MAXIMUM ENTROPY BASED NUMERICAL ALGORITHMS FOR APPROXIMATION OF PROBABILITY DENSITY FUNCTIONS

A. Balestrino, A. Caiti, A. Noe', F. Parenti

Dept. Electrical Systems and Automation (DSEA), Univ. of Pisa, Italy; e-mail: aldo.balestrino,andrea.caiti@dsea.unipi.it

Keywords: Maximum Entropy, probability density functions,estimation, identification, Tchebicheff functions.

Abstract

This paper describes several fast algorithms for approximation of the maximum entropy estimate of probability density functions on the basis of a finite number of sampled data. The proposed algorithms are compared with the exact maximum entropy estimate in terms of approximation accuracy and computational efficiency. Some application examples are given.

1 Introduction

In several scientific and engineering applications the problem of estimating the probability density function (p.d.f.) of a random variable on the basis of a finite number of realizations is of crucial interest. In measurement systems, for instance, such estimate gives a complete characterization of a sensor capabilities at different operating conditions. Moreover, whenever one is interested in estimating the interval of variation with a prescribed confidence level (as, for instance, ”the 95% confidence interval”), the whole probability density function of the variable has to be known a priori (e.g., gaussianity of the p.d.f.) or estimated from the available data. In its seminal work, Jaynes [1] has introduced the principle of maximum entropy as the underlying theoretical basis to tackle the p.d.f. estimation problem when the only a priori knowledge available is through moments of the p.d.f itself. Jaynes approach leads to the most uniform (or unbiased) p.d.f. estimate conditioned on the available a priori information. From a computational point of view, the application of the maximum entropy principle leads to the casting and solution of a nonlinear optimization problem of polynomial complexity. Several variations of standard optimization algorithms have been implemented for the solution of such problem, see for instance [2], [3], [4]. Related problems have also been investigated: [2] discusses the inverse problem of determining the set of constraints that optimally describes the observed samples accordingly to the MinMax measure [5]; [6] and [7] have discussed conditions on the moment constraints that guarantees the existence and uniqueness of a maximum entropy p.d.f. Much less explored, at least to the authors knowledge, is the study of numerical schemes for the *approximation* of the maximum entropy estimate. Such study can be of interest in situations in which the p.d.f. is required on-line: if this is the case, the solution of the nonlinear optimization problem may be too computational demanding, while approximated solutions, obtained with faster computational schemes, may be more appropriate. One such situation, that has motivated the present research, is that of on-line localization and tracking of autonomous vehicles when measurement errors are unknown but bounded, with known worst case bound. Standard algorithms from set-membership theory are employed to determine the feasible set in which the vehicle is located [8], [9]; however, more information could be obtained by estimating on line the p.d.f. of some of the observed variables within the bounds determined by the set-membership algorithms. Since the estimate has to be produced on-line, efficiency in the numerical computation is critical for the proper integration of the p.d.f. estimate in the localization and tracking algorithms.

With this background and motivations, in this paper a suite of approximating numerical schemes for the maximum entropy estimate of the p.d.f. from a finite number of samples are proposed and compared with the exact estimate. The proposed algorithms are all based on the construction of the approximating function as a linear combination of basis functions; the functions may be problem-specific, and selected on the basis of the available data. The results obtained from numerical simulations show the validity of the proposed approach.

The paper is organized as follows: in the next section the problem is formally stated, the maximum entropy p.d.f. estimation approach is reviewed, and the numerical algorithm for exact estimation described; in section 3 the implemented approximating algorithms are described; in section 4 the algorithms are compared among themselves and with the exact estimate in terms of accuracy and computational efficiency using simulated data; finally, some conclusions are given.

2 Background and problem statement

Let $f(x)$ be an unknown p.d.f. defined over a finite real interval $[a, b]$ and subject to the natural probabilty constraints:

$$\int_a^b f(x)dx = 1, \quad f(x) \geq 0 \; \forall x \in [a, b] \tag{1}$$

Let us suppose that k additional moment constraints on f are known in the form:

$$\int_a^b g_r(x)f(x)dx = a_r \quad r = 1, \ldots, k \tag{2}$$

with known functions $g_r(x)$ and known real constants a_r.

