Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2009, Article ID 563281, 13 pages
doi:10.1155/2009/563281

Research Article

Recovering Decay Rates from Noisy Measurements with Maximum Entropy in the Mean

Henryk Gzyl and Enrique Ter Horst

Facultad de Ciencias, Instituto de Estudios Superiores de Administración IESA, Universidad Central de Venezuela, Caracas 1010, DF, Venezuela

Correspondence should be addressed to Henryk Gzyl, henryk.gzyl@iesa.edu.ve

Received 5 December 2008; Revised 26 February 2009; Accepted 30 March 2009

Recommended by Nikolaos Limnios

We present a new method, based on the method of maximum entropy in the mean, which builds upon the standard method of maximum entropy, to improve the parametric estimation of a decay rate when the measurements are corrupted by large level of noise and, more importantly, when the number of measurements is small. The method is developed in the context on a concrete example: that of estimation of the parameter in an exponential distribution. We show how to obtain an estimator with the noise filtered out, and using simulated data, we compare the performance of our method with the Bayesian and maximum likelihood approaches.

1. Introduction

Suppose that you want to measure the half-life of a decaying nucleus or the life-time of some elementary particle, or some other random variable modeled by an exponential distribution describing, say a decay time or the life time of a process. Assume as well that the noise in the measurement process can be modeled by a centered Gaussian random variable whose variance may be of the same order of magnitude as that of the decay rate to be measured. We assume that the variance δ of noise is known. To make things worse, assume that you can only collect very few measurements.

We want to emphasize, that the method developed is tailored to this one important model in applications, and on the other hand, to the fact that the samples have to be small and contaminated by observational noise (on top of their inherent randomness). And what the method provides in general, is a technique for filtering noise out.

Thus, if x_i denotes a realized value of the variable, one can only measure $y_i = x_i + e_i$, for $i = 1, 2, \ldots, n$, where n is a small number, say 2 or 3, and e_1 denotes the additive measurement noise. When one knows that the distribution of X is exponential, the parameter should be

estimated by $(1/n)\sum y_i = (1/n)\sum x_i + (1/n)\sum e_i$. In other words, assume that you know that the sample comes from a specific parametric distribution but is contaminated by additive noise, then your estimator of the relevant parameters is contaminated by the measurement noise. What to do? One possible approach is to apply to the small sample the standard statistical estimation procedures like maximum likelihood. But these work well when the sample size is larger than what concerns us here. In our particular example, the MLE is $(1/n)\sum y_i$ in which the noise may be important (unless n is large.) Thus apart from the issues arising from the smallness of the sample, we have the issue of the presence of the observational noise. We should direct the reader to the work of Rousseeuw and Verboven [1], in which issues relating to estimation in small samples are discussed.

Still another possibility, the one we want to explore here, is to apply a maxentropic filtering method, to estimate both the unknown variable and the noise level. For this we recast the problem as a typical inverse problem consisting of solving for $\mathbf{x}$ in

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \quad \mathbf{x} \in \mathbf{K}, \tag{1.1}$$

where $\mathbf{K}$ is a convex set in $\mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^k$ and for some d and k, and $\mathbf{A}$ is an $k \times d$-matrix which depends on how we rephrase the our problem. We could, for example, consider the following problem. Find $\widehat{x} \in [0, \infty)$ such that

$$\widehat{y} = \widehat{x} + \widehat{e}. \tag{1.2}$$

In our case $\mathbf{K} = [0, \infty)$, and we set $\widehat{y} = (1/n)\Sigma_j\, y_j$. Or we could consider a collection of n such problems, one for every measurement, and then proceed to carry on the estimation. Once we have solved the generic problem (1.1), the variations on the theme are easy to write down. What is important to keep in mind here, is that the output of the method is a filtered estimator $\widehat{x}^*$ of $\widehat{x}$, which itself is an estimator of the unknown parameter. The novelty then is to filter out the noise in (1.2).

The method of maximum entropy in the mean (MEM) is rather well suited for solving problems like (1.1). See Navaza's [2] for an early development and Dacunha-Castelle and Gamboa [3] for full mathematical treatment. We shall briefly review what the method is about and then apply it to obtain an estimator $\widehat{x}$ from (1.2). In Section 3 we obtain the maxentropic estimator, and in Section 4 we examine some of its properties, in particular we examine what the results would be if either the noise level were small or the number of measurements were large. We devote Section 4 to some simulations in which the method is compared with a Bayesian and a maximum likelihood approaches.

