Probab. Th. Rel. Fields 86, 403–422 (1990)

Probability Theory and Related Fields

Discrimination distance bounds and statistical applications

Marek Kanter
1216 Monterey Avenue, Berkeley, CA 94707, USA
Received February 2, 1990; in revised form February 22, 1990

Summary. Bounds are obtained for the Kullback-Leibler discrimination distance between two random vectors X and Y. If X is a sequence of independent random variables whose densities have similar tail behavior and $Y=AX$, where A is an invertible matrix, then the bounds are a product of terms depending on A and X separately. We apply these bounds to obtain the best possible rate of convergence for any estimator of the parameters of an autoregressive process with innovations in the domain of attraction of a stable law. We provide a general theorem establishing the link between total variation proximity of measures and the rate of convergence of statistical estimates to complete the exposition for this application.

1. Introduction

1.1. The application

The past two decades have witnessed an increase of interest in the statistical analysis of linear processes with infinite variance. Such stochastic processes have the property that each observation is a linear combination of independent random variables with infinite variance. From a theoretical point of view these processes are a natural extension of finite variance models; in practice, they provide a better fit to certain time series, e.g. stock market prices.

The statistical theory of parameter estimation for linear processes with infinite variance is complicated by the fact that the Fisher information of the parameter to be estimated is generally infinite. In particular no theory of "efficient" estimators has been developed for such processes, except, as in [8], when observations are independent.

Autoregressive models with innovations in the domain of attraction of a stable law with characteristic exponent $\lambda \in (0, 2)$ constitute a useful subclass of linear processes with infinite variance. For any metric ρ on $\mathbb{R}^p$ which is uniformly

equivalent to the usual Euclidean metric (as in (3.7)), the results in [6], [7], and [9] show that for all α in a suitable subset of $\mathbb{R}^P$

$$\lim_{n\to\infty} n^{1/\delta}\rho(W_n, \alpha)=0 \quad \text{(a.s. } P_\alpha\text{)}, \tag{1.1}$$

where $\delta>\lambda$ and W_n is the usual least squares estimate of the vector parameter α defining the autoregressive process with corresponding measure P_α. If $\lambda=2$, then (1.1) still holds by virtue of classical finite information, finite variance theory which completely specifies the limit in (1.1) if $\delta=\lambda=2$. Recently [4], an advance has been made in the case $\lambda<2$; the closely related Yule-Walker estimate W_n' was shown to satisfy

$$(n/\log(n))^{1/\lambda}(W_n'-\alpha)\xrightarrow{d} F_\alpha, \tag{1.2}$$

where $\xrightarrow{d}$ stands for convergence in distribution and F_α is a continuous distribution function. Based on these results, one might conjecture that the „best" rate of convergence for *any* estimator V_n is offerred by sequences of the form $c_n=n^{1/\lambda}s_n$, where s_n is slowly varying as $n\to\infty$, i.e. $\lim_{n\to\infty} s_n/s_{n+1}=1$.

Our present state of knowledge is such that we can only demonstrate a comparatively primitive version of this conjecture; i.e. for V_n and c_n as above, we will show that the limiting distribution $G_\alpha(t)$ of $c_n\rho(V_n, \alpha)$ satisfies

$$(G_\alpha(+\infty)\leqq 1/2 \quad \text{if } \delta<\lambda; \quad G_\alpha(0+)\leqq 1/2 \quad \text{if } \delta=\lambda) \tag{1.3}$$

for all but countably many α. The interpretation is that for $\delta<\lambda$ the sequence $n^{1/\delta}$ increases so quickly that at least half the mass in the distribution of $c_n\rho(V_n, \alpha)$ is sent to ∞, while in the boundary case $\delta=\lambda$ no more than half the mass is sent to 0.

Our method for proving (1.3) consists of the following steps:

(a) Tying together the rate of convergence of parameter estimates with the total variation distance of the corresponding measures.

(b) Noting that the Kullback-Leibler discrimination distance dominates the total variation metric.

(c) Bounding the discrimination distance.

Step (a) is made in Theorem 4.1. Step (b) is just the well known inequality (3.19). Step (c) is our main result. It is accomplished by developing discrimination distance bounds for random variables, and then applying these inequalities to autoregressive processes. These one-dimensional bounds are extended in Theorem 2.3, where the discrimination distance between two random vectors is expressed in terms of the distance between a sequence of independent random variables and the new sequence gotten by scaling the original sequence by constants and then adding random shifts.

1.2. Preliminary definitions

We start by defining the Kullback-Leibler discrimination distance $K(P:Q)$ in the simple case when P and Q are probability measures on $\mathbb{R}^n$ with density functions f and g, respectively:

$$K(P:Q)=\int f(x)\log(f(x)/g(x))\,dx. \tag{1.4}$$

If X and Y are random vectors with distributions P and Q respectively, we shall write $K(X:Y)$ for $K(P:Q)$. If X and Y are infinite dimensional random vectors, we define

$$K(X:Y)=\lim_{n\to\infty} K(X^{[n]}:Y^{[n]}), \tag{1.5}$$

where $X^{[n]}=(X_1, \ldots, X_n)$ if $X=(X_1, \ldots, X_n, X_{n+1}, \ldots)$.

Suppose now that (V, X) and (W, Y) are random vectors in $\mathbb{R}^{m+n}$. We need to define $K_V((X \mid V):(Y \mid W))$, the P_V average discrimination distance between X and Y given V and W, where P_V is the distribution of V. If v is a point in $\mathbb{R}^m$, we put

$$K_v(X:Y)=\int \log(f_v(x)/g_v(x))\, f_v(x)\, dx, \tag{1.6}$$

where f_v and g_v are the conditional densities of X given $V=v$ and Y given $W=v$, respectively. We define

$$K_V((X \mid V):(Y \mid W))=\int K_v(X:Y)\, dP_V(v). \tag{1.7}$$

We simplify notation and write $K_V(X:Y)$ in place of $K_V((X \mid V):(Y \mid W))$ if $P_V=P_W$ and the joint distributions of (V, X) and (W, Y) are clear from the context. We note that m may equal ∞ in the above definition, i.e. V may be an infinite dimensional random vector.

If (V, X) and (W, Y) are random vectors in $\mathbb{R}^{m+n}$, then we have

$$K((V, X):(W, Y))=K(V:W)+K_V((X \mid V):(Y \mid W)). \tag{1.8}$$

This well known additivity property for K follows directly from (1.6) and is the basis for our success in computing $K(Y:X)$ for linear processes X, Y.

The other properties of discrimination distance that we use are shared by the other standard distances between measures such as the total variation distance. Among these properties is the important inequality

$$K(X:Y)\geqq K(T(X):T(Y)), \tag{1.9}$$

where T is Borel measurable.

1.3. Summary of results

Since our results are easier to describe in the case when all random variables are symmetric, we will implicitly make this restriction in this section in order to summarize our work with less clutter. We emphasize that our actual results are *not* so restricted.

Upper bounds for the discrimination distance between random variables constitute the core of this paper and comprise most of Sect. 2. We open Sect. 2 with the following inequalities for a random variable X with absolutely continuous density f:

$$|E(\log(f(aX)/f(aX+z))|\leqq(1/2)\, M(\partial h)\, z^2 \tag{1.10}$$

$$E(|\log(f(aX)/f(aX+z)|)\leqq M(h/r)\log(1+|z|), \tag{1.11}$$

where $h=-f'/f$ and $r(x)=(1+|x|)^{-1}$. (See the beginning of Sect. 2.1 for the definitions of $M(\partial h)$ and $M(h/r)$.) As a direct consequence, we show that

$$|K(aX+z:X)-K(aX:X)| \leqq J_1(X)\log(1+z^2), \tag{1.12}$$

where $J_1(X)$ is defined in (2.19). (See (2.25) for the general version of (1.12).)

