Generalized Correntropy for Robust Adaptive Filtering

Badong Chen, *Senior Member, IEEE*, Lei Xing, Haiquan Zhao, *Member, IEEE*, Nanning Zheng, *Fellow, IEEE*, and José C. Príncipe, *Fellow, IEEE*

***Abstract*—As a robust nonlinear similarity measure in kernel space, correntropy has received increasing attention in domains of machine learning and signal processing. In particular, the maximum correntropy criterion (MCC) has recently been successfully applied in robust regression and filtering. The default kernel function in correntropy is the Gaussian kernel, which is, of course, not always the best choice. In this paper, we propose a generalized correntropy that adopts the generalized Gaussian density (GGD) function as the kernel, and present some important properties. We further propose the generalized maximum correntropy criterion (GMCC) and apply it to adaptive filtering. An adaptive algorithm, called the GMCC algorithm, is derived, and the stability problem and steady-state performance are studied. We show that the proposed algorithm is very stable and can achieve zero probability of divergence (POD). Simulation results confirm the theoretical expectations and demonstrate the desirable performance of the new algorithm.**

***Index Terms*—Correntropy, generalized correntropy, adaptive filtering, GMCC algorithm.**

I. Introduction

SELECTING a proper cost function (usually a statistical measure of the error signals) is a key issue in adaptive filtering theory and applications [1]–[3]. The mean square error (MSE) is widely used as a cost function since it has attractive features, such as smoothness, convexity, mathematical tractability, low computational burden and optimality under Gaussian assumption. The well-known least mean square (LMS) algorithm and its variants, such as normalized LMS (NLMS) and variable step-size LMS (VSSLMS), were developed under this criterion [1], [2]. The MSE is desirable if the signals are Gaussian distributed. In non-Gaussian situations, however, its performance may degrade considerably and in these cases, a non-quadratic cost will be, in general, better than MSE [3].

Generally speaking, there are two types of non-Gaussian distributions: light-tailed (e.g., uniform, binary, etc.) and heavy-tailed (e.g., Laplace, α-stable, etc.) distributions. When the desired signals are disturbed by light-tailed non-Gaussian noises, a higher-order statistical (HOS) measure of the error is usually more desirable. A typical example is the least mean fourth (LMF) family algorithms, which use the mean even power of the error as the cost function [4]. Compared with the LMS algorithm, the LMF may achieve a faster convergence speed and a lower steady-state mean square deviation (MSD) especially in light-tailed noises. One drawback of the LMF algorithm however is that the stability is not guaranteed, which depends on the input and noise powers, and the initial values of the weights. A more general class of algorithms are the least mean p-power (LMP) family algorithms, which adopt the p-order absolute moment of the error as the adaptation cost [5].

When the desired signals are disturbed by heavy-tailed impulsive noises (which may cause large outliers), a lower-order statistical (LOS) measure of the error is usually more robust (i.e., less sensitive to impulsive interferences). For example, the sign algorithm (SA), which employs the mean absolute value of the error as the cost function, is rather robust to the presence of large noises [6]–[8]. The convergence speed and steady-state performance of the SA algorithm is however not so good in general. In the literature many other robust cost functions have been proposed to develop robust adaptive filtering algorithms. Typical examples include mixed-norm [9], [10], M-estimate cost [11], [12], and error entropy [13]–[18]. Particularly in recent years, the maximum correntropy criterion (MCC) has been successfully used in robust adaptive filtering, wherein the filter weights are adapted such that the correntropy between the desired signal and filter output is maximized [19]–[24]. The correntropy is a nonlinear and local similarity measure directly related to the probability of how similar two random variables are in a neighborhood of the joint space controlled by the kernel bandwidth, which also has its root in Renyi's entropy (hence the name "correntropy") [13], [19]. Since correntropy is insensitive to outliers especially with a small kernel bandwidth, it is naturally a robust adaptation cost in presence of heavy-tailed impulsive noises.

The kernel function in correntropy is usually a Gaussian kernel, which is desirable due to its smoothness and *strict positive-definiteness*. With a Gaussian kernel, the correntropy induces a nonlinear metric called the correntropy induced

Manuscript received April 12, 2015; revised September 27, 2015 and December 22, 2015; accepted February 23, 2016. Date of publication March 07, 2016; date of current version May 18, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gustau Camps-Valls. This work was supported by 973 Program (No. 2015CB351703) and National NSF of China (No. 61372152).

B. Chen, L. Xing, and N. Zheng are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: chenbd@mail.xjtu.edu.cn; nnzheng@mail.xjtu.edu.cn; xl2010@stu.xjtu.edu.cn).

H. Zhao is with the School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China (e-mail: hqzhao@home.swjtu.edu.cn).

J. C. Príncipe is with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China, is also with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: principe@cnel.ufl.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2539127

metric (CIM) which behaves like an L_2 norm when data are relatively small compared with the kernel bandwidth, an L_1 norm as data get larger, and an L_0 norm when data are far away from the origin [19], [25]. However, Gaussian kernel is, of course, not always the best choice. In the present work, we propose to use the *generalized Gaussian density* (GGD) [26], [27] function as a kernel function in correntropy, and the new correntropy is called the *generalized correntropy*. Some important properties of the generalized correntropy are presented. In particular, we show that the order-α *generalized correntropy induced metric* (GCIM) or *generalized correntropic loss* (GC-loss) function behaves like different norms (from L_α to L_0) of the data in different regions.

Similar to the usual correntropy with Gaussian kernel, the generalized correntropy can also be used as an optimization cost in estimation-related problems. In this work, we focus mainly on applying the generalized maximum correntropy criterion (GMCC) to adaptive filtering. We show that the optimal solution of GMCC filtering is in form similar to the well-known Wiener solution, except that the autocorrelation matrix and cross-correlation vector are weighted by an error nonlinearity. If the signals involved are zero-mean Gaussian, the optimal solution will equal to the Wiener solution. Under the GMCC criterion, a stochastic gradient based adaptive filtering algorithm, called the *GMCC algorithm*, is developed. The stability problem and the steady-state performance of the GMCC algorithm are analyzed. In particular, we present a simple example to show that the GMCC may have a zero probability of divergence (POD). A theoretical value of the steady-state *excess mean square error* (EMSE) of the GMCC algorithm is also derived. Simulation results confirm the theoretical expectations and the desirable performance of the GMCC.

The rest of the paper is organized as follows. In Section II, we define the generalized correntropy, and present some important properties. In Section III, we apply the generalized maximum correntropy criterion (GMCC) to adaptive filtering and develop the GMCC algorithm. In Section IV, we analyze the stability and steady-state performance of the GMCC algorithm. In Section V, we present illustrative examples to verify the theoretical results and demonstrate the good performance of the new algorithm. Finally in Section VI, we give the conclusion.

II. Generalized Correntropy

A. Definition

Given two random variables X and Y, the correntropy is defined by [19], [28]

$$V(X,Y)=\mathbf{E}[\kappa(X,Y)]=\int \kappa(x,y)dF_{XY}(x,y) \tag{1}$$

where $\mathbf{E}$ denotes the expectation operator, $\kappa(\cdot,\cdot)$ is a *shift-invariant Mercer kernel*, and $F_{XY}(x,y)$ denotes the joint distribution function of (X,Y). Without mentioned otherwise, the kernel function of correntropy is the Gaussian kernel:

$$\begin{aligned}\kappa(x,y)=G_\sigma(e)&=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{e^2}{2\sigma^2}\right)\\&=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\lambda e^2\right)\end{aligned} \tag{2}$$

where $e=x-y$, $\sigma>0$ is the kernel bandwidth, and $\lambda=1/2\sigma^2$ is the kernel parameter. The correntropy $V(X,Y)$ can also be expressed as

$$V(X,Y)=\mathbf{E}[\varphi(X)^T\varphi(Y)]=\mathrm{trace}(\mathbf{R}_{\varphi(X)\varphi(Y)}) \tag{3}$$

where $\mathbf{R}_{\varphi(X)\varphi(Y)}=\mathbf{E}[\varphi(X)\varphi(Y)^T]$, and $\varphi(\cdot)$ denotes a nonlinear mapping induced by κ, which transforms its argument into a high-dimensional (infinite for Gaussian kernels) Hilbert space $\boldsymbol{F}_\kappa$, satisfying $\varphi(X)^T\varphi(Y)=\kappa(X,Y)$ [19]. The correntropy is therefore essentially a second-order statistic of the mapped feature space data.

There is a well-known generalization of Gaussian density function, called the *generalized Gaussian density* (GGD) function, which with zero-mean is given by [26], [27]

$$\begin{aligned}G_{\alpha,\beta}(e)&=\frac{\alpha}{2\beta\Gamma(1/\alpha)}\exp\left(-\left|\frac{e}{\beta}\right|^\alpha\right)\\&=\gamma_{\alpha,\beta}\exp\left(-\lambda|e|^\alpha\right)\end{aligned} \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function, $\alpha>0$ is the shape parameter, $\beta>0$ is the scale (bandwidth) parameter, $\lambda=1/\beta^\alpha$ is the kernel parameter, and $\gamma_{\alpha,\beta}=\alpha/(2\beta\Gamma(1/\alpha))$ is the normalization constant. This parametric family of symmetric distributions include the Gaussian ($\alpha=2$) and Laplace ($\alpha=1$) distributions as the special cases. As $\alpha\to\infty$, the GGD density converges point-wise to a uniform density on $(-\beta,\beta)$.

In this work, we use the GGD density function as the kernel function of correntropy, and define

$$V_{\alpha,\beta}(X,Y)=\mathbf{E}[G_{\alpha,\beta}(X-Y)] \tag{5}$$

To make a distinction between (5) and the correntropy with Gaussian kernel, we call it the generalized correntropy. Clearly, the correntropy with Gaussian kernel corresponds to the generalized correntropy with $\alpha=2$.

