A Switch Kernel Width Method of Correntropy for Channel Estimation

WeiHua Wang, Jihong Zhao, Hua Qu, Badong Chen
School of Electronic and Information Engineering,
Xi'an Jiaotong University
Xi'an, China
chenbd@mail.xjtu.edu.cn

Jose C. Principe
Department of Electrical and Computer Engineering
University of Florida
Gainesville, USA
principe@cnl.ufl.edu

Abstract—Correntropy has been successfully applied in non-Gaussian signal processing, but the superior performance achieved is depend on appropriate selection of the kernel width. How to select a proper kernel width is a crucial problem in correntropy applications. In this paper, we propose an adaptive algorithm to update the kernel width, which is set at a maximum between the absolute value of instantaneous error divided by square root of 2 and a predetermined kernel width. The new algorithm involves no extra free parameters and keeps the simplicity and robustness of the original maximum correntropy criterion (MCC) algorithm. Simulation results confirm that the proposed algorithm can achieve excellent performance in channel estimation under impulsive noises.

Keywords—correntropy; MCC; adaptive kernel width; channel estimation

I. INTRODUCTION

Adaptive filters are widely used in signal processing applications such as channel estimation, noise cancellation, system identification etc. The least mean square (LMS) [1] is one of the most popular adaptive algorithms due to its simplicity and effectiveness. However, the superior performance of the LMS algorithm depends heavily on Gaussian assumptions since it minimizes only a second-order statistics called the mean squared error (MSE) between the desired signal and the filter output. Many other criteria such as least mean fourth (LMF) [2] exploit a higher-order moment of the error, but their performance deteriorates rapidly in the presence of heavy-tailed impulsive noises.

Many studies have shown that lower-order norms lead to robustness against impulsive and intensive interferences. The least mean p-norm (LPM) algorithm based on an $L_p$ ($p < 2$) norm of the error was proposed in [3]. The sign algorithm (SA) [4] based on the $L_1$-norm minimization has attracted more attention due to its considerably low computational cost and easy implementation, where only the sign of the error signal is involved in the updating process.

In recent years, information theoretic learning (ITL) [5,6] methods have been shown to be efficient approaches in non-Gaussian signal processing since they capture the whole distribution rather than a specific moment of the data. The minimum error entropy (MEE) [7-9] and the maximum correntropy criterion (MCC) [10-12] are two popular optimization criteria in ITL. The MEE cost function has shown robustness in presence of non-Gaussian noises, however, it is computationally much expensive than an LMS-type algorithm. The MCC is a robust optimality criterion for non-Gaussian (especially heavy-tailed) signal processing, and the adaptive algorithms under MCC are computationally very simple (with a stochastic gradient method the computational complexity is almost same as the LMS algorithm). Thus, MCC is more suitable for practical implementation. Recently, MCC has been successfully applied to various signal processing and machine learning problems [13-15].

The kernel width of the correntropy is an important free parameter, which determines the performance of the MCC learning. How to choose a proper kernel width is thus a crucial problem in correntropy applications. Silverman’s rule [16] is one of the most widely used methods for kernel selection, but in general, it is only effective for probability density function (PDF) estimation problem. Other PDF estimation based methods include the statistical method [17] and cross-validation techniques [18]. In order to optimize the kernel width in general ITL learning, an adaptive update algorithm by minimizing a Kullback-Leibler (KL) divergence was derived in [19]. A fixed-point update rule with no free parameters was further developed in [20]. However, all the methods mentioned above are not directly related to the goodness criteria of an adaptive filtering algorithm, which are usually measured by the convergence speed or misadjustment. In [21], an adaptive algorithm was proposed for the selection of kernel width in correntropy, which utilizes an appropriate kernel width under Gaussian condition as a standard to search a proper kernel width for other conditions, and updates the kernel width at every iteration based on the shape of the error distribution. However, estimating the shape of the error distribution is in general complex and inaccurate.

