Robust Estimation of Shift–Invariant Patterns Exploiting Correntropy

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

Jose C. Principe
Computational NeuroEngineering Laboratory (CNEL)
University of Florida
Gainesville, FL
principe@cnel.ufl.edu

***Abstract*—We propose a novel framework for robust estimation of recurring patterns in time series. Particularly, we utilize correntropy and a shift–invariant adaptation of sparse modeling techniques as the underpinnings of a data–driven scheme where potential outliers, such as spikes, dropouts, high–amplitude impulsive noise, gaps, and overlaps are managed in a principled manner. The Maximum Correntropy Criterion (MCC) is applied to the estimation paradigms and solved via the Half–Quadratic (HQ) technique, which allows a fast and efficient computation of the optimal projection vectors without adding extra free parameters. We also posit a heuristic regarding the initial set of functions to be estimated; specifically, we restrict the search space to patterns with modulatory activity only. We then implement a robust clustering routine to provide a principled initial seed for the greedy algorithms. This heuristic is proved to alleviate the computational burden that shift–invariant unsupervised learning usually entails. The framework is tested on synthetic time series built from weighted Discrete Cosine Transform (DCT) atoms under four different variants of outliers. In addition, we present preliminary results on winding data that illustrate the clear advantages of the methods.**

***Index Terms*—Correntropy, Dictionary Learning, K-SVD, Shift–Invariant, Robust Estimation**

I. Introduction

Robust estimation plays a key role in applications prone to measurement errors, external perturbations, or outliers in general. More specifically, in the time series domain it is not uncommon to find instance perturbation scenarios where outliers are usually the result of artifacts, sensor malfunctioning, communication channel failures and so on. These perturbations can easily bias any statistical estimator based on the available samples and, thus, yield erroneous conclusions regarding the data. An epitome of such case is the unsupervised learning of patterns in time series, also known as time series clustering.

An alternative view to the time series clustering problem is the estimation of a generative set of functions or atoms, i.e. a dictionary, in a data–driven scheme. This approach resembles the dictionary learning techniques developed in the sparse modeling community [1]–[3] with the added constraint of shift–invariance. Yet, this extra condition makes the classic decomposition algorithms improper when working in the time domain: the learned patterns would be multiple shifted replicas of a few true generative atoms. Also, the extra computational burden of working in such a high–dimensional space (theoretically infinite for time series) is basically impractical in most cases. Thus, tractable dictionary learning algorithms that harness the shift–invariance property of the generative functions have been proposed in the literature [4], [5].

Mailhé et al. [4] extended K-SVD [1], arguably one of the most widely utilized dictionary learning algorithm [6]–[8], to the shift–invariant scenario. In particular, instead of regular sparse decomposition algorithms, they proposed a fast convolution–based technique to constraint the support of the active atoms and, subsequently, exploited the regular K-SVD algorithm to update the pattern in question. However, it was never fully addressed how to choose a proper initial dictionary for the greedy optimization. In addition as the name suggests, K-SVD is based on iterative Singular Value Decompositions (SVD), which implicitly optimize a second order cost function, i.e. minimum Mean Squared Error (MSE). This approach, although principled and with closed form solutions, is sensitive to small perturbations in the data that can easily bias the learned generative shift-invariant dictionary. The main contribution of this manuscript is twofold: first, we circumvent the limitations of MSE–based estimation by exploiting correntropy [9], a similarity measure that provides robustness when used as a cost function, and second, we propose a heuristic to choose a suitable initial dictionary that is robust to outliers with the added benefit of significant reduction of the computational burden that regular random samplers usually entail.

In practice, correntropy–based optimization does not have closed–form solutions. Therefore, we exploit the Half–Quadratic technique [10], [11] as an efficient and suitable alternative to find optimal solutions. This approach has been exploited in image processing applications and has even outperformed L1–norm–based algorithms [12], [13]. Consequently, we incorporate this advantageous feature into the parsimonious constraints inherent to sparse modeling and introduce a novel robust framework to the time series clustering literature.

