A Novel ℓ_0-constrained Gaussian Graphical Model for Anomaly Localization

Dzung T. Phan, Tsuyoshi Idé, Jayant Kalagnanam
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{phandu, tide,jayant}@us.ibm.com

Matt Menickelly, Katya Scheinberg
Lehigh University
Bethlehem, PA 18015
{mjm412,katyas}@lehigh.edu

***Abstract*—We consider the problem of anomaly localization in a sensor network for multivariate time-series data by computing anomaly scores for each variable separately. To estimate the sparse Gaussian graphical models (GGMs) learned from different sliding windows of the dataset, we propose a new model wherein we constrain sparsity directly through ℓ_0 constraint and apply an additional ℓ_2 regularization in the objective. We then introduce a proximal gradient algorithm to efficiently solve this difficult nonconvex problem. Numerical evidence is provided to show the benefits of using our model and method over the usual convex relaxations for learning sparse GGMs using a real dataset.**

I. Introduction

The ever-increasing pervasiveness of digital technologies makes it possible to collect enormous amounts of information about the physical world. Data mining techniques play a critical role in converting low-level sensor data collected from IoT (Internet-of-Things) devices into actionable insights. Anomaly and change detection is of primary importance in IoT since identifying something as unusual is almost always the first step towards additional actions by humans.

In the traditional problem setting, the goal of anomaly *detection* is to compute the degree of anomalousness for a multivariate measurement, giving an overall anomaly score. We are instead interested in the task of anomaly *localization*, where a variable-wise anomaly score is desired. In wireless sensor networks, for example, it is often inadequate simply to indicate whether or not a network is behaving anomalously, *i.e.* it is not enough to simply perform anomaly detection. If a sensor network takes measurements from different car parts of a prototype automobile on a test track, then it is far more valuable to the field engineers to know *which* car parts are contributing to an anomalous behavior, rather than simply knowing that the car as a whole is behaving anomalously [1]. Similar problem settings can be found across different application domains: monitoring mechanical stresses in buildings and bridges [2], pinpointing the change points in stock price time series [3], identifying car thefts [4], and finding sources of serious threats in computer networks [5].

For anomaly localization, two main lines of research have been proposed to date. In the first, Jiang *et al.* [6] used sparse principal components analysis (PCA) to identify a set of variables that have nonzero weights in a subspace corresponding to the distribution of abnormal samples in the training data. Their problem setting is more like in-sample data cleansing, which identifies anomalous samples and variables in the training data. In the second, Idé *et al.* [1], [7] proposed a graph-based anomaly localization approach, where two separate *sparse* dependency graphs are inferred from training and testing data. In the training phase, a normal state model encoded by a sparse Gaussian graphical model (GGM) is crated based on a training data set. Then, some measure (*anomaly score*) of each node (variable) in the dependency graphs is computed to determine how responsible each node is for the difference between the two graphs.

Although sparsity is of utmost importance in terms of identification of responsible variables as well as robustness to the noise of real-world sensor data, little is known about what is the best approach to learn sparse dependency graphs in the context of anomaly localization. In the literature, only an ℓ_1-regularized model has been investigated [7], [8]. In this paper, we demonstrate that the proposed GGM-learning model employing both an ℓ_0 constraint and an ℓ_2 regularization, along with a novel optimization algorithm to solve it, outperforms other models including typical ℓ_1-regularized models. In particular, we leverage the conditional expected Kullback-Liebler (KL) divergence method proposed in [7] when computing anomaly scores for each variable. A comparative study is conducted on various GGM-learning models and scoring methods for assigning anomaly scores.

The layout of the paper is as follows. In § II-A, we will describe in detail the models that we propose for learning sparse GGMs. In § II-B, we will discuss methods of anomaly localization for identifying anomalous variables between different GGMs. In § III, we will introduce our proposed proximal gradient algorithm. Finally, in § IV, we provide some numerical results illustrating a preference for using the proposed sparsity-constrained optimization model in the anomaly localization setting.

II. Learning Graphical Models and Anomaly Scoring Methods

We will now describe optimization models that we use to learn sparse precision matrices, and then review methods for performing anomaly localization on pairs of sparse precision matrices.

