Minimax Risk Inequalities for the Location-Parameter Classification Problem

Pieter C. Allaart

December 16, 2000

Department of Mathematics and Computer Science
Vrije Universiteit Amsterdam

Abstract

Minimax risk inequalities are obtained for the location-parameter classification problem. For the classical single observation case with continuous distributions, best possible bounds are given in terms of their Lévy concentration, establishing a conjecture of Hill and Tong (1989). In addition, sharp bounds for the minimax risk are derived for the multiple (i.i.d.) observations case, based on the tail concentration and the Lévy concentration. Some fairly sharp bounds for discontinuous distributions are also obtained.

1 Introduction

In the classification problem, in its standard form, an observation X is given and the statistician's task is to guess from which of several specified distributions it comes. More precisely, if F_1,...,F_n are probability distributions on the real line and X is a random variable having unknown distribution F, then a test for testing the hypotheses

$$H_i : F = F_i, \qquad i = 1, ..., n$$

is sought which achieves the minimax risk, i.e. minimizes the largest probability of misclassification.

As a practical example, one might think of scoring systems used in mental health, where it is assumed that patients from different socio-psychological backgrounds score essentially differently on a certain psychological test, and the psychiatrist's task is to recover the patient's background from his score.

Another example is the so-called *two-armed slot machine problem*: if the arms of a two-armed slot machine have different payoff distributions, then the gambler at some point has to decide which arm gives him the best average payoff, and pull this arm only from then on.

The two-armed slot machine problem has far-reaching implications in the medical world, where the two arms represent two different drugs that are to be tested, and giving the wrong drug to a patient may have fatal consequences.

This paper focuses on the special case of the classification problem where the shape and scale of the distribution F are known, but the location is unknown. It answers affirmatively a question raised by Hill and Tong [6] (see also Open Problem no. 11 in [7]), who give best possible bounds for the minimax risk in terms of the tail d-concentration ρ (definition 2.3 below) in case F is continuous, and ask whether the same inequality holds when ρ is replaced by the Lévy concentration λ. Hill and Tong show that the inequality does hold when $n = 2$, and Section 3 of this paper combines their ideas and an induction principle to prove the result for general n. As a corollary, a non-trivial bound is obtained for the minimax risk in terms of variance.

The second main result of this paper is Theorem 5.1, which gives a sharp bound for the minimax risk in terms of the tail d-concentration in case several i.i.d. observations are available. From this inequality, a sufficient condition can be derived on the number of observations in order for the minimax risk to be less than a given confidence level.

The organization of this paper is as follows. Section 2 contains preliminaries and relates multi-hypotheses testing to the theory of optimal-partitioning, an important tool of which is the convexity result by Dvoretzky, Wald and Wolfowitz [2], which is stated in Proposition 2.2.

Section 3 then solves the conjecture by Hill and Tong, with their theorem for the tail d-concentration as a corollary. In addition, a non-trivial bound is given in terms of the variance.

The case of measures with atoms is considered in Section 4. First it is shown that if *randomized* decision rules are allowed, then the bounds from Section 3 still hold. A minimax risk inequality is then given for the case where randomizing is not allowed. This bound follows from Theorem 3.1 and a generalization of Proposition 2.2 to measures with atoms.

Section 5, which studies classification based on multiple observations, gives a sharp minimax risk inequality in terms of the tail d-concentration, and shows that the same bound fails for Lévy concentration if the number of observations is at least two.

2 Notation and basic tools

Throughout this paper, $\mu, \mu_1, \ldots, \mu_n$ will always denote (countably additive) probability measures on ($\mathbb{R}$,$\mathcal{B}$), the real line equipped with the Borel σ-algebra. The corresponding distribution functions of $\mu, \mu_1,...,\mu_n$ will be denoted by $F, F_1, \ldots, F_n$, respectively.

For a sequence of measures $\mu_1,...,\mu_n$, the *vector measure* $\vec{\mu} = (\mu_1,...,\mu_n)$ is defined by

$$\vec{\mu}(A) := (\mu_1(A), ..., \mu_n(A)) \in \mathbb{R}^n, \qquad A \in \mathcal{B}.$$

For a probability measure μ, a set $E \in \mathcal{B}$ is an *atom* of μ if $\mu(E) > 0$ and for all $F \subset E, F \in \mathcal{B} : \mu(F) = 0$ or $\mu(F) = \mu(E)$; A measure is called *atomless* if it does not have any atoms. Note that a measure μ on $(\mathbb{R},\mathcal{B})$ is atomless if and only if $\mu(\{x\}) = 0$ for all $x \in \mathbb{R}$. By a *general* probability measure will be meant *any* probability measure on $(\mathbb{R},\mathcal{B})$, atomless or not.

A (measurable) *partition* is an ordered collection $(A_i)_{i=1}^n$ of Borel-measurable subsets of $\mathbb{R}$ such that $A_i \cap A_j = \emptyset$ for all $i \neq j$, and $\bigcup_{i=1}^n A_i = \mathbb{R}$.