The maximum entropy estimate of $f(x)$ is obtained by maximization of the Shannon entropy $S(f)$ associated to f subject to the constraints given by equations (1) and (2), where the Shannon entropy is given by:

$$S(f) = -\int_a^b f(x)\ln f(x)dx \tag{3}$$

Jaynes [1] has shown that the maximization of $S(f)$ with respect to f, subject to the constraints (1) and (2), leads to the following analytical solution:

$$f(x) = e^{-\lambda_0-\lambda_1 g_1(x)-\cdots-\lambda_k g_k(x)} \tag{4}$$

where the Lagrangian multipliers $\lambda_0, \cdots, \lambda_k$ satisfy the following relations:

$$\int_a^b \exp\left(-\sum_{j=1}^k \lambda_j g_j(x)\right) dx = \exp\left(\lambda_0\right) \tag{5}$$

$$\frac{\int_a^b g_r(x)\exp\left(-\sum_{j=1}^k \lambda_j g_j(x)\right)dx}{\int_a^b \exp\left(-\sum_{j=1}^k \lambda_j g_j(x)\right)}dx = a_r \quad j = 1, \cdots, k \tag{6}$$

From a practical point of view, the determination of a maximum entropy p.d.f. from available data is reduced to the solution of the nonlinear system of k equations (6). Although it has been shown that this a nonlinear programming problem of polynomial complexity, and that known methods are available for its solution, the computational cost associated to the determination of the maximum entropy p.d.f is such to preclude an on-line use of the estimate (see [3] for a thourough discussion of several computational approaches to the determination of the maximum entropy p.d.f). The results obtained in this paper have been obtained by applying a standard Newton-Raphson method. In the next section some fast computational algorithms for the *approximated* solution of the system (6) are proposed.

3 Fast approximating algorithms

Three computationally efficient algorithms to obtain approximated solutions to the system (6) are now described. Moreover, it is assumed throughout the section that the functions g_r in equation (2) have the following form:

$$g_r(x) = x^r, \quad r = 1, \cdots, k \tag{7}$$

The available data are the real constants a_r in equation (2) The proposed algorithms are based on the approximation of the p.d.f. f (see equation (4)) with a linear combination of basis functions:

$$f(x) \approx \sum_{i=1}^k \beta_i f_i(x) = \hat{f}(x) \tag{8}$$

For any given set of basis functions $\{f_i\}$, the coefficients β_i are determined by solving the following system of *linear* equations:

$$\begin{cases} \beta_1 \int_a^b f_1(x)dx + \ldots + \beta_k \int_a^b f_n(x)dx = 1 \\ \beta_1 \int_a^b x f_1(x)dx + \ldots + \beta_k \int_a^b x f_n(x)dx = a_1 \\ \vdots \\ \beta_1 \int_a^b x^n f_1(x)dx + \ldots + \beta_k \int_a^b x^n f_n(x)dx = a_k \end{cases} \tag{9}$$

The three algorithms differ in the choice of the set of basis functions. Before describing the three possible choices implemented, it is important to underline the relation between the approximating function $\hat{f}$ and the true maximume entropy p.d.f. estimate f. Let the following notation be used, for any generic function $g(x)$ and any integer i:

$$E(g, x^i) = \int_a^b x^i g(x)dx \tag{10}$$

Then by construction:

$$E(f, x^i) = E(\hat{f}, i), \quad i = 1, \cdots, k \tag{11}$$

Since two functions are equal (but for a set of null measure) if all their moments are equal, it follows that $\hat{f} \to f$ as $k \to \infty$.

Algorithm 1: the basis functions f_i are taken as the Tchebycheff polynomials, after normalization of the $[a, b]$ interval to the $[0, 1]$ interval:

$$\begin{cases} f_1(x) = 1 \\ f_2(x) = x \\ f_j(x) = 2x f_{j-1}(x) - f_{j-2}(x) \qquad j = 2, 3, \ldots \end{cases} \tag{12}$$

With this choice of basis functions, which is independent from the available data (i.e., from the coefficients a_r), the linear system (9) can be directly solved.

