CHAPTER 9

Finite de Finetti Style Theorems for Linear Models

The implications of invariance assumptions for linear models are investigated here via a finite de Finetti style theorem. Before discussing linear models, we describe a general method for approximating projected measures. Of course, the origins of the method are in the four examples described in Chapter 8. All of the material in this chapter is from Diaconis, Eaton and Lauritzen (1987).

9.1. Approximating extendable probabilities. Consider a measurable space $(\mathbf{X}_2, \mathbf{B}_2)$ which is acted on by a compact group G. Let $\mathscr{P}$ be the set of all G-invariant probability measures defined on $\mathbf{B}_2$. The symbol U denotes a random element of G which has the uniform distribution on G. For each $x \in \mathbf{X}_2$, let

$$H_x = \mathscr{L}(Ux).$$

It was pointed out in Chapter 4 that each $P \in \mathscr{P}$ has the representation

$$P = \int H_x P(dx). \tag{9.1}$$

In other words, every element of the convex set $\mathscr{P}$ can be represented as an average of the H_x's. It is clear that for any $x \in \mathbf{X}_2$,

$$kH_{gx} = H_x \quad \text{for } k, g \in G,$$

because

$$\mathscr{L}(U) = \mathscr{L}(k^{-1}Ug) \quad \text{for } k, g \in G.$$

Now, let Q_x be an "approximation" to H_x. In applications, Q_x is often taken to be a normal distribution whose mean and covariance match those of H_x. Further, in all the applications that I know Q_x satisfies

$$kQ_{gx} = Q_x \quad \text{for } k, g \in G,$$

just as H_x does.

Next consider a second measurable space $(\mathbf{X}_1, \mathbf{B}_1)$ and a measurable map

$$\pi: \mathbf{X}_2 \to \mathbf{X}_1.$$

Think of a map π as a projection (as it was in Chapter 8). Then, let

$$\mathscr{P}_{21} = \{\pi P | P \in \mathscr{P}\}$$

be the set of projected invariant measures on $(\mathbf{X}_1, \mathbf{B}_1)$. Because of (9.1),

$$\pi P = \int \pi H_x P(dx) \tag{9.2}$$

so that all the elements in the convex set $\mathscr{P}_{21}$ are averages of the πH_x's. The elements of $\mathscr{P}_{21}$ are just those probabilities P_1 on $(\mathbf{X}_1, \mathbf{B}_1)$ which have G invariant extensions to $(\mathbf{X}_2, \mathbf{B}_2)$, extensions in the sense that there is a $P \in \mathscr{P}$ such that $\pi P = P_1$.

The following result, which captures the essence of the argument used throughout Chapter 8, provides an upper bound on the variation distance between (i) an element πP of $\mathscr{P}_{21}$ and (ii) the closest approximation to πP based on averages of the πQ_x's.

THEOREM 9.1. *Given* $P \in \mathscr{P}$,

$$\inf_{\mu} \left\| \pi P - \int \pi Q_x \mu(dx) \right\| \leq \sup_x \|\pi H_x - \pi Q_x\| \equiv D, \tag{9.3}$$

where the inf *ranges over all probabilities on* $\mathbf{X}_2$.

PROOF. From (9.2),

$$\begin{aligned} \inf_{\mu} \left\| \pi P - \int \pi Q_x \mu(dx) \right\| &= \inf_{\mu} \left\| \int \pi H_x P(dx) - \int \pi Q_x \mu(dx) \right\| \\ &\leq \left\| \int \pi H_x P(dx) - \int \pi Q_x P(dx) \right\| \leq \int \|\pi H_x - \pi Q_x\| P(dx) \\ &\leq \sup_x \|\pi H_x - \pi Q_x\| = D. \end{aligned}$$

□

The upper bound D in (9.3) should be quite good (as a universal bound) because $H_x \in \mathscr{P}$ for each $x \in \mathbf{X}_2$. Thus, if πQ_x is a reasonable approximation to πH_x, then D should provide a reasonable bound in (9.3). Here is the example of Section 8.2 reworked in the above notation.

EXAMPLE 9.1. Take $\mathbf{X}_2 = R^n$ and let $G = O_n$. Then, for $x \in R^n$,

$$H_x = \mathscr{L}(Ux),$$

where U is uniform on O_n. Since

$$Ux = \|x\| U\left(\frac{x}{\|x\|}\right),$$

H_x is the uniform distribution on

$$\{y | y \in R^n, \|y\| = \|x\|\}.$$

An easy calculation shows that

$$\mathbf{E}Ux = 0$$

and

$$\mathrm{Cov}(Ux) = n^{-1}\|x\|^2 I_n.$$

For this example, take

$$Q_x = N\big(0, n^{-1}\|x\|^2 I_n\big),$$

which is the normal distribution on R^n with the same mean and covariance matrix as H_x.

Now, let $\mathbf{X}_1 = R^k$ with $k < n$ and consider

$$\pi = (I_k \quad 0)\colon k \times n$$

as the projection from R^n to R^k. Given an O_n-invariant probability P on R^n, its projection on R^k is πP. According to Theorem 9.1, the variation distance between πP and the closest average of the πQ_x's is bounded above by

$$D = \sup_x \|\pi H_x - \pi Q_x\|.$$

But

$$\pi H_x = \mathscr{L}(\pi Ux) = \mathscr{L}(\|x\|V),$$

where V is the vector in R^k which is the first k coordinates of a random vector in R^n which is uniform on the n sphere. Also,

$$\pi Q_x = N\big(0, n^{-1}\|x\|^2 I_k\big)$$

because

$$\pi\pi' = I_k.$$

Thus,

$$\begin{aligned} D &= \sup_x \big\| \mathscr{L}(\|x\|V) - N\big(0, n^{-1}\|x\|^2 I_k\big) \big\| \\ &= \big\| \mathscr{L}(V) - N\big(0, n^{-1} I_k\big) \big\| \\ &= \big\| \mathscr{L}(\sqrt{n}\, V) - N(0, I_k) \big\|. \end{aligned}$$

Therefore the calculation of D in this example reduces to finding the variation distance between the $N(0, I_k)$ distribution and $\mathscr{L}(\sqrt{n}\, V)$. An upper bound on this distance was given in Proposition 7.6 (with p replaced by k). This completes the example. □

Here are a few remarks concerning the above example which are also valid in other examples. First, Q_x is a terrible approximation to H_x; they are in fact mutually singular. However, πQ_x is a good approximation to πH_x. This is what matters in applications. Second, both Q_x and H_x are invariant functions of x so

that the averages

$$\int Q_x \mu(dx), \qquad \int H_x \mu(dx)$$

can be written as averages over a maximal invariant under the group action on $\mathbf{X}_2$. In Example 9.1, a maximal invariant is $x \to \|x\|$ and the above averages are obviously just averages over $\|x\|$. Further,

$$x \to \|\pi H_x - \pi Q_x\|$$

is also an invariant function of x. This often makes the calculation of D (or bounding D above) an achievable task. In Example 9.1, the sup was in fact calculated explicitly by observing that for all $x \neq 0$,

$$\|\pi H_x - \pi Q_x\| = \|\pi H x_0 - \pi Q x_0\|,$$

where x_0 is a fixed nonzero vector.

The method described above is from Diaconis, Eaton and Lauritzen (1987) where it is applied to a variety of univariate and multivariate examples. In the examples discussed thus far, it is clear that the "appropriate" Qx's come from an associated "infinite" theorem. This is not at all clear in the description of the above method, but in every example that I know, there is some "infinite" theorem lurking in the background.

9.2. The general univariate linear model. The goal of this section is to discover the implications of extendability and invariance in the context of linear models. In a finite dimensional inner product space $(V, (\cdot, \cdot))$ consider an observation vector Y whose mean vector μ lies in a known linear subspace $M \subseteq V$ and whose covariance is $\sigma^2 I$ where I is the identity linear transformation on V. When Y is $N(\mu, \sigma^2 I)$, it is clear that

$$\mathscr{L}(Y) = \mathscr{L}(gY) \tag{9.4}$$

for all orthogonal transformations g such that

$$gx = x \quad \text{for all } x \in M.$$

The group of all such orthogonal transformations is denoted by $O(M)$. Rather than assume Y has a normal distribution, it is only assumed that (9.4) holds, namely, that the distribution of Y is $O(M)$-invariant.

Now, let $(V_1, (\cdot, \cdot)_1)$ be another finite dimensional inner product space and assume that π: $V \to V_1$ is a linear transformation which satisfies

$$\pi\pi' = I_1,$$

where I_1 is the identity on V_1. This π is an example of what we have been calling a "projection." With $Y_1 = \pi Y$, the mean vector of Y_1 is $\mu_1 = \pi\mu$ which is an element of $M_1 = \pi(M)$. Obviously M_1 is a linear subspace of V_1.

Given that $P = \mathscr{L}(Y)$ satisfies

$$gP = P, \qquad g \in O(M), \tag{9.5}$$

the problem we discuss here is: What can we say about $\mathscr{L}(Y_1) = \pi P$? The example discussed in Section 8.4 is a special case of this problem. In essence, the

result below provides a bound on the variation distance between πP and a mixture of the distributions $\{N(\mu_1, \sigma^2 I_1);\ \mu_1 \in M_1,\ \sigma^2 \geq 0\}$.

The approach to this problem is that described in the previous section. To this end, let U have a uniform distribution on $O(M)$ and for each $x \in V$, write $x = x_1 + x_2$ with $x_1 \in M$ and $x_2 \in M^{\perp}$ ($M^{\perp}$ is the orthogonal complement of M). Then

$$H_x = \mathscr{L}(Ux) = \mathscr{L}(x_1 + Ux_2)$$

since $U \in O(M)$. It can be shown [see Diaconis, Eaton and Lauritzen (1987)] that

$$\mathbf{E}Ux = x_1$$

and

$$\operatorname{Cov}(Ux) = \frac{\|x_2\|^2}{n - m} C,$$

where n is the dimension of V, m is the dimension of M and C is the orthogonal projection onto $M^{\perp}$. Therefore

$$\begin{aligned} \mathbf{E}\pi Ux &= \pi x_1, \\ \operatorname{Cov}(\pi Ux) &= \frac{\|x_2\|^2}{n - m} \pi C \pi'. \end{aligned} \tag{9.6}$$

To apply Theorem 9.1, pick Q_x to be the $N(x_1, \|x_2\|^2 (n - m)^{-1} I_2)$ distribution on V. Then

$$\pi Q_x = N\left(\pi x_1, \|x_2\|^2 (n - m)^{-1} I_1\right)$$

is a normal distribution on V. Therefore

$$\begin{aligned} D &= \sup_x \|\pi H_x - \pi Q_x\| \\ &= \sup_x \left\| \mathscr{L}(\pi x_1 + \pi U x_2) - N\left(\pi x_1, \|x_2\|^2 (n - m)^{-1} I_1\right) \right\| \\ &= \sup_x \left\| \mathscr{L}\left(\pi U \frac{x_2}{\|x_2\|}\right) - N\left(0, (n - m)^{-1} I_1\right) \right\| \\ &= \left\| \mathscr{L}(\pi U x_0) - N\left(0, (n - m)^{-1} I_1\right) \right\|, \end{aligned} \tag{9.7}$$

where x_0 is any fixed vector of length 1 in $M^{\perp}$.

LEMMA 9.1. *Let k be the dimension of V_1 and set*

$$A_0 = \pi C \pi'. \tag{9.8}$$

For $k \leq n - m - 4$,

$$\left\| \mathscr{L}(\pi U x_0) - N\left(0, (n - m)^{-1} A_0\right) \right\| \leq 2 \frac{k + 3}{n - m - k - 3}. \tag{9.9}$$

PROOF. See Diaconis, Eaton and Lauritzen (1987), Proposition A.1. □

We now assume $k \leq n - m - 4$ and A_0 in (9.8) has rank k.

THEOREM 9.2. *The variation distance between πP and the closest mixture of the normal distributions $N(\mu_1, \sigma^2 I_1)$ with $\mu_1 \in M_1$ and $\sigma^2 \geq 0$ is bounded above by*

$$\beta_n(A_0) = 2\frac{k+3}{n-m-k-3} + 2\left[(\det A_0)^{-1/2} - 1\right]. \tag{9.10}$$

PROOF. It suffices to show that D in (9.7) is bounded above by $\beta_n(A_0)$. But

$$\begin{aligned} D &= \left\| \mathscr{L}(\pi U x_0) - N\left(0, (n-m)^{-1} I_1\right) \right\| \\ &\leq \left\| \mathscr{L}(\pi U x_0) - N\left(0, (n-m)^{-1} A_0\right) \right\| \\ &\quad + \left\| N\left(0, (n-m)^{-1} A_0\right) - N\left(0, (n-m)^{-1} I_1\right) \right\|. \end{aligned}$$

The first of the two summands is bounded above by $2(k+3)/(n-m-k-3)^{-1}$ according to Lemma 9.1. Because A_0 has all its eigenvalues in $(0, 1]$, it follows easily that

$$\begin{aligned} &\left\| N\left(0, (n-m)^{-1} A_0\right) - N\left(0, (n-m)^{-1} I_1\right) \right\| \\ &\quad = \| N(0, A_0) - N(0, I_1) \| \leq 2\left[(\det A_0)^{-1/2} - 1\right]. \end{aligned}$$

This completes the proof. □

When $\beta_n(A_0)$ is small, Theorem 9.2 implies that πP is close to a distribution generated by first selecting (μ_1, σ^2) according to some distribution and then selecting Y_1 from a $N(\mu_1, \sigma^2 I_1)$ distribution. In other words, the smallness of $\beta_n(A_0)$ implies that Y_1 looks like it was drawn from a normal distribution. Thus the original invariance assumptions on Y have very strong implications for $\mathscr{L}(Y_1)$. These issues are discussed more fully in the next section where the standard univariate regression model is treated.

9.3. The regression model. In R^n consider the usual regression model

$$Y = X\beta + \varepsilon, \tag{9.11}$$

where X is a known $n \times q$ matrix of rank q, β is a $q \times 1$ vector of unknown parameters and ε is the error vector. Thus the regression subspace

$$M = \{\mu | \mu = X\beta, \beta \in R^q\}$$

is of dimension q. It is assumed that

$$\mathscr{L}(Y) = \mathscr{L}(gY) \tag{9.12}$$

for all $g \in O(M)$. This is equivalent to the assumption that

$$\mathscr{L}(\varepsilon) = \mathscr{L}(g\varepsilon)$$

for all $g \in O(M)$.

In this example, the "projection" π is taken to be

$$\pi = (I_k \quad 0)\colon k \times n,$$

where I_k is the $k \times k$ identity and $k > q$. Partition Y as

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

with $Y_1 = \pi Y$. Also partition X as

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

where X_1 is $k \times q$ so X_2 is $(n - k) \times q$. Finally partition ε as

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

with $\varepsilon_1 = \pi\varepsilon$ in R^k. Then the projected regression model is

$$\pi Y = Y_1 = X_1\beta + \varepsilon_1. \tag{9.13}$$

The statistical interpretation of this model description is the following. We do an experiment in which model (9.13) is assumed for Y_1. But we imagine that a larger experiment could have been performed and the invariance assumption (9.12) for the model (9.11) is assumed to hold. Then, the implications of this model assumption [namely (9.11) and (9.12)] are of concern. This is what Theorem 9.2 yields.

We now turn to the evaluation of $\beta_n(A_0)$ in (9.10) for the above regression model. First, the orthogonal projection onto the orthogonal complement of M is

$$C = I_n - X(X'X)^{-1}X'$$

so that

$$A_0 = \pi C\pi' = I_k - X_1(X'X)^{-1}X_1'.$$

Since the dimension of M is q in this example,

$$\beta_n(A_0) = 2\frac{k+3}{n-q-k-3} + \left[(\det A_0)^{-1/2} - 1\right].$$

Now, fix k and q, and let $n \to \infty$. In order to obtain an "infinite" theorem [i.e., $\beta_n(A_0) \to 0$ as $n \to \infty$] for this example, $X\colon n \times q$ must satisfy

$$\lim_{n\to\infty} \det\left(I_k - X_1(X'X)^{-1}X_1'\right) = 1. \tag{9.14}$$

Recall that k and q are fixed so we are thinking of $X_1\colon k \times q$ as a fixed matrix, namely, the design matrix of the experiment actually performed. With X_1 fixed, a necessary and sufficient condition for (9.14) to hold is that

$$\lim_{n\to\infty} (X'X)^{-1} = 0. \tag{9.15}$$

Equation (9.15) means that each element of the $q \times q$ matrix $(X'X)^{-1}$ converges

to 0 as $n \to \infty$. The statistical interpretation of (9.15) is that the parameter vector β is consistently estimated (by least squares) since the covariance matrix of the least squares estimator of β is $\sigma^2(X'X)^{-1}$. This assumes that Y and hence ε has a covariance matrix which is $\sigma^2 I_n$ when (9.11) holds.

The point of the above discussion is that the conditions for the existence of an "infinite" theorem have a direct statistical interpretation in terms of the estimation of β. For a further discussion of the relationship between "infinite" theorems and statistical interpretations, see Lauritzen (1988).