Computing Maximum Entropy Densities: A Hybrid Approach

Badong Chen chenbd04@mails.tsinghua.edu.cn
Department of Precision Instruments and Mechanology
Tsinghua University
Beijing, 100084, P. R. China

Jinchun Hu hujinchun@ tsinghua.edu.cn
Department of Precision Instruments and Mechanology
Tsinghua University
Beijing, 100084, P. R. China

Yu Zhu zhuyu@tsinghua.edu.cn
Department of Precision Instruments and Mechanology
Tsinghua University
Beijing, 100084, P. R. China

Abstract

This paper proposes a hybrid method to calculate the maximum entropy (MaxEnt) density subject to known moment constraints, which combines the linear equation (LE) method and Newton's method together. The new approach is more computationally efficient than ordinary Newton's method as it usually takes fewer Newton iterations to reach the final solution. Compared with the simple LE method, the hybrid algorithm will produce a more accurate solution. Numerical examples confirm the excellent performance of the proposed method.

Keywords: Maximum entropy principle (MEP), maximum entropy density, Lagrange multiplier, Newton's method, hybrid algorithm.

1. INTRODUCTION

The maximum entropy (MaxEnt) density is obtained by maximizing an entropy measure subject to known moment constraints. The underlying theoretical basis of the MaxEnt density is Jaynes' Maximum Entropy Principle (MEP), which states that among all the distributions that satisfy certain constraints (say the moment constraints), one should choose the distribution that maximizes the entropy [1, 2, 3]. The distribution determined by MEP fits the known data without committing extensively to unseen data.

The MaxEnt density provides flexible and powerful tool for density approximation and estimation since it nests a wide range of statistical distributions as special cases, yet, according to the MEP, it yields the most uniform (unbiased) density estimation conditioned on the available a priori knowledge. In fact, most known distributions can be regarded as the MaxEnt distribution with certain moment constrains [2]. The MaxEnt density has found applications in many areas (see [2] for typical examples).

The computation of the MaxEnt density, however, is not an easy task. This is in part, because that the maximization of the entropy is usually achieved through the use of Lagrange multipliers

whose numerical solution requires involved nonlinear optimization. Most existing approaches adopt the iterative Newton's method [4-7]. The Newton's method is too computationally demanding and is sensitive to the choice of the initial values. To reduce the computational complexity, Erdogmus et al. [8] have proposed a simple method to compute the Lagrange multipliers by solving a set of linear equations (LE). The main drawback of Erdogmus' LE method is its poor accuracy. In [9], an efficient numerical approximation of the MaxEnt density has been proposed by Balestrino et al. This method, however, does not compute the exact maximum entropy density (i.e. generalized exponential family).

In this work, we propose a hybrid approach to compute the MaxEnt density, which combines together LE and Newton's methods. The new approach is more computationally efficient than standard Newton's method while produces more accurate solution than LE method. The organization of the paper is as follows. In section II, the MaxEnt densities and some theoretical background are briefly described. In section III some existing algorithms for computing the MaxEnt density are reviewed, and a hybrid method is proposed. In section IV, numerical examples are presented to demonstrate the efficiency of the new approach. Finally, section V gives the conclusion.

2. MAXIMUM ENTROPY DENSITIES

This section gives some theoretical background about the maximum entropy (MaxEnt) densities. The entropy definition in this work is the Shannon's information entropy, which is given by [1]

$$H_S(p) = -\int_{\Omega} p(x)\log p(x)\,dx \tag{1}$$

where $p(x)$ is the probability density function (PDF) of a random variable X. Shannon's entropy measures the uncertainty (or dispersion) contained in $p(x)$.

