MINIMUM ENTROPY ESTIMATION AS A NEAR MAXIMUM-LIKELIHOOD METHOD AND ITS APPLICATION IN SYSTEM IDENTIFICATION WITH NON-GAUSSIAN NOISE

MINH TA AND VICTOR DEBRUNNER
School of Electrical and Computer Engineering
The University of Oklahoma
202 West Boyd Street, Room 219
Norman, OK 73019, USA
{taqminh,vdebrunn}@ ou.edu

ABSTRACT

We derive the Minimum Entropy Estimation (MEE) method from Information Theory to show the similarity of this method to the Maximum Likelihood Method for the linear regression problem. The result is a nonparametric-based identification technique that can be applied in any case with iid noise that outperforms estimators in this case, including the popular LS method and a recently-developed (and limited) version of the MEE. Performance-wise, the MEE method is comparable to the Expectation-Maximization (EM) method. Its application to FIR system identification produces a very efficient implementation of this technique.

1. INTRODUCTION

Estimation of unknown parameters based on available data is an important problem that finds applications in numerous fields. This paper investigates a particular form of this problem, namely the linear regression problem, which is the foundation of the field of system identification.

In the case when the characteristic of the output noise is known, the best solution to the linear regression problem is undoubtedly the classical Maximum Likelihood (ML) estimator. However, this knowledge is not realistically available in most situations. Thus, a straightforward application of the ML method is not possible. The widely-used (sub-optimal) approach is to assume that the noise is Gaussian and the ML estimator becomes the Least-Squares (LS) estimator. Other methods such as Expectation Maximization (EM) promise some results asymptotic to the ML solution [1]. In this paper, we consider an alternative – the method of Minimum Entropy – that also shares some desirable properties with the ML method. At this point of research, this Minimum Entropy estimator produces comparable results with the method of EM while having apparently lower computational complexity.

The idea of Minimum Entropy Estimation has been considered by several researchers (namely, [2], [3], and [4]). However, previous attempts either lack precise mathematical rationales [3], [4]; or were derived for the limited case where the output noise has an even *pdf* [2]. To surmount these shortcomings, we derive the Minimum Entropy Estimator (MEE) from the framework of Information Theory to show that this method can be applied to *iid* noises with any *pdf*. Our derived method is, therefore, far better than the popular LS estimator in this case. Furthermore, our proposed method outperforms the somewhat similar algorithm given in [2].

2. PROBLEM FORMULATION AND ANALYSIS

Consider the linear regression problem:

$$\underline{y} = \overline{\boldsymbol{\theta}}^T \underline{\mathbf{u}} + \underline{\varepsilon} \tag{1}$$

where the random variable (RV) $\underline{y}$ and random vector $\underline{\mathbf{u}}$ can be observed and $\underline{\varepsilon}$ is the immeasurable random noise. We wish to estimate the true parameter $\overline{\boldsymbol{\theta}}$ given the finite observations $\{y_i, u_i\}$, i=1…N, where $N >> \dim(\overline{\boldsymbol{\theta}})$. We assume that the noise $\{\varepsilon_i\}$ are *iid* with the *pdf* $f_{\underline{\varepsilon}}(\varepsilon)$ and the sequence of vectors $\{\mathbf{u}_i\}$ are also *iid* with the *pdf* $f_{\underline{\mathbf{u}}}(\mathbf{u})$. We further assume that $\underline{\mathbf{u}}$ possesses sufficient randomness to identify $\overline{\boldsymbol{\theta}}$ from the data. If $\boldsymbol{\theta}$ is any arbitrary estimate of $\overline{\boldsymbol{\theta}}$ then the residual noise is:

$$\underline{e} = \underline{y} - \boldsymbol{\theta}^T \underline{\mathbf{u}} = (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \underline{\mathbf{u}} + \underline{\varepsilon} \tag{2}$$

Since we assume that both $\{\varepsilon_i\}$ and $\{u_i\}$ are independent sequences, the sequence $\{e_i\}$ is also *iid* with the *pdf*:

$$f_{\underline{e}}(e) = \int_{-\infty}^{\infty} f_{\underline{\mathbf{u}}}(\mathbf{u}) f_{\underline{\varepsilon}}(e - (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{u}) d\mathbf{u} \tag{3}$$

When $f_{\underline{\varepsilon}}(\varepsilon)$ is known, the ML estimator $\hat{\boldsymbol{\theta}}_{ML}$ is

