Entropy **2015**, *17*, 3419-3437; doi:10.3390/e17053419

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Minimum Error Entropy Algorithms with Sparsity Penalty Constraints

Zongze Wu [1], Siyuan Peng [1], Wentao Ma [2], Badong Chen [2],* and Jose C. Principe [2,3]

[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China; E-Mails: zzwu@scut.edu.cn (Z.W.); pengsiyuan9@gmail.com (S.P.)

[2] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China; E-Mail: xjtu.wentaoma@gmail.com

[3] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA; E-Mail: principe@cnel.ufl.edu

* Author to whom correspondence should be addressed; E-Mail: chenbd@mail.xjtu.edu.cn; Tel.: +86-29-82668802 ext. 8009; Fax: +86-29-82668672.

Academic Editor: Kevin H. Knuth

Received: 30 January 2015 / Accepted: 5 May 2015 / Published: 18 May 2015

Abstract: Recently, sparse adaptive learning algorithms have been developed to exploit system sparsity as well as to mitigate various noise disturbances in many applications. In particular, in sparse channel estimation, the parameter vector with sparsity characteristic can be well estimated from noisy measurements through a sparse adaptive filter. In previous studies, most works use the mean square error (MSE) based cost to develop sparse filters, which is rational under the assumption of Gaussian distributions. However, Gaussian assumption does not always hold in real-world environments. To address this issue, we incorporate in this work an l_1-norm or a reweighted l_1-norm into the minimum error entropy (MEE) criterion to develop new sparse adaptive filters, which may perform much better than the MSE based methods, especially in heavy-tailed non-Gaussian situations, since the error entropy can capture higher-order statistics of the errors. In addition, a new approximator of l_0-norm, based on the correntropy induced metric (CIM), is also used as a sparsity penalty term (SPT). We analyze the mean square convergence of the proposed new sparse adaptive filters. An energy conservation relation is derived and a sufficient condition is obtained, which ensures the mean square convergence. Simulation results confirm the superior performance of the new algorithms.

Keywords: sparse estimation; minimum error entropy; correntropy induced metric; mean square convergence; impulsive noise

MSC Codes: 62B10

1. Introduction

In recent years, sparsity aware learning methods have received a lot of attention due to their broad applicability. In sparse channel estimation, the goal is usually to estimate a parameter vector of an unknown channel with most zero tap under noise disturbances. So far many sparsity aware adaptive filtering algorithms have been developed to solve the problem of sparse channel estimation. In general, a sparse adaptive filtering algorithm can be derived by incorporating a sparsity penalty term (SPT), such as the l_0-norm, into a traditional adaptive algorithm. Typical examples of sparse adaptive filtering algorithms include sparse least mean square (LMS) [1–4], sparse affine projection algorithms (APA) [5], sparse recursive least squares (RLS) [6], and their variations [7–12].

However, there are some limitations of the existing sparse adaptive filters. Specifically, when data are non-Gaussian (especially when data are disturbed by impulsive noise or containing large outliers), they may perform very poorly. The main reason for this is that most of the existing algorithms are developed based on the well-known mean square error (MSE) criterion, which relies heavily on the assumption of Gaussian distributions. This assumption does not always hold, particularly in most practical applications. For instance, different types of artificial noise in electronic devices, atmospheric noises, and lighting spikes in natural phenomena, can be described more accurately using non-Gaussian noise models [13,14]. When sparse filters are applied in such situations, the performance will become much worse due to the sensitivity to the impulsive noises or outliers [15].

Information theoretic learning (ITL), on the other hand, provides a nice approach for dealing with non-Gaussian signal processing [16,17]. The minimum error entropy (MEE) [18–27] criterion in ITL was successfully used in adaptive filtering to improve the learning performance in non-Gaussian noises. Basically, the MEE aims at minimizing the entropy of the training error such that the adaptive model becomes as close as possible to the unknown system. Since the MEE can capture higher-order statistics and information content of signals rather than simply their energy, it is particularly useful for non-Gaussian machine learning and signal processing. In this work, we will use the MEE instead of the MSE to develop sparse adaptive filtering algorithms. The new adaptive filters are very robust to impulsive noises.

As an important part, the SPT in sparse adaptive filters enables them to fit well the sparse structures of the unknown systems. Finding the sparsest solution leads to the l_0-norm minimization, which is an NP-hard problem. In existing methods, the l_1-norm and reweighted l_1-norm are frequently used as the SPT. As a nice approximator of the l_0-norm, the Correntropy Induced Metric (CIM) can also be used as a sparsity penalty term in sparse channel estimation [28,29]. In the present paper, we will incorporate the above-mentioned SPTs into the sparsity aware MEE algorithms, and develop three sparse MEE algorithms, namely the sparse MEE with zero-attracting (l_1-norm) penalty term (ZAMEE), sparse

MEE with the logarithmic (reweighted l_1-norm [30]) penalty term (RZAMEE), and sparse MEE with the CIM penalty term (CIMMEE).

The organization of the rest of the paper is as follows. In Section 2, we briefly introduce the MEE criterion and the CIM. In Section 3, we derive the ZAMEE, RZAMEE and CIMMEE algorithms. In Section 4, we establish an energy conservation relation and derive a sufficient condition that ensures the mean square convergence of the sparse MEE algorithms. In Section 5, we present simulation results to demonstrate the performance of the developed methods. Finally in Section 6, we give the conclusion.

2. MEE and CIM

2.1. Minimum Error Entropy Criterion

Figure 1 shows a general scheme of adaptive system training under MEE criterion. As entropy measures the average uncertainty or diversity of a random variable, minimizing the error entropy will make the error distribution more concentrated (usually with higher peaks), and the discrepancy between the unknown system and adaptive model will be minimized. In supervised learning, the error signal is, in general, defined as the difference between the outputs of unknown system and adaptive model.

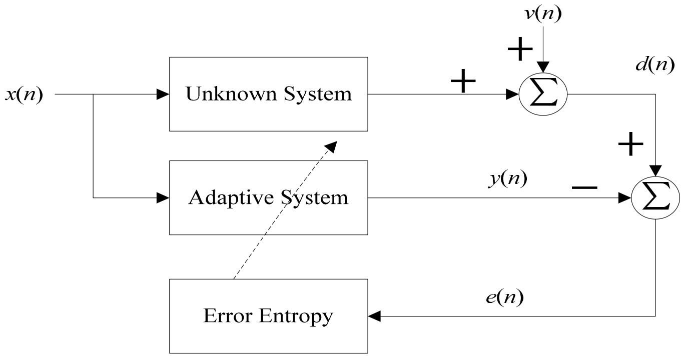

Figure 1. Adaptive system training under minimum error entropy (MEE) criterion.