In testing the multiple hypotheses

$$H_i : F = F_i, \qquad i = 1, ..., n, \tag{1}$$

a *decision rule* corresponds to a partition $(A_i)_{i=1}^n$ such that H_i is accepted if and only if $X \in A_i$. The i-th *risk* of a decision rule $(A_i)_{i=1}^n$ is defined by

$$R_i((A_j)_{j=1}^n) := \text{Prob}(X \notin A_i \mid H_i) = 1 - \mu_i(A_i),$$

and the *minimax risk* for the hypotheses (1) by

$$\inf\{\max_{1\leq i\leq n} R_i((A_j)_{j=1}^n) \mid (A_j)_{j=1}^n \text{ is a decision rule}\} = 1 - C_n^*(\vec{\mu}),$$

where

$$C_n^*(\vec{\mu}) = \sup\{\min_{1\leq i\leq n} \mu_i(A_i) \mid (A_i)_{i=1}^n \text{ is a partition of } \mathbb{R}\}.$$

Thus the problem of testing multiple hypotheses is equivalent to one of *fair division*, i.e. partitioning an object (in this case the real line) among n persons so that the minimum share of all persons, according to their own respective measures, is as large as possible. It is in this setting that most of the results in this paper will be stated and proved.

An important notion in the theory of fair division is the *partition range*:

Definition 2.1 *For a vector measure $\vec{\mu} = (\mu_1,...,\mu_n)$, the* partition range $\mathcal{PR}(\vec{\mu})$ *is defined by*

$$\mathcal{PR}(\vec{\mu}) := \{(\mu_1(A_1), ..., \mu_n(A_n)) \mid (A_i)_{i=1}^n \textit{ is a partition of } \mathbb{R}\}.$$

The following result is a fundamental tool in this article.

Proposition 2.2 [Dvoretzky, Wald and Wolfowitz (1951)]. *If $\mu_1,...,\mu_n$ are atomless, then $\mathcal{PR}(\vec{\mu})$ is compact and convex.*

Most of the bounds in this paper will be some function of one of the following two concentrations.

Definition 2.3 *For a probability measure μ and a positive real number d,*

(i) the tail d-concentration $\rho(\mu, d)$ *is defined by*

$$\rho(\mu, d) = \max\{\mu((-\infty, \text{ess inf } \mu + d)), \mu((\text{ess sup } \mu - d, \infty))\}, \textit{ and}$$

(ii) the Lévy d-concentration $\lambda(\mu, d)$ *is defined by*

$$\lambda(\mu, d) = \sup_{x\in\mathbb{R}} \mu((x, x + d)).$$

Note that by definition, $\lambda(\mu,d) \geq \rho(\mu,d)$. Furthermore, it is not difficult to see that $\lambda(\mu,d) > 0$ for all $d > 0$. Although $\lambda := \lambda(\mu,d)$ need not be attained in general, the next lemma says that λ is always attained by *half open* intervals (see, e.g. Theorem 1.1.8 and the remark at the bottom of p.9 of Hengartner and Theodorescu [5]).

Lemma 2.4 *For all $\lambda := \lambda(\mu,d) > 0$, there exists an $x \in \mathbb{R}$ such that either $\mu((x,x+d] = \lambda$ or $\mu([x,x+d)) = \lambda$*

If λ is close to zero, then the distribution μ is very flat. On the other hand, if λ is close to one, then μ is essentially concentrated on an interval of length d. Thus Lévy concentration, like variance, provides a measure of how spread-out a distribution is. In fact, using a slightly different definition of Lévy concentration, the following concentration-variance inequality holds; here $\lambda^c(\mu,d) := \sup_{x\in\mathbb{R}} \mu([x,x+d])$.

Proposition 2.5 [Lévy (1937) - see also [5], p. 27] *For every probability measure μ with finite variance σ_μ^2,*

$$\sigma_\mu^2 \geq \frac{d^2}{12} m(m+1)(3 - \lambda^c(\mu,d)\cdot(2m+1)),$$

where $m = \max\{j \in \mathbb{N} : \mathrm{j} < [\lambda^c(\mu, \mathrm{d})]^{-1}\}$. Equivalently,

$$\lambda^c(\mu,d) \geq f(\sigma_\mu^2, d) := \frac{3}{2m+1}(1 - \frac{4}{m(m+1)} \cdot \frac{\sigma_\mu^2}{d^2}), \tag{2}$$

where $m \in \mathbb{N}$ is such that $\frac{d^2}{12}(m^2-1) < \sigma_\mu^2 \leq \frac{d^2}{12}(m^2+2m)$.

Note that for atomless distributions, the two definitions of Lévy concentration coincide, and the above inequalities hold also for $\lambda(\mu,d)$. It is not known to the author whether λ^c can be replaced by λ in Proposition 2.5 for general distributions.

