Best Subset Selection: Optimal Pursuit for Feature Selection and Elimination

Zhihan Zhu, Yanhao Zhang

School of Mathematical Sciences, Beihang University

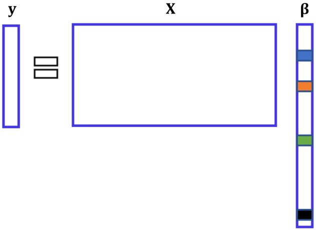
June 15, 2025

Joint work with Prof. Yong Xia

Code Paper

Best Subset Selection

Subset Selection : Selecting the most important features in high-dimensional data is a fundamental challenge in statistics and machine learning.



Application :

- Feature selection [Kohavi and John, 1997, Das and Kempe, 2011]
- Sparse regression [Miller, 2002, Das and Kempe, 2018]
- Compressed sensing [Chen et al., 2001]
- Maximum coverage [Feige, 1998]
- Large language model [Wang et al., 2024]

Best Subset Selection

The fundamental multivariate linear regression model with coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is expressed as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{n\times 1}$ represents the response vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times 1}$ denotes the measurement noise.

The goal of **best subset selection** is to select a subset of features by identifying nonzero coefficients (i.e., Active / Support Set S) in $\boldsymbol{\beta}$ that achieves a balance between accuracy and model simplicity:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{L}_n(\boldsymbol{\beta}) \triangleq \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.}\ \|\boldsymbol{\beta}\|_0 \leq K, \tag{2}$$

where K is maximum allowed sparsity level.

Best subset selection (BSS) is considered the gold standard for feature selection, but the problem is NP-hard!

Subset Selection Algorithms

Relaxation-based Methods

Since problem (2) is NP-hard [Davis et al., 1997], significant efforts have been directed toward developing polynomial-time approximation algorithms.

Relaxation-based methods:

- Least Absolute Shrinkage and Selection Operator (LASSO) [Tibshirani, 1996]
- Adaptive LASSO [Zou, 2006]
- Smoothly Clipped Absolute Deviation (SCAD) [Fan and Li, 2001]
- Minimax Concave Penalty (MCP) [Zhang, 2010].

However, these methods could be **computational burdensome** [Hazimeh and Mazumder, 2020, Needell and Tropp, 2009] and are also difficult to **control the number of selected features**.

Subset Selection Algorithms

Greedy Algorithms

Another widely used class of methods is greedy algorithms, known for their high computational efficiency and simplicity.

Greedy algorithms:

- Perform subset selection directly by selecting and eliminating basis based on **feature importance**.
- The criteria for feature selection and elimination in this category **are generally consistent**.
- Differ only in the underlying combination strategies.

Greedy Algorithms

Feature Selection Criterion

Correlation-based Selection. Greedy algorithms typically select features based on their correlation with residuals, calculated as follows:

$$\mathbf{r}^k = \mathbf{y} - \mathbf{X}\beta^{k-1}, \ j^* = \underset{j \in S^c}{\text{argmax}} \frac{|{\mathbf{r}^k}^T \mathbf{X}_j|}{\|\mathbf{X}_j\|_2}, \tag{3}$$

where $\mathbf{X}_j$ is the j-th column of $\mathbf{X}$, S^c is the complement of support S, β^{k-1} denotes the updated coefficient on S, and $\mathbf{r}^k$ represents the residual at step k.

Representative methods:

- Matching Pursuit (MP) [Mallat and Zhang, 1993].
- Orthogonal Matching Pursuit (OMP) [Pati et al., 1993].
- CoSaMP [Needell and Tropp, 2009].
- Least Angle Regression (LARS) [Efron et al., 2004].
- Adaptive Best-Subset Selection (ABESS) [Zhu et al., 2020].

Greedy Algorithms

Feature Elimination Criterion

Wald-T Test Statistics-based Elimination. Feature elimination often relies on the absolute value of Wald-T test statistics, defined as (here we assume the columns of **X** are centralized with zero mean for convenience):

$$|T_j| = \frac{|\beta_j^{k-1}|}{M_{\beta_j^{k-1}}}, \text{ where } M_{\beta_j^{k-1}} = \frac{\|\mathbf{r}^k\|/\sqrt{df}}{\sqrt{\mathbf{X}_j^T\mathbf{X}_j}}, \; j \in S, \tag{4}$$

where *df* serves as degree of freedom. Elimination is often based on minimizing the T-statistic or setting a threshold for deletion.

Representative methods:

- Iterative Hard Thresholding (IHT) [Blumensath and Davies, 2009].
- Hard Thresholding Pursuit (HTP) [Foucart, 2011].
- CoSaMP [Needell and Tropp, 2009].
- Adaptive Best-Subset Selection (ABESS) [Zhu et al., 2020].

Classic Criteria

In general, greedy methods can be regarded as a combination of correlation-based selection and T-statistic-based elimination, with different strategies integrated to perform subset selection.

- Criteria (3) and (4), according to their formulas, **focus solely on the individual significance of features**, neglecting their interaction with other features.
- A feature that appears important within the current active set might become less significant when the active set changes, and conversely, a feature deemed less critical could gain importance under a different active set configuration. The current criteria fail to **capture these dynamic properties**.
- How to interpret these criteria from an optimization perspective?

