Kybernetika

László Györfi; Adam Krzyżak
Why L_1 view and what is next?

Kybernetika, Vol. 47 (2011), No. 6, 840--854

Persistent URL: http://dml.cz/dmlcz/141728

KYBERNETIKA — VOLUME 47 (2011), NUMBER 6, PAGES 840–854

WHY L_1 VIEW AND WHAT IS NEXT?

László Györfi and Adam Krzyżak

N. N. Cencov wrote a commentary chapter included in the Appendix of the Russian translation of the Devroye and Györfi book [15] collecting some arguments supporting the L_1 view of density estimation. The Cencov's work is available in Russian only and it hasn't been translated, so late Igor Vajda decided to translate the Cencov's paper and to add some remarks on the occasion of organizing the session "25 Years of the L_1 Density Estimation" at the Prague Stochastics 2010 Symposium. In this paper we complete his task, i. e., we translate the Cencov's chapter and insert some remarks on the related literature focusing primarily on Igor's results. We would also like to acknowledge the excellent work of Alexandre Tsybakov who translated the Devroye and Györfi book in Russian, annotated it with valuable comments and included some related references published in Russian only.

Keywords: Cencov's comments, inverse problems in distribution estimation, L_1 density estimation, variational distance, ϕ-divergence

Classification: 62G08, 62G20

1. INTRODUCTION

The monograph by Devroye and Györfi [15] is devoted to density estimation from random observations. The book surveys state of the art of the field up to 1983, discusses in detail the most important results and poses open problems. The title closely reflects the content, so it may seem to an unexperienced reader that the book by two young scientists is devoted to a narrow and exotic problem. Nevertheless the L_1 approach seems the simplest and the most natural for the stated problem and this claim can be rigorously proven. To this end it is sufficient to apply the theory of statistical inference proposed by Wald [43]. In the short account presented here we place the subject of the book in the framework of mathematical statistics and thus complement and make authors' results more precise, in particular regarding the choice of the L_1 approach and its relation to the L_2 approach.

In probability theory we define a probability space $(\Omega, \mathcal{A}, P)$ to describe random events, where Ω is the space of elementary events $\omega \in \Omega$, $\mathcal{A}$ is the σ-algebra in Ω, and P is a probability measure on $\mathcal{A}$ called probability distribution function. The measurable space $(\Omega, \mathcal{A})$ provides a qualitative description of a random variable and P is a quantitative measure. Let $\omega_1, \ldots, \omega_N$ be independent measurements of a given event, i. e., $P^N\{\mathrm{d}\omega\} = P\{\mathrm{d}\omega_1\} \ldots P\{\mathrm{d}\omega_N\}$. By the law of large numbers the frequency $\nu(A)/N \approx P(A)$ for any event $A \in \mathcal{A}$, where $\nu(A) = card\{\omega_k, \omega_k \in A, k =$

$1, \ldots, N\}$ with estimation accuracy (measured by probability P^N) improving with N. The study of behavior of an empirical mean is one of the fundamental problems of probability theory and it often determines its practical importance. In mathematical statistics we encounter an inverse problem: given the sequence of random variables $\omega_k,\ k = 1, \ldots, N$ from a given probability space $(\Omega, \mathcal{A})$ we estimate the "observed" probability distribution P or any of its characteristics or determine the properties of an unknown distribution. In the former case we talk about pointwise estimation problem (PEP) which is the subject of Devroye and Györfi studies and of this article. We would like to add that pointwise estimation of P is the first step in solving of many statistical estimation problems, where estimate P^* is used to infer properties of P.

Inverse physics problems are by their nature ill-conditioned. The same is true for the inverse problems in probability theory. For existence of a strong solution additional information is needed. For example, consider a family of probability distributions $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$. If the parameter space Θ is finite then we have the hypothesis testing problem. If P_θ smoothly depends on a finite dimensional parameter vector then we face a parametric estimation problem. In this book particular attention is given to the estimation problem when family $\mathcal{P}$ is smoothly parameterized by a countable dimensional vector of real coordinates. One example is when random variable is finite with density $p(x)$ in $C^{(2)}$. Finally, family $\mathcal{P}$ may be so large that it cannot be smoothly parameterized: an example is the case of two-dimensional random variables where the only prior knowledge is assumption about independence of the components. Traditionally the latter two approaches belong to the class of nonparametric estimation problems even though the approaches and solution methodologies are entirely different. In particular, the latter approach is only correct in the weak sense. Naturally this classification is not exhaustive. We only listed the most important approaches ordering them according to diminishing prior information along with accuracy. Thus in testing simple hypotheses error probability goes down exponentially, in a finite parametric case the pointwise rate is of order $N^{-1/2}$ and in a countable parametric case the rate is even slower. The latter is the case we will consider in detail.

2. DENSITY ESTIMATION AND THE CHOICE OF METRICS

According to Wald [43] in each statistical problem in addition to input data we should be given a measurable space $(\Delta, \mathcal{B})$ from which we choose a decision $\delta \in \Delta$ based on experimental data. The resulting deterministic decision, that is data processing rule chosen by a statistician, is expressed by a function $\delta = f(\omega_1, \ldots, \omega_N)$ depending only on observations and a randomized decision function is expressed by $\delta = f(\omega_1, \ldots, \omega_N; \eta)$ depending also on random parameter η which needs to be determined as well. The randomized and deterministic solutions can be conveniently described implicitly by a probability distribution $M(\omega_1, \ldots, \omega_N; d\delta)$ without specifying the space of random experiments.

