Research Article

Sample Dependence in the Maximum Entropy Solution to the Generalized Moment Problem

Henryk Gzyl

Centro de Finanzas, IESA, Caracas, Capital District, Venezuela

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

Received 7 July 2016; Accepted 11 October 2016

Academic Editor: Ramón M. Rodríguez-Dagnino

Copyright © 2016 Henryk Gzyl. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The method of maximum entropy is quite a powerful tool to solve the generalized moment problem, which consists in determining the probability density of a random variable X from the knowledge of the expected values of a few functions of the variable. In actual practice, such expected values are determined from empirical samples, leaving open the question of the dependence of the solution upon the sample. It is the purpose of this note to take a few steps towards the analysis of such dependence.

1. Introduction and Preliminaries

To state what the generalized moment problem is about, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(S, \mathcal{B}, m)$ be a measure space, with m a finite or *sigma*-finite measure. Let X be an S -valued random variable, such that its distribution has a density with respect to the measure m . The generalized moment problem consists in determining a density $f(x)$ such that

$$\int_S h_k(x) f(x) m(dx) = d_k \quad \text{for } k = 0, 1, \dots, M, \quad (1)$$

where h_k is a collection of measurable functions $h_k : S \rightarrow \mathbb{R}$ and d_k are given real numbers, and we set $h_0 \equiv 1$ and $d_0 = 1$ to take care of the natural requirement on f . A typical example is the following. X stands for a positive random variable (a stopping time, or perhaps being total risk severity) and we can compute $E[\exp(-\alpha_k X)] = d_k$ by some Monte Carlo procedure at a finite number of points α_k . The problem that we need to solve amounts to inverting the Laplace transform from such finite collection of values of the transform parameter α .

Actually this last problem is of much interest in the banking and insurance industries, where the density is necessary to compute risk premia and regulatory capital of various types, samples may be small, and the estimation of d_k reflects that. We direct the reader to Gomez-Goncalves et al. [1], where

this issue was addressed in the context of risk modeling and Laplace transform inversion.

Let us denote by f^* the solution to problem (1) obtained by the maximum entropy method as explained in Section 3 below when the moments in the right hand side are known exactly. As in many situations the moments are to be estimated empirically as detailed in (2) below, it is to be expected that the maxentropic distribution obtained will depend on the sample $X_1, \dots, X_N$ used to compute $\hat{d}_k$. We shall denote this maxentropic density by $\hat{f}_N^*$ to emphasize its dependence on the empirical moments $\hat{d}_k$.

The problem that we address in this note is the convergence on the $\hat{f}_N^*$ to f^* as well as that of the oscillations of mean values computed with $\hat{f}_N^*$ about the oscillations of mean values computed with respect to f^* .

When we have a sample $X_1, \dots, X_N$, the empirical generalized moments (the sample averages) are given by

$$\hat{d}_k = \frac{1}{N} \sum_{n=1}^N h_k(X_n) \quad (2)$$

which fluctuate around the exact moments d_k ; we thus expect the output of the maxentropic procedure to somehow reflect this variability.

For that, in the next section we recall in a (short) historical survey the notion of entropy of a density, and in the following

section we present the basics of the maximum entropy method.

In Section 4 we take up the main theme of this work: the variability of $\hat{f}_N^*$ that comes in through $\hat{d}_k$. There we prove that $\hat{f}_N^*$ converges pointwise and in L_1 to the maxentropic density f^* obtained from the exact data, and we shall examine how $\hat{f}_N^*$ deviates from f^* in terms of the difference between the true and the estimated (sample) moments. We examine as well the deviations of expected values like $\int g(x) \hat{f}_N^*(x) dm(x)$ from $\int g(x) f^*(x) dm(x)$. That the density reconstruction from empirical moments has to depend on the sample seems to be intuitive, but neither the behavior of the maxentropic density as the sample size increases nor the fluctuations of the expected values with the densities $\hat{f}_N^*$ seem not to have been studied before.

