The Annals of Statistics
1992, Vol. 20, No. 3, 1637–1646

MINIMUM IMPURITY PARTITIONS

BY DAVID BURSHTEIN, VINCENT DELLA PIETRA, DIMITRI KANEVSKY AND ARTHUR NÁDAS

Tel Aviv University and IBM, T. J. Watson Research Center

Let (X, U) be jointly distributed on $\mathscr{X} \times \mathscr{R}^n$. Let $Y = E(U|X)$ and let $\mathscr{U}$ be the convex hull of the range of U. Let $C: \mathscr{X} \to \mathscr{C} = \{1, 2, \ldots, k\}$, $k \geq 1$, induce a measurable k way partition $\{\mathscr{X}_1, \ldots, \mathscr{X}_k\}$ of $\mathscr{X}$. Define the impurity of $\mathscr{X}_c = C^{-1}(c)$ to be $\phi(c, E(U|C(X) = c))$, where $\phi: \mathscr{C} \times \mathscr{U} \to \mathscr{R}^1$ is a concave function in its second argument. Define the impurity Ψ of the partition as the average impurity of its members: $\Psi(C) = E\phi(C(X), E(U|C(X)))$. We show that for any $C: \mathscr{X} \to \mathscr{C}$ there exists a mapping $\tilde{C}: \mathscr{U} \to \mathscr{C}$, such that $\Psi(\tilde{C}(Y)) \leq \Psi(C)$ and such that $\tilde{C}^{-1}(c)$ is convex, that is, for each $i, j \in C$, $i \neq j$, there exists a separating hyperplane between $\tilde{C}^{-1}(i)$ and $\tilde{C}^{-1}(j)$. This generalizes some results in statistics and information theory. Suitable choices of U and ϕ lead to optimal partitions of simple form useful in the construction of classification trees and multidimensional regression trees.

1. Introduction. Let (X, U) be jointly distributed random variables with values in $\mathscr{X} \times \mathscr{R}^n$, and let $\mathscr{U} \subset \mathscr{R}^n$ be the convex hull of the range of U. Let $\mathscr{C}$ be a finite set $\mathscr{C} = \{1, 2, \ldots, k\}$, and let $\phi: \mathscr{C} \times \mathscr{U} \to \mathscr{R}^1$ be concave in its second argument, that is,

$$\phi(c, ty_0 + (1 - t)y_1) \geq t\phi(c, y_0) + (1 - t)\phi(c, y_1)$$

for any $y_0, y_1 \in \mathscr{U}$ and $t \in [0, 1]$.

Let $Y = E\{U|X\}$ so that $Y: \mathscr{X} \to \mathscr{U}$ is a random variable on $\mathscr{X}$. For any measurable partition $C: \mathscr{X} \to \mathscr{C}$, define the objective function $\Psi(C)$ by

$$\Psi(C) = E\phi(C(X), E(U|C(X))) = E\phi(C(X), E(Y|C(X))).$$

Explicitly,

$$\begin{aligned} \Psi(C) &= \sum_{c \in \mathscr{C}} P(C^{-1}(c))\phi(c, E(U|C(X) = c)) \\ &= \sum_{c \in \mathscr{C}} P(C^{-1}(c))\phi\left(c, \frac{\int_{C^{-1}(c)} Y(x)P(dx)}{P(C^{-1}(c))}\right). \end{aligned} \tag{1}$$

We interpret C as partitioning $\mathscr{X}$ into k subsets labeled by $\mathscr{C}$, and we interpret $\Psi(C)$ as the cost of partitioning. We are interested in finding a C

Received August 1989; revised June 1991.

AMS 1980 *subject classifications*. Primary 62H30, 62C05, 62C10, 62J02; secondary 68T05, 68T10.

Key words and phrases. Classification, discrimination, decision theory, regression, partitioning, decision trees, CART.

with minimum cost. We note that an important special case of (1) is

$$\Psi(C) = \sum_{c \in \mathscr{C}} P(C^{-1}(c))\phi(c, P(\Theta = 1|C(X) = c), \ldots, P(\Theta = n|C(X) = c)), \tag{2}$$

where Θ is a random variable with values in the finite set: $\{1, 2, \ldots, n\}$. (2) is a special case of (1) obtained by taking $U = [\delta(\Theta, 1), \ldots, \delta(\Theta, n)]$, where $\delta(\Theta, i) = 1$ if $\Theta = i$ and 0 otherwise.

