DECISION THEORY APPLIED TO AN INSTRUMENTAL VARIABLES MODEL [1]

Gary Chamberlain

ABSTRACT

This paper applies some general concepts in decision theory to a simple instrumental variables model. There are two endogenous variables linked by a single structural equation; k of the exogenous variables are excluded from this structural equation and provide the instrumental variables (IV). The reduced-form distribution of the endogenous variables conditional on the exogenous variables corresponds to independent draws from a bivariate normal distribution with linear regression functions and a known covariance matrix. A canonical form of the model has parameter vector (ρ, ϕ, ω), where ϕ is the parameter of interest and is normalized to be a point on the unit circle. The reduced-form coefficients on the instrumental variables are split into a scalar parameter ρ and a parameter vector ω, which is normalized to be a point on the $(k-1)$-dimensional unit sphere; ρ measures the strength of the association between the endogenous variables and the instrumental variables, and ω is a measure of direction. A prior distribution is introduced for the IV model. The parameters ϕ, ρ, and ω are treated as independent random variables. The distribution for ϕ is uniform on the unit circle; the distribution for ω is uniform on the unit sphere with dimension $k-1$. These choices arise from the solution of a minimax problem. The prior for ρ is left general. It turns out that given any positive value for ρ, the Bayes estimator of ϕ does not depend upon ρ; it equals the maximum-likelihood estimator. This Bayes estimator has constant risk; since it minimizes average risk with respect to a proper prior, it is minimax.

The same general concepts are applied to obtain confidence intervals. The prior distribution is used in two ways. The first way is to integrate out the nuisance parameter ω in the IV model. That gives an integrated likelihood function with two scalar parameters, ϕ and ρ. Inverting a likelihood ratio test, based on the integrated likelihood function, provides a confidence interval for ϕ. This lacks finite sample optimality, but invariance arguments show that the risk function only depends upon ρ and not on ϕ or ω. The second approach to confidence sets aims for finite sample optimality by setting up a loss function that trades off coverage against the length of the interval. The automatic uniform priors are used for ϕ and ω, but a prior is also needed for the scalar ρ, and no guidance is offered on this choice. The Bayes rule is a highest posterior density set. Invariance arguments show that the risk function only depends upon ρ and not on ϕ or ω. The optimality result combines average risk and maximum risk. The confidence set minimizes the average–with respect to the prior distribution for ρ–of the maximum risk, where the maximization is with respect to ϕ and ω.

KEYWORDS: Instrumental variables, risk function, Bayes decision rule, minimax, invariance

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

[1] This article is based on my Fisher-Schultz Lecture presented at the 2001 Econometric Society European Meeting in Lausanne. An earlier version is Chamberlain (2003). Financial support was provided by the National Science Foundation.

September 2006

DECISION THEORY APPLIED TO AN INSTRUMENTAL VARIABLES MODEL

1. INTRODUCTION

I shall apply some general concepts in decision theory to a simple instrumental variables model. There are two endogenous variables linked by a single structural equation; k of the exogenous variables are excluded from this structural equation and provide the instrumental (IV) variables. The reduced-form distribution of the endogenous variables conditional on the exogenous variables corresponds to independent draws from a bivariate normal distribution with linear regression functions and a known covariance matrix. Section 2 sets up the model and puts it in a canonical form that will simplify derivations. In the canonical form, there is a key parameter of interest, ϕ, which corresponds to the ratio of coefficients of the two endogenous variables in the single structural equation. The normalization is that ϕ is a point on the unit circle. The reduced-form coefficients on the instrumental variables are split into a scalar parameter ρ and a parameter vector ω, which is normalized to be a point on the $(k-1)$-dimensional unit sphere; ρ measures the strength of the association between the endogenous variables and the instrumental variables, and ω is a measure of direction. Section 3 lays out the basic elements of a statistical decision problem, which combine to provide the risk function. Section 4 provides a simple version of the complete class theorem, relating admissible decision rules to Bayes decision rules, which minimize average risk.

Section 5 relates the minimization of average risk to the minimization of posterior expected loss, conditional on the sample. This provides a link between Bayesian statistics and frequentist decision theory. A prior distribution is introduced for the IV model. The parameters ϕ, ρ, and ω are treated as independent random variables. The distribution for ϕ is uniform on the unit circle; the distribution for ω is uniform on the unit sphere with dimension $k-1$. These choices arise from the solution of a minimax problem. The prior for ρ is left general. It turns out that given any positive value for ρ, the Bayes estimator of ϕ does not depend upon ρ, and this Bayes estimator equals the maximum-likelihood estimator. So, for a particular prior distribution and loss function, the maximum-likelihood estimator has a finite sample optimality property in this simple IV model.

This optimality, however, does not extend to all the parameters. There is an even simpler model, a version of the k-means model of Stein (1956, 1959) and James and Stein (1961), in which the parameters consist of the scalar ρ and the point ω on the unit sphere with dimension $(k-1)$. The maximum-likelihood estimator of ρ is not admissible in this model. The problem becomes more serious as the dimension of the parameter space increases. Alternative estimators are considered, including a Bayes estimator based on a uniform prior distribution for ω. Now, however, the prior distribution for ρ matters, and there is no specific guidance on the choice of the prior distribution for ρ. The risk function simplifies, so that it only depends upon ρ and not on ω.

A Bayes decision rule requires a prior distribution on the parameter space. A careful, thoughtful specification for this distribution may be sufficiently costly that one is interested in alternative criteria for working with a risk function. An alternative to average risk is maximum risk. Section 6 develops the minimax criterion and its relationship to invariance. The statistical decision problem, as developed in Section 3, involves three spaces: sample space, parameter space, and action space. Invariance involves transformations on each of these spaces and leads to a simplification of the risk function, generalizing the result for the k-means model in Section 5. This simplification of the risk function leads to an optimality result that combines the average risk and maximum risk criteria. The Bayes estimator in the k-means model minimizes the average—with respect to the prior distribution for ρ—of the maximum, over ω, risk. This estimator requires paying the cost of constructing a prior distribution for ρ. Since ρ is a scalar, perhaps the cost will not be too high. For ω, which may be of high dimension, the "automatic" prior is used (uniform on the unit sphere of dimension $k-1$). In the IV model, the prior distribution for ρ turns out not to matter for point estimation of ϕ. Only the "automatic" uniform priors for ϕ (on the unit circle) and for ω (on the unit sphere) are used. The resulting Bayes estimator for ϕ, which equals the ML estimator, is optimal in the minimax sense.

Section 7 considers confidence sets. The prior distribution is used in two ways. The first way is to integrate out the nuisance parameter ω in the IV model. That gives an integrated likelihood function with two scalar parameters, ϕ and ρ. Inverting a likelihood ratio test, based on the integrated likelihood function, provides a confidence interval for ϕ. This lacks finite sample optimality but the invariance arguments show that the risk function only depends upon ρ and not on ϕ or ω. It follows that the integrated likelihood function can be used to evaluate the risk of this procedure, even though risk is defined with respect to the original parameter space that includes ω. This suggests that large-sample approximations will be more accurate using the integrated likelihood function, since the dimension of the parameter space is reduced from $k+1$ to 2.

The second approach to confidence sets aims for finite sample optimality by setting up a loss function that trades off coverage against the length of the interval. The automatic uniform priors are used for ϕ and ω, but a prior is also needed for the scalar ρ, and, as before, no guidance is offered on this choice. The Bayes rule is a highest posterior density set. Invariance arguments show that the risk function only depends upon ρ and not on ϕ or ω. The optimality result again combines average risk and maximum risk. The confidence set minimizes the average–with respect to the prior distribution for ρ–of the maximum risk, where the maximization is with respect to ϕ and ω.

Section 8 considers hypothesis tests. The null hypothesis is that $\phi \in A$, where A is a given subset of the unit circle. We can use the automatic uniform prior for ω but not for ϕ. For example, if A is a discrete set then the uniform prior assigns it zero probability, whereas the null hypothesis must be assigned positive prior probability in order to obtain a nontrivial Bayes test. So we must pay the cost of constructing prior distributions for ϕ and ρ. Then the risk function for the Bayes test depends only upon ϕ and ρ and not on ω. The Bayes test minimizes the average—with respect to the prior distribution for ϕ and ρ—of the maximum, over ω, risk. If the primary goal is a confidence interval, then one could consider inverting the Bayes test to find the set of null hypotheses for ϕ that are not rejected. This would be different from the approach to optimal confidence intervals in Section 7; each of the null hypotheses would involve a different prior distribution for ϕ, whereas the Bayes interval in Section 7 uses a single prior distribution for ϕ, which is uniform on the unit circle.

I have tried to make the paper self-contained. The theory I draw on is essentially contained in Ferguson (1967), which owes much to Wald (1950). The finite-sample optimality results for the IV model are to my knowledge new. There is some overlap in the Section 8 result on hypothesis tests with independent work by Andrews, Moreira, and Stock (2006). The first version of the confidence interval procedure in Section 7, which inverts a likelihood ratio test based on the integrated likelihood function, is in Chamberlain and Imbens (2004).

Gilboa and Schmeidler (1989) provide axioms that are related to Wald's minimax risk criterion. Hansen, Sargent, and coauthors use minimax ideas in their work on robust estimation and control; see, for example, Hansen, Sargent, Turmuhambetova, and Williams (2005) and Hansen and Sargent (2005). Manski (2004) uses a minimax regret criterion in his work on statistical treatment rules. Sims (2001) discusses pitfalls in a minimax approach to model uncertainty, but does see a potential use for automatic procedures to generate priors and associated decision rules.

2. THE MODEL

This section sets up the basic model. Suppose that we have a sample of individuals ($i = 1, \ldots, n$) with the following specification for a regression function:

$$E(Y_{i2} \,|\, x_{i1}, x_{i2}, Y_{i1}, A_i) = x_{i1}\alpha_1 + \gamma Y_{i1} + \alpha_2 A_i.$$

Here Y_{i1}, Y_{i2}, and A_i are scalars; x_{i1} and x_{i2} are row vectors, α_1 is a column vector, and γ and α_2 are scalars. We are interested in the coefficient γ on Y_{i1}. The specification imposes an exclusion restriction: it assumes that x_{i2} does not help in predicting Y_{i2}, conditional on x_{i1}, Y_{i1}, and A_i. The regression function for Y_{i1} conditional on x_{i1}, x_{i2}, and A_i is not restricted, apart from linearity:

$$E(Y_{i1} \,|\, x_{i1}, x_{i2}, A_i) = x_{i1}\alpha_3 + x_{i2}\pi_1^* + \alpha_4 A_i.$$

The regression function for A_i conditional on x_{i1} and x_{i2} is restricted to exclude x_{i2}:

$$E(A_i \,|\, x_{i1}, x_{i2}) = x_{i1}\alpha_5.$$

Here α_3, π_1^*, and α_5 are column vectors, and α_4 is a scalar.

There is a missing data problem: A_i is not observed, and so, for example, estimation based on a least-squares regression of Y_2 on x_1, Y_1, and A is not feasible. We can, however, exploit the exclusion of x_2 from the regression functions for Y_2 and A. The motivation for the exclusion restriction could be based on random assignment. Suppose that Y_{i2} is a measure of earnings for individual i, Y_{i1} is a measure of his education, and A_i is a measure of ability. The variables in x_1 could include a constant, age, and measures of family background. The variables in x_2 reflect a randomly assigned encouragement (subsidy) for the individual to obtain more education. The exclusion restriction is that the encouragement itself does not help to predict earnings if we condition on the actual education attained (and on x_1 and A); also the encouragement variables do not help to predict ability (given x_1), due to the random assignment.

Consider the regression functions of Y_{i1} and Y_{i2} conditional on x_{i1} and x_{i2}, so that the latent variable A_i is not being used:

$$\begin{aligned} E(Y_{i1}\,|\,x_{i1},x_{i2}) &= E[E(Y_{i1}\,|\,x_{i1},x_{i2},A_i)\,|\,x_{i1},x_{i2}] \\ &= x_{i1}\alpha_3 + x_{i2}\pi_1^* + x_{i1}\alpha_5\alpha_4 \\ &= x_{i1}\alpha_1^* + x_{i2}\pi_1^* \end{aligned}$$

(with $\alpha_1^* = \alpha_3 + \alpha_5\alpha_4$). Likewise,

$$E(Y_{i2}\,|\,x_{i1},x_{i2}) = x_{i1}\alpha_2^* + x_{i2}\pi_2^*,$$

where

$$\pi_2^* = \gamma\pi_1^*$$

(and we shall not need the formula for α_2^*). The key point here is that these regression functions only involve observed variables, and we can obtain γ from the x_{i2} coefficient vectors π_1^* and π_2^*, provided that $\pi_1^* \neq 0$. We shall refer to the variables in x_{i2} as *instrumental variables*.

Suppose that there are k variables in x_{i2}. Let $\Pi^* = (\,\pi_1^* \quad \pi_2^*\,)$. Π^* is $k \times 2$ but has rank at most equal to one, since π_2^* equals a scalar (γ) times π_1^*. It will be useful to express Π^* as the product of a column vector and a row vector:

$$\Pi^* = \omega^* \phi^{*\prime},$$

where ω^* is $k \times 1$ and ϕ^* is 2×1; for example:

$$\omega^* = \pi_1^*, \quad \phi^* = \begin{pmatrix} 1 \\ \gamma \end{pmatrix}.$$

This decomposition is not unique, since we can multiply ω^* by a nonzero constant c and multiply ϕ^* by c^{-1}. Nevertheless, given ϕ^* up to scale, we can recover γ from the ratio of the second component of ϕ^* divided by the first component: $\gamma = \phi_2^*/\phi_1^*$. (If $\phi_1^* = 0$ then $\pi_1^* = 0$; in that case we cannot recover γ from π_1^* and π_2^*.)