In general, the MaxEnt density is obtained by maximizing the entropy (1) subject to certain moment constraints. Specifically, we have to solve the constrained non-linear optimization problem:

$$\begin{cases} \max\limits_{p} H_S(p) = -\int_{\Omega} p(x)\log p(x)\,dx \\ \text{s.t.} \begin{cases} \int_{\Omega} p(x)dx = 1 \\ \int_{\Omega} g_k(x)\,p(x)dx = \mu_k, \quad k = 1,2,\cdots,K \end{cases} \end{cases} \tag{2}$$

where $\int_{\Omega} p(x)dx = 1$ is the normalization constraint, $\int_{\Omega} g_k(x)\,p(x)dx = \mu_k$ ($k = 1,2,\cdots,K$) are the K moment constraints, with known functions $g_k(x)$ and known real constants μ_k. Usually the moment constraints take the form ($g_k(x) = x^k$)

$$\int_{\Omega} x^k p(x)dx = \mu_k, \quad k = 1,2,\cdots,K \tag{3}$$

which are the so called arithmetic moment (or power moment) constraints.

By Lagrange's method and the calculus of variation, we can easily derive the analytical solution of the optimization (2), which is expressed as [2]

$$p_{MaxEnt}(x) = \exp\left(-\lambda_0 - \sum_{k=1}^{K} \lambda_k g_k(x)\right) \tag{4}$$

where λ_0, λ_1,..., λ_K are the Lagrange multipliers that satisfy

$$\begin{cases} \int_{\Omega} \exp\left(-\sum_{k=1}^{K}\lambda_k g_k(x)\right)dx = \exp(\lambda_0) \\ \dfrac{\int_{\Omega} g_i(x)\exp\left(-\sum_{k=1}^{K}\lambda_k g_k(x)\right)dx}{\exp(\lambda_0)} = \mu_i, \quad i=1,2,\cdots,K \end{cases} \tag{5}$$

The existence and uniqueness of the MaxEnt density is not guaranteed when arbitrary combinations of moments are used as the side conditions. For the Hausdorff moment problem in which the PDF is defined over $[0,1]$, and the moment constraints are restricted to the arithmetic moments, a necessary and sufficient condition for the existence and uniqueness of the MaxEnt density is as follows [10]

$$\sum_{k=0}^{m}(-1)^k\binom{m}{k}\mu_k > 0\,,\ m=0,1,2,\cdots \tag{6}$$

The above condition is not restrictive as it is satisfied by almost all the PDF defined over $[0,1]$. In [6], it is shown that the arithmetic sample moments for any finite sample always satisfy this condition.

Another important problem involved in MaxEnt density is the convergence problem. Let $p_0(x)$ be a nonnegative function integrable in $[0,1]$ whose arithmetic moments are $\mu_0, \mu_1, \cdots$, and let $p_K(x)$, $K=1,2,\cdots$ be the MaxEnt density sequence associated with the same moments, then [10]

$$\lim_{K\to\infty}\int_0^1 F(x)\,p_K(x)\,dx = \int_0^1 F(x)\,p_0(x)\,dx \tag{7}$$

where $F(x)$ is some continuous function in $[0,1]$. This convergence result suggests that for any finite sample, the MaxEnt density can be used to approximate the underlying distribution arbitrarily well provided that the sample size is large enough to allow precise estimates of the moments.

3. A HYBRID ALGORITHM TO COMPUTING THE MAXIMUM ENTROPY DENSITIES

To compute a MaxEnt density p_{MaxEnt}, we have to determine the Lagrange multipliers λ_0, λ_1,..., λ_K by solving the non-linear system of $K+1$ equations (5). One may use the familiar Newton's method which solves the Lagrange multipliers by iteratively updating [4-6]

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \boldsymbol{H}^{-1}\,\partial\Gamma/\partial\boldsymbol{\lambda} \tag{8}$$

where $\boldsymbol{\lambda} = [\lambda_1,\cdots,\lambda_K]^T$ denotes the Lagrange multipliers vector, $\Gamma = \lambda_0 + \left(\sum_{k=1}^{K}\lambda_k\mu_k\right)$ is the dual objective function, $\partial\Gamma/\partial\boldsymbol{\lambda}$ and $\boldsymbol{H}$ are the gradient and Hessian that take the form ($i, j = 1,2,\cdots,K$)

