May 5, 1999

ECONOMETRIC APPLICATIONS OF MAXMIN EXPECTED UTILITY

Gary Chamberlain*
Harvard University

ABSTRACT

Gilboa and Schmeidler (1989) provide axioms on preferences that imply a set of distributions and a preference ordering based on the minimum expected utility with respect to this set. We consider joint distributions for data and for the random variables that, together with the agent's choice, determine utility-relevant outcomes; for example, a joint distribution for data that will be available when a portfolio decision is made and for future returns that will determine the value of the portfolio. The set of distributions is generated by combining a parametric model with a set of prior distributions. We seek a decision rule (a function of the data) that maximizes the minimum expected utility (or, equivalently, minimizes maximum risk) over the set of prior distributions. An algorithm is provided for the case of a finite set of prior distributions. It is based on finding the Bayes rule for a given prior and then solving a concave program to find the least-favorable prior distribution. The minmax value we obtain for the finite set of priors is a lower bound on the minmax risk for a larger set, such as the infinite set that includes all point masses on a Euclidean space, as in Wald (1950). An upper bound can be obtained by fixing a decision rule and finding its maximum risk. These bounds are applied to an estimation problem in an autoregressive model for panel data.

Key words: Mixture models; Minmax risk; Bayes decision rule; Least-favorable prior distribution; Concave program; Autoregression; Panel data

*I am grateful to Moshe Buchinsky, Jinyong Hahn, Guido Imbens, Keisuke Hirano, Peter Klibanoff, Charles Manski, Ariel Pakes, and Jack Porter for discussions. Financial support was provided by the National Science Foundation.

Department of Economics, Harvard University, Cambridge, MA 02138
gary_chamberlain@harvard.edu

ECONOMETRIC APPLICATIONS OF MAXMIN EXPECTED UTILITY

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 F_θ for some value of the parameter θ in the parameter space Θ. He is interested in a function $g(\theta)$ and would like an estimator that is optimal under a mean-square error criterion. How shall he choose an estimator?

In Section 2 we develop a framework that covers both of these problems. An algorithm for implementing the framework is developed in Section 3, and there is an application to an autoregressive model for panel data in Section 4.

2. FRAMEWORK

2.1 *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 is not 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 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 $\alpha f(z, w) + (1-\alpha)g(z, w)$.

A (randomized) decision rule 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 all such decision rules. The mapping $h : \mathcal{W} \times \mathcal{A}^* \to \mathcal{Y}$ 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 $l_d \in \mathcal{L}$: $l_d(z, w) = h(w, d(z))$.

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 a 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 $u: \mathcal{Y} \to \mathcal{R}$ and a non-empty, closed, convex set $\mathcal{S}$ of probability measures on $\mathcal{Z} \times \mathcal{W}$ such that: for all $f, g \in \mathcal{L}$,

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

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.

The preference relation on $\mathcal{L}$ induces a preference relation on the set $\mathcal{D}$ of decision rules, and we shall take the decision maker's problem to be:

$$\max_{d \in \mathcal{D}} \min_{Q \in \mathcal{S}} \int_{\mathcal{Z} \times \mathcal{W}} u(l_d(z, w)) \, dQ(z, w).$$

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.

2.2 *Mixture Models*

In order to make the maxmin problem operational, we shall consider mixture models in which the distribution Q for the random vector (Z, W) has the following form:

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

where π is a (prior) probability distribution on the parameter space Θ. The probability distribution P_θ can be decomposed into a marginal distribution F_θ for Z and a conditional distribution G_θ for W given Z:

$$P_\theta(A \times B) = \int_A G_\theta(B \,|\, z) \, dF_\theta(z).$$

We shall assume that F_θ has density $f(z \,|\, \theta)$ with respect to the measure μ:

$$F_\theta(A) = \int_A f(z \,|\, \theta) \, d\mu(z)$$

for all $\theta \in \Theta$.

We shall consider a set Γ of prior distributions π. Then the set $\mathcal{S}$ of distributions for (Z, W) is

$$\mathcal{S} = \{\int_\Theta P_\theta \, d\pi(\theta) : \pi \in \Gamma\}.$$

Now the decision maker's problem can be written as:

$$\min_{d \in \mathcal{D}} \max_{\pi \in \Gamma} r(\pi, d)$$

with risk function r:

$$r(\pi, d) = \int_\Theta \int_{\mathcal{Z}} L(\theta, z, d(z)) f(z \,|\, \theta) \, d\mu(z) \, d\pi(\theta)$$

and loss function L:

$$L(\theta, z, a^*) = -\int_{\mathcal{W}} u(h(w, a^*)) \, dG_\theta(w \,|\, z).$$

The use of loss, with a minus sign, and hence a minmax criterion is traditional, dating back to Wald (1950).

