Entropy **2014**, *16*, 814-824; doi:10.3390/e16020814

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

A Note on the W-S Lower Bound of the MEE Estimation

Badong Chen [1],*, Guangmin Wang [1], Nanning Zheng [1] and Jose C. Principe [2]

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China; E-Mails: gm.wang@stu.xjtu.edu.cn (G.W.); nnzheng@mail.xjtu.edu.cn (N.Z.)

[2] 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.: +86-29-82668672.

Received: 5 November 2013; in revised form: 7 January 2014 / Accepted: 26 January 2014 / Published: 10 February 2014

Abstract: The minimum error entropy (MEE) estimation is concerned with the estimation of a certain random variable (unknown variable) based on another random variable (observation), so that the entropy of the estimation error is minimized. This estimation method may outperform the well-known minimum mean square error (MMSE) estimation especially for non-Gaussian situations. There is an important performance bound on the MEE estimation, namely the W-S lower bound, which is computed as the conditional entropy of the unknown variable given observation. Though it has been known in the literature for a considerable time, up to now there is little study on this performance bound. In this paper, we reexamine the W-S lower bound. Some basic properties of the W-S lower bound are presented, and the characterization of Gaussian distribution using the W-S lower bound is investigated.

Keywords: estimation; entropy; MEE estimation; W-S lower bound

MSC Codes: 62B10

1. Introduction

Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be two random vectors, with joint probability density function (PDF) $p_{XY}(x, y)$, where X represents a unknown variable and Y stands for the observation. An optimal estimator of X based on the observation is a function of Y that minimizes a certain cost function. Under the well-known minimum mean square error (MMSE) criterion, the optimal estimator is:

$$\begin{aligned} g^* &= \arg\min_{g\in \boldsymbol{G}} \mathbf{E}\left[E^2\right] \\ &= \arg\min_{g\in \boldsymbol{G}} \mathbf{E}\left[\left(X-g(Y)\right)^2\right] \\ &= \arg\min_{g\in \boldsymbol{G}} \int_{\mathbb{R}^m}\int_{\mathbb{R}^n} \left(x-g(y)\right)^2 p_{XY}(x,y)dxdy \end{aligned} \tag{1}$$

where **E**[.] denotes the expectation operator, $E = X - g(Y)$ denotes the estimation error, and $\boldsymbol{G}$ denotes the collection of all measurable functions of *Y*. The MMSE criterion is prevalent in estimation theory due to its mathematical tractability. Under Gaussian assumption, the MMSE criterion yields a linear optimal estimator, which requires only a simple matrix-vector operation [1]. When the data are non-Gaussian, however, the MMSE estimator will be suboptimal and even unacceptable, since it considers only up to second-order statistics for their design.

In order to take into account higher-order statistics in design of estimators, researchers proposed many non-MMSE criteria. The minimum error entropy (MEE) is one of such criteria [2–9]. Under MEE criterion, the optimal estimator is obtained by minimizing the error entropy, that is:

$$\begin{aligned} g^\dagger &= \arg\min_{g\in \boldsymbol{G}} H\left(E\right) \\ &= \arg\min_{g\in \boldsymbol{G}} -\int_{\mathbb{R}^n} p_E(e)\log p_E(e)de \\ &= \arg\min_{g\in \boldsymbol{G}} \mathbf{E}\left[-\log p_E(E)\right] \end{aligned} \tag{2}$$

where $H\left(E\right)$ denotes the Shannon entropy [10] of the error$E$, and $p_E(.)$ denotes the error's PDF, which is:

$$p_E(e) = \int_{\mathbb{R}^m} p_{X|Y}(e+g(y)\,|\,y)p_Y(y)dy \tag{3}$$

where $p_{X|Y}(.\,|\,y)$ denotes the conditional PDF of X given $Y = y$, and $p_Y(y)$ is the marginal PDF of Y. The MEE criterion is invariant with respect to the error's mean. In practice, the MEE estimator is usually restricted to an unbiased one with zero-mean error. The entropy is a measure of concentration of a distribution, minimizing the error entropy forces the error to gather.

