Sankhyā : The Indian Journal of Statistics
2004, Volume 66, Part 2, pp 292-305
© 2004, Indian Statistical Institute

Variations on a Theme by Neyman and Pearson

V.S. Borkar
Tata Institute of Fundamental Research, Mumbai, INDIA
S.K. Mitter
Laboratory for Information and Decision Systems, Cambridge, USA
S.R. Venkatesh
Boston University, Boston, USA

Abstract

A symmetric version of the Neyman-Pearson test is developed for discriminating between sets of hypotheses and is extended to encompass a new formulation of the problem of parameter estimation based on finite data sets. Such problems can arise in distributed sensing and localization problems in sensor networks, where sensor data must be compressed to account for communication constraints. In this setting it is natural to focus on methods that balance coarse resolution of the estimates for achieving higher reliability.

AMS (2000) *subject classification.* 62F03, 62C20.
Keywords and phrases. Multiple hypothesis testing, parametric inference, minmax, convex optimization.

1 Introduction

In this paper we present a new approach to statistical modeling and estimation with finite data. This problem is motivated by the need to provide a framework for highly nonstationary situations where the complexity of the environment often exceeds the ability to collect meaningful data. For example, in communications this situation arises when the coherence time is significant relative to the delay spread. Historically, statistical methods address this issue by appealing to Occam's razor, which has lead to many approaches which in general result in choosing the estimates by optimizing a combination of model complexity and empirical error Barron et. al. (1998). Our approach here, in contrast, is to treat the parameter estimation problem as a 'continuum' hypothesis testing problem and consider the minimization of an appropriate 'worst case' risk, leading to a minmax problem. This in turn leads to a convex optimization problem resulting in a test that can be

viewed as a direct generalization of the celebrated Neyman-Pearson lemma Lehmann (1997). Another interesting feature of our treatment is that we choose among sets of hypotheses rather than individual hypotheses, the idea being to make inference with prescribed accuracy in the parameter space (in contrast with the classical 'confidence intervals' which specify accuracy with reference to the underlying probability space).

The paper is organized as follows: The next section gives a novel treatment of composite hypothesis testing using convex programming tools. Classical hypothesis testing has a special role for the null hypothesis. A symmetrized version that avoids this and treats all hypothesis sets on an equal footing is developed in section 3. This results in a 'minmax' problem reminiscent of the minmax formulation in Bayesian statistics. Section 4 contains our main result, viz., an extension of the above to the problem of parameter estimation based on finite data sets, casting it as a 'continuum hypothesis testing' problem. While these results are not surprising in view of their similarity to Bayesian minmax, they are arrived at from a somewhat different perspective.

2 Revisiting the Neyman-Pearson Lemma

Consider $M+1$ disjoint compact subsets $\Theta_0, \cdots, \Theta_M$ of a Polish space S, with $\Theta \stackrel{\Delta}{=} \bigcup_i \Theta_i$. Θ will be our parameter space. We associate with it a family of probability densities $f_\theta(\cdot), \theta \in \Theta$, on R^d (say) for some $d \geq 1$. We shall assume that $f_\theta(x)$ is jointly continuous in θ, x. The problem we address is: Given an observation Y generated according to one of these densities, say $f_\theta(\cdot)$ for some $\theta \in \Theta_i$, come up with the best guess $\gamma(Y) \in \{0, 1, 2, \cdots, M\}$ of the index i. To define what we mean by 'best' here, we introduce the following notation: Let $P_\theta(\cdot)$ denote the probability under $\theta \in \Theta$ and let $\alpha_1, \cdots, \alpha_M$ be given positive scalars. Following the classical hypothesis testing framework, we seek to find $\gamma(Y)$ so as to

$$\begin{array}{ll} \text{Minimize} & \sup\limits_{\theta \in \Theta_0} P_\theta(\gamma(Y) \neq 0) \\ \text{subject to} & \sup\limits_{\theta \in \Theta_i} P_\theta(\gamma(Y) \neq i) \leq \alpha_i, \quad 1 \leq i \leq M. \end{array}$$

Note that if Θ_i's were singletons, this would reduce to the classical multiple hypothesis testing problem Lehmann (1997). In particular, Θ_0 plays the special role of the 'null hypothesis'. Let $\mathcal{P}(\cdots)$ denote the Polish space of probability measures on the Polish space '$\cdots$' with Prohorov topology (Borkar (1995), Chapter 2). Let $Q \stackrel{\Delta}{=} \{\lambda = [\lambda_0, \cdots, \lambda_M] : \lambda_i \geq 0 \ \forall i, \lambda_0 = 1\}$

and $\Gamma \triangleq Q \times \Pi_{i=0}^{M} \mathcal{P}(\Theta_i)$. Our main result is:

THEOREM 2.1 *There exist $\lambda^* \in Q, \pi_i^* \in \mathcal{P}(\Theta_i), 0 \leq i \leq M$, such that the optimal $\gamma(Y)$ is given by: $\gamma(Y) = i$ if*

$$\begin{aligned} \lambda_i^* \int \pi_i^*(d\theta) f_\theta(Y) &> \max_{j<i} \lambda_j^* \int \pi_j^*(d\theta) f_\theta(Y), \\ \lambda_i^* \int \pi_i^*(d\theta) f_\theta(Y) &\geq \max_{j>i} \lambda_j^* \int \pi_j^*(d\theta) f_\theta(Y). \end{aligned}$$

In turn, $\lambda^*, \{\pi^*\}$ are the solutions of the convex programming problem

$$\min_{(\lambda, \{\pi_i\}) \in \Gamma} \left(\int \max_{0 \leq i \leq M} \left(\lambda_i \int \pi_i(d\theta) f_\theta(x) \right) dx - \sum_{i=1}^{M} \lambda_i (1 - \alpha_i) \right).$$

Before proving this theorem, we introduce some relaxations of the above optimization problem. The first is:

$$\begin{aligned} \text{Minimize} \quad & \sup_{\pi \in \mathcal{P}(\Theta_0)} \int \pi(d\theta) P_\theta(\gamma(Y) \neq 0) \\ \text{subject to} \quad & \sup_{\pi \in \mathcal{P}(\Theta_i)} \int \pi(d\theta) P_\theta(\gamma(Y) \neq i) \leq \alpha_i, \quad 1 \leq i \leq M. \end{aligned}$$

This is clearly equivalent to the original problem. Define $\varphi_i(x) = I\{\gamma(x) = i\}$ for $0 \leq i \leq M$. Then $\varphi(\cdot) = [\varphi_0(\cdot), \cdots, \varphi_M(\cdot)] \in H_0$ where

$$\begin{aligned} H_0 \triangleq \; & \{\eta(\cdot) = [\eta_0(\cdot), \cdots, \eta_M(\cdot)] : \eta_i \textit{ is measurable } R^d \to \{0,1\} \\ & \textit{for } 0 \leq i \leq M, \; \sum_i \eta_i(\cdot) \equiv 1\}. \end{aligned}$$

We view H_0 as a subset of

$$\begin{aligned} H \triangleq \; & \{[\eta_0(\cdot), \cdots, \eta_M(\cdot)] : \eta_i \textit{ is measurable } R^d \to [0,1] \\ & \textit{for } 0 \leq i \leq M, \; \sum_i \eta_i(\cdot) \equiv 1\}, \end{aligned}$$

viewed as a convex compact subset of $L_\infty(R^d)^{M+1}$ endowed with the *weak** topology. The above optimization problem can now be restated as:

$$\begin{aligned} \text{Minimize} \quad & \sup_{\pi \in \mathcal{P}(\Theta_0)} \int \int \pi(d\theta) f_\theta(x)(1 - \varphi_0(x)) dx \\ \text{subject to} \quad & \sup_{\pi \in \mathcal{P}(\Theta_i)} \int \int \pi(d\theta) f_\theta(x)(1 - \varphi_i(x)) dx \leq \alpha_i, \quad 1 \leq i \leq M, \end{aligned}$$

where the minimization is over H_0.

PROOF OF THEOREM 1. Letting $\lambda \in Q$ denote the Lagrange multipliers associated with the above convex programming problem, we consider the associated saddle point problem

$$\min_{\varphi \in H_0} \max_{(\lambda, \{\pi_i\}) \in \Gamma} \left(\int \int \pi_0(d\theta) f_\theta(x)(1 - \varphi_0(x)) dx \right.$$
$$\left. + \sum_{i=1}^{M} \lambda_i \left(\int \int \pi_i(d\theta) f_\theta(x)(1 - \varphi_i(x)) dx - \alpha_i \right) \right).$$

We consider the relaxation

$$\min_{\varphi \in H} \max_{(\lambda, \{\pi_i\}) \in \Gamma} \left(\int \int \pi_0(d\theta) f_\theta(x)(1 - \varphi_0(x)) dx \right.$$
$$\left. + \sum_{i=1}^{M} \lambda_i \left(\int \int \pi_i(d\theta) f_\theta(x)(1 - \varphi_i(x)) dx - \alpha_i \right) \right). \tag{1}$$