Notation:

$$Y_1 = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{n1} \end{pmatrix}, \quad Y_2 = \begin{pmatrix} Y_{12} \\ \vdots \\ Y_{n2} \end{pmatrix}, \quad X_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_{12} \\ \vdots \\ x_{n2} \end{pmatrix}.$$

A classical regression model for (Y_1, Y_2) conditional on (X_1, X_2) has the following form:

$$Y_1 = X_1\alpha_1^* + X_2\pi_1^* + U_1$$

$$Y_2 = X_1\alpha_2^* + X_2\pi_2^* + U_2,$$

where the disturbances (U_1, U_2) have the following multivariate normal distribution:

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \mid (X_1, X_2) \sim \mathcal{N}(0, \Sigma \otimes I_n).$$

See, for example, Goldberger (1991). (Here the disturbances are just notation for the difference between a variable and its conditional expectation: $U_1 = Y_1 - E(Y_1 \mid X_1, X_2)$, $U_2 = Y_2 - E(Y_2 \mid X_1, X_2)$. The role of the latent variable A in the missing data problem is quite different.) Consider estimating π_1^* and π_2^* with the coefficient vectors on X_2 in the least-squares regressions of Y_1 and Y_2 on (X_1, X_2). Using the residual regression result (Goldberger, Chapter 17), we can write this as

$$\hat{\pi}^* = \begin{pmatrix} \hat{\pi}_1^* \\ \hat{\pi}_2^* \end{pmatrix} = (I_2 \otimes (\tilde{X}_2' \tilde{X}_2)^{-1} \tilde{X}_2') \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix},$$

where $\tilde{X}_2$ is formed from the residuals in least-squares regressions of the columns of X_2 on X_1:

$$\tilde{X}_2 = (I_n - X_1(X_1'X_1)^{-1}X_1')X_2.$$

The distribution of $\hat{\pi}^*$ conditional on (X_1, X_2) is

$$\hat{\pi}^* \,|\, X_1, X_2 \,\sim\, \mathcal{N}(\pi^*, \Sigma \otimes (\tilde{X}_2'\tilde{X}_2)^{-1}),$$

where

$$\pi^* = \begin{pmatrix} \pi_1^* \\ \pi_2^* \end{pmatrix} = \begin{pmatrix} \phi_1^*\omega^* \\ \phi_2^*\omega^* \end{pmatrix} = \phi^* \otimes \omega^*.$$

Even if the covariance matrix Σ is given, procedures based on standard large sample approximations can lead to poor inferences for γ.[2] So I shall simplify the analysis by assuming that Σ is given. There will still be problems in making inferences on γ, and we will have a simpler setting in which to examine decision theoretic solutions to these problems. With Σ given, we can simplify the model by choosing nonsingular matrices C and F such that

$$C\Sigma C' = I_2, \quad F(\tilde{X}_2'\tilde{X}_2)^{-1}F' = I_k,$$

and defining

$$\pi = (C \otimes F)\pi^*, \quad \hat{\pi} = (C \otimes F)\hat{\pi}^*, \quad \phi = C\phi^*/||C\phi^*||, \quad \omega = F\omega^*/||F\omega^*||.$$

($||s||$ denotes the norm of the column vector s: $||s|| = (s's)^{1/2}$.) Then we have

$$\begin{aligned} \hat{\pi} \,&\sim\, \mathcal{N}(C\phi^* \otimes F\omega^*, C\Sigma C' \otimes F(\tilde{X}_2'\tilde{X}_2)^{-1}F') \\ &= \mathcal{N}(\rho\phi \otimes \omega, I_{2k}), \end{aligned}$$

where $\rho = ||\pi||$ is a scalar, ϕ is a 2×1 vector with $||\phi|| = 1$, and ω is a $k \times 1$ vector with $||\omega|| = 1$. Note that ϕ and ω are not separately identified since $(-\phi) \otimes (-\omega) = \phi \otimes \omega$. The set $\{\phi, -\phi\}$ is determined from $\phi \otimes \omega$ since $\omega_j\phi$ is determined, where ω_j is a nonzero element of ω, and $||\omega_j\phi|| = |\omega_j|$. The normalization of the structural coefficients to a point on the unit circle is used in Hillier (1990).

[2] See, for example, Bekker (1994) and Staiger and Stock (1997).

3. BASIC ELEMENTS OF A STATISTICAL DECISION PROBLEM

This section sets up the basic concepts of frequentist decision theory. The IV model is used to illustrate.

Sample Space: $\mathcal{Z}$. The observation z is some point in the sample space $\mathcal{Z}$. In the IV example, with $\hat{\pi}$ as the observation, the sample space is $\mathcal{R}^{2k}$.

Parameter Space: Θ. There is a set of distributions on the sample space that is indexed by a parameter θ, which takes on values in the parameter space Θ.

Statistical Model: The statistical model is a mapping from the parameter space Θ into probability distributions on the sample space $\mathcal{Z}$. A point $\theta \in \Theta$ is mapped into a distribution P_θ. Let $\mathcal{P}(\mathcal{Z})$ denote the set of probability distributions on $\mathcal{Z}$. The statistical model is the mapping $P\colon \Theta \to \mathcal{P}(\mathcal{Z})$. The model implies a set of distributions: $\{P_\theta : \theta \in \Theta\}$ (which is the image of the parameter space under the mapping).

In the IV example, the parameter space is

$$\Theta = \{(\rho, \phi, \omega) : \rho \in \mathcal{R}, \rho \geq 0; \phi \in \mathcal{R}^2, ||\phi|| = 1; \omega \in \mathcal{R}^k, ||\omega|| = 1\}.$$

Let $\mathcal{R}_+$ denote the nonnegative real numbers, and let S^m denote the unit sphere of dimension m in $\mathcal{R}^{m+1}$:

$$S^m = \{x \in \mathcal{R}^{m+1} : ||x|| = 1\}.$$

Then we can write the parameter space for the IV model as

$$\Theta = \mathcal{R}_+ \times S^1 \times S^{k-1}.$$

With $\theta = (\rho, \phi, \omega)$, the distribution P_θ is $\mathcal{N}(\rho\phi \otimes \omega, I_{2k})$.

Action Space: $\mathcal{A}$. This is the set of actions (choices) available to the statistician. Consider estimation of ϕ in the IV model. We shall say that a_1 and $a_2 \in S^1$ are equivalent if $a_1 = a_2$ or $a_1 = -a_2$. This is an equivalence relation, and it partitions S^1 into equivalence classes. Let the action space $\mathcal{A}$ consist of these equivalence classes.

Loss Function: $L\colon \Theta \times \mathcal{A} \to \mathcal{R}$. The loss function is a real-valued function defined on $\Theta \times \mathcal{A}$. In the IV model, a loss function for estimating ϕ could be

$$L((\rho, \phi, \omega), a) = 1 - (\phi' a)^2.$$

Since $||\phi|| = ||a|| = 1$, $\phi' a$ is the cosine of the angle formed by ϕ and the estimate a. The minimal value of the loss is zero, which is attained when $a = \pm\phi$. This loss function can be expressed in terms of the original parametrization. Use

$$\phi = \Sigma^{-1/2} \begin{pmatrix} 1 \\ \gamma \end{pmatrix} / b(\gamma)^{1/2}, \quad a = \Sigma^{-1/2} \begin{pmatrix} 1 \\ \hat{\gamma} \end{pmatrix} / b(\hat{\gamma})^{1/2},$$

where

$$b(\gamma) = \begin{pmatrix} 1 \\ \gamma \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} 1 \\ \gamma \end{pmatrix} \quad \text{and} \quad b(\hat{\gamma}) = \begin{pmatrix} 1 \\ \hat{\gamma} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} 1 \\ \hat{\gamma} \end{pmatrix}.$$

Then some algebra gives

$$1 - (\phi' a)^2 = \det(\Sigma^{-1})(\gamma - \hat{\gamma})^2 / [b(\gamma) b(\hat{\gamma})].$$

This loss function differs from squared error, $(\gamma - \hat{\gamma})^2$, by dividing by $b(\gamma)b(\hat{\gamma})$. So it puts less weight on large values of γ or $\hat{\gamma}$.

Decision Rule: $d\colon \mathcal{Z} \to \mathcal{A}$. A decision rule (or strategy) is a mapping from the sample space to the action space. Given an observation z, the statistician chooses an action $a = d(z)$ in $\mathcal{A}$. The set of feasible decision rules is $\mathcal{D}$. In the IV model, the decision rule d could be some estimator for ϕ.

Risk Function: $R : \Theta \times \mathcal{D} \to \mathcal{R}$. The risk function gives the expected loss from using the decision rule d when the parameter value is θ:

$$R(\theta, d) = \int L(\theta, d(z))\, dP_\theta(z).$$

An expectation notation is useful, and we shall use E_θ to denote expectation with respect to the P_θ distribution. So if Z is a random variable that takes on values in the sample space $\mathcal{Z}$, we can write the risk function as

$$R(\theta, d) = E_\theta[L(\theta, d(Z))].$$

Example: Estimation with Quadratic Loss

$$\Theta = \mathcal{R}, \quad \mathcal{A} = \mathcal{R}, \quad L(\theta, a) = (\theta - a)^2, \quad Z = (Z_1, ..., Z_n) \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$$

Consider $d(z) = c\bar{z}$, where $0 < c \leq 1$ and $\bar{z} = \sum_{i=1}^{n} z_i / n$.

$$R(\theta, d) = E_\theta[\theta - d(Z)]^2 = \frac{c^2}{n} + (1-c)^2\theta^2 \tag{3.1}$$

4. ADMISSIBLE DECISION RULES AND THE COMPLETE CLASS THEOREM

This section follows Ferguson (1967, Chapter 2) in providing a simple version of the fundamental complete class theorem.

Given a statistical model $P\colon \Theta \to \mathcal{P}(\mathcal{Z})$ and a loss function L, what decision rule should the statistician use? Consider solving $\min_d R(\theta, d)$ for the risk function in equation (3.1):

$$\frac{\partial}{\partial c}[\frac{c^2}{n} + (1-c)^2\theta^2] = 0 \quad \Rightarrow \quad c = \frac{n\theta^2}{1 + n\theta^2}$$

—but this is not operational since θ is unknown. Or consider using the risk function to compare two specific decision rules: $d(z) = \bar{z}$, $d^*(z) = \frac{1}{2}\bar{z}$:

$$R(\theta, d) = 1/n, \quad R(\theta, d^*) = \frac{1}{4}(\frac{1}{n} + \theta^2).$$

d is better if $\theta^2 > 3/n$; d^* is better if $\theta^2 < 3/n$. This is the typical situation: risk functions for different rules cross each other. We can, however, obtain a partial ordering through the concept of admissibility.

Admissibility: a decision rule d is *admissible* if there exists no rule that (weakly) dominates d; i.e., no rule d^* with $R(\theta, d^*) \leq R(\theta, d)$ for all $\theta \in \Theta$ and $R(\theta, d^*) < R(\theta, d)$ for some $\theta \in \Theta$. A decision rule d is weakly admissible if there exists no rule d^* that strongly dominates d; i.e., no rule d^* with $R(\theta, d^*) < R(\theta, d)$ for all $\theta \in \Theta$.

In order to characterize the class of admissible rules, we shall consider putting a distribution on the parameter space. A distribution on the parameter space is called a *prior distribution*.

Average Risk: Given a prior distribution ψ on the parameter space Θ, the average risk of a decision rule d is

$$R^*(\psi, d) = \int_\Theta R(\theta, d)\, d\psi(\theta),$$

which is formed by averaging over the different values for the parameter θ, using the weights supplied by the prior distribution ψ.

Given a prior distribution ψ, we can try to find a decision rule that minimizes the average risk.

Bayes Rule: a decision rule d_ψ is *Bayes* with respect to the prior distribution ψ if

$$R^*(\psi, d_\psi) = \inf_{d \in D} R^*(\psi, d).$$

Theorem 4.1. If d_ψ is a Bayes rule with respect to the prior distribution ψ, then d_ψ is weakly admissible.[3]

Proof. The proof is by contradiction. Suppose that there is a $d^* \in \mathcal{D}$ that strictly dominates d_ψ, so that

$$R(\theta, d^*) < R(\theta, d_\psi) \quad \text{for all } \theta \in \Theta.$$

It follows that

$$R^*(\psi, d^*) = \int_\Theta R(\theta, d^*)\, d\psi(\theta) < \int_\Theta R(\theta, d_\psi)\, d\psi(\theta) = R^*(\psi, d_\psi);$$

but then d_ψ is not Bayes with respect to ψ. $\diamond$

It will be convenient to allow for randomization over a finite set of decision rules. For any finite set $\{d_1, \ldots, d_m\} \subset \mathcal{D}$ and probability vector $(\alpha_1, \ldots, \alpha_m)$ (with $\alpha_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_{i=1}^m \alpha_i = 1$), we allow the *randomized decision rule* δ that employs d_i with probability α_i, and we define

$$R(\theta, \delta) = \sum_{i=1}^m \alpha_i R(\theta, d_i).$$

[3] There is a related result in Ferguson (1967, Theorem 2, p. 60).

$\mathcal{D}^*$ is the set consisting of all such randomized decision rules. (Note that $\mathcal{D} \subset \mathcal{D}^*$, since a randomized rule can assign probability one to a single element of $\mathcal{D}$.)