The early work in MEE estimation can be traced back to the late 1960s when Weidemann and Stear [2] studied the use of error entropy as a cost function for analyzing the performance of a general sampled-data estimating systems. Minamide [5] extended Weidemann and Stear's results to a continuous-time estimating system. Tomita, Kalata, Minamide *et al.* applied the MEE estimation to linear Gaussian systems, and studied filtering (state estimation), smoothing, and predicting problems from the information theory viewpoint [3–5]. Some important properties of the MEE estimation were also reported in [11–16]. In recent years, MEE has become a popular optimization criterion in the areas of signal processing and machine learning [8,9,17–22]. Combining kernel density estimation (KDE) and Renyi's quadratic entropy yields a computationally simple, nonparametric entropy estimator that has been successfully used in *information theoretic learning* (ITL) [8].

There is a performance bound on the MEE estimation, which was originally derived by Weidemann and Stear [2], and later was rederived and named the W-S lower bound by Janzura *et al.* [6]. The W-S lower bound provides a lower bound on the error entropy, although it is not necessarily attained by the MEE estimator for a given joint distribution p_{XY}. This performance bound is nothing but the conditional entropy of the unknown variable X given the observation Y, that is, we have:

$$H(E) \geq H(X \mid Y) \tag{4}$$

The above inequality can be easily derived using Jensen's inequality. Let $\Phi(x) = -x \log x$. We have:

$$\begin{aligned} H(E) &= \int_{\mathbb{R}^n} \Phi\left(p_E(x)\right) dx \\ &= \int_{\mathbb{R}^n} \Phi\left(\int_{\mathbb{R}^m} p_{X|Y}(x+g(y) \mid y) p_Y(y) dy\right) dx \\ &\overset{(a)}{\geq} \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}^m} \Phi\left(p_{X|Y}(x+g(y) \mid y)\right) p_Y(y) dy\right] dx \\ &= \int_{\mathbb{R}^m} \left[\int_{\mathbb{R}^n} \Phi\left(p_{X|Y}(x+g(y) \mid y)\right) dx\right] p_Y(y) dy \\ &= \int_{\mathbb{R}^m} \left[\int_{\mathbb{R}^n} \Phi\left(p_{X|Y}(x \mid y)\right) dx\right] p_Y(y) dy \\ &= \int_{\mathbb{R}^m} H(X|Y=y) p_Y(y) dy \\ &= H(X|Y) \end{aligned} \tag{5}$$

where (a) comes from the concavity of $\Phi(x)$ and Jensen's inequality, and $H(X|Y=y)$ denotes the conditional entropy of X given $Y=y$.

The performance bounds are very important in estimation theory. So far there is, however, little study on the W-S lower bound of the MEE estimation. In this paper, we will present some important properties of the W-S lower bound, and show that this performance bound can be applied to characterize the Gaussian distribution. The rest of the paper is organized as follows: in Section 2, some basic properties of the W-S lower bound are presented. In Section 3, the characterization of the Gaussian distribution using W-S lower bound is investigated. Finally, the conclusions are given in Section 4.

2. Some Properties of the W-S Lower Bound

In the following, we present some properties of the W-S lower bound. First, we present several sufficient and necessary conditions under which the W-S lower bound can be achieved.

Theorem 1: Let X and Y be two random vectors, $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$. The MEE estimator $\hat{X} = g^\dagger(Y)$ of X based on Y achieves the W-S lower bound $H(X|Y)$ if and only if any one of the following properties holds:

(1) the error $E = X - g^\dagger(Y)$ is independent of Y;

(2) $X = g^\dagger(Y) + Z$, where $Z \in \mathbb{R}^n$ is a random vector that is independent of Y;

(3) $p_{XY}(x,y) = p_Z(x - g^\dagger(y)) p_Y(y)$, where $p_Z(.)$ is a density function that is independent of Y;

(4) $p_E(x) = p_{X|Y}(x + g^\dagger(y) \mid y)$.

