Robust K-SVD: A Novel Approach for Dictionary Learning

Carlos A. Loza[1][0000−0002−0509−5104]

Department of Mathematics, Universidad San Francisco de Quito, Quito, Ecuador
cloza@usfq.edu.ec

Abstract. A novel criterion to the well-known dictionary learning technique, K-SVD, is proposed. The approach exploits the L1-norm as the cost function for the dictionary update stage of K-SVD in order to provide robustness against impulsive noise and outlier input samples. The optimization algorithm successfully retrieves the first principal component of the input samples via greedy search methods and a parameter-free implementation. The final product is Robust K-SVD, a fast, reliable and intuitive algorithm. The results thoroughly detail how, under a wide range of noisy scenarios, the proposed technique outperforms K-SVD in terms of dictionary estimation and processing time. Recovery of Discrete Cosine Transform (DCT) bases and estimation of intrinsic dictionaries from noisy grayscale patches highlight the enhanced performance of Robust K-SVD and illustrate the circumvention of a misplaced assumption in sparse modeling problems: the availability of untampered, noiseless, and outlier-free input samples for training.

Keywords: Dictionary Learning · K-SVD · Robust Estimation.

1 Introduction

Sparse modeling constitutes an advantageous framework for applications where sparsity and parsimonious representations are favored, e.g. data compression, image processing and high-dimensional statistics are some of the fields that exploit its inherent concepts. The modeling itself is usually compartmentalized into two very distinctive stages: sparse coding and dictionary learning; while the former strives to represent the input signal as a combination (usually linear) of a few elements, known as bases or atoms (e.g. the JPEG compression standard), the latter learns the set of such overcomplete generating atoms, i.e. a dictionary, in a data-driven scheme.

The sparse coding problem is usually solved by either exploiting a surrogate of the L0-pseudonorm that characterizes sparse decompositions, e.g. L1-norm-based convex optimization programs, such as Basis Pursuit [4], or a greedy approach that usually yields a suboptimal, but rather more tractable, solution: Matching Pursuit (MP) or any of its variants [13, 15]. The dictionary learning estimation is generally solved via probabilistic approaches [11] or generalized clustering [7, 1]. K-SVD, one of the clustering-based methods, is arguably the most widely utilized and recognized dictionary learning algorithm in the literature [6, 12, 3]. Its core optimization can be summarized as an alternation between sparse coding and dictionary update stages.

However, K-SVD implicitly relies on second-order statistics via the Singular Value Decomposition (SVD) or Principal Component Analysis (PCA) alternating update stage. This approach, although principled and practical, might yield erroneous estimations under the presence of additive non-Gaussian noise, e.g. impulsive noise. In addition, outlier samples can easily bias the dictionaries due to the equal weight policy of the Minimum Squared Error (MSE) criterion. A well-known alternative against outliers is to substitute MSE fidelity terms by L1-norms of the estimation errors. Therefore, the main contribution of this manuscript is to incorporate robustness into the dictionary learning framework by exploiting the L1-norm as the optimization criterion in the SVD update stage of K-SVD.

In practical terms, unlike regular SVD, L1-norm-based PCA does not have a closed form solution and numerical methods are usually required. Ding et al. proposed R1-PCA in a successful attempt to obtain robust principal components, however, the method is highly dependent on the dimension m of a surrogate subspace [5]. On the other hand, in [2, 9], the authors exploit a probabilistic approach with Laplacian priors to perform a L1-norm-based decomposition; nevertheless, they are both limited in practice due to reliance on particular heuristics or use of linear and quadratic programs, respectively. In terms of sparse modeling, Mukherjee et al. developed a L1-based K-SVD variant by solving a reweighted L2-norm problem; yet, the comparison to baseline K-SVD might be biased due to the disparity in sparse coding algorithms, i.e. iteratively reweighted least squares (IRLS) for the proposed method and Orthogonal Matching Pursuit (OMP) for K-SVD [14]. In the present work, the fact that K-SVD needs to only estimate the first principal component (for each atom) is exploited by using the algorithm proposed by Kwak [10]; this technique provides a fast, efficient and reliable methodology to estimate the eigenvector with largest L1 dispersion (in feature space). In this way, the proposed final dictionary learning technique, known as Robust K-SVD, is able to estimate overcomplete patterns in a robust and fast scheme that empirically seems to even alleviate the computational burdens that the state of the art entails.

