JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **15**: 625–644 (2000)

ECONOMETRIC APPLICATIONS OF MAXMIN EXPECTED UTILITY

GARY CHAMBERLAIN*
Department of Economics, Harvard University, Cambridge, MA 02138, USA

SUMMARY

Gilboa and Schmeidler (1989) develop a set of axioms for decision making under uncertainty. The axioms imply a utility function and a *set* of distributions such that the preference ordering is obtained by calculating expected utility with respect to each distribution in the set, and then taking the minimum of expected utility over the set. In a portfolio choice problem, the distributions are joint distributions for the data that will be available when the choice is made and for the future returns that will determine the value of the portfolio. The set of distributions could be generated by combining a parametric model with a set of prior distributions. We apply this framework to obtain a preference ordering over decision rules, which map the data into a choice. We seek a decision rule that maximizes the minimum expected utility (or, equivalently, minimizes maximum risk) over the set of distributions. An algorithm is provided for the case of a finite set of distributions. It is based on solving a concave programme to find the least-favourable mixture of these distributions. The minimax rule is a Bayes rule with respect to this least-favourable distribution. The minimax value is a lower bound for minimax risk relative to a larger set of distributions. An upper bound can be found by fixing a decision rule and calculating its maximum risk. We apply the algorithm to an estimation problem in an autoregressive, random-effects model for panel data.

1. INTRODUCTION

Consider an individual making a portfolio choice at date T involving two assets. The (gross) returns at t per unit invested at $t-1$ are y_{1t} and y_{2t}. The individual has observed these returns from $t=0$ to $t=T$. He has also observed the values of the variables $y_{3t}, \ldots, y_{Kt}$, which are thought to be relevant in forecasting future returns. So the information available to him when he makes his portfolio choice is $z \equiv \{(y_{1t}, \ldots, y_{Kt})\}_{t=0}^{T}$. He invests one unit, divided between an amount a in asset one and $1-a$ in asset two, and then holds on to the portfolio until date H. Let $w = \{(y_{1t}, y_{2t})\}_{t=T+1}^{H}$ and let $h(w,a)$ denote the value of the portfolio at $t=H$:

$$h(w,a) = a \prod_{t=T+1}^{H} y_{1t} + (1-a) \prod_{t=T+1}^{H} y_{2t} \tag{1}$$

How should a be chosen?

Consider an econometrician who observes a sample vector z drawn from a distribution P_θ for some value of the parameter θ in the parameter space $\Theta \subset \mathcal{R}^p$. He is interested in a real-valued function $g(\theta)$ and would like an estimator that is optimal under a mean-square error criterion. How should he choose an estimator?

*Correspondence to: Gary Chamberlain, Department of Economics, Harvard University, Cambridge, MA 02138, USA. e-mail: gary_chamberlain@harvard.edu

Received 30 September 1999
Revised 31 August 2000

Gilboa and Schmeidler (1989) develop a set of axioms for decision making under uncertainty. The axioms imply a utility function and a *set* of distributions such that the preference ordering is obtained by calculating expected utility with respect to each distribution in the set, and then taking the minimum of expected utility over the set. In Section 2 we apply this framework to obtain a preference ordering over decision rules, which map the observation z into a choice a. The decision maker's problem is to choose a decision rule that maximizes the minimum expected utility. In the portfolio choice problem, this gives

$$\max_{d\in\mathcal{D}} \min_{Q\in\mathcal{S}} \int u(h(w, d(z)))dQ(z, w)$$

where d is a decision rule, $\mathcal{D}$ is the set of feasible decision rules, and u is the utility function. The value (z,w) is regarded as the realization of a random variable (Z,W) with distribution Q, and $\mathcal{S}$ is the set of distributions.

With risk defined as the negative of expected utility, this framework corresponds to Wald's (1950) minimax risk criterion. In the estimation problem, we could take the set $\mathcal{S}$ to be $\{P_\theta : \theta \in \Theta\}$. This gives

$$\min_{d\in\mathcal{D}} \max_{\theta\in\Theta} \int (g(\theta) - d(z))^2 dP_\theta(z) \tag{2}$$

The set $\mathcal{S}$ of distributions is a key element of this framework. A possible criticism of the criterion in equation (2) is that it puts too much weight on parts of the parameter space that are *a priori* unlikely. A response is to consider a set Γ of prior distributions on Θ. Let $w = \theta$, $dQ_\pi(z,\theta) = dP_\theta(z)d\pi(\theta)$, and $\mathcal{S} = \{Q_\pi : \pi \in \Gamma\}$. The estimation problem becomes:

$$\min_{d\in\mathcal{D}} \max_{\pi\in\Gamma} \int\int (g(\theta) - d(z))^2 dP_\theta(z)d\pi(\theta) \tag{3}$$

This corresponds to Good's (1952) argument that a minimax solution is reasonable provided that only reasonable subjective distributions are entertained. If Γ (and hence $\mathcal{S}$) consists of a single distribution, then the solution to equation (3) is the Bayes rule for that prior distribution, and the framework reduces to Bayesian decision theory.

Note that for any prior distribution π,

$$\int\int (g(\theta) - d(z))^2 dP_\theta(z)d\pi(\theta) \leq \max_{\theta\in\Theta} \int ((g(\theta) - d(z))^2 dP_\theta(z)$$

So the minimax risk value in equation (2) provides an upper bound on equation (3), for any set Γ of prior distributions. If Γ consists of all prior distributions on Θ, including point masses that assign probability one to a single point, then equation (3) reduces to equation (2). Even if the set of all priors is thought to be too big, in that it contains distributions that are not subjectively reasonable, it may still be of interest to calculate the minimax risk value in equation (2). It may be that a Bayes rule d_π for a subjectively reasonable prior π has a maximum risk (over $\theta \in \Theta$) that is close to the minimax value in equation (2). Then that decision rule is attractive in terms of average risk with respect to the prior π and in terms of maximum risk over Θ. In addition, if the set $\mathcal{D}$ is unrestricted, then the posterior risk (with respect to π) is minimized by the choice $d_\pi(z)$ for any value of the observation z.

Section 3 develops an algorithm for computing a minimax decision rule. We consider the case in which $\mathcal{S}$ is the convex hull of a finite set of distributions. For example, in equation (3), we could have Γ equal to the convex hull of $\{\pi_1,\ldots,\pi_J\}$, so that the set of prior distributions consists of mixtures of a finite set of priors. The algorithm is based on Blackwell and Girshick's (1954) minimax theorem for $\mathcal{S}$-games, in which nature has a finite set of pure strategies. The optimal mixed strategy for nature corresponds to a least-favourable distribution, and the minimax rule is a Bayes rule with respect to this least-favourable distribution. This Bayes rule is based on a subjectively reasonable prior to the extent that $\pi_1,\ldots\pi_J$ are subjectively reasonable priors. The least-favourable distribution is obtained numerically by solving a concave programming problem. We use a sequential quadratic programming algorithm.

The minimax risk value that we obtain provides a lower bound on the minimax value relative to a larger set of distributions. For example, suppose that Θ in equation (2) contains an open set of $\mathcal{R}^p$. We can obtain a lower bound for the minimax value in equation (2) by selecting a finite subset $\{\theta_1,\ldots,\theta_J\} \subset \Theta$, and applying our algorithm with $\mathcal{S}$ equal to the convex hull of $\{P_{\theta_1},\ldots,P_{\theta_J}\}$. We can obtain an upper bound on the minimax value in equation (2) by fixing a decision rule and calculating its maximum risk over Θ. For example, we can calculate the maximum risk over Θ of the maximum likelihood estimator.

If the set $\mathcal{D}$ of feasible decision rules is unrestricted, then a Bayes rule can be obtained by minimizing posterior risk. Section 4 develops this case, and also considers a problem in nonparametric estimation in which there are restrictions on the set of feasible rules.

Section 5 considers an autoregressive, random-effects model for panel data. We focus on estimation of the autoregressive parameter, with mean-square error as the risk function. We obtain a lower bound on the minimax risk in equation (2) by using a finite subset $\{\theta_1,\ldots\theta_J\} \subset \Theta$. The maximum risk over Θ of the maximum likelihood estimator provides an upper bound, which turns out to be fairly close to the lower bound.

2. PREFERENCES

Consider an individual making a decision under uncertainty. Suppose that he will observe the value z of a random variable Z before making his choice. The outcome given choice a depends upon a random variable W, whose value w may not be known when the choice is made. $\mathcal{Z} \times \mathcal{W}$ is the range of (Z,W), $\mathcal{A}$ is the set of possible choices, and $\mathcal{X}$ is the set of outcomes.

Let $\mathcal{Y}$ denote the set of probability distributions over $\mathcal{X}$ with finite support, corresponding to lotteries with prizes in $\mathcal{X}$. The probabilities in these lotteries are exogenously given, as in a roulette lottery. Consider y_1 and y_2 in $\mathcal{Y}$, with the union of their supports equal to $\{x_j\}_{j=1}^k$; y_1 assigns probabilities $\{p_j\}_{j=1}^k$ to these outcomes, and y_2 assigns probabilities $\{q_j\}_{j=1}^k$. Then for $\alpha \in (0,1)$, the mixture $\alpha y_1 + (1-\alpha)y_2 \in \mathcal{Y}$ assigns probabilities $\{\alpha p_j + (1-\alpha)q_j\}_{j=1}^k$ to these outcomes.