$$\begin{cases} \dfrac{\partial \Gamma}{\partial \lambda_i} = \mu_i - \mu_{g_i}(\boldsymbol{\lambda}) \\ \boldsymbol{H}_{ij} = \dfrac{\partial^2 \Gamma}{\partial \lambda_i \partial \lambda_j} = \mu_{g_i g_j}(\boldsymbol{\lambda}) - \mu_{g_i}(\boldsymbol{\lambda})\mu_{g_j}(\boldsymbol{\lambda}) \end{cases} \tag{9}$$

where

$$\begin{cases} \mu_{g_i}(\boldsymbol{\lambda}) = \dfrac{\int_{\Omega} g_i(x) \exp\left(-\sum_{k=1}^{K} \lambda_k g_k(x)\right) dx}{\int_{\Omega} \exp\left(-\sum_{k=1}^{K} \lambda_k g_k(x)\right) dx} \\ \mu_{g_i g_j}(\boldsymbol{\lambda}) = \dfrac{\int_{\Omega} g_i(x) g_j(x) \exp\left(-\sum_{k=1}^{K} \lambda_k g_k(x)\right) dx}{\int_{\Omega} \exp\left(-\sum_{k=1}^{K} \lambda_k g_k(x)\right) dx} \end{cases} \tag{10}$$

The Newton's method described above is straightforward but has at least two disadvantages: (1) it is computationally expensive due to a lot of numerical integrals involved, (2) it is sensitive to the choice of the initial values ($\boldsymbol{\lambda}_0$) and only works for a limited set of moment constraints[1]. Wu proposed in [6] a sequential updating method (sequential Newton's method) to calculate the MaxEnt density, which increases the chance of converging to the optimum solution but is still computationally costly.

In order to reduce the computational burden for computing the MaxEnt density, Erdogmus et al. [8] proposed an approximate approach to calculate the Lagrange multipliers by solving a linear system of equations. Specifically, let $\partial \Gamma / \partial \lambda_i = 0$, we have

$$\mu_i = \int_{\Omega} g_i(x) \exp\left(-\lambda_0 - \sum_{k=1}^{K} \lambda_k g_k(x)\right) dx \tag{11}$$

Applying the integrating by parts method and assuming the function $G_k(x) = \int g_k(x) dx$ satisfies $G_k(x)\, p(x)\big|_{-\infty}^{\infty} = 0$, it follows that

$$\mu_i = \sum_{j=1}^{K} \lambda_j \mathbf{E}_p\left[G_i(x) g_j'(x)\right] = \sum_{j=1}^{K} \lambda_j \beta_{ij} \tag{12}$$

where $\beta_{ij} = \mathbf{E}_p\left[G_i(x) g_j'(x)\right]$, $\mathbf{E}_p$ denotes the expectation operator over $p(x)$. Thus the Lagrange multipliers $\boldsymbol{\lambda}$ can be expressed as the solution of a linear system of equations, that is

$$\boldsymbol{\lambda} = \boldsymbol{\beta}^{-1} \boldsymbol{\mu} \tag{13}$$

where $\boldsymbol{\mu} = \left[\mu_1, \cdots, \mu_K\right]^T$, $\boldsymbol{\beta} = \left[\beta_{ij}\right]$.

In this work we refer to (13) as the linear equations (LE) method for solving the Lagrange multipliers of MaxEnt density. It should be noted that in practical applications, one may obtain only approximate solution of (13). In fact, as the expectation in β_{ij} is over the maximum entropy

[1] For this disadvantage, Ormoneit and White [5] suggested two possible reasons: (a) numerical errors may build up during the updating process; (b) near-singularity of the Hessian may occur for a large range of $\boldsymbol{\lambda}$ space.

distribution which is unknown (to be solved), we have to approximate it using the sample mean from actual data distribution:

$$\hat{\beta}_{ij} = \frac{1}{L}\sum_{l=1}^{L} G_i(x_l)\, g'_j(x_l) \tag{14}$$

Therefore, the LE method may produce a solution with less accuracy although it is very computationally simple.