Roughly speaking, Proposition 2.5 says that a small variance implies a large value of the Lévy concentration. The converse is false: a distribution with Lévy concentration close (but not equal) to one can still have arbitrarily large variance, as the following example shows:

Example 2.6 Let $\varepsilon > 0$ be given, and let μ be the measure $\mu = (1-\varepsilon)\delta_{\{0\}} + \varepsilon\delta_{\{M\}}$, for some $M \gg 0$ where $\delta_{\{x\}}$ is the Dirac measure on x. Then $\lambda(\mu,d) = 1-\varepsilon$ for all $d > 0$, but $\sigma_\mu^2 = \varepsilon(1-\varepsilon)M^2$. Since M is arbitrary, it follows that the variance of μ can be arbitrarily large.

Inequality (2) will be used in the next section to derive a bound on the minimax risk in terms of the variance.

For the remainder of this article, it will be assumed that $\mu_1,\ldots,\mu_n$ belong to the same location-parameter family, and have equally spaced location parameters. That is, there exist a probability measure μ and a real number $d > 0$ such that

$$\mu_i(A) = \mu(A - (i-1)d), \qquad i = 1,\ldots,n, \quad A \in \mathcal{B}, \tag{3}$$

where $A - x \equiv \{a - x : a \in A\}$.

3 Atomless distributions

The main result of this section is the following theorem.

Theorem 3.1 *If μ is atomless, then*

$$C_n^*(\vec{\mu}) \geq \left[\sum_{j=0}^{n-1}(1-\lambda)^j\right]^{-1}, \tag{4}$$

where $\lambda = \lambda(\mu, d)$ is the Lévy d-concentration of μ. This bound is attained for all n, λ and d.

Remark 3.2 Inequality (4) was proved by Hill and Tong [6] for the special case $n = 2$.

Corollary 3.3 [Hill and Tong (1989), Theorem 2.2] *If μ is atomless, then*

$$C_n^*(\vec{\mu}) \geq \left[\sum_{j=0}^{n-1}(1-\rho)^j\right]^{-1}, \tag{5}$$

where $\rho = \rho(\mu, d)$ is the tail d-concentration of μ, and this bound is attained for all n, ρ and d.

Proof: (5) follows immediately from (4) since the right hand side in (4) is increasing in λ, and since $\lambda \geq \rho$. The sharpness of (5), and hence of (4), will be demonstrated in Section 5 as a special case of Example 5.2. □

Corollary 3.4 *If μ is atomless, then*

$$C_n^*(\vec{\mu}) \geq \left[\sum_{j=0}^{n-1}(1-f(\sigma_\mu^2, d))^j\right]^{-1},$$

where σ_μ^2 is the variance of μ and $f(\sigma_\mu^2, d)$ is defined as in (2).

Proof: Immediate from Proposition 2.5 (and the remark following it) and Theorem 3.1. □

Example 3.5 Let μ be the normal distribution with mean 0 and variance 1, and $d = 1$. Then $\lambda = \lambda(\mu, d) = 0.3829$, and the bounds in (4) for $n = 2, 3, 4$ and 5 are $0.618, 0.500, 0.448$ and 0.421, respectively.

If $d = 1$ and μ is *any* continuous distribution with variance 1, then (2) yields $\lambda = \lambda(\mu, d) \geq 2/7$, hence the right hand sides in (4) for $n = 2, 3, 4$ and 5 are at least $0.583, 0.450, 0.386$ and 0.350, respectively.

The proof of Theorem 3.1 uses the following lemma, which also holds for measures with atoms (a fact that will be needed in Section 4).

Lemma 3.6 *For each $n \in \mathrm{I\!N}$, $\mathcal{PR}(\vec{\mu})$ contains a vector $\vec{s}$ of the form $\vec{s} := (r+\lambda, \lambda, ..., \lambda, 1-r) \in \mathrm{I\!R}^n$ for some $r \in [0, 1-\lambda]$, where $\lambda = \lambda(\mu, d)$.*

Proof: By Lemma 2.4 there exists a $\gamma \in \mathrm{I\!R}$ such that either $\mu((\gamma, \gamma+d]) = \lambda$ or $\mu([\gamma, \gamma+d)) = \lambda$. If $\mu((\gamma, \gamma+d] = \lambda$ let $r = \mu((-\infty, \gamma])$ and consider the partition

$$((-\infty, \gamma+d],\ (\gamma+d, \gamma+2d],\ ...,\ (\gamma+(n-2)d, \gamma+(n-1)d],\ (\gamma+(n-1)d, \infty)),$$

else let $r = \mu((-\infty, \gamma))$ and consider the partition

$$((-\infty, \gamma+d),\ [\gamma+d, \gamma+2d),\ ...,\ [\gamma+(n-2)d, \gamma+(n-1)d),\ [\gamma+(n-1)d, \infty)).$$

