Minimax Estimation and Forecasting in a Stationary Autoregression Model

By GARY CHAMBERLAIN*

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. Thus, 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 1 and an amount $1-a$ in asset 2, and he then holds on to the portfolio until date $T+H$. Let $w = \{(y_{1t}, y_{2t})\}_{t=T+1}^{T+H}$ and let $h(w, a)$ denote the value of the portfolio at $t = T+H$:

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

How should a be chosen?

Itzhak Gilboa and David 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. Chamberlain (2000) applies 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

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

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 S is the set of distributions.

Let $L(w, a)$ denote the loss function, and define the risk function as expected loss from using decision rule d when Q is the joint distribution of the observation Z and the utility-relevant variable W:

$$(1) \quad r(Q, d) = \int L(w, d(z))\, dQ(z, w).$$

In the portfolio-choice problem, loss would be the negative of utility: $L(w, a) = -u(h(w, a))$. Then, the decision-maker's problem is

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

The use of risk, and hence a minimax criterion, is traditional, dating back to Abraham Wald (1950).

Chamberlain (2000) develops an algorithm for computing a minimax decision rule. Section I describes this algorithm and Section II applies it to a stationary autoregression.

I. Algorithm

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

$$S = \left\{ \sum_{j=1}^{J} \delta_j Q_j : 0 \leq \delta_j \leq 1, \sum_{j=1}^{J} \delta_j = 1 \right\}.$$

Consider a zero-sum game in which the decision-maker chooses $d \in \mathcal{D}$, nature chooses

* Department of Economics, Harvard University, Cambridge, MA 02138. I thank James Stock for comments.

$Q \in S$, and the payoff to the decision-maker is $-r(Q, d)$. The minimax (or upper) value of the game is $\bar{V} = \inf_{d \in \mathcal{D}} \sup_{Q \in S} r(Q, d)$. A minimax decision rule d_0 satisfies $\sup_{Q \in S} r(Q, d_0) = \bar{V}$. The maxmin (or lower) value of the game is $\underline{V} = \sup_{Q \in S} \inf_{d \in \mathcal{D}} r(Q, d)$. A least-favorable 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 $\mathcal{D}$:

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

One 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. David Blackwell and M. A. Girshick (1954 Ch. 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

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

and there exists a least-favorable 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. I shall assume that the risk set is convex, closed, and bounded (see Blackwell and Girshick, 1954 [theorem 2.4.2]; Thomas 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,

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

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 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:

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

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

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

The least favorable distribution is $Q_0 = \sum_{j=1}^{J} \delta_{0j} Q_j$. The concave program can be solved using a sequential quadratic programming algorithm, as in Robert Wilson (1963). (The routine used in the application in Section II is nag_nlp_sol, from the NAG Fortran 90 library.)

The minimax value $r(Q_0, d_0)$ is with respect to the set S of distributions. If one considers a larger set of distributions $S' \supset S$, then

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

Thus, the minimax value relative to S provides a lower bound for the minimax value relative to the larger set S'.

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

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

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

II. Application: Stationary Autoregression

I shall work with the following parametric family: $\mathbf{Y} = (Y_0, \ldots, Y_{T+H}) \sim \{P_\theta : \theta \in \Theta\}$, with

$$Y_0 \sim \mathcal{N}(0, 1)$$

$$Y_t | Y_0 = y_0, \ldots, Y_{t-1} = y_{t-1} \sim \mathcal{N}(\theta y_{t-1}, 1 - \theta^2) \qquad t = 1, \ldots, T + H.$$

The parameter space Θ will be a subset of the open interval $\{\theta \in \mathcal{R} : -1 < \theta < 1\}$. The marginal distribution for Y_t is stationary: $Y_t \sim \mathcal{N}(0, 1)$ for $0 \leq t \leq T + H$. The observation is the realized value of $Z = (Y_0, \ldots, Y_T)$.