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \left(\sum_{i=1}^{N} \log f(y_i, u_i \mid \boldsymbol{\theta}) \right) \tag{4}$$

As pointed out by Akaike [5], the maximization of the Log-Likelihood Function is in fact the minimization of the (sampled version of) the Kullback-Leibler (K-L) distance from $f(\mathbf{u}, y \mid \boldsymbol{\theta})$ to $f(\mathbf{u}, y \mid \overline{\boldsymbol{\theta}})$:

$$D = E\left\{ \frac{f(\mathbf{u}, y \mid \overline{\boldsymbol{\theta}})}{f(\mathbf{u}, y \mid \boldsymbol{\theta})} \right\} = \int f(\mathbf{u}, y \mid \overline{\boldsymbol{\theta}}) \log \frac{f(\mathbf{u}, y \mid \overline{\boldsymbol{\theta}})}{f(\mathbf{u}, y \mid \boldsymbol{\theta})} d\mathbf{u} dy \tag{5}$$

Note that:

$$f(\mathbf{u}, y \mid \boldsymbol{\theta}) = f_{\underline{\mathbf{u}}}(\mathbf{u}) f_{\underline{\varepsilon}}(y - \boldsymbol{\theta}^T \mathbf{u})$$

Rewriting D in terms of the independent random variables $\underline{\mathbf{u}}$ and $\underline{\varepsilon}$, we have:

$$D = \int f_{\underline{\mathbf{u}}}(\mathbf{u}) f_{\underline{\varepsilon}}(y - \overline{\boldsymbol{\theta}}^T \mathbf{u}) \log \frac{f_{\underline{\varepsilon}}(y - \overline{\boldsymbol{\theta}}^T \mathbf{u})}{f_{\underline{\varepsilon}}(y - \boldsymbol{\theta}^T \mathbf{u})} d\mathbf{u} dy$$

$$= \int f_{\underline{\mathbf{u}}}(\mathbf{u}) f_{\underline{\varepsilon}}(\varepsilon) \log \frac{f_{\underline{\varepsilon}}(\varepsilon)}{f_{\underline{\varepsilon}}(\varepsilon + (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{u})} d\mathbf{u} d\varepsilon$$

$$= -H(\varepsilon) - \int f_{\underline{\mathbf{u}}}(\mathbf{u}) f_{\underline{\varepsilon}}(\varepsilon) \log f_{\underline{\varepsilon}}(\varepsilon + (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{u}) d\mathbf{u} d\varepsilon$$

where $H(\varepsilon) = -\int f_{\underline{\varepsilon}}(\varepsilon) \log f_{\underline{\varepsilon}}(\varepsilon) d\varepsilon$ is the entropy of $\underline{\varepsilon}$. Performing the change of variable: $\varepsilon \leftarrow \varepsilon + (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{u}$, we get:

$$D = -H(\varepsilon) - \int f_{\underline{\mathbf{u}}}(\mathbf{u}) f_{\underline{\varepsilon}}(\varepsilon - (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{u}) \log f_{\underline{\varepsilon}}(\varepsilon) d\mathbf{u} d\varepsilon \qquad (6)$$

Applying (3) into (6):

$$D = -H(\varepsilon) - \int f_{\underline{\varepsilon}}(\varepsilon) \log f_{\underline{\varepsilon}}(\varepsilon) d\varepsilon$$

$$= -H(\varepsilon) - \int f_{\underline{e}}(\varepsilon) \log \frac{f_{\underline{e}}(\varepsilon)}{f_{\underline{\varepsilon}}(\varepsilon)} d\varepsilon - \int f_{\underline{e}}(\varepsilon) \log f_{\underline{e}}(\varepsilon) d\varepsilon \qquad (7)$$

$$= D(f_{\underline{e}} \,\|\, f_{\underline{\varepsilon}}) + H(e) - H(\varepsilon)$$

Note that the $\|$ denotes the K-L distance. Let $\underline{v} = (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \underline{\mathbf{u}}$ then $\underline{v}$ is also an RV. Note that:

$$H(\underline{e}) = I(\underline{e}, \underline{v}) + H(\underline{e} \,|\, \underline{v})$$

where $I(\underline{e}, \underline{v})$ is the mutual entropy between $\underline{e}$ and $\underline{v}$, and:

$$H(e \,|\, v) = -\int f_v(v) \left(\int f(e \,|\, \underline{v} = v) \log f(e \,|\, \underline{v} = v) de \right) dv$$