2. The Entropy of a Density

As there seem to be several notions of entropy, it is the aim of this section to point out that they are all variations on the theme of one single definition. Let us begin by spelling out what it is that we call the entropy of a density. Let P be a measure on $(\Omega, \mathcal{B})$. Suppose that $P \ll m$ and let f denote its density. The entropy $H_m(P)$ (when we want to emphasize the density we write $H_m(f)$) is defined by

$$H_m(P) = - \int_{\Omega} f \ln(f) dm \quad (3)$$

whenever $\ln(f)$ is P -integrable, or $-\infty$ if not. $H_m(P)$ is called the entropy of P (with respect to m) and $H_m(f)$ is called the entropy of f . Actually, we can also define $H_m(f)$ for $f \in \{g \in L_1(dm), g \geq 0\}$ as follows. When P is not necessarily a probability measure having a density f with respect to m , (3) is to be modified as follows:

$$H_m(P) (= H_m(f)) = - \int_{\Omega} f \ln(f) dm + \left(\int_{\Omega} f dm - 1 \right). \quad (4)$$

When m is a probability measure, denote it by Q , and both P and Q are equivalent to a measure n , with densities given, respectively, by $f = dP/dn$ and $g = dQ/dn$; then (3) can be written as

$$\begin{aligned} H_n(P, Q) &= - \int_{\Omega} \left(\frac{f}{g} \right) \ln \left(\frac{f}{g} \right) g dn \\ &= - \int_{\Omega} \ln \left(\frac{f}{g} \right) f dn = - \int_{\Omega} \ln \left(\frac{f}{g} \right) dP \end{aligned} \quad (5)$$

and we call it the entropy of f with respect to g .

Comment. For the applications that we shall be dealing with, Ω will stand for a closed, convex subset in some $\mathbb{R}^n$, and m will be the usual Lebesgue measure. We also mention that when m is a discrete measure, then the integrals would become sums.

The expression (3) seems to have made its first appearance in the work of Boltzmann in the last quarter of the XIXth

century. There it was defined in $\mathbb{R}^3 \times \mathbb{R}^3$, where $f(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}$ was to be interpreted as the number of particles with position within $d\mathbf{x}$ and velocity within $d\mathbf{v}$. The function happened to be a Lyapunov functional for the dynamics that Boltzmann proposed for the evolution of a gas, which grew as the gas evolved towards equilibrium. Not much later Gibbs used the same function, but now defined on $\mathbb{R}^{6N} \times \mathbb{R}^{6N}$, whose points $(\mathbf{q}, \mathbf{p})$ denote the joint position and momenta of a system of N particles. This time $dm = d\mathbf{q} d\mathbf{p}$, and $f(\mathbf{q}, \mathbf{p}) d\mathbf{q} d\mathbf{p}$ is the probability of finding the system within the specified “volume” element. Motivated by earlier work in thermodynamics, it was postulated that in equilibrium the density of the system yielded a maximum value to the entropy $H_m(f)$. These remarks explain the name of the method.

The expression (3) (with a reversed sign) made its appearance in field of information transmission under the name of information content in the density f ; that is why it is sometimes called the Shannon-Boltzmann entropy. Also, expression (5) appeared in the statistical literature under the name of (Kullback-) divergence of the density f with respect to the density g , and it is denoted by $K(f, g)$ and equal to $-H_n(P, Q)$. See Cover and Thomas [2] or Kullback [3] for a detailed study of the properties of the entropy functions.

Having made those historical remarks and having stated those equivalent definitions, we mention that we shall be working mostly with (3). In what comes below we make use of some interesting and well-known properties of (3) and (5), which we gather under the following.

Theorem 1. *With the notation introduced above, one gets the following:*

- (i) *The function $f \rightarrow H_m(f)$ is strictly concave.*
- (ii) *For any two densities f and g , $H_n(f, g) \leq 0$, and $H_n(f, g) = 0$ if and only if $f = g$ a.e. n .*
- (iii) *For any two densities f, g such that $H_n(f, g)$ is finite, one has (Kullback's inequality)*

$$\frac{1}{2} (\|f - g\|_1)^2 \leq -H_n(f, g). \quad (6)$$

The reader is directed to either Cover and Thomas [2] or Kullback [3] for proofs.

3. The Standard Maximum Entropy Method

Here we recall some well-known results about the standard maximum entropy (SME) method along with some historical remarks. Even though the core idea seems to have been first made in the work of Esscher [4], where he introduced what nowadays is called the Esscher transform, it was not until the mid-1950s that it became part of the methods used in statistics, through the work of Kullback [3]. It seems to have been first formulated as a variational procedure by Jaynes [5] to solve the (inverse) problem consisting in finding a probability density $f(y)$ (on the phase space of a mechanical system), satisfying the following integral constraints:

$$\int_{\Omega} h_k(x) f(x) dm(x) = d_k \quad \text{for } k = 0, 1, \dots, M, \quad (7)$$

where d_k are observed (measured) expected values of some functions (“observables” in the physicist’s terminology), of the random variable X . That problem appears in many fields; see Kapur [6] and Jaynes [7], for example.