Consider a linear channel model, where the input vector $X(n)=\left[x_{n-M+1},\cdots,x_{n-1},x_n)\right]^T$ at time n is sent over an FIR channel with parameter vector $W^*=[w_1^*,w_2^*,\cdots,w_M^*]^T$ (M is the size of the channel memory). Assume that the channel parameters are real-valued, and most of them are zero. The desired signal $d(n)$ is then

$$d(n)=W^{*T}X(n)+v(n) \tag{1}$$

where $v(n)$ denotes an interference noise. Let $W(n)=[w_1(n),w_2(n),\cdots,w_M(n)]^T$ be the weight vector of an adaptive filter. The instantaneous error can be calculated as $e(n)=d(n)-W^T(n)X(n)$. Assume that the error $e(n)$ is a random variable with probability density function (PDF) $f_e(e)$. Let $\hat{f}_e(e)$ be an estimator of $f_e(e)$ based on a set of error samples. Then an estimator of Renyi's quadratic entropy for the error signal can be expressed as [16,17]

$$H_{R2}(e)=-\log\int\hat{f}_e^2(\xi)d\xi=-\log V(e) \tag{2}$$

where $V(\mathrm{e})=\int \hat{f}_e^2(\xi)d\xi$ is called the information potential (IP) [16–18]. Based on Parzen window approach, the probability density function of the error takes the following form [16,17]

$$\hat{f}_e(e)=\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(e-e(i)), \tag{3}$$

where N is the samples number, $\kappa_\sigma(\cdot)$ denotes a kernel function with bandwidth σ, and the N error samples are $\{e(1),e(2),\cdots,e(N)\}$. The Gaussian kernel function is one of the most popular kernels, which is given by

$$\kappa_\sigma(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{x^2}{2\sigma^2}). \tag{4}$$

In this work, without mentioned otherwise, the kernel function is a Gaussian kernel. Combining (2) and (3), one can derive

$$\begin{aligned}H_{R2}(e)&=-\log\int \hat{f}_e^2(e)de\\&=-\log\int(\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(e-e(i)))^2de\\&=-\log\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int\kappa_\sigma(e-e(i))\kappa_\sigma(e-e(j))de\\&=-\log\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa_{\sigma\sqrt{2}}(e(i)-e(j)).\end{aligned} \tag{5}$$

It follows easily that

$$V(e)=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa_{\sigma\sqrt{2}}(e(i)-e(j)). \tag{6}$$

Obviously, minimizing the error entropy is equivalent to maximizing the information potential [31,32]. Thus, the optimization criterion for MEE training can be

$$J_{MEE}=\max_{W}V(e). \tag{7}$$

From (7), a steepest ascent algorithm for estimating the weight vector can be derived as

$$W(n+1)=W(n)+\eta\nabla V(e(n)), \tag{8}$$

where η denotes a step size, and $\nabla V(e(n))$ stands for the gradient of the information potential with respect to the weight vector, expressed as

$$\begin{aligned}\nabla V(e(n))&=\frac{\partial V(e(n))}{\partial W(n)}=\frac{\partial}{\partial W(n)}\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa_{\sigma\sqrt{2}}(e(i)-e(j))\right)\\&=\frac{1}{2N^2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\kappa_{\sigma\sqrt{2}}(e(i)-e(j))(e(i)-e(j))\left(\frac{\partial y(i)}{\partial W(n)}-\frac{\partial y(j)}{\partial W(n)}\right)\right],\end{aligned} \tag{9}$$

where $y(i)$ and $y(j)$ denote the outputs of the system at i and j time, respectively.

2.2. Correntropy Induced Metric

Correntropy is a novel nonlinear similarity measure between two random variables, quantifying how similar two random variables in a neighborhood of the joint space [28,33,34]. Given two vectors of samples: $X=\left[x_1,\cdots,x_N\right]^T$, $Y=\left[y_1,\cdots,y_N\right]^T$, a sample mean estimator of the correntropy between X and Y is defined by

$$\hat{\mathrm{V}}(X,Y)=\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_i-y_i). \tag{10}$$

In order to find the sparsest vector (minimum l_0-norm) satisfying a series of linear constrains, one can use the CIM as an approximation of the l_0-norm. Based on the correntropy, the CIM is defined as [28]

$$CIM(\mathrm{X},Y)=\left(\kappa(0)-\hat{V}(X,Y)\right)^{1/2}, \tag{11}$$

which is a metric in sample space and satisfies

(1) Non-negativity: $CIM(X,Y)\geq 0$.
(2) Identity of indiscernible: $CIM(X,Y)=0$ if and only if $X=Y$.
(3) Symmetry: $CIM(X,Y)=CIM(Y,X)$.
(4) Triangle inequality: $CIM(X,Z)\leq CIM(X,Y)+CIM(Y,Z)$.

The CIM provides a nice approximation for the l_0-norm. Given a vector $X=\left[x_1,\cdots,x_N\right]^T$, the l_0-norm of X can be approximated by [28,29]

$$\| X \|_0 \sim CIM^2(\mathrm{X},0)=\frac{\kappa(0)}{N}\sum_{i=1}^{N}(1-\exp(-\frac{x_i^2}{2\sigma^2})). \tag{12}$$

Figure 2 shows the contours of the CIM in a 3-D space, from which one can observe that this metric divides the space in three regions, namely Euclidean region, Transition region and Rectification region. The CIM behaves like an l_2-norm (convex function) in the Euclidean region, like an l_1-norm in the Transition region and like an l_0-norm (non-convex function) in the Rectification region. It can be shown that if $|x_i|>\delta$, $\forall x_i\neq 0$, then as $\sigma\rightarrow 0$, one can get a solution arbitrarily close to that of the l_0-norm, where δ is a small positive number (see [29] for details). Therefore, with a smaller kernel width, the CIM will favor sparsity and can be used as a penalty term in sparse channel estimation.

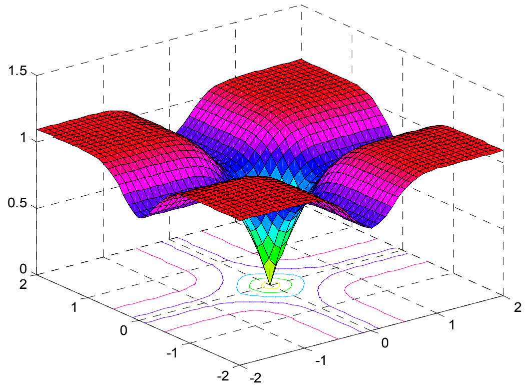

Figure 2. Contours of $CIM(\mathrm{X},0)$ in 3-D space (where kernel size is 0.4).

3. Sparse MEE Algorithms

3.1. Sparse MEE with Zero-Attracting (l_1-norm) Penalty Term (ZAMEE)

To develop a sparse MEE algorithm with zero-attracting (l_1-norm) penalty term [4], we introduce the cost function

$$J_{ZAMEE}(n) = -J_{MEE}(n) + \lambda J_{ZA}(n) = -\frac{1}{L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j)) + \lambda \| \mathrm{W(n)} \|_1, \tag{13}$$

where $J_{ZA}(\mathrm{n}) = \| \mathrm{W(n))} \|_1$ denotes the l_1-norm of the estimated parameter vector, L is the sliding data length, and σ_1 is the kernel width in MEE. In (13), the MEE term is robust to impulsive noises, and the ZA penalty term is a sparsity inducing term, and the two terms are balanced by a weight factor $\lambda \geq 0$.