The rest of the paper is organized as follows: Section 2 details the learning algorithm for the time series case alongside the robust nature of it. Section 3 focuses on the heuristic developed for the seed dictionary. Section 4 describes the experimental results on synthetic and real data, and lastly,

This work was supported by Universidad San Francisco de Quito.

Section 5 concludes the paper and discusses further work.

II. Robust Shift–Invariant Dictionary Learning

A. The Shift–Invariant Case

In classic sparse modeling framework, the main goal is to estimate a set of patterns or atoms that are usually overcomplete, i.e. a generative dictionary $\Delta \in \mathbb{R}^{M\times K}$, in a data–driven scheme. The resulting set is able to encode or linearly decompose the inputs $Y = \{\mathbf{y}_i\}_{i=1}^N$, $(\mathbf{y}_i \in \mathbb{R}^M)$ in a sparse code, X, under the following constrained cost function:

$$\min_{\Delta,X} ||Y - \Delta X||_2^2 \qquad \text{subject to} \qquad \forall i,\ ||\mathbf{x}_i||_0 \leq T \quad (1)$$

where T denotes the number of non–zero entries in the sparse representation vector $\mathbf{x}_i$, i.e. i-th column of X.

In the shift–invariant case, the estimation is performed on a single long time series, s, and the dictionary is the result of shifting a set of generative patterns, D, i.e. $D = \{\mathbf{d}_k\}_{k=1}^K$ where $\mathbf{d}_k \in \mathbb{R}^M$. This modified setting yields the following new objective function:

$$\min_{D,\mathbf{c}} \left|\left| s - \sum_k \sum_\tau c_{k,\tau}\mathbf{d}_k(t-\tau) \right|\right|_2^2 \quad \text{s.t.} \quad ||\mathbf{c}||_0 \leq T \quad (2)$$

$$= \min_{D,\mathbf{c}} \left|\left| s - \sum_k \sum_\tau c_{k,\tau} T_\tau \mathbf{d}_k \right|\right|_2^2 \quad \text{s.t.} \quad ||\mathbf{c}||_0 \leq T \quad (3)$$

where T_τ is a shift operator that places a pattern $\mathbf{d}$ at the time instance $t = \tau$ and zeros everywhere else. In this way, $\Delta = \{T_\tau \mathbf{d}_k\}_{k,\tau}$ and the new goal is to estimate the generative set, D, in an unsupervised, data–driven framework.

B. A Greedy Solution

As many sparse decomposition problems, finding the optimal solution is essentially combinatorial. Hence, it is common practice to appeal to greedy algorithms to obtain a more tractable, albeit suboptimal, solution. Here, Matching Pursuit (MP) [14] is the preferred technique: it basically decomposes the input signal in a sequential manner by finding the mostly correlated atom to the current residue. The modified version of MP for time series invokes the cross–correlation operator to find the sparse code alongside the temporal support of the active atoms: $\sigma_k = \{\tau | c_{k,\tau} \neq 0\}$. In addition, the implementation of such cross–correlation–based algorithms can be optimized exploiting Fast Fourier Transform (FFT) heuristics. In this way, the resulting sparse decomposition algorithm in the time domain is both tractable and principled.

Once the sparse codes and temporal supports are estimated, it is necessary to update the current dictionary likewise K-SVD, i.e. successively for each generative function based on the residue, r, and decomposition parameters. For a given atom, let $\hat{s}_\kappa = r + \sum_\tau c_{k,\tau} T_\tau \mathbf{d}_\kappa$ be the signal without the contributions of the set $\{\mathbf{d}_k\}$ for $k \neq \kappa$. Then, the optimal update for the atom in question and its sparse code is:

$$(\mathbf{d}_\kappa^{\text{opt}}, \mathbf{c}_\kappa^{\text{opt}}) = \arg\min_{||\mathbf{d}||_2=1} \left|\left| \hat{s}_\kappa - \sum_{\tau\in\sigma_\kappa} c_\tau T_\tau \mathbf{d} \right|\right|_2^2 \quad (4)$$