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 (1) (for $a \in \mathcal{A}$), and u is a von Neumann-Morgenstern utility function defined over roulette lotteries with monetary prizes. The specification of the parametric 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. In this application, the focus would not be 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 typically is on a function of the parameter, which we shall denote by $g(\theta)$. In this case, we set W equal to θ. The function h is given by: $h(\theta, a)$ $= (g(\theta), a)$ for $a \in \mathcal{A}$. The loss function could be $L(\theta, z, a) = (g(\theta) - a)^2$, with mean-square error for the risk function. Or the loss function could have a piecewise linear form:

$$L(\theta, z, a) = \begin{cases} c_1 |g(\theta) - a|, & \text{if } a \leq g(\theta); \\ c_2 |g(\theta) - a|, & \text{otherwise,} \end{cases} \tag{2}$$

with $c_1, c_2 > 0$. Then choosing $c_1/(c_1 + c_2) = .025$ and $.975$ could give estimates corresponding to a traditional confidence interval.

3. ALGORITHM

We shall consider a finite set of prior distributions: $\{\pi_1, \ldots, \pi_J\}$, and Γ is the convex hull:

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

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

$$\overline{V} = \inf_{d \in \mathcal{D}} \sup_{\pi \in \Gamma} r(\pi, d).$$

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

$$\underline{V} = \sup_{\pi \in \Gamma} \inf_{d \in \mathcal{D}} r(\pi, d).$$

A least-favorable distribution π_0 satisfies $\inf_{d \in \mathcal{D}} r(\pi_0, d) = \underline{V}$. A decision rule d_0 is Bayes with respect to the distribution π if

$$r(\pi, d_0) = \inf_{d \in \mathcal{D}} r(\pi, d).$$

A decision rule d generates a vector of risk values $(r(\pi_1, d), \ldots, r(\pi_J, d))$. The risk set S consists of all such vectors as d varies over $\mathcal{D}$:

$$S = \{(r(\pi_1, d), \ldots, r(\pi_J, d)) \in \mathcal{R}^J : d \in \mathcal{D}\}.$$

The risk set is convex, since we have allowed randomized decision rules. The minmax theorem states that if the risk set is bounded, then

$$\inf_{d \in \mathcal{D}} \sup_{\pi \in \Gamma} r(\pi, d) = \sup_{\pi \in \Gamma} \inf_{d \in \mathcal{D}} r(\pi, d),$$

and there exists a least favorable distribution π_0. If in addition the risk set is closed, then there exists a minmax decision rule d_0, and it is Bayes with respect to π_0. [See Blackwell and Girshick (1954, Theorem 2.4.2) and Ferguson (1967, Theorem 1, p. 82)]. We shall assume that the risk set is closed and bounded.

The first step in our algorithm is to find a Bayes rule with respect to a given prior distribution π. Note that

$$r(\pi, d) = \int_{\mathcal{Z}} \int_{\Theta} L(\theta, z, d(z)) f(z \mid \theta) \, d\pi(\theta) \, d\mu(z) \geq \int_{\mathcal{Z}} [\inf_{a \in \mathcal{A}} \int_{\Theta} L(\theta, z, a) f(z \mid \theta) \, d\pi(\theta)] \, d\mu(z)$$

for all $d \in \mathcal{D}$. We shall assume that the infimum of the inner integral is in fact obtained for some choice $a \in \mathcal{A}$, so that a Bayes rule with respect to π satisfies

$$d_\pi(z) = \arg\min_{a \in \mathcal{A}} \int_{\Theta} L(\theta, z, a) \, d\bar{\pi}(\theta \mid z), \tag{3}$$

where $\bar{\pi}$ is the posterior distribution of θ conditional on Z:

$$\bar{\pi}(B \mid z) = \int_{B} f(z \mid \theta) \, d\pi(\theta) \Big/ \int_{\Theta} f(z \mid \theta) \, d\pi(\theta). \tag{4}$$

The optimal choice under π minimizes the posterior expected loss. See Wald (1950, chap. 5.1), Blackwell and Girshick (1954, chap. 7.3), and Ferguson (1967, chap. 1.8). If the minimizer in (3) is not unique, then a Bayes rule could randomize over the set of minimizers.

