CHAPTER 6

Invariant Decision Problems

This rather lengthy chapter provides an introduction to invariant decision problems. After describing the basic ingredients in a decision problem, invariance is introduced and used to define an invariant decision problem. A main result in this chapter shows how to construct a best invariant rule when the group action is transitive on the parameter space and when the dominating measure is decomposable [That is, the integral J defined by the measure satisfies Equation (5.14) in Theorem 5.5.] Applications of this result to the construction of best invariant estimators are given.

Finally, invariant testing problems are discussed. Wijsman's theorem (described in Section 5.3) is used to derive an invariant test with some optimum properties.

6.1. Decision problems and invariance. In this section the basic objects of a decision problem are first reviewed and then invariance is introduced into the problem. Here are the ingredients of a decision problem:

(i) A sample space $(\mathbf{X}, \mathbf{B}_1)$.
(ii) A parameter space $(\Theta, \mathbf{B}_2)$.
(iii) An action space $(A, \mathbf{B}_3)$.
(iv) A statistical model $\{P_\theta | \theta \in \Theta\}$ which consists of a family of probability measures defined on the sample space.
(v) A loss function L defined on $A \times \Theta$ to $[0, \infty)$ and assumed to be jointly measurable.

To describe the decision rules, let $M_1(A)$ denote the set of all probability measures on the action space.

DEFINITION 6.1. A *decision rule* δ is a function defined on $\mathbf{X}$ with values in $M_1(A)$ such that δ is a Markov kernel (as in Example 2.19).

The value of a decision rule δ at x is denoted by $\delta(\cdot|x)$ and because δ is assumed to be a Markov kernel, the map

$$x \to \delta(B|x)$$

is Borel measurable for each fixed set $B \in \mathbf{B}_3$. In some of the literature δ is called a randomized decision rule since $\delta(\cdot|x)$ is a probability measure on the action space for each $x \in \mathbf{X}$ [for example, see Berger (1985)]. The decision rule δ is called *nonrandomized* if for each x, $\delta(\cdot|x)$ puts probability 1 at a single point, say $a(x)$, in A. It is easy to show that when δ is a nonrandomized rule, then the corresponding induced function $x \to a(x)$ is measurable on $(\mathbf{X}, \mathbf{B}_1)$ to $(A, \mathbf{B}_3)$. Conversely given a measurable function $x \to a(x)$, the corresponding δ which puts probability 1 at $a(x)$ for $x \in \mathbf{X}$ is a decision rule.

Given a decision rule δ,

$$R(\delta, \theta) = \iint L(a, \theta)\delta(da|x)P_\theta(dx) \tag{6.1}$$

is the *risk* of δ at θ. The function $\theta \to R(\delta, \theta)$ is called the *risk function* of δ. The risk function is used to compare decision rules with the goal being to find decision rules with "small" risk functions.

To introduce invariance into the decision problem, let G be a topological group which acts on the left of the three spaces $\mathbf{X}$, Θ and A. The word "space" here is being used as described in Section 1.1. The three group actions are not distinguished notationally, but the group action under consideration will be clear. For example, gx means the action of G on $\mathbf{X}$ evaluated at (g, x), while $g\theta$ means the action of G on Θ evaluated at (g, θ) and similarly for the action of G on A. It is emphasized that it is the group action which changes from space to space and not the group which changes. In other words, the group G remains fixed in our discussion and G acts in perhaps different ways on the different spaces. This is the reason for not adopting the more common notation used to describe invariant decision problems [for example, see Berger (1985), Chapter 6]. Given the ingredients of the decision problem listed above, here is the definition of a G-invariant decision problem when G acts on $\mathbf{X}$, Θ and A.

DEFINITION 6.2. The decision problem above is G-invariant if:

(i) The model $\{P_\theta | \theta \in \Theta\}$ is invariant, that is,

$$gP_\theta = P_{g\theta} \quad \text{for } g \in G,\ \theta \in \Theta.$$

(ii) The loss function L is invariant, that is

$$L(ga, g\theta) = L(a, \theta) \quad \text{for } g \in G,\ \theta \in \Theta,\ a \in A.$$

Now we turn to the invariance of decision rules. As in Example 2.9, the group G acts on $\mathbf{B}_3 \times \mathbf{X}$. Given a decision rule δ and $g \in G$, the decision rule $g\delta$ is defined by

$$(g\delta)(B|x) = \delta(g^{-1}B|g^{-1}x)$$

for $B \in \mathbf{B}_3$ and $x \in \mathbf{X}$.

DEFINITION 6.3. A decision rule δ is *invariant* if $g\delta = \delta$ for all $g \in G$.

When a decision rule δ is nonrandomized, say δ corresponds to a measurable function $x \to a_0(x)$, it is easy to show that δ is invariant iff the function a_0 is equivariant, that is, iff a_0 satisfies

$$a_0(gx) = ga_0(x)$$

for all $g \in G$ and $x \in \mathbf{X}$. This fact is used below without mention.

The primary focus of this chapter is to describe some techniques for finding "good" invariant decision rules. To this end we first describe some transformation formulas which are used later in the chapter.

THEOREM 6.1. *When the model $\{P_\theta | \theta \in \Theta\}$ is G-invariant, the formula*

$$\int f(gx) P_\theta(dx) = \int f(x) P_{g\theta}(dx) \tag{6.2}$$

holds for any function f which is integrable. For any decision rule δ, the formula

$$\int k(a)(g\delta)(da|x) = \int k(ga)\delta(da|g^{-1}x) \tag{6.3}$$

holds for any function k for which the integrals are well defined.

PROOF. Verify the formulas for the indicator functions of sets and then extend in the usual way. □

Throughout the remainder of this section and in the next section, it is assumed that we have a given invariant decision problem with the ingredients specified above. Here is a basic risk function identify for such problems.