Proof: (1) Denote $I(E;Y)$ the mutual information between E and Y. It is easy to derive:

$$\begin{aligned} I(E;Y) &= H(E) - H(E|Y) \\ &= H(E) - H(X - g^{\dagger}(Y)|Y) \\ &= H(E) - H(X|Y) \end{aligned} \tag{6}$$

Hence $H(E) = H(X|Y) \Leftrightarrow I(E;Y) = 0$. The mutual information $I(E;Y)$ equals zero if and only if E and Y are independent, so we conclude that the MEE estimator achieves the W-S lower bound if and only if the error is independent of Y.

(2) As $E = X - g^{\dagger}(Y)$, we have $X = g^{\dagger}(Y) + E$. The error entropy $H(E)$ achieves the W-S lower bound $H(X|Y)$ if and only if E is independent of Y (*i.e.*, $I(E;Y) = 0$), so we have $X = g^{\dagger}(Y) + Z$, where $Z \in \mathbb{R}^n$ is independent of Y (let $Z = E$).

(3) The error entropy achieves the W-S lower bound if and only if $X = g^{\dagger}(Y) + Z$, where $Z \in \mathbb{R}^n$ is independent of Y. Denote the density function of Z by $p_Z(.)$. The conditional density function of X given $Y=y$ will be $p_{X|Y}(x|y) = p_Z(x - g^{\dagger}(y))$. Thus $p_{XY}(x,y) = p_{X|Y}(x|y)p_Y(y) = p_Z(x - g^{\dagger}(y))p_Y(y)$.

(4) Since the conditional density function of X given Y is $p_{X|Y}(x|y) = p_Z(x - g^{\dagger}(y)) = p_E(x - g^{\dagger}(y))$, we have $p_E(x) = p_{X|Y}(x + g^{\dagger}(y) \mid y)$.

Remark: The properties (2)~(4) of Theorem 1 suggest that if the error entropy achieves the W-S lower bound, only the location (or mean) of the conditional density of X given $Y = y$ will depend on y through function $g^{\dagger}(y)$, while the shape of the conditional density is always the same as the shape of the error density, which is independent of y.

Theorem 2: Let X and Y be two random vectors, $X = [X_1, X_2, \cdots, X_n] \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$. If there exists an MEE estimator of X based on Y that achieves the W-S lower bound $H(X|Y)$, then $\forall r = [r_1, r_2, \cdots, r_n] \in \mathbb{N}^n$, the r-th conditional central moment of X given $Y=y$ is constant over $\mathbb{R}^m$, that is, $\mathbf{E}\left[(X_1 - \mu_1)^{r_1}(X_2 - \mu_2)^{r_2} \cdots (X_n - \mu_n)^{r_n} | Y = y\right]$ does not depend on y, where μ_i denotes the conditional mean value of X_i given $Y=y$.

Proof: If the error entropy achieves the W-S lower bound, the shape of the conditional density of X given $Y=y$ will not depend on y. Thus, the theorem holds since the shape of a density function determines its central moments (Note that the central moments depend only on the shape of a density, and are independent of the location of distribution).

Theorem 3: Let X and Y be two random vectors, $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$. If there exists an MEE estimator of X based on Y that achieves the W-S lower bound $H(X|Y)$, then it will be:

$$\hat{X} = g^{\dagger}(Y) = \mathbf{E}\left[X|Y\right] + c \tag{7}$$

where $c \in \mathbb{R}^n$ is a n-dimensional constant vector.

Proof: According to the property (2) of Theorem 1, we have $X = g^{\dagger}(Y) + Z$, where $Z \in \mathbb{R}^n$ is a random vector that is independent of Y. It follows easily that:

$$\mathbf{E}\left[X\middle|Y\right]=\mathbf{E}\left[g^{\dagger}(Y)+Z\middle|Y\right]=g^{\dagger}(Y)+\mathbf{E}\left[Z\right] \tag{8}$$

And hence, $g^{\dagger}(Y)=\mathbf{E}\left[X\middle|Y\right]+c$, where $c=-\mathbf{E}\left[Z\right]$.

It has been shown in [15] that the MEE estimator may be non-unique even if the error distribution is restricted to zero-mean (unbiased). However, the following corollary holds.

