Greedy vs. L1 convex optimization in sparse coding

Ren, Huamin; Pan, Hong; Olsen, Søren Ingvor; Moeslund, Thomas B.

Published in:
Journal of Machine Learning Research

Publication date:
2015

Document Version
Peer reviewed version

Citation for published version (APA):
Ren, H., Pan, H., Olsen, S. I., & Moeslund, T. B. (2015). Greedy vs. L1 convex optimization in sparse coding: comparative study in abnormal event detection. Journal of Machine Learning Research, 37.

Greedy vs. L1 Convex Optimization in Sparse Coding

Ren, Huamin; Pan, Hong; Olsen, Søren Ingvor; Moeslund, Thomas B.

Publication date:
2015

Document Version
Peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Ren, H., Pan, H., Olsen, S. I., & Moeslund, T. B. (2015). Greedy vs. L1 Convex Optimization in Sparse Coding: Comparative Study in Abnormal Event Detection. Paper presented at ICML '15 Workshop, Lille, France.

Greedy vs. L1 Convex Optimization in Sparse Coding: Comparative Study in Abnormal Event Detection

Huamin Ren HR@CREATE.AAU.DK
Aalborg University, Aalborg, Denmark

Hong Pan HONG.PAN@DI.KU.DK
University of Copenhagen, Copenhagen, Denmark

Søren Ingvor Olsen INGVOR@DI.KU.DK
University of Copenhagen, Copenhagen, Denmark

Thomas B. Moeslund TBM@CREATE.AAU.DK
Aalborg University, Aalborg, Denmark

Abstract

Sparse representation has been applied successfully in many image analysis applications, including abnormal event detection, in which a baseline is to learn a dictionary from the training data and detect anomalies from its sparse codes. During this procedure, sparse codes which can be achieved through finding the L0-norm solution of the problem: $\min \|Y - D\alpha\|_2^2 + \|\alpha\|_0$, is crucial. Note that D refers to the dictionary and α refers to the sparse codes. This L0-norm solution, however, is known as a NP-hard problem. Despite of the research achievements in some classification fields, such as face and action recognition, a comparative study of codes in abnormal event detection is less studied and hence no conclusion is gained on the effect of codes in detecting abnormalities. We constrict our comparison in two types of the above L0-norm solutions: greedy algorithms and convex L1-norm solutions. Considering the property of abnormal event detection, i.e., only normal videos are used as training data due to practical reasons, effective codes in classification application may not perform well in abnormality detection. Therefore, we compare the sparse codes and comprehensively evaluate their performance from various aspects to better understand their applicability, including computation time, reconstruction error, sparsity, detection accuracy on the UCSD Anomaly Dataset. Experiments show that greedy algorithms, especially MP and StOMP algorithm could achieve better abnormality detection with relatively less computations.

Proceedings of the 31^{st} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

1. Introduction

Sparse representation has gained a great deal of attention since being applied effectively in many image analysis applications, e.g., image denoising (Donoho & Elad, 2002; Elad & Aharon, 2006), compression (Zepeda et al., 2011), face recognition (Wright et al., 2009; Yang et al., 2011), and action recognition (Qiu et al., 2011; Jiang et al., 2013). Its success stems from the discovery of underlying properties from low-level to mid-level human vision: many neurons in the visual pathway are selective for a variety of specific stimuli, such as color, texture, orientation, scale, and even view tuned object images (Olshausen & Fieldt, 1997). Attributing to this finding, many researchers use sparse representation, either to construct a sparse dictionary or search for sparse coefficients given by a predefined dictionary.

When applying sparse representations to abnormal event detection (Cong et al., 2011; Jiang et al., 2011; Qiu et al., 2011; Lu et al., 2013), visual features are first extracted either on spatial- or temporal- domain, then a dictionary D is learned based on these visual features, which consists of basis vectors capturing high-level patterns in the input features. A sparse representation of a feature is a linear combination of a few elements or atoms from a dictionary. Mathematically, it can be expressed as $y = Dx$, where $y \in R^p$ is a feature of interest, $D \in R^{p \times m}$ is a dictionary and $x \in R^m$ is the sparse representation of y in D. Typically $m \gg p$ results in an overcomplete or redundant dictionary. Then, each feature has a reconstruction error based on the dictionary and its coefficients. As a result, a testing feature can be determined as normal or anomaly based on its reconstruction error.