If σ_κ does not contain overlapping occurrences, it is possible to exploit the unitary nature of the shift operator, T_τ, and utilize its adjoint, T_τ^*, for reformulating (4). The optimal updated atom (according to MSE criterion) is the singular vector corresponding to the largest singular value of a matrix of patches (from $\hat{s}_\kappa$) where the pattern in question is active:

$$\mathbf{d}_\kappa \leftarrow \arg\max_{||\mathbf{d}||_2=1} \langle \mathbf{d}, T_\tau^* \hat{s}_\kappa \rangle^2 \quad (5)$$

This greedy solution was first proposed in [4]. However, the optimal updated atom (and therefore the dictionary) can be easily biased by outliers, such as spikes, dropouts, gaps, high–amplitude impulsive noise, or even overlaps. Hence, we incorporate robustness into the shift–invariant dictionary learning framework by exploiting correntropy as the cost function in the SVD decompositions.

C. MCC–based Shift–Invariant Dictionary Learning

Correntropy was introduced in [9] as a novel similarity measure that goes beyond second–order statistics and Gaussianity assumptions. Cross–correntropy, or simply correntropy, for two random variables, X and Y, is defined as:

$$V_\gamma(X,Y) = \mathbb{E}[G_\gamma(X-Y)] \quad (6)$$

where $G_\gamma(\cdot)$ is the Gaussian kernel with shape parameter γ. In particular, γ modulates the interactions between samples, e.g. a very large value results in L2-norm interplay, while a very low well-tuned γ mimics L0-pseudonorm type of interactions. For this reason, the induced metric by correntropy is able to handle outliers in a much more principled way than the regular MSE approach which intrinsically gives equal weights to all input samples.

As suggested in [12], the following expression exploits correntropy as the cost function between the input samples, $\{\mathbf{x}_i\}_{i=1}^N$ and their corresponding projection in a lower dimensional space, $\{\mathbf{v}_i\}_{i=1}^N$ via the matrix U:

$$J(U) = \sum_{i=1}^N G_\gamma(\mathbf{x}_i - U\mathbf{v}_i) \quad (7)$$

Maximizing (7) is known as the Maximum Correntropy Criterion or MCC. Provided that U is orthonormal and substituting $\mathbf{v}_i = U^T\mathbf{x}_i$ into (7) utilizing the projection theorem, the goal is to optimize the following cost function:

$$\max_U J_{HQ}(U) = \sum_{i=1}^N G_\gamma\left(\sqrt{\mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T UU^T\mathbf{x}_i}\right) \quad (8)$$

One alternative to solve (8) is to exploit the Half–Quadratic technique that was pioneered in [10] as a plausible regularization for image denoising problems. In particular, it is necessary to introduce an enlarged parameter space defined as:

$$\hat{J}_{HQ}(U,p) = \sum_{i=1}^{N} \left(p_i \left(\mathbf{x}_i^T \mathbf{x}_i - \mathbf{x}_i^T U U^T \mathbf{x}_i \right) - \phi(p_i) \right) \quad (9)$$

where p is an auxiliary variable and $\phi(\mathbf{x})$ is a convex conjugated function of $G_\gamma(\mathbf{x})$ [11]. Particularly, for a fixed $\mathbf{x}$, the optimum value is obtained when $p = -G_\gamma(\mathbf{x})$. Moreover for a fixed U, maximizing $\hat{J}_{HQ}(U,p)$ is equivalent to maximizing $J_{HQ}(U)$. Hence, the final algorithm (Algorithm 1) alternates between optimizing for either p or U while keeping the other parameter fixed. The result is a robust estimator of the projection matrix (set of singular vectors) while maximizing the correntropy between input and features spaces.

Algorithm 1 utilizes a stopping threshold, ϵ, for the norm of the difference between successive estimated singular vectors. It is also worth mentioning that the proposed technique deals with zero-mean input samples and it only learns the first principal component, i.e. the projection vector corresponding to the largest dispersion in the feature space. Also, the initial estimation of U can be provided via regular eigendecomposition routines (EIG$(\cdot)$ in Algorithm 1). Lastly, we utilize the Silverman's rule [15] for the kernel width, γ, in order to mimic kernel annealing and avoid the need for an additional free parameter in the framework. Algorithm 1 has been utilized for robust principal analysis in images and non–shift–invariant dictionary learning [8], [12], [13].

