MAXIMUM ENTROPY ESTIMATION OF THE PROBABILITY DENSITY FUNCTION FROM THE HISTOGRAM USING ORDER STATISTIC CONSTRAINTS

R. Lynn Kirlin, IEEE Fellow

University of Victoria
Department of Electrical Engineering
Victoria, B.C., CANADA

Ali M. Reza, IEEE Senior Member

U.S. Coast Guard Academy
Department of Engineering
New London, CT, U.S.A.

ABSTRACT

An analytical expression for a probability density is usually required in detection and estimation problems, yet it is usually only assumed or selected from contenders by parameter estimation, or the histogram is smoothed with an arbitrary window function. In contrast, given a histogram containing R sample points, we derive a nonlinear differential equation (NDEQ) whose solution is a maximum entropy density given constraints that arise from assumptions that the samples are means of the order statistics of the parent distribution. We solve the NDEQ for R=1 and approximate the solution for general R using the fact that order means partition the density into equal probability regions, which we require to independently be maximum entropy. Finally we show with a Rayleigh density example what errors may result.

Index Terms— estimation of probability density function; maximum entropy; order statistics; histogram

1. INTRODUCTION

Maximum entropy (Maxent) is an optimization technique that is used to produce analytical functions that have maximum uncertainty while simultaneously satisfying apriori knowledge. If the unknown density (pdf) is $f(x)$ and its distribution function (cdf) is $F(x)$, the density of the r-th ordered sample [1] from $f(x)$ is $f_{(r)}(x)$, where, if R is the sample size,

$$f_{(r)}(x) = \frac{R!}{(r-1)!(R-r)!} F^{r-1}(x)[1-F(x)]^{R-r} f(x) \quad (1)$$

In the following we maximize the pdf entropy,

$$H = -\int_{-\infty}^{\infty} f(x) \log f(x) dx \quad (2)$$

by assuming that the samples are the means of the respective order statistics of the parent pdf. The result is a nonlinear differential equation, which, when $R = 1$, is linear and an exponential pdf is the solution. Further, by utilizing the density-partitioning property of means of order statistics and simply constraining the cdf at the sample points we produce a piecewise contiguous Maxent solution and an example when the parent distribution is Rayleigh.

2. ORDER STATISTIC CONSTRAINTS

2.1. Constraints and implications that the samples are the means of the order statistics

The entropy H in (2) is to be maximized over $f(x)$ **subject to the constraints**

$$E\{x_{(r)}\} = \int_{-\infty}^{\infty} x f_{(r)}(x) dx = x_{(r)}\ ,\ r = 1, 2, \ldots, R \quad (3)$$

where $x_{(r)}$ is the r-th ordered value of $\{x_i\}, i = 1, 2, \ldots, R$, and $f_{(r)}(x)$ is the density of $x_{(r)}$. These constraints are much like those given in [2], where the moments of x are assumed known. Some approaches proceed by assuming that the sample moments are the true moments. We offer an alternative.

The constraints we have chosen also imply that the probability distribution function will necessarily be equal to $r/(R+1)$ at $x_{(r)}$ ([1], sec. 3.1); i.e.,

$$F(x_{(r)}) = \frac{r}{R+1} \quad (4)$$

This result, (4), may be useful for an initial approximation to $F(x)$ in an iterative solution of the differential equation which results from the general maximum entropy optimization. Further it is possible to use only these constraints and omit those in (3). However, use of constraints (3) allows a continuous solution, which encompasses those constraints in (4) but not vice versa. Excluding constraints (3) but requiring (4) leads to uniform densities between the samples, each contiguous (sub-) density having area $F(x_{(r)}) - F(x_{(r-1)}) = 1/(R+1)$. That becomes clear when we discard the terms in the resulting differential equation, (8), arising from (3) and add terms according to (4). However, how then to adjust for semi-infinite intervals (tails) in the distribution is not clear at this point, though asymptotic results from (1) allow possible solutions.

2.2. Area and spread constraints

We also require that $f(x)$ have unit area, and we may possibly constrain that var[x] = σ^2. Finally we require boundary conditions $F(a) = 0, F(b) = 1, a < b$, where a and b may be $-\infty$ and ∞ respectively.