By the saddle point theorem for Lagrange multipliers (Luenberger (1969), p. 219), (1) equals

$$\max_{(\lambda, \{\pi_i\}) \in \Gamma} \min_{\varphi \in H} \left(\int \int \pi_0(d\theta) f_\theta(x)(1 - \varphi_0(x)) dx \right.$$
$$\left. + \sum_{i=1}^{M} \lambda_i \left(\int \int \pi_i(d\theta) f_\theta(x)(1 - \varphi_i(x)) dx - \alpha_i \right) \right).$$

This in turn equals

$$- \min_{(\lambda, \{\pi_i\}) \in \Gamma} \max_{\varphi \in H} \left(\int \int \pi_0(d\theta) f_\theta(x) \varphi_0(x) \, dx \right.$$
$$\left. + \sum_{i=1}^{M} \lambda_i \int \int \pi_i(d\theta) f_\theta(x) \varphi_i(x) dx - 1 - \sum_{i=1}^{M} \lambda_i (1 - \alpha_i) \right). \tag{2}$$

The inner maximum is clearly attained at φ^* given by

$$\begin{aligned} \varphi_i^*(x) \quad = \quad & I \left\{ \lambda_i \int \pi_i(d\theta) f_\theta(x) > \max_{j<i} \lambda_j \int \pi_j(d\theta) f_\theta(x), \right. \\ & \left. \lambda_i \int \pi_i(d\theta) f_\theta(x) \geq \max_{j>i} \lambda_j \int \pi_j(d\theta) f_\theta(x) \right\} \end{aligned}$$

Thus (2) equals

$$-\min_{(\lambda,\{\pi_i\})\in\Gamma}\left(\int\max_{0\le i\le M}\left(\lambda_i\int\pi_i(d\theta)f_\theta(x)\right)dx-1-\sum_{i=1}^{M}\lambda_i(1-\alpha_i)\right).$$

The expression being minimized is easily see to be continuous. Since the domain is compact, a minimum is attained at some $(\lambda^*,\{\pi_i^*\})$. This completes the proof. □

The saddle point (λ^*,φ^*) satisfies $\varphi^*\in H_0$, so that the relaxation is equivalent to the original problem. For $M=1$, the theorem is seen to reduce to the 'ratio test': Pick the null hypothesis if

$$\frac{\int\pi_0(d\theta)f_\theta(Y)}{\int\pi_1(d\theta)f_\theta(Y)}\ge\lambda_1,$$

and not otherwise. If in addition $\Theta_i, i=0,1$, are singletons, this reduces to the familiar Neyman-Pearson test.

3 A Symmetrized Problem

The classical hypothesis testing formulation accords a special status to the null hypothesis. We now consider a variation where all hypotheses are treated the same. Thus we consider the minmax problem

$$\min_{0\le i\le M}\max_{\theta\in\Theta_i}P_\theta(\gamma(Y)\ne i).$$

Redefine Q by $Q\stackrel{\Delta}{=}\{\lambda=[\lambda_0,\cdots,\lambda_M]:\lambda_i\ge 0\ \forall i,\sum_{i=0}^{M}\lambda_i=1\}$. Define Γ as before. By familiar arguments, we consider the relaxation

$$\begin{aligned}
&\min_{\varphi\in H}\max_{(\lambda,\{\pi_i\})\in\Gamma}\int\sum_i\lambda_i\int\pi_i(d\theta)f_\theta(x)(1-\varphi_i(x))dx\\
&=\max_{(\lambda,\{\pi_i\})\in\Gamma}\min_{\varphi\in H}\int\sum_i\lambda_i\int\pi_i(d\theta)f_\theta(x)(1-\varphi_i(x))dx\\
&=\max_{(\lambda,\{\pi_i\})\in\Gamma}\min_{\varphi\in H}\left(1-\int\sum_i\lambda_i\int\pi_i(d\theta)f_\theta(x)\varphi_i(x)dx\right)\\
&=1-\min_{(\lambda,\{\pi_i\})\in\Gamma}\max_{\varphi\in H}\int\sum_i\lambda_i\int\pi_i(d\theta)f_\theta(x)\varphi_i(x)dx\\
&=1-\min_{(\lambda,\{\pi_i\})\in\Gamma}\int\max_i\left(\lambda_i\int\pi_i(d\theta)f_\theta(x)\right)dx.
\end{aligned}$$

Thus exactly as in Theorem 1 above, we are lead to:

THEOREM 3.1 *For the symmetrized problem above, there exist* $\lambda^* \in Q, \pi_i^* \in \mathcal{P}(\Theta_i), 0 \leq i \leq M$, *such that the optimal* $\gamma(Y)$ *is given by:* $\gamma(Y) = i$ *if*

$$\begin{aligned} \lambda_i^* \int \pi_i^*(d\theta) f_\theta(Y) &> \max_{j<i} \lambda_j^* \int \pi_j^*(d\theta) f_\theta(Y), \\ \lambda_i^* \int \pi_i^*(d\theta) f_\theta(Y) &\geq \max_{j>i} \lambda_j^* \int \pi_j^*(d\theta) f_\theta(Y). \end{aligned}$$