Usually, we set $h_0 \equiv 1$ and $d_0 = 1$ to take care of the natural requirement on $f(x)$. It actually takes a standard computation to see that when the problem has a solution, it is of the type

$$f_M^*(x) = \exp \left(- \sum_{k=0}^M \lambda_k^* h_k(x) \right) \quad (8)$$

in which the number of moments M appears explicitly. It is usually customary to write $e^{-\lambda^*} = Z(\lambda^*)^{-1}$, where $\lambda^* = (\lambda_1^*, \dots, \lambda_M^*)$ is an M -dimensional vector. Clearly, the generic form of the normalization factor is given by

$$Z(\lambda) = \int_0^1 e^{-\sum_{k=1}^M \lambda_k h_k(x)} m(dx). \quad (9)$$

With this notation the generic form of the solution can be rewritten as

$$f^*(x) = \frac{1}{Z(\lambda^*)} e^{-\langle \lambda^*, \mathbf{h}(x) \rangle} = e^{-\sum_{k=0}^M \lambda_k^* h_k(x)}. \quad (10)$$

Here $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes the standard Euclidean scalar product in $\mathbb{R}^M$, and $\mathbf{h}(x)$ is the vector with components $h_k(x)$. At this point we mention that the simple minded proof appearing in many applied mathematics or physics textbooks is not really correct. That is because the set of densities is not open in $L_1(dm)$. There are many alternative proofs. Consider, for example, the work by Csiszar [8] and Cherny and Maslov [9].

The heuristics behind (10) and what comes next are the following. If in statement (ii) of Theorem 1 we take $g(x)$ to be any member of the exponential family $\lambda \rightarrow e^{-\sum_{k=0}^M \lambda_k h_k(x)}$, the inequality becomes

$$H_m(f) \leq \ln Z(\lambda) + \langle \lambda, \mathbf{d} \rangle \quad (11)$$

which suggests that if we find a minimizer λ^* such that the inequality becomes an equality, by Theorem 1 we conclude that (10) is the desired solution. This dualization argument seems to have been first proposed in Mead and Papanicolaou [10] and is expounded in full rigor in Borwein and Lewis [11]. The vector λ^* can be found minimizing the dual entropy:

$$\Sigma(\lambda, \mathbf{d}) = \ln Z(\lambda) + \langle \lambda, \mathbf{d} \rangle, \quad (12)$$

where $\mathbf{d}$ is the M -vector with components d_k , and obviously the dependence of f^* on $\mathbf{d}$ is through the minimizer λ^* . We add that, technically speaking, the minimization of $\Sigma(\lambda, \mathbf{d})$ is over the domain of $Z(\lambda)$ which is a convex set with a nonempty interior, and usually the minimum is achieved in its interior. In many applications the domain of $Z(\lambda)$ is $\mathbb{R}^M$. And for the record, we state the result of the duality argument as follows.

Lemma 2. *With the notations introduced above, if the minimizer λ^* of (12) lies in the interior of the domain of $Z(\lambda)$, then*

$$H_m(f^*) = \Sigma(\lambda^*, \mathbf{d}) = \ln Z(\lambda^*) + \langle \lambda^*, \mathbf{d} \rangle. \quad (13)$$

The proof goes as follows. Note that if λ^* is a minimizer of (12), the first order condition is $-\nabla_\lambda \ln Z(\lambda^*) = \mathbf{d}$, which written explicitly states that (10) satisfies the constraints (7). Since the entropy of this density is given by the right hand side of (13), it must be the density that maximizes the entropy.

4. Mathematical Complement

In this section we gather some results about $Z(\lambda)$ that we need as follows.