Based on the cost function (13), one can derive the following adaptive algorithm:

$$\begin{aligned} W(n+1) &= W(n) - \eta\frac{\partial J_{ZAMEE}(n)}{\partial W(n)} \\ &= W(n) - \eta\left[-\frac{1}{2\sigma_1^2L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}\left[[e(i)-e(j)]\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j))\left[\frac{\partial y(i)}{\partial W(\mathrm{n})}-\frac{\partial y(j)}{\partial W(\mathrm{n})}\right]\right] + \lambda sign(\mathrm{W(n)})\right] \\ &= W(n) + \frac{\eta}{2\sigma_1^2L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}\Big[[e(i)-e(j)]\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j))[X(i)-X(j)]\Big] - \rho sign(\mathrm{W(n)}), \end{aligned} \tag{14}$$

where $\rho = \eta\lambda$ is the zero-attractor control factor, and $sign(\cdot)$ is a component-wise sign function [2–4], with $sign(x) = 1$ for $x > 0$, $sign(x) = -1$ for $x < 0$, and $sign(x) = 0$ for $x = 0$. The algorithm (14) is referred to as the ZAMEE algorithm.

3.2. Sparse MEE with the Logarithmic Penalty Term (RZAMEE)

In this part, we derive a sparse MEE algorithm with a logarithmic penalty term [1,2], which can also generate a zero attractor. The corresponding cost function is given by

$$\begin{aligned} J_{RZAMEE}(n) &= -J_{MEE}(n) + \lambda J_{RZA}(n) \\ &= \frac{1}{L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j)) + \lambda\sum_{i=1}^{M}\log(1+|w_i(n)|/\delta), \end{aligned} \tag{15}$$

where the log-sum penalty $\sum_{i=1}^{M}\log(1+|\mathrm{w}_i(n)|/\delta)$ behaves more similarly to the l_0-norm than the l_1-norm $\|\mathrm{W}\|_1$, and δ is a positive number. Then, a gradient-based adaptive algorithm can be easily derived as

$$\begin{aligned} w_i(n+1) &= w_i(n) - \eta\frac{\partial J_{RZAMEE}(n)}{\partial w_i(n)} \\ &= w_i(n) - \eta\left[-\frac{1}{2\sigma_1^2L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}[e(i)-e(j)]\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j))[X(i)-X(j)] + \lambda\frac{\mathrm{sign}(w_i(\mathrm{n}))}{1+\delta'|w_i(\mathrm{n})|}\right] \\ &= w_i(n) + \frac{\eta}{2\sigma_1^2L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}\Big[[e(i)-e(j)]\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j))[X(i)-X(j)]\Big] - \rho\frac{\mathrm{sign}(w_i(\mathrm{n}))}{1+\delta'|w_i(\mathrm{n})|} \end{aligned} \tag{16}$$

where $\delta' = \frac{1}{\delta}$. This algorithm is referred to as the RZAMEE algorithm.

3.3. Sparse MEE with CIM Penalty Term (CIMMEE)

One can also employ the CIM as a sparsity penalty term to develop a sparse MEE algorithm. A new cost function can be defined by

$$\begin{aligned} J_{CIMMEE}(n) &= -J_{MEE}(n) + \lambda J_{CIM}(n) \\ &= -\frac{1}{L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j)) + \lambda\frac{1}{M\sigma_2\sqrt{2\pi}}\sum_{i=1}^{M}(1-\exp(-\frac{w_i(n)^2}{2\sigma_2^{\,2}})), \end{aligned} \tag{17}$$

where σ_2 denotes the kernel width in CIM. The second term (*i.e.*, the CIM) with a smaller kernel width will become a sparsity inducing term. Based on the new cost function of (17), we derive a gradient-based adaptive algorithm:

$$\begin{aligned} W(n+1) &= W(n) - \eta\frac{\partial J_{CIMMEE}(n)}{\partial W(n)} \\ &= W(n) - \eta\left[-\frac{1}{2\sigma_1^2 L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}[e(i)-e(j)]\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j))[X(i)-X(j)] + \lambda\frac{1}{M\sigma_2^3\sqrt{2\pi}}W(n).*\exp(-\frac{W(n).*W(n)}{2\sigma_2^2})\right] \\ &= W(n) + \frac{\eta}{2\sigma_1^2 L^2}\sum_{i=n-L+1}^{n}\sum_{j=n-L+1}^{n}[e(i)-e(j)]\kappa_{\sqrt{2}\sigma_1}(e(i)-e(j))[X(i)-X(j)] - \rho\frac{1}{M\sigma_2^3\sqrt{2\pi}}W(n).*\exp(-\frac{W(n).*W(n)}{2\sigma_2^2}), \end{aligned} \tag{18}$$

where .* denotes element-wise product. The above algorithm is referred to as the CIMMEE algorithm. The kernel width σ_2 is a key parameter in the penalty term. A proper kernel width will make the CIM approximate well the l_0-norm.

The derived sparsity aware MEE algorithms can be written in a unifying form:

$$\begin{aligned} W(n+1) &= W(n) - \eta\frac{\partial J_{MEE}(\vec{e}(n))}{\partial W(n)} - G(W(n)) \\ &= W(n) + \eta\chi^T(n)\vec{h}(\vec{e}(n)) - G(W(n)) \end{aligned} \tag{19}$$

where $\vec{e}(n) = [e(n-L+1), e(n-L+2), \cdots, e(n)]^T$ is an $L\times 1$ error vector, $\chi(n)$ = [$X(n$ – L + 1), $X(n - L + 2)$, …, $X(n)$]$^{\mathrm{T}}$ is an $L\times M$ input matrix, $\vec{h}(\vec{e}(n)) = [h_1(\vec{e}(n)), h_2(\vec{e}(n)), \cdots, h_L(\vec{e}(n))]^T$, in which

$$h_i(\vec{e}(n)) = \frac{\partial J_{MEE}(\vec{e}(n))}{\partial e(n-L+i)} \tag{20}$$

and $G(W(n))$ is the derivative of the sparsity penalty term with respect to $W(n)$, which is an $M\times 1$ vector. For ZAMEE, RZAMEE and CIMMEE, $G(W(n))$ can be expressed respectively as $G(W(n))$ = $\rho sign(W(n))$, $G(W(n)) = \rho\frac{\mathrm{sign}(W(n))}{1+\delta'|W(n)|}$ and $G(W(n)) = \rho\frac{1}{M\sigma_2^{\,3}\sqrt{2\pi}}W(n).*\exp(-\frac{W(n).*W(n)}{2\sigma_2^{\,2}})$.

4. Mean Square Convergence Analysis

Now we analyze the mean square convergence of the algorithm (19). For simplicity and rigorous analysis, we only consider the case in which $G(W(n))=\rho\frac{1}{M\sigma_2^3\sqrt{2\pi}}W(n).*\exp(-\frac{W(n).*W(n)}{2\sigma_2^2})$. First, we derive a fundamental energy conservation relation [24,35,36].