To proceed further we treat z in (1.12) as a random variable Z and take the expectation. We characterize the tail behavior of the distribution of Z via the constant $\sigma_\lambda(Z)$. This constant is explicitly defined in terms of the characteristic function $\phi_Z(t)=E\exp(\mathrm{i}tZ)$ in (2.17); when $\lambda=2$, it equals the variance of Z. We get

$$|K_Z(aX+cZ:X)-K(aX:X)| \leqq \Gamma(\lambda) J_1(X)\,\sigma_\lambda(Z)\,|c|^\lambda. \tag{1.13}$$

If cZ is a sum $\sum_1^n c_i X_i$ with independent random variables X_i, we show that $\sigma_\lambda(Z)|c|^\lambda$ may be replaced by $\sum_1^n \sigma_\lambda(X_i)|c_i|^\lambda$ in (1.13). This important property follows by virtue of the subadditivity of the Q functional defined in (2.27). We conclude the section with the multivariate inequality

$$\begin{aligned} K(AX:X)+\log(|\det(A)|) \leqq \sum_1^n (K(a_{ii}X_i:X_i)+\log(|a_{ii}|)) \\ +\Gamma(\lambda) J(X)\,\sigma_\lambda(X)\,N_\lambda(\hat{A}), \end{aligned} \tag{1.14}$$

where $X=(X_1, \ldots, X_n)$ is a vector with independent random components and A is an invertible matrix. (See Def. 2.4 and 2.5 for the definitions of $J(X)$ and $N_\lambda(\hat{A})$.)

Section 2 also contains a parallel stream of inequalities for which the Laplace transform of $|Z|$ replaces the Fourier transform ϕ_Z if $\lambda\in(0, 1)$. This parallel development is not possible in the case $\lambda\in[1, 2)$; in that case we note that the condition that $\sigma_\lambda(Z)$ be finite acts as an important centering device for non-symmetric random variables.

Section 3 contains the statistical application of the bound (1.13) and its non-symmetric cousins; for a p-th order autoregressive process parametrized by $\alpha=(\alpha_1, \ldots, \alpha_p)$ via (3.3) we show that (1.3) holds for all but countably many α in the parameter space M_p defined in (3.6), if the innovations X_i of the autoregressive process satisfy $\sigma_\lambda(X_i)<\infty$.

In Sect. 4 we derive a general result tying together convergence rates for parameter estimators with the total variation distance of the corresponding measures (Theorem 4.1). Since this section has more in common with the existing statistical literature than the other sections, it will be expedient to describe Theorem 4.1 in that broader context.

1.4. Relation to previous work

Many other authors have departed from the classical case of independent observations and finite Fisher information. The recent book by Ibragimov and

Has'Minkii [8], for instance, develops quite extensively the situation when observations about a parameter with infinite Fisher information are taken independently. By contrast, Roussas [16] and Basawa and Rao [2] consider estimation of a parameter when observations have Markov dependence; however, these authors hold to the restriction that the Fisher information of the parameter be finite. On a more general note, the works [13] and [14] of Le Cam contain important background information with regards to general estimators in the non-classical setting.

Previous work which most directly impinges on this paper involves the derivation of general results linking convergence rates of estimators to the proximity of the corresponding measures. Given a parameter space M with metric ρ, Akahira and Takeuchi [1] define "c_n-consistent" estimators as measurable functions V_n such that for ε sufficiently small

$$\lim_{t\to\infty} \limsup_{n\to\infty} \sup_{\beta} \{P_\beta[c_n \rho(V_n, \beta) \geqq t] : \rho(\alpha, \beta) \leqq \varepsilon\} = 0, \tag{1.15}$$

for all $\alpha \in M$, where c_n is a given sequence strictly increasing to ∞. For such estimators their Theorem 2.3.1 implies

$$\lim_{r\to\infty} \limsup_{n\to\infty} \inf_{\beta} \{D_n(P_\alpha, P_\beta) : \rho(\alpha, \beta) \geqq r/c_n\} = 2. \tag{1.16}$$

(See Sect. 4.1 for the definition of the total variation distance $D_n(P_\alpha, P_\beta)$.) It is not clear that the least squares and Yule-Walker estimates W_n considered in [7] and [4] can be modified to satisfy (1.15) with $c_n = n^{1/\delta}$ and $c_n = (n/\log(n))^{1/\lambda}$, respectively. For this reason we consider more general parameter estimators V_n satisfying (1.15) *without* the sup over β (i.e. with $\varepsilon = 0$). If such an estimator exists, our Theorem 4.1 has as consequence that $d_\alpha(+\infty) = 1$ for all but countably many α, where

$$d_\alpha(r) = (1/2) \limsup_{n\to\infty} (\sup_{\beta} \{D_n(P_\alpha, P_\beta) : r c_n^{-1} \leqq \rho(\alpha, \beta) < r c_{n-1}^{-1}\}). \tag{1.17}$$

The relationship between [1] and Theorem 4.1 can be brought into sharper focus if we paraphrase Theorem 2.3.1 in [1] as

$$\liminf_{n\to\infty} (\inf_{\rho(\alpha,\beta)\leqq\varepsilon} P_\beta[c_n \rho(V_n, \beta) < r/2]) \leqq \tfrac{1}{2}(1 + 2^{-1} \liminf_{n\to\infty} D_n(P_\alpha, P_{\beta_n})) \tag{1.18}$$

for all α in M, where β_n is a sequence in M satisfying $\rho(\alpha, \beta_n) \geqq r c_n^{-1}$. In fact, examination of the proof in [1] reveals that ε may be replaced by $c_n^{-1} r$, i.e. fixed neighborhoods of α may be replaced by shrinking neighborhoods. By contrast, the inequality proved in Theorem 4.1 replaces ε with 0 and $2^{-1} \liminf D_n(P_\alpha, P_{\beta_n})$ with $d_\alpha(r)$ in (1.18), and allows for a countable exceptional set of α. The proof of (1.18) is comparatively simple; it can be achieved by replacing $F_\alpha^n(t) + F_\beta^n(t)$ with $2(\min(F_\alpha^n(t), F_\beta^n(t))$ in Lemma 4.1. For our purposes Theorem 4.1 is more appropriate, because our discrimination distance inequalities give us control over the quantity $d_\alpha(r)$ and thus we need not restrict our consideration to "c_n-consistent" estimates. We should also mention that Vostrikova and Birgé prove results similar to (1.18) in [18] and [3], respectively.

The bounds developed in this paper were first presented in [10] in the context of applications to the theory of measures on infinite dimensional linear spaces as well as to statistical estimation. In this paper we have reworked [10] with an increased focus on statistical applications; in particular we treat infinite variance autoregressive processes of any order p rather than just $p=1$ as in [10]. We have also removed the restriction of symmetry previously imposed on probability distributions in [10]. On the other hand, [10] contains material on multidimensional discrimination distance inequalities which extends the result indicated in (1.14). This material will be reworked in a future paper, wherein applications to equivalence-singularity dichotomies for measure on infinite dimensional spaces will be given.

2. Information discrimination inequalities

2.1. One-dimensional inequalities

Conventions. All random variables X considered in this section will have an absolutely continuous density function, usually denoted by $f(x)$, and a score function $h(x)=-f'(x)/f(x)$. (We interpret $0/0=0$.) We shall let $I(X)$ stand for $E(h(X)^2)$, the Fisher information of X; and we shall let $H(X)$ stand for $\int -f(x)\log f(x)\,dx$, the Shannon differential entropy of X. We define the functional $\eta(X)=1$ if X is symmetrically distributed; otherwise we set $\eta(X)=2$. (The functional η will be used to give our inequalities a concise form, applicable to both symmetric and non-symmetric random variables.)

For any complex valued function g defined on the real line $\mathbb{R}$, we let $M(g)=\sup\{|g(x)|: x\in\mathbb{R}\}$ and $M(\partial g)=\sup\{|g(x)-g(y)|/|x-y|: x\neq y\}$. If $x, y\in\mathbb{R}$, we let $x\wedge y=\min(x, y)$ and $x\vee y=\max(x, y)$. We let $r(x)=(1+|x|)^{-1}$ and note that $R(x)=\operatorname{sign}(x)\log(1+|x|)$ is the indefinite integral of $r(x)$ specified by $R(0)=0$.