2. The Basics of MEM

MEM is a technique for transforming a possibly ill-posed, linear problem with convex constraints into a simpler (usually unconstrained) but non-linear minimization problem. The number of variables in the auxiliary problem is equal to the number of equations in the original problem, k in the case of example 1. To carry out the transformation one thinks of the $\mathbf{x}$ there as the expected value of a random variable $\mathbf{X}$ with respect to some measure $\mathbb{P}$ to be determined. The basic datum is a sample space $(\Omega_s, \mathcal{F}_s)$ on which $\mathbf{X}$ is to be defined. In our setup the natural choice is to take $\Omega_s = \mathbf{K}$, $\mathcal{F}_s = \mathcal{B}(\mathbf{K})$, the Borel subsets of $\mathbf{K}$, and $\mathbf{X} = \mathbf{id}_{\mathbf{K}}$ the

identity map. Similarly, we think of $\mathbf{e}$ as the expected value of a random variable $\mathbf{V}$ taking values in $\mathbb{R}^k$. The natural choice of sample space here is $\Omega_n = \mathbb{R}^k$ and $\mathcal{F}_n = \mathcal{B}(\mathbb{R}^k)$ the Borel subsets.

To continue we need to select a prior measures $dQ_s(\xi)$ and $dQ_n(v)$ on $(\Omega_s, \mathcal{F}_s)$ and $(\Omega_n, \mathcal{F}_n)$. The only restriction that we impose on them is that the closure of the convex hull of both $\operatorname{supp}(Q_s)$ (resp., of $\operatorname{supp}(Q_n)$) is $\mathbf{K}$ (resp., $\mathbb{R}^k$). These prior measures embody knowledge that we may have about $\mathbf{x}$ and $\mathbf{e}$ but are not priors in the Bayesian sense. Actually, the model for the noise component describes the characteristics of the measurement device or process, and it is a datum. The two pieces are put together setting $\Omega = \Omega_s \times \Omega_n$; $\mathcal{F} = \mathcal{F}_s \otimes \mathcal{F}_n$, and $dQ(\xi, v) = dQ_s(\xi)dQ_n(v)$. And to get going we define the class

$$\mathbb{P} = \{P \mid P \ll Q;\ AE_P[\mathbf{X}] + E_P[\mathbf{V}] = \mathbf{y}\}. \tag{2.1}$$

Note that for any $P \in \mathbb{P}$ having a strictly positive density $\rho = dP/dQ$, then $E_P[\mathbf{X}] \in \operatorname{int}(\mathbf{K})$. For this standard result in analysis check in Rudin's book [4]. The procedure to explicitly produce such P's is known as the maximum entropy method. The first step of which is to assume that $\mathbb{P} \neq \emptyset$, which amounts to say that our inverse problem (1.1) has a solution, and we define

$$S_Q : \mathbb{P} \longrightarrow [-\infty, \infty) \tag{2.2}$$

by the rule

$$S_Q(P) = -\int_\Omega \ln\left(\frac{dP}{dQ}\right) dP \tag{2.3}$$

whenever the function $\ln(dP/dQ)$ is P-integrable and $S_Q(P) = -\infty$ otherwise. This entropy functional is concave on the convex set $\mathbb{P}$. To guess the form of the density of the measure P^* that maximizes S_Q is to consider the class of exponential measures on Ω defined by

$$dP_{\boldsymbol{\lambda}} = \frac{e^{-\langle \boldsymbol{\lambda}, \mathbf{A}\xi\rangle - \langle \boldsymbol{\lambda}, v\rangle}}{Z(\boldsymbol{\lambda})} dQ, \tag{2.4}$$

where the normalization factor is

$$Z(\boldsymbol{\lambda}) = E_Q\left[e^{-\langle \boldsymbol{\lambda}, \mathbf{A}\xi\rangle - \langle \boldsymbol{\lambda}, v\rangle}\right]. \tag{2.5}$$

Here $\boldsymbol{\lambda} \in \mathbb{R}^k$, if we define the dual entropy function

$$\Sigma(\boldsymbol{\lambda}) : \mathfrak{D}(Q) \longrightarrow (-\infty, \infty] \tag{2.6}$$

by the rule

$$\Sigma(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \mathbf{y} \rangle \tag{2.7}$$

or $\Sigma(\boldsymbol{\lambda}) = \infty$ whenever $\boldsymbol{\lambda} \notin \mathfrak{D}(Q) \equiv \{\mu \in \mathbb{R}^k \mid Z(\mu) < \infty\}$.

It is easy to prove that, $\Sigma(\boldsymbol{\lambda}) \geq S_Q(P)$ for any $\boldsymbol{\lambda} \in \mathfrak{D}(Q)$, and any $P \in \mathbb{P}$. Thus if we were able to find a $\boldsymbol{\lambda}^* \in \mathfrak{D}(Q)$ such that $P_{\boldsymbol{\lambda}^*} \in \mathbb{P}$, we are done. To find such a $\boldsymbol{\lambda}^*$ it suffices to minimize (the convex function) $\Sigma(\boldsymbol{\lambda})$ over (the convex set) $\mathfrak{D}(Q)$. We leave for the reader to verify that if the minimum is reached in the interior of $\mathfrak{D}(Q)$, then $P_{\boldsymbol{\lambda}^*} \in \mathbb{P}$. We direct the reader to [4, 5] for all about this, and much more. For a review of the use of maximum entropy on the mean for solving linear inverse problems, the reader might want to look at [6].