Now, MCC–based SVD can be incorporated into the greedy solution proposed in the previous subsection in order to provide robustness against potential outliers. In particular likewise k-means, the learning framework alternates between sparse coding of the time series and dictionary update stages. It is the latter stage that harnesses Algorithm 1 to estimate in a robust and sequential manner each one of the atoms. The resulting algorithm is referred to as MCC–based Shift–Invariant K-SVD, or MCCKSVD for short (Algorithm 2) where L is the number of sequential decompositions for Matching Pursuit. As previously mentioned, the Sparse Coding routine can be efficiently computed exploiting parallel FFT–based heuristics.

In practice, either a fixed number of alternating optimizations (stages) are performed or a convergence criterion based on successive estimated dictionaries is applied. A similar approach has been used to estimate relevant recurring patterns in single–channel EEG recordings [16].

Algorithm 1 MCC-SVD

Input: $X \in \mathbb{R}^{M \times N}, \epsilon$
Output: $U \in \mathbb{R}^{M}$
  $U^1 \leftarrow \text{EIG}(XX^T, 1)$
  $J \leftarrow 1$
  while convergence == FALSE **do**
    $\mathbf{r} \leftarrow ||\mathbf{x}_i - U^J (U^J)^T \mathbf{x}_i||^2 \qquad i = 1, \ldots, N$
    $\gamma \leftarrow 1.06 \times \min\{\text{std}(\mathbf{r}), \text{IQR}(\mathbf{r})/1.34\} \times N^{-1/5}$
    $p_i \leftarrow -G_\gamma(\sqrt{\mathbf{x}_i^T \mathbf{x}_i - \mathbf{x}_i^T (U^J)(U^J)^T \mathbf{x}_i}) \quad i = 1, \ldots, N$
    $P \leftarrow \text{diag}(-p)$
    $U^{J+1} \leftarrow \text{EIG}(XPX^T, 1)$
    if $||U^J - U^{J+1}|| < \epsilon$ **then**
      convergence = TRUE
      $U \leftarrow U^{J+1}$
    else
      convergence = FALSE
      $J \leftarrow J + 1$
    end if
  end while

Algorithm 2 MCCKSVD

Input: s, M, K, L, ϵ
Output: $D \in \mathbb{R}^{M \times K}$
  $D = \text{InitialDictionary}(s, M, K)$
  while convergence == FALSE **do**
    Sparse Coding Stage (Time Series MP):
    $r_0 \leftarrow s$
    for $i = 1, \ldots, L$ **do**
      $(k_i, \tau_i) \leftarrow \arg\max_{(k,\tau)} |\langle r_{i-1}, T_\tau \mathbf{d}_k \rangle|$
      $c_{k,\tau} \leftarrow [c_{k,\tau}, \langle r_{i-1}, T_{\tau_i} \mathbf{d}_{k_i} \rangle]$
      $r_i \leftarrow r_{i-1} - \langle r_{i-1}, T_{\tau_i} \mathbf{d}_{k_i} \rangle T_{\tau_i} \mathbf{d}_{k_i}$
    end for
    $r \leftarrow r_i$
    Dictionary Update Stage:
    for $\kappa = 1, \ldots, K$ **do**
      $(c, \tau) = \{(c, \tau) | c_{k,\tau} = c_{\kappa,\tau}\}$
      for $i = 1, \ldots, \text{size}(c)$ **do**
        $E_\kappa \leftarrow [E_\kappa, T_{\tau_i}^* r + T_{\tau_i} c_i \mathbf{d}_\kappa]$
      end for
      $\mathbf{d}_\kappa \leftarrow \text{MCC-SVD}(E_\kappa, \epsilon)$
    end for
  end while

III. A Heuristic for a Suitable Initial Dictionary

Likewise k-means, Algorithm 2 is a greedy technique; therefore, it usually requires multiple initializations and a criterion to select the best subset of clusters out of all the possible replicates. In the shift–invariant case, both requirements pose their own particular challenges.

