MINIMAX ESTIMATION OF A MEAN VECTOR FOR DISTRIBUTIONS ON A COMPACT SET

By Richard Dykstra[1]

University of Iowa, USA

Abstract

Minimax estimation procedures for the mean vector of a distribution on a compact set under squared error type loss functions are considered. In particular, a Dirichlet process prior is used to show that a linear function of $\bar{X}$ is a minimax estimator in the class of all measurable estimators and all possible distributions. This effort extends some earlier work of Bühlmann to a more general setting.

Keywords and phrases

Minimax decision rule; squared error loss; Dirichlet process; compact sets; Bayes rules; isotonic regression.

1. introduction

In an often cited paper, Bühlmann (1976) has considered linear minimax estimators for the mean of a univariate distribution in a nonparametric setting under squared error loss.

To be precise, Bühlmann assumes that $F_\theta(x)$ is a family of CDF's indexed by the one-dimensional parameter θ and that $U(\theta)$ is a CDF. Bühlmann also assumes that after observing the random variable X, distributed as $F_\theta(x)$, the actuary chooses a *linear* estimator of the form $d(X) = \gamma X + \delta$ to estimate the mean of X. Nature chooses (a) a family of distributions $F_\theta(x)$ and (b) a CDF $U(\theta)$ for Θ. Because (a) and (b) together determine a joint distribution for X and Θ, a natural loss function is given by

$$\begin{aligned} L[(F, U), (\gamma, \delta)] &= E[\gamma X + \delta - \mu(\Theta)]^2 \\ &= \int (\gamma x + \delta - \mu(\theta))^2 \, F_\theta(dx)\, U(d\theta) \\ &= \gamma^2 v + (1-\gamma)^2 w + [(1-\gamma) m - \delta]^2 \end{aligned}$$

where

$$v = E[\sigma^2(\Theta)], \quad w = \text{Var}\,[\mu(\Theta)], \quad \text{and} \quad m = E[\mu(\Theta)].$$

[1] This work was partially supported by Office of Naval Research Contract N00014-83-K-0249.

https://doi.org/10.2143/AST.20.2.2005440 Published online by Cambridge University Press

Since the loss function depends only upon v, w, and m (when considering linear estimators), BÜHLMANN takes nature's action space to be $\{(v, w, m);\ v \in I_1, w \in I_2,\ m \in I_3\}$ where I_1, I_2 and I_3 are finite, closed intervals. BÜHLMANN has shown that under mild restrictions there exists a unique pure minimax strategy for the actuary, and a mixed minimax strategy for nature. BÜHLMANN also identifies the form of the actuary's minimax estimator, and the value of the minimax risk.

While elegantly presented and carefully done, there are some restrictions in Professor BÜHLMANN'S work which one would like to remove. In particular, estimators are only considered in the class of linear estimators, and then only on one observation. Conceivably, in the larger class of all estimators, the minimax risk could be substantially reduced. Even if one were satisfied to only consider linear estimators, BÜHLMANN'S suggestion to base the estimator on $\bar{X}$ (to reduce matters to one random variable), while reasonable, does not appear to be necessarily optimal in any sense. Finally, one might wish to estimate several means simultaneously if observing multivariate data.

In this paper, we show (in a necessarily slightly different setting) that when i.i.d. multivariate observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are observed in a compact region $\mathscr{X}$, a minimax estimator of $\boldsymbol{\mu}$ (the vector of means) of the form $\gamma\bar{\boldsymbol{X}}+\delta$ exists in the class of all measurable estimators for the family of all distributions over $\mathscr{X}$ when using a squared error type loss function.

