2016 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 13–16, 2016, SALERNO, ITALY

A ROBUST MAXIMUM CORRENTROPY CRITERION FOR DICTIONARY LEARNING

Carlos A. Loza, Jose C. Principe

Computational NeuroEngineering Laboratory
University of Florida
Gainesville, FL

ABSTRACT

We introduce a method that incorporates robustness to one of the main building blocks of sparse modeling: dictionary learning. Particularly, we exploit correntropy to compute the principal components in cases where outliers might be detrimental without proper care. This is further added to one of the most utilized dictionary learning tools: K-SVD; the result is Correntropy K-SVD, or CK-SVD, a method that is based on a Maximum Correntropy Criterion (MCC) instead of the somewhat limited Minimum Squared Error (MSE) approach. The optimization is performed using the well-known Half-Quadratic (HQ) technique, which allows a fast and efficient implementation. The results show the importance of this work not only by outperforming K-SVD, but also by circumventing one of the main assumptions during learning overcomplete representations: the availability of untampered, noiseless and outlier-free samples for training stages.

Index Terms— Correntropy, Dictionary Learning, Half-Quadratic Optimization, Singular Value Decomposition, Sparse Modeling

1. INTRODUCTION

Sparse Modeling has become one of the most attractive frameworks for applications such as signal processing, computer vision, and data mining where a parsimonious representation is considered advantageous. For instance, image compression, denoising, and optimization algorithms are some of the areas that have successfully incorporated sparsity as a paramount feature [1, 2, 3]. Sparse coding and dictionary learning constitute the two main building blocks in sparse modeling. The former strives to find a minimal decomposition given a set of bases or dictionary, e.g. JPEG compression using a Discrete Cosine Transform (DCT) basis, while the latter aims to learn usually overcomplete, high-dimensional vectors given a set of training samples and additional constraints that favor sparsity.

Furthermore, the sparse coding problem is well-known combinatorial; hence, a possible alternative is finding a local optimum solution via greedy methods, such as Matching Pursuit (MP) [4] or one of its many variants [5, 6, 7]. These algorithms are usually simple and effective due to the sequential inner product operations they require. On the other hand, the dictionary learning estimation is usually performed using a probabilistic approach [8, 9] or a generalized clustering scheme [10, 11]. Of these approaches, K-SVD is one of the most efficient and widely utilized algorithms for dictionary learning due to its flexibility and evident generalization of the well-known K-means clustering technique [12, 13].

However, K-SVD heavily relies on Singular Value Decompositions (SVD) to iteratively update the dictionary elements or atoms. This approach, although principled, might yield erroneous estimations when non Additive White Gaussian Noise (AWGN) is present. In particular, outliers could potentially deviate the estimated principal components and sequentially affect all the subsequent bases and iterations in a domino effect. This is a direct consequence of working under a second-order estimation framework, i.e. Minimum Squared Error (MSE), where SVD decompositions yield analytical solutions and optimal results. Hence, we propose correntropy [14] as a feasible alternative that exploits higher-order statistics of the data and overcomes the MSE limitations. In this way, robustness against high-tailed impulsive noise can be guaranteed in the K-SVD dictionary learning procedure.

In practice, unlike SVD, there is no closed-form solution to the correntropy-based principal component decomposition; hence, we utilize the Half-Quadratic technique [15] as a fast and efficient optimization approach. In particular, HQ optimization has successfully worked in a MCC-based principal component scheme before according to He el at. [16]. Moreover, this method has proved to outperform robust techniques that utilize a L_1 norm regularization in image processing applications [17]. Consequently, we incorporate this inherent robustness of correntropy to the parsimonious constraints of the dictionary learning problem and introduce a novel addition to the sparse modeling literature.

The rest of the paper is organized as follows: Section 2 introduces the correntropy measure and describes the HQ implementation of a robust Principal Component Analysis (PCA) using correntropy as cost function. Section 3 discusses

This work was supported by Michael J. Fox Grant 9558

978-1-5090-0746-2/16/$31.00

the modifications of the K-SVD algorithm to incorporate correntropy as part of the dictionary learning scheme. Section 4 presents results on two types of data with different noise environments, and finally, Section 5 concludes the paper with further research directions.

