Information Theoretic Methods in Probability and Statistics*

I. Csiszár, Budapest

Abstract

Ideas of information theory have found fruitful applications not only in various fields of science and engineering but also within mathematics, both pure and applied. This is illustrated by several typical applications of information theory specifically in probability and statistics.

1 Introduction

In its early years, information theory (IT) "has perhaps been ballooned to an importance beyond its actual accomplishments" (Shannon 1956), being applied "to biology, psychology, linguistics, fundamental physics, economics, the theory of organizations, and many others." While criticizing these superficial applications, Shannon did believe that serious applications of IT concepts in other fields were forthcoming, "indeed, some results are already quite promising – but the establishing of such applications is not a trivial matter ... but rather the slow and tedious process of hypothesis and experimental verification." Shannon (loc. cit.) also emphasized that "the hard core of IT is, essentially, a branch of mathematics" and "a thorough understanding of the mathematical foundation ... is surely a prerequisite to other applications."

As "the hard core of IT is a branch of mathematics," one could expect a natural *two-way interaction* of IT with other branches of mathematics that, in addition to enriching IT, also leads to significant applications of IT ideas *within mathematics.* Indeed, such applications had already been around in 1956, such as Kullback's information theoretic approach to statistics, and others were to follow soon. A celebrated example (Kolmogorov 1958) was to use the IT fact that stationary coding does not increase entropy rate to show that stationary processes of different entropy rate are never isomorphic in the sense of ergodic theory. This demonstrated that not all i.i.d. processes are mutually isomorphic, solving a long-standing problem. Kolmogorov's work initiated spectacular developments in ergodic theory, and entropy became a basic concept in that field.

The times when some scientists regarded IT as a panacea have long passed, but today's information theorists are proudly aware of well established and substantial applications of their discipline

*This work has been partially supported by the Hungarian National Foundation for Scientific Research, Grant T016386.

in quite a few other ones. This author, a mathematician, is particularly fascinated by the many applications of IT in various branches of pure and applied mathematics, including combinatorics, ergodic theory, algebra, operations research, systems theory, and perhaps primarily probability and statistics. The goal of this paper is to give a flavor of such applications, surveying some typical ones specifically in probability and statistics.

For the applications treated in this paper, the main IT tools are the properties of information measures, the method of types, and the concept of coding. Applications of IT in probability will be treated in Section 2, and those in statistics in Section 3.

1.1 Preliminaries on I-divergence

The I-divergence (information divergence, also called relative entropy or Kullback-Leibler distance) of probability distributions (PD's) P, Q on a finite set $\mathcal{X}$ is defined as

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} \tag{1.1}$$

(in this paper, we use natural logarithms). The I-divergence of PD's on an arbitrary measurable space $(\mathcal{X},\mathcal{F})$, i.e., of probability measures P, Q on $(\mathcal{X},\mathcal{F})$, is defined as

$$D(P\|Q) = \sup_{\mathcal{A}} D(P^{\mathcal{A}}\|Q^{\mathcal{A}}), \tag{1.2}$$

the sup taken for all $\mathcal{F}$-measurable partitions $\mathcal{A} = (A_1,\dots,A_k)$ of $\mathcal{X}$. Here $P^{\mathcal{A}}$ denotes the $\mathcal{A}$-quantization of P defined as the PD $P^{\mathcal{A}} = (P(A_1),\dots,P(A_k))$ on $\{1,\dots,k\}$. A well known integral formula for $D(P\|Q)$ is

$$D(P\|Q) = \int p(x)\log\frac{p(x)}{q(x)}\lambda(dx) \tag{1.3}$$

where $p(x)$ and $q(x)$ are the densities of P and Q with respect to an arbitrary dominating measure λ.

I-divergence is a (non-symmetric) information theoretic measure of distance of P from Q. A key property is that $D(P\|Q)\geq 0$, with equality iff $P=Q$. A stronger property known as Pinsker's inequality is

$$|P-Q| \leq \sqrt{2D(P\|Q)} \tag{1.4}$$

where

$$|P-Q| = \int |p(x)-q(x)|\lambda(dx) \tag{1.5}$$

is the variation distance of P and Q.

While not a true metric, I-divergence is in many respects an analogue of squared Euclidean distance. In particular, if Π is a convex set of PD's and the minimum of $D(P\|Q)$ subject to $P \in \Pi$ is attained then the minimizer P^*, called the I-projection of Q onto Π, is unique and

$$D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q) \quad \text{for each} \quad P \in \Pi \tag{1.6}$$

(Csiszár 1975). If Π is defined by a finite number of linear constraints then (1.6) holds with equality. This is an analogue of the Pythagorean theorem, while (1.6) is an analogue of the cosine theorem in Euclidean geometry.

2 Applications of IT in probability

2.1 Gaussian measures dichotomy theorem

Let $\mathcal{X}$ consist of functions $x(t)$, $t \in T$, and let $\mathcal{F}$ be the σ-algebra spanned by the cylinder sets. A PD P on $(\mathcal{X}, \mathcal{F})$ is called a Gaussian measure if all of its finite dimensional distributions $P_{t_1 \dots t_k}$ are Gaussian. Here, for $\{t_1, \dots, t_k\} \subset T$, $P_{t_1 \dots t_k}$ is the image of P under the mapping $x(\cdot) \to (x(t_1), \dots, x(t_k))$.

The dichotomy theorem says that two Gaussian measures P and Q on $(\mathcal{X}, \mathcal{F})$ are either equivalent (mutually absolutely continuous) or orthogonal (singular detection is possible: there exists $A \in \mathcal{F}$ with $P(A) = 1$, $Q(A) = 0$).

The first proof of this important result was obtained via IT (Hájek 1958), by showing that for Gaussian measures, $D(P\|Q) + D(Q\|P) = \infty$ implies orthogonality. Of course, $D(P\|Q) + D(Q\|P) < \infty$ implies equivalence, even if P and Q are non-Gaussian.

Sketch of Hájek's proof:

"(i)". $D(P\|Q) = \sup_{\{t_1, \dots, t_k\}} D(P_{t_1 \dots t_k} \| Q_{t_1 \dots t_k})$
(an easy consequence of eq. (1.2)).

"(ii)". I-divergence is invariant under one-to-one transformations, this permits us to reduce calculation of $D(P_{t_1 \dots t_k} \| Q_{t_1 \dots t_k})$ to the "easy case" when $P_{t_1 \dots t_k}$ is (k-dimensional) standard Gaussian and also $Q_{t_1 \dots t_k}$ is of product form.