In this paper, a new adaptive algorithm for kernel width selection is proposed. The basic idea of the proposed algorithm is adaptively choose a kernel width which makes the prediction error decrease along the largest slop when the error is larger than a threshold. This method updates the kernel width at every iteration based on the instantaneous error, and switches it to a predetermined kernel width when the adaptive kernel width is smaller than it. The developed new algorithm achieves excellent performance while keeps the simplicity and robustness of the original MCC algorithm. This makes it...
Correntropy and Maximum Correntropy Criterion

A. Correntropy

Correntropy is a nonlinear similarity measure between two random variables $X$ and $Y$ in kernel spaces [10]:

$$V_n(X,Y) = E[k_n(x,y)] = \int k_n(x,y) dF_{xy}(x,y)$$  \hspace{1cm} (1)

where $E[]$ denotes an expectation operator, $k_n(\cdot,\cdot)$ is a translation-invariant Mercer kernel, with a bandwidth $\sigma$, and $F_{xy}(\cdot)$ stands for the joint distribution of $(X,Y)$. In practice, the data distribution is usually known, and only a finite number of samples $(x_i,y_i)$, for $X$ and $Y$ are available. In this case, it is common to use a sample mean estimator for the expectation operator, and the correntropy can be estimated as:

$$\hat{V}_{N,n}(X,Y) = \frac{1}{N} \sum_{i=1}^{N} k_n(x_i,y_i)$$  \hspace{1cm} (2)

In this work, without mentioned otherwise, the kernel function in correntropy is a Gaussian kernel, defined by

$$k_n(x,y) = \exp(-\frac{||x-y||^2}{2\sigma^2})$$  \hspace{1cm} (3)

Using a Taylor series expansion of the exponential function in the Gaussian kernel and substituting in Eq. (2),

$$\hat{V}_{N,n}(X,Y) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n n!} E[(X-Y)^n]$$  \hspace{1cm} (4)

as we can see, correntropy can be viewed as a generalized correlation function between two random variables, it containing higher (even) order moments of the error $(X-Y)$ between them. Correntropy is a local measure of how similar two random variables are, within a small neighborhood determined by the kernel width $\sigma$. As a new measure of similarity, the correntropy can be used as an objective function for many application. Bayesian estimation under maximum correntropy criterion has been studied in [22], it has been shown that maximum correntropy (MC) estimation is a smoothed maximum a posteriori (MAP) estimation, including the MAP and the minimum mean square error (MMSE) estimation as the extreme case. When the kernel size in correntropy is larger than some value, the MC estimation will have a unique optimal solution lying in a strictly concave region of the smoothed posterior distribution.

The localization provided by the kernel width proves to be very useful in reducing the detrimental effects of outliers and impulsive noise. The correntropy is always bounded for any distribution and is robust to impulsive noises. A measure based on just second order statistics, like MSE, can easily get biased in such conditions.

B. Maximum Correntropy Criterion

Like MSE criterion, correntropy can be used as a cost function in adaptive filters [13]. The objective of the filter is to optimize a cost function in such a way that the filter output resembles as closely as possible to the desired signal. Let define our cost function to be the correntropy between the desired signal $d_i$ and the filter output $y_i$. We will use a normalized Gaussian kernel to compute correntropy.

$$J_n = \frac{1}{N} \sum_{i=n-N+1}^{n} \exp(-\frac{|d_i-y_i|^2}{2\sigma^2})$$  \hspace{1cm} (5)

If the filter weights at the nth time instant are $w_n$, the cost function can be written as,

$$J_n = \frac{1}{N} \sum_{i=n-N+1}^{n} \exp(-\frac{|d_i-w_i^T X_i|^2}{2\sigma^2})$$  \hspace{1cm} (6)

the goal is to maximize correntropy between the desired signal $d_i$ and the filter output $y_i$.

Analytically, it is difficult to find the set of weights $w_n$ which can maximize the cost function $J_n$. Similar with MSE criterion, we can use an iterative gradient ascent approach to search the optimal solution, which is the next set of filter weights are corrected by taking a value proper to the positive gradient to the cost function in the weight space. Therefore,

$$w_{n+1} = w_n + \mu \nabla J_n$$  \hspace{1cm} (7)

Substituting $J_n$ and computing the gradient with regard to $w_n$, we obtain,

$$w_{n+1} = w_n + \frac{\mu}{\sigma^2} \frac{1}{N} \sum_{i=n-N+1}^{n} \left[ \exp(-\frac{|e_i|^2}{2\sigma^2}) \epsilon_i X_i \right]$$  \hspace{1cm} (8)

Approximating the sum by the current value inspired by the stochastic gradient, the weight update equation using the correntropy cost function has therefore been reduced to this simple form.

$$w_{n+1} = w_n + \frac{\mu}{\sigma^2} \frac{1}{N} \sum_{i=n-N+1}^{n} \exp(-\frac{|e_i|^2}{2\sigma^2}) e_i X_i$$  \hspace{1cm} (9)

in which $e_i = d_i - w_i^T X_i$ is the prediction error.