Let $\mathcal{L}$ denote the set of mappings from $\mathcal{Z} \times \mathcal{W}$ to $\mathcal{Y}$. An element $l \in \mathcal{L}$ can be regarded as a lottery in which the prize corresponding to state (of nature), (z,w) is a roulette lottery. l resembles a horse lottery in that the probabilities of the states are not exogenously given. Let $\mathcal{L}_c$ denote the set of constant functions in $\mathcal{L}$. We shall identify the roulette lotteries $\mathcal{Y}$ with $\mathcal{L}_c$. If $\alpha \in (0,1)$ and $f,g \in \mathcal{L}$, then $\alpha f + (1-\alpha)g$ denotes the horse lottery in $\mathcal{L}$ whose prize in state (z,w) is the roulette lottery in $\mathcal{Y}$ corresponding to the mixture $\alpha f(z,w) + (1-\alpha)g(z,w)$.

Gilboa and Schmeidler (1989) consider a preference relation $\succeq$ over $\mathcal{L}$ that satisfies certain axioms. A key axiom is certainty-independence: for all f, g in $\mathcal{L}$ and r in $\mathcal{L}_c$ and for all $\alpha \in (0,1)$, f $\succ g$ if and only if $\alpha f + (1-\alpha)r \succ \alpha g + (1-\alpha)r$. So the horse lottery f is strictly preferred to the horse lottery g if and only if the (element by element) α-mixture of f with any roulette lottery r is strictly preferred to the corresponding mixture of g with r. Gilboa and Schmeidler show that their axioms are equivalent to the existence of an affine function u: $\mathcal{Y} \rightarrow \mathcal{R}$ and a non-empty, closed, convex set $\mathcal{S}$ of probability measures on $\mathcal{Z} \times \mathcal{W}$ such that: for all $f,g \in \mathcal{L}$,

$$f \succeq g \text{ iff } \min_{Q \in \mathcal{S}} \int u \circ f \, dQ \geq \min_{Q \in \mathcal{S}} \int u \circ g \, dQ.$$

If the certainty-independence axiom is strengthened so that it holds not just for the constant functions but for all r in $\mathcal{L}$, then we have the Anscombe and Aumann (1963) version of the Savage (1954) axioms, and the set $\mathcal{S}$ consists of a single distribution.

A (randomized) decision rule d is a mapping from $\mathcal{Z}$ to $\mathcal{A}^*$, the set of probability distributions on $\mathcal{A}$ with finite support. (We shall identify $\mathcal{A}$ with the subset of $\mathcal{A}^*$ consisting of degenerate distributions). Let $\mathcal{D}$ denote the set of feasible decision rules. The mapping: h: $\mathcal{W} \times \mathcal{A}^* \rightarrow \mathcal{Y}$ determines the outcome distribution as a function of (w,a^*). For example, if a^* assigns probabilities $\{p_j\}_{j=1}^{k}$ to the choices $\{a_j\}_{j=1}^{k}$, then $h(w,a^*)$ is the roulette lottery that assigns probabilities $\{p_j\}_{j=1}^{k}$ to the outcomes $\{h(w,a_j)\}_{j=1}^{k}$. A decision rule $d \in \mathcal{D}$ corresponds to a horse lottery $l_d \in \mathcal{L}$: $l_d(z,w) = h(w,d(z))$.

The preference relation on $\mathcal{L}$ induces a preference relation on the set $\mathcal{D}$ of decision rules. Define the *risk function* as the negative of expected utility:

$$r(Q,d) = -\int_{\mathcal{Z} \times \mathcal{W}} u(l_d(z,w)) dQ(z,w)$$

Then the decision maker's problem is:

$$\min_{d \in \mathcal{D}} \max_{Q \in \mathcal{S}} r(Q,d)$$

The use of risk, and hence a minimax criterion, is traditional, dating back to Wald (1950). We shall not be explicit about measurability and integrability restrictions. Such issues can be avoided by taking the state space $\mathcal{Z} \times \mathcal{W}$ to be a finite set.

The connection of this framework to the portfolio choice problem is quite direct. Z corresponds to the data available when the portfolio is chosen. W is a vector of future returns on the assets, and Q is the joint distribution for (Z,W). The function h is given in equation (1) (for $a \in \mathcal{A}$), and u is a von Neumann–Morgenstern utility function defined over roulette lotteries with monetary prizes. The decision rule d determines the amount a invested in asset one as a function of the data z. The set of feasible rules $\mathcal{D}$ could include all such functions mapping $\mathcal{Z}$ into $\mathcal{A}$. (It could also include all randomized rules, mapping $\mathcal{Z}$ into $\mathcal{A}^*$).

In the estimation problem in equation (3), we set w equal to θ, $u(h(\theta,a)) = -(g(\theta)-a)^2$, $Q_\pi(A \times B) = \int_B P_\theta(A) d\pi(\theta)$, and $\mathcal{S} = \{Q_\pi : \pi \in \Gamma\}$. A decision rule d maps the data z into an estimate a. An example of a restriction on the set $\mathcal{D}$ of feasible rules is an unbiasedness restriction: $\mathcal{D} = \{d$: $\mathcal{Z} \rightarrow \mathcal{R} : \int d(z) dP_\theta(z) = g(\theta), \theta \in \Theta\}$.

An alternative approach for working with a set of priors is to calculate the corresponding set of posterior distributions. Some principle would be needed to make a choice based on this set of posterior distributions. Then we would have a decision rule, and we could ask whether it is optimal under a preference relation that satisfies certain axioms.

3. ALGORITHM

We shall consider a finite set of distributions: $\{Q_1,\ldots,Q_J\}$, and $\mathcal{S}$ is the convex hull:

$$\mathcal{S} = \left\{ \sum_{j=1}^{J} \delta_j Q_j : 0 \le \delta_j \le 1, \sum_{j=1}^{J} \delta_j = 1 \right\} \tag{4}$$

Consider a zero-sum game in which the decision maker chooses $d \in \mathcal{D}$, nature chooses $Q \in \mathcal{S}$, and the payoff to the decision maker is $-r(Q,d)$. The minimax (or upper) value of the game is

$$\overline{V} = \inf_{d \in \mathcal{D}} \sup_{Q \in \mathcal{S}} r(Q, d)$$

A minimax decision rule d_0 satisfies $\sup_{Q \in \mathcal{S}} r(Q,d_0) = \overline{V}$. The maxmin (or lower) value of the game is

$$\underline{V} = \sup_{Q \in \mathcal{S}} \inf_{d \in \mathcal{D}} r(Q, d)$$

A least-favourable distribution Q_0 satisfies $\inf_{d \in \mathcal{D}} r(Q_0, d) = \underline{V}$. A decision rule d_Q is Bayes with respect to the distribution Q if

$$r(Q, d_Q) = \inf_{d \in \mathcal{D}} r(Q, d)$$

A decision rule d generates a vector of risk values $(r(Q_1,d),\ldots,r(Q_J,d))$. The risk set consists of all such vectors as d varies over $\mathcal{D}$:

$$S_r = \left\{ (r(Q_1, d), \ldots, r(Q_J, d)) \in \mathcal{R}^J : d \in \mathcal{D} \right\}$$

We can regard the game as being played as follows. The decision maker chooses a point $s = (s_1,\ldots,s_J) \in S_r$. Independently of his choice, nature chooses a coordinate j with probability δ_j. Blackwell and Girshick (1954, Chapter 2.4) refer to such games, in which nature has a finite number of pure strategies, as S-games. The minimax theorem for S-games states that if the risk set is bounded, then

$$\inf_{d \in \mathcal{D}} \sup_{Q \in \mathcal{S}} r(Q, d) = \sup_{Q \in \mathcal{S}} \inf_{d \in \mathcal{D}} r(Q, d)$$

and there exists a least favourable distribution Q_0. If in addition the risk set is convex and closed, then there exists a minimax decision rule d_0, and it is Bayes with respect to Q_0. We shall assume that the risk set is convex, closed, and bounded. (See Blackwell and Girshick, 1954, Theorem 2.4.2, and Ferguson, 1967, Theorem 1, p. 82. The mixed extension of the game allows the decision maker to use mixed strategies, in which case the risk set is automatically convex since it is the convex hull of S_r; see Blackwell and Girshick, 1954, Theorem 2.4.1.)

Let Σ_J denote the $J-1$ dimensional simplex:

$$\Sigma_J = \left\{ \delta \in \mathcal{R}^J : \delta_j \ge 0, \sum_{j=1}^{J} \delta_j = 1 \right\}$$

and let Q^δ denote the mixture distribution: $Q^\delta = \sum_{j=1}^{J} \delta_j Q_j$. As δ varies over Σ_J, Q^δ varies over $\mathcal{S}$. Note that the risk function is affine in its first argument:

$$r(Q^{\delta}, d) = \sum_{j=1}^{J} \delta_j r(Q_j, d)$$

Let d^{δ} denote the Bayes rule with respect to Q^{δ}. Consider the minimized risk:

$$\rho(\delta) \equiv \min_{d \in \mathcal{D}} r(Q^{\delta}, d) = r(Q^{\delta}, d^{\delta}) \tag{5}$$

Since $r(Q^{\delta},d)$ is an affine function of δ for each d, it follows that ρ is a concave function. So maximizing ρ over the convex set Σ_J is a concave program:

$$\delta_0 = \arg \max_{\delta \in \Sigma_J} \rho(\delta) \tag{6}$$

The least favourable distribution is $Q_0 = \sum_{j=1}^{J} \delta_{0j} Q_j$. The concave program in equation (6) can be solved using a sequential quadratic programming algorithm, as in Wilson (1963). (The routine used in the application in Section 5 is nag_nlp_sol, from the NAG Fortran 90 library; it is based on the subroutine NPSOL described in Gill *et al.*, 1986.)