Corollary 1: Let X and Y be two random vectors, $X\in\mathbb{R}^{n}$, $Y\in\mathbb{R}^{m}$. If there exists an MEE estimator of X based on Y that achieves the W-S lower bound $H(X|Y)$, then the unbiased MEE estimator will be unique, and be identical to the MMSE estimator.

Proof: If error E is restricted to zero-mean (*i.e.*, $\mathbf{E}\left[E\right]=\mathbf{0}$), by Theorem 3, we have $g^{\dagger}(Y)=\mathbf{E}\left[X\middle|Y\right]$ (*i.e.*, $c=\mathbf{0}$). In this case, the MEE estimator becomes the conditional mean of X given Y (*i.e.*, the MMSE estimator), which is, obviously, unique.

Theorem 4: Let X and Y be two random vectors, $X\in\mathbb{R}^{n}$, $Y\in\mathbb{R}^{m}$. If there exists an MEE estimator of X based on Y that achieves the W-S lower bound $H(X|Y)$, then the MEE estimator and the smoothed MEE (SMEE) estimator of X based on Y will be identical.

Proof: According to [16], the SMEE estimator is obtained by minimizing the smoothed MEE criterion $H\left(E+\lambda U\right)$, where λ is the smoothing factor, and U is a smoothing variable (see [16] for the detailed description of the smoothing variable) that is independent of X, Y and E. Clearly, the SMEE estimator of X based on Y is identical to the MEE estimator of $X+\lambda U$ based on Y. Since the MEE estimator of X based on Y achieves the W-S lower bound, we have $X=g^{\dagger}(Y)+Z$, where Z is a random vector that is independent of Y. It follows that $X+\lambda U=g^{\dagger}(Y)+Z+\lambda U=g^{\dagger}(Y)+Z'$, where $Z'=Z+\lambda U$. Because U is independent of X, Y and E, the variable Z' will also be independent of Y. By property (2) of Theorem 1, one may easily conclude that the MEE estimator of X based on Y is identical to the MEE estimator of $X+\lambda U$ based on Y. This completes the proof.

Theorem 5: Let the random vector $\begin{pmatrix}X\\Y\end{pmatrix}$ has a joint (multivariate) Gaussian distribution, $\begin{pmatrix}X\\Y\end{pmatrix}\sim\mathcal{N}\left(\mu,\Sigma\right)$, where $\mu=\begin{pmatrix}\mu_X\\\mu_Y\end{pmatrix}$, $\Sigma=\begin{pmatrix}\Sigma_{XX} & \Sigma_{XY}\\\Sigma_{YX} & \Sigma_{YY}\end{pmatrix}$, then the MEE estimator of X based on Y will achieve the W-S lower bound, and it will be an affine linear function of Y.

Proof: It is easy to prove that the conditional distribution of X given Y has a Gaussian distribution with mean vector $\mathbf{E}\left[X\middle|Y\right]=\mu_X+\Sigma_{XY}\Sigma_{YY}^{-1}\left(Y-\mu_Y\right)$ and covariance matrix $\Sigma_{XX}-\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$. That is, the conditional mean of X given Y is an affine linear function of Y, and the conditional covariance matrix of X given Y is constant (*i.e.*, does not depend on Y). Since the shape of the Gaussian distribution depends only on the covariance matrix, the conditional density of X given Y has a fixed shape. And hence, the MEE estimator of X based on Y will achieve the W-S lower bound. By Theorem 3, the MEE estimator of X will also be an affine linear function of Y.

