Table 4.2

Grand R^2's

		$M=2$	4	10	100
$R_2^2=0$	$T=5$	.87	.64	.50	.43
	10	.89	.87	.84	.75
	100	.9	.9	.9	.89
$R_2^2=.8$	$T=5$	.89	.85	.83	.8
	10	.89	.87	.85	.81
	100	.9	.9	.9	.89

M = number of models, T = number of observations
$R^2=.9$; $R_2^2=R_3^2=\cdots=R_M^2$

sample size, $T=5$. For reasonably large sample sizes compared with the number of models, the grand R^2 is almost equal to the maximum R^2. This result derives from the fact that for reasonably large T, the marginal likelihoods are extremely sensitive to R^2. Imperceptible differences in R^2s translate into large differences in posterior probabilities, and all but the best model will be assigned nearly zero posterior probability.

CHAPTER 5

INTERPRETIVE SEARCHES

In this chapter we discuss searches designed to inter pret multi-dimensional evidence. A completely speci fied model usually contains a large number of collinea explanatory variables, and the least-squares estimate that result are rarely "acceptable." Various constraint on the parameters may be imposed to "improve" th estimate, and one among many constrained least squares estimates is usually selected to convey the dat evidence.

A pair of fictitious examples illustrate the phenome non.

Example 1. The demand for oranges D is thought t depend negatively on the price of oranges P, positivel on the price of grapefruit π, and positively on mone income Y. The following regression is estimated (wit standard errors in parentheses)

$$\log D = 7.0 + \underset{(.3)}{.1}\log P + \underset{(.4)}{.2}\log \pi + \underset{(.2)}{.6}\log Y.$$

Unhappily, the direct price elasticity—the coefficient o $\log P$—has the wrong sign. Neither price coefficient i

significantly different from zero. In an effort to "improve" these results, the following constrained regression is estimated:

$$\log D = 4.0 - \underset{(.1)}{.2} \log\left(\frac{P}{\pi}\right) + \underset{(.2)}{.5} \log\left(\frac{Y}{P}\right).$$

The constraint that implies this equation is suggested by economic theory; it is not rejected at the 5% level, and it yields estimates that are both the right sign and also statistically significant. The second regression is, therefore, selected to convey the content of the data.

Example 2. Consumption expenditures C_t in quarter t are thought to depend especially on income receipts in that quarter Y_t, but also somewhat on receipts in previous quarters $Y_{t-1}, Y_{t-2}, \ldots$. To estimate this relationship, the explanatory variables $Y_t, Y_{t-1}, \ldots$ are sequentially included in the equation until one of the variables becomes statistically insignificant. This procedure yields the estimated relationship

$$C_t = \underset{(.2)}{4.0} + \underset{(.1)}{.3} Y_t + \underset{(.1)}{.4} Y_{t-1} + \underset{(.1)}{.1} Y_{t-2}.$$

These two examples illustrate, first, a *contracting search*, in which a series of constraints is imposed on a general model and, second, an *expanding search*, in which the assumptions implicit in an initial model are relaxed and a variety of more general models are estimated. In both cases the most general model leads to, or would lead to "unacceptable" results. In the case of the expanding search it is assumed at the outset that the data could not "support" the most general model. A severe set of constraints is initially imposed and then gradually loosened. The contracting search, on the other hand, occurs when the researcher discovers that, in fact, the data would not support the most general model, and he imposes a series of constraints designed to "improve" the results.

The most important feature of these examples is the fact that the constraints are thought to be, a priori, likely. If the constraints were certain, they would have been imposed without testing. The researcher is, in fact, less confident than this. He feels that the constraint may be "approximately true," but he checks with the data to "make sure." If the constraint "works," he will impose it; otherwise he will not. To put this another way, the researcher has a priori knowledge about some parameter or some linear combination. If the sample evidence is sufficiently strong, he will disregard that information. Given weak evidence, he may prefer to

use his a priori estimate. By definition, then, the intent of an interpretiv search is to integrate into the data analysis uncertain a priori informatio In the absence of such information, no interpretive search should b performed.

The Bayesian solution to this problem is quite straightforward. The dat evidence is summarized in terms of the unconstrained equation and i sufficient statistics. The evidence is interpreted by bouncing the data off prior distribution where the word "interpret" refers to the process updating one's opinions from prior to posterior distribution in response t the data evidence. A Bayesian evaluation of interpretive searches thu amounts to the question: does an interpretive search lead to a descriptic of the uncertainty similar to a posterior distribution corresponding to son prior? A secondary issue is whether anyone actually holds that pri opinion.

The failure of ad hoc interpretive searches is twofold. First, it may b difficult to find a prior distribution that makes a search seem reasonabl But much more important is the fact that the output of an interpretiv search is an interpretation of the data evidence built on some implicit pri information. This interpretation is relevant to the reader only to the exte that he accepts the implicit prior information as his own, and only then he understands that it is already built into the result. Publication of th output of an interpretive search is thus equivalent to publication of posterior distribution without either the sample result or the prior. Public tion of the search process is useful only in simple cases when the procedu simply and unambiguously reveals the prior. Most interpretive searches a terribly complex and would be almost impossible to comprehend even they were fully reported. An interpretive search is thus an inefficient wa to use ill-defined, uncommunicable prior information.

That uncertain prior information is used in the evaluation of nonexpe mental evidence is incontrovertible. Nonexperimental models worthy degrees of belief almost always have large numbers of collinear explan tory variables. The amount we can learn from the data about individu parameters in these models would be almost nil if there were no pri information that effectively constrained the ranges of at least some of th parameters. The mining of data that is common among nonexperiment scientists constitutes prima facie evidence of the existence of prior inform tion. Arguments concerning the use of prior information should th address the question of how rather than whether prior information shou be used.

There are at least three alternative approaches that may be taken wi respect to the use of prior information in a regression model.

1. *Complete and understandable description of the sample likelihood function.* We may decide that a researcher should report only the likelihood function. Prior information is difficult to specify personally and may vary considerably among intended readers. We may prefer, therefore, to report the evidence and not to interpret it.
2. *Bayesian analysis.* In one or two dimensions, a likelihood function may be straightforwardly described. In higher dimensions, a likelihood function defies intelligible reporting. Described least pretentiously, a Bayesian analysis is merely a tool for exploring likelihood functions. Difficulties in specifying a personally or a publicly acceptable prior distribution should be dealt with by performing a sensitivity analysis designed to characterize as generally as possible the mapping from prior to posterior distribution. In fact, unless he has a strong reason to believe that his priors are somehow superior to his readers', a researcher's only obligation is to report this mapping as informatively as possible.
3. *Interpretive search.* The unintelligibility of the complete likelihood function has led most researchers to use interpretive searches that involve fitting and refitting the equation with various a priori likely constraints. One of the perhaps hundreds of equations is selected and reported, often as if the others had never been estimated. The resulting estimate involves an unknown and perhaps an undesirable mixture of prior and sample information. It, furthermore, constitutes an interpretation of the evidence built, surprisingly, without a theory of interpretation.

The choice between a complete description of the likelihood function and the Bayesian approach involves *only* a disagreement over how to report results. A Bayesian merely explores and reports the region of the parameter space that is favored by the data by computing how the likelihood function affects various prior distributions. In higher dimensional problems, the Bayesian approach seems viable, but the likelihood approach does not, which is another way of saying that I think it is possible to identify and to choose the critical features of multidimensional priors. The choice between the Bayesian approach and the interpretive search approach is, however, a choice between theory and "ad hockery." Interpretative searches lead to ill-defined use of ill-defined prior information. They are an abuse that has led many to discount completely all statistical analyses with nonexperimental data. It is highly misleading, however, to regard them to be complete evils. Rather, they are a commonsense solution designed by intelligent men to complete an unworkable incomplete theory of inference. As we see in this chapter and again in later

chapters, intuition and common sense often lead in a desirable direction What we are proposing is only a formal structure to police our intuitive instincts and to help avoid judgmental errors. Never do we desire cessation of common sense.

The rules that are used to direct an interpretive search are rarely sufficiently well defined to be written mathematically. An incomplete list of hypothetical rules will give some flavor of the great menu of search strategies. For *contracting sequential searches*, estimate the complete unconstrained model and do one of the following:

a. Drop all variables that have t values less than some cutoff point.
b. Drop the variable with the lowest t value, refit the equation, and continue until all coefficients have significant t's.
c. Specify an a priori sequence of variables. If the first is insignificant drop it and refit. Proceed similarly with the second, and terminate the process when a variable is reached that has a significant t.
d. Apply a linear transformation to the explanatory variables that make them orthogonal, and drop any of the new variables with insignificant coefficients.
e. Proceed as in either (b)–(d) but terminate the search when the coefficient on a particular variable is (1) positive or (2) significantly positive or (3) not significantly negative.

For *expanding sequential searches*, estimate a constrained model and do one of the following:

a. Add sequentially the omitted variables in a predetermined order and terminate when a variable has an insignificant coefficient.
b. Add only one other variable selected to maximize the R^2.
c. Proceed as in either (a) or (b), but seek to find an equation that yields a significantly positive coefficient on a particular variable.

For *nonsequential searches*, select the regression equation that yields:

a. The highest $\bar{R}^2$.
b. The biggest percentage of statistically significant coefficients.
c. Estimated residuals that are not autocorrelated.
d. The most number of coefficients with the "right" signs.
e. Some complex combination of (a)–(d).

Only the simplest of these rules yields to a theoretical classical analysis In this chapter we explore a few of them from that point of view and ma

others from the Bayesian view. As far as a Bayesian is concerned the effectiveness of a rule depends on how well it implements prior information and how relevant that prior information may be. We argue that this is, in fact, the only question that should be asked. Since classical inference includes no theory of learning and no prior information, classical analysis instead evaluates rules in terms of their sampling properties. This material is discussed in Sections 5.2 and 5.3, the first dealing with the analysis of several simple *ad hoc* rules, the second dealing with the Stein–James estimators and "ridge" estimators.

This classical approach supposes that the researcher has a formal point estimation problem with a conveniently chosen quadratic loss function. Interpretive search rules are then evaluated in terms of the expected loss they imply. The formal shortcoming of this approach is that, at best, it can determine only whether a search estimator is admissible or not—an estimator being inadmissible if there exists another estimator that yields smaller expected loss regardless of the true parameter value. But since the class of admissible estimators is enormous, ruling out inadmissible estimators is only modestly useful. A classical approach often concludes (rather sheepishly?) that choice among the set of admissible estimators depends in some vague way on prior information.