Research on sparse representation in image and video processing applications can be generally categorized into dictionary learning (Cong et al., 2011; Jiang et al., 2011; Qiu et al., 2011; Lu et al., 2013) (Engan et al., 1999) (Aharon et al., 2006; Michal et al., 2005), and sparse coding (Donoho et al., 2006a; Donoho & Elad, 2006; Needell & Vershynin, 2010; B. et al., 2011; Asif & Romberg, 2013). Dictionary learning aims to obtain atoms (or basis vectors) for a dictionary. Such atoms could be either predefined, e.g., undecimated Wavelets, steerable Wavelets, Contourlets, Curvelets, and more variants of Wavelets), or learned from the data itself (Mairal et al., 2008a) (Mairal et al., 2008b). Sparse coding, on the other hand, attempts to find sparse codes (or coefficients) by giving a dictionary, i.e., finding the solution to the underdetermined system of equations $y = Dx$ either by greedy algorithms or convex algorithms. Through sparse coding, input features can be approximately represented as a weighted linear combination of a small number of (unknown) basis vectors. We only consider sparse coding in this paper, and put special attention on comparison of greedy algorithms with L1-norm minimization algorithms.

L1-norm minimization has become a popular tool to solve sparse coding, which benefits both from efficient algorithms (Beck & Teboulle, 2009) (Yuan et al., 2010) and a well-developed theory for generalization properties and variable selection consistency (Zhang, 2009). We list two common L1 minimization formulation in E.q. 1 and E.q. 2. Since the problem is convex, there are efficient and accurate numerical solvers.

$$\hat{x} = \operatorname*{argmin}_x \frac{1}{2}\|Dx - y\|_2^2 + \lambda\|x\|_1 \tag{1}$$

$$\hat{x} = \operatorname*{argmin}_x \|x\|_1 \quad \text{subject to} \quad \|Dx - y\|_2 \leq \epsilon \tag{2}$$

Meanwhile, there is also a variety of greedy/iterative methods for solving such problems. Greedy algorithms rely on interactive approximation of the feature coefficients and supports, either by iteratively identifying the support of the feature until a convergence criterion is met, or alternatively by obtaining an improved estimate of the sparse signal at each iteration that attempts to account for the mismatch to the measured data. Compared to L1-norm minimization methods, greedy algorithms are much fasters, therefore are more applicable to very large problems.

To evaluate the performance of these two types of solutions, we in this paper compare the performance of codes on abnormal event detection, especially from aspects such as computation complexity, reconstruction error, sparsity of codes, and their detection performance. The remainder of this paper is organized as follows. We give a brief review of greedy algorithms and L1-norm solutions in Sec. 2 and Sec. 3, then show our comparative results in Sec. 4. Sec. 5 concludes the paper with discussions and future work.

2. Greedy Algorithms

We review two broad categories of greedy methods to reconstruct y, which is called 'greedy pursuits' and 'threshold' algorithms. Greedy pursuits can be defined as a set of methods that iteratively build up an estimate x. They contains three basic steps. Firstly, the x is set to a zero vector. Secondly, these methods estimate a set of non zero components of x by iteratively adding new components that are deemed to be non zeros. Thirdly, the values for all non zeros components are optimized. In contrast, thresholding algorithms alternate both element selection and element pruning steps.

There is a large and growing family of greedy pursuit methods. The general framework in greedy pursuit techniques is 1) to select an element and 2) to update the coefficients. In Matching Pursuit (MP) (Mallat & Zhang, 1993). Mallat and Zhang discussed a general method for approximate decomposition in E.q. 3 that addresses the sparsity issue directly. The algorithm selects one column from D at a time and, at each iteration, only the coefficient associated with the selected column is updated. More concretely, it starts from an initial approximation x (0) = 0 and residual $R(0) = x$, then builds up a sequence of sparse approximations stepwise. At stage k, it identifies the dictionary atom that best correlates with the residual and then adds to the current approximation a scalar multiple of that atom. After m steps, one has a sparse code in E.q. 3 with residual $R = R^{(m)}$.

$$y = \sum_{i=1}^{m} x_{r_i} d_{r_i} + R^{(m)} \tag{3}$$

Orthogonal Matching Pursuit (OMP) (Pati et al., 1993), updates x in each iteration by projecting y orthogonally onto the columns of D associated with the current support atoms. Different from MP, OMP never reselects an atom and the residual at any iteration is always orthogonal to all currently selected atoms in the dictionary. Another difference is that OMP minimizes the coefficients for all selected atoms at iteration k, while MP only updates the coefficient of the most recently selected atom. In order to speed up pursuit algorithms, it is necessary to select multiple atoms at a time, therefore, the algorithms are proposed to keep computational costs low enough for applying to large-scale problems, such as Stagewise Orthogonal Matching Pursuit (StOMP) (Donoho et al., 2006b) (Donoho et al., 2012) and Stagewise Weak Gradient Pursuit (StWGP) (Blumensath & Davies, 2009). These algorithms choose the element that meets some threshold criterion at the atom selection step, and have demonstrated both theoretical and empirical effectiveness for the large-system.