2. MCC-BASED ROBUST PCA VIA HALF-QUADRATIC OPTIMIZATION

Correntropy was proposed by Liu, Pokharel, and Principe as an alternative measure beyond second-order statistics paradigms and Gaussian environments. Specifically, cross-correntropy, or simply correntropy for two random variables X and Y is defined as:

$$V_\gamma(X, Y) = \mathbf{E}[\kappa_\gamma(X - Y)] \tag{1}$$

where $\kappa_\gamma(X - Y)$ represents a kernel operator with parameter vector γ. In most applications, the Gaussian kernel is preferred due to its computational tractability, single-parameter property ($\gamma = \sigma$), and compliance with the Mercer's Theorem [18]. Furthermore, by non-linearly mapping the input data to a reproducing kernel Hilbert space (RKHS), correntropy is considered as a second-order statistic in a higher-dimensional space and is able to incorporate statistical moments beyond variance. This allows to assess independence in a more strict sense, i.e. uncorrelatedness in the mapped space is theoretically translated to independence in the input space. These properties have placed correntropy as an attractive tool in the signal processing and machine learning fields with relevant applications including non-linear analysis, robust filtering and image processing [17, 19, 20, 21, 22].

In practice, correntropy is estimated utilizing averages of the available samples of the random variables X and Y, i.e. $\{(x_i, y_i)_{i=1}^N\}$ and the Gaussian kernel, $G_\sigma(x)$:

$$\hat{V}_\sigma(X, Y) = \frac{1}{N} \sum_{i=1}^{i=N} G_\sigma(x_i - y_i) \tag{2}$$

Now in order to incorporate correntropy in the PCA formulation, we follow a similar approach as He et. al [16] and define the cost function $J(\theta)$ as:

$$J(\theta) = \sum_{i=1}^{N} G_\sigma(x_i - \mu - U v_i) \tag{3}$$

where $y_i = \mu - U v_i$ and $\theta \triangleq (\mu, U)$. Particularly, θ encompasses the estimated mean of the data along with the projection matrix U that will contain the principal eigenvectors of the distribution. Maximizing (3) is also known as the Maximum Correntropy Criterion or MCC [14]. Provided that U is orthonormal and substituting $v_i = U^T(x_i - \mu)$ into (3) using the projection theorem, we obtain the following optimization:

$$\max_\theta J_{HQ}(\theta) = \sum_{i=1}^{N} G_\sigma\Big(\sqrt{x_i^{\mu T} x_i^\mu - x_i^{\mu T} U U^T x_i^\mu}\Big) \tag{4}$$

where $x_i^\mu = x_i - \mu$. In order to solve (4), we exploit the advantages of the Half-Quadratic technique [15] and reformulate the cost function to introduce an enlarged parameter space:

$$\hat{J}_{HQ}(\theta, p) = \sum_{i=1}^{N} \Big(p_i\Big(x_i^{\mu T} x_i^\mu - x_i^{\mu T} U U^T x_i^\mu\Big) - \phi(p_i)\Big) \tag{5}$$

where p is a scalar variable and $\phi(x)$ is a convex conjugated function of $G_\sigma(x)$. In particular, for a fixed x, the optimum value for this constraint is obtained at $p = -G_\sigma(x)$. Moreover, for a fixed θ, the following equation holds true:

$$\max_\theta J_{HQ}(\theta) = \max_{\theta, p} \hat{J}_{HQ}(\theta, p) \tag{6}$$

Hence, maximizing J_{HQ} is equivalent to maximizing $\hat{J}_{HQ}$. Furthermore, it is possible to optimize (5) in an efficient alternating scheme that is fully described in Algorithm 1. We have used $E \in R^{n \times d}$ instead of X and d instead of N to comply with the notation of the next section. Also, $g_\sigma(x)$ is the normalized Gaussian kernel, i.e. $g_\sigma(x) \triangleq \exp(-x^2/2\sigma^2)$, and the m_r parameter is introduced to control the number of principal components to be computed, i.e. $1 < m_r < d$.