Our first lemma is a useful compendium of basic results. First, a definition:

Definition 2.1. The random variable X is called *regular* if either (a) $M(\partial h)\vee E(|X|)<\infty$ or (b) $M(h)<\infty$.

Lemma 2.1. *Let X be a random variable with density f and score function h. Let $a\neq 0$.*

(i) *If h is locally integrable with respect to Lebesgue measure, then f is strictly positive everywhere.*

(ii) *If X is regular, then h is locally integrable with respect to Lebesgue measure, $h(aX+z)$ has finite expectation, and the decomposition*

$$K(aX+z:X)=E(\log(f(aX)/f(aX+z))+K(aX:X) \tag{2.1}$$

holds for all $z\in\mathbb{R}$. Furthermore

$$|K(aX+z:X)+K(aX-z:X)-2K(aX:X)|\leqq\eta((a-1)X)M(\partial h)z^2. \tag{2.2}$$

Proof. To prove (i) write

$$\log(f(ax)/f(ax+z))=\int_0^z h(ax+y)\,dy, \tag{2.3}$$

and conclude that f does not vanish anywhere. (Otherwise the left hand side of (2.3) would be $\pm\infty$ for some x, z in $\mathbb{R}$, in contradiction to the local integrability of h.)

To prove (ii) note that if X is regular, then h is locally integrable and $h(aX+z)$ has finite expectation in case (b) of Definition 2.1; furthermore, the same conclusions follow in case (a) by virtue of the inequality

$$|h(s+t)-h(s)| \leqq M(\partial h)\,|t| \quad (\text{for } s, t \in \mathbb{R}). \tag{2.4}$$

We prove (2.1) by noting that $|a|\,f(ax+z)$ is the density of $a^{-1}(X-z)$; thus

$$\begin{aligned} &\int f(x)\log(f(x)/|a|\,f(ax+z))\,dx \\ &\quad = \int f(x)\,[\log(f(ax)/f(ax+z)) + \log(f(x)/|a|\,f(ax))]\,dx. \end{aligned} \tag{2.5}$$

We now define $v_a(y) = E(h(aX+y))$ and use (2.3) to write

$$K(aX+z:X)+K(aX-z:X) = \int_0^z v_a(y)-v_a(-y)\,dy + 2K(aX:X). \tag{2.6}$$

The bound (2.2) is a consequence of (2.6) and

$$|v_a(y)-v_a(-y)| \leqq 2M(\partial h)\,y. \tag{2.7}$$

(The last inequality follows from (2.3) and (2.4).)

To prove (2.2) when $\eta((a-1)X)=1$ (i.e. if X is symmetric or $a=1$) we need to show that $v_a(0)=0$. (We can then rewrite (2.6) and (2.7) without $v_a(-y)$, without $K(aX-z:X)$, and without the factor of 2 in front of $K(aX:X)$ and $M(\partial h)$.) If X is symmetric then $v_a(0)=0$ without computation. If $a=1$ we write

$$v_1(0) = \int f'(x)\,dx = \lim_{t\to\infty}(f(t)-f(-t)) = 0. \tag{2.8}$$

(Note $\int |f'(x)|\,dx < \infty$ because $h(X)$ has finite expectation.) Q.E.D.

Remark 2.1. It is appropriate to recall the known result [11] that

$$\lim_{z\to 0} z^{-2} K(X-z:X) = (1/2)\,I(X). \tag{2.9}$$

We conclude that (2.2) is not sharp in the limit as $z\to 0$. In fact, variants of (2.2) are available in which the right hand side is replaced by $(2a)^{-1}M(h)(I(X))^{1/2}z^2$ or by $a^{-1}\{(\tfrac{1}{2})I(f)\,E[\exp(zh(X))-1-zh(X)]\}^{1/2}$ (see [10]). The factor a^{-1} renders these bounds unsuitable for our use, though they may be sharper as $z\to 0$.

Remark 2.2. If h is differentiable, then $M(h')=M(\partial h)$ by the mean value theorem. Using the inequality $|h'(x)| \leqq |f''(x)/f(x)| + (h(x))^2$, we can verify that $M(h') \leqq \infty$ if the density f is strictly positive everywhere with $M(f'')<\infty$ and $f''(x) \sim |x|^{-(3+\lambda)}$ as $|x|\to\infty$ for some $\lambda\in(0,2)$. Such densities will have the property that $f(x)\sim|x|^{-(1+\lambda)}$ as $|x|\to\infty$ and will have infinite variance. Stable and Pareto densities of index λ are examples.

Remark 2.3. In future applications of Lemma 2.1 it will be useful to replace $K(aX+z:X)$ and $K(aX:X)$ in (2.2) by $K^*(aX+z:X)$ and $K^*(aX:X)$, where $K^*(aX+z:X)=K(aX+z:X)+\log(|a|)$ for any random variable X. K^* is the renormalized form of K. If the differential entropy $H(X)$ exists and is finite, then we can define

$$K^*(z:X)=\lim_{a\to 0} K^*(aX+z:X)=H(X)-\log(f(z)), \tag{2.10}$$

and (2.2) is valid as a limit at $a=0$.

Lemma 2.1 will suffice for our future needs only in the special case when X has finite variance. Otherwise we need a sharper inequality in the region $|z|\geqq 1$. The following lemma provides the first step in this program.

Lemma 2.2. *Let $g(x)=g(|x|)\geqq 0$ be decreasing as a function of $|x|$. Let G be the indefinite integral of g specified by $G(0)=0$. Then for all $z\geqq 0$,*

$$(\tfrac{1}{2})(G(x+z)+G(z-x))\leqq G(z) \quad (\textit{for } x\in\mathbb{R}). \tag{2.11}$$

Proof. We may assume without loss of generality that $x\geqq 0$. If $z\geqq x$ we rewrite (2.11) as $G(x+z)-G(z)\leqq G(z)-G(z-x)$. If $0\leqq z\leqq x$, we rewrite (2.11) as $G(x+z)-G(x-z)\leqq 2G(z)$. Now use the fact that g is decreasing on $[0,\infty)$ to argue each case. Q.E.D.

Remark 2.4. Any symmetric random variable X is a mixture of symmetric random variables taking on the two values $\pm x$. It follows that $E(G(X+z))\leqq G(|z|)$ for such random variables.

We now establish our sharper bound for large z. We let $r(x)$ and $R(x)$ be defined as preceding Definition 2.1.

Lemma 2.3. *Let X be a regular random variable. Then for $z\in\mathbb{R}$,*

$$|K(aX+z:X)+K(aX-z:X)-2K(aX:X)|\leqq 2M\left(\frac{h}{r}\right)\log(1+|z|). \tag{2.12}$$

Proof. We may assume $z\geqq 0$ without loss of generality, since X may be exchanged with $-X$. We note

$$K(aX-z:X)-K(aX:X)=K(-aX+z:-X)-K(-aX:-X) \tag{2.13}$$

and use (2.1) to write

$$\begin{aligned} &K(aX+z:X)+K(aX-z:X)-2K(aX:X)\\ &\quad=E[\log(f(aX)/f(aX+z))+\log(f(-aX)/f(-aX+z))]. \end{aligned} \tag{2.14}$$

We recall (2.3) to get the estimate

$$\begin{aligned} |\log(f(ax)/f(ax+z))|&\leqq M(h/r)\int_0^z r(ax+y)\,dy\\ &=M(h/r)(R(ax+z)-R(ax)). \end{aligned} \tag{2.15}$$

Using the obvious relation

$$E[R(aX)+R(-aX)]=0, \tag{2.16}$$

and applying Lemma 2.2, we get (2.12). Q.E.D.