Now we propose a more efficient hybrid method for computing the MaxEnt density, which combines the previous two methods (Newton's and the LE). Specifically, the hybrid method consists of two steps:

Step 1: use the simple LE method to calculate an approximate solution $\tilde{\boldsymbol{\lambda}}$ ($\tilde{\boldsymbol{\lambda}} = \hat{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}$) for the Lagrange multipliers $\boldsymbol{\lambda}$.

Step 2: apply Newton's method to search a more accurate solution, with the estimated Lagrange multipliers $\tilde{\boldsymbol{\lambda}}$ as the initial values for iteration.

The obvious advantages of the proposed approach over previous methods are as follows:
(i) *More computationally efficient than the standard Newton's method:* As the estimated Lagrange multipliers are close to the optimum values, it takes only few (usually one or two) Newton iterations to reach the final solution.
(ii) *No choice of the initial values:* In the hybrid method, the initial values for Newton's iteration are not chosen randomly but instead, are calculated by the LE method. Then there is no problem such as sensitivity to the choice of the initial values and not converging to the optimum solution.
(iii) *More accurate than the LE method:* Due to the refinements by Newton's method, the new approach will produce more accurate solution than the simple LE method.

The hybrid algorithm can be interpreted as a Newton's method for computing the minimum cross-entropy [2] (MinxEnt) density [2]. The MinxEnt density is obtained by minimizing the cross-entropy between $p(x)$ and an a priori density $q(x)$, subject to the same moment constraints on $p(x)$. This yields the following optimization:

$$\begin{cases} \min\limits_{p} D_{KL}(p\|q) = \int_{\Omega} p(x)\log\left(\dfrac{p(x)}{q(x)}\right)dx \\ \text{s.t.} \begin{cases} \int_{\Omega} p(x)dx = 1 \\ \int_{\Omega} g_k(x)\,p(x)dx = \mu_k, \quad k = 1,2,\cdots,K \end{cases} \end{cases} \tag{15}$$

where $D_{KL}(p\|q)$ denotes the cross-entropy (or Kullback-Leibler information divergence, KLID). For the case in which the a priori distribution is a uniform distribution, the MinxEnt density will reduce to the MaxEnt density. The analytical solution for optimization (15) takes the form [2]

$$p_{MinxEnt}(x) = q(x)\exp\left(-\bar{\lambda}_0 - \sum_{k=1}^{K}\bar{\lambda}_k g_k(x)\right) \tag{16}$$

If the a priori density $q(x)$ is chosen as the MaxEnt density produced by the LE method, we have

$$q(x) = \exp\left(-\tilde{\lambda}_0 - \sum_{k=1}^{K}\tilde{\lambda}_k g_k(x)\right) \tag{17}$$

[2] The cross-entropy is also named the Kullback entropy, relative entropy, discrimination information, directed divergence, or the Kullback-Leibler information divergence (KLID).

where $\tilde{\boldsymbol{\lambda}} = \hat{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}$. And hence

$$p_{MinxEnt}(x) = \exp\left(-\gamma_0 - \sum_{k=1}^{K}\gamma_k g_k(x)\right) \tag{18}$$

where $\gamma_i = \overline{\lambda}_i + \tilde{\lambda}_i$, $i = 0,1,\cdots,K$. In this case, the parameters $\{\gamma_i\}$ can be solved by Newton's method with initial values $\boldsymbol{\gamma}_0 = \tilde{\boldsymbol{\lambda}}$, that is

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\gamma}_t - \boldsymbol{H}^{-1}\partial\Gamma'/\partial\boldsymbol{\gamma} \ , \ \boldsymbol{\gamma}_0 = \hat{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu} \tag{19}$$

where $\Gamma' = \gamma_0 + \left(\sum_{k=1}^{K}\gamma_k\mu_k\right)$ is the dual objective function. This is actually the proposed hybrid algorithm (i.e. Newton's method with initial values calculated by the LE method).

4. NUMERICAL EXAMPLES

This section presents numerical examples to demonstrate the efficiency of the hybrid algorithm for computing the MaxEnt density estimates. Consider the case in which the sample data is generated from a mixed-Gaussian distribution with two peaks located at ± 1. Fig. 1 plots the empirical density (histogram) of 5000 samples generated.