Optimization Perspective

From an optimization standpoint, the existing criteria can be interpreted as the variation of objective function achieved by updating the support set with fixed coefficients in the first step of **block coordinate descent**. Specifically:

Optimization Perspective

Step 1 (Support update with fixed coefficient): Maximize the correlation in (3) (or minimize the T-statistics in (4)) for support set updating is equivalent to solving (P0) (or (Q0)). The classical criteria (3) and (4) corresponds to the variation of objective function value in this part exactly.

$$\begin{aligned} \operatorname*{argmin}_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{s. t.} \quad & \|\boldsymbol{\beta} - \boldsymbol{\beta}^{k-1}\|_0 \leq 1,\ \operatorname{supp}(\boldsymbol{\beta}^{k-1}) \subset \operatorname{supp}(\boldsymbol{\beta}) . \end{aligned} \tag{P0}$$

$$\begin{aligned} \operatorname*{argmin}_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{s. t.} \quad & \|\boldsymbol{\beta} - \boldsymbol{\beta}^{k-1}\|_0 \leq 1,\ \operatorname{supp}(\boldsymbol{\beta}) \subset \operatorname{supp}(\boldsymbol{\beta}^{k-1}). \end{aligned} \tag{Q0}$$

Optimization Perspective

Step 2 (Coefficient update with fixed support): It is followed by refining the coefficients on the updated support set. This step also leads to a change in the function value, which, however, is not captured by the classical criteria.

Limitations of classic criteria:

- The constraints in (P0) and (Q0) (or equivalently, criteria (3) and (4)) only allows one change in the support set of β while the coefficients on the remaining support set are fixed.
- In the next step when coefficients are updated on newly selected support set, the influence of newly chosen feature on the remaining coefficients is not considered.

Optimization Perspective

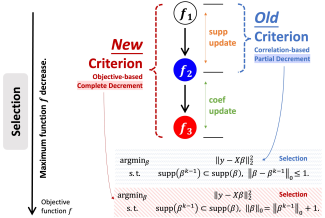
Modeling feature individual significance and interaction via block coordinate descent:

- **Step 1 measures the individual significance of the features.**
- **Step 2 assesses the interaction between features**, where classical criteria fail to capture.

Therefore, the update strategy in (P0) and (Q0) (or criteria (3) and (4)) can be understood as **the objective of performing one step of block coordinate descent**, rather than an objective that takes into account the overall descent.

A. SUBSET SELECTION

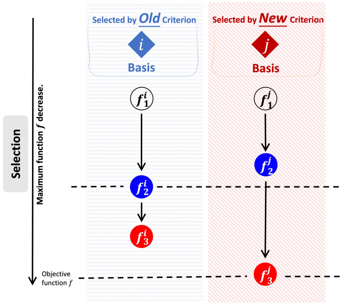
A1. The old correlation-based criterion corresponds to the partial decrement of the objective function



$$\operatorname{argmin}_\beta \quad \|y - X\beta\|_2^2 \qquad \text{Selection}$$
$$\text{s.t.} \quad \operatorname{supp}(\beta^{k-1}) \subset \operatorname{supp}(\beta),\ \|\beta - \beta^{k-1}\|_0 \le 1.$$

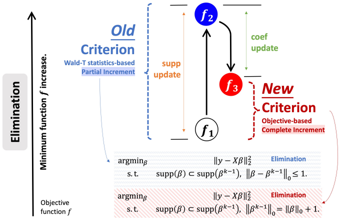
$$\operatorname{argmin}_\beta \quad \|y - X\beta\|_2^2 \qquad \text{Selection}$$
$$\text{s.t.} \quad \operatorname{supp}(\beta^{k-1}) \subset \operatorname{supp}(\beta),\ \|\beta\|_0 = \|\beta^{k-1}\|_0 + 1.$$

A2. The proposed new objective-based criterion corresponds to the complete decrement of the objective function



B. SUBSET ELIMINATION

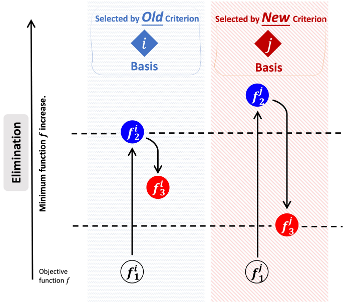
B1. The old Wald-T statistics-based criterion corresponds to the partial increment of the objective function



$$\operatorname{argmin}_\beta \quad \|y - X\beta\|_2^2 \qquad \text{Elimination}$$
$$\text{s.t.} \quad \operatorname{supp}(\beta) \subset \operatorname{supp}(\beta^{k-1}),\ \|\beta - \beta^{k-1}\|_0 \le 1.$$

$$\operatorname{argmin}_\beta \quad \|y - X\beta\|_2^2 \qquad \text{Elimination}$$
$$\text{s.t.} \quad \operatorname{supp}(\beta) \subset \operatorname{supp}(\beta^{k-1}),\ \|\beta^{k-1}\|_0 = \|\beta\|_0 + 1.$$

