CS539 Convex Optimization - Final Project: Sparse Modeling for Finding Representative Objects

Woonki Kim
Department of Computer Science
Oregon State University
Corvallis, OR 97331
kimwoon@oregonstate.edu

Abstract

In this paper, we explore insights gained from the paper "See All by Looking at a Few: Sparse Modeling for Finding Representative Objects" and examine how its method identifies representative data points from an original dataset. We further formulate the SMRS (Sparse Modeling Representative Selection) algorithm using the ADMM (Alternating Direction Method of Multipliers) framework. Through both synthetic and practical experiments, we validate the effectiveness of this approach and investigate its impact on classification performance.

1 Introduction

The importance of this technique lies in its ability to efficiently reduce the complexity of large datasets by identifying a subset of representative objects. In fields like machine learning, computer vision, and data analysis, datasets often contain large amounts of high-dimensional data, making it challenging to process and interpret them effectively. By finding a small number of representative points, this method saves memory and computational resources and also enhances the performance of algorithms for tasks like classification and clustering. Furthermore, it enables better understanding of the data's underlying structure without relying on assumptions like low-rank or cluster-based distributions, which are common limitations of traditional approaches. This flexibility makes the technique widely applicable to various applications, such as summarizing videos and images[Elhamifar et al., 2012].

1.1 Self-Representation

Previously, there has been a study focused on approximating a dataset $Y \in \mathbb{R}^{m\times n}$ through the use of a dictionary $D \in \mathbb{R}^{m\times k}$ and a sparse coefficient matrix $X \in \mathbb{R}^{k\times n}$, aiming to minimize the reconstruction error $\|Y - DX\|_F^2$ while enforcing a sparsity constraint on X[Ramirez et al., 2010, Elhamifar et al., 2012].

Mathematically, this is expressed as:

$$\min_{D,X} \|Y - DX\|_F^2 \quad \text{subject to} \quad \|x_i\|_0 \leq s,\ \|d_j\|_2 \leq 1,$$

However, dictionary D is not restricted to elements of the original dataset Y, leading to representations that may not directly correspond to original data points. To address this limitation from Dictionary method, Elhamifar et al., introduces self-representation framework. Instead of optimizing over an arbitrary dictionary D, this approach seeks to represent Y using its own data points. By introducing a sparse coefficient matrix C, the dataset is approximated as YC, where C holds the relationships among data points. The optimization problem is then reformulated as:

$$\min_{C} \|Y - YC\|_F^2 \quad \text{subject to} \quad \|C\|_{row,0} \leq k,\ 1^\top C = 1^\top.$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Here, the constraint $\|C\|_{row,0}$ enforces row-wise sparsity, ensuring that each data point is represented is limited to k, while $1^\top C = 1^\top$ ensures affine constraints for invariance under translation. This formulation enables direct identification of representative data points from the dataset, improving the drawback of arbitrary dictionary selection[Elhamifar et al., 2012].

For instance, given a dataset Y with 1000 data points, this approach might identify a subset of only 50 representative points. These selected points preserve the dataset's structure, ensuring efficient data summarization and dimensionality reduction.

1.2 Advantages of Sparse Self-Representation Algorithm

This technique addresses several challenges in representing large and complex datasets with reduced dataset, providing an efficient and flexible solution for various applications. Unlike traditional methods such as K-medoids or Affinity Propagation, which rely on assumptions like low-rank structure or clustering, this approach can handle datasets without requiring them to form a specific geometric or statistical models. Moreover, the method is particularly effective in detecting and rejecting outliers by analyzing their contribution to the dataset's reconstruction, ensuring that only meaningful points are retained. [Elhamifar et al., 2012].

2 Method

Elhamifar et al., aims to find representative points from original dataset using the characteristics of Frobenius norm and gain sparsity by ℓ_0 norm, with affinity using affine constraints.