The rest of the paper is organized as follows: Section 2 introduces the concept of robustness to the K-SVD formulation and details the necessary conditions, optimization criteria and algorithms. Section 3 focuses on two types of experiments alongside their results and discussion. Lastly, Section 4 concludes the paper.

2 Robust K-SVD

Let $X = [\mathbf{x_1}, \ldots, \mathbf{x_n}] \in \mathrm{I\!R}^{d \times n}$ be the collection of n zero-mean d-dimensional vectors. L2-norm-based PCA attempts to find an m-dimensional linear subspace

($m < d$), such that the variance is maximized (or the MSE is minimized); this task is accomplished by solving (1):

$$J_2(W,V) = ||X - WV||_2^2 \tag{1}$$

where $W \in \mathbb{R}^{d\times m}$ is a projection matrix with columns $\{\mathbf{w}_k\}_{k=1}^{m}$ known as bases of the m-dimensional subspace, i.e. feature space. $V \in \mathbb{R}^{m\times n}$ is the corresponding coefficient matrix, and $|| \cdot ||_2$ denotes the L2-norm operator. The global minimum of (1) is achieved via SVD which also corresponds to the solution of the dual problem:

$$W^* = \arg\max_W ||W^T S_x W||_2 = \arg\max_W ||W^T X||_2 \tag{2}$$
$$\text{subject to} \quad W^T W = I_m$$

where S_x is the covariance matrix of X and I_m is the $m \times m$ identity matrix. Usually, the solution of (2) is proper and tractable. Nevertheless, it is well-known that the squared L2-norm is sensitive to outliers; thus, it is necessary to appeal to L1-norm-based optimization to mitigate such effect.

2.1 SVD Based on L1-norm Maximization

Instead of minimizing the squared L2-norm of the error, the following criterion is adopted:

$$J_1(W,V) = ||X - WV||_1 \tag{3}$$

where $||\cdot||_1$ denotes the L1-norm operator. In [10] it was noted that optimizing (3) is rather difficult; thus, it is posited that instead of minimizing J_1 in the original d-dimensional input space, it would be more advantageous to maximize the L1 dispersion in the feature space as follows:

$$W^* = \arg\max_W ||W^T X||_1 \tag{4}$$
$$\text{subject to} \quad W^T W = I_m$$

To obtain a local minimizer, [10] proposes a greedy method where, for $m = 1$, Equation (4) becomes:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} ||\mathbf{w}^T X||_1 = \arg\max_{\mathbf{w}} \sum_{i=1}^{n} |\mathbf{w}^T \mathbf{x}_i| \tag{5}$$
$$\text{subject to} \quad ||\mathbf{w}||_2 = 1$$

For the remaining $m - 1$ principal components, another greedy search is proposed, however, it is not addressed here. Algorithm 1 guarantees finding a local minimizer of (5) by leveraging the fact that $\sum_{i=1}^{n} |\mathbf{w}^T(t)\mathbf{x}_i|$ is a non-decreasing function of t (for further details, refer to [10]). In practice, a stopping threshold that tracks the norm of successive estimations is utilized.

Algorithm 1 PCA-L1 [10].

```
Input: $X = [\mathbf{x_1}, \ldots, \mathbf{x_n}] \in \mathbb{R}^{d \times n}$
Output: $\mathbf{w} \in \mathbb{R}^d$
  $\mathbf{w}(0) \leftarrow \mathbf{w}(0)/||\mathbf{w}(0)||$    $t \leftarrow 0$
  repeat
    for $i = 1, \ldots, n$ do
      if $\mathbf{w}^T(t)\mathbf{x}_i < 0$ then
        $p_i(t) = -1$
      else
        $p_i(t) = 1$
      end if
    end for
    $t \leftarrow t + 1$
    $\mathbf{w}(t) = \sum_{i=1}^n p_i(t-1)\mathbf{x}_i$
    $\mathbf{w}(t) \leftarrow \mathbf{w}(t)/||\mathbf{w}(t)||_2$
  until convergence
```