Risk Set: with $\Theta = \{\theta_1, \ldots, \theta_k\}$ finite, the *risk set* consists of the risk vectors that correspond to some decision rule:

$$S = \{(y_1, \ldots, y_k) \in \mathcal{R}^k : \text{for some } \delta \in \mathcal{D}^*,\ y_j = R(\theta_j, \delta) \text{ for } j = 1, \ldots, k\}.$$

Lemma. The risk set S is convex.[4]

Proof. Given any $y \in S$ and $y^* \in S$, and any number β between zero and one, we need to show that $\beta y + (1-\beta)y^* \in S$. There are decision rules δ and δ^* in $\mathcal{D}^*$ for which $y_j = R(\theta_j, \delta)$ and $y_j^* = R(\theta_j, \delta^*)$ for $j = 1, \ldots, k$. By combining the finite sets of nonrandom decision rules used by δ and δ^*, we have a set $\{d_1, \ldots, d_m\} \subset \mathcal{D}$ such that δ assigns probabilities $(\alpha_1, \ldots, \alpha_m)$ to these rules and δ^* assigns probabilities $(\alpha_1^*, \ldots, \alpha_m^*)$. Consider the randomized rule δ_β that assigns probabilities $(\beta\alpha_1 + (1-\beta)\alpha_1^*, \ldots, \beta\alpha_m + (1-\beta)\alpha_m^*)$ to these rules. (We can regard δ_β as randomizing over δ and δ^* with probabilities β and $1-\beta$.) Then

$$\begin{aligned} R(\theta, \delta_\beta) &= \sum_{i=1}^{m} (\beta\alpha_i + (1-\beta)\alpha_i^*) R(\theta, d_i) \\ &= \beta \sum_{i=1}^{m} \alpha_i R(\theta, d_i) + (1-\beta) \sum_{i=1}^{m} \alpha_i^* R(\theta, d_i) \\ &= \beta R(\theta, \delta) + (1-\beta) R(\theta, \delta^*). \end{aligned}$$

So if w is the risk vector for δ_β, with $w_j = R(\theta_j, \delta_\beta)$, then $w = \beta y + (1-\beta)y^* \in S$. $\diamond$

Theorem 4.2. (Complete Class) If Θ is finite and $\delta \in \mathcal{D}^*$ is admissible, then δ is Bayes (with respect to some prior distribution).[5]

Proof. Let $a = (R(\theta_1, \delta), \ldots, R(\theta_k, \delta))$ denote the risk vector for δ, and let Q_a denote the set of risk vectors that are at least as good as a:

$$Q_a = \{x \in \mathcal{R}^k : x_j \le a_j \text{ for } j = 1, \ldots, k\}.$$

[4] See Ferguson (1967, Lemma 1, p. 35).

[5] See Ferguson (1967, Theorem 1, p. 86).

Since δ is admissible, $Q_a \cap S = \{a\}$. Since $Q_a - \{a\}$ and S are disjoint convex sets, the separating hyperplane theorem asserts that there is a $p \neq 0$ in $\mathcal{R}^k$ such that $p'x \leq p'y$ for all $x \in Q_a - \{a\}$ and $y \in S$ (and hence for $x = a$). If some coordinate p_l of the p vector were negative, then $\sum_j p_j x_j > \sum_j p_j y_j$ by taking x_l sufficiently negative. Hence $p_j \geq 0$ for $j = 1, \ldots, k$, and we can normalize p so that $\sum_j p_j = 1$. Now p corresponds to a probability distribution ψ over Θ (with $\psi\{\theta_j\} = p_j$), and

$$R^*(\psi, \delta) = \sum_j p_j R(\theta_j, \delta) = p'a \leq p'y$$

for all $y \in S$ implies that δ is a Bayes rule with respect to ψ. $\diamond$

There are general versions of the complete class theorem in Wald (1950, Chapter 3.6), Le Cam (1986, Chapter 2.2), and Strasser (1985, Chapter 8, Section 47).

5. AVERAGE RISK OPTIMALITY

We have seen that a Bayes rule plays a fundamental role in achieving admissibility. This section develops a result on the calculation of Bayes rules.

5.1. *Calculation of Bayes Rules*

Once we have a definition of risk, it is natural to want to minimize risk. There is an obstacle, however, because risk depends upon the unknown distribution that is generating the data. The risk $R(\theta, d)$ of a decision rule d depends upon the value of the parameter θ that indexes the distribution P_θ. Our observation z is drawn from the distribution P_θ for some value of the parameter θ in the parameter space Θ, and $R(\theta, d) = \int_{\mathcal{Z}} L(\theta, d(z))\, dP_\theta(z)$ depends (in general) on the value of θ.

One response to this obstacle is to consider a weaker notion of optimality, namely admissibility. But that only gives us a partial ordering of decision rules. It eliminates ones that are dominated for all θ, but many decision rules remain. The admissible set will typically include some decision rules that are not very appealing. In the problem of estimating θ under quadratic loss, the decision rule (estimator) $d(z) = 4$ would be admissible (if $4 \in \Theta$), even though it does not use the data at all. No other estimator will be as good at $\theta = 4$.

One way to choose a single, optimal decision rule is to focus on average risk. We reduce the risk function $R(\,\cdot\,, d)$ to a single number (for a given d) by averaging over the parameter space. This requires a set of weights to attach to the different values of θ—a prior distribution ψ. Then average risk with respect to ψ is

$$R^*(\psi, d) = \int_{\Theta} R(\theta, d)\, d\psi(\theta).$$

Now we have a real-valued objective function, and we can try to solve the problem

$$\min_{d \in \mathcal{D}} R^*(\psi, d).$$

This problem has a solution, at least in the sense of there being decision rules with average risk arbitrarily close to $\inf_{d\in\mathcal{D}} R^*(\psi, d)$. The problem now is computational: $d(\cdot)$ is a function defined on the sample space $\mathcal{Z}$, and the optimization is over a space of functions.

Fortunately, this function-space optimization can be reduced to an "ordinary" optimization over a subset of a finite-dimensional Euclidean space. The algorithm for minimizing average risk requires a *likelihood function*. We shall assume that each distribution P_θ in the statistical model has a density with respect to the same measure m.

Likelihood Function: $f\colon \mathcal{Z} \times \Theta \to \mathcal{R}$. For any value of the parameter $\theta \in \Theta$, $f(\cdot \,|\, \theta)$ is the density function (with respect to the measure m) of the distribution P_θ, so that for any (measurable) subset B of the sample space $\mathcal{Z}$:

$$P_\theta(B) = \int_B f(z \,|\, \theta) dm(z).$$

In the discrete case, m is counting measure: $P_\theta(B) = \sum_{z\in B} f(z \,|\, \theta)$. In the (absolutely) continuous case, m is Lebesgue measure: $P_\theta(B) = \int_B f(z \,|\, \theta)\, dz$. The IV model has this form, with m equal to Lebesgue measure on $\mathcal{R}^{2k}$ and

$$f(z \,|\, (\rho, \phi, \omega)) = (2\pi)^{-k} \exp(-\frac{1}{2}||z - \rho(\phi \otimes \omega)||^2).$$

The key step in minimizing the average risk is to reverse the order of integration, so we first average over Θ for a given value z of the observation, and then average over the sample space:[6]

$$\begin{aligned} R^*(\psi, d) &= \int_\Theta \biggl[\int_{\mathcal{Z}} L(\theta, d(z)) f(z \,|\, \theta)\, dm(z) \biggr] d\psi(\theta) \\ &= \int_{\mathcal{Z}} \biggl[\int_\Theta L(\theta, d(z)) f(z \,|\, \theta)\, d\psi(\theta) \biggr] dm(z). \end{aligned}$$

Now consider minimizing the inner integral for fixed z:

$$\int_\Theta L(\theta, d(z)) f(z \,|\, \theta)\, d\psi(\theta) \geq \inf_{a \in \mathcal{A}} \int_\Theta L(\theta, a) f(z \,|\, \theta)\, d\psi(\theta).$$

This inequality holds for every z, and so it holds when we average over z:

$$R^*(\psi, d) \geq \int_{\mathcal{Z}} \biggl[\inf_{a \in \mathcal{A}} \int_\Theta L(\theta, a) f(z \,|\, \theta)\, d\psi(\theta) \biggr] dm(z).$$

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

$$d_\psi(z) = \arg\min_{a \in \mathcal{A}} \int_\Theta L(\theta, a) f(z \,|\, \theta)\, d\psi(\theta). \tag{5.1}$$

In our applications the action space $\mathcal{A}$ will have finite dimension, and so (5.1) reduces the function minimization problem to an ordinary minimization problem. We need to assume that the set of decision rules is unrestricted to ensure that the solution in (5.1) is in $\mathcal{D}$. For example, an unbiasedness restriction in an estimation problem would require that $E_\theta d(Z) = \theta$ for all $\theta \in \Theta$. The decision rule in (5.1) would not in general satisfy that restriction.

For a given value z of the observation Z, define the *posterior distribution* on Θ as follows:

$$\bar{\psi}(B \,|\, z) = \int_B f(z \,|\, \theta)\, d\psi(\theta) \biggm/ \int_\Theta f(z \,|\, \theta)\, d\psi(\theta),$$

for $B \subset \Theta$. It is convenient to write this as

$$\bar{\psi}(B \,|\, z) = c(z) \int_B f(z \,|\, \theta)\, d\psi(\theta),$$

[6] See Ferguson (1967, p. 44).

where the function $c(z)$ does not depend upon θ. For a fixed value of z, the minimizing value for a in (5.1) is not affected if we multiply $f(z \,|\, \theta)$ by $c(z)$, and so

$$d_\psi(z) = \arg\min_{a \in \mathcal{A}} \int_\Theta L(\theta, a)\, d\bar{\psi}(\theta \,|\, z). \tag{5.1$'$}$$

The Bayes rule evaluated at the observation value z is obtained by choosing the action to minimize expected loss, where the expectation is taken with respect to the posterior distribution.

Suppose that the prior distribution ψ has a density π with respect to a measure υ: for $B \subset \Theta$,

$$\psi(B) = \int_B \pi(\theta)\, d\upsilon(\theta)$$

($= \int_B \pi(\theta)\, d\theta$ for Lebesgue measure, and $= \sum_B p(\theta)$ for counting measure). Then the density (with respect to υ) of the posterior distribution is

$$\bar{\pi}(\theta \,|\, z) = c(z) f(z \,|\, \theta) \pi(\theta). \tag{5.2}$$

A convenient restatement of (5.2) is

$$\bar{\pi}(\theta \,|\, z) \propto f(z \,|\, \theta) \pi(\theta) \tag{5.2$'$}$$

—the posterior density is proportional to the likelihood times the prior density, where the proportionality constant is $c(z)$, which does not depend upon θ. Here likelihood refers to $f(z \,|\, \theta)$, which is regarded as a function of θ with z fixed.

It is significant that a conditional distribution on the parameter space Θ emerges from the minimization of average risk. The complete class theorem motivates the focus on minimizing average risk. But that does not by itself lead to a posterior distribution on the parameter space. It is necessary that the form of the decision rule not be restricted. Average risk is a double integral, over the parameter space and the sample space. If the form of the decision rule is not restricted, then the optimal rule is obtained by minimizing the integral over the parameter space, separately for each point in the sample space. It is this minimization that corresponds to the conditional (posterior) distribution on the parameter space. The optimal rule evaluated at a point z in the sample space minimizes expected loss, where the expectation is with respect to the conditional (posterior) distribution on the parameter space, given z.

An alternative starting point is to regard the prior distribution as representing subjective opinion or beliefs about θ before observing z. The prior density $\pi(\theta)$ gives a density for the marginal distribution on Θ, the likelihood $f(z \,|\, \theta)$ gives a density for the conditional distribution on the sample space $\mathcal{Z}$ given θ, and their product $f(z \,|\, \theta)\pi(\theta)$ is the joint density for the distribution on $\Theta \times \mathcal{Z}$. The density for the marginal distribution on $\mathcal{Z}$ is obtained by integrating θ out of the joint density: $q(z) = \int f(z \,|\, \theta)\pi(\theta)\, d\upsilon(\theta) = 1/c(z)$. Then we obtain the density for the conditional distribution on Θ given z from Bayes' Theorem: $\bar{\pi}(\theta \,|\, z) = f(z \,|\, \theta)\pi(\theta)/q(z) = c(z)f(z \,|\, \theta)\pi(\theta)$.

In Bayesian statistics, the conditional distribution on Θ given z is treated as a complete description of beliefs about θ that combines the prior beliefs with the information from the observation. Given a loss function, the optimal action is chosen to minimize posterior (or conditional) expected loss. This corresponds exactly with the Bayes rule d_ψ for minimizing average risk, when d_ψ is evaluated at the observation z. It also provides a link between frequentist decision theory, which focuses on the risk function $R(\theta, d)$, and Bayesian statistics. The frequentist aspect is that $R(\theta, d)$ $= E_\theta[L(\theta, d(Z)]$ involves averaging over the sample space $\mathcal{Z}$. One can think of repeated samples $\{Z^{(j)}\}_{j=1}^J$ drawn from P_θ, and an average loss $\frac{1}{J}\sum_{j=1}^J L(\theta, d(Z^{(j)}))$. Then, by a law of large numbers, the long-run average loss in repeated samples from P_θ would converge to our risk $R(\theta, d)$ as $J \to \infty$. A Bayesian statistician with prior beliefs ψ will choose the same action as a frequentist decision theorist who minimizes average risk, using ψ for the weights.

5.2. *Application: IV Model*

The likelihood function is

$$f(z \,|\, (\rho, \phi, \omega)) = (2\pi)^{-k} \exp(-\frac{1}{2}||z - \rho(\phi \otimes \omega)||^2),$$

where

$$\rho \geq 0, \quad \phi \in S^1 = \{x \in \mathcal{R}^2 : ||x|| = 1\}, \quad \omega \in S^{k-1} = \{x \in \mathcal{R}^k : ||x|| = 1\}.$$

Let ψ_1 denote the prior distribution for ρ on $\mathcal{R}_+$; I shall leave this general, and not make a specific choice. I will, however, make specific choices for prior distributions for ϕ and ω; namely, uniform distributions on the unit circle in $\mathcal{R}^2$ and on the unit sphere in $\mathcal{R}^k$. The joint prior distribution will be formed from the product of these distributions. Let

$$\lambda_{S^j} = \text{surface measure on } S^j.$$

Then for $A \subset \mathcal{R}_+$, $B \subset S^1$, $C \subset S^{k-1}$, we have

$$\psi(A \times B \times C) = \psi_1(A)\tau_1(B)\tau_{k-1}(C)$$

where

$$\tau_1(B) = \frac{\lambda_{S^1}(B)}{\lambda_{S^1}(S^1)} \quad \text{and} \quad \tau_{k-1}(C) = \frac{\lambda_{S^{k-1}}(C)}{\lambda_{S^{k-1}}(S^{k-1})}.$$

Our result on calculating the Bayes rule gives

$$d_\psi(z) = \arg\min_{a \in S^1} \int [1 - (\phi' a)^2] f(z \,|\, (\rho, \phi, \omega))\, d\psi_1(\rho)\, d\tau_1(\phi)\, d\tau_{k-1}(\omega). \tag{5.3}$$

Let $d_{\psi,\rho}$ denote the Bayes rule when the prior ψ_1 for ρ is a point mass:

$$d_{\psi,\rho}(z) = \arg\min_{a \in S^1} \int [1 - (\phi' a)^2] f(z \,|\, (\rho, \phi, \omega))\, d\tau_1(\phi)\, d\tau_{k-1}(\omega). \tag{5.4}$$

We shall see that this Bayes rule in fact does not depend upon the value of ρ as long as $\rho > 0$.

