Variations of a Neyman and Pearson theme∗

V. S. Borkar † S. K. Mitter ‡ S. R. Venkatesh §

September 9, 2003

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.

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

∗Work supported in part by ARO under the MURI Grant DAAD19-00-1-0466, under the Center for Imaging Science Subcontract DAAD19-99-1-0012 (Johns Hopkins University), the MURI grant subcontract no. 654-21256 (Brown Uni.) and ONR Young Investigator Grant No. N00014-02-1-0362
†The author is with the School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, INDIA.
‡Laboratory for Information and Decision Systems, M.I.T., 77, Mass. Ave., Cambridge, MA 02139, USA.
§Dept. of Electrical and Computer Engineering, Boston University, 8, St. Mary’s St., Boston, MA 02215, USA.
approaches which in general result in choosing the estimates by optimizing a combination of model complexity and empirical error [2]. 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 [5]. 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 $\Theta \triangleq \bigcup_i \Theta_i$. $\Theta$ will be our parameter space. We associate with it a family of probability densities $f_\theta(\cdot), \theta \in \Theta$, on $\mathbb{R}^d$ (say) for some $d \geq 1$. We shall assume that $f_\theta(x)$ is jointly continuous in $\theta, x$. 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

Minimize $\sup_{\theta \in \Theta_0} P_\theta(\gamma(Y) \neq 0)$
subject to
\[ \sup_{\theta \in \Theta_i} P_{\theta}(\gamma(Y) \neq i) \leq \alpha_i, \]
\[ 1 \leq i \leq M. \]

Note that if \( \Theta_i \)'s were singletons, this would reduce to the classical multiple hypothesis testing problem [5]. In particular, \( \Theta_0 \) 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 ([3], Chapter 2). Let \( Q \triangleq \{ \lambda = [\lambda_0, \cdots, \lambda_M] : \lambda_i \geq 0 \ \forall i, \lambda_0 = 1 \} \) and \( \Gamma \triangleq Q \times \Pi_{i=0}^M \mathcal{P}(\Theta_i) \). Our main result is:

**Theorem 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:
\[ \gamma(Y) = i \]
\[ \lambda^*_i \int \pi^*_i(d\theta)f_\theta(Y) > \max_{j<i} \lambda^*_j \int \pi^*_j(d\theta)f_\theta(Y), \]
\[ \lambda^*_i \int \pi^*_i(d\theta)f_\theta(Y) \geq \max_{j>i} \lambda^*_j \int \pi^*_j(d\theta)f_\theta(Y). \]

In turn, \( \lambda^*, \{ \pi^* \} \) are the solutions of the convex programming problem
\[ \min_{(\lambda, \{ \pi_i \}) \in \Gamma} (\int \max_{0 \leq i \leq M} \lambda_i \int \pi_i(d\theta)f_\theta(x))dx - \sum_{i=1}^M \lambda_i(1 - \alpha_i)). \]

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

\[ \text{Minimize } \sup_{\pi \in \mathcal{P}(\Theta_0)} \int \pi(d\theta)P_{\theta}(\gamma(Y) \neq 0) \]
subject to
\[ \sup_{\pi \in \mathcal{P}(\Theta_i)} \int \pi(d\theta)P_{\theta}(\gamma(Y) \neq i) \leq \alpha_i, \]
\[ 1 \leq i \leq M. \]

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
\[ H_0 \triangleq \{ \eta(\cdot) = [\eta_0(\cdot), \cdots, \eta_M(\cdot)] : \eta_i \text{ is measurable } R^d \to \{0,1\} \text{ for } 0 \leq i \leq M, \sum_i \eta_i(\cdot) \equiv 1 \}. \]
We view $H_0$ as a subset of
$$ H \triangleq \{[\eta_0(\cdot), \cdots, \eta_M(\cdot)] : \eta_i \text{ is measurable } \mathbb{R}^d \to [0,1] \text{ for } 0 \leq i \leq M, \sum_i \eta_i(\cdot) \equiv 1 \} $$
viewed as a convex compact subset of $L_\infty(\mathbb{R}^d)^{M+1}$ endowed with the weak* topology. The above optimization problem can now be restated as:

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

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

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

We consider the relaxation

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

Viewing the $\lambda_i$’s as extended positive real-valued, this has the advantage of being a minmax problem for a separately linear functional defined on a product of compact convex sets, so that the minmax theorem ([4], Theorem 3) applies. As we show later, the saddle point $(\lambda^*, \varphi^*)$ satisfies $\varphi^* \in H_0$, so that the relaxation is equivalent to the original problem. By minmax theorem, (1) equals

$$\begin{align*}
&\max_{(\lambda, (\pi_i)) \in \Gamma} \min_{\varphi \in H} \left( \int \int \pi_0(d\theta)f_\theta(x)(1-\varphi_0(x))dx \\
&\quad + \sum_{i=1}^M \lambda_i \left( \int \int \pi_i(d\theta)f_\theta(x)(1-\varphi_i(x))dx - \alpha_i \right) \right).
\end{align*}$$
This in turn equals
\[-\min_{(\lambda, \{\pi_i\}) \in \Gamma} \max_{\varphi \in H} \left( \int \int \pi_0(d\theta) f_\theta(x) \varphi_0(x) \right) dx \]
\[+ \sum_{i=1}^{M} \lambda_i \int \int \pi_i(d\theta) f_\theta(x) \varphi_i(x) dx - 1 - \sum_{i=1}^{M} \lambda_i(1 - \alpha_i). \tag{2} \]