"(iii)". In the "easy case" above, direct calculation shows that if $D(P_{t_1 \dots t_k} \| Q_{t_1 \dots t_k}) + D(Q_{t_1 \dots t_k} \| P_{t_1 \dots t_k})$ is "large" then $P_{t_1 \dots t_k}$ and $Q_{t_1 \dots t_k}$ are "almost concentrated on disjoint sets".

2.2 Large deviations: Sanov's theorem, Gibbs' conditioning principle

Let $X_1, X_2, \ldots$ be i.i.d. random variables with values in an arbitrary set $\mathcal{X}$ (equipped with a σ-algebra $\mathcal{F}$), with common distribution Q. The empirical distribution $\widehat{P}_n$ of $X^n = (X_1, \ldots, X_n)$ is the random probability measure on $\mathcal{X}$ defined by $\widehat{P}_n(A) = \frac{1}{n}|\{i\colon X_i \in A\}|$. Sanov's theorem (Sanov 1957) says, intuitively, that for any PD $P \neq Q$,

$$Pr\{\widehat{P}_n \text{ is close to } P\} \sim \exp\{-nD(P\|Q)\}. \tag{2.1}$$

If $\mathcal{X}$ is a finite set, we have more exactly

$$Pr\{\widehat{P}_n = P\} = \exp\{-nD(P\|Q) + O(\log n)\} \tag{2.2}$$

whenever P is a possible type for block-length n (a basic fact of the method of types, cf. Csiszár and Körner 1981, p.32). Clearly, (2.2) implies that

$$\lim_{n\to\infty} \frac{1}{n} \log Pr\{\widehat{P}_n \in \Pi\} = -\inf_{P\in\Pi} D(P\|Q) \tag{2.3}$$

for any set Π of PD's on $\mathcal{X}$ in which n-types become dense as $n \to \infty$.

For arbitrary $\mathcal{X}$, a general form of Sanov's theorem says that for any set Π of PD's on $\mathcal{X}$ for which the probabilities $Pr\{\widehat{P}_n \in \Pi\}$ are defined, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log Pr\{\widehat{P}_n \in \Pi\} \geq -\inf_{P\in\Pi^\circ} D(P\|Q) \tag{2.4}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log Pr\{\widehat{P}_n \in \Pi\} \leq -\inf_{P\in\overline{\Pi}} D(P\|Q). \tag{2.5}$$

Here Π° and $\overline{\Pi}$ denote the interior and closure of Π in the τ-topology, i.e., the topology of setwise convergence of probability measures (the τ-topology is weaker than the topology of variation distance). Of course, the equality of the right hand sides of (2.4) and (2.5) is a sufficient condition for the limit relation (2.3). In particular, (2.3) always holds if Π is convex and $D(P\|Q) < \infty$ for some $P \in \Pi^\circ$.

Of particular interest is the choice

$$\Pi = \{P\colon \int f dP \in C\} \tag{2.6}$$

where f is a given (possibly vector valued) function on $\mathcal{X}$ and C is a given subset of the range of f; then $\widehat{P}_n \in \Pi$ means that $\frac{1}{n}\sum_{i=1}^{n} f(X_i) \in C$.

In the parlance of large deviations theory (cf. Dembo and Zeitouni 1993), the asymptotic bounds (2.4), (2.5), together with the easily checked compactness (in τ-topology) of the divergence balls

$$B(P,a) = \{P' \colon D(P'\|P) \le a\}, \tag{2.7}$$

mean that $\widehat{P}_n$ satisfies the large deviation principle, with good rate function $D(P\|Q)$.

The simplest available proof of this important result is inherently information theoretic (Groeneboom, Oosterhoff and Ruymgaart 1979; they acknowledge using a suggestion of this author in proving (2.8) below). We sketch the proof of the more difficult part (2.5).

For any (measurable) partition $\mathcal{A} = (A_1, \dots, A_k)$ of $\mathcal{X}$, $Pr\{\widehat{P}_n \in \Pi\}$ is bounded above by the sum of probabilities $Pr\{\widehat{P}_n^{\mathcal{A}} = P'\}$ for all PD's P' on $\{1, \dots, k\}$ such that $P' = P^{\mathcal{A}}$ for some $P \in \Pi$ and P' is an n-type. This implies by (2.2) that

$$Pr\{\widehat{P}_n \in \Pi\} \le \exp\{-n \inf_{P \in \Pi} D(P^{\mathcal{A}}\|Q^{\mathcal{A}}) + O(\log n)\}.$$

This gives

$$\limsup_{n \to \infty} \frac{1}{n} \log Pr\{\widehat{P}_n \in \Pi\} \le -\sup_{\mathcal{A}} \inf_{P \in \Pi} D(P^{\mathcal{A}}\|Q^{\mathcal{A}}),$$

and (2.5) follows by (1.2) if one checks that

$$\sup_{\mathcal{A}} \inf_{P \in \Pi} D(P^{\mathcal{A}}\|Q^{\mathcal{A}}) = \inf_{P \in \overline{\Pi}} \sup_{\mathcal{A}} D(P^{\mathcal{A}}\|Q^{\mathcal{A}}). \tag{2.8}$$

The latter is non-trivial but not too hard.

Suppose next that Π is a convex set of PD's on $\mathcal{X}$ and the conditional distribution of X_1 on the condition $\widehat{P}_n \in \Pi$ belongs to Π. This is always the case for Π given by (2.6) if C is a convex set and $Ef(X_1)$ exists. Under the above hypotheses, the asymptotic bound (2.5) may be sharpened to a non-asymptotic bound. More importantly, under mild additional hypotheses, a strong form of Gibbs' conditioning principle (cf. Dembo and Zeitouni 1993) may be established by a simple IT reasoning (Csiszár 1984).

The following identity

$$D(\widetilde{P}\|Q^n) = D(\widetilde{P}\|P^n) + nD(P\|Q) \tag{2.9}$$

holds for any PD $\widetilde{P}$ on $\mathcal{X}^n$ whose marginals on $\mathcal{X}$ are equal to P, and for any PD Q on $\mathcal{X}$. Take here $\widetilde{P} = P_{X^n|\widehat{P}_n \in \Pi}$ (the conditional distribution of X^n on the condition $\widehat{P}_n \in \Pi$), then $P = P_{X_1|\widehat{P}_n \in \Pi} \in \Pi$ by assumption. It follows that

$$-\log Pr\{\widehat{P}_n \in \Pi\} = D(\widetilde{P}\|Q^n) = D(\widetilde{P}\|P^n) + nD(P\|Q) \ge n \inf_{P \in \Pi} D(P\|Q) \tag{2.10}$$

proving the claimed non-asymptotic bound. If P^* is the I-projection of Q onto Π, from (2.10) and (1.6) we further obtain

$$-\log Pr\{\widehat{P}_n \in \Pi\} \quad \geq D(\widetilde{P} \| P^n) + nD(P \| P^*) + nD(P^* \| Q) = D(\widetilde{P} \| P^{*n}) + nD(P^* \| Q), \tag{2.11}$$

where the last step follows from (2.9) with Q replaced by P^*.