In turn, $\lambda^*, \{\pi^*\}$ *are the solutions of the convex programming problem*

$$\min_{(\lambda, \{\pi_i\}) \in \Gamma} \int \max_{0 \leq i \leq M} \left(\lambda_i \int \pi_i(d\theta) f_\theta(x) \right) dx.$$

EXAMPLE 1 Consider four hypotheses, H_k, $k = 0, 1, 2, 3$. Suppose we are given the following family of distributions:

$$f_{H_k}(y) = \mathcal{N}(k, 1), \ k \leq 2, \ \text{and} \ f_{H_3}(y) = \mathcal{N}(1.8, .01).$$

Consider hypothesis sets $\Theta_1 = \{H_0, H_1\}$, $\Theta_2 = \{H_2, H_3\}$, that must be discriminated based on observation $y \in \mathbb{R}$. It follows by direct numerical verification that

$$\{y \mid f_{H_3}(y) \geq f_{H_k}(y)\} \subset \{y \mid f_{H_2}(y) \geq f_{H_k}(y)\}, \ k \in \{0, 1\}.$$

Therefore the hypothesis H_2 is not a factor and the optimal solution is given by comparing H_3 against the set Θ_1. The optimal solution turns out to be:

$$\gamma(y) = \begin{cases} \Theta_1 & \text{if } y \leq \frac{3}{2} \\ \Theta_2 & \text{if } y > \frac{3}{2} \end{cases}$$

EXAMPLE 2 Consider four hypotheses as in the previous example, but now with distributions defined by:

$$f_{H_k}(y) = \mathcal{N}(k, 1).$$

Consider overlapping hypothesis sets $\Theta_1 = \{H_0, H_1, H_2\}$, $\Theta_2 = \{H_2, H_3, H_4\}$, that must be discriminated based on observation $y \in \mathbb{R}$. (This does not exactly fit the formulation above because of lack of disjointness, but is included here in order to motivate the developments of the next section.) In this case,

irrespective of how the observations are partitioned into the two sets, the worst-case hypothesis is the one that is common to both the sets, which is H_2 in the example. Therefore, it follows that the optimal solution in this situation is given by:

$$\lambda_j = 1, \forall j, \ \pi_0(H_2) = 1, \ \pi_1(H_2) = 1.$$

The worst-case probability of error is 1/2 for this situation. Therefore overlapping hypothesis structures do not provide improvements in error probability for this formulation. In the next section we formulate an estimation problem where such structures lead to significant improvements in error probability.

4 Parameter Estimation with Finite Data

We now consider the problem of estimating a parameter $\theta \in \Theta \stackrel{\Delta}{=}$ a compact convex subset of some $R^m, m \geq 1$, with an associated family of densities $f_\theta(\cdot)$ as above. The difference with the foregoing will be that instead of choosing among a finite family $\Theta_i, 0 \leq i \leq M$, of subsets of Θ as above, we now seek to find the 'best' Borel subset of Θ of diameter not exceeding a prescribed $\epsilon \geq 0$ so as to minimize the worst case risk. Specifically, let $\psi(Y)$ denote one such set, chosen as a function of the observation Y. Then we seek to make the choice thereof so as to minimize

$$\sup_{\theta \in \Theta} P_\theta(\theta \notin \psi(Y)). \tag{3}$$

This formulation is from Venkatesh et al. (2002).

Since a larger set can only reduce the risk and the largest set of diameter not exceeding ϵ is the closed ball of radius ϵ, it suffices to consider only $\psi(Y)$ of the form $\{x \in \Theta : ||x - \eta(Y)|| \leq \epsilon\}$. For $\epsilon > 0$, this problem may be viewed as a dual of the traditional interval estimation problem, insofar as we prescribe the confidence level ϵ in terms of the natural metric of the parameter space rather than in terms of probabilities over its inverse image in the underlying sample space. For $\epsilon = 0$, we recover the point estimation problem.

We strengthen our continuity assumption on $f_\theta, \theta \in \Theta$, to:

(†) The maps $\theta \to f_\theta(x), x \in R^d$, are bounded and continuously differentiable with

$$\sup_x ||\nabla^\theta f_\theta(x)|| \leq q(x)$$

for some $q(\cdot) \geq 0$ with $\int q(x)dx < \infty$. (As will be clear later, this assumption could be relaxed further.)