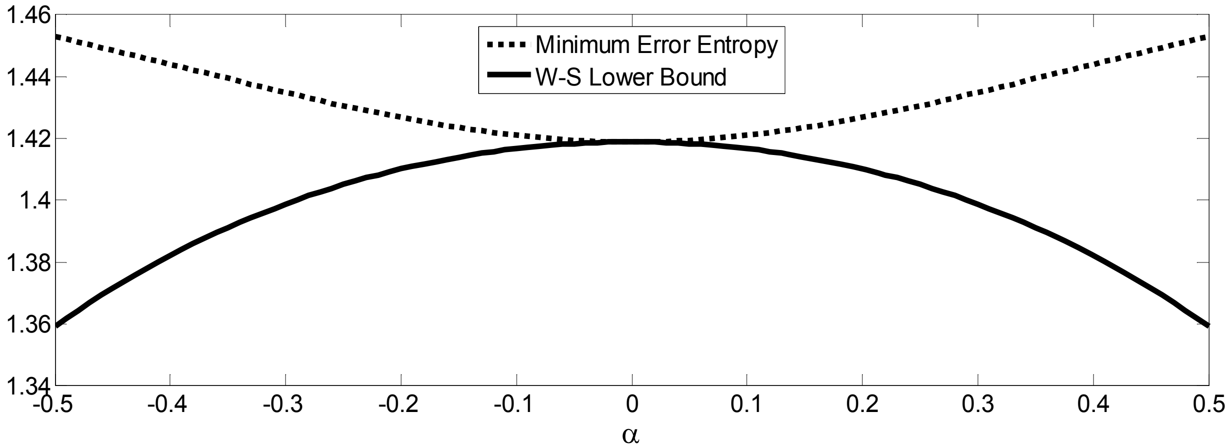
Remark: If X and Y are joint Gaussian, the MEE estimator of X based on Y will achieve the W-S lower bound. However, it should be noted that for most cases, the MEE estimator cannot achieve this performance bound. A simple example is given below.

Example 1: Consider the joint PDF $p_{XY}(x,y) = p_{X|Y}(x|y)\,p_Y(y)$, where:

$$\begin{cases} p_Y(y) = \dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{y^2}{2}\right) \\ p_{X|Y}(x|y) = \dfrac{1}{\sqrt{2\pi}\sigma_y}\exp\left(-\dfrac{x^2}{2\sigma_y^2}\right) \end{cases} \tag{9}$$

where $\sigma_y = 1 + \alpha\sin(y)$, $|\alpha| < 1$. The joint density p_{XY} is joint Gaussian only when $\alpha = 0$, and in this case σ_y is independent of Y. For any y and α ($|\alpha| < 1$), the conditional distribution of X given $Y = y$ is Gaussian, which is *symmetric and unimodal* (SUM). According to the Theorem 1 in [12], the MEE estimator of X based on Y will be the conditional median of X given Y. For different α values, we can calculate the minimum error entropy and the W-S lower bound $H(X|Y)$. The results are shown in Figure 1. As one can see clearly, when $\alpha \neq 0$ (joint non-Gaussian), the minimum error entropy is always above the W-S lower bound.

Figure 1. The minimum error entropy and the W-S lower bound.



3. Characterization of the Gaussian Distribution

The W-S lower bound can be applied to characterize the Gaussian distribution, *i.e.*, constructing some conditions under which a distribution is Gaussian (or joint Gaussian). The problem of characterization of the Gaussian distribution is an interesting problem, which has been extensively studied in the literature [23–27]. First, we introduce a lemma.

Lemma 1 [24]: Let X and Y be two random vectors, $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$. If $X \sim \mathcal{N}(\mu_X, \Sigma_{XX})$, and the distribution of Y given X is a q-dimensional Gaussian distribution with mean vector $a + BX$ ($a \in \mathbb{R}^q$, $B \in \mathbb{R}^{q\times p}$), and constant covariance matrix $\Sigma_0 \in \mathbb{R}^{q\times q}$, then the joint distribution of $\begin{pmatrix} X \\ Y \end{pmatrix}$ will be a $(p+q)$-dimensional Gaussian distribution, whose mean vector and covariance matrix are:

$$\begin{cases} \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} \mu_X \\ a + B\mu_X \end{pmatrix} \\ \Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_{YY} \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XX} B^T \\ B\Sigma_{XX} & \Sigma_0 + B\Sigma_{XX} B^T \end{pmatrix} \end{cases} \tag{10}$$

Based on Lemma 1, we can state the following theorem.