Supposing finally that the given Π satisfies (2.3), it follows from (2.11) – recalling $\widetilde{P} = P_{X^n | \widehat{P}_n \in \Pi}$ – that

$$\frac{1}{n} D(P_{X^n | \widehat{P}_n \in \Pi} \| P^{*n}) \to 0 \quad \text{as } n \to \infty. \tag{2.12}$$

This says, intuitively, that conditionally on $\widehat{P}_n \in \Pi$, the random variables $X_1, \ldots, X_n$ "almost behave" as independent ones with common distribution P^*. In particular, (2.12) implies that for any fixed k

$$D(P_{X^k | \widehat{P}_n \in \Pi} \| P^{*k}) \to 0 \quad \text{as } n \to \infty.$$

Thus the conditional joint distribution of $X_1, \ldots, X_k$ on the condition $\widehat{P}_n \in \Pi$ converges to the product distribution P^{*k} in a stronger sense than in variation distance, cf. (1.4). This is the promised strong version of Gibbs' conditioning principle.

The above result may be extended to the case when the I-projection of Q onto Π does not exist. Namely, a unique P^* (not necessarily in Π) always exists such that $D(P_n \| Q) \to \inf_{P \in \Pi} D(P \| Q)$ with $P_n \in \Pi$ implies $D(P_n \| P^*) \to 0$; then (2.12) holds with this "generalized I-projection" P^* (Csiszár 1984).

2.3 Measure concentration

Measure concentration is currently a hot topic in probability theory (cf. Talagrand 1995, 1996). A general description of problems pertinent to that topic would lead too far, but one result now considered as an early example of a measure concentration theorem is well known to information theorists. It is the blowing up lemma (Margulis 1974, Ahlswede, Gács and Körner 1976) that says, intuitively, that by slightly "blowing up" any set $A \subset \mathcal{X}^n$ of not exponentially small probability, one gets a set of probability close to 1.

In IT, the blowing up lemma was originally just a mathematical tool, particularly useful in multiuser Shannon theory (cf. Csiszár and Körner 1981). Today it is an integral part of IT due to its information theoretic proof (Marton 1986) that was, actually, its first simple proof. Recently, Marton extended her approach to prove also other measure concentration results, providing beautiful examples of applications of IT in probability theory. Here we sketch some main ideas of that approach, restricting attention to the simplest case.

Let $Q_1, Q_2, \ldots$ be PD's on a finite set $\mathcal{X}$, let $Q^{(n)} = Q_1 \times \ldots \times Q_n$, and let

$$A^{\varepsilon} = \{y^n \colon d(x^n, y^n) \leq \varepsilon \quad \text{for some} \quad x^n \in A\} \tag{2.13}$$

denote the ε-blow up of a set $A \subset \mathcal{X}^n$ (where d denotes normalized Hamming distance).

Marton (1996) proved that the distance

$$d(A, B) = \min\{d(x^n, y^n) \colon x^n \in A,\, y^n \in B\} \tag{2.14}$$

of two subsets of $\mathcal{X}^n$ can be bounded in terms of their probabilities as

$$d(A, B) \leq \sqrt{\frac{1}{2n} \log \frac{1}{Q^{(n)}(A)}} + \sqrt{\frac{1}{2n} \log \frac{1}{Q^{(n)}(B)}}. \tag{2.15}$$

With the choice $B = (A^{\varepsilon})^c$, when $d(A, B) \geq \varepsilon$, (2.15) gives the following strong version of the blowing up lemma:

$$Q^{(n)}(A^{\varepsilon}) \geq 1 - \exp\left\{-2n \left[\varepsilon - \sqrt{\frac{1}{2n} \log \frac{1}{Q^{(n)}(A)}}\,\right]\right\}. \tag{2.16}$$

Sketch of proof of (2.15):

(i) The key idea is to show that given $Q^{(n)} = Q_1 \times \ldots \times Q_n$ and any PD $P^{(n)}$ on $\mathcal{X}^n$ not necessarily of product form, there exist random variables $X^n = (X_1, \ldots, X_n)$ with distribution $Q^{(n)}$ and $Y^n = (Y_1, \ldots, Y_n)$ with distribution $P^{(n)}$ such that

$$Ed(X^n, Y^n) \leq \sqrt{\frac{1}{2n} D(P^{(n)} \| Q^{(n)})}. \tag{2.17}$$

For $n = 1$, this is obvious by Pinsker's inequality (1.4), since the minimum of $Pr\{X \neq Y\}$ subject to $P_X = Q$, $P_Y = P$ equals $\frac{1}{2}|P - Q|$.

For $n > 1$, the proof of (2.17) goes by induction, using a coupling argument to extend X^n and Y^n satisfying the induction hypothesis by suitable new components X_{n+1}, Y_{n+1}.

(ii) The result (i) implies, by the triangle inequality, that given arbitrary PD's $P^{(n)}$ and $\widetilde{P}^{(n)}$ on $\mathcal{X}^n$, there exist Y^n and $\widetilde{Y}^n$ with distributions $P^{(n)}$ and $\widetilde{P}^{(n)}$ such that

$$Ed(Y^n, \widetilde{Y}^n) \leq \sqrt{\frac{1}{2n} D(P^{(n)} \| Q^{(n)})} + \sqrt{\frac{1}{2n} D(\widetilde{P}^{(n)} \| Q^{(n)})}. \tag{2.18}$$

(iii) Finally, (2.15) follows from (2.18), letting $P^{(n)}$ and $\widetilde{P}^{(n)}$ be the conditional distributions obtained from $Q^{(n)}$ by conditioning on A and B, respectively. Indeed, with that choice, the left hand side of (2.18) is lower bounded by $d(A, B)$ since $d(Y^n, \widetilde{Y}^n) \geq d(A, B)$ with probability 1.

This approach to derive bounds like (2.15) has been extended to distributions of certain processes with memory in the role of $Q^{(n)}$, including mixing Markov chains (Marton 1996). So far, other approaches to measure concentration could not be so extended. It should be noted that a weaker asymptotic form of the blowing up lemma has been established for a rather broad class of processes, again with substantial use of IT ideas (Marton and Shields 1994).