The problem of minimizing special cases of (1) and (2) has been studied both in the literature of statistics, [Anderson (1984), Breiman, Friedman, Olshen and Stone (1984) and Fisher (1958)] and information theory [Chou (1988) and Nemetz (1967)].

In particular (2) includes the objective function of the Bayesian classification problem with possibly different costs assigned to different types of classification errors [e.g., Anderson (1984), page 224], as a special case: that is, suppose that Θ is a random variable taking n possible values, and suppose that based on the value of X, we wish to decide on the value of Θ. Let C: $\mathscr{X} \to \mathscr{C}$ be the decision mapping. If $\Theta = i$ and $C(X) = j$, then the classification cost is $d_{i,j}$. The cost function is

$$\Psi(C) = \sum_{c \in \mathscr{C}} P(C^{-1}(c)) \sum_{c' \in \mathscr{C}} d_{c',c} P(\Theta = c'|C(X) = c),$$

which is a special instance of (2), since linear functions are concave.

Breiman, Friedman, Olshen and Stone (1984), Chou (1988), Fisher (1958) and Nemetz (1967) studied special cases of (1) and (2). They all assume $\phi(c, \cdot) = \phi(\cdot)$ is independent of c. The problem of minimizing (1) was studied by Fisher (1958) for the case where $\mathscr{X}$ is a finite set, k is arbitrary, $n = 1$, and where the objective function $\phi(y)$ is $-y^2$, the motivation being to group (cluster) members with similar characteristics. In connection with the construction of decision trees, Breiman, Friedman, Olshen and Stone (1984) studied (1) for the case where $\mathscr{X}$ is a finite set, ϕ is any concave function, $k = 2$ and $n = 1$. They also studied (2) for the case $k = 2$, $n = 2$. The construction technique of Breiman, Friedman, Olshen and Stone consists of an initial tree growing step and a second pruning step. Tree growing is performed by recursive partitioning of the predictor space $\mathscr{X}$, both for regression trees via (1) and for classification trees via (2).

Chou (1988) has studied the objective function (2) for the case where $\mathscr{X}$ is a finite set, k and n are arbitrary and where ϕ is the Shannon entropy function:

$$\phi(u_1, u_2, \ldots, u_n) = -\sum_{i=1}^{n} u_i \log u_i.$$

Nemetz (1967) has treated the case where $k = 2$, $n = 2$, Ψ has the form (2) and ϕ is the Shannon entropy function. In Section 2 we study the general case, thus generalizing the results of Breiman, Friedman, Olshen and Stone (1984), Chou (1988), Fisher (1958) and Nemetz (1967). One of the consequences of our results for finite $\mathscr{X}$ is an algorithm with complexity polynomial

in the size of $\mathscr{X}$, for finding a partition C that minimizes $\Psi(C)$. In Section 3 we characterize $\Psi(C)$ in decision-theoretic terms and give examples.

2. Results.

THEOREM 1. *For any* C: $\mathscr{X} \to \mathscr{C}$ *there exists a* $\tilde{C}$: $\mathscr{U} \to \mathscr{C}$ *such that* $\Psi(\tilde{C}(Y)) \leq \Psi(C)$ *and such that* $\tilde{C}^{-1}(c)$ *is convex for all* $c \in \mathscr{C}$.

PROOF. Let $p = (p(1), p(2), \ldots, p(k)) \in \mathscr{R}^k$, $y(i) \in \mathscr{R}^n$ for $i = 1, 2, \ldots, k$ and $y = (y(1), y(2), \ldots, y(k)) \in \mathscr{R}^{kn}$. Now define the convex subset $\mathscr{W} \subset \mathscr{R}^{k(n+1)}$ by

$$\mathscr{W} = \left\{ (p, y) : \text{for each } c \in \mathscr{C} \text{ either } p(c) = 0 \text{ or } p(c) > 0 \text{ and } \frac{y(c)}{p(c)} \in \mathscr{U} \right\}.$$

Define η: $\mathscr{W} \to \mathscr{R}^1$ by

$$\eta(p, y) = \sum_{c \in \mathscr{C}} p(c) \phi\left(c, \frac{y(c)}{p(c)} \right) \quad \text{for } (p, y) \in \mathscr{W}.$$