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

$$M_J = \{\delta \in \mathcal{R}^{J-1} : \delta_j \geq 0, \sum_{j=1}^{J-1} \delta_j \leq 1\},$$

and let π^δ denote the mixture distribution:

$$\pi^\delta = \sum_{j=1}^{J} \delta_j \pi_j,$$

with $\delta_J = 1 - \sum_{j=1}^{J-1} \delta_j$. As δ varies over M_J, π^δ varies over Γ. The posterior distribution of θ conditional on Z under the mixture model (i.e., under the prior distribution π^δ) is

$$\bar{\pi}^\delta(B \mid z) = \sum_{j=1}^{J} \delta_j f_j(z) \bar{\pi}_j(B \mid z) \Bigg/ \sum_{j=1}^{J} \delta_j f_j(z), \tag{5}$$

where $\bar{\pi}_j$ is the posterior distribution of θ given Z under model j (i.e., under prior distribution π_j), and f_j is the likelihood under model j:

$$f_j(z) = \int_\Theta f(z \mid \theta)\, d\pi_j(\theta).$$

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

$$\rho(\delta) \equiv \min_{d \in \mathcal{D}} r(\pi^\delta, d) = r(\pi^\delta, d^\delta).$$

Since $r(\pi^\delta, d)$ is a linear function of δ for each d, it follows that ρ is a concave function. So maximizing ρ over the convex set M_J is a concave program:

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

The least favorable prior distribution is $\pi_0 = \sum_{j=1}^{J} \delta_{0j} \pi_j$.

Let $\partial\rho(\delta)$ denote the subgradient of ρ at δ. Let $\zeta_\delta \in \mathcal{R}^{J-1}$ have j^{th} component equal to $r(\pi_j, d^\delta) - r(\pi_J, d^\delta)$. We shall show that ζ_δ is a subgradient of ρ at δ. Note that

$$\rho(\delta) = \langle \zeta_\delta, \delta \rangle + r(\pi_J, d^\delta),$$

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

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

Hence $\zeta_\delta \in \partial\rho(\delta)$.

We can write the program in (6) as

$$\min_{\delta} -\rho(\delta) \quad \text{subject to} \quad g_1(\delta) \leq 0, \ldots, g_J(\delta) \leq 0, \tag{P}$$

where $g_j(\delta) = -\delta_j$ $(j = 1, \ldots, J-1)$ and $g_J(\delta) = \sum_{j=1}^{J-1} \delta_j - 1$. The Lagrangian of (P) is the following function $\mathcal{J}$ on $\mathcal{R}^J \times \mathcal{R}^{J-1}$:

$$\mathcal{J}(v, \delta) = -\rho(\delta) + \sum_{j=1}^{J} v_j g_j(\delta)$$

if $v \in \mathcal{R}_+^J$ (nonnegative components), with $\mathcal{J} = -\infty$ if $v \notin \mathcal{R}_+^J$.

In order for δ_0 to be an optimal solution to (P), it is necessary and sufficient that there exist an v_0 such that (v_0, δ_0) is a saddle point of the Lagrangian $\mathcal{J}$ of (P). Equivalently, δ_0 is an optimal solution if and only if there exist Lagrange multiplier values λ_j which, together with δ_0, satisfy the Kuhn-Tucker conditions for (P):

$$\lambda_j \geq 0, \quad g_j(\delta_0) \leq 0, \quad \lambda_j g_j(\delta_0) = 0 \qquad (j = 1, \ldots, J) \tag{a}$$

$$0 \in -\partial\rho(\delta_0) + \sum_{j=1}^{J} \lambda_j \partial g_j(\delta_0) \tag{b}$$

[Rockafellar (1970), Corollary 28.3.1].