THEOREM 6.2. *For any decision rule δ,*

$$R(g\delta, g\theta) = R(\delta, \theta) \tag{6.4}$$

for all $g \in G$ and $\theta \in \Theta$.

PROOF. We use (6.2) and (6.3) and calculate as

$$\begin{aligned} R(g\delta, \theta) &= \iint L(a, \theta)(g\delta)(da|x) P_\theta(dx) \\ &= \iint L(ga, \theta)\delta(da|g^{-1}x) P_\theta(dx) \\ &= \iint L(a, g^{-1}\theta)\delta(da|x) P_{g^{-1}\theta}(dx) = R(\delta, g^{-1}\theta). \end{aligned}$$ □

An important consequence of (6.4) is:

THEOREM 6.3. *If δ is an invariant decision rule, then*

$$R(\delta, g\theta) = R(\delta, \theta) \tag{6.5}$$

for all $g \in G$ and $\theta \in \Theta$.

PROOF. When δ is invariant, $g\delta = \delta$ so (6.4) yields (6.5). □

Equation (6.5) says that the risk function of an invariant decision rule δ is an invariant function of θ. In particular, when the action of G is transitive on Θ and δ is invariant,

$$R(\delta, \theta) = R(\delta, \theta_0), \qquad \theta \in \Theta,$$

for any fixed $\theta_0 \in \Theta$. In other words, when G is transitive on Θ, invariant decision rules have constant risk functions. In this situation, we expect a *best invariant decision rule* to exist because we can fix θ_0 and then try to minimize $R(\delta, \theta_0)$ over the class of all invariant decision rules. More precisely, δ_0 is called a *best invariant decision rule* if

$$R(\delta_0, \theta_0) \le R(\delta, \theta_0)$$

for all invariant rules δ. Of course, the choice of θ_0 is irrelevant because G is assumed to act transitively on Θ.

In the next section we describe one method of finding a best invariant decision rule when G is transitive on Θ. The method is originally due to Stein (unpublished) but the treatment here is rather different than I've seen in the literature. A related work is Zidek (1969).

6.2. Best invariant rules in the transitive case. Consider an invariant decision problem as described in the previous section. It is assumed throughout this section that G acts transitively on Θ. Hence all invariant decision rules have constant risk functions. Also assume that the Radon measure μ on $(\mathbf{X}, \mathbf{B}_1)$ dominates each P_θ in the statistical model $\{P_\theta | \theta \in \Theta\}$. The dominating measure μ is assumed to be relatively invariant with multiplier Δ^{-1} (Δ is the right hand modulus of G) as in Theorem 5.5. The densities

$$p(x|\theta) = \frac{dP_\theta}{d\mu}(x)$$

are, as usual, assumed to satisfy

$$p(x|\theta) = p(gx|g\theta)\Delta^{-1}(g) \tag{6.6}$$

for all x, θ and g.

The main assumption of this section, is that G acts properly on $\mathbf{X}$ (Definition 5.1). Therefore, the representation of J (the integral defined by μ) described in Theorem 5.5 holds. This representation involves the function T defined on $K(\mathbf{X})$ to $K(\mathbf{X}/G)$ by

$$T(f)(\pi(x)) = \int f(gx)\nu_r(dg), \tag{6.7}$$

where ν_r is right Haar measure on G and π is the natural projection from $\mathbf{X}$ to $\mathbf{X}/G$. Then Theorem 5.5 shows that

$$J(f) = J_1(T(f)) \tag{6.8}$$

for some integral J_1 on $K(\mathbf{X}/G)$. Of course (6.8) holds for all f which are μ-integrable.

We now apply (6.7) and (6.8) to the expression for the risk function of an invariant decision rule δ. Fix δ, fix $\theta \in \Theta$ and set

$$f_0(x) = \int_A L(a, \theta)\delta(da|x)p(x|\theta)$$

so that f_0 is nonnegative and

$$R(\delta, \theta) = \int f_0(x)\mu(dx) = J(f_0). \tag{6.9}$$

From (6.8), we have

$$R(\delta, \theta) = J_1(T(f_0)), \tag{6.10}$$

where

$$\begin{aligned} T(f_0)(\pi(x)) &= \int_G f_0(gx)\nu_r(dg) \\ &= \int_G \int_A L(a, \theta)\delta(da|gx)p(gx|\theta)\nu_r(dg). \end{aligned} \tag{6.11}$$

THEOREM 6.4. *The function $T(f_0)$ in* (6.11) *satisfies the equation*

$$T(f_0)(\pi(x)) = \int_G \int_A L(a, g\theta)\delta(da|x)p(x|g\theta)\nu_r(dg). \tag{6.12}$$

PROOF. Using (6.3) and (6.6) in (6.11) yields

$$T(f_0)(\pi(x)) = \int_G \int_A L(ga, \theta)\delta(da|x)p(x|g^{-1}\theta)\Delta(g)\nu_r(dg). \tag{6.13}$$

The invariance of L, the identity $\nu_l = \Delta\nu_r$, together with the fact that

$$\nu_l(dg^{-1}) = \nu_r(dg)$$

applied to (6.13) show that

$$\begin{aligned} T(f_0)(\pi(x)) &= \int_G \int_A L(a, g^{-1}\theta)\delta(da|x)p(x|g^{-1}\theta)\nu_l(dg) \\ &= \int_G \int_A L(a, g\theta)\delta(da|x)p(x|g\theta)\nu_r(dg). \end{aligned}$$

Thus (6.12) holds. □

Here is the description of how to find a best invariant decision rule. Define H on $A \times \mathbf{X}$ by

$$H(a, x) = \int_G L(a, g\theta)p(x|g\theta)\nu_r(dg). \tag{6.14}$$

Because G is transitive on Θ, the function H does not depend on $\theta \in \Theta$.