B2. The proposed new objective-based criterion corresponds to the complete increment of the objective function



Optimal Selection and Elimination Problem

As analyzed above, to obtain the optimal solution at each step, we consider the following optimization problems:

$$\begin{aligned} \underset{\beta}{\operatorname{argmin}} \quad & \|\mathbf{y}-\mathbf{X}\beta\|_2^2 \\ \text{s. t.} \quad & \|\beta\|_0 = \|\beta^{k-1}\|_0 + 1,\ \operatorname{supp}(\beta^{k-1}) \subset \operatorname{supp}(\beta), \end{aligned} \tag{P1}$$

$$\begin{aligned} \underset{\beta}{\operatorname{argmin}} \quad & \|\mathbf{y}-\mathbf{X}\beta\|_2^2 \\ \text{s. t.} \quad & \|\beta\|_0 = \|\beta^{k-1}\|_0 - 1,\ \operatorname{supp}(\beta) \subset \operatorname{supp}(\beta^{k-1}), \end{aligned} \tag{Q1}$$

where (P1) and (Q1) correspond to selection and elimination subproblems for each step exactly.

- Unlike the constraint in (P0), the constraint in (P1) does not require the coefficients fixed on the remaining support set.
- The new optimization problem consider a complete descent (including both step 1&2), providing the optimal criterion.

Optimal Selection and Elimination Criteria

Optimal Selection Criterion

Theorem 1

Problem (P1) *is equivalent (in the sense of identifying the true subset) to:*

$$\operatorname*{argmax}_{j \in S_{k-1}^c} \frac{\left(\mathbf{r}^{k^T} \mathbf{X}_j\right)^2}{\mathbf{X}_j^T \left(\mathbf{I} - \mathbf{X}_{S_{k-1}} \left(\mathbf{X}_{S_{k-1}}^T \mathbf{X}_{S_{k-1}}\right)^{-1} \mathbf{X}_{S_{k-1}}^T\right) \mathbf{X}_j}, \tag{5}$$

where $S_{k-1} = \operatorname{supp}(\beta^{k-1})$.

Definition 2 (**Objective-based Selection**)

By Theorem 1, the new criterion for feature importance outside the support set could be formulated as criterion (5).

Optimal Selection and Elimination Criteria

Optimal Elimination Criterion

Theorem 3

Let $\mathbf{C}_{k-1} = \left(\mathbf{X}_{S_{k-1}}^T \mathbf{X}_{S_{k-1}}\right)^{-1}$, $\mathbf{e}_j = (\delta_{1i}, \delta_{2i}, \cdots, \delta_{ii}, \cdots, \delta_{|S_{k-1}|i})^T \in \mathbb{R}^{|S_{k-1}|}$, *where j represents the i-th element of S_{k-1} for $i = 1, 2, \ldots, |S_{k-1}|$. The Kronecker delta function δ_{ab} is defined as $\delta_{ab} = 1$ if $a = b$, and $\delta_{ab} = 0$ otherwise. Then, problem* (Q1) *is equivalent (in the sense of identifying the true subset) to*

$$\underset{j \in S_{k-1}}{\operatorname{argmax}} \ \mathbf{y}^T \mathbf{X}_{S_{k-1}} \left(\mathbf{I} - \mathbf{e}_j \mathbf{e}_j^T\right) \left(\mathbf{C}_{k-1} - \frac{\mathbf{C}_{k-1} \mathbf{e}_j \mathbf{e}_j^T \mathbf{C}_{k-1}}{\mathbf{e}_j^T \mathbf{C}_{k-1} \mathbf{e}_j}\right) \left(\mathbf{I} - \mathbf{e}_j \mathbf{e}_j^T\right) \mathbf{X}_{S_{k-1}}^T \mathbf{y}. \tag{6}$$

Definition 4 (**Objective-based Elimination**)

By Theorem 3, the new criterion for feature importance inside the support set could be formulated as criterion (6).

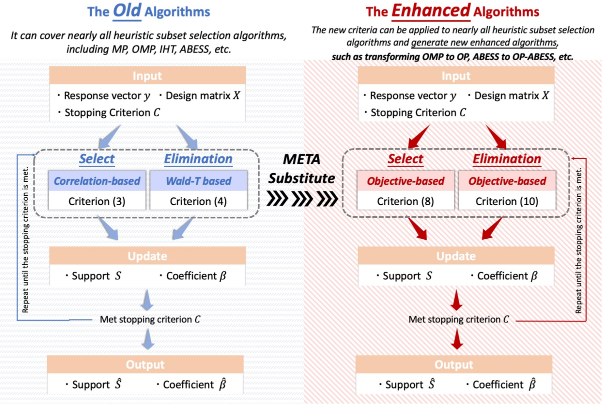
Optimal Selection and Elimination Criteria

A comprehensive derivation and discussion of the proposed criteria (5) and (6) can be found in Section 3 of our paper [Zhu et al., 2025].

