CHAPTER 3

Invariant Statistical Models

In this lecture, invariant statistical models are introduced and a variety of examples is given. Invariant testing problems and equivariant estimators are introduced. Univariate and multivariate linear models provide a host of standard examples.

3.1. Invariant models. Given a measurable space $(\mathbf{X}, \mathscr{B})$, a family of probability measures $\mathscr{P}$ defined on $\mathscr{B}$ is a *statistical model.* If the random variable X takes values in $\mathbf{X}$ and $\mathscr{L}(X) \in \mathscr{P}$, then we say $\mathscr{P}$ is a model for $\mathscr{L}(X)$.

DEFINITION 3.1. Suppose the group G acts measurably on $\mathbf{X}$. The statistical model $\mathscr{P}$ is *G-invariant* if for each $P \in \mathscr{P}$, $gP \in \mathscr{P}$ for all $g \in G$.

When the model $\mathscr{P}$ is G-invariant, G acts on $\mathscr{P}$ according to Definition 2.1. Further, when $\mathscr{P}$ is a model for $\mathscr{L}(X)$, then $\mathscr{P}$ is G-invariant means that $\mathscr{L}(gX) \in \mathscr{P}$ whenever $\mathscr{L}(X) \in \mathscr{P}$ for all $g \in G$.

EXAMPLE 3.1. Consider $\mathbf{X} = R^n$ with the Borel σ-algebra and let $f_0(\|x\|^2)$ be any probability density on R^n. Then the probability measure P_0 defined by

$$P_0(B) = \int I_B(x) f_0(\|x\|^2)\, dx$$

is *orthogonally invariant* as described in Example 2.6. That is, for each $g \in O_n$, the group of $n \times n$ orthogonal matrices, $gP_0 = P_0$. Thus $\mathscr{P} = \{P_0\}$ is O_n-invariant.

Of course, there are other orthogonally invariant probabilities than those defined by such a density. In fact, given any probability measure Q on R^n, define P by

$$P(B) = \int_{O_n} (gQ)(B)\nu(dg) = \int Q(g^{-1}B)\nu(dg),$$

where ν is invariant probability measure on O_n. Clearly P is O_n-invariant. Thus, averaging gQ over g with respect to the Haar measure on O_n always produces an O_n-invariant probability.

It is easy to see that this procedure is valid for any compact group G acting measurably on a space $(\mathbf{X}, \mathscr{B})$. That is, let Q be a probability on $(\mathbf{X}, \mathscr{B})$ and define P by

$$P = \int gQ\nu(dg),$$

where ν is the Haar probability measure on G. The above equation means

$$P(B) = \int (gQ)(B)\nu(dg)$$

for $B \in \mathscr{B}$. That P is G-invariant follows from

$$(hP)(B) = P(h^{-1}B) = \int (gQ)(h^{-1}B)\nu(dg) = \int Q(g^{-1}h^{-1}B)\nu(dg)$$

$$= \int Q\big((hg)^{-1}B\big)\nu(dg) = \int Q(g^{-1}B)\nu(dg) = P(B). \qquad \square$$

More detail concerning the structure of probabilities invariant under compact groups is given in the next lecture. Here is a standard parametric example where the group is not compact.

EXAMPLE 3.2. Consider $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$ on R^1. Then the random vector

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in R^n$$

has the distribution

$$\mathscr{L}(X) = N_n\big(\mu e_n, \sigma^2 I_n\big),$$

where e_n is the vector of 1's in R^n. A statistical model for $\mathscr{L}(X)$ is

$$\mathscr{P} = \big\{N\big(\mu e_n, \sigma^2 I_n\big) | \mu \in R^1, \sigma^2 > 0\big\}.$$

The appropriate group G for this example has group elements which are triples (γ, a, b) with $a > 0$, $b \in R^1$ and $\gamma \in O_n$ such that $\gamma e_n = e_n$. The group action on R^n is

$$x \to a\gamma x + be_n$$

and the group operation is

$$(\gamma_1, a_1, b_1)(\gamma_2, a_2, b_2) = (\gamma_1\gamma_2, a_1 a_2, a_1 b_2 + b_1).$$

That $\mathscr{P}$ is G-invariant follows from the observation that when

$$\mathscr{L}(X) = N\big(\mu e_n, \sigma^2 I_n\big),$$

then

$$\mathscr{L}((\gamma, a, b)X) = \mathscr{L}(a\gamma X + be_n) = N(a\mu + b)e_n, a^2\sigma^2 I_n\big)$$

which is an element of the model $\mathscr{P}$. □

Example 3.2 is one instance of a standard method of generating an invariant model. Given a fixed probability Q_0 on $(\mathbf{X}, \mathscr{B})$ and a group G acting measurably on $(\mathbf{X}, \mathbf{B})$, let

$$\mathscr{P} = \{gQ_0 | g \in G\}.$$

Obviously $\mathscr{P}$ is G-invariant because $h(gQ_0) = (hg)Q_0$ which is in $\mathscr{P}$ for $gQ_0 \in \mathscr{P}$. In the previous example, Q_0 is the $N(0, I_n)$ distribution on R^n, G is the group given in the example and $\mathscr{P}$ is just the family $\{N(\mu e_n, \sigma^2 I) | \mu \in R^1, \sigma^2 > 0\}$. Here is a slightly more complicated linear model example.