Multiple initializations consume computational resources by running the same algorithm with different initial dictionary seeds (usually chosen at random). The main problem with this approach is the possibility of initializing the dictionary with close replicas of a single function, i.e. by choosing slightly shifted versions of an atom; this would clearly yield suboptimal dictionaries. Another alternative comes from the data mining literature in the form of clustering of time series subsequences, i.e. isolate all the possible M–dimensional snippets from s and exploit regular k-means to obtain the prototypical patterns of the initial dictionary. However, as posited by Keogh and Lin, such approach is troublesome without proper constraints; for instance, the resulting clusters are sinusoidal in essence and do not reflect the local structure of the data [17]. This is a direct consequence of disregarding portions of the time series that are considered meaningless, e.g.

noise, random fluctuations, or novelty patterns. We, therefore, opt for isolating only the M–dimensional patterns that exhibit a clear modulatory behavior or prominent temporal envelopes. This is achieved by locating the peaks of the running variance version of the time series and selecting the patterns around such peaks. Those patterns will constitute the M–dimensional samples to be clustered via classic techniques, e.g. k-means.

Even though this heuristic is more principled than random sampling, it is still sensitive to outliers in the time series. Our proposed solution exploits correntropy again as the cost function in a clustering scheme that resembles a dictionary learning problem with sparsity pattern equal to 1, i.e. each input sample is represented by a single weighted version of an atom. The complete heuristic is detailed in Algorithm 3 (where $\mathbf{x}_T^k$ represents the k-th row in X). Specifically, the initial seed for the dictionary is computed via k-means and then, we exploit correntropy to update the clusters in a robust manner. Hence, we address the first limitation of the shift–invariant case by providing a robust initial dictionary (which alleviates the computational burden of the subsequent algorithms). In this way, instead of running all the algorithms with different initializations, the proposed framework devotes more time to the initial dictionary and perform a single replicate of Algorithm 2.

The second limitation deals with the selection of an optimal initial dictionary. Here, it is not suitable to use reconstruction error as the metric of success due to the presence of potential outliers. Therefore in practice, we opted to run several replicates of Algorithm 3 and choose the set with minimal mutual coherence, $\mu(D)$, as the initial dictionary of the entire framework [18]:

Algorithm 3 Heuristic for Initial Dictionary

Input: s, M, K
Output: $D \in \mathbb{R}^{M \times K}$

$s_p = \text{RunningVariance}(s, M)$
$\Pi = \text{Peaks}(s_p)$
$Y \leftarrow \{T_\tau^* s | \tau + M/2 \in \Pi\}$
$D \leftarrow \text{KMEANS}(Y, K)$
while convergence == FALSE **do**
  Assignment Step:
  $k_i \leftarrow \arg\max_k |\langle \mathbf{y}_i, \mathbf{d}_k \rangle|$ $\quad i = 1, \ldots, |\Pi|$
  $\alpha_i \leftarrow \langle \mathbf{y}_i, \mathbf{d}_{k_i} \rangle$ $\quad i = 1, \ldots, |\Pi|$
  $\mathbf{x}_i[k_i] = \alpha_i$ $\quad i = 1, \ldots, |\Pi|$
  Update Stage:
  for $k = 1, \ldots K$ **do**
    $w_k \leftarrow \{i | 1 \leq i \leq |\Pi|, \mathbf{x}_T^k(i) \neq 0\}$
    $E_k \leftarrow Y - \sum_{j \neq k} \mathbf{d}_j \mathbf{x}_T^j$
    $\Omega_k \in \mathbb{R}^{|\Pi| \times |w_k|}$ s.t. $\Omega_k(w_k(i), i) = 1$
    $E_k^R \leftarrow E_k \Omega_k$
    $\mathbf{d}_k \leftarrow \text{MCC-SVD}(E_k^R)$
  end for
end while