A. Sparse Graph Learning Models

Given a sample covariance matrix $\mathbf{S}$, the authors of [7] compute the sparse dependency graph $\mathbf{X}$ by solving the ℓ_1-regularized maximum likelihood problem, i.e.,

$$\min_{\mathbf{X}\succ\mathbf{0}} \ \mathrm{tr}(\mathbf{SX}) - \log\det(\mathbf{X}) + \lambda\|\mathbf{X}\|_1, \tag{1}$$

where $\|\mathbf{X}\|_1 = \sum_{i,j=1}^n |X_{ij}|$ is the component-wise ℓ_1 norm of the matrix $\mathbf{X}$, and $\lambda > 0$ is a regularization parameter to control sparsity. Zeros in the precision matrix $\mathbf{X}$ indicate conditional independence between two variables.

In this section, we propose a new model for estimating sparse graphs. In particular, rather than regularizing an ℓ_1 term to control sparsity as in (1), we will directly *constrain* sparsity by specifying a maximally allowable number of nonzeros κ in the optimal solution and add a quadratic penalty

$$\begin{array}{ll}\min\limits_{\mathbf{X}\succ\mathbf{0}} & \mathrm{tr}(\mathbf{SX}) - \log\det(\mathbf{X}) + \frac{\lambda}{2}\|\mathbf{X}\|_F^2 \\ \text{s.t.} & \|\mathbf{X}\|_0 \le \kappa,\end{array} \tag{2}$$

where $\lambda > 0$ is once again a regularization parameter and the Frobenius norm term $\|\mathbf{X}\|_F^2 = \sum_{i,j=1}^n |X_{ij}|^2$ is also referred to as ℓ_2 regularization in our paper. We denote by $\|\cdot\|_0$ the number of nonzeros of the argument, the so-called ℓ_0 norm. The ℓ_0 constraint guarantees that the solution will admit a certain level of sparsity. The quadratic penalty encourages the capacity of selecting groups in the presence of highly correlated variables [9]. Without the regularization, some entries of the purely ℓ_0-constrained model can have relatively large magnitudes. The associated anomaly scores significantly dominate the others, and thus some faulty variables can be overlooked. This is not a desirable property of sparse precision matrices in the context of anomaly localization, even if such matrices yield meaningful sparsity patterns. Hence the ℓ_2 regularization term keeps the magnitude of all entries uniformly similar.

We also investigate some other precision matrix estimation models that can be compared with our model (2) in the context of anomaly localization. In the literature, some authors have studied the cardinality-constrained model [10], [11]

$$\begin{array}{ll}\min\limits_{\mathbf{X}\succ\mathbf{0}} & \mathrm{tr}(\mathbf{SX}) - \log\det(\mathbf{X}) \\ \text{s.t.} & \|\mathbf{X}\|_0 \le \kappa.\end{array} \tag{3}$$

We will also consider an ℓ_1-based "elastic net" version for GGM model,

$$\min_{\mathbf{X}\succ\mathbf{0}} \ \mathrm{tr}(\mathbf{SX}) - \log\det(\mathbf{X}) + \lambda_1\|\mathbf{X}\|_1 + \lambda_2\|\mathbf{X}\|_F^2. \tag{4}$$

We note that among the four models (1)-(4), only the ℓ_1 model (1) has been applied in learning a sparse GGM for anomaly detection algorithms. We will conduct empirical experiments using all of these models, and show that our ℓ_0-based model (2) performs consistently better than the others.

B. Methods for Computing Anomaly Score of Each Variable

We briefly summarize three previously proposed techniques for performing the change analysis, *i.e.* for assigning *anomaly scores* to individual variables to measure the magnitude of their contributions to the observed differences between two sparse dependency graphs from training and testing datasets.