Greedy algorithms are easy to implement and use and can be extremely fast. However, they do not have recovery guarantees, i.e., how well each sample can be reconstructed by the dictionary and their sparse codes, as strong as methods based on convex relaxation such as L1-norm approximation. Recent methods, including Iterative Hard Thresholding (IHT) (Blumensath & Davies, 2008), the Compressive Sampling Matching Pursuit (CoSaMP) (Needell & Tropp, 2008), Subspace Pursuit (SP), aim to bridge this gap. They are fairly easy to implement and can be extremely fast but also show the strong performance guarantees.

3. L1-norm Approximation

L1-norm approximation replaces the L0 constraint by a relaxed L1-norm. For example, in the Basis Pursuit method (BP) (Chen et al., 1998; 2001), an almost everywhere differentiable and often convex cost function is applied, in the Focal Underdetermined System Solver (FOCUSS) algorithm (Murray & Kreutz-Delgado, 2001), a more general model is optimzed.

Donoho and etc. (Donoho & Elad, 2002) suggested that for some measurement matrices D, the generally NP-Hard problem (L0 norm) should be equivalent to its convex relaxation: L1 norm, see E.q. 1 and 2. The convex L1 problem can be solved using methods of linear programming. Representative work includes Basis Pursuit (BP). Instead of seeking sparse representations directly, it seeks representations that minimize the L1 norm of the coefficients. Furthermore, BP can compute sparse solutions in situations where greedy algorithms fail. The Lasso algorithm (Tibshirani, 1996) is quite similar to BP and is in fact know as Basis Pursuit De-Noising (BPDN) in some areas. The Lasso rather than trying to minimizing the L1-norm like BP places a restriction on its value.

The FOCUSS algorithm has two integral parts: a low-resolution initial estimate of the real signal and the iteration process that refines the initial estimate to the final localized energy solution. The iterations are based on weighted norm minimization of the dependent variable with the weights being a function of the preceding iterative solutions. The algorithm is presented as a general estimation tool usable across different applications. In general, L1-norm methods offer better performance in many cases, but are also more demanding with respect to computation.

4. Experimental Results

There are intensive studies on different sparse coding algorithms, see surveys in (Baraniuk, 2007). In this paper we highlight these aspects in evaluating the sparse codes for abnormal event detection applications: computation time, reconstruction error, the ratio of sparsity in codes, and their performance on abnormal event detection.

4.1. Dataset and Settings

We use a public anomaly dataset: UCSD Ped1 dataset (Mahadevan et al., 2010), which has been popular used in detecting abnormal behaviors. UCSD Ped1 includes clips of groups of people walking towards and away from the camera with some amount of perspective distortion. There are 34 training videos and 36 testing videos with a resolution of 238×158. Training videos contain only normal behaviors. Testing videos are abnormal behaviors where there are either non-pedestrian entities in the walkways or anomalous pedestrian motion patterns.

We use the spatial-temporal cubes, where 3D gradient features are computed, following the setting in (Kratz & Nishino, 2009). Each frame is divided into patches with a size of 23×15, consecutive 5 frames are used to form 3D patches, and gradients features are extracted in each patch. See details in (Kratz & Nishino, 2009). Through this, we obtain 500-dimensional visual features and reduce them to 100 dimension by using PCA algorithm.

4.2. Codes Evaluation

We randomly select 1% of the training features (238000 features in total), then use K-SVD algorithm (Michal et al., 2005) to construct a dictionary consisting of 1000 atoms, and generate sparse codes by applying various algorithms. There are many algorithms available, we only select representative greedy algorithms (OMP, MP, StOMP) and compare them with representative L1-norm solutions (BP and Lasso algorithm). The reconstruction error is calculated by $Re = \|y - Dx\|_2^2$. We also calculate the mean ratio of sparsity in the codes, i.e., the average percentage of non zeros in the dimension of the codes (1000). We report these results as well as computation time in Tab. 1. Greedy algorithms need far less time to compute, among which OMP achieves the fast computation, following by StOMP algorithm. OMP is approx. 180 times faster than Lasso algorithm. OMP and StOMP could achieve sparser solutions, while BP could obtain an extremely dense solution with an exact recovery.