Remark 1: It is worth noting that in the generalized correntropy, the kernel function does not necessarily satisfy the Mercer's condition. Actually, the kernel function $\kappa(x,y)=G_{\alpha,\beta}(x-y)$ is positive definite if and only if $0<\alpha\le 2$ (see [29], page 434).

In practice, the joint distribution of X and Y is usually unknown, and only a finite number of samples $\{(x_i,y_i)\}_{i=1}^N$ are available. In this case, the sample mean estimator of the generalized correntropy is

$$\hat{V}_{\alpha,\beta}(X,Y)=\frac{1}{N}\sum_{i=1}^{N}G_{\alpha,\beta}(x_i-y_i) \tag{6}$$

B. Properties

Below we present several basic properties of the generalized correntropy. Some of them are simple extensions of the properties in [19], and hence will not be proved here.

Property 1: $V_{\alpha,\beta}(X,Y)$ is symmetric, i.e., $V_{\alpha,\beta}(X,Y)=V_{\alpha,\beta}(Y,X)$.

Property 2: $V_{\alpha,\beta}(X,Y)$ is positive and bounded: $0<V_{\alpha,\beta}(X,Y)\le G_{\alpha,\beta}(0)=\gamma_{\alpha,\beta}$, and it reaches its maximum if and only if $X=Y$.

Property 3: The generalized correntropy involves higher-order absolute moments of the error $X-Y$: $V_{\alpha,\beta}(X,Y) = \gamma_{\alpha,\beta}\sum_{n=0}^{\infty}\frac{(-\lambda)^n}{n!}\mathbf{E}\left[|X-Y|^{\alpha n}\right]$.

Remark 2: When the kernel parameter λ is small enough, we have $V_{\alpha,\beta}(X,Y) \approx \gamma_{\alpha,\beta}\left(1-\lambda\mathbf{E}\left[|X-Y|^{\alpha}\right]\right)$. In this case, the generalized correntropy is, approximately, an affine linear function of the α-order absolute moment of the error.

Property 4: Assume that the samples $\{(x_i,y_i)\}_{i=1}^N$ are drawn from the joint PDF $p_{XY}(x,y)$. Let $\hat{p}_e(\cdot)$ be the Parzen estimate of the error PDF from the samples $\{e_i = x_i - y_i\}_{i=1}^N$, with the GGD density function $G_{\alpha,\beta}$ as the Parzen window kernel. Then $\hat{V}_{\alpha,\beta}(X,Y)$ is the value of $\hat{p}_e(\cdot)$ evaluated at the point zero, that is

$$\hat{V}_{\alpha,\beta}(X,Y) = \hat{p}_e(0) \tag{7}$$

where $\hat{p}_e(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N} G_{\alpha,\beta}(\varepsilon - e_i)$

Property 5: For the case $0<\alpha\leq 2$, the generalized correntropy is a second-order statistic of the mapped feature space data.

Proof: When $0<\alpha\leq 2$, the kernel function $\kappa(x,y) = G_{\alpha,\beta}(x-y)$ is a Mercer kernel, and hence we have $V_{\alpha,\beta}(X,Y) = \mathbf{E}\left[\varphi_{\alpha,\beta}(X)^T\varphi_{\alpha,\beta}(Y)\right]$ where $\varphi_{\alpha,\beta}(\cdot)$ is a nonlinear mapping induced by $G_{\alpha,\beta}$.

In data analysis such as regression and classification, a measure called the *correntropic loss* (C-loss) can be used instead of using the correntropy [30], [31]. A generalized C-loss (GC-loss) function between X and Y can be defined as

$$J_{\text{GC-loss}}(X,Y) = G_{\alpha,\beta}(0) - V_{\alpha,\beta}(X,Y) \tag{8}$$

the GC-loss satisfies $J_{\text{GC-loss}}(X,Y)\geq 0$, and when $0<\alpha\leq 2$, it can be expressed as

$$J_{\text{GC-loss}}(X,Y) = \frac{1}{2}\mathbf{E}\left[\left\|\varphi_{\alpha,\beta}(X)-\varphi_{\alpha,\beta}(Y)\right\|^2\right] \tag{9}$$

which is a mean-square loss in the feature space $\boldsymbol{F}_\kappa$ induced by the Mercer kernel $\kappa(x,y) = G_{\alpha,\beta}(x-y)$. Clearly, minimizing the GC-loss will be equivalent to maximizing the generalized correntropy.

Let $\{(x_i,y_i)\}_{i=1}^N$ be N samples drawn from p_{XY}. An estimator of the GC-loss is

$$\begin{aligned}\hat{J}_{\text{GC-loss}}(X,Y) &= G_{\alpha,\beta}(0) - \hat{V}_{\alpha,\beta}(X,Y)\\ &= \gamma_{\alpha,\beta} - \frac{1}{N}\sum_{i=1}^{N} G_{\alpha,\beta}(x_i-y_i)\\ &= \gamma_{\alpha,\beta} - \frac{1}{N}\sum_{i=1}^{N} G_{\alpha,\beta}(e_i)\end{aligned} \tag{10}$$

Property 6: Let $\mathbf{X} = [x_1,\ldots,x_N]^T, \mathbf{Y} = [y_1,\ldots,y_N]^T$. Then the function $\text{GCIM}(\mathbf{X},\mathbf{Y}) = \sqrt{\hat{J}_{\text{GC-loss}}(X,Y)}$, called the *generalized correntropy induced metric* (GCIM), defines a metric in the N-dimensional sample vector space when $0<\alpha\leq 2$.

Proof: See Appendix A.

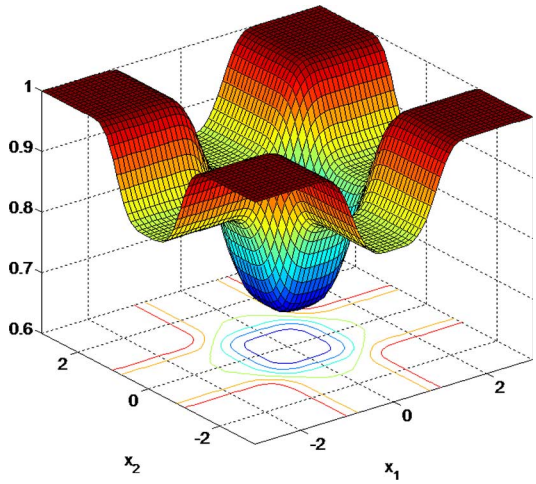


Fig. 1. Surface of the GCIM in 3D space ($\alpha = 4, \lambda = 1$).

For the case in which $\mathbf{X} = [x_1,x_2]^T, \mathbf{Y} = [0,0]^T, \alpha = 4$, and $\lambda = 1$,the surface of the $\text{GCIM}(\mathbf{X},\mathbf{Y})$ with respect to x_1 and x_2 is shown in Fig. 1. As one can see, the GCIM behaves like different norms (from L_α to L_0) of the data in different regions. This observation, which is similar to that obtained with Gaussian kernel [19], is confirmed by *Properties 7* and *8* below.

Property 7: As $\lambda\to 0+$ (or $x_i\to 0, i=1,\ldots,N$), the function $L_{\alpha,\beta}(\mathbf{X}) = \left(\frac{N}{\lambda\gamma_{\alpha,\beta}}\hat{J}_{\text{GC-loss}}(X,0)\right)^{1/\alpha}$ will approach the l_α-norm of $\mathbf{X}$, that is

$$L_{\alpha,\beta}(\mathbf{X}) \approx \|\mathbf{X}\|_\alpha = \left(\sum_{i=1}^{N}|x_i|^\alpha\right)^{1/\alpha}, \quad \text{as } \lambda\to 0 \tag{11}$$

Proof: See Appendix B.

Property 8: Assume that $|x_i|>\delta, \forall i: x_i\neq 0$, where δ is a small positive number. As $\lambda\to\infty$ (or $\beta\to 0+$), minimizing the function $L_{\alpha,\beta}(\mathbf{X})$ will be, approximately, equivalent to minimizing the l_0-norm of $\mathbf{X}$, that is

$$\min_{\mathbf{X}\in\Omega} L_{\alpha,\beta}(\mathbf{X}) \sim \min_{\mathbf{X}\in\Omega}\|\mathbf{X}\|_0, \quad \text{as } \lambda\to\infty \tag{12}$$

where Ω denotes a feasible set of $\mathbf{X}$.

Proof: Similar to the one presented in [32]. See Appendix C for the detailed derivation.

Below we present some optimization properties of the GC-loss. Similar results for the C-loss can be found in [31].

Property 9: Let $\mathbf{e} = [e_1,\ldots,e_N]^T$. Then the following statements hold:

1) if $0<\alpha\leq 1$, then the GC-loss $\hat{J}_{\text{GC-loss}}$ is concave at any $\mathbf{e}$ with $e_i\neq 0$ $(i=1,\ldots,N)$;
2) if $\alpha>1$, then the GC-loss $\hat{J}_{\text{GC-loss}}$ is convex at any $\mathbf{e}$ with $0<|e_i|\leq[(\alpha-1)/\alpha\lambda]^{1/\alpha}(i=1,\ldots,N)$;
3) if $\lambda\to 0+$, then for any $\mathbf{e}$ with $e_i\neq 0$ $(i=1,\ldots,N)$, the GC-loss $\hat{J}_{\text{GC-loss}}$ is concave at $\mathbf{e}$ for $0<\alpha\leq 1$, and convex at $\mathbf{e}$ for $\alpha>1$.

Proof: See Appendix D.

Property 10: For $\alpha>1$, the GC-loss $\hat{J}_{\text{GC-loss}}$ is a differentiable invex function of $e = [e_1,\ldots,e_N]^T$ with $e_i\leq M(i=1,\ldots,N)$, where M is an arbitrary positive number.

Proof: See Appendix E.