2.3. Derivation of the optimal density function

The $R+2$ constraints are appended with Lagrangian multipliers λ_j to the functional H to be maximized giving

$$J(f) = \int \left(G_0 + \sum \lambda_j G_j \right) dx \quad (5)$$

where

$$G_j(x,F,f) = \begin{cases} -f\log f & j=0 \\ xF^{j-1}(1-F)^{R-j}f & 1\le j\le R \\ f & j=R+1 \\ x^2 f & j=R+2 \end{cases} \quad (6)$$

With the boundary condition constraints also applied, Gelfand and Fomin [3] show that the optimal F-satisfies

$$\frac{\partial}{\partial F}\left(G_o+\sum_{j=1}^{R+2}\lambda_j G_j\right) - \frac{d}{dx}\left[\frac{\partial}{\partial f}\left(G_o+\sum_{j=1}^{R+2}\lambda_j G_j\right)\right] \quad (7)$$

The required partials are

$$\frac{\partial G_j}{\partial F} = \begin{cases} 0 & j=0 \\ \frac{1-2F}{F(1-F)}G_j & 1\le j\le R \\ 0 & j=R+1,R+2 \end{cases}$$

$$\frac{\partial G_j}{\partial f} = \begin{cases} -(1+\log f) & j=0 \\ f^{-1}G_j & 1\le j\le R \\ 1 & j=R+1 \\ x^2 & j=R+2 \end{cases}$$

These partials and their derivatives with respect to x are used in (7) to give after considerable algebra the following nonlinear differential equation in F:

$$\frac{F''}{F'} + \sum_{j=1}^{R}\lambda_j F^{j-1}(1-F)^{R-j} - 2\lambda_{R+2}x = 0 \quad (8)$$

Alternate forms are given in the extended version of the paper.

2.4. The differential equation with only CDF (i.e., F) constraints (4) and area = 1

If instead of the means integral constraints (3), only the implied values of $F(x_{(r)})$ are used for constraints, i.e.,

$$F(x_{(r)}) = \int_a^{x_{(r)}} f(x)dx = \frac{r}{R+1} \quad (9)$$

then (8) becomes in each interval,

$$\frac{f'}{f} = 0 \quad (10)$$

and the solutions therein must be the constants

$$f(x) = \frac{1}{(R+1)(x_{(r+1)}-x_{(r)})},\ \begin{cases} x_{(r)} < x \le x_{(r+1)} \\ r = 1,2,\ldots,R \end{cases} \quad (11)$$

This is a simple result, adequate, when a continuous solution is not required. It remains to determine results for the tail.

3. ANALYTICAL SOLUTION FOR $\mathbf{R=1}$

An analytical solution of (8) for $R>1$ has not been found. However, four special cases for $R=1$ have been found and are informative and useful.

1. No constraint on $\sigma^2, a=0, b=\infty$. (The density allows values on the positive real line and we in no way fix the variance or use it to further maximize the average entropy H).
2. No constraint on $\sigma^2, a=-\infty, b=\infty$. (The density allows values on the whole real line and we in no way fix the variance or use it to further maximize H). We do allow one point of discontinuity in $f(x)$. (See [2]).
3. No constraint on σ^2, a, b real and finite.
4. Constrain σ^2 such that H is further optimized, $a=-\infty, b=\infty$.

The differential equation for all four cases is ($\mathbf{R=1}$, $R=1$),

$$\begin{cases} F'' + \lambda_1 F' - 2\lambda_{R+2}xF' = 0, \text{ or} \\ f' + \lambda_1 f - 2\lambda_{R+2}xf = 0 \end{cases} \quad (12)$$

For cases 1, 2 and 3, $\lambda_{R+2}=0$ (no σ^2 constraint):

3.1. Case 1 Solution

Eq. (12) reduces to $f'/f=\lambda_1$ and the constraints lead to the exponential distribution

$$f(x) = \begin{cases} \lambda_1 e^{-\lambda_1 x} & \lambda_1 = 1/(x_{(1)}-a) \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

3.2. Case 2 Solution

This problem is the same as for Case 1, except that a point of discontinuous $f(x)$ is allowed; $f(x)$ must be continuous. These considerations and the constraints lead to

$$f(x) = \alpha e^{-\alpha|x-x_{(1)}|} \quad (14)$$

Not enough is known to determine α.