4.1. Energy Conservation Relation

In order to presenting a unifying formulation for the sparsity under MEE criterion, we rewrite $e(n)=d(n)-W^T(n)X(n)$ as follows

$$\vec{e}(n)=\vec{d}(n)-\chi(n)W(n), \tag{21}$$

where $\vec{d}(n)=[d(n-L+1),d(n-L+2),\cdots,d(n)]^T$ is the $L\times 1$ desired signal vector. From (1), we derive

$$\vec{d}(n)=\chi(n)W^*+\vec{v}(n), \tag{22}$$

where $\vec{v}(n)=[v(n-L+1),v(n-L+2),\cdots,v(n)]^T$ is the noise vector. Obviously, combining (21) and (22), the error vector $\vec{e}(n)$ can be expressed as

$$\vec{e}(n)=\chi(n)\tilde{W}(n)+\vec{v}(n), \tag{23}$$

where $\tilde{W}(n)=W^*-W(n)$ is the weight error vector. Let us define the *a priori* error vector $\vec{e}_a(\mathrm{n})$ and *a posteriori* error vector $\vec{e}_p(\mathrm{n})$ as follows:

$$\begin{cases}\vec{e}_a(n)=\chi(n)\tilde{W}(n)\\ \vec{e}_p(n)=\chi(n)\tilde{W}(n+1)\end{cases} \tag{24}$$

In addition, $\vec{e}_a(\mathrm{n})$ and $\vec{e}_p(\mathrm{n})$ have the following relationship:

$$\begin{aligned}\vec{e}_p(n)&=\vec{e}_a(n)+\chi(n)(\tilde{W}(n+1)-\tilde{W}(n))\\ &=\vec{e}_a(n)-\chi(n)(W(n+1)-W(n)).\end{aligned} \tag{25}$$

To simplify the analysis, here we assume $L=M$. Then, combining (19) and (25), we have

$$\begin{aligned}&\vec{e}_p(n)=\vec{e}_a(n)-\chi(n)(\eta\chi^T(n)\vec{h}(\vec{e}(n))-G(W(n)))\\ &\Rightarrow\vec{e}_p(n)-\vec{e}_a(n)=-(\eta\Re(n)\vec{h}(\vec{e}(n))-\chi(n)G(W(n)))\\ &\Rightarrow\Re^{-1}(n)(\vec{e}_p(n)-\vec{e}_a(n))=-(\eta\vec{h}(\vec{e}(n))+\Re^{-1}(n)\chi(n)G(W(n)))\\ &\Rightarrow\chi^T(n)\Re^{-1}(n)(\vec{e}_p(n)-\vec{e}_a(n))=-(W(n+1)-W(n))\\ &\Rightarrow\chi^T(n)\Re^{-1}(n)(\vec{e}_p(n)-\vec{e}_a(n))=\tilde{W}(n+1)-\tilde{W}(n),\end{aligned} \tag{26}$$

where $\Re(n)=\chi(n)\chi^T(n)$ is an $L\times L$ -dimensional symmetric matrix and is assumed to be invertible. Therefore, we have

$$\tilde{W}(n+1)=\tilde{W}(n)+\chi^T(n)\Re^{-1}(n)(\vec{e}_p(n)-\vec{e}_a(n)). \tag{27}$$

Squaring both sides of (27), we obtain

$$\tilde{W}^T(n+1)\tilde{W}(n+1)=[\tilde{W}(n)+\chi^T(n)\Re^{-1}(n)(\vec{e}_p(n)-\vec{e}_a(n))]^T\times[\tilde{W}(n)+\chi^T(n)\Re^{-1}(n)(\vec{e}_p(n)-\vec{e}_a(n))]. \tag{28}$$

After some simple manipulations, we derive

$$\left\|\tilde{W}(n+1)\right\|^2+\left\|\vec{e}_a(n)\right\|^2_{\Re^{-1}(n)}=\left\|\tilde{W}(n)\right\|^2+\left\|\vec{e}_p(n)\right\|^2_{\Re^{-1}(n)}, \tag{29}$$

where $\left\|\tilde{W}(n)\right\|^2=\tilde{W}^T(n)\tilde{W}(n)$, $\left\|\vec{e}_a(n)\right\|^2_{\Re^{-1}(n)}=\vec{e}_a^{\,T}(n)\Re^{-1}(n)\vec{e}_a(n)$ and $\left\|\vec{e}_p(n)\right\|^2_{\Re^{-1}(n)}=\vec{e}_p^{\,T}(n)\Re^{-1}(n)\vec{e}_p(n)$. Taking the expectations of the both sides of (29), we have

$$E\left[\left\|\tilde{W}(n+1)\right\|^2\right]+E\left[\left\|\vec{e}_a(n)\right\|^2_{\Re^{-1}(n)}\right]=E\left[\left\|\tilde{W}(n)\right\|^2\right]+E\left[\left\|\vec{e}_p(n)\right\|^2_{\Re^{-1}(n)}\right] \tag{30}$$

where $E[\bullet]$ denotes the expectation operator, $E\left[\left\|\tilde{W}(n)\right\|^2\right]$ is the weight error power (WEP) at iteration n.

Remark: *Equation (30) is referred to as the energy conservation relation for the proposed sparsity aware MEE algorithms, which is, interestingly, the same as the energy conservation relation derived in* [24]. *In fact, the sparsity penalty terms have no influence on the energy conservation relation. Similar extensions of the energy conservation relation to multi-dimensional error can be found in* [37,38].

4.2. Sufficient Condition for Mean Square Convergence

Based on the energy conservation relation (30), a sufficient condition can be derived that ensures the mean square convergence. By substituting $\vec{e}_p(n)=\vec{e}_a(n)-(\eta\Re(n)\vec{h}(\vec{e}(n))-\chi(n)G(W(n)))$ into (30), we obtain

$$\begin{aligned}E\left[\left\|\tilde{W}(n+1)\right\|^2\right]=&E\left[\left\|\tilde{W}(n)\right\|^2\right]-2\eta E\left[\vec{e}_a^{\,T}(\mathrm{n})\vec{h}(\vec{e}(n))\right]+\eta^2E\left[\vec{h}^T(\vec{e}(n))\Re(n)\vec{h}(\vec{e}(n))\right]\\&+E\left[G^T(W(n)G(W(n))\right]-2E\left[\vec{e}_a^{\,T}(\mathrm{n})\Re^{-1}(n)\chi(n)G(W(n))\right]\\&+2\eta E\left[\vec{h}^T(\vec{e}(n))\chi(n)G(W(n))\right]\end{aligned} \tag{31}$$

Before evaluating the expectations $E\left[G^T(W(n))G(W(n))\right]$, $E\left[\vec{e}_a^{\,T}(\mathrm{n})\Re^{-1}(n)\chi(n)G(W(n))\right]$, $E\left[\vec{e}_a^{\,T}(\mathrm{n})\vec{h}(\vec{e}(n))\right]$, $E\left[\vec{h}^T(\vec{e}(n))\chi(n)G(W(n))\right]$, and $E\left[\vec{h}^T(\vec{e}(n))\Re(n)\vec{h}(\vec{e}(n))\right]$, we give the following assumptions.

Assumptions:

(A) The noise $\{v(n)\}$ is independent, identically distributed, and independent of the input $\{X(n)\}$.

(B) The *a priori* error vector $\vec{e}_a(n)$ is jointly Gaussian distributed.

(C) The input vectors $\{X(n)\}$ are zero-mean independent, identically distributed.

(D) $\forall i,j\in\{n-L+1,\cdots,n\}$, $\Re_{i,j}(n)$ is independent of $\{e(i),e(j)\}$.

(E) The vectors $\{G(W(n))\}$ are zero-mean independent, identically distributed, and independent of the input $\{X(n)\}$.