Wald suggests to measure the accuracy of decision δ by the loss function $\mathcal{L}(\theta, \delta)$, which a statistician faces when he makes decision δ upon observing P_θ. The quality

of the decision rule M can be measured by an expected loss or risk

$$\mathcal{R}_{M(N)} = E\mathcal{L}(\theta, \delta) = \int_{\Delta} \mathcal{L}(\theta, \delta) P_{\theta,M}^{N}(\mathrm{d}\delta) \tag{1}$$

where probability measure $M = P_{\theta,M}^{N}$ is defined by

$$P_{\theta,M}^{N}\{\cdot\} = \int_{\Omega^N} M(\omega_1, \ldots, \omega_N; \cdot) P_\theta\{\mathrm{d}\omega_1\} \ldots P_\theta\{\mathrm{d}\omega_N\}. \tag{2}$$

Note that the risk can be replaced by any quantile of random losses. Nevertheless knowing family $\mathcal{P}$ we can determine distribution $P_{\theta,M}^{N}$ in advance and then choose the most convenient rule M.

In the inverse problem of probability theory we take as set Δ of all possible estimates P^* a complete family of all probability distributions P on $(\Omega, \mathcal{A})$ denoted by $Cap(\Omega, \mathcal{A})$. For the Lebesgue measurable spaces $(\Omega, \mathcal{A})$ after overcoming some difficulties following the approach of Prokhorov one may embed this set with a σ-algebra of events and design a measurable space of random probability measures $Cap(Cap(\Omega, \mathcal{A}), \mathcal{K}(\mathcal{A}))$, where $\mathcal{K}(\mathcal{A})$ is the σ-algebra on $Cap(\Omega, \mathcal{A})$. In this we use $P(A_i)$ as parameters where A_i are some elements of $\mathcal{A}$ and dependence of P on parameters $\mathcal{K}(\mathcal{A})$ is measurable but not smooth. We adopt the variational distance criterion for measuring the accuracy of estimate P^*, namely

$$|P^* - P| = \sup_{A \in \mathcal{A}} [P^*(A) - P(A)] - \inf_{A \in \mathcal{A}} [P^*(A) - P(A)], \tag{3}$$

i. e., the strong metrics in a linear space $L_1(\Omega, \mathcal{A})$ of all countable additive measures (positive-negative in general) on $(\Omega, \mathcal{A})$; $L_1(\Omega, \mathcal{A}) = Lin\, Cap(\Omega, \mathcal{A})$.

We have the following result (see Cencov [11], Theorem 4)

Theorem 1. Let the loss function for a pointwise estimation problem be specified by the strong norm. Then the pointwise estimation problem for family $Cap(E, \mathcal{A}^*)$ without prior information is ill-posed, where E is the unit interval and $\mathcal{A}^*$ is an algebra of its all Lebesgue measurable subsets.

It turns out that for any sequence of decision functions

$$M(N) : Cap(E^N, \mathcal{A}^{*N}) \to Cap(Cap(\mathcal{E}, \mathcal{A}^*), \mathcal{K}(\mathcal{A}^*))$$

there exists a probability distribution $P \in Cap(E, \mathcal{A}^*)$ such that

$$\lim_{N \to \infty} \mathcal{R}_{M(N)}(P) \geq 1, \tag{4}$$

where E^N is the Cartesian product of N sets E.

Remark 1. *In this respect Devroye and Györfi [16] proved that, for any estimator P_N and $1 > \delta > 0$, there exists a singular probability distribution P^* defined on the Borel σ-algebra of the interval $[0,1]$ such that for all N*

$$|P_N - P^*| > 1 - \delta$$

almost surely, which means that without any additional information on an unknown probability distribution P, it is impossible to estimate it consistently in variational distance. However, if there is a known σ-finite measure Q dominating the non-atomic part of P, then there is a partitioning-based estimate P_N such that

$$|P_N - P| \to 0$$

almost surely, as $N \to \infty$ (cf. Barron, Györfi and van der Meulen [3]).

Besides strong metrics on $Cap(E, \mathcal{A}^*)$ there exist weak metrics expressed by distribution functions, where E is the unit interval. A simple one is a uniform distance metric

$$\rho(P, Q) = \sup_x |F(x) - G(x)|,$$

where $F(x) = P\{[0, x)\}$, $G(x) = Q\{[0, x)\}$. According to the famous Kolmogorov theorem [25]

$$\sup_x |F_N^*(x) - F(x)| \to 0$$

almost surely as $N \to \infty$ so that empirical distribution function $F_N^*(x)$ is a consistent estimate of the theoretical distribution F. Thus ill-conditioning of the inverse theory of probability problem is not very strong, likewise numerical differentiation problem in real analysis.