Theorem 6: Let X and Y be two random vectors, $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$. Assume that $Y \sim \mathcal{N}\left(\mu_Y, \Sigma_{YY}\right)$, and the distribution of X given Y is a n-dimensional Gaussian distribution. If there exists a linear estimator $\hat{X} = BY$ such that the error entropy $H(E)$ achieves the W-S lower bound $H(X|Y)$, where $B \in \mathbb{R}^{n\times m}$, $E = X - BY$, then $\begin{pmatrix} X \\ Y \end{pmatrix}$ will be a $(n+m)$-dimensional multivariate Gaussian random vector.

Proof: Since the linear estimator $\hat{X} = BY$ achieves the W-S lower bound, by Theorem 1, we have:

$$X = BY + Z \tag{11}$$

where $Z \in \mathbb{R}^n$ is a random vector that is independent of Y. And hence, the conditional mean vector of X given Y is $a + BY$, where $a = \mathbf{E}[Z]$. In addition, the conditional covariance matrix of X given Y is equal to the covariance matrix of Z, which is a constant matrix (*i.e.*, independent of Y). By applying Lemma 1, we complete the proof.

The next lemma is needed in the proof of Theorem 7.

Lemma 2 [25]: Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be two random vectors. Suppose the conditional densities $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$ are both (multivariate) Gaussian. Denote the covariance matrix of the conditional density $p_{X|Y}(x|y)$ by $\Sigma_{X|Y}(y)$. Then the following two statements are equivalent:

(i) the joint density function $p_{XY}(x,y)$ is multivariate Gaussian;
(ii) $\Sigma_{X|Y}(y)$ is constant on $\mathbb{R}^m$.

Theorem 7: Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be two random vectors. Suppose the conditional densities $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$ are both (multivariate) Gaussian. If the MEE estimator of X based on Y achieves the W-S lower bound $H(X|Y)$, then the joint density function $p_{XY}(x,y)$ will be multivariate Gaussian.

Proof: Since the MEE estimator of X based on Y achieves the W-S lower bound, by Theorem 2, the conditional covariance matrix $\Sigma_{X|Y}(y)$ will not depend on y (*i.e.*, be constant on $\mathbb{R}^m$). By Lemma 2, the joint density p_{XY} will be multivariate Gaussian.

Before presenting Theorem 8, we introduce the third lemma, which is an extended version of Ghurye and Olkin's theorem [23,26].

Lemma 3 [26]: Let $U_1 \in \mathbb{R}^p$ and $U_2 \in \mathbb{R}^q$ be two independent non-degenerate random vectors, and let $X_1 \in \mathbb{R}^p$ and $X_2 \in \mathbb{R}^q$ be two independent random vectors such that:

$$\begin{cases} X_1 = A_1 U_1 + A_2 U_2 \\ X_2 = C_1 U_1 + C_2 U_2 \end{cases} \tag{12}$$

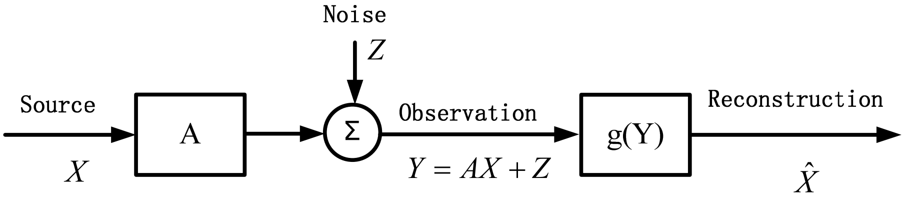
where $A_1 \in \mathbb{R}^{p\times p}$, $A_2 \in \mathbb{R}^{p\times q}$, $C_1 \in \mathbb{R}^{q\times p}$, $C_2 \in \mathbb{R}^{q\times q}$, among which A_1 and C_2 are square nonsingular matrices. Then U_1 and U_2 are multivariate Gaussian random vectors if the following conditions on A_1, A_2, C_1 and C_2 are satisfies:

(i) $P = C_2 - C_1 A_1^{-1} A_2$ is nonsingular,

(ii) none of the rows of $\Gamma_1 = A_1^{-1} A_2$ and $\Gamma_2 = P^{-1} C_1$ are null vectors.