2.2 Robust L1-norm-based Dictionary Learning

K-SVD was proposed by Aharon et al. as a generalization of K-means [1]. Given a set of observations $Y = \{\mathbf{y}_i\}_{i=1}^N$, ($\mathbf{y}_i \in \mathbb{R}^n$), the estimation objective is to find a set of overcomplete patterns, atoms, or bases, i.e. a dictionary $D \in \mathbb{R}^{n \times K}$, that is able to sparsely encode the inputs (in a linear fashion):

$$\min_{D,X}\{||Y - DX||_2^2\} \qquad \text{subject to} \qquad \forall i,\ ||\mathbf{x}_i||_0 \leq T_0 \tag{6}$$

where T_0 denotes the number of non-zero entries in the sparse representation vector $\mathbf{x}_i$, i.e. i-th column of X. Likewise K-means, K-SVD alternates between input assignment and centroids update stages. The former utilizes any of the standard sparse coding algorithms, e.g. MP, OMP, while the latter updates the dictionary atoms via SVD operations. The update stage assumes that both X and $k-1$ columns of D are fixed, then, the atom in question, $\mathbf{d}_k$, alongside its support in X, i.e. $\mathbf{x}_T^k$ (k-th row in X) are jointly updated as follows:

$$||Y - DX||_2^2 = \left|\left|Y - \sum_{j=1}^{K} \mathbf{d}_j \mathbf{x}_T^j\right|\right|_2^2 = \left|\left|(Y - \sum_{j \neq k} \mathbf{d}_j \mathbf{x}_T^j) - \mathbf{d}_k \mathbf{x}_T^k\right|\right|_2^2 = ||E_k - \mathbf{d}_k \mathbf{x}_T^k||_2^2 \tag{7}$$

where E_k is the error when the k-th atom is removed. In order to preserve the sparsity of the solution, it is necessary to restrict the support of E_k to the columns that are currently using the atom $\mathbf{d}_k$; this shrinking operation results in E_k^R which is the matrix to be linearly decomposed via SVD. The resulting updated atom is the first eigenvector (sorted by largest variance). In order to guarantee robustness against impulsive noise and outliers, the MSE criterion is

Algorithm 2 Robust K-SVD

```
Input: $Y \in \mathrm{I\!R}^{n\times N}, D \in \mathrm{I\!R}^{n\times K}, T_0$
Output: $D \in \mathrm{I\!R}^{n\times K}$
  repeat
    Sparse Coding Stage (standard algorithms can be utilized, e.g. MP, OMP):
    $X \leftarrow \mathrm{SpCod}(Y, D, T_0)$
    Dictionary Update Stage:
    for $k = 1, 2 \ldots K$ do
      $w_k \leftarrow \{i|1 \leq i \leq N, \mathbf{x}_T^k(i) \neq 0\}$
      $E_k \leftarrow Y - \sum_{j\neq k} \mathbf{d}_j \mathbf{x}_T^j$
      $\Omega_k \in \mathrm{I\!R}^{N\times|w_k|}$    s.t.    $\Omega_k(w_k(i), i) = 1$
      $E_k^R \leftarrow E_k \Omega_k$
      $\mathbf{d}_k \leftarrow \text{PCA-L1}(E_k^R)$
    end for
  until convergence
```

replaced by the L1-norm-based PCA algorithm described in the previous subsection [10]. The final result, Robust K-SVD, is summarized in Algorithm 2. In practice, the algorithm stops after either surpassing a threshold in successive estimations or reaching a fixed number of iterations. One of the main advantages of the proposed approach is that K-SVD, by only exploiting the first eigenvector, does not require the remaining $m - 1$ bases, which is exactly what Algorithm 1 provides in a fast and efficient implementation. This suggests a clear leverage over other techniques that require the full estimation of the bases.

3 Results and Discussion

3.1 Recovery of Orthogonal Bases

The first set of experiments focuses on linear combinations of 16-dimensional DCT bases. Specifically, the atoms are linearly combined ($T_0 = 4$) in a random fashion using coefficients from a uniform distribution between -1 and 1. Out of the 5000 generated samples, impulsive noise is added to 10% of them (SNR from -30 dB to -15 dB). Then, both K-SVD and Robust K-SVD estimate the dictionary with the following model parameters: $K = 16$, $T_0 = 4$, 10^{-3} as convergence criterion for PCA-L1, 20 alternate iterations between sparse coding and dictionary update stages, and 25 different trials for each noise scenario. Figure 1 displays the average normalized cross-correlation coefficient for all the noise cases and combinations between sparse coding mechanisms and dictionary update approaches. It is clear that Robust K-SVD outperforms K-SVD and provides a principled scheme to deal with impulsive noise.