It makes sense to me to begin the analysis with prior information—Bayes estimators derived from proper prior distributions are always admissible, anyway. More importantly, I think the estimation framework does not capture the essential motivation for interpretive searches, that is the pooling of prior information with the data information. It thereby encourages the arbitrary distortion of the data evidence, since it suggests that a "better" estimator results by adjusting the least-squares estimator toward a location not necessarily related to prior information. Further discussion of these points is reported in Section 5.5.

In the first section the problem of this chapter is described as the choice of constrained least-squares estimates, and it is shown that all constrained least-squares estimates lie on an ellipsoid. One shortcoming of a classical analysis is that it considers the very restricted problem of selecting one of only two (arbitrarily chosen) points on this ellipsoid.

As mentioned before, classical analysis of search rules is reported in Sections 5.2 and 5.3. A Bayesian analysis of the same estimation problem is discussed in Section 5.4, and comments are given in Section 5.5. In Section 5.6 we develop a correspondence between search strategies and prior distributions. It is shown that certain classes of priors are implied by certain search strategies in the sense that a Bayesian with such a prior can (loosely) approximate his posterior distribution with a set of constrained least-squares points.

Multicollinearity is discussed in Section 5.7. A distinction is drawn

between the weak evidence problem and the interpretation problem. Multi collinearity implies weak evidence in the sense that coefficient standar errors are large, but nothing can be done about that except getting mor data. A more confusing consequence of collinearity is that the apparer sample evidence about one parameter depends on the prior informatio about other parameters. Collinearity, therefore, creates an incentive to us carefully formulated prior information.

I do not believe that anyone could meaningfully specify a complet multivariate prior distribution. Furthermore, readers are certain to vary i their judgments. For both reasons, it is necessary to perform a sensitivit analysis that determines the sensitivity of features of the posterior distribu tion to changes in the prior distribution. Local sensitivity analysis discussed in Section 5.7 and global sensitivity analysis in Section 5.8.

A game of definitions is reported in Section 5.9. The words identifiabl estimable, publicly informative, and the phrase "leads to a consensus" a shown to be equivalent. It is also pointed out that although a parameter may not be identified, the experiment may nonetheless yield informatic about θ (because of prior dependence between θ and some other param ters). Lastly, an example is reported in Section 5.10.

Before proceeding, one shortcoming of this chapter must be acknow edged. Almost exclusively, our attention focuses on the choice of poi estimates, and tends to ignore the choice of interval around the estimat or other measures of dispersion. This reflects the state of theoretic developments, not the importance of measures of dispersion.

5.1 The Family of Constrained Estimates

The problem under study in this chapter is the choice of one or mo constrained regression estimates that jointly imply an "adequate" interpr tation of the data evidence. As a preliminary it is useful identify the set all constrained estimates. It is trivial to show that if all constraints of t form $\mathbf{R}\boldsymbol{\beta}=\mathbf{r}$ are allowed, then any value of $\boldsymbol{\beta}$ is a constrained regressi for some value of $\mathbf{R}$ and $\mathbf{r}$. If, however, we consider only constraints of t form $\mathbf{R}\boldsymbol{\beta}=\mathbf{0}$, the family of constrained estimates is an ellipsoid describ in Theorem 5.1 and pictured in Figure 5.1. No interpretation of the choi $\mathbf{r}=\mathbf{0}$ seems possible classically, but from the Bayesian point of view t amounts to assuming a prior distribution that is located at the origin.

THEOREM 5.1 (FEASIBLE ELLIPSOID). *A constrained least-squares es mate computed subject to a set of constraints* $\mathbf{R}\boldsymbol{\beta}=\mathbf{0}$ *lies on the ellipsc*

$$\left(\boldsymbol{\beta}-\frac{\mathbf{b}}{2}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}-\frac{\mathbf{b}}{2}\right)=\frac{\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}}{4} \qquad (5.$$

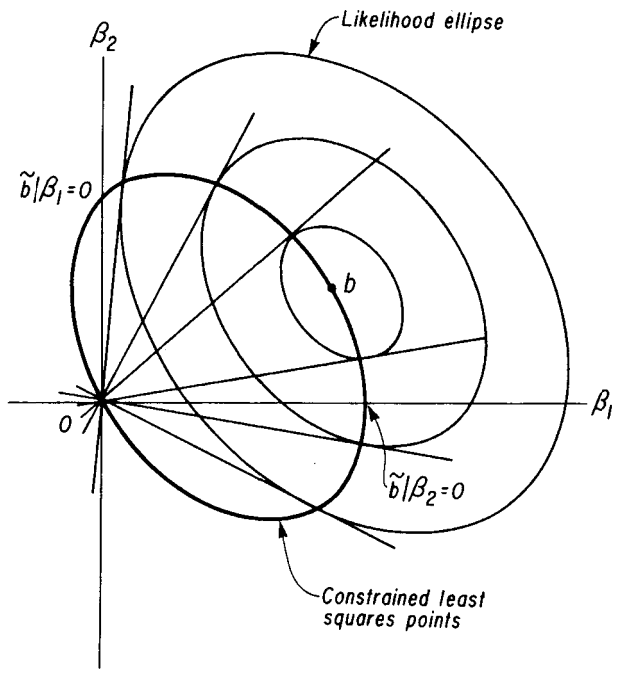

Fig. 5.1 The ellipse of constrained least-squares points.

where **b** *is the unconstrained least-squares vector. Furthermore, any point on this ellipsoid is a constrained estimate for some* **R**.

Proof: Equation (3.14) in Chapter 3 is the formula for computing a constrained estimate. Simply insert that value into (5.1). Conversely, any point on the ellipsoid (5.1), say, $\tilde{\boldsymbol{\beta}}$, is a least-squares estimate subject to the constraint $(\mathbf{b}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}=\mathbf{0}$. (Verification left to reader.)

The set of constrained estimates described in Equation (5.1) is merely a translated likelihood ellipsoid. It is located at half the least-squares vector and travels through the origin and **b**. We argue below that if a researcher can select the location but is unable or unwilling to describe more fully his prior, then he should be interested in all points in this ellipsoid, but no other points. Incidentally, any origin $\mathbf{b}^*$ may be selected or, equivalently, constraints of the form $\mathbf{R}(\boldsymbol{\beta}-\mathbf{b}^*)=\mathbf{0}$ may be considered. It is easy to show that (5.1) continues to apply but with $\boldsymbol{\beta}-\mathbf{b}^*$ and $\mathbf{b}-\mathbf{b}^*$ replacing $\boldsymbol{\beta}$ and **b**.

An interpretive search is a procedure for selecting points from among the set of feasible points (5.1). An *interpretive-search estimator* can thus be

generally defined as a weighted average of points on the feasible ellipsoi

$$\hat{\boldsymbol{\beta}}^s=\int_{\mathbf{R}}\omega(\mathbf{R})\,\hat{\boldsymbol{\beta}}\,(\mathbf{R})\,d\mathbf{R}, \qquad \int_{\mathbf{R}}\omega(\mathbf{R})\,d\mathbf{R}=1,$$

where $\hat{\boldsymbol{\beta}}(\mathbf{R})$ is a constrained estimate and $\omega(\mathbf{R})$ is a weight functio Usually, only a finite subset of the feasible points is considered. Sometime a set of at most k linearly independent constraints is identified, fo example, $\beta_i=0, i=1,\ldots,k$, and only estimates that involve subsets of thi set of constraints are computed. In two dimensions the constraints $\beta_1=$ and $\beta_2=0$ imply the four different estimates illustrated in Figure 5.1: $\mathbf{b}_{\{2\}}$, $\mathbf{b}_{\{1\}}$, and the origin, where $\mathbf{b}_{\{i\}}$ is the estimate subject to the it constraint. Given k constraints, there are 2^k ways of imposing them, and 2 different constrained estimates.[1] Let I be a subset of the first k integers; le $\mathbf{b}_I$ be a constrained estimate, then a (discrete) interpretative search estima tor is

$$\hat{\boldsymbol{\beta}}^s=\sum_I \omega_I(\mathbf{Y},\mathbf{X})\mathbf{b}_I, \qquad (5.2$$

where we have written the weight function ω_I to indicate its possibl dependence on the data, **Y** and **X**.

The interpretive search estimator (5.2) can thus be built in three steps (1) An *origin* is selected. This restricts the set of constrained estimates to li on the ellipsoid (5.1). (2) A set of k linearly independent constraints (*coordinate system*) is chosen. This further restricts the set of interestin constrained least squares points from the ellipsoidal continuum to a set o 2^k points. (3) Last, a weighting function $\omega_I(\mathbf{Y},\mathbf{X})$ is identified.

Measured in terms of its effect on reducing the set of interestin constrained least-squares points, the choice of origin is the most critica decision. After that, the choice of coordinate system is important. A shortcoming of the classical analysis of interpretative searches now to b discussed is its failure to comment meaningfully on either the choice o origin or the choice of coordinate system.

5.2 Classical Evaluation of Ad Hoc Rules for Interpretive Searches

In this section we report a classical analysis of interpretive searches. Bot an origin and a coordinate system for imposing constraints are assumed t be known before the analysis begins. Furthermore, only very specia weight functions may be considered.

[1]Computation of the 2^k regressions is discussed by Schatzoff et al. (1968), Garside (1965), ar Furnival (1971).

