In numerous real-world applications, features are not entirely independent; instead, they tend to cluster together into distinct groups. This prior knowledge has been applied to different problems, such as simultaneous sparse approximation [16], multi-task compressive sensing [17], and group Lasso [18,19]. Its benefit in sparse recovery over standard algorithms has been theoretical proven by [20–22]. Given the salience of group structure, the challenge of feature selection evolves into

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \|y - X\beta\|_2^2, \quad \text{s.t.} \|\beta\|_{2,0} \leq k, \tag{2}$$

where $\|\cdot\|_{2,0}$ represents the $l_{2,0}$ norm (i.e., compute the l_2 norm within every group and then apply the l_0 norm among groups). Herein, k denotes the quantity of nonzero groups. Within the context of this paper, k serves to represent either the degree of sparsity or the number of nonzero groups, with the precise interpretation being readily deducible from the context.

To harness the advantages of group structure, Group-OMP is explored in [23,24] using two particular models: least squares and logistic regression. Analogous to conventional OMP, Group-OMP retains simplicity, efficiency, and the explicit preservation of sparsity. Nevertheless, it is also impeded by erroneous selections made in its initial stages. Moreover, addressing the application of Group-OMP to more general functions remains an unresolved issue.

In this paper, we focus on the application of Group-OMP to the feature selection problem, considering the following general smooth objective function $Q(\beta)$, i.e.,

$$\beta^* = \underset{\bar{\beta}}{\arg\min}\, Q(\beta) \qquad \text{s.t.} \quad \|\beta\|_{2,0} \leq k. \tag{3}$$

Evidently, our model encompasses the preceding work of Group-OMP presented in [23,24] as two specific cases. To address this general problem formulated in Equation (3), we introduce an innovative algorithm termed Group Forward–Backward OMP (Group-FoBa-OMP). Our algorithm integrates backward steps and gradient information to mitigate the limitations of existing Group-OMP algorithms. Furthermore, we establish theoretical bounds for the proposed algorithm to ascertain its performance.

The main contributions of this paper can be summarized in four aspects: (1) We introduce backward steps into Group-OMP to eliminate incorrectly selected features. (2) We investigate a general model that encompasses other models from previous research on group OMP as its special cases. (3) We utilize gradient information to supplant the reduction of the objective function as the criterion for selecting potential features, thereby enhancing the efficiency of our algorithm. (4) We establish theoretical bounds for our algorithm concerning support set recovery, explanatory variable estimation, and the objective

function value. Lastly, the efficacy of the proposed Group-FoBa-OMP is validated on both synthetic and real-world data.

2. Materials and Methods

2.1. Algorithm

This subsection introduces our Group-FoBa-OMP algorithm. By incorporating backward steps and gradient information, it enhances the capabilities of Group-OMP, yielding improved effectiveness and robustness for general convex smooth models. The algorithm is formally presented in Algorithm 1.

Algorithm 1: Group-FoBa-OMP

Input : $\epsilon_G > 0$
Output: $\beta^{(k)}$
Initial : $F^{(0)} = \emptyset, \quad \beta^{(0)} = 0, \quad k = 0;$
while $\|\nabla Q(\beta^{(k)})\|_{2,\infty} \geq \epsilon_G$ **do**
  `// forward step`
  $G_{i^{(k)}} = \arg\max_{G_i \notin F^{(k)}} : \|\nabla Q(\beta^{(k)})_{G_i}\|_2;$
  $F^{(k+1)} = F^{(k)} \cup G_{i^{(k)}};$
  $\beta^{(k+1)} = \arg\min_{\mathrm{supp}(\beta) \subset F^{(k+1)}} Q(\beta);$
  $\delta^{(k+1)} = Q(\beta^{(k)}) - Q(\beta^{(k+1)});$
  $k = k + 1;$
  `// backward correction step`
  while $\min_{G_i \in F^{(k+1)}} Q(\beta^{(k)} - E_{G_i}\beta^{(k)}_{G_i}) - Q(\beta^{(k)}) < \frac{\delta^{(k)}}{2}$ **do**
    $G_{i^{(k)}} = \arg\min_{G_i} Q(\beta^{(k)} - E_{G_i}\beta^{(k)}_{G_i});$
    $k = k - 1;$
    $F^{(k)} = F^{(k+1)} \cap \bar{G}_{i^{(k+1)}};$
    $\beta^{(k)} = \arg\min_{\mathrm{supp}(\beta) \subset F^{(k)}} Q(\beta);$
  end while
end while

Our notations are as follows: $\epsilon_G > 0$ is the stop condition. $\|\cdot\|_{2,\infty}$ represents the $l_{2,\infty}$ norm (i.e., it first computes the l_2 norm within every group and then computes the l_∞ norm among groups). G_i is a set containing the index of the features which belongs to the i-th group. The set $\mathrm{supp}(\beta)$ is the support set of β, which is defined as $\mathrm{supp}(\beta) = \{j : \beta_j \neq 0\}$. The matrix $E_{G_i} \in \mathbb{R}^{d \times d}$ is diagonal. Its diagonal elements comprise $e_{ll} = 1$ if $l \in G_i$ or $e_{ll} = 0$ otherwise. For two sets of A and B, we use $A - B$ to denote $\overline{(A \cap B)} \cap A$, i.e., a set consists of elements only belong to A but do not belong to B. $\nabla Q(\beta)$ is the gradient of objective function $Q(\beta)$ with respect to β. $\nabla Q(\beta)_{G_i} \in \mathbb{R}^{|G_i|}$ is a reduced gradient vector, consisting of the entries indexed by G_i, where $|G_i|$ is the number of vectors in set G_i.

Throughout our proofs, we assume that the objective function $Q(\beta)$ satisfies the restricted strong convexity constants (RSCCs), where $\rho_-(s)$ and $\rho_+(s)$ for any given integer $s > 0$, which generalizes the restricted isometry property (RIP) in [25]. Then, the values of the objective function at β and β' satisfy the following lower and upper bounds:

$$\frac{\rho_-(s)}{2}\|t\|^2 \leq Q(\beta') - Q(\beta) - \langle \nabla Q(\beta), t \rangle \leq \frac{\rho_+(s)}{2}\|t\|^2,$$

where $t = \beta' - \beta$ and $\|t\|_0 \leq s$. The geometric interpretation of RSCCs is illustrated in Figure 1 for the case of $\beta \in \mathbb{R}^1$. According to this figure, RSCCs intend to constrain the derivation of $Q(\beta)$, which should be neither too gentle nor too steep. Definitions similar to RSCCs are also used in [15,26–29].

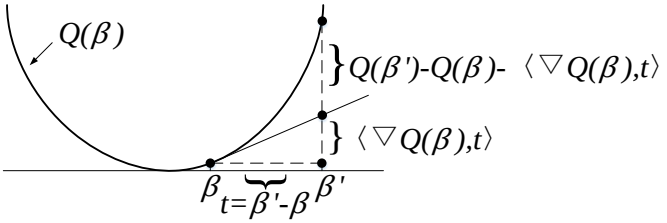


Figure 1. Geometry interpretation of RSCCs for $\beta \in \mathbb{R}^1$. The expression $\nabla Q(\beta)$ denotes the gradient of the function $Q(\beta)$ evaluated at the point β, where t is defined as $\beta' - \beta$. Consequently, $Q(\beta') - Q(\beta) - \langle \nabla Q(\beta), t \rangle$ signifies the discrepancy at point β between the value of the function $Q(\beta)$ and the slope originating from point β. Within the RSCC framework, we employ a pair of lower-bound $\rho_-(s)$ and upper-bound $\rho_+(s)$ to ensure that the gradient of the cost function $Q(\beta)$ remains within reasonable bounds, neither excessively flat nor overly steep.