Let d_0 be a Bayes rule with respect to Q_0. In order for d_0 to be minimax and Q_0 to be least favourable (so δ_0 solves equation (6)), it is necessary and sufficient that

$$r(Q_k, d_0) = \max_{1 \leq j \leq J} r(Q_j, d_0) \quad \text{if} \quad \delta_{0k} > 0, \quad k = 1, \ldots, J \tag{7}$$

So assume that equation (7) holds. Then for any decision rule d,

$$\max_{1 \leq j \leq J} r(Q_j, d) \geq r(Q_0, d) \geq r(Q_0, d_0) = \max_{1 \leq j \leq J} r(Q_j, d_0)$$

So d_0 is minimax. For any $Q \in \mathcal{S}$,

$$r(Q, d_Q) \leq r(Q, d_0) \leq \max_{1 \leq j \leq J} r(Q_j, d_0) = r(Q_0, d_0)$$

So Q_0 is least favourable. This sufficiency argument does not use the minimax theorem.

To see that equation (7) is necessary, assume that d_0 is minimax and Q_0 is least favourable:

$$\max_{1 \leq j \leq J} r(Q_j, d_0) = \inf_{d \in \mathcal{D}} \max_{1 \leq j \leq J} r(Q_j, d) = \overline{V}$$

$$\inf_{d \in \mathcal{D}} r(Q_0, d) = \sup_{Q \in \mathcal{S}} \inf_{d \in \mathcal{D}} r(Q, d) = \underline{V}$$

Since d_0 is a Bayes rule with respect to Q_0, the conclusion of the minimax theorem implies that

$$\sum_{j=1}^{J} \delta_{0j} r(Q_j, d_0) = r(Q_0, d_0) = \inf_{d \in \mathcal{D}} r(Q_0, d) = \underline{V} = \overline{V} = \max_{1 \leq j \leq J} r(Q_j, d_0)$$

So the maximum risk of d_0 equals the average risk under Q_0, which implies equation (7).

In the numerical algorithm for the concave program, we use a subgradient of ρ. It is convenient to solve for $\delta_J = 1 - \sum_{j=1}^{J-1} \delta_j$ and regard ρ as defined on a subset of $\mathcal{R}^{J-1}$: $M_{J-1} = \{\delta \in \mathcal{R}^{J-1}\colon\ \delta_j \geq 0, \sum_{j=1}^{J-1} \delta_j \leq 1\}$. Let $\zeta_\delta \in \mathcal{R}^{J-1}$ have jth component equal to $r(Q_j,d^{\delta}) - r(Q_J,d^{\delta})$. We will show that ζ_δ is a subgradient of ρ at δ. Note that

$$\rho(\delta) = \langle \zeta_\delta, \delta \rangle + r(Q_J, d^{\delta})$$

where $\langle a,b \rangle$ denotes $\sum_{i=1}^{k} a_i b_i$ for $a,b \in \mathcal{R}^k$. For any $\delta' \in M_{J-1}$,

$$\begin{aligned}\rho(\delta') &= \min_{d\in\mathcal{D}}\left(\sum_{j=1}^{J-1}\delta'_j\big(r(Q_j,d)-r(Q_J,d)\big)+r(Q_J,d)\right)\\ &\leq \sum_{j=1}^{j-1}\delta'_j\big(r(Q_j,d^{\delta})-r(Q_J,d^{\delta})\big)+r(Q_J,d^{\delta})\\ &= \langle\zeta_{\delta},\delta'\rangle + r(Q_J,d^{\delta}) = \langle\zeta_{\delta},\delta\rangle + \langle\zeta_{\delta},\delta'-\delta\rangle + r(Q_J,d^{\delta})\\ &= \rho(\delta) + \langle\zeta_{\delta},\delta'-\delta\rangle\end{aligned}$$

and so ζ_{δ} is a subgradient.

3.1 Minimax Bounds

The minimax value $r(Q_0,d_0)$ is with respect to the set $\mathcal{S}$ of distributions. If we consider a larger set of distributions $\mathcal{S}' \supset \mathcal{S}$, then

$$\overline{V} = \inf_{d\in\mathcal{D}}\sup_{Q\in\mathcal{S}} r(Q,d) \leq \inf_{d\in\mathcal{D}}\sup_{Q\in\mathcal{S}'} r(Q,d) = \overline{V}'$$

So the minimax value relative to $\mathcal{S}$ provides a lower bound for the minimax value relative to the larger set $\mathcal{S}'$.

Now fix a decision rule d, and construct an upper bound:

$$\overline{V}' \leq \sup_{Q\in\mathcal{S}'} r(Q,d)$$

This upper bound is useful in that it may be feasible to maximize $r(Q,d)$ over $Q \in \mathcal{S}'$ for a fixed d, even though it is not feasible to compute the minimax value for $\mathcal{S}'$.

Kempthorne (1987) develops an algorithm for the case in which the least-favourable prior distribution is known to have finite support, but the location of the support points is not known. Suppose, for example, in equation (2) that Θ is a closed interval on the real line; the risk function is an analytic function of θ for any decision rule; and the risk of the Bayes procedure for the least-favourable prior distribution is not constant on Θ. Then the least-favourable prior distribution has finite support, and the algorithm converges to this distribution. The algorithm constructs a sequence of discrete prior distributions whose successive support sets change by adjusting the locations of the existing support points as well as adding new points to the support. For a given number of support points, the algorithm finds a local maximum of the Bayes risk of the Bayes rule, maximizing over the locations of the support points as well as the probabilities attached to the support points. This is not a concave program. The optimisation is done using an unconstrained maximization procedure with differences to approximate derivatives.

4. THE SET $\mathcal{D}$ OF DECISION RULES

A step in our algorithm requires finding the Bayes rule d^{δ}, which maximizes the risk $r(Q^{\delta},d)$ over the set $\mathcal{D}$ of feasible decision rules. Consider first the case in which $\mathcal{D}$ is unrestricted.

4.1 $\mathcal{D}$ Is Unrestricted

In this case, we can obtain the Bayes rule by minimizing posterior risk — see Wald (1950), Chap. 5.1), Blackwell and Girshick (1954, Chap. 7.3), and Ferguson (1967, Chap. 1.8). The distribution Q for (Z,W) can be decomposed into Q_m, which is the marginal distribution for Z, and Q_c, which is the conditional distribution for W given Z: $Q(A \times B) = \int_A Q_c(B|z)dQ_m(z)$. Let f_j be the density of Q_{jm} with respect to the measure μ: $Q_{jm}(A) = \int_A f_j(z)d\mu(z)$ $(j = 1, \ldots, J)$. Define the *loss function* as the negative of conditional expected utility given $Z = z$ for the choice a:

$$L(Q, z, a) = -\int_{\mathcal{W}} u(h(w, a))dQ_c(w \mid z) \tag{8}$$

Then we have

$$\begin{aligned} r(Q^{\delta}, d) = \sum_{j=1}^{J} \delta_j r(Q_j, d) &= \int_{\mathcal{Z}} \left[\sum_{j=1}^{J} L(Q_j, z, d(z)) f_j(z) \delta_j \right] d\mu(z) \\ &\geq \int_{\mathcal{Z}} \left[\inf_{a \in \mathcal{A}} \sum_{j=1}^{J} L(Q_j, z, a) f_j(z) \delta_j \right] d\mu(z) \end{aligned}$$

We shall assume that the infimum is in fact obtained for some choice $a \in \mathcal{A}$. We can regard δ as providing prior probabilities on the discrete parameter space $\{1, \ldots, J\}$, and calculate the posterior probabilities as

$$\bar{\delta}_j(z) = f_j(z)\delta_j \bigg/ \sum_{k=1}^{J} f_k(z)\delta_k \tag{9}$$

Then a Bayes rule d^{δ} for the prior δ can be obtained by minimizing posterior risk, which equals posterior expected loss:

$$d^{\delta}(z) = \arg\min_{a \in \mathcal{A}} \sum_{j=1}^{J} L(Q_j, z, a)\bar{\delta}_j(z) \tag{10}$$

Mixture models

Consider a mixture model in which the distribution Q for the vector (Z,W) has the following form:

$$Q_{\pi}(A \times B) = \int_{\Theta} P_{\theta}(A \times B)d\pi(\theta)$$

We start with a parameter space Θ, and $\{P_{\theta} : \theta \in \Theta\}$ is a set of distributions for (Z,W). We introduce a family Γ of prior distributions π on Θ. This gives a set $\{Q_{\pi} : \pi \in \Gamma\}$ of distributions for (Z,W), in which the prior π plays the role of a parameter. Consider a finite set of prior distributions: $\{\pi_1, \ldots, \pi_J\}$, and Γ is the convex hull:

$$\Gamma = \left\{ \sum_{j=1}^{J} \delta_j \pi_j : 0 \leq \delta_j \leq 1, \sum_{j=1}^{J} \delta_j = 1 \right\}$$

Let $Q_j = Q_{\pi_j}$. Then the set $\mathcal{S} = \{Q_\pi : \pi \in \Gamma\}$ is the convex hull of $\{Q_1, \ldots, Q_J\}$, as in equation (4), which is the basis for our algorithm in Section 3.

The probability distribution P_θ can be decomposed into a marginal distribution $P_{\theta m}$ and a conditional distribution $P_{\theta c} : P_\theta(A \times B) = \int_A P_{\theta c}(B|z)\, dP_{\theta m}(z)$. We shall assume that $P_{\theta m}$ has density $f(z|\theta)$ with respect to the measure μ: $P_{\theta m}(A) = \int_A f(z|\theta) d\mu(z)$ for all $\theta \in \Theta$. Since $\mathrm{Q}_{jm} = \int P_{\theta m} d\pi_j(\theta)$, the density f_j for Q_{jm} is given by

$$f_j(z) = \int_\Theta f(z \mid \theta) d\pi_j(\theta) \tag{11}$$

Let $\bar{\pi}_j$ denote the posterior distribution of θ conditional on Z:

$$\bar{\pi}_j(B \mid z) = \left[f_j(z)\right]^{-1} \int_B f(z \mid \theta) d\pi_j(\theta)$$

Then the loss function in equation (8) can be obtained by integrating the loss for P_θ with respect to the posterior distribution of θ:

$$L(Q_j, z, a) = \int_\Theta L(P_\theta, z, a) d\bar{\pi}_j(\theta \mid z) \tag{12}$$

These formulas for the marginal density f_j and the loss $L(Q_j,z,a)$ in equations (11) and (12) are useful for calculating the posterior risk, as in equations (9) and (10).