2.1 Frobenius norm

Self-representation algorithm ensures minimal reconstruction error through the Frobenius norm $\|Y - YC\|_F^2$ while selecting representative points that correspond directly to the dataset. The Frobenius norm is a measure of the size or magnitude of a matrix, similar to the Euclidean norm for vectors. For $m \times n$ matrix S, with elements s_{ij}, the Frobenius norm $\|S\|_F$, is defined as:

$$\|S\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|s_{ij}|^2}.$$

This formula calculates the square root of the sum of the absolute squares of all entries in the matrix, effectively treating the matrix elements as components of a vector. By aggregating the contributions of all individual elements in a matrix, the Frobenius norm provides a single value that reflects the overall weight of the matrix.

In applications involving approximation, such as $\|X - \hat{X}\|_F^2$, the Frobenius norm measures the overall difference between matrices X and $\hat{X}$. It evaluates the sum of squared differences between corresponding elements, capturing the approximation error across all elements. Minimizing this norm ensures accurate and reliable reconstructions, and since it is differentiable making it particularly effective on approximations in optimization problems. [Golub and Van Loan, 2013] [Herzog et al., 2023]

2.2 Sparsity Constraints

Sparse constraint ensures the selection of a minimal but still informative subset of representatives. Elhamifar et al., has ensured sparsity using ℓ_0-norm.

The ℓ_0-norm counts the number of nonzero entries:

$$\|x_i\|_0 = \sum_{j=1}^{n} I(x_{ij} \neq 0),$$

where $I(\cdot)$ is the indicator function. To encourage sparsity at each row, this concept is extended to the rows of the coefficient matrix C, leading to the mixed ℓ_0/ℓ_q-norm:

$$\|C\|_{0,q} = \sum_{i=1}^{N} I(\|c_i\|_q > 0),$$

where c_i is the i-th row of C, and $\|c_i\|_q$ measures the contribution of the i-th row. The constraint $\|C\|_{0,q} \leq k$ ensures that only k rows of C are nonzero, selecting k representatives from the dataset. [Wang et al., 2013]

Row-wise sparsity plays an important role since it makes representations easier to understand and more efficient. By limiting the number of active rows, it prevents overly complex combinations, ensuring that relationships among data points are sparse in structure of C.

Moreover, the row-sparsity constraint is also a powerful mechanism for managing outliers. The sparsity constraint limits the number of nonzero rows in C, ensuring that most rows remain sparse. Outliers, due to their incoherence with the main dataset, naturally form rows with very few nonzero entries, reflecting their limited connection to the dataset. In contrast, inliers, being integral to the dataset, are represented by combinations of other inliers. As a result, their rows in C usually have more nonzero entries, indicating the use of multiple representatives. [Elhamifar et al., 2012].

2.3 Affine Constraint

The affine constraint $1^\top C = 1^\top$ ensures affine invariance in the selection of representatives, maintaining consistency in data representation even when global translations has been done. When the dataset undergoes a shift, such as $y_i' = y_i - T$, this constraint ensures that the representation $y_i = Yc_i$ remains valid for the translated data by enforcing that the coefficients in each row of C sum to one. As a result, the structure of the dataset is preserved rather than being influenced by arbitrary translations. [Li et al., 2018]

Furthermore, this constraint ensures that each data point is represented as an affine combination of other points, which discourages outliers. Inliers, being consistent with the dataset, naturally select other inliers as representatives, creating meaningful combinations under the affine constraint. Outliers tend to represent themselves, leading to distinct structures in their corresponding rows of C. This distinction highlights outliers while preserving the internal structure of the dataset. [Elhamifar et al., 2012].

3 Author's Contribution

Elhamifar et al., has reformulated the problem of SMRS to bridge theoretical foundations with practical use cases, simplifying the challenge.