EXAMPLE 3.3. On R^n, fix a distribution Q_0 which we think of as a *standardized error distribution* for a linear model [e.g., $Q_0 = N(0, I_n)$]. Let M be a linear subspace of R^n which is regarded as the *regression subspace* of the linear model. Elements of a group G are pairs (a, b) with $a \neq 0$, $a \in R^1$ and $b \in M$. The group operation is

$$(a_1, b_1)(a_2, b_2) = (a_1 a_2, a_1 b_2 + b_1)$$

and the action on R^n is

$$x \to ax + b.$$

The G-invariant statistical model is

$$\mathscr{P} = \{gQ_0 | g \in G\}.$$

To describe this model in more standard terminology, let ε_0 have distribution Q_0 so $\mathscr{L}(\varepsilon_0) = Q_0$. For $(a, b) \in G$,

$$\mathscr{L}((a, b)\varepsilon_0) = \mathscr{L}(b + a\varepsilon_0).$$

Hence an observation from this model can be written

$$Y = b + a\varepsilon_0,$$

where $b \in M$. Assuming ε_0 has mean 0, the mean of Y is $b \in M$ and the covariance matrix of Y is

$$a^2 \operatorname{Cov}(\varepsilon_0),$$

where $\operatorname{Cov}(\varepsilon_0)$ is the covariance matrix of ε_0. Setting $\varepsilon = a\varepsilon_0$ and $\mu = b$, the model for Y is

$$Y = \mu + \varepsilon$$

which is the standard "Y equals mean vector plus error" model common in linear regression. When $\operatorname{Cov}(\varepsilon_0) = I_n$, we are in the case when $\operatorname{Cov}(Y) = \sigma^2 I_n$ for some $\sigma^2 > 0$. Thus, the usual simple regression models are group generated models when it is assumed that the error distribution is some scaled version of a fixed distribution on R^n. □

In many situations, invariant statistical models $\mathscr{P}$ are parametric statistical models having parametric density functions with respect to a fixed σ-finite measure. The proper context to discuss the expression of the invariance in terms of the densities is the following. Consider a topological group G acting measurably on $(\mathbf{X}, \mathscr{B})$ and assume that μ is a σ-finite measure on $(\mathbf{X}, \mathscr{B})$ which is relatively invariant with multiplier χ on G. That is,

$$\int f(g^{-1}x)\mu(dx) = \chi(g)\int f(x)\mu(dx)$$

for $g \in G$ and μ-integrable f.

THEOREM 3.1. *Assume the group G acts on the parameter space Θ and that $\{p(\cdot|\theta)|\theta \in \Theta\}$ is a family of densities with respect to μ. If the densities satisfy*

$$p(x|\theta) = p(gx|g\theta)\chi(g), \tag{3.1}$$

then the parametric family of probability measures $\mathscr{P} = \{P_\theta|\theta \in \Theta\}$ defined by the densities is G-invariant. Further, $gP_\theta = P_{g\theta}$.

PROOF. For each $\theta \in \Theta$ and $g \in G$, it suffices to verify that

$$gP_\theta = P_{g\theta}$$

since this implies $\mathscr{P}$ is G-invariant. For $B \in \mathscr{B}$,

$$\begin{aligned}(gP_\theta)(B) = P_\theta(g^{-1}B) &= \int I_B(gx)p(x|\theta)\mu(dx)\\ &= \chi(g)\int I_B(gx)p(gx|g\theta)\mu(dx)\\ &= \chi(g)\chi(g^{-1})\int I_B(x)p(x|g\theta)\mu(dx)\\ &= P_{g\theta}(B).\end{aligned}$$ □

A converse to Theorem 3.1 is "almost true." That is, consider a parametric family $\mathscr{P} = \{P_\theta|\theta \in \Theta\}$ which is G-invariant and satisfies

$$gP_\theta = P_{g\theta}, \qquad g \in G,\ \theta \in \Theta.$$

If each P_θ has a density $p(x|\theta)$ with respect to a σ-finite measure μ which is relatively invariant with multiplier χ, then the argument used in Theorem 3.1 shows that (3.1) holds a.e. for each $g \in G$ and $\theta \in \Theta$. Unfortunately, the null set where (3.1) does not hold can depend on both g and θ. However, in all of the interesting cases that I know, a version of the density exists so that (3.1) holds for all x, θ and g. When (3.1) holds for the density $p(\cdot|\theta)$, we say that the family of densities is *invariant* (the multiplier χ is understood to be given by the context), although a better word might be χ-invariant.