In both cases it follows that $\vec{s} \in \mathcal{PR}(\vec{\mu})$. □

Proof of Theorem 3.1 Let $c_n := c_n(\lambda) := [\sum_{j=0}^{n-1}(1-\lambda)^j]^{-1}$. If $n = 1$, then $c_n = c_1 = 1$ and the conclusion of Theorem 3.1 is trivial. The proof now proceeds by induction. For each $n \in \mathbb{N}$, let $\vec{\mu}_n := (\mu_1, ..., \mu_n)$. Suppose that (4) holds for some $n \in \mathbb{N}$. By the compactness of $\mathcal{PR}(\vec{\mu}_n)$ (Proposition 2.2) there exist measurable partitions $(A_i)_{i=1}^n$ and $(B_i)_{i=2}^{n+1}$ such that

$$\begin{aligned} a_i &:= \mu_i(A_i) \geq c_n, \quad i = 1, ..., n, \text{ and} \\ b_i &:= \mu_i(B_i) \geq c_n, \quad i = 2, ..., n+1. \end{aligned}$$

Setting $A_{n+1} = B_1 = \emptyset$, it follows that $\vec{a} := (a_1, ..., a_n, 0)$ and $\vec{b} := (0, b_2, ..., b_{n+1})$ are both in $\mathcal{PR}(\vec{\mu}_{n+1})$.

By Lemma 3.6 (applied to $n+1$), $\mathcal{PR}(\vec{\mu}_{n+1})$ also contains the vector $\vec{s} := (r+\lambda, \lambda, ..., \lambda, 1-r) \in \mathbb{R}^{n+1}$, for some $r \in [0, 1-\lambda]$. Now let $t_1 = (1-r-\lambda)c_n^{-1}c_{n+1}$, $t_2 = rc_n^{-1}c_{n+1}$ and $t_3 = c_{n+1}$. Then $t_i \geq 0$, $i = 1, 2, 3$ and

$$t_1 + t_2 + t_3 = \{(1-\lambda)c_n^{-1} + 1\}c_{n+1} = 1,$$

and hence Proposition 2.2 implies that

$$\vec{v} := t_1\vec{a} + t_2\vec{b} + t_3\vec{s} \in \mathcal{PR}(\vec{\mu}_{n+1}),$$

i.e. there exists a partition $(E_i)_{i=1}^{n+1}$ of $\mathbb{R}$ such that

$$\mu_i(E_i) = v_i, \qquad i = 1, ..., n+1,$$

where

$$\begin{aligned} v_1 &= t_1a_1 + t_3(r+\lambda) \geq (1-r-\lambda)c_{n+1} + c_{n+1}(r+\lambda) = c_{n+1}; \\ v_i &= t_1a_i + t_2b_i + t_3\lambda \geq (1-r-\lambda)c_{n+1} + rc_{n+1} + c_{n+1}\lambda = c_{n+1}, \quad i = 2, ..., n; \end{aligned}$$

and

$$v_{n+1} = t_2b_{n+1} + t_3(1-r) \geq rc_{n+1} + (1-r)c_{n+1} = c_{n+1}.$$

Thus the conclusion of Theorem 3.1 also holds for $n+1$. □

Remark 3.7 Note that the key idea of the proof of Theorem 3.1 was to find several 'good' vectors in the partition range of $\vec{\mu}$, and then use convexity to find a point on the diagonal $\{(\eta, \eta, \ldots, \eta) : \eta \in [0, 1]\}$ lying 'far' away from the origin. This idea, which is a characteristic aspect of the proofs of many partitioning inequalities (e.g. [6], [8]), will return many times throughout the remainder of this paper. Finding vectors which work is often a matter of trial and error.

4 General distributions

If the measure μ has atoms, then the conclusion of Theorem 3.1 may fail in general, as the following simple example shows.

Example 4.1 Let $n = 2$ and $d = 1$, and let μ be the Bernoulli distribution $\mu = \frac{1}{2}\delta_{\{0\}} + \frac{1}{2}\delta\{1\}$. It is clear that $C_2^*(\vec{\mu}) = \frac{1}{2}$, while $\lambda = \lambda(\mu, d) = \frac{1}{2}$ and hence $c_2(\lambda) = \frac{2}{3} > \frac{1}{2}$.

However, if the statistician is allowed to base his decision not only on the observation X, but also on the outcome of some external experiment, like picking a number at random from the unit interval, then he can do just as well as in the atomless case.

To make this more precise, let Q be the uniform distribution on $(0, 1)$ and let U be a random variable with distribution Q, independent of X. A *randomized decision rule* corresponds to a partition $(A_i)_{i=1}^n$ of $\mathbb{R} \times (0, 1)$ such that H_i is accepted if and only if $(X, U) \in A_i$. The *randomized risk set* is the partition range of the vector measure $(\mu_1 \times Q, ..., \mu_n \times Q)$ on $\mathbb{R} \times (0, 1)$.

Lemma 4.2 [Ferguson (1967), Lemma 1.7.1] *The randomized risk set is convex.*

Theorem 4.3 *Let μ be a general probability measure with Lévy d-concentration λ, X a random variable with unknown distribution F, and U a uniform $(0,1)$ variable independent of X. Then there exists a test based on the pair (X,U) for testing the hypotheses (1), which has minimax risk at most*

$$1-\left[\sum_{j=0}^{n-1}(1-\lambda)^j\right]^{-1},$$

and this bound is attained.