Consider, first, the two variable linear regression

$$\mathbf{Y}=\mathbf{x}\beta+\mathbf{z}\gamma+\mathbf{u}, \qquad E(\mathbf{u})=\mathbf{0}, V(\mathbf{u})=\sigma^2\mathbf{I},$$

where $\mathbf{Y},\mathbf{x},\mathbf{z}$ and $\mathbf{u}$ are $T\times 1$ vectors and β and γ are scalar parameters. The least-squares estimator is

$$\begin{bmatrix} b \\ g \end{bmatrix}=\begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{z} \\ \mathbf{z}'\mathbf{x} & \mathbf{z}'\mathbf{z} \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{x}'\mathbf{Y} \\ \mathbf{z}'\mathbf{Y} \end{bmatrix}$$

which is unbiased with variance-covariance matrix

$$V(b,g)=\sigma^2\begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{z} \\ \mathbf{z}'\mathbf{x} & \mathbf{z}'\mathbf{z} \end{bmatrix}^{-1}. \tag{5.3}$$

An alternative estimator with γ set to zero (i.e., with $\mathbf{z}$ omitted) is

$$\begin{bmatrix} \tilde{\beta} \\ \tilde{\gamma} \end{bmatrix}=\begin{bmatrix} (\mathbf{x}'\mathbf{x})^{-1} & \mathbf{x}'\mathbf{Y} \\ 0 & \end{bmatrix}.$$

The expected value of $\tilde{\beta}$ is

$$E(\tilde{\beta})=(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'E(\mathbf{Y})$$

$$=(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'(\mathbf{x}\beta+\mathbf{z}\gamma)=\beta+(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{z}\gamma.$$

The bias of $(\tilde{\beta},\tilde{\gamma})$ is, therefore,

$$E\begin{bmatrix} \tilde{\beta} \\ \tilde{\gamma} \end{bmatrix}-\begin{bmatrix} \beta \\ \gamma \end{bmatrix}=\begin{bmatrix} (\mathbf{x}'\mathbf{x})^{-1} & \mathbf{x}'\mathbf{z}\gamma \\ & -\gamma \end{bmatrix}, \tag{5.4}$$

a linear function of γ, taking on the value of zero only at $\gamma=0$. The variance is straightforwardly calculated as

$$V(\tilde{\beta},\tilde{\gamma})=\sigma^2\begin{bmatrix} (\mathbf{x}'\mathbf{x})^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.5}$$

The classical theory of estimation suggests choosing between these two estimators on the basis of their sampling properties. This used to mean discarding $(\tilde{\beta},\tilde{\gamma})$ because of its bias given in (5.4). That counsel has been disregarded in practice; researchers often report $(\tilde{\beta},\tilde{\gamma})$ even when, with essential certainty, γ is not equal to zero. Although the least-squares estimator has minimum variance among linear unbiased estimators, few researchers are willing to accept "peculiar" estimates, and the standard operating procedure is to search for constraints that yield "acceptable" estimates. The fact that the resulting estimator is neither unbiased, linear, nor "best" is no large deterrent to a person whose research project would be dubbed "fruitless" if it were summarized in a nonsensical estimate.

The overwhelming body of nonexperimental data analysis that rests on the obviously shaky foundation of interpretative searches has understand-

ably generated interest among theoretical statisticians. It is currently popular now to discount unbiasedness as an irrelevant criterion conjured up to make the problem soluble. Instead of an unbiased estimator, current wisdom suggests an estimator that yields an estimate close to the true parameter on the average. A tractable distance function for measuring closeness is the squared deviation from the true value of the parameter, and the resulting criterion is the mean squared error. For a one-dimensional parameter θ with estimator $\hat{\theta}$, the mean squared error is defined by

$$MSE(\hat{\theta},\theta)=E\left[(\theta-\hat{\theta})^2|\theta\right]$$

which is informatively rewritten as

$$MSE(\hat{\theta},\theta)=E\left[(\theta-E\hat{\theta}+E\hat{\theta}-\hat{\theta})^2|\theta\right]$$

$$=V(\hat{\theta})+(\theta-E[\hat{\theta}|\theta])^2$$

the sampling variance plus the square of the bias. An estimator according to this criterion is judged desirable if it has small mean squared error. It is readily seen that an estimator may be deemed desirable even though it is biased, that is if the variance is reduced enough to offset the (square of the) bias.

The multivariate generalization of this criterion is the mean squared error matrix

$$MSE(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})=E\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'|\boldsymbol{\theta}\right]$$

$$=V(\hat{\boldsymbol{\theta}})+(\boldsymbol{\theta}-E[\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}])(\boldsymbol{\theta}-E[\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}])'$$

where $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ are vectors. The reader may verify that the mean-square error of any linear combination $\boldsymbol{\lambda}'\hat{\boldsymbol{\theta}}$ of the estimators is

$$MSE(\boldsymbol{\lambda}'\hat{\boldsymbol{\theta}},\boldsymbol{\lambda}'\boldsymbol{\theta})=\boldsymbol{\lambda}'MSE(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\boldsymbol{\lambda}$$

and it is desirable to have a mean squared error matrix be small in a matrix sense.

Returning to our problem the mean squared error matrix of the least-squares estimator is

$$MSE(b,g)=\sigma^2\begin{bmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{z} \\ \mathbf{z}'\mathbf{x} & \mathbf{z}'\mathbf{z} \end{bmatrix}^{-1} \tag{5.6}$$

whereas the mean-square error matrix of constrained least squares is

$$MSE(\tilde{\beta},\tilde{\gamma})=\sigma^2\begin{bmatrix} (\mathbf{x}'\mathbf{x})^{-1} & 0 \\ 0 & 0 \end{bmatrix}+\gamma^2\begin{bmatrix} (\mathbf{x}'\mathbf{x})^{-2}(\mathbf{x}'\mathbf{z})^2 & -(\mathbf{x}'\mathbf{x})^{-1}(\mathbf{x}'\mathbf{z}) \\ -(\mathbf{x}'\mathbf{x})^{-1}(\mathbf{x}'\mathbf{z}) & 1 \end{bmatrix}.$$

(5.7)

If interest centers on β, we say the estimator $\tilde{\beta}$ is preferred to b according to the mean squared error criterion if

$$\sigma^2(\mathbf{x}'\mathbf{x})^{-1}+\gamma^2(\mathbf{x}'\mathbf{x})^{-2}(\mathbf{x}'\mathbf{z})^2<\sigma^2\left(\mathbf{x}'\mathbf{x}-\mathbf{x}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}\right)^{-1}.$$

Feldstein (1973) and Wallace (1964) write this inequality informatively in terms of the ratio of the mean-square errors

$$\frac{MSE(\tilde{\beta})}{MSE(b)}=1+r_{xz}^2\left(\tau_\gamma^2-1\right) \tag{5.8}$$

where r_{xz} is the correlation between $\mathbf{x}$ and $\mathbf{z}$[2]

$$r_{xz}^2=(\mathbf{x}'\mathbf{z})^2(\mathbf{x}'\mathbf{x})^{-1}(\mathbf{z}'\mathbf{z})^{-1}$$

and τ_γ is the "true t" for testing $\gamma=0$

$$\tau_\gamma^2=\frac{\gamma^2\left[(\mathbf{x}'\mathbf{x})(\mathbf{z}'\mathbf{z})-(\mathbf{x}'\mathbf{z})^2\right]}{\sigma^2(\mathbf{x}'\mathbf{x})}.$$

It is readily seen from (5.8) that the MSE of $\tilde{\beta}$ is less than the MSE of b if and only if

$$\tau_\gamma^2<1. \tag{5.9}$$

This identifies the region in the two-dimensional parameter space within which $MSE(\tilde{\beta})<MSE(b)$. Since the mean squared errors do not depend on β, we may draw a one-dimensional graph of the mean squared error function, Figure 5.2.

The reader should take note of the following features of this figure:

i. Neither $\tilde{\beta}$ nor b dominates the other in the sense of yielding uniformly smaller mean squared error.

ii. $\tilde{\beta}$ does best around the origin $\gamma=0$ but since $MSE(\tilde{\beta})$ is a quadratic function of γ whereas $MSE(b)$ is just a constant, the relative inferiority of $\tilde{\beta}$ is unbounded as γ grows. The difference at the origin is a function of r_{xz}^2.

iii. The origin of this figure is completely arbitrary. That is, there is a whole class of estimators $\tilde{\beta}_{g^*}$ estimated by setting γ to some arbitrary value and calculating

$$\tilde{\beta}_{g^*}=(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'(\mathbf{Y}-\mathbf{z}g^*).$$

This estimator will have a relatively low MSE in the neighborhood of $\gamma=g^*$ but a relatively high MSE elsewhere. Figure 5.2 applies with g^*

[2]If desired, the reader may consider the variables to have had their means removed.

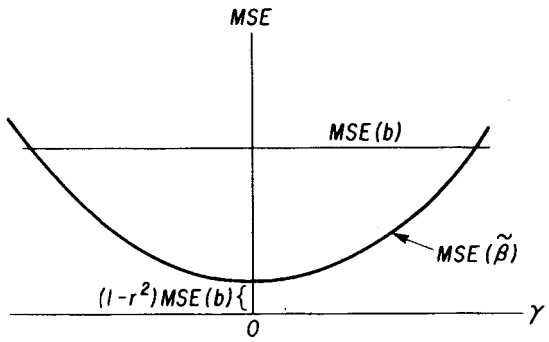

Fig. 5.2 Mean squared error functions.

as the origin and

$$\tau_{\gamma,g^*}^2=\frac{(\gamma-g^*)^2\left[(\mathbf{x}'\mathbf{x})(\mathbf{z}'\mathbf{z})-(\mathbf{x}'\mathbf{z})^2\right]}{\sigma^2(\mathbf{x}'\mathbf{x})}$$

replacing τ_γ^2.

We have yet to discuss an interpretive search strategy, since neith nor $\tilde{\beta}_{g^*}$ requires a search over estimates. The simplest interpretive se that has been analyzed involves calculating both b and $\tilde{\beta}_{g^*}$ (for some v of g^*) and picking the "better" estimate. No doubt, every conceiv criterion of choice has been used in practice. Theoretically, one parti criterion has been subject to much analysis: pick b if the least-squ estimate of γ is significantly different from g^*; otherwise pick $\tilde{\beta}_{g^*}$. T statistic for testing the restriction $\gamma=g^*$ is $t_{g^*}^2=(g-g^*)^2/s^2[$ $\mathbf{z}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{z}]^{-1}$ and the "pretest" estimator more formally is

$$\hat{\beta}^p(g^*,\alpha)=\begin{cases} b & \text{if } t_{g^*}^2>t_\alpha^2 \\ \tilde{\beta}_{g^*} & \text{otherwise} \end{cases} \tag{5}$$

where t_α is the $\alpha/2$ percentage point of the Student's distribution for s arbitrarily chosen value of α.