Because $\underline{e} = \underline{\varepsilon} + \underline{v}$ and the entropy is translation invariant, we get:

$$\int f(e \,|\, \underline{v} = v) \log f(e \,|\, \underline{v} = v) de$$

$$= \int f_{\underline{\varepsilon}}(\varepsilon + v) \log f_{\underline{\varepsilon}}(\varepsilon + v) d\varepsilon$$

$$= \int f_{\underline{\varepsilon}}(\varepsilon) \log f_{\underline{\varepsilon}}(\varepsilon) d\varepsilon$$

Thus:

$$H(e) = I(\underline{e}, \underline{v}) + H(\varepsilon) \qquad (8)$$

Substitution of (8) into (7) yields:

$$D = D(f_{\underline{e}} \,\|\, f_{\underline{\varepsilon}}) + I(\underline{e}, \underline{v})$$

Since both $D(f_{\underline{e}} \,\|\, f_{\underline{\varepsilon}})$ and $I(\underline{e}, \underline{v})$ are non-negative, this equation reveals that the purpose of minimizing D is to get $f_{\underline{e}}$ as close to $f_{\underline{\varepsilon}}$ as possible while making the RV $\underline{e}$ and $\underline{v}$ as independent as possible. Note that both conditions can only happen when $f_{\underline{e}}$ and $f_{\underline{\varepsilon}}$ are the same, hence the minimization of one of the two aforementioned quantities $D(f_{\underline{e}} \,\|\, f_{\underline{\varepsilon}})$ and $I(\underline{e}, \underline{v})$ will lead to the minimization of the other.

In the absence of knowledge about $f_{\underline{\varepsilon}}$, we can't minimize $D(f_{\underline{e}} \,\|\, f_{\underline{\varepsilon}})$. However, knowledge of $\underline{e}$ is available because we measure both y and $\mathbf{u}$. Coupling (8) with the fact that $H(\varepsilon)$ is a constant, we arrive at the conclusion that the minimization of *H(e)* will lead to the minimization of $I(\underline{e}, \underline{v})$ and hence D. This notion forms the basis of the Minimum Entropy method.

3. THE MINIMUM ENTROPY METHOD

As noted in section (2), minimization of *H(e)* w.r.t. to $\boldsymbol{\theta}$ will lead to a close estimate of the true parameter $\overline{\boldsymbol{\theta}}$. However, *H(e)* is not readily available since we don't know $f_{\underline{e}}$ (e) (as a function of unknown $f_{\underline{\mathbf{u}}}(\mathbf{u})$ and $f_{\underline{\varepsilon}}(\varepsilon)$). Thus, an estimate of *H(e)* based on the available data $\{e_i, i=1\ldots N\}$ is needed.

The current way to estimate *H(e)* is based on the estimation of the function $f_{\underline{e}}$ (e) and this, in turn, is a nonparametric problem (see e.g. [6], [7]). The classical approach is Parzen's kernel estimator:

$$\hat{f}_{\underline{e}}(e) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{e - e_i}{h} \right) \qquad (9)$$

where K is any function integrable to 1 that approximates the Dirac delta function, such as the Gaussian kernel $(1/\sqrt{2\pi}) \exp(-e^2/2)$ (see more details in [6] [7]).

Once the approximated *pdf* $\hat{f}_{\underline{e}}(e)$ is established, the entropy *H(e)* can be estimated in several ways (see [8], [9]). Two popular entropy estimators are:

- The estimator based on the Law of Large Number [8]:

$$\hat{H}(e) = -\frac{1}{N} \sum_{i=1}^{N} \log\left(\hat{f}_{\underline{e}}(e_i) \right)$$

- The "plug-in" estimator [9]:

$$\hat{H}(e) = -\int_{-A}^{A} \hat{f}_{\underline{e}}(e) \log \hat{f}_{\underline{e}}(e) de$$

Let $\hat{H}(e)$ be an entropy estimator, the Minimum Entropy Estimator for the parameter $\overline{\boldsymbol{\theta}}$ is then defined to be:

$$\hat{\boldsymbol{\theta}}_{MEE} \triangleq \arg\min\left(\hat{H}(e) \right) = \arg\min\left(\hat{H}(y - \boldsymbol{\theta}^T \mathbf{u}) \right) \qquad (10)$$