Comparing with the LMS weight update rule, we can see that the weight update at each iteration in Eq.(9) contains just
an extra scaling factor which is an exponential function of the value of the error in that iteration. This factor reflects the outlier rejection property of the correntropy similarity measure. We therefore expect the adaption of weights using MCC to be more stable if the desired signal has strong outliers or impulsive characteristics. The steady-state excess mean square error (EMSE) of the adaptive filtering under the maximum correntropy criterion in presence of Gaussian and non-Gaussian noise environment has been studied [23].

From the definition of correntropy, the kernel width is the window within which the similarity of the two random variables is computed. A very large kernel width will therefore yield a similarity measure close to MSE value. A small value, of the order of the error, is useful in exploiting the advantageous properties of correntropy, but will lead to a slower convergence speed when the error is larger at beginning. Thus the kernel width of correntropy is a very important parameter that must be select by user at the initial stage of the MCC algorithm.

III. SWITCH KERNEL WIDTH METHOD

The kernel width has important effects in the process of the MCC learning. How to choose a proper kernel width is a crucial problem in correntropy applications. In the original MCC algorithm, the kernel width is fixed throughout the whole process, but the error in an adaptive system decreases gradually as the system weights continuously update to a steady state. A fixed kernel width may affect the adaptation dynamics and would be a compromise between fast learning initially, and fast learning near the optimum point. Therefore, adaptively update the kernel width according to the error can potentially improve the overall learning rate.

The basic idea of the proposed algorithm is to adaptively update the kernel width according to the instantaneous error, select the kernel width which makes the prediction error decrease along the largest slop, that make sure the algorithm converge most quickly at each iteration, and switches to a predetermined kernel width when the adaptive kernel width is smaller than it, this helps to keep the robustness property of the conventional MCC algorithm.

When we adaptively update the kernel width, then the weight update expression would be:

\[
W_{n+1} = W_n + \frac{\mu}{\sigma_n^2} \text{exp} \left( \frac{-|e_n|^2}{2\sigma_n^2} \right) e_n X_n
\]

where \(\sigma_n\) is the updated kernel width in nth iteration calculated according to the error.

First of all, when we got the error \(e_n\) at nth iteration, then calculate the kernel width that makes error decrease along the largest slop.

\[
\max_{\sigma_n} J'_n = \frac{1}{\sigma_n} \exp \left( \frac{-|e_n|^2}{2\sigma_n^2} \right) e_n
\]

For compute simplicity, we denote

\[
g(z) = \exp(-z^2) \cdot z^2
\]

in which

\[
z^2 = \frac{|e_n|^2}{2\sigma_n^2}
\]

then we can derive (11) as

\[
\max J'_n = 2g(z)/\epsilon_n
\]

When the error \(\epsilon_n\) at nth iteration is calculated, the kernel width which makes error decrease along the largest slop equivalent the \(z\) makes \(g(z)\) maximum. \(g(z)\) is a bounded function of \(z\), the curve of \(g(z)\) is shown is Fig. 1.

![Fig. 1. The curve of \(g(z)\)](image)

when \(z^2 = 1\), namely,

\[
\sigma_n^2 = \epsilon_n^2 / 2
\]

we got

\[
\max J'_n = 2\exp(-1)/\epsilon_n
\]

It is important to notice that if we adopt adaptive kernel width \(\sigma_n^2 = \epsilon_n^2 / 2\), then the exponential part of correntropy will be a constant, and this algorithm is no longer a correntropy based algorithm, but it still keeps the robustness property in a way different with correntropy. The maximum of \(J'_n\) is a reciprocal function of instantaneous error, it would be small when the error is large, so the weight wouldn’t change much when the input signal is interrupted by an impulsive noise, this helps to keep the robustness property.

It also should be noticed that the maximum of \(J'_n\) growing larger as the error decrease, which means the increment of the weight would be increase and the weight would change a lot as the error decrease, that would lead the divergence of the algorithm. So this algorithm would converge to a small error at first and then diverge when the error is smaller than a threshold.
Here we introduce a switch method to deal with this problem. We adopt the adaptive kernel width at the beginning to speed up the convergence rate, and then switches to a predetermined kernel width when the adaptive kernel width is smaller than it, this make sure the algorithm would converge faster while achieve the almost same steady state misjudgement as original MCC. So the adaptive kernel width update expression is

$$\sigma^2_e = \max(e^2_c/2, \sigma^2)$$

(17)

where $\sigma$ denotes the predetermined kernel width which could be calculated by Silverman’s rule [16] or other method. Since the updated kernel width is switches between an adaptive kernel width and a predetermined kernel width, so we name it after switch kernel width maximum correntropy criterion (SMCC) algorithm.