Let $\mathcal{D}(\Theta)$ denote the set of continuous probability densities over Θ. Then the set $\mathcal{P}_0(\Theta) \triangleq \{\phi(\theta)d\theta : \phi(\cdot) \in \mathcal{D}(\Theta)\}$ is dense in $\mathcal{P}(\Theta)$. Let B denote the set of closed balls of diameter ϵ with centers in Θ, endowed with the metric topology of the Hausdorff metric. Let $\mathcal{M}$ denote the set of measurable maps $\psi : R^d \to B$ and $\mathcal{C}$ the set of measurable maps $\eta : R^d \to \Theta$. We may identify an element ψ of the former family with an element η of the latter such that $\eta(y)$ is the centre of $\psi(y)$. Thus we shall use these interchangeably. Let $\mathcal{H}$ denote the collection of random variables $\{\eta(Y) : \eta \in \mathcal{C}\}$. We metrize $\mathcal{H}$ by the metric

$$\rho(\eta_1(Y), \eta_2(Y)) \triangleq \sup_{\theta \in \Theta} E_\theta[||\eta_1(Y) - \eta_2(Y)|| \wedge 1],$$

where $E_\theta[\,\cdot\,]$ denotes the expectation under P_θ for $\theta \in \Theta$.

LEMMA 1. $(\mathcal{H}, \rho)$ is a complete separable metric space.

PROOF. ρ is clearly a metric. Recalling that Θ is a bounded set, observe that $\mathcal{H}$ is a complete metric space under each of the metrics

$$\rho_\theta(\eta_1(Y), \eta_2(Y)) \triangleq E_\theta[||\eta_1(Y) - \eta_2(Y)||]$$

by Theorem 1.5.1, p. 13, of Borkar (1995), with the metric convergence corresponding to the convergence in probability under P_θ. A Cauchy sequence $\{\eta_n(Y)\}$ under ρ will be Cauchy with respect to each ρ_θ and hence it converges w.r.t. each. Since convergence in probability implies convergence a.s. along a subsequence, it follows that the respective limits must agree a.e. on the intersection of the supports of $f_\theta, f_{\theta'}$ for any two distinct θ, θ'. Thus we can consistently define a random variable $\hat{\eta}_\infty$ such that $\eta_n(Y) \to \hat{\eta}_\infty$ under each ρ_θ, θ rational. To claim that it also does so under ρ, we need to prove that this convergence is uniform in θ. This follows from assumption (†) above (which implies pointwise boundedness and equicontinuity of $\rho_\theta(\eta_n(Y), \eta_\infty(Y))$ in θ) and the Arzela-Ascoli theorem. Since $\eta_n(Y)$ is $\sigma(Y)-$measurable for each n, so will be $\hat{\eta}_\infty$ and hence $\hat{\eta}_\infty = \eta_\infty(Y)$ a.s. for a suitable $\eta_\infty \in \mathcal{C}$ by Theorem 1.1.4, p. 5, of Borkar (1995). Separability follows from the density of η of the type

$$\eta(y) = \sum_{m=1}^{n} y_m I\{y \in A_m\},$$

where $n \geq 1$, the y_i's have rational components and A_m's are disjoint axis-parallel rectangles in R^d with rational corners. □

The infimum of (3) equals

$$\begin{aligned}
&\inf_{\psi \in \mathcal{M}} \max_{\theta \in \Theta} P_\theta(\theta \notin \psi(Y)) \\
&= 1 - \sup_{\psi \in \mathcal{M}} \min_{\theta \in \Theta} P_\theta(\theta \in \psi(Y)) \\
&= 1 - \sup_{\psi \in \mathcal{M}} \min_{\pi \in \mathcal{P}(\Theta)} \int \pi(d\theta) P_\theta(\theta \in \psi(Y)) \\
&= 1 - \sup_{\eta(Y) \in \mathcal{H}} \min_{\pi \in \mathcal{P}(\Theta)} \int \pi(d\theta) P_\theta(||\theta - \eta(Y)|| \leq \epsilon) \\
&= 1 - \sup_{\eta(Y) \in \mathcal{H}} \min_{\pi \in \mathcal{P}(\Theta)} \int \int \pi(d\theta) f_\theta(x) I\{||\theta - \eta(x)|| \leq \epsilon\} dx.
\end{aligned}$$

Consider

$$F(\pi, \eta(Y)) \stackrel{\Delta}{=} \int \int \pi(d\theta) f_\theta(x) I\{||\theta - \eta(x)|| \leq \epsilon\} dx. \tag{4}$$