Proposition 3. *With the notations introduced above, suppose that the matrix $\mathbf{C}$ which we use to denote the covariance of $\mathbf{h}(x)$ computed with respect to the density f^* is strictly positive definite. Let one suppose as well that the set $\{\lambda \in \mathbb{R}^M \mid Z(\lambda) < \infty\}$ is an open set. Then one gets the following:*

- (1) *The function $Z(\lambda)$ defined above is log-convex; that is, $\ln Z(\lambda)$ is convex.*
- (2) *$Z(\lambda)$ is continuously differentiable as many times as one needs.*
- (3) *If one sets $\phi(\lambda) = -\nabla_\lambda \ln Z(\lambda)$, then $\lambda = \phi^{-1}(\mathbf{d})$ is continuously differentiable in $\mathbf{d}$.*
- (4) *The Jacobian $\mathbf{D}$ of ϕ^{-1} at $\mathbf{d}$ equals the (negative) covariance matrix of $\mathbf{h}(x)$ computed with respect to $f^*(x)$.*

The first two assertions are proved in Kullback’s book. Actually, the log-convexity of $Z(\lambda)$ is a consequence of Hölder’s inequality, and the analyticity of $Z(\lambda)$ involves a systematic estimation procedure. The third drops out from the inverse function theorem in calculus. See Fleming [12], and the last one follows from the fact that the Jacobian of ϕ^{-1} equals the negative of the inverse of the Hessian matrix of $\ln Z(\lambda)$, which is (minus) the covariance matrix $\mathbf{C}$. As a simple consequence of item (4) in Proposition 3 we have the following result which is relevant for the arguments in the next section.

Theorem 4. *With the notations introduced above, setting $\delta \mathbf{d} = \hat{\mathbf{d}}_N - \mathbf{d}$, the following assertions hold. The change in λ^* as $\mathbf{d} \rightarrow \mathbf{d} + \delta \mathbf{d}$ up to terms $o(\delta \mathbf{d})$ is given by*

$$\delta \lambda = \mathbf{D} \delta \mathbf{d} \quad (14)$$

and, more importantly, using (10) and, again, up to terms $o(\delta \mathbf{d})$,

$$\hat{f}_N^*(x) - f^*(x) = -f^*(x) \langle (\mathbf{h}(x) - \mathbf{d}), \mathbf{D}(\delta \mathbf{d}) \rangle. \quad (15)$$

To sketch a proof of (15) we proceed as follows. Let $\delta \lambda = \lambda_N^* - \lambda^*$; then

$$\hat{f}_N^*(x) = \frac{e^{-\langle \lambda^*, \mathbf{h}(x) \rangle}}{Z(\lambda^*)} = \frac{e^{-\langle (\lambda^* + \delta \lambda), \mathbf{h}(x) \rangle}}{Z(\lambda^* + \delta \lambda)}. \quad (16)$$

Now, neglecting terms of second order in $\delta \lambda$, we approximate the numerator by

$$e^{-\langle (\lambda^* + \delta \lambda), \mathbf{h}(x) \rangle} = e^{-\langle \lambda^*, \mathbf{h}(x) \rangle} (1 + \langle \delta \lambda, \mathbf{h}(x) \rangle) \quad (17)$$

and the denominator by

$$\begin{aligned} Z(\lambda^* + \delta\lambda) &= Z(\lambda^*) (1 + \langle \nabla_{\lambda} \ln Z(\lambda^*), \delta\lambda \rangle) \\ &= Z(\lambda^*) (1 + \langle \mathbf{d}, \delta\lambda \rangle), \end{aligned} \quad (18)$$

where we used the fact that at the minimum

$$-\nabla_{\lambda} \ln Z(\lambda^*) = \mathbf{d}. \quad (19)$$

Therefore

$$\hat{f}_N^*(x) = \frac{e^{-\langle \lambda, \mathbf{h}(x) \rangle} (1 + \langle \delta\lambda, \mathbf{h}(x) \rangle)}{Z(\lambda^*) (1 + \langle \mathbf{d}, \delta\lambda \rangle)} \quad (20)$$

from which the desired result readily drops out after neglecting terms of second order in $\delta\lambda$.

5. Sample Dependence

Throughout this section, we shall consider a sample $\{X_1, \dots, X_N\}$ of size N of the random variable X . Here we shall relate the fluctuations of $\hat{d}_k = (1/N) \sum_{n=1}^N h_k(X_n)$ around its mean to the fluctuations of the density. The following is obtained from an application of the strong law of large numbers.

Theorem 5. Suppose that $\mathbf{h}$ is integrable, with mean $\mathbf{d}$ and covariance matrix $\Sigma(\mathbf{h})$. Then, for each N , (2) is an unbiased estimator of $\mathbf{d}$ and

$$\hat{d}_k = \frac{1}{N} \sum_{n=1}^N H_k(X_n) \longrightarrow \mathbf{d} \text{ a.s. } \mathbb{P}, \quad \text{when } N \longrightarrow \infty. \quad (21)$$

Consider now the following.