Here we will not consider the question how much less informative is estimation in weak metrics from estimation in strong metrics. For more details the reader is referred to [36]. We will instead consider an example. It is well-known that every measure on the real line can be decomposed into linear combination of three components: discrete, continuous and singular. A discrete measure is concentrated on a countable set of discrete points with positive measure. A continuous measure possesses Lebesgue density and a singular measure is concentrated on a subset with Lebesgue measure zero. All three measures are strictly separated by distributions. They are also separated in the strong metrics: variational distance between two measures of different types is equal to 2. In weak metrics continuous and singular measures are not separable. One may show that there is no consistent rule which allows to decide whether a given distribution is continuous or singular. It is worth noting that discrete distributions can be separated from two other types of distributions if and only if the equality between identical results of experiments can be established with infinite precision.

One can just conclude from preceding discussion that the maximal set for which PEP is valid in strong metrics sense is a subset of all dominated measures $Capd(\Omega, \mathcal{A}, Z)$ on any measurable space $(\Omega, \mathcal{A})$ with fixed ideal Z of zero measure sets and, in particular, a subset of all distributions on real line or unit interval which have

Lebesgue density. Here $Capd(\Omega, \mathcal{A}, Z)$ denotes the collection of all probability measures on $(\Omega, \mathcal{A})$ that vanish on the ideal Z. For such subsets of distributions the variational distance reduces to the standard L_1 distance between two densities

$$|Q - P| = \int_{\Omega} |q(\omega) - p(\omega)| \mu(\mathrm{d}\omega), \tag{5}$$

where μ is a dominating measure, e. g., $\mu(dx) = dx$ for a unit interval, and the density estimate converges to the density. Well-posedness of this problem in pointwise case has been established in 1976 independently by Abu-Jaoude [1] (his results are proven in the Devroye and Györfi monograph [15]) and Nadaraya [33]. The former author used the histogram density estimate while the latter authors the kernel density estimate. From invariance of the family of all continuous mutually measurable distributions with respect to the length preserving mappings of the interval onto itself it follows that any universally consistent decision rule on the interval cannot be uniformly consistent on this interval [11]. At the same time the risk of the same decision procedures can converge uniformly over family $\mathcal{P}$ for prior distributions P satisfying stricter conditions.

An important task for a theoretician is quest for the most general approach. We will now try to elucidate why the L_2 approach is inferior to the L_1 approach. Each statistical decision rule M yields an affine mapping from $Cap(\Omega^N, \mathcal{A}^N)$ to $Cap(\Delta, \mathcal{B})$. This is a consequence of the fact that the rule M is determined by a transitive distribution function. Consider all possible families $Cap(\Omega, \mathcal{A})$ and all possible transitive distributions M from any measurable space $(\Omega, \mathcal{A})$ to any other measurable space $(\Omega', \mathcal{A}')$. They represent an algebraic category CAP, which consists of subsets $Cap(\Omega, \mathcal{A})$. The morphisms (or Markov morphisms) are defined by the transitive distributions

$$Q\{\cdot\} = (PM)\{\cdot\} = \int_{\Omega} M(\omega; \cdot) P(\mathrm{d}\omega).$$

With this all requirements of categories are fulfilled.

1. The identity $\mathcal{T}$ mapping each object onto itself is a category, i. e., $\mathcal{T}(\omega, B) = I_B(\omega)$, where I_B is an indicator of set B.

2. The composition ST of two Markov mappings is a Markov mapping Π

$$\Pi(\omega; \cdot) = \int_{\Omega'} T(\omega'; \cdot) S(\omega; \mathrm{d}\omega')$$

3. The composition is associative, i. e., $(ST)R = S(TR)$.

This fact was first noted by Cencov [5] and independently by Morse and Sacksteder [29]. In addition, multiplication operation for members as well as multiplication and averaging operations for morphisms are defined in CAP as well, see Cencov [9]. This has the following implication [9, 10]: two families of distributions $\{P_\theta, \theta \in \Theta\}$ and

$\{Q_\theta, \theta \in \Theta\}$ sharing parameter set Θ have the same statistical properties if and only if there exist two Markov morphisms S and T such that

$$P_\theta S = Q_\theta, \quad Q_\theta T = P_\theta, \quad \forall\, \theta \in \Theta.$$

In any theory a common law allows equivalent formulation. In other words the implications of the law should not change when we move from one case to another provided that they are equivalent in that theory. In classical geometry equivalent transformations form groups, in statistics they form categories entailing original geometries. The families of distributions play role of "figures" and Markov morphisms of "movements". Many basic notions of mathematical statistics could be interpreted as invariants, co-invariants or more complex equivariants of that geometry. Since Markov morphisms are generally not invertible except for a typical for group geometry notion of invariance for congruent shapes (i. e., statistical equivalent families) we introduce the notion of monotone invariant. We will apply it only for a pair of probabilistic laws, but it can be easily generalized to a more general family.

Definition. A real-valued function $f(P, Q)$ whose two arguments are probability distributions defined on the Cartesian product of all elements of $Cap(\Omega, \mathcal{A})$ is called a monotone invariant (relative to the category of Markov morphisms) if

$$f(P\mathcal{M}, Q\mathcal{M}) \le f(P, Q) \tag{6}$$

for all $P, Q, \mathcal{M} \in \Omega$.