Then for any C: $\mathscr{X} \to \mathscr{C}$,

$$\Psi(C) = \eta(p_C, y_C), \quad \text{where } p_C(c) = P(C^{-1}(c))$$

and

$$y_C(c) = \int_{C^{-1}(c)} Y(x) P(dx).$$

Now η is concave on $\mathscr{W}$. Indeed, suppose $w_0, w_1 \in \mathscr{W}$ and $0 \leq t \leq 1$. Set $w_t = tw_0 + (1 - t)w_1 = (p_t, y_t)$. Then

$$\begin{aligned}
\eta(w_t) &= \sum_{c \in \mathscr{C}} p_t(c) \phi\left(c, \frac{y_t(c)}{p_t(c)} \right) \\
&= \sum_{c \in \mathscr{C}} p_t(c) \phi\left(c, \frac{tp_0(c)}{p_t(c)} \frac{y_0(c)}{p_0(c)} + \frac{(1 - t)p_1(c)}{p_t(c)} \frac{y_1(c)}{p_1(c)} \right) \\
&\geq \sum_{c \in \mathscr{C}} tp_0(c) \phi\left(c, \frac{y_0(c)}{p_0(c)} \right) + \sum_{c \in \mathscr{C}} (1 - t)p_1(c) \phi\left(c, \frac{y_1(c)}{p_1(c)} \right) \\
&= t\eta(w_0) + (1 - t)\eta(w_1).
\end{aligned}$$

Now, for two vectors $u, v \in \mathscr{R}^m$ we define the ordering relation $u \leq v$ if $u = v$ or if $u_i < v_i$, where i is the first coordinate j for which $u_j \neq v_j$. Thus, by Proposition 2, there exist an integer m and an $m \times k(n + 1)$ matrix Λ such that

$$\eta(w') \leq \eta(w) \quad \text{if } \Lambda w' \leq \Lambda w.$$

Λw has the form

$$\Lambda w = \Lambda(p, y) = \sum_{c \in \mathscr{C}} (\alpha(c)p(c) + \Omega(c)y(c))$$

for some $\alpha(c) \in \mathscr{R}^m$ and some $m \times n$ matrices $\Omega(c)$. In particular, for C: $\mathscr{X} \to \mathscr{C}$,

$$\Lambda(p_C, y_C) = \int_{x \in \mathscr{X}} P(dx)(\alpha(C(x)) + \Omega(C(x))Y(x)).$$

Define $\tilde{C}$: $\mathscr{U} \to \mathscr{C}$ by

$$\tilde{C}(u) = \arg\min_{c \in \mathscr{C}}(\alpha(c) + \Omega(c)u),$$

where for f: $\mathscr{C} \to \mathscr{R}^m$, $\arg\min_{c \in \mathscr{C}} f$ is the smallest $c \in \mathscr{C}$ at which f attains its minimum.

It is clear that $\tilde{C}^{-1}(c)$ is convex for any c. Moreover, since the ordering in $\mathscr{R}^m$ respects addition and multiplication by nonnegative scalars, it is clear that $\Lambda(p_{C'}, y_{C'}) \leq \Lambda(p_C, y_C)$,where $C' = \tilde{C}(Y)$. Thus $\eta(p_{C'}, y_{C'}) \leq \eta(p_C, y_C)$, so

$$\Psi(\tilde{C}(Y)) = \Psi(C') = \eta(p_{C'}, y_{C'}) \leq \eta(p_C, y_C) = \Psi(C)$$

and the proof is complete. □

The geometrical interpretation of Theorem 1 is that for any $i, j \in \mathscr{C}$, $i < j$, there exists a hyperplane that separates the set $\tilde{C}^{-1}(i)$ from the set $\tilde{C}^{-1}(j)$. It follows that $\tilde{C}^{-1}(i)$ is the convex set obtained as the intersection of $\mathscr{U}$ with an intersection of half-spaces. For the case $\mathscr{U} = \mathscr{R}^1$ (k arbitrary), we have that $\tilde{C}^{-1}(c)$, $c = 1, 2, \ldots, k$, are disjoint line segments, that is, $\tilde{C}^{-1}(c)$ has one of the four forms:

$$(y^{c-1}, y^c) \quad \text{or} \quad [y^{c-1}, y^c) \quad \text{or} \quad (y^{c-1}, y^c] \quad \text{or} \quad [y^{c-1}, y^c],$$

where if $\mathscr{U} = (y_1, y_2)$ or $[y_1, y_2)$ or $(y_1, y_2]$ or $[y_1, y_2]$ then $y^0 = y_1$ and $y^k = y_2$.