THEOREM 6.5. *Assume a measurable function a_0, defined on $\mathbf{X}$ with values in A, exists which satisfies*

$$(6.15)\qquad \begin{aligned} &\text{(i)}\quad H(a,x) \geq H(a_0(x),x) \quad \text{for all } a, x,\\ &\text{(ii)}\quad a_0(gx) = ga_0(x) \qquad\quad \text{for all } g, x. \end{aligned}$$

Then a_0 defines a best invariant decision rule δ_0 [that is, $\delta_0(\cdot|x)$ puts probability 1 at $a_0(x)$ and δ_0 is invariant by (ii)].

PROOF. Fix an invariant decision rule δ and fix $\theta \in \Theta$. Because G is transitive on Θ, it suffices to show that

$$R(\delta_0, \theta) \leq R(\delta, \theta).$$

By definition of a_0 and δ_0, first observe that for any decision rule δ,

$$(6.16)\qquad \begin{aligned} &\int_G\int_A L(a, g\theta)\delta(da|x)p(x|g\theta)\nu_r(dg)\\ &\geq \int L(a_0(x), g\theta)p(x|g\theta)\nu_r(dg)\\ &= \int_G\int_A L(a, g\theta)\delta_0(da|x)p(x|g\theta)\nu_r(dg). \end{aligned}$$

Now, apply the integral J_1 to the first and last expressions in (6.16) and use (6.10) and (6.11) to yield $R(\delta_0, \theta) \leq R(\delta, \theta)$. This completes the proof. □

The argument above provides a constructive method for finding a best invariant decision rule (under the stated assumptions). Namely, for each $x \in \mathbf{X}$, we minimize (over a) the function

$$(6.17)\qquad H(a,x) = \int L(a, g\theta)p(x|g\theta)\nu_r(dg)$$

to get the minimizer $a_0(x)$ (assuming it exists). It is an easy exercise to show that

$$H(ha, hx) = H(a, x) \quad \text{for } h \in G$$

and for all a and x. Thus the discussion of the orbit-by-orbit method given after Theorem 3.2 is valid. Under the regularity conditions of Theorem 3.2 applied to H, the resulting minimizer a_0 will satisfy (6.15)(ii). For simplicity of exposition, this invariance condition is simply assumed in the statement of Theorem 6.5.

There is a Bayesian interpretation of Theorem 6.5 which provides a partial answer to the question raised in the discussion after Theorem 3.2. Fix θ_0 in Θ and define the function τ on G to Θ by

$$\tau(g) = g\theta_0.$$

The map τ is onto because G is transitive. Now define the "induced" measure ξ on measurable subsets of Θ by

$$\xi(B) = \nu_r(\tau^{-1}(B)),$$

or equivalently

$$\int f(\theta)\xi(d\theta) = \int f(g\theta_0)\nu_r(dg),$$

for $f \in K(\Theta)$. When ξ is a well defined Radon measure [it may not be in some examples when $\xi(B) = +\infty$ for some compact B], then an easy calculation shows that ξ is relatively invariant with multiplier Δ^{-1}. Assuming ξ is well defined, the definition of ξ shows that

$$H(a, x) = \int L(a, \theta)p(x|\theta)\xi(d\theta). \tag{6.18}$$

But (6.18) is proportional to the posterior loss for taking action a when the prior (possibly improper) is ξ. Thus, a_0 can be interpreted as a formal Bayes rule for the prior distribution ξ. Note that ξ is a proper prior iff G is compact.

In some examples, the "natural" dominating measure for the probability measures of the model is not relatively invariant with multiplier Δ^{-1}, but is relatively invariant with some other multiplier. This situation was discussed earlier in the context of Theorem 5.6. To discuss this in the present context, assume that μ_0 is the dominating measure and μ_0 is relatively invariant with multiplier χ_0. From (5.16), the new dominating measure

$$\mu(dx) = \frac{1}{\phi(x)}\mu_0(dx)$$

is relatively invariant with multiplier Δ^{-1}. Here, ϕ is a strictly positive function on $\mathbf{X}$ which satisfies

$$\phi(gx) = \chi_0(g)\Delta(g)\phi(x) \quad \text{for } x \in X,\ g \in G.$$

The existence of such a ϕ was discussed in Chapter 5. Relative to this new dominating measure μ, the densities become

$$\tilde{p}(x|\theta) = p(x|\theta)\phi(x), \tag{6.19}$$

where $p(x|\theta)$ is the density of P_θ with respect to μ_0. Theorem 6.5 applied to $\tilde{p}(x|\theta)$ shows that a best invariant rule is formed by minimizing

$$\tilde{H}(a, x) = \int L(a, g\theta)\tilde{p}(x|g\theta)\nu_r(dg).$$

Using (6.19), we have

$$\tilde{H}(a, x) = \phi(x)H(a, x),$$

where

$$H(a, x) = \int L(a, g\theta)p(x|\theta)\nu_r(dg). \tag{6.20}$$

Because ϕ is strictly positive, minimizing $\tilde{H}$ is the same as minimizing H. The point is that as long as the dominating measure is relatively invariant with respect to some multiplier [and the densities satisfy (3.1) for that multiplier], then a best invariant rule is found by minimizing (6.20) for each $x \in \mathbf{X}$. Examples of this are given in the next section.

6.3. Examples of best equivariant estimators. Three examples in which best equivariant estimators are derived are presented here. The first concerns normal data with a known coefficient of variation. In this example the best equivariant estimator under normalized quadratic loss is different than the maximum likelihood estimator. The last two examples deal with the estimation of a covariance matrix based on data from a normal distribution. This case is quite interesting because the problem is invariant under two different groups and the resulting estimators are different for the loss functions considered.