Algorithm 1 Robust PCA via HQ

Input: $E \in R^{n \times d}, \epsilon, \mu \in R^n, U^1 \in R^{n \times m_r}$
Output: $\mu, U \in R^{n \times m_r}$

$J \leftarrow 1$
while convergence == FALSE **do**
  $r \leftarrow \|(e_i - \mu) - U^J (U^J)^T (e_i - \mu))\|^2 \qquad i = 1, \ldots, d$
  $\sigma \leftarrow 1.06 \times \min\{\text{std}(r), \text{IQR}(r)/1.34\} \times d^{-1/5}$
  $e_i^\mu \leftarrow e_i - \mu \qquad i = 1, \ldots, d$
  $p_i \leftarrow -g_\sigma(\sqrt{e_i^{\mu T} e_i^\mu - e_i^{\mu T} (U^J)(U^J)^T e_i^\mu}) \qquad i = 1, \ldots, d$
  $\mu \leftarrow (\sum_{i=1}^d p_i x_i)/(\sum_{i=1}^d p_i)$
  $E_c \leftarrow [x_1 - \mu, x_2 - \mu, \ldots, x_d - \mu]$
  $P \leftarrow \text{diag}(-p)$
  $U^{J+1} \leftarrow \text{PCA}(E_c P E_c^T, m_r)$
  if $\|u_i^J - u_i^{J+1}\| < \epsilon \qquad i = 1, \ldots, m_r$ **then**
    convergence = TRUE
    $U \leftarrow U^{J+1}$
  else
    convergence = FALSE
    $J \leftarrow J + 1$
  end if
end while

Additionally, Algorithm 1 requires a stopping criterion threshold that, in this case, is based on L_2 norms of successive

estimated principal components. Also, the initializations of μ and U are necessary as well. These can be, for instance, sample estimators of the mean and bases obtained using regular PCA, respectively. Lastly, it is worth mentioning that Algorithm 1 includes a recurrent estimator of the kernel parameter, σ, that is based on the Silverman's rule [23]. In this way, σ is incorporated in the HQ optimization. This has been proved to work well in practice by mimicking kernel annealing and, at the same time, eliminating an additional free parameter of the model.

Fig. 1 shows an illustrative example of the method for $n = 2, d = 200, m_r = 1, \epsilon = 10^{-4}$. Even though the number of outliers is relatively small (10 out of 200 samples), it is clear that regular PCA erroneously estimates the principal component or eigenvector of the original, outlier-free distribution. On the other hand, robust MCC-based PCA accurately recovers the maximum direction of spread in the data. This opens the possibility to utilize this robust framework in applications where PCA or Singular Value Decomposition (SVD) are the gold standard, namely K-SVD, one of the most utilized dictionary learning tools in the sparse modeling community.

3. MCC-BASED K-SVD

K-SVD was proposed by Aaron, Elad, and Bruckstein [11] as a generalization of the well-known K-means clustering algorithm [24]. Basically, given a set of observations $Y = \{y_i\}_{i=1}^{N}$, ($y_i \in R^n$), the goal is to find a set of bases or atoms, also known as dictionary, $D \in R^{n \times K}$, along with a sparse representation, $X \in R^{K \times N}$, given a particular constraint, i.e.:

$$\min_{D,X}\{\|Y - DX\|_F^2\} \qquad \text{subject to} \qquad \forall i,\ \|x_i\|_0 \leqslant T_0 \tag{7}$$

where T_0 determines the number of non-zero entries in the sparse representation vector x_i, i.e. i-th column of X.

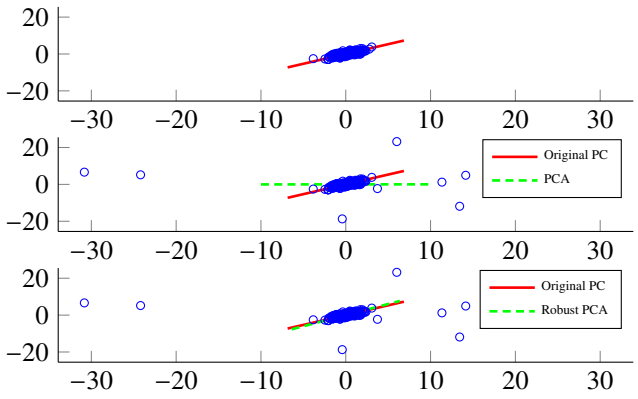

Fig. 1: MCC-based Robust PCA example. From top to bottom: Original Gaussian pdf (First Principal Component in red). 20 outliers (5%) are added and PCA is computed. Same distribution with MCC-based Robust PCA eigenvector.