3.3. Case 3 Solution

The algebra is much as in Case 1 giving

$$\lambda_1 = \frac{e^{-\lambda_1 a} - e^{-\lambda_1 b}}{x_{(1)}(e^{-\lambda_1 a} - e^{-\lambda_1 b}) - ae^{-\lambda_1 a} + be^{-\lambda_1 b}} \quad (15)$$

from which λ_1 can be determined numerically once a and b are given. The density in x is

$$f(x) = \begin{cases} \frac{\lambda_1 e^{-\lambda_1 x}}{e^{-\lambda_1 a} - e^{-\lambda_1 b}} & a < x \le b \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

The corresponding maximum entropy densities for the solutions λ_1 in (16) are shown in Figure 1.

3.4. Case 4 Solution

Adding the constraint associated with a known variance leads to the Gaussian density:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-x_{(1)})^2/2\sigma^2} \quad (17)$$

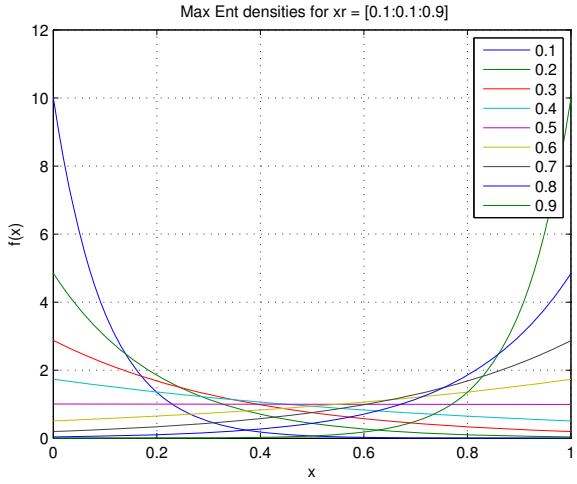


Fig. 1. Maximum Entropy densities corresponding to solutions λ_1 in (15), when the domain of x (and $x_{(r)}$) is normed to $(a, b] = (0, 1]$.

3.5. Asymptotic Solution

Because we are often interested in setting decision boundaries near extremes of the a-posteriori density, the asymptotes of the density $f(x)$ are highly important. Therefore, consider again (8) with no variance constraint ($\lambda_{R+2} = 0$) and for the cases of $a \to -\infty$ or $b \to \infty$ or both. In those cases we ask what is the differential equation (8) as $x \to \pm\infty$. Because

$$F^{j-1}(x)[1 - F(x)]^{R-j} \to \begin{cases} 1, \ j = R, \ x \to \infty \\ 1, \ j = 1, \ x \to \infty \\ 0, \ \text{Otherwise} \end{cases} \tag{18}$$

Equation (8) becomes

$$\begin{cases} f'(x) - \lambda_R f(x) = 0, \ \ x \to \infty \\ f'(x) - \lambda_1 f(x) = 0, \ \ x \to -\infty \end{cases} \tag{19}$$

The solutions for these cases are those of Cases 1 or 2. I.e., the max-ent $f(x)$ is asymptotically an exponential density as if there were just the single sample, either $x_{(1)}$ or $x_{(R)}$, with the given boundary conditions (a, b).

However, the Cases 1 and 2 give density solutions for a single sample. In the general case, $R > 1$, the asymptotic solution must incorporate this difference. In particular, the total probability of x beyond $x_{(R)}$ and below $x_{(1)}$; $P(x < x_{(1)})$ or $P(x > x_{(R)})$ must equal $1/(R + 1)$. Thus we must scale the asymptotic solution accordingly.