Maximum Likelihood. I want to compare the Bayes estimator in (5.4) with the maximum-likelihood estimator. The general results for maximum likelihood are based on large-sample approximations. The maximum-likelihood estimator for θ is obtained from

$$\begin{aligned}(\hat{\rho}_{\mathrm{ML}}(z), \hat{\phi}_{\mathrm{ML}}(z), \hat{\omega}_{\mathrm{ML}}(z)) &= \arg \max_{\rho \geq 0, ||\phi||=1, ||\omega||=1} f(z \,|\, (\rho, \phi, \omega)) \\ &= \arg \min_{\rho \geq 0, ||\phi||=1, ||\omega||=1} (z - \rho\phi \otimes \omega)'(z - \rho\phi \otimes \omega) \\ &= \arg \min_{\rho \geq 0, ||\phi||=1, ||\omega||=1} [z'z + \rho^2 - 2\rho z'(\phi \otimes \omega)].\end{aligned}$$

So the maximum-likelihood estimator for ϕ and ω is obtained from

$$(\hat{\phi}_{\mathrm{ML}}(z), \hat{\omega}_{\mathrm{ML}}(z)) = \arg \max_{||\phi||=1, ||\omega||=1} z'(\phi \otimes \omega).$$

Note that this is true *even if ρ is given*. The maximum-likelihood estimator for (ϕ, ω) given ρ does not depend upon ρ.

We shall use the following notation:

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \quad D(z) = \begin{pmatrix} z_1 & z_2 \end{pmatrix},$$

where z_1 and z_2 are $k \times 1$ and $D(z)$ is $k \times 2$, and

$$D(z)'D(z) = \begin{pmatrix} z_1'z_1 & z_1'z_2 \\ z_2'z_1 & z_2'z_2 \end{pmatrix}.$$

Let $Q(z) = \begin{pmatrix} q_1(z) & q_2(z) \end{pmatrix}$ be an orthogonal matrix whose columns are the eigenvectors of $D(z)'D(z)$:

$$Q(z)'(D(z)'D(z))Q(z) = \begin{pmatrix} \zeta_1(z) & 0 \\ 0 & \zeta_2(z) \end{pmatrix}, \quad Q(z)'Q(z) = I_2,$$

where the eigenvalues $\zeta_1(z)$ and $\zeta_2(z)$ are ordered so that $\zeta_1(z) \geq \zeta_2(z)$. The maximum likelihood estimator of ϕ is the eigenvector q_1 corresponding to the largest eigenvalue.[7] To see this, note that

$$\max_{\alpha \in S^1} \alpha' D(z)'D(z)\alpha = q_1(z)'D(z)'D(z)q_1(z) = \zeta_1(z).$$

A bit of algebra shows that

$$z'(\phi \otimes \omega) = (D(z)\phi)'\omega. \tag{5.5}$$

The Cauchy-Schwarz inequality gives (suppressing the z argument)

$$(D\phi)'\omega \leq [\phi'D'D\phi]^{1/2}(\omega'\omega)^{1/2} = [\phi'D'D\phi]^{1/2}.$$

[7] See Anderson and Rubin (1949) and Goldberger and Olkin (1971) for eigenvector solutions to similar problems.

So we have

$$z'(\phi \otimes \omega) = (D\phi)'\omega \leq [\phi'(D'D)\phi]^{1/2} \leq [q_1'(D'D)q_1]^{1/2} = \sqrt{\zeta}_1.$$

Setting $\phi = q_1$ and $\omega = Dq_1/\sqrt{\zeta}_1$ gives $(D\phi)'\omega = \sqrt{\zeta}_1$. So $\hat{\phi}_{\mathrm{ML}}(z) = q_1(z)$ and $\hat{\omega}_{\mathrm{ML}}(z) = D(z)q_1(z)/\sqrt{\zeta_1(z)}$.

It is shown in the Appendix that the Bayes estimator of ϕ in (5.4) does not depend upon ρ; it equals the maximum-likelihood estimator:

Theorem 5.1. For any $\rho > 0$, $d_{\psi,\rho}(z) = q_1(z)$.

So the ML estimator of ϕ in the IV model has a finite sample optimality property. If we consider estimation of (ϕ, ω) for a given value of ρ, the ML estimator of ϕ minimizes average risk for the prior distribution $\tau_1 \times \tau_{k-1}$. This ML estimator does not depend upon ρ, and equals the eigenvector q_1 corresponding to the maximum eigenvalue.

The ML estimator is not so attractive if we consider estimating ρ in the IV model. We can make this point in a simpler model, the k-means model, which we shall develop next.

5.3 *Application: k-Means Model*

One version of the k-means model has sample space $\mathcal{Z} = \mathcal{R}^k$, parameter space $\Theta = \mathcal{R}^k$, and $P_\theta = \mathcal{N}(\theta, I_k)$. The observation is a $k \times 1$ vector z, which is regarded as the realized value of the random variable Z, with

$$Z \sim \mathcal{N}(\theta, I_k) \quad \text{for some } \theta \in \Theta.$$

This is the model in Stein (1956) and James and Stein (1961). I shall use a slightly different version, following Stein (1959), in order to make connections with the IV model. The sample space is still $\mathcal{Z} = \mathcal{R}^k$, but the parameter space is $\Theta = \mathcal{R}_+ \times S^{k-1}$, and $P_\theta = \mathcal{N}(\rho\omega, I_k)$ with $\theta = (\rho, \omega)$. So now

$$Z \sim \mathcal{N}(\rho\omega, I_k) \quad \text{for some } \rho \in \mathcal{R}_+ \text{ and } \omega \in S^{k-1}.$$

The likelihood function is

$$f(z \,|\, (\rho, \omega)) = (2\pi)^{-k/2} \exp(-\frac{1}{2}||z - \rho\omega||^2).$$

The ML estimator of (ρ, ω) is obtained from

$$(\hat{\rho}_{\mathrm{ML}}(z), \hat{\omega}_{\mathrm{ML}}(z)) = \arg \min_{\rho \geq 0, ||\omega||=1} ||z - \rho\omega||.$$

We can attain the lower bound of 0 for $||z - \rho\omega||$ by setting

$$\hat{\rho}_{\mathrm{ML}}(z) = ||z||, \quad \hat{\omega}_{\mathrm{ML}}(z) = z/||z||.$$

Suppose that the loss function is

$$L(\theta, a) = (\rho^2 - a)^2$$

(with $\theta = (\rho, \omega)$). Then the risk function is the mean-square error in estimating ρ^2:

$$\begin{aligned} R(\theta, d) &= E_\theta[\rho^2 - d(Z)]^2 \\ &= \mathrm{Var}_\theta[d(Z)] + [\rho^2 - E_\theta d(Z)]^2. \end{aligned}$$

The ML estimator of ρ^2 is $\hat{\rho}^2_{\mathrm{ML}}$, with

$$\begin{aligned} E_\theta[\hat{\rho}_{\mathrm{ML}}(Z)^2] &= \sum_{j=1}^{k} E_\theta(Z_j^2) = \sum_{j=1}^{k} [\mathrm{Var}_\theta(Z_j) + (E_\theta Z_j)^2] \\ &= \sum_{j=1}^{k} (1 + \rho^2\omega_j^2) = k + \rho^2. \end{aligned}$$

So the ML estimator of ρ^2 is biased upward by the amount k. The risk function of this ML estimator is

$$R(\theta, \hat{\rho}^2_{\mathrm{ML}}) = \mathrm{Var}_\theta(\hat{\rho}^2_{\mathrm{ML}}) + k^2.$$

Consider an alternative estimator for ρ^2 that removes the bias from the ML estimator:

$$d^*(z) = \hat{\rho}_{\mathrm{ML}}(z)^2 - k.$$

Compare the risk functions for the two estimators:

$$R(\theta, d^*) = \mathrm{Var}_\theta(\hat{\rho}^2_{\mathrm{ML}}) < R(\theta, \hat{\rho}^2_{\mathrm{ML}}) \quad \text{for all } \theta \in \Theta.$$

So the ML estimator of ρ^2 is not admissible; it is dominated by d^*.

A problem with the unbiased estimator d^* is that it can be negative. This suggests using the biased estimator $\max\{d^*(z), 0\}$. I would like to compare these estimators with Bayes estimators that minimize average risk. The prior distribution is similar to the one used in the IV model: $\psi = \psi_1 \times \tau_{k-1}$, where τ_{k-1} denotes the uniform distribution on S^{k-1}, the unit sphere in $\mathcal{R}^k$. Such a uniform distribution on the unit sphere is used in Stein (1962, p. 281). The prior distribution for ρ is ψ_1; it is left general. The loss function depends on θ only through ρ: $L((\rho, \omega), a) = \tilde{L}(\rho, a)$. The Bayes estimator (if $\mathcal{D}$ is unrestricted) is obtained from

$$\begin{aligned} d_\psi(z) &= \arg\min_{a \geq 0} \int_{\mathcal{R}_+} \tilde{L}(\rho, a) \left[\int_{S^{k-1}} f(z \,|\, (\rho, \omega))\, d\tau_{k-1}(\omega) \right] d\psi_1(\rho) \\ &= \arg\min_{a \geq 0} \int_{\mathcal{R}_+} \tilde{L}(\rho, a) f_{\mathrm{I}}(z \,|\, \rho)\, d\psi_1(\rho), \end{aligned}$$

where

$$f_{\mathrm{I}}(z \,|\, \rho) = \int_{S^{k-1}} f(z \,|\, (\rho, \omega))\, d\tau_{k-1}(\omega).$$

We shall refer to f_{I} as the integrated likelihood function.[8] Note that it behaves like a standard likelihood function, in that for any $\rho \geq 0$, $f_{\mathrm{I}}(\cdot \,|\, \rho)$ is the density function for a distribution on the sample space:

$$f_{\mathrm{I}}(z \,|\, \rho) \geq 0 \quad \text{for all } z \in \mathcal{R}^k \quad \text{and} \quad \int_{\mathcal{R}^k} f_{\mathrm{I}}(z \,|\, \rho)\, dz = 1.$$

This corresponds to a statistical model in which a value for ω is drawn from the uniform distribution on S^{k-1}, and then Z is drawn from the distribution with density function $f(z \,|\, (\rho, \omega))$. So having chosen the uniform prior distribution for ω, the problem of finding a Bayes estimator reduces to a problem with a one-dimensional parameter space, with $f_{\mathrm{I}}(z \,|\, \rho)$ as the likelihood function and ψ_1 as the prior distribution for ρ.

We can simplify the evaluation of f_{I}. First note that

$$f(z \,|\, (\rho, \omega)) = (2\pi)^{-k/2} \exp[-\frac{1}{2}||z - \rho\omega||^2]$$

[8] The use of an integrated likelihood function in this model is discussed in Berger, Liseo, and Wolpert (1999, p. 9).

$$= (2\pi)^{-k/2} \exp[-\frac{1}{2}(z'z + \rho^2 - 2\rho z'\omega)]$$

$$= c(z) \exp[-\frac{1}{2}(\rho^2 - 2\rho z'\omega)],$$

where $c(z)$ denotes some function of z that does not depend upon θ. So

$$f_{\mathrm{I}}(z \mid \rho) = c(z) \exp(-\frac{1}{2}\rho^2) \int_{S^{k-1}} \exp(\rho||z||(z/||z||)'\omega) \, d\lambda_{S^{k-1}}(\omega).$$

The integral can be simplified using the following result:

$$\int_{S^N} h(\alpha'\omega) \, d\lambda_{S^N}(\omega) = \lambda_{S^{N-1}}(S^{N-1}) \int_{[-1,1]} h(s)(1-s^2)^{\frac{N}{2}-1} \, ds \tag{5.6}$$

for all $\alpha \in S^N$, $N \geq 1$, and all measurable h on $([-1,1], \mathcal{B}_{[-1,1]})$ which are either bounded or non-negative. ($\mathcal{B}_E$ denotes the Borel σ-algebra over the topological space E.) See Stroock (1999), pages 88, 89, 213–215. Applying the result in (5.6) with $k \geq 2$ gives

$$f_{\mathrm{I}}(z \mid \rho) = c(z) \exp(-\frac{1}{2}\rho^2) G_k(\rho||z||), \tag{5.7}$$

where $G_k \colon \mathcal{R}_+ \to \mathcal{R}$ is given by

$$G_k(t) = \int_{[-1,1]} \exp(ts)(1-s^2)^{(k-3)/2} \, ds. \tag{5.8}$$

Define $G_1 \colon \mathcal{R}_+ \to \mathcal{R}$ by

$$G_1(t) = \exp(t) + \exp(-t). \tag{5.9}$$

Then (5.7) also holds for $k = 1$ (with S^0 equal to $\{-1, 1\}$ and $\lambda_{S^0}\{-1\} = \lambda_{S^0}\{1\} = 1$).

Now the Bayes estimator can be obtained from

$$d_\psi(z) = \arg\min_{a \geq 0} \int_{\mathcal{R}_+} \tilde{L}(\rho, a) \exp(-\frac{1}{2}\rho^2) G_k(\rho||z||) \, d\psi_1(\rho).$$

Note that this estimator only depends on the observation z through its norm: $d_\psi(z) = \tilde{d}_\psi(||z||)$. The ML estimator and the modification of it to remove bias also depend on z only through $||z||$.

This results in a simplification of the risk function. This simplification is related to a general result in Section 6.2 on the risk function of an invariant decision rule.

Theorem 5.2. If $d(z) = \tilde{d}(||z||)$, then the risk function

$$R((\rho,\omega),d) = \int \tilde{L}(\rho, d(z)) f(z \,|\, (\rho,\omega))\, dz$$

does not depend upon ω: $R((\rho,\omega),d) = \tilde{R}(\rho,d)$.