2.4 Other topics

There are many other applications of IT to probability that can not be covered here.

Let us just mention that various limit theorems of probability theory have been given information theoretic proofs. These include:
Central limit theorem (Linnik 1959, Barron 1986);
Ergodicity of Markov chains (Rényi 1961, Kendall 1963, Fritz 1973);
Limit theorem for the convolution powers of a PD on a topological group (Csiszár 1965).

A promising recent idea is to prove bounds for recurrence and matching problems, utilizing the non-existence of codes beating the entropy bound (Shields 1996, Section II.5).

3 Information theoretic methods in statistics

Statistics, the science of extracting information from data, appears the most natural field of applications of IT, besides communication theory. Historically, an information measure had been used by statisticians prior to Shannon (Fisher's information, Fisher 1925). I-divergence was first explicitly introduced for purposes of statistics, though motivated by Shannon's work (Kullback and Leibler 1951). Implicitly it had played a role also in earlier statistical works (Wald 1947, Good 1950), and Kullback soon developed a unified approach to testing statistical hypotheses based on this information measure (Kullback 1959).

Several results considered in retrospect as applications of IT in statistics were actually established by statisticians independently of IT. "Although Wald did not explicitly mention information in his treatment of sequential analysis, it should be noted that his work must be considered a major contribution to the statistical applications of IT" (Kullback 1959, p.2). This author shares this view, and he also considers the results in Subsection 3.2 below as applications of a typical IT tool, viz. the method of types. The proof of these results, however, preceded the development of the method of types in IT; indeed, it represented one of the origins of that method. Some would prefer to speak in this context about interplay of statistics and IT rather than statistical applications of IT.

There are two major inference methods motivated by IT: The methods of maximizing entropy (or minimizing I-divergence) and of minimizing "description length." Their coverage is impossible here, for lack of space, but perhaps not necessary, either, since most information theorists have at least some familiarity with these methods. We will but illustrate them by simple examples.

3.1 Early results

Let us start with Wald's inequality relating the expected sample size of a sequential test to the type 1 and type 2 error probabilities.

Assuming i.i.d. sampling from a distribution known to be either P or Q , a sequential test accepts one of these hypotheses on the basis of a sample $X^N = (X_1, \ldots, X_N)$ of random length N. Here N is a stopping time, i.e., knowledge of $X_1, \ldots, X_n$ determines whether or not $N = n$. Wald (1947) proved that

$$E_P(N)D(P\|Q) \geq (1-\alpha)\log\frac{1-\alpha}{\beta} + \alpha\log\frac{\alpha}{1-\beta} \tag{3.1}$$

where α is the probability under P of accepting Q and β is the probability under Q of accepting P. Moreover, Wald showed that his sequential probability ratio test nearly attains the equality in (3.1).

The IT interpretation makes this result easy to understand: Denoting by P^N and Q^N the distribution of X^N under P and Q, the left hand side of (3.1) equals $D(P^N\|Q^N)$ (this can be checked using Wald's identity), whereas the right hand side is the I-divergence of the $\mathcal{A}$-quantizations of P^N and Q^N for $\mathcal{A} = (A_1, A_2)$, A_1 and A_2 being the acceptance regions of P and Q. Were the likelihood ratio constant on both A_1 and A_2, the equality would hold in (3.1). While no test can achieve this exactly, in general, the sequential probability ratio test comes close.

Another early result in statistical IT is the celebrated "Stein's lemma" (Chernoff 1952; Stein apparently disowns it). It provides an operational meaning to I-divergence: For testing a simple hypothesis P against a simple alternative Q, the best test of sample size n and type 1 error probability $\leq \varepsilon$ (for any $0 < \varepsilon < 1$) has type 2 error probability $\exp\{-nD(P\|Q)+o(n)\}$. Notice that if the type 1 error were required to go to zero, rather than just $\leq \varepsilon$, the special case $N = \text{const} = n$ of Wald's inequality (3.1) would already imply that the type 2 error probability exponent can not exceed $D(P\|Q)$.

3.2 Hypothesis testing: exponential rate optimal tests

Let $\mathcal{X}$ be a finite set and P a given PD on $\mathcal{X}$. Suppose the null-hypothesis that an i.i.d. sample of size n comes from P is to be tested; the alternative hypothesis Q is not specified.

Then the test that accepts the null-hypothesis when the empirical distribution $\widehat{P}_n$ of the sample belongs to the divergence ball $B(P,a) = \{P' : D(P'\|P) \le a\}$, is universally exponential rate optimal in the following sense.

The probability of type 1 error goes to 0 exponentially, with exponent a, and for any alternative hypothesis Q such that $b(P,Q,a)$ below is positive, the probability of type 2 error goes to zero with exponent

$$b = b(P,Q,a) = \min_{P' \in B(P,a)} D(P'\|Q). \tag{3.2}$$

Even if the alternative hypothesis Q were specified, no tests with type 1 error probability exponent $\ge a$ could have type 2 error probability that decreases with a larger exponent than $b(P,Q,a)$; in particular, against an alternative Q with $b(P,Q,a) = 0$, an exponential decrease of type 2 error is not achievable.

Also of interest is the modification of the above test replacing the constant a by a sequence $a_n \to 0$ such that $\frac{n}{\log n} a_n \to \infty$. Then the type 1 error probability still goes to 0 (though no longer exponentially), and the probability of type 2 error against any alternative Q goes to zero with exponent $D(P\|Q)$. The latter is best possible, by Stein's lemma.

The above results are due to Hoeffding (1965). For today's information theorists, their proof is an easy exercise in the method of types (cf. Csiszár and Körner 1981, p.44).

The extension to testing composite hypotheses is straightforward. To test the null-hypothesis that the true distribution belongs to a given set Π of PD's on $\mathcal{X}$, take the union of the divergence balls $B(P,a)$, $P \in \Pi$, and accept the null-hypothesis when $\widehat{P}_n$ is in that union. Then the type 1 error probability still goes to 0 with exponent a, and for any (simple) alternative Q, the type 2 error probability exponent will be the infimum of $b(P,Q,a)$ subject to $P \in \Pi$, the best possible.

Notice that the acceptance criterion $\widehat{P}_n \in \cup_{P\in\Pi} B(P,a)$, i.e., $\inf_{P\in\Pi} D(\widehat{P}_n\|P) \le a$, is equivalent to

$$\frac{\sup_{P\in\Pi} \prod P(x)^{n\widehat{P}_n(x)}}{\prod \widehat{P}_n(x)^{n\widehat{P}_n(x)}} \ge \exp(-na).$$

Thus the above universally rate optimal tests are what statisticians call likelihood ratio tests.