Remark: *The assumptions (A), (B), (C) and (D) are commonly used in the literature* [35,36]. *In this work, the unknown system is assumed to be a sparse system, of which most coefficients are zero or close to zero, so the weight vector $W(n)$ of the adaptive filter is also sparse, especially at the final stage of convergence when the filter gets very close to the unknown system. Since $W(n)$ is sparse, the* *vector* $G(W(n)) = \rho \frac{1}{M\sigma_2{}^3\sqrt{2\pi}} W(n).*\exp(-\frac{W(n).*W(n)}{2\sigma_2{}^2})$ *will be close to a null vector, because we have* $\phi(x) = \rho \frac{1}{M\sigma_2{}^3\sqrt{2\pi}} x\exp(-\frac{x^2}{2\sigma_2{}^2}) \approx 0$ *when x is very close to or far from zero. Thus, for the CIMMEE, the assumption (E) is reasonable.*

If the above assumptions hold, in a similar way to [24,35], one can derive

$$E\left[\vec{e}_a^T(\mathrm{n})\vec{h}(\vec{e}(n))\right] = \gamma^2(n)\theta_G(\gamma^2(n)) \tag{32}$$

$$E\left[\vec{h}^T(\vec{e}(n))\Re(n)\vec{h}(\vec{e}(n))\right] = \theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right] \tag{33}$$

where $\gamma^2(n)=E[(e_a(n-L+i))^2]$, $\theta_G(\gamma^2(n))$ and $\theta_I(\gamma^2(n))$ denote two functions of $\gamma^2(n)$. The subscript G in θ_G points to the fact that the Gaussian assumption (B) is the main assumption for the Equation (32) and the subscript I in θ_I indicates that the independence assumption (D) is the major assumption leading to the expression (33). For more details about (32) and (33), interested readers are referred to [24]. By assumption (E), it follows easily that

$$\begin{cases} E\left[\vec{e}_a^T(\mathrm{n})\Re^{-1}(n)\chi(n)G(W(n))\right] = E\left[\vec{e}_a^T(\mathrm{n})\Re^{-1}(n)\chi(n)\right]E\left[G(W(n))\right] = 0 \\ E\left[\vec{h}^T(\vec{e}(n))\chi(n)G(W(n))\right] = E\left[\vec{h}^T(\vec{e}(n))\chi(n)\right]E\left[G(W(n))\right] = 0 \end{cases} \tag{34}$$

Let the variance of $\{G(W(n))\}$ be ς^2. Then we derive

$$E\left[G^T(W(n))G(W(n))\right] = \varsigma^2 \tag{35}$$

Substituting (32), (33), (34) and (35) into (31), we obtain

$$E\left[\|\tilde{W}(n+1)\|^2\right] = E\left[\|\tilde{W}(n)\|^2\right] - 2\eta\gamma^2(n)\theta_G(\gamma^2(n)) + \eta^2\theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right] + \varsigma^2 \tag{36}$$

From (36), we observe

$$E\left[\|\tilde{W}(n+1)\|^2\right] \le E\left[\|\tilde{W}(n)\|^2\right] \Leftrightarrow \eta^2\theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right] - 2\eta\gamma^2(n)\theta_G(\gamma^2(n)) + \varsigma^2 \le 0 \tag{37}$$

Since $\Re(n)$ is assumed to be invertible, we have

$$\theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right] > 0 \tag{38}$$

Thus, to make the weight error power monotonically decreased (hence converge), the step size η should satisfy the following inequality:

$$\frac{\gamma^2(n)\theta_G(\gamma^2(n)) - \sqrt{\Upsilon}}{\theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right]} \le \eta \le \frac{\gamma^2(n)\theta_G(\gamma^2(n)) + \sqrt{\Upsilon}}{\theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right]} \tag{39}$$

where $\Upsilon = (\gamma^2(n)\theta_G(\gamma^2(n)))^2 - \theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right]\varsigma^2$. As $\eta > 0$, the above inequality implies

$$\theta_G(\gamma^2(n)) > 0 \tag{40}$$

$$\Upsilon \geq 0 \tag{41}$$

Therefore, a sufficient condition for the mean square convergence will be

$$\begin{cases} \theta_G(\gamma^2(n)) > 0 \\ \dfrac{\gamma^2(n)\theta_G(\gamma^2(n)) - \sqrt{\Upsilon}}{\theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right]} \leq \eta \leq \dfrac{\gamma^2(n)\theta_G(\gamma^2(n)) + \sqrt{\Upsilon}}{\theta_I(\gamma^2(n))E\left[\|X(n)\|^2\right]}, \qquad \forall n \\ \Upsilon \geq 0 \end{cases} \tag{42}$$

Remark: *It is worth noting that the sufficient condition of (42) does not ensure that the WEP will converge to zero. Actually, for a stochastic gradient-based algorithm, there are always misadjustments. Even so, the derived condition will guarantee the monotonic decrease of WEP and ensure that the algorithm does not diverge.*

5. Simulation Results

In this section, we perform simulations on time-varying channel estimation to demonstrate the performance of the proposed sparse aware MEE algorithms (ZAMEE, RZAMEE, and CIMMEE), compared with several other algorithms, including least absolute deviation (LAD) [39], MEE, ZALMS, and RZALMS, in a sparse system identification setting. In all the simulations, the performance measure adopted is the mean square deviation (MSD), defined as

$$\mathrm{MSD}{=}E\left[\left\|W^* - W(n)\right\|^2\right] \tag{43}$$

5.1. Experiment 1

In the first experiment, in order to identify the sparsity of the system, we use a filter of 20 coefficients in the time varying system. The parameter vector of the unknown channel is assumed to be

$$W^* = \begin{cases} [0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0] & n \leq 2000 \\ [1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0] & 2000 < n \leq 3000 \\ [1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1] & 3000 < n \end{cases} \tag{44}$$

In (44), the channel memory size, *M*, is 20. The channel model has a sparsity of 1/20 during 1 to 2000 iterations, while the sparsity changes to 1/4 when the iteration is from 2000 to 3000, and it is non-sparsity after 3000 iterations. The input signal $\{x(n)\}$ is a white Gaussian random sequence with zero mean and unit variance. Simulation results below are obtained by averaging over 100 independent Monte Carlo runs, and each run has 5000 iterations.

We employ the alpha-stable distribution [40] as impulsive noise model, which has been widely applied in the literature [41–43]. The characteristic function of the alpha-stable distribution is given by

$$f(t) = \exp\{\mathrm{j}\delta t - \gamma|t|^{\alpha}[1 + \mathrm{j}\beta \operatorname{sgn}(t) \mathrm{S}(t,\alpha)]\} \tag{45}$$

in which

$$S(t,\alpha)=\begin{cases}\tan\dfrac{\alpha\pi}{2} & if\,\alpha\neq 1\\ \dfrac{2}{\pi}\log|t| & if\,\alpha=1\end{cases} \tag{46}$$

where $\alpha\in(0,2]$ is the characteristic factor, $-\infty<\delta<+\infty$ is the location parameter, $\beta\in[-1,1]$ is the symmetry parameter, and $\gamma>0$ is the dispersion parameter. Such a distribution is called a symmetric alpha-stable ($S\alpha S$) distribution when $\beta=0$. We define the parameters vector as $V=(\alpha,\beta,\gamma,\delta)$.