Proof. Given any $\omega \in S^{k-1}$, there is an orthogonal matrix Q_ω such that

$$Q_\omega \omega = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \equiv e_1, \quad Q_\omega Q'_\omega = I_k,$$

where e_1 is a $k \times 1$ vector whose elements are all zero except for the first element which equals 1. If $Z \sim \mathcal{N}(\rho\omega, I_k)$, then $Q_\omega Z \sim \mathcal{N}(\rho e_1, I_k)$, and so the distribution of $Q_\omega Z$ does not depend upon ω. Since Q_ω is an orthogonal matrix, $||Q_\omega Z|| = ||Z||$. So the distribution of $||Z||$ does not depend upon ω. Hence

$$R((\rho,\omega),d) = E_{(\rho,\omega)}[\tilde{L}(\rho, \tilde{d}(||Z||)]$$

does not depend upon ω. $\diamond$

If the risk function for d does not depend upon ω, then we can use the integrated likelihood function to evaluate risk:

$$\begin{aligned} R((\rho,\omega),d) = \tilde{R}(\rho,d) &= \int_{S^{k-1}} \tilde{R}(\rho,d)\, d\tau_{k-1}(\omega) \qquad (5.10) \\ &= \int_{S^{k-1}} \left[\int_{\mathcal{Z}} \tilde{L}(\rho, d(z)) f(z \,|\, (\rho,\omega))\, dz \right] d\tau_{k-1}(\omega) \\ &= \int_{\mathcal{Z}} \tilde{L}(\rho, d(z)) \left[\int_{S^{k-1}} f(z \,|\, (\rho,\omega))\, d\tau_{k-1}(\omega) \right] dz \\ &= \int_{\mathcal{Z}} \tilde{L}(\rho, d(z)) f_{\mathrm{I}}(z \,|\, \rho)\, dz. \end{aligned}$$

So we can use the integrated likelihood function to evaluate the risk of an invariant decision rule. This suggests that large-sample approximations based on a likelihood function can be applied using the integrated likelihood function, even though we are using the original model with parameter space $\mathcal{R}_+ \times S^{k-1}$ and likelihood function $f(z \,|\, (\rho, \omega))$. To the extent that we can treat f_{I} as a likelihood function, there is reason to expect that large-sample approximations will be more accurate with f_{I} than with f, since the parameter space for f_{I} has lower dimension. This difference could be dramatic if k is large.

6. MINIMAX AND INVARIANCE

6.1. *Minimax in the k-Means Model*

A Bayes decision rule d_ψ requires a prior distribution ψ on the parameter space Θ. A careful, thoughtful specification for this distribution may be sufficiently costly that one is interested in alternative criteria for working with a risk function. We still face the basic issue that risk depends upon the distribution P_θ our observation z came from, and our statistical model only assumes that θ is some point in the parameter space Θ. An alternative to average risk is maximum risk. We reduce the risk function $R(\cdot, d)$ to a single number (for a given d) by maximizing over the parameter space:

$$\sup_{\theta \in \Theta} R(\theta, d).$$

Now we have a real-valued objective function, and we can try to solve the problem

$$\min_{d \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, d).$$

In Section 5, we considered estimating ρ in the k-means model:

$$Z \sim \mathcal{N}(\rho\omega, I_k) \quad \text{for some } \rho \in \mathcal{R}_+ \text{ and } \omega \in S^{k-1}.$$

We developed the Bayes estimator

$$d_\psi = \arg\min_{d \in \mathcal{D}} \int R(\theta, d)\, d\psi(\theta)$$

for the prior distribution $\psi = \psi_1 \times \tau_{k-1}$, where ψ_1 is some distribution on $\mathcal{R}_+$ and τ_{k-1} is the uniform distribution on S^{k-1}. The loss function depended on $\theta = (\rho, \omega)$ only through ρ: $L((\rho, \omega), a)$

$= \tilde{L}(\rho, a)$. We showed that, for any choice of the prior distribution ψ_1, the risk function for this Bayes estimator does not depend upon ω: $R((\rho, \omega), d_\psi) = \tilde{R}(\rho, d_\psi)$. These results lead to a simple argument that relates this Bayes estimator to minimax.[9]

Theorem 6.1. If $R((\rho, \omega), d_\psi) = \tilde{R}(\rho, d_\psi)$, then d_ψ solves the following problem, which combines the average risk and maximum risk criteria:

$$d_\psi = \arg\min_{d \in \mathcal{D}} \int_{\mathcal{R}_+} [\sup_{\omega \in S^{k-1}} R((\rho, \omega), d)]\, d\psi_1(\rho).$$

Proof. For any $d \in \mathcal{D}$,

$$\begin{aligned}
&\int_{\mathcal{R}_+} [\sup_{\omega \in S^{k-1}} R((\rho, \omega), d)]\, d\psi_1(\rho) \\
&\qquad \geq \int_{\mathcal{R}_+} \left[\int_{S^{k-1}} R((\rho, \omega), d)\, d\tau_{k-1}(\omega) \right] d\psi_1(\rho) \\
&\qquad \geq \int_{\mathcal{R}_+} \left[\int_{S^{k-1}} R((\rho, \omega), d_\psi)\, d\tau_{k-1}(\omega) \right] d\psi_1(\rho) \\
&\qquad = \int_{\mathcal{R}_+} \tilde{R}(\rho, d_\psi)\, d\psi_1(\rho) \\
&\qquad = \int_{\mathcal{R}_+} [\sup_{\omega \in S^{k-1}} R((\rho, \omega), d_\psi)]\, d\psi_1(\rho). \quad \diamond
\end{aligned}$$

The use of minimax here does not eliminate the choice of a prior distribution; the average risk criteria on the parameter space $\mathcal{R}_+$ for ρ requires that we specify a prior distribution ψ_1. But we can replace the choice of a prior distribution on the parameter space S^{k-1} for ω by the maximum risk criterion. It turns out that the solution to the minimax problem calls for a particular, least-favorable, distribution on S^{k-1}: the uniform distribution τ_{k-1}.

A key part of the argument is that the risk function for d_ψ simplifies: the risk at $\theta = (\rho, \omega)$ is constant for all θ with the same value for ρ. There is a method, based on invariance, for simplifying a risk function in this way. This method is developed next; the approach is based on Ferguson (1967,

[9] The proof is based on Ferguson (1967, Theorem 1, p. 90).

Chapter 4). For a general treatment of invariant prior distributions and their role in invariant decision problems, see Eaton (1989).

6.2. *Invariance*

In Section 3, we saw that a statistical decision problem involves three spaces: sample space, parameter space, and action space. Invariance involves transformations on each of these three spaces. The transformations are connected through an index set G:

$$m_1: G \times \mathcal{Z} \to \mathcal{Z}$$

$$m_2: G \times \Theta \to \Theta$$

$$m_3: G \times \mathcal{A} \to \mathcal{A}.$$

For each element $g \in G$, $m_1(g, \cdot)$ maps the sample space $\mathcal{Z}$ into $\mathcal{Z}$; $m_2(g, \cdot)$ maps the parameter space Θ into Θ; and $m_3(g, \cdot)$ maps the action space $\mathcal{A}$ into $\mathcal{A}$.

Invariant Model. The statistical model $(P: \Theta \to \mathcal{P}(\mathcal{Z}))$ is invariant if, for any $g \in G$ and $\theta \in \Theta$, $Z \sim P_\theta$ implies that $m_1(g, Z) \sim P_{m_2(g,\theta)}$.

Invariant Loss. The loss function is invariant if, for all $g \in G$, $\theta \in \Theta$, and $a \in \mathcal{A}$,

$$L(m_2(g,\theta), m_3(g,a)) = L(\theta, a).$$

Invariant Decision Rule. The decision rule $d \in \mathcal{D}$ is invariant if, for all $g \in G$ and $z \in \mathcal{Z}$,

$$d(m_1(g,z)) = m_3(g, d(z)).$$

Theorem 6.2.[10] If the statistical model $P: \Theta \to \mathcal{P}(\mathcal{Z})$, the loss function L, and the decision rule d are invariant, then, for all $g \in G$ and $\theta \in \Theta$,

$$R(\theta, d) = R(m_2(g,\theta), d).$$

[10] See Ferguson (1967, Theorem 1, p. 150).

Proof.

$$\begin{aligned}R(\theta, d) &= \int_{\mathcal{Z}} L(\theta, d(z))\, dP_\theta(z) = E_\theta[L(\theta, d(Z))] \\ &= E_\theta[L(m_2(g,\theta), m_3(g, d(Z)))] \\ &= E_\theta[L(m_2(g,\theta), d(m_1(g,Z)))] \\ &= E_{m_2(g,\theta)}[L(m_2(g,\theta), d(Z))] \\ &= R(m_2(g,\theta), d). \quad \diamond\end{aligned}$$

The theorem shows how invariance leads to a simplification of the risk function. Risk for an invariant decision rule is constant for all parameter values of the form $m_2(g,\theta)$ as g varies over the index set G.

6.3. *Application: k-Means Model*

At the end of Section 5, we obtained a simplification of the risk function for estimating ρ in the k-means model. This simplification can also be obtained by applying the invariance theorem. The index set G is the set of $k \times k$ orthogonal matrices:

$$G = O(k) = \{k \times k \text{ matrices } g : gg' = I_k\}.$$

The transformation on the sample space is

$$m_1(g, z) = gz \quad \text{for} \quad z \in \mathcal{R}^k.$$

If $Z \sim P_\theta = \mathcal{N}(\rho\omega, I_k)$ (with $\rho \in \mathcal{R}_+$ and $\omega \in S^{k-1}$), then

$$m_1(g, Z) = gZ \sim \mathcal{N}(\rho g\omega, I_k) = P_{m_2(g,\theta)}$$

with

$$m_2(g, (\rho, \omega)) = (\rho, g\omega).$$

So the model is invariant. The loss function for estimating ρ does not depend upon ω: $L((\rho,\omega), a) = \tilde{L}(\rho, a)$. So we can set $m_3(g, a) = a$ and L is invariant. Then an estimator d is invariant if $d(gz) = d(z)$ for all $g \in O(k)$ and $z \in \mathcal{R}^k$.

For any $z \in \mathcal{R}^k$, there is a $g_z \in O(k)$ such that

$$g_z z = ||z|| \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = ||z|| e_1,$$

where e_1 is a vector whose elements are all 0 except for the first element which equals 1; here e_1 is $k \times 1$. So if d is an invariant estimator, then $d(z) = d(g_z z) = d(||z|| e_1) \equiv \tilde{d}(||z||)$. If $d(z)$ only depends on $||z||$, so that $d(z) = \tilde{d}(||z||)$, then d is invariant since $d(gz) = \tilde{d}(||gz||) = \tilde{d}(||z||) = d(z)$. So d is invariant if and only if $d(z) = \tilde{d}(||z||)$.

If d is invariant, then Theorem 6.2 implies that

$$R((\rho, \omega), d) = R((\rho, g_\omega \omega), d) = R((\rho, e_1), d) \equiv \tilde{R}(\rho, d).$$

So the risk function does not depend upon ω.

6.4. *Application: IV-Model*

The index set is the Cartesian product of $O(2)$ and $O(k)$:

$$G = O(2) \times O(k) = \{(g_1, g_2) : g_1 g_1' = I_2,\ g_2 g_2' = I_k\}.$$

The transformation on the sample space is (with $g = (g_1, g_2)$)

$$m_1(g, z) = (g_1 \otimes g_2) z \quad \text{for} \quad z \in \mathcal{R}^{2k}.$$

If $Z \sim P_\theta = \mathcal{N}(\rho\phi \otimes \omega, I_{2k})$ (with $\rho \in \mathcal{R}_+$, $\phi \in S^1$, $\omega \in S^{k-1}$), then

$$m_1(g, Z) \sim \mathcal{N}(\rho g_1 \phi \otimes g_2 \omega, I_{2k}) = P_{m_2(g,\theta)}$$

with

$$m_2(g, (\rho, \phi, \omega)) = (\rho, g_1\phi, g_2\omega).$$

So the model is invariant. Let $m_3(g, a) = g_1 a$ for $a \in S^1$. Then

$$L(m_2(g, \theta), m_3(g, a)) = 1 - [(g_1\phi)' g_1 a]^2 = 1 - (\phi' a)^2,$$

and the loss function is invariant. An estimator $d\colon \mathcal{R}^{2k} \to S^1$ is invariant if $d((g_1 \otimes g_2)z) = g_1 d(z)$ for all $g_1 \in O(2)$, $g_2 \in O(k)$, and $z \in \mathcal{R}^{2k}$.

Given any $\phi \in S^1$, there is a $g_\phi \in O(2)$ such that $g_\phi \phi = (\,1 \quad 0\,)'$; given any $\omega \in S^{k-1}$, there is a $g_\omega \in O(k)$ such that $g_\omega \omega = (\,1 \quad 0 \quad \ldots \quad 0\,)' \equiv e_1$. So Theorem 6.2 implies that if d is invariant, then

$$R((\rho, \phi, \omega), d) = R((\rho, g_\phi \phi, g_\omega \omega), d) = R((\rho, (\,1 \quad 0\,)', e_1), d) \equiv \tilde{R}(\rho, d).$$

So the risk function for an invariant estimator depends upon $\theta = (\rho, \phi, \omega)$ only through ρ.

Theorem 6.3. $\hat{\phi}_{\text{ML}}$ is invariant.