III. Adaptive Filtering Under GMCC Criterion

A. Cost Function

Similar to the correntropy, the generalized correntropy can also be used as an optimization cost in estimation-related problems. In the context of linear adaptive filtering, under the *generalized maximum correntropy criterion* (GMCC), the optimal weight vector of the filter can be solved by minimizing the following GC-loss function:

$$\begin{aligned} J_{\text{GC-loss}} &= \gamma_{\alpha,\beta} - \mathbf{E}\left[G_{\alpha,\beta}(e(i))\right] \\ &= \gamma_{\alpha,\beta}\left\{1 - \mathbf{E}\left[\exp\left(-\lambda|e(i)|^{\alpha}\right)\right]\right\} \end{aligned} \tag{13}$$

where the error

$$e(i) = d(i) - y(i) = d(i) - W^T X(i) \tag{14}$$

with $d(i) \in \mathbb{R}$ being the desired value at time i, $y(i) = W^T X(i)$ the output of the filter, $W = [w_1, w_2, \ldots, w_m]^T \in \mathbb{R}^m$ the weight vector, and $X(i) \in \mathbb{R}$ the input vector, usually given by

$$X(i) = [x(i), x(i-1), \ldots, x(i-m+1)]^T \tag{15}$$

where $x(i)$ is the input signal.

B. Optimal Solution

On the optimal weight vector of the GMCC adaptive filter, we summarize the main results as the following theorems.

Theorem 1: The optimal weight vector that minimizes $J_{\text{GC-loss}}$ can be expressed as

$$W_{\text{opt}} = \left[\mathbf{R}_{XX}^h\right]^{-1} P_{dX}^h \tag{16}$$

where $\mathbf{R}_{XX}^h = \mathbf{E}\left[h(e(i))X(i)X(i)^T\right]$ is a weighted autocorrelation matrix of the input signal, in which the weighting is a function of the error $h(e(i)) = \exp\left(-\lambda|e(i)|^{\alpha}\right)|e(i)|^{\alpha-2}$, and $P_{dX}^h = \mathbf{E}\left[h\left(e(i)\right)d(i)X(i)\right]$ is a weighted cross-correlation vector between the desired and the input vector.

Proof: Let $\frac{\partial}{\partial W} J_{\text{GC-loss}} = 0$, we have

$$\begin{aligned} &\mathbf{E}\left[\exp\left(-\lambda|e(i)|^{\alpha}\right)|e(i)|^{\alpha-1}\text{sign}\left(e(i)\right)X(i)\right] = 0 \\ &\Rightarrow \mathbf{E}\left[h\left(e(i)\right)\left(d(i) - W^T X(i)\right)X(i)\right] = 0 \\ &\Rightarrow \mathbf{E}\left[h\left(e(i)\right)X(i)X(i)^T\right]W = \mathbf{E}\left[h\left(e(i)\right)d(i)X(i)\right] \\ &\Rightarrow W = \left[\mathbf{R}_{XX}^h\right]^{-1} P_{dX}^h \end{aligned} \tag{17}$$

Remark 3: Note that the above solution is not a closed-form solution, since both matrices $\mathbf{R}_{XX}^h$ and P_{dX}^h depend on the weight vector W through the error $e(i)$. The solution (16) is actually a *fixed-point* equation.

For the case $\alpha = 2$, we have $h(e(i)) = \exp\left(-\lambda e^2(i)\right)$. In this case, as $\lambda \to 0+$, we have $h(e(i)) \to 1$,and $W_{\text{opt}} \to \mathbf{R}_{XX}^{-1} P_{dX}$, with $\mathbf{R}_{XX} = \mathbf{E}\left[X(i)X(i)^T\right]$, $P_{dX} = \mathbf{E}\left[d(i)X(i)\right]$, which corresponds to the well-known *Wiener solution*.

Theorem 2: If and are both zero-mean Gaussian processes, then the optimal solution under GMCC criterion is equal to the Wiener solution.

Proof: See Appendix F.

C. Adaptive Algorithm

Based on the cost function (13), a stochastic gradient based adaptive algorithm, called in this work the GMCC algorithm, can be simply derived as

$$\begin{aligned} W(i+1) &= W(i) + \mu\frac{\partial}{\partial W(i)}\exp\left(-\lambda|e(i)|^{\alpha}\right) \\ &= W(i) - \mu\lambda\alpha\exp\left(-\lambda|e(i)|^{\alpha}\right)|e(i)|^{\alpha-1}\text{sign}(e(i))\frac{\partial e(i)}{\partial W(i)} \\ &= W(i) + \eta\exp\left(-\lambda|e(i)|^{\alpha}\right)|e(i)|^{\alpha-1}\text{sign}(e(i))X(i) \end{aligned} \tag{18}$$

where $\eta = \mu\lambda\alpha$ is the step-size parameter.

We have the following observations:

1) When $\alpha = 2$, the GMCC algorithm becomes
$$W(i+1) = W(i) + \eta\exp\left(-\lambda e^2(i)\right)e(i)X(i) \tag{19}$$
which is the original MCC algorithm [20].
2) The weight update equation (18) can be rewritten as
$$W(i+1) = W(i) + \eta(i)|e(i)|^{\alpha-1}\text{sign}(e(i))X(i) \tag{20}$$
where $\eta(i) = \eta\exp\left(-\lambda|e(i)|^{\alpha}\right)$. Therefore, the GMCC algorithm can be viewed as an LMP algorithm with $p = \alpha$ and a variable step-size $\eta(i)$. The LMP algorithm is derived under the LMP (*least mean P-power*) criterion, including the SA ($p = 1$), LMS ($p = 2$) and LMF ($p = 4$) as special cases (see [5] for the details).
3) When $\lambda \to 0+$, we have $\eta(i) \to \eta$. In this case, the GMCC algorithm reduces to the traditional LMP algorithm with $p = \alpha$:
$$W(i) = W(i-1) + \eta|e(i)|^{\alpha-1}\text{sign}(e(i))X(i) \tag{21}$$
In particular, when $\alpha = 2$, (21) becomes the well-known LMS algorithm:
$$W(i) = W(i-1) + \eta e(i)X(i) \tag{22}$$
4) When $|e(i)| \to \infty$, we have $\eta(i) \to 0$. Thus, a large error will have little influence on the filter weights. This implies that the GMCC algorithm will be robust to large outliers (or impulsive noises), which often cause large errors.

Remark 4: One can derive various variants of the GMCC algorithm, such as the variable step-size GMCC and normalized GMCC, where the step-size is changed across iterations or divided by the squared norm of the input vector.

The computational complexity of the GMCC algorithm is almost the same as that of the LMP algorithm, and the only extra computational effort needed is to calculate the term $\exp\left(-\lambda|e(i)|^{\alpha}\right)$, which is obviously not expensive.

IV. Stability and Steady-State Performance

In this section, we analyze the stability and steady-state mean-square performance of the proposed GMCC algorithm. The algorithm (18) can be written in a general form:

$$W(i) = W(i-1) + \eta f\left(e(i)\right)X(i) \tag{23}$$

where $f\left(e(i)\right)$ is a nonlinear function of $e(i)$, that is,

$$f\left(e(i)\right) = \exp\left(-\lambda\left|e(i)\right|^{\alpha}\right)\left|e(i)\right|^{\alpha-1}\text{sign}(e(i)) \tag{24}$$

Assume that the desired signal $d(i)$ can be expressed as

$$d(i) = W_0{}^T X(i) + v(i) \tag{25}$$

where W_0 denotes an unknown weight vector that needs to be estimated, and $v(i)$ stands for the disturbance noise, with variance σ_v^2. Then, we have

$$e(i) = \tilde{W}(i-1)^T X(i) + v(i) = e_a(i) + v(i) \tag{26}$$

where $\tilde{W}(i-1) = W_0 - W(i-1)$ is the *weigh error vector* at iteration $i-1$, and $e_a(i) = \tilde{W}(i-1)^T X(i)$ is referred to as the *a priori* error. After some simple algebra, one can obtain [33]:

$$\mathbf{E}\left[\left\|\tilde{W}(i)\right\|^2\right] = \mathbf{E}\left[\left\|\tilde{W}(i-1)\right\|^2\right] - 2\eta\mathbf{E}\left[e_a(i)f(e(i))\right] + \eta^2\mathbf{E}\left[\|X(i)\|^2 f^2(e(i))\right] \tag{27}$$

A. Stability Analysis

From (27), if the step-size η is chosen such that for all i

$$\eta \le \frac{2\mathbf{E}\left[e_a(i)f(e(i))\right]}{\mathbf{E}\left[\|X(i)\|^2 f^2(e(i))\right]} = \frac{2\mathbf{E}\left[e_a(i)\exp\left(-\lambda|e(i)|^\alpha\right)|e(i)|^{\alpha-1}\mathrm{sign}(e(i))\right]}{\mathbf{E}\left[\|X(i)\|^2\exp\left(-2\lambda|e(i)|^\alpha\right)|e(i)|^{2(\alpha-1)}\right]} \tag{28}$$

then the sequence of weight error power $\mathbf{E}\left[\left\|\tilde{W}(i)\right\|^2\right]$ will be decreasing and converging.