It is readily seen that both b and $\tilde{\beta}_{g^*}$ are in the class of pretest estim with $\alpha=1$ and $\alpha=0$, respectively. The mean square error of other mem of the class $\tilde{\beta}^p(g^*,\alpha)$ tend to be like the mean squared error of $\tilde{\beta}_{g^*}$ those values of γ that are highly likely to yield insignificant values o that is, for small values of $(\gamma-g^*)^2$. Conversely, $MSE(\tilde{\beta}^p)$ approxim

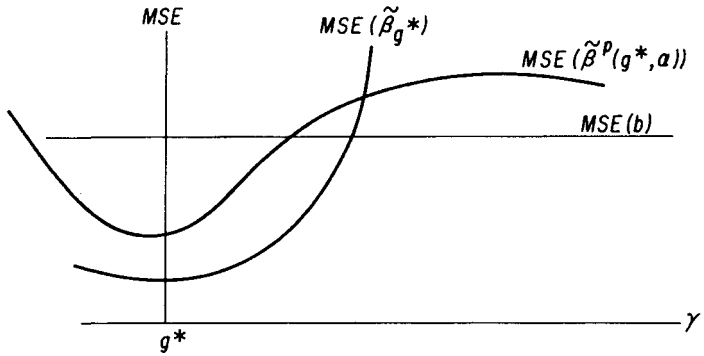

Fig. 5.3 Mean squared error functions.

$MSE(b)$ when b is most likely to be selected, that is, for large values of $(\gamma - g^*)^2$. A typical mean squared error function is depicted in Figure 5.3.

One desirable feature of $MSE(\tilde{\beta}^p)$ is that it is bounded. It exceeds $MSE(\tilde{\beta}_{g^*})$ at $\gamma = g^*$ but attains its minimum there. There is a finite region in which it is the worst of the three estimators. [For other properties of this function, see Wallace and Ashar (1972) and Feldstein (1973).]

We have now identified a two-dimensional infinity of estimates of $\beta, \tilde{\beta}^p(g^*, \alpha)$, $-\infty < g^* < \infty$, and $0 \leqslant \alpha \leqslant 1$. Choice of g^* locates the region in which we will do relatively well, and choice of α determines the amount by which and the region within which this estimator "does better" than ordinary least squares. Which are you to choose? There is simply no answer to this question within the confines of classical inference. That body of statistical theory determines the class of admissible estimators, but does not provide a method of choice from this class. Clearly, however, the choice must depend on where you want to do better than least squares. This surely means the values of γ that you think are most likely and that means prior information about γ. To quote Wallace and Ashar (1972, p. 177):

(1) If one has a strong prior that τ_γ^2 is either greater or less than one half, then no pre-testing is called for. One simply treates $\gamma = 0$ in the latter and $\gamma \neq 0$ in the former if the mean squared error loss function is taken as a guide.

(2) Any intermediate prior about τ_γ^2 can be reflected through choice of α. The stronger the belief that $\gamma \neq 0$, the smaller should be the choice of α and conversely, α should be larger the stronger the prior doubts about the inclusion of z. Of course, if one could cast priors into a precise distributional form, there are both classical and Bayesian procedures to by-pass the pre-testing altogether.

This last sentence brings up the question to be discussed subseque "what is the best way of picking a prior: directly or indirectly by pick search strategy?"[3]

Note that this pretest estimator is a discontinuous function of the since a small change in the data **Y** that is large enough to shift the esti of g from "insignificance" to significance induces a discrete change i estimate from $\tilde{\beta}$ to b. This discontinuity has been shown by Cohen (to imply that this pretest estimator is inadmissible; there is, necess another estimator that has smaller MSE for all values of (β, γ). We ca our procedure of this discontinuity by having a continuous mixing sch such as the following.

Let a weighted-average estimator be

$$\beta_{g^*}^w = \lambda(\mathbf{Y}, \mathbf{x}, \mathbf{z}) b + \left[1 - \lambda(\mathbf{Y}, \mathbf{x}, \mathbf{z})\right] \tilde{\beta}_{g^*}$$

where λ is a continuous function of the data. Feldstein (1970) deri value of λ independent of $(\mathbf{Y}, \mathbf{x}, \mathbf{z})$ that minimizes the mean squared of $\beta_{g^*}^w$:

$$E(\beta_{g^*}^w - \beta)^2 = E\left[\lambda(b - \beta) + (1 - \lambda)(\tilde{\beta}_{g^*} - \beta)\right]^2$$

$$= \lambda^2 MSE(b) + (1 - \lambda)^2 MSE(\tilde{\beta}_{g^*}) + 2\lambda(1 - \lambda) E(b - \beta)(\tilde{\beta}_{g^*} - \beta).$$

Setting the derivative of this with respect to λ equal to zero yields

$$0 = 2\lambda MSE(b) - 2(1 - \lambda) MSE(\tilde{\beta}_{g^*}) + 2(1 - 2\lambda) E(b - \beta)(\tilde{\beta}_{g^*} - \beta$$

which simplifies to

$$\lambda = \frac{\tau_{\gamma, g^*}^2}{1 + \tau_{\gamma, g^*}^2}$$

a function of γ. This suggests using a weight proportional to the squa

[3]There are a couple other developments relating to this model that are worth reportin Note that the usual α-level test of the hypothesis $H_0: \gamma = g^*$ against the alternative H_1: is not an α-level test of the hypothesis that $\tilde{\beta}_{g^*}$ is a better estimator than b. If we wan test those hypotheses we would test

$$H_0': \tau_{\gamma, g^*}^2 = \frac{(\gamma - g^*)^2}{\sigma^2 \left(\mathbf{z}'\mathbf{z} - \mathbf{z}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{z}\right)^{-1}} < 1$$

$$H_1': \tau_{\gamma, g^*}^2 > 1$$

Toro-Vizcarrondo and Wallace (1968) argue that H_0 and H_1 are irrelevant in this conte propose instead α-level tests of H_0' versus H_1'.

the t statistic used to test $\gamma = g^*$

$$\lambda(\mathbf{Y}, \mathbf{x}, \mathbf{z}) = \frac{t_{g^*}^2}{1 + t_{g^*}^2}$$

where

$$t_{g^*}^2 = \frac{(g - g^*)^2}{s^2\left(\mathbf{z}'\mathbf{z} - \mathbf{x}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}\right)^{-1}}.$$

These weights were first suggested by Huntsberger (1955) and have been explored in a Monte Carlo study by Feldstein (1973).

5.3 "Stein" Estimators and Ridge Regression

It is possible to dismiss the "pretest" estimators discussed in the previous section, since they do not, in fact, dominate the least-squares estimator. However, a most provocative result of modern statistical theory is that the least-squares estimator is, in fact, inadmissible when there are more than two coefficients and when the loss function takes a special form.[4] Consider the k-means model

$$Y_i = \xi_i + u_i \qquad i = 1, 2, \ldots, k \tag{5.11}$$

with the u_is having independent normal distributions with mean 0 and variance σ^2. The least-squares estimator and also the maximum likelihood estimator are $\hat{\xi}_{io} = Y_i$, $i = 1, 2, \ldots, k$, or, in vector notation, $\hat{\boldsymbol{\xi}}_0 = \mathbf{Y}$. Assuming the quadratic loss function $L(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}) = (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}})'(\boldsymbol{\xi} - \hat{\boldsymbol{\xi}})$ Stein (1956) and James and Stein (1961) have shown that

- The least-squares estimator $\hat{\boldsymbol{\xi}}_0 = \mathbf{Y}$ is admissible for $k \leqslant 2$. That is, there is no estimator that provides uniformly smaller risk (expected loss) than $\hat{\boldsymbol{\xi}}_0$.
- For $k \geqslant 3$, an alternative estimator

$$\hat{\boldsymbol{\xi}}_1 = \left(1 - \frac{(k-2)\sigma^2}{\mathbf{Y}'\mathbf{Y}}\right)\mathbf{Y}$$

has uniformly smaller risk than $\hat{\boldsymbol{\xi}}_0$. Thus $\hat{\boldsymbol{\xi}}_0$ is inadmissible.

[4] An excellent summary of this literature is given in Zellner and Vandaele (1975). Our discussion is very abbreviated, and the reader should consult Zellner and Vandaele for a fuller treatment. We discuss here the analysis when σ^2 is known. There are similar developments for σ^2 unknown.

"Stein" Estimators and Ridge Regression

The general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ with $\mathbf{u}$ distributed normally mean $\mathbf{0}$ and variance matrix $\sigma^2\mathbf{I}$ can be transformed neatly into k-means problem as in Sclove (1968). Let us find a matrix $\mathbf{P}$ such $\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P} = T\mathbf{I}$, where T is the number of observations. The linear model be written

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\mathbf{P}\mathbf{P}^{-1}\boldsymbol{\beta} + \mathbf{u}. \\ &= \mathbf{W}\boldsymbol{\theta} + \mathbf{u}\end{aligned}$$

where $\mathbf{W} = \mathbf{X}\mathbf{P}$ and $\boldsymbol{\theta} = \mathbf{P}^{-1}\boldsymbol{\beta}$. Premultiplying now by $\mathbf{W}'$, we obtain

$$\begin{aligned}\mathbf{W}'\mathbf{Y} &= \mathbf{W}'\mathbf{W}\boldsymbol{\theta} + \mathbf{W}'\mathbf{u} \\ &= T\boldsymbol{\theta} + T\boldsymbol{\varepsilon}\end{aligned}$$

with $T\boldsymbol{\varepsilon}_n \sim N(\mathbf{0}, \sigma^2\mathbf{W}'\mathbf{W}) = N(\mathbf{0}, \sigma^2 T\mathbf{I})$. Dividing by T, we obtain

$$\frac{\mathbf{W}'\mathbf{Y}}{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

which is precisely the same form as the k-means problem with variance σ^2/T. Thus the estimator

$$\hat{\boldsymbol{\theta}}_1 = \left(1 - \frac{(k-2)T\sigma^2}{\mathbf{Y}'\mathbf{W}'\mathbf{W}\mathbf{Y}}\right)\frac{\mathbf{W}'\mathbf{Y}}{T}$$

dominates least squares

$$\hat{\boldsymbol{\theta}}_0 = \frac{\mathbf{W}'\mathbf{Y}}{T}$$

when the loss function is $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$. The corresponding estimators loss function in terms of the natural parameters $\boldsymbol{\beta}$ are[5]

$$\hat{\boldsymbol{\beta}}_0 = \mathbf{P}\hat{\boldsymbol{\theta}}_0 = \frac{\mathbf{P}\mathbf{W}'\mathbf{Y}}{T} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{b}$$

$$\hat{\boldsymbol{\beta}}_1 = \mathbf{P}\hat{\boldsymbol{\theta}}_1 = \left(1 - \frac{(k-2)\sigma^2}{\mathbf{Y}'\mathbf{Y}}\right)\frac{\mathbf{P}\mathbf{W}'\mathbf{Y}}{T} = \left(1 - \frac{(k-2)\sigma^2}{\mathbf{Y}'\mathbf{Y}}\right)\hat{\boldsymbol{\beta}}_0$$

$$L(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{P}^{-1\prime}\mathbf{P}^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = T^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