The examples of monotone invariants of a pair of distributions include variational distance $|P - Q|$ (3), relative entropy

$$\mathcal{H}(P, Q) = \int_\Omega \left[\frac{\mathrm{d}Q}{\mathrm{d}P}(\omega) \ln \frac{\mathrm{d}Q}{\mathrm{d}P}(\omega)\right] P(\mathrm{d}\omega) = \int_\Omega \left[\ln \frac{\mathrm{d}P}{\mathrm{d}Q}(\omega)\right] Q(\mathrm{d}\omega),$$

and Bhattacharya distance

$$s(P, Q) = 2 \arccos \int_\Omega \sqrt{P(\mathrm{d}\omega) Q(\mathrm{d}\omega)} \tag{7}$$

entailing quadratic Fisher information. As a matter of fact this distance is a unique invariant Riemann distance metrics up to a multiplicative constant on members of category CAP. When speaking of smooth sets $\{P_\theta, \theta \in \Theta\}$ of finite measures it is always tacitly assumed that P_θ is a differentiable function with respect to $\theta \in \Theta \subset \mathcal{R}^d$ in a sense of the given metrics. It is easy to show that $\phi - divergence$ of Csiszár [13, 14]

$$I_\phi(P, Q) = \int \phi\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(\omega)\right) Q(\mathrm{d}\omega)$$

also belongs to the class of monotone invariants, where ϕ is a convex function on R^+. For mutually absolutely continuous measures the notion of entropy reduces to

$$\mathcal{H}(P, Q) = \int_\Omega \left[\ln \frac{q(\omega)}{p(\omega)}\right] q(\omega) \mu(\mathrm{d}\omega), \tag{8}$$

and the Fisher form for family $\{P_\theta\}$ is

$$\mathrm{d}s^2 = \sum_{\alpha,\beta} \mathrm{d}\theta_\alpha \mathrm{d}\theta_\beta \int_\Omega \frac{\partial \ln p(\omega;\theta)}{\partial \theta_\alpha} \cdot \frac{\partial \ln p(\omega;\theta)}{\partial \theta_\beta}. \tag{9}$$

An exceptional role of the L_1 norm in the class of invariant metrics is elucidated in the following theorem, see [32].

Theorem 2. If metrics ρ defined on the members of category CAP is monotone, then

$$\rho(P,Q) \geq \frac{1}{8}\rho(R_{1/2}, R_{1/4}) \cdot |P - Q|, \tag{10}$$

where R_θ is probability distribution on $\Omega_2 = \{\omega_1, \omega_2\}$, $R_\theta(\omega_1) = \theta$, $R_\theta(\omega_2) = 1 - \theta$, $0 \leq \theta \leq 1$.

Thus if statistical problem with variational distance loss function is ill-conditioned then it remains ill-conditioned whenever the loss function is replaced by any other invariant metrics (the converse is false). Likewise the variational distance is the unique (up to multiplicative constant) invariant distance determined by the norm of the difference, see [8]. More precisely if the metrics ρ on elements of CAP is invariant with respect to the category and homogeneous, i. e.,

$$P - Q = \lambda(P' - Q') \Rightarrow \rho(P,Q) = \lambda\rho(P',Q'),$$

then

$$\rho(P,Q) = \frac{1}{2}\rho(R_{1/2}, R_{1/4}) \cdot |P - Q|.$$

Similar property holds for general loss functions (for ϕ-divergence it was proved by Csiszár) if they are monotone invariants, see [32].

Theorem 3. For a monotone invariant function $\mathcal{L}(P,Q)$ satisfying

$$Q \neq P \Rightarrow \mathcal{L}(P,Q) \neq \mathcal{L}(P,P)$$

there exists a constant $c = \mathcal{L}(P,P), \forall\, P$ and a monotone real function

$$g(z) = \inf_{2|x-y| \geq z} \rho(R_x, R_y), \tag{11}$$

where $0 \leq z \leq 2$, $g(z) > 0$ with $z > 0$, $g(0) = 0$ such that

$$\mathcal{L}(P,Q) \geq c + g(|P - Q|).$$

Thus we have shown that the L_1 approach investigated by the authors of [15] is the most general invariant approach to density estimation, the claim which we have embarked to show.

Remark 2. *For the convex function ϕ with $\phi(1) = 0$, no ϕ-divergences except the constant multiples of variational distance are metrics (cf. Khosravifard, Fooladivanda and Gulliver [24]). Powers of ϕ-divergences may be metrics (cf. Csiszár and Fischer [12], Kafka, Österreicher and Vincze [22], Österreicher and Vajda [35], Vajda [42]). Vajda [41] proved that if function ϕ is strictly convex then a lower bound*

$$I_\phi(P, Q) \geq c_\phi |P - Q|$$

is impossible for any finite constant c_ϕ, while the sharp lower bound

$$I_\phi(P, Q) \geq c_\phi |P - Q|^2$$

holds.