3. Entropic Estimators

Let us now turn our attention to (1.2). Since our estimator is a sample mean of an exponential (of unknown parameter), it is natural to assume, for the method described in Section 2, that the prior Q_s for $\mathbf{X}$ is a $\Gamma(n, \alpha/n)$, where $\alpha > 0$ is our best (or prior) guess of the unknown parameter. Here in after we propose a criterion for the best choice of α. Similarly, we shall chose Q_n to be the distribution of a $N(0, \delta^2/n)$ random variable as prior for the noise component. Things are rather easy under these assumptions. To begin with, note that

$$Z(\lambda) = \frac{e^{\lambda^2\delta^2/2n}}{(\lambda/n\alpha + 1)^n}, \tag{3.1}$$

and the typical member $dP_\lambda(\xi, v)$ of the exponential family is now

$$dP_\lambda(\xi, v) = (\lambda + n\alpha)^n \frac{\xi^{n-1}}{\Gamma(n)} e^{-(\lambda+n\alpha)\xi} \frac{e^{-(v+\delta^2\lambda/n)^2(n/2\delta^2)}}{(2\pi\delta^2/n)^{1/2}} d\xi\, dv. \tag{3.2}$$

It is also easy to verify that the dual entropy function $\Sigma(\lambda)$ is given by

$$\Sigma(\lambda) = \frac{\lambda^2\delta^2}{2n} - n\ln\left(\frac{\lambda}{n\alpha} + 1\right) + \lambda\widehat{y}, \tag{3.3}$$

the minimum value of which is reached at λ^* satisfying

$$\frac{\lambda^*\delta^2}{n} - \frac{1/\alpha}{\lambda^*/n\alpha + 1} + \widehat{y} = 0, \tag{3.4}$$

and, discarding one of the solutions (because it leads to a negative estimator of a positive quantity), we are left with

$$\frac{\lambda^*}{n\alpha} = \frac{1}{2}\left(-\left(1 + \frac{\widehat{y}}{\alpha\delta^2}\right) + \left(\left(1 - \frac{\widehat{y}}{\alpha\delta^2}\right)^2 + \frac{4}{\alpha^2\delta^2}\right)^{1/2}\right), \tag{3.5}$$

from which we obtain that

$$\frac{\lambda^*}{n\alpha} + 1 = \frac{1}{2}\left(\left(1 - \frac{\widehat{y}}{\alpha\delta^2}\right) + \left(\left(1 - \frac{\widehat{y}}{\alpha\delta^2}\right)^2 + \frac{4}{\alpha^2\delta^2}\right)^{1/2}\right) \tag{3.6}$$

as well as

$$\widehat{x}^* = E_{P(\lambda^*)}[\mathbf{X}] = \frac{n}{(\lambda^* + n\alpha)} = \left[\frac{\alpha}{2}\left(\left(1 - \frac{\widehat{y}}{\alpha\delta}\right) + \sqrt{\left(1 - \frac{\widehat{y}}{\alpha\delta}\right)^2 + \frac{4}{\alpha^2\delta^2}}\right)^{1/2}\right]^{-1},$$
$$\widehat{e}^* = E_{P(\lambda^*)}[\mathbf{V}] = -\frac{\delta^2\lambda^*}{n}. \tag{3.7}$$

Comment 1. Clearly, from (3.4) it follows that $\widehat{y} = \widehat{x}^* + \widehat{e}^*$. Thus it makes sense to think of $\widehat{x}^*$ as the estimator with the noise filtered out, and to think of $\widehat{e}^*$ as the residual noise.

4. Properties of $\widehat{x}^*$

Let us now spell out some of the notation underlying the probabilistic model behind (1.1). We shall assume that the x_i and the e_i in the first section are values of random variables X^i and ε^i defined on a sample space $(\mathbb{W}, \mathcal{W})$. For each $\theta > 0$, we assume to be given a probability law $P(\theta)$ on $(\mathbb{W}, \mathcal{W})$, with respect to which the sequences $\{X^k \mid k = 1, 2, \ldots\}$ and $\{\varepsilon^k \mid k = 1, 2, \ldots\}$ are both i.i.d. and independent of each other, and with respect to $P(\theta)$, $X^k \sim \exp(\theta)$ and $\varepsilon^k \sim N(0, \delta^2)$. That is, we consider the underlying model for the noise as our prior model for it. Minimal consistency is all right. From (3.6) and (3.7) , the following basic results are easy to obtain.

Lemma 4.1. *If we take $\alpha = 1/\widehat{y}$, then $\lambda^* = 0$, $\widehat{x}^* = \widehat{y}$, and $\widehat{e}^* = 0$.*

Comment 2. Actually it is easy to verify that the solution to $\widehat{x}^*(\alpha) = 1/\alpha$ is $\alpha = 1/\widehat{y}$.

To examine the case in which large data sets were available, let us add a superscript n and write $\widehat{y}(n)$ to emphasize the size of the sample. If $\widehat{x}^{(n)}$ denotes the arithmetic mean of an i.i.d. sequence of random variables having $\exp(\theta)$ as common law, it will follow from the LLN the following.