The proposed Algorithm 1 of Group-FoBa-OMP runs as follows:

- It begins with an empty index set $F = \emptyset$ to contain selected features, and $\beta = 0$. The rest of the algorithm iterates with forward and backward steps.
- In forward steps (step 2 to 6), firstly, the best group of features outside the current F is added to F.
- Secondly, we minimize β based on the new F.
- Finally, a parameter δ is computed as the stopping threshold for the following backward steps.
- In backward steps (step 8 to 11), every group within F is evaluated based on its contribution to reduce function value $Q(\beta)$.
- Then, the groups that contribute the least and less than the threshold will be removed from F.
- Again, β is computed by using the new F.
- Backward step terminates after all of these insignificant groups are removed.
- The whole algorithm stops if the gradient of $Q(\beta)$ satisfies $\|\nabla Q(\beta^{(k)})\|_{2,\infty} \leq \epsilon_G$.

In this algorithm, gradients are used in step 1 and step 2 to accelerate them, instead of computing the function values using the standard OMP. The reason is that gradients are one order lower than their original functions, and they require less effort to compute.

The beauty of the proposed algorithm lies in its simplicity and efficiency. However, to prove its performance with theoretical bounds is far from straightforward. We will devote the next subsection to prove theoretical guarantees for the proposed Group-FoBa-OMP.

2.2. Theoretical Results

In this subsection, we prove theoretical bounds for the proposed Group-FoBa-OMP in terms of support set recovery, explanatory variable estimation, and the value of the objective function.

Theorem 1. *Let $s \leq n$ be any positive integer and choose ϵ_G as*

$$\epsilon_G > \frac{2\sqrt{2}\rho_+(|G_0|)}{\rho_-(s)} \|\nabla Q(\bar{\beta})\|_{2,\infty}. \tag{4}$$

where $|G_0|$ is the largest group size. By denoting $\beta^{(k)}$ and $F^{(k)}$ as the explanatory variable and support set of Algorithm 1 after iteration k, respectively, we have the following bounds:

$$\|\beta^{(k)} - \beta^*\|^2 \leq \frac{8\mu^2\epsilon_G^2}{\rho_-(s)^2}\bar{\triangle} \tag{5}$$

$$Q(\beta^{(k)}) - Q(\beta^*) \leq \frac{\mu^2\epsilon_G^2}{\rho_-(s)}\bar{\triangle} \tag{6}$$

$$\frac{\rho_-(s)^2|G_0|}{8\mu^2\rho_+(|G_0|)^2}|F^{(k)}-F^*|_G = \frac{\rho_-(s)^2}{8\mu^2\rho_+(|G_0|)^2}|F^{(k)}-F^*| \leq |F^*-F^{(k)}| \leq 2\bar{\triangle} \tag{7}$$

where $\mu = \frac{1}{\sqrt{|G_0|}}$, $\gamma = \frac{2\sqrt{2}\mu\epsilon_G}{\rho_-(s)}$, $\bar{\triangle} := |\{j \in F^* - F^{(k)} : |\bar{\beta}_j| < \gamma\}|$, $F^* - F^{(k)} = F^* \cap \bar{F}^*$, F^* *and* β^* *are corresponding ground truths.*

To prove the above Theorem 1, we need to prove several lemmas in advance.

Firstly, Lemma 1 establishes result to show that Algorithm 1 will terminate with $Q(\beta^{(k)}) \geq Q(\beta^*)$.

Lemma 1. *Let* $\bar{\beta} = \arg\min_{\text{supp}(\beta)\in\bar{F}} Q(\beta)$. *Consider* $\beta^{(k)}$ *in the beginning of each iteration of Algorithm 1. Denote its support set as* $F^{(k)}$. *Let* s *be any integer larger than* $|F^{(k)} - \bar{F}|$. *If takes* $\epsilon_G > \frac{2\sqrt{2|G_0|}\mu\rho_+(|G_0|)}{\rho_-(s)}\|\nabla Q(\bar{\beta})\|_{2,\infty}$, *then we have* $Q(\beta^{(k)}) \geq Q(\bar{\beta})$, *where* $\mu = \frac{1}{\sqrt{|G_0|}}$.

Proof. Since we assume $\bar{\beta} = \arg\min_{\text{supp}(\beta)\in\bar{F}} Q(\beta)$, then we have

$$Q(\beta^{(k)}) - Q(\bar{\beta}) \geq \langle \nabla Q(\bar{\beta}), \beta^{(k)} - \bar{\beta}\rangle + \frac{\rho_-(s)}{2}\|\beta^{(k)} - \bar{\beta}\|^2 \tag{8}$$

Meanwhile, because $Q(\beta)$ is a convex function in terms of β, it is easy to know that

$$\nabla Q(\bar{\beta})_{\bar{F}} = 0, \tag{9}$$

which further makes Equation (8) become

$$\begin{aligned} Q(\beta^{(k)}) - Q(\bar{\beta}) \geq & \langle \nabla Q(\bar{\beta})_{F^{(k)}-\bar{F}}, (\beta^{(k)} - \bar{\beta})_{F^{(k)}-\bar{F}}\rangle + \frac{\rho_-(s)}{2}\|\beta^{(k)} - \bar{\beta}\|^2 \\ \geq & -\mu\|\nabla Q(\bar{\beta})\|_{2,\infty}\|(\beta^{(k)} - \bar{\beta})_{F^{(k)}-\bar{F}}\|_1 + \frac{\rho_-(s)}{2}\|\beta^{(k)} - \bar{\beta}\|^2. \end{aligned} \tag{10}$$

Considering the fact that $\|A\|_1 \leq \sqrt{m}\|A\|_2$ for $A \in \mathbb{R}^{m\times n}$, then Equation (10) becomes

$$\begin{aligned} Q(\beta^{(k)}) - Q(\bar{\beta}) \geq & -\mu\|\nabla Q(\bar{\beta})\|_{2,\infty}\sqrt{|F^{(k)} - \bar{F}|}\|\beta^{(k)} - \bar{\beta}\| + \frac{\rho_-(s)}{2}\|\beta^{(k)} - \bar{\beta}\|^2 \\ = & -\mu\|\nabla Q(\bar{\beta})\|_{2,\infty}\sqrt{|G_0|}\sqrt{|F^{(k)} - \bar{F}|_G}\|\beta^{(k)} - \bar{\beta}\| + \frac{\rho_-(s)}{2}\|\beta^{(k)} - \bar{\beta}\|^2 \\ \geq & -\mu\|\nabla Q(\bar{\beta})\|_{2,\infty}\|\beta^{(k)} - \bar{\beta}\|\|(\beta^{(k)} - \bar{\beta})_{F^{(k)}-\bar{F}}\|\frac{\sqrt{2}\rho_+(|G_0|)\sqrt{|G_0|}}{\epsilon_G} \\ & + \frac{\rho_-(s)}{2}\|\beta^{(k)} - \bar{\beta}\|^2 \quad \text{(Lemma 3)} \\ \geq & -\mu\|\nabla Q(\bar{\beta})\|_{2,\infty}\|\beta^{(k)} - \bar{\beta}\|^2\frac{\sqrt{2}\rho_+(|G_0|)\sqrt{|G_0|}}{\epsilon_G} + \frac{\rho_-(s)}{2}\|\beta^{(k)} - \bar{\beta}\|^2 \\ = & \left(\frac{\rho_-(s)}{2} - \frac{\sqrt{2|G_0|}\mu\rho_+(|G_0|)}{\epsilon_G}\|\nabla Q(\bar{\beta})\|_{2,\infty}\right)\|\beta^{(k)} - \bar{\beta}\|^2. \end{aligned} \tag{11}$$