4. THE MINIMUM OF *H(e)*

The function *H(e)* can be shown to have only one global minimum at $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$. From (8), the minimum of *H(e)* is obtained when $I(\underline{e}, \underline{v}) = 0$ or $\underline{e}$ and $\underline{v}$ are independent. Since $\underline{e} = \underline{\varepsilon} + \underline{v}$ and $\underline{\varepsilon}$ and $\underline{v}$ are independent, the above condition can only occur when $\underline{v} = (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \underline{\mathbf{u}}$ is a constant. Since we assume that $\underline{\mathbf{u}}$ is sufficiently random, $(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \underline{\mathbf{u}}$ can't be a constant unless $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$. At this minimum, the empirical error $\underline{e}$ is certainly equal to the true noise $\underline{\varepsilon}$, which shows that minimizing *H(e)* will lead to the minimization of $D(f_{\underline{e}} \,\|\, f_{\underline{\varepsilon}})$.

5. EXTENSION OF THE MEE TO NONLINEAR REGRESSION

The extension of the MEE to nonlinear regression is straightforward. The regression problem is now cast in this form:

$$\underline{y} = G(\overline{\boldsymbol{\theta}}, \underline{\mathbf{u}}) + \underline{\varepsilon} \tag{11}$$

where G is a known function. Since the analysis for the linear regression still holds, the function G must satisfy the following condition in order to estimate $\overline{\boldsymbol{\theta}}$:

$$G(\overline{\boldsymbol{\theta}}, \mathbf{u}) - G(\boldsymbol{\theta}, \mathbf{u}) = \text{a constant}, \ \forall \mathbf{u} \ \Leftrightarrow \ \boldsymbol{\theta} = \overline{\boldsymbol{\theta}} \tag{12}$$

The MEE in this case can be defined as:

$$\hat{\boldsymbol{\theta}}_{MEE} \triangleq \arg\min \hat{H}(\underline{y} - G(\boldsymbol{\theta}, \underline{\mathbf{u}})) \tag{13}$$

where $H(\cdot)$ is again an entropy estimator.

When $G(\cdot)$ doesn't satisfy (12), we consider two cases:

- If $G(\cdot)$ can be decomposed into:

$$G(\overline{\boldsymbol{\theta}}, \mathbf{u}) = g_1(\overline{\boldsymbol{\theta}}_1) + g_2(\overline{\boldsymbol{\theta}}_2, \mathbf{u})$$

where g_1 has a unique inverse and g_2 satisfies (12). Note that the regression equation can be rewritten as:

$$\underline{y} = g_2(\overline{\boldsymbol{\theta}}_2, \underline{\mathbf{u}}) + (g_1(\overline{\boldsymbol{\theta}}_1) + \underline{\varepsilon})$$

If we define $\underline{\varepsilon}' = g_1(\theta_1) + \underline{\varepsilon}$ then $\{ \varepsilon_i' = g_1(\overline{\boldsymbol{\theta}}_1) + \varepsilon_i \}$ is also an *iid* noise sequence. Therefore, $\overline{\boldsymbol{\theta}}_2$ can be estimated by the MEE. If we know furthermore that $\underline{\varepsilon}$ has zero mean then $\overline{\boldsymbol{\theta}}_1$ can be estimated as the mean of the residual noise:

$$\hat{\boldsymbol{\theta}}_1 = g_1^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i - g_2(\hat{\boldsymbol{\theta}}_2, \mathbf{u}_i)\right)\right]$$

- If the function g_1 doesn't have a unique inverse, such as

$$g_1(\overline{\boldsymbol{\theta}}_1) = \overline{\theta}_{11} + \overline{\theta}_{12}$$

then this is the unidentifiable case and $\overline{\boldsymbol{\theta}}_1$ simply can't be estimated from the available data.

Note that in [10], an extension of the work described in [2], the authors also considered the application of their MEE to the nonlinear case. However, their extension still suffers from the same limitation because the output noise is restricted to the class of symmetric *pdf*'s. By following the same development as given in section 2, we remove this artificial restriction.