If $\mathscr{X}$ is a finite set with m members, then the complexity of a full search for a partition C that minimizes $\Psi(C)$ is k^m. On the other hand, Theorem 1 implies that there exists an algorithm, polynomial in m, for finding a mapping C that minimizes Ψ. This statement relies on a well-known theorem [e.g., Cover (1965)], which states that given some set of m points in $\mathscr{R}^n$, the number of possible divisions of the set into two linearly separable subsets (i.e., into two subsets that may be separated from each other by an hyperplane) is bounded by

$$2\sum_{k=0}^{n} \binom{m-1}{k},$$

which is polynomial in m. This result implies that the number of possible divisions of the m points into k linearly separable subsets (i.e., k subsets, where each may be linearly separated from any other subset) is also polynomial in m. Thus, by Theorem 1, when searching for the best partition, one

needs to compute the objective function Ψ, only for these polynomial number of subsets.

3. Characterization and examples. In this section we give a decision-theoretic characterization of the objective function (1). We formulate a general partitioning problem in decision-theoretic terms and show that both the appearance of a concave function and the reduction of U (or Y) to its conditional expectation in (1) follow naturally from the assumption of a linear loss function. We shall also describe some specific instances of this problem in clustering, statistical decision theory, classification and regression.

Let $\mathscr{X}$ be a sample space, P be a probability distribution on $\mathscr{X}$, and Y: $\mathscr{X} \to \mathscr{U}$ be any random variable. Let $\mathscr{A}$ be a set of actions and let L: $\mathscr{A} \times \mathscr{U} \to \mathscr{R}^1$ be a loss (cost) function. Suppose that when we observe $x \in \mathscr{X}$ we select an action $d(x) \in \mathscr{A}$ according to some decision function d: $\mathscr{X} \to \mathscr{A}$. The expected cost of this process is

$$r(d) = EL(d(X), Y(X)) = \int_{x \in \mathscr{X}} P(dx) L(d(x), Y(x)). \tag{3}$$

Now let $\mathscr{C}$ be a finite set of classes and for each class $c \in \mathscr{C}$, let $\mathscr{A}_c \subset \mathscr{A}$ be some set which specifies the actions which are allowed for c. Suppose that we require the strategy d to be of the form $d = \tilde{d}(C)$, where C: $\mathscr{X} \to \mathscr{C}$ is some partition of $\mathscr{X}$ into classes labeled by $\mathscr{C}$, and $\tilde{d}$: $\mathscr{C} \to \mathscr{A}$ is some choice of allowed action $\tilde{d}(c) \in \mathscr{A}_c$ for each class $c \in \mathscr{C}$. In other words, for a given partitioning function C, we only allow strategies (decision functions) d which select the same action $\tilde{d}(c)$ for all x assigned to the class $c = C(x)$. The lowest possible cost for a given C is

$$\Psi(C) = \inf\left\{ r\left(\tilde{d}(C)\right) \middle| \tilde{d}(c) \in \mathscr{A}_c \text{ for all } c \in \mathscr{C} \right\},$$

where the infimum is taken over the set of allowed decision functions $\tilde{d}$: $\mathscr{C} \to \mathscr{A}$.

The *partitioning problem* is to find a partition C of small or minimum cost $\Psi(C)$. We shall investigate partitioning assuming the following *linear* structure: $L(a, y) = L_0(a, y) + L_1(y)$ where L_0 is affine linear in y; that is,

$$L_0(a, ty_0 + (1-t)y_1) = tL_0(a, y_0) + (1-t)L_0(a, y_1)$$

for all $y_0, y_1 \in \mathscr{U}$ and $0 \le t \le 1$.

We shall show in the examples below how various questions in clustering, decision theory, classification and regression can be viewed as instances of this general problem for appropriate choices of $\mathscr{U}$, $\mathscr{A}$ and L.