In the portfolio choice problem, Z corresponds to the data available when the portfolio is chosen, and W is a vector of future returns on the assets. The specification of the family $\{P_\theta : \theta \in \Theta\}$ might be based on a vector autoregression with multivariate normal innovations, and Γ would be a family of prior distributions for the parameters of the vector autoregression. Barberis (2000) uses such a specification with a single prior distribution. In the portfolio choice problem, the focus is not on the parameter vector θ; the role of the parametric model is to generate a joint distribution for the observables Z and W.

Now consider an estimation problem. Here the focus is on a function of the parameter, which we shall denote by $g(\theta)$. In this case we set w equal to θ, and the choice a is the estimate of $g(\theta)$. If g is real valued, the loss function could be $L(P_\theta,z,a) = (g(\theta) - a)^2$, with mean-square error for the risk function. Or the loss function could have a piecewise linear form:

$$L(P_\theta, z, a) = \begin{cases} c_1 \mid g(\theta) - a \mid & \text{if } a \leq g(\theta) \\ c_2 \mid g(\theta) - a \mid & \text{otherwise} \end{cases}$$

with $c_1, c_2 > 0$. Then choosing $c_1/(c_1 + c_2) = 0.025$ and 0.975 could give estimates corresponding to a traditional confidence interval.

4.2 $\mathcal{D}$ is Restricted

We may be interested in a restricted class of decision rules. Consider, for example, the mixture model, $\{\int_\Theta P_\theta\, d\pi(\theta) : \pi \in \Gamma\}$, where the prior distributions in Γ are indexed by a parameter $\tau \in \mathcal{R}^m$: $\Gamma = \{\pi_\tau : \tau \in \mathcal{R}^m\}$. For a given value of τ, there is a decision rule d_τ that minimizes posterior risk:

$$d_\tau(z) = \arg\min_{a \in \mathcal{A}} \int_\Theta L(P_\theta, z, a) d\bar{\pi}_\tau(\theta \mid z)$$

where $\bar{\pi}_\tau$ is the posterior distribution corresponding to the prior π_τ. Suppose that we set $Q_j = \int P_\theta \, d\pi_{\tau_j}$ $(j = 1, \ldots, J)$, and as a step in our minimax algorithm, minimize the risk $r(Q^\delta, d) = \sum_{j=1}^{J} \delta_j r(Q_j, d)$. The unrestricted Bayes rule d^δ in equation (10) minimizes posterior risk for the prior distribution $\sum_{j=1}^{J} \delta_j \pi_{\tau_j}$. This distribution is not, in general, in Γ unless Γ is convex. So if Γ is not convex, the unrestricted Bayes rule will generally not belong to $\{d_\tau : \tau \in \mathcal{R}^m\}$. Or consider setting $Q_j = P_{\theta_j}$ with $\{\theta_1, \ldots, \theta_J\} \subset \Theta$. Then the Q_j will not generally be in the mixture model unless Γ includes all point masses on Θ. Nevertheless, we may want to restrict the set of decision rules to $\mathcal{D} = \{d_\tau : \tau \in \mathcal{R}^m\}$ because the mixture model with the family Γ of prior distributions is tractable or familiar. Then we need to solve

$$\min_{\tau \in \mathcal{R}^m} r(Q^\delta, d_\tau)$$

in order to obtain the minimized risk $\rho(\delta)$ in equation (5).

Another application could arise in non-parametric estimation. Based on asymptotic theory or other arguments, we may be interested in a class of estimators, such as orthogonal series estimators. Obtaining a particular estimator within the class requires a value for a vector of parameters τ, which may govern how much smoothness is imposed on the estimator. Efromovich (1999, Chap. 3) develops a data-driven orthogonal series estimator for a univariate density based on a random sample of size n. The procedure for determining which terms appear in the series requires setting a value for τ. For example, terms in the series beyond a certain cutoff point are dropped if a test statistic does not exceed a threshold that depends upon τ. Default settings are suggested for τ. To evaluate the performance of the estimator, Efromovich constructs eight 'corner' densities that represent features of interest that are expected to occur in practice. Monte Carlo simulation is used with data sets generated according to each of the corner densities. The risk measure is expected integrated squared error. Efromovich notes that the optimal choice of the parameter τ involves a tradeoff between better estimation for some of the corner densities and worse estimation for others. Within our framework, a way to make an optimal tradeoff is to let $\{Q_1, \ldots, Q_J\}$ be the distributions for the sample corresponding to the corner densities, and choose the parameter τ by solving

$$\min_{\tau \in \mathcal{R}^m} \max_{j \in \{1, \ldots, J\}} r(Q_j, d_\tau) \tag{13}$$

Let $g_j(\tau) = r(Q_j, d_\tau)$ and suppose that $g_j : \mathcal{R}^m \to \mathcal{R}$ is continuously differentiable with gradient ∇g_j. Applying our algorithm in Section 3 to equation (13) gives $\tau_0 \in \mathcal{R}^m$ and $\delta_0 \in \Sigma_J$ such that

$$\sum_{j=1}^{J} \delta_{0j} \nabla g_j(\tau_0) = 0$$

and

$$g_k(\tau_0) = \max_{j \in \{1, \ldots, J\}} g_j(\tau_0) \quad \text{if} \quad \delta_{0k} > 0, \quad k = 1, \ldots, J$$

An alternative algorithm for equation (13) is developed in Polak (1997, Chap. 2.4). For a given value of τ, solve the following quadratic program:

$$\delta^\tau = \arg\min_{\delta \in \Sigma_J} \left(\sum_{j=1}^{J} \delta_j \left[\Psi(\tau) - g_j(\tau) \right] + \frac{1}{2} \| \sum_{j=1}^{J} \delta_j \nabla g_j(\tau) \|^2 \right) \tag{14}$$

where $\Psi(\tau) = \max_{j \in \{1,\dots,J\}} g_j(\tau)$ and $||a||^2 = \langle a,a \rangle$. Obtain the search direction

$$h = -\sum_{j=1}^{J} \delta_j^\tau \nabla g_j(\tau)$$

A new value for τ is chosen according to $\tau^* = \tau + \lambda h$, where the scalar λ is determined by a step-size algorithm. Then replace τ by τ^* and repeat until the minimum value of the quadratic program in equation (14) is zero. Polak (1997, Chap. 2) develops additional algorithms for equation (13) and provides references to related work.

5. APPLICATION: AUTOREGRESSIVE MODELS FOR PANEL DATA

We will work with the following parametric family:

$$\begin{aligned} &Z_{it} = \gamma Z_{i,t-1} + \alpha_i + U_{it} \\ &\alpha_i \mid \{Z_{i0} = z_{i0}\}_{i=1}^{N} \overset{\text{ind}}{\sim} \mathcal{N}(\tau_1 + \tau_2 z_{i0}, \sigma_\nu^2) \\ &U_{it} \mid \{\alpha_i, Z_{i0} = z_{i0}\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (i = 1, \dots, N;\ t = 1, \dots, T) \end{aligned} \tag{15}$$

The parameter vector is $\zeta = (\theta,\psi)$, with $\theta = (\gamma,\lambda)$, $\psi = (\tau_1,\tau_2,\sigma)$, and $\lambda \equiv \sigma_\nu/\sigma$. We obtain a family of distributions $\{P_\theta : \theta \in \Theta\}$ by specifying a single prior distribution for ψ and integrating ψ out of the model. The prior distribution for ψ is motivated by work in Chamberlain and Hirano (1999) using residuals from regressions of log earnings on education and age in the Panel Study of Income Dynamics. It specifies that $1/\sigma^2 \sim \mathcal{X}^2(10)/0.9$, so that the 0.1 and 0.9 quantiles for σ are 0.24 and 0.43. Conditional on σ, the components of (τ_1,τ_2) are independent normals with variances proportional to σ^2. The mean of τ_1 is 0, the mean of τ_2 is 0.25, and the standard deviations of τ_1 and τ_2 in the (unconditional) t-distribution are 0.20.

We obtain the same P_θ family if we start out with $\alpha = (\alpha_1,\dots,\alpha_N)$ as part of the parameter vector: $\zeta = (\theta,\psi,\alpha)$. The random effects model in equation (15) provides a prior distribution for α given (θ,ψ). Combining this with our prior distribution for ψ gives the P_θ distribution.