3.1 NP-Hardness and Relaxation

The initial problem seeks to find a representative points and mark it in C from Y that are actual data points, ensuring that the reconstruction error is minimized:

$$\min \|Y - YC\|_F^2 \quad \text{s.t.} \quad \|C\|_{0,q} \leq k, \quad 1^\top C = 1^\top.$$

The problem is inherently NP-hard because this formulation searches over all possible subsets of k representatives using ℓ_0 norm, making the tasks to grow exponentially as the dataset size increases. To make the problem computationally feasible, a common approach is to replace the ℓ_0-norm constraint, with a convex relaxation using the ℓ_1-norm. This relaxation still has the ability to maintain sparsity while significantly reducing computational complexity and making the constrain a convex, as the ℓ_1-norm is both convex and efficiently solvable. Thus rewriting the formula as:

$$\min \|Y - YC\|_F^2 \quad \text{s.t.} \quad \|C\|_{1,q} \leq \tau, \quad 1^\top C = 1^\top.$$

In this version, instead of using the ℓ_0-norm, which counts the exact number of non-zero values in C, Elhamifar et al., use the mixed ℓ_1-norm $\|C\|_{1,q}$. The parameter $\tau > 0$ allows us to adjust the trade-off between accurately reconstructing Y and keeping C sparse since ℓ_1-norm does not guarantee sparsity with zero. A smaller τ leads to more sparsity, while a larger τ allows for less sparsity and more flexibility in C.

This relaxed version of the problem replaces the non-convex and NP-Hardness of ℓ_0-norm with the easier ℓ_1-norm. This makes the problem more practical to solve while still achieving a good balance between approximation accuracy and flexibility. [Ramirez et al., 2013, Feng et al., 2018].

3.2 Reformulating as a Compression Problem

The problem of identifying representative points from Y can also be reformulated as a compression problem, shifting the focus from fixing the number of representatives k to minimizing the number of representatives, while maintaining a reconstruction error within a specified threshold ϵ. This leads to the formulation:

$$\min \|C\|_{1,q} \quad \text{s.t.} \quad \|Y - YC\|_F^2 \leq \epsilon, \quad 1^\top C = 1^\top.$$

This approach automatically adjusts the number of representatives based on the required level of reconstruction accuracy, rather than forcing engineers to choose a fixed k number of representatives. It focuses on meeting the quality of the reconstruction while minimizing the number of representatives, making it more adaptable to the data and less reliant on arbitrary choices. [Elhamifar et al., 2012]

4 Algorithm

The final algorithm is formulated as follows:

$$\min_C \|C\|_{1,q}, \quad \text{subject to } \|Y - YC\|_F \leq \epsilon, \ \mathbf{1}^T C = \mathbf{1}^T.$$

To solve this optimization problem, we employed the ADMM method. ADMM is particularly well-suited for this problem because it efficiently manages the combination of the non-smooth term $\|C\|_{1,q}$ and the convex constraints. [Hong and Luo, 2013].

4.1 Solving Optimization Problem with ADMM

To implement this using ADMM, we introduce splitting variables $C_1 = C$ and $C_2 = C$, separating the norm term $\|C_1\|_{1,q}$ from the constraints $\|Y - YC_2\|_F \leq \epsilon$ and $\mathbf{1}^\top C_2 = \mathbf{1}^\top$. This reformulates the problem as:

$$\min_{C,C_1,C_2} \|C_1\|_{1,q}, \quad \text{subject to } \|Y - YC_2\|_F \leq \epsilon, \ \mathbf{1}^\top C_2 = \mathbf{1}^\top, \ C = C_1 = C_2.$$

The augmented Lagrangian for this problem is:

$$\mathcal{L}(C, C_1, C_2, \Lambda_1, \Lambda_2) = \|C_1\|_{1,q} + \frac{\rho}{2}\|C - C_1 + \Lambda_1\|_F^2 + \frac{\rho}{2}\|C - C_2 + \Lambda_2\|_F^2,$$

where Λ_1 and Λ_2 are scaled dual variables associated with the constraints $C = C_1$ and $C = C_2$, respectively, and $\rho > 0$ is a penalty parameter. This ensures convergence to a solution that minimizes $\|C\|_{1,q}$, satisfies $\|Y - YC\|_F \leq \epsilon$, and $\mathbf{1}^T C = \mathbf{1}^T$.