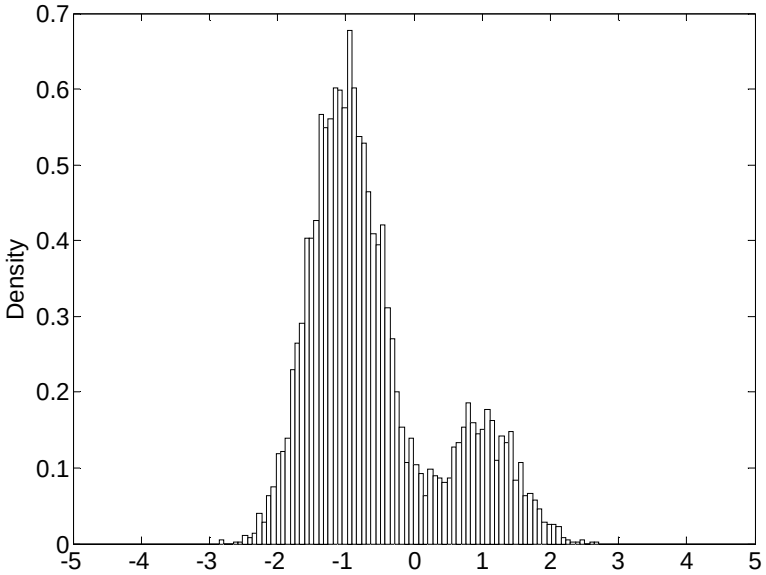

FIGURE 1: Empirical density of data sample obtained from a mixed-Gaussian distribution with peaks located at ± 1.

From the sample, the first eight empirical power moments ($g_k(x) = x^k, k = 1,2,\cdots,8$) have been calculated, and used as the known constants, μ_k , in computing the MaxEnt density. Fig. 2 illustrates a comparison among LE, Newton and hybrid (LE+Newton) algorithms. In Fig. 2, the number N stands for the total number of Newton iterations. The initial values of Newton's method are set as parameters of the standard Gaussian distribution ($\lambda_0 = \log\sqrt{2\pi}$, $\lambda_2 = 0.5$, $\lambda_1 = \lambda_3 = \lambda_4 = \cdots = \lambda_8 = 0$). From Fig. 2, it is evident that the hybrid algorithm converges much faster than ordinary Newton's method and achieves better accuracy than LE method in density

estimation. Note that the hybrid algorithm takes just one Newton iteration to approach the final solution. To further compare the convergence performance between Newton's and hybrid methods, we plot in Fig. 3 the Kullback-Leibler information divergence (KLID) between the estimated and final MaxEnt density.