I shall consider estimating a power of θ with squared-error loss, so that the risk function is

$$r_1(\theta, d) = E_\theta[\theta^H - d(Z)]^2.$$

This corresponds to $L(w, a) = (w^H - a)^2$ in (1), with $w = \theta$. (To simplify notation, I shall use $r(\theta, d)$ interchangeably with $r(P_\theta, d)$.) I shall also be interested in forecasting Y_{T+H} using squared-error loss, with risk function

$$r_2(\theta, d) = E_\theta[Y_{T+H} - d(Z)]^2.$$

This corresponds to $L(w, a) = (w - a)^2$, with $w = y_{T+H}$. Note that

$$Y_{T+H} | Z = z \sim \mathcal{N}(\theta^H y_T, 1 - \theta^{2H}).$$

Hence,

$$r_2(\theta, d) = 1 - \theta^{2H} + E_\theta[\theta^H Y_T - d(Z)]^2.$$

I shall let the set $\mathcal{D}$ of decision rules be unrestricted.

Let $Y_t^* = \mu + \sigma Y_t$, so that the stationary distribution of Y_t^* is a general normal distribution with mean μ and variance σ^2, where μ and $\sigma > 0$ are known: $Y_t^* \sim \mathcal{N}(\mu, \sigma^2)$. Consider the risk function for point estimation when the observation is $Z^* = (Y_0^*, \ldots, Y_T^*)$: $r_1^*(\theta, d) = E_\theta[\theta^H - d(Z^*)]^2$. Define a function g mapping $\mathcal{D}$ onto $\mathcal{D}$: $g \circ d(z) = d(\sigma z + \mu)$. Then $r_1^*(\theta, d) = r_1(\theta, g \circ d)$ and

$$\inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} r_1^*(\theta, d) = \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} r_1(\theta, d).$$

Thus, the minimax value calculated for the $\mathcal{N}(0, 1)$ case applies to the general $\mathcal{N}(\mu, \sigma^2)$ case (with μ and σ known).

Now consider the risk function for forecasting Y_{T+H}^*: $r_2^*(\theta, d) = E_\theta[Y_{T+H}^* - d(Z^*)]^2$. Define a function g mapping $\mathcal{D}$ onto $\mathcal{D}$ as follows: $g \circ d(z) = \sigma^{-1}[d(\sigma z + \mu) - \mu]$. Then, $r_2^*(\theta, d) = \sigma^2 r_2(\theta, g \circ d)$, and

$$\inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} r_2^*(\theta, d) = \sigma^2 \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} r_2(\theta, d).$$

Thus, one can obtain the minimax value for $\mu = 0$ and $\sigma = 1$ and then simply multiply by σ^2 to obtain the minimax value for general μ and σ.

I shall work with a finite set of prior distributions on Θ: $\{\pi_j\}_{j=1}^J$. Then, Q_j is formed as the joint distribution with P_θ as the distribution for Y conditional on θ and with π_j as the marginal distribution on Θ. The set S is the convex hull of $\{Q_1, \ldots, Q_J\}$. Consider the case $T = 15$ with the parameter space Θ equal to the interval $0 \leq \theta < 1$. Initially, I let π_j be a point mass on θ_j and set up a grid with 100 equally spaced values for θ_j: 0, 0.01, ... , 0.99. For estimating θ (r_1 with $H = 1$), the solution to the concave program gives a minimax value for root mean-square-error (MSE) of $\rho^{0.5}(\delta_0) = 0.169$. Consider the minimax value corresponding to the entire parameter space: $\bar{V}_\Theta = \inf_d \sup_{\theta \in \Theta} r_1(\theta, d)$. The minimax value relative to $\{\theta_1, \ldots, \theta_J\}$ provides a lower bound on $\bar{V}_\Theta$. One can obtain an upper bound by calculating the maximum risk over Θ of the minimax estimator d_0 based on $\{\theta_1, \ldots, \theta_J\}$. The maximum value for root MSE is 0.171. Thus, the minimax value is tightly bounded at 0.17, and the minimax estimator based on the finite set with 100 points gives a close approximation to the minimax estimator for the interval [0, 1).

Now consider estimating θ^H. The lower bound on $\bar{V}_\Theta$ based on the discrete minimax estimator has root MSE equal to 0.113 for $H = 5$ and equal to 0.108 for $H = 10$. The upper bounds based on the maximum risk over [0, 1) of the discrete minimax estimator are 0.114 for $H = 5$ and 0.109 for $H = 10$. Thus, again the minimax values are tightly bounded, and the minimax estimator based on the finite set with 100 points gives a close approximation to the minimax estimator for the interval [0, 1).