EXAMPLE 3.4. Consider a random vector $X \in R^p$ which is multivariate normal with mean vector $\mu \in R^p$ and positive definite covariance $\Sigma = \text{Cov}(X) \in S_p^+$. The density of X with respect to Lebesgue measure dx is

$$p(x|\mu, \Sigma) = \frac{|\Sigma|^{-1/2}}{(\sqrt{2\pi})^p}\exp\left[-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]. \tag{3.2}$$

Thus, the sample space is R^p and the parameter space is $\Theta = R^p \times S_p^+$. The affine group Al_p acts on R^p by

$$x \to gx + a,$$

where $(g, a) \in \mathrm{Al}_p$ with $g \in \mathrm{Gl}_p$ and $a \in R^p$. When $\mathscr{L}(X) = N(\mu, \Sigma)$, then

$$\mathscr{L}((g, a)X) = N(g\mu + a, g\Sigma g'),$$

so the appropriate action on θ is

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

An easy calculation shows that dx is relatively invariant with multiplier

$$\chi(g, a) = |\det(g)|, \qquad (g, a) \in \mathrm{Al}_p.$$

The direct verification of (3.1) entails showing that

$$p(x|\mu, \Sigma) = p(gx + a|g\mu + a, g\Sigma g')|\det(g)|,$$

which is routine.

The parametric family of this example is

$$\mathscr{P} = \{N(\mu, \Sigma)|\mu \in R^p, \Sigma \in S_p^+\}.$$

With $Q_0 = N(0, I_p)$, it is clear that Al_p acting on Q_0 generates the family $\mathscr{P}$, that is,

$$\mathscr{P} = \{(g, a)Q_0|(g, a) \in \mathrm{Al}_p\}.$$

Notice that Al_p does not give a one-to-one indexing for this parametric family. There is nothing special about the normal distribution in the example above. Given any fixed density f_1 (with respect to dx) on R^p, let Q_1 be the probability measure defined by f_1. Then the parametric family

$$\mathscr{P}_1 = \{(g, a)Q_1|(g, a) \in \mathrm{Al}_p\}$$

is generated by Al_p acting on Q_1. A direct calculation shows that $(g, a)Q$ has a density

$$p(x|(g, a)) = f_1(g^{-1}(x - a))|\det(g)|^{-1},$$

which clearly satisfies (3.1). Again, the group Al_p ordinarily does not provide a one-to-one indexing of the parametric family $\mathscr{P}_1$. That is, it is usually the case that

$$(g_1, a_1)Q_1 = (g_2, a_2)Q_1$$

does not imply that $(g_1, a_1) = (g_2, a_2)$.

In the case of the normal distribution, the mean and covariance provided a one-to-one indexing of $\mathscr{P}$. However, this is not the case in general, but when the density f_1 has the form $f_1(x) = k_1(\|x\|^2)$, then the distribution $(g, a)Q_1$ depends on (g, a) only through gg' and a, as in the normal case. □

3.2. Invariant testing problems. In this section, invariant testing problems are described and invariant tests are discussed. The setting for this discussion is a statistical model $\mathscr{P}$ which is invariant under a group G acting on a sample space $(\mathbf{X}, \mathscr{B})$. Thus, the observed random variable X satisfies $\mathscr{L}(X) \in \mathscr{P}$. Consider a testing problem in which the null hypothesis H_0 is that a submodel $\mathscr{P}_0$ of $\mathscr{P}$ actually obtains. In other words, the null hypothesis is that $\mathscr{L}(X) \in \mathscr{P}_0 \in \mathscr{P}$ as opposed to the alternative that $\mathscr{L}(X) \in \mathscr{P} - \mathscr{P}_0$.

Definition 3.2. The above hypothesis testing problem is *invariant under* G if both $\mathscr{P}_0$ and $\mathscr{P}$ are G invariant.

When the testing problem is invariant under G, then it is clear that the alternative $\mathscr{P} - \mathscr{P}_0$ is also invariant under G.

Following standard terminology [e.g., Lehmann (1986)], a test function ϕ is a measurable function from $(\mathbf{X}, \mathscr{B})$ to $[0,1]$ and $\phi(x)$ is interpreted as the conditional probability of rejecting H_0 when the observation is x. The behavior of a test function ϕ is ordinarily described in terms of the *power function*

$$\beta(P) = \mathbf{E}_P\phi(X), \qquad P \in \mathscr{P}.$$

Ideally, one would like to choose ϕ to make $\beta(P) = 0$ for $P \in \mathscr{P}_0$ and $\beta(P) = 1$ for $P \in \mathscr{P} - \mathscr{P}_0$.