$$\mu(D) = \max_{i \neq j} |\mathbf{d}_i^T \mathbf{d}_j| \tag{10}$$

IV. Results

A. 4 Different Variants of Outliers in Synthetic Data

The first set of validation experiments utilizes a pre–established orthornormal basis (15 zero–mean Discrete Cosine Transform atoms in a 16–dimensional space) to build synthetic time series with random decomposition amplitudes uniformly distributed between 0 and 1. This basis was chosen due to its multiscale nature and its unambiguous decompositions. We simulate 4 different scenarios of feature perturbation in the form of impulsive noise in single time instances and patterns, gaps in the time series, and overlap between generative functions. For all the experiments, the algorithms use the following parameters: $K = 15$, $M = 16$, $\epsilon = 10^{-4}$, $L = 500$, and 25 alternating iterations between sparse coding and dictionary update stages.

The first variant simulates 500 non–overlapping DCT atoms in a time series with baseline SNR of 50 dB. Then, a rate of single time instances are affected by impulsive noise (SNR = -20 dB). We compare the performance of 6 different algorithms: Shift–Invariant K-SVD with Initial Dictionary chosen randomly from the peaks of the running variance version of the time series (KSVD-IniRandom) and its MCCKSVD counterpart (MCCKSVD-IniRandom), Shift–Invariant K-SVD with Initial Dictionary exploiting K-SVD as the estimation technique in Algorithm 3 (KSVD-IniKSVD), i.e. $\mathbf{d}_k \leftarrow \text{SVD}(E_k^R)$ instead of $\mathbf{d}_k \leftarrow \text{MCC-SVD}(E_k^R)$ and its corresponding MCCKSVD algorithm (MCCKSVD-IniKSVD), and lastly, the proposed approach utilizing MCC–based SVD for all the algorithms (MCCKSVD-IniMCCKSVD) and its counterpart (KSVD-IniMCCKSVD). A total of 20 different runs are simulated for each case of impulsive noise rate. The average cross–correlation between the atoms in estimated and original dictionaries is reported in Fig. 1.

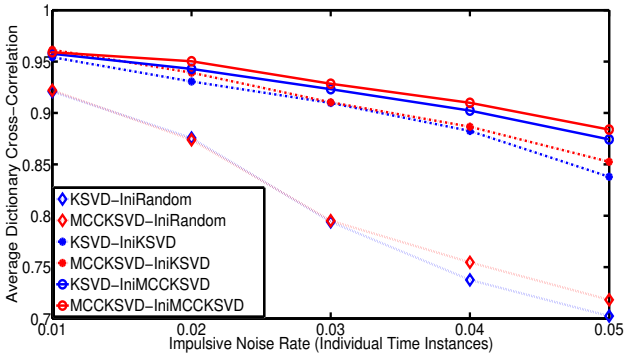

Fig. 1. Average dictionary cross–correlation for several rates of impulsive noise in individual time instances. 15 generative DCT atoms. Baseline SNR = 50 dB. Outlier SNR = -20 dB.

The second case is very similar to the first one with the difference of the impulsive noise being applied to 16–sample patches in the time series, i.e. the perturbation affects whole

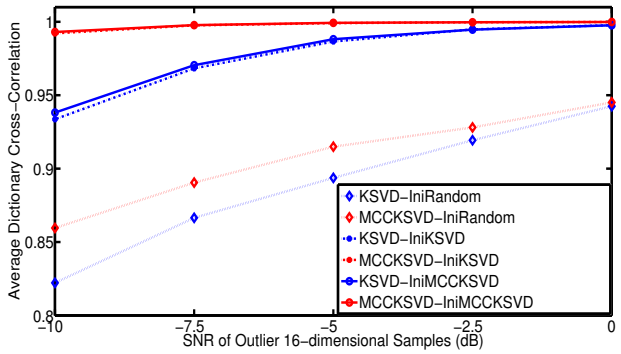

Fig. 2. Average dictionary cross–correlation for several SNRs of impulsive noise in individual 16–dimensional sample instances. 15 generative DCT atoms. Baseline SNR = 50 dB. Outlier rate = 0.2.