Algorithm 1 ADMM for SMRS

Require:

Data matrix Y
Tolerance $\epsilon > 0$
Penalty parameter $\rho > 0$
Maximum iterations maxIter
Initialize : $C^{(0)}, C_1^{(0)}, C_2^{(0)}, \Lambda_1^{(0)}, \Lambda_2^{(0)}$
$k = 0$
while not converged and $k <$ maxIter **do**
    Update C_1 (Proximal operator for $\|\cdot\|_{1,q}$):

$$C_1^{(k+1)} = \text{Prox}_{\|\cdot\|_{1,q}/\rho}\left(C^{(k)} + \Lambda_1^{(k)}\right)$$

    Update C_2 (Projection onto constraints)

$$C_2^{(k+1)} = \text{Proj}_{\|Y - YC_2\|_F \leq \epsilon,\ \mathbf{1}^\top C_2 = \mathbf{1}^\top}\left(C^{(k)} + \Lambda_2^{(k)}\right)$$

    Update C:

$$C^{(k+1)} = \frac{1}{2}\left(C_1^{(k+1)} - \Lambda_1^{(k)} + C_2^{(k+1)} - \Lambda_2^{(k)}\right)$$

    Update Lagrange multipliers

$$\Lambda_1^{(k+1)} = \Lambda_1^{(k)} + C^{(k+1)} - C_1^{(k+1)}$$

$$\Lambda_2^{(k+1)} = \Lambda_2^{(k)} + C^{(k+1)} - C_2^{(k+1)}$$

    Convergence check:
    if $\|C^{(k+1)} - C_1^{(k+1)}\|_F \leq \epsilon$ and $\|C^{(k+1)} - C_2^{(k+1)}\|_F \leq \epsilon$ **then**
        break
    end if
    $k \leftarrow k + 1$
end while

4.2 Finding representative data

Now we have found sparse C that constructs Y most likely. We need to extract the most representing indices from C, which is the highest value in C. The main idea is that rows with higher norms carry more weight, making them better candidates for representing the dataset. Finding representative index using C is:

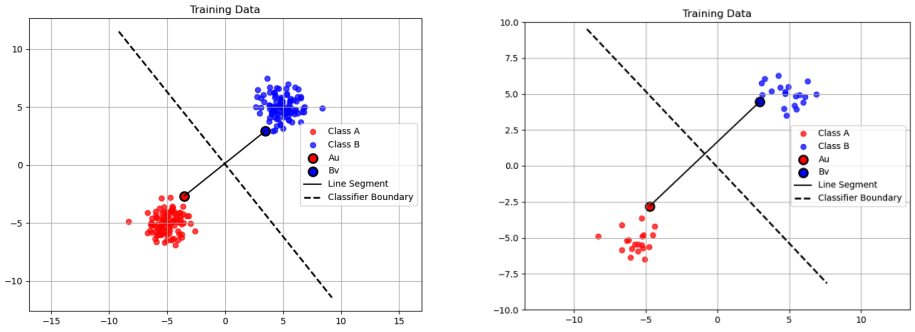
$$row[i] = \|C[i,:]\|_2$$

The formula $row[i] = \|C[i,:]\|_2$ helps find representative indices from C by applying the norm of each row $C[i,:]$ to quantify its significance in constructing Y. Rows with higher norms indicate greater overall magnitude of their entries, which implies that they contribute more to the representation of Y. Since C is sparse, most rows contain small or zero values, but rows with large norms stand out as key contributors. By identifying and selecting the indices i with the highest $row[i]$ values, we focus on the most informative rows, which effectively represent the dataset Y.

5 Experiment with Synthetic Data

To evaluate the effectiveness of representative data points in classification tasks, we conducted experiments on two synthetic datasets: one with separable two convex hulls and another with overlapping two convex hulls each convex hulls with 100 points. Using Python and the cvx library, we employed Support Vector Machines (SVMs) to classify the datasets. Our objective was to determine whether representative data points could efficiently encapsulate the characteristics of the entire dataset while maintaining a similar classification error rate to models trained on the full dataset.