6. APPLICATION OF THE MEE TO FIR FILTER ESTIMATION

Suppose we need to estimate a p^{th}-order FIR filter whose output is corrupted by unknown non-Gaussian noise:

$$y(n) = \sum_{i=0}^{p-1} w(i)u(n-i) + \varepsilon(n) \tag{14}$$

We can define $\overline{\boldsymbol{\theta}} = [w(1), ..., w(p)]^T$ and $\mathbf{u} = [u(n), ..., u(n-p+1)]^T$ to represent this problem as a linear regression problem. The MEE in this case is:

$$\hat{\boldsymbol{\theta}}_{MEE} = sol\left\{\frac{\partial \hat{H}(y - \boldsymbol{\theta}^T u)}{\partial \theta} = 0\right\} \tag{15}$$

where we consider the entropy estimate:

$$\hat{H}(e = y - \boldsymbol{\theta}^T u) = -\int_{-A}^{A} \hat{f}_{\underline{e}}(e)\log(\hat{f}_{\underline{e}}(e))de$$

and $\hat{f}_{\underline{e}}(e)$ is the Parzen's kernel estimate:

$$\hat{f}_{\underline{e}}(e) = \frac{1}{Nh}\sum_{i=1}^{N}\left(\frac{e - (y_i - \boldsymbol{\theta}^T \mathbf{u}_i)}{h}\right)$$

In general, this solution provides an adequate estimate of the true parameters. However, in some specific cases, it can be significantly improved. We consider in this paper one such case where the input $\{u(n)\}$ has non-zero mean.

Since the goal is to get $\hat{\boldsymbol{\theta}}_{MEE}$ as close to $\overline{\boldsymbol{\theta}}$ as possible, one indirect method is to make $\partial \hat{H}(y - \boldsymbol{\theta}^T u)/\partial \boldsymbol{\theta}\big|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}$ as close to 0 as possible. Computation of the expectation of this quantity yields:

$$E\left\{\frac{\partial \hat{H}(e)}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\right\} = E\{\mathbf{u}\}\frac{1}{N}\sum_{i=1}^{N} E\left\{-\int_{-A}^{A}(\log(\hat{f}_{\varepsilon}(e)) + 1)K'\left(\frac{e - \varepsilon_i}{h}\right)de\right\} \tag{16}$$

Thus, in order to force this expectation to zero, it is desirable to have a sequence of zero-mean input. This can be accomplished by considering a change of variable in the regression equation:

$$y(n) = \overline{\boldsymbol{\theta}}^T \mathbf{u}(n) + \varepsilon(n) = \overline{\boldsymbol{\theta}}^T(\mathbf{u}(n) - \overline{\mathbf{u}}) + (\varepsilon(n) + \overline{\boldsymbol{\theta}}^T \overline{\mathbf{u}}) \tag{17}$$

where $\overline{\mathbf{u}}$ is the sample mean of $\{\mathbf{u}(n)\}$. If we introduce $\mathbf{u}'(n) = \mathbf{u}(n) - \overline{\mathbf{u}}$ and $\varepsilon'(n) = \varepsilon(n) - \overline{\boldsymbol{\theta}}^T \overline{\mathbf{u}}$ then the regression equation is rewritten as:

$$y(n) = \overline{\boldsymbol{\theta}}^T \mathbf{u}'(n) + \varepsilon'(n)$$

with the input sequence $\mathbf{u}'(n)$ having zero-mean and the noise $\varepsilon'(n)$ being an *iid* sequence. The MEE for this regression equation will then have the desired property of $E\left\{\frac{\partial \hat{H}(e)}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\right\} = 0$, making the parameter estimates much closer to the true parameters. The simple operation of removing the sample mean from the vector sequence $\{\mathbf{u}(n)\}$ is equivalent to removing the mean from the input sequence $\{u(n)\}$

We illustrate this algorithm by considering a simulation with an actual case: $\overline{\boldsymbol{\theta}} = [1, -0.5, 0.2, -0.3, 0.1]^T$. The input signal $\{u(n)\}$ is an *iid* sequence uniformly distributed in [0,2] and the simulated noise $\{\varepsilon(n)\}$ is a bimodal distribution with nonzero mean (see Figure 1). We conduct 200 experiments, each has N=512 input/output data samples.

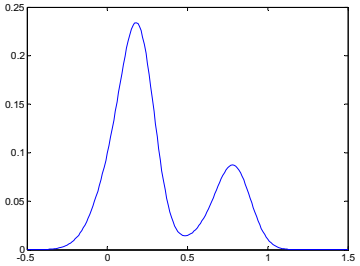

Figure 1. Bimodal density distribution of the true noise

The particular algorithms that we are comparing are the Least Squares (LS) estimator, the MEE introduced by Pronzato and Thierry in [2] (MEE-PT), the general MEE developed in section 2-3, and the modified MEE developed in section 6 (Adj. MEE). We also try the LS algorithm on the adjusted input (Adj. LS) as described in Equation (17).