Proof. For all $g_1 \in O(2)$, $g_2 \in O(k)$, and $z \in \mathcal{R}^{2k}$,

$$\begin{aligned} ||(g_1 \otimes g_2)z - \rho\phi \otimes \omega|| &= ||(g_1 \otimes g_2)(z - \rho g_1^{-1}\phi \otimes g_2^{-1}\omega)|| \\ &= ||z - \rho g_1^{-1}\phi \otimes g_2^{-1}\omega||. \end{aligned}$$

Since every point in S^{k-1} equals $g_2^{-1}\omega$ for some $\omega \in S^{k-1}$,

$$\min_{\rho \in \mathcal{R}_+} \min_{\omega \in S^{k-1}} ||z - \rho g_1^{-1}\phi \otimes g_2^{-1}\omega|| = \min_{\rho \in \mathcal{R}_+} \min_{\omega \in S^{k-1}} ||z - \rho g_1^{-1}\phi \otimes \omega||.$$

If ϕ^* is a value for ϕ that minimizes

$$\min_{\rho \in \mathcal{R}_+} \min_{\omega \in S^{k-1}} ||z - \rho\phi \otimes \omega||,$$

then $g_1\phi^*$ is a value for ϕ that minimizes

$$\min_{\rho \in \mathcal{R}_+} \min_{\omega \in S^{k-1}} ||z - \rho g_1^{-1}\phi \otimes \omega||.$$

So

$$\hat{\phi}_{\text{ML}}((g_1 \otimes g_2)z) = g_1 \hat{\phi}_{\text{ML}}(z). \quad \diamond$$

We saw in Section 5.2 that if we consider estimation of (ϕ, ω) for a given value of ρ, then the ML estimator of ϕ does not depend upon ρ. If the prior distribution for (ϕ, ω) is $\tau = \tau_1 \times \tau_{k-1}$, which is a product of uniform distributions on S^1 and S^{k-1}, then Theorem 5.1 shows that the Bayes estimator, with $\rho > 0$ given, equals this ML estimator (= the eigenvector q_1 corresponding to the maximum eigenvalue). Given ρ, the risk of the ML estimator is constant, since $\hat{\phi}_{\mathrm{ML}}$ is invariant. Since $\hat{\phi}_{\mathrm{ML}}$ is a Bayes estimator with constant risk, it is a minimax estimator.

Theorem 6.4. Given $\rho > 0$, $\hat{\phi}_{\mathrm{ML}}$ is minimax: for any estimator $d \in \mathcal{D}$,

$$\sup_{\phi \in S^1} \sup_{\omega \in S^{k-1}} R((\rho, \phi, \omega), d) \geq \sup_{\phi \in S^1} \sup_{\omega \in S^{k-1}} R((\rho, \phi, \omega), \hat{\phi}_{\mathrm{ML}}).$$

Proof.

$$\begin{aligned} \sup_{(\phi,\omega) \in S^1 \times S^{k-1}} R((\rho, \phi, \omega), d) &\geq \int_{S^1 \times S^{k-1}} R((\rho, \phi, \omega), d)\, d\tau(\phi, \omega) \\ &\geq \int_{S^1 \times S^{k-1}} R((\rho, \phi, \omega), \hat{\phi}_{\mathrm{ML}})\, d\tau(\phi, \omega) \\ &= \sup_{(\phi,\omega) \in S^1 \times S^{k-1}} R((\rho, \phi, \omega), \hat{\phi}_{\mathrm{ML}}). \quad \diamond \end{aligned}$$

Here the use of minimax eliminates the choice of a prior distribution (although a particular prior distribution emerges in the solution to the minimax problem). For any given positive value of ρ, the ML estimator of ϕ is optimal in the minimax sense. Since the ML estimator does not depend upon ρ, it remains a feasible estimator when ρ is not given.

This minimax solution depends upon the loss function, and the solution is associated with a least favorable prior distribution. In a particular context, such as the example in Section 2 using earnings and education, this loss function may not be appealing and the least favorable prior distribution may not be subjectively plausible. Minimax solutions could be computed for other loss functions, but they may not have an explicit, closed form. Apart from tractability, an argument for the loss function I have chosen, and for the least favorable prior in the minimax solution, is that they do not depend upon a specific context, and so can provide an automatic procedure for generating priors and associated decision rules. In a specific context, more relevant loss functions and prior distributions may certainly be available.

7. CONFIDENCE SETS

7.1. *Confidence Interval for ϕ in the IV Model*

A confidence set can be regarded as a decision rule where the action space consists of subsets of the parameter space. The parameter space for ϕ in the IV model is the unit circle S^1. So the action space $\mathcal{A}$ will consist of subsets of S^1: $a \subset S^1$ for $a \in \mathcal{A}$. We shall use the following loss function:

$$L((\rho, \phi, \omega), a) = \begin{cases} 0, & \text{if } \phi \in a; \\ 1, & \text{if } \phi \notin a \end{cases}. \tag{7.1}$$

Let $\theta = (\rho, \phi, \omega)$. The corresponding risk function is

$$\begin{aligned} R(\theta, d) &= P_\theta(\phi \notin d(Z)) = 1 - P_\theta(\phi \in d(Z)) \\ &= 1 - \text{coverage rate.} \end{aligned}$$

In Section 7.4, we shall extend this loss function to depend also on the length of a.

Invariance arguments can be used to simplify the risk function. As in Section 6.4, the index set is the Cartesian product of the set of 2×2 orthogonal matrices and the set of $k \times k$ orthogonal matrices:

$$G = O(2) \times O(k) = \{(g_1, g_2) : g_1 g_1' = I_2,\ g_2 g_2' = I_k\}.$$

The transformation on the sample space is (with $g = (g_1, g_2)$)

$$m_1(g, z) = (g_1 \otimes g_2) z \quad \text{for} \quad z \in \mathcal{R}^{2k}.$$

The transformation on the parameter space is

$$m_2(g, (\rho, \phi, \omega)) = (\rho, g_1 \phi, g_2 \omega).$$

The model is invariant under these transformations on the sample space and the parameter space. The transformation on the action space is

$$m_3(g, a) = g_1 a = \{g_1 s : s \in a\} \quad \text{for} \quad a \subset S^1.$$

So $m_3(g,\cdot)$ maps subsets of S^1 into subsets of S^1, by multiplying each element of the subset by g_1 (which preserves the unit length and so gives a point in S^1). Then the loss function is invariant: for $a \subset S^1$,

$$\phi \in a \quad \text{iff} \quad g_1\phi \in g_1 a,$$

and so

$$L((\rho,\phi,\omega),a) = L((\rho, g_1\phi, g_2\omega), g_1 a).$$

A decision rule $d \in \mathcal{D}$ is invariant if

$$d((g_1 \otimes g_2)z) = g_1 d(z) = \{g_1 s : s \in d(z)\}.$$

Given any $\phi \in S^1$, there is a $g_\phi \in O(2)$ such that $g_\phi\phi = (\,1 \quad 0\,)'$; given any $\omega \in S^{k-1}$, there is a $g_\omega \in O(k)$ such that $g_\omega\omega = (\,1 \quad 0 \quad \ldots \quad 0\,)' \equiv e_1$. So Theorem 6.2 implies that if d is invariant, then

$$R((\rho,\phi,\omega),d) = R((\rho, g_\phi\phi, g_\omega\omega),d) = R((\rho,(\,1 \quad 0\,)', e_1),d) \equiv \tilde{R}(\rho,d).$$

So the coverage rate for an invariant confidence set depends upon $\theta = (\rho,\phi,\omega)$ only through ρ.

I should stress that the loss function here is only getting at part of the problem. Coverage is important and is often the focus in evaluating the performance of a confidence set procedure. But keep in mind that if coverage were the *only* consideration, then we could achieve a .95 (for example) coverage rate without using the data. Simply use a randomized decision rule that selects S^1 95% of the time, and selects the null set 5% of the time.

7.2. *Invert a Likelihood Ratio Test*

Since our loss function only gets at part of the problem, we shall not try to develop confidence sets that are optimal under this loss function. Instead we shall focus on some particular confidence set procedures, and use the invariance result to simplify evaluation of their finite sample coverage rates. The procedures involve inverting a likelihood ratio test. Inverting the standard likelihood ratio test gives the following set:

$$d_{\text{LR}}(z) = \left\{ s \in S^1 : \frac{\max_{\rho \ge 0, \phi \in S^1, \omega \in S^{k-1}} f(z \,|\, (\rho,\phi,\omega))}{\max_{\rho \ge 0, \omega \in S^{k-1}} f(z \,|\, (\rho, s, \omega))} \le \text{crit} \right\}.$$

The idea is that if $\phi = s$ is the null hypothesis, then we compare the maximized likelihood, with no restrictions, to the maximized likelihood with ϕ restricted to equal s. This ratio of maximized likelihoods has to be greater than or equal to one. If it is bigger than a critical value (crit), then we conclude that the restriction is not favored by the data and we reject the hypothesis. Then the confidence set is simply the values for the null hypothesis that are not rejected.

If we were only comparing two values for ϕ, say $\phi = s_0$ and $\phi = s_1$, with ρ and ω given, then we literally would have a likelihood ratio: $f(z \,|\, (\rho, s_1, \omega))/f(z \,|\, (\rho, s_0, \omega))$, and the likelihood ratio test would have a finite sample optimality property (Neyman-Pearson Lemma). A test based on a ratio of *maximized* likelihoods does not, in general, have finite-sample optimality properties. It does, however, have large-sample optimality properties and is widely used (and is commonly referred to as a likelihood ratio test).[11] A common choice of the critical value is based on a large-sample approximation under which two times the log of the ratio of maximized likelihoods has a chi-square distribution under the null hypothesis; the degrees of freedom for the chi-square distribution is the difference in dimensions between the unrestricted parameter space and the restricted parameter space. So in our case it would be a chi-square distribution with one degree of freedom. To achieve a .95 coverage rate, the approximation suggests setting crit so that $2\log(\text{crit})$ equals the .95 quantile of a $\chi^2(1)$ distribution ($= 3.8415$), or crit $= 6.83$.

We have seen in Section 6 that a uniform prior distribution (for ω) on S^{k-1} leads to estimators with invariance and minimax properties. So, for $k \geq 2$, it might be of interest to consider a likelihood ratio test based on the following integrated likelihood function:

$$f_{\mathrm{I}}(z \,|\, (\rho, \phi)) = \int_{S^{k-1}} f(z \,|\, (\rho, \phi, \omega)) \, d\tau_{k-1}(\omega), \tag{7.2}$$

where τ_{k-1} is the uniform distribution on S^{k-1}. This is the likelihood function for the statistical model in which a value for ω is drawn from the uniform distribution on S^{k-1}, and then Z is drawn from the distribution with density function $f(z \,|\, (\rho, \phi, \omega))$. We shall be evaluating risk under the original model with likelihood function $f(z \,|\, (\rho, \phi, \omega))$. But if a procedure based on the integrated

[11] See, for example, van der Vaart (1998, Chapter 16).

likelihood function turns out to have a risk function that does not depend upon ω, then its risk function under the original model coincides with its risk function under the integrated model. So good properties under the integrated model will carry over to the original model.

Inverting the integrated likelihood ratio test gives the following set:

$$d_{\mathrm{I,LR}}(z) = \left\{ s \in S^1 : \frac{\max_{\rho \geq 0, \phi \in S^1} f_{\mathrm{I}}(z \,|\, (\rho, \phi))}{\max_{\rho \geq 0} f_{\mathrm{I}}(z \,|\, (\rho, s))} \leq \mathrm{crit} \right\}.$$

7.3. *Invariance*

We shall show that d_{LR} and $d_{\mathrm{I,LR}}$ are both invariant. It then follows that the coverage rates only depend upon ρ. This greatly simplifies the numerical evaluation of finite-sample coverage rates. We can set ϕ and ω at arbitrary values such as $\phi = (\,1 \quad 0\,)'$ and $\omega = e_1'$. Choose a set of values for ρ. For each value in this set, form $\theta = (\rho, \phi, \omega)$ and generate a large number of independent samples from P_θ: $Z^j \sim P_\theta$ $(j = 1, \ldots, J)$. Then the coverage rate for that value of ρ is approximately the fraction of the samples for which $\phi \in d(Z^j)$. The accuracy of this Monte Carlo approximation becomes arbitrarily good as the number of Monte Carlo samples $J \to \infty$.

Theorem 7.1. d_{LR} is invariant.

Proof. To evaluate $d_{\mathrm{LR}}((g_1 \otimes g_2)z)$, note that

$$f(z \,|\, (\rho, \phi, \omega)) = (2\pi)^{-k} \exp(-\frac{1}{2} ||z - \rho\phi \otimes \omega||^2),$$

$$||(g_1 \otimes g_2)z - \rho\phi \otimes \omega|| = ||(g_1 \otimes g_2)(z - \rho g_1^{-1}\phi \otimes g_2^{-1}\omega)|| = ||z - \rho g_1^{-1}\phi \otimes g_2^{-1}\omega||.$$

Since multiplication by the orthogonal matrix g_1^{-1} maps S^1 onto itself, and multiplication by the orthogonal matrix g_2^{-1} maps S^{k-1} onto itself,

$$\min_{\rho \geq 0, \phi \in S^1, \omega \in S^{k-1}} ||z - \rho g_1^{-1}\phi \otimes g_2^{-1}\omega|| = \min_{\rho \geq 0, \phi \in S^1, \omega \in S^{k-1}} ||z - \rho\phi \otimes \omega||$$

and

$$\min_{\rho \geq 0, \omega \in S^{k-1}} ||z - \rho g_1^{-1} s \otimes g_2^{-1}\omega|| = \min_{\rho \geq 0, \omega \in S^{k-1}} ||z - \rho g_1^{-1} s \otimes \omega||.$$

So

$$s \in d_{\mathrm{LR}}((g_1 \otimes g_2)z) \quad \text{iff} \quad g_1^{-1}s \in d_{\mathrm{LR}}(z) \quad \text{iff} \quad s \in g_1 d_{\mathrm{LR}}(z).$$

Hence

$$d_{\mathrm{LR}}((g_1 \otimes g_2)z) = g_1 d_{\mathrm{LR}}(z). \quad \diamond$$

Theorem 7.2. $d_{\mathrm{I,LR}}$ is invariant.