Consider next hypothesis testing for PD's on an arbitrary set $\mathcal{X}$ (equipped with a σ-algebra $\mathcal{F}$). Then the previous acceptance criterion does not make sense, as for a continuous distribution

P always $D(\widehat{P}_n \| P) = \infty$. One way out is to consider a refining sequence of partitions $\mathcal{A}_n = (A_{n1}, \ldots, A_{nm(n)})$ of $\mathcal{X}$ that generates $\mathcal{F}$, and accept the null-hypothesis that the true distribution belongs to a given set Π of PD's when $\inf_{P \in \Pi} D(\widehat{P}_n^{\mathcal{A}_n} \| P^{\mathcal{A}_n}) \leq a$. Assuming that $m(n) = o(n)$ when the number $\begin{pmatrix} n + m(n) - 1 \\ m(n) - 1 \end{pmatrix}$ of n-types for alphabet size $m(n)$ is $\exp(o(n))$, the type 1 error probability of this test is still $\leq \exp(-na + o(n))$, while the type 2 error probability (against any given Q) will be $\leq \exp(-nb_n + o(n))$ where

$$b_n = \inf_{P \in \Pi} \min_{P' : D(P'^{\mathcal{A}_n} \| P^{\mathcal{A}_n}) \leq a} D(P'^{\mathcal{A}_n} \| Q^{\mathcal{A}_n}). \tag{3.3}$$

This approach is due to Tusnády 1977. He showed under a compactness assumption on Π that b_n in (3.3) approaches $\inf_{P \in \Pi} b(P, Q, a)$, and then the above test is universally exponential rate optimal. In particular, rate optimality always holds when the null-hypothesis is simple.

Notice the relationship of the results in this Subsection to those in Subsection 2.2.

3.3 Iterative scaling, EM algorithm

Iterative scaling is a familiar procedure to infer a matrix with non-negative entries p_{ij} when the row and column sums

$$p_{i\cdot} = \sum_j p_{ij} \qquad p_{\cdot j} = \sum_i p_{ij} \tag{3.4}$$

are known, say $p_{i\cdot} = \overline{p}_i$, $p_{\cdot j} = \overline{\overline{p}}_j$, and a prior guess $p_{ij}^0 = q_{ij}$ is available. The procedure consists in iteratively adjusting the row and column sums, setting

$$p_{ij}^k = \begin{cases} p_{ij}^{k-1} \frac{p_{i\cdot}}{\overline{p}_i} & k \text{ odd} \\ p_{ij}^{k-1} \frac{p_{\cdot j}}{\overline{\overline{p}}_i} & k \text{ even} \end{cases} \tag{3.5}$$

and taking the limit

$$p_{ij}^* = \lim_{k \to \infty} p_{ij}^k. \tag{3.6}$$

To this author's knowledge, iterative scaling was first used (Kruithof 1937) to estimate telephone traffic, viz. the number p_{ij} of calls from exchange i to exchange j on a given day, from the counts $p_{i\cdot}$ and $p_{\cdot j}$ of outgoing and incoming calls, and from the exact knowledge of the traffic q_{ij} on some previous day. In statistics, the same procedure was first proposed by Deming and Stephan 1943 to infer a two dimensional distribution $P = (p_{ij})$ with known marginals $\overline{P} = (p_{i\cdot})$, $\overline{\overline{P}} = (p_{\cdot j})$, using the empirical distribution of the observed sample ("contingency table") as prior guess Q.

The IT nature of this procedure has been recognized by Ireland and Kullback 1968. Suppose w.l.o.g. that $\overline{P}, \overline{\overline{P}}$ and Q are PD's, let Π_1 and Π_2 denote the set of (two-dimensional) PD's with first marginal $\overline{P}$, respectively with second marginal $\overline{\overline{P}}$. Then P^k belongs to Π_1 or Π_2 according as k is odd or even, and

$$D(P\|P^{k-1}) = D(P\|P^k) + D(P^k\|P^{k-1}) \tag{3.7}$$

for each $P \in \Pi_1$ (k odd) or $P \in \Pi_2$ (k even). Due to this Pythagorean identity, P^k is the I-projection of P^{k-1} onto Π_1 or Π_2, respectively.

We show (following Csiszár 1975, filling a gap in the proof of Ireland and Kullback 1968) that if $\Pi_1 \cap \Pi_2 \cap \{P\colon D(P\|Q) < \infty\} \neq \emptyset$, the limits (3.6) exist and $P^* = (p^*_{ij})$ equals the I-projection of Q onto $\Pi_1 \cap \Pi_2$.

By the uniqueness of I-projection, it suffices to show that for any convergent subsequence $P^{n_i} \to P'$, say, we have $P' \in \Pi_1 \cap \Pi_2$ and for each $P \in \Pi_1 \cap \Pi_2$

$$D(P\|Q) = D(P\|P') + D(P'\|Q). \tag{3.8}$$

Since (3.7) holds for each k if $P \in \Pi_1 \cap \Pi_2$, summing these identities from $k = 1$ to n_i and letting $i \to \infty$ gives

$$D(P\|Q) = D(P\|P') + \sum_{k=1}^{\infty} D(P^k\|P^{k-1}). \tag{3.9}$$

If $D(P\|Q) < \infty$, (3.9) implies that $D(P^k\|P^{k-1}) \to 0$, consequently $\lim P^{n_i+1} = \lim P^{n_i} = P'$, establishing $P' \in \Pi_1 \cap \Pi_2$. Thus (3.9) applies to $P = P'$, yielding that the sum in (3.9) equals $D(P'\|Q)$. This completes the proof.

A similar convergence result had been claimed (Kullback 1968) also for iterative scaling of densities. That case, however, is much harder; the above proof essentially relies upon the continuity of I-divergence as a function of its second variable, and this no longer holds if the underlying set is infinite. A convergence proof for iterative scaling of densities appears available under additional assumptions only (Rüschendorf 1995).