Algorithm 2: the basis functions f_i are taken so that each of them is the solution of a simplified maximum entropy estimation problem involving one of the known moments; in particular, the basis function f_i is taken as solution of the following problem:

$$\left\{\begin{array}{l} \max(-\int_a^b f_i(x)\ln f_i(x)dx) \\ \qquad \int_a^b f_i(x)dx = 1 \\ \int_a^b x^i f_i(x)dx = a_i \end{array}\right. \tag{13}$$

Applying Jaynes result to the problem (13), one obtains:

$$f_j(x) = e^{-\lambda_{0_j} - \lambda_j x^j} \quad j = 1, \cdots, k \tag{14}$$

where the λ_{0_j} and λ_j are obtained as the solution of the following nonlinear system:

$$\left\{\begin{array}{rcl} \int_a^b e^{-\lambda_j x^j} dx & = & e^{\lambda_{0_j}} \\ a_j \int_a^b e^{-\lambda_j x^j} dx - \int_a^b x^j e^{-\lambda_j x^j} dx & = & 0 \end{array}\right. \tag{15}$$

The computational advantage of this approach is that, instead of solving one nonlinear system in k unknowns, one has to solve k nonlinear equations in one unknown, each one independent from the others. These equations can be potentially solved in parallel, though we have implemented the algorithm sequentially. After the k nonlinear equations have been solved, the functions f_i are determined, and the linear system (9) can be solved.

Algorithm 3: the basis functions f_i are taken so that each of them is the solution of a simplified maximum entropy estimation problem involving two of the known constraints; in particular, any basis function f_i is taken as solution of the following problem:

$$\left\{\begin{array}{l} \max(-\int_a^b f_i(x)\ln f_i(x)dx) \\ \qquad \int_a^b f_i(x)dx = 1 \\ \int_a^b x^p f_i(x)dx = a_p \qquad\qquad p, q \in [1, 2, \cdots, k] \\ \int_a^b x^q f_i(x)dx = a_q \end{array}\right. \tag{16}$$

Of course for each f_i a different couple (a_p, a_q) must be chosen. Each function $f_i, \; i = 1, \cdots, k$ solution of the problem (16) is again given through Jaynes formalism and the solution of a nonlinear system of dimension two. After each f_i has been determined, the linear system (9) can be solved. As compared to Algorithm 2, Algorithm 3 has an additional computational burden due to the need of solving k systems of nonlinear equations of dimension two instead of one; moreover it requires the choice of the couple (a_p, a_q) to be associated to each f_i. In our implementation this choice has been arbitrarily made; however, it may well be the case that some choices are to be preferred in terms of approximating precision or computational efficiency. Note that if only two moments are known, Algorithms 2 and 3 are coincident.

To summarize: the algorithm proposed are all based on the use of basis functions to approximate the maximum entropy estimate. The simplest choice is to use *a priori* defined basis functions (Tchebycheff functions, in our case) and then solve a linear system of equations to obtain the approximation $\hat{f}$. Algorithm 2 increases the computational effort by selecting basis functions on the basis of the available data, and in particular linking each basis function to a simplified maximum entropy problem with one single moment constraint. Algorithm 3 increases even more the computational effort by linking each basis function to a maximum entropy problem with two moment constraints, where the two moment constraints are arbitrarily chosen among the k available constraints. It is clear that one could proceed and define an Algorithm 4, etc. increasing the moment constraints to which the basis functions are related. When all the k moments are employed, one comes back to the original maximum entropy estimate (4).

One important point to note here is that in determining $\hat{f}$ with the procedures described above, the natural probability constraint of equation (1) is not enforced anymore. In particular, for Algorithms 2 and 3 each function f_i respects the constraint, since it is a Jaynes solution of a maximum entropy problem, but their linear combination does not, since the coefficients β_i can assume arbitrary real values. This loss of the probability constraints is due to the fact that $\hat{f}$ is an approximation of f. It has to be remarked, though, that $\hat{f}$ is convergent to f as the number of known moments k increases, and f does respect the natural probabilty constraints.

In the following section the approximating capabilities of the proposed algorithm and their computational efficiency will be investigated through simulative examples and comparison with the true maximum entropy estimate.