The inner maximum is clearly attained at \( \varphi^* \) given by
\[\varphi_i^*(x) = I\{\lambda_i \int \pi_i(d\theta) f_\theta(x) > \max_{j<i} \lambda_j \int \pi_j(d\theta) f_\theta(x), \]
\[\lambda_i \int \pi_i(d\theta) f_\theta(x) \geq \max_{j>i} \lambda_j \int \pi_j(d\theta) f_\theta(x) \}\]

Thus (2) equals
\[-\min_{(\lambda, \{\pi_i\}) \in \Gamma} \left( \int \max_{0 \leq i \leq M} \lambda_i \int \pi_i(d\theta) f_\theta(x) \right) dx - 1 - \sum_{i=1}^{M} \lambda_i(1 - \alpha_i). \]

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

For \(M = 1\), this is seen to reduce to the ‘ratio test’: Pick the null hypothesis if
\[\frac{\int \pi_0(d\theta) f_\theta(Y)}{\int \pi_1(d\theta) f_\theta(Y)} \geq \lambda_1, \]
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
\[\min_{0 \leq i \leq M} \max_{\theta \in \Theta} P_\theta(\gamma(Y) \neq i).\]
Redefine $Q$ by $Q \triangleq \{\lambda = [\lambda_0, \cdots, \lambda_M] : \lambda_i \geq 0 \ \forall i, \sum_{i=0}^M \lambda_i = 1\}$. Define $\Gamma$ as before. By familiar arguments, we consider the relaxation

$$
\min_{\varphi \in \mathcal{H}} \max_{(\lambda, \{\pi_i\}) \in \Gamma} \int \sum_i \lambda_i \int \pi_i(d\theta)f_\theta(x)(1 - \varphi_i(x))dx
$$

$$
= \max_{(\lambda, \{\pi_i\}) \in \Gamma} \min_{\varphi \in \mathcal{H}} \int \sum_i \lambda_i \int \pi_i(d\theta)f_\theta(x)(1 - \varphi_i(x))dx
$$

$$
= \max_{(\lambda, \{\pi_i\}) \in \Gamma} \min_{\varphi \in \mathcal{H}} (1 - \int \sum_i \lambda_i \int \pi_i(d\theta)f_\theta(x)\varphi_i(x)dx)
$$

$$
= 1 - \min_{(\lambda, \{\pi_i\}) \in \Gamma} \max_{\varphi \in \mathcal{H}} \int \sum_i \lambda_i \int \pi_i(d\theta)f_\theta(x)\varphi_i(x)dx
$$

$$
= 1 - \min_{(\lambda, \{\pi_i\}) \in \Gamma} \int \max_i(\lambda_i \int \pi_i(d\theta)f_\theta(x))dx.
$$

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

**Theorem 2** 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

$$
\lambda^*_i \int \pi^*_i(d\theta)f_\theta(Y) > \max_{j<i} \lambda^*_j \int \pi^*_j(d\theta)f_\theta(Y),
$$

$$
\lambda^*_i \int \pi^*_i(d\theta)f_\theta(Y) \geq \max_{j>i} \lambda^*_j \int \pi^*_j(d\theta)f_\theta(Y).
$$

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

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

**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 $H_3$ is not a factor and the optimal solution is given by comparing $H_3$ against the set $\Theta_1$. 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:

\[ \lambda_j = 1, \quad \forall j, \quad \pi_0(H_2) = 1, \quad \pi_1(H_2) = 1. \]

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 \( \theta \in \Theta \) a compact convex subset of some \( \mathbb{R}^m \), \( m \geq 1 \), 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 \( \Theta \) as above, we now seek to find the ‘best’ Borel subset of \( \Theta \) 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

\[ \sup_{\theta \in \Theta} P_\theta(\theta \notin \psi(Y)). \]  

(3)

This formulation is from [6].