2. RESULTS

To set notation, we let $\boldsymbol{X} = \{X(t, \omega);\ t \in T\}$ indicate a stochastic process defined on the measurable space $(\Omega, \mathscr{F})$ whose range is the set $\mathscr{X}$, where T is assumed to be a bounded Borel set in R^k. We assume that $\|\cdot\|$ is a norm defined on a linear vector space containing $\mathscr{X}$, and that

$$\|\boldsymbol{x}\| = \left[\int_T x^2(t)\, W(dt)\right]^{\frac{1}{2}}$$

for all $\boldsymbol{x} \in \mathscr{X}$ where W is a finite measure over T. Our stochastic process $\boldsymbol{X}$ is assumed to be measurable with respect to the σ-field generated by the open sets of $\|\cdot\|$, and $\mathscr{X}$ and all singleton subsets of $\mathscr{X}$ are assumed to belong to this σ-field. We let $\mathscr{P}$ denote the set of all possible distributions of $\boldsymbol{X}$, and also assume that $\mathscr{X}$ is compact in the topology associated with $\|\cdot\|$. (Typically, T will be the finite set $\{1, 2, \dots, n\}$, so that $\|\boldsymbol{x}\| = \left(\sum_1^n x_i^2\, \omega_i\right)^{\frac{1}{2}}$ for positive weights $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)$. This substitution can be made in the proofs to make them more intuitive. However the additional generality is often useful, e.g., see example 2).

We observe i.i.d. random vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, and wish to estimate $\boldsymbol{\mu}_P = E_P \boldsymbol{X}$ under the loss function

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2, \tag{2.1}$$

https://doi.org/10.2143/AST.20.2.2005440 Published online by Cambridge University Press

where $\hat{\boldsymbol{\mu}}$ is required to take values in the convex hull of $\mathscr{X}$, and be such that all appropriate expectations exist.

The following lemma, while straightforward, is crucial.

Lemma 2.1. There exists $P_0 \in \mathscr{P}$ such that

$$E_{P_0} \|\boldsymbol{X} - \boldsymbol{\mu}_{P_0}\|^2 = \sup_{P \in \mathscr{P}} E_P \|\boldsymbol{X} - \boldsymbol{\mu}_P\|^2 = c_0 .$$

Proof. Since $\mathscr{X}$ is compact, $\mathscr{P}$ is tight and hence relatively compact (BILLINGSLEY 1968, p. 37). Thus if $\{P_n\}$ is a sequence of distributions such that $E_{P_n} \|\boldsymbol{X} - \boldsymbol{\mu}_{P_n}\|^2 \to c_0$, there must exist a distribution P_0 and a subsequence n_j such that P_{n_j} converges weakly to P_0. However, since the integrands are bounded, uniform integrability arguments (SERFLING (1980), p. 14) imply the desired result. □

Moreover, c_0 is an extreme value in the following sense:

Lemma 2.2. For all $\boldsymbol{x} \in \mathscr{X}$, $\|\boldsymbol{x} - \boldsymbol{\mu}_{P_0}\|^2 \leq c_0$.

Proof. Suppose there exists $\hat{\boldsymbol{x}} \in \mathscr{X}$ such that $\|\hat{\boldsymbol{x}} - \boldsymbol{\mu}_{P_0}\|^2 = c_1 > c_0$. Let $\hat{P}$ be the distribution degenerate at $\hat{\boldsymbol{x}}$, and let $P^* = \alpha P_0 + (1-\alpha)\hat{P}$. Then

$$\begin{aligned} E_{P^*} \|\boldsymbol{X} - \boldsymbol{\mu}_{P^*}\|^2 &= \int \|\boldsymbol{x} - \boldsymbol{\mu}_{P^*}\|^2 \, dP^* \\ &= \int \|\alpha(\boldsymbol{x} - \boldsymbol{\mu}_{P_0}) + (1-\alpha)\,(\boldsymbol{x} - \boldsymbol{\mu}_{\hat{P}})\|^2 \, dP^* \\ &= \alpha \int \|\mathbf{x} - \boldsymbol{\mu}_{P_0}\|^2 \, dP_0 + \alpha(1-\alpha)\,\|\boldsymbol{\mu}_{P_0} - \boldsymbol{\mu}_{\hat{P}}\|^2 \\ &= \alpha\, c_0 + \alpha(1-\alpha)\, c_1 \end{aligned}$$

by expanding the integrand and using linearity properties of integrals. However, for $0 < \alpha < 1$,

$$\alpha\, c_0 + \alpha(1-\alpha)\, c_1 > c_0 = \alpha\, c_0 + (1-\alpha)\, c_0$$

if

$$\alpha > \frac{c_0}{c_1},$$

which clearly is inconsistent with the definition of c_0. □

The following lemma now follows easily from Lemma 2.1 and Lemma 2.2.