4. CONTIGUOUS INTERVAL SOLUTIONS

Because the solution for $R = 1$ is analytic, we can now use that solution to support a contiguous solution for $f(x)$ when $R > 1$, and the constraints are taken on the cdf as in (4) rather than from the integrals for the means as in (3). We now use the following notation, where $[\cdot]$ subscripted in $f_{[r]}$ indicates contiguous indexed density, not order statistic:

$$f(x) = \textstyle\sum_{r=0}^{R} f_{[r]}(x), \quad f_{[r]} = f(x); \quad a_r \le x < b_r$$
$$a_r = \begin{cases} a \text{ or } -\infty, \ r = 0 \\ x_{(r)}, \ 1 \le r \le R \end{cases} \quad b_r = \begin{cases} x_{(r+1)}, \ 1 \le r \le R - 1 \\ b \text{ or } \infty, \ r = R \end{cases} \tag{20}$$

The contiguous approach to the solution for the maximum entropy density is based on the following four points:

1. The samples are assumed to be the means of their respective order statistics.
2. Assumption 1 implies[1] that the density $f(x)$ is a sum of $R + 1$ mutually exclusive and contiguous densities $f_{[r]}(x)$, (see (20)) , each having equal area (probability) $1/(R + 1)$.
3. Excluding the extreme lower and upper (tail) densities (if they are not finite), each of the interior contiguous densities $f_{[r]}(x)$, $1 \le r \le R - 1$, spans the interval between the ordered samples $x_{(r)}$ and $x_{(r+1)}$.
4. The maximum entropy distribution $f_{[r]}(x)$ in a finite interval with no samples is flat; in that case

$$f_{[r]}(x) = 1/\left[(R + 1)(x_{(r+1)} - x_{(r)})\right] \tag{21}$$

If either or both the domain extremes are infinite, the tail densities for $-\infty < x \le x_{(1)}$ and/or $x_{(R)} < x < \infty$, must be found another way , as we will now suggest.

4.1. Asymptotic Solutions

Now we ask the question: what about infinite domain extremes? It is still true that the first and last intervals must each contain a total probability of $1/(R + 1)$. Thus from (21)

$$f_{[R]}(x) = \begin{cases} \gamma e^{-\beta(x-a)}, \ \ x_{(R)} < x \\ 0, \ \ \text{Otherwise} \end{cases} \tag{22}$$
$$\beta = 1/(x_{(R)} - a), \ \ \gamma = e/(R + 1)$$

A similar argument can be made for an exponential density over the segment closest to the least value, $x = a$.

4.2. Examples

We have purposely chosen the Rayleigh density because its large x asymptote drops off as $e^{-\eta x^2}$, much faster than our proposed simple exponential. Figure 2 shows the Rayleigh density and samples, the Max entropy approximation out to $b = 5$, and a Parzen smoothed histogram density estimate. The $x-$axis increment dx is set to $1/40$ of the minimum distance between data samples. The Parzen window size has been chosen to be the maximum of either 40 plot increments or the minimum number of plot increments between closest samples. The difference between Max entropy and Parzen-smoothed varies of course with choice of the Parzen window size. The chosen size here gives some smoothing while also retaining high resolution. Figure 3 gives an example with $R = 9$. Note in Figure 3 (b) that the cdf exactly fits the data, as it is forced to do. However, the Max-entropy contiguous approximate density gives order statistics that have 0.1091 standard normalized error from the true Raylegh order statistics and 0.1832 standard normalized error from the samples

[1] In support of this assumption, we note that ([1], Example 3.1.1) shows that the means of the order statistics divide the density into $R + 1$ equally likely regions. He also notes that only densities which have impulsive parts at the extremes cannot be addressed this way.