The bound in (28) seems very complicated and is not easy to be verified in a practical situation. To better understand the stability of the GMCC, it's more important to investigate the *probability of divergence* (POD) [34]. Here the divergence means $\lim_{i\to\infty}\left\|\tilde{W}(i)\right\|^2 = \infty$ in a realization of an adaptive algorithm. For the LMF algorithm, the POD is nonzero when the input distribution has infinite support, no matter how small the step-size is chosen [34]. Below we present a simple example to show that the GMCC is very stable and may have zero POD, no matter what input distribution. Let's consider a scalar filtering case as in [34], in which the desired signal is $d(i) = W_0 X(i)$, where W_0 and $X(i)$ are both scalars, and the noise $v(i)$ is assumed to be zero for simplicity. Then, we have

$$\begin{aligned}\tilde{W}(i) &= \tilde{W}(i-1) - \eta f\left(e(i)\right)X(i)\\ &= \left[1-\eta\exp\left(-\lambda|e(i)|^\alpha\right)|e(i)|^{\alpha-2}X^2(i)\right]\tilde{W}(i-1)\\ &= \left[1-\eta\exp\left(-\lambda|e(i)|^\alpha\right)|e(i)|^\alpha\left|\tilde{W}(i-1)\right|^{-2}\right]\tilde{W}(i-1)\end{aligned} \tag{29}$$

It is easy to show

$$0 \le \exp\left(-\lambda|e(i)|^\alpha\right)|e(i)|^\alpha \le \frac{1}{\lambda}\exp(-1) \tag{30}$$

So it holds that if $\left|\tilde{W}(i-1)\right|^2 \ge \frac{\eta}{2\lambda}\exp(-1)$, then

$$0 \le \eta\exp\left(-\lambda|e(i)|^\alpha\right)|e(i)|^\alpha\left|\tilde{W}(i-1)\right|^{-2} \le 2 \tag{31}$$

In this case, we have

$$\left|1-\eta\exp\left(-\lambda|e(i)|^\alpha\right)|e(i)|^\alpha\left|\tilde{W}(i-1)\right|^{-2}\right| \le 1, \text{ and}$$

$$\left|\tilde{W}(i)\right|^2 = \left|1-\eta\exp\left(-\lambda|e(i)|^\alpha\right)|e(i)|^\alpha\left|\tilde{W}(i-1)\right|^{-2}\right|^2 \times\left|\tilde{W}(i-1)\right|^2 \le \left|\tilde{W}(i-1)\right|^2 \tag{32}$$

Therefore, the limit (if exists) of $\left|\tilde{W}(i)\right|$ always satisfies $\lim_{i\to\infty}\left|\tilde{W}(i)\right| < \infty$, which implies that in this simple example the GMCC will never diverge (or its POD is zero). Note that in this case the LMF will have a non-zero POD [34].

When $W(i)$ is a vector and there is a noise $v(i)$, a rigorous analysis of the POD of GMCC is very difficult and is left open in this work. However, we believe that the above scalar example explains clearly what is the mechanism of the zero POD of GMCC. Our simulation results also suggest that in most cases the POD of GMCC is zero, even when the noise signal contains large outliers.

B. Steady-State Mean Square Performance

With a similar derivation presented in [35], one can analyze the mean square transient behaviors of the algorithm (18). This is a trivial but quite tedious task since we have to evaluate the expectations $\mathbf{E}\left[e_a(i)\exp\left(-\lambda|e(i)|^\alpha\right)|e(i)|^{\alpha-1}\mathrm{sign}(e(i))\right]$ and $\mathbf{E}\left[\|X(i)\|^2\exp\left(-2\lambda|e(i)|^\alpha\right)|e(i)|^{2(\alpha-1)}\right]$. In the following, we only analyze the steady-state mean square performance by using a Taylor expansion method [23].

As the filter reaches the steady-state, we have $\mathbf{E}\left[\left\|\tilde{W}(i)\right\|^2\right] = \mathbf{E}\left[\left\|\tilde{W}(i-1)\right\|^2\right]$. By (27), it holds that

$$2\mathbf{E}\left[e_a(i)f(e(i))\right] = \eta\mathbf{E}\left[\|X(i)\|^2 f^2(e(i))\right] \tag{33}$$

Assume that $\|X(i)\|^2$ is asymptotically uncorrelated with $f^2(e(i))$ (the rationality of this assumption has been discussed in [33]). Then (33) becomes

$$2\mathbf{E}\left[e_a(i)f(e(i))\right] = \eta\mathrm{Tr}\left(\mathbf{R}_{XX}\right)\mathbf{E}\left[f^2(e(i))\right] \tag{34}$$

In the steady-state, the distributions of $e_a(i)$ and $e(i)$ are independent of i, thus one can omit the time index and simply write (34) as

$$2\mathbf{E}\left[e_a f(e)\right] = \eta Tr\left(\mathbf{R}_{XX}\right)\mathbf{E}\left[f^2(e)\right] \tag{35}$$

Let $S = \lim_{i\to\infty}\mathbf{E}\left[e_a^2(i)\right] = \mathbf{E}\left[e_a^2\right]$ be the steady-state *excess mean square error* (EMSE). An approximate analytical expression of S can be derived. Before proceeding we give two common assumptions:

A1: The noise $v(i)$ is zero-mean, independent, identically distributed, and is independent of the input.

A2: The *a priori* error $e_a(i)$ is zero-mean and independent of the noise.

Taking the Taylor expansion of $f(e)$ with respect to e_a around v, we obtain

$$f(e) = f(e_a+v) = f(v) + f'(v)e_a + \frac{1}{2}f''(v)e_a^2 + o\left(e_a^2\right) \tag{36}$$

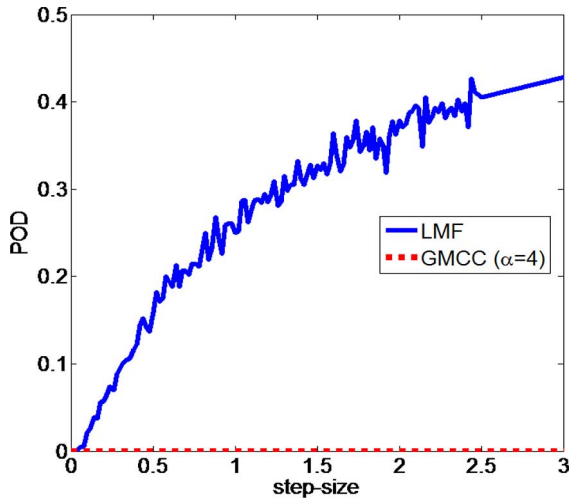


Fig. 2. PODs with different step-sizes.

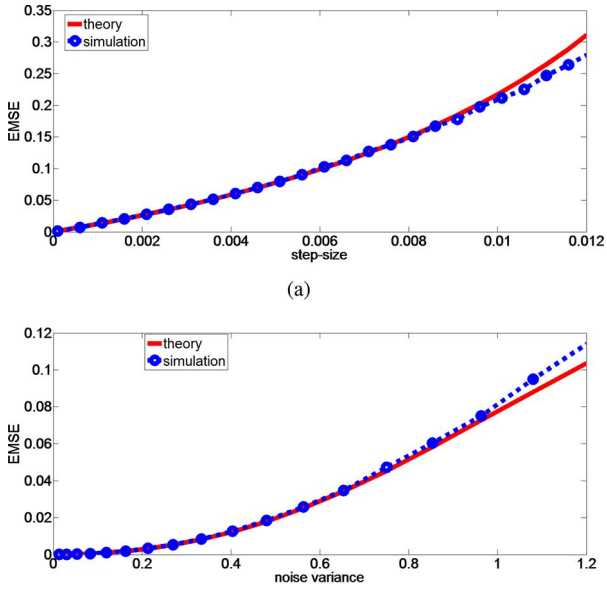


Fig. 3. Theoretical and simulated EMSEs: (a) with different step-sizes; (b) with different noise variances.

where $f'(v)$ and $f''(v)$ are the first and second derivatives, and $o\left(e_a^2\right)$ denotes the third and higher-order terms. If $\mathbf{E}\left[o(e_a^2)\right]$ is small enough, then based on the assumptions **A1** and **A2** we can derive

$$\begin{aligned}\mathbf{E}\left[e_a f(e)\right] &= \mathbf{E}\left[e_a f(v) + f'(v)e_a^2 + o(e_a^2)\right] \\ &\approx \mathbf{E}\left[f'(v)\right] S \qquad (37)\\ \mathbf{E}\left[f^2(e)\right] &\approx \mathbf{E}\left[f^2(v)\right] + \mathbf{E}\left[f(v)f''(v) + |f'(v)|^2\right] S \qquad (38)\end{aligned}$$

Substituting (37) and (38) into (35), we obtain (39) at the bottom of the page. in which

$$\begin{aligned}\zeta(v) &= \left[f(v)f''(v) + |f'(v)|^2\right] \\ &= \exp\left(-2\lambda|v|^\alpha\right)|v|^{2\alpha-4} \\ &\quad \times \left[(\alpha-1)(2\alpha-3) - 5\lambda\alpha(\alpha-1)|v|^\alpha + 2\lambda^2\alpha^2|v|^{2\alpha}\right]\end{aligned} \qquad (40)$$

When the step-size η is small enough, (39) can be simplified to

$$S \approx \frac{\eta \mathrm{Tr}(\mathbf{R}_{XX})\mathbf{E}\left[\exp\left(-2\lambda|v|^\alpha\right)|v|^{2\alpha-2}\right]}{2\mathbf{E}\left[\exp\left(-\lambda|v|^\alpha\right)|v|^{\alpha-2}\left((\alpha-1) - \lambda\alpha|v|^\alpha\right)\right]} \qquad (41)$$

Remark 5: Given a noise distribution, one can evaluate the expectations in (39) and obtain a theoretical value of the steady-state EMSE. It is, however, worth noting that the steady-state EMSE of (39) is derived under the assumption that the steady-state *a priori* error e_a is small such that its third and higher-order terms are negligible. When the step-size or noise power is too large, the *a priori* error will also be large. In this case, the derived EMSE value will not accurately enough characterize the performance.

V. Simulation Results

Now we present simulation results to confirm the theoretical predictions and demonstrate the desirable performance of the proposed GMCC algorithm.