When a hypothesis testing problem is invariant under a group G, it is common to see the following "soft" argument to support the use of an *invariant test function*, that is, a test function ϕ which satisfies $\phi(x) = \phi(gx)$ for $x \in \mathbf{X}$ and $g \in G$. This argument is:

> Consider $x \in \mathbf{X}$ and suppose $X = x$ supports H_0. Then we tend to believe $\mathscr{L}(X) \in \mathscr{P}_0$. If we had observed $X = gx$ instead, then $x = g^{-1}X$ and $\mathscr{L}(g^{-1}X) \in \mathscr{P}_0$ when $\mathscr{L}(X) \in \mathscr{P}_0$. Hence we should also believe H_0 if gx obtains. In other words, x and gx should carry the same weight of evidence for H_0. Exactly the same argument holds if x supports H_1.

We now turn to a discussion of two ways of obtaining an invariant test in the special case that $\mathscr{P}$ is a parametric family $\mathscr{P} = \{P_\theta | \theta \in \Theta\}$, G acts on Θ and there is a density $p(\cdot|\theta)$ for P_θ with respect to a σ-finite measure μ. It is assumed further that μ is relatively invariant with multiplier χ and the density $p(\cdot|\theta)$ satisfies Equation (3.1), i.e.,

$$p(x|\theta) = p(gx|g\theta)\chi(g).$$

In this notation, the invariance of the hypothesis testing problem means that

$$\mathscr{P}_0 = \{P_\theta | \theta \in \Theta_0\}$$

is G invariant so the set $\Theta_0 \in \Theta$ is G invariant. The likelihood ratio statistic for testing H_0: $\theta \in \Theta_0$ versus H_1: $\theta \in \Theta - \Theta_0 = \Theta_1$ is ordinarily defined by

$$\Lambda(x) = \frac{\sup_{\theta \in \Theta_0} p(x|\theta)}{\sup_{\theta \in \Theta} p(x|\theta)}.$$

This statistic is then used to define a test function ϕ via

$$\phi(x) = \begin{cases} 1, & \text{if } \Lambda(x) < c, \\ 0, & \text{if } \Lambda(x) \geq c, \end{cases}$$

where c is some appropriately chosen constant. Because of our assumptions on $p(\cdot|\theta)$ and the invariance of Θ_0 and Θ, it is readily shown that $\Lambda(x) = \Lambda(gx)$ and hence that $\phi(x) = \phi(gx)$. In other words, the test defined by the likelihood ratio statistic is invariant under any group for which the testing problem is invariant.

A second method which can sometimes be employed to define an invariant test involves the use of relatively invariant measures defined on Θ_0 and Θ_1. More precisely, assume we can find measures ξ_0 and ξ_1 on Θ_0 and Θ_1, respectively, which are relatively invariant with the same multiplier χ_1. Assuming the following expression makes sense, let

$$T(x) = \frac{\int_{\Theta_0} p(x|\theta)\xi_0(d\theta)}{\int_{\Theta_1} p(x|\theta)\xi_1(d\theta)}.$$

Then T can be used to define a test function ϕ via

$$\phi(x) = \begin{cases} 1, & \text{if } T(x) < c, \\ 0, & \text{if } T(x) \geq c, \end{cases}$$

where c is the omnipresent constant. The assumed invariance of the density $p(\cdot|\theta)$ and the relative invariance of ξ_0 and ξ_1 combine to imply that $T(x) = T(gx)$ and hence that the test ϕ is an invariant test.

The main reason for introducing the invariant tests described above is to raise some questions for which answers are provided in later lectures—in particular, how to select ξ_0 and ξ_1. Under some rather restrictive conditions, it is shown in a later lecture how to choose the measures ξ_0 and ξ_1 so that the test defined by the statistic T is a "good" invariant test. It is also shown that the likelihood ratio test does not necessarily yield a "good" invariant test when one exists.

3.3. Equivariant estimators.

As in the last section, consider a group G which acts on a sample space $(\mathbf{X}, \mathscr{B})$ and a parameter space Θ in such a way that the parametric model $\mathscr{P} = \{P_\theta | \theta \in \Theta\}$ is invariant. A density $p(x|\theta)$ is assumed to exist and the invariance condition (3.1) is to hold throughout this section. Thus, the dominating measure μ is relatively invariant with multiplier χ. In this context, a point estimator t mapping $\mathbf{X}$ to Θ is *equivariant* if

$$t(gx) = gt(x). \tag{3.3}$$

A soft argument leading to the consideration of equivariant estimators is:

> Consider t_0 as an estimator of θ. When $\mathscr{L}(X) = P_\theta$, then $t_0(X)$ is supposed to estimate θ. But, when $\mathscr{L}(X) = P_\theta$, $\mathscr{L}(gX) = P_{g\theta}$ so $t_0(gX)$ should estimate $g\theta$. However, $t_0(X)$ estimates θ, so $gt_0(X)$ estimates $g\theta$. Equating these two estimators of θ leads to $t_0(gX) = gt_0(X)$ and hence estimators satisfying (3.3).