Proposition 6. Define the empirical moments as in (2). Denote by $\hat{\lambda}_N^*$ the Lagrange multiplier determined as explained in Section 2. Then, as $N \rightarrow \infty$, $\hat{\mathbf{d}}_N \rightarrow \mathbf{d}$ and therefore $\hat{\lambda}_N^* \rightarrow \lambda^*$ (a.s.) $\mathbb{P}$.

If $\hat{f}_N^*$ and f^* are the maxentropic densities given by (10), corresponding, respectively, to $\hat{\mathbf{d}}$ and $\mathbf{d}$, then $\hat{f}_N^* \rightarrow f^*$ pointwise for every x and almost surely $\mathbb{P}$.

The proof hinges in the following arguments. From Theorems 5 and 4 we obtain the first assertion. The rest follows from the continuous dependence of the densities on the parameter λ^* . Also, taking limits as $N \rightarrow \infty$ in (15), we obtain another proof of the convergence of $\hat{f}_N^*$ to f^* .

The next result concerns the convergence of $\hat{f}_N^*$ to f^* in $L_1(dm)$.

Theorem 7. With the notations introduced above, one has

$$\hat{f}_N^* \xrightarrow{L_1(dm)} f^* \quad \text{as } N \longrightarrow \infty. \quad (22)$$

Proof. The proof is a consequence of the continuous dependence of $\Sigma(\lambda, \mathbf{d})$ on its arguments, of the identity (13) and item

(iii) in Theorem 1 with $\hat{f}_N^*$ playing the role of f and f^* playing the role of g . In this case $-H_m(\hat{f}_N^*, f^*)$ happens to be

$$\begin{aligned} -H_m(\hat{f}_N^*, f^*) &= \langle \lambda^*, \hat{\mathbf{d}}_N \rangle - \langle \lambda_N^*, \hat{\mathbf{d}}_N \rangle + \ln Z(\lambda_N^*) \\ &\quad - \ln Z(\lambda^*), \end{aligned} \quad (23)$$

which, as mentioned, tends to 0 as $N \rightarrow \infty$. $\square$

To continue, consider the following.

Theorem 8. With the notations introduced above, one has the following:

- (1) $\hat{f}_N^*$ is an unbiased estimator of f^* .
- (2) For any bounded, Borel measurable function $g(x)$,

$$\begin{aligned} &\left(\int g(x) \hat{f}_N^*(x) dm(x) - \int g(x) f^*(x) dm(x) \right)^2 \\ &\leq \left\| \int \mathbf{h}(x) g(x) f^*(x) - \mathbf{d} \int g(x) f^*(x) \right\|^2 \\ &\quad \cdot \langle \delta \mathbf{d}, \mathbf{D} \delta \mathbf{d} \rangle. \end{aligned} \quad (24)$$

The proof follows from (15). Multiply both sides of that identity by $g(x)$ and integrate with respect to $dm(x)$ and then invoke the Cauchy-Schwartz inequality in $\mathbb{R}^K$ to obtain the inequality.

What is interesting about (2) in Theorem 8 is the possibility of combining it with Chebyshev's inequality to obtain rates of convergence. It is not hard to verify that

$$\mathbb{P}(\|\mathbf{D}^{1/2}(\hat{\mathbf{d}}_N - \mathbf{d})\| > a) \leq \frac{\text{tr}(\Sigma_N \mathbf{D})}{a^2} = \frac{\text{tr}(\mathbf{D} \Sigma)}{Na^2}, \quad (25)$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^K$ and

$$\begin{aligned} \Sigma_N &= E_{\mathbb{P}}[\delta \mathbf{d} \delta \mathbf{d}^t] = \frac{1}{N} E_{\mathbb{P}}[(\mathbf{h}(X) - \mathbf{d})(\mathbf{h}(X) - \mathbf{d})^t] \\ &= \frac{1}{N} \Sigma(\mathbf{h}). \end{aligned} \quad (26)$$

Corollary 9. With the notations introduced in Theorem 8 and the two lines above,

$$\begin{aligned} &\mathbb{P}\left(\left| \int g(x) \hat{f}_N^*(x) dm(x) - \int g(x) f^*(x) dm(x) \right| \leq a\right) \\ &\geq 1 - \frac{\left\| \int \mathbf{h}(x) g(x) f^*(x) - \mathbf{d} \int g(x) f^*(x) \right\|^2 \text{tr}(\mathbf{D} \Sigma(\mathbf{h}))}{Na^2}. \end{aligned} \quad (27)$$

Comment. If we take $g = I_A$, we obtain a simple estimate of the speed of decay of $|\int_A \hat{f}_N^*(x) dm(x) - \int_A f^*(x) dm(x)|$ to zero, or of the speed of convergence of $\int_A \hat{f}_N^*(x) dm(x)$ to $\int_A f^*(x) dm(x)$ if you prefer. Regarding the fluctuations around the mean, consider the following two possibilities.