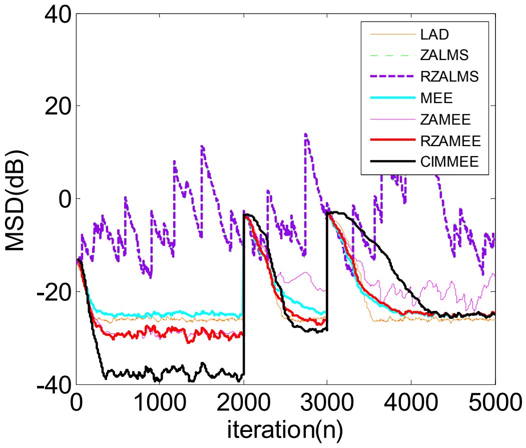

Figure 3. Tracking and steady-state behaviors of 20-order adaptive filters.

First, we investigate the convergence behaviors of the proposed methods in impulsive alpha-stable noise, where the noise parameters vector is $V=(1.2,0,0.2,0)$. The sliding data length for MEE is L = 20. The step size is set at 0.03 for all algorithms. The kernel widths in MEE and CIM are 2.0 and 0.04, respectively. For all sparse aware algorithms, ρ is set at 0.0001. The parameter δ' for RZALMS and RZAMEE is 10. The average convergence curves in terms of the MSD are shown in Figure 3. As one can see from the MSD results, when the channel system is very sparse (before the 2000th iteration), the sparse aware MEE achieve faster convergence rate and better steady-state performance than the other robust algorithms (LAD, MEE), while ZALMS and RZALMS work poorly, as they are sensitive to the impulsive noises. Thus, we only consider the MEE, LAD algorithms comparing with the proposed algorithm in the next experiment case. In addition, CIMMEE achieves lower MSD than ZAMEE and RZAMEE, since the CIM provides a nice approximation for the l_0-norm. After the 2000th iteration, as the number of non-zero taps increases to ten, the performance of the ZAMEE and RZAMEE deteriorates while the CIMMEE maintains the best performance among all the sparse aware filters. After 3000 iterations, the sparse aware MEE algorithms still perform comparable with the MEE, even though the system is now completely non-sparse.

Second, we conduct the simulation with different γ (0.2, 0.4, 0.6, 0.8, 1) and α (1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7) to further demonstrate the performance of the proposed method. In this simulation, we mainly focus on the fully sparse channel system in the first stage of the proposed model. The step size is set at 0.02 for all algorithms, and other parameter settings are the same as in the previous simulation for all algorithms. The MSD, *versus* different γ and α, are illustrated in Figures 5 and 6, respectively. Evidently, the sparse aware MEE algorithms perform well with the different parameter of the impulsive noise model. Moreover, we see that the CIMMEE achieves much lower MSDs in all the cases. Simulation results confirm that the proposed sparse aware MEE algorithms, especially CIMMEE, can efficiently estimate a sparse channel in impulsive noise environment.

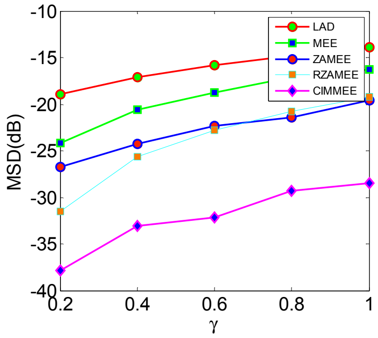

Figure 4. Steady-state mean square deviation (MSD) *versus* different values of γ .

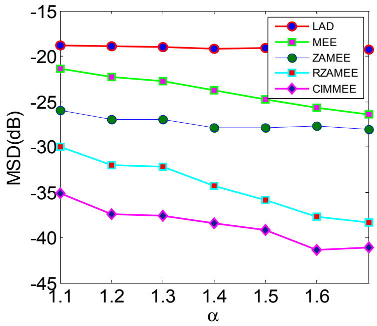

Figure 5. Steady-state mean square deviation (MSD) *versus* different values of α.

Third, we perform simulations to investigate how the kernel width σ_1 affects the performance, which is an important parameter for the sparse aware MEE. Here, the steady-state MSDs of the CIMMEE with different σ_1 (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5) and different α (1, 1.2, 1.4, 1.6, 1.8, and 2) are computed. Other parameters are set as: $\gamma = 1$, $\eta = 0.01$, $\rho = 0.0001$, $\sigma_2 = 0.04$ and $\delta' = 10$. The results are given in Figure 6. One can see that the CIMMEE achieves different MSDs with different σ_1 and under different noise distributions. In this example, the lowest MSD will be obtained around $\sigma_1 = 1.5$. From the simulation results, we may conclude that the kernel width in MEE has a significant influence on the performance.

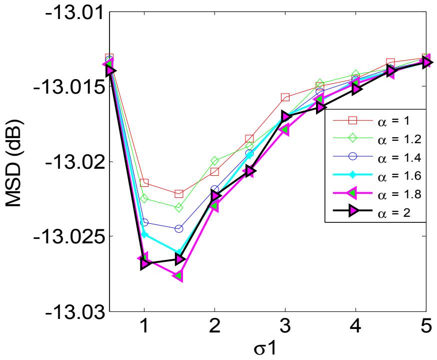

Figure 6. Steady-state mean square deviation (MSD) of sparse minimum error entropy (MEE) with the correntropy induced metric (CIM) penalty term (CIMMEE) with different kernel size σ_1 for different α .

5.2. Experiment 2

In the second experiment, the system is the same as the first experiment, except for the switching times. The parameter vector of the unknown channel is assumed to be

$$W^* = \begin{cases} [0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0] & n \le 5000 \\ [1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0] & 5000 < n \le 10000 \\ [1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1,1,-1] & 10000 < n \end{cases} \tag{47}$$

and the channel memory size, *M*, is 20. The input signal $\{x(n)\}$ is now a correlated signal generated by the process $(x(n) = 0.8x(n-1) + v(n))$ and then normalized to variance 1, where $v(n)$ is a white Gaussian process. The observed noise is the same noise assumed in the first experiment with the same parameters. All simulation results are obtained by averaging over 100 independent Monte Carlo runs, and each run performs 15,000 iterations. The sliding data length is *L* = 20. The step size is set at 0.04 for all algorithms. The kernel widths in MEE and CIM are 3.0 and 0.05, respectively. For all sparse MEE algorithms, ρ is set at 0.0001. The parameter δ' for RZAMEE is 10. Figure 7 shows the average MSD estimate of the three sparse MEE filters. As seen from the MSD results, similar performance trends are observed as in the first experiment. When the system is very sparse, the CIMMEE achieves better steady-state performance than ZAMEE and RZAMEE. As the number of non-zero taps increases to 10, even 20 (completely non-sparse), the CIMMEE algorithms still performs better than the other sparse MEE filters because the CIM has a nice approximation for the l_0-norm.

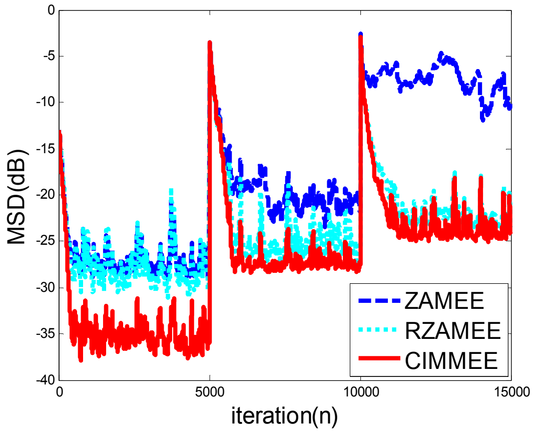

Figure 7. Tracking and steady-state behaviors of 20-order adaptive filters with correlated input.