Remark 2.5. The motivation for defining $r(x)=(1+|x|)^{-1}$ comes from the simple fact that if $0<\lambda<2$ and f is a density with $f(x)\sim|x|^{-(1+\lambda)}$ as $|x|\to\infty$, then $M(h/r)$ is finite in most cases. (For example, the extra conditions that $f'(x)\sim|x|^{-(2+\lambda)}$ and $M(hr)<\infty$ will ensure that $M(h/r)<\infty$.) Another possibility for defining r would be to use $r(x)=1$, thereby getting $2M(h)|z|$ in the right hand side of (2.12). This choice of $r(x)$ is unsatisfactory because we need to replace z by random variable Z to proceed further, and Z will generally not have a finite expectation. We will categorize the tail behavior of Z via the quantity $\sigma_\lambda(Z)$:

Definition 2.2. Let $\phi_Z(t)=E(\exp(itZ))$ be the characteristic function of Z. For $0<\lambda\leqq 2$ define

$$\sigma_\lambda(Z)=2\sup_{t>0} t^{-\lambda}|1-\phi_Z(t)|. \tag{2.17}$$

Remark 2.6. It is a classical result that $\sigma_\lambda(Z)$ will be finite if Z is in the domain of attraction of a stable distribution with norming coefficients $n^{1/\lambda}$. If $\sigma_\lambda(Z)<\infty$ for $1<\lambda\leqq 2$, then $E(|Z|)<\infty$ and $\phi'_Z(0)=iE(Z)=0$. (The finiteness of $E(|Z|)$ follows from the truncation inequality in Loève [15, p. 196], which easily implies that $P(|Z|\geqq s]$ is bounded by $(7/(1+\lambda))\,\sigma_\lambda(Z)\,s^{-\lambda}$ for any $s>0$.) If $\lambda=2$ it is easy to check, using the elementary inequality $|e^{ix}-1-ix|\leqq x^2/2$, that $\sigma_2(Z)=E(Z^2)-(E(Z))^2$.

In the cases $0<\lambda<1$, the random variable Z can be categorized in an alternative fashion via the following definition:

Definition 2.3. Let $\psi_Z(t)$ stand for the Laplace transform $E(\exp(-t|Z|))$. We define $\mu_\lambda(Z)$ via

$$\mu_\lambda(Z)=\sup_{t>0} t^{-\lambda}(1-\psi_Z(t)). \tag{2.18}$$

Remark 2.7. It follows from the well known relation $-\psi'_Z(0)=E(|Z|)$ that $\mu_\lambda(Z)$ will be infinite if $\lambda>1$ and Z is not identically 0. In most cases $\mu_\lambda(Z)$ and $\sigma_\lambda(Z)$ will be simultaneously finite or infinite when $0<\lambda<1$, but we will not spare the space to give a precise rendering of this assertion.

Our work thus far culminates in the following theorem. Given a random variable X with score function h, we define

$$J_a(X)=\{2M(h/r)\}\vee\{(\log(2))^{-1}\,\eta((a-1)X)\,M(\partial h)\}, \tag{2.19}$$

where $a\in\mathbb{R}$ and η is defined at the beginning of Sect. 2.1.

Theorem 2.1. *Let $X_0, X_1, \ldots, X_n, \ldots$ be random variables such that X_0 is regular and such that $\sum_1^\infty c_i X_i = Z$ converges a.s. for a given sequence c_i in $\mathbb{R}$. Let P_Z stand for the distribution of Z and let $\Gamma(\lambda) = \int_0^\infty t^{\lambda-1} e^{-t}\, dt$. Define*

$$\Delta_Z(K_a X_0)) = \int |K(aX_0 + z : X_0) + K(aX_0 - z : X_0) - 2K(aX_0 : X_0)|\, dP_Z(z). \tag{2.20}$$

Then for $a \neq 0$,

$$\Delta_Z(K_a X_0)) \leqq 2M(h_0/r) \sum_1^\infty \Gamma(\lambda_i)\, \mu_{\lambda_i}(X_i)\, |c_i|^{\lambda_i} \quad (\text{if } \lambda_i \in (0, 1)\, \forall i). \tag{2.21}$$

Furthermore, if $(X_1, \ldots, X_N, \ldots)$ are independent,

$$\Delta_Z(K_a X_0)) \leqq J_a(X_0) \sum_1^\infty \Gamma(\lambda_i)\, \sigma_{\lambda_i}(X_i)\, |c_i|^{\lambda_i} \quad (\text{if } \lambda_i \in (0, 2]\, \forall i). \tag{2.22}$$

Proof. To prove (2.21) we use the trivial inequality $\log(1 + |x + y|) \leqq \log(1 + |x|) + \log(1 + |y|)$ and (2.12) to conclude that

$$\Delta_Z(K_a X_0)) \leqq 2M(h/r) \sum_1^\infty E(\log(1 + |c_i X_i|)). \tag{2.23}$$

Furthermore, for any random variable X

$$E(\log(1 + |cX|)) = \int_0^\infty (1 - \psi_X(|c|\, t))\, e^{-t} t^{-1}\, dt, \tag{2.24}$$

and (2.21) follows immediately. (Check (2.24) by letting X be constant and then randomizing.)

To prove (2.22) we start by combining (2.2) and (2.12) via the inequality

$$|K(aX_0 + z : X_0) + K(aX_0 - z : X_0) - 2K(aX_0 : X_0)| \leqq J_a(X_0) \log(1 + z^2). \tag{2.25}$$

(Note that $\log(1 + z^2) \geqq \log(2)\, z^2$ for $|z| \leqq 1$, while $\log(1 + z^2) \geqq \log(1 + |z|)$ for $|z| \geqq 1$, in checking (2.25).) Companion to (2.24) is the identity

$$E(\log(1 + c^2 X^2)) = 2 \int_0^\infty (1 - \operatorname{Re}(\phi_X(t))\, e^{-t} t^{-1}\, dt, \tag{2.26}$$

from which follows

$$E(\log(1 + c^2 X^2)) \leqq Q(X) \equiv 2 \int_0^\infty |1 - \phi_X(t)|\, e^{-t} t^{-1}\, dt. \tag{2.27}$$

The functional Q defined in (2.27) satisfies the subadditive property

$$Q\left(\sum_1^\infty X_i\right) \leqq \sum_1^\infty Q(X_i) \tag{2.28}$$

as a consequence of the relations

$$|1-\phi_Z(t)| = |1-\Pi\phi_{X_i}(t)| \leqq \sum_1^\infty |1-\phi_{X_i}(t)|. \tag{2.29}$$

We conclude (2.22) from (2.25), (2.27), and (2.28). Q.E.D.

Remark 2.8. If $\lambda_i = 2$ for all i in Theorem 2.1, we may replace $J_a(X_0)$ in (2.22) by the smaller constant $\eta((a-1)X_0)M(\partial h_0)$. (Note that $\log(1+z^2)$ is replaced by z^2 in (2.25) and that we use the additivity of variances instead of (2.28), since the X_i all have expectation 0 by Remark 2.6.)

Remark 2.9. The inequality (2.22) will be more useful than (2.21), because (2.22) is applicable over the full range $\lambda \in (0, 2]$. The fact that (2.21) is valid without assuming independence of the random variables X_i is basically a curiosity in our present work, since the applications we have in mind will impose independence.

Remark 2.10. If X is symmetric then $Q(X)$ reduces to $E(\log(1+X^2))$ and the subadditivity property (2.28) can be straightforwardly proven without taking Fourier transforms. (See [10].)

Given the random variables X_0 and Z, we shall use the notation $Y_a = aX_0 + Z$ and $\bar{Y}_a = aX_0 - Z$. If $X = (Z, X_0)$ and $Y = (Z, Y_a)$, then the relations

$$K(Y:X) = K(Z:Z) + K_Z(Y_a:X_0) = K_Z(Y_a:X_0) \tag{2.30}$$

show that

$$K(Y_a:X_0) \leqq K_Z(Y_a:X_0), \tag{2.31}$$

(since $K(Y_a:X_0) \leqq K(Y, X)$ by virtue of (1.9)). We now develop a bound for the right hand side of (2.31).