4 Examples

The first simulative example considers data generated from an hyperbolic distribution in the interval $[0\ 1]$; 1500 samples have been generated and are reported in the histogram in figure 1. From the samples, the empirical moments of the data set have been computed, and used as known terms a_i in the algorithms previously described. In figs. 2, 3, and 4 the results obtained with the three algorithms using moment information with two, three and four moments respectively are reported and compared with the exact maximum entropy estimate obtained using four moments information. The convergence process of all the three algorithms is rather evident. The closeness between the true estimate and the approximation obtained with Algorithm 3 using the same moment constraints is remarkable.

The second example presented here is related to the classic case of data obtained from a truncated gaussian distribution. In this

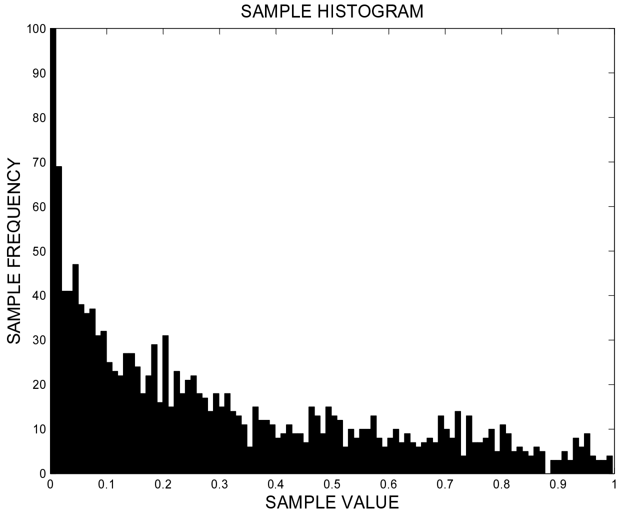

Figure 1: *Data samples obtained from an hyperbolic distribution.*

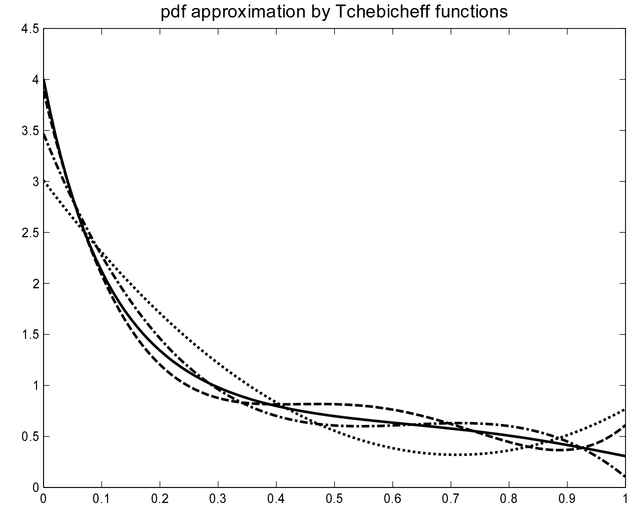

Figure 2: *Probability density function estimates from the data reported in figure 1 obtained with Algorithm 1, based on the use of Tchebycheff functions. Black line is the exact maximum entropy estimate obtained using four moments obtained from the data. Dotted, dash-dotted and dashed lines are the Tchebicheff approximations obtained using two, three and four data moments, respectively.*

case the maximum entropy estimate requires only the knowledge of two moments, and coincides with the true distribution (if the moments are known exactly). First Algorithms 2 and 3 are considered, and compared with the exact maximum entropy estimate as the number of moments used by the algorithm is progressively increased from two to four; the moments have been computed also in this case from a sample of 1500 data in the interval $[0 \;\; 1]$, as in the previous case. In figure 5 the results obtained with Algorithm 2 are shown. The enlargement in the figures show the convergent behaviour of the algorithm as the number of moments increases. Similar results are obtained with Algorithm 3. Note, however, that when only two moments are used, the resulting approximation gets slightly below zero toward the end of the interval; increasing the number of moments the failure in satisfying the probability constraints disappears. The results obtained with Algorithm 1 are shown in figure 6, where a much slower convergence toward the exact estimate can be noted; moreover, the failure in satisfying the probability constraints is present in all the estimates obtained using up to six moments.