Second, we perform simulations to investigate how the kernel width σ_1 and the characteristic factor α affect the performance, which are important parameters for the sparse aware MEE. Here, the steady-state MSDs of the CIMMEE with different σ_1 (1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5) and different α (1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, and 1.9) are computed. Other filter parameters are set as: $\gamma = 1$, $\eta = 0.01$, $\rho = 0.0001$, $\sigma_2 = 0.05$ and $\delta' = 10$. Figure 8 shows the simulation result in 3-D space. As one can observe clearly, the best performance of the CIMMEE can be obtained at about $\sigma_1 = 3$. If σ_1 is too small or too large, the convergence performance will become worse. However, the MSD is little affected by the characteristic factor α. This implies that the MEE is an extremely robust principle in impulsive non-Gaussian noises.

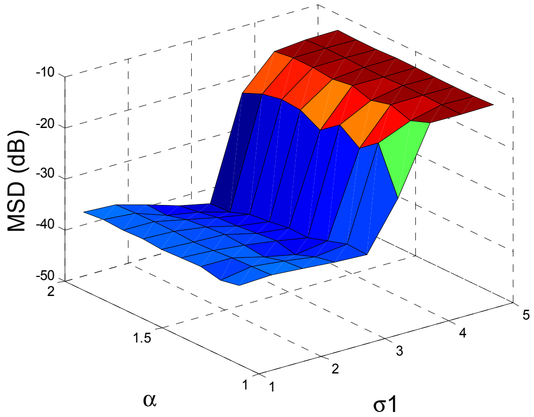

Figure 8. Steady-state mean square deviation (MSD) of sparse minimum error entropy (MEE) with the correntropy induced metric (CIM) penalty term (CIMMEE) with different kernel size σ_1 and different α in 3-D space.

5.3. Experiment 3

In the third experiment, we demonstrate the performance when the input signal is a fragment of 2 s of real speech, sampled at 8kHZ [4,8]. Figure 9 shows an acoustic echo path of a 1024-tap system with 52 non-zero coefficients, which can be considered to be very sparse and is used in the simulation. The output is still disturbed by an alpha-stable noise and the noise parameters vector is $V = (1.4, 0, 0.2, 0)$. All simulation results are obtained by averaging over 100 independent Monte Carlo runs. The sliding data length is L = 20. The other parameters are set as: $\eta = 0.0015$, $\rho = 0.0001$, $\sigma_1 = 1.0$, $\sigma_2 = 0.05$ and $\delta' = 10$. The convergence curves for the sparse MEE algorithms are shown in Figure 10. Compared with the ZAMEE and RZAMEE, the CIMMEE algorithm achieves a smaller MSD.

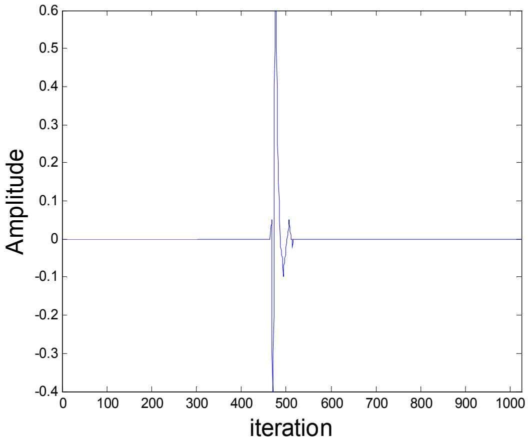

Figure 9. Acoustic echo path with length M = 1024.

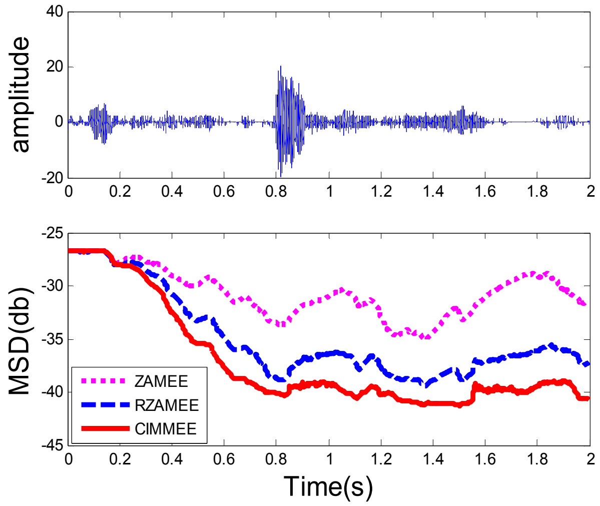

Figure 10. Convergence behaviors with speech signal input. The speech signal is shown in the upper plot.

6. Conclusion

The MEE, as an adaptation criterion, has been successfully applied in many fields because of its desirable performance in non-Gaussian situations. In this work, we develop several sparsity aware MEE algorithms, including ZAMEE, RZAMEE, and CIMMEE, which are derived by incorporating different sparsity penalty terms into the MEE criterion. The mean square convergence properties of the proposed algorithms have been analyzed. Based on an energy conservation relation, we derive a sufficient condition that guarantees the mean square stability. Simulation results show that the new algorithms can achieve excellent performance, especially when the measurements are disturbed by impulsive non-Gaussian noises. How to select proper parameters, such as the kernel bandwidth, is an important issue. This will be an interesting topic for future study.

Acknowledgments

This work was supported by 973 Program (No. 2015CB351703) and National Natural Science Foundation of China (No. 61372152, no. 61271210).