Lemma 4.2. *As $n \to \infty$ then*

$$\left(\widehat{x}^{(n)}\right)^* \longrightarrow \widetilde{x}(\alpha) \equiv \left[\frac{\alpha}{2}\left(\left(1 - \frac{\theta}{\alpha\delta^2}\right) + \left(\left(1 - \frac{\theta}{\alpha\delta^2}\right)^2 + \frac{4}{\alpha^2\delta^2}\right)^{1/2}\right)\right]^{(-1)}. \tag{4.1}$$

Proof. Start from (3.7), invoke the LLN to conclude that $\widehat{y}(n)$ tends to θ and obtain (4.1). □

Corollary 4.3. *The true parameter is the solution of* $\widetilde{x}(\alpha) - 1/\alpha = 0$.

Proof. Just look at the right hand-side of (4.1) to conclude that $\widetilde{x}(1/\theta) = \theta$. □

Comment 3. What this asserts is that when the number of measurements is large, to find the right value of the parameter it suffices to solve $\widetilde{x}(\alpha) - 1/\alpha = 0$. And when the noise level goes to zero, we have the following.

Lemma 4.4. *With the notations introduced above,* $\widehat{x}^* \to \widehat{y}$ *as* $\delta \to 0$.

Proof. When $\delta \to 0$, the $dQ_n(v) \to \epsilon_0(dv)$, the Dirac point mass at 0. In this case, we just set $\delta = 0$ in (3.4) and the conclusion follows. □

When we choose $\alpha = 1/\widehat{y}$, the estimator $\widehat{x}^*$ happens to be unbiased.

Lemma 4.5. *Let* θ *denote the true but unknown parameter of the exponential, and* $P_\theta(dy)$ *have density*

$$f_\theta(y) = \int_{-\infty}^{y} \theta^n (y-s)^{n-1} \frac{e^{-\theta(y-s)} e^{-s^2/2\delta^2}}{\Gamma(n)\sqrt{2\pi}\delta} ds \tag{4.2}$$

for $y > 0$ *and* 0 *otherwise. With the notations introduced above, one has* $E_{P(\theta)}[(\widehat{x}^{(n)})^*] = 1/\theta$ *whenever the prior* α *for the maxent is the sample mean* $\widehat{y}$.

Proof. It drops out easily from Lemma 4.1, from (1.2), and the fact that the joint density f_θ of $\widehat{y}$ is a convolution. □

But the right choice of the parameter α is still a pending issue. To settle it we consider once more the identity $|\widehat{y} - \widehat{x}^*| = |\widehat{e}^*|$. In our particular case we shall see that $\alpha = 0$ minimizes the right-hand side of the previous identity. Thus, we propose to choose α to minimize the residual or reconstruction error.

Lemma 4.6. *With the same notations as above,* $\widehat{e}^*$ *happens to be a monotone function of* α *and* $\widehat{e}^*(\alpha = 0) = (1/2)(\widehat{y} - \sqrt{\widehat{y}^2 + 4\delta^2})$ *and* $\widehat{e}^*(\alpha \to \infty) = \widehat{y}$. *In the first case* $\widehat{x}^*(\alpha = 0) = (1/2)(\widehat{y} + \sqrt{\widehat{y}^2 + 4\delta^2})$, *whereas in the second* $\widehat{x}^*(\alpha \to \infty) = 0$.

Proof. Recall from the first lemma that when $\alpha\widehat{y} = 1$, then $\widehat{e}^* = 0$. A simple algebraic manipulation shows that when $\alpha\widehat{y} > 1$ then $\widehat{e}^* > 0$, and that when $\alpha\widehat{y} < 1$ then $\widehat{e}^* < 0$. To compute the limit of $\widehat{e}^*$ as $\alpha \to \infty$, note that for large α we can neglect the term $4/\delta^2$ under the square root sign, and then the result drops out. It is also easy to check the positivity of the derivative of $\widehat{e}^*$ with respect to α. Also clearly $|\widehat{e}^*(0)| < |\widehat{e}^*(\infty)|$. □

To sum up, with the choice $\alpha = 0$, the entropic estimator and residual error are

$$\widehat{x}^*(0) = \frac{1}{2}\left(\widehat{y} + \sqrt{\widehat{y}^2 + 4\delta^2}\right), \qquad \widehat{e}^*(0) = \frac{1}{2}\left(\widehat{y} - \sqrt{\widehat{y}^2 + 4\delta^2}\right). \tag{4.3}$$