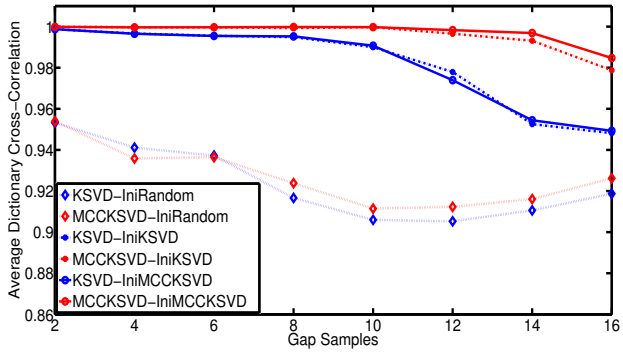

Fig. 3. Average dictionary cross–correlation for different lengths of inserted gaps (concatenated zeros). 15 generative DCT atoms. Baseline SNR = 50 dB. Gap rate = 0.5.

16–dimensional patterns, instead of single time instances. Here, the outlier rate was fixed to 0.2 and their SNR was varied from -10 to 0 dB. Again, a total of 20 different scenarios are simulated (500 non–overlapping DCT atoms in the time series) and the average results are illustrated in Fig. 2.

In both impulsive noise experiments MCCKSVD-IniMCCKSVD outperforms all of the other algorithms while, on the other hand, the random initializers clearly bias the estimated dictionaries. It is worth noting that the proposed heuristic for the seed dictionary is able to set the proper path even in the K-SVD implementations; however, they fail to surpass MCC–based methods due to their lack of robustness.

The third case introduces a different type of perturbation: gaps. In particular, 50 % of randomly selected patterns in the time series are replaced by gaps (concatenated zeros). The length of such gaps is varied from 2 to 16 samples while the unaffected instances have a baseline SNR of 50 dB. The result is a time series where 50 % out of the 500 DCT patterns are incomplete. The 6 approaches are simulated for 20 different runs. Fig. 3 summarizes the results. Again, the MCCbased algorithms outperform their counterparts. Also in general, as the gaps become wider, the performance degrades accordingly.

Lastly, we introduce overlapping patterns as the last variant of outliers. The overlapping between pairs of adjacent patterns is set to 50 %, i.e. 8 samples. Then the percentage of overlaps is varied from 10 to 70 %. For this case, no baseline noise is added to the time series. The result is an input where a fixed rate (per case) of adjacent patterns are allowed an 8–sample overlap. Table I summarizes the average results for three algorithms. Once again, the proposed approach is robust for several overlap rates. This suggests that the MCC–based framework is able to implicitly handle overlapping patterns up to a certain degree; however, more evidence is needed in order to assess proper key parameters, such as breakdown points.

Utilizing four different sources of outliers, we demonstrate that MCC–based algorithms are comprehensively more robust than the state of the art in shift–invariant dictionary learning. In general, the MCCKSVD variants yield higher dictionary cross–correlations than their K-SVD counterparts. Also, the heuristic proposed for the initial dictionary consistently surpasses the random sampling usually utilized in these settings.

TABLE I
AVERAGE DICTIONARY CROSS–CORRELATION WITH RESPECT TO PERCENTAGE OF OVERLAPPING PATTERNS. 15 GENERATIVE DCT ATOMS. 8–SAMPLE OVERLAP BETWEEN ADJACENT PATTERNS.

Algorithm	**Percentage of Overlapping Patterns**						
	10	20	30	40	50	60	70
MCCKSVD-IniMCCKSVD	1.00	1.00	1.00	0.99	0.99	0.99	0.99
MCCKSVD-IniRandom	0.97	0.96	0.96	0.95	0.95	0.92	0.91
KSVD-IniRandom	0.97	0.96	0.95	0.95	0.95	0.91	0.90

B. Winding Data

We present preliminary results on the winding dataset [1]. The data come from an industrial wire winding process. Here, we perform the dictionary estimation on the first dimension alone (U1). The time series has significant impulsive–noise–like patterns, i.e. spikes and dropouts (Fig. 4).