Consider now the problem of estimating the source $X \in \mathbb{R}^n$ given the observation $Y = AX + Z$, where $Y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, and $Z \in \mathbb{R}^m$ is the additive noise that is independent of X, as shown in Figure 2.

Figure 2. General setup of the source estimating problem.



For the above estimation problem, the following theorem holds:

Theorem 8: For the estimation problem in Figure 2, if there exists a linear estimator $\hat{X} = BY$ such that the error entrpy $H(E)$ achieves the W-S lower bound $H(X|Y)$, where $B \in \mathbb{R}^{n\times m}$, $E = X - BY$, then X and Z are multivariate Gaussian random vectors if the following conditions on A and B are satisfied:

(i) the matrices $I_n - BA$ and $I_m + A(I_n - BA)^{-1}B$ are nonsingular, where I_n and I_m are identity matrices of order n and m respectively,

(ii) none of the rows of $(I_n - BA)^{-1}B$ and $\left[I_m + A(I_n - BA)^{-1}B\right]^{-1}A$ are null vectors.

Proof: Since $Y = AX + Z$, the error E can be expressed as:

$$\begin{aligned} E &= X - BY \\ &= X - B(AX + Z) \\ &= (I_n - BA)X - BZ \end{aligned} \tag{13}$$

Thus we have:

$$\begin{cases} E = A_1 X + A_2 Z \\ Y = C_1 X + C_2 Z \end{cases} \tag{14}$$

where $A_1 = I_n - BA$, $A_2 = -B$, $C_1 = A$, $C_2 = I_m$. On the other hand, the error entropy $H(E)$ achieves the W-S lower bound $H(X|Y)$ if and only if E is independent of the observation Y. Then by applying Lemma 3, we arrive easily at the results. When $m = n = 1$, the conditions (i) and (ii) in Theorem 8 will be equivalent to the condition that $A \neq 0$, $B \neq 0$, and $B \neq 1/A$. Therefore, we have the following corollary:

Corollary 2: For the estimation problem in Figure 2, if X, Y, Z are all scalar random variables ($m = n = 1$), and there exists a linear estimator $\hat{X} = BY$ such that the error entropy $H(E)$ achieves the

W-S lower bound $H(X|Y)$, where $B \in \mathbb{R}$, $E = X - BY$, then X and Z are scalar Gaussian random variables if $A \neq 0$, $B \neq 0$, and $B \neq 1/A$.

Remark: It is worth noting that when the condition that $A \neq 0$, $B \neq 0$, and $B \neq 1/A$ does not hold, some of the variables in the estimating system will become degenerate random variables. Specifically, when $A = 0$, the variable AX will be a degenerate random variable; when $B = 0$, the estimator $\hat{X} = BY$ will become a degenerate random variable; when $B = 1/A$ ($A \neq 0$), the variable $(1 - BA)X$ will be a degenerate random variable.

4. Conclusions

MEE estimation provides an appealing approach to design optimal estimators in the framework of information theory. Recent successful applications of the MEE criterion in the areas of signal processing and machine learning suggest that this estimation method has significant potential advantages over traditional MMSE estimation, especially when data possess non-Gaussian distributions. Though it has shown remarkable success in many applications, some theoretical aspects of the MEE estimation need further study. In this work, we reexamine the W-S lower bound on the MEE estimation, which is a Shannon theory-type performance bound on the estimation error entropy. We present some basic properties of the W-S lower bound, and give some interesting results on the characterization of Gaussian distribution using this performance bound. It is hoped that the results of this work will help us to gain insights into the attainment of the W-S lower bound in MEE estimation.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61372152, No. 90920301) and 973 Program (No. 2012CB316400, No. 2012CB316402).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Kailath, T.; Sayed, A.H.; Hassibi, B. *Linear Estimation*; Prentice Hall: Upper Saddle River, NJ, USA, 2000.
2. Weidemann, H.L.; Stear, E.B. Entropy analysis of estimating systems. *IEEE T. Inform. Theory* **1970**, *16*, 264–270.
3. Tomita, Y.; Ohmatsu, S.; Soeda, T. An application of the information theory to estimation problems. *Inform. Contr* **1976**, *32*, 101–111.
4. Kalata, P.; Priemer, R. Linear prediction, filtering and smoothing: An information theoretic approach. *Inform. Sci.* **1979**, *17*, 1–14.
5. Minamide, N. An extension of the entropy theorem for parameter estimation. *Inform. Contr.* **1982**, *53*, 81–90.
6. Janzura, M.; Koski, T.; Otahal, A. Minimum entropy of error principle in estimation. *Inform. Sci.* **1994**, *79*, 123–144.