Given the form of the constraint functions g_j, these Kuhn-Tucker conditions become

$$\lambda_j \geq 0, \quad \delta_{0j} \geq 0, \quad \lambda_j \delta_{0j} = 0 \qquad (j = 1, \ldots, J-1), \tag{a}$$

$$\lambda_J \geq 0, \quad \sum_{j=1}^{J-1} \delta_{0j} \leq 1, \quad \lambda_J \left(\sum_{j=1}^{J-1} \delta_{0j} - 1 \right) = 0,$$

$$0 \in -\partial\rho(\delta_0) + \begin{pmatrix} \lambda_J - \lambda_1 \\ \vdots \\ \lambda_J - \lambda_{J-1} \end{pmatrix}. \tag{b}$$

Since $\zeta_{\delta_0} \in \partial\rho(\delta_0)$, (b) is implied by

$$r(\pi_j, d_0) - r(\pi_J, d_0) = \lambda_J - \lambda_j \qquad (j = 1, \ldots, J-1), \tag{b'}$$

where d_0 is a Bayes rule with respect to π^{δ_0}. So (a) and (b') are sufficient for δ_0 to be an optimal solution to (P). If ρ is differentiable at δ_0, then ζ_{δ_0} is the unique subgradient [Rockafellar (1970), Theorem 25.1], and so (a) and (b') are necessary for δ_0 to be an optimal solution to (P). Since ρ is concave, the subset of the interior of M_J where ρ is not differentiable has Lebesgue measure zero [Rockafellar (1970), Theorem 25.5].

Let B be the set of integers j such that: $1 \le j \le J-1$ and $\delta_{0j} > 0$, or $j = J$ and $\sum_{j=1}^{J-1} \delta_{0j} < 1$. It follows from (a) that $\lambda_j = 0$ if $j \in B$. Then it follows from (b') that

$$\begin{aligned} r(\pi_j, d_0) &= r(\pi_0, d_0) \quad \text{if} \quad j \in B \\ r(\pi_j, d_0) &\le r(\pi_0, d_0) \quad \text{if} \quad j \notin B, \end{aligned} \tag{c}$$

where $\pi_0 = \pi^{\delta_0}$. Conversely, (c) implies that there exist Lagrange multiplier values λ_j which, together with δ_0, satisfy (a) and (b') (hence (a) and (b)). To see this, set

$$\lambda_j = r(\pi_0, d_0) - r(\pi_j, d_0) \qquad (j = 1, \ldots, J),$$

so (b') holds. Also $\lambda_j \ge 0$ $(1 \le j \le J)$ and $\lambda_j = 0$ if $j \in B$, so (a) holds.

Once we have obtained the least favorable prior π_0, the minmax rule d_0 is a Bayes rule with respect to π_0. So $d_0(z)$ solves

$$d_0(z) = \arg\min_{a \in \mathcal{A}} \int_\Theta L(\theta, z, a)\, d\bar{\pi}_0(\theta \,|\, z), \tag{7}$$

where $\bar{\pi}_0$ is the posterior distribution corresponding to the least favorable prior π_0. If the minimizing value in (7) is not unique, then the minmax rule d_0 may involve randomization.

Minmax Bounds

The minmax value $r(\pi_0, d_0)$ is with respect to the set Γ of prior distributions. If we consider a larger set of distributions Γ', with $\Gamma' = \Gamma \cup \Lambda$, then

$$\overline{V} = \inf_{d \in \mathcal{D}} \sup_{\pi \in \Gamma} r(\pi, d) \le \inf_{d \in \mathcal{D}} \sup_{\pi \in \Gamma'} r(\pi, d) = \overline{V}'.$$

So the minmax value relative to Γ provides a lower bound for the minmax value relative to Γ'.

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

$$\overline{V}' \leq \sup_{\pi \in \Gamma'} r(\pi, d) = \max\{\max_{1 \leq j \leq J} r(\pi_j, d), \sup_{\pi \in \Lambda} r(\pi, d)\}.$$

This upper bound is useful in that it may be feasible to maximize $r(\pi, d)$ over $\pi \in \Lambda$ for a fixed d, even though it is not feasible to compute the inf sup over $\mathcal{D}$ and Λ.

Suppose, for example, that the parameter space Θ equals $\mathcal{R}^p \times \Theta_2$. Let ϕ_β denote the point-mass distribution that assigns probability one to the point β. Let Λ consist of the set of product measures formed from point masses on $\mathcal{R}^p$ and a fixed distribution η on Θ_2:

$$\Lambda = \{\phi_\beta \times \eta : \beta \in \mathcal{R}^p\}.$$

Then by identifying θ_1 with the distribution $\phi_{\theta_1} \times \eta$, we can index the risk function by θ_1:

$$r(\theta_1, d) = \int_{\Theta_2} \int_{\mathcal{Z}} L((\theta_1, \theta_2), z, d(z)) f(z \,|\, (\theta_1, \theta_2))\, d\mu(z)\, d\eta(\theta_2), \qquad (\theta_1 \in \mathcal{R}^p).$$

Maximizing $r(\pi, d)$ over Λ reduces to maximizing $r(\theta_1, d)$ over $\mathcal{R}^p$:

$$\sup_{\pi \in \Lambda} r(\pi, d) = \sup_{\theta_1 \in \mathcal{R}^p} r(\theta_1, d).$$

This is a finite-dimensional maximization problem, and gradient methods may be effective provided that $r(\theta_1, d)$ is a smooth function of its first argument.