The observation is $Z = (Z_{11},\dots,Z_{1T},\dots,Z_{N1},\dots,Z_{NT})'$. The P_θ distribution for Z is conditional on $\{z_{i0}\}_{i=1}^{N}$, which is observed. The values for $\{z_{i0}\}_{i=1}^{N}$ in our risk calculations are obtained by drawing from a normal distribution with mean 0 and standard deviation 0.45; these values for z_{i0} are then kept fixed in evaluating risk. The density (for $\lambda > 0$) is

$$f(z \mid \theta) = c(z)\det(H)^{1/2}\det(\bar{H})^{-1/2}(m'Hm - \bar{m}'\bar{H}\bar{m} + z^{*\prime}z^* + b_2)^{-(NT+b_1)/2}$$

where $1/\sigma^2 \sim \text{Gamma}\,(b_1/2,2,/b_2)$, $(\tau_1,\tau_2)\,|\,\sigma \sim \mathcal{N}(m_1,\sigma^2 B^{-1})$,

$$H = \begin{pmatrix} B & 0 \\ 0 & \lambda^{-2} I_N \end{pmatrix} \quad m = \begin{pmatrix} m_1 \\ 0 \end{pmatrix} \quad \bar{H} = X'X + H \quad \bar{m} = \bar{H}^{-1}(X'z^* + Hm)$$

$X = (R\ \ I_N \otimes l_T)$, l_T is a $T \times 1$ vector of ones, $z^* = (z_{11}^*,\dots,z_{1T}^*,\dots,z_{N1}^*,\dots,z_{NT}^*)'$ with $z_{it}^* = z_{it} - \gamma z_{i,t-1}$, R is a $NT \times 2$ matrix with the row corresponding to z_{it}^* equal to $(1\ z_{i0})$, and $c(z)$ is some function of z that does not depend upon θ. It is useful to simplify $f(z\,|\,\theta)$, since it is evaluated repeatedly for different values of θ. The computational simplification is similar to the one described in Chamberlain (2000).

We will focus on the estimation of γ, using a squared-error loss function: $L(P_\theta,z,a) = (\gamma - a)^2$. We set $W = \theta$, since θ combined with a choice (an estimate) determines the utility-relevant outcome. We shall work with a finite subset $\{\theta_1,\ldots,\theta_J\}$ of Θ, with $Q_j = P_{\theta_j}$ and $\mathcal{S}$ equal to the convex hull of $\{Q_1,\ldots,Q_J\}$. (This corresponds to a mixture model in which the prior distribution π_j assigns unit mass to the point θ_j.) We let the set of decision rules $\mathcal{D}$ be unrestricted, so that the Bayes rule d^δ is the posterior mean of γ:

$$d^{\delta}(z) = \sum_{j=1}^{J} \gamma(\theta_j) f(z \mid \theta_j)\delta_j \Bigg/ \sum_{j=1}^{J} f(z \mid \theta_j)\delta_j \tag{16}$$

where $\gamma(\theta)$ denotes the first component of θ. This follows from equations (9) and (10), with $f_j(z) = f(z \mid \theta_j)$ (as in equation (11) with π_j assigning unit mass to the point θ_j). The risk under Q^δ for an estimator d is

$$r(Q^{\delta}, d) = \sum_{j=1}^{J} \delta_j r(P_{\theta_j}, d) \tag{17}$$

where $r(P_\theta,d) = \int_{\mathcal{Z}} [\gamma(\theta) - d(z)]^2 f(z \mid \theta) dz$.

We can approximate $r(P_\theta,d)$ by Monte Carlo simulation. Obtain independent and identically distributed draws $\{Z(\theta,k)\}_{k=1}^{K}$ from $f(\cdot \mid \theta)$. Then we have

$$r(P_\theta, d) \cong \frac{1}{K}\sum_{k=1}^{K} [\gamma(\theta) - d(Z(\theta,k))]^2 \tag{18}$$

We use the same set of pseudo-random numbers to construct $\{Z(\theta,k)\}_{k=1}^{K}$ for each θ. This ensures that the simulated $r(\theta,d)$ varies smoothly in θ. Then we calculate $\rho(\delta) = r(\mathrm{Q}^{\delta},d^{\delta})$ from equations (16) and (17). A numerical optimisation routine is used for the constrained maximization of ρ over the $J-1$ dimensional simplex. (The routine is nag_nlp_sol, from the NAG Fortran 90 library). The maximizing value δ_0 gives the least-favourable prior, and $\rho(\delta_0)$ is the minimax value for risk, relative to the set $\{\theta_1,\ldots,\theta_J\}$.

Consider the case $N = 100$ and $T = 2$. Preliminary work indicated that most of the mass in the least-favourable prior is concentrated in the rectangle with $0 \le \gamma \le 1.4$ and $0 \le \lambda \le 1.4$. Then I set up a grid with 15 values for γ: $0, 0.1, \ldots, 1.4$, and 8 values for λ: $10^{-4}, 0.2, 0.4, \ldots, 1.4$, giving $J = 120$ values for θ. The solution to the concave program gives a minimax value for root mean-square error (MSE) of $\rho(\delta_0)^{1/2} = 0.115$. (The Monte Carlo simulation uses $K = 8000$ samples.) The least-favourable prior assigns 0 probability to almost 60% of the points: $\delta_{0j} = 0$ for 69 points, $0 \le \delta_{0j} \le 10^{-6}$ for 5 points, and $\delta_{0j} > 10^{-6}$ for 46 points. The maximum root MSE of the minimax estimator over the 120 points is equal to the minimax value: $\max_j r(P_{\theta_j},d_0)^{1/2} = 0.115$. The root MSE is equalized at 0.115 across the 46 points that are assigned probability greater than 10^{-6}. (The variation is in the fourth decimal place, between 0.1149 and 0.1151.) So the solution satisfies the necessary and sufficient condition in equation (7) very well.

The upper panels of Figure 1 show the root MSE at the 15 values of γ for $\lambda = 10^{-4}$, 0.6, and 1.2. The lower panels show the least-favourable prior probabilities, δ_{0j}. We see that the root MSE is equalized at the minimax value of 0.115 for the points that receive positive probability. The support of the conditional distribution of γ given λ is fairly concentrated, and it shifts to the left as λ increases. This negative correlation between γ and λ under δ_0 is visible in Figure 2,

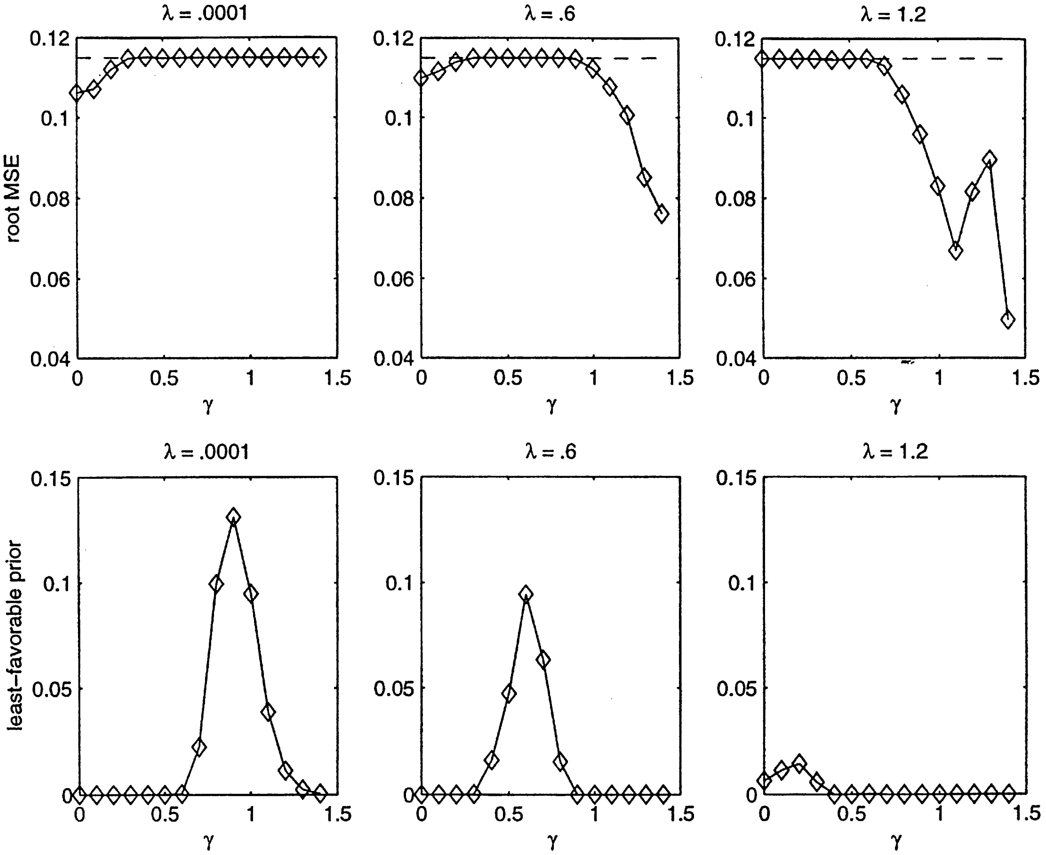

Figure 1. Root mean-square error for minimax estimator (upper panel): $r(\theta,d_0)^{1/2}$; the dashed line indicates the minimax value of 0.115. Least-favourable prior probabilities (lower panel): δ_0. Evaluated at $\theta=(\gamma,\lambda)\in\{\theta_1,\dots,\theta_J\}$ for $\gamma=\{0,.1,\dots,1.4\}$ and for λ as shown. $N=100$ and $T=2$

which shows the joint distribution. The negative correlation is also visible in Figure 3, which shows equal probability contours of the joint distribution. The support of the joint distribution is fairly concentrated along a negatively sloped diagonal. We may have a precise estimate of a positive covariance between Z_{i2} and Z_{i1}, but different values for the (γ,λ) pair can imply the same covariance, with γ decreasing as λ increases. So knowing that (γ,λ) is distributed along a negatively sloped diagonal in the first quadrant does not help in the difficult choice of a point on that diagonal. (Figs. 2 and 3 use bicubic spline interpolation to the probability values at the 120 points, with negative values in the interpolating spline set to 0.)