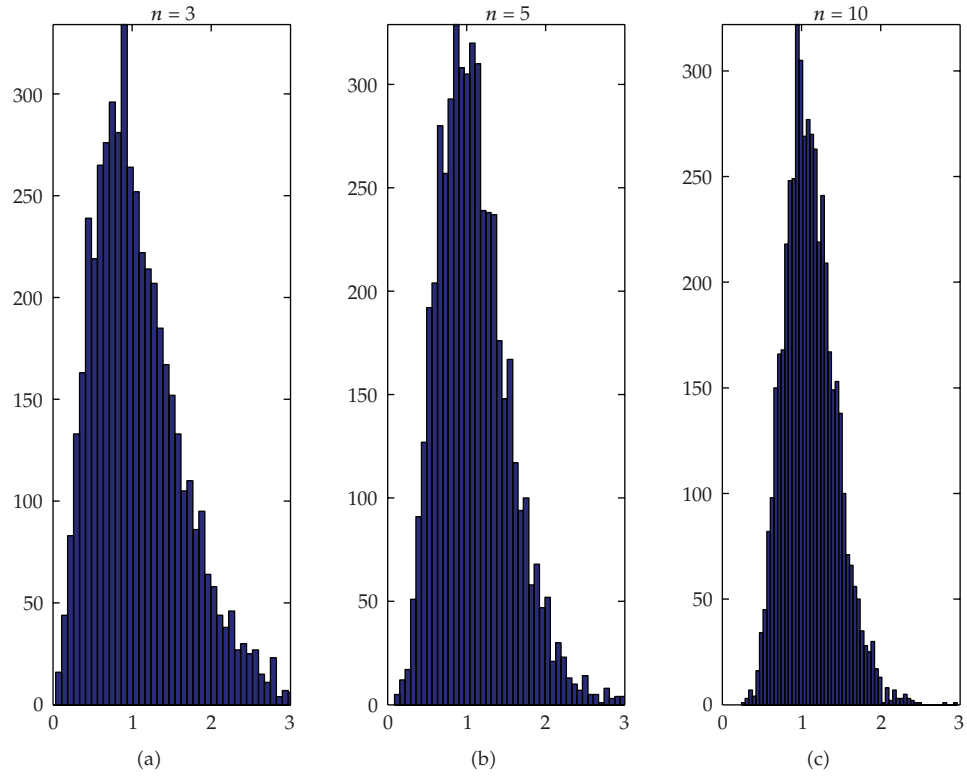

Figure 1: The simple MLE for different n.

5. Simulation and Comparison with the Bayesian and Maximum Likelihood Approaches

In this section we compare the proposed maximum entropy in the mean procedure with the Bayesian and maximum likelihood estimation procedures. We do that by simulating data and carrying out the three procedures and plotting the histograms of the corresponding estimators. First, we generate histograms that describe the statistical nature of $\widehat{x}^*$ as a function of the parameter α. For that we generate a data set of 5000 samples of 1, 3, 5, and 10 measurements, and for each of them we obtain $\widehat{x}^*$ from (4.3). Also, for each data point we apply both a Bayesian estimation method, a simple-minded maximum likelihood estimation and a maximum likelihood method and plot the resulting histograms.

5.1. The Simple-Minded MLE

This consists of an application of the MLE method as if there was no measurement noise. We carried out this for the sake of comparison, to verify that when the sample size becomes larger, the effect of the measurement noise is washed away on the average. The plot of the results for $n = 1$ is too scattered, and we do not display it. The result of the simulations is displayed in Figure 1.

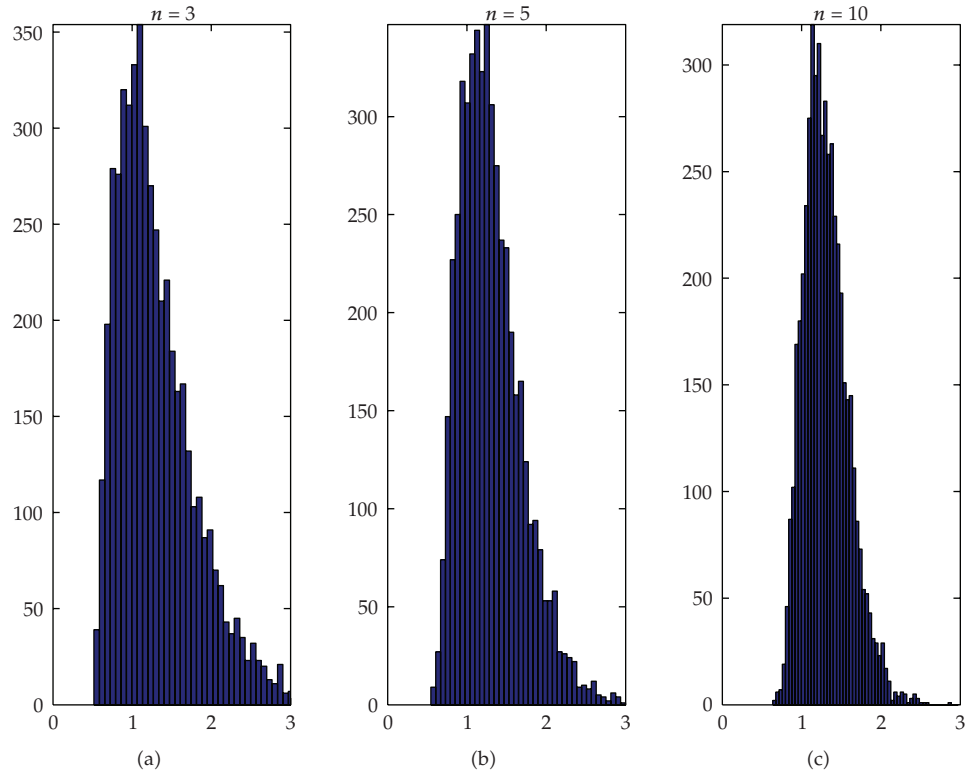

Figure 2: Histogram for $E(x)$ with MEM for different n.