Since we assume $\epsilon_G > \frac{2\sqrt{2|G_0|}\mu\rho_+(|G_0|)}{\rho_-(s)}\|\nabla Q(\bar{\beta})\|_{2,\infty}$, it is easy to conclude that

$$Q(\beta^{(k)}) - Q(\bar{\beta}) > 0. \tag{12}$$

□

Secondly, Lemma 2 proves a lower bound to measure the contribution of a group of features in terms of reducing $Q(\beta)$. This lemma will be used in the proof of Lemma 3.

Lemma 2. *If* $\|\nabla Q(\beta)_{G_j}\|_2 \geq \epsilon_G$, *then we have*

$$Q(\beta) - \min_{\eta_{G_j}} Q(\beta + E_{G_j}\eta_{G_j}) \geq \frac{\epsilon_G^2}{2\rho_+(|G_j|)} \tag{13}$$

where $\eta_{G_j} \in \mathbb{R}^{|G_j|}$ *and* $E_{G_j} = [e_{G_{j_1}}, e_{G_{j_2}}, \cdots, e_{G_{j_{|G_j|}}}] \in \mathbb{R}^{d\times|G_j|}$, $j = 1, \cdots, J$, $G_i \cap G_j = \emptyset$ *for* $i \neq j$, *and* $G_1 \cup G_2 \cup \cdots \cup G_J = \{1, 2, \cdots, d\}$.

Proof. Consider the left hand side, as follows:

$$\begin{aligned}
Q(\beta) - \min_{\eta_{G_j}} Q(\beta + E_{G_j}\eta_{G_j}) =& \max_{\eta_{G_j}} Q(\beta) - Q(\beta + E_{G_j}\eta_{G_j}) \\
\geq& \max_{\eta_{G_j}} -\langle \nabla Q(\beta), E_{G_j}\eta_{G_j}\rangle - \frac{\rho_+(|G_j|)}{2}\|E_{G_j}\eta_{G_j}\|^2 \\
=& \max_{\eta_{G_j}} -\nabla Q(\beta)^T E_{G_j}\eta_{G_j} - \frac{\rho_+(|G_j|)}{2}\eta_{G_j}^T E_{G_j}^T E_{G_j}\eta_{G_j} \\
\geq& \frac{1}{2\rho_+(|G_j|)}\nabla Q(\beta)^T E_{G_j}\left(E_{G_j}^T E_{G_j}\right)^{-1} E_{G_j}^T \nabla Q(\beta) \\
=& \frac{1}{2\rho_+(|G_j|)}\nabla Q(\beta)^T E_{G_j} E_{G_j}^T \nabla Q(\beta) \\
=& \frac{\|\nabla Q(\beta)_{G_j}\|_2^2}{2\rho_+(|G_j|)} \\
\geq& \frac{\epsilon_G^2}{2\rho_+(|G_j|)}.
\end{aligned} \tag{14}$$

□

Thirdly, Lemma 3 gives an upper bound of $|F^{(k)} - F^*|_G$ to show the effectiveness of backward steps.

Lemma 3. *Consider* $\beta^{(k)}$ *with the support* $F^{(k)}$ *in the beginning of each iteration in Algorithm 1. Then, we have*

$$\|\beta^{(k)}_{F^{(k)}-F^*}\|^2 = \|(\beta^{(k)} - \beta^*)_{F^{(k)}-F^*}\|^2 \geq \frac{\delta_G^{(k)}}{\rho_+(|G_0|)}|F^{(k)} - F^*|_G \geq \frac{\epsilon_G^2}{2\rho_+(|G_0|)^2}|F^{(k)} - F^*|_G \tag{15}$$

Proof. According to RSCCs, we have

$$\begin{aligned}
&\frac{\rho_-(s)}{2}\|t\|^2 \leq Q(\beta') - Q(\beta) - \langle \nabla Q(\beta), t\rangle \leq \frac{\rho_+(s)}{2}\|t\|^2 \\
&\implies Q(\beta') \leq Q(\beta) + \langle \nabla Q(\beta), \beta' - \beta\rangle + \frac{\rho_+(s)}{2}\|\beta' - \beta\|^2,
\end{aligned} \tag{16}$$

and then we have

$$Q(\beta^{(k)} - e_{G_j}\beta^{(k)}_{G_j}) \leq Q(\beta^{(k)}) + \nabla Q(\beta)^T_{G_j}\beta^{(k)}_{G_j} + \frac{\rho_+(|G_j|)}{2}\|\beta^{(k)}_{G_j}\|^2. \tag{17}$$

Furthermore, we can have

$$\begin{aligned}|F^{(k)}-\bar{F}|_G \min_{j\in F_G^{(k)}} Q(\beta^{(k)}-e_{G_j}\beta_{G_j}^{(k)}) \leq & \sum_{j\in(F^{(k)}-\bar{F})_G} Q(\beta^{(k)}-e_{G_j}\beta_{G_j}^{(k)}) \\ \leq & \sum_{j\in(F^{(k)}-\bar{F})_G} Q(\beta^{(k)})+\nabla Q(\beta)_{G_j}^T\beta_{G_j}^{(k)}+\frac{\rho_+(|G_j|)}{2}\|\beta_{G_j}^{(k)}\|^2 \\ = & |F^{(k)}-\bar{F}|_G Q(\beta^{(k)})+\sum_{j\in(F^{(k)}-\bar{F})_G}\nabla Q(\beta)_{G_j}^T\beta_{G_j}^{(k)} \\ & +\frac{\rho_+(|G_0|)}{2}\|(\beta^{(k)}-\bar{\beta})_{F^{(k)}-\bar{F}}\|^2 \\ \leq & |F^{(k)}-\bar{F}|_G Q(\beta^{(k)})+\frac{\rho_+(|G_0|)}{2}\|(\beta^{(k)}-\bar{\beta})_{F^{(k)}-\bar{F}}\|^2, \end{aligned} \tag{18}$$

which leads to

$$\begin{aligned}\frac{\rho_+(|G_0|)}{2}\|(\beta^{(k)}-\bar{\beta})_{F^{(k)}-\bar{F}}\|^2 \geq & |F^{(k)}-\bar{F}|_G\left(\min_{j\in F_G^{(k)}} Q(\beta^{(k)}-e_{G_j}\beta_{G_j}^{(k)})-Q(\beta^{(k)})\right) \\ \geq & |F^{(k)}-\bar{F}|_G\frac{\delta_G^{(k)}}{2}. \end{aligned} \tag{19}$$