An alternative to the minimax estimator is maximum likelihood (ML). Let $\hat{\theta}_{\mathrm{ML}}(z)$ denote the ML estimator of θ, so that $d_{\mathrm{ML}}(z) = [\hat{\theta}_{\mathrm{ML}}(z)]^H$ is the ML estimator of θ^H. The risk function $r_1(\theta, d_{\mathrm{ML}})$ for the ML estimator is smooth and unimodal on the interval [0, 1). The maximum values for root MSE with $H = 1, 5, 10$ are 0.21, 0.15, and 0.14. These maximum values are attained at $\theta = 0$, 0.78, and 0.89. Thus, the minimax estimator shows a noticeable improvement over ML in terms of maximum risk when $T = 15$.

The discrete minimax estimator is a Bayes estimator for the least-favorable prior. Let $F(\theta)$ denote the distribution function for the least-favorable prior. When $T = 15$ and $H = 1$, there is substantial mass at $\theta = 0$ with $F(0) = 0.45$. The rest of the distribution is concentrated on the interval [0.43, 0.71], with $F(0.42) = 0.45$, $F(0.43) = 0.58$, and $F(0.71) = 0.98$. As H increases, the distribution puts relatively more weight on larger values of θ. With $H = 5$, $F(0.58) = 0.00$, $F(0.59) = 0.05$, $F(0.82) = 0.50$, and $F(0.95) = 0.98$. With $H = 10$, $F(0.76) = 0.00$, $F(0.77) = 0.02$, $F(0.89) = 0.52$, and $F(0.98) = 1.00$.

Now consider the forecasting problem, with squared-error loss and risk function r_2. Let the parameter space be the interval $[\alpha, \beta)$, where $0 \leq \alpha < \beta \leq 1$. The surprising result is that the least favorable prior assigns probability 1 to the point α, so that the minimax forecasting rule is $d_0(z) = \alpha^H y_T$. The risk of this rule is

$$r_2(\theta, d_0) = 1 - \theta^{2H} + (\theta^H - \alpha^H)^2$$
$$= 1 - 2\theta^H\alpha^H + \alpha^{2H}.$$

The maximum risk is $1 - \alpha^{2H}$, which is attained at $\theta = \alpha$. Since $r_2(\alpha, d) > 1 - \alpha^{2H}$ unless $d(z) = \alpha^H y_T$ (almost everywhere with respect to Lebesgue measure on $\mathcal{R}^{T+1}$), it follows that d_0 is the unique minimax forecasting rule. If the lower bound of the parameter space has $-1 < \alpha < 0$, then the unique minimax forecasting rule is $d_0(z) = 0$.

The ML forecast rule is $d_{\mathrm{ML}}(z) = [\hat{\theta}_{\mathrm{ML}}(z)]^H y_T$. Consider the maximum value of risk for this rule, with $\Theta = [0, 1)$ and horizons $H = 1, 5, 10$. The root MSE values are 1.019, 1.004, and 1.003. Thus, the percentage increase over the minimax value of 1 is quite small.

I have been working with a model in which the variance of the stationary distribution is given, and fixed at 1. Now consider a model in which the innovation variance is given, and fixed at 1:

$$Y_0 \sim \mathcal{N}(0, 1/[1-\theta^2])$$

$$Y_t | Y_0 = y_0, \ldots, Y_{t-1} = y_{t-1} \sim \mathcal{N}(\theta y_{t-1}, 1)$$
$$t = 1, \ldots, T+H.$$

The marginal distribution of Y_t is stationary: $Y_t \sim \mathcal{N}(0, 1/[1 - \theta^2])$. Note that now

$$Y_{t+H} | Z = z \sim \mathcal{N}(\theta^H y_T, [1-\theta^{2H}]/[1-\theta^2])$$

and

$$r_2(\theta, d) = (1 - \theta^{2H})/(1 - \theta^2)$$
$$+ E_\theta[\theta^H Y_T - d(Z)]^2.$$

Consider the problem of forecasting Y_{T+H}. Let the parameter space be the interval $[0, \beta]$, with $\beta < 1$. We have the surprising result that, when $H \geq 2$, the least favorable prior assigns probability 1 to the point β, so that the minimax forecasting rule is $d_0(z) = \beta^H y_T$. The risk of this rule is

$$r_2(\theta, d_0) = [(1 - \theta^{2H})$$
$$+ (\theta^H - \beta^H)^2]/(1 - \theta^2).$$