The second scenario fixes the impulsive noise SNR to -20 dB and varies the impulsive noise rate. Again, Figure 1 confirms that Robust K-SVD is less sensitive to outlier-like, contaminated samples. It is worth mentioning that these comparisons are not biased by the choice of sparse coding mechanisms, i.e. both

MP and OMP are used for both dictionary learning techniques; thus, it is explicit that the improved performance is solely due to the Robust K-SVD algorithm. Also for this case, there is marginal differences between MP and OMP because the original generating dictionary is orthogonal, i.e. MP reduces to OMP.

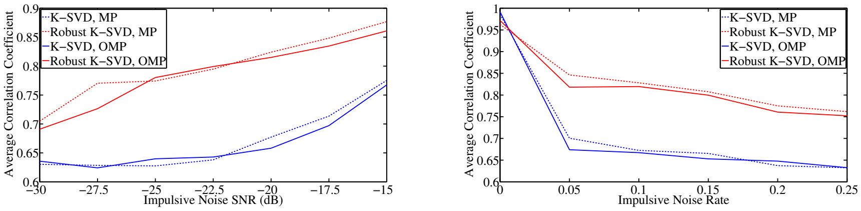

Fig. 1: Recovery of DCT bases. Average normalized cross-correlation between original and estimated dictionaries under impulsive noise. Left: Variable SNR, noise rate of 0.1. Right: Variable rates, noise SNR = -20 dB.

3.2 Dictionary Learning on Grayscale Image Patches

The second set of experiments utilizes the Yale Face Database [8] to extract 10000 random 8×8 grayscale patches in order to estimate an intrinsic generating dictionary. For this case, outlier samples were simulated by blocks of salt and pepper noise, i.e. blocks with black and white pixels. The size of the outlier blocks was varied from 1×1 (only one pixel is affected in the sample) until 8×8 (all pixels in the sample are distorted). The number of outlier samples are modified as well for rate values between 0 (no noise), until 0.25 (25% of the samples are perturbed by noise). In addition, each 8×8 grayscale patch is vectorized into a 64-dimensional zero-mean sample. Lastly, out of the 10000 available blocks, 80% go under the salt and pepper noise treatment, while the remaining clean 20% are reserved for testing and model quantification.

Both K-SVD and Robust K-SVD begin the estimation process with the same initial seed dictionary and utilize the same sparse coding algorithm (OMP). Values of $K = 400$ and $T_0 = 10$ are set. A total of 20 trials per noise scenario are simulated. Lastly, 30 sequential sparse coding and dictionary update stages are ran for each case. The normalized L2-norm of the reconstruction error on the test set is chosen as the metric of success. Figure 2 illustrates the effect of outliers in the dictionary learning process for the 1×1 and 8×8 noisy blocks cases: K-SVD is clearly more biased than Robust K-SVD when outliers are present. The size of the outlier block has a clear effect on the estimation as well. Lastly, it is remarkable how Robust K-SVD outperforms the baseline even for the 0 rate case (no salt and pepper noise), i.e. the proposed method is even able to discard the effect of potential outliers inherent to real-world signals. This clearly suggests that Robust K-SVD is a suitable alternative to K-SVD even when no impulsive noise is explicitly present.

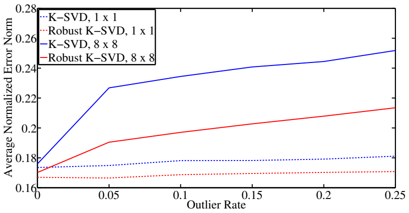

Fig. 2: Estimation of intrinsic dictionary for grayscale image patches. Average L2-norm of reconstruction error (normalized by L2-norm of input) as function of sizes of salt and pepper noisy blocks and rates.