There is a large variety of problems requiring to infer a PD or, more generally, a non-negative valued function, when the available information consists in certain linear constraints. The popular "maximum entropy" method suggests to take the feasible P closest in I-divergence to a default model Q, i.e., the I-projection of Q onto the feasible set of P's satisfying the given constraints. The I-divergence of non-negative valued functions P and Q, not necessarily PD's, on a finite set $\mathcal{X}$, is defined by the following extension of eq. (1.1):

$$D(P\|Q) = \sum \left[P(x) \log \frac{P(x)}{Q(x)} - P(x) + Q(x) \right]. \tag{3.10}$$

Often, the feasible set can be represented as $\Pi_1 \cap \ldots \cap \Pi_m$ such that I-projection onto each Π_i is easy to compute. Then, as in iterative scaling, the desired I-projection ("maxent solution") can be computed by iterating successive I-projections onto $\Pi_1, \ldots, \Pi_m$ starting with $P^0 = Q$. Convergence to the I-projection P^* of Q onto $\Pi_1 \cap \ldots \cap \Pi_m$ follows in the same way as above, using the Pythagorean theorem for I-projections (eq. (1.6) with equality), provided that $\Pi_1 \cap \ldots \cap \Pi_m \cap \{P \colon D(P\|Q) < \infty\} \neq \emptyset$.

Some other iterative algorithms often used in statistics and elsewhere can also be given intuitive IT interpretations. One such algorithm, designed to compute I-projection onto an arbitrary feasible set defined by linear constraints (when the underlying set is finite) is generalized iterative scaling (Darroch and Ratcliff 1972), also known as SMART algorithm (Byrne 1993). This has been shown (Csiszár 1989) equivalent to iterative I-projection performed in a suitable product space.

The so-called EM algorithm (Dempster, Laird and Rubin 1977), designed to compute maximum likelihood estimates from incomplete data, has been shown (Csiszár and Tusnády 1984) equivalent to an alternating minimization of $D(P\|Q)$ for $P \in \Pi_1$ and $Q \in \Pi_2$ with suitably constructed Π_1 and Π_2. Convergence of the latter was proved under some technical conditions, the most important being convexity of Π_1 and Π_2. The general result implies convergence of the EM algorithm in the particular case of decomposition of mixtures. Remarkably, the same general result implies also the convergence of familiar algorithms for computing channel capacity (Arimoto 1972, Blahut 1972), rate-distortion functions (Blahut 1972) and optimum portfolios (Cover 1984).

Recent works related to the topics in this Subsection include Byrne 1993, 1996, Della Pietra, Della Pietra and Lafferty 1997, Matus 1997.

3.4 Minimum description length (MDL)

MDL is a statistical inference principle motivated by IT (Rissanen 1978, 1989). It says that among various possible stochastic models (or model classes) for a data sequence $x^n = x_1 \ldots x_n$, one should select that yielding the shortest code for x^n, taking into account also the bits needed to describe the model (model class) that has been used for the encoding. MDL has naturally lead to a strong interplay with statistics of the theory of universal data compression in IT. This would deserve detailed coverage, but because of limited space we will consider just one example, the MDL approach to Markov order estimation.

For binary sequences x^n, consider the model classes "i.i.d.," "first order Markov," "second order Markov," ..., order k_0 Markov. A (prefix condition) code may be regarded optimal for a model class if the maximum over the model class of either its mean-redundancy or its max-redundancy is the smallest possible. It is known from IT that a code $f_k \colon \{0,1\}^n \to \{0,1\}^*$ optimal in either sense for the “order k Markov” model class ($k = 0$ meaning i.i.d.), has codeword length

$$\ell\left(f_k(x^n)\right) = -\log_2 P_M^{(k)}(x^n) + 2^{k-1}\log_2 n + O(1). \tag{3.11}$$

Here $P_M^{(k)}$ is the maximum of the probability of x^n for order k Markov sources. Hence, disregarding the $O(1)$ term, the order k yielding minimum codelength for x^n will be

$$k^* = \arg\max\left[\log P_M^{(k)}(x^n) - 2^{k-1}\log n\right]. \tag{3.12}$$

This k^* is taken as the MDL estimate of the Markov order k. Notice that now the description length for the model class is constant over the considered (finite number of) classes, hence it does not enter the above comparison.

Eq. (3.12) is an instance of a more general result that to chose among model classes involving different number of parameters, the criterion given by MDL is maximized log-likelihood minus the number of parameters times $\frac{1}{2}\log n$. In statistics, this is known as the BIC criterion, and it enjoys desirable properties.

It is interesting to note that a previous “penalized maximum likelihood criterion” known as AIC had also been derived using IT considerations (Akaike 1973).

3.5 Mutual information in statistics

Using IT ideas in statistics comes most naturally when adopting the Bayesian approach. Indeed, suppose the joint distribution of $X^n = (X_1, \ldots, X_n)$ depends on an unknown parameter ϑ, say $X_1, \ldots, X_n$ are i.i.d. with distribution P_ϑ. Then, in order the amount of information provided by the observation X^n about the parameter ϑ, viz. the mutual information $I(\vartheta \wedge X^n)$, be defined, it is necessary that ϑ be assigned a distribution (called prior distribution). The latter plays the role of input distribution for the channel defined by the possible distributions of X^n, corresponding to the possible values of ϑ. Of course, as an input distribution is often assigned in IT just as a technical tool, a prior can always be so assigned. For this reason, the statistical applicability of mutual information and related IT tools – such as Fano’s inequality – is by no means restricted to Bayesian statistics.

In the sixties, Rényi studied in several papers the asymptotics of $I(\vartheta \wedge X^n)$ when the set Θ of possible values of ϑ was finite, cf. Rényi 1969. He showed that $I(\vartheta \wedge X^n) \to H(\vartheta)$ exponentially fast, and related this to the asymptotic behavior of the error probability of the Bayesian (maximum a posteriori probability) estimate of ϑ.

When Θ is a subset of $\mathrm{I\!R}^k$ of positive Lebesgue measure, the mutual information $I(\vartheta \wedge X^n)$ typically goes to infinity as $n \to \infty$. Its asymptotics was studied in the seventies by Russian researchers (Pinsker 1972, Ibragimov and Hasminskii 1973, and others). Recently, Clarke and Barron 1994 obtained sharp results. In Bayesian statistics, Bernardo 1979 suggested to use a so-called reference prior selected by the IT criterion of yielding maximum $I(\vartheta \wedge X^n)$ is the limit $n \to \infty$. He argued that this criterion leads to the familiar Jeffreys prior; Clarke and Barron 1994 provide a rigorous proof, under not too restrictive hypotheses on the family $\{P_\vartheta\}$.