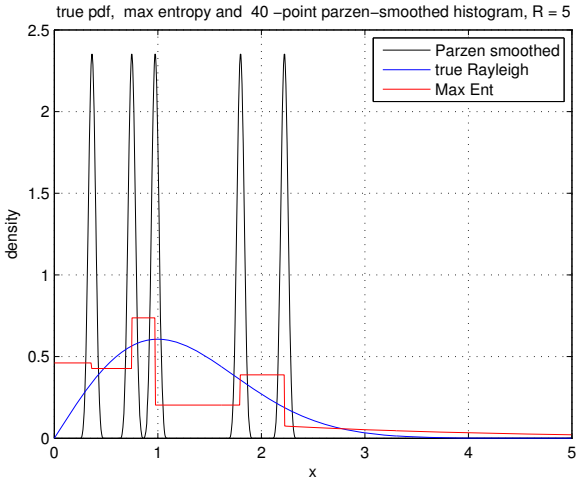


Fig. 2. Rayleigh density and 5 samples Max entropy approximation and a 40 axis-point Parzen smoothed histogram density estimate.

(assumed order statics). More data along these lines can be found for other densities and R values. The counterpart for $a = -\infty$ is straightforward, since it is symmetrical. The method can be applied to both tails simultaneously if both $a = -\infty$ and $b = \infty$.

5. CONCLUSIONS

A nonlinear differential equation has been derived for the maximum entropy density function estimate which would yield the R ordered samples, given in a histogram, assuming that each of those samples is the mean of its associated order statistic density. An analytical solution to the NDEQ (8) might yield a continuous density function, but the solution that meets the alternate constraints (9) may often be just as useful. It gives flat densities of equal probability weight between samples. If the domain is infinite, the tail domains cannot have a flat density, and exponential asymptotic solutions based on (18) are recommended.

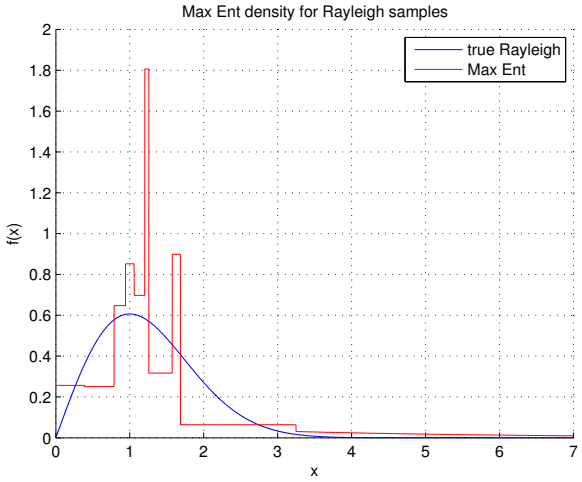
The NDEQ and analytical solutions we have provided may allow further insights to the problem. Examples are shown to be reasonable under the assumptions and compare favorably to Parzen smoothing window estimates. Other fixed window smoothers would have similar differences from the maxent.

6. REFERENCES

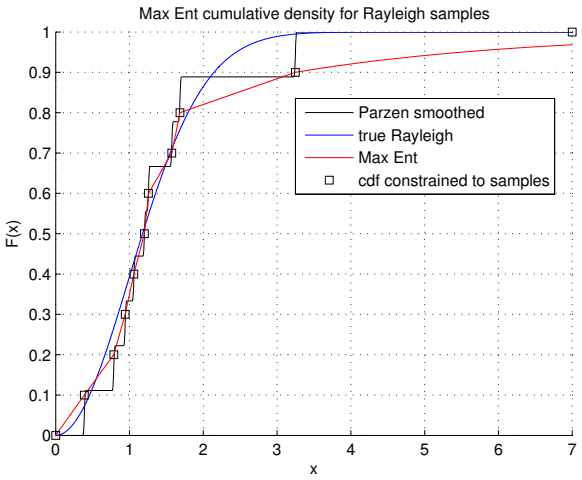
[1] H.A. David, *Order Statistics*, Wiley Series in Probability and Mathematics, 1980.

[2] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw Hill, 1984.

[3] I.M. Gelfand and S.V. Fomin, *Calculus of Variations*, Prentice Hall, 1963.



(a)



(b)

Fig. 3. Example application, $R = 9$, tail (essentially to $b = \infty$, with the exponential density in the last intervals $x_{(R)} < x < \infty$. The exponential tail has weight $1/(R+1)$. In (b) are shown the respective cumulative density functions and the sample point cdf values.