5.2. The Maxentropic Estimator

The simulated data process goes as follows. For $n = 3$ the data points y_1, y_2, y_3 are obtained in the following way.

(i) Simulate a value for x_i from an exponential distribution with parameter $\theta(= 1)$.

(ii) Simulate a value for e_i from a normal distribution $N(0, \delta^2 = 0.25)$.

(iii) Sum x_i with e_i to get y_i, if $y_i < 0$, repeat first two steps until $y_i > 0$.

(iv) Do this for $i = 1, 2, 3$.

(v) Compute the maximum entropy estimator given by (4.3).

We then display the resulting histogram in Figure 2.

5.3. The Bayesian Estimator

In this section we derive the algorithm for a Bayesian inference of the model given by $y_i = x + e_i$, for $i = 1, 2, \ldots, n$. The classical likelihood estimator of x is given by $\widehat{y} = (1/n) \sum_{i=1}^{n} y_i$. As we know that the unknown mean x has an exponential probability distribution with parameter

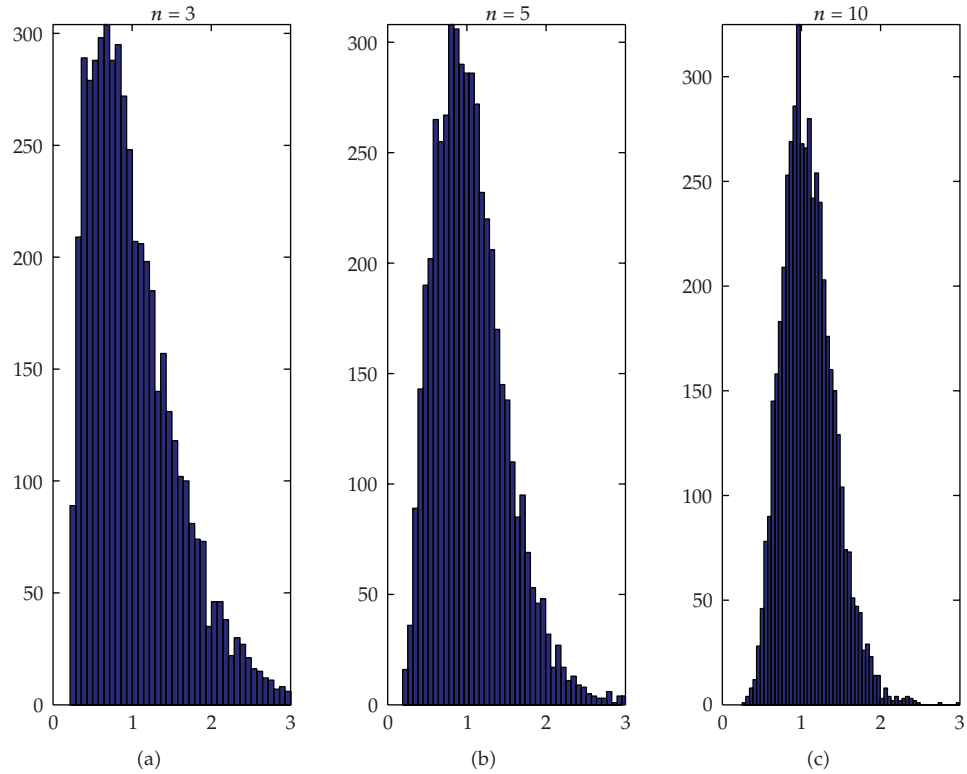

Figure 3: Histogram for $E(x)$ with Bayes Method for different n.

$\theta(x \sim \mathbb{E}(\theta))$, therefore the joint density of the y_i and x is proportional to

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\delta^2}} \exp\left\{-\frac{(y_i - x)^2}{2\delta^2}\right\} \theta \exp(-\theta x)\pi(\theta), \tag{5.1}$$

where $\theta \exp(-\theta x)$ is the density of the unknown mean x and where $\pi(\theta) \propto \theta^{-1}$ is the Jeffrey's noninformative prior distribution for the parameter θ [5].

In order to derive the Bayesian estimator, we need to get the posterior probability distribution for θ, which we do with the following Gibbs sampling scheme [7].

(i) Draw $x \sim N(\widehat{y} - \theta\delta^2/n, \delta^2/n)1_{x>0}$.

(ii) Draw $\theta \sim \mathbb{E}(x)$.