In many statistical problems, the parameter set Θ is infinite dimensional, e.g., it may be the set of all probability densities on $\mathbb{R}$ or $\mathbb{R}^{\mathrm{d}}$, or the class of densities satisfying some smoothness conditions. In this context, it is a good idea to consider $I(\vartheta \wedge X^n)$ for ϑ restricted to and having uniform distribution on a suitable finite subset Θ_n of Θ. Suppose the problem is to estimate $\vartheta \in \Theta$ from the observations X^n, an estimator $T(X^n)$ being evaluated by the supremum over Θ of the expected loss $E_\vartheta d(\vartheta, T(X^n))$, for a given loss function d. Subject to suitable assumptions, $E_\vartheta d(\vartheta, T(X^n))$ may be bounded below, for $\vartheta \in \Theta_n$, in terms of $P_\vartheta(T_n(X^n) \neq \vartheta)$, where T_n is a Θ_n-valued approximation of the estimator T. Then the sup expected loss may be bounded below in terms of

$$\sup_{\vartheta \in \Theta_n} P_\vartheta(T_n(X^n) \neq \vartheta) \geq \frac{1}{|\Theta_n|} \sum_{\vartheta \in \Theta_n} P_\vartheta(T_n(X^n) \neq \vartheta).$$

Here the right hand side equals $Pr\{T_n(X^n) \neq \vartheta\}$, for ϑ uniformly distributed on Θ_n. Hence by Fano's inequality, it is bounded below by

$$\frac{1}{\log|\Theta_n|} \quad [H(\vartheta|T_n(X^n)) - \log 2] = 1 - \tfrac{I(\vartheta \wedge T_n(X^n))}{\log|\Theta_n|} - \tfrac{\log 2}{\log|\Theta_n|} \geq \tag{3.13}$$

$$\geq 1 - \tfrac{I(\vartheta \wedge X^n)}{\log|\Theta_n|} - \tfrac{\log 2}{\log|\Theta_n|}. \tag{3.14}$$

If here $I(\vartheta \wedge X^n) < c \log|\Theta_n|$, for some $c < 1$, one arrives at a useful lower bound to $\sup_{\vartheta \in \Theta} Ed(\vartheta, T(X^n))$, valid for any estimator $T(X^n)$.

Ideas as hinted to above have been used to derive risk bounds tight up to a constant factor in non-parametric density estimation. Works in this direction include Hasminskii 1978, Ibragimov and Hasminskii 1982, Efroimovich and Pinsker 1982, and recently Yu 1995, Yang and Barron 1997; the results of the latter are particularly impressive.

3.6 Other topics

In this Section, only a fraction of statistical applications of IT could be covered. For others, and for more information about those only tangentially mentioned here, let me refer to the excellent survey Barron 1997.

Let me just mention one field that jointly belongs to IT and statistics, and obviously requires methods of both disciplines: hypothesis testing and estimation based on remote observations, subject to rate constraints on permissible communication. Works about hypothesis testing and estimation problems, respectively, with communication constraints, include Ahlswede and Csiszár 1986, Han 1987, Shalaby and Papamarcou 1992, respectively Zhang and Berger 1988, Ahlswede and Burnashev 1990, Han and Amari 1995.

References

Ahlswede, R. and Burnashev, M. (1990) "On minimax estimation in presence of side information about remote data," *Ann. Statist.,* vol.18, pp.141–171.

Ahlswede, R. and Csiszár, I. (1986) "Hypothesis testing with communication constraints," *IEEE Trans. IT*, vol.32, pp.533–542.

Ahlswede, R., Gács, P. and Körner, J. (1976) "Bounds on conditional probabilities with applications in multiuser communication," *Z. Wahrscheinlichkeitsth. verw. Gebiete,* vol.34, pp.157–177.

Akaike, H. (1973) "Information theory and an extension of the maximum likelihood principle," *Second Int'l Symp. Inform. Theory,* pp.267–281, B. N. Petrov and F. Csáki, eds., Akadémiai Kiadó, Budapest.

Arimoto, S. (1972) "An algorithm for computing the capacity of discrete memoryless channels," *IEEE Trans. IT,* Vol.18, pp.14–20.

Barron, A. R. (1986) "Entropy and the central limit theorem," *Ann. Probab.*, vol.14, pp.336–342.

Barron, A. R. (1997) "Information theory in probability, statistics, learning, and neural nets," Manuscript.

Bernardo, J. M. (1979) "Reference posterior for Bayesian inference," *J. Roy. Statist. Soc.* B, vol.41, pp.113–147.

Blahut, R. (1972) "Computation of channel capacity and rate-distortion functions," *IEEE Trans. IT,* vol.18, pp.460–473.

Byrne, C. (1993) "Iterative image reconstruction algorithms based on cross-entropy minimization," *IEEE Trans. Image Processing,* vol.2, pp.96–103.

Byrne, C. (1996) "Alternating minimization, generalized orthogonality and Pythagorean identities in iterative image reconstruction," *SIAM J. Optimization,* submitted.

Chernoff, H. (1952) “A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations,” *Ann. Math. Statist.*, vol.23, pp. 493–507.

Clarke, B. and Barron, A. (1994), “Jeffreys’ prior is asymptotically least favorable under entropy risk,” *J. Statist. Planning and Inference,* vol.41, pp. 37–60.

Cover, T. (1984) “An algorithm for maximizing expected log investment return,” *IEEE Trans. IT,* vol.30, pp. 369–373.

Csiszár, I. (1965) “A note on limiting distributions on topological groups,” *Publ. Math. Inst. Hungar. Acad. Sci.,* vol.9, pp.595–599.

Csiszár, I. (1975) “I-divergence geometry of probability distributions and minimization problems,” *Ann. Probab.,* vol.3, pp.146–158.

Csiszár, I. (1984) “Sanov property, generalized I-projections, and a conditional limit theorem,” *Ann. Probab.,* vol.12, pp.768–793.

Csiszár, I. (1989) “A geometric interpretation of Darroch and Ratcliff’s generalized iterative scaling,” *Ann. Statist.,* vol.17, pp.1409–1413.

Csiszár, I. and Körner, J. (1981) *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic.

Csiszár, I. and Tusnády, G. (1984) “Information geometry and alternating minimization procedures,” *Statistics and Decisions,* Suppl.1, pp.205–237.

Darroch, J. N. and Ratcliff, D. (1972) “Generalized iterative scaling for log-linear models,” *Ann. Math. Statist.,* vol.43, pp.1470–1480.

Della Pietra, S., Della Pietra, V. and Lafferty, J. (1997), “Bregman distances, iterative scaling, and auxiliary functions,” Manuscript.

Dembo, A. and Zeitouni, O. (1993), *Large Deviations Techniques and Applications,* Jones and Bartlett.

Deming, W. E. and Stephan, F. F. (1943) “On a least squares adjustment of a sampled frequency table when the expected marginal totals are known,” *Ann. Math. Statist.,* vol.11, pp.427–444.