Consider the minimax value corresponding to the entire parameter space:

$$\overline{V}_\Theta = \inf_d \sup_{\theta\in\Theta} r(\theta,d)$$

The minimax value relative to $\{\theta_1,\dots,\theta_J\}$ provides a lower bound on $\overline{V}_\Theta$. We can obtain an

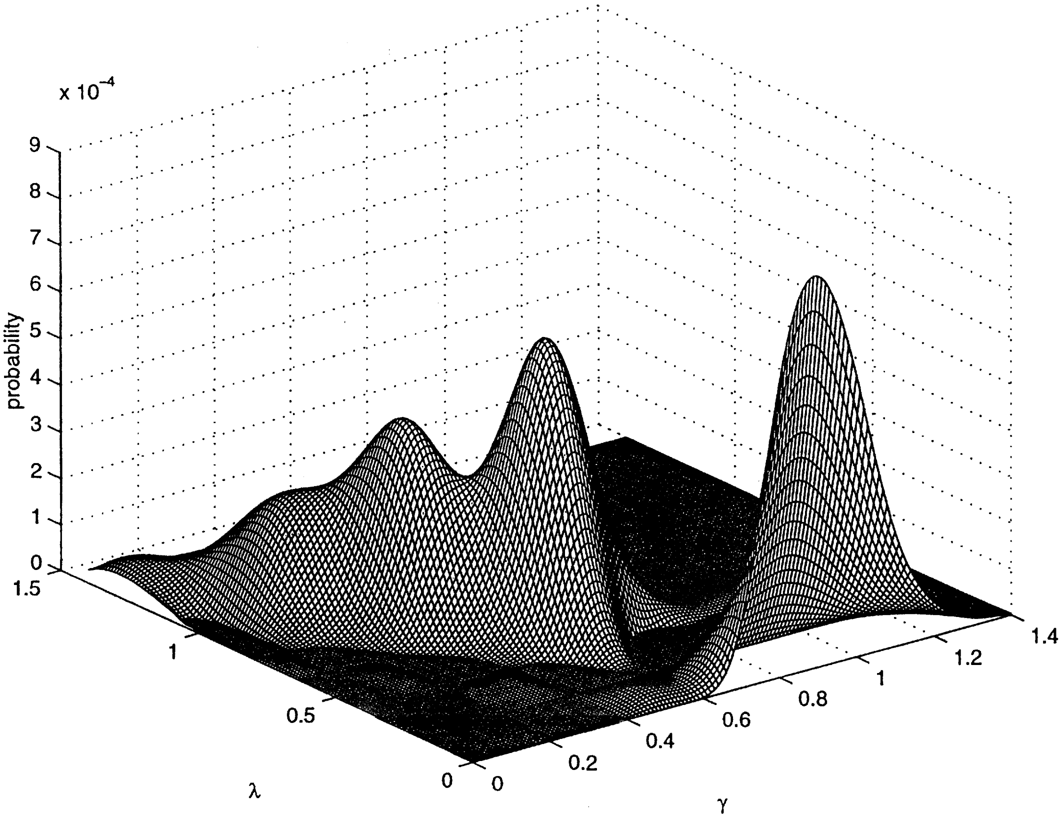

Figure 2. Least-favourable prior probabilities δ_0; interpolation. $N = 100$ and $T = 2$

upper bound by calculating the maximum risk over Θ of the maximum likelihood (ML) estimator, which solves $\max_{\theta \in \Theta} f(z \,|\, \theta)$. The risk at a given θ is calculated using equation (18), where now θ is not restricted to the finite set of J values. The risk function for the ML estimator is smooth and unimodal. With $\Theta = \mathcal{R} \times [10^{-4},\infty)$, the maximum value for root MSE is 0.132, which is attained at $(\gamma,\lambda) = (0.903, 0.007)$. So we can bound the minimax value for the whole parameter space between 0.115 and 0.132. The maximum likelihood estimator is attractive in terms of risk, with a maximum root MSE that is quite close to the lower bound.

Now consider using the minimax estimator d_0 based on $\{\theta_1, \ldots ,\theta_J\}$ to provide an upper bound on $\overline{V}_\Theta$. First we shall examine the maximum risk of d_0 for (γ,λ) in the rectangle $[0,1.4] \times [10^{-4},1.4]$. This will indicate how well the minimax value for an infinite, bounded set can be approximated by the minimax value for a finite set. The maximum risk is obtained using a grid search combined with a numerical optimisation routine. The grid has 0.01 increments for γ and 0.05 increments for λ. The maximum value for root MSE is 0.117, which is attained at $(\gamma,\lambda) = (0.913, 0.712)$. So the minimax value for the rectangle is very tightly bounded between 0.115 and 0.117.

Figure 4 compares the risk functions for the ML and minimax estimators, for $0 \le \gamma \le 1.4$ and

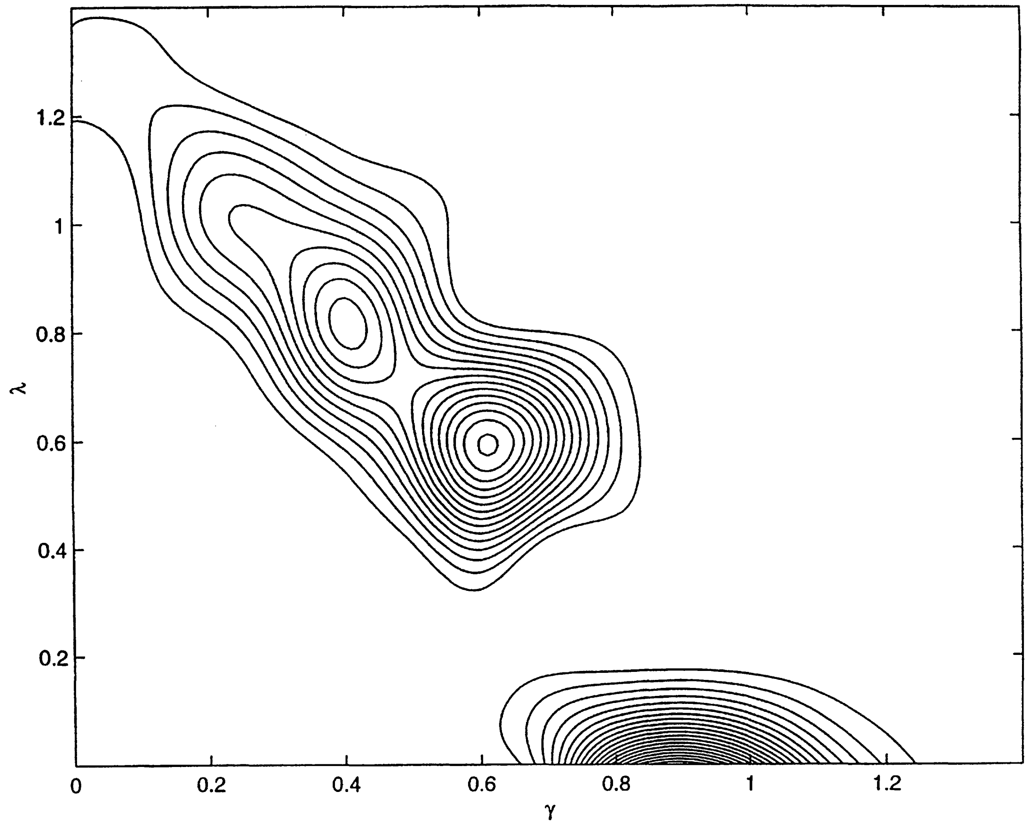

Figure 3. Equal probability contours of the least-favourable prior δ_0; interpolation. The heights of the contours are at equal probability increments: Δ, 2Δ, 3Δ, ... $N = 100$ and $T = 2$

$\lambda = 10^{-4}$, 0.4, 0.8, and 1.2. The risk of the minimax estimator is evaluated at 0.01 increments for γ. The minimax estimator d_0 based on the 120 points is close to being a minimax estimator for the rectangle. However, for values of λ greater than 1.4, there are points where the root MSE for d_0 exceeds the maximum value for ML of 0.132. This is so even though adding such points to $\{\theta_1, \ldots, \theta_J\}$ has very little effect on the minimax value, which remains close to 0.115.

5.1 Local Parameter Space: A Connection with Asymptotic Theory

A key idea in asymptotic statistics, due to Le Cam (1986), is that a sequence of statistical experiments (or models) can be approximated by a limit experiment. The observation Z has i.i.d. components $(Z_1, \ldots, Z_n)$. The joint distribution of Z is P_θ^n for some value of the parameter θ in the parameter space Θ, which is an open subset of $\mathcal{R}^p$. The approximation involves a local parameter $h = \sqrt{n}\,(\theta - \theta_0)$, where θ_0 is a fixed point in the interior of Θ and is regarded as known. Under certain conditions, for large n the experiments