Remark 3. *General inequalities have been proven in the Liese and Vajda book [31]. For example, if ϕ-divergence is symmetric, which means the convex function ϕ with $\phi(1) = 0$ is self-adjoint in the sense*

$$\phi(t) = \phi^*(t),$$

where $\phi^*(t) = t\phi(1/t)$, then

$$I_\phi(P, Q) \geq g_\phi(|P - Q|),$$

where

$$g_\phi(t) = \frac{2 + t}{2} \phi\left(\frac{2 - t}{2 + t}\right), \quad 0 \leq t < 2$$

(cf. Proposition 8.28 in [31]). There are similar inequalities for non-symmetric ϕ-divergences such that g_ϕ is replaced by the convex envelope of g_ϕ and g_{ϕ^*} (cf. (8.26) in [31]).

Remark 4. *The best known inequality involving the variational distance $|P - Q|$ and the relative entropy $\mathcal{H}(P, Q)$ is the Pinsker inequality:*

$$|P - Q|^2/2 \leq \mathcal{H}(P, Q)$$

(cf. Csiszár [13], Kullback [27] and Kemperman [23]), so the relative entropy is a more demanding divergence. Kullback [27], [28] sharpened the Pinsker inequality adding a fourth power term:

$$|P - Q|^2/2 + |P - Q|^4/36 \leq \mathcal{H}(P, Q),$$

while the best known lower bound is due to Toussaint [39]:

$$|P - Q|^2/2 + |P - Q|^4/36 + |P - Q|^6/288 \leq \mathcal{H}(P, Q).$$

Vajda [40] proved a slightly different lower bound:

$$\log \frac{2 + |P - Q|}{2 - |P - Q|} - \frac{|P - Q|}{2 + |P - Q|} \leq \mathcal{H}(P, Q).$$

Vajda [41] gave further upper and lower bounds of $\mathcal{H}(P, Q)$ in terms of $|P - Q|$.

Remark 5. *In order to have distribution estimate consistent in relative entropy, one needs more information on the underlying distribution. For example, for any estimator P_N, there is a probability distribution P^* on the set of positive integers such that its Shannon entropy is finite and for all N*

$$\mathcal{H}(P_N, P^*) = \infty$$

almost surely (cf. Györfi, Páli and van der Meulen [18]). If there is a known probability measure Q dominating P such that

$$\mathcal{H}(P, Q) < \infty,$$

then there is a partitioning-based estimate P_N such that

$$\mathcal{H}(P_N, P^*) \to 0 \quad as \quad N \to \infty$$

almost surely (cf. Barron, Györfi and van der Meulen [3] and Györfi, Páli and van der Meulen [19]).

The L_2 approach also considered by Devroye and Györfi [15] is not invariant since the squared norm

$$\int_\Omega [p(\omega) - q(\omega)]^2 \mu(\mathrm{d}\omega) \tag{12}$$

depends on the choice of dominating measure μ. It may happen that upon change of measure the finite norm of the difference becomes infinite, when at the same time (5) and (8) remain unchanged. Nevertheless for some classes of prior distributions $\mathcal{P}$ it is possible to obtain quasi-invariance by using a set of equivalent L_2 norms. We will consider this approach a bit later.

3. L_2 NORM AND EXPONENTIAL DENSITY ESTIMATION APPROACH

Regardless of the lack of invariance the L_2 approach is convenient as it allows to design simple density estimates and analysis of those is straightforward. For instance projection-based density estimates are intrinsically related to L_2 norms. These estimates were introduced in 1958 by Cencov and further investigated in [4, 17]. As shown experimentally by Statultavicius [37] the projection approach is more efficient by an order of magnitude than the Rosenblatt–Parzen kernel approach with a fixed kernel and storage requirements are of order of magnitude less demanding. Projection techniques as well as kernel methods with non-positive kernels also suffer from serious problems, namely the plug-in partial Fourier series estimate may assume negative values and the projected density estimate may turn out not to be a density. Naturally one may set negative parts to zero and normalize the resulting estimate so that its integral is one. As Devroye and Györfi have shown [15, p. 269] this leads to reduction of L_1 error, but the simplicity of the projection approach and its low memory requirements are lost. Thus need arises for an equivariant approach. The essential requirement is attainment of good precision. It is well-known that the histogram and kernel approaches with positive kernels converge weakly to densities in $C^{(2)}$ with the rate $N^{-2/5}$.

An exponential density estimation approach proposed by Stratonovich [38] and Cencov [8] is free of these deficiencies. The idea is not to estimate a density directly but its logarithm instead. Since the integral of $p_0(\omega)\exp[s^j q_j(\omega)]$ need not be a unit integral we need to normalize it yielding

$$p(\omega;\mathbf{s}) = p_0(\omega)\exp[s^j q_j(\omega) - \Psi(\mathbf{s})], \tag{13}$$

$$\Psi(\mathbf{s}) = \ln\int_\Omega \exp[s^j q_j(\omega)]p_0(\omega)\mu(\mathrm{d}\omega), \tag{14}$$

where $\Psi(\mathbf{s})$ is the logarithm of the normalizing factor. Equation (13) is valid when (14) is finite. It is easy to show that $\Psi(\mathbf{s})$ is convex and consequently the set $\{\mathbf{s} : \Psi(\mathbf{s}) < \infty\}$ is convex (possibly empty).