Define ϕ: $\mathscr{C} \times \mathscr{U} \to \mathscr{R}^1$ by

$$\phi(c, y) = \inf\{L_0(a, y) | a \in \mathscr{A}_c\},$$

where the infimum is taken over the set of allowed actions a for the class c. The linearity property of L implies that $\phi(c, y)$ is *concave* in y. It is easy to

verify [simply note that $r(\tilde{d}(C)) = \sum_{c \in C} P(C^{-1}(c)) L_0(\tilde{d}(c), E(Y|C(X) = c))$ + const.)] that $\Psi(C)$ can be expressed as

$$\Psi(C) = \int_{x \in \mathcal{X}} P(dx) \phi(C(x), E(Y|C(x))) + \text{const.},$$

where const. $= \int_{x \in \mathcal{X}} P(dx) L_1(Y(x))$ and where $E(Y|C(x))$ is a shorthand notation for $E(Y|C)(C(x))$. $E(Y|C)$: $\mathcal{C} \to \mathcal{U}$ is the conditional expectation

$$E(Y|C)(c) = E(Y|C(X) = c) = \frac{\int_{C^{-1}(c)} P(dx) Y(x)}{P(C^{-1}(c))}. \tag{4}$$

We thus see that the objective function (1) is a natural choice for the partitioning problem.

In many of the examples below, ϕ can be interpreted as a measure of impurity or uncertainty, so that $\phi(c, E(Y|C(X) = c))$ can be viewed as the impurity of the class c, and $\Psi(C)$ can be viewed as the average impurity of the partition C.

4. Examples.

Clustering. Let Y: $\mathcal{X} \to \mathcal{U}$ be a measurable function. In clustering we wish to partition $\mathcal{X}$ into classes labeled by $\mathcal{C}$, and for each class choose a centroid $a = \tilde{d}(c) \in \mathcal{U}$ so as to minimize the average within-class squared deviation

$$r(\tilde{d}(C)) = E\|Y(X) - \tilde{d}(C(X))\|^2 = \int_{x \in \mathcal{X}} P(dx) \|Y(x) - \tilde{d}(C(x))\|^2.$$

The clustering cost can be viewed as an instance of the general cost (3) if we take $\mathcal{A} = \mathcal{U}$, and let $L(a, y) = \|y - a\|^2$.

This L satisfies the linearity condition with $L_0(a, y) = \|a - y\|^2 - \|y\|^2$. If we allow all maps $\tilde{d}$: $\mathcal{C} \to \mathcal{A}$ (so that $\mathcal{A}_c = \mathcal{A}$ for all c), then $\phi(y) = -\|y\|^2$. This is the multidimensional extension of Fisher's problem [Fisher (1958)].

Bayesian decision theory. Let $\mathcal{X}$ be a measure space, let Θ be a finite set, and let $\tilde{P}$ be a probability distribution on $\mathcal{X} \times \Theta$. Let $\mathcal{A}$ be a space of actions and let $\tilde{L}$: $\mathcal{A} \times \Theta \to \mathcal{R}^1$ be a loss function which gives the cost $\tilde{L}(a, \theta)$ of θ if action a is selected.

We wish to choose a decision function d: $\mathcal{X} \to \mathcal{A}$ which minimizes the Bayes risk

$$r(d) = \int_{(x, \theta) \in \mathcal{X} \times \Theta} \tilde{P}(dx, d\theta) \tilde{L}(d(x), \theta). \tag{5}$$

The Bayes risk (5) can be expressed as an instance of the general cost (3) as follows. Let $\mathcal{U}$ be the $n - 1$ simplex in $\mathcal{R}^n$ for n the cardinality of Θ; we view $\mathcal{U}$ as the space of probability vectors on Θ. Let P be the marginal distribution of $\tilde{P}$ on $\mathcal{X}$ and let Y: $\mathcal{X} \to \mathcal{U}$ assign to $x \in \mathcal{X}$ the conditional