A. Probability of Divergence

First, we investigate the stability problem of the GMCC algorithm. The steady-state performance is valid only when the algorithm does not diverge. In many cases, however, an adaptive algorithm may diverge especially at the initial convergence stage. Below we present some simulation results about the probability of divergence (POD) of the GMCC (with $\alpha = 4.0, \lambda = 2.0$), compared with that of the LMF algorithm, whose probability of divergence has been studied in [34]. The weight vector of the unknown system is assumed to be $W_0 = [0.1, 0.2, 0.3, 0.4, 0.5, 0.4, 0.3, 0.2, 0.1]$, and the initial weight vector of the adaptive filter is a null vector. The input

$$\begin{aligned}S &= \frac{\eta Tr(\mathbf{R}_{XX})\mathbf{E}\left[f^2(v)\right]}{2\mathbf{E}\left[f'(v)\right] - \eta \mathrm{Tr}(\mathbf{R}_{XX})\mathbf{E}\left[f(v)f''(v) + |f'(v)|^2\right]} \\ &= \frac{\eta Tr(\mathbf{R}_{XX})\mathbf{E}\left[\exp(-2\lambda|v|^\alpha)|v|^{2\alpha-2}\right]}{2\mathbf{E}\left[\exp(-\lambda|v|^\alpha)|v|^{\alpha-2}((\alpha-1) - \lambda\alpha|v|^\alpha)\right] - \eta Tr(\mathbf{R}_{XX})\mathbf{E}\left[\zeta(v)\right]}\end{aligned} \qquad (39)$$

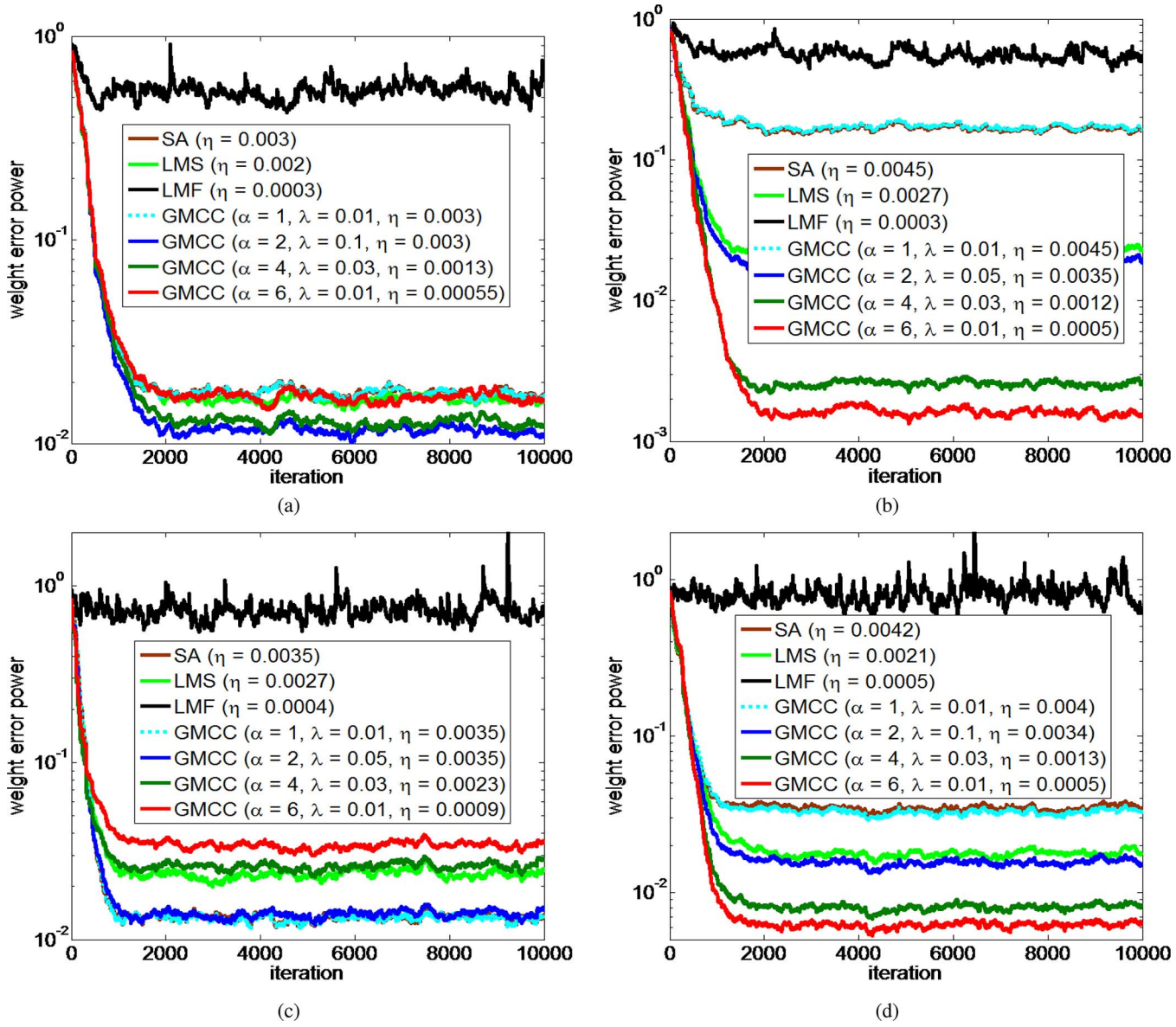


Fig. 4. Convergence curves with different distributions of $A(i)$: (a) Gaussian; (b) Binary; (c) Laplace; (d) Uniform.

signal and the disturbance noise are both zero-mean Gaussian with variance 1.0. The PODs with different step-sizes are illustrated in Fig. 2. To evaluate the PODs, 1000 independent Monte Carlo simulations were performed and in each simulation, 1000 iterations were run. We labeled a learning curve as "diverging" if at the last iteration the weight error power $\|W_0 - W(i)\|^2$ is larger than 100. As one can see clearly, compared with the LMF, the GMCC is very stable and does not diverge in this particular situation.

B. Steady-State Performance

Second, we show the theoretical and simulated steady-state performance of the GMCC. In the simulation, we set $\alpha = 4$, and $\lambda = 0.03$. The filter length is 20, the input signal is a zero-mean white Gaussian process with variance 1.0, and the disturbance noise is assumed to be zero-mean and uniform distributed over $\left[-\sqrt{3}, \sqrt{3}\right]$. Fig. 3 shows the steady-state EMSEs with different step-sizes and the noise variances, where the simulated EMSEs are computed as an average over 100 independent Monte Carlo simulations, and in each simulation, 50000 iterations were run to ensure the algorithm to reach the steady state, and the steady-state EMSE was obtained as an average over the last 1000 iterations. One can observe: i) the steady-state EMSEs are increasing with step-size and noise variance; ii) when the step-size and noise variance are small, the steady-state EMSEs computed by simulations match very well the theoretical values computed by (39); iii) when the step-size and noise variance become large, the experimental results will, however, gradually differ from the theoretical values, and this coincides with the theoretical prediction.

C. Performance Comparison With Other Algorithms

Third, we compare the performance of the GMCC and the LMP family algorithms with different p values, namely SA ($p = 1$), LMS ($p = 2$), and the LMF ($p = 4$). The

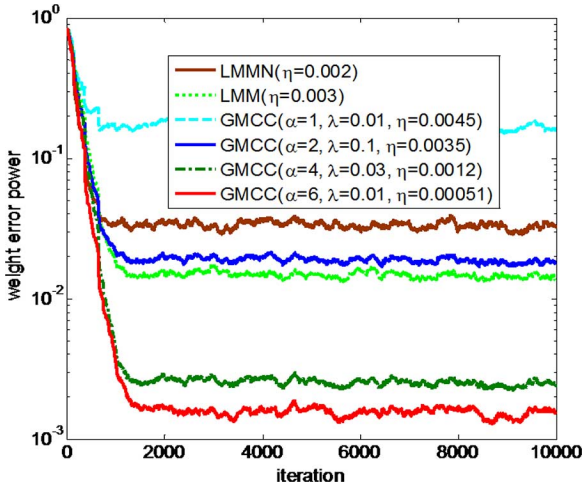


Fig. 5. Convergence curves of LMMN, LMM and GMCC.

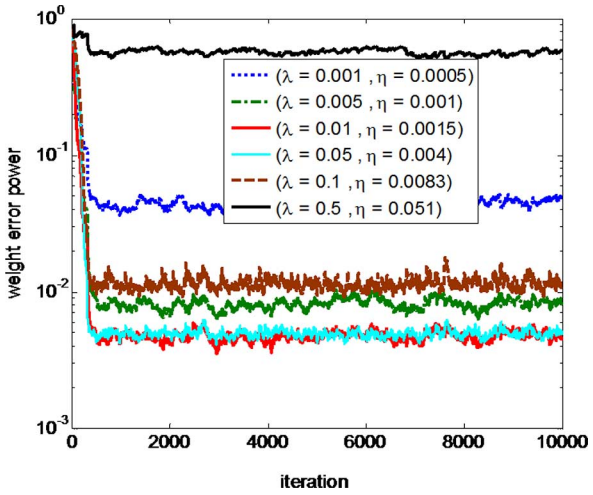


Fig. 6. Convergence curves with different λ ($\alpha = 6$).