[5] The dependence of the loss function on $\mathbf{X}'\mathbf{X}$ is apparently peculiar. It does make sense the following prediction problem. Let Y_f be a future outcome of the dependent variable $Y_f = \mathbf{x}_f'\boldsymbol{\beta} + u_f$, and let $\hat{Y}_f$ be a conditional forecast $\hat{Y}_f = \mathbf{x}_f'\hat{\boldsymbol{\beta}}$. Then the squared prediction error is $(Y_f - \hat{Y}_f)^2 = (\mathbf{x}_f'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + u_f)^2$. Taking the expected value of this squared error and using $E\mathbf{x}_f\mathbf{x}_f' = \mathbf{X}'\mathbf{X}/T$ we obtain

$$E\left[(Y_f - \hat{Y}_f)^2 | \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}\right] = T^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \sigma^2.$$

The equation $E\mathbf{x}_f\mathbf{x}_f' = \mathbf{X}'\mathbf{X}/T$ makes sense if the explanatory variable vectors are drawn independently from a fixed multivariate distribution.

The Stein–James estimator $\hat{\boldsymbol{\beta}}_1$ is just least squares times a shrinkage factor; to put it somewhat differently, it is a "weighted average" of the zero vector and the least-squares vector with weights depending on the χ^2 statistics for testing the restriction $\boldsymbol{\beta}=0$, $\chi^2=\mathbf{Y}'\mathbf{Y}/\sigma^2$.[6] The possibility of negative weights has led Baranchik (1964) to propose a "positive part estimator." Except for the possibility of negative weights, the Stein–James estimator is an interpretative search estimator (5.2).

There is another class of estimators mysteriously known as ridge regression estimators that are shown in Section 5.5 to be interpretive search estimators (5.2). Hoerl and Kennard (1970a, b) note that when $\mathbf{X}'\mathbf{X}$ is nearly singular, calculation of the least-squares estimate $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is subject to "a number of 'errors'." They propose the "ridge estimator"

$$\hat{\boldsymbol{\beta}}^r(c)=(\mathbf{X}'\mathbf{X}+c\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}=(\mathbf{X}'\mathbf{X}+c\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X}\mathbf{b}+c\mathbf{I}\mathbf{0}),$$

where $\mathbf{0}$ is a zero vector and c is some mysteriously chosen scalar. It is readily seen that this is a conditional Bayesian posterior mean with a spherical prior centered at zero and thus has a Bayesian justification.

Hoerl and Kennard's (1970) non-Bayesian arguments in favor of "ridge regression" are difficult for this author to understand. They prove that for any $\boldsymbol{\beta}$ there is a constant c greater than zero such that the mean squared error of $\hat{\boldsymbol{\beta}}^r(c)$ is smaller than the mean squared error of least squares, $\hat{\boldsymbol{\beta}}^r(0)$. Even if we accept the mean-squared-error logic, this result has limited applicability. The constant c is a function of $\boldsymbol{\beta}$, but if $\boldsymbol{\beta}$ is known, there is an even better estimator than the ridge estimator. Hoerl and Kennard also offer informal arguments in favor of ridge regression. They (1970, p. 56) point out that when $\mathbf{X}'\mathbf{X}$ has one or more small eigenvalues "the distance from $\mathbf{b}$ to $\boldsymbol{\beta}$ will tend to be large. Estimated coefficients that are large in absolute value have been observed by all who have tackled live nonorthogonal data problems." It would seem that the average distance between $\mathbf{b}$ and $\boldsymbol{\beta}$ is fully reflected in the sampling variance of $\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, which does indeed have large elements when $\mathbf{X}'\mathbf{X}$ has small eigenvalues. It is hard to imagine that we could cure this disease by shrinking to some arbitrary point. The outcome of such a procedure is, of course, to improve the mean squared error at the arbitrary point, but the cost would usually be worsened mean squared error elsewhere. Whether we want to do this must depend on prior information. This suggests that the last sentence of the quotation in this paragraph might better read: estimated coefficients that are far from a priori likely coefficients have been observed by all who have tackled live nonorthogonal data problems.

[6]The weights need not sum to one. In fact, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$ may be on opposite sides of the origin.

Bayes' Decisions and the Admissibility of Bayes' Rules

Note also that the origin and the "metric" are arbitrary.[7] "Shrinki $\mathbf{A}\boldsymbol{\beta}$ to $\mathbf{A}\mathbf{b}^*$ with observations generated by $\mathbf{Y}-\mathbf{X}\mathbf{b}^*=(\mathbf{X}\mathbf{A}^{-1})(\mathbf{A}\boldsymbol{\beta}-\mathbf{A}\mathbf{b}^*)$ yields $\mathbf{A}\hat{\boldsymbol{\beta}}-\mathbf{A}\mathbf{b}^*=(\mathbf{A}'^{-1}\mathbf{X}'\mathbf{X}\mathbf{A}^{-1}+c\mathbf{I})^{-1}\mathbf{A}'^{-1}\mathbf{X}'(\mathbf{Y}-\mathbf{X}\mathbf{b}^*)=\mathbf{A}(\mathbf{X}'\mathbf{X}$ $c\mathbf{A}'\mathbf{A})^{-1}\mathbf{X}'(\mathbf{Y}-\mathbf{X}\mathbf{b}^*)$. In terms of $\hat{\boldsymbol{\beta}}$ this is

$$\begin{aligned}\hat{\boldsymbol{\beta}}&=\mathbf{b}^*+(\mathbf{X}'\mathbf{X}+c\mathbf{A}'\mathbf{A})^{-1}(\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{X}\mathbf{b}^*)\\&=(\mathbf{X}'\mathbf{X}+c\mathbf{A}'\mathbf{A})^{-1}(\mathbf{X}'\mathbf{X}\mathbf{b}^*+c\mathbf{A}'\mathbf{A}\mathbf{b}^*+\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{X}\mathbf{b}^*)\\&=(\mathbf{X}'\mathbf{X}+c\mathbf{A}'\mathbf{A})^{-1}(\mathbf{X}'\mathbf{X}\mathbf{b}+c\mathbf{A}'\mathbf{A}\mathbf{b}^*)\end{aligned}$$

where $\mathbf{A}'\mathbf{A}$ is an arbitrary symmetric positive definite matrix. Thus class of ridge estimators is as wide as the class of conditional poste means (3.28), but the mean-square error logic has nothing to say ab either the choice of origin $\mathbf{b}^*$ or the choice of "metric" $\mathbf{A}'\mathbf{A}$.

5.4 Bayes' Decisions and the Admissibility of Bayes' Rules

Although I do not think practical pretesting is intended to solve estimation problems just discussed, it is important to understand ho Bayesian would solve them, if he had to. It is simple to show that he wc estimate the parameters with his posterior mean. An important resul that the posterior mean is necessarily an admissible estimator provi that the prior is proper. That an improper prior may lead to an inadmi ble estimator should already be clear from the discussion of Stein's (19 result on the inadmissibility of least-squares, the posterior mean implied an improper diffuse prior.

A general quadratic loss function can be written $l(\boldsymbol{\beta},\hat{\boldsymbol{\beta}})=(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{Q}$ $-\hat{\boldsymbol{\beta}})$, where $\boldsymbol{\beta}$ is the $k\times 1$ vector of regression parameters, $\mathbf{Q}$ is a symm ric positive definite $(k\times k)$ matrix, and $\hat{\boldsymbol{\beta}}$ is a $k\times 1$ decision ve representing the estimate of $\boldsymbol{\beta}$.[8] Making use of the data $\mathbf{Y}$, a Baye would select an estimate $\hat{\boldsymbol{\beta}}$ that minimizes expected posterior loss

$$\min_{\hat{\boldsymbol{\beta}}} E\left[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{Q}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})|\mathbf{Y}\right].$$

It is easy to show that *regardless of the choice of positive definite* $\mathbf{Q}$,

[7]In Chapter 3 it was shown that a posterior mean is different from the least-squares esti because it attempts to satisfy the prior that asserts that $(\boldsymbol{\beta}-\mathbf{b}^*)\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)$ is small. The "origin" refers to the prior mean b^*, and the word "metric" refers to the prior preci matrix $\mathbf{N}^*$ that determines the distance function $(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b})$.

[8]For convenience, $\mathbf{Q}$ is not allowed to be positive semi-definite. If $\mathbf{Q}$ were semi-defini Bayes estimator is nonunique, and the results reported below would not formally a However, even if $\mathbf{Q}$ is semi-definite, the posterior mean can be shown to be admissible.

expression is minimized at the mean of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}(\mathbf{Y}) = E(\boldsymbol{\beta}|\mathbf{Y}),$$

where $\hat{\boldsymbol{\beta}}$ is written as a function of $\mathbf{Y}$ to emphasize the fact that the posterior mean of $\boldsymbol{\beta}$ is a function of the data $\mathbf{Y}$.