Corollary 2.1. *Let X_0 and Z be given as in Theorem* 2.1, *with the extra hypothesis that X_0 and Z are independent. Then for $a \neq 0$*

$$K_Z(Y_a:X_0) + K_Z(\bar{Y}_a:X_0) \leqq 2K(aX_0:X_0) + M(h/r)\sum_1^\infty \Gamma(\lambda_i)\,\mu_{\lambda_i}(X_i)\,|c_i|^{\lambda_i} \tag{2.32}$$

if $\lambda_i \in (0, 1)$ for all i. If $(X_1, \ldots, X_n, \ldots)$ are mutually independent, then

$$K_Z(Y_a:X_0) + K_Z(\bar{Y}_a:X_0) \leqq 2K(aX_0:X_0) + J_a(X_0)\sum_1^\infty \Gamma(\lambda_i)\,\sigma_{\lambda_i}(X_i)\,|c_i|^{\lambda_i}. \tag{2.33}$$

Proof. We can write

$$K_Z(Y_a:X_0) = \int K(aX_0 + z:X_0)\,dP_Z(z) \tag{2.34}$$

because Z is independent of X_0. We conclude that

$$K_Z(Y_a:X_0)+K_Z(\bar{Y}_a:X_0)\leqq 2K(aX_0:X_0)+\Delta_Z(K(aX_0)), \tag{2.35}$$

so (2.32) and (2.33) follow from (2.21) and (2.22), respectively. Q.E.D.

It will be useful to remove the restriction that $a\neq 0$ for future applications of these results.

Theorem 2.2. *Suppose either* (a) $M(h_0/r)\vee E(\log(1+|X_0|))<\infty$, *or* (b) $M(\partial h_0)\vee E(|X_0|)<\infty$. *Then the restriction that* $a\neq 0$ *in Theorem* 2.1 *and its corollary can be removed by replacing* K *with the renormalization* K^* *throughout. (See Remark* 2.3.)

Proof. It suffices to show that the differential entropy $H(X_0)$ is defined via an absolutely convergent integral. We write

$$|H(X_0)+\log(f_0(0))|=\int \log(f_0(x)/f_0(0))\, f_0(x)\,dx. \tag{2.36}$$

(Here f_0, the density of X_0, is strictly positive everywhere on account of Lemma 2.1.) Using (2.3), the theorem follows from (2.15) in case (a), and from (2.4) in case (b). Q.E.D.

Remark 2.11. Let $0<\lambda\leqq 2$. If $\sigma_\lambda(X)<\infty$, then $E(\log(1+|X|))$ is finite by virtue of (2.27) and the simple inequality $E(\log(1+|X|))\leqq\log(2)+E(\log(1+X^2))$. If $\mu_\lambda(X)<\infty$ then $E(\log(1+|X|))$ is finite on account of (2.24). This remark will be useful in showing that the hypotheses of Theorem 2.2 hold in the context to be established in Sect. 2.2.

It is interesting to present two examples for which we can calculate $K^*(aX+z:X)$ exactly.

Example 2.1. Let X be a Cauchy random variable with density function $f(x)=\pi^{-1}(1+x^2)^{-1}$ and characteristic function $\phi(t)=\exp(-|t|)$. We evaluate $K^*(aX:X)$ by noting that $K^*(aX:X)=V(a)-V(1)$, where $V(a)=E(\log(1+a^2X^2))$. Using (2.26), we get $K^*(aX:X)=\log((1+|a|)^2/4)$.

To evaluate $K^*(aX+z:X)$ when $z\neq 0$, we let $w_a(z)$ stand for the partial derivative of $K^*(aX+z:X)$ with respect to z. Using (2.1), we get

$$w_a(z)=\int h(ax+z)\, f(x)\,dx, \tag{2.37}$$

where $h(x)=2\pi x f(x)$. We apply Parseval's relation to get

$$w_a(z)=|a|^{-1}\int e^{-|t/a|}\sin(tz/|a|)\, e^{-|t|}\,dt=2z/(z^2+(1+|a|)^2). \tag{2.38}$$

We conclude that $K^*(aX+z:X)=\log((z^2+(1+|a|)^2/4)$.

Example 2.2. Let X be a Gaussian random variable with density $f(x)=(2\pi)^{-1/2}\exp(-(1/2)x^2)$. It is easy to show that $K^*(aX+z:X)=\frac{1}{2}(z^2+a^2-1)$.

2.2. *Multi-dimensional inequalities*

Our final result of the section will be the application of the preceding one-dimensional discrimination bounds to yield a discrimination distance bound

for random vectors. This theorem will form the basis of a future paper on the applications of discrimination inequalities to probability theory on infinite dimensional linear spaces. Lack of space prevents the further development of these ideas here, though a preliminary exposition can be found in [10].

Definition 2.4. Let $X=(X_1, \ldots, X_n, \ldots)$ be a finite or infinite sequence of regular random variables with score functions h_n, respectively. For λ in $(0, 2]$ define $\sigma_\lambda(X)=\sup_n \sigma_\lambda(X_n)$. We shall say X is *tame of order 2* if $\sigma_2(X)<\infty$ and we shall set $J(2, X)=\sup_n \eta(X_n) M(\partial h_n)$ in that case. (See Remark 2.8.) If $\sigma_2(X)=\infty$, but $\sigma_\lambda(X)<\infty$ for some $\lambda\in(0, 2)$ we shall say X is *tame of order* λ and we shall define $J(\lambda, X)=\sup_n 2M(h_n/r)\vee(\log(2))^{-1}\eta(X_n)M(\partial h_n)$.

Given the matrix $A=(a_{ij})$, we let a_A and $\hat{A}$ stand for its diagonal part $(a_{ij}\delta_{ij})$ and non-diagonal part $(a_{ij}(1-\delta_{ij}))$, respectively. We define $N_\lambda(A)=\sum_{i,j}|a_{ij}|^\lambda$ and $N_\lambda(\hat{A})=\sum_{i\neq j}|a_{ij}|^\lambda$. (We shall also use the notation $N_\lambda(x)$ for vectors x, i.e., $N_\lambda(x)=\sum_i|x_i|^\lambda$.)

Theorem 2.3. *Let $X=(X_1, \ldots, X_n)$ be a finite sequence of independent random variables such that X is tame of order λ for some $\lambda\in(0, 2]$. Given the invertible matrix $A=(a_{ij})$, form the random vector $Y=AX$. Then*

$$K^*(Y:X)=\sum_1^n K^*_{Z_i}(Y_i:X_i), \tag{2.39}$$

where $Z_i=\sum_{j\neq i} a_{ij}X_j$, $K^*(Y:X)=K(Y, X)+\log(|\det(A)|)$ *and* $K^*_{Z_i}(Y_i:X_i)=K_{Z_i}(Y_i:X_i)+\log(|a_{ii}|)$. *Setting* $\bar{Y}_i=a_{ii}X_i-Z_i$ *and* $\bar{Y}=(\bar{Y}_1, \ldots, \bar{Y}_n)$, *we get*

$$K^*(Y:X)+K^*(\bar{Y}:X)\leqq 2\sum_1^n K^*_{Z_i}(a_{ii}X_i:X_i)+\Gamma(\lambda)J(\lambda, X)\sigma_\lambda(X)N_\lambda(\hat{A}). \tag{2.40}$$

Proof. Letting $W=A^{-1}X$, we have $K(Y:X)=K(X:W)$. The density of X is $f(x)\equiv\Pi f_i(x_i)$, where f_i is the density of X_i; hence the density of W is $|\det(A)|f(Ax)$. It follows that

$$K(Y:X)=E[\log(f(X)/f(AX))]-\log(|\det(A)|). \tag{2.41}$$

We can write

$$E(\log(f(X)/f(AX))=\sum_1^n E[\log(f_i(X_i)/f_i(a_{ii}X_i+Z_i))]. \tag{2.42}$$

It is clear from definition that $K^*(a_{ii}X_i+z:X_i)$ is the conditional expectation of $\log(f_i(X_i)/f_i(a_{ii}X_i+Z_i))$ given that $Z_i=z$; hence (2.39) holds by virtue of (2.34). (Note that the differential entropy of X_i is finite on account of Theorem 2.2, so all of the above goes through even if some a_{ii} are 0.)