Proof: Follow the proof of Theorem 3.1, but now lift all partitions to partitions of $\mathrm{I\!R}\times(0,1)$ via the mapping $A\rightarrow A\times(0,1)$. Lemma 4.2 plus the same arguments as before imply the existence of a partition $(\bar{A}_i)_{i=1}^n)$ of $\mathrm{I\!R}\times(0,1)$ such that

$$(\mu_i\times Q)(\bar{A}_i)\geq c_n(\lambda),\quad i=1,...,n,$$

so that the corresponding risks are

$$\begin{aligned} R_i((\bar{A}_j)_{j=1}^n) &= \text{Prob}((X,U)\notin\bar{A}_i\mid H_i)=1-(\mu_i\times Q)(\bar{A}_i)\\ &\leq 1-c_n(\lambda),\quad i=1,...,n. \end{aligned}$$

The bound is attained by the same distribution which attains the bound of Theorem 3.1 (see Example 5.2). This follows from the well-known fact that for continuous distributions randomized decision rules do not perform better than non-randomized decision rules (e.g. [2], §4). □

If randomized decision rules are not allowed, then in many cases non-trivial bounds can still be obtained. For example, Gouweleeuw [4] gives sufficient conditions on the atoms of a vector measure $\vec{\mu}$ such that $\mathcal{PR}(\vec{\mu})$ is convex, in which case the bound in (4) still holds. A different approach may be based on the following idea: if the atoms of $\vec{\mu}$ are small, then the partition range $\mathcal{PR}(\vec{\mu})$ is "almost convex", and the bound in Theorem 3.1 is *almost* attained.

For a vector $x\in\mathrm{I\!R}^n$, let $\|x\|_\infty:=\max_{1\leq i\leq n}|x_i|$ denote the l^∞-norm of x. For a set $A\subset\mathrm{I\!R}^n$ and a point $x\in\mathrm{I\!R}^n$, let $d_\infty(x,A):=\inf_{y\in A}\|x-y\|_\infty$ denote the l^∞-distance from x to A, and for any subset $A\subset\mathrm{I\!R}^n$, let $D_\infty(A):=\sup\{d_\infty(x,A):x\in\text{Conv}(A)\}$ denote the Hausdorff-∞-distance from A to its convex closure.

The next theorem, which is taken from [1], improves the bound of Theorem 3.2 in [6].

Theorem 4.4 *Let $\vec{\nu}=(\nu_1,\dots,\nu_n)$, where $\nu_1,\dots,\nu_n$ are finite non-negative measures. If $\nu_i(E)\leq\alpha$ for every atom E of ν_i and all i, then*

$$D_\infty(\mathcal{PR}(\vec{\nu}))\leq\frac{n-1}{n}\alpha,$$

and this bound is attained for all $\alpha>0$.

Together with a slightly modified proof of Theorem 3.1, this implies the following inequality for probability measures with a bounded atom size.

Theorem 4.5 *Let μ be a general probability measure. If $\mu(E)\leq\alpha$ for every atom E of μ, then*

$$C_n^*(\vec{\mu})\quad\geq\quad c_n(\lambda)-\frac{n-1}{n}\alpha,$$

where $\lambda=\lambda(\mu,d)$.

Remark 4.6 The author does not know if this inequality is sharp.

Example 4.7 Let $n=3,\lambda=1/2$ and $\alpha=1/10$. Then the right hand side above evaluates to $c_3(1/2)-1/15\approx 0.505$.

5 Multiple observations

In the previous sections, only the classical form of the classification problem with a single observation was studied. In many practical situations, however, a sequence $X_1, ..., X_k$ of observations will be available, and one would expect to be able to reduce the maximum risk by using the information contained in the full vector $(X_1, ..., X_k)$ instead of only that of X_1.

What happens in the classification problem when several observations are available? It turns out that optimal partitioning is again the proper background, but now decision rules correspond to partitions of $\mathbb{R}^k$.

For the remainder of this section, $X_1, ..., X_k$ are independent, identically distributed random variables with common distribution F. Let $\mathcal{B}^k$ denote the product σ-algebra $\mathcal{B} \times \mathcal{B} \times \ldots \times \mathcal{B}$ on $\mathbb{R}^k$. A decision rule for the hypotheses (1) corresponds to a $\mathcal{B}^k$- measurable partition $(A_i)_{i=1}^n$ of $\mathbb{R}^k$ such that H_i is accepted if and only if $(X_1, ..., X_k) \in A_i$. For a measure μ on $\mathbb{R}$, and $k \in \mathbb{N}$, let μ^k denote the k-dimensional product measure determined by

$$\mu^k(E_1 \times ... \times E_k) = \mu(E_1) \cdot ... \cdot \mu(E_k),$$

for all $E_1, ..., E_k \in \mathcal{B}$. Then the i-th risk of a decision rule $(A_i)_{i=1}^n$ is

$$R_i((A_j)_{j=1}^n) = \text{Prob}((X_1, ..., X_k) \notin A_i \mid H_i) = 1 - \mu_i^k(A_i),$$

and the minimax risk is

$$\inf\{\max_{1 \le i \le n} R_i((A_j)_{j=1}^n) \mid (A_j)_{j=1}^n \text{ is a decision rule}\} = 1 - C_{n,k}^*(\vec{\mu}),$$

where

$$C_{n,k}^*(\vec{\mu}) = \sup\{\min_{1 \le i \le n} \mu_i^k(A_i) \mid (A_j)_{j=1}^n \text{ is a partition of } \mathbb{R}^k\}.$$