(posterior) distribution (probability vector) $y = \{y_\theta\}_{\theta \in \Theta} = \{\tilde{P}(\theta|X = x)\}_{\theta \in \Theta}$ on Θ. Finally, let $L: \mathscr{A} \times \mathscr{U} \to \mathscr{R}^1$ be the average cost

$$L(a, y) = \sum_{\theta \in \Theta} y_\theta \tilde{L}(a, \theta).$$

Since this L is linear in y, we can take $L_0 = L$. From (4) and the definition of Y it follows that $E(Y|C)(c) = \tilde{P}(\cdot|C)(c)$ and so the cost $\Psi(C)$ is given by

$$\Psi(C) = E\Phi\big(C(X), \tilde{P}(\cdot|C(X))\big) = \int_{x \in \mathscr{X}} P(dx)\phi\big(C(x), \tilde{P}(\cdot|C(x))\big).$$

Some specific choices of $\mathscr{A}$ and $\tilde{L}$ are:

1. *Error rate*. Let $\mathscr{A} = \Theta$ and let $\tilde{L}(\theta', \theta) = 0$ if $\theta' = \theta$ and 1 otherwise. Then $r(d)$ is the probability of error if we guess that θ equals $d(x)$ when we observe $x \in \mathscr{X}$.
2. *Cross entropy*. Let $\mathscr{A} = \mathscr{U}$ be the space of probability vectors on Θ and let $\tilde{L}(a, \theta) = -\log a_\theta$. Then the cross entropy between y and a is
$$L(a, y) = -\sum_{\theta \in \Theta} y_\theta \log a_\theta.$$
3. *Coding length*. Let $\mathscr{A}$ be the set of valid encodings of Θ and let $\tilde{L}(a, \theta)$ be the length of the codeword for θ in the encoding a. Then $r(d)$ is the expected codeword length. This example is, of course, closely related to the previous one.

If we allow all maps $\tilde{d}: \mathscr{C} \to \mathscr{A}$ (so that $\mathscr{A}_c = \mathscr{A}$ for all c), then it is easy to calculate $\phi: \mathscr{C} \times \mathscr{U} \to \mathscr{R}^1$ in these examples:

1. *Error rate*. $\phi(c, y) = \min_{\theta \in \Theta}(1 - y_\theta)$.
2. *Entropy*. $\phi(c, y) = -\sum_{\theta \in \Theta} y_\theta \log y_\theta$.

These ϕ have well-known interpretations as measures of uncertainty (impurity). Note that in example 2, the cost $\Psi(C)$ is just the familiar conditional entropy of $\{y_\theta\}_{\theta \in \Theta}$ given the class of x.

Classification. In the setup of the previous example, let $\mathscr{A} = \mathscr{C}$ and assume that the only allowed action for class c is $a = c$ (so that $\mathscr{A}_c$ consists of the single element c). Then $\phi = L$ and

$$\Psi(C) = \int_{(x, \theta) \in \mathscr{X} \times \Theta} \tilde{P}(dx, d\theta)\tilde{L}(C(x), \theta)$$

is the usual classification cost, for example, the probability of misclassification in the case of zero–one losses.

Regression. Let $\mathscr{X}$ be a measure space and let $\tilde{P}$ be a probability measure on $\mathscr{X} \times \mathscr{R}^n$. Let

$$\mathscr{A} \subset \big\{(B, \beta) | B \in \mathscr{R}^{n^2}, \beta \in \mathscr{R}^n\big\},$$

where B is arranged as an $n \times n$ matrix.

Let $\tilde{L}$: $\mathscr{A} \times \mathscr{R}^n \to \mathscr{R}^1$ be the following loss function which gives the cost $\tilde{L}(a, w)$ of w if action $a = (B, \beta)$ is selected:

$$\tilde{L}(a, w) = \| B(w - \beta) \|^2 \quad \text{for } w \in \mathscr{R}^n.$$

This is the cost of the affine transformation, represented by $a = (B, \beta)$ on w.

We want to choose a class $C(x)$ for each $x \in \mathscr{X}$, and choose an affine transformation $\tilde{d}(c) = (\tilde{B}(c), \tilde{\beta}(c)) \in \mathscr{A}$ for each class $c \in \mathscr{C}$ so as to minimize the average squared deviation

$$(6) \quad r(\tilde{d}(C)) = \int_{(x,w) \in \mathscr{X} \times \mathscr{R}^n} \tilde{P}(dx, dw) \| \tilde{B}(C(x))(w - \tilde{\beta}(C(x))) \|^2.$$

(Note that this generalizes the clustering example.) Some specific choices of $\mathscr{A}$ are:

1. *Translations.* $\mathscr{A} = \{a = (B, \beta)$: $B =$ Identity$\}$.
2. *Volume preserving maps.* $\mathscr{A} = \{a = (B, \beta)$: det $B = 1\}$, where det is the determinant.