The proof of (2.40) follows by applying (2.33) and (2.39) to Y and $\bar{Y}$. Q.E.D.

Remark 2.12. If a_A and $\hat{A}$ commute, then we can rewrite (2.40) as

$$K(Y:X)+K(\bar{Y}:X)\leqq \log(\det(a_A^2))-\log(\det(a_A^2-\hat{A}^2)) + 2\sum_1^n K_{Z_i}(a_{ii}X_i:X_i)+\Gamma(\lambda)J(\lambda,X)\sigma_\lambda(X)N_\lambda(\hat{A}), \tag{2.43}$$

wherein we have used the identity $(a_A-\hat{A})(a_A+\hat{A})=a_A^2-\hat{A}^2$. Since $\log(\det(a_A^2))-\log(\det(a_A^2-\hat{A}^2))=\log(\det(I-a_A^{-1}\hat{A}^2a_A^{-1}\hat{A}^2a_A^{-1})^{-1})$, we conclude that the right hand side of (2.43) decomposes $K(Y:X)+K(\bar{Y}:X)$ into non-negative components, under the condition that $I\geqq a_A^{-1}\hat{A}^2a_A^{-1}$. If $\alpha I\geqq a_A^{-1}\hat{A}^2a_A^{-1}$ for some $\alpha\in(0,1)$, then we can use the inequality

$$\log(\det(I-a_A^{-1}\hat{A}^2a_A^{-1})^{-1})\leqq \operatorname{trace}(a_A^{-1}\hat{A}^2a_A^{-1}(I-a_A^{-1}\hat{A}^2a_A^{-1})^{-1}) \tag{2.44}$$

to simplify (2.43), since the right hand side of (2.44) is bounded by

$$(1-\alpha)^{-1}\operatorname{trace}(a_A^{-1}\hat{A}^2a_A^{-1})=(1-\alpha)^{-1}\sum_{i\neq j}a_i^{-1}a_{ij}^2a_j^{-1}. \tag{2.45}$$

(The inequality (2.44) is essentially a consequence of Hadamard's inequality and can be found in Simon [17, p. 47].)

3. Parameter estimation for auto-regressive processes

In this section we will combine the one-dimensional discrimination distance bounds from Sect. 2 with the general results concerning statistical estimation from Sect. 4 in order to study a particular "non-regular" example where a finite dimensional parameter has infinite Fisher information. The example we will study is the class of p-th order autoregressive processes with infinite variance. We start by establishing notations and conventions about general statistical estimation.

3.1. Conventions

Our results about statistical estimation will be developed in the context of a collection of probability measures $(P_\alpha:\alpha\in M)$ where M is a separable metric space equipped with a metric ρ. The measures P_α will all be defined on $\mathbb{R}^\infty$. In the typical situation we observe the first n coordinates $y^{[n]}$ of a point $y=(y_1, y_2, \ldots, y_n, \ldots)$ in $\mathbb{R}^\infty$ and attempt to estimate the parameter $\alpha\in M$ by a sequence V_n of Borel measurable functions from $\mathbb{R}^n$ to M. We shall treat V_n as a random variable V_n^* on $(\mathbb{R}^\infty, P_\alpha)$ by setting

$$V_n^*(y)=V_n(y^{[n]}) \quad (\text{for } y\in R^\infty). \tag{3.1}$$

We shall denote the distribution function of $\rho(V_n^*,\alpha)$ under P_α by F_α^n, i.e.

$$F_\alpha^n(t)=P_\alpha[\rho(V_n^*,\alpha)<t], \tag{3.2}$$

for t in $\mathbb{R}$. (Note that $F_\alpha^n(t)=0$ for $t\leqq 0$ and that F_α^n is continuous from the left.) We shall write $P_\alpha^{[n]}$ to stand for the image of the measure P_α under the restriction mapping $y \to y^{[n]}$.

3.2. *Auto-regressive processes*

The class of p-th order autoregressive processes is defined by the real parameters $\alpha=(\alpha_j: 1\leqq j\leqq p)$ and the equation

$$Y_n+\alpha_1 Y_{n-1}+\ldots+\alpha_p Y_{n-p}=X_n, \tag{3.3}$$

where $X=(X_k: k=0, \pm 1, \pm 2, \ldots)$ is an infinite two-sided sequence of independent identically distributed random variables. We shall assume for the rest of this section that X is tame of order $\lambda\in(0, 2]$ with associated constants $J_1(X)=J_1(X_0)$ and $\sigma_\lambda(X)$ defined in (2.19) and (Def. 2.4), respectively.

Let ω be any non-negative integer or $+\infty$. Given w, the corresponding solution of (3.3) can be written as

$$Y_n(\alpha)=\sum_{k=-\omega}^{n} A_{n-k}^{(p)} X_k \quad (\text{for } n\geqq\omega). \tag{3.4}$$

The coefficients $A_r^{(p)}$ have the representation

$$A_r^{(p)}=\sum_{\{r_i\}}\prod_{i=1}^{p}(a_i)^{r_i} \quad (\text{summed over } r_1+\ldots+r_p=r\geqq 0), \tag{3.5}$$

where $r_1, \ldots, r_p$ are non-negative integers and $a_1, \ldots, a_p$ stands for the complex roots of the characteristic equation $z^p+\alpha_1 z^{p-1}+\ldots+\alpha_p=0$. If ω is finite, we have $Y_{-\omega}(\alpha)=X_{-\omega}$ and $Y_n(\alpha)=0$ for $n<-\omega$. If ω is infinite, then we need the condition that $\max(|a_j|)<1$ to insure that the infinite sum in (3.4) converges a.s. We shall observe the random variables $Y_n(\alpha)$ only for $n\geqq 1$; hence ω is a way of specifying the initial values $(Y_1(\alpha), \ldots, Y_p(\alpha))$ of the p-th order Markov process $Y(\alpha)=(Y_1(\alpha), \ldots, Y_n(\alpha), \ldots)$. If $\omega=\infty$, then $Y(\alpha)$ is stationary.

It turns out that the condition $\max(|a_j|)<1$ is necessary for the application of our discrimination distance inequalities even when ω is finite. For this reason it is necessary to express the condition $\max(|a_j|)<1$ in terms of the actual parameters $\alpha=(\alpha_j: 1\leqq j\leqq p)$. We let τ denote the mapping from $\mathbb{R}^p$ into itself such that for $1\leqq j\leqq p$, $\tau(a)_j=\alpha_j$ is the coefficient of z^{p-j} in $\prod_{j=1}^{p}(z-a_j)$. We define

$$M_p=\tau(\{a: a\in R^p, \max(|a_j|)<1\}). \tag{3.6}$$

We shall treat M_p as a metric space with metric ρ satisfying

$$c\rho(\alpha, \beta)\leqq\rho_2(\alpha, \beta)\leqq C\rho(\alpha, \beta) \tag{3.7}$$

for all $\alpha, \beta\in M_p$, where $\rho_2(\alpha, \beta)$ is the usual Euclidean metric $(N_2(\alpha-\beta))^{1/2}$ and $0<c<C$ are constants. (See Def. 2.5 for N_2.)