EXAMPLE 6.1. Consider $X_1, \dots, X_n$ which are iid $N(\theta, \theta^2)$ with $\theta > 0$. Thus the random vector X with coordinates $X_1, \dots, X_n$ has the distribution

$$\mathscr{L}(X) = N(\theta e, \theta^2 I_n),$$

where e is the vector of 1's in R^n. With $\Theta = A = (0, \infty)$ and

$$L(a, \theta) = \frac{(a-\theta)^2}{\theta^2},$$

the ingredients of the decision problem are specified. Take G to be the multiplicative group $(0, \infty)$ so G acts on Θ and A in the obvious way. Clearly G is transitive on Θ. Further G acts on sample vectors $x \in R^n$ by coordinatewise multiplication. That the decision problem is invariant is easily checked.

The group G does not act properly on R^n, but G does act properly on the modified sample space $\mathbf{X} = R^n - \{0\}$. Thus the results of the previous section show that a best equivariant estimator is found by minimizing

$$H(a) = \int_0^\infty L(a, g\theta) f(x|g\theta) \frac{dg}{g},$$

where $f(\cdot|\theta)$ is the density of the data X with respect to Lebesgue measure on $\mathbf{X}$. Of course dg/g is a right Haar measure on G. Since $H(a)$ does not depend on the choice of θ, we take $\theta = 1$ for convenience. Thus, for each x, the function

$$H(a) = \int_0^\infty L(a, g) f(x|g) \frac{dg}{g}$$

needs to be minimized. With

$$q(\theta|x) = \frac{\theta^{-1} f(x|\theta)}{\int_0^\infty \theta^{-1} f(x|\theta)\, d\theta}.$$

$q(\cdot|x)$ can be viewed as a posterior density of θ given x obtained from the improper prior $d\theta/\theta$. Clearly, minimizing H is equivalent to minimizing

$$\begin{aligned} H_1(a) &= \int_0^\infty L(a, \theta) q(\theta|x)\, d\theta \\ &= \mathbf{E}\left[\frac{(a-\theta)^2}{\theta^2} \middle| X = x \right]. \end{aligned}$$

The minimum is easily shown to be

$$a_0(x) = \frac{\mathbf{E}[\theta^{-1}|X = x]}{\mathbf{E}[\theta^{-2}|X = x]},$$

which satisfies the equivariance condition (6.15)(ii). The best equivariant estimator a_0 is not known in closed form, but can be computed numerically. See Kariya (1984) for some further discussion.

For comparative purposes, the maximum likelihood estimator of θ is

$$\hat{\theta}(x) = \frac{-\bar{x} + \left[(\bar{x})^2 + 4\bar{x^2}\right]^{1/2}}{2},$$

where

$$\bar{x^2} = \frac{1}{n}\sum_1^n x_i^2.$$

Obviously $\hat{\theta}$ is equivariant, but it is not too hard to show that $\hat{\theta} \neq a_0$. This completes the first example. □

EXAMPLE 6.2. Consider iid p-dimensional random vectors $X_1, \ldots, X_n$ which have a multivariate normal distribution $N_p(0, \Sigma)$. The problem considered here is the estimation of the $p \times p$ covariance matrix Σ which is assumed to be nonsingular, but otherwise unknown. Further, it is assumed that $n \geq p$ so that the sufficient statistic

$$S = \sum_{i=1}^n X_i X_i'$$

is positive definite with probability 1. Without loss of generality, estimators of Σ are functions of S. Thus the sample space $\mathbf{X}$, the parameter space Θ and the action space A are taken to be the set of $p \times p$ positive definite matrices. Obviously S has a Wishart distribution $W(\Sigma, p, n)$ and the maximum likelihood estimator of Σ is

$$\hat{\Sigma} = n^{-1}S.$$

The group Gl_p acts on $\mathbf{X}$, Θ and A in the obvious way:

$$S \to gSg',$$
$$\Sigma \to g\Sigma g',$$
$$a \to gag',$$

for $g \in \mathrm{Gl}_p$. The model for S is invariant under this group action.

Now, consider a loss function L which is invariant, that is, L satisfies

$$L(gag', g\Sigma g') = L(a, \Sigma) \tag{6.21}$$

for all g, a and Σ. A standard invariance argument shows that L satisfies (6.21) iff L can be written as a function of the eigenvalues of $a\Sigma^{-1}$, say $\lambda_1 \geq \cdots \geq$

$\lambda_p > 0$. Two interesting examples are

$$(6.22)\qquad \begin{aligned} L_1(a, \Sigma) &= \operatorname{tr} \Sigma^{-1}(a - \Sigma)\Sigma^{-1}(a - \Sigma) = \operatorname{tr}\left(\Sigma^{-1/2}a\Sigma^{-1/2} - I_p\right)^2 \\ &= \sum_{i=1}^{p} (\lambda_i - 1)^2 \end{aligned}$$

and

$$(6.23)\qquad \begin{aligned} L_2(a, \Sigma) &= \operatorname{tr} a\Sigma^{-1} - \log\det(a\Sigma^{-1}) - p \\ &= \sum_{i=1}^{p} (\lambda_i - \log \lambda_i - 1). \end{aligned}$$

The loss function L_1 was used by Selliah (1964) in his Stanford thesis [also see Olkin and Selliah (1977)]. Stein (1956) introduced L_2 in his study of covariance estimation [also see James and Stein (1960)].