With a reasonable assumption about the prior for $\boldsymbol{\beta}$ it can be shown that the posterior mean is an admissible estimator, where the word admissible is now to be defined precisely. The *risk function* given the decision rule $\mathbf{a}(\mathbf{Y})$ is the expected loss conditional on $\boldsymbol{\beta}$,

$$R(\boldsymbol{\beta},\mathbf{a}) = E\left[(\boldsymbol{\beta}-\mathbf{a})'\mathbf{Q}(\boldsymbol{\beta}-\mathbf{a})|\,\boldsymbol{\beta}\right].$$

An estimator $\mathbf{a}_1(\mathbf{Y})$ is said to be *inadmissible* if there is another estimator $\mathbf{a}_2(\mathbf{Y})$ such that $R(\boldsymbol{\beta},\mathbf{a}_2) \leqslant R(\boldsymbol{\beta},\mathbf{a}_1)$ for all $\boldsymbol{\beta}$ with strict inequality $R(\boldsymbol{\beta},\mathbf{a}_2) < R(\boldsymbol{\beta},\mathbf{a}_1)$ for at least one value of $\boldsymbol{\beta}$. Otherwise, $\mathbf{a}_1(\mathbf{Y})$ is said to be *admissible*.

The risk function integrated with respect to the prior on $\boldsymbol{\beta}$ is known as the *Bayes risk*:

$$B(\mathbf{a}) = E\left[R(\boldsymbol{\beta},\mathbf{a})\right].$$

The Bayes estimator $E(\boldsymbol{\beta}|\mathbf{Y})$ minimizes Bayes risk. To verify this, write the Bayes risk as

$$B(\mathbf{a}) = E\left[E(\boldsymbol{\beta}-\mathbf{a})'\mathbf{Q}(\boldsymbol{\beta}-\mathbf{a})|\mathbf{Y}\right]$$

and observe that the expression in the inner brackets is minimized for each value of Y by setting $\mathbf{a} = E(\boldsymbol{\beta}|\mathbf{Y})$.

The following lemma on the uniqueness of the posterior mean is necessary to prove the admissibility of the Bayes decision.

LEMMA. *If an estimator* $\mathbf{a}(\mathbf{Y})$ *has the same Bayes risk as the estimator* $E(\boldsymbol{\beta}|\mathbf{Y})$, *then* $\mathbf{a}(\mathbf{Y})$ *is identically equal to* $E(\boldsymbol{\beta}|\mathbf{Y})$.

Proof: Write $\mathbf{a}(\mathbf{Y})$ as $\mathbf{a}(\mathbf{Y}) = \mathrm{E}(\boldsymbol{\beta}|\mathbf{Y}) + \mathbf{d}(\mathbf{Y})$. The Bayes risk of $\mathbf{a}$ can be written as

$$\begin{aligned} B(\mathbf{a}) &= E\left(\left[E(\boldsymbol{\beta}|\mathbf{Y}) + d(\mathbf{Y}) - \boldsymbol{\beta}\right]'\mathbf{Q}\left[E(\boldsymbol{\beta}|\mathbf{Y}) + d(\mathbf{Y}) - \boldsymbol{\beta}\right]\right) \\ &= E\left(\left[E(\boldsymbol{\beta}|\mathbf{Y}) - \boldsymbol{\beta}\right]'\mathbf{Q}\left[E(\boldsymbol{\beta}|\mathbf{Y}) - \boldsymbol{\beta}\right]\right) + E\left(\left[\mathbf{d}(\mathbf{Y})\right]'\mathbf{Q}\left[\mathbf{d}(\mathbf{Y})\right]\right) \\ &= B(E(\boldsymbol{\beta}|\mathbf{Y})) + E\left(\left[\mathbf{d}(\mathbf{Y})\right]'\mathbf{Q}\left[\mathbf{d}(\mathbf{Y})\right)\right). \end{aligned}$$

Thus $B(a)$ exceeds $B(E(\boldsymbol{\beta}|\mathbf{Y}))$ unless the last term vanishes, which can occur only if $\mathbf{d}(\mathbf{Y}) = \mathbf{0}$ if $\mathbf{Q}$ is positive definite.

THEOREM 5.2 (ADMISSIBILITY OF THE POSTERIOR MEAN) *quadratic loss function* $l(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{Q}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$ *with* $\mathbf{Q}$ *p nite, a proper prior distribution for* $\boldsymbol{\beta}$, *and the normal linea model, the posterior mean* $\hat{\boldsymbol{\beta}}(\mathbf{Y}) = E(\boldsymbol{\beta}|\mathbf{Y})$ *is an admissible est*

Proof: Assume that $\hat{\boldsymbol{\beta}}(\mathbf{Y})$ is not admissible. Then there is estimator $\mathbf{a}$ such that

$$R(\boldsymbol{\beta},\mathbf{a}) \leqslant R(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) \quad \text{for all} \quad \boldsymbol{\beta}$$

$$R(\boldsymbol{\beta},\mathbf{a}) < R(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) \quad \text{for some} \quad \boldsymbol{\beta}.$$

Integrating these risk functions with respect to the prior yields t ity

$$E(R(\boldsymbol{\beta},\mathbf{a})) \leqslant E\left(R(\boldsymbol{\beta},\hat{\boldsymbol{\beta}})\right),$$

but by assumption $\hat{\boldsymbol{\beta}}$ minimizes Bayes risk, and this last inequal an equality. But by the previous lemma, if the Bayes risks are e

For a discussion of the admissibility of Bayes decision rules, may consult Ferguson (1967, Section 2.3).

5.5 Comments on Interpretive Searches

There are two other approaches toward the problem of interp mensional evidence—the likelihood approach and the Bayesiar Comments on interpretive searches from both these viewpoin given.

LIKELIHOOD COMMENTS

An examination of the likelihood function reveals that there point that is unambiguously most favored by the data. This i squares point, the maximum likelihood point. Reporting any constitutes a distortion of the evidence, an interpretation dep plicitly or explicitly on prior information. A likelihood advoc not to interpret evidence; he prefers only to summarize it. He general to the estimation framework which implicitly or explici a well-specified decision problem that requires us to summarize tainty about the parameters in a single point or estimate. Orc such decision problem is even envisaged. Selecting a tractable lc in such circumstances is as arbitrary as selecting a tractable pri tion. In the absence of any loss structure (even a vague one) the framework seems irrelevant. Of course, we tend to use the l

estimation but what we call "estimates" are usually data summaries, not decisions. That is to say, interest centers on the least-squares estimators not because they are best for some decision but rather because the estimates together with confidence ellipsoids provide a useful data summary.

For the simple, two-variable linear-regression problem $\mathbf{Y}=\mathbf{x}_1\beta_1+\mathbf{x}_2\beta_2+\mathbf{u}$ the most straightforward way of summarizing the data evidence is to draw two-dimensional likelihood contours as in Figure 5.4. There are two useful results on the geometrical relationship between a confidence ellipsoid and various confidence intervals. The first is that the projection of a suitably chosen ellipsoid is a confidence interval. In Figure 5.4, the interval $[b_1^-, b_1^+]$ is a 95% interval for β_1. The second result is that the length of a conditional confidence interval given some linear restriction can be found by intersecting a confidence ellipsoid with a suitably chosen linear manifold. In Figure 5.4, the length of a confidence interval for β_1 given a value of β_2 is w^*, found by drawing a line through the center of the ellipsoid perpendicular to the β_2 axis. These two results are stated first, and then further discussed.

THEOREM 5.3 (SUPPORTING HYPERPLANES). *The region between any pair of parallel supporting hyperplanes of the ellipsoid* $(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})=\sigma^2\chi_\alpha^2(1)$ *is a* $1-\alpha$ *percent confidence region for* $\boldsymbol{\beta}$.

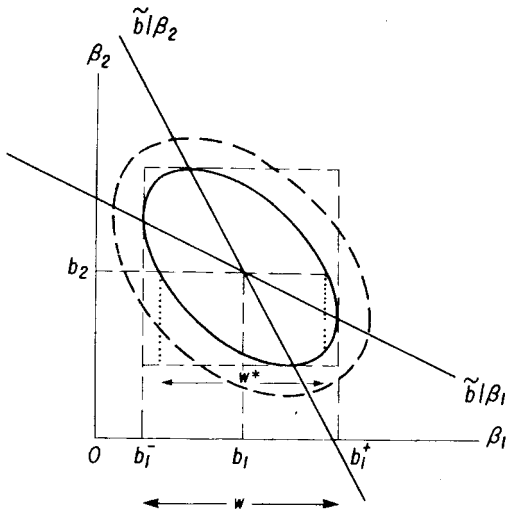

Fig. 5.4 Confidence ellipse.

Proof: A pair of supporting hyperplanes, $\psi'\beta=c_1$ and $\psi'\beta=c_2$ found by determining the extreme values of the function $\psi'\beta$ co to the ellipsoid. Solving this simple constrained optimization prob to the supporting hyperplanes

$$\psi'\beta=\psi'\mathbf{b}\pm\left[\psi'(\mathbf{X}'\mathbf{X})^{-1}\psi\sigma^2\chi_\alpha^2(1)\right]^{1/2}$$

which is a $1-\alpha$% interval for $\psi'\beta$.

THEOREM 5.4 (CONSTRAINED CONFIDENCE INTERVAL). *The fun evaluated on the intersection of the ellipsoid* $(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})$ *and the linear manifold* $\mathbf{R}(\boldsymbol{\beta}-\mathbf{b})=0$ *attains an interval of value length to the conditional* $1-\alpha$% *confidence interval for* $\psi'\beta$ *give*

Proof: This is also a constrained optimization problem left to t Incidentally, Theorem 5.3 is a direct consequence of Theorem 5.

These two results illustrate the value of exact prior knowledge combinations of parameters. In Figure 5.4 the outer ellipse confidence ellipse for the parameter vector (β_1,β_2). It, furthern tains 95% of the volume under the likelihood surface and is a 95% region if the prior distribution is diffuse. The interior ellipse is that its projection onto the β_1 axis, $[b_1^-, b_1^+]$ is a 95% interval f Theorem 5.3, any pair of parallel lines tangent to this ellipsoid 95% region.[9]

The two lines $\tilde{\mathbf{b}}|\beta_1$ and $\tilde{\mathbf{b}}|\beta_2$ are the locus of tangencies bet family of likelihood ellipses and horizontal and vertical lines, res These are the estimates given exact knowledge of one of the pa The width of the interval for β_1 given β_2 is, by Theorem 5.4 Letting the correlation between x_1 and x_2 be $r^2=(\mathbf{x}_1'\mathbf{x}_2)^2/(\mathbf{x}_1'\mathbf{x}_1)($ ratio of the two widths is $w^*/w=(1-r^2)^{1/2}$. This is illustrated 5.5 and the multidimensional analogue of $(1-r^2)^{1/2}$ is suggested i 5.7 as a measure of the collinearity problem.