Thus we have designed an exponential family ρ of densities with canonical parameter $\mathbf{s}$ and domain $Dom\ \gamma = \{\mathbf{s} : \Psi(\mathbf{s}) < \infty\}$. Naturally canonical parametrization is determined up to affine transformations of parameters and statistics $q_j(\omega)$.

The exponential families are popular in mathematical statistics (and statistical physics). For more detailed account refer to [2, 6, 8]. Introduce a vector function

$$\mathbf{t} = \mathbf{T}(\mathbf{s}) = \int_\Omega \mathbf{q}(\omega)p(\omega;\mathbf{s})\mu(\mathrm{d}\omega). \tag{15}$$

It is well-known that one-to-one and analytical vector function $\mathbf{T}(\mathbf{s}) = \mathrm{grad}\ \Psi(\mathbf{s})$ inside the region $Dom\ \gamma$ is the Legendre transformation such that

$$\mathbf{s} = \mathrm{grad}[s^j(\mathbf{t})t_j - \Psi(\mathbf{s}(\mathbf{t}))].$$

The dependence between $\mathbf{s}$ and $\mathbf{t}$ is a bit more complex on the boundary of $Dom\ \gamma$, see [8]. The parameter $\mathbf{t}$ will be called a natural parameter of the exponential family. There exists a simple estimate for this parameter

$$\mathbf{t}^* = N^{-1}[\mathbf{q}(\omega_1) + \ldots + \mathbf{q}(\omega_N)]. \tag{16}$$

This estimate is efficient, i. e. its Fisher information inequality becomes equality. The converse is also true: if parameters of the family have efficient estimates then the family is exponential. The most complicated aspect of the estimation procedure is derivation of $\mathbf{s}^*$ corresponding to $\mathbf{t}^*$ as the probability density is defined via $\mathbf{s}$. However, this difficulty is not essential as we need to solve the system of equations $\mathrm{grad}\Psi(\mathbf{s}) = \mathbf{t}^*$ once at the end. The density $p_0(\omega)$ is determined up to a multiplicative constant and parameters $q_j(\omega)$ up to a constant. If we take $p_0(\omega) = p(\omega;\mathbf{0})$, $\mathbf{q}(\omega) = \mathbf{q}(\omega;\mathbf{0})$, where $\int p(\omega;\mathbf{0})\mathbf{q}(\omega;\mathbf{0})\,\mathrm{d}\mu = \mathbf{0}$, then $\Psi(\mathbf{s}) = H(P_s, P_0)$, where entropy is given by (8) and

$$p(\omega;\mathbf{s}) = p(\omega;\mathbf{0})\exp[s^j q_j(\omega;\mathbf{0}) - H(P_s, P_0)]. \tag{17}$$

Compare (17) with the multivariate gaussian density with identity covariance matrix and mean vector $\mathbf{s}$

$$p(x,\mathbf{s}) = p(x,\mathbf{0})\exp[s^j x_j - (\mathbf{s},\mathbf{s})/2]. \tag{18}$$

We can draw two conclusions. First, we see that the gaussian family is a unique family with canonical parametrization corresponding to natural parametrization [8]. For instance in statistical physics a natural parameter temperature $t = s^{-1}$. Second, the relative entropy is generalization of the L_2 distance. The analogy goes so far that for the relative entropy non-symmetric Pythagoras theorem holds, see [7].

Theorem 4. If $P_\sigma = \arg\min_s \mathcal{H}(R, P_s)$, where $\gamma = \{P_s\}$ is an exponential family, $\sigma \in Int\ Dom\ \gamma$ then

$$\mathcal{H}(R, P_s) = \mathcal{H}(R, P_\sigma) + \mathcal{H}(P_\sigma, P_s) \quad \forall\, s \in Dom\ \gamma. \tag{19}$$

When parameters $q_j(\omega; \mathbf{0})$ in (17) are bounded then the additional condition on σ can be dropped.

Laplace suggested to measure the loss by a distance between the estimate and the true value of a parameter. Gauss noted that theory is significantly simplified for the squared error loss function [30]. For non-gaussian distributions it is natural to take the loss $\mathcal{L} = 2\mathcal{H}(P^*, P_\theta)$ instead of any $L_2(\mu)$ norm dependent on μ. The main goal of a statistician is to choose a priori reasonable families guaranteeing good rates of convergence of loss functions (maximum risk, Bayes risk, etc.) Note that unlike the L_1 approach pointwise estimation problem with entropy loss function for all dominating subsets $Capd(\mathcal{E}, \mathcal{E}^*, Z)$ is already ill-conditioned, see [11, Theorem 6].

We have already observed that density estimates in general are not uniformly consistent [11]. Uniformly consistent estimates can only be constructed for a narrow class $\mathcal{P}$ of distributions. First of all one has to assume that

$$C^{-1} \le \frac{\mathrm{d}Q}{\mathrm{d}P}(\omega) \le C, \quad \forall\, \omega \in \Omega,\ \forall\, P, Q \in \mathcal{P}, \tag{20}$$

for some positive constant C. Observe that condition (20) is rather strong. All $L_p(R)$ norms, $1 \le p < \infty$, $R \in \mathcal{P}$ define the unique topology on $\mathcal{P}$ and all L_2 norms are equivalent when $R \in \mathcal{P}$. In light of quasi-invariance of measures we can derive lower bounds for density estimates and we can get efficient algorithms [8, 21]. Unfortunately the family of distributions satisfying (20) is quite small. For instance it does not contain gaussian distributions (18).