Thus the problem of testing the multiple hypotheses (1) using an i.i.d. sequence of k observations is equivalent to partitioning the space $\mathbb{R}^k$ among the n measures $\mu_1^k, ..., \mu_n^k$.

The next theorem is the k-dimensional analogue of Corollary 3.3.

Theorem 5.1 *If μ is atomless with tail d-concentration ρ, then*

$$C_{n,k}^*(\vec{\mu}) \ge \left[\sum_{j=0}^{n-1} (1-\rho)^{kj}\right]^{-1}, \tag{6}$$

and this bound is attained for all n, k, d and ρ.

Proof: The proof is analogous to the proof of Theorem 2.2 of Hill and Tong [6].

If $\rho = 0$, then the bound in (6) is $1/n$ and inequality (6) holds for any atomless distributions $\mu_1, ..., \mu_n$ (cf. [7], §3). Suppose that $0 < \rho \le 1$. Then *ess inf* $\mu > -\infty$ or *ess sup* $\mu < \infty$, and by translation it may be assumed that one of these, say *ess inf* μ, is zero and that $\mu_1([0, d)) = \mu_1((-\infty, d)) = \rho$.

For $i = 1, ..., n$, let B_i be defined by

$$B_i := [(i-1)d, \infty)^k,$$

and for each $m, 1 \le m \le n$, let the partition $(A_i^m)_{i=1}^n$ be defined by

$$A_i^m = \begin{cases} B_i \backslash B_{i+1}, & i = 1, ..., m-1 \\ B_m, & i = m \\ \emptyset, & i = m+1, ..., n. \end{cases}$$

Then $\{\vec{a_1}, ..., \vec{a_n}\} \subset \mathcal{PR}(\vec{\mu}^k)$, where

$$\vec{a_m} = (\mu_1^k(A_1^m), ..., \mu_n^k(A_n^m)) = (1-(1-\rho)^k, ..., 1-(1-\rho)^k, 1, 0, ..., 0)$$

is the vector in $\mathrm{I\!R}^{\mathrm{n}}$ with 1 in the m-th coordinate and preceded by $m-1$ entries of $1-(1-\rho)^k$.

Let $\beta_m = (1-\rho)^{k(n-m)} / \sum_{j=0}^{n-1}(1-\rho)^{kj}, m = 1, ..., n$. Then by Proposition 2.2,

$$\vec{a} := \sum_{m=1}^{n} \beta_m \vec{a_m} \in \mathcal{PR}(\vec{\mu}^k),$$

and a straightforward calculation shows that each entry of $\vec{a}$ is $[\sum_{j=0}^{n-1}(1-\rho)^{kj}]^{-1}$.

To see that (6) is best possible for $\rho = 0$, let $\mu = \mu_M$ be the uniform distribution on $[-M, M]$. Then as $M \to \infty$, $\rho(\mu, d) \to 0$ and $C_{n,k}^*(\vec{\mu}) \to 1/n$. For $\rho = 1$, the bound in (6) is 1 and therefore trivially attained. That (6) is attained for all n, k and d and all $\rho \in (0, 1)$ is shown by the next example. □

Example 5.2 Let $\rho \in (0,1)$, $d > 0$, $k \in \mathrm{I\!N}$ and $n \geq 2$ be given and let α be the unique number such that $1 - e^{-\alpha d} = \rho$. Let $F(x) = 1 - e^{-\alpha x}$ for $x > 0$ and let $F_i(x) = F(x-(i-1)d)$, $i = 1, ..., n$. Then the corresponding density functions are

$$f_i(x) = e^{-\alpha(x-(i-1)d)} \quad \text{for} \quad x \geq (i-1)d$$

and zero otherwise for $i = 1, ..., n$. Clearly F has $\rho(F, d) = \lambda(F, d) = \rho$. Moreover, it is easily checked that, letting $q := 1 - \rho$,

$$f_1(x) \geq q^{i-1} f_i(x) \quad \text{for all} \quad x \in \mathrm{I\!R},\ \mathrm{i = 1, ..., n}. \tag{7}$$

Now let $(A_i)_{i=1}^n$ be any partition of $\mathrm{I\!R}^{\mathrm{k}}$. Then by (7),

$$q^{k(i-1)} \mu_i^k(A_i) \leq \mu_1^k(A_i), \quad i = 1, ..., n.$$

Summing over i gives

$$\sum_{i=1}^{n} q^{k(i-1)} \mu_i^k(A_i) \leq 1,$$

but this implies that

$$\min_{1 \leq i \leq n} \mu_i^k(A_i) \leq \left[\sum_{i=1}^{n} q^{k(i-1)}\right]^{-1}. \quad \Box$$

Analogous bounds in terms of Lévy concentration for general n and k are not known to the author. However, the following special case shows that things may change radically in the multiple observations case.