- **Degeneration to classic criteria in easy case:** When the features are orthogonal, proposed criteria (5) and (6) degenerate to the classical correlation-based criterion (3) and the Wald-T based criterion (4), respectively.
- **Comprehensive combination effect:** Criteria (5) and (6) take into account the interaction between features, i.e., the impact on the other features in S_{k-1} resulting from the selection and elimination of feature j.
- **Identical computational efficiency:** The matrix $\mathbf{C}_{k-1}$ or Cholesky decomposition of $\mathbf{X}_{S_{k-1}}^T \mathbf{X}_{S_{k-1}}$ has been already computed in the previous step when updating the coefficients on the support set S_{k-1}, so the inversion term in criteria (5) and (6) does not incur additional magnitude of computational cost.

Enhanced Algorithms for Best Subset Selection

By leveraging the optimal criteria above, we can perform **Meta-Substitution** of the objective-based criteria (5) and (6) into classical algorithms like MP, OMP, CoSaMP, IHT, and (A)BESS, resulting in an enhanced algorithm family.


The Old Algorithms
It can cover nearly all heuristic subset selection algorithms, including MP, OMP, IHT, ABESS, etc.
Input
· Response vector y · Design matrix X
· Stopping Criterion C
Select
Elimination
Correlation-based
Criterion (3)
Wald-T based
Criterion (4)
Update
· Support S · Coefficient β
Met stopping criterion C
Repeat until the stopping criterion is met.
Output
· Support $\hat{S}$ · Coefficient $\hat{\beta}$
META Substitute
The Enhanced Algorithms
The new criteria can be applied to nearly all heuristic subset selection algorithms and generate new enhanced algorithms,
such as transforming OMP to OP, ABESS to OP-ABESS, etc.
Input
· Response vector y · Design matrix X
· Stopping Criterion C
Select
Elimination
Objective-based
Criterion (8)
Objective-based
Criterion (10)
Update
· Support S · Coefficient β
Met stopping criterion C
Repeat until the stopping criterion is met.
Output
· Support $\hat{S}$ · Coefficient $\hat{\beta}$

Enhanced Algorithms for Best Subset Selection

We classify subset selection algorithms into three categories based on their combination strategies for feature selection and elimination, providing **one representative for each** to show how Meta-Substitution generates new algorithms:

- **Select-Only**: This type of algorithm greedily selects feature at each step. Example: OMP $\rightarrow$ OP.
- **Select-First, Eliminate-Next**: This type of algorithms first selects the features and then removes the irrelevant ones. Example: CoSaMP $\rightarrow$ CoSaOP.
- **Exchange-Based**: This class of algorithms swaps irrelevant features in the active set with significant features outside the active set. Example: (A)BESS $\rightarrow$ OP-(A)BESS.

Beyond these examples, other greedy subset selection algorithms can also be enhanced through meta substitution scheme. These enhanced algorithms not only retain the original theoretical properties but also achieve significant meta-gains across various tasks, scenarios and evaluation metrics.

Theory of Optimal Subset Selection Criteria

Define the function $f(S)$ as:

$$f(S) \triangleq \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \quad \text{s.t.} \ \operatorname{supp}(\beta) = S.$$

With this definition, we present the following theorems.

Theorem 5

For index j^* selected by criterion (5),

$$f\left(S \cup \{j^*\}\right) \leq f\left(S \cup \{j\}\right), \ \ \forall j \in S^c.$$

Theorem 6

For index j^* selected by criterion (6),

$$f\left(S \backslash \{j^*\}\right) \leq f\left(S \backslash \{j\}\right), \ \ \forall j \in S.$$

Theory of Optimal Subset Selection Criteria

Theorems 5 and 6 summarize the previous discussion, demonstrating that criteria (5) and (6) serve as the optimal decisions in the current subset selection process.

Theorem 7

The computational complexities of OMP and OP, CoSaMP and CoSaOP, as well as (A)BESS and OP-(A)BESS, are of the same order of magnitude.

Theorem 7 indicates that the enhanced algorithms have the same computational complexity as the original algorithms.

We also show **how the enhanced algorithms retain the theoretical properties of the original algorithms**, see Section 4 in our paper [Zhu et al., 2025].

Theory of Optimal Subset Selection Criteria

Theorems 8 and 9 further demonstrate the significant advantages of criteria (5) and (6) in the presence of high feature correlation.

Theorem 8

Suppose the true subset S^* *contains indices* (i, j)*, where the correlation between feature* $\mathbf{X}_i$ *and* $\mathbf{X}_j$ *is* $\rho = \frac{|\mathbf{X}_i^T \mathbf{X}_j|}{||\mathbf{X}_i||_2 ||\mathbf{X}_j||_2}$*. Assuming the current support set* S *already includes feature i, then the classical correlation-based criterion* (3) *for feature j satisfies:*

$$\frac{|{\mathbf{r}^k}^T \mathbf{X}_j|}{||\mathbf{X}_j||_2} \leq \sqrt{1-\rho^2}||\mathbf{r}^k||_2, \tag{7}$$

while the objective-based criterion (5) *satisfies*

$$\frac{\left({\mathbf{r}^k}^T \mathbf{X}_j\right)^2}{\mathbf{X}_j^T \left(\mathbf{I} - \mathbf{X}_S \left(\mathbf{X}_S^T \mathbf{X}_S\right)^{-1} \mathbf{X}_S^T\right) \mathbf{X}_j} \geq \frac{1}{1-\rho^2} \left(\frac{{\mathbf{r}^k}^T \mathbf{X}_j}{||\mathbf{X}_j||_2}\right)^2. \tag{8}$$