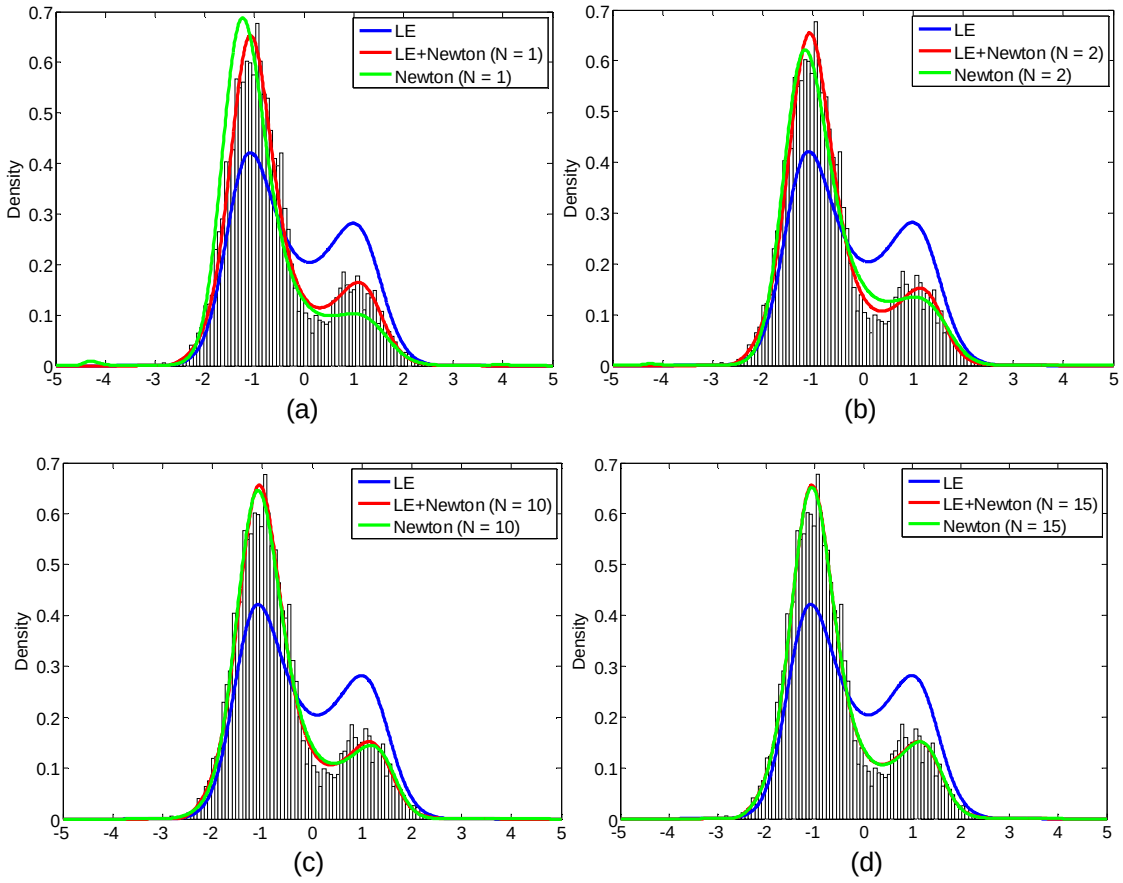

FIGURE 2: MaxEnt density estimates from the data reported in Fig.1.

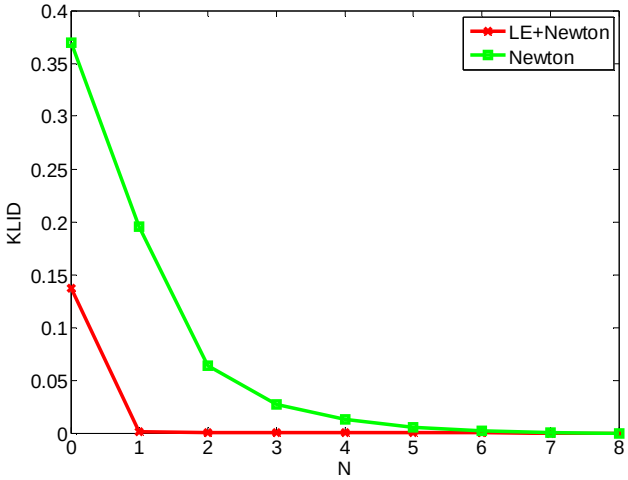

FIGURE 3: Convergence performance of LE+Newton and standard Newton's method.

Consider another set of sample data which are generated from mixed-Gaussian distribution with peaks located at ± 2 (see Fig. 4 for the empirical density). In this case, Newton's method fails to

converge. However, the LE and hybrid methods still work. As shown in Fig. 5, the hybrid algorithm takes only one Newton iteration to reach the final solution and again, achieves better results than LE method.