This is linear and continuous in π on $\mathcal{P}(\Theta)$, where the continuity follows from that of $(\theta, x) \to f_\theta(x)$. (The latter implies the continuity of $\theta \to P_\theta$ in total variation norm by Scheffe's theorem. Thus $\theta \to P_\theta(||\theta - \eta(Y)|| \leq \epsilon)$ is continuous.) $F(\pi, \eta(Y))$ is upper semicontinuous in $\eta(Y)$ on $\mathcal{H}$, because $B(z) \stackrel{\Delta}{=} \{z : ||x - z|| \leq \epsilon\}$ is a closed set. Therefore, taking the standard relaxation of the above minmax problem, we consider

$$\begin{aligned}
&\sup_{\Phi \in \mathcal{P}(\mathcal{H})} \min_{\pi \in \mathcal{P}(\Theta)} \int \Phi(d\eta) F(\pi, \eta) \\
&= \inf_{\pi \in \mathcal{P}(\Theta)} \sup_{\Phi \in \mathcal{P}(\mathcal{H})} \int \Phi(d\eta) F(\pi, \eta) \\
&= \inf_{\lambda \in \mathcal{D}(\Theta)} \sup_{\eta \in \mathcal{C}} \int \int \lambda(\theta) f_\theta(x) I\{||\theta - \eta(x)|| \leq \epsilon\} dx d\theta \\
&= \inf_{\lambda} \left(\int \sup_z \left[\int_{B(z)} \lambda(\theta) f_\theta(x) d\theta \right] dx \right),
\end{aligned}$$

where the first equality follows from the minmax theorem (Theorem 3 of Fan (1953)), the second follows from the density of $\mathcal{P}_0(\Theta)$ in $\mathcal{P}(\Theta)$, and the final equality follows by taking the optimum choice

$$\eta(Y) \in \arg\max_z \int_{B(z)} \lambda(\theta) f_\theta(Y) d\theta.$$

Thus we have:

THEOREM 4.1 *If the convex minimization problem on the r.h.s. above has a solution (i.e., the infimum is a minimum attained at) $\lambda^*(\cdot)$, then the closed $\epsilon-$ball centered at $\hat{\theta} \in \arg\max_z \int_{B(z)} \lambda^*(\theta) f_\theta(Y) d\theta$ is the optimum choice for the relaxed problem.*

Unfortunately the existence of a minimizer λ^* is not guaranteed. Nevertheless, the foregoing suggests an approximation procedure whereby we may replace the minimization over $\mathcal{D}(\Theta)$ by minimization over a sufficiently rich compact (in $C(\Theta)$) subset thereof. The minimizing λ^* will then be near-optimal rather than optimal. For example, a computationally appealing choice could be a finitely parametrized family.

Note also that we may replace $\mathcal{P}_0(\Theta)$ by any other convenient dense subset of $\mathcal{P}(\Theta)$, e.g., by $\mathcal{P}_1(\Theta) \stackrel{\Delta}{=}$ the set of finitely supported probability measures on Θ. Minimization over $\pi \in \mathcal{P}_1(\Theta)$ may then be approximated by minimization over the set of finitely supported probability measures on Θ supported on at most N points for some large $N \geq 1$. Suppose $\pi^* \in \mathcal{P}_1(\Theta)$ is a minimizer. A test analogous to the above ensues, leading to the choice

$$\eta(y) \in \arg \max_{z \in \ support(\pi^*)} \sum_{\theta \in B(z) \cap support(\pi^*)} \pi^*(\{\theta\}) f_\theta(Y).$$

Of course, this applies to the case when Θ itself is a finite set (see the example below).

If the variation of $f_\theta(x)$ in θ over an $\epsilon-$ball is small, the optimal decision in Theorem 4.1 may be approximated by

$$\eta(Y) \in \arg \max_\theta \lambda(\theta) f_\theta(Y).$$

The π^* above and $\sum_i \lambda_i^* \pi_i^*$ in the preceding section may be viewed as 'worst case priors', thus making contact with Bayesian hypothesis testing Lehmann (1997). Finally, observe that the computational aspects hinge on a convex programming problem for which many effective algorithms are available Bertsekas (1999).

EXAMPLE 3 This example illustrates the salient difference between maximum-likelihood estimates and the proposed formulation. Consider a discrete parameter set,

$$\Theta = \{1, 2, \ldots, 6\}$$

Let $f_\theta(x)$, $\theta \in \Theta$, $x \in \mathbb{R}$ be a family of gaussian distributions over the reals for each parameter, θ, i.e.,

$$f_\theta(x) = \mathcal{N}(\theta, \sigma)$$

We are interested in picking estimates $\hat{\theta}(x)$ so that

$$\max_\theta \text{Prob}_\theta\{\theta \notin \hat{\theta}(x) \pm 2\}$$

is minimized. Nevertheless, the finite parameter set here ensures that the expression in (4) is linear and continuous in π. It can also be verified that $F(\cdot,\cdot)$ in (4) is upper semi-continuous in $\eta(Y) \in \mathcal{H}$. The statement of the Theorem 4.1 then follows. Furthermore, the probability simplex $\mathcal{P}(\Theta)$ being compact guarantees the existence of the minimum.