Theory of Optimal Subset Selection Criteria

Theorem 9

(1) The upper bound of the objective-based criterion (6) *is* $||\mathbf{y}||_2^2$*. If the true subset* S^* *is contained within the current subset* S*, then for all* $j_m \in S \setminus S^*$*,*

$$||\mathbf{y}||_2^2 - ||\epsilon||_2^2 \leq (\textit{criterion}\,(6)\textit{ for } j_m) \leq ||\mathbf{y}||_2^2.$$

And in noiseless scenario,

$$j_m \in \mathrm{argmax}_{j \in S}\ \textit{objective-based criterion}\,(6).$$

(2) Suppose a feature $\mathbf{X}_p$ *in the current subset is pesudo-correlated with an important feature* $\mathbf{X}_i$ *in the true subset (with correlation* $1 - \mu$*). When* μ *is sufficiently small, classical T-statistics based criteria* (4) *could erroneously discard true features even in simple cases like* $S = S^* \cup \{p\}$*, whereas proposed criterion* (6) *could correctly identify and remove the spurious feature* $\mathbf{X}_p$*.*

Experiment

The comparison involves representative algorithms from the three categories: OP, CoSaOP, and OP-(A)BESS, evaluated against classical subset selection methods: OMP, CoSaMP, and (A)BESS, highlighting the superiority of the enhanced algorithms from various perspectives.

Task: Compressed sensing and sparse regression.
Scenario:

- Measurement rate
- Noise level (SNR)
- Number of features

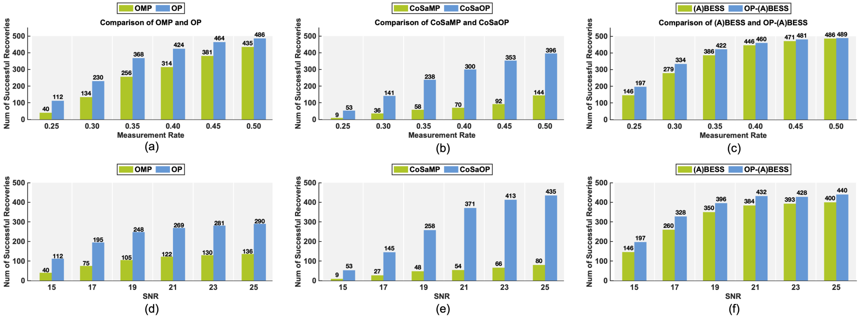
Evaluation metrics: The number of successful recoveries, NMSE, R^2 , and runtime.

Experiment 1: Compressed Sensing (Synthetic Sparse Data)

In this experiment, we randomly generate β with a dimensionality of $p = 200$ and a sparsity level of $K = 10$. The design matrix $\mathbf{X}$ is a $n \times p$ random Gaussian matrix. Let S^* represent the true support set of the signal and $\hat{S}$ the estimated support set, **recovery is deemed successful if $\hat{S} = S^*$**. For each algorithm, we conduct 500 independent runs and record the number of successful recoveries.

Multiple tests are conducted as the sampling rate increases from 25% to 50% and SNR increases from 15 to 25.

Experiment 1: Compressed Sensing (Synthetic Sparse Data)



(a) (b) (c)

(d) (e) (f)

Figure: Meta-gain comparison of three kinds of subset selection algorithms. Row one: different sampling rates (SNR = 15). Row two: Varying SNRs (measurement rate = 0.25).

Experiment 1: Compressed Sensing (Synthetic Sparse Data)

In this experiment, we conduct additional comparisons of the algorithm under extreme scenarios:

- **Small-sample rate and high-dimensional vectors:** $p = 2000$, with n/p varying from 0.05 to 0.1.
- **High noise:** SNR varies from 5 to 15.
- **Highly correlated features (RIP violated):** The covariance matrix of the row vectors of $\mathbf{X}$ follows a Toeplitz structure, where the correlation between position i, j is $corr_{ij} = \rho^{|i-j|}$, with $\rho = 0.7$.

Experiment 1: Compressed Sensing (Synthetic Sparse Data)

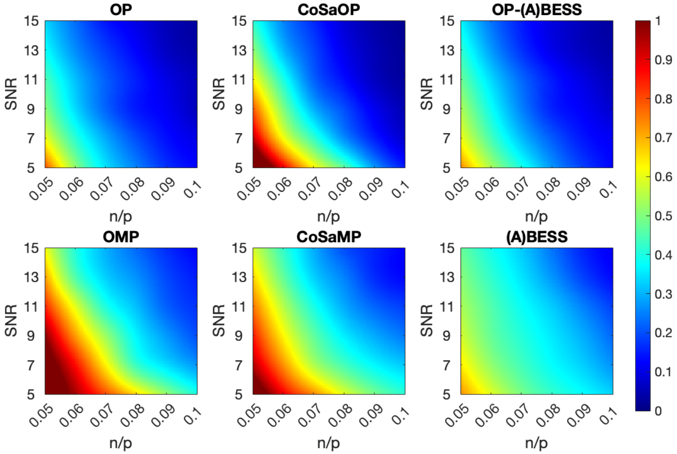


Figure: Phase transition with correlated features.