The three proposed algorithms and the algorithm for the exact maximum entropy estimate are now compared in terms of computational efficiency. An ensemble of 15 different data realizations have been generated, with different distributions including gaussian, hyperbolic, exponential, and some combinations of the above, in order to include probability densities with more than one maximum. For each of the distributions in the set, the exact maximum entropy algorithm, and the algorithms 1, 2 and 3 have been run using a number of moment constraints ranging from 2 to 6. The mean computational time and the variance obtained from the various methods over this set of distributions is reported in figure 7. The computational time reported is in seconds, and obtained from the implementation of all of the above methods in Matlab on an AMD 1.1 MHz PC. It can be seen that, starting from the fourth moment on, the computational saving obtained with the use of the fast approximating algorithms becomes consistent.

5 Conclusions

Three different algorithms have been presented for a computationally efficient approximation of the maximum entropy estimate of a p.d.f. from a finite sample. The algorithms are all based on approximating the p.d.f. estimate through a set of basis functions. The basis functions may be determined by solving a reduced complexity maximum entropy problem linked to the original problem. The closer the link to the original problem, the fastest the convergence of the approximation to the true maximum entropy estimate, and the lower the computational benefit of the approach. Simulation results show the convergence properties of the algorithm proposed as the number of data moments considered in the computation increases, and the computational advantages with respect to a standard Newton-Raphson method for the determination of the true estimate.

Current work is focusing toward the application of the proposed approach in localization and tracking problems where the data interval is determined through the application of set-membership algorithms.

Acknowledgements

This work has been partially supported by MIUR (project ”Self-localization algorithms for air and maritime vehicles with guaranteed uncertainty bounds”).

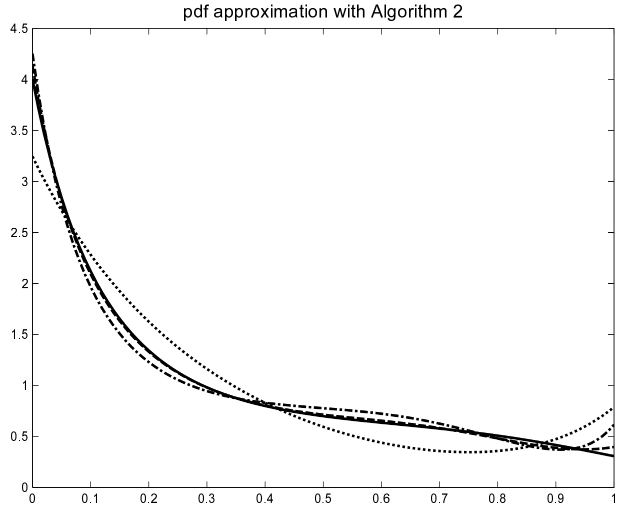

Figure 3: *Probability density function estimates from the data reported in figure 1 obtained with Algorithm 2. Black line is the exact maximum entropy estimate obtained using four moments obtained from the data. Dotted, dash-dotted and dashed lines are the Algorithm 2 approximations obtained using two, three and four data moments, respectively. The dashed line is barely distinguishable from the black line in the figure.*

References

[1] E.T. Jaynes, ”Information theory and statistical mechanics”, *Phys. Rev*, vol. 106, pp. 361-373, 1957.

[2] M. Srikanth, H.K. Kesavan, P.H. Roe, ”Probability density function estimation using the MinMax measure”, *IEEE Trans. Sys. Man Cyber. - part C*, vol.30, n.1, pp. 77-82, 2000.

[3] X. Wu, ”Calculation of maximum entropy densities with application to income distribution”, in revision to *J. Econometrics*, available on line at http://are.berkeley.edu/x̃iming/, 2002.

[4] D. Ormoneit, H. White, ”An efficient algorithm to compute maximum entropy densities”, *Econometric reviews*, vol.18, n.2, pp.127-140, 1999.

[5] J.N. Kapur, G. Baciu, H.K. Kesavan, ”The MinMax information measure”, *Int. J. Sys. Sci.*, vol. 26, n.1, pp. 1-12, 1995.