For $\alpha \in M_p$ we let P_α denote the probability measure on $\mathbb{R}^\infty$ induced by the stochastic process $Y(\alpha)$ satisfying (3.4). (The parameter ω, which will *not* be estimated, is suppressed.) In order to explain how the results of [4] and [7] indicate that the rate $n^{1/\lambda}$ is best possible in (1.1), we define the distribution function

$$G_\alpha(t) = \liminf_{n \to \infty} F_\alpha^n(c_n^{-1} t), \tag{3.8}$$

given any sequence of constants c_n strictly increasing to ∞ and any sequence of estimates V_n as in (3.2). We define $G_\alpha(+\infty) = \lim_{t \uparrow \infty} G_\alpha(t)$ and $G_\alpha(0+) = \lim_{t \downarrow 0} G_\alpha(t)$. When the sequence c_n is of the form $n^{1/\delta} s_n$, where s_n is slowly varying at ∞, the following theorem shows that no sequence of estimators V_n can converge to α so quickly that $c_n(V_n - \alpha)$ converges properly in distribution for an uncountable set of α. If $\delta = \lambda$, the theorem shows that no sequence of estimators can converge to α so quickly that $c_n(V_n - \alpha)$ converges in distribution to 0 for an uncountable set of α. In the course of the theorem we will sometimes refer to material in Sect. 4 concerning the total variation metric on measures.

Theorem 3.1. *Let the linear process $Y(\alpha) = (Y_n(\alpha): n \geqq 1)$ be defined in terms of X as in* (3.4) *and let P_α denote the corresponding probability measure on $\mathbb{R}^\infty$. Let s_n be slowly varying at ∞ and assume $c_n = n^{1/\delta} s_n$ is strictly increasing with n, where $0 < \delta \leqq \lambda$. Given any sequence V_n of Borel measurable functions, let G_α be defined via* (3.2) *and* (3.8), *where ρ satisfies* (3.7). *Then there exists a countable set M_p' such that* (1.3) *holds for $\alpha \in M_p / M_p'$, where A/B stands for the set of points in A but not in B.*

Proof. We will first prove that

$$K(P_\alpha^{[n]}: P_\beta^{[n]}) \leqq n \Gamma(\lambda) \Omega_\lambda(a) J_1(X) \sigma_\lambda(X) N_\lambda(\beta - \alpha), \tag{3.9}$$

where $\Omega_\lambda(a) = \prod_j (1 - |a_j|^{\lambda \wedge 1})^{-(\lambda \vee 1)}$ and $a = \tau^{-1}(\alpha)$ for τ as in (3.6). We note that

$$K(P_\alpha^{[n]}: P_\beta^{[n]}) \leqq K(W_\alpha^{[n]}: W_\beta^{[n]}) \tag{3.10}$$

by (1.9), where $W_\alpha^{[n]}$ stands for the infinite vector $(\ldots, X_{-n}, \ldots, X_{-1}, X_0, Y_1(\alpha), \ldots, Y_n(\alpha))$. We can apply (1.8) inductively to get

$$K(W_\alpha^{[n]}: W_\beta^{[n]}) = \sum_{m=0}^{n-1} K_m(W_\alpha^{[m+1]}: W_\beta^{[m+1]}), \tag{3.11}$$

where $K_m(W_\alpha^{[m+1]}: W_\beta^{[m+1]}) \equiv K_{W_\alpha^{[m]}}((Y_{m+1}(\alpha) \mid W_\alpha^{[m]}): (Y_{m+1}(\beta) \mid W_\beta^{[m]}))$.

The conditional distribution of $Y_{m+1}(\alpha)$ given $W_\alpha^{[m]}$ is identical with the conditional distribution of $Y_{m+1}(\alpha)$ given $(Y_{m-p+1}(\alpha), Y_{m-p+2}(\alpha), \ldots, Y_m(\alpha))$, because $Y(\alpha)$ is a p-th order Markov process. We use convolution notation

$$(\alpha * Y(\alpha))_m = \sum_{j=1}^{p} \alpha_j Y_{m-j}(\alpha), \tag{3.12}$$

to write

$$(3.13)\quad K_m(W_\alpha^{[m+1]}: W_\beta^{[m+1]}) = E\{\int \log [f(x-(\alpha * Y(\alpha))_m)/f(x-(\beta * Y(\alpha))_m)]\, f(x-(\alpha * Y(\alpha))_m)\, dx\},$$

where f is the common density of the random variables X_k. The simple change of variables $x-(\alpha * Y(\alpha))_m = y$ then yields

$$(3.14)\quad K_m(W_\alpha^{[m+1]}: W_\beta^{[m+1]}) = K_{Z_m}(X_m + Z_m : X_m),$$

where $Z_m = ((\beta-\alpha) * Y(\alpha))_m$. Using (2.32), we bound the right hand side of (3.14):

$$(3.15)\quad K_{Z_m}(X_m+Z_m:Z_m) \leqq \Gamma(\lambda) J_1(X)\, \sigma_\lambda(X) \left(\sum_{k=1}^{m+\omega} \left| \sum_{j=1}^{p \wedge k} (\beta_j - \alpha_j) A_{k-j}^{(p)} \right|^\lambda \right).$$

It is convenient to define $(\beta_j-\alpha_j)=0$ for $j \notin \{1, \ldots, p\}$ and to define $A_r^{(p)}=0$ for $r<0$. We then recognize the sum from 1 to $p \wedge k$ in (3.15) as a true convolution (which we shall denote as $(\beta-\alpha) * A^{(p)}$, in accord with the notation in (3.12)). We now borrow from harmonic analysis (see [5]) the well known inequality

$$(3.16)\quad N_\lambda(c * A) \leqq N_\lambda(c) (N_{\lambda \wedge 1}(A))^{\lambda \vee 1},$$

to get

$$(3.17)\quad K_{Z_m}(X_m+Z_m:Z_m) \leqq \Gamma(\lambda) J_1(X)\, \sigma_\lambda(X)\, N_\lambda(\beta-\alpha) (N_{\lambda \wedge 1}(A^{(p)}))^{\lambda \vee 1}.$$

(If $\lambda \geqq 1$, then (3.16) is a special case of Young's inequality. If $\lambda \in (0, 1)$ then (3.16) follows directly from the sub-additivity of $x \to |x|^\lambda$.) Since the sequence $A^{(p)} = (A_r^{(p)}: r \geqq 0)$ defined in (3.5) is the p-fold convolution of the sequences $A(i) = (1, a_i, a_i^2, \ldots, a_i^r, \ldots)$ for $1 \leqq i \leqq p$, we can apply (3.16) p times to get

$$(3.18)\quad N_{\lambda \wedge 1}(A^{(p)}) \leqq (\Omega_\lambda(a))^{1/(\lambda \vee 1)}.$$

The inequality (3.9) follows from (3.11), (3.14), (3.17) and (3.18).

We now recall [11, p. 69, Problem 7.32] the relation

$$(3.19)\quad D^2(P, Q) \leqq 4K(P, Q)$$

valid for any probability measures P and Q. For $\gamma = (\gamma_1, \ldots, \gamma_p)$ with $0 < \gamma_i < 1$ for all i, (3.9) and (3.19) yield

$$(3.20)\quad D_n^2(P_\alpha, P_\beta) \leqq 4n\Gamma(\lambda) \Omega_\lambda(\gamma) J_1(X)\, \sigma_\lambda(X)\, N_\lambda(\beta-\alpha)$$

if $|a_i| \leqq \gamma_i$ for all i. (See the paragraph containing (4.1) for the definition of D and D_n.)

If we assume temporarily that $\rho(\alpha, \beta) = \rho_2(\alpha, \beta)$, then $\rho(\alpha, \beta)^\lambda\, p^{(2-\lambda)/2} \geqq N_\lambda(\beta-\alpha)$; whence by (1.17)

$$(3.21)\quad d_\alpha(r) \leqq r^{\lambda/2} (\Gamma(\lambda) \Omega_\lambda(\gamma) J_1(X)\, \sigma_\lambda(X))^{1/2}\, p^{(2-\lambda)/4} \liminf_{n \to \infty} t_n,$$

where $t_n = (n-1)^{((\delta-\lambda)/2\delta)} (s_{n-1})^{-\lambda/2}$. Using Theorem 4.1 and letting $\gamma_i \uparrow 1$ for all i, we conclude that there exists a countable set M_p' such that (4.5) holds for

$\alpha \in M_p/M_p'$ and all rational r. The relations in (1.3) then follow immediately from (3.21).