7. Wolsztynski, E.; Thierry, E.; Pronzato, L. Minimum-entropy estimation in semi-parametric models. *Signal Process.* **2005**, *85*, 937–949.
8. Principe, J.C. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer: New York, NY, USA, 2010.
9. Chen, B.; Zhu, Y.; Hu, J.; Principe, J.C. *System Parameter Identification: Information Criteria and Algorithms*; Elsevier Inc.: London, UK, 2013.
10. Cover, T.M.; Thomas, J.A. *Element of Information Theory*; Wiley & Sons, Inc.: New York, NY, USA, 1991.
11. Otahal, A. Minimum entropy of error estimate for multi-dimensional parameter and finite-state-space observations. *Kybernetika* **1995**, *31*, 331–335.
12. Chen, T.-L.; Geman, S. On the minimum entropy of a mixture of unimodal and symmetric distributions. *IEEE Trans. Inform. Theory* **2008**, *54*, 3166–3174.
13. Chen, B.; Zhu, Y.; Hu, J.; Zhang, M. On optimal estimations with minimum error entropy criterion. *J. Franklin Inst.* **2010**, *347*, 545–558.
14. Chen, B.; Zhu, Y.; Hu, J.; Zhang, M. A new interpretation on the MMSE as a robust MEE criterion. *Signal Process.* **2010**, *90*, 3313–3316.
15. Chen, B.; Principe, J.C. Some further results on the minimum error entropy estimation. *Entropy* **2012**, *14*, 966–977.
16. Chen, B.; Principe, J.C. On the smoothed minimum error entropy criterion. *Entropy* **2012**, *14*, 2311–2323.
17. Erdogmus, D.; Principe, J.C. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Trans. Signal Process.* **2002**, *50*, 1780–1786.
18. Erdogmus, D.; Principe, J.C. Generalized information potential criterion for adaptive system training. *IEEE Trans. Neural Network* **2002**, *13*, 1035–1044.
19. Erdogmus, D.; Principe, J.C. Convergence properties and data efficiency of the minimum error entropy criterion in Adaline training. *IEEE Trans. Signal Process.* **2003**, *51*, 1966–1978.
20. Erdogmus, D.; Principe, J.C. From linear adaptive filtering to nonlinear information processing—The design and analysis of information processing systems. *IEEE Signal Process. Mag.* **2006**, *23*, 14–33.
21. Santamaria, I.; Erdogmus, D.; Principe, J.C. Entropy minimization for supervised digital communications channel equalization. *IEEE Trans. Signal Process.* **2002**, *50*, 1184–1192.
22. Chen, B.; Hu, J.; Pu, L.; Sun, Z. Stochastic gradient algorithm under (h, ϕ)-entropy criterion. *Circuits Syst. Signal Process.* **2007**, *26*, 941–960.
23. Ghurye, S.G.; Olkin, I. A characterization of the multivariate normal distribution. *Ann. Math. Stat.* **1962**, *33*, 533–541.
24. Fidalgo, J.L.; Albajar, R.A. Characterizing the general multivariate normal distribution through the conditional distributions. *Extr. Math.* **1997**, *12*, 15–18.
25. Bischoff, W.; Fieger, W. Characterization of the multivariate normal distribution by conditional normal distributions. *Metrika* **1991**, *38*, 239–248.
26. Fisk, P.R. A note on a characterization of the multivariate normal distribution. *Ann. Math. Stat.* **1970**, *41*, 486–494.

27. Hamedani, G.G. On a recent characterization of the bivariate normal distribution. *Metrika* **1991**, *38*, 255–258.