If we replace the $L_2(R)$ norm in the definition of the loss function by the entropy for which the following always holds

$$\mathcal{H}(P, Q) \le ||P - Q||_P^2. \tag{21}$$

A family $\mathcal{P}$ is quasi-homogeneousness if

$$C^{-1} \le D_{R'} \left[\ln \frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) \right] / D_{R''} \left[\ln \frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) \right] \le C \tag{22}$$

for some fixed constant C and any $P, Q, R', R'' \in \mathcal{P}$, where D_R is the differentiation operator, see [8]. This condition is locally satisfied by the regular smooth families in the Cramér–Rao sense. Following [8] we recall a basic result of the optimal estimation theory.

Theorem 5. Let $\{P_\theta, \theta \in \Theta\}$ be a compact smooth family of probability distributions and let Θ be a closed region in R^d. Then for $\mathcal{L}(\theta, P^*) = 2\mathcal{H}(P^*, P)$

$$\lim_{N\to\infty} N \cdot \inf_{M(N)} \sup_{\theta\in\Theta} \mathcal{R}_{M(N)}(\theta) = dim\ \Theta, \tag{23}$$

where we do not have to restrict ourselves to estimates $P^* \in \{P_\theta\}$ and the maximum likelihood estimate $\Pi(N)$ is asymptotically optimal

$$\sup_{\theta\in\Theta} |N \cdot \mathcal{R}_{\Pi(N)}(\theta) - dim\ \Theta| \to 0, \quad \text{as} \quad N \to \infty. \tag{24}$$

One may interpret the maximum likelihood estimate as an estimate obtained by minimizing the relative entropy, see [20, 26] and also [8].

Combining Theorem 4 and Theorem 5 Cencov [8] developed theory of almost optimal nonparametric density estimation (specifically countably parametric) for quasi-homogeneous families with properly decreasing information width. As we have already seen Cencov approach is close to the L_2 approach and coincides with it at the stage of constructing averages of basis (control) functions $q_j(\omega)$, however it differs in "interpretation" of the constructed averages (both approaches are equivalent only for histograms). For details the reader is referred to [8]. Thus it appears to Cencov that thanks to its equivariance the exponential approach to density estimation has no less potential than traditional methods discussed in Nadaraya [34] or smoothing kernel and orthogonal series approaches discussed in the book of Devroye and Györfi.

4. CONCLUSIONS

We have translated into English Cencov's comments to the Russian translation of the monograph by Devroye and Györfi [15] and complemented it by the remarks on the related results concerning the variational distance, ϕ-divergences and relative entropy. Cencov used decision theory and the framework of inverse problems in statistics to which he substantially contributed over the years, see [8] to justify the validity of L_1 approach to density estimation taken by the authors of [15] and its advantage over L_2 approach. He also described an exponential density estimation approach and showed its many interesting properties by exploiting its relation to the well-known family of exponential distributions.

(Received October 18, 2010)

REFERENCES

[1] S. Abou-Jaoude: Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité. Ann. Inst. H. Poincaré *XII* (1976), 213–231.

[2] O. Barndorff-Nielsen: Information and Exponential Families in Statistical Theory. Wiley, 1978.

[3] A. R. Barron, L. Györfi, and E. C. van der Meulen: Distribution estimation consistent in total variation and two types of information divergence. IEEE Trans. Inform. Theory *38* (1992), 1437–1454.

[4] N. N. Cencov: Estimation of unknown density function from observations. (in Russian) Trans. SSSR Acad. Sci. *147* (1962), 45–48.

[5] N. N. Cencov: Categories of mathematical statistics. (in Russian) Trans. SSSR Acad. Sci. *164* (1965), 511–514.

[6] N. N. Cencov: General theory of exponential families of distribution functions. Theory Probab. Appl. *11* (1966), 483–494.

[7] N. N. Cencov: Asymmetric distance between distribution functions, entropy and Pithagoras theorem. (in Russian) Math. Notes *4* (1968), 323–332.

[8] N. N. Cencov: Statistical Decision Rules and Optimal Inference. (in Russian) Nauka, Moscow 1972.

[9] N. N. Cencov: Algebraic foundation of mathematical statistics. Math. Operationsforsch. Statist., Ser. Statistics *9* (1978), 267–276.

[10] N. N. Cencov: On basic concepts of mathematical statistics. Banach Center Publ. *6* (1980), 85-94.

[11] N. N. Cencov: On correctness of the pointwise estimation problem. (in Russian) Theory Probab. Appl. *26* (1981) 15–31.

[12] I. Csiszár and J. Fischer: Informationsentfernungen im Raum der Wahscheinlichkeitsverteilungen. Publ. Math. Inst. Hungar. Acad. Sci. *7* (1962), 159–180.