The $J'$ determines the increment of the weight, which will controls the convergence rate and the steady state misjudgement of the algorithm. The ideal case is $J'$ keeps steady when the error is large, and decrease with the error when it get smaller. The curve of $J'$ versus $|e|$ with different fixed kernel width is shown in Fig. 2.

![Fig. 2. The curve of $J'$ versus $|e|$](image)

The $J'$ curve with adaptive kernel width $\sigma^2 = e^2_c/2$ is always above all the curve with different fixed kernel width, this would helps the algorithm converge faster at the beginning when the error is large, but it growing larger as the error decrease, that would lead the algorithm cannot converge. It also should be noticed that the curve with fixed kernel width tangent to the curve with adaptive kernel width $\sigma^2 = e^2_c/2$. The point of tangency is where the switch is happened, from there the curve would decrease along the curve with fixed kernel width as the error decrease, and the algorithm would converge to the same steady state misjudgement as original MCC algorithm. The lager the fixed kernel width is, the earlier the switch will happen.

It is obvious that the proposed method could speed up convergence rate without involving any extra free parameter while keep the computational simplicity of original MCC algorithm.

For a better understanding, the proposed SMCC algorithm is summarized as Tab. I.

<table>
<thead>
<tr>
<th>Switch kernel width maximum correntropy criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initialization</strong></td>
</tr>
<tr>
<td>set step-size $\mu$ and predetermined kernel width $\sigma$</td>
</tr>
<tr>
<td>$W_0 = 0, e_i = d_i$</td>
</tr>
<tr>
<td><strong>Computation</strong></td>
</tr>
<tr>
<td>While ${X_n, d_n}$ available do</td>
</tr>
<tr>
<td>% compute the error</td>
</tr>
<tr>
<td>$e_n = d_n - \hat{W}_n^2X_n$</td>
</tr>
<tr>
<td>% update the kernel width</td>
</tr>
<tr>
<td>$\sigma^2_e = \max(e^2_c/2, \sigma^2)$</td>
</tr>
<tr>
<td>% update the weight</td>
</tr>
<tr>
<td>$W_{n+1} = W_n + \frac{\mu}{\sigma_n} \exp\left(-\frac{\left</td>
</tr>
<tr>
<td>end while</td>
</tr>
</tbody>
</table>

**IV. SIMULATION RESULTS**

To validate the performance of the proposed SMCC algorithm, in this section, we compare the performance of MCC, SA, adaptive kernel width selection algorithm proposed in [21] (denote as AMCC) and switch kernel width MCC (SMCC) algorithm proposed in this paper for Multiple Input Multiple Output (MIMO)-Orthogonal Frequency Domain Multiple Xing (OFDM) channel estimation.

Channel estimation process is an important technical issue for MIMO-OFDM wireless communication systems. Adaptive channel estimation is a widely used estimation technique for MIMO-OFDM systems. Among numerous adaptive techniques that exist in the open literature, the popular category of approaches which are obtain form the minimization of the MSE between the output of the filter and desired signal to perform channel estimation. Pilot based channel estimation schemes have been proposed in [24-26], all of this performance is done under the assumption of additive white Gaussian noise (AWGN), but in practical communication system with impulsive noise, most method would deteriorate rapidly. Here we introduce correntropy as the cost function to deal with this situation.

When the proposed SMCC algorithm is employed in adaptive channel estimation, the overall channel estimation framework is shown in Fig. 3. The QPSK signal is transmitted via an unknown and time-varying multipath Rayleigh channels, and corrupted by an impulsive environment noise. The aim of channel estimation algorithm is to minimize the mean square deviation of real channel coefficients and coefficients of adaptive filter. There are two adaptation loops in the system, one for weights update, and the other for the kernel width...
update. The kernel width update is performed according to the instantaneous error, and then update the weight based on the prediction error and the updated kernel width.

Let's say there are 8 antennas at the base station, 1 antenna at the user equipment, so there will be 8 unknown and time-varying multipath Rayleigh channel coefficients need to be estimate. There 2000 QPSK training samples get through the 8 unknown and time-varying multipath Rayleigh channels, and corrupted by mixture Gaussian noise, the channels changed in the middle of the training sequence.