[6] L.R. Mead, N. Papanicolau, ”Maximum entropy in the problem of moments”, *J. Math. Phys.*, vol.25, n.8, pp.2404-2417, 1984.

[7] A. Tagliani, ”Maximum entropy in the discrete generalized moment problem”, *Statistica*, vol. LX, pp. 59-72, 2000.

[8] M.Milanese, A. Vicino, ”Information based complexity and nonparametric worst-case system identification”, *J. Complexity*, vol. 9, pp. 427-446, 1993.

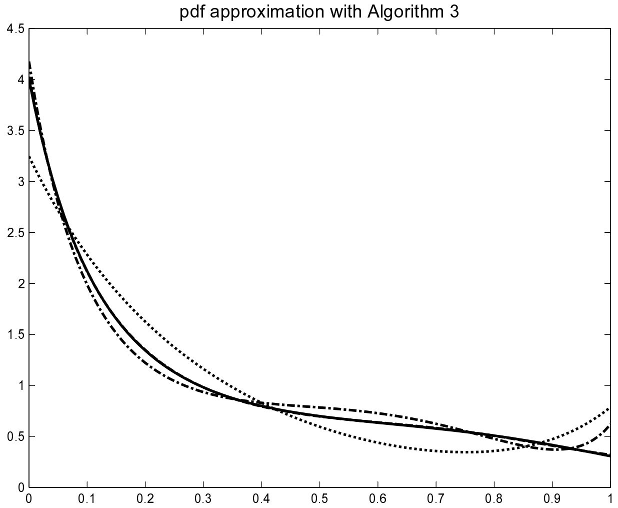

Figure 4: *Probability density function estimates from the data reported in figure 1 obtained with Algorithm 3. Black line is the exact maximum entropy estimate obtained using four moments obtained from the data. Dotted, dash-dotted and dashed lines are the Algorithm 3 approximations obtained using two, three and four data moments, respectively. The dashed line is undistinguishable from the black line in the figure.*

[9] A.Caiti, A.Garulli, F.Livide, D. Prattichizzo, ”Set-membership acoustic tracking of autonomous underwater vehicles”, *Acta Acustica/Acustica*, vol. 88, pp. 648-652, 2002.