Theorem 5.3 *If μ is atomless, then*

$$C_{2,2}^*(\vec{\mu}) \geq \frac{1+\lambda}{2},$$

where $\lambda = \lambda(\mu, d)$, and this bound is attained for all $\lambda \in (0, 1]$.

Proof: Let γ and r be defined as in the proof of Lemma 3.6. By symmetry it can be assumed that $r \geq (1-\lambda)/2$. Considering the partitions (A, A^c) of $\mathbb{R}^2$ given by $A = \emptyset, A = (-\infty, \gamma+d] \times \mathbb{R}$, and $A = (-\infty, \gamma+d]^2$, respectively shows that $\mathcal{PR}(\vec{\mu}^2)$ contains the vectors $\vec{v}_1 = (0,1)$, $\vec{v}_2 = (r+\lambda, 1-r)$ and $\vec{v}_3 = ((r+\lambda)^2, 1-r^2)$. Now distinguish two cases.

Case 1, $(r+\lambda)^2 \leq 1-r^2$. Let

$$\alpha_2 = \frac{1-r^2-(r+\lambda)^2}{r(1-r)+(r+\lambda)(1-r-\lambda)} \quad \text{and} \quad \alpha_3 = \frac{r+\lambda-(1-r)}{r(1-r)+(r+\lambda)(1-r-\lambda)}.$$

Then clearly $\alpha_2+\alpha_3 = 1$, and by the assumptions on r, $\alpha_2 \geq 0$ and $\alpha_3 \geq 0$. Hence by Proposition 2.2 $\mathcal{PR}(\vec{\mu}^2)$ contains the vector $\vec{v} := \alpha_2\vec{v}_2 + \alpha_3\vec{v}_3$. An easy calculation gives $\vec{v} = (c,c)$, where

$$c = c(r) = \frac{(r+\lambda)(1-r)(1-\lambda)}{r(1-r)+(r+\lambda)(1-r-\lambda)} =: \frac{t(r)}{n(r)} \cdot (1-\lambda),$$

where $t(r)$ and $n(r)$ are defined in the obvious manner.

To show that this is at least $\frac{1+\lambda}{2}$, it is enough to show that $c(\frac{1-\lambda}{2}) = \frac{1+\lambda}{2}$ and that $c(r)$ is increasing in r for $r \geq \frac{1-\lambda}{2}$. The first is an easy substitution; the second follows since $n'(r) = 2t'(r) = 2(1-2r-\lambda)$ (where $'$ denotes derivative with respect to r) and hence $(nt' - tn')(r) = (1-2r-\lambda)(n(r)-2t(r)) \geq 0$, since both factors are non-positive. Thus $c'(r) \geq 0$ for all $r \geq \frac{1-\lambda}{2}$.

Case 2, $(r+\lambda)^2 \geq 1-r^2$. Let

$$\beta_1 = \frac{(r+\lambda)^2-(1-r^2)}{(r+\lambda)^2+r^2} \quad \text{and} \quad \beta_3 = \frac{1}{(r+\lambda)^2+r^2}.$$

Then $\beta_1 \geq 0, \beta_3 \geq 0$ and $\beta_1 + \beta_3 = 1$, and Proposition 2.2 gives $\vec{w} = \beta_1\vec{v}_1 + \beta_3\vec{v}_3 \in \mathcal{PR}(\vec{\mu}^2)$. Computing $\vec{w}$ yields $\vec{w} = (\bar{c}, \bar{c})$, where

$$\begin{aligned}
\bar{c} = \bar{c}(r) &= \left\{1 + \left(\frac{r}{r+\lambda}\right)^2\right\}^{-1} = \left\{1 + \left(1 - \frac{\lambda}{r+\lambda}\right)^2\right\}^{-1} \\
&\geq \{1+(1-\lambda)^2\}^{-1} = \frac{1+\lambda}{1+\lambda+(1-\lambda)(1-\lambda^2)} \\
&\geq \frac{1+\lambda}{2},
\end{aligned}$$

where the first inequality follows since $r \leq 1-\lambda$.