The method of maximum likelihood leads to an equivariant estimator when the maximum likelihood estimator is unique. This result is a consequence of the following constructive method for finding a maximum likelihood estimator.

THEOREM 3.2. *Consider X with G-invariant density $p(\cdot|\theta)$ and fix $x_0 \in \mathbf{X}$. Assume that $\theta_0 \in \Theta$ uniquely maximizes $p(x_0|\theta)$ as θ varies over Θ, so $p(x_0|\theta) \le p(x_0|\theta_0)$ with equality iff $\theta = \theta_0$. For $x \in O_{x_0}$, the orbit of x_0, write $x = g_x x_0$ for some $g_x \in G$ and set*

$$\hat{\theta}(x) = g_x\theta_0.$$

Then for $x \in O_{x_0}$, $\hat{\theta}$ is the maximum likelihood estimator of θ, is unique and is equivariant.

PROOF. The proof is not hard and can be found in Eaton [1983, page 259–260]. □

The import of Theorem 3.2 is that, in invariant situations, the maximum likelihood estimators can be found by simply selecting some convenient point x_0 from each orbit in $\mathbf{X}$ and then calculating the maximum likelihood estimator, say θ_0, for that x_0. The value of $\hat{\theta}(x)$ for other x's in the same orbit is calculated by finding a g_x such that $g_x x_0 = x$ and setting $\hat{\theta}(x) = g_x\theta_0$. This orbit-by-orbit method of solution arises in other contexts later.

The results of Theorem 3.2 are valid for other methods of estimation also. For example, certain nonparametric methods can be characterized as choosing an estimator $t_0(x)$ so as to maximize a function

$$H(x|\theta) \ge 0$$

as θ ranges over Θ. If the function H satisfies

$$H(x|\theta) = H(gx|g\theta)\chi_0(g)$$

for some multiplier χ_0, then Theorem 2.3 applies [just replace the density $p(x|\theta)$ by $H(x|\theta)$ in the statement of Theorem 3.2]. In other words, the orbit-by-orbit method applies and the resulting estimator is equivariant and unique.

In a Bayesian context, inferential statements about θ are in the form of probability distributions on Θ which depend on x. These probability distributions are often obtained from a measure ξ on Θ. The measure ξ is a *prior distribution* if $0 < \xi(\Theta) < +\infty$ and is an *improper prior distribution* if $\xi(\Theta) = +\infty$. Given ξ, let

$$m(x) = \int p(x|\theta)\xi(d\theta)$$

and assume that $0 < m(x) < +\infty$ for all $x \in \mathbf{X}$. Then define

$$q(\theta|x) = \frac{p(x|\theta)}{m(x)}$$

so that

$$Q(B|x) = \int I_B(\theta)q(\theta|x)\xi(d\theta)$$

determines a probability measure $Q(\cdot|x)$ on Θ. For B fixed $Q(B|\cdot)$ is a measurable function defined on $(\mathbf{X}, B)$. Thus Q is a Markov kernel (randomized decision rule, posterior distribution, etc.) as discussed in Example 2.19. In this context, the appropriate notion of invariance is that defined in Example 2.19, namely Q is invariant if $gQ = Q$. That is, if

$$Q(B|x) = Q(g^{-1}B|g^{-1}x) \tag{3.4}$$

for measurable sets $B \subset \Theta$, $x \in \mathbf{X}$ and $g \in G$. Here is a condition which implies that (3.4) holds:

THEOREM 3.3. *Assume the measure ξ on Θ is relatively invariant with some multiplier χ_1. Then Q satisfies* (3.4).

PROOF. For $g \in G$,

$$\begin{aligned} Q(g^{-1}B|g^{-1}x) &= \int I_B(g\theta)q(\theta|g^{-1}x)\xi(d\theta) \\ &= \frac{\int I_B(g\theta)p(g^{-1}x|\theta)\xi(d\theta)}{\int p(g^{-1}x|\theta)\xi(d\theta)}. \end{aligned}$$

Using (3.1), we have

$$Q(g^{-1}B|g^{-1}x) = \frac{\int I_B(g\theta)p(x|g\theta)\xi(d\theta)}{\int p(x|g\theta)\xi(d\theta)}.$$

The assumed invariance of ξ now yields (3.4). □

Again, the question of how to choose ξ from the class of all relatively invariant measures naturally arises. This is discussed in later lectures.