Author Contributions

The contributions of each author are as follows: Zongze Wu and Badong Chen proved the main results and wrote the draft; Siyuan Peng and Wentao Ma carried out the simulations; Jose C. Principe polished the language and was in charge of technical checking. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Chen, Y.; Gu, Y.; Hero, A.O. Sparse LMS for system identification. In Proceedings of 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2009), Taipei, Taiwan, 19–24 April 2009; pp. 3125–3128.
2. Gu, Y.; Jin, J.; Mei, S. l_0 norm constraint LMS algorithm for sparse system identification. *IEEE Signal Process. Lett.* **2009**, *16*, 774–777.
3. Jin, J.; Qu, Q.; Gu, Y. Robust Zero-point Attraction Least Mean Square Algorithm on Near Sparse System Identification. *IET Signal Process.* **2013**, *7*, 210–218.
4. Shi, K.; Shi, P. Convergence analysis of sparse LMS algorithms with l_1-norm penalty based on white input signal. *Signal Process.* **2010**, *90*, 3289–3293.
5. Yin, D.; So, H.C.; Gu, Y. Sparse Constraint Affine Projection Algorithm with Parallel Implementation and Application in Compressive Sensing. In Proceedings of 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2014), Florence, Italy, 4–9 May 2014; pp. 7288–7292.
6. Babadi, B.; Kalouptsidis, N.; Tarokh, V. SPARLS: The sparse RLS algorithm. *IEEE Trans. Signal Process.* **2010**, *58*, 4013–4025.
7. Wu, F.Y.; Tong, F. Gradient optimization p-norm-like constraint LMS algorithm for sparse system estimation. *Signal Process*. **2013**, *93*, 967–971.
8. Salman, M.S. Sparse leaky LMS algorithm for system identification and its convergence analysis. *Int. J. Adapt. Control Signal Process*. **2014**, *28*, 1065–1072.
9. Aliyu, M.L.; Alkassim, M.A.; Salman, M.S. A p-norm variable step-size LMS algorithm for sparse system identification. *Signal Image Video Process*. **2014**, doi: 10.1007/s11760-013-0610-7.
10. Wu, F.Y.; Tong, F. Non-Uniform Norm Constraint LMS Algorithm for Sparse System Identification. *IEEE Commun. Lett.* **2013**, *17*, 385–388.
11. Das, B.K.; Chakraborty, M. Sparse Adaptive Filtering by an Adaptive Convex Combination of the LMS and the ZA-LMS Algorithms. *IEEE Trans. Circuits Syst*. **2014**, *61*, 1499–1507.
12. Liu, Y.; Li, C.; Zhang, Z. Diffusion sparse least-mean squares over networks. *IEEE Trans. Signal Process.* **2012**, *60*, 4480–4485.
13. Plataniotis, K.N.; Androutsos, D.; Venetsanopoulos, A.N. Nonlinear filtering of non-Gaussian noise. *J. Intell. Robot. Syst*. **1997**, *19*, 207–231.
14. Weng, B.; Barner, K.E. Nonlinear system identification in impulsive environments. *IEEE Trans. Signal Process*. **2005**, *53*, 2588–2594.
15. Golub, G.H.; van Loan, C.F. *Matrix Computation*; the Johns Hopkins University Press: Baltimore, MD, USA, 1983.
16. Principe, J.C. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer: New York, NY, USA, 2010.
17. Chen, B.; Zhu, Y.; Hu, J.; Principe, J.C. *System Parameter Identification: Information Criteria and Algorithms*; Elsevier: Amsterdam, The Netherlands, 2013.
18. Erdogmus, D.; Principe, J.C. From linear adaptive filtering to nonlinear information processing. *IEEE Signal Process. Mag*. **2006**, *23*, 15–33.

19. Erdogmus, D.; Principe, J.C. An error-entropy minimization for supervised training of nonlinear adaptive systems. *IEEE Trans. Signal Process.* **2002**, *50*, 1780–1786.
20. Chen, B.; Hu, J.; Pu, L.; Sun, Z. Stochastic gradient algorithm under (h, ϕ)-entropy criterion. *Circuit Syst. Signal Process*. **2007**, *26*, 941–960.
21. Wolsztynski, E.; Thierry, E.; Pronzato, L. Minimum-entropy estimation in semi-parametric models. *Signal Process*. **2005**, *85*, 937–949.
22. Song, A.; Qiu, T. The Equivalency of Minimum Error Entropy Criterion and Minimum Dispersion Criterion for Symmetric Stable Signal Processing. *IEEE Signal Process. Lett*. **2010**, *17*, 32–35.
23. Chen, B.; Principe, J.C. Some further results on the minimum error entropy estimation. *Entropy* **2012**, *14*, 966–977.
24. Chen, B.; Zhu, Y.; Hu, J. Mean-square convergence analysis of ADALINE training with minimum error entropy criterion. *IEEE Trans. Neural Netw*. **2010**, *21*, 1168–1179.
25. Chen, B.; Principe, J.C. On the Smoothed Minimum Error Entropy Criterion. *Entropy* **2012**, *14*, 2311–2323.
26. Li, C.; Shen, P.; Liu, Y.; Zhang, Z. Diffusion information theoretic learning for distributed estimation over network. *IEEE Trans. Signal Process*. **2013**, *61*, 4011–4024.
27. Xue, Y.; Zhu, X. The minimum error entropy based robust wireless channel tracking in impulsive noise. *IEEE Commun. Lett*. **2002**, *6*, 228–230.
28. Liu, W.F.; Pokharel, P.P.; Principe, J.C. Correntropy: Properties and Applications in Non-Gaussian Signal Processing. *IEEE Trans. Signal Process*. **2007**, *55*, 5286–5298.
29. Seth, S.; Principe, J.C. Compressed signal reconstruction using the correntropy induced metric. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2008), Las Vegas, NV, USA, 31 March–4 April 2008; pp. 3845–3848.
30. Wipf, D.P.; Nagarajan, S.S. A new view of automatic relevance determination. *Adv. Neural Inf. Process. Syst*. **2008**, Available online: http://papers.nips.cc/paper/3372-a-new-view-of-automatic-relevance-determination (accessed on 5 May 2015).
31. Chen, B.; Zhu, P.; Principe, J.C. Survival information potential: A new criterion for adaptive system training. *IEEE Trans. Signal Process*. **2012**, *60*, 1184–1194.
32. Principe, J.C.; Xu, D.; Zhao, Q.; John, F. Learning from examples with information theoretic criteria. *J. VLSI Signal Process. Syst. Signal Image Video Technol*. **2000**, *26*, 61–77.
33. Chen, B.; Xing, L.; Liang, J.; Zheng, N.; Principe, J.C. Steady-state Mean-square Error Analysis for Adaptive Filtering under the Maximum Correntropy Criterion. *IEEE Signal Process. Lett*. **2014**, *21*, 880–884.
34. Chen, B.; Principe, J.C. Maximum correntropy estimation is a smoothed MAP estimation. *IEEE Signal Process. Lett*. **2012**, *19*, 491–494.
35. Al-Naffouri, T.Y.; Sayed, A.H. Adaptive filters with error nonlinearities: Mean-square analysis and optimum design. *EURASIP J. Appl. Signal Process*. **2001**, *4*, 192–205.
36. Douglas, S.C.; Meng, T.H.Y. Stochastic gradient adaptation under general error criteria. *IEEE Trans. Signal Process*. **1994**, *42*, 1335–1351.
37. Sayed, A.H. *Fundamentals of Adaptive Filtering*; Wiley: New York, NY, USA, 2003.

38. Shin, H.-C.; Sayed, A.H. Mean-square performance of a family of affine projection algorithms. *IEEE Trans. Signal Process.* **2004**, *52*, 90–102.
39. Papoulis, E.V.; Stathaki, T. A normalized robust mixed-norm adaptive algorithm for system identification. *Signal Process. Lett*. **2004**, *11*, 5286–5298.
40. Shao, M.; Nikias, C.L. Signal processing with fractional lower order moments: Stable processes and their applications. *Proc. IEEE*. **1993**, *81*, 986–1010.
41. Weng, B.; Barner, K.E. Nonlinear system identification in impulsive environments. *IEEE Trans. Signal Process*. **2005**, *53*, 2588–2594.
42. Georgiadis, A.T.; Mulgrew, B. A family of recursive algorithms for channel identification in alpha-stable noise. In Proceedings of the Fifth Bayona Workshop on Emerging Technologies in Telecommunications, Bayona, Spain, 6–8 September 1999; pp. 153–157.
43. Wang, J.; Kuruoglu, E.E.; Zhou, T. Alpha-stable channel capacity. *IEEE Commun. Lett*. **2011**, *15*, 1107–1109.