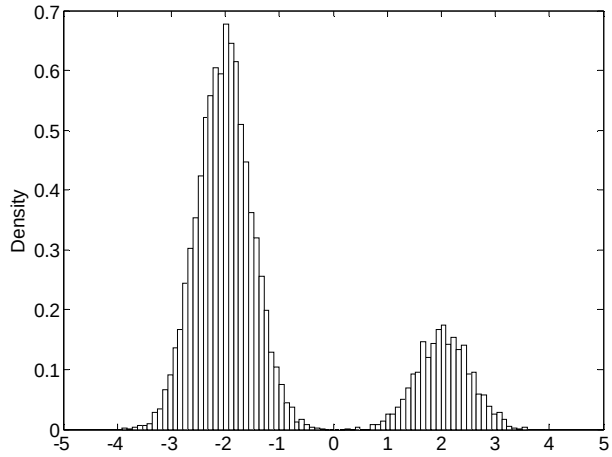

FIGURE 4: Empirical density of data sample obtained from a mixed-Gaussian distribution with peaks located at ± 2 .

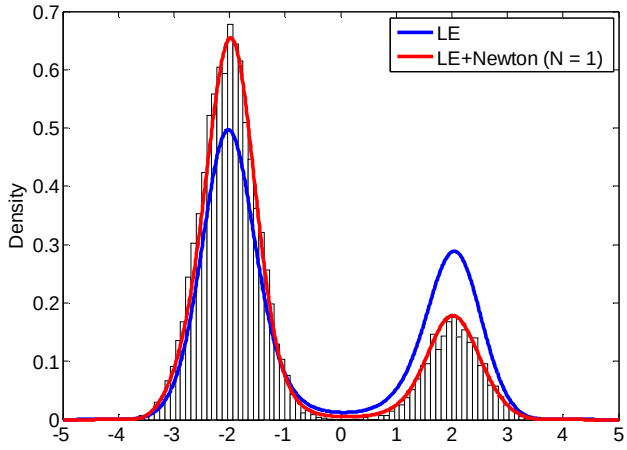

FIGURE 5: MaxEnt density estimates from the data reported in Fig.4.

5. CONCLUSION

A hybrid approach has been proposed to compute the MaxEnt density, which combines the linear equation (LE) method and Newton's method together. First, a rough solution is calculated by the simple LE method. Afterward, a more precise solution is obtained by Newton's method with the rough solution as starting values. The advantages of the proposed approach over standard Newton's method are: (i) increased computational efficiency (faster convergence), (ii) no need for the choice of initial values, (iii) higher numerical stability (converging to the optimum solution). The efficiency of the hybrid algorithm has been confirmed by numerical examples.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 60904054), National Key Basic Research and Development Program (973) of China (No. 2009CB724205), and China Postdoctoral Science Foundation Funded Project (20080440384).

7. REFERENCES

1. T. M. Cover and J. A. Thomas, *"Element of Information Theory"*, Chichester: Wiley & Son, Inc., 1991
2. J. N. Kapur, H. K. Kesavan, *"Entropy Optimization Principles with Applications"*, Academic Press, Inc., 1992
3. E. T. Jaynes, *"Information theory and statistical mechanics"*, Phys. Rev., 106: 620-630, 1957
4. A. Zellner, R. A. Highfield, *"Calculation of maximum entropy distributions and approximation of marginal posterior distributions"*, Journal of Econometrics, 37: 195-209, 1988
5. D. Ormoneit, H. White, *"An efficient algorithm to compute maximum entropy densities"*, Econometrics Reviews, 18(2): 127-140, 1999
6. X. Wu, *"Calculation of maximum entropy densities with application to income distribution"*, Journal of Econometrics, 115(2): 347-354, 2003
7. X. Wu, T. Stengos, *"Partially adaptive estimation via the maximum entropy densities"*, Econometrics Journal, 8: 352-366, 2005
8. D. Erdogmus, K. E. Hild II, Y. N. Rao, J. C. Principe, *"Minimax mutual information approach for independent component analysis"*, Neural Computation, 16: 1235-1252, 2004
9. A. Balestrino, A. Caiti, E. Crisostomi, *"Efficient numerical approximation of maximum entropy estimates"*, International Journal of Control, 79(9): 1145-1155, 2006
10. L. R. Mead, N. Papanicolaou, *"Maximum entropy in the problem of moments"*, Journal of Mathematical Physics, 25(8): 2404-2417, 1984