5.1 Separable Case

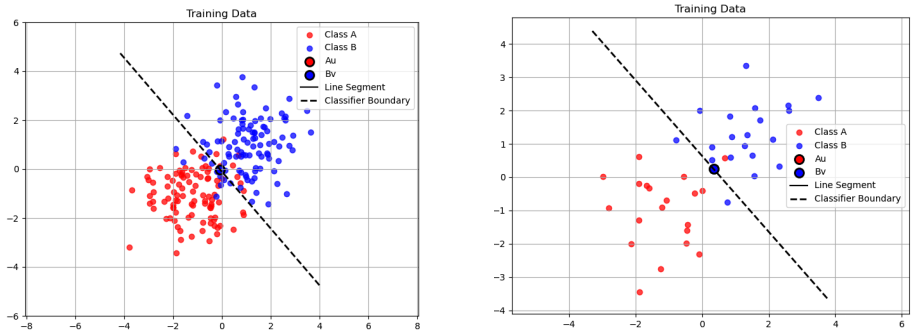


(a) Original Dataset (b) Representative Dataset

Figure 1: Comparison between classifiers trained on the original dataset and a representative dataset constructed from separable convex hulls.

The results demonstrated that these representative points successfully outlined the decision boundary of the complete dataset. Training a classifier on the representative data achieved a classification error rate of 0.00, identical to the performance of a model trained on the entire dataset. This outcome is unsurprising, as the representatives are selected from separated convex hulls, inherently ensuring perfect separability. However, the more significant result lies in the shape and structure retained by the representatives. These points effectively capture the geometric features of the dataset, preserving its essential characteristics while reducing redundancy.

5.2 Overlapping Case



(a) First image caption. (b) Second image caption.

Figure 2: Comparison between classifiers trained on the original dataset and a representative dataset constructed from overlapping convex hulls.

For the overlapping case, where the convex hulls of two classes intersect, a common approach is to reduce the convex hulls so that they do not overlap [Bennett and Bredensteiner, 2000]. We hypothesized that, representative points picked from SMRS algorithm would yield similar results to directly reducing the convex hulls by limiting u and v with an upper bound d, as expressed in the following formulation:

$$f(u,v) = \frac{1}{2}\|Au - Bv\|_2^2, \quad \text{subject to } 1^\top u = 1, \quad d \geq u \geq 0, \quad 1^\top v = 1, \quad d \geq v \geq 0.$$

This hypothesis was based on the assumption that the SMRS algorithm's ability to reject outliers would simplify the data representation, effectively managing the overlap without the need for reducing the convex hulls. Assuming that simplification reduces redundancy and overlap by focusing on the

most representative points while preserving the characteristics of the data. However, contrary to our hypothesis, this approach resulted in a higher error rate compared to using the original convex hulls.

The increased error rate suggests that relying solely on the representative points selected by SMRS algorithm may fail to fully remove the intersecting data. Future studies will explore alternative methods for handling overlapping convex hull scenarios, including strategies for classifying data by selectively removing representatives to acheive structural simplification with classification performance.

6 Experiment with Practical Data

Following the validation of methodology using synthetic data, we extended our experiments to a real-world dataset to assess the practical applicability of the proposed approach. The experiments were conducted using the FashionMNIST dataset, a widely used benchmark for image classification tasks.

6.1 Experimental Setup

The FashionMNIST dataset consists of 60,000 training images and 10,000 testing images, each with dimensions 28×28. We preprocessed the dataset by flattening the images into vectors and applying Principal Component Analysis (PCA) to reduce dimensionality. This preprocessing step ensured efficient representation of the data and consistent scaling across all samples. Our primary objective was to evaluate the efficacy of the SMRS algorithm in identifying representative subsets and to measure the classification performance of a Support Vector Machine trained on these subsets.

The algorithm generated a set of representative indices, which we used to extract the corresponding samples and their labels from the original dataset. To assess the quality of the selected representatives, we visualized the representative images, grouped by their respective labels. Using these representative data points, identified by SMRS, we trained an SVM and evaluated its performance. This allowed us to quantify the effectiveness of the representative selection process and its impact on classification outcomes.