Table 1. Comparison of greedy algorithms and L1-norm solutions on sparse code generation.

ALGORITHMS	COMPUTATION TIME (S)	RECONSTRUCTION ERROR	SPARSITY (%)
MP	166.00	0	31.8%
OMP	1.83	0.4236	1.9%
STOMP	15.79	0	10%
BP	114.20	0	100%
LASSO	333.49	0.0005	9.9%

Table 2. Comparative results on UCSD Ped1: frame-level evaluation results (AUC and EER) and pixel-level evaluation results (AUC and EDR) are reported.

ALGORITHMS	AUC (FRAME-LEVEL)	EER	AUC (PIXEL-LEVEL)	EDR	COMPUTATION TIME (S)
MP	0.6956	0.3547	0.3898	0.5716	13342
OMP	0.5003	0.5052	0.2849	0.6637	527
STOMP	0.5415	0.465	0.3494	0.6190	4668
BP	0.5454	0.4764	0.3057	0.6479	38949
LASSO	0.5305	0.5173	0.3132	0.6383	56400

4.3. Abnormal Event Detection Evaluation

To measure the abnormality detection accuracy, we calculate the reconstruction error of each feature, and detect features with large reconstruction errors as anomalies. A frame with an abnormal feature is considered as a positive frame. To compare the performance, we adopt two popular evaluation criterion in abnormality detection: frame-level evaluation and pixel-level evaluation, which are defined in (Mahadevan et al., 2010). We exactly follow their setting in our evaluation, that is: on frame-level evaluation, a frame is considered an abnormal frame if it contains at least one anomaly feature; while on pixel-level evaluation, a frame is marked as a correctly detected abnormality if at least 40% of the truly abnormal pixels are detected. Ground truth on frame-level and pixel-level annotation is available, therefore, we calculate true positive rate and false positive rate to draw ROC curves. The Area Under the Curve (AUC) is reported. Following (Mahadevan et al., 2010), we obtain the value when false positive number equals to the missing value. They are called equal error rate (EER) and equal detected rate (EDR) on frame and pixel-level evaluation, respectively. See Tab. 2 for details. On frame-level evaluation, MP algorithm achieves the best results with a moderate computation time. Remarkablely, StOMP algorithm is relatively fast, and the AUC is quite satisfactory.

It is notable that pixel-level AUC is lower than frame-level AUC in general, which is because pixel-level evaluation is stricter and has taken location into consideration. On frame-level evaluation, there could be a coincidental detection - a normal feature could be wrongly detected as an anomaly in an abnormal frame, and this incorrect detection of feature ends up with a correct detection of that frame. While in pixel-level evaluation, a frame is marked as a correctly detected abnormality only if sufficient number of anomaly features have been found. The StOMP algorithm can achieve a competitive detection results on pixel-level evaluation compared with MP algorithm, but it is 3 times faster than MP algorithm. BP algorithm also performs well on pixel-level detection, however, its high computation cost will hamper its application on real detection problems.

In summary, greedy algorithms are fast to compute, however, their reconstruction errors are relatively larger than L1-norm solutions. Convex relaxations, such as BP and Lasso algorithm, have better theoretical guarantees and recovery ability, but are more time consuming. When applied in abnormal event detection, surprisingly, greedy algorithms, especially StOMP algorithm, seem to perform better on pixel-level detection, which means that they could localize anomaly feature more accurately.

5. Discussions and Conclusion

In this paper, we compare greedy algorithms with L1-norm solutions from different aspects: computation cost, recovery ability, sparsity, and their performance on abnormal event detection. Experimental results show that greedy algorithms can obtain good detection results, with less computations. Remarkably, among the top three best detection results, two are greedy algorithms. Considering the computation requirement, which limits some L1-norm algorithms from being applied in real surveillance applications, greedy algorithms are really promising. However, there are more consideration that need to be addressed in future. How is the performance of greedy algorithms compared to L1-norm solutions when the size of the training data is much larger? How do they perform in other domains when the amount of the training data is relatively small? In which case does greedy algorithms suit better than L1-norm solutions? These interesting questions will be considered in our future work.

References

Aharon, M., Elad, M., and Bruckstein, A. Svdd: An algorithm for designing overcomplete dictionaries for sparse representation. *Trans. Sig. Proc.*, 54(11):4311–4322, 2006.

Asif, M.S. and Romberg, J. K. Sparse recovery of streaming signals using l1-homotopy. *CoRR*, abs/1306.3331, 2013.

B., S., Candès, E. J., and Grant, M. C. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3(3):165–218, 2011.