Proof. The likelihood function simplifies:

$$f(z \,|\, (\rho, \phi, \omega)) = c(z) \exp[-\rho^2/2 + \rho z'(\phi \otimes \omega)],$$

where we shall use $c(z)$ to denote any function of z that does not depend upon θ. So we need to evaluate

$$\int \exp(\rho z'(\phi \otimes \omega)) \, d\lambda_{S^{k-1}}(\omega).$$

Note that, as in (5.5),

$$z'(\phi \otimes \omega) = (D(z)\phi)'\omega,$$

where $D(z) = (\, z_1 \quad z_2 \,)$. The integral can be simplified by applying (5.6) (with $k \geq 2$):

$$\int \exp(\rho z'(\phi \otimes \omega)) \, d\lambda_{S^{k-1}}(\omega) = \lambda_{S^{k-2}}(S^{k-2}) G_k(\rho ||D\phi||),$$

where $G_k \colon \mathcal{R}_+ \to \mathcal{R}$ is given by

$$G_k(t) = \int_{[-1,1]} \exp(ts)(1 - s^2)^{(k-3)/2} \, ds.$$

For $k = 1$, use

$$G_1(t) = \exp(t) + \exp(-t).$$

Then we have

$$f_{\mathrm{I}}(z \,|\, (\rho, \phi)) = c(z) \exp(-\rho^2/2) G_k(\rho ||D(z)\phi||). \tag{7.3}$$

Note that

$$D((g_1 \otimes g_2)z) = g_2 D(z) g_1' \quad \text{and} \quad ||g_2 D(z) g_1' \phi|| = ||D(z) g_1^{-1} \phi||. \tag{7.4}$$

Hence

$$\begin{aligned}\max_{\rho \geq 0, \phi \in S^1} f_{\mathrm{I}}((g_1 \otimes g_2)z \,|\, (\rho, \phi)) &= c((g_1 \otimes g_2)z) \max_{\rho \geq 0, \phi \in S^1} \exp(-\rho^2/2) G_k(\rho ||D(z) g_1^{-1} \phi||) \\ &= c((g_1 \otimes g_2)z) \max_{\rho \geq 0, \phi \in S^1} \exp(-\rho^2/2) G_k(\rho ||D(z) \phi||)\end{aligned}$$

and

$$\max_{\rho \geq 0} f_{\mathrm{I}}((g_1 \otimes g_2)z \,|\, (\rho, s)) = c((g_1 \otimes g_2)z) \max_{\rho \geq 0} \exp(-\rho^2/2) G_k(\rho ||D(z) g_1^{-1} s||).$$

So

$$s \in d_{\mathrm{I,LR}}((g_1 \otimes g_2)z) \quad \text{iff} \quad g_1^{-1} s \in d_{\mathrm{I,LR}}(z) \quad \text{iff} \quad s \in g_1 d_{\mathrm{I,LR}}(z),$$

which implies that

$$d_{\mathrm{I,LR}}((g_1 \otimes g_2)z) = g_1 d_{\mathrm{I,LR}}(z). \quad \diamond$$

If d is an invariant decision rule, then its risk function does not depend upon ϕ or ω. Since it does not depend upon ω, we have

$$R((\rho, \phi, \omega), d) = \int_{\mathcal{Z}} \tilde{L}(\phi, d(z)) f_{\mathrm{I}}(z \,|\, (\rho, \phi)) \, dz$$

(as in (5.10)). So we can use the integrated likelihood function to evaluate the risk of an invariant decision rule. This suggests that large-sample approximations based on a likelihood function can be applied using the integrated likelihood function, even though we are using the original model with parameter space $\mathcal{R}_+ \times S^1 \times S^{k-1}$ and likelihood function $f(z \,|\, (\rho, \phi, \omega))$. In particular, this suggests setting the critical value for $d_{\mathrm{I,LR}}$ based on the same $\chi^2(1)$ approximation used for d_{LR}. To the extent that we can treat f_{I} as a likelihood function, there is reason to expect that large-sample approximations will be more accurate with f_{I} than with f, since the parameter space for f_{I} has lower dimension.

7.4 *Posterior Interval*

Consider the following loss function:

$$L_b((\rho,\phi,\omega),a) = 1(\phi \notin a) + b\int_a d\lambda_{S^1}, \tag{7.5}$$

where b is a nonnegative number. The loss is b times the length of the set plus an indicator function that equals one if the set fails to cover ϕ and equals zero otherwise. The loss function in (7.1) is a special case with $b = 0$. As in Section 5.2, consider the prior distribution $\psi_1 \times \tau$, where ψ_1 is the prior distribution for ρ and the prior distribution for (ϕ,ω) is $\tau = \tau_1 \times \tau_{k-1}$, the uniform distribution on $S^1 \times S^{k-1}$. The marginal posterior distribution for ϕ has a density $\bar{\pi}_2(\phi \,|\, z)$ with respect to λ_{S^1}:

$$\bar{\pi}_2(\phi \,|\, z) = \int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho,\phi))\, d\psi_1(\rho) \Big/ \int_{S^1}\int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho,\phi))\, d\psi_1(\rho)\, d\lambda_{S^1}(\phi),$$

where f_{I} from (7.2) is the integrated likelihood based on the uniform distribution for ω on S^{k-1}.

With the loss function in (7.5), the posterior expected loss is

$$1 - \int_a [\bar{\pi}_2(\phi \,|\, z) - b]\, d\lambda_{S^1}(\phi).$$

The Bayes rule d_ψ is obtained by minimizing the posterior expected loss; this gives a highest posterior density set:

$$d_\psi(z) = \{\phi \in S^1 : \bar{\pi}_2(\phi \,|\, z) \geq b\}.$$

Theorem 7.3. d_ψ is invariant.

Proof. For any $g_1 \in O(2)$ and $g_2 \in O(k)$, it follows from (7.3) and (7.4) that

$$\begin{aligned}
&\bar{\pi}_2(\phi \,|\, (g_1 \otimes g_2)z) \\
&\quad = \int_{\mathcal{R}_+} f_{\mathrm{I}}((g_1 \otimes g_2)z \,|\, (\rho,\phi))\, d\psi_1(\rho) \Big/ \int_{S^1}\int_{\mathcal{R}_+} f_{\mathrm{I}}((g_1 \otimes g_2)z \,|\, (\rho,\phi))\, d\psi_1(\rho)\, d\lambda_{S^1}(\phi) \\
&\quad = \int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho, g_1^{-1}\phi))\, d\psi_1(\rho) \Big/ \int_{S^1}\int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho, g_1^{-1}\phi))\, d\psi_1(\rho)\, d\lambda_{S^1}(\phi).
\end{aligned}$$

If ϕ has a uniform distribution on S^1, then $g_1^{-1}\phi$ also has a uniform distribution on S^1. Hence

$$\begin{aligned}\bar{\pi}_2(\phi \,|\, (g_1 \otimes g_2)z) &= \int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho, g_1^{-1}\phi))\, d\psi_1(\rho) \Big/ \int_{S^1}\int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho, \phi))\, d\psi_1(\rho)\, d\lambda_{S^1}(\phi) \\ &= \bar{\pi}_2(g_1^{-1}\phi \,|\, z).\end{aligned}$$

So

$$s \in d_\psi((g_1 \otimes g_2)z) \quad \text{iff} \quad g_1^{-1}s \in d_\psi(z) \quad \text{iff} \quad s \in g_1 d_\psi(z),$$

which implies that

$$d_\psi((g_1 \otimes g_2)z) = g_1 d_\psi(z). \quad \diamond$$

Then, as in the proof of Theorem 6.1,

$$d_\psi = \arg\min_{d \in \mathcal{D}} \int_{\mathcal{R}_+} [\sup_{(\phi,\omega) \in S^1 \times S^{k-1}} R((\rho, \phi, \omega), d)]\, d\psi_1(\rho).$$

The use of minimax here does not eliminate the choice of a prior distribution; the average risk criteria on the parameter space $\mathcal{R}_+$ for ρ requires that we specify a prior distribution ψ_1. But we can replace the choice of a prior distribution on the parameter space $S^1 \times S^{k-1}$ for (ϕ, ω) by the maximum risk criterion.

8. HYPOTHESIS TESTS

8.1. *Bayes Tests*

The statistical model specifies that

$$Z \sim P_\theta \quad \text{for some} \quad \theta \in \Theta.$$

The hypothesis H is a subset of the parameter space: $H \subset \Theta$. The only actions available are to accept H $(a = 0)$ or to reject H $(a = 1)$. So the action space is $\mathcal{A} = \{0, 1\}$. We shall use the following loss function:

$$L(\theta, a) = \begin{cases} 0, & \text{if } \theta \in H, a = 0; \\ 1, & \text{if } \theta \in H, a = 1; \\ b, & \text{if } \theta \notin H, a = 0; \\ 0, & \text{if } \theta \notin H, a = 1, \end{cases}$$

with $b > 0$. There are two types of error: rejecting H when it is true (type 1 error) and accepting H when it is false (type 2 error); the magnitude of b reflects the relative importance of these two types of error.

This loss function implies the following risk function:

$$R(\theta, d) = E_\theta[L(\theta, d(Z))] = \begin{cases} P_\theta\{d(Z) = 1\}, & \text{if } \theta \in H; \\ b[1 - P_\theta\{d(Z) = 1\}], & \text{if } \theta \notin H. \end{cases}$$

The decision rule or test d corresponds to a critical region: $\{z \in \mathcal{Z} : d(z) = 1\}$; this is the subset of the sample space where the test rejects the hypothesis. The probability that P_θ attaches to this set is the power function of the test: $P_\theta\{d(Z) = 1\}$. The risk function depends upon the statistical model only through this power function.

Given a prior distribution ψ on the parameter space Θ, the average risk is

$$\begin{aligned} R^*(\psi, d) &= \int_\Theta R(\theta, d)\, d\psi(\theta) \\ &= \int_H P_\theta\{d(Z) = 1\}\, d\psi(\theta) + b \int_{\Theta - H} [1 - P_\theta\{d(Z) = 1\}]\, d\psi(\theta). \end{aligned}$$

A Bayes test corresponding to the prior distribution ψ minimizes the average risk:

$$d_\psi = \arg\min_{d \in \mathcal{D}} R^*(\psi, d).$$

If the set of feasible tests $\mathcal{D}$ is unrestricted, then the Bayes test can be obtained by minimizing posterior expected loss. This gives

$$d_\psi(z) = \arg\min_{a \in \{0,1\}} \int_\Theta L(\theta, a) f(z \,|\, \theta)\, d\psi(\theta).$$

Note that

$$\int_\Theta L(\theta, 0) f(z \,|\, \theta)\, d\psi(\theta) = b \int_{\Theta - H} f(z \,|\, \theta)\, d\psi(\theta)$$

and

$$\int_\Theta L(\theta, 1) f(z \,|\, \theta)\, d\psi(\theta) = \int_H f(z \,|\, \theta)\, d\psi(\theta).$$

So

$$d_\psi(z) = 1 \quad \text{if} \quad \frac{\int_H f(z \,|\, \theta)\, d\psi(\theta)}{\int_{\Theta - H} f(z \,|\, \theta)\, d\psi(\theta)} \le b, \tag{8.1}$$

and $d_\psi(z) = 0$ otherwise.[12]

8.2. *Neyman-Pearson Lemma*

Suppose that the parameter space consists of two points: $\Theta = \{\theta_H, \theta_J\}$ and $H = \{\theta_H\}$. Then the Bayes test in (8.1) becomes

$$d_\psi(z) = 1 \quad \text{if} \quad \frac{f(z \,|\, \theta_H)\psi(H)}{f(z \,|\, \theta_J)(1 - \psi(H))} \le b,$$

and $d_\psi(z) = 0$ otherwise. So we reject H if the likelihood ratio for θ_H compared to θ_J is less than a critical value:

$$\frac{f(z \,|\, \theta_H)}{f(z \,|\, \theta_J)} \le b \cdot \frac{1 - \psi(H)}{\psi(H)}. \tag{8.2}$$

Since Θ is a finite set, the simple version of the complete class theorem in Theorem 4.2 applies here. The admissible tests correspond to Bayes tests for some prior distribution ψ. As $\psi(H)$ varies from 0 to 1, the right-hand side of the inequality in (8.2) varies from ∞ to 0. So the admissible tests have the form

$$d_\psi(z) = 1 \quad \text{if} \quad \frac{f(z \,|\, \theta_H)}{f(z \,|\, \theta_J)} \le \text{crit}$$

for crit $\in [0, \infty]$.[13]

8.3. *Application: IV Model*

The model is

$$Z \sim \mathcal{N}(\rho\phi \otimes \omega, I_{2k}) \quad \text{for some} \quad \rho \in \mathcal{R}_+,\ \phi \in S^1,\ \omega \in S^{k-1}.$$

The hypothesis is that $\phi \in A \subset S^1$:

$$H = \{(\rho, \phi, \omega) \in \mathcal{R}_+ \times S^1 \times S^{k-1} : \phi \in A\}.$$

[12] See Wald (1950, p. 132).

[13] See Wald (1950, p. 127).

We shall develop a Bayes test for a prior distribution of the form

$$\psi = \psi_1 \times \psi_2 \times \tau_{k-1} \quad \text{where} \quad \tau_{k-1} = \text{Uniform}(S^{k-1}).$$

The prior distributions ψ_1 for ρ and ψ_2 for ϕ are left unspecified. The numerator in (8.1) is

$$\int_H f(z \,|\, (\rho, \phi, \omega))\, d\psi(\rho, \phi, \omega) = \int_A \int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho, \phi))\, d\psi_1(\rho)\, d\psi_2(\phi)$$

using the integrated likelihood function f_{I} from (7.3):

$$f_{\mathrm{I}}(z \,|\, (\rho, \phi)) = c(z) \exp(-\rho^2/2) G_k(\rho ||D(z)\phi||).$$

The denominator in (8.1) is

$$\int_{\Theta - H} f(z \,|\, (\rho, \phi, \omega))\, d\psi(\rho, \phi, \omega) = \int_{S^1 - A} \int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho, \phi))\, d\psi_1(\rho)\, d\psi_2(\phi).$$

So the Bayes test is

$$d_\psi(z) = 1 \quad \text{if} \quad \frac{\int_A \int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho, \phi))\, d\psi_1(\rho)\, d\psi_2(\phi)}{\int_{S^1 - A} \int_{\mathcal{R}_+} f_{\mathrm{I}}(z \,|\, (\rho, \phi))\, d\psi_1(\rho)\, d\psi_2(\phi)} \leq b. \tag{8.3}$$

Invariance can be used to simplify the risk function. The index set G is the set of $k \times k$ orthogonal matrices:

$$G = O(k) = \{k \times k \text{ matrices } g : gg' = I_k\}.$$

The transformation on the sample space is

$$m_1(g, z) = (I_2 \otimes g) z \quad \text{for} \quad z \in \mathcal{R}^{2k}.$$

If $Z \sim P_\theta = \mathcal{N}(\rho\phi \otimes \omega, I_{2k})$, then

$$m_1(g, Z) = (I_2 \otimes g) Z \sim \mathcal{N}(\rho\phi \otimes g\omega, I_{2k}) = P_{m_2(g,\theta)}$$

with

$$m_2(g, (\rho, \phi, \omega)) = (\rho, \phi, g\omega).$$

So the model is invariant. The loss function for testing $\phi \in A$ does not depend upon ω (or ρ): $L((\rho, \phi, \omega), a) = \tilde{L}(\phi, a)$. So we can set $m_3(g, a) = a$ and L is invariant. Then a test d is invariant if $d((I_2 \otimes g)z) = d(z)$ for all $g \in O(k)$ and $z \in \mathcal{R}^{2k}$.