6.2 Result

Figure 3: Representatives of FashionMNIST datset Label1 found by SMRS algorithm.

The results of the experiment are summarized in Tables 12, highlighting the performance of the SMRS-selected representative subset compared to the full dataset. 3,000 representative points were selected and achieved an accuracy of 84.03% on the validation set, which is slightly lower than the 89% achieved when training on the full dataset. This minor drop in accuracy demonstrates that the SMRS-selected data effectively captured the dataset's essential structure, even though only 5% of the data was used for training.

The most significant advantage of the SMRS approach is the substantial reduction in computational time. Training the model on the representative subset took 4.29 seconds , compared to 145.89 seconds required when using the full dataset. This drastic decrease in training time demonstrates the effectiveness of SMRS for the situations where computational efficiency is critical.

Overall, these findings underscore the potential of SMRS as a data-efficient training strategy. While the representative subset does not fully match the performance of the full dataset, it offers a significant reduction in computational cost and storage requirements.

Table 1: Accuracy Measured

Metric	**SMRS**	**Original**
Accuracy (%)	84	89
Precision (%)	85	89
Recall (%)	84.03	89

Table 2: Time Measured

Metric	**SMRS**	**Original**
Time (Seconds)	4.29	145.89

7 Conclusion

In this study, we explored a method for selecting representative data points using the Frobenius norm with sparsity constraints, which effectively excluded outliers and enabled efficient optimization through the ℓ_1-norm relaxation. Experiments conducted on both synthetic and practical datasets showed the algorithm's ability to represent the essential structure of the datasets while providing computational efficiency, making it useful for situations with limited computational resources.

However, a limitation of the method lies in the relaxation of the ℓ_0-norm to the ℓ_1-norm. Specifically, achieving sparsity required iterative tuning of the sparsity parameter when searching for representative points in C , as the ℓ_1-norm does not guarantee exact sparsity of 0. This made us to test multiple times of our experiments to obtain the desired sparsity, highlighting the need for more efficient sparsity promoting method.

Future work will aim to address these challenges by exploring more advanced classification techniques beyond SVM to evaluate the performance of representatives selected through SMRS on more complex and challenging classification tasks. Additionally, future studies will investigate strategies to effectively handle datasets with overlapping convex hulls.

References

Kristin Bennett and Erin Bredensteiner. Duality and geometry in svm classifiers. 09 2000.

Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2848–2855. IEEE, 2012.

Mingbin Feng, John J. Mitchell, Jong Shi Pang, Xin Shen, and Andreas Waechter. Complementarity formulations of l0-norm optimization. *Pacific Journal of Optimization*, 14(2):273–305, 2018. ISSN 1348-9151.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4th edition, 2013.

Roland Herzog, Frederik Köhne, Leonie Kreis, and Anton Schiela. Frobenius-type norms and inner products of matrices and linear maps with applications to neural network training, 2023. URL `https://arxiv.org/abs/2311.15419`.

Mingyi Hong and Zhi-Quan Luo. On the linear convergence of the alternating direction method of multipliers, 2013. URL `https://arxiv.org/abs/1208.3922`.

Chun-Guang Li, Chong You, and Rene Vidal. On geometric analysis of affine sparse subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1520–1533, December 2018. ISSN 1941-0484. doi: 10.1109/jstsp.2018.2867446. URL `http://dx.doi.org/10.1109/JSTSP.2018.2867446`.

Carlos Ramirez, Vladik Kreinovich, and Miguel Argaez. Why l1 is a good approximation to l0: A geometric explanation. 7:203–207, 01 2013.

Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3508. IEEE, 2010. doi: 10.1109/CVPR.2010.5539964.

Yilun Wang, Jing Wang, and Zongben Xu. On recovery of block-sparse signals via mixed l2/lq (0 < q <= 1) norm minimization. *EURASIP Journal on Advances in Signal Processing*, 2013(1):76, 2013. doi: 10.1186/1687-6180-2013-76. URL `https://doi.org/10.1186/1687-6180-2013-76`.