Having made these statements, we may now return to the pr interpreting the data summarized in Figure 5.4. Assume tha centers on characterizing the sample evidence about β_1. For simp desirable to have a one-dimensional description of that evidence ably, the sample evidence does not allow an unambiguous one-di

[9] A region with minimal area is generated by lines parallel to the major axis of the is a geometrical interpretation of Silvey's (1969) result that a linear combination estimated relatively precisely, if ψ can be expressed as a linear combination of t tors of $\mathbf{X}'\mathbf{X}$ with the largest eigenvalues.

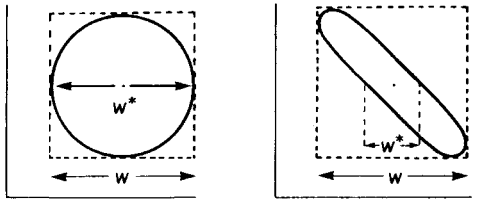

Fig. 5.5 Constrained and unconstrained confidence intervals.

data summary, since the evidence about β_1 depends on the prior information about β_2. For example, if prior information is relatively vague about both β_1 and β_2, the interval $b_1 \pm w$ is a useful summary of the evidence about β_1. If we know something about β_2, the effective sample information about β_1 may change drastically. In particular, if we know $\beta_2 = 0$, a useful summary of the evidence about β_1 is $(\tilde{b}_1 | \beta_2 = 0) \pm w^*/2$. This interval is both shorter and recentered. The discrepancy between the two intervals will increase with the correlation r (see Figure 5.5). Thus when r is large, the sample evidence cannot be meaningfully interpreted without referring to prior information; to put this somewhat differently, the box implied by the point estimates and their standard errors gives a most inaccurate picture of the region favored by the data, the confidence ellipse. A two-dimensional likelihood function cannot unambiguously be collapsed into two one-dimensional functions.

BAYESIAN COMMENTS

A Bayesian analysis (somewhat magically) implies a one-dimensional summary of the data evidence about β_1, since the marginal posterior distribution of β_1 can be written using Bayes' rule as

$$f(\beta_1|\mathbf{Y}) \propto \left[\int_{\beta_2} f(\mathbf{Y}|\beta_1,\beta_2) f(\beta_2|\beta_1) d\beta_2 \right] f(\beta_1) = f(\mathbf{Y}|\beta_1) f(\beta_1).$$

The last part of this expression appears to be a one-dimensional application of Bayes' rule with the integrated or marginal (one-dimensional) likelihood function $f(\mathbf{Y}|\beta_1)$ summarizing in the usual way the sample evidence about β_1. The first line indicates, however, that this marginal likelihood depends on the conditional prior $f(\beta_2|\beta_1)$, which is a formal Bayesian way of saying what is already clear: the sample evidence about β_1 depends on the prior information about β_2.

Although a complete Bayesian analysis of this problem requires a fully

specified prior distribution, some progress can be made with a crudely specified distribution. Pretesting on β_2 is meaningdul when β_1 is relatively uncertain and when β_2 is thought to be near zero. In this case the prio contours effectively parallel the β_1 axis with the most likely value bein $\beta_2 = 0$. A posterior distribution mixes this information with the sampl information according to Bayes' rule $f(\beta_1,\beta_2|\mathbf{Y}) \propto f(\mathbf{Y}|\beta_1,\beta_2) f(\beta_2)$. Th modes of the posterior must be on the locus of tangencies between th likelihood contours and the prior contours (see Figure 5.6). This locus o points is just a line $\tilde{\mathbf{b}}|\beta_2$. Independent of any further distributional assump tions, the most likely value of the couple (β_1,β_2) after we have seen th data is on the line $\tilde{\mathbf{b}}|\beta_2$ between the least-squares estimate $\mathbf{b}$ and th constrained estimate given $\beta_2 = 0$.

The position of the mode on this line, as well as the complete join posterior distribution, depends on a completely specified prior distributio for β_2. The reader may convince himself that there are four posterio distributions that imply the marginal posteriors (and in this case th marginal likelihoods) in Figure 5.7. Cases (a) and (b) occur when there i overwhelming sample or overwhelming prior information. In case (c) th two sources of information are roughly comparable in content and w react by mixing them into a unimodal distribution. This would occur wit the conjugate normal prior that is widely discussed. If prior informatio about β_2 is steeper around zero, we may get the antimixture case (d). Thi is not an atypical situation with (my?) meaningful prior distributions. Not incidentally that cases (a) and (b) are, in fact, special cases of (c) and (d

In light of the preceding, a Bayesian might make the following com ments:

a. The theory of pretesting is misleading. It suggests that one ma implement his prior information without carefully specifying it. I usually leads to pretesting at an arbitrary level of significance, say, .0 This may or may not capture the essential features of your prior.
b. Pretesting works only in the extreme cases when b_1 or $\tilde{b}_1|\beta_2 = 0$ ar appropriate summaries. The mixture and antimixture cases in Figur 5.7 are excluded. (Incidentally, the exclusion of these cases means tha small adjustments in the data evidence may imply jumping from the summary to the $\tilde{\mathbf{b}}$ summary. This discontinuity ordinarily implies tha the estimators are inadmissible. This situation can, of course, b improved by continuous mixing, but the analysis gets quite complex.
c. Pretesting does not clearly distinguish sample from nonsample infor mation. Ordinarily, our belief in the output of a study should depen on the judgmental inputs. When these inputs are disguised, we have n way of evaluating an empirical study.

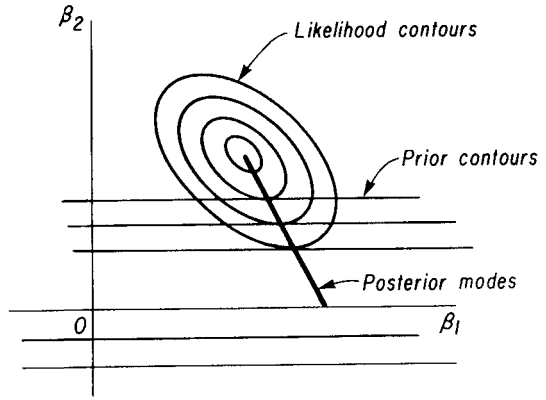

Fig. 5.6 Locus of posterior modes.

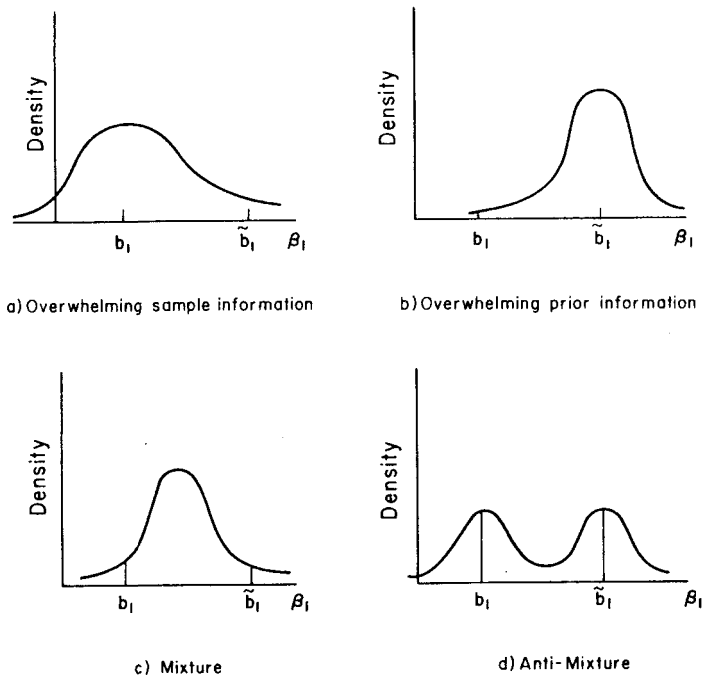

Fig. 5.7 Posterior distributions.

d. The pretesting theory just discussed is inappropriate when there is prior information about the other parameters. It is very difficult to make pretesting meaningful in higher dimensional problems, since perusal of multidimensional risk functions (expected losses) is terribly difficult. More generally, the ambiguities inherent in implementing prior opinion are more easily discussed in terms of alternative prior distributions than in terms of alternative risk functions (or sampling distributions).

In summary, the formal estimation framework does not focus on theoretical questions posed by practical pretesting. Practical pretesting is not designed or should not be designed to improve estimators independently of prior information, as suggested by this literature. In the first place, the estimation framework misstates the problem unless there is a real point decision to be made (a most rare event). Usually, an interpretive search is designed to characterize the information contained in the data set, and the estimation framework suggests the erroneous conclusion that the sample evidence somehow depends on what decisions are to be made. Second, even if you did have such a decision problem, in the context of quadratic loss, a Bayes estimate of one coefficient is just its posterior mean, which does not depend on the weight given another coefficient in the loss function. Third, the resultant estimators rarely dominate least squares and, furthermore, have generally unknown properties (though sampling properties are rarely relevant for interpreting the data). In the fourth place, in the context of the more usual problem of inference, the least-squares point, James and Stein notwithstanding, is the unique point most favored by the data. In the absence of prior information, it is the only point that has an unambiguous claim to be reported.

To put this slightly differently, data analysis can usefully be regarded to include three distinct phases. First, the data evidence is *summarized.* Second, it is *interpreted.* Finally, *decisions* may be made. The first phase requires a theory of sufficient statistics. The second phase requires a theory of learning. And the third phase requires a theory of decision making under uncertainty. The essential differences between Bayesian and classical inference arise in the second or interpretative phase. The Bayesian model of learning is described by Bayes' rule: let data evidence incrementally affect your opinions. The classical theory of learning, however, is either incomplete or is trivially described by: completely believe the data evidence.