1) Conditional Expected KL-Divergence [7]: For the ith variable, and for a learned GGM $\mathbf{X}$, let

$$\mathbf{X} = \begin{bmatrix} \mathbf{L} & \mathbf{l} \\ \mathbf{l}^\top & \alpha \end{bmatrix}, \ \mathbf{X}^{-1} = \begin{bmatrix} \mathbf{W} & \mathbf{w} \\ \mathbf{w}^\top & \beta \end{bmatrix}$$

be permuted such that the last row and column of $\mathbf{X}, \mathbf{X}^{-1}$ correspond to the ith variable. Then, letting $\mathbf{X}^A$ and $\mathbf{X}^B$ denote the GGMs learned from datasets A and B respectively, one can show

$$\begin{aligned} d_i^{AB} &= \mathbf{w}^{A\top}(\mathbf{l}^B - \mathbf{l}^A) + \frac{1}{2}\left[\frac{\mathbf{l}^{B\top}\mathbf{W}^A\mathbf{l}^B}{\alpha^B} - \frac{\mathbf{l}^{A\top}\mathbf{W}^A\mathbf{l}^A}{\alpha^A}\right] \\ &+ \frac{1}{2}\left[\ln\frac{\alpha^A}{\alpha^B} + \beta^A(\alpha^B - \alpha^A)\right]. \end{aligned} \tag{5}$$

We then define the anomaly score of the ith variable as

$$d_i = \max\left\{d_i^{AB}, d_i^{BA}\right\}. \tag{6}$$

2) Stochastic Nearest Neighbors [1] : Let $\mathbf{S}^A$ and $\mathbf{S}^B$ denote the sample covariance matrices of datasets A and B respectively, and let $\mathbf{S}_i^A$,$\mathbf{S}_i^B$ denote the ith columns of $\mathbf{S}^A$ and $\mathbf{S}^B$ respectively. For a sparse GGM $\mathbf{X}^A$ learned for dataset A, define an indicator vector associated with the ith variable $\mathbf{l}_{A,i}$ coordinate-wise by

$$[\mathbf{l}_{A,i}]_j = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are adjacent in } \mathbf{X}^A \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Then, a measure of the dissimilarity between the neighborhoods of the ith variable between the GGMs for A and B, weighted by the sample covariances of the two datasets, is given by

$$d_i^{AB} = \left|\frac{\mathbf{l}_{A,i}^\top(\mathbf{S}_i^A - \mathbf{S}_i^B)}{(1+\mathbf{l}_{A,i}^\top\mathbf{S}_i^A)(1+\mathbf{l}_{A,i}^\top\mathbf{S}_i^B)}\right| \tag{8}$$

Symmetrically, we define d_i^{BA} and then compute an anomaly score d_i as in (6).

3) Sparsest Subgraph Approximation: Given two (sparse) dependency graphs, *e.g.* $\mathbf{X}^A$ and $\mathbf{X}^B$, consider a graph given by an adjacency matrix $\mathbf{\Lambda}$ defined entrywise by

$$\Lambda_{ij} = |X_{ij}^A - X_{ij}^B|$$

The authors of [8] proposed a convex quadratic program

$$\min_{\mathbf{d}\in\mathbb{R}^n} \mathbf{d}^\top\mathbf{\Lambda}_\mu\mathbf{d} \quad \text{ s. to } \mathbf{1}_n^\top\mathbf{d} = 1, \mathbf{d} \ge \mathbf{0}_n, \tag{9}$$

where $\mathbf{\Lambda}_\mu = \mathbf{\Lambda} + \mu\mathbf{I}_n$ is the original matrix $\mathbf{\Lambda}$ with an added scaled identity with $\mu \ge 0$ sufficiently bounded away from zero to enforce the positive definiteness of $\mathbf{\Lambda}_\mu$. The solution $\mathbf{d}^*$ to (9) can be interpreted as anomaly scores.

III. Optimization Algorithm for ℓ_0 Sparse Models

This section proposes a proximal gradient algorithm for solving the ℓ_0-constrained problem (2). We will consider a general problem of the form

$$\min\{f(\mathbf{X}) : \mathbf{X} \in \mathbb{R}^{n\times n}, \|\mathbf{X}\|_0 \le \kappa, \mathbf{X} \succ \mathbf{0}, \mathbf{X} = \mathbf{X}^\top\}, \tag{10}$$

where f is a smooth convex function. We define the sparsity constraint set

$$\Omega = \{\mathbf{X} : \|\mathbf{X}\|_0 \leq \kappa\}$$

and the projection operator

$$P_\Omega(\mathbf{X}) = \arg\min_{Y \in \Omega} \|\mathbf{X} - \mathbf{Y}\|_F.$$

A possible strategy for solving a constrained optimization problem like (10) is to use a proximal gradient method. One issue here is that the constraint set in our problem $\Omega \cap \{\mathbf{X} \succ \mathbf{0}, \mathbf{X} = \mathbf{X}^\top\}$ is an intersection of Ω and the symmetric positive-definite (PD) cone without its boundary, making it nonconvex and not closed. In our method, feasibility with respect to membership in Ω is handled via projection, while symmetric positive-definiteness of the iterates in ensured through a line-search procedure.