For any choice of ψ_1 and ψ_2, our Bayes test d_ψ is invariant. This follows since $d_\psi(z)$ depends upon z only through $||D(z)\phi||$, and (as in (7.4))

$$D((I_2 \otimes g)z) = gD(z)$$

$$||D((I_2 \otimes g)z)\phi|| = ||gD(z)\phi|| = ||D(z)\phi||.$$

The invariance Theorem 6.2 implies that

$$R((\rho, \phi, \omega), d_\psi) = R((\rho, \phi, g_\omega \omega), d_\psi) = R((\rho, \phi, e_1), d_\psi) \equiv \tilde{R}((\rho, \phi), d_\psi),$$

where g_ω is an orthogonal matrix with $g_\omega \omega = e_1$, and e_1 is a $k \times 1$ vector whose elements are all 0 except for the first element which equals 1. So the risk function does not depend upon ω. Then, as in the proof of Theorem 6.1, it follows that

$$d_\psi = \arg\min_{d \in \mathcal{D}} \int_{S^1} \int_{\mathcal{R}_+} [\sup_{\omega \in S^{k-1}} R((\rho, \phi, \omega), d)]\, d\psi_1(\rho)\, d\psi_2(\phi). \tag{8.4}$$

This is true for any choice of ψ_1 and ψ_2, provided that ψ_3 equals the uniform distribution on S^{k-1}.

9. CONCLUSION

In the simple IV model, the dimension $(k+1)$ of the parameter space can be arbitrarily large. The minimax results are most useful when they provide a way of dealing with a large number of nuisance parameters in ω. A related application might include incidental parameters in a panel data model, where the minimax criterion bears some resemblance to a "fixed-effects" approach, that seeks to protect against any sequence of incidental parameters. There are recent discussions of incidental parameters and panel data in Lancaster (2002) and Arellano (2003).

Even in the simple IV model, with the loss functions I have chosen, the optimality results typically cannot be applied without further input: for example, a prior distribution for the scalar parameter ρ is needed for the optimal confidence set. Point estimation of ϕ is the exception, where the minimax result does not require additional input. I expect a minimax treatment of part of the parameter space to be useful in other applications, but that, typically, finite sample optimality will involve a combination of the average risk and maximum risk criteria.

As for invariance, my preference is to regard invariance arguments as a step in the derivation of minimax results. Other models need not have the invariance structure, but, if minimax is appealing for part of the parameter space, then numerical methods can be used. An algorithm is developed in Chamberlain (2000), using the minimax theorem for S-games (Blackwell and Girshick, 1954) and a concave programming algorithm, as in Wilson (1963). An appeal of the average risk and maximum risk criteria is that there is an explicit objective function to be minimized. Approximations may be needed due to computational constraints, but those constraints should become less binding as computational costs continue to decline.

Once one focuses on a risk function, it is natural to think about criteria, like average risk and maximum risk, that lend themselves to optimization. But decision theory can guide the evaluation of procedures, whether or not optimality plays a role. Consider for example the first approach to confidence sets in Section 7. The risk function is simply one minus the coverage rate, so the focus is not on an optimal procedure. But given a candidate procedure, such as inverting a likelihood ratio test, evaluating the risk function calls for determining the coverage rate at each point in the parameter space. A procedure may be motivated by an asymptotic argument that the limiting coverage rate is .95, without the need to specify a likelihood function. Nevertheless, it would be good to know how the finite sample coverage rate varies over the parameter space, which requires that there *be* a parameter space. A key component of decision theory is the evaluation of the risk of a decision rule over a set of distributions provided by the image under the model of the parameter space.

APPENDIX

Proof of Theorem 5.1. We can use (7.3) to simplify the integral in (5.4):

$$\int [1-(\phi' a)^2]\left[\int f(z\,|\,(\rho,\phi,\omega))\,d\tau_{k-1}(\omega)\right]d\tau_1(\phi)$$

$$= \int [1-(\phi' a)^2] f_{\mathrm{I}}(z\,|\,(\rho,\phi))\,d\tau_1(\phi)$$

$$= c(z)\exp(-\rho^2/2)\int [1-(\phi' a)^2] G_k(\rho||D(z)\phi||)\,d\tau_1(\phi),$$

with

$$G_1(t) = \exp(t) + \exp(-t)$$

and

$$G_k(t) = \int_{[-1,1]} \exp(ts)(1-s^2)^{(k-3)/2}\,ds$$

$$= \int_{[0,1]} [\exp(ts)+\exp(-ts)](1-s^2)^{(k-3)/2}\,ds$$

for $k > 1$. Note that G_k is an increasing function.

Let $Q(z) = (\,q_1(z) \quad q_2(z)\,)$ be an orthogonal matrix whose columns are the eigenvectors of $D(z)'D(z)$:

$$Q(z)'(D(z)'D(z))Q(z) = \begin{pmatrix} \zeta_1(z) & 0 \\ 0 & \zeta_2(z) \end{pmatrix}, \quad Q(z)'Q(z) = I_2,$$

where the eigenvalues $\zeta_1(z)$ and $\zeta_2(z)$ are ordered so that $\zeta_1(z) \geq \zeta_2(z)$. If ϕ has a uniform distribution on S^1, then $Q^{-1}(z)\phi$ has a uniform distribution on S^1 (for any value of z). So we have

$$\int (\phi' a)^2 G_k(\rho||D\phi||)\,d\tau_1(\phi) = \int [(Q^{-1}\phi)'(Q^{-1}a)]^2 G_k(\rho||(DQ)(Q^{-1}\phi)||)\,d\tau_1(\phi)$$

$$= \int [\phi'(Q^{-1}a)]^2 G_k(\rho||(DQ)\phi||)\,d\tau_1(\phi)$$

$$= \int [\phi'(Q^{-1}a)]^2 G_k(\rho(\zeta_1\phi_1^2+\zeta_2\phi_2^2)^{1/2})\,d\tau_1(\phi),$$

where $\phi' = (\,\phi_1 \quad \phi_2\,)$ and we have simplified the notation by suppressing the z argument in $D(z)$, $Q(z)$, $\zeta_1(z)$, and $\zeta_2(z)$.

$d_{\psi,\rho}(\mathrm{z})$ is a solution to

$$\max_{a \in S^1} \int [\phi'(Q^{-1}a)]^2 G_k(\rho(\zeta_1 \phi_1^2 + \zeta_2 \phi_2^2)^{1/2})\, d\tau_1(\phi). \tag{A.1}$$

We shall show that

$$\max_{a \in S^1} \int (\phi' a)^2 G_k(\rho(\zeta_1 \phi_1^2 + \zeta_2 \phi_2^2)^{1/2})\, d\tau_1(\phi) \tag{A.2}$$

is attained at $a = (\,1 \quad 0\,)'$. Then $d_{\psi,\rho}(z) = Q(z)\,(\,1 \quad 0\,)' = q_1(z)$ is a solution to (A.1).

Define $b_{\rho k}(t) = G_k(\rho\sqrt{t})$ for $k \geq 1$, and note that for a given value of $\rho > 0$, $b_{\rho k}(\cdot)\colon \mathcal{R}_+ \to \mathcal{R}$ is an increasing function. Represent ϕ and a in S^1 as follows:

$$\phi = \begin{pmatrix} \cos(s) \\ \sin(s) \end{pmatrix}, \quad a = \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix},$$

and note that

$$\phi' a = \cos(s)\cos(t) + \sin(s)\sin(t) = \cos(s-t).$$

Then we have

$$\begin{aligned} &\max_{a \in S^1} \int_{S^1} (\phi' a)^2 G_k(\rho(\zeta_1 \phi_1^2 + \zeta_2 \phi_2^2)^{1/2})\, d\lambda_{S^1}(\phi) \\ &\qquad = \max_{t \in [-\pi,\pi]} \int_{[-\pi,\pi]} \cos(s-t)^2 b_{\rho k}(\zeta_1 \cos(s)^2 + \zeta_2 \sin(s)^2)\, ds. \end{aligned} \tag{A.3}$$

Lemma.

$$\begin{aligned} &\int_{[-\pi,\pi]} \cos(s)^2 b_{\rho k}(\zeta_1 \cos(s)^2 + \zeta_2 \sin(s)^2)\, ds \\ &\qquad \geq \int_{[-\pi,\pi]} \cos(s-t)^2 b_{\rho k}(\zeta_1 \cos(s)^2 + \zeta_2 \sin(s)^2)\, ds \quad \text{for} \quad t \in [-\pi,\pi],\ \rho > 0,\ k \geq 1. \end{aligned}$$

Proof.

$$\cos(s)^2 - \cos(s-t)^2 = (1 - \cos(t)^2)(2\cos(s)^2 - 1) - 2\cos(t)\sin(t)\cos(s)\sin(s).$$

Let

$$w(s) = b_{\rho k}(\zeta_1 \cos(s)^2 + \zeta_2 \sin(s)^2) = b_{\rho k}((\zeta_1 - \zeta_2)\cos(s)^2 + \zeta_2).$$

Since $\cos(-s)\sin(-s)w(-s) = -\cos(s)\sin(s)w(s)$,

$$\int_{[-\pi,\pi]} \cos(s)\sin(s)w(s)\,ds = 0$$

and so

$$\int_{[-\pi,\pi]} (\cos(s)^2 - \cos(s-t)^2)w(s)\,ds = (1-\cos(t)^2)\int_{[-\pi,\pi]} (2\cos(s)^2 - 1)w(s)\,ds.$$

Consider the interval $[\pi/2, \pi]$. The functions $2(\cos(\cdot))^2 - 1$ and $w(\cdot)$ are increasing on this interval, and

$$\int_{[\pi/2,\pi]} (2\cos(s)^2 - 1)\,ds = \int_{[\pi/2,\pi]} \cos(2s)\,ds = \frac{1}{2}\sin(2s)\,\Big|_{\pi/2}^{\pi} = 0.$$

Hence

$$\int_{[\pi/2,\pi]} (2\cos(s)^2 - 1)w(s)\,ds \geq 0.$$

A similar argument applies to the intervals $[-\pi, -\pi/2]$, $[-\pi/2, 0]$, and $[0, \pi/2]$, since, on each of these intervals, the functions $2(\cos(\cdot))^2 - 1$ and $w(\cdot)$ are either both increasing or both decreasing, and the integral of $2(\cos(\cdot))^2 - 1$ is zero. (When both functions are decreasing, multiply each of them by -1 to obtain two increasing functions.) This completes the proof of the Lemma.

It follows from the Lemma that the maximizing value for t in (A.3) is $t = 0$, and so the maximizing value for a in (A.2) is $a = (\,1 \quad 0\,)'$. Hence

$$d_{\psi,\rho}(z) = Q(z)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = q_1(z)$$

is a solution to (A.1). $\diamond$

REFERENCES

Anderson, T. W., and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D., M. Moreira, and J. Stock (2006): "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74, 715–752.

Arellano, M. (2003): *Panel Data Econometrics*. Oxford: Oxford University Press.

Bekker, P. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62, 657–681.

Berger, J., B. Liseo, and R. Wolpert (1999): "Integrated Likelihood Methods for Eliminating Nuisance Parameters," *Statistical Science*, 14, 1–22.

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

Chamberlain, G. (2000): "Econometric Applications of Maxmin Expected Utility," *Journal of Applied Econometrics*, 15, 625–644.

Chamberlain, G. (2003): "Instrumental Variables, Invariance, and Minimax," unpublished manuscript, Department of Economics, Harvard University.

Chamberlain, G., and G. Imbens (2004): "Random Effects Estimators with Many Instrumental Variables," *Econometrica*, 72, 295–306.

Eaton, M. (1989): *Group Invariance Applications in Statistics*. Regional Conference Series in Probability and Statistics, Volume 1, Institute of Mathematical Statistics.

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.

Goldberger, A. S. (1991): *A Course in Econometrics*. Cambridge: Harvard University Press.

Goldberger, A. S., and I. Olkin (1971): "A Minimum-Distance Interpretation of Limited-Information Estimation," *Econometrica*, 39, 635–639.

Hansen, L. P., T. Sargent, G. Turmuhambetova, and N. Williams (2005): "Robust Control and Model Misspecification," *Journal of Economic Theory*, forthcoming.

Hansen, L. P., and T. Sargent (2005): "Robust Estimation and Control Under Commitment," *Journal of Economic Theory*, 124, 258–301.

Hillier, G. (1990): "On the Normalization of Structural Equations: Properties of Direction Estimators," *Econometrica*, 58, 1181–1194.

James, W., and C. Stein (1961): "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–380. Berkeley: University of California Press.

Lancaster, T. (2002): "Orthogonal Parameters and Panel Data," *Review of Economic Studies*, 69, 647–666.

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

Manski, C. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

Sims, C. (2001): "Pitfalls of a Minimax Approach to Model Uncertainty," *American Economic Review*, 91, 51–54.

Staiger, D., and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stein, C. (1956): "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 197–206. Berkeley: University of California Press.

Stein, C. (1959): "An Example of Wide Discrepancy Between Fiducial and Confidence Intervals," *The Annals of Mathematical Statistics*, 30, 877–880.

Stein, C. (1962): "Confidence Sets for the Mean of a Multivariate Normal Distribution," *Journal of the Royal Statistical Society, Series B*, 24, 265–296.

Strasser, H. (1985): *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*. Berlin: Walter de Gruyter.

Stroock, D. W. (1999): *A Concise Introduction to the Theory of Integration*. Boston: Birkhäuser.

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

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

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