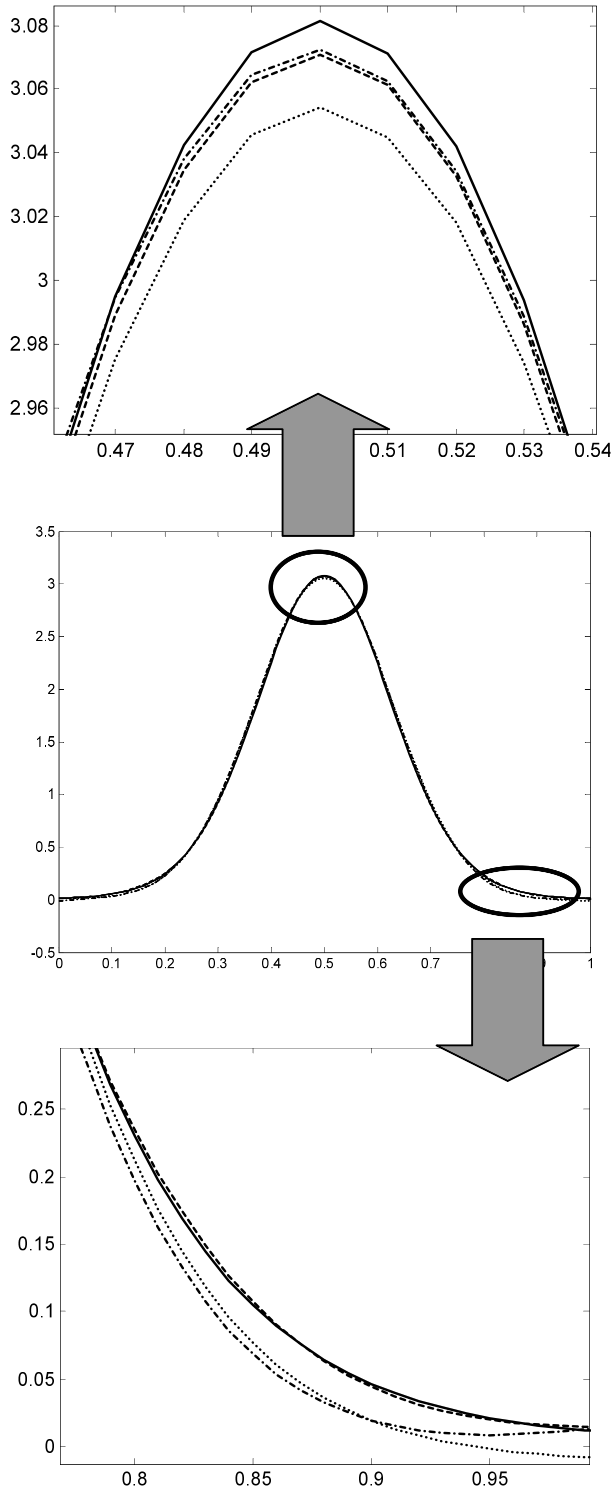

Figure 5: *Probability density function estimates from gaussian distributed data. Black line is the exact maximum entropy estimate obtained using two moments obtained from the data. Dotted, dash-dotted and dashed lines are the Algorithm 2 approximations obtained using two, three and four data moments, respectively.*

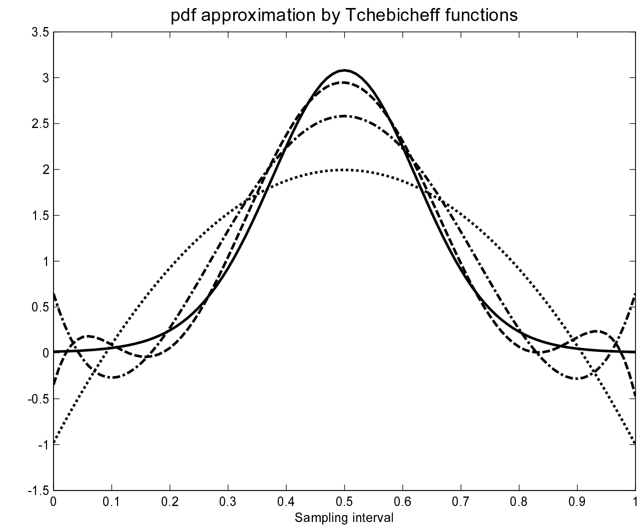

Figure 6: *Probability density function estimates from gaussian distributed data. Black line is the exact maximum entropy estimate obtained using two moments obtained from the data. Dotted, dash-dotted and dashed lines are the Tchebycheff approximations obtained using two, four and six data moments, respectively.*

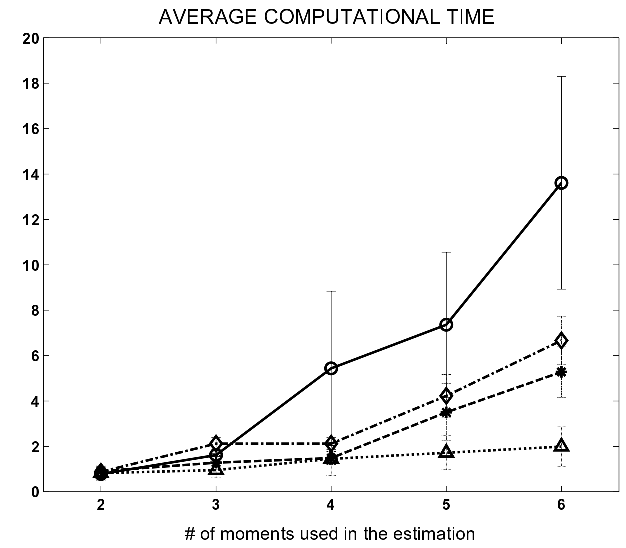

Figure 7: *Average computational time (in seconds) as function of the number of moments employed for the exact maximum entropy estimate (continuous line, circles), Algorithm 1 (dotted line, triangles), Algorithm 2 (dashed line, stars), Algorithm 3 (dash-dotted line, diamonds).*