Experiment 2: Compressed Sensing (Audio Data)

Audio signal exhibits transform sparsity in the DCT domain[Donoho, 2006]. Therefore, we test our method using real-world audio data. The data presented here is randomly sampled from the *AudioSet* dataset [Gemmeke et al., 2017], consisting of 4 audio signals with full dimensionality $p = 480$. These signals have approximately 20–40 non-zero entries (K) in the DCT domain, with the number of observations fixed at $n = 150$.

The normalized mean squared error (NMSE), defined as $\text{NMSE} = \|\hat{\beta} - \beta^*\|_2^2 / \|\beta^*\|_2^2$, is used to quantify the recovery performance. We conducted 100 random experiments, and the results are summarized in Table 1.

Experiment 2: Compressed Sensing (Audio Data)

Table: Reconstruction NMSE (mean ± std) for Signals from *AudioSet*. Meta-gains are highlighted in **red**, with the best in **bold** and the second-best underlined. (Audio 1-4: -0SdAVK79lg.wav, _qxgIqI0uA.wav, _0bN5mYLXb0.wav, _0Jd6JJeyJ4.wav)

	NMSE (OMP & OP)			**NMSE (CoSaMP & CoSaOP)**			**NMSE ((A)BESS & OP-(A)BESS)**		
Audio Set	**OMP**	**OP**	**Gains**	**CoSaMP**	**CoSaOP**	**Gains**	**(A)BESS**	**OP-(A)BESS**	**Gains**
Audio 1	9.45E-04 (1.35E-03)	7.69E-04 (5.78E-04)	**19%**	2.36E-03 (1.22E-03)	1.64E-03 (1.36E-03)	**31%**	1.41E-03 (7.89E-03)	**6.22E-04** (**2.64E-04**)	**56%**
Audio 2	5.79E-03 (7.89E-03)	5.49E-03 (8.87E-03)	**5%**	1.87E-03 (6.05E-04)	**6.36E-04** (**2.01E-04**)	**66%**	6.71E-03 (6.07E-02)	**6.36E-04** (**2.01E-04**)	**91%**
Audio 3	1.46E-03 (3.13E-03)	1.18E-03 (2.14E-03)	**19%**	1.54E-03 (5.86E-04)	**5.52E-04** (**2.03E-04**)	**64%**	2.84E-03 (2.29E-02)	**5.52E-04** (**2.03E-04**)	**81%**
Audio 4	4.05E-02 (1.03E-01)	3.55E-02 (9.70E-02)	**12%**	2.85E-03 (9.84E-04)	**8.22E-04** (**2.22E-04**)	**71%**	2.65E-02 (1.07E-01)	1.86E-02 (1.12E-01)	**30%**

Experiment 3: Sparse Regression

In sparse regression tasks, β does not have a ground truth. The goal is to select sparse features that provide a better explanation of target variable **y**. Therefore, similar to the metric used in sparse regression evaluations in [Qian et al., 2017, Das and Kempe, 2018], we **quantify the explanatory ability of the features using Coefficient of Determination:** $R^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2$.

We utilize six real-world datasets in our experiments:

(1) Boston Housing Data [Pedregosa et al., 2011],

(2) California Housing Data [Pedregosa et al., 2011],

(3) Superconductivity Data [Hamidieh, 2018],

(4) House 16H [Vanschoren, 2014],

(5) Prostate.v8.egen [Lin and Pan, 2024, Hastie et al., 2017],

(6) Spectra [The MathWorks, Inc., 2025].

Experiment 3: Sparse Regression

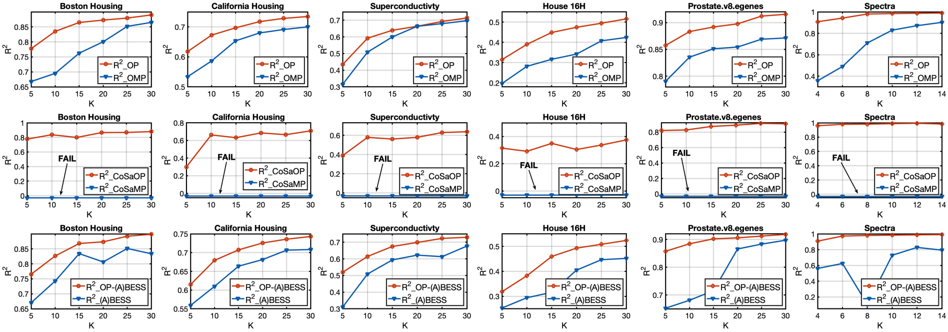


Figure: Rows 1–3 present the meta-gains in feature representation capability (R^2, closer to 1 is better) for the Boston Housing, California Housing, Superconductivity datasets, House 16H, Prostate.v8.egenes, and Spectra datasets, respectively, across three algorithms as the number of selected features K varies.