Table 1 summarizes the results for the rest of the noisy block sizes and a particular outlier rate. Again, Robust K-SVD outperforms the corresponding baseline case and shows a natural degradation as the outlier samples grow in influence. It is worth mentioning that Robust K-SVD consistently showcased a faster processing time that regular K-SVD; e.g. in the 8×8 case, K-SVD needed an average of 1.58 seconds in the dictionary update stage (*svds* MATLAB routine) while Robust K-SVD spent 0.59 seconds in the same task (iMac 2.7 GHz Intel Core i5, 8 GB memory). This empirically suggests that Robust K-SVD is not only superior in performance, but also less computationally demanding than the state of the art. In the spirit of openness, the MATLAB code of the proposed algorithm can be found in `https://github.com/carlosloza/Robust_KSVD`.

Table 1: Average L2-norm of reconstruction error as function of sizes of salt and pepper noisy blocks (0.1 rate case).

Noisy Block Size	**1×1**	**2×2**	**3×3**	**4×4**	**5×5**	**6×6**	**7×7**	**8×8**
K-SVD	0.178	0.183	0.189	0.199	0.209	0.220	0.229	0.234
Robust K-SVD	0.168	0.169	0.171	0.175	0.182	0.190	0.195	0.197

4 Conclusion

Robust K-SVD is able to incorporate robustness into the sparse modeling framework by substituting MSE-based SVD operations with robust and fast estimation of principal components via L1-norm optimization. The results not only illustrate the expected safeguard against impulsive noise and outliers, but they also indicate that Robust K-SVD is suitable for problems framed as noiseless. In addition, we empirically prove that Robust K-SVD is faster and computationally less demanding than K-SVD. This opens the possibility of exploiting the proposed algorithm in high-dimensional scenarios where SVD computations are clearly prohibitive.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing **54**(11), 4311–4322 (2006). https://doi.org/10.1109/TSP.2006.881199
2. Baccini, A., Besse, P., Falguerolles, A.: A L1-norm pca and a heuristic approach. Ordinal and symbolic data analysis **1**(1), 359–368 (1996)
3. Bryt, O., Elad, M.: Compression of facial images using the K-SVD algorithm. Journal of Visual Communication and Image Representation **19**(4), 270–282 (2008). https://doi.org/10.1016/j.jvcir.2008.03.001
4. Chen, S., Donoho, D., Saunders, M.: Atomic decomposition by basis pursuit. SIAM review **43**(1), 129–159 (2001). https://doi.org/10.1.1/jpb001
5. Ding, C., Zhou, D., He, X., Zha, H.: R1-pca: rotational invariant L1-norm principal component analysis for robust subspace factorization. In: Proceedings of the 23rd international conference on Machine learning. pp. 281–288. ACM (2006). https://doi.org/10.1145/1143844.1143880
6. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image processing **15**(12), 3736–3745 (2006). https://doi.org/10.1109/TIP.2006.881969
7. Engan, K., Aase, S.O., Husoy, J.H.: Method of optimal directions for frame design. In: Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on. vol. 5, pp. 2443–2446. IEEE (1999). https://doi.org/10.1109/ICASSP.1999.760624
8. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE transactions on pattern analysis and machine intelligence **23**(6), 643–660 (2001). https://doi.org/10.1109/34.927464
9. Ke, Q., Kanade, T.: Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 739–746. IEEE (2005). https://doi.org/10.1109/CVPR.2005.309
10. Kwak, N.: Principal component analysis based on L1-norm maximization. IEEE transactions on pattern analysis and machine intelligence **30**(9), 1672–1680 (2008). https://doi.org/10.1109/TPAMI.2008.114
11. Lewicki, M.S., Olshausen, B.A.: Probabilistic framework for the adaptation and comparison of image codes. JOSA A **16**(7), 1587–1601 (1999). https://doi.org/10.1364/JOSAA.16.001587
12. Loza, C.A., Principe, J.C.: A robust maximum correntropy criterion for dictionary learning. In: Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on. pp. 1–6. IEEE (2016). https://doi.org/10.1109/MLSP.2016.7738898
13. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Transactions on signal processing **41**(12), 3397–3415 (1993). https://doi.org/10.1109/78.258082
14. Mukherjee, S., Basu, R., Seelamantula, C.S.: L1-K-SVD: A robust dictionary learning algorithm with simultaneous update. Signal Processing **123**, 42–52 (2016). https://doi.org/10.1016/j.sigpro.2015.12.008
15. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Transactions on information theory **53**(12), 4655–4666 (2007). https://doi.org/10.1109/TIT.2007.909108

View publication stats