According to Lemma 2, we have the following conclusion:

$$\delta_G^{(k)} \geq \frac{\epsilon_G^2}{2\rho_+(|G_0|)}. \tag{20}$$

□

As a result, having these lemmas on hand, we can prove Theorem 1 as follows:

Proof of Theorem 1. According to Lemma 1 and RSCCs, we have

$$0 \geq Q(\bar{\beta})-Q(\beta^{(k)}) \geq \langle\nabla Q(\beta^{(k)}),\bar{\beta}-\beta^{(k)}\rangle+\frac{\rho_-(s)}{2}\|\bar{\beta}-\beta^{(k)}\|^2, \tag{21}$$

which implies that

$$\begin{aligned}\frac{\rho_-(s)}{2}\|\bar{\beta}-\beta^{(k)}\|^2 \leq & -\langle\nabla Q(\beta^{(k)}),\bar{\beta}-\beta^{(k)}\rangle \\ = & -\langle\nabla Q(\beta^{(k)})_{\bar{F}-F^{(k)}},(\bar{\beta}-\beta^{(k)})_{\bar{F}-F^{(k)}}\rangle \\ \leq & \mu\|\nabla Q(\beta^{(k)})_{\bar{F}-F^{(k)}}\|_{2,\infty}\|(\bar{\beta}-\beta^{(k)})_{\bar{F}-F^{(k)}}\|_1 \\ \leq & \mu\sqrt{|\bar{F}-F^{(k)}|}\|\nabla Q(\beta^{(k)})_{\bar{F}-F^{(k)}}\|_{2,\infty}\|(\bar{\beta}-\beta^{(k)})_{\bar{F}-F^{(k)}}\|_2 \\ \leq & \mu\epsilon_G\sqrt{|\bar{F}-F^{(k)}|}\|\bar{\beta}-\beta^{(k)}\| \end{aligned} \tag{22}$$

So, we obtain

$$\|\bar{\beta}-\beta^{(k)}\| \leq \frac{2\mu\epsilon_G}{\rho_-(s)}\sqrt{|\bar{F}-F^{(k)}|}. \tag{23}$$

It follows that

$$\begin{aligned}\frac{4\mu^2\epsilon_G^2}{\rho_-(s)^2}|\bar{F}-F^{(k)}| &\geq \|\bar{\beta}-\beta^{(k)}\|^2 \\ &\geq \|(\bar{\beta}-\beta^{(k)})_{\bar{F}-F^{(k)}}\|^2 = \|\bar{\beta}_{\bar{F}-F^{(k)}}\|^2 \\ &\geq \gamma^2|\{j\in\bar{F}-F^{(k)}:|\bar{\beta}_j|\geq\gamma\}| \\ &= \frac{8\mu^2\epsilon_G^2}{\rho_-(s)^2}|\{j\in\bar{F}-F^{(k)}:|\bar{\beta}_j|\geq\gamma\}| \end{aligned} \tag{24}$$

which implies that

$$\begin{aligned}&|\bar{F}-F^{(k)}|\geq 2|\{j\in\bar{F}-F^{(k)}:|\bar{\beta}_j|\geq\gamma\}| = 2(|\bar{F}-F^{(k)}|-|\{j\in\bar{F}-F^{(k)}:|\bar{\beta}_j|<\gamma\}|) \\ \Longrightarrow &|\bar{F}-F^{(k)}|\leq 2|\{j\in\bar{F}-F^{(k)}:|\bar{\beta}_j|<\gamma\}| \end{aligned} \tag{25}$$

Applying the above inequality to Equation (23), we have

$$\|\bar{\beta}-\beta^{(k)}\|\leq\frac{2\mu\epsilon_G}{\rho_-(s)}\sqrt{|\bar{F}-F^{(k)}|}\leq\frac{2\sqrt{2}\mu\epsilon_G}{\rho_-(s)}\sqrt{|\{j\in\bar{F}-F^{(k)}:|\bar{\beta}_j|<\gamma\}|}. \tag{26}$$

This proves the first inequality. Next, we consider

$$\begin{aligned}Q(\bar{\beta})-Q(\beta^{(k)}) \geq& \langle\nabla Q(\beta^{(k)}),\bar{\beta}-\beta^{(k)}\rangle+\frac{\rho_-(s)}{2}\|\bar{\beta}-\beta^{(k)}\|^2 \\ =& \langle\nabla Q(\beta^{(k)})_{\bar{F}-F^{(k)}},(\bar{\beta}-\beta^{(k)})_{\bar{F}-F^{(k)}}\rangle+\frac{\rho_-(s)}{2}\|\bar{\beta}-\beta^{(k)}\|^2 \\ \geq& -\mu\|\nabla Q(\beta^{(k)})\|_{2,\infty}\|\bar{\beta}-\beta^{(k)}\|_1+\frac{\rho_-(s)}{2}\|\bar{\beta}-\beta^{(k)}\|^2 \\ \geq& -\mu\epsilon_G\sqrt{|\bar{F}-F^{(k)}|}\|\bar{\beta}-\beta^{(k)}\|_2+\frac{\rho_-(s)}{2}\|\bar{\beta}-\beta^{(k)}\|^2 \\ \geq& \frac{\rho_-(s)}{2}\left(\|\bar{\beta}-\beta^{(k)}\|-\frac{\mu\epsilon_G\sqrt{|\bar{F}-F^{(k)}|}}{\rho_-(s)}\right)^2-\frac{\mu^2\epsilon_G^2|\bar{F}-F^{(k)}|}{2\rho_-(s)} \\ \geq& -\frac{\mu^2\epsilon_G^2|\bar{F}-F^{(k)}|}{2\rho_-(s)} \\ \geq& -\frac{\mu^2\epsilon_G^2}{\rho_-(s)}|\{j\in\bar{F}-F^{(k)}:|\bar{\beta}_j|<\gamma\}| \quad \text{(use Equation (25))} \end{aligned} \tag{27}$$

This proves the second inequality. The last inequality in this theorem is obtained as follows:

$$\begin{aligned}\frac{\epsilon_G^2|G_0|}{2\rho_+(|G_0|)^2}|F^{(k)}-\bar{F}|_G=\frac{\epsilon_G^2}{2\rho_+(|G_0|)^2}|F^{(k)}-\bar{F}| \leq& \|(\beta^{(k)}-\bar{\beta})_{F^{(k)}-\bar{F}}\|^2 \quad \text{(Lemma 3)} \\ \leq& \|\beta^{(k)}-\bar{\beta}\|^2\leq\frac{4\mu^2\epsilon_G^2}{\rho_-(s)^2}|\bar{F}-F^{(k)}| \quad \text{(use Equation (23))} \\ \leq& \frac{8\mu^2\epsilon_G^2}{\rho_-(s)^2}|\{j\in\bar{F}-F^{(k)}:|\bar{\beta}_j|<\gamma\}|. \end{aligned} \tag{28}$$

It proves the third inequality. □

Obviously, the inequalities of Equations (5)–(7) in Theorem 1 establish upper bounds for support set recovery, explanatory variable estimation, and the value of the objective function, respectively.