Some discussion about the previous approaches can be useful. The LS estimator is optimal when the noise $\underline{\varepsilon}$ is Gaussian distributed, thus a non-Gaussian noise yields sub-optimal performance. The algorithms developed by Pronzato and Thierry, on the other hand, make an explicit assumption that the *pdf* of the noise is even symmetric, thus as this assumption is untrue in this case, that method won't perform well either.

Figure 2 shows the average mean square errors (MSE) of the respective algorithms. It is quite clear that the MEE is much better than both the popular LS method and the algorithm proposed by Pronzato and Thierry. To be more precise, Pronzato and Thierry's algorithm will produce a poor estimate with large MSE either if the unknown output noise has non-even symmetric *pdf* or if the input sequence has nonzero mean. Both of these limitations have been addressed in the generalized MEE presented in this paper.

7. CONCLUSIONS

In this paper, we derive the generalized version of a regression technique – the Minimum Entropy Estimation method – from Information Theory to show the kinship of this method with the optimal method of Maximum Likelihood. The results are quite positive – it is much better than the currently popular Least Square method as well as a (somewhat limited) version of the MEE developed recently by Pronzato and Thierry. Our technique has also been shown to be extensible to a class of nonlinear regressions. The application of this method to estimate FIR filters whose output is corrupted by non-Gaussian additive noise is shown to be highly successful, especially with the modified version of our method is used to address the particular case of nonzero-mean input.

At this particular point of research, we note that the MEE does produce a comparable estimate to the well-established Expectation-Maximization method (an iterative ML scheme where the parameters and the noise distribution are estimated alternatively). Because the MEE needs to minimize only one cost function, while the EM method has to minimize a similar

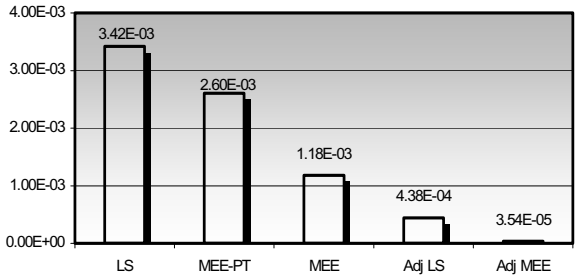

Figure 2. Average MSE's of several algorithms

cost function several times, the computational complexity of MEE is much better than that of the EM method. We will investigate and compare the relative performances between these two methods at a more detailed level in the near future.

The convergence rates of our proposed MEE estimators, as well as their application adaptive filters, are also among several directions of research that we are conducting.

REFERECES

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical Society* B, 39, 1-38, 1977.

[2] L. Pronzato and E. Thierry. "Entropy minimization for parameter estimation problems with unknown distribution of the output noise," *Proc. ICASSP'2001* , 6-11 May 2001, Salt Lake City (USA).

[3] Paul Viola, Nicol N. Schraudolph, Terrence J. Sejnowski, "Empirical Entropy Manipulation for Real-World Problems", *Advances in Neural Information Processing Systems 8*, MIT Press, 1996.

[4] H. Y. Kim, J. H. Kim, "Minimum Entropy Estimation of Hierarchical Random Graph Parameters for Character Recognition," *Proceedings of 15th International Conference on Pattern Recognition*, Vol. 2 , 2000.

[5] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," *Proc. 2nd Symposium on Information Theory*, pp. 267-281, 1973.

[6] E. Parzen, "On estimation of a probability density function and mode," *Annals of Math. Stat*., vol. 35, pp. 1065-1076, 1962.

[7] P. P. B. Eggermont and V. N. LaRiccia, *Maximum Penalized Likelihood Estimation*, Springer, New York, 2001.

[8] I. A. Ahmad and P. E. Lin, "A Nonparametric Estimation of the Entropy for Absolutely Continuous Distributions," *IEEE Trans. on Information Theory*, May 1976.

[9] J. Beirland, E. J. Dudewicz, L. Gyorfi, and E. C. Van der Meulen, "Nonparametric Entropy Estimation: An Overview," *International J. Mathematical and Statistical Sciences*, vol. 6, pp. 17-39, 1997.

[10] L. Pronzato and E. Thierry, "A minimum- entropy estimator for regression problems with unknown distribution of observation errors," Tech. Rep. 00-08, Laboratoire I3S, CNRS- Universite de Nice-Sophia Antipolis, 06903 Sophia Antipolis, France 2000, http://www.i3s.unice.fr/~pronzato/ .