The regression cost (6) can be expressed as an instance of the general cost (3) as follows. Let P be the marginal distribution of $\tilde{P}$ on χ, and let Y: $\mathscr{X} \to \mathscr{U}$ assign to $x \in \mathscr{X}$ the cross-products matrix and the mean vector of the conditional distribution of $\tilde{P}$ given x:

$$Y(x) = (Y_1(x), Y_2(x)) = \left(\int_{w \in \mathscr{R}^n} ww^\dagger \tilde{P}(dw|x), \int_{w \in \mathscr{R}^n} w\tilde{P}(dw|x) \right) \in \mathscr{U},$$

$$\mathscr{U} = \left\{ (U, \mu) | U \in \mathscr{R}^{n^2}, \mu \in \mathscr{R}^n, U - \mu\mu^\dagger \text{ is nonnegative definite} \right\},$$

where $w^\dagger$ is the transpose of w. Then it is easy to see that $r(d)$ is given by (3), with L: $\mathscr{A} \times \mathscr{U} \to \mathscr{R}^1$ given by

$$(7) \quad L(a, y) = \operatorname{tr}\{B \cdot \operatorname{cov}(y) \cdot B^\dagger\} + \| B(y_2 - \beta) \|^2,$$

$$\text{where } \operatorname{cov}(y) = y_1 - y_2 y_2^\dagger,$$

where tr is the trace and $B^\dagger$ is the transpose of B, for $y = (y_1, y_2) \in \mathscr{U}$ and $a = (B, \beta) \in \mathscr{A}$. Since L is linear in y, we can take $L_0 = L$.

For the examples 1 and 2 we can calculate ϕ using (7) and (for example 2) Proposition 3:

1. *Trace.* $\phi(y) = \operatorname{tr} \operatorname{cov}(y)$.
2. *Normalized generalized variance.* $\phi(y) = n \det^{1/n} \operatorname{cov}(y)$ [Anderson (1984) defines $\det \operatorname{cov}(y)$ as the generalized variance of y].

These ϕ has well-known interpretations as measures of uncertainty.

APPENDIX A

We collect here some facts about convexity which were needed in the body of the paper. All our results are simple corollaries of the following separation theorem.

THEOREM 2. *Let $\mathscr{T}$ be a nonempty convex subset of $\mathscr{R}^n$ and let $w \in \mathscr{R}^n \setminus \mathscr{T}$. Then there exists a nonzero $\lambda \in \mathscr{R}^n$ such that $\lambda \cdot t \geq \lambda \cdot w$ for all $t \in \mathscr{T}$.*

For a proof see, for example, Rockafellar (1970), Theorem 11.3.

We begin with a version of Theorem 2 which involves strong inequalities. We first define the following ordering relation on the elements of $\mathscr{R}^m$:

> For any two vectors $u, v \in \mathscr{R}^m$, $u \leq v$ if $u = v$ or if $u_i < v_i$, where i is the first coordinate j for which $u_j \neq v_j$.

PROPOSITION 1. *Let $\mathscr{T}$ be a nonempty convex subset of $\mathscr{R}^n$ and let $w \in \mathscr{R}^n \setminus \mathscr{T}$. Then there exist an integer $m \leq n$ and an $m \times n$ matrix Λ such that $\Lambda t > \Lambda w$ if $t \in \mathscr{T}$.*

PROOF. We use induction on $\dim(\mathscr{T})$, the dimension of $\mathscr{T}[\dim(\mathscr{T}) \leq n]$. The case $\dim(\mathscr{T}) = 1$ is easy since then $\mathscr{T}$ is an interval. For arbitrary $\dim(\mathscr{T})$, we use Theorem 2 to find a nonzero $\lambda \in \mathscr{R}^n$ such that $\lambda \cdot t \geq \lambda \cdot w$ for all $t \in \mathscr{T}$. Let $\mathscr{H} = \{u \in \mathscr{R}^n | \lambda \cdot u = \lambda \cdot w\}$, so that $\dim \mathscr{H} = n - 1$. Then by the inductive hypothesis applied to the convex set $\mathscr{T} \cap \mathscr{H}$ whose dimension is less then n, there exist an $m \leq n - 1$ and an $m \times n - 1$ matrix Λ' such that $\Lambda' t > \Lambda' w$ for all $t \in \mathscr{T} \cap \mathscr{H}$.