Similarly to K-means, in K-SVD there are two well-defined stages: Sparse Coding and Dictionary Update. The first one utilizes any of the standard sparse coding algorithms, e.g. Matching Pursuit, Orthogonal Matching Pursuit (OMP), and a fixed dictionary to find a sparse representation of the samples in Y; this is analogous to finding the nearest cluster to the input data using a particular metric. However, K-SVD allows contributions from multiple clusters in a linear combination fashion similar to fuzzy clustering.

The Dictionary Update stage resembles the centroid recalculation procedure in K-means, nevertheless, in K-SVD this task is accomplished by Singular Value Decompositions that attempt to minimize the L_2 norm of the residual error for a given basis. Particularly, it is necessary to assume that both X and $k-1$ columns of D are fixed; then, the k-th atom, i.e. d_k, and the elements in X currently associated to such atom, i.e. x_T^k (k-th row in X), are jointly updated via the following formulation:

$$\begin{aligned} \|Y - DX\|_F^2 &= \left\| Y - \sum_{j=1}^{K} d_j x_T^j \right\|_F^2 \\ &= \left\| (Y - \sum_{j \neq k} d_j x_T^j) - d_k x_T^k \right\|_F^2 \\ &= \|E_k - d_k x_T^k\|_F^2 \end{aligned} \tag{8}$$

where E_k is the error when the k-th atom is removed. The optimal solution for (8) in a MSE sense is given by the SVD: $E_k = U\Lambda V^T$. Specifically, the updated d_k will be equal to the first column of U and the sparse representation coefficients will be the result of the first column of V multiplied by the first diagonal element of Λ. This is all assuming the principal components are sorted in a decreasing represented power fashion. Lastly, it is necessary to take into account only the atoms currently using the dictionary element in question; hence, the matrix $\Omega_k \in R^{N \times |w_k|}$ isolates such elements placing ones on its $(w_k(i), i)$th entries and zeros elsewhere. This guarantees the sparsity constraint is not violated.

Although K-SVD has been proved to be efficient under certain environments, it heavily relies on second-order statistics, namely SVD operations which are only optimal under MSE assumptions and Gaussian scenarios. Especially, it is well known that during the learning phase, the input samples are assumed to be high-SNR, untampered, and outlier-free. If these conditions are not met, the dictionary algorithm might yield erroneous estimations as a direct result of lacking robustness. We propose incorporating the MCC-based robust PCA approach to the dictionary learning scheme in order to deal with outliers in a principled way. This is accomplished by implementing MCC-based decomposition instead of the traditional SVD for each one of the dictionary elements updates. The final technique, Correntropy K-SVD, or CK-SVD, is detailed in Algorithm 2.

Algorithm 2 CK-SVD

Input: $Y \in R^{n\times N}, D \in R^{n\times K}, T_0, K$
Output: $D \in R^{n\times K}, X \in R^{K\times N}$

repeat
  Sparse Coding Stage
  $X \leftarrow \text{SpCod}(Y, D, T_0)$
  Dictionary Update Stage
  for $k = 1, 2 \dots K$ **do**
    $w_k \leftarrow \{i | 1 \le i \le N, x_T^k(i) \neq 0\}$
    $E_k \leftarrow Y - \sum_{j\neq k} d_j x_T^j$
    $\Omega_k \in R^{N\times|w_k|}$ s.t. $\Omega_k(w_k(i), i) = 1$
    $E_k^R \leftarrow E_k \Omega_k$
    Solve $E_k^R = U\Lambda V^T$ via Robust PCA
    $d_k \leftarrow$ first column of U
  end for
until convergence