Theorem 10. *With the notations introduced in Theorem 4 and in the identity (15), one has*

$$\begin{aligned} & \sqrt{N} \left(\hat{f}_N^*(x) - f^*(x) \right) \\ &= -f^*(x) \left\langle (\mathbf{h}(x) - \mathbf{d}), \mathbf{D} \left(\sqrt{N}(\delta \mathbf{d}) \right) \right\rangle \\ &\sim N \left(0, \sigma^2(x) \right) \end{aligned} \quad (28)$$

in law as $N \rightarrow \infty$. Also, for any bounded, Borel measurable $g(x)$,

$$\begin{aligned} & \sqrt{N} \left(\int g(x) \hat{f}_N^*(x) dm(x) \right. \\ & \quad \left. - \int g(x) f^*(x) dm(x) \right) \\ &= \left\langle \left(\int \mathbf{h}(x) g(x) f^*(x) - \mathbf{d} \int g(x) f^*(x) \right), \right. \\ & \quad \left. \mathbf{D} \left(\sqrt{N}(\delta \mathbf{d}) \right) \right\rangle \sim N \left(0, \sigma^2(g) \right) \end{aligned} \quad (29)$$

in law. Above, $\sigma^2(x) = \langle \mathbf{v}(x), \mathbf{\Sigma}(\mathbf{h})\mathbf{v}(x) \rangle$, $\sigma^2(g) = \langle \mathbf{v}(g), \mathbf{\Sigma}(\mathbf{h})\mathbf{v}(g) \rangle$, where $\mathbf{v}(x) = f^*(x)\mathbf{D}^{1/2}(\mathbf{h}(x) - \mathbf{d})$ and $\mathbf{v}(g) = \int g(x)\mathbf{v}(x)dm(x)$.

The proof of the assertions involves applying the central limit theorem to the vector variable $\sqrt{N}(\hat{\mathbf{h}}_N - \mathbf{d})$.

Final Comments. Some numerical results illustrating the results presented here appear in the paper by Gomez-Gonçalves et al. [1]. There, they display graphically a “cloud of densities” corresponding to samples on various sizes and see how they shrink to the plot of the true (or exact) density as the sample size increases.

Competing Interests

The author declares that they have no competing interests.

References

- [1] E. Gomez- Gonçalves, H. Gzyl, and S. Mayoral, “Loss data analysis Analysis of the sample dependence in density reconstruction by maximum entropy methods,” *Insurance: Mathematics and Economics*, vol. 71, pp. 154–153, 2016.
- [2] T. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2003.
- [3] S. Kullback, *Information Theory and Statistics*, Dover Publications, New York, NY, USA, 2nd edition, 1968.
- [4] F. Esscher, “On the probability function in the collective theory of risk,” *Skandinavisk Aktuarietidskrift*, vol. 15, pp. 175–195, 1932.
- [5] E. T. Jaynes, “Information theory and statistical mechanics,” *Physical Review*, vol. 106, pp. 620–630, 1957.
- [6] N. Kapur, *Maximum entropy Models in Science and Engineering*, Wiley Eastern, New Delhi, India, 1998.
- [7] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
- [8] I. Csiszar, “I-divergence geometry of probability distributions and minimization problems,” *Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.
- [9] A. S. Cherny and V. P. Maslov, “On maximization and minimization of entropy functionals in various disciplines,” *Theory of Probability and Its Applications*, vol. 3, pp. 447–464, 2003.
- [10] L. R. Mead and N. Papanicolaou, “Maximum entropy in the problem of moments,” *Journal of Mathematical Physics*, vol. 25, no. 8, pp. 2404–2417, 1984.
- [11] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization*, CMS Books in Mathematics, Springer, New York, NY, USA, 2000.
- [12] W. Fleming, *Functions of Several Variables*, Springer, Berlin, Germany, 1987.

Hindawi

Submit your manuscripts at
<http://www.hindawi.com>