We compare the proposed MCC–based approach with a shift–invariant version of K-SVD that uses a random sampler for the initial dictionary ($K = 3, M = 80, L = 50, \epsilon = 10^{-4}$, 50 alternating iterations between sparse coding and dictionary update stages). As Fig. 4 suggests, the estimated atoms for MCCKSVD are less affected by the outlier patterns, i.e. more robust than the K-SVD scheme. Once again, due to the spikes and dropouts of the time series, the reconstructed error would be heavily biased in this case. Therefore, it is not possible to judge the success of the algorithm solely based on reconstruction error (as suggested in [19]). However, visual inspection clearly favors the proposed framework over the MSE–based optimizations.

In the spirit of openness and to encourage reproducibility, the MATLAB code corresponding to the proposed algorithms and datasets is available at https://github.com/carlosloza/ShiftInvariantMCCKSVD.

[1] http://alumni.cs.ucr.edu/~rakthant/TSEpenthesis/

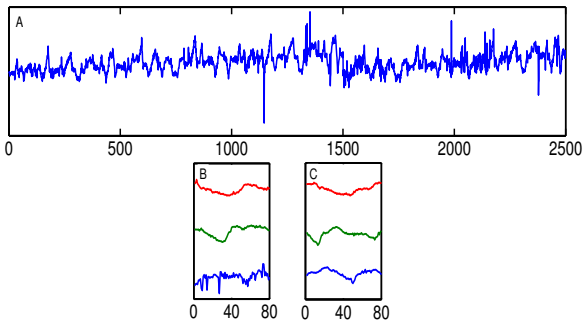

Fig. 4. A. Dimension U1 of Winding Dataset. B. Resulting atoms corresponding to Shift–Invariant K-SVD. C. Resulting atoms corresponding to MCC–based Shift–Invariant K-SVD.

V. Conclusion

We have derived and implemented a shift–invariant dictionary learning scheme fully based on the Maximum Correntropy Criterion (MCC). First, correntropy was utilized to obtain a robust initial dictionary that would set the "right direction" for the greedy algorithms to follow. Second, correntropy was also exploited to update the learned dictionary atoms in an iterative scheme that provides robustness against outliers in the time series, e.g. spikes, dropouts, gaps, high–amplitude impulsive noise, and even overlap. In this way, this second asset keeps the greedy algorithm in the "right path" and prevents possible bias towards suboptimal solutions.

The proposed framework can be easily adapted to different knowledge discovery areas such as communications prone to channel failure, robust blind source separation, and generative models in general. In the future, we would like to explore techniques to select the dimensionality of the dictionary in a principled way in order to render a fully data–driven framework for clustering of relevant time series subsequences.

References

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[2] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5. IEEE, 1999, pp. 2443–2446.

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

[4] B. Mailhé, S. Lesage, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Shift-invariant dictionary learning for sparse representations: extending K-SVD," in *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–5.

[5] P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval, "Motif: An efficient algorithm for learning translation invariant dictionaries," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.

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

[7] Q. Zhang and B. 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.

[8] C. A. Loza and J. C. Principe, "A robust maximum correntropy criterion for dictionary learning," in *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*. IEEE, 2016, pp. 1–6.

[9] W. Liu, P. P. Pokharel, and J. C. Príncipe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[10] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[11] R. T. Rockafellar, *Convex analysis*. Princeton university press, 2015.

[12] R. He, B.-G. Hu, W.-S. Zheng, and X.-W. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1485–1494, 2011.

[13] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2011.

[14] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[15] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[16] C. A. Loza, M. S. Okun, and J. C. Príncipe, "A marked point process framework for extracellular electrical potentials," *Frontiers in systems neuroscience*, vol. 11, p. 95, 2017.

[17] E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless: implications for previous and future research," *Knowledge and information systems*, vol. 8, no. 2, pp. 154–177, 2005.

[18] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[19] T. Rakthanmanon, E. J. Keogh, S. Lonardi, and S. Evans, "Time series epenthesis: Clustering time series streams requires ignoring some data," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 547–556.