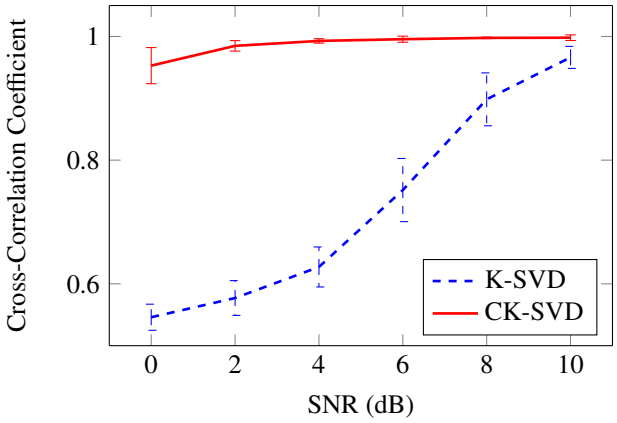

Fig. 2: Comparison of K-SVD and CK-SVD for different baseline SNR values and high-tailed impulsive noise of −20dB in 5% of the samples. Average cross-correlation is computed with respect to noiseless 16-dimensional DCT dictionary.

4. RESULTS

Two sets of results are presented, the first one deals with complete dictionaries and two different types of high-tailed impulsive noise, while the second set shows the potential of the novel algorithm when working with grayscale image patches.

4.1. Recovering Orthogonal Bases

The first experiment uses 16-dimensional DCT bases as the generating, complete dictionary. These atoms are linearly combined in sets of three ($T_0 = 3$) in a random fashion with coefficients drawn from a uniform distribution with range $(-1, 1)$. A total of 1500 cases are generated, i.e. $n = 16$, $N = 1500$. Then, we add impulsive noise of SNR= −20dB to 5% of the samples while the rest 95% is kept at a baseline SNR well above −20dB.

Subsequently, K-SVD and CK-SVD estimate the dictionary using Matching Pursuit, model parameters $K = 16$, $T_0 = 3$, $\epsilon = 10^{-4}$, $m_r = 1$, initial dictionary estimates from regular PCA, and a total of 30 sequential dictionary update iterations for each case. Finally, the baseline SNR was varied from 0 to 10dB while the outlier SNR was kept at −20dB. A total of 100 realizations were simulated per each baseline SNR value. Fig. 2 illustrates the average cross-correlation coefficient between the estimated dictionary atoms and the DCT elements as ground truth. It is evident that CK-SVD outperforms K-SVD for every single baseline SNR. Most importantly, CK-SVD displays less variability across trials, i.e. smaller standard deviations, and consistently high cross-correlation coefficients across different noise floors. This confirms the robustness and efficiency of the proposed algorithm.

For the next experiment, a more challenging scenario is posed: instead of having 5% of the samples influenced by outliers, every single one of the 1500 samples will present a different type of impulsive noise, namely 1 of the 16 entries of each input data sample will be influenced by outliers. This is equivalent to say that roughly 5% of the individual entries of each sample will be an outlier. Another way of presenting this scenario is to say that 1 out of the 16 dimensions of the input space will be randomly affected by impulsive noise in every single sample; however, this dimension will be random for each vector. In comparison, the first problem was equivalent to affecting all 16 dimensions at once for 5% of the population.

The same parameters concerning sparsity were utilized, i.e. $T_0 = 3$, $\epsilon = 10^{-4}$, $m_r = 1$ and Matching Pursuit. Once again, $N = 1500$ and we simulated 100 different trials per outlier scenario along with 30 sequential dictionary update iterations per case. For this experiment, the baseline SNR was kept at 10dB while the high-tailed impulsive noise SNR is varied from −20 to 0dB. Fig. 3 depicts the final results using cross-correlation to the ground truth dictionary. Similarly to the first case, CK-SVD consistently outperforms K-SVD. Furthermore, it is remarkable how CK-SVD constantly remains closer to 1 starting at the −10dB mark, which is the point where the outlier samples are theoretically unaffected by noise (outlier noise is matched by baseline noise), however, K-SVD is unable to fully recover the dictionary atoms and displays higher variability. This again confirms the robustness of CK-SVD. It is worth mentioning that results using OMP were very similar for both experiments due to the fact that the generating dictionary is complete; hence, MP reduces to OMP.