Finally, the relationship between the equivariance of a point estimator and (3.4) requires a comment. Given any point estimator $t_0(x)$, the natural way to identify t_0 with a Markov kernel Q_0 is to let $Q_0(\cdot|x)$ be degenerate at the point $t_0(x) \in \Theta$, that is,

$$Q_0(B|x) = \begin{cases} 1, & \text{if } t_0(x) \in B, \\ 0, & \text{otherwise.} \end{cases}$$

With this identification, it is routine to show that (3.4) holds iff t_0 is equivariant.

Here is a simple example. More interesting and complicated examples appear in later lectures.

EXAMPLE 3.5. Suppose $X \in R^1$ is $N(\theta, 1)$, so $\Theta = R^1$. Obviously the model $\mathscr{P} = \{N(\theta, 1)|\theta \in R^1\}$ is invariant under the group $G = R^1$ acting on $\mathbf{X}$ and Θ via translation. For this example t is an equivariant point estimator iff

$$t(x + g) = t(x) + g$$

for all $x, g \in R^1$. Setting $g = -x$, we see that t is equivariant iff

$$t(x) = x + a,$$

where a is some fixed real number. The choice $a = 0$ gives the maximum likelihood estimator. To generate invariant distributions on Θ, consider the relatively invariant measure

$$\xi(d\theta) = e^{\theta b}\,d\theta,$$

where b is a fixed real number. (These are all the relatively invariant Radon measures on R^1 up to positive multiples.) For x given, this ξ gives

$$Q(\cdot\,|x) = N(x+b,1) \quad \text{on } \Theta = R^1.$$

In words, the formal posterior distribution is normal with mean $x + b$ and variance 1. Of course, the "usual" choice is $b = 0$ in which case $\xi(d\theta) = d\theta$ is the Haar measure on $G = \Theta$. In more complicated examples where G is not unimodular, the choice is not so obvious. □

3.4. Linear models. In order to motivate our discussion of linear models, first consider what is commonly called the multivariate analysis of variance model (the MANOVA model)

$$Y = XB + E. \tag{3.5}$$

Here, X: $n \times k$ is a known matrix of rank k, B: $k \times p$ is a matrix of unknown regression parameters and E: $n \times p$ is a matrix of random variables (errors). A standard assumption concerning E is that the rows of E are iid multivariate normal with mean 0 and common positive definite covariance, say C: $p \times p$. This assumption on E is often written

$$\mathscr{L}(E) = N(0, I_n \otimes C), \tag{3.6}$$

where $I_n \otimes C$ denotes the Kronecker product of the $n \times n$ identity matrix I_n and C. Thus the $n \times p$ error matrix has a normal distribution (with mean 0 and the specified covariance) on the vector space of $n \times p$ real matrices $\mathscr{L}_{p,n}$. The standard coordinate inner product on $\mathscr{L}_{p,n}$ is

$$\langle x, y\rangle = \operatorname{tr} xy' = \sum_i \sum_j x_{ij} y_{ij},$$

where tr denotes the trace. That E has covariance $I_n \otimes C$ means that

$$\operatorname{cov}\{\langle x, E\rangle, \langle y, E\rangle\} = \langle x, (I_n \otimes C)y\rangle,$$

where cov denotes ordinary covariance between real random variables. As usual, the Kronecker product $I_n \otimes C$ is the linear transformation on $\mathscr{L}_{p,n}$ to $\mathscr{L}_{p,n}$ defined by

$$(I_n \otimes C)x = I_n x C' = xC'.$$

An alternative way to write (3.5) is

$$Y = \mu + E, \tag{3.7}$$

where E is the error vector as before and $\mu = XB$ is the mean vector for Y. Thus, the space of possible values for μ is the linear subspace of $\mathscr{L}_{p,n}$,

$$M = \{\mu | \mu = XB,\ B\colon k \times p\} \subset \mathscr{L}_{p,n}. \tag{3.8}$$

When the distributional assumption (3.6) holds, then

$$\mathscr{L}(Y) = N(\mu, I_n \otimes C), \tag{3.9}$$

where $\mu \in M$ and C is some $p \times p$ positive definite covariance matrix. Thus the parametric model for $\mathscr{L}(Y)$ is

$$\mathscr{P} = \{N(\mu, I_n \otimes C) | \mu \in M, C \in S_p^+\}.$$

To describe the invariance of this model, let $M_0 \subset R^n$ be the linear subspace spanned by the columns of the matrix X and set

$$G_0 = \{g | g \in O_n, g(M_0) \subset M_0\}.$$

Elements of G_0 are orthogonal transformations on R^n to R^n which have M_0 and hence the orthogonal complement $M_0^{\perp}$, as invariant subspaces. Note that if $\mu \in M$, then the matrix product $g\mu$ is also in M for $g \in G_0$, because the columns of μ are elements of M_0.