Define the $m + 1 \times n$ matrix Λ by $\Lambda = \begin{pmatrix} \lambda \\ \Lambda' \end{pmatrix}$. Then clearly $\Lambda t > \Lambda w$ for all $t \in \mathscr{T}$. □

We next apply Proposition 1 to concave functions.

PROPOSITION 2. *Let $\eta: \mathscr{D} \to \mathscr{R}^1$ be a concave function defined on a convex subset $\mathscr{D}$ of $\mathscr{R}^n$ and let $d_0 \in \mathscr{D}$. Then there exist an integer $m \leq n$ and an $m \times n$ matrix Λ such that $\eta(d) \leq \eta(d_0)$ if $\Lambda d \leq \Lambda d_0$.*

Proposition 2 follows by taking $\mathscr{T} = \{d \in \mathscr{D} | \eta(d) > \eta(d_0)\}$, $w = d_0$ in Proposition 1 (and noting that if $\mathscr{T}$ is empty then Proposition 2 holds trivially).

APPENDIX B

Let T be a nonnegative definite $n \times n$ symmetric matrix. For $\mathscr{F}$ a set of $n \times n$ matrices, let

$$\rho_{\mathscr{F}}(T) = \inf_{U \in \mathscr{F}} \operatorname{tr} U^{\dagger} T U,$$

where tr is the trace and $U^{\dagger}$ is the transpose of U.

PROPOSITION 3. $\rho_{\mathscr{F}}(T) = n \det^{1/n}(T)$ *for* $\mathscr{F} = \{U | \det U = 1\}$.

Here det is the determinant.

PROOF. If $\det U = 1$ and $\{\lambda_i\}$ is a complete set of eigenvalues for the symmetric matrix $U^\dagger TU$, then

$$(8) \qquad \frac{1}{n}\operatorname{tr} U^\dagger TU = \frac{1}{n}\sum_i \lambda_i \geq \left(\prod_i \lambda_i\right)^{1/n} = \det^{1/n} U^\dagger TU = \det^{1/n} T.$$

The middle inequality is the familiar one between arithmetic and geometric means, and the last equality holds since det is multiplicative. On the other hand, since T is nonnegative definite, there exists an $n \times n$ matrix S such that $T = S^\dagger S$. Then if $U = (\det^{1/n} S)S^{-1}$, $\det U = 1$ and

$$(9) \qquad \operatorname{tr} U^\dagger TU = \operatorname{tr}(\det^{2/n} S)I = n \det^{2/n} S = n \det^{1/n} T.$$

The statement follows from (8) and (9). □

Acknowledgments. The authors wish to thank Dr. R. Polyak of the IBM, T. J. Watson Research Center for some helpful discussions about convexity, and an anonymous referee for his careful reading of the paper and for his helpful comments.

REFERENCES

ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. (1984). *Classification and Regression Trees*. Wadsworth, Monterey, Calif.

CHOU, P. (1988). Applications of information theory to pattern recognition and the design of decision trees and trellises. Ph.D. dissertation, Stanford Univ.

COVER, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Comput.* **EC-14** 326–334.

FISHER, W. (1958). On grouping for maximum homogeneity. *J. Amer. Statist. Assoc.* **53** 789–798.

NEMETZ, T. (1967). Information theory and testing of a hypothesis. In *Proceedings of the Colloquium on Information Theory* 283–293. Bolyai Math. Soc., Debrecen, Hungary.

ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton Univ. Press.

D. BURSHTEIN
DEPARTMENT OF ELECTRICAL
ENGINEERING–SYSTEMS
FACULTY OF ENGINEERING
TEL AVIV UNIVERSITY
TEL AVIV 69978
ISRAEL

V. DELLA PIETRA
D. KANEVSKY
A. NÁDAS
IBM, T. J. WATSON RESEARCH CENTER
P.O. BOX 704
YORKTOWN HEIGHTS, NEW YORK, 10598