The following example shows that the bound $\frac{1+\lambda}{2}$ is attained for every $\lambda \in (0,1]$. □

Example 5.4 Let $\mathbb{H}^-$ be the left halfplane $\{(x,y) \in \mathbb{R}^2 : \text{x} \leq 0\}$ and $\mathbb{H}^+ := \mathbb{R}^2 \backslash \mathbb{H}^-$. Let μ be the distribution with density f given by $f(x) = \sum_{j \in \mathbf{Z}} \lambda(\frac{1-\lambda}{1+\lambda})^{|j|} 1_{[j,j+1)}(x)$. Then $\lambda(\mu, 1) = \lambda$ and for any partition (A, A^c) of $\mathbb{R}^2$,

$$\begin{aligned}
2\min\{\mu_1^2(A), \mu_2^2(A^c)\} &\leq \mu_1^2(A) + \mu_2^2(A^c) \\
&= \iint_A f_1(x)f_1(y)\ dx\ dy + \iint_{A^c} f_2(x)f_2(y)\ dx\ dy \\
&\leq \iint_{\mathbb{R}^2} \max\{f_1(x)f_1(y), f_2(x)f_2(y)\}\ dx\ dy \\
&= \iint_{\mathbb{H}^-} f_1(x)f_1(y)\ dx\ dy + \iint_{\mathbb{H}^+} f_2(x)f_2(y)\ dx\ dy \\
&= \frac{1+\lambda}{2} + \frac{1+\lambda}{2} = 1+\lambda,
\end{aligned}$$

where the second equality follows by the definition of f and the third equality is an easy computation.

Hence

$$\min\{\mu_1^2(A), \mu_2^2(A^c)\} \leq \frac{1+\lambda}{2} \quad \Box.$$

It should be noted here that the critical distributions for the case $n = k = 2$ are symmetric, whereas in the single-observation case they were skewed (cf. Example 5.2). In fact, for the distribution from Example 5.4 the advantage of having a second observation is completely absent, as follows by Example 5.4 and Theorem 2.6 of Hill and Tong [6].

Another consequence is, that the bound in (6) may fail in general when ρ is replaced by λ, since $(1+\lambda)/2 < \{1+(1-\lambda)\}^{-1}$ for all $\lambda < 1$.

For measures with atoms, there is a similar result as in the single-observation case.

Corollary 5.5 *Let μ be a general probability measure with tail d-concentration ρ. If $\mu(E) \leq \alpha$ for every atom E of μ, then*

$$C^*_{n,k}(\vec{\mu}) \geq [\sum_{j=0}^{n-1}(1-\rho)^{kj}]^{-1} - \frac{n-1}{n}\alpha^k.$$

Proof: Follows easily from Theorem 5.1, Theorem 4.4 and the fact that if all atoms of μ have mass at most α, then the atoms of μ^k have mass at most α^k. $\Box$

Remark 5.6 Let $c_{n,k}(\rho)$ denote the right hand side in (6), $c_{n,k}(\rho) = [\sum_{j=0}^{n-1}(1-\rho)^{kj}]^{-1}$. The bound $c_{n,k}$ has the following easily verified properties:

(i) $c_{n,k}(\rho) \downarrow$ as $n \to \infty$, and $\lim_{n\to\infty} c_{n,k}(\rho) = 1-(1-\rho)^k$.

(ii) $c_{n,k}(\rho) \uparrow$ as $k \to \infty$, and $\lim_{k\to\infty} c_{n,k}(\rho) = 1$.

(iii) If $k \geq \log\alpha/\log(1-\rho)$, then $c_{n,k}(\rho) \geq 1-\alpha$ for all $n \in \mathbb{N}$.

Example 5.7 Let F be the exponential distribution with mean 1, so $F(x) = 1-e^{-x}$ $(x > 0)$. Then $\rho(F,d) = 1-e^{-d}$, so if $d = \frac{1}{2}$ (for example), then (iii) above implies that 6 observations suffice for the minimax risk to be less than 5%, regardless of n.

Acknowledgement

I wish to express my gratitude to Professor Ted Hill for showing me the problem and for many fruitful discussions; to Georgia Tech for their hospitality during the winter of 1996; and to Barry Schouten for test-reading my drafts and for some useful suggestions.

References

[1] Allaart, P. C. (1997). *In preparation*

[2] Dvoretzky, A., Wald, A., and Wolfowitz, J. (1951). Relations among certain ranges of vector measures. *Pacific J. Math.* **1**, 59-74.

[3] Ferguson, T. (1967). *Mathematical statistics: a decision theoretic approach.* Academic Press, New York.

[4] Gouweleeuw, J. M. (1995). A characterization of measures with convex range. *Proc. London Math. Soc.* (3) **70**, 336-362.

[5] Hengartner, W. and Theodorescu, R. (1973). *Concentration functions.* Academic Press, New York.

[6] Hill, T. P. and Tong, Y. L. (1989). Optimal-partitioning inequalities in classification and multi-hypotheses testing. *Ann. Stat.* **17**, 1325-1334.

[7] Hill, T. P. (1993). Partitioning inequalities in probability and statistics. *Stochastic Inequalities, IMS Lecture Notes* **22**, 116-132.

[8] Legut, J. (1988). Inequalities for α-optimal partitioning of a measurable space. *Proc. Amer. Math. Soc.* **104**, 1249-1251.

[9] Lévy, P. (1937). *Théorie de l'addition des variables aléatoires.* Gauthier-Villars, Paris.

[10] Lyapounov, A. (1940). Sur les fonctions-vecteurs complètement additives. *Bull. Acad. Sci. URSS* **4**, 465-478.