[13] I. Csiszár: Information-type measures of divergence of probability distributions and indirect observations. Studia Sci. Math. Hungar. *2* (1967), 299–318.

[14] I. Csiszár: On topological properties of f-divergence. Studia Sci. Math. Hungar. *2* (1967), 329–339.

[15] L. Devroye and L. Györfi: Nonparametric Density Estimation: The L_1 View. Wiley, 1985. Russian translation: Mir, Moscow, 1988 (Translated from English to Russian by A. Tsybakov).

[16] L. Devroye and L. Györfi: No empirical measure can converge in the total variation sense for all distribution. Ann. Statist. *18* (1990), 1496–1499.

[17] A. S. Frolov and N. N. Cencov: Application of dependent observations in the Monte Carlo method for recovering smooth curves. (in Russian) In: Proc. 6th Russian Conference on Probability Theory and Mathematical Statistics, Vilnus 1962, pp. 425–437.

[18] L. Györfi, I. Páli, and E. C. van der Meulen: There is no universal source code for infinite alphabet. IEEE Trans. Inform. Theory *40* (1994), 267–271.

[19] L. Györfi, I. Páli, and E. C. van der Meulen: On universal noiseless source coding for infinite source alphabets. Europ. Trans. Telecomm. *4* (1993), 9–16.

[20] J. A. Hartigan: The likelihood and invariance principles. Annals Math. Statist. *38* (1967), 533–539.

[21] I. A. Ibragimov and R. Z. Hasminski: On estimation of density. (in Russian) Scientific Notes of LOMI Seminars *98* (1980), 61–86.

[22] P. Kafka, F. Österreicher, and I. Vincze: On powers of f-divergences defining a distance. Studia Sci. Math. Hungar. *26* (1991), 415–422.

[23] J. H. B. Kemperman: An optimum rate of transmitting information. Ann. Math. Statist. *40* (1969), 2156–2177.

[24] M. Khosravifard, D. Fooladivanda, and T. A. Gulliver: Confliction of the convexity and metric properties in f-divergences. IEICE Trans. Fundamentals *E90-A* (2007), 1848–1853.

[25] A. L. Kolmogorov: Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari *4* (1933), 83-91.

[26] T. A. Kriz and J. V. Talacko: Equivalence of the maximum likelihood estimator to a minimum entropy estimator. Trab. Estadist. Invest. Oper. *19* (1968), 55-65.

[27] S. Kullback: A lower bound for discrimination in terms of variation. IEEE Trans. Inform, Theory *13* (1967), 126–127.

[28] S. Kullback: Correction to "A lower bound for discrimination in terms of variation". IEEE Trans. Inform. Theory *16* (1970), 652.

[29] N. Morse and R. Sacksteder: Statistical isomorphism. Ann. Math. Statist. *37* (1966), 203–214.

[30] L. LeCam: On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. Univ. Calif. Publ. Statist. *1* (1953), 267–329.

[31] F. Liese and I. Vajda: Convex Statistical Distances. Teubner, Leipzig 1987.

[32] E. A. Morozova and N. N. Cencov: Markov maps in noncommutative probability theory and mathematical statistics. (in Russian) In: Proc. 4th Internat. Vilnius Conf. Probability Theory and Mathematical Statistics, VNU Science Press *2* (1987), pp. 287–310.

[33] E. A. Nadaraya: On nonparametric estimation of Bayes risk in classification problems. (in Russian) Trans. Georgian Acad. Sci. *82* (1976), 277–280.

[34] E. A. Nadaraya: Nonparametric Estimation of Probability Density and Regression Curve. (in Russian) Tbilisi State University, Georgia 1983.

[35] F. Österreicher and I. Vajda: A new class of metric divergences on probability spaces and its statistical applications. Ann. Inst. Statist. Math. *55* (2003), 639–653.

[36] I. M. Sobol: Multidimensional Quadratic Formulas and Haar Functions. (in Russian) Nauka, Moscow 1969.

[37] W. W. Statulavicius: On Some Asymptotic Properties of Minimax Density Estimates. (in Russian) PhD. Thesis, Vilnus State University 1986.

[38] R. L. Stratonovich: Rate of convergence of probability density estimates. (in Russian) Trans. SSSR Acad. Sci., Ser. Technical Cybernetics *6* (1969), 3–15.

[39] G. T. Toussaint: Sharper lower bounds for information in term of variation. IEEE Trans. Inform. Theory *21* (1975), 99–103.

[40] I. Vajda: Note on discrimination information and variation. IEEE Trans. Inform. Theory *IT-16* (1970), 771–773.

[41] I. Vajda: On the f-divergence and singularity of probability measures. Period. Math. Hungar. *2* (1972), 223–234.

[42] I. Vajda: On metric divergences of probability measures. Kybernetika *45* (2009), 885–900.

[43] A. Wald: Contributions to the theory of statistical estimation and testing hypotheses. Ann. Math. Statist. *10* (1939), 299–326.

László Györfi, Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Magyar Tudósok körútja 2., Budapest, H-1117. Hungary.

e-mail: gyorfi@szit.bme.hu

Adam Krzyżak, Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, H3G 1M8. Canada.

e-mail: krzyzak@cs.concordia.ca