Repeat this algorithm many times in order to obtain a large sample from the posterior distribution of θ in order to obtain the posterior distribution of $E(x) = 1/\theta$. For our application, we simulate data with $\theta = 1$, which gives an expected value for x equal to $E(x) = 1$.

We get the histogram displayed in Figure 3 for the estimations of $E(x)$ after 5000 iterations when simulating data for $\theta = 1$.

5.4. The Maximum Likelihood Estimator

The problem of obtaining a ML estimator is complicated in this setup because data points are distributed like

$$f_\theta(t) = \int_{-\infty}^{t} \frac{\theta e^{-\theta(t-s)} e^{-s^2/2\delta^2} ds}{\sqrt{(2\pi\delta^2)}},$$
$$f_\theta(t) = \theta e^{-\theta t + (\theta\delta)^2/2} \mathbb{P}(S < t), \tag{5.2}$$

where $S \sim N(\theta\delta^2, \delta^2)$. Therefore, after observing t_1, t_2, and t_3, we get the following likelihood that we maximize numerically:

$$\theta^3 e^{-\theta \sum_{i=1}^{3} t_i + 3(\theta\delta)^2/2} \prod_{i=1}^{3} \mathbb{P}(S < t_i). \tag{5.3}$$

If we attempted to obtain the ML estimator analytically, we would need to solve

$$\frac{n}{\theta} - \sum_{j=1}^{n} \frac{\int_{-\infty}^{t_j} \theta e^{-\theta(t_j-s)} e^{-s^2/2\delta^2} ds/\sqrt{(2\pi\delta^2)}}{\int_{-\infty}^{t_j} \theta e^{-\theta(t_j-s)} e^{-s^2/2\delta^2} ds/\sqrt{(2\pi\delta^2)}} = 0. \tag{5.4}$$

Notice that as $\delta \to 0$ this equation tends to $(n/\theta) - \sum_{j=1}^{n} t_j = 0$ as expected. We can move forward a bit, and integrate by parts each numerator, and after some calculations we arrive to

$$\frac{n}{\theta} - \sum_{j=1}^{n} t_j + n\delta^2\theta - \sum_{j=1}^{n} \frac{\delta e^{-t_j^2/2\delta^2}}{\int_{-\infty}^{t_j} \theta e^{-\theta(t_j-s)} e^{-s^2/2\delta^2} ds/\sqrt{(2\pi\delta^2)}} = 0. \tag{5.5}$$

Trying to solve this equation in θ is rather hopeless. That is the reason why we carried on a numerical maximization procedure on (5.3). To understand what happens when the noise is small, we drop the last term in the last equation and we are left with

$$\frac{n}{\theta} - \sum_{j=1}^{n} t_j + n\delta^2\theta \tag{5.6}$$

the solution of which is

$$\frac{1}{\theta}^{*} = \frac{1}{2}\left(\widehat{y} + \sqrt{\widehat{y}^2 - 4\delta^2}\right) \tag{5.7}$$

or $\theta^* = 2(\widehat{y} + \sqrt{\widehat{y}^2 - 4\delta^2})^{-1}$, and we see that the effect of noise is to increase the ML estimator. In Figure 4 we plot the histogram of $(1/\theta)^*$ obtained by numerically maximizing (5.3) for each simulated data point.