Since a larger set can only reduce the risk and the largest set of diameter not exceeding \( \epsilon \) is the closed ball of radius \( \epsilon \), 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 \( \epsilon \) 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:

\[ (\dagger) \text{The maps } \theta \to f_\theta(x), x \in \mathbb{R}^d, \text{ are bounded and continuously differentiable with} \]
\[ \sup_x ||\nabla^\theta f_\theta(x)|| \leq q(x) \]

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

Let \( D(\Theta) \) denote the set of continuous probability densities over \( \Theta \). Then the set \( \mathcal{P}_0(\Theta) \Delta = \{ \phi(\cdot)d\theta : \phi(\cdot) \in D(\Theta) \} \) is dense in \( \mathcal{P}(\Theta) \). Let \( B \) denote the set of closed balls of diameter \( \epsilon \) with centers in \( \Theta \), endowed with the metric topology of the Hausdorff metric. Let \( \mathcal{M} \) denote the set of measurable maps \( \psi : \mathbb{R}^d \to B \) and \( \mathcal{C} \) the set of measurable maps \( \eta : \mathbb{R}^d \to \Theta \). We may identify an element \( \psi \) of the former family with an element \( \eta \) 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 \( P_\theta \) for \( \theta \in \Theta \).

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

**Proof** \( \rho \) is clearly a metric. Recalling that \( \Theta \) 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 [3], with the metric convergence corresponding to the convergence in probability under \( P_\theta \). A Cauchy sequence \( \{ \eta_n(Y) \} \) under \( \rho \) will be Cauchy with respect to each \( \rho_\theta \) 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.s. on the
intersection of the supports of $f_{\theta}, f_{\theta'}$ for any two distinct $\theta, \theta'$. Thus we can consistently define a random variable $\hat{\eta}_\infty$ such that $\eta_n(Y) \to \hat{\eta}_\infty$ under each $\rho_{\theta}$. To claim that it also does so under $\rho$, we need to prove that this convergence is uniform in $\theta$. This follows from assumption (†) above (which implies pointwise boundedness and equicontinuity of $\rho_{\theta}(\eta_n(Y), \eta_\infty(Y))$ in $\theta$) 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 $\eta_\infty \in C$ by Theorem 1.1.4, p. 5, of [3]. Separability follows from the density of $\eta$ 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. \hfill \Box

The infimum of (3) equals

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

Consider

$$F(\pi, \eta(Y)) \triangleq \int \int \pi(d\theta)f_\theta(x)I\{|\theta - \eta(x)| \leq \epsilon\}dx.$$ 

This is linear and continuous in $\pi$ 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 $B(z) \triangleq \{z : \|x - z\| \leq \epsilon\}$ is a closed set. Therefore, taking the standard relaxation of the above minmax problem, we consider

$$\sup_{\Phi \in \mathcal{P}(\mathcal{H})} \min_{\pi \in \mathcal{P}(\Theta)} \int \Phi(d\eta)F(\pi, \eta)$$
\[
\inf_{\pi \in \mathcal{P}(\Theta)} \sup_{\eta \in \mathcal{P}(\mathcal{H})} \int \Phi(d\eta) F(\pi, \eta)
\]
\[
= \inf_{\lambda \in \mathcal{D}(\Theta)} \sup_{\eta \in \mathcal{C}} \int \lambda(\theta) f_\theta(x) I\{|\theta - \eta(x)| \leq \epsilon\} dx d\theta
\]
\[
= \inf_{\lambda} \left( \int \sup_z \left[ \int B(z) \lambda(\theta) f_\theta(x) d\theta \right] dx \right),
\]
where the first equality follows from the minmax theorem (Theorem 3 of [4]), 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
\[
\eta(Y) \in \text{Argmax}_z \int_{B(z)} \lambda(\theta) f_\theta(Y) d\theta.
\]

Thus we have:

**Theorem 3** 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 \(\hat{\theta} \in \text{Argmax}_z \int_{B(z)} \lambda^*(\theta) f_\theta(Y) d\theta\) is the optimum choice for the relaxed problem.

Unfortunately the existence of a minimizer \(\lambda^*\) 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 \(\lambda^*\) will then be near-optimal rather than optimal.

Note also that we may replace \(\mathcal{P}_0(\Theta)\) by any other convenient dense subset of \(\mathcal{P}(\Theta)\), e.g., by \(\mathcal{P}_1(\Theta) \triangleq \) the set of finitely supported probability measures on \(\Theta\). Minimization over \(\pi \in \mathcal{P}_1(\Theta)\) may then be approximated by minimization over the set of finitely supported probability measures on \(\Theta\) 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
\[
\eta(y) \in \text{Argmax}_{z \in \text{support}(\pi^*)} \sum_{\theta \in B(z) \cap \text{support}(\pi^*)} \pi^*(\{\theta\}) f_\theta(Y).
\]

If the variation of \(f_\theta(x)\) in \(\theta\) over an \(\epsilon\)-ball is small, the optimal decision in Theorem 3 may be approximated by
\[
\eta(Y) \in \text{Argmax}_\theta \lambda(\theta) f_\theta(Y).
\]

The \(\pi^*\) 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 [5]. Finally, observe that the computational aspects hinge on a convex programming problem for which many effective algorithms are available [1].
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, \( \theta \), 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. 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 \( \Theta \) is picked based on the following rule:

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

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 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:

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

which is considerably smaller than that obtained for ML.

Our next step is to show as to 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.,

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

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

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

In the new algorithm, 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:

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

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.

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
Figure 1: Illustration of the tradeoff between resolution and accuracy for different noise variances.
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.

References