https://doi.org/10.2143/AST.20.2.2005440 Published online by Cambridge University Press

Lemma 2.3. It must be the the case that $\|x-\mu_{P_0}\|^2 = c_0$ a.s. (P) if $P \ll P_0$.

A prior distribution which can be used to place probability over a large class of distributions is the Dirichlet process prior discussed in FERGUSON (1973). In particular, if α is a finite measure over $\mathscr{X}$ and $\mathscr{D}(\alpha)$ an associated Dirichlet process prior, the posterior distribution given observations $x_1, \ldots, x_n$ will be $\mathscr{D}\left(\alpha + \sum_1^n \delta_{x_i}\right)$ (where δ_x indicates a degenerate distribution at x), from FERGUSON (1973). (A Dirichlet process is an extension of the Dirichlet distribution to the continuous case. It has many attractive mathematical properties).

Since the loss function is of a squared error type, a Bayes rule for this prior would be

$$\frac{\alpha(\mathscr{X})\mu_\alpha + n\bar{X}}{\alpha(\mathscr{X})+n}$$

where $\mu_\alpha(t)$ is the mean of $X(t)$ with respect to the probability distribution $\alpha/\alpha(\mathscr{X})$ (see FERGUSON, 1973).

Now the risk function of such an estimator can be written as

$$\begin{aligned}
E_P & \left\| \frac{n\bar{X}+\alpha(\mathscr{X})\mu_\alpha}{\alpha(\mathscr{X})+n} - \mu_P \right\|^2 \\
& = E_P \left\| \frac{n(\bar{X}-\mu_p)}{\alpha(\mathscr{X})+n} + \frac{\alpha(\mathscr{X})}{\alpha(\mathscr{X})+n}(\mu_\alpha - \mu_P) \right\|^2 \\
& = \left\| \left[\left(\frac{n}{\alpha(\mathscr{X})+n} \right)^2 \frac{\sigma_P^2}{n} + \left(\frac{\alpha(\mathscr{X})}{\alpha(\mathscr{X})+n} \right)^2 (\mu_\alpha - \mu_P)^2 \right]^{1/2} \right\|^2 \\
& = \frac{n}{(n+\sqrt{n})^2} \left\| \left[\sigma_P^2 + (\mu_\alpha - \mu_P)^2 \right]^{1/2} \right\|^2 \qquad (\text{if } \alpha(\mathscr{X}) = \sqrt{n}) \\
& = \frac{n}{(n+\sqrt{n})^2} \|[E_P(X-\mu_\alpha)^2]^{1/2}\|^2 \\
& = \frac{n}{(n+\sqrt{n})^2} E_P\|(X-\mu_\alpha)\|^2 .
\end{aligned}$$

Now, if α is taken to be $\sqrt{n}\, P_0$, we have $\mu_\alpha = \mu_{P_0}$, and the risk function must be bounded above by $\frac{n}{(n+\sqrt{n})^2} c_0$ from Lemma 2.2.

However, this prior puts probability on distributions whose support is contained in the support of P_0 with probability one, so that the Bayes risk of the decision rule

$$\frac{n\bar{\boldsymbol{X}}+\sqrt{n}\,\boldsymbol{\mu}_{P_0}}{n+\sqrt{n}}$$

must be $\dfrac{n}{(n+\sqrt{n})^2}\,c_0$ (by Lemma 2.3). Since we have a Bayes rule whose risk function is bounded above by its Bayes risk, it must be minimax. This proves the following theorem.