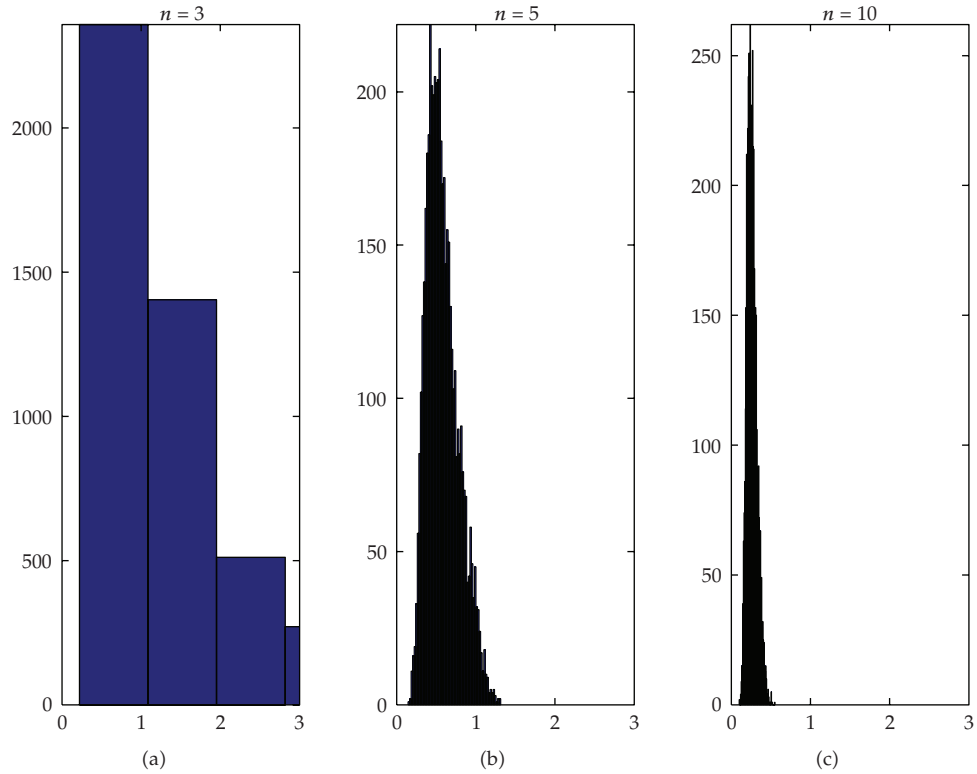

Figure 4: Histogram for $E(x)$ with the MLE for different n.

Table 1: Means and standard deviations for different methods when $n = 3$.

Method	Mean	Standard deviation
Maximum entropy	1.3112	0.5193
Bayesian	1.0271	0.5701
Maximum likelihood	1.8085	2.4630
Sample average	1.0928	0.5904

Table 2: Means and standard deviations for different methods when $n = 5$.

Method	Mean	Standard deviation
Maximum entropy	1.3090	0.4034
Bayesian	1.0532	0.4561
Maximum likelihood	0.5817	0.2016
Sample average	1.1009	0.4596

The results are summarized in Tables 1, 2, and 3.

When simulating data for $\theta = 1$, the MEM, Maximum likelihood and Bayesian histograms are all skewed to the right and yield a mean under the three simulated histograms

Table 3: Means and standard deviations for different methods when $n = 10$.

Method	Mean	Standard deviation
Maximum entropy	1.3093	0.2825
Bayesian	1.0846	0.3239
Maximum likelihood	0.2614	0.0627
Sample average	1.1097	0.3230

close to 1. As shown in the table, compiled for $n = 3$, the MEM method yields a sample mean of 1.3112 with a sample standard deviation of 0.5193, the Bayesian yields a sample mean equal to 1.0271 and sample standard deviation of 0.5701, and the Maximum Likelihood method yields a sample mean of 1.808 with a sample standard deviation of 2.463. All the three methods produce right skewed histograms for $E(x)$. The MEM and Bayesian method provide better and similar results and more accurate than the Maximum Likelihood method, but keep in mind that we carried out an approximate calculation.

Note as well that for $\ln \sim 10$ the variability in the (approximate) MLE decreases, but it is far away from the true value. This could be due to the numerical approximation that we used, whereas for $n \sim 25$ the estimator improves considerably. We owe this observation to one of our referee's, who pointed out that for very small samples, the MEM estimator outperforms the MLE estimator because it is biased, and that this advantage disappears as n becomes large.

6. Concluding Remarks

We exhibited the usefulness of the method of maximum entropy of the mean for dealing with estimation problems in which the samples are small and contaminated by noise, thus adding and extra source of randomness which has to be filtered out. The problem we chose, while being real, it is a problem in which the Lagrange multiplier λ could be estimated analytically and the properties of the resulting estimators could be studied analytically as well. This possibility appears as well when the observed signal is Gaussian. In general, to obtain the filtered estimator, one has to determine the Lagrange multipliers numerically.

On one hand, MEM backs up the intuitive belief, according to which, if the y_i are all the data that you have, it is all right to compute your estimator of the mean for $\alpha = 0$. The MEM and Bayesian methods yield closer results to the true parameter value than the maximum likelihood estimator for small number of measurements.

On the other hand, and this depends on your choice of priors, MEM provides us with a way of modifying those priors, and obtain representations like $\widehat{y} = \widehat{x}^* + \widehat{e}^*$; where of course $\widehat{x}^* = \widehat{x}^*(\widehat{y})$. What we saw above, is that there is a choice of prior distributions such that $\widehat{x}^* = \widehat{y}$ and $\widehat{e}^* = 0$.

The important thing is that this is actually true regardless of what the "true" probability describing the x_i is.

Aknowledgment

The authors want to than two referees for their comments and suggestions, which improved their presentation.

References

[1] R. Rousseeuw and S. Verboven, "Robust estimation in small samples," *Computational Statistics & Data Analysis*, vol. 40, no. 4, pp. 741–758, 2002.
[2] J. Navaza, "The use of non-local constraints in maximum-entropy electron density reconstruction," *Acta Crystallographica. Section A*, vol. 42, no. 4, pp. 212–223, 1986.
[3] D. Dacunha-Castelle and F. Gamboa, "Maximum d'entropie et probleme des moments," *Annales de l'Institut Henri Poincaré*, vol. 26, pp. 567–596, 1990.
[4] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, NY, USA, 1973.
[5] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, Berlin, Germany, 2nd edition, 1985.
[6] H. Gzyl, "Ill-posed linear inverse problems and maximum entropy in the mean," *Acta Científica Venezolana*, vol. 53, no. 2, pp. 74–93, 2002.
[7] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer, Berlin, Germany, 2005.