Dempster, A., Laird, N. and Rubin, D. (1977) “Maximum likelihood from incomplete data via the EM algorithm,” *J. Roy. Statist. Soc.,* B, vol.39, pp.1–38.

Efroimovich, S. Y. and Pinsker, M. S. (1982) “Estimation of square-integrable probability density of a random variable” (in Russian), *Probl. Pered. Inform.*, vol.18, no.3, pp.19–38.

Fisher, R. A. (1925) "Theory of statistical estimation," *Proc. Camb. Phil. Soc.*, vol.22, pp.700–725.

Fritz, J. (1973) "An information theoretic proof of limit theorems for reversible Markov processes," *Trans. Sixth Prague Conference on Inform. Theory etc.,* pp.183–197, Academia, Prague.

Good, I. J. (1950) *Probability and the Weighing of Evidence*, Griffin, London.

Groeneboom, P., Oosterhoff, J. and Ruymgaart, F. H. (1979) "Large deviation theorems for empirical probability measures," *Ann. Probab.*, vol.7, pp.553–586.

Hájek, J. (1958) "On a property of normal distributions of any stochastic process" (in Russian), *Czechoslovak Math. J.*, vol.8, pp.610–617.

Han, T. S. (1987) "Hypothesis testing with multiterminal data compression," *IEEE Trans. IT*, vol.33, pp.759–772.

Han, T. S. and Amari, S. (1995) "Parameter estimation with multiterminal data compression," *IEEE Trans. IT*, vol.41, pp.1802–1833.

Hasminskii, R. Z. (1978) "A lower bound on the risks of nonparametric estimates of densities in the uniform metric," *Theory Probab. Appl.*, vol.23, pp.794–796.

Hoeffding, W. (1965) "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol.36, pp.369–400.

Ibragimov, I. A. and Hasminskii, R. Z. (1973), "On the information in a sample about a parameter," *Second Int'l Symp. Inform. Theory*, pp. 295–309, B. N. Petrov and F. Csáki, eds., Akadémiai Kiadó, Budapest.

Ibragimov, I. A. and Hasminskii, R. Z. (1982) "Bounds for the risks of non-parametric regression estimates," *Theory Probab. Appl.*, vol.27, pp. 84–99.

Ireland, C. T. and Kullback, S. (1968) "Contingency tables with given marginals," *Biometrika*, vol.55, pp.179–188.

Kendall, D. G. (1963) "Information theory and the limit theorem for Markov chains and processes with a countable infinity of states," *Ann. Inst. Stat. Math.*, vol.15, pp.137–143.

Kolmogorov, A. N. (1958) "A new invariant for transitive dynamical systems" (in Russian), *Dokl. A.N.SSSR*, vol.119, pp.861–864.

Kruithof, R. (1937) "Telefoonverkeersrekening," *De Ingenieur*, vol.52, pp.E15–E25.

Kullback, S. (1959) *Information Theory and Statistics*, Wiley.

Kullback, S. (1968) “Probability densities with given marginals,” *Ann. Math. Statist.*, vol.39, pp. 1236–1243.

Kullback, S. and Leibler, R. A. (1951) “On information and sufficiency,” *Ann. Math. Statist.*, vol.22, pp.79–86.

Linnik, Yu. V. (1959) “An information theoretic proof of the central limit theorem on Lindeberg conditions” (in Russian), *Teor. Veroyat. Primen.*, vol.4, pp. 311–321.

Margulis, G. A. (1974) “Probabilistic characteristics of graphs with large connectivity” (in Russian), *Probl. Pered. Inform.*, vol.10, no.2, pp.101–108.

Marton, K. (1986) “A simple proof of the blowing up lemma,” *IEEE Trans. IT*, vol.32, pp.445–446.

Marton, K. (1996) “Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration,” *Ann. Probab.*, vol.24, pp.857–866.

Marton, K. and Shields, P. (1994) “The positive divergence and blowing up properties,” *Israeli J. Math.*, vol.86, pp.331–348.

Matus, F. (1997) “On iterated averages of I-projections,” *Ann. Statist.*, submitted.

Pinsker, M. S. (1972), “Information contained in observations, and asymptotically sufficient statistics” (in Russian), *Probl. Pered. Inform.*, vol.8, pp.45–61.

Rényi, A. (1961) “On measures of entropy and information,” *Proc. Fourth Berkeley Symposium Math. Statist. Probab.*, vol.1, pp.547–561, Univ. Calif. Press.

Rényi, A. (1969) “On some problems of statistics from the point of view of information theory,” *Proc. Coll. Inform. Theory*, pp.343–357, J. Bolyai Math. Soc., Budapest.

Rissanen, J. (1978) “Modeling by shortest data description,” *Automatica*, vol.14, pp. 465–471.

Rissanen, J. (1989) *Stochastic Complexity in Statistical Inquiry*, World Scientific.

Rüschendorf, L. (1995) “Convergence of the iterative proportional fitting procedure,” *Ann. Statist.*, vol.23, pp.1160–1174.

Sanov, I. N. (1957) “On the probability of large deviations of random variables” (in Russian), *Math. Sbornik*, vol.42, pp.11–44.

Shalaby, H. and Papamarcou, A. (1992) “Multiterminal detection with zero-rate data compression,” *IEEE Trans. IT*, vol.38, pp.254–267.

Shannon, C. E. (1956) “The bandwagon,” *IRE Trans. IT*, vol.2, p.3.

Shields, P. C. (1996) *The Ergodic Theory of Discrete Sample Paths*, Graduate Studies in Math., vol.13, Amer. Math. Soc.

Talagrand, M. (1995) “Concentration of measure and isoperimetric inequalities in product spaces,” *Publ. Math. IHES*, vol.81, pp.73–205.

Talagrand, M. (1996) “A new look at independence,” Special Invited Paper, *Ann. Probab.*, vol.24, pp.1-34.

Tusnády, G. (1977) “On asymptotically optimal tests,” *Ann. Statist.*, vol.5, pp.385–393.

Wald, A. (1947) *Sequential Analysis*, Wiley.

Yang, Y. and Barron, A. R. (1997) “Information-theoretic determination of minimax rates of convergence,” to appear in *Ann. Statist.*

Yu, B. (1995) “Assouad, Fano, and Le Cam,” to appear in *Festschrift in honor of Lucien Le Cam.*

Zhang, Z. and Berger, T. (1988) “Estimation via compressed information,” *IEEE Trans. IT*, vol.34, pp.198–211.