unknown system is assumed to be the same as that in the first simulation. In particular, we consider a noise model with form $v(i) = (1 - a(i))A(i) + a(i)B(i)$, where $a(i)$ is a binary independent and identically distributed process with $\Pr\{a(i) = 1\} = c, \Pr\{a(i) = 0\} = 1 - c$, and $0 \leq c \leq 1$ is an occurrence probability; whereas $A(i)$ is a noise process with smaller variance, and $B(i)$ is another noise process with substantially much larger variance to represent large outliers (or impulsive disturbances). The noise processes $A(i)$ and $B(i)$ are mutually independent and they are both independent of $a(i)$. In the simulation, c is set at 0.06, and $B(i)$ is a white Gaussian process with zero-mean and variance 15. For the noise $A(i)$, we consider four distributions: a) Gaussian distribution with zero-mean and unit variance; b) Binary distribution over $\{-1, 1\}$ with probability mass $\Pr\{x = -1\} = \Pr\{x = 1\} = 0.5$; c) Laplace distribution with zero-mean and unit variance; d) Uniform distribution over $\left[-\sqrt{3}, \sqrt{3}\right]$. The convergence curves in terms of the weight error power $\|W_0 - W(i)\|^2$ averaged over 100 independent Monte Carlo runs are shown in Fig. 4. In the simulation, the step-sizes are chosen such that all the algorithms have almost the same initial convergence speed, and the parameter λ in GMCC is experimentally chosen such that the algorithm achieves desirable performance. The final selected values of these parameters are reported in the figures. From simulation results we can observe: 1) the GMCC family algorithms are much more stable (robust) than the LMP family algorithms (In this example, when $p > 4$, the LMP will not converge); 2) the GMCC with $\alpha \neq 2$ may outperform significantly the original MCC ($\alpha = 2$) algorithm. In particular, the GMCC with $\alpha = 6$ achieves the best performance when $A(i)$ is of Binary or Uniform distribution. Additionally, in Fig. 5, we demonstrate the performance comparison among GMCC, least mean mixed-norm (LMMN) [9] and least mean M-estimate (LMM) [11] algorithms, where $A(i)$ is Binary distributed. The parameters in LMMN and LMM are selected by scanning for the best results. Evidently, the GMCC with a proper α value can outperform both LMMN and LMM significantly. To investigate how the parameter λ affects the performance, we show in Fig. 6 the convergence curves of the GMCC with different λ, where $A(i)$ is Binary distributed and $\alpha = 6$. As one can see, the performance will become worse when λ is too small or too large. In this example, the GMCC algorithm achieves the best performance when λ is about 0.01. In order to further demonstrate the robustness of the GMCC against large outliers, we increase the variance of the outlier noise $B(i)$ from 15 to 100. In this case, the LMP family algorithms, except SA and LMS, will diverge, while the GMCC family algorithms can still work well. We present in Fig. 7 the simulation results when $A(i)$ is Uniform distributed.

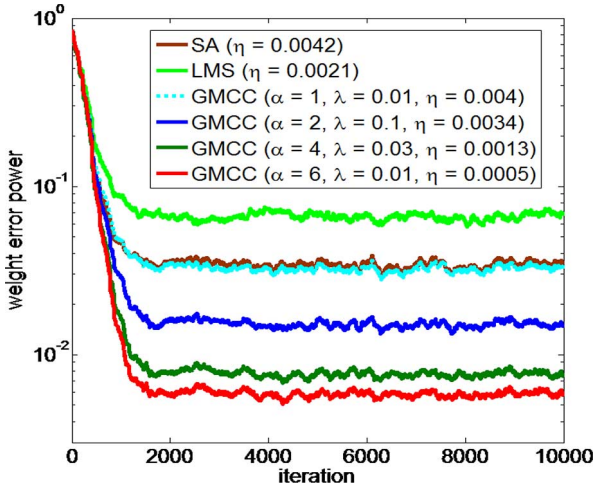


Fig. 7. Convergence curves with larger outliers.

D. Application to EEG Denoising

Finally, we consider the denoising of an electroencephalograph (EEG) signal that contains interferences caused by the EOG signal. The task is to remove the interferences by subtracting the output of an adaptive filter from the original EEG signal. In our approach, the adaptive filter is trained by the GMCC algorithm, with the EOG signal being the input and the original EEG being the desired signal. The filter length is set to ten. We use the power ratio $\left(10 \log\left(E\left[d_i^2\right] / E\left(e_i^2\right)\right)\right)$ between the original and cleaned EEGs as a measure of performance.

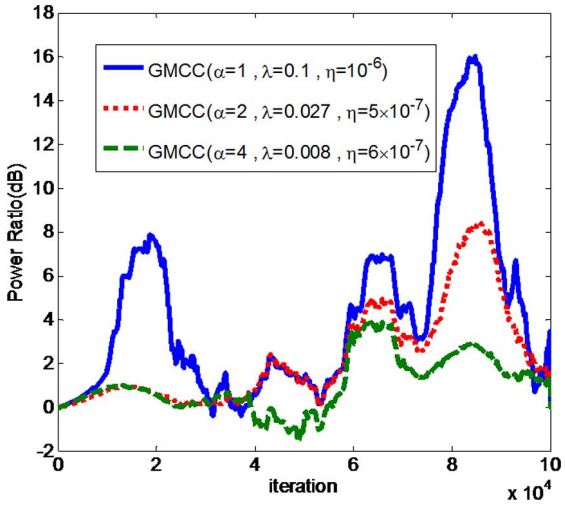


Fig. 8. Power ratio between the original and cleaned EEGs.

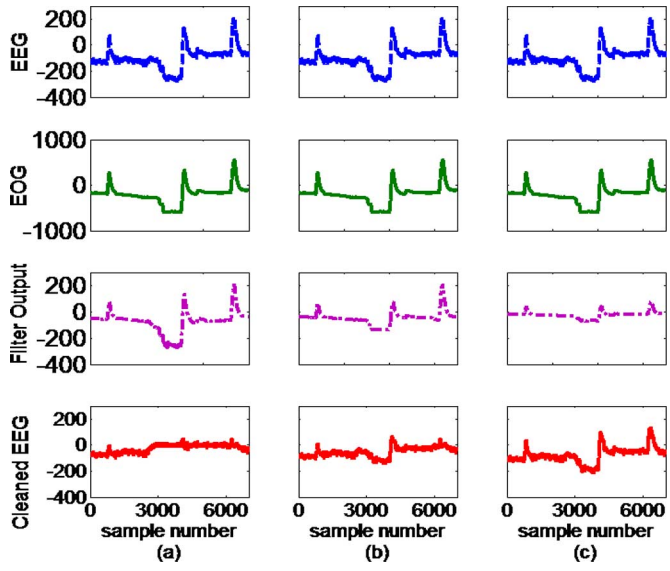


Fig. 9. EEG, EOG, filter output and cleaned EEG: (a) GMCC ($\alpha = 1$); (b) GMCC ($\alpha = 2$); (c) GMCC ($\alpha = 4$).

A larger power ratio means more interferences are removed from the original signal. For GMCC with different α values ($\alpha = 1, 2, 4$), the power ratios (after smoothing using a running window of 10000 samples) are illustrated in Fig. 8. In this figure, the GMCC with $\alpha = 1$ achieves the best performance. The results confirm again the fact that the GMCC with $\alpha \neq 2$ may outperform the original MCC (corresponding to the GMCC with $\alpha = 2$). The original EEG, EOG, filter output and cleaned EEG (all within a segment of 6000 samples) are shown in Fig. 9, in which the sample rate is 1000 Hz and the unit is mV.

VI. Conclusion

Correntropy is a recently proposed similarity measure, and the maximum correntropy criterion (MCC) has been widely applied in domains of machine learning and signal processing. In previous studies, the kernel function in correntropy has been however limited to the Gaussian kernel. Gaussian kernel is desirable in many cases but obviously, it is not always the best choice. In this work, we proposed a generalized correntropy, using the generalized Gaussian density (GGD) function as the kernel. The new definition is very general and flexible, which includes the original correntropy with Gaussian kernel as a special case. Some important properties of the generalized correntropy were presented. The generalized maximum correntropy criterion (GMCC) was also proposed as an optimality criterion in estimation related problems. In particular, we applied the GMCC criterion to adaptive filtering. The optimal solution under GMCC and an adaptive algorithm, called the GMCC algorithm, were derived. Further, we investigated the stability and steady-state performance of the developed algorithm. A simple example was presented to show that the GMCC algorithm is very stable and will have a zero probability of divergence (POD). A theoretical value of the steady-state *excess mean square error* (EMSE) was also derived. Theoretical predictions and excellent performance of the GMCC were confirmed by simulation results. Of course, similar to the original correntropy, the generalized correntropy can also be applied to other domains such as classification, principal components analysis (PCA), period estimation, spectral characterization and so on. This will be an interesting area of future study.

Appendix A
Proof of Property 6

Proof: When $0 < \alpha \leq 2$, we can construct a nonlinear mapping: $\Phi_{\alpha,\beta}(\mathbf{X}) = \left[\varphi_{\alpha,\beta}(x_1)^T, \ldots, \varphi_{\alpha,\beta}(x_N)^T\right]^T$, such that $\|\Phi_{\alpha,\beta}(\mathbf{X}) - \Phi_{\alpha,\beta}(\mathbf{Y})\| = \sqrt{2N} \times GCIM(\mathbf{X}, \mathbf{Y})$. In this case the GCIM function $\mathrm{GCIM}(\mathbf{X}, \mathbf{Y})$ defines a "*Euclidean distance*" in the Hilbert space $\boldsymbol{F}_\kappa^N$, and hence is a metric in the sample vector space since it satisfies: i) $\mathrm{GCIM}(\mathbf{X}, \mathbf{Y}) \geq 0$; ii) $\mathrm{GCIM}(\mathbf{X}, \mathbf{Y}) = \mathrm{GCIM}(\mathbf{Y}, \mathbf{X})$; iii) $\mathrm{GCIM}(\mathbf{X}, \mathbf{Z}) \leq \mathrm{GCIM}(\mathbf{X}, \mathbf{Y}) + \mathrm{GCIM}(\mathbf{Y}, \mathbf{Z})$.