Now, let G have elements which are triples (g, a, α) with $g \in G_0$, $a \in \mathrm{Gl}_p$ and $\alpha \in M$. The action of (g, a, α) on $\mathscr{L}_{p,n}$ is

$$(g, a, \alpha)x = gxa' + \alpha$$

and the group composition is

$$(g_1, a_1, \alpha_1)(g_2, a_2, \alpha_2) = (g_1g_2, a_1a_2, g_1\alpha_2a_1' + \alpha_1).$$

When (3.9) holds, then

$$\mathscr{L}((g, a, \alpha)Y) = N(g\mu a' + \alpha, I_n \otimes (aCa')),$$

which is again in $\mathscr{P}$. Thus the MANOVA model $\mathscr{P}$ is invariant under G. It is well known that the maximum likelihood estimator for μ in this model is

$$\hat{\mu} = P_0Y,$$

where P_0: $n \times n$ is the orthogonal projection onto $M_0 \subset R^n$. Further, $\hat{\mu}$ is the unique unbiased estimator of μ based on the sufficient statistic for $\mathscr{P}$ and $\hat{\mu}$ is the best linear unbiased estimator of μ. The invariance of the model $\mathscr{P}$ implies that $\hat{\mu}$ is an equivariant estimator of μ where the action of G on M is

$$\mu \to g\mu a' + \alpha.$$

It should be mentioned that the linear transformation on $\mathscr{L}_{p,n}$ to $\mathscr{L}_{p,n}$ defined by

$$x \to P_0x$$

is just the orthogonal projection onto M in the inner product space $(\mathscr{L}_{p,n}, \langle \cdot, \cdot \rangle)$.

The aspect of the MANOVA model with which the rest of this section deals is the equivariance of the estimator of the mean vector. For this discussion, we first describe the Gauss–Markov theorem for the so-called *regular linear models*. Consider finite dimensional inner product space $(V, (\cdot, \cdot))$. By a linear model for a random vector Y with values in V, we mean a model of the form

$$Y = \mu + \varepsilon, \tag{3.10}$$

where

(i) the random vector ε has mean 0 and covariance $\Sigma = \mathrm{Cov}(\varepsilon)$ assumed to lie in some known set γ of positive definite linear transformations;

(ii) the mean vector μ of Y lies in a known subspace M of V which is called the *regression subspace*.

Thus, the pair (M, γ) determines the assumed mean and covariance structure for Y taking values in $(V, (\cdot, \cdot))$. In what follows it is assumed that the identity covariance I is in γ. This assumption is without loss of generality since any pair (M, γ) can be transformed via one element of γ to another pair (M_1, γ_1) with $I \in \gamma_1$.

DEFINITION 3.3. The pair (M, γ) (with $I \in \gamma$) is a *regular linear model* for Y if

$$\Sigma(M) \subset M, \qquad \Sigma \in \gamma. \tag{3.11}$$

In other words, the model is regular if the regression subspace is an invariant subspace under each possible covariance $\Sigma \in \gamma$. That the MANOVA model is regular is readily verified. The condition (3.11) is equivalent to the condition

$$\Sigma A_0 = A_0 \Sigma, \qquad \Sigma \in \gamma, \tag{3.12}$$

where A_0 is the orthogonal projection onto the subspace M.

Consider any linear unbiased estimator AY of $\mu \in M$, that is, A is a linear transformation from V to V which satisfies

$$Ax = x, \qquad x \in M. \tag{3.13}$$

Here is one version of the Gauss–Markov theorem which compares linear unbiased estimators in terms of covariance:

THEOREM 3.4 (Gauss–Markov theorem). *Assume the pair (M, γ) determines a regular linear model for Y and let A_0 be the orthogonal projection onto M. Then, for any linear unbiased estimator AY of $\mu \in M$,*

$$\operatorname{Cov}(A_0 Y) \le \operatorname{Cov}(AY),$$

where $\le$ is in the sense of nonnegative definiteness. That is, Cov(AY) − Cov$(A_0 Y)$ is nonnegative definite.