We remove the assumption that $\rho = \rho_2$ by noting that (1.3) holds for $\rho = \rho_2$ if and only if (1.3) holds for ρ satisfying (3.7). Q.E.D.

4. Parameter estimation and total variation proximity

In this section we will present some general results linking asymptotic parameter estimation with the total variation distance between probability measures.

If P and Q are probability measures on $\mathbb{R}^\infty$ we define the total variation distance $D(P, Q)$ by

$$D(P, Q) = 2 \sup |P(C) - Q(C)| \tag{4.1}$$

where the supremum is taken over all Borel subsets C of $\mathbb{R}^n$. We shall write $D_n(P_\alpha, P_\beta)$ to stand for $D(P_\alpha^{[n]}, P_\beta^{[n]})$. Other notation will follow the conventions in Paragraph 3.1.

Our first result is almost a triviality, yet is fundamental.

Lemma 4.1. *Let $(P_\alpha : \alpha \in M)$ be a collection of probability measures on $\mathbb{R}^\infty$ and let V_n be a Borel measurable function from $\mathbb{R}^n$ into M. Then, for $t \leqq \rho(\alpha, b)/2$:*

$$F_\alpha^n(t) + F_\beta^n(t) \leqq 1 + (1/2) D_n(P_\alpha, P_\beta). \tag{4.2}$$

Proof. The triangle inequality for ρ implies

$$[\rho(V_n^*, \alpha) \geqq t] \cup [\rho(V_n^*, \beta) \geqq t] = \mathbb{R}^\infty, \tag{4.3}$$

whence it follows that

$$(1 - F_\alpha^n(t)) + P_\alpha[\rho(V_n^*, \beta) \geqq t] \geqq 1. \tag{4.4}$$

We use (4.1) to compare the P_α and P_β measure of $[\rho(V_n^*, \beta) \geqq t]$, and thereby get (4.2). Q.E.D.

Theorem 4.1. *Let M be a separable metric space and let $(P_\alpha : \alpha \in M)$ be a collection of probability measures on $\mathbb{R}^\infty$. Let c_n be a sequence of positive numbers strictly increasing to ∞. Define $d_\alpha(r)$ as in (1.17) and $G_\alpha(r)$ as in (3.8). Then, for any sequence V_n of Borel measurable functions from $\mathbb{R}^n$ to M, there exists a countable subset M' such that*

$$G_\alpha(r/2) \leqq (1 + d_\alpha(r))/2 \tag{4.5}$$

for all $\alpha \in M/M'$ and all rational $r > 0$.

Proof. For any $v > 1$, define

$$B_n^v(r) = \{\alpha \in M : F_\alpha^n(c_n^{-1} r/2) > (v + d_\alpha(r))/2\} \tag{4.6}$$

and let $B^v(r)$ stand for the set of α in M which belong to all but finitely many $B_n^v(r)$ as n varies. We claim that $B^v(r)$ contains no limit points. Suppose, by way of contradiction, that the claim is false and that there exists a sequence β_k in $B^v(r)$ and a point α in $B^v(r)$ such that the distance $\rho(\alpha, \beta_k)$ is strictly

decreasing to 0. It follows that there exists a sequence n_k tending to ∞ such that

$$(r/c_{n_k}) \leqq \rho(\alpha, \beta_k) < (r/c_{n_k-1}) \tag{4.7}$$

for k sufficiently large, whence

$$D_{n_k}(\alpha, \beta_k) \leqq 2(v-1+d_\alpha(r)) \tag{4.8}$$

for k sufficiently large. Using Lemma 4.1, we conclude that

$$F_\alpha^{n_k}(\rho(\alpha, \beta_k)/2) + F_{\beta_k}^{n_k}(\rho(\alpha, \beta_k)/2) \leqq v + d_\alpha(r) \tag{4.9}$$

for k sufficiently large. Since $\rho(\alpha, \beta_k) \geqq r/c_{n_k}$, (4.9) is in contradiction to the assumption that α and β_k are in $B^v(r)$; thus the claim is proved.

We refer now to the considerations in [12, Th. 1, p. 161] leading up to the Cantor-Bendixson Theorem and conclude that $B^v(r)$ must be countable. The proof of the theorem is completed by letting v tend to 1. Q.E.D.

Remark 4.1. Some of the arguments in Theorem 4.1 are similar in spirit to those in Proposition 9 of [13, p. 191], though the details and applications are quite different.

Remark 4.2. A more measure theoretic version of Theorem 4.1 was proved in [10]. That proof applies in the case when M is a subset of any locally compact group, μ is Haar measure, and ρ is a group invariant metric. Under extra measurability assumptions, (4.5) is demonstrated to hold for μ almost all α in M; in that version the supremum over β defining $d_\alpha(r)$ in (1.17) is taken over β such that $\rho(\alpha, \beta) = r c_n^{-1}$.

Remark 4.3. If $G_\alpha(+\infty) = 1$, then $d_\alpha(+\infty) = 1$ as a consequence of (4.5). This remark facilitates understanding the relationship of Theorem 4.1 to "c_n-consistent" estimates defined in (1.15).

Acknowledgements. I would like to thank Professor Lucien Le Cam for a critical reading of the first version [10] of this work and for the subsequent discussions we enjoyed on these topics.

References

1. Akahira, M., Takeuchi, K.: The concept of asymptotic efficiency and higher order efficiency in statistical estimation theory. (Lect. Notes Stat., vol. 7). Berlin Heidelberg New York: Springer 1981
2. Basawa, I.V., Prakasa Rao, B.L.S.: Statistical inference for stochastic processes. New York: Academic Press 1980
3. Birgé, L.: Vitesses optimales de convergence des estimateurs, in grandes deviations et applications statistiques. Société Math. de France. Astérisque 6, Paris 1979
4. Davis, R., Resnick, S.: Limit theory for the sample covariance and correlation functions of moving averages. Ann. Stat. **14**, 532–558 (1986)
5. Edwards, R.: Fourier series: a modern introduction, vol. II. Berlin Heidelberg New York: Springer 1982
6. Hannan, E.J., Hesse, C.H.: Rates of convergence for the quantile function of a linear process. Aust. J. Stat. **30** A, 283–295 (1988)

7. Hannan, E.J., Kanter, M.: Autoregressive processes with infinite variance. J. Appl. Probab. **14**, 411–415 (1977)
8. Ibragimov, I.A., Has'Minskii, R.Z.: Statistical estimation, asymptotic theory. (Applications of math. series vol. 16.) Berlin Heidelberg New York: Springer 1981
9. Kanter, M., Steiger, W.L.: Regression and autoregression with infinite variance. Adv. Appl. Probab. **6**, 768–783 (1974)
10. Kanter, M.: Entropy bounds for stochastic processes and statistical applications. (1982 unpublished)
11. Kullback, S.: Information theory and statistics. New York: Wiley 1959
12. Kuratowski, K.: Introduction to set theory and topology. Reading, Mass.: Pergamon Press/ Addison Wesley 1962
13. Le Cam, L.: Notes on asymptotic methods in statistical decision theory. Centre de recherches mathématiques, Université de Montréal, Montréal
14. Le Cam, L.: Asymptotic methods in statistical decision theory. Berlin Heidelberg New York: Springer 1986
15. Loéve, M.: Probability theory, vol. 1. Berlin Heidelberg New York: Springer 1977
16. Roussas, G.C.: Contiguity of probability measures. Cambridge: Cambridge University Press 1972
17. Simon, B.: Trace ideals and their applications. London mathematical lecture notes series 35. Cambridge: Cambridge University Press 1979
18. Vostrikova, L.J.: On criteria for c_n-consistency of estimators. Stochastics **11**, 265–290 (1984)