Theorem 2.1. For the loss function defined in (2.1), $(n+\sqrt{n})^{-1}(n\bar{\boldsymbol{X}}+\sqrt{n}\,\boldsymbol{\mu}_{P_0})$ is a minimax estimator of $\boldsymbol{\mu}_P$ in the set of all measurable estimators for the class of all distributions $\mathscr{P}$.

3. EXAMPLES

(1) Suppose one observes $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ which are i.i.d. multinomial $(1, p_1, \ldots, p_k)$ and we wish to estimate $\boldsymbol{p} = (p_1, \ldots, p_k)$ under the loss function

$$L(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \sum_1^k (p_i - \hat{p}_i)^2 w_i .$$

Since X_{1i} is binomial $(1, p_i)$, choosing $\boldsymbol{p}_0$ to solve $\sup\limits_{\boldsymbol{p}} E\|\boldsymbol{X}-\boldsymbol{\mu}_{\boldsymbol{p}}\|^2$ is equivalent to the problem

$$\sup_{\substack{\Sigma p_i = 1\\ p_i \geq 0}} \sum_{i=1}^k p_i(1-p_i)\,w_i .$$

It is straightforward to show that this is solved by

$$\tilde{p}_i = \frac{1}{2} - \frac{(k-2)\left(\sum\limits_1^k w_i^{-1}\right)^{-1}}{2w_i}, \qquad i = 1, \ldots, k,$$

if these values are nonnegative (if not, further adjustments are necessary). In this case, our earlier work ensures that if

$$p_i^* = \frac{n\bar{X}_i + \sqrt{n}\,\tilde{p}_i}{n+\sqrt{n}}, \qquad i = 1, \ldots, k,$$

then $\boldsymbol{p}^* = (p_1^*, \ldots, p_k^*)$ is a minimax estimator of $\boldsymbol{p}$.

(2) Suppose i.i.d. observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are taken from a k-variate distribution. It is assumed that the support of the distribution is contained in the set $\mathscr{X} = [a_1, b_1] \times \ldots \times [a_k, b_k]$, and we wish to estimate the joint CDF under the loss function

https://doi.org/10.2143/AST.20.2.2005440 Published online by Cambridge University Press

$$L(F, \hat{F}) = \int_{\mathscr{X}} (\hat{F}(\boldsymbol{t}) - F(\boldsymbol{t}))^2 \, W(d\boldsymbol{t})$$

where W is an arbitrary finite measure over $\mathscr{X}$.

If we let

$$Y(\boldsymbol{t}) = I_{(-\infty, t_1] \times \ldots \times (-\infty, t_k]}(\boldsymbol{X}),$$

then $E(Y(\boldsymbol{t})) = F(\boldsymbol{t})$, and the problem fits nicely into our earlier framework. Clearly the solution to

$$\sup_{P \in \mathscr{P}} E_P \| \boldsymbol{Y} - \boldsymbol{\mu}_P \|^2$$

is given by the distribution that puts probability one-half on $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$. For this distribution, the mean of $Y(\boldsymbol{t})$ is $\frac{1}{2}$, $\boldsymbol{t} \in \mathscr{X}$, $\boldsymbol{t} \neq \boldsymbol{b}$, and it follows that

$$F^*(\boldsymbol{t}) = \frac{n\hat{F}_n(\boldsymbol{t}) + \sqrt{n}\,\frac{1}{2}}{n + \sqrt{n}}, \qquad \boldsymbol{t} \in \mathscr{X} - \{\boldsymbol{b}\}$$

(with the obvious extensions elsewhere) is a minimax estimator of F where $\hat{F}_n(\boldsymbol{t})$ is the standard empirical CDF. Note that $\boldsymbol{a}$ and $\boldsymbol{b}$ only enter into F^* by defining where certain jumps occur, and that the estimator is totally free of W. PHADIA (1973) has obtained minimax results of this type, and HJORT (1976) has taken a similar approach in his thesis.