We use Gaussian-mixture noise models to simulate the impulsive environment noise,
\[ p_{\text{noise}} = (1-\theta)N(\mu_i,\sigma_i^2) + \theta N(\mu_i,\sigma_i^2) \]  
(18)
where \( N(\mu_i,\sigma_i^2) \) denotes the Gaussian distributions with mean \( \mu_i \) and variances \( \sigma_i^2 \), and \( \theta \) is the mixture coefficient. parameters \( (\mu_i,\mu_i,\sigma_i^2,\sigma_i^2,\theta) \) in the mixed Gaussian distribution are set as \( (0,0,0.001,10,0.05) \). The Gaussian distribution with variance 10 creates stronger impulsive outliers.

Since the main contribution of the proposed SMCC algorithm is to speed up the convergence rate of MCC, while keep the same steady state behavior as MCC, so we first compare the convergence curves of those algorithm, in terms of mean square deviation (MSD) of the real channel coefficients and the coefficients of the adaptive filter. The step size of MCC and SMCC are set to 0.4, and the predetermined kernel width for MCC and SMCC are set to 2 for this case. We adjust the parameters of AMCC to meet the same level of steady state MSD to compare its convergence rate with MCC and SMCC. 50 Monte-Carlo simulations are run for the different date with different noises, the simulation results are shown in Fig. 4.

We can see a significant improvement in term of MSD convergence rate compare SMCC with MCC and AMCC, and the MSD convergence curve of SMCC overlap with MCC at the steady state MSD. Since the SMCC algorithm only adopt instantaneous error to update the kernel width, so it is computational simpler than AMCC, and it has a better track ability than AMCC.

The sign algorithm (SA) is less accurate and with higher misadjustment, but robust to the presence of outliers [4]. We adjust the parameters of SA and SMCC to meet the same level of steady state MSD and then compare their convergence rate in different level of environment noise. The environment noises are set as \( (0,0,0.001,10,0.05) \) and \( (0,0,0.001,1,0.05) \) respectively. The simulation results is shown in Fig. 5.

It is obvious that the convergence curve of SA and SMCC almost have the same performance in the strong noise environment, but in weak noise environment the SMCC converge faster than SA. That means the SMCC algorithm performs better in time-varying cases.

Like MCC algorithm, there are two free parameter in the proposed SMCC algorithm. In the following part, the effect of the predetermined kernel width \( \sigma \) and the step size \( \mu \) were
demonstrated. The step size is set to 0.4, 50 Monte-Carlo simulations are run for the different training data and impulsive noise with different kernel width. The simulation results are shown in Fig. 6.

![Fig. 6. Convergence curves of SMCC with different predetermined kernel widths in a mixture Gaussian noise environment](image)

The adaptive kernel width switched to the predetermined kernel width when the instantaneous error is smaller enough, the larger the predetermined kernel width is, the earlier the adaptive kernel width switched to it. The scalar factor $g(z)$ growing smaller as the predetermined kernel width increase, so the MSD converge slower and the steady state MSD is smaller.

Step size is another important parameter that has to be set by user. To examine the effect of step size in the proposed SMCC algorithm, set predetermined kernel width to 2, 50 Monte-Carlo simulations are run for the different training data and impulsive noise with different step size. The simulation results in term of MSD convergence curve are shown in Fig. 7.

![Fig. 7. Convergence curves of SMCC with different step sizes in a mixture Gaussian noise environment](image)

The step size is a compromise of convergence rate and steady state misjudgement. The algorithm converge faster as the step size growing larger, and the steady state MSD growing larger as well.

V. CONCLUSION

How to select a kernel width is a crucial problem in correlation applications. In this paper, we developed an adaptive algorithm to update the kernel width, which is the maximum between the absolute value of instantaneous error divided by square root of 2 and a predetermined kernel width. The proposed new algorithm converge most fast, and have the almost same steady state behavior as original MCC algorithm. The proposed new algorithm involves no extra free parameters and keeps the simplicity and robustness of the original MCC algorithm. Simulation results confirm that the proposed new algorithm achieve excellent performance in channel estimation under impulsive noise.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61372152, No. 61372092), the National High Technology Research and Development Program of China (863 Program) (No.2014AA01A701), and the Key Project of major national science and technology on new generation of broadband wireless mobile communication network (No. 2013ZX0302010-003).

References