Because Gl_p acts transitively on Θ, a best equivariant (nonrandomized) estimator should exist. Rather than use the results of Section 6.2, it is a bit easier for this particular example to proceed as follows. An estimator $\tau(S)$ is equivariant iff

$$(6.24)\qquad \tau(gSg') = g\tau(S)g'$$

for all S and $g \in \mathrm{Gl}_p$. Picking $g = S^{-1/2}$, (6.24) implies that

$$(6.25)\qquad \tau(S) = S^{1/2}\tau(I)S^{1/2}.$$

But, for any $\gamma \in O_p$, (6.24) implies

$$(6.26)\qquad \tau(I) = \tau(\gamma I \gamma') = \gamma\tau(I)\gamma',$$

which implies that the matrix $\tau(I)$ must be a multiple of the identity matrix, say $\tau(I) = \alpha I$. Combining this with (6.25) shows that an estimator is equivariant iff

$$(6.27)\qquad \tau(S) = \alpha S$$

for some real α. Thus, to find a best equivariant estimator, the constant α is selected to minimize the risk at some (any) fixed point in Θ. For L_1, α is to be selected to minimize the risk

$$\begin{aligned} R_1(\alpha S, I) &= \mathbf{E}_I \operatorname{tr}(\alpha S - I)^2 \\ &= \alpha^2 \mathbf{E}_I \operatorname{tr} S^2 - 2\alpha \mathbf{E}_I \operatorname{tr} S + p. \end{aligned}$$

The minimizer is

$$\alpha_1 = \frac{\mathbf{E}_I \operatorname{tr} S}{\mathbf{E}_I \operatorname{tr} S^2} = \frac{np}{(n^2 - 2n)p + n(p-1)} = \frac{1}{n + 2 + (p-1)/p}.$$

Using L_2, a similar calculation shows that the minimizing α is

$$\alpha_2 = \frac{1}{n}.$$

Hence for L_2, the maximum likelihood estimator is a best equivariant estimator.

Because the group of $p \times p$ lower triangular matrices with positive diagonal elements G_T^+ is a subgroup of Gl_p, the above estimation problem (with loss functions L_1 or L_2) is also invariant under G_T^+. Since G_T^+ acts transitively on Θ, a best equivariant estimator should exist for this problem. In this case an estimator $\tau(S)$ is equivariant iff

$$(6.28) \qquad \tau(gSg') = g\tau(S)g', \qquad g \in G_T^+,$$

for each $S \in \mathbf{X}$. Standard arguments show that (6.28) holds iff

$$(6.29) \qquad \tau(S) = TAT',$$

where T is the unique element in G_T^+ such that $S = TT'$ and A is a fixed symmetric matrix. Thus finding a best equivariant (under G_T^+) estimator involves finding A to minimize the risk at a fixed point in Θ. This of course depends on the loss function.

For loss function L_1, A is chosen to minimize

$$R_1 = \mathbf{E}_I \operatorname{tr}(TAT' - I)^2.$$

The minimizing A is a diagonal matrix, but the minimizer is not known explicitly. For the equations determining the minimizer see Selliah (1964) and Olkin and Selliah (1977).

When the loss function is L_2, A is chosen to minimize

$$\begin{aligned} R_2 &= \mathbf{E}_I[\operatorname{tr} TAT' - \log\det(TAT') - p] \\ &= \mathbf{E}_I \operatorname{tr} T'TA - \log\det(A) + c_0, \end{aligned}$$

where c_0 is a constant. Thus, it suffices to minimize

$$\tilde{R}_2 = \operatorname{tr}\mathbf{E}(T'T)A - \log\det(A) = \sum_{i=1}^{p} d_{ii}a_{ii} - \log\det(A),$$

where $a_{ii}, \ldots, a_{pp}$ are the diagonal elements of A and

$$d_{ii} = n + p - 2i + 1, \qquad i = 1, \ldots, p.$$

The minimum is achieved at

$$A_0 = D^{-1},$$

where D is diagonal with diagonal elements $d_{11}, \ldots, d_{pp}$. Therefore the best equivariant estimator in this case is

$$\tau_2(S) = TD^{-1}T'$$

as was established by Stein [see James and Stein (1960)]. It also follows from results in Kiefer (1957) that this estimator τ_2 is minimax [because the group G_T^+ is solvable; see Bondar and Milnes (1981) for a survey], but the maximum likelihood estimator $\hat{\Sigma}$ is not minimax when L_2 is the loss function. For some related results, see Eaton and Olkin (1987). This completes Example 6.2. □

It should be noted that the results established in Section 6.2 were not used to find the best equivariant estimators in Example 6.2. Rather, the procedure was to first characterize the functional form of the equivariant estimators and then minimize the risk (at a fixed point in Θ) over the class of equivariant estimators. This was feasible because the functional form was rather simple for the two cases considered. In the next example, which also concerns covariance estimation, this procedure seems not to be feasible because the class of equivariant estimators is rather large. For this reason, the method described in Section 6.2 is used in the following example.

EXAMPLE 6.3. As in Example 6.2, assume that data S which is $W(\Sigma, p, n)$ with $n \geq p$, is available. Partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} is $q \times q$ and Σ_{22} is $r \times r$, so $p = q + r$. Consider "extra" data $X_1, \ldots, X_m$ which consists of iid random vectors in R^q such that each X_i has a $N(0, \Sigma_{11})$ distribution. The problem is to estimate Σ using S and the data $X_1, \ldots, X_m$. Note that (S, V) is a sufficient statistic where

$$V = \sum_1^m X_i X_i'.$$

To describe the invariance of this problem, let G be the subgroup of Gl_p whose elements g have the form

$$g = \begin{pmatrix} g_{11} & 0 \\ g_{21} & g_{22} \end{pmatrix}$$

with $g_{11} \in \mathrm{Gl}_q$ and $g_{22} \in \mathrm{Gl}_r$. The model is easily shown to be invariant under the group actions

$$(S, V) \to (gSg', g_{11}Vg_{11}'),$$
$$\Sigma \to g\Sigma g'.$$

This invariance together with the orbit-by-orbit method described in Theorem 3.2 can be used to derive the maximum likelihood estimator for Σ. The details are not given here [see Eaton (1970) for one derivation of $\hat{\Sigma}$].