On each iteration, we begin from a feasible iterate $\mathbf{X}^k$ and then backtrack along the projection arc defined by $P_\Omega(\mathbf{X}^k - \alpha_k \nabla f(\mathbf{X}^k))$, initializing the stepsize $\alpha_k > 0$ with a Barzilai-Borwein stepsize [12]. We terminate the backtracking procedure for computing $\mathbf{X}^{k+1}$ once the following conditions (C1) and (C2) are both satisfied:

(C1) A sufficient decrease condition for the objective function value is attained, i.e.

$$f(\mathbf{X}^{k+1}) \leq f(\mathbf{X}^k) - \frac{\delta}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2,$$

for some algorithmic parameter $\delta > 0$.

(C2) The next iterate is feasible with respect to the positive definite constraint, i.e. $\mathbf{X}^{k+1} \succ \mathbf{0}$.

A detailed description of an algorithm for solving (10) is given in Algorithm 1.

As mentioned, Ω is a nonconvex set, and as such, the operator P_Ω is generally set-valued, *i.e.* projections are non-unique. It is well-known that a point in $P_\Omega(\mathbf{X})$ can be quickly obtained [13]. Indeed, if $\mathbf{X} \in \mathbb{R}^{n \times n}$, then $P_\Omega(\mathbf{X})$ is computed by sorting the absolute values of the magnitudes of the n^2 many entries in $\mathbf{X}$ and setting all but the κ largest values in $\mathbf{X}$ to 0, where ties can be broken arbitrarily. For the sake of convergence, when we compute P_Ω, ties are not broken arbitrarily, but by a dictionary order on the matrix entries. The dictionary order is chosen in a way to ensure that the projected matrix is symmetric. We remark that as a consequence of sorting n^2 matrix entries, the projection operation can be performed in $\mathcal{O}(n^2 \log(n))$ time.

IV. Experimental Results

In this section, we study the empirical performance of the sparse GGMs in Section II-A. Three anomaly scoring method were used: conditional expected KL-divergence [7], stochastic nearest neighbors (NN) [1], and sparsest subgraph approximation (SA) [8]. We use a particular ROC curve and the area under the ROC curve (AUC) as described in [7]. A detailed usage and notations are given in Table I.

Algorithm 1: Proximal Gradient Algorithm

Given parameters $\sigma \in (0,1), \delta > 0, [\alpha_{min}, \alpha_{max}] \subset (0, \infty)$, an initial feasible point $\mathbf{X}^0$.
Set $k = 0$.
while *some stopping criteria not satisfied* **do**
  Step 1: BB step size:
  if $k = 0$ **then**
    $\alpha \leftarrow 1$
  else
    $\alpha \leftarrow \min(\alpha_{max}, \max(\alpha_{min}, \frac{\mathrm{tr}((\nabla f(\mathbf{X}^k) - \nabla f(\mathbf{X}^{k-1}))(\mathbf{X}^k - \mathbf{X}^{k-1}))}{\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F^2}))$
  Step 2: Line search along projection arc
  $j \leftarrow 0,\ ls \leftarrow$ **true**
  while $(ls =$ **true**$)$ **do**
    $\beta_k \leftarrow \sigma^j \alpha$
    $\mathbf{X}^{k+1} \leftarrow P_\Omega(\mathbf{X}^k - \beta_k \nabla f(\mathbf{X}^k))$
    if $\Big(f(\mathbf{X}^{k+1}) \leq f(\mathbf{X}^k) - \frac{\delta}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2$ *and* $\mathbf{X}^{k+1} \succ \mathbf{0}\Big)$ **then**
      $ls \leftarrow$ **false**
    else
      $j \leftarrow j + 1$
  Step 3: Iterate
  $k \leftarrow k + 1$