2.3. The Benefit of Group-FoBa-OMP

The proposed Group-FoBa-OMP algorithm offers two primary advantages. Firstly, it more effectively manages data exhibiting group structures compared to the conventional FoBa-OMP. Secondly, in contrast to the Group-OMP, the suggested method utilizes backward steps to refine sparse recovery by rectifying errors made during initial iterations. Moreover, the algorithm can be applied to a broad range of convex smooth models, including least squares and logistic regression.

2.3.1. Comparison with standard FoBa-OMP

By setting the group size $|G_0| = 1$ within Theorem 1, the resulting bounds simplify to those associated with FoBa-OMP, disregarding any group structure. For clarity in our discussion, we employ ϵ to signify the stopping criterion for the standard FoBa-OMP and ϵ_G to represent the corresponding criterion for the proposed Group-FoBa-OMP.

Evidently, the distinctions between the bounds for the Group-FoBa-OMP and FoBa-OMP reside in their respective stopping conditions. To elucidate the advantages of the Group-FoBa-OMP over FoBa-OMP, we will demonstrate that the value of ϵ_G will not exceed $|G_0|\epsilon$. Initially, relying on the RSCCs, we establish the following inequalities:

$$\begin{aligned} Q(\beta') - Q(\beta) - \langle \nabla Q(\beta), E_{G_j}\eta_{G_j}\rangle &\leq \frac{\rho_+(|G_0|)}{2}\|\eta_{G_j}\|^2 \\ &\leq \frac{\rho_+(1)}{2}\|\eta_{G_j}\|^2 \end{aligned}$$

which implies that $\rho_+(|G_0|) \leq \rho_+(1)$. Furthermore, when combined with another straightforward observation, namely $\|\nabla Q(\bar{\beta})\|_{2,\infty} \leq |G_0|\|\nabla Q(\bar{\beta})\|_\infty$, we arrive at

$$\epsilon_G > \frac{2\sqrt{2}\rho_+(|G_0|)}{\rho_-(s)}\|\nabla Q(\bar{\beta})\|_{2,\infty},$$

$$\epsilon > \frac{2\sqrt{2}\rho_+(1)}{\rho_-(s)}\|\nabla Q(\bar{\beta})\|_\infty \geq \frac{2\sqrt{2}\rho_+(|G_0|)}{\rho_-(s)|G_0|}\|\nabla Q(\bar{\beta})\|_{2,\infty},$$

which suggests that we can establish ϵ_G to be no greater than $|G_0|\epsilon$. Consequently, the performance of the Group-FoBa-OMP will not be inferior to that of the standard FoBa-OMP when dealing with data that exhibit group structures.

2.3.2. Comparison with Group-OMP

In this subsection, we compare our algorithm with the existing group OMP algorithms. As per [23,24], these studies only consider particular instances of least squares regression and logistic regression. Adhering to the notation used in these two publications, we introduce $\mathcal{G}_{bad}$ and g_{bad} to denote bad groups and bad indices, respectively. Furthermore, we define

$$\begin{aligned} \mu_X(\mathcal{G}_{good}) &= \|X^+_{g_{good}} X_{g_{bad}}\|^{good/bad}_{(2,1)} \\ &= \sup_{\|v\|^{bad}_{(2,1)}=1} \|X^+_{g_{good}} X_{g_{bad}} v\|^{good}_{(2,1)} \\ &= \sup_{\|v\|^{bad}_{(2,1)}=1} \sum_{G_i \in \mathcal{G}_{good}} \sqrt{\sum_{j\in G_i} (X^+_{g_{good}} X_{g_{bad}} v)^2_j} \end{aligned}$$

where X^+ is the Moore–Penrose pseudo-inverse [30,31] of X and

$$\|v\|^{bad}_{(2,1)} = \sum_{G_i \in \mathcal{G}_{bad}} \sqrt{\sum_{j\in G_i} v_j^2}. \tag{29}$$

The theoretical bounds presented in [23,24] necessitate a condition

$$\mu_X(\mathcal{G}_{good}) < 1, \tag{30}$$

which aligns with the target independent irrepresentable condition outlined in [9,26] for the standard OMP, or bears resemblance to the irrepresentable condition in [32] for Lasso. Nonetheless, this condition is sufficient yet not indispensable for ensuring the success of OMP. Additionally, it proves challenging to fulfill with real-world data. In contrast, by leveraging backward steps and gradient information, our proposed Group-FoBa-OMP can achieve success without being confined to such stringent conditions.

3. Results

In this section, we evaluate the proposed Group-FoBa-OMP on both synthetic and real-world data.

In sparse learning problems, the main difficulty lies in support set recovery. Once the support set is known, β can usually be solved efficiently by standard algorithms, such as those in [33]. Take the least squares as an example, in which β can be expressed analytically after knowing its support set. Thus, we evaluate the experimental results based on support set recovery. More specifically, we use the F-measure, $F = \frac{2PR}{P+R} = \frac{2|F^* \cap F^{(k)}|}{|F^*|+|F^{(k)}|}$, where $P = \frac{|F^* \cap F^{(k)}|}{|F^{(k)}|}$ and $R = \frac{|F^* \cap F^{(k)}|}{|F^*|}$ denote precision and recall of support sets, respectively.

In practice, it is difficult to choose stopping thresholds of ϵ_G and ϵ. Instead, we use $\epsilon_G = \|\nabla Q(\beta^*)\|_{2,\infty}$ and $\epsilon = \|\nabla Q(\beta^*)\|_\infty$ in synthetic data. In real-world data, the algorithm terminates once the number of nonzero entries in $\beta^{(k)}$ exceeds a given sparsity.

3.1. Least Squares

In this subsection, we conduct an experiment utilizing synthetic data, employing the least squares model as delineated in Equation (2).

In this example of least squares, we generate synthetic data according to the following procedures:

- We will generate $n = 128$ data points, each of which has $d = 512$ features.
- Sparsity number s ranges from 30 to 50, which means that the number of nonzero features will be between 30 and 50.
- Each data point consists of 240 feature groups with group size of either 2 or 3.
- We produce m independent auxiliary variables $z_1, \cdots, z_m \in \mathbb{R}^n$, corresponding to m nonzero groups. Each element within these variables adheres to an independently and identically distributed (i.i.d.) Gaussian distribution, specifically $z_{lj} \sim \mathcal{N}(0, 350)$, where l ranges from 1 to m and j ranges from 1 to n.
- Compute d predictors $x_i \in \mathbb{R}^n$ as follows:

$$x_i = z_{\lfloor (i-1)/p_i \rfloor + 1} + \nu_i, \tag{31}$$

  where $i = 1, \cdots, d$, $\nu_i \in \mathbb{R}^n$ represents the noise vector, with each of its elements $\nu_{ij} \in \mathbb{R}^1$ being generated according to an i.i.d. Gaussian distribution, specifically $\nu_{ij} \sim \mathcal{N}(0, 5)$.
- The ground truth of the explanatory variables β are randomly assigned from a discrete set ranging from 2 to 6, following an independently and identically distributed (i.i.d.) uniform distribution.
- Introduce observation noise $\nu \in \mathbb{R}^n$, which follows an i.i.d. Gaussian distribution, with each of its elements satisfying $\mathcal{N}(0, 300)$. Subsequently, our objective is to estimate the explanatory variables β within the following true model:

$$y = \beta_1 \sum_{i=1}^{p_1} x_i + \beta_2 \sum_{i=p_1+1}^{p_1+p_2} x_i + \cdots + \beta_m \sum_{i=p_1+\cdots+p_{m-1}+1}^{p_1+\cdots+p_m} x_i + \nu. \tag{32}$$

Experimental outcomes are presented in Figures 2–6, averaged across 50 trials. We have identified several intriguing insights:

Firstly, in terms of the F-measure, the Group-FoBa-OMP outperforms the standard OMP and FoBa-OMP, as evidenced in Figure 2. This is attributed to the fact that the Group-FoBa-OMP leverages prior knowledge of group structures to iteratively select and eliminate feature groups. Furthermore, as depicted in Figure 4, the Group-FoBa-OMP surpasses the Group-OMP due to its incorporation of backward steps to rectify the errors made in earlier stages.