Because the group G_T^+ of $p \times p$ lower triangular matrices with positive diagonal elements is a subgroup of G, the model for the data is G_T^+ invariant. Also, G_T^+ acts transitively on Θ so a best equivariant estimator should exist. In the remainder of this example we focus on finding a best equivariant (under G_T^+) estimator when the loss function is L_2 given in Example 6.2. First, we characterize the equivariant estimators, that is, those estimators τ which satisfy

$$\tau(gSg', g_{11}Vg_{11}') = g\tau(S, V)g' \tag{6.30}$$

for $g \in G_T^+$ and for sample points (S, V) for which S is positive definite and V is nonnegative definite. Let T be the unique element in G_T^+ satisfying $S = TT'$. Picking $g = T^{-1}$ in (6.30) shows that τ satisfying (6.30) must satisfy

$$\tau(TT', V) = T\tau\left(I, T_{11}^{-1}V\left(T_{11}^{-1}\right)'\right)T' = T\tau_0\left(T_{11}^{-1}V\left(T_{11}^{-1}\right)'\right)T',$$

where τ_0 maps $q \times q$ nonnegative definite matrices into $p \times p$ positive definite matrices (since τ is assumed to take values in Θ). Conversely, if τ is given by

$$\tau(TT', V) = T\tau_0\big(T_{11}^{-1}V(T_{11}^{-1})'\big)T', \tag{6.31}$$

then it is easy to show that τ satisfies (6.30). Hence, all equivariant estimators have the form (6.31) where τ_0 is "arbitrary." Thus, to find a best equivariant estimator, we must minimize the risk over all τ_0's:

$$R_2 = \mathbf{E}_I L_2\big[T\tau_0\big(T_{11}^{-1}V(T_{11}^{-1})'\big)T, I\big].$$

A straightforward approach to this problem, such as that used in Example 6.2, seems very difficult to carry out. For this reason, the method described in Section 6.2 is used.

To show that the method of Section 6.2 is applicable, first note that the sample space $\mathbf{X} \times (R^q)^m$ is acted on by G_T^+ via

$$(S, x_1, \ldots, x_m) \to (gSg', g_{11}x_1, \ldots, g_{11}x_m),$$

where $S \in \mathbf{X}$ (as in Example 6.2), each x_i is in R^q and $g \in G_T^+$. Lebesgue measure on $\mathbf{X} \times (R^q)^m$ is relatively invariant under the group action and the action is proper. Thus a best equivariant estimator is found by minimizing [from (6.20)]

$$H(a) = \int L_2(a, g\theta_0)p(z|g\theta_0)\nu_r(dg), \tag{6.32}$$

where θ_0 is a fixed point in Θ, $z \in \mathbf{X} \times (R^q)^m$ is a sample point and ν_r is a right-invariant measure on G_T^+. For convenience, pick $\theta_0 = I_p \in \Theta$. Substituting the explicit form of the density p in (6.32), the function which needs to be minimized is

$$\begin{aligned} H(a) = \int L_2(a, gg')&\big|(gg')^{-1}S\big|^{n/2}|g_{11}g_{11}'|^{-m/2} \\ &\times \exp\big[-\tfrac{1}{2}\operatorname{tr}(gg')^{-1}S - \tfrac{1}{2}\operatorname{tr}(g_{11}g_{11}')^{-1}V\big]\nu_r(dg). \end{aligned} \tag{6.33}$$

The orbit-by-orbit method shows it is sufficient to minimize (6.33) when $S = I_p$. Setting $S = I_p$ and making the change of variable $g \to g^{-1}$, the function we want to minimize is

$$H_1(a) = \int L_2(a, gg')|g'g|^{n/2}|g_{11}'g_{11}|^{m/2}\exp\big[-\tfrac{1}{2}\operatorname{tr} g'g - \tfrac{1}{2}\operatorname{tr} g_{11}'g_{11}V\big]\nu_l(dg),$$

where ν_l is a left-invariant measure on G_T^+. The details of this minimization are given in Eaton (1970) and are not described further here. The best equivariant estimator obtained from this minimization is as follows. First write the data (S, V) as

$$S = TT', \qquad U = T_{11}^{-1}V(T_{11}^{-1})',$$

where $T \in G_T^+$. Further, write

$$I_q + U = WW', \tag{6.34}$$

where W is the unique $q \times q$ lower triangular matrix with positive diagonals satisfying (6.34). The best equivariant estimator (for loss function L_2) can be written

$$\tau(S) = T\tau_0(U)T',$$

where

$$\tau_0(U) = \begin{bmatrix} [(W^{-1})'DW^{-1} + (p-q)I]^{-1} & 0 \\ 0 & E^{-1} \end{bmatrix}$$

and D: $q \times q$ and E: $r \times r$ are fixed diagonal matrices with diagonal elements

$$d_{ii} = m + n + q - 2i + 1, \qquad i = 1, \dots, q,$$

and

$$e_{ii} = n + p - 2q - 2i + 1, \qquad i = 1, \dots, r.$$

This estimator is minimax because of results in Kiefer (1957). □

6.4. Invariant testing examples. Here we consider an invariant testing problem where the representation theorem, Theorem 5.9, is applicable. On a space **X** which is acted on *properly* by a locally compact group G, assume that a Radon measure μ is relatively invariant with multiplier χ. A family of densities with respect to μ, say $\{p(\cdot|\theta)|\theta \in \Theta\}$, is given. Further, the group G acts on Θ and the basic invariance condition

$$p(x|\theta) = p(gx|g\theta)\chi(g) \quad \text{for } x, \theta, g$$

is assumed to hold. Thus the probability model on **X** determined by the family of densities is invariant.