(3) Suppose i.i.d. observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are taken from a distribution with support $[a, b] \times [a, b] \times \ldots \times [a, b]$. It is desired to estimate $\boldsymbol{\mu}$ under the restrictions $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_k$ with loss function $L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{k} (\mu_i - \hat{\mu}_i)^2 w_i$. Clearly, the distribution with probability one-half on $\boldsymbol{a} = (a, \ldots, a)$ and $\boldsymbol{b} = (b, \ldots, b)$ solves $\sup_{P \in \mathscr{P}} E \| \boldsymbol{X} - \boldsymbol{\mu}_P \|^2$.

Then

$$d(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{n\bar{\boldsymbol{X}} + \sqrt{n}\,(\boldsymbol{a} + \boldsymbol{b})/2}{n + \sqrt{n}}$$

is a minimax estimator (ignoring the order restrictions) and every other estimator has a risk value at least as large as that of d for some distribution which puts probability only on $\boldsymbol{a} = (a, \ldots, a)$ and $\boldsymbol{b} = (b, \ldots, b)$ (and hence has correctly ordered means). However, if $\boldsymbol{\mu}^*$ is the isotonic regression of $\bar{\boldsymbol{X}}$ with weights $\boldsymbol{w}$ and the linear increasing order (see ROBERTSON et al., Chap. 1), then the isotonic regression of $d(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is

$$d^*(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{n\boldsymbol{\mu}^* + \sqrt{n}\,(\boldsymbol{a} + \boldsymbol{b})/2}{n + \sqrt{n}},$$

https://doi.org/10.2143/AST.20.2.2005440 Published online by Cambridge University Press

and

$$E_P\,L(\boldsymbol{\mu}_P, d) \geq E_P\,L(\boldsymbol{\mu}_P, d^*)$$

for all distributions P with correctly ordered means (ROBERTSON et al., 1988, Section 1.6).

It follows that d^* must be a minimax estimator of $\boldsymbol{\mu}$ within the class of all estimators for the set of all distributions with nondecreasing means and support contained in $[a, b] \times \ldots \times [a, b]$.

(4) In some situations, a Mahalanobis type loss function

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})'\, D\,(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})$$

(where D is a $k \times k$ symmetric, positive definite matrix) may be appropriate. However, if D is written as $O'\Lambda O$ where O is an orthogonal matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues, the loss function may be expressed as

$$\sum_{i=1}^{k} [(O\boldsymbol{\mu})_i - (O\hat{\boldsymbol{\mu}})_i]^2\, \lambda_i\,.$$

If one makes the change of variables $Y_i = OX_i$, then $E(Y_i) = O\boldsymbol{\mu}$, and Y_i must take values in the set $O\mathscr{X} = \{O\boldsymbol{x};\, \boldsymbol{x} \in \mathscr{X}\}$. Then if $d(Y_1, \ldots, Y_n)$ is a minimax decision rule for the transformed problem, $O'd(Y_1, \ldots, Y_n)$ must be a minimax decision rule for the original problem. Thus our results also apply to Mahalanobis type loss functions as well as weighted squared error loss.

4. ACKNOWLEDGEMENT

The author would like to thank the referees for a careful reading and valuable comments on an earlier version of the manuscript.

REFERENCES

BÜHLMANN, HANS (1976) Minimax Credibility. *Scand. Actuarial J.* 65–78.

BILLINGSLEY, PATRICK (1968) *Convergence of Probability Measures.* Wiley, New York.

FERGUSON, T.S. (1973) A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.* **1**, 209–230.

HJORT, N.L. (1976) Dirichletprosessen anvendt på noen ikke-parametriske estimeringsproblemer. Unpublished thesis from the University of Tromsø, Norway.

PHADIA, E.G. (1973) Minimax estimation of a cumulative distribution function. *Annals of Statistics,* 1149–1157.

ROBERTSON, TIM, WRIGHT, F.T. and DYKSTRA, R.L. (1988) *Order Restricted Statistical Inference.* Wiley, Chichester.

SERFLING, ROBERT J. (1980) *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

RICHARD DYKSTRA
Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242/USA.

https://doi.org/10.2143/AST.20.2.2005440 Published online by Cambridge University Press