Secondly, the proposed methodology effectively recovers the correct number of nonzero entries, as illustrated in Figures 3 and 5, following the application of backward steps. In contrast, the standard OMP and FoBa-OMP tend to incorporate irrelevant features. Although the Group-OMP exhibits superior performance compared to the standard OMP and FoBa-OMP, it still incurs more errors at higher levels of sparsity than our proposed approach.

Thirdly, as shown in Figure 6, the computational overhead associated with the Group-FoBa-OMP is not prohibitive. Although the inclusion of additional backward steps marginally increases the computational time relative to the Group-OMP, it remains significantly faster than both the standard OMP and FoBa-OMP. This efficiency is achieved through the group-wise selection and deletion of entries, which reduces the overall number of operations required. Moreover, the utilization of gradient information further accelerates the proposed algorithm.

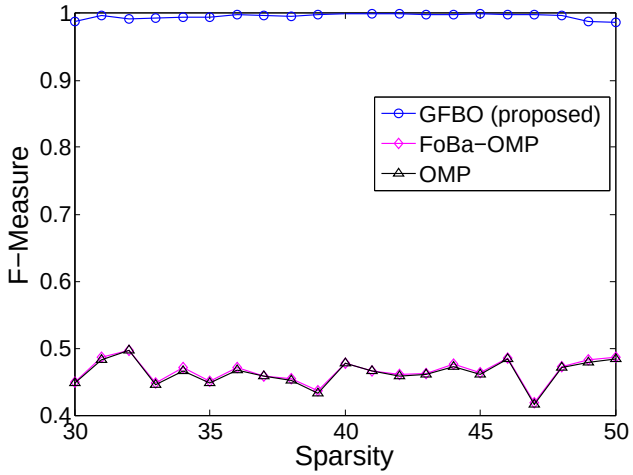


Figure 2. F-measure on the least squares model in comparison to OMP-based algorithms. GFBO denotes the proposed Group-FoBa-OMP algorithm.

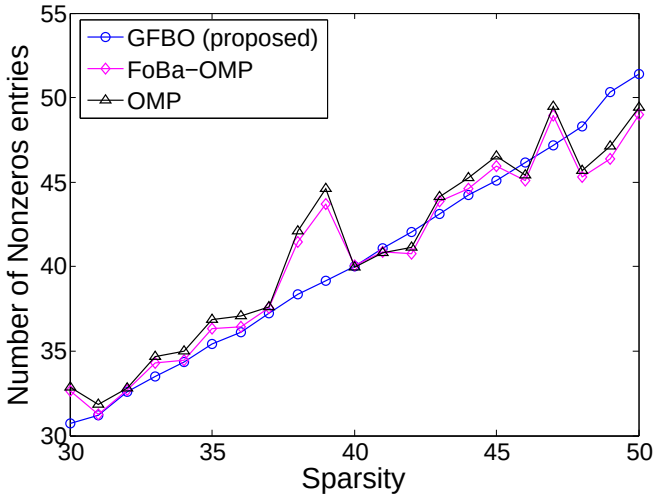


Figure 3. The number of estimated nonzero entries (i.e., the sparsity) on the least squares model in comparison to OMP-based algorithms. GFBO denotes the proposed Group-FoBa-OMP algorithm.

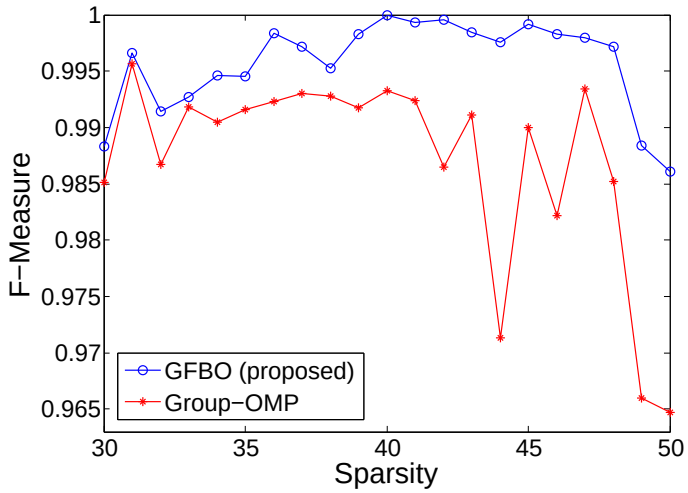


Figure 4. F-measure on the least squares model in comparison to Group-OMP. GFBO denotes the proposed Group-FoBa-OMP algorithm.

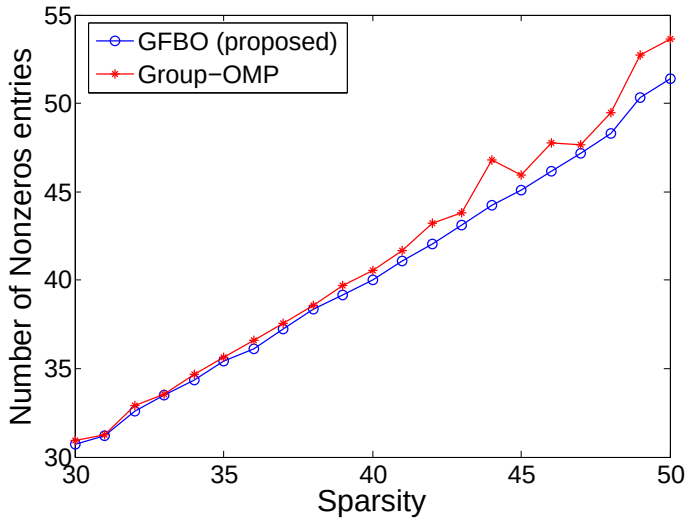


Figure 5. The number of estimated nonzero entries (i.e., the sparsity) on the least squares model in comparison to Group-OMP. GFBO denotes the proposed Group-FoBa-OMP algorithm.