Baraniuk, Richard. Compressive sensing. *IEEE Signal Processing Mag*, pp. 118–120, 2007.

Beck, Amir and Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, March 2009. ISSN 1936-4954.

Blumensath, T. and Davies, M.E. Stagewise weak gradient pursuits. *Signal Processing, IEEE Transactions on*, 57 (11):4333–4346, Nov 2009.

Blumensath, Thomas and Davies, MikeE. Iterative Thresholding for Sparse Approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, December 2008.

Chen, Scott Shaobing, Donoho, David L., and Saunders, Michael A. Atomic decomposition by basis pursuit. *SIAM JOURNAL ON SCIENTIFIC COMPUTING*, 20: 33–61, 1998.

Chen, Scott Shaobing, Donoho, David L., and Saunders, Michael A. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, January 2001. ISSN 0036-1445.

Cong, Y., Yuan, J., and Liu, J. Sparse reconstruction cost for abnormal event detection. In *CVPR*, pp. 3449–3456, 2011.

Donoho, D. L. and Elad, M. *Optimally Sparse Representation in General (non-orthogonal) Dictionaries Via L1 Minimization*. Department of Statistics, Stanford University, 2002.

Donoho, D. L. and Elad, M. On the stability of the basis pursuit in the presence of noise. *Signal Process.*, 86(3): 511–532, 2006.

Donoho, D. L., Elad, M., and Temlyakov, V. N. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1): 6–18, 2006a.

Donoho, David L., Tsaig, Yaakov, Drori, Iddo, and luc Starck, Jean. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. Technical report, 2006b.

Donoho, D.L., Tsaig, Y., Drori, I., and Starck, J.-L. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 58(2):1094–1121, Feb 2012.

Elad, M. and Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15:3736–3745, 2006.

Engan, K., Aase, S. O., and Husoy, J. H. Frame based signal compression using method of optimal directions (mod). In *ISCAS*, volume 4, pp. 1–4, 1999.

Jiang, Z., Lin, Z., and Davis, L. S. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR*, pp. 1697–1704, 2011.

Jiang, Z., Lin, Z., and Davis, L. S. Label consistent k-svd: Learning a discriminative dictionary for recognition. *PAMI*, 35(11):2651–2664, 2013.

Kratz, L. and Nishino, K. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, pp. 1446–1453, 2009.

Lu, C., Shi, J., and Jia, J. Abnormal event detection at 150 fps in matlab. In *ICCV*, pp. 2720–2727, 2013.

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. Anomaly detection in crowded scenes. In *CVPR*, pp. 1975–1981, 2010.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. Supervised dictionary learning. In *NIPS*, pp. 1033–1040, 2008a.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. Discriminative learned dictionaries for local image analysis. In *CVPR*, pp. 1–8, 2008b.

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

Michal, A., Michael, E., and Alfred, B. K-svd: Design of dictionaries for sparse representation. In *SPARS'05*, pp. 9–12, 2005.

Murray, J.F. and Kreutz-Delgado, K. An improved focuss-based learning algorithm for solving sparse linear inverse problems. In *Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on*, volume 1, pp. 347–351 vol.1, Nov 2001.

Needell, D. and Tropp, J. A. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. ACM Technical Report 2008-01, California Institute of Technology, Pasadena, July 2008.

Needell, D. and Vershynin, R. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):310–316, 2010.

Olshausen, B. A. and Fieldt, D. J. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research*, 37:3311–3325, 1997.

Pati, Y.C., Rezaiifar, R., and Krishnaprasad, P.S. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pp. 40–44 vol.1, Nov 1993.

Qiu, Q., Jiang, Z., and Chellappa, R. Sparse dictionary-based representation and recognition of action attributes. In *ICCV*, pp. 707–714, 2011.

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

Wright, J., Yang, A.Y., Ganesh, A., Sastr, S.S., and Ma, Y. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2009.

Yang, M., Zhang, L., Yang, J., and Zhang, D. Robust sparse coding for face recognition. In *CVPR*, pp. 625–632, 2011.

Yuan, Guo-Xun, Chang, Kai-Wei, Hsieh, Cho-Jui, and Lin, Chih-Jen. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *Journal of Machine Learning Research*, 11:3183–3234, 2010.

Zepeda, J., Guillemot, C., and Kijak, E. Image compression using sparse representations and the iteration-tuned and aligned dictionary. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):1061–1073, 2011.

Zhang, Tong. Some sharp performance bounds for least squares regression with l1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 10 2009. doi: 10.1214/08-AOS659.