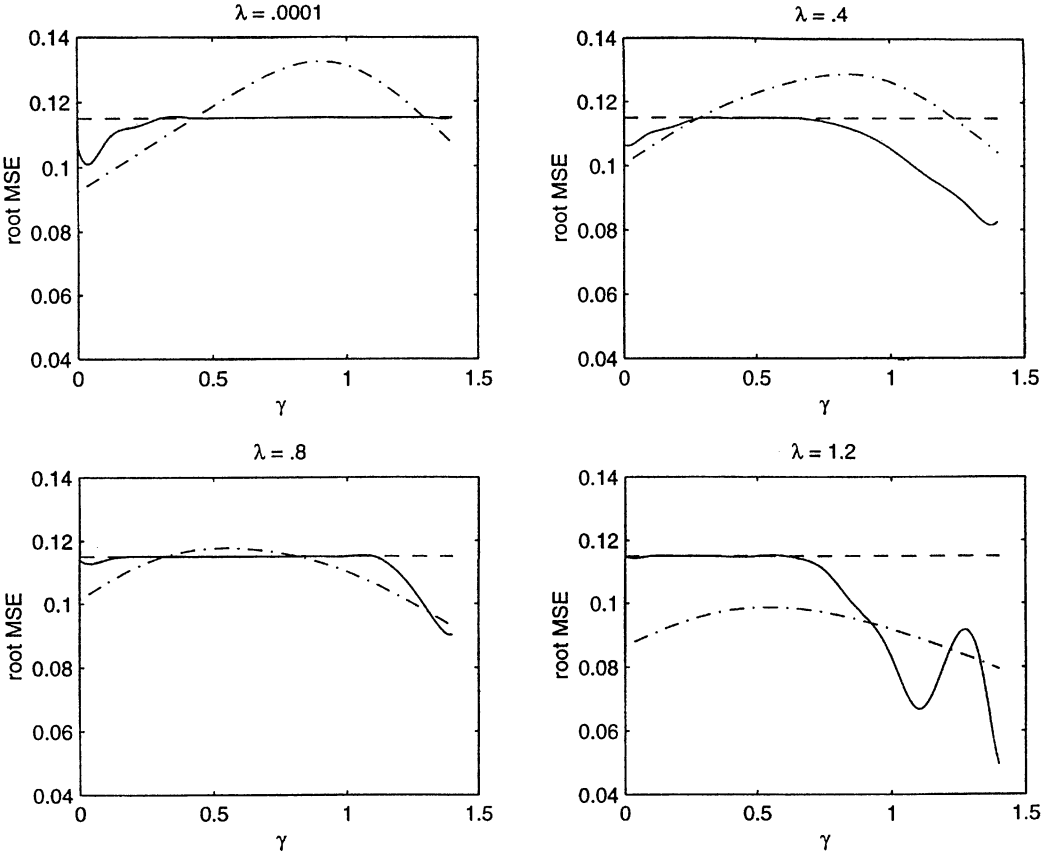

Figure 4. Root mean-square error for minimax estimator (solid line): $r(\theta,d_0)^{1/2}$, evaluated for $\gamma \in [0,1.4]$ in 0.01 increments, and for λ as shown; the dashed line indicates the minimax value of 0.115. Root mean-square error for maximum likelihood estimator (dash dot line), evaluated for $\gamma \in [0,1.4]$ in 0.05 increments, and for λ as shown. $N = 100$ and $T = 2$

$$\left\{P^n_{\theta_0+h/\sqrt{n}} : h \in \mathcal{R}^p\right\} \quad \text{and} \quad \left\{\mathcal{N}(h, I^{-1}_{\theta_0}) : h \in \mathcal{R}^p\right\}$$

have similar statistical properties. Here the limit experiment consists of observing a single observation from a normal distribution with mean h and variance matrix equal to the inverse of the Fisher information matrix. See van der Vaart (1998, Chap. 7) for an exposition. He observes that (p. 97): 'A motivation for studying a local approximation is that, usually, asymptotically, the "true" parameter can be known with unlimited precision. The true statistical difficulty is therefore determined by the nature of the measures P_θ in a small neighbourhood of the true value. In the present situation "small" turns out to be "of size $O(1/\sqrt{n})$".' One version of optimality in this framework is provided by the local asymptotic minimax theorem. It gives a lower bound for the maximum risk over a small neighbourhood of θ_0 (van der Vaart, Chap. 8.7).

This suggests examining the maximum risk of an estimator, such as maximum likelihood, not over all of Θ but over a neighbourhood around some point, such as the maximum likelihood estimate. Likewise, we can consider the minimax value and minimax estimator corresponding to this neighbourhood.

For an example, I will use data on log earnings residuals for a sample of high school graduates from the PSID. The data are described in Chamberlain and Hirano (1999) and have $N = 100$, $T = 9$. (As above, we condition on the $t = 0$ observation.) The maximum likelihood estimates are $\hat{\gamma}_{\mathrm{ML}} = 0.426$ and $\hat{\lambda}_{\mathrm{ML}} = 0.508$. Profile likelihood intervals at an approximate 0.99 confidence level are [0.327, 0.532] for γ and [0.329, 0.708] for λ. I will set the local parameter space equal to the rectangle [0.32, 0.54] × [0.32, 0.71].

For the minimax analysis over this rectangle, I set up a grid with 10 equally spaced values for γ from 0.32 to 0.54, and 7 equally spaced values for λ from 0.32 to 0.71. The solution to the concave program gives a minimax value for root MSE of 0.033. (The Monte Carlo simulation uses $K = 8000$ samples.) The least-favourable prior assigns 0 probability to 70% of the points: $\delta_{0j} = 0$ for 49 points, $0 \leq \delta_{0j} \leq 10^{-6}$ for 2 points, and $\delta_{0j} > 10^{-6}$ for 19 points. The maximum root MSE of the discrete minimax estimator over the 70 points is equal to the minimax value 0.033. The root MSE is equalized at 0.033 across the 19 points that are assigned probability greater than 10^{-6}. (The variation is in the fourth decimal place, between 0.0327 and 0.0328.) So the solution satisfies condition (7) very well.

Figure 5 shows the least-favourable prior probabilities δ_0 and Figure 6 equal probability

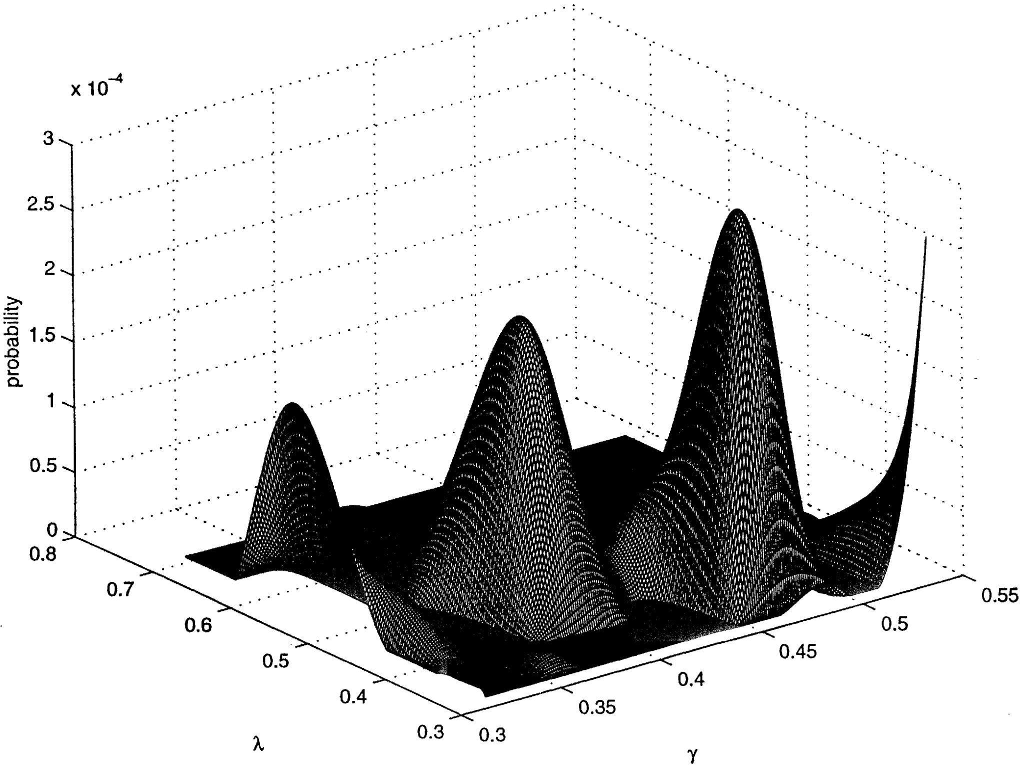

Figure 5. Least-favourable prior probabilities δ_0; interpolation. $N = 100$ and $T = 9$

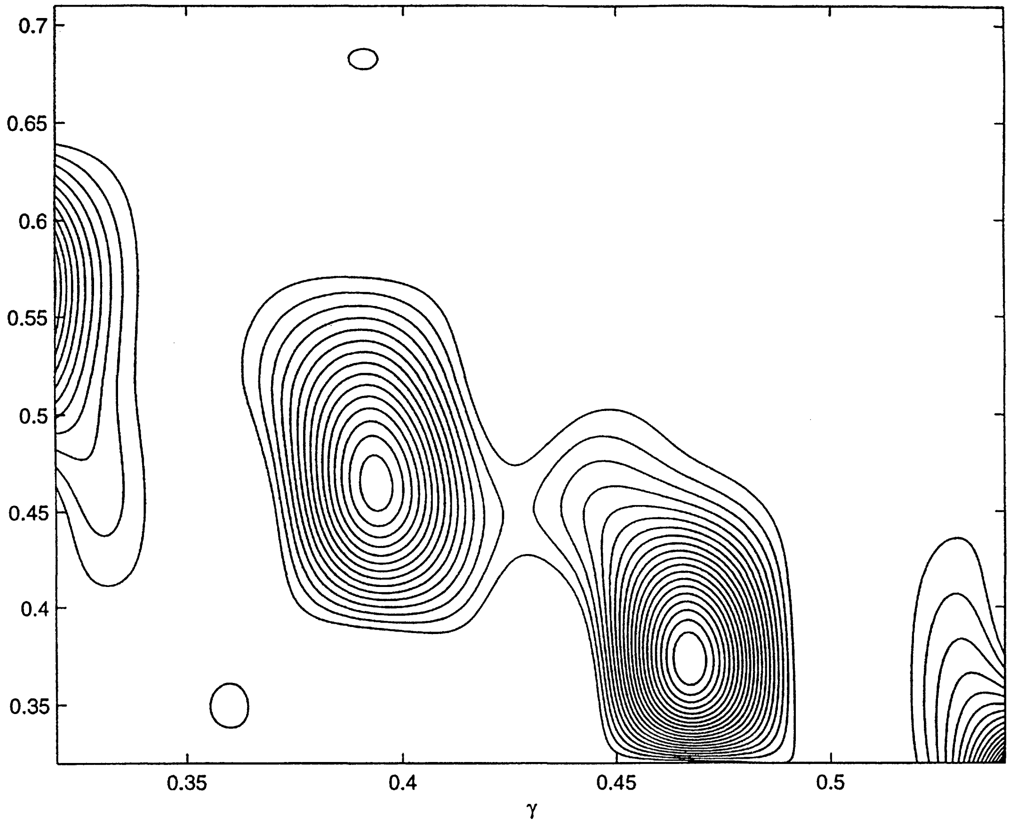

Figure 6. Equal probability contours of the least-favourable prior δ_0; interpolation. The heights of the contours are at equal probability increments: Δ, 2Δ, 3Δ, ... $N = 100$ and $T = 9$