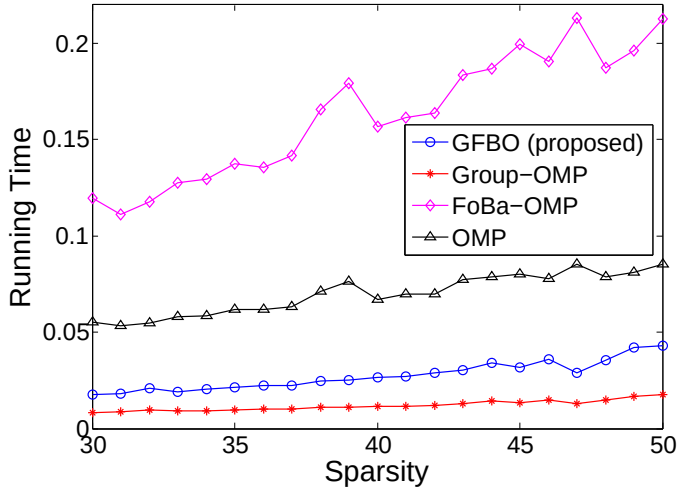


Figure 6. Running time on the least squares model. GFBO denotes the proposed Group-FoBa-OMP algorithm.

In summary, the proposed Group-FoBa-OMP exhibits its advantages over competing methods by integrating backward steps and gradient information into its framework.

3.2. Logistic Regression

In this subsection, we consider L_2-regularized logistic regression, which has a smooth convex loss function:

$$Q(\beta) = \frac{1}{n}\sum_i \log(1+\exp(-y_i X_{i.}\beta)) + \frac{1}{2}\lambda\|\beta\|^2 \tag{33}$$

where

- $X_{i.} \in \mathbb{R}^d$ represents the i-th row (or sample) in $X \in \mathbb{R}^{n\times d}$, where $n = 1024$, $d = 4096$. The group size is 4. The data matrix X consists of two classes, each of which is generated via an i.i.d. Gaussian distribution $\mathcal{N}(\beta^*, I_{d\times d})$ or $\mathcal{N}(-\beta^*, I_{d\times d})$. Here, β^* is the true sparse coefficient, whose nonzero entries are generated by uniform distribution.
- $y_i \in \{1, -1\}$ is the corresponding label,
- ℓ_2-norm regularization is used to mitigate overfitting, where $\lambda = 0.0001$.
- The gradient of $Q(\beta)$ is

$$\frac{\partial Q(\beta)}{\partial \beta} = \frac{1}{n}\sum_i \frac{-y_i \exp(-y_i X_{i.}\beta)}{1+\exp(-y_i X_{i.}\beta)} X_{i.}^T + \lambda\beta \tag{34}$$

The experimental findings are documented in Tables 1 and 2. We have drawn the following conclusions: with respect to the F-measure and the accurate recovery of the sparsity level, the Group-FoBa-OMP yields superior results compared to all other algorithms, with a particular emphasis on outperforming FoBa-OMP and the standard OMP. The underlying reasons for this are consistent with those observed in the least squares scenario.

These observations affirm the advantages of incorporating backward steps when dealing with data that exhibit group structure, aligning with our theoretical predictions.

Table 1. F-measure of logistic regression.

s [1]	**Proposed** [2]	GOMP [3]	FoBa-OMP	OMP
8	**0.9860**	0.9660	0.7026	0.6669
16	**0.9460**	0.9221	0.6740	0.6462
32	**0.9434**	0.9356	0.6596	0.6421
64	**0.9317**	0.9269	0.6435	0.6311
128	**0.9189**	0.9166	0.5959	0.5870

[1] s: sparsity. [2] Proposed: Group-FoBa-OMP (GFBO). [3] GOMP: Group-OMP.

Table 2. Number of estimated nonzero entries.

s [1]	**Proposed** [2]	GOMP [3]	FoBa-OMP	OMP
8	**8.0800**	8.4800	5.8600	6.6000
16	**15.9200**	16.8000	9.8400	10.7000
32	**32.0800**	32.6400	18.1600	19.7000

[1] s: sparsity. [2] Proposed: Group-FoBa-OMP (GFBO). [3] GOMP: Group-OMP.

3.3. A Real-World Data Set

In this subsection, we consider a real-world data set of "Communities and Crime" (http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime, accessed on 15 June 2024), which is a collection in the UCI Machine Learning Repository.

The data set consists of 319 complete instances (data), with 120 explanatory variables. The response variable is the "total number of violent crimes per 100K population ". Following similar procedures in [18,23], we use the least squares model with a third-order polynomial expansion to approximate the unknown true model, i.e., every group consists of X_i, X_i^2 and X_i^3. Then, we have 360 explanatory variables with a group size of 3.

In contrast to synthetic data, we do not know the ground truth of β nor the targeted sparsity number in the real-world case. Thus, we vary the sparsity number to find the smallest value of the test error. Here, both the training error and testing error are defined as $error = \frac{\|y_{true} - y_{est}\|_2}{\|y_{true}\|_2}$, where y_{est} is the response variable.

The experiment is conducted 50 times by utilizing a five-fold cross-validation approach. The findings are illustrated in Figures 7 and 8, where the range of sparsity levels spans from 18 to 30. From these results, we make several observations: (1) As shown in Figure 7, the proposed Group-FoBa-OMP attains the minimum error rate at a sparsity of 21. Notably, neither FoBa-OMP nor Group-OMP surpass the standard OMP in terms of achieving the lowest testing error, suggesting that neither the presence of group structures nor the implementation of backward steps alone can ensure superior performance. (2) In terms of the objective function value (as depicted in Figure 8), Group-FoBa-OMP emerges as the best one.

Once again, the experimental results highlight the benefits of incorporating backward steps within Group-OMP.

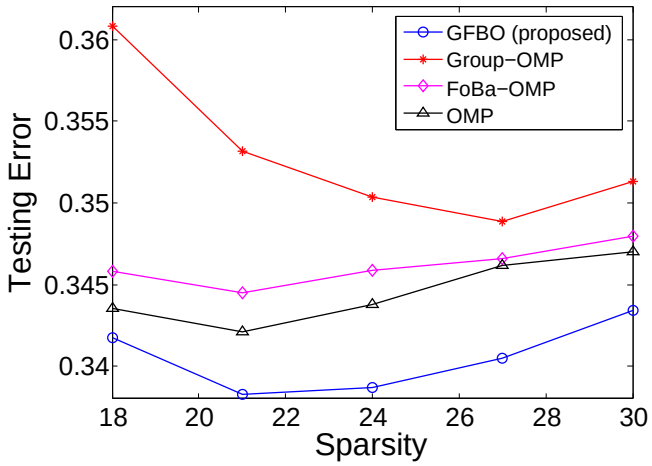


Figure 7. Testing error of regression results of "Communities and Crime Data Set". $n = 319$, $d = 120$. The group size is 3. Results are averaged over 50 trails.