Appendix B
Proof of Property 7

Proof: As $\lambda \to 0+$ (or $x_i \to 0, i = 1, \ldots, N$), we have

$$\begin{aligned} L_{\alpha,\beta}(\mathbf{X}) &= \left(\frac{N}{\lambda\gamma_{\alpha,\beta}} \hat{J}_{\text{GC-loss}}(X, 0)\right)^{1/\alpha} \\ &= \left[\frac{N}{\lambda\gamma_{\alpha,\beta}}\left(G_{\alpha,\beta}(0) - \frac{1}{N}\sum_{i=1}^{N} G_{\alpha,\beta}(x_i)\right)\right]^{1/\alpha} \\ &\approx \left[\frac{N}{\lambda\gamma_{\alpha,\beta}}\left(\gamma_{\alpha,\beta} - \frac{1}{N}\sum_{i=1}^{N} \gamma_{\alpha,\beta}\left(1 - \lambda|x_i|^\alpha\right)\right)\right]^{1/\alpha} \\ &= \left[\sum_{i=1}^{N} |x_i|^\alpha\right]^{1/\alpha} \end{aligned} \tag{42}$$

APPENDIX C
PROOF OF PROPERTY 8

Proof: Let $\mathbf{X}_0$ be the solution obtained by minimizing $\|X\|_0$ over $\mathbf{\Omega}$ and $\mathbf{X}_l$ the solution achieved by minimizing $L_{\alpha,\beta}(\mathbf{X})$. Then $L_{\alpha,\beta}(\mathbf{X}_l) \leq L_{\alpha,\beta}(\mathbf{X}_0)$, and hence

$$\sum_{i=1}^{N} G_{\alpha,\beta}\left((\mathbf{X}_l)_i\right) \geq \sum_{i=1}^{N} G_{\alpha,\beta}\left((\mathbf{X}_0)_i\right) \tag{43}$$

where $(\mathbf{X}_l)_i$ denotes the ith component of $\mathbf{X}_l$. It follows that

$$\begin{aligned}(N - \|\mathbf{X}_l\|_0) + \sum_{i=1,(\mathbf{X}_l)_i\neq 0}^{N} \exp\left(-\lambda|(\mathbf{X}_l)_i|^{\alpha}\right) \\ \geq (N - \|\mathbf{X}_0\|_0) + \sum_{i=1,(\mathbf{X}_0)_i\neq 0}^{N} \exp\left(-\lambda|(\mathbf{X}_0)_i|^{\alpha}\right)\end{aligned} \tag{44}$$

Thus we have

$$\begin{aligned}\|\mathbf{X}_l\|_0 - \|\mathbf{X}_0\|_0 \leq \sum_{i=1,(\mathbf{X}_l)_i\neq 0}^{N} \exp\left(-\lambda|(\mathbf{X}_l)_i|^{\alpha}\right) \\ - \sum_{i=1,(\mathbf{X}_0)_i\neq 0}^{N} \exp\left(-\lambda|(\mathbf{X}_0)_i|^{\alpha}\right)\end{aligned} \tag{45}$$

Since $|x_i| > \delta, \forall i : x_i \neq 0$, as $\lambda \to \infty$ the right hand side of (45) will approach zero. Therefore, if λ is large enough, it holds that

$$\|\mathbf{X}_0\|_0 \leq \|\mathbf{X}_l\|_0 \leq \|\mathbf{X}_0\|_0 + \varepsilon \tag{46}$$

where ε is a small positive number arbitrarily close to zero.

APPENDIX D
PROOF OF PROPERTY 9

Proof: The Hessian (if exists) of $\hat{J}_{\text{GC-loss}}$ with respect to $\mathbf{e}$ is

$$\begin{aligned}\mathbf{H}_{\hat{J}_{\text{GC-loss}}}(\mathbf{e}) = -\frac{\alpha\lambda\gamma_{\alpha,\beta}}{N} \\ \times \begin{pmatrix} T(e_1)\left(\alpha\lambda|e_1|^{\alpha} - (\alpha - 1)\right) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T(e_N)\left(\alpha\lambda|e_N|^{\alpha} - (\alpha - 1)\right) \end{pmatrix}\end{aligned} \tag{47}$$

where $T(x) = \exp\left(-\lambda|x|^{\alpha}\right)|x|^{\alpha-2}$. From (47) one can see:

i) if $0 < \alpha \leq 1$, then $\mathbf{H}_{\hat{J}_{\text{GC-loss}}}(\mathbf{e}) \leq \mathbf{0}$ for any $\mathbf{e}$ with $e_i \neq 0 (i = 1, \ldots, N)$;
ii) if $\alpha > 1$, then $\mathbf{H}_{\hat{J}_{\text{GC-loss}}}(\mathbf{e}) \geq \mathbf{0}$ for any $\mathbf{e}$ with $0 < |e_i| \leq [(\alpha - 1)/\alpha\lambda]^{1/\alpha} (i = 1, \ldots, N)$;
iii) if $\lambda \to 0+$, then for any $\mathbf{e}$ with $e_i \neq 0 (i = 1, \ldots, N)$, we have $\mathbf{H}_{\hat{J}_{\text{GC-loss}}}(\mathbf{e}) \leq \mathbf{0}$ for $0 < \alpha \leq 1$, and $\mathbf{H}_{\hat{J}_{\text{GC-loss}}}(\mathbf{e}) \geq \mathbf{0}$ for $\alpha > 1$.

APPENDIX E
PROOF OF PROPERTY 10

Proof: A differentiable function $f : S \mapsto \mathbb{R}(S \subset \mathbb{R}^N)$ is said to be invex, if and only if [31]

$$f(x_2) \geq f(x_1) + q(x_1, x_2)^T \nabla f(x_1) \tag{48}$$

where $\nabla f(x)$ denotes the gradient of f with respect to x, and $q(x_1, x_2)$ is some vector valued function. For $\alpha > 1$, the GC-loss $\hat{J}_{\text{GC-loss}}$ is a differentiable function of $\mathbf{e}$, and the gradient $\nabla\hat{J}_{\text{GC-loss}}(\mathbf{e})$ is (49) at the bottom of the page, where $\text{sign}(\cdot)$ is the sign function. Since $e_i \leq M$, we have $\nabla\hat{J}_{\text{GC-loss}}(\mathbf{e}) = \mathbf{0}$ if and only if $\mathbf{e} = \mathbf{0}$. On the other hand, we have $\hat{J}_{\text{GC-loss}}(\mathbf{e}) \geq \hat{J}_{\text{GC-loss}}(\mathbf{0}) = \mathbf{0}$. So we can construct the function shown in (50) at the bottom of the page, such that it holds

$$\hat{J}_{\text{GC-loss}}(\mathbf{e}_2) \geq \hat{J}_{\text{GC-loss}}(\mathbf{e}_1) + q(\mathbf{e}_1, \mathbf{e}_2)^T \nabla\hat{J}_{\text{GC-loss}}(\mathbf{e}_1) \tag{51}$$

APPENDIX F
PROOF OF THEOREM 2

Proof: First, we derive

$$\begin{aligned}\mathbf{E}[G_{\alpha,\beta}(e(i))] &= \int_{\mathbb{R}} G_{\alpha,\beta}(e) p_{e(i)}(e) de \\ &= \int_{\mathbb{R}} p_{e(i)}(e) \left(\int_0^{\gamma_{\alpha,\beta}} \mathbb{I}\left(\xi \leq G_{\alpha,\beta}(e)\right) d\xi \right) de \\ &= \int_0^{\gamma_{\alpha,\beta}} \left(\int_{-\infty}^{\infty} p_{e(i)}(e) \mathbb{I}\left(\xi \leq G_{\alpha,\beta}(e)\right) de \right) d\xi \\ &= \int_0^{\gamma_{\alpha,\beta}} \left(\int_{\{e: G_{\alpha,\beta}(e) \geq \xi\}} p_{e(i)}(e) de \right) d\xi \\ &= \int_0^{\gamma_{\alpha,\beta}} \left(\int_{-\varepsilon}^{\varepsilon} p_{e(i)}(e) de \right) d\xi \end{aligned} \tag{52}$$

where $\mathbb{I}(\cdot)$ denotes an indicator function, and ε is a certain positive number satisfying $G_{\alpha,\beta}(\varepsilon) = G_{\alpha,\beta}(-\varepsilon) = \xi$. Since $x(i)$

$$\nabla\hat{J}_{\text{GC-loss}}(\mathbf{e}) = \frac{\lambda\alpha\gamma_{\alpha,\beta}}{N}\left[\exp\left(-\lambda|e_1|^{\alpha}\right)|e_1|^{\alpha-1}\text{sign}(e_1) \quad \ldots \quad \exp\left(-\lambda|e_N|^{\alpha}\right)|e_N|^{\alpha-1} sign(e_N)\right]^T \tag{49}$$

$$q(\mathbf{e}_1, \mathbf{e}_2) = \begin{cases} \frac{\hat{J}_{\text{GC-loss}}(\mathbf{e}_2) - \hat{J}_{\text{GC-loss}}(\mathbf{e}_1)}{\nabla\hat{J}_{\text{GC-loss}}(\mathbf{e}_1)^T \nabla\hat{J}_{\text{GC-loss}}(\mathbf{e}_1)} \nabla\hat{J}_{\text{GC-loss}}(\mathbf{e}_1) & \text{if } \mathbf{e} \neq \mathbf{0} \\ \mathbf{0} & \text{if } \mathbf{e} = \mathbf{0} \end{cases} \tag{50}$$

and $d(i)$ are both zero-mean Gaussian processes, then error $e(i)$ is also a zero-mean Gaussian process, with PDF

$$p_{e(i)}(e) = \frac{1}{\sqrt{2\pi}\sigma_{e(i)}} \exp\left(-\frac{e^2}{2\sigma_{e(i)}^2}\right) \tag{53}$$

where $\sigma_{e(i)}^2 = \mathbf{E}[e^2(i)]$ is the error variance. Hence

$$\int_{-\varepsilon}^{\varepsilon} p_{e(i)}(e)de = erf\left(\frac{\varepsilon}{\sqrt{2}\sigma_{e(i)}}\right) \tag{54}$$

with $erf(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} \exp(-t^2)dt$ being the *error function*, which is a monotonically increasing function of x. Therefore, $\int_{-\varepsilon}^{\varepsilon} p_{e(i)}(e)de$ is a monotonically decreasing function of $\sigma_{e(i)}^2$. It follows that

$$\max \mathbf{E}\left[G_{\alpha,\beta}\left(e(i)\right)\right] \Leftrightarrow \max \int_{-\varepsilon}^{\varepsilon} p_{e(i)}(e)de \Leftrightarrow \min \sigma_{e(i)}^2 \tag{55}$$

That is, the maximization of the generalized correntropy $\mathbf{E}[G_{\alpha,\beta}(e(i))]$ will be equivalent to the minimization of the mean square error $\sigma_{e(i)}^2 = \mathbf{E}[e^2(i)]$. In this case, the optimal solution under GMCC will be equal to the Wiener solution.

REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1996.

[2] Sayed and H. Ali, "Fundamentals of adaptive filtering," *IEEE Control Syst.*, vol. 25, no. 4, pp. 77–79, 2003.

[3] B. Chen, Y. Zhu, J. Hu, and J. C. Principe, *System Parameter Identification: Information Criteria and Algorithms*. New York, NY, USA: Newnes, 2013.

[4] E. Walach and B. Widrow, "The least mean fourth (LMF) adaptive algorithm and its family," *IEEE Trans. Inf. Theory*, vol. 30, no. 2, pp. 275–283, 1984.

[5] S.-C. Pei and C.-C. Tseng, "Least mean p-power error criterion for adaptive fir filter," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 9, pp. 1540–1547, 1994.

[6] V. J. Mathews and S. H. Cho, "Improved convergence analysis of stochastic gradient adaptive filters using the sign algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 4, pp. 450–454, 1987.

[7] T. Shao, Y. R. Zheng, and J. Benesty, "An affine projection sign algorithm robust against impulsive interferences," *IEEE Signal Process. Lett.*, vol. 17, no. 4, pp. 327–330, 2010.

[8] M. Shao and C. L. Nikias, "Signal processing with fractional lower order moments: Stable processes and their applications," *Proc. IEEE*, vol. 81, no. 7, pp. 986–1010, 1993.

[9] J. Chambers *et al.*, "Least mean mixed-norm adaptive filtering," *Electron. Lett.*, vol. 30, no. 19, pp. 1574–1575, 1994.

[10] J. Chambers and A. Avlonitis, "A robust mixed-norm adaptive filter algorithm," *IEEE Signal Process. Lett.*, vol. 4, no. 2, pp. 46–48, 1997.

[11] Y. Zou, S.-C. Chan, and T.-S. Ng, "Least mean m-estimate algorithms for robust adaptive filtering in impulse noise," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 47, no. 12, pp. 1564–1569, 2000.

[12] S.-C. Chan and Y.-X. Zou, "A recursive least m-estimate algorithm for robust adaptive filtering in impulsive noise: Fast algorithm and convergence performance analysis," *IEEE Trans. Signal Process.*, vol. 52, no. 4, pp. 975–991, 2004.

[13] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. New York, NY, USA: Springer Science & Business Media, 2010.

[14] D. Erdogmus and J. C. Principe, "From linear adaptive filtering to nonlinear information processing—The design and analysis of information processing systems," *IEEE Signal Process. Mag.*, vol. 23, no. 6, pp. 14–33, 2006.

[15] D. Erdogmus and J. C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1780–1786, 2002.

[16] B. Chen, J. Hu, L. Pu, and Z. Sun, "Stochastic gradient algorithm under (h, φ)-entropy criterion," *Circuits, Syst., Signal Process.*, vol. 26, no. 6, pp. 941–960, 2007.

[17] B. Chen, Y. Zhu, and J. Hu, "Mean-square convergence analysis of Adaline training with minimum error entropy criterion," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1168–1179, 2010.

[18] B. Chen, P. Zhu, and J. C. Príncipe, "Survival information potential: A new criterion for adaptive system training," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1184–1194, 2012.

[19] W. Liu, P. P. Pokharel, and J. Príncipe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, 2007.

[20] A. Singh and J. C. Principe, "Using correntropy as a cost function in linear adaptive filters," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2009, pp. 2950–2955.

[21] S. Zhao, B. Chen, and J. C. Principe, "Kernel adaptive filtering with maximum correntropy criterion," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2011, pp. 2012–2017.

[22] A. Singh and J. C. Principe, "A loss function for classification based on a robust similarity metric," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2010, pp. 1–6.

[23] B. Chen, L. Xing, J. Liang, N. Zheng, and J. C. Principe, "Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion," *IEEE Signal Process. Lett.*, vol. 21, no. 7, pp. 880–884, 2014.

[24] L. Shi and Y. Lin, "Convex combination of adaptive filters under the maximum correntropy criterion in impulsive interference," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1385–1388, 2014.

[25] S. Seth and J. Príncipe, "Compressed signal reconstruction using the correntropy induced metric," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 3845–3848.

[26] M. K. Varanasi and B. Aazhang, "Parametric generalized Gaussian density estimation," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1404–1415, 1989.

[27] B. Chen, J. C. Principe, J. Hu, and Y. Zhu, "Stochastic information gradient algorithm with generalized Gaussian distribution model," *J. Circuits, Syst., Comput.*, vol. 21, no. 01, p. 1250006, 2012.

[28] B. Chen and J. C. Príncipe, "Maximum correntropy estimation is a smoothed map estimation," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 491–494, 2012.

[29] V. N. Vapnik and V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998, vol. 1.

[30] M. N. Syed, J. C. Principe, and P. M. Pardalos, "Correntropy in data classification," in *Dynamics of Information Systems: Mathematical Foundations*. New York, NY, USA: Springer, 2012, pp. 81–117.

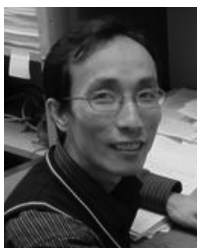
[31] M. N. Syed, P. M. Pardalos, and J. C. Principe, "On the optimization properties of the correntropic loss function in data analysis," *Optim. Lett.*, vol. 8, no. 3, pp. 823–839, 2014.

[32] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, 2003.

[33] T. Y. Al, "Adaptive filters with error nonlinearities: Mean-square analysis and optimum design," *EURASIP J. Adv. Signal Process.*, vol. 2001, no. 4, pp. 192–205, 2001.

[34] V. H. Nascimento and J. C. M. Bermudez, "Probability of divergence for the least-mean fourth algorithm," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1376–1385, 2006.

[35] T. Y. Al-Naffouri and A. H. Sayed, "Transient analysis of adaptive filters with error nonlinearities," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 653–663, 2003.



Badong Chen (M'10–SM'13) received the B.S. and M.S. degrees in control theory and engineering from Chongqing University, in 1997 and 2003, respectively, and the Ph.D. degree in computer science and technology from Tsinghua University in 2008. He was a Post-Doctoral Researcher with Tsinghua University from 2008 to 2010, and a Post-Doctoral Associate at the University of Florida Computational NeuroEngineering Laboratory (CNEL) during the period October, 2010 to September, 2012. He visited the Nanyang Technological University (NTU) as a visiting research scientist during July to August 2015. He is currently a professor at the Institute of Artificial Intelligence and Robotics (IAIR), Xi'an Jiaotong University. His research interests are in signal processing, information theory, machine learning, and their applications in cognitive science and the engineering. He has published 2 books, 3 chapters, and over 100 papers in various journals and conference proceedings. Dr. Chen is associate editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and *Journal of The Franklin Institute*, and has been on the editorial board of *Entropy*.

Lei Xing is a Ph.D. student with the school of electronic and information engineering, Xi'an Jiaotong University, Xi'an, China. His current research focuses mainly on the information theoretic learning (ITL), in particular, robust machine learning under maximum correntropy criterion (MCC).

Haiquan Zhao (M'11) was born in Henan Province, China, in 1974. He received the B.S. degree in applied mathematics in 1998, the M.S. degree and the Ph.D. degree in signal and information processing all at Southwest Jiaotong University, Chengdu, China, in 2005 and 2011, respectively. Since August 2012, he was a Professor with the School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China. From 2015 to 2016, as a visiting scholar, he worked at University of Florida, USA. His current research interests include adaptive filtering algorithm, adaptive Volterra filter, nonlinear active noise control, nonlinear system identification and chaotic signal processing. At present, he is the author or coauthor of more than 70 journal papers and the owner of 20 invention patents. Prof. Zhao has served as an active reviewer for several IEEE Transactions, IET and other international journals.

Nanning Zheng (SM'93–F'06) graduated from the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China, in 1975, and received the M.S. degree in information and control engineering from Xi'an Jiaotong University in 1981 and the Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1985. He is currently a professor and director of the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, pattern recognition and image processing, and hardware implementation of intelligent systems. Prof. Zheng became a member of the Chinese Academy of Engineering in 1999, and he is the Chinese Representative on the Governing Board of the International Association for Pattern Recognition. He serves as an executive deputy editor of the Chinese Science Bulletin.

José C. Prıncipe (M'83–SM'90–F'00) is currently the Distinguished Professor of Electrical and Biomedical Engineering at the University of Florida, Gainesville, where he teaches advanced signal processing and artificial neural networks (ANNs) modeling. He is BellSouth Professor and Founder and Director of the University of Florida Computational Neuro-Engineering Laboratory (CNEL). He is involved in biomedical signal processing, in particular, the electroencephalogram (EEG) and the modeling and applications of adaptive systems. Dr. Príncipe is the past Editor-in-Chief of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, past President of the International Neural Network Society, and former Secretary of the Technical Committee on Neural Networks of the IEEE Signal Processing Society. He is an AIMBE Fellow and a recipient of the IEEE Engineering in Medicine and Biology Society Career Service Award. He is also a former member of the Scientific Board of the Food and Drug Administration, and a member of the Advisory Board of the McKnight Brain Institute at the University of Florida. Since 2014, he has also been a Country "Thousand Talents Program" Distinguished Expert with Xi'an Jiaotong University, Xi'an, China.