Graph learning	anomaly score	symbol
Eq. (1)	KL divergence (5)	$\ell_1 + \mathsf{KL}$
Eq. (1)	stochastic NN (8)	$\ell_1 + \mathsf{SNN}$
Eq. (1)	sparsest SA (9)	$\ell_1 + \mathsf{SSA}$
Eq. (3)	KL divergence (5)	$\ell_0 + \mathsf{KL}$
Eq. (3)	stochastic NN (8)	$\ell_0 + \mathsf{SNN}$
Eq. (3)	sparsest SA (9)	$\ell_0 + \mathsf{SSA}$
Eq. (2)	KL divergence (5)	$\ell_0 + \ell_2 + \mathsf{KL}$
Eq. (2)	stochastic NN (8)	$\ell_0 + \ell_2 + \mathsf{SNN}$
Eq. (2)	sparsest SA (9)	$\ell_0 + \ell_2 + \mathsf{SSA}$
Eq. (4)	KL divergence (5)	$\ell_1 + \ell_2 + \mathsf{KL}$
Eq. (4)	stochastic NN (8)	$\ell_1 + \ell_2 + \mathsf{SNN}$
Eq. (4)	sparsest SA (9)	$\ell_1 + \ell_2 + \mathsf{SSA}$

TABLE I: Combination of graph learning methods and anomaly scoring methods.

A. Data Set

We experimented a real-world dataset to assess the models. The *Sensor Error* data is a 42-dimensional dataset of sensor signals, which comes from several experimental runs with prototype cars. It consists of one run under a normal system operation and two runs under abnormal conditions. The anomalies in the latter two runs are caused by sensor miswiring errors. There are 350 and 360 time series samples respectively for the two abnormal runs and roughly 550 time series samples in the normal reference run. There were two miswired sensors in the anomalous runs known to the experimenters, and hence a good method of anomaly localization should be able to identify these two sensors as faulty. Data was generated by taking pairs

	ℓ_0+ℓ_2+KL	ℓ_0+ℓ_2+SSA	ℓ_0+ℓ_2+SNN	ℓ_1+KL	ℓ_1+SSA	ℓ_1+SNN	ℓ_0+KL	ℓ_0+SSA	ℓ_0+SNN	ℓ_1+ℓ_2+KL	ℓ_1+ℓ_2+SSA	ℓ_1+ℓ_2+SNN
mean	**0.9767**	0.9412	0.9637	0.9631	0.8334	0.9388	0.9448	0.8606	0.9551	0.9627	0.9536	0.9454
std	**0.0545**	0.1318	0.0606	0.1043	0.1933	0.0747	0.1039	0.1399	0.0650	0.1055	0.1095	0.0683

TABLE II: The mean and standard deviation for AUC values

of sliding windows of length 50, one from each of the normal and abnormal runs.

B. Anomaly Localization Evaluation

The ROC curves are reported in Fig. 1. As we can see, for each change analysis method, our $\ell_0+\ell_2$-based model (2) outperforms the ℓ_1-based models and the pure ℓ_0 model (3). One exception is for the sparsest subgraph approximation scoring method (9), our method is slightly behind the $\ell_1+\ell_2$ model. The performance of the $\ell_1+\ell_2$ model is slightly better than that of the ℓ_1 model. The combination of the ℓ_0 constraint with the ℓ_2 regularization is more advantageous than the use of the $\ell_1+\ell_2$ regularization.

We also see the benefit of adding the Frobenius norm term to our model (2) when comparing it against the pure ℓ_0 model (3). The regularization term helps to improve the accuracy in all cases. The pure ℓ_0 model (3) is generally defeated by the proposed model (2). As evidence for the power of using an ℓ_0 constraint as opposed to an ℓ_1 regularization in terms of sparsity pattern recovery, when using the stochastic nearest neighbors as a scoring method, a method inherently only concerned with sparsity patterns as opposed to relative magnitudes, the pure ℓ_0 model has a good performance.

We see that $\ell_0+\ell_2+\text{KL}$ is often the top performer. The use of the $\ell_0+\ell_2$ model with other anomaly scoring methods still gives a very competitive result. The sparse graphs learned by the pure ℓ_0 model (3) tend to do worse than other graph learning methods if the associated scoring method also uses the magnitude values of precision matrix.