4. APPLICATION: AUTOREGRESSIVE MODELS FOR PANEL DATA

We shall work with the following parametric family:

$$\begin{aligned}
&Y_{it} = \gamma Y_{i,t-1} + \alpha_i + U_{it} \\
&\alpha_i \,|\, \{Y_{i0} = y_{i0}\}_{i=1}^N \overset{\text{ind}}{\sim} \mathcal{N}(\tau_1 + \tau_2 y_{i0}, \sigma_v^2) \\
&U_{it} \,|\, \{\alpha_i, Y_{i0} = y_{i0}\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \qquad (i = 1, \ldots, N;\, t = 1, \ldots, T),
\end{aligned} \tag{8}$$

with $\theta = (\gamma, \tau_1, \tau_2, \sigma^2, \psi)$ and $\psi \equiv \sigma_v^2/\sigma^2$. We shall focus on the estimation of γ, using a squared-error loss function. The observation Z is $\{Y_{it}, 1 \leq t \leq T\}_{i=1}^N$. The F_θ distribution for Z is given by (8); it is conditional on $\{y_{i0}\}_{i=1}^N$, which is observed. We set $W = \theta$, since θ combined with an action (an estimate) determines the utility-relevant outcome. Given a prior distribution π on Θ, the joint distribution Q of (Z, θ) is given by

$$Q(A, B) = \int_B F_\theta(A)\, d\pi(\theta) = \int_B \int_A f(z \,|\, \theta)\, d\mu(z)\, d\pi(\theta).$$

The loss function is $L(\theta, z, a) = (\gamma - a)^2$. The Bayes rule $d_\pi(z)$ in (3) is the posterior mean of γ:

$$d_\pi(z) = E_{\bar{\pi}}(\gamma \,|\, z) = \int \gamma\, d\bar{\pi}(\theta \,|\, z),$$

where $\bar{\pi}(\cdot \,|\, z)$, the posterior distribution of θ given $Z = z$, is given in (4). When the prior distribution is a mixture: $\pi^\delta = \sum_{j=1}^J \delta_j \pi_j$, it follows from (5) that the posterior mean is a convex combination of the posterior means under the components, π_j, of the mixture:

$$d^\delta(z) = E_{\bar{\pi}^\delta}(\gamma \,|\, z) = \sum_{j=1}^J \delta_j f_j(z) E_{\bar{\pi}_j}(\gamma \,|\, z) \Big/ \sum_{j=1}^J \delta_j f_j(z), \tag{9}$$

where $f_j(z) = \int f(z \,|\, \theta)\, d\pi_j(\theta)$ is the likelihood under model j. The risk under π^δ for an estimator d is

$$r(\pi^\delta, d) = \sum_{j=1}^J \delta_j r(\pi_j, d), \tag{10}$$

where

$$\begin{aligned} r(\pi_j, d) &= \int \int [\gamma - d(z)]^2 f(z \,|\, \theta)\, d\mu(z)\, d\pi_j(\theta) \qquad (11) \\ &= \int \left[\int [\gamma - d(z)]^2\, d\bar{\pi}_j(\theta \,|\, z) \right] f_j(z)\, d\mu(z) \\ &= \int \left[\mathrm{Var}_{\bar{\pi}_j}(\gamma \,|\, z) + [E_{\bar{\pi}_j}(\gamma \,|\, z) - d(z)]^2 \right] f_j(z)\, d\mu(z), \end{aligned}$$

and $\mathrm{Var}_{\bar{\pi}_j}(\cdot \,|\, z)$ is the conditional variance under the posterior distribution $\bar{\pi}_j(\cdot \,|\, z)$.