Best Subset Selection with Ultra-high Dimensions: Optimal Gradient Pursuit

In ultra-high-dimensional settings where solving the least squares problem over a given subset, i.e., solving a linear system, can be computationally prohibitive, we propose an acceleration scheme for Optimal Pursuit: Optimal Gradient Pursuit (OGP).

Comparison Between Gradient Pursuit and Optimal Gradient Pursuit

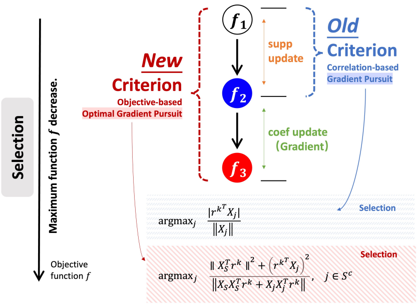

Selection
Maximum function f decrease.
Objective function f
New
Criterion
Objective-based
Optimal Gradient Pursuit
f_1
f_2
f_3
supp update
coef update (Gradient)
Old
Criterion
Correlation-based
Gradient Pursuit
Selection
$\operatorname{argmax}_j \frac{|r^{k^T} X_j|}{\|X_j\|}$
Selection
$\operatorname{argmax}_j \frac{\| X_S^T r^k \|^2 + \left(r^{k^T} X_j\right)^2}{\|X_S X_S^T r^k + X_j X_j^T r^k\|}, \quad j \in S^c$

Figure: Optimal Gradient Pursuit.

Unsupervised Learning: Column Subset Selection

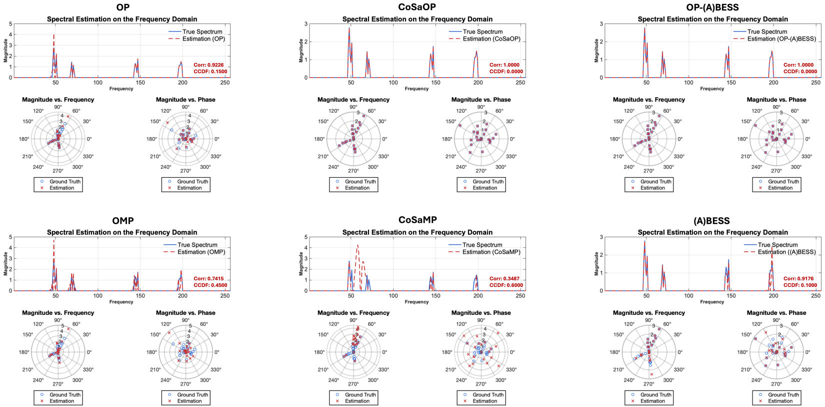
Column Subset Selection (CSS) and PCA are both important dimensionality reduction methods with widespread applications in unsupervised learning [Belhadji et al., 2020]. The goal of CSS is to select a subset of important columns (features) from a dataset that can better represent the entire dataset, formulated as:

$$\begin{aligned} \min_{S,\mathbf{B}} \ & \|\mathbf{X} - \mathbf{X}_S\mathbf{B}\|_F^2 \\ \text{s. t.} \ & |S| \leq K. \end{aligned}$$

In fact, this problem can also be viewed as a special case of best subset selection problem. We have extended the optimal pursuit criterion to the CSS task, demonstrating the advantages of our proposed criteria over classical criteria in this setting.

Complex Signal Processing

Although our paper primarily discusses these theories and methods in the real domain, they can be directly extended to the complex domain. A classic example in complex signal processing is line spectrum estimation, widely applied in modern wireless communications.



Conclusion

- By revisiting classical criteria in traditional algorithms through the lens of block coordinate descent, we revealed that they only **reflect a one-step variation of the objective function**.
- Building on this, we formulated exact optimization subproblems for feature selection and elimination.
- We derive **explicit solutions using forward and backward matrix inversion**.
- The proposed criteria account for **both individual feature significance and interactions**, proving optimal for subset selection.
- Replacing classical criteria with the proposed ones, we developed **enhanced algorithms** that **retain the original theoretical guarantees while achieving significant performance gains** across various tasks, scenarios and evaluation metrics, **all without added computational cost**.

Future Work

The results affirm the advantages of the new criteria both theoretically and practically, opening new avenues for improving best subset selection algorithms. Future work may consider:

- integrating the proposed criteria into arbitrary greedy subset selection algorithms to develop enhanced methods and application on structured sparse learning [Huang et al., 2009],
- developing optimal selection and elimination criteria for general objective functions,
- investigating statistical inference theories of the new criteria.

References I

[Belhadji et al., 2020] Belhadji, A., Bardenet, R., and Chainais, P. (2020).
A determinantal point process for column subset selection.
Journal of Machine Learning Research, 21(197):1–62.

[Blumensath and Davies, 2009] Blumensath, T. and Davies, M. E. (2009).
Iterative hard thresholding for compressed sensing.
Applied and Computational Harmonic Analysis, 27(3):265–274.