PROOF. With $\Sigma = \operatorname{Cov}(Y)$,

$$\begin{aligned} \operatorname{Cov}(AY) &= A\Sigma A' \\ &= A_0\Sigma A_0 + (A - A_0)\Sigma A_0 \\ &\quad + A_0\Sigma(A - A_0)' + (A - A_0)\Sigma(A - A_0)'. \end{aligned}$$

However, the term $(A - A_0)\Sigma A_0$ is 0. To see this, first notice that $\Sigma A_0 = A_0 \Sigma$ due to the regularity of the linear model. But (3.13) implies that $(A - A_0)A_0 = 0$, so

$$(A - A_0)\Sigma A_0 = (A - A_0)A_0\Sigma = 0.$$

Thus $A_0\Sigma(A - A_0)'$ is also 0 as it is the transpose of $(A - A_0)\Sigma A_0$. Hence

$$\begin{aligned} \operatorname{Cov}(AY) &= A_0\Sigma A_0 + (A - A_0)\Sigma(A - A_0)' \\ &= \operatorname{Cov}(A_0 Y) + (A - A_0)\Sigma(A - A_0)'. \end{aligned}$$ □

Now consider a linear model

$$Y = \mu + \varepsilon,$$

where again $\mu \in M$, the error vector ε has mean zero and satisfies the following invariance assumption:

$$\mathscr{L}(\varepsilon) = \mathscr{L}(g_0\varepsilon),$$

where

$$g_0 = I - 2A_0.$$

As usual, A_0 is the orthogonal projection onto M, so $g_0^2 = I$ and $g_0^{-1} = g_0$. This model for Y is invariant under the transformations

$$x \to gx + a, \qquad x \in V,$$

where g is either g_0 or I and $a \in M$. In other words when $Y = \mu + \varepsilon$,

$$gY + a = (g\mu + a) + g\varepsilon = \mu^* + \varepsilon^*,$$

where $\mu^* \in M$ and ε^* has the same distribution as ε. Hence the mean of $gY + a$ is still in M and the error vector has the same distribution for $gY + a$ as for Y. The group in question, say G, has elements which are pairs (g, a) with $a \in M$ and g is g_0 or I. The group operation is

$$(g_1, a_1)(g_2, a_2) = (g_1g_2, g_1a_2 + a_1).$$

The appropriate equivariance of a point estimator t of μ is

$$t(gx + a) = gt(x) + a \tag{3.14}$$

because μ is mapped into $g\mu + a$ by the group element (g, a). The following result shows that an equivariant estimator of μ is just the Gauss–Markov estimator A_0Y of Theorem 3.4.

THEOREM 3.5. *Suppose t: $V \to M$ satisfies* (3.14) *for $(g, a) \in G$. Then $t(x) = A_0x$ for $x \in V$.*

PROOF. For $x \in V$, write $x = x_1 + x_2$ where $x_1 \in M$ and $x_2 \in M^\perp$. Here, $M^\perp$ is the orthogonal complement of M. Choosing $g = I$ and $a = -x_1$ in (3.14) yields

$$t(x_1 + x_2 - x_1) = t(x_1 + x_2) - x_1$$

so that

$$t(x_1 + x_2) = x_1 + t(x_2).$$

Now, in (3.14) pick $a = 0$, $x = x_2 \in M^\perp$ and $g = g_0$. From (3.14) we have

$$t(g_0x_2) = g_0t(x_2).$$

But $g_0x_2 = x_2$ as $x_2 \in M^\perp$, and $g_0t(x_2) = -t(x_2)$ because t takes values in M. Thus,

$$t(x_2) = -t(x_2)$$

so $t(x_2) = 0$ for all $x_2 \in M^\perp$. Therefore

$$t(x) = t(x_1 + x_2) = x_1 + t(x_2) = A_0x. \qquad \square$$

The invariance argument used to characterize the equivariant estimator A_0Y depends on the assumption

$$\mathscr{L}(\varepsilon) = \mathscr{L}(g_0\varepsilon),$$

where $g_0 = I - 2A_0$. Hence, if ε has a covariance, say $\Sigma = \text{Cov}(\varepsilon)$, this invariance assumption on $\mathscr{L}(\varepsilon)$ implies that

$$\Sigma = g_0\Sigma g_0. \tag{3.15}$$

However, this condition is exactly the same as the regularity assumption which led to Theorem 3.2. To see this, (3.15) is

$$\begin{aligned}\Sigma &= (I - 2A_0)\Sigma(I - 2A_0\Sigma)\\ &= \Sigma - 2A_0\Sigma - 2\Sigma A_0 + 4A_0\Sigma A_0,\end{aligned}$$

which yields

$$2A_0\Sigma A_0 = A_0\Sigma + \Sigma A_0.$$

Now, multiplying this on the left by A_0 gives

$$A_0\Sigma A_0 = A_0\Sigma$$

while multiplication on the right gives

$$A_0\Sigma A_0 = \Sigma A_0$$

so that $\Sigma A_0 = A_0\Sigma$. This is just (3.12) which is equivalent to (3.11). Thus, the invariance assumption on $\mathscr{L}(\varepsilon)$ is very closely connected with the assumption on γ in the definition of regularity.