The prior distributions we consider have the following structure: $1/\sigma^2$ has a gamma distribution; conditional on σ^2, the components of (γ, τ_1, τ_2) are independent normals with variances proportional to σ^2; the prior distribution for ψ assigns unit mass to a single point. This family of priors is convenient in that $E_{\bar{\pi}_j}(\gamma \,|\, z)$, $\mathrm{Var}_{\bar{\pi}_j}(\gamma \,|\, z)$, and $f_j(z)$ have closed-form expressions. They are simple to compute with no need for numerical integration.

We can approximate $r(\pi_j, d)$ by Monte Carlo simulation. Obtain independent and identically distributed (i.i.d.) draws $\{Z(j,k)\}_{k=1}^K$ by drawing $\theta(j,k)$ from the prior distribution π_j and then drawing $Z(j,k)$ from $f(\cdot \,|\, \theta(j,k))$. Then we have

$$r(\pi_j, d) \cong \frac{1}{K} \sum_{k=1}^{K} \Big[\mathrm{Var}_{\bar{\pi}_j}(\gamma \,|\, Z(j,k)) + [E_{\bar{\pi}_j}(\gamma \,|\, Z(j,k)) - d(Z(j,k))]^2 \Big]. \tag{12}$$

Now we can approximate $r(\pi^\delta, d)$ using (10). This is how we calculate $\rho(\delta) = r(\pi^\delta, d^\delta)$, with d^δ obtained from (9). A numerical optimization 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; it is based on the subroutine NPSOL described in Gill et al. (1986)]. The maximizing value δ_0 gives the least favorable prior, $\pi_0 = \sum_{j=1}^J \delta_{0j} \pi_j$, and $\rho(\delta_0)$ is the minmax value for risk, relative to the set of models $\{\pi_j\}_{j=1}^J$.

In order to start with a more tractable problem, I decompose the parameter space Θ into the product $\Theta_1 \times \Theta_2$, put a prior distribution on Θ_2, which is held fixed, and do the minmax analysis with respect to Θ_1. Judging γ and ψ to be particularly important, I set $\theta_1 = (\gamma, \psi)$, $\theta_2 = (\tau_1, \tau_2, \sigma^2)$, with $\Theta_1 = \mathcal{R} \times [0, \infty)$ and $\Theta_2 = \mathcal{R}^2 \times (0, \infty)$. The prior distribution for $(\tau_1, \tau_2, \sigma^2)$ is the same in all the models. The minmax analysis is with respect to (γ, ψ). The prior distribution for $(\tau_1, \tau_2, \sigma^2)$ 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 $1/\sigma^2 \sim \chi^2(10)/.9$, so that the .1 and .9 quantiles for σ are .24 and .43. The mean of τ_1 is 0, the mean of τ_2 is .25, and the standard deviations of τ_1 and τ_2 in the (unconditional) t-distribution are .20. The values for $\{y_{i0}\}_{i=1}^N$ are obtained by drawing from a normal distribution with mean 0 and standard deviation .45; these values for y_{i0} are then kept fixed in evaluating risk.

I began by calculating minmax risk values over sets of forty to fifty models. I found that a similar minmax value could be obtained using considerably fewer models. Consider the case with

$N = 100$, $T = 2$, and nine models formed by combining three priors for γ with three priors for ψ. The priors for γ have means equal to .2, .5, .8 and standard deviations equal to .2. The priors for ψ have unit mass at .0, .4, and 1.0. These models were chosen based on examining the δ_0 weights in the least-favorable prior based on larger sets of models. Panel (a) of Table 1 gives the δ_0 weights for these nine models in the least favorable prior. (The Monte Carlo simulation uses $K = 10{,}000$ draws.) Note that the solution is on the boundary of the simplex, with four of the models receiving zero weight. Panel (b) of the table gives the square root of the mean-square error (MSE) of the minmax estimator d_0, under each of the nine models. Note that the risk is equalized (at $.111^2$) for the models that receive positive weight in the least-favorable prior. The risk is less for the other models. So the Kuhn-Tucker conditions are in fact satisfied by this solution.