4.2. Gray Scale Image Patches

For the second set of experiments, we utilized 1000 8×8 patches of grayscale images randomly extracted from the Yale Face Database [25]. Particularly for this type of signals, the dynamic range restricts the possibilities for high-tailed impul-

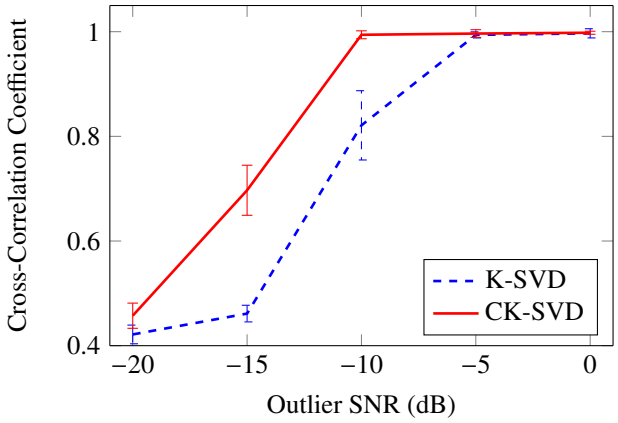

Fig. 3: Comparison of K-SVD and CK-SVD for different high-tailed impulsive noise cases and a constant baseline SNR of 10dB. Every single input data sample has one entry considered as an outlier. Average cross-correlation is computed with respect to noiseless 16-dimensional DCT dictionary.

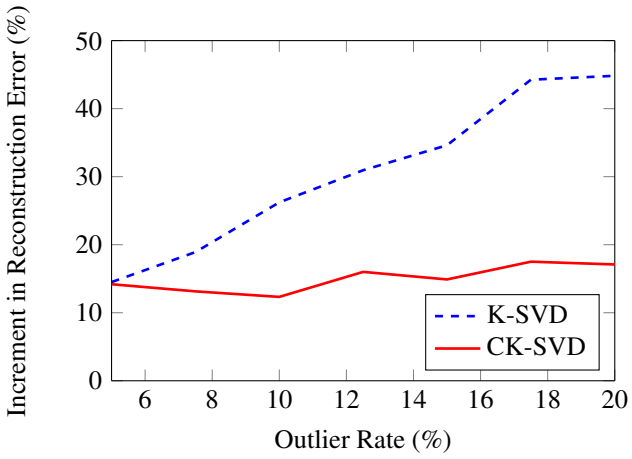

Fig. 4: Percentage of increment in reconstruction error as a function of outlier presence in 8×8 grayscale image patches (4×4 noisy blocks). Reconstruction error is compared to the noiseless case where K-SVD was utilized.

sive noise cases, however, it is possible to simulate outliers with blocks consisting of salt and pepper noise, i.e. blocks with black and white pixels. The number of outlier pixels was varied depending on the size of the noisy square block, e.g. 4, 9, and 16 outlier pixels, and the percentage of outlier-influenced patches was set to 20%.

Next, we estimated the generating dictionary applying K-SVD to the noiseless samples and correlating the atoms to the corresponding K-SVD and CK-SVD results with the outlier-influenced samples as input. Specifically, the parameters were the following: $N = 1000$, $n = 64$, $K = 100$, $T_0 = 5$, $\epsilon = 10^{-4}$, $m_r = 1$. We ran a total of 20 different realizations per outlier scenario along with 50 sequential dictionary update iterations per case. Both MP and OMP were utilized. Table 1 summarizes the average percentage in increase in the reconstruction error, i.e. $\|Y - DX\|_F^2$, with respect to the noiseless K-SVD case. It is evident that CK-SVD performs better than K-SVD, i.e. it is able to recover the noiseless dictionary with minimal error. Another interesting remark is the fact that the OMP cases have a higher increase in reconstruction error than their MP counterparts, regardless of the utilized dictionary learning framework (K-SVD or CK-SVD). This is, however, expected because OMP relies on sequential orthogonal projections that are optimal only under the Gaussian assumption and the MSE cost function, hence, they will be heavily affected by the outliers in the input space.

Finally, we also simulated the cases where the percentage of outliers is varied. For this scenario, we chose blocks of 16 salt and pepper pixels per outlier-influenced sample and compared the resulting dictionaries with the noiseless case. The results are shown in Fig. 4 and indicate a clear consistency across outlier presence for CK-SVD, while, on the other hand, K-SVD increases the reconstruction error in proportion to the percentage of altered pixels.