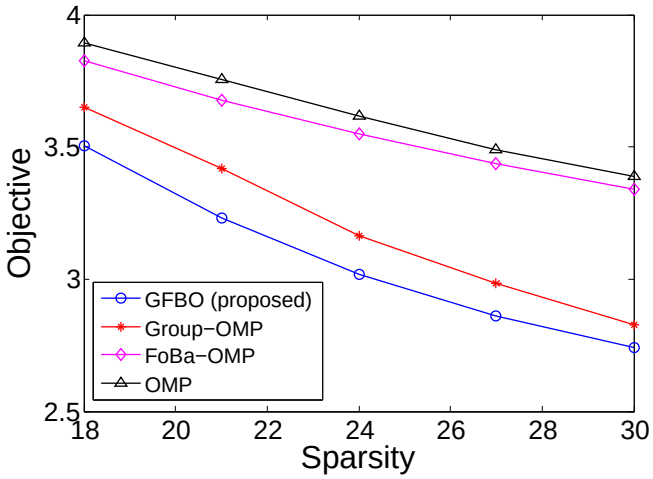


Figure 8. The value of objective function of regression results of "Communities and Crime Data Set". $n = 319$, $d = 120$. The group size is 3. Results are averaged over 50 trails.

4. Discussion

In this paper, we introduce a novel algorithm, termed Group-FoBa-OMP, which extends the Group-OMP framework by incorporating backward steps and leveraging gradient information. Whereas the existing convergence proofs for Group-OMP hinge on the irrepresentable condition, our derivations rely on more moderate assumptions. This paper establishes theoretical bounds to ascertain the performance of the proposed algorithm concerning support set recovery, coefficient estimation, and the objective function value. Empirical evidence derived from both synthetic and real-world data sets corroborates the theoretical advantages of the proposed Group-FoBa-OMP algorithm over its counterparts, namely the standard OMP, FoBa-OMP, and Group-OMP.

Author Contributions: Conceptualization, Z.P.; methodology, Z.P.; software, Z.P.; validation, Z.P., G.Z. and W.H.; formal analysis, Z.P.; investigation, Z.P.; resources, G.Z. and W.H.; data curation, Z.P.; writing—original draft preparation, Z.P.; writing—review and editing, Z.P., G.Z. and W.H.; visualization, Z.P.; supervision, G.Z. and W.H.; project administration, G.Z. and W.H.; funding acquisition, G.Z. and W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific Research Platform Project of Education Department of Guangdong Province (Grant Nos. 2021KCXTD038 and 2022KSYS003), Discipline Construction and Promotion Project of Guangdong Province (Grant No. 2022ZDJS065), Education and Teaching Reform Project of Hanshan Normal University (Grant No. PX-161241546).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: The authors would like to thank the referees and the editors for their useful suggestions, which significantly improved this paper.

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

References

1. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
2. Ng, A.Y. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 78.
3. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
4. Wang, R.; Qin, Y.; Wang, Z.; Zheng, H. Group-Based Sparse Representation for Compressed Sensing Image Reconstruction with Joint Regularization. *Electronics* **2022**, *11*, 182. [CrossRef]
5. Zhang, T. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inf. Theory* **2011**, *57*, 4689–4708. [CrossRef]
6. Candes, E.; Wakin, M.; Boyd, S. Enhancing sparsity by reweighted L1 minimization. *J. Fourier Anal. Appl.* **2008**, *14*, 877–905. [CrossRef]
7. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. Methodol.* **1996**, *58*, 267–288. [CrossRef]
8. Zhang, T. Some sharp performance bounds for least squares regression with L1 regularization. *Ann. Stat.* **2009**, *37*, 2109–2144. [CrossRef]
9. Tropp, J. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inf. Theory* **2004**, *50*, 2231–2242. [CrossRef]
10. Tropp, J.; Gilbert, A. Signal recovery from random measurements via Orthogonal Matching Pursuit. *IEEE Trans. Inf. Theory* **2007**, *53*, 4655–4666. [CrossRef]
11. Zhang, T. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans. Inf. Theory* **2011**, *57*, 6215–6221. [CrossRef]
12. Bi, X.; Leng, L.; Kim, C.; Liu, X.; Du, Y.; Liu, F. Constrained Backtracking Matching Pursuit Algorithm for Image Reconstruction in Compressed Sensing. *Appl. Sci.* **2021**, *11*, 1435. [CrossRef]
13. Shao, C.; Wei, X.; Ye, P.; Xing, S. Efficiency of Orthogonal Matching Pursuit for Group Sparse Recovery. *Axioms* **2023**, *12*, 389. [CrossRef]
14. Buhlmann, P. Boosting for high-dimensional linear models. *Ann. Stat.* **2006**, *34*, 559–583. [CrossRef]
15. Liu, J.; Fujimaki, R.; Ye, J. Forward-Backward Greedy Algorithms for General Convex Smooth Functions over A Cardinality Constraint. In Proceedings of the International Conference on Machine Learning (ICML-2014), Beijing, China, 21–26 June 2014.
16. Wipf, D.; Rao, B. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Trans. Signal Process* **2007**, *55*, 3704–3716. [CrossRef]
17. Ji, S.; Dunson, D.; Carin, L. Multi-task compressive sensing. *IEEE Trans. Signal Process* **2009**, *57*, 92–106. [CrossRef]

18. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* **2006**, *68*, 49–67. [CrossRef]
19. Bach, F.R. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **2008**, *9*, 1179–1225.
20. Huang, J.; Zhang, T. The Benefit of Group Sparsity. *Ann. Stat.* **2010**, *38*, 1978–2004. [CrossRef]
21. Huang, J.; Zhang, T.; Metaxas, D. Learning with structured sparsity. *J. Mach. Learn. Res.* **2011**, *12*, 3371–3412.
22. Louici, K.; Pontil, M.; Tsybakov, A.B.; Van De Deer, S.A. Taking advantage of sparsity in multi-task learning. In Proceedings of the Conference on Learning Theory (COLT-2009), Montreal, QC, Canada, 18–21 June 2009.
23. Lozano, A.; Swirszcz, G.; Abe, N. Group Orthogonal Matching Pursuit for Variable Selection and Prediction. In Proceedings of the Advances in Neural Information Processing Systems (NIPS-2009), Vancouver, BC, Canada, 7–10 December 2009.
24. Lozano, A.; Swirszcz, G.; Abe, N. Group Orthogonal Matching Pursuit for Logistic Regression. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS-2011), Fort Lauderdale, FL, USA, 11–13 April 2011.
25. Candes, E.; Tao, T.; Boyd, S. Decoding by linear programming. *IEEE Trans. Inf. Theory* **2005**, *51*, 4203–4215. [CrossRef]
26. Zhao, P.; Yu, B. On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.* **2009**, *10*, 555–568.
27. Jalali, A.; Johnson, C.C.; Ravikumar, P.D. On learning discrete graphical models using greedy methods. In Proceedings of the Advances in Neural Information Processing Systems (NIPS-2011), Granada, Spain, 12–15 December 2011.
28. Bahmani, S.; Boufounos, P.; Raj, B. Greedy sparsityconstrained optimization. *J. Mach. Learn. Res.* **2013**, *14*, 807–841.
29. Negahban, S.; Ravikumar, P.D.; Wainwright, M.J.; Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In Proceedings of the Advances in Neural Information Processing Systems (NIPS-2009), Vancouver, BC, Canada, 7–10 December 2009.
30. Moore, E. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.* **1920**, *26*, 394–395.
31. Penrose, R. A generalized inverse for matrices. *Proc. Camb. Philos. Soc.* **1955**, *51*, 406–413. [CrossRef]
32. Zhao, P.; Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res.* **2006**, *7*, 2541–2567.
33. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge Univesity Press: Cambridge, UK, 2004.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.