Let Θ_0 and Θ_1 be G invariant subsets of Θ and consider testing

$$H_0\colon \theta \in \Theta_0 \quad \text{versus} \quad H_1\colon \theta \in \Theta_1.$$

This testing problem is, according to the discussion given in Section 3.2, invariant. In what follows, attention is restricted to invariant test functions ϕ. Thus, if $\tau(X)$ is a maximal invariant, an invariant test function ϕ can be written as

$$\phi(X) = \psi(\tau(X)).$$

Now, we add the final assumption that *G acts transitively on both Θ_0 and Θ_1*. Under this assumption (plus those above), a most powerful level α invariant test is given below. To describe this test, first observe that the power function of any invariant test, say

$$\beta_\phi(\theta) = \mathbf{E}_\theta \phi(X),$$

is an invariant function of θ. Because G is transitive on Θ_0 and Θ_1, this power function takes on only two values, namely,

$$a_i = \beta_\phi(\theta), \qquad \theta \in \Theta_i, \qquad i = 0, 1.$$

Thus, a_0 is the level of ϕ and a_1 is the power of ϕ. But, for fixed a_0, the Neyman–Pearson lemma tells us that the most powerful test rejects for large values of the ratio

$$r(t) = \frac{q_1(t)}{q_0(t)},$$

where q_i is the density of $\tau(X)$ when X has density $p(\cdot|\theta)$, $\theta \in \Theta_i$ for $i = 0, 1$. Of course, q_i does not depend on $\theta \in \Theta_i$ because G is transitive on Θ_i. Note that densities always exist in the present situation because $\tau(X)$ has only two possible distributions—one under H_0 and one under H_1. However, Theorem 5.9 tells us how to compute the ratio $r(t)$. In the notation of this section, fix $\theta_i \in \Theta_i$, $i = 0, 1$. Then according to (5.20) [using the maximal invariant $\pi(x)$],

$$r(\pi(x)) = \frac{\int p(gx|\theta_1)\chi(g)\nu_l(dg)}{\int p(gx|\theta_0)\chi(g)\nu_l(dg)}, \tag{6.35}$$

where χ is the multiplier specified in the model and ν_l is a left-invariant measure on G. Naturally the denominator is assumed to be positive. Also, the expression (6.35) does not depend on the choices of θ_0 and θ_1 because of the transitivity assumption. The above discussion implies that the most powerful level α-invariant test rejects for large values of $r(\pi(x))$ given in (6.35).

Before turning to some examples, it is useful to contrast the hypothesis testing problem considered here with the estimation problem of the last section. It was assumed that G acted transitively in the estimation problem, and this assumption implied that equivariant estimators had constant risk. Thus the risk of decision rule in the estimation problem is a single number, because there is only one orbit in the parameter space. In the testing problem above, the parameter space is $\Theta_0 \cup \Theta_1$ and there are two orbits (assuming $\Theta_0 \neq \Theta_1$). Thus, the risk function of an invariant test is determined by two numbers. In this situation, the Neyman–Pearson lemma tells us what the good invariant decision rules are.

EXAMPLE 6.4. This example concerns Hotelling's T^2 test for certain types of nonnormal data and comes from Kariya (1981). Consider a random matrix X: $n \times p$ which has a density with respect to Lebesgue measure, on np dimensional space given by

$$p(x|\theta) = |\Sigma|^{-n/2} f\left[\operatorname{tr}(x - e\mu')\Sigma^{-1}(x - e\mu')'\right]. \tag{6.36}$$

Here x is $n \times p$, e is the vector of 1's in R^n, μ is an unknown vector in R^p, Σ is a $p \times p$ positive definite matrix, tr denotes the trace and f is some nonnegative function defined on $[0, \infty)$ which satisfies

$$\int f\left[\operatorname{tr} x'x\right] dx = 1.$$

The parameter θ stands for the pair (μ, Σ). In the case that

$$f(z) = (\sqrt{2\pi})^{-np} \exp\left[-\tfrac{1}{2}z\right], \qquad z \in [0, \infty],$$

then of course the rows of X are iid $N_p(\mu, \Sigma)$.

Here is the classical hypothesis testing problem which Hotelling's T^2 test solves for normal data:

$$H_0: \mu = 0 \quad \text{versus} \quad H_1: \mu \neq 0.$$

The matrix Σ is unrestricted under both H_0 and H_1. With a rotation of coordinates and some relabelling, we can (and do) replace the vector e in (6.36) with the vector

$$\varepsilon_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in R^n.$$

This relabelling results in some notational simplification. It is assumed that $n \geq p + 1$ and matrices x: $n \times p$ are partitioned as

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

where x_1 is $1 \times p$ and x_2 is $(n - 1) \times p$. The sample space for this problem is taken to be the set of x's such that x_2: $(n - 1) \times p$ has rank p.

To describe a group under which this problem is invariant, let G be the product group $G_0 \times \mathrm{Gl}_p$ where G_0 is the subgroup of O_n whose elements have the form

$$g = \begin{pmatrix} 1 & 0 \\ 0 & \gamma \end{pmatrix}, \qquad \gamma \in O_{n-1}.$$

Given $(g, a) \in G_0 \times \mathrm{Gl}_p$, the group action on $\mathbf{X}$ is

$$x \to gxa'$$

and the group action on the parameter $\theta = (\mu, \Sigma)$ is

$$(\mu, \Sigma) \to (a\mu, a\Sigma a').$$

Note that Lebesgue measure is relatively invariant with multiplier

$$\chi(g, a) = |\det(a)|^n.$$

Routine calculations show that a maximal invariant function on $\mathbf{X}$ is

$$T^2 = X_1(X_2'X_2)^{-1}X_1',$$

where

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Further, a maximal invariant function on the parameter space is

$$\delta = \mu'\Sigma^{-1}\mu.$$

Now, fix $\delta_1 > 0$ and consider

$$\Theta_0 = \{(\mu, \Sigma) | \delta = 0\}$$

and

$$\Theta_1 = \{(\mu, \Sigma) | \delta = \delta_1\}.$$

The group G acts transitively on Θ_0 and Θ_1 and G acts properly on $\mathbf{X}$. Thus, a best invariant level α test for testing the null hypothesis