Table 1: Relative Increment (%) in Reconstruction Error for 20% of outliers

outlier pixels	K-SVD + MP	CK-SVD + MP	K-SVD + OMP	CK-SVD + OMP
4	14.34	11.48	18.58	14.85
9	31.49	13.84	35.33	16.06
16	44.69	15.18	51.43	21.30

5. CONCLUSIONS AND FURTHER WORK

We have derived and implemented a MCC-based dictionary learning approach that exploits the generalization of the K-means clustering algorithm, namely K-SVD. We showed synthetic and digital image examples under different noise environments, and the results highlight the robustness of the algorithm along with a straightforward implementation. Our next interests involve incorporating correntropy to the full sparse modeling framework by taking advantage of previous work that implemented a greedy, robust sparse decomposition method based on the Generalized Correntropy measure [26].

6. REFERENCES

[1] Richard G Baraniuk, "Compressive sensing," *IEEE signal processing magazine*, vol. 24, no. 4, 2007.

[2] Jean-Luc Starck, Emmanuel J Candès, and David L Donoho, "The curvelet transform for image denoising," *Image Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 670–684, 2002.

[3] Peter Bühlmann and Sara Van De Geer, *Statistics for*

high-dimensional data: methods, theory and applications, Springer Science & Business Media, 2011.

[4] Stéphane G Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[5] Joel A Tropp and Anna C Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, 2007.

[6] Deanna Needell and Joel A Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[7] David L Donoho, Yaakov Tsaig, Iddo Drori, and Jean-Luc Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 58, no. 2, pp. 1094–1121, 2012.

[8] Michael S Lewicki and Bruno A Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *JOSA A*, vol. 16, no. 7, pp. 1587–1601, 1999.

[9] Michael S Lewicki and Terrence J Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000.

[10] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. IEEE, 1999, vol. 5, pp. 2443–2446.

[11] Michal Aharon, Michael Elad, and Alfred Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[12] Michael Elad and Michal Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, 2006.

[13] Qiang Zhang and Baoxin Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2691–2698.

[14] Weifeng Liu, Puskal P Pokharel, and José C Príncipe, "Correntropy: properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[15] R Tyrrell Rockafellar, "Convex analysis (princeton mathematical series)," *Princeton University Press*, vol. 46, pp. 49, 1970.

[16] Ran He, Bao-Gang Hu, Wei-Shi Zheng, and Xiang-Wei Kong, "Robust principal component analysis based on maximum correntropy criterion," *Image Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1485–1494, 2011.

[17] Ran He, Wei Shi Zheng, and Bao Gang Hu, "Maximum correntropy criterion for robust face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1561–1576, 2011.

[18] Vladimir Vapnik, *The nature of statistical learning theory*, Springer Science & Business Media, 2013.

[19] Abhishek Singh and Jose C Principe, "Using correntropy as a cost function in linear adaptive filters," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. IEEE, 2009, pp. 2950–2955.

[20] Aysegul Gunduz and Jose C Principe, "Correntropy as a novel measure for nonlinearity tests," *Signal Processing*, vol. 89, no. 1, pp. 14–23, 2009.

[21] Kyu-Hwa Jeong, Weifeng Liu, Seungju Han, Erion Hasanbelliu, and Jose C Principe, "The correntropy mace filter," *Pattern Recognition*, vol. 42, no. 5, pp. 871–885, 2009.

[22] Songlin Zhao, Badong Chen, and Jose C Principe, "Kernel adaptive filtering with maximum correntropy criterion," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011, pp. 2012–2017.

[23] Bernard W Silverman, *Density estimation for statistics and data analysis*, vol. 26, CRC press, 1986.

[24] Allen Gersho and Robert M Gray, *Vector quantization and signal compression*, vol. 159, Springer Science & Business Media, 2012.

[25] Athinodoros S Georghiades, Peter N Belhumeur, and David J Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[26] Carlos A Loza and Jose C Principe, "Generalized correntropy matching pursuit: A novel, robust algorithm for sparse decomposition," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, p. to appear.