[Chen et al., 2001] Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001).
Atomic decomposition by basis pursuit.
SIAM review, 43(1):129–159.

[Das and Kempe, 2011] Das, A. and Kempe, D. (2011).
Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection.
In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 1057–1064.

[Das and Kempe, 2018] Das, A. and Kempe, D. (2018).
Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection.
Journal of Machine Learning Research, 19(3):1–34.

[Davis et al., 1997] Davis, G., Mallat, S., and Avellaneda, M. (1997).
Adaptive greedy approximations.
Constructive Approximation, 13:57–98.

[Donoho, 2006] Donoho, D. L. (2006).
Compressed sensing.
IEEE Transactions on information theory, 52(4):1289–1306.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004).
Least angle regression.
The Annals of Statistics, 32(2):407–451.

[Fan and Li, 2001] Fan, J. and Li, R. (2001).
Variable selection via nonconcave penalized likelihood and its oracle properties.
Journal of the American statistical Association, 96(456):1348–1360.

References II

[Feige, 1998] Feige, U. (1998).
A threshold of ln n for approximating set cover.
Journal of the ACM (JACM), 45(4):634–652.

[Foucart, 2011] Foucart, S. (2011).
Hard thresholding pursuit: An algorithm for compressive sensing.
SIAM Journal on Numerical Analysis, 49(6):2543–2563.

[Gemmeke et al., 2017] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017).
Audio set: An ontology and human-labeled dataset for audio events.
In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

[Hamidieh, 2018] Hamidieh, K. (2018).
Superconductivty Data.
UCI Machine Learning Repository.
DOI: https://doi.org/10.24432/C53P47.

[Hastie et al., 2017] Hastie, T., Tibshirani, R., and Friedman, J. (2017).
The elements of statistical learning: data mining, inference, and prediction.

[Hazimeh and Mazumder, 2020] Hazimeh, H. and Mazumder, R. (2020).
Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms.
Operations Research, 68(5):1517–1537.

[Huang et al., 2009] Huang, J., Zhang, T., and Metaxas, D. (2009).
Learning with structured sparsity.
In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 417–424.

[Kohavi and John, 1997] Kohavi, R. and John, G. H. (1997).
Wrappers for feature subset selection.
Artificial Intelligence, 97(1-2):273–324.

[Lin and Pan, 2024] Lin, Z. and Pan, W. (2024).
A robust cis-mendelian randomization method with application to drug target discovery.
Nature Communications, 15(1):6072.

References III

[Mallat and Zhang, 1993] Mallat, S. G. and Zhang, Z. (1993).
Matching pursuits with time-frequency dictionaries.
IEEE Transactions on Signal Processing, 41(12):3397–3415.

[Miller, 2002] Miller, A. (2002).
Subset selection in regression.
Chapman and Hall/CRC.

[Needell and Tropp, 2009] Needell, D. and Tropp, J. (2009).
Cosamp: Iterative signal recovery from incomplete and inaccurate samples.
Applied and Computational Harmonic Analysis, 26(3):301–321.

[Pati et al., 1993] Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. S. (1993).
Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition.
In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44. IEEE.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011).
Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830.

[Qian et al., 2017] Qian, C., Shi, J.-C., Yu, Y., Tang, K., and Zhou, Z.-H. (2017).
Subset selection under noise.
Advances in Neural Information Processing Systems, 30.

[The MathWorks, Inc., 2025] The MathWorks, Inc. (2025).
Statistics and Machine Learning Toolbox.

[Tibshirani, 1996] Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1):267–288.

[Vanschoren, 2014] Vanschoren, J. (2014).
OpenML Dataset 574: House 16H.

References IV

[Wang et al., 2024] Wang, J., Zhang, B., Du, Q., Zhang, J., and Chu, D. (2024).
A survey on data selection for LLM instruction tuning.
arXiv preprint arXiv:2402.05123.

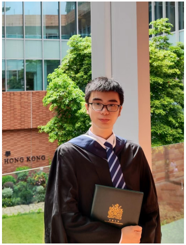
[Zhang, 2010] Zhang, C.-H. (2010).
Nearly unbiased variable selection under minimax concave penalty.
The Annals of Statistics, 38(2):894 – 942.

[Zhu et al., 2020] Zhu, J., Wen, C., Zhu, J., Zhang, H., and Wang, X. (2020).
A polynomial algorithm for best-subset selection problem.
Proceedings of the National Academy of Sciences, 117(52):33117–33123.

[Zhu et al., 2025] Zhu, Z., Zhang, Y., and Xia, Y. (2025).
Best subset selection: Optimal pursuit for feature selection and elimination.
Proceedings of the 42nd International Conference on Machine Learning.

[Zou, 2006] Zou, H. (2006).
The adaptive lasso and its oracle properties.
Journal of the American Statistical Association, 101(476):1418–1429.

Thanks to co-authors



Zhu, Zhihan
Beihang University

Zhang, Yanhao
Beihang University

Prof. Xia, Yong
Beihang University

Thanks for Your Attention!

Code

Paper