Table 1. Minmax: $N = 100$, $T = 2$

(a) least-favorable prior: $\delta_{0,j}$				(b) root-MSE: $\sqrt{r(\pi_j, d_0)}$		
		$\psi = \sigma_v^2/\sigma^2$			$\psi = \sigma_v^2/\sigma^2$	
$E(\gamma), \mathrm{std}(\gamma)$	.0	.4	1.0	.0	.4	1.0
.2, .2	.000	.000	.463	.102	.103	.111
.5, .2	.008	.350	.000	.111	.111	.107
.8, .2	.147	.032	.000	.111	.111	.100

I have tried augmenting this set of nine models in various ways. For example, keep the three priors for γ and obtain forty-five models by combining them with the following fifteen point mass priors for ψ: $\psi = (.0, .1, \ldots, 1.0, 1.5, 2.0, 3.0, 5.0)$. Or keep the original three point mass priors for ψ and obtain forty-five models by combining them with the following fifteen priors for γ: $E(\gamma) = (.0, .1, \ldots, 1.4)$ with standard deviations equal to .2. Then decrease the standard deviation to .05, and then to essentially the point mass case with $\mathrm{std}(\gamma) = 10^{-6}$. These four sets of forty-five models all give minmax values for root-MSE that are close to the .111 value based on the nine models in Table 1. The increase in maximal root-MSE was in all cases less than .002. This was also the case in checking a wide range for $E(\gamma)$, with fifteen values between -10 and 10, and $\mathrm{std}(\gamma) = .2$.

Table 2 has the results for $N = 100$, $T = 4$. The three priors for γ again have means equal to .2, .5, .8 and standard deviations equal to .2. The priors for ψ have unit mass at .0, .3, and .8. The solution for the weights δ_0 in the least-favorable prior again occurs on the boundary of the simplex, with four of the nine models receiving zero weight. The root-MSE of the minmax estimator is equalized across the models that receive positive weight (at .065 or .064), with a lower root-MSE for the other models. There are similar results in Table 3 for $N = 100$, $T = 10$, with a minmax root-MSE of .032, and in Table 4 for $N = 1000$, $T = 2$, with a minmax root-MSE of .061. As before, I tried augmenting the nine models in various ways, and found little increase in the maximal root-MSE. For example, with $N = 100$ and $T = 4$, the increase in maximal root-MSE across the four sets of forty-five models was at most .002.

Table 2. Minmax: $N = 100$, $T = 4$

(a) least-favorable prior: $\delta_{0,j}$				(b) root-MSE: $\sqrt{r(\pi_j, d_0)}$		
	$\psi = \sigma_v^2/\sigma^2$			$\psi = \sigma_v^2/\sigma^2$		
$E(\gamma), \text{std}(\gamma)$	.0	.3	.8	.0	.3	.8
$.2, .2$	.000	.000	.415	.064	.064	.065
$.5, .2$	.000	.445	.007	.061	.065	.065
$.8, .2$	.066	.068	.000	.064	.065	.054

Table 3. Minmax: $N = 100$, $T = 10$

(a) least-favorable prior: $\delta_{0,j}$				(b) root-MSE: $\sqrt{r(\pi_j, d_0)}$		
	$\psi = \sigma_v^2/\sigma^2$			$\psi = \sigma_v^2/\sigma^2$		
$E(\gamma), \text{std}(\gamma)$	.05	.2	.4	.05	.2	.4
$.1, .2$	.000	.042	.128	.032	.032	.032
$.2, .2$	.000	.468	.021	.032	.032	.032
$.4, .2$	.340	.000	.000	.032	.032	.032

Table 4. Minmax: $N = 1000$, $T = 2$

(a) least-favorable prior: $\delta_{0,j}$				(b) root-MSE: $\sqrt{r(\pi_j, d_0)}$		
	$\psi = \sigma_v^2/\sigma^2$			$\psi = \sigma_v^2/\sigma^2$		
$E(\gamma), \text{std}(\gamma)$	.0	.2	.5	.0	.2	.5
$.5, .1$	.000	.000	.398	.053	.060	.061
$.7, .1$	.000	.355	.014	.060	.061	.061
$.9, .1$	.195	.038	.000	.061	.061	.050

These minmax risk values are in each case relative to the finite set of nine models. The risk values are lower bounds relative to a larger set of models, such as the infinite set that includes point masses on every point $(\gamma, \psi) \in \Theta_1$. In order to obtain an upper bound for that case, I shall consider some particular estimators and calculate their maximum risk over Θ_1. This calculation is based on (12), where now the prior π (which replaces π_j) is not restricted to a finite set. The prior for $\theta_2 = (\tau_1, \tau_2, \sigma^2)$ is fixed as before, but the prior for $\theta_1 = (\gamma, \psi)$ can place unit mass on any point in $\Theta_1 = \mathcal{R} \times [0, \infty)$. So we can index π by θ_1 and (12) becomes

$$r(\theta_1, d) \cong \frac{1}{K} \sum_{k=1}^{K} [\gamma - d(Z(\theta_1, k))]^2. \tag{13}$$