There are a number of data windows, and each of them can give an AUC value based on the standard ROC curve definition as described in [6], [8]. Table II gives the mean and standard deviation of AUC values for all possible tests. The best value is highlighted in bold. The mean AUC of $\ell_0+\ell_2+\text{KL}$ is the highest. The standard deviations of our proposed $\ell_0+\ell_2$ model are often the smallest for the same scoring method.

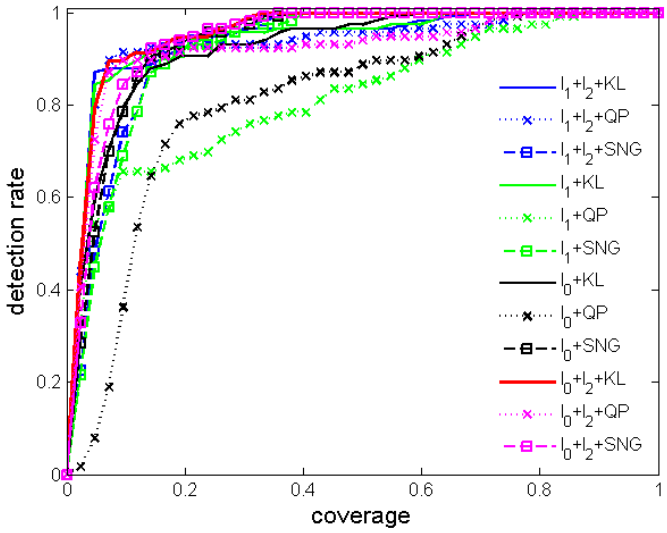

Fig. 1: Comparison of ROC curves

V. Conclusions

This paper introduces a new method for anomaly localization for multivariate time series datasets. We proposed an ℓ_0-constrained GGM model with an ℓ_2 regularization in the objective to learn sparse dependency graphs and developed a proximal gradient algorithm for the ℓ_0-constrained problem. It has been shown that our ℓ_2-regularized ℓ_0-constrained model, combined with the conditional expected Kullback-Liebler divergence anomaly scoring, outperforms other methods for detecting anomalies.

References

[1] T. Idé, S. Papadimitriou, and M. Vlachos, "Computing correlation anomaly scores using stochastic nearest neighbors," in *Proceedings of IEEE International Conference on Data Mining*, 2007, pp. 523–528.

[2] N. Xu, S. Rangwala, K. K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D. Estrin, "A wireless sensor network for structural monitoring," in *SENSYS*, 2004, pp. 13–24.

[3] X. Liu, X. Wu, H. Wang, R. Zhang, J. Bailey, and K. Ramamohanarao, "Mining distribution change in stock order streams," in *ICDE*, 2010, pp. 105–108.

[4] H. Song, S. Zhu, and G. Cao, "Svats: A sensor-network-based vehicle anti-theft system," in *INFOCOM*, 2008, pp. 2128–2136.

[5] A. Lahkina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM*, 2004, pp. 219–230.

[6] R. Jiang, H. Fei, and J. Huan, "A family of joint sparse pca algorithms for anomaly localization in network data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2421–2433, November 2011.

[7] T. Idé, A. C. Lozano, N. Abe, and Y. Liu, "Proximity-based anomaly detection using sparse structure learning," in *Proceedings of 2009 SIAM International Conference on Data Mining*, 2009, pp. 97–108.

[8] S. Hara, T. Morimura, T. Takahashi, and H. Yanagisawa, "A consistent method for graph based anomaly detection," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[9] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[10] Z. Lu and Y. Zhang, "Sparse approximation via penalty decomposition methods," *SIAM Journal on Optimization*, vol. 23, pp. 2448–2478, 2013.

[11] X. Yuan, P. Li, and T. Zhang, "Gradient hard thresholding pursuit for sparsity-constrained optimization," in *ICML'14*, 2014.

[12] J. Barzilai and J. M. Borwein, "Two point step size gradient methods," *IMA Journal of Numerical Analysis*, vol. 8, pp. 141–148, 1988.

[13] W. W. Hager, D. T. Phan, and J. Zhu, "Projection algorithms for nonconvex minimization with application to sparse principal components analysis," *Journal of Global Optimization*, vol. 65, pp. 657–676, 2016.