contours for the least-favourable prior. The support of the distribution is quite concentrated along a negatively sloped diagonal. Figure 7 compares the risk functions for the ML and discrete minimax estimators, for $0.32 \leq \gamma \leq 0.54$ and $\lambda = 0.32$, 0.45, 0.58, and 0.71. The risk of the minimax estimator is evaluated at 0.01 increments for γ.

The risk of the ML estimator is fairly constant over the rectangle. The maximum root MSE is 0.039, which is attained at $(\gamma,\lambda) = (0.540, 0.320)$. The minimum root MSE over the rectangle is 0.035. Now consider the minimax estimator d_0 based on $\{\theta_1, \ldots, \theta_J\}$. The maximum risk of d_0 over the rectangle is obtained using a grid search combined with a numerical optimisation routine. The grid has 0.01 increments for γ and 0.01 increments for λ. The maximum value for root MSE of d_0 is 0.033, which is attained at $(\gamma,\lambda) = (0.442, 0.669)$. So the minimax value for the rectangle is 0.033, and the minimax estimator d_0 based on the $J = 70$ points is close to being a minimax estimator for the rectangle.

If one were given the rectangle as the parameter space, then the minimal value for maximum root MSE over this rectangle is 0.033, and there is a minimax estimator based on 70 points that essentially attains this value. The maximum likelihood estimator has root MSE that varies over

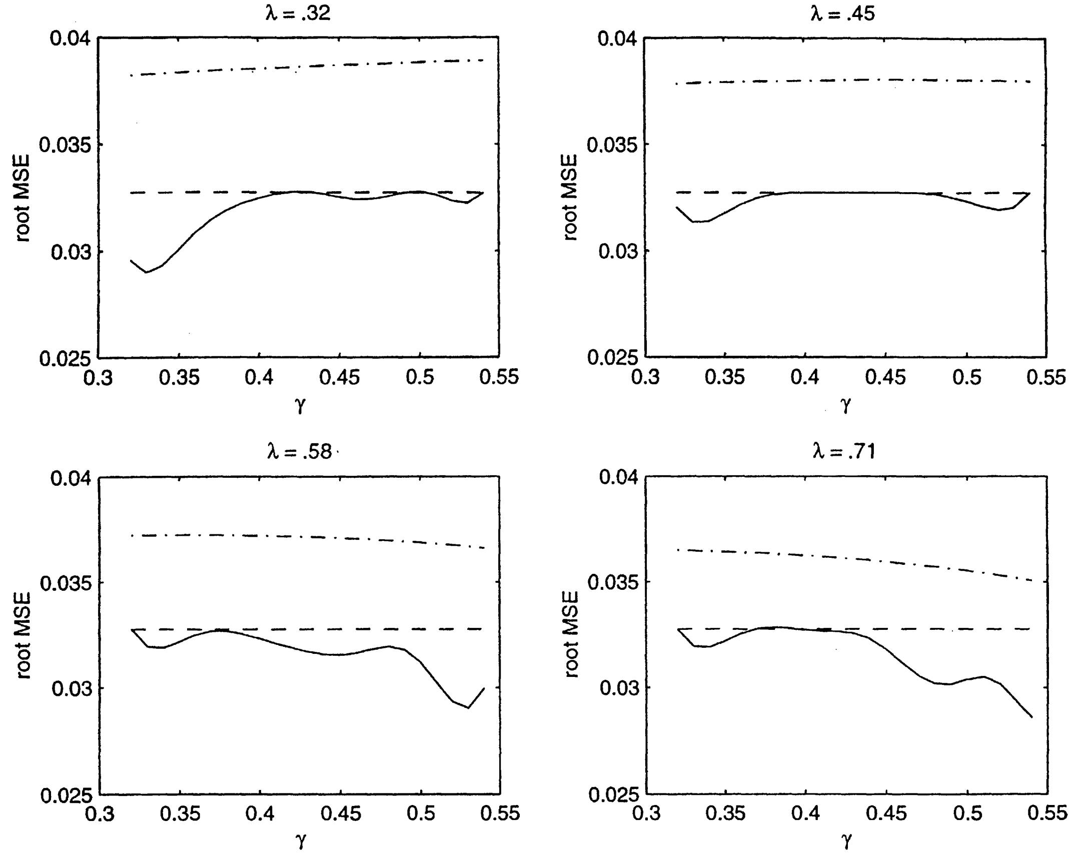

Figure 7. Root mean-square error for minimax estimator (solid line): $r(\theta,d_0)^{1/2}$, evaluated for $\gamma \in [0.32,0.54]$ in 0.01 increments, and for λ as shown; the dashed line indicates the minimax value of 0.033. Root mean-square error for maximum likelihood estimator (dash dot line), evaluated for $\gamma \in$ [0.32,0.54] in 0.01 increments, and for λ as shown. $N = 100$ and $T = 9$

the rectangle between 0.035 and 0.039. So the ML estimator is dominated over this local parameter space, but not by much. There is not much room here for improving on the risk of the ML estimator. In other applications, however, it might be of interest to evaluate the risk of the two-step estimator that applies maximum likelihood in the first step and then applies the minimax procedure over a neighbourhood around the maximum likelihood estimate.

6. CONCLUSION

We have developed an algorithm for calculating minimax decision rules with respect to the convex hull of a finite set of distributions. The minimax rule is a Bayes rule for the least-favourable distribution in the convex hull. The corresponding minimax value provides a lower bound on the minimax risk with respect to a larger set of distributions.

In our application, we compared the maximum risk of the maximum likelihood estimator with our lower bound. The comparison indicates that there is not a great deal to be gained from an alternative estimator in terms of reducing maximum risk. The application uses a set of point masses on the parameter space Θ. In other applications, it may be useful to work with a set of non-degenerate prior distributions. Our algorithm can be used to find the least-favourable mixture of these prior distributions. This least-favourable mixture will be subjectively reasonable if each prior in the set is subjectively reasonable. The Bayes estimator for this least-favourable prior will minimize the maximum risk over the set of priors. We can calculate the maximum risk of this Bayes estimator over all of Θ, and check whether it has lower maximum risk than an alternative such as the maximum likelihood estimator. We can also compare the maximum risk over Θ of this Bayes estimator with the lower bound based on the set of priors. If the maximum risk is close to the lower bound, then this Bayes estimator is attractive in terms of average risk with respect to the least-favourable prior and with respect to maximum risk over Θ.

ACKNOWLEDGEMENTS

I am grateful to Moshe Buchinsky, John Geweke, Jinyong Hahn, Keisuke Hirano, Guido Imbens, Frank Kleibergen, Peter Klibanoff, Thomas Knox, Charles Manski, Ariel Pakes, Benjamin Polak, Jack Porter, and John Rust for discussions. Financial support was provided by the National Science Foundation.

REFERENCES

Anscombe, F. and R. Aumann (1963), ‘A definition of subjective probability’, *Annals of Mathematical Statistics*, **34**, 199–205.

Barberis, N. (2000), ‘Investing for the long run when returns are predictable’, *Journal of Finance*, **55**, 225–264.

Blackwell, D. and M. Girshick (1954), *Theory of Games and Statistical Decisions*, John Wiley, New York.

Chamberlain, G. (2000), ‘Econometrics and decision theory’, *Journal of Econometrics*, **95**, 255–283.

Chamberlain, G. and K. Hirano (1999), ‘Predictive distributions based on longitudinal earnings data’, *Annales d’Économie et de Statistique*, **55–56**, 211–242.

Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer-Verlag, New York.

Ferguson, T. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, New York.

Gilboa, I. and D. Schmeidler (1989), ‘Maxmin expected utility with non-unique prior’, *Journal of Mathematical Economics*, **18**, 141–153.

Gill, P., W. Murray, M. Saunders and M. Wright (1986), ‘User’s guide for NPSOL’, Version 4.0, Report SOL 86-2, Department of Operations Research, Stanford University.

Good, I. J. (1952), ‘Rational decisions’, *Journal of the Royal Statistical Society, Series B*, **14**, 107–114.

Kempthorne, P. (1987), ‘Numerical specification of discrete least favourable prior distributions’, *SIAM J. Sci. Stat. Comput.*, **8**, 171–184.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Polak, E. (1997), *Optimization: Algorithms and Consistent Approximations*, Springer Verlag, New York.

Savage, L. J. (1954), *The Foundations of Statistics*, John Wiley, New York.

Van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Wald, A. (1950), *Statistical Decision Functions*, John Wiley, New York.

Wilson, R. B. (1963), *A Simplicial Algorithm for Concave Programming*, PhD dissertation, Harvard University, Cambridge, MA.