The i.i.d. draws $\{Z(\theta_1, k)\}_{k=1}^{K}$ are obtained by drawing $\theta_2(k)$ from its (fixed) prior, and then drawing $Z(\theta_1, k)$ from $f(\cdot \,|\, (\theta_1, \theta_2(k)))$. Then for a given estimator d, we maximize $r(\theta_1, d)$ over $\theta_1 \in \Theta_1$. This gives an upper bound on the minmax risk over Θ_1, given the fixed prior on Θ_2.

We consider three estimators for γ. (1) *Empirical Bayes*. For a given value of ψ, we impose the fixed prior on $\theta_2 = (\tau_1, \tau_2, \sigma^2)$ and an essentially uniform prior on γ [std(γ) = 1000]. We integrate over this prior distribution for (γ, θ_2) to obtain a marginal likelihood function for ψ, which is maximized to obtain $\hat{\psi}$. Then the estimate of γ is the posterior mean conditional on the data and on ψ, evaluated at $\psi = \hat{\psi}$. (2) *Uniform Prior*. The second estimator for γ is the posterior mean based on the fixed prior for θ_2 with an essentially uniform prior for (γ, ψ) [std(γ) = 1000, constant density for ψ on $[0, \infty)$]. (3) *Point-Mass Prior*. The third estimator for γ is the posterior mean based on the fixed prior for θ_2, an essentially uniform prior for γ [std(γ) = 1000], and a prior distribution for ψ that assigns point mass of .10 to $\psi = 0$, and probability of .90 to a gamma distribution with mean 12.5 and standard deviation 11.9. (The shape parameter is 1.1.) This prior distribution for ψ was motivated by examining the least-favorable priors, which tend to place substantial probability near 0.

The computation of the three estimators is described in Chamberlain (1998). The risk at a given value of (γ, ψ) is calculated using (13). Then the maximum risk over $(\gamma, \psi) \in \mathcal{R} \times [0, \infty)$ is calculated using a grid search. We ensure that the approximation in (13) gives a smooth function of θ_1 by using a single set of normal and gamma draws, which are then used to calculate $r(\theta_1, d)$ for the various values of θ_1. In fact the risk function appears to be very well-behaved in θ_1, and gradient maximization routines should be effective in working with a higher dimensional problem.

The maximum root-MSE values for the three estimators are shown in Table 5.

Table 5. Maximum root-MSE for $(\gamma, \psi) \in \mathcal{R} \times [0, \infty)$

	(N,T)			
Estimator	$(100, 2)$	$(100, 4)$	$(100, 10)$	$(1000, 2)$
(1) *empirical Bayes*	.136	.079	.035	.074
(2) *uniform prior*	.160	.097	.039	.090
(3) *point-mass prior*	.137	.098	.051	.097
minmax lower bound	.111	.065	.032	.061

These values are greater than the maximum root-MSE values for these estimators over the nine models that were used to obtain the minmax lower bounds. The minmax lower bounds, from Tables 1–4, are shown again in Table 5. The empirical Bayes estimator and the point-mass prior estimator have similar maximal risk when $N = 100$ and $T = 2$, but the point-mass prior estimator does not do as well at the other sample sizes. Of the three estimators, the empirical Bayes estimator provides the sharpest upper bound on root-MSE. Combining that upper bound with the minmax lower bound, we have a fairly good bound on the minmax value for $(\gamma, \psi) \in \mathcal{R} \times [0, \infty)$. Given the focus on the two-dimensional subproblem for (γ, ψ), and given the fixed prior for the other parameters, our tentative conclusion is that the maximum risk of the empirical Bayes estimator is fairly close to the minmax value.

REFERENCES

Anscombe, F. and R. Aumann (1963): "A Definition of Subjective Probability," *Annals of Mathematical Statistics*, 34, 199–205.

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

Chamberlain, G. (1998): "Econometrics and Decision Theory," *Journal of Econometrics*, forthcoming.

Chamberlain, G. and K. Hirano (1999): "Predictive Distributions Based on Longitudinal Earnings Data," *Annales d'Économie et de Statistique*, forthcoming.

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

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.

Rockafellar, R. T. (1970): *Convex Analysis*, Princeton: Princeton University Press.

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

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