$$H_0\colon \theta \in \Theta_0 \quad \text{versus} \quad \tilde{H}\colon \theta \in \Theta_1$$

is found by calculating r in (6.35). Clearly the original null hypothesis is equivalent to $\theta \in H_0$, but the original alternative is H_1: $\delta > 0$. Our first goal is to find a best invariant test for H_0 versus $\tilde{H}_1$. To this end, take $\theta_0 \in \Theta$ to be

$$\theta_0 = (0, I_p)$$

and θ_1 to be

$$\theta_1 = \left(\sqrt{\delta_1}\,\xi, I_p\right),$$

where

$$\xi = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in R^p.$$

A left-invariant measure on G is

$$\nu_l(dh) = \nu(dg)\frac{da}{|\det(a)|^p},$$

where ν is the invariant probability measure on G_0 and da is Lebesgue measure on Gl_p. Substituting these and (6.36) into (6.35), and changing variables shows that

$$(6.37) \qquad r(v) = \frac{\int f\left[\operatorname{tr} a'a + \delta_1 - 2a_{11}v\delta_1^{1/2}\right]|a'a|^{(n-p)/2}\,da}{\int f\left[\operatorname{tr} a'a\right]|a'a|^{(n-p)/2}\,da},$$

where both of the integrals are over Gl_p, a_{11} is the (1, 1) element of a and

$$v = \frac{T^2}{(1+T^2)^{1/2}}.$$

Obviously v is an increasing function of T^2 and v is also a maximal invariant statistic since it is a one-to-one function of T^2.

PROPOSITION 6.1. *Assume f is a convex function on $[0,\infty)$. Then the best invariant level α test of H_0 versus $\tilde{H}_1$ rejects for large values of T^2. Further, the null distribution of T^2 is that when*

$$f(z) = \left(\sqrt{2\pi}\right)^{-np}\exp\left[-\tfrac{1}{2}z\right].$$

PROOF. Because Lebesgue measure on Gl_p is invariant under the transformation $a \to -a$, it follows that

$$r(v) = r(-v), \qquad v \ge 0,$$

where r is given in (6.37). The convexity of f implies that for $\frac{1}{2} \le \beta \le 1$ and

$v \geq 0$, we have

$$r((2\beta - 1)v) = r(\beta v + (1 - \beta)(-v))$$
$$\leq \beta r(v) + (1 - \beta) r(-v) = r(v).$$

Hence $r(v)$ is nondecreasing in v so rejecting for large values of r is equivalent to rejecting for large values of v, which is, in turn, equivalent to rejecting for large values of T^2. The first part of the proposition is proved.

That the null distribution of T^2 does not depend on the particular f in (6.36) is a consequence of the null robustness results described in Section 4.3. □

An immediate consequence of Proposition 6.1 is that the test which rejects for large values of T^2 is a best invariant test (of its level) for testing H_0 versus H_1. This follows because the best invariant test of H_0 versus $\tilde{H}_1$ did not depend on the particular alternative $\tilde{H}_1$. Thus, as long as f is convex, Hotelling's T^2 test as a best invariant test of H_0 versus H_1 and the null distribution of T^2 is known for each f. This completes our discussion of Example 6.4. □

Example 6.5. Here we briefly discuss a relatively smooth example where a best invariant test exists (Wijsman's theorem applies) and this test is different from the likelihood ratio test. This example is essentially due to Stein and was originally constructed to show that the Hunt–Stein theorem is not valid for the group Gl_p, $p \geq 2$. See Lehmann [(1959), Problem 10, page 344] for a related result.

Consider two independent Wishart matrices S_1 with

$$\mathscr{L}(S_1) = W(\Sigma, p, n)$$

and S_2 with

$$\mathscr{L}(S_2) = W(c\Sigma, p, n).$$

It is assumed $n \geq p \geq 2$, Σ is an unknown $p \times p$ positive definite matrix and the real constant c is positive. The problem is to test

$$H_0\colon c = 1 \quad \text{versus} \quad H_1\colon c = 2$$

so Σ is a nuisance parameter. In our previous notation,

$$\Theta_0 = \{(c, \Sigma) | c = 1\}$$

and

$$\Theta_1 = \{(c, \Sigma) | c = 2\}.$$

It is easily verified that this problem is invariant under Gl_p acting on S_i by

$$S_i \to g S_i g', \qquad g \in \mathrm{Gl}_p$$

and on Σ by

$$\Sigma \to g \Sigma g', \qquad g \in \mathrm{Gl}_p.$$

Clearly Gl_p is transitive on Θ_0 and Θ_1. The other conditions necessary to apply the argument at the beginning of this section are easily checked. Fix a level α in

(0, 1). The ratio in (6.35) defines a best level α Gl_p-invariant test, say ϕ_0. Note that the likelihood ratio test is Gl_p-invariant so can be no better than ϕ_0.

Now, G_T^+ is a subgroup of Gl_p and so the testing problem above is also invariant under G_T^+. Also G_T^+ acts transitively on Θ_0 and Θ_1. Again the conditions necessary to apply the argument at the beginning of this section can be verified. Thus again the ratio in (6.35) computed using G_T^+ defines a best G_T^+-invariant test, say ϕ_1. The test ϕ_1 is at least as good as ϕ_0 because ϕ_0 is G_T^+-invariant. In fact ϕ_1 is a better test than ϕ_0 and hence ϕ_1 dominates the likelihood ratio test.

The point of this example is that there is a bit "too much" invariance in the problem above. Fully invariant procedures such as the likelihood ratio test can be improved upon by simply requiring less invariance. This completes Example 6.5. □

Finally we mention the book by Kariya and Sinha (1988) which contains material on null and nonnull robustness as well as further applications of Wijsman's theorem.