The sign of the derivative with respect to θ of $r_2(\theta, d_0)$ is the sign of

$$\theta[(1-\beta^H)^2 + \beta^H[2 + (H-2)\theta^H - H\theta^{H-2}]]$$

for $0 \leq \theta \leq \beta < 1$. Thus, the derivative is nonnegative if $H \geq 2$, and the maximum value for $r_2(\theta, d_0)$ is $(1-\beta^{2H})/(1-\beta^2)$, which is attained at $\theta = \beta$. Since $r_2(\beta, d) > (1-\beta^{2H})/(1-\beta^2)$ unless $d(z) = \beta^H y_T$, it follows that d_0 is the unique minimax forecasting rule. As β increases to 1, the minimax risk approaches H.

When $H = 1$, the lower bound based on the discrete (100-point) minimax estimator gives a forecast root MSE equal to 1.02. When $H = 5$ and 10, the minimax values for root MSE are $\sqrt{H} = 2.24$ and 3.16, respectively. Now consider the maximum risk of the ML forecast rule, with $\Theta = [0, 1)$ and horizons $H = 1, 5$, and 10. The root MSE values are 1.03, 2.27, and 3.20. Thus, the percentage increase over the minimax value is quite small.

In the model where the stationary variance is given, the minimax forecast rule is based on assigning probability 1 to a single point in the parameter space. This is also true when the innovation variance is given, for forecast horizons $H \geq 2$. This aspect of the minimax solution can be avoided by allowing only subjectively reasonable distributions in the set S. One can specify a set of nondegenerate prior distributions on Θ: $\{\pi_j\}_{j=1}^J$. Then Q_j is formed as the joint distribution with P_θ as the distribution for Y conditional on θ and with π_j as the marginal distribution on Θ. For example, with $\Theta = [0, 1)$, let π_1 be a beta(1, 19) distribution (with mean 0.05 and standard deviation 0.048); let $\pi_2, \ldots, \pi_5$ be beta distributions with means equal to 0.2, 0.4, 0.6, and 0.8, and with standard deviations equal to 0.1; and let π_6 be a beta(19, 1) distribution (with mean 0.95 and standard deviation 0.048).

Consider estimating θ^H. The lower bound on $\bar{V}_\Theta$ based on the (100-point) discrete minimax estimator has root MSE equal to 0.19, 0.22, and 0.25, respectively, for $H = 1, 5, 10$. The corresponding maximum values of root MSE for the ML estimator are 0.22, 0.32, and 0.37. Now consider the minimax estimator based on the six beta distributions. The maximum values for root MSE (over $\Theta = [0, 1)$) are 0.19, 0.23, and 0.28. Thus, the minimax estimator based on the six beta priors shows a noticeable improvement over ML in terms of maximum risk when $T = 15$. The least-favorable δ weights on the beta priors are 0.39, 0, 0.06, 0.33, 0.22, and 0 for $H = 1$; 0, 0, 0, 0.33, 0.04, and 0.63 for $H = 5$; and 0, 0, 0, 0.06, 0.27, and 0.68 for $H = 10$.

In the forecast problem, the least favorable δ weights put all the weight on a single distribution (π_6) for horizons $H = 5$ and 10. This is reasonable to the extent that the six beta distributions are each judged to be subjectively reasonable. It would be interesting to make similar comparisons between minimax and ML decision rules in the portfolio-choice problem.

REFERENCES

Blackwell, David and Girshick, M. A. *Theory of games and statistical decisions*. New York: Wiley, 1954.

Chamberlain, Gary. "Econometric Applications of Maxmin Expected Utility." Mimeo, Harvard University, 2000.

Ferguson, Thomas S. *Mathematical statistics: A decision theoretic approach*. New York: Academic Press, 1967.

Gilboa, Itzhak and Schmeidler, David. "Maxmin Expected Utility with Non-unique Prior." *Journal of Mathematical Economics*, 1989, *18*(2), pp. 141–53.

Wald, Abraham. *Statistical decision functions*. New York: Wiley, 1950.

Wilson, Robert B. "A Simplicial Algorithm for Concave Programming." Ph.D. dissertation, Harvard University, 1963.