We now discuss the example in more detail. The maximum likelihood estimate is given by:

$$\theta_{ML}(x) = \text{argmax}_\theta f_\theta(x)$$

From this it follows that each parameter in the set Θ is picked based on the following rule:

$$\theta_{ML}(x) = \begin{cases} 1 & \text{if, } x \in (-\infty,\ 1 + \frac{1}{2\sigma}] \\ 2 & \text{if, } x \in (1 + \frac{1}{2\sigma}, 2 + \frac{1}{2\sigma}] \\ \vdots & \qquad \vdots \\ 6 & \text{if, } x \in (5 + \frac{1}{2\sigma}, \infty) \end{cases}$$

The worst-case probability of error for the ML estimate is when the parameter 3 is picked while the actual value is 6. The probability of error in this case is given by:

$$P_e = \text{erfc}(3/2\sigma)$$

For our formulation, the task is reduced to mapping each value of $x \in \mathbb{R}$ to one of two possible subsets: $\Theta_1 = 3 \pm 2$; $\Theta_2 = 4 \pm 2$. Now, consider the set-valued estimate, $\Gamma(x) \in \{\Theta_1,\ \Theta_2\}$. The optimal solution for the problem:

$$\min_{\hat{\theta}(x)} \max_\theta \text{Prob}_\theta\{\theta \notin \hat{\theta}(x) \pm 2\}$$

is given by:

$$\Gamma(x) = \begin{cases} \Theta_1 & \text{if } x \leq \frac{7}{2\sigma} \\ \Theta_2 & \text{if } x > \frac{7}{2\sigma} \end{cases}$$

This follows from the fact that — on account of significant overlap between the two sets — the parameters ranging from $2 \leq \theta \leq 5$ will always be identified (with probability one) within a diameter 2 irrespective of the decision strategy. Therefore an error occurs only when either the parameter 1 or 6 is the correct value. The probability of error is now given by:

$$P_e = \mathrm{erfc}(5/2\sigma)$$

which is considerably smaller than that obtained for ML. The next step is to check whether the relaxation of the problem also yields the same risk. Notice that the statement of the Theorem 4.1 requires us to pick a distribution π over θ that solves the relaxed problem.

$$\min_{\pi} \max_{\hat{\theta}} \sum_k \pi_k \int f_k(x) I\{k \notin \hat{\theta}(x) \pm 2\} dx.$$

Upon closer observation the following values for the distribution π is optimal:

$$\pi_1 = 1/2; \ \pi_6 = 1/2.$$

This will ensure that the estimator $\theta(x)$ will pick the midpoint between 1 and 6 as a threshold. In turn, this threshold serves as the decision strategy to choose among the two parameter sets. This is the same strategy obtained for the primal problem.

In our next step we illustrate how the resolution affects the error probability. For this purpose, consider the same problem except that we want to find estimators that minimize the error probability for a diameter equal to one, i.e.,

$$\max_{\theta} \mathrm{Prob}_{\theta}\{\theta \notin \hat{\theta}(x) \pm 1\}.$$

The ML estimator is unchanged and the probability of error increases as:

$$P_e = \mathrm{erfc}(1/2\sigma).$$

In the proposed scheme, the task is reduced to mapping the observations to one of four possible choices:

$$\{1, 2, 3\}, \{2, 3, 4\}, \{3, 4, 5\}, \{4, 5, 6\}.$$

It follows based on arguments presented earlier that the largest error grows as:

$$P_e = \mathrm{erfc}(1/\sigma).$$

The example serves to show how the choice of a resolution has a significant impact on the error probability. Figure 1 illustrates these aspects as a function of increasing variance.

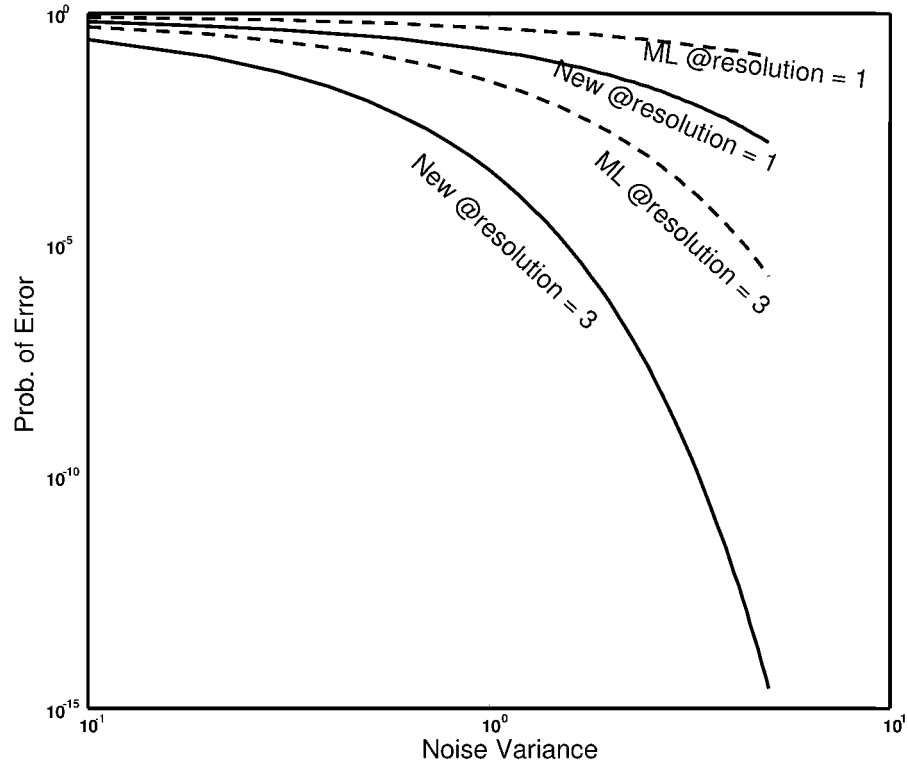

FIGURE 1. ILLUSTRATION OF THE TRADEOFF BETWEEN RESOLUTION AND ACCURACY FOR DIFFERENT NOISE VARIANCES FOR ML AND 'NEW' SCHEMES

5 Conclusions

We have considered the problem of estimating parameters up to a prescribed accuracy in the parameter space based on a finite amount of data. By casting it as a 'continuum hypothesis testing' problem, we are lead to a minmax problem reminiscent of the classical Bayesian minmax. This formulation has the advantage of a clear notion of optimality in finite data set-up without appeal to any asymptotics. Computationally, it offers the possibility of using standard nonlinear optimization tools such as primal-dual methods for the purpose. The other notable feature has been our a priori specification of accuracy in the parameter space with reference to which the inference is done. This framework allows trading resolution for higher accuracy in a natural way.

Acknowledgements. Support for this work was provided by the Army Research Office under the MURI Grant: Data Fusion in Large Arrays of Mi-

crosensors DAAD19-00-1-0466 and the Department of Defense Muri Grant: Complex Adaptive Networks for Cooperative Control Subaward #03-132. The third author was also supported in part by theONR Young Investigator Award N00014-02-100362.

References

BERTSEKAS D.P. (1999). *Nonlinear Programming*, second edition, Athena Scientific, Belmont, Mass.

BARRON, A., RISSANEN, J. and YU, B. (1998). The minimum description length principle in modeling and coding, *IEEE Trans. Information Theory*, **44**, 2743-2760.

BORKAR, V.S. (1995). *Probability Theory: An Advanced Course*, Springer Verlag, New York.

FAN, K. (1953). Minmax theorems, *Proc. Nat. Academy of Sciences* **39**, 42-47.

LEHMANN, E.L. (1997). *Testing Statistical Hypotheses*, second edition, Springer Verlag, New York.

LUENBERGER, D.G. (1969). *Optimization by Vector Space Methods*, Wiley, New York.

VENKATESH, S.R. and MITTER, S.K. (2002). Statistical estimation and modeling with finite data, *International Symp. on Information Theory*, Lausanne, June-July 2002.

V.S. BORKAR
SCHOOL OF TECHNOLOGY
AND COMPUTER SCIENCE
TATA INSTITUTE OF FUNDAMENTAL RESEARCH
HOMI BHABHA ROAD
MUMBAI 400005, INDIA
E-mail: borkar@tifr.res.in

S.K. MITTER
LABORATORY FOR INFORMATION
AND DECISION SYSTEMS
M.I.T., 77, MASSACHUSSETTS AVENUE
CAMBRIDGE, MA 02139, USA
E-mail: mitter@mit.edu

S.R. VENKATESH
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
BOSTON UNIVERSITY, ST. MARY'S STREET
BOSTON, MA 02215, USA
E-mail: srv@bu.edu

Paper received: January 2003; revised September 2003.