The least-squares point and the variance-covariance matrix are the appropriate summary of the data evidence. Practical pretesting occurs at the interpretive stage and constitutes a de facto rejection of the trivial theory

of learning implicit in classical inference and/or an ad hoc completion of that incomplete theory. A loss function occurs only in the third (decision-making) phase, and inadmissibility results are relevant to the learning phase only to the extent that we would prefer not to have a model of learning that, when coupled with a *particular* model of decision making under uncertainty leads to decisions that could be improved on, on the average. Inadmissibility results do not, however, have implications for the summarization phase and, for example, the Stein–James result has nothing to do with the adequacy of the least-squares point as a summary of the data evidence. The result *should* be disconcerting to those Bayesians who are fond of improper prior distributions, however, since it suggests that there is a serious inadequacy in their learning model.

5.6 Regression Selection Strategies and Revealed Priors

By definition, an interpretive search is an ad hoc covert method of introducing uncertain prior information. It is hardly remarkable that overt formal Bayesian methods could cast light on these murky goings-on. This section develops a bridge between searchers and Bayesians by answering the question: what kind of prior information does an interpretive search seem to be based on? This discussion is not intended merely to apologize for interpretive searches. It is possible to find priors that justify many procedures. In so doing we hope also to encourage a more careful review of the available prior information, a more overt use of that which is available, and finally, a clearer communication of the prior on which the data analysis rests.

As discussed in Section 5.1, an interpretive search strategy involves three decisions: (a) the choice of an origin, which determines the ellipsoid of constrained estimates, (b) the choice of a coordinate system, which selects 2^k or fewer points on the feasible ellipsoid, and (c) the choice of a weighting function that mixes the 2^k points into a single estimate. The features of the prior distribution that implicitly determine the first two of these choices are the surfaces upon which the prior density is constant. The third choice—the weighting function—is implied by the function that assigns to each prior isodensity surface a particular density value.

In this section, we first explore the relationship between prior isodensities and choice of constraints, and then, for one special case, analyze the relationship between the weighting function and the density labeling function. A correspondence is developed between priors that are uniform on ellipsoids and regression selection strategies that either (a) compute 2^k regressions or (b) compute principal component regressions. Similarly,

priors that are uniform on hyperbolas are associated with strategies that omit insignificant variables. Lexicographic priors and priors that are uniform on cones are also discussed, the latter intended to capture the notion that parameters are equal.

5.6.1 *Choice of Constraints*

The language and method of thinking about this problem is borrowed from the Edgeworth–Bowley analysis of trade between a pair of consumers. In that problem, a fixed quantity of a set of k commodities is to be distributed between two consumers. Let $\mathbf{q}$ be a k-dimensional vector indicating the available quantities of the k commodities; let $\boldsymbol{\beta}$ be a k-dimensional vector indicating the allocation of commodities to the first consumer, who thereby attains utility level $U_1(\boldsymbol{\beta})$. The remainder, $\mathbf{q}-\boldsymbol{\beta}$, is allocated to the second consumer, who thereby attains utility level $U_2(\mathbf{q}-\boldsymbol{\beta})$. An allocation $\boldsymbol{\beta}_a$ is said to dominate another allocation $\boldsymbol{\beta}_b$, if $U_1(\boldsymbol{\beta}_a) \geqslant U_1(\boldsymbol{\beta}_b)$ and $U_2(\mathbf{q}-\boldsymbol{\beta}_a) \geqslant U_2(\mathbf{q}-\boldsymbol{\beta}_b)$ with at least one strict inequality. In words, one person is better off at $\boldsymbol{\beta}_a$ and no one is worse off. The undominated set of allocations is called the Pareto efficient set. Under differentiability and convexity assumptions, the Pareto efficient set is a curve formed by maximizing one of the consumer's utility levels subject to a given utility level of the other. This curve is called a *contract curve*, since given an initial allocation $\boldsymbol{\beta}$ off the contract curve it is likely that the consumers would trade to a suitable point on it, thereby making at least one better off and neither worse off. It is enough that utility be ordinal to define the contract curve. Picking an optimal point on it requires cardinal utility functions and a social welfare function $W(U_1, U_2)$ to be maximized.

Consider now the analogous problem of Bayesian inference with a k dimensional parameter $\boldsymbol{\beta}$. The data communicates its information through a likelihood function, say, $U_1(\boldsymbol{\beta})$, and the researcher communicates his information through a prior density, say, $U_2(\boldsymbol{\beta})$. Posterior modes are found by maximizing the "social information function", $W(U_1, U_2) = U_1 U_2$. If the prior density is ordinal, that is, if it is defined only up to the surfaces on which it is constant, all that can be said is that the posterior mode is confined to a curve, which is called the *information contract curve*. In developing the correspondence between regression selection strategies and priors, we hypothesize that a researcher is attempting to approximate the information contract curve with a set of constrained regression estimates, where the word approximates means to find a set of points that contains the curve.

DEFINITIONS: ISODENSITY SURFACES AND LABELING FUNCTIONS

Let $h(\mathbf{x})=z$ be a convenient representation of a family of surfaces indexed by z, and let a family of density functions be written

$$f(\mathbf{x})=g(h(\mathbf{x})),$$

where h is given and g is any monotonically decreasing differentiable function with the restriction that $\int g(h(\mathbf{x}))d\mathbf{x}=1$. The surface $h(\mathbf{x})=z$ is called an *isodensity* surface, and the function g that assigns a density value to each of these surfaces is called a *labeling function*.

Example. A multivariate normal distribution

$$f(\mathbf{x})=(2\pi)^{-k/2}|\mathbf{H}|^{1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{H}(\mathbf{x}-\bar{\mathbf{x}})\right\}$$

is uniform on the ellipsoids

$$h(\mathbf{x})=(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{H}(\mathbf{x}-\bar{\mathbf{x}})$$

with labeling function

$$g(z)\propto\exp-\left(\frac{z}{2}\right).$$

A multivariate Student distribution has the same elliptical isodensity surfaces but has the labeling function

$$g(z|\nu)\propto(\nu+z)^{-(k+\nu)/2}.$$

Johnson and Kotz [1972, p. 296] refer to densities that are uniform on ellipsoids as elliptically symmetric distributions. We call them elliptically uniform distributions to emphasize the fact that the ellipsoids are isodensity surfaces.

Prior-to-posterior analysis of a k-dimensional parameter $\boldsymbol{\beta}$ depends on the data $\mathbf{Y}$ through Bayes' rule

$$f_2(\boldsymbol{\beta}|\mathbf{Y})\propto f_y(\mathbf{Y}|\boldsymbol{\beta})f_1(\boldsymbol{\beta})$$

where the subscripts 1, 2, and y indicate prior, posterior, and sample, respectively. This may be rewritten in terms of isodensity and labeling functions as

$$f_2(\boldsymbol{\beta}|\mathbf{Y})\propto g_y(h_y(\boldsymbol{\beta}))g_1(h_1(\boldsymbol{\beta})).$$

Modes of the posterior distribution require the derivatives of this function to be set to zero

$$\mathbf{0}=g_y g_1'\frac{\partial h_1}{\partial\boldsymbol{\beta}}+g_y' g_1\frac{\partial h_y}{\partial\boldsymbol{\beta}}$$

or

$$\mathbf{0}=\lambda\frac{\partial h_1}{\partial\boldsymbol{\beta}}+\frac{\partial h_y}{\partial\boldsymbol{\beta}} \tag{5.12}$$

where

$$\lambda(\boldsymbol{\beta})=\frac{g_1'/g_1}{g_y'/g_y} \tag{5.13}$$

and where λ is assumed to be positive, since $g_y'<0$, $g_1'<0$.

Notice now that as a function of λ, Equation (5.12) defines a k-dimensional space curve that depends on the isodensity functions h_1 and h_y only. In particular, the labeling distributions are irrelevant. To the extent that joint modes of the posterior are of interest, it is, therefore, possible to perform a Bayesian analysis in two steps: (1) select isodensity functions h_1 and h_y and compute the *information contract curve* given by Equation (5.12); (2) specify the labeling functions g_1 and g_y that can be used to compute the modes of the posterior. (All modes necessarily lie on the information contract curve.)

Interest in this two-step approach derives from the following assertion. It is impossible to measure degrees of belief about continuous random variables with enough accuracy that we would be content with a Bayesian analysis based on a single distribution (prior or data). Instead, ambiguities in the choice of distribution are properly dealt with by performing the analysis with many different distributions. A class of distributions that is of interest is the class with fixed isodensity surfaces and with varying labeling functions. This class is an appropriate focus of attention when the choice of isodensity curves is relatively unambiguous, but the choice of labeling functions is relatively ambiguous. In the case of complete ambiguity over choice of labeling function, the Bayesian analysis can at best specify the information contract curve. The value of such a limited statement is not to be underassessed, however. The restriction of modes to a k-dimensional curve is a very significant restriction.

Incidentally, the case of complete ambiguity over choice of labeling distribution represents the continuous analogue of Keynes' (1921) suggestion that probabilities are sometimes only ordinally ordered. In that case values of the random variable may be said to be more likely, less likely, or equally likely to other values, but it is impossible to say how much more likely. More formally, choice of isodensity function with a monotonicity assumption on the labeling function implies an ordinal ranking of points in the outcome space of the random variable. The cardinality of this ordering is determined by the labeling function.

THE DATA INDIFFERENCE SURFACES

The isolikelihood surfaces implied by the normal linear regression model are a family of concentric ellipsoids

$$z=(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b}),$$

where $\mathbf{b}$ is the least-squares estimate and $\mathbf{X}'\mathbf{X}$ is the design matrix. The data set is indifferent among all values of $\boldsymbol{\beta}$ lying on such an ellipsoid in the sense that each is assigned the same likelihood value. If the residual variance σ^2 were known, the data would assign to each ellipsoidal surface a particular likelihood value. If σ^2 is unknown, relative likelihoods (but not absolute likelihoods) can be computed. For the purposes of this section, only the indifference surfaces are needed, and what is discussed applies to any sampling process that determines elliptical isolikelihood surfaces.

ELLIPTICALLY UNIFORM PRIORS AND PRINCIPAL COMPONENT REGRESSION

The information of the researcher can be packaged in a prior density function. A prior density that is uniform on concentric ellipsoids can be written as[10]

$$f_1(\boldsymbol{\beta})=cg_1[(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)],$$

thereby indicating indifference among all points on an ellipsoid $(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)$.

The information contract curve formed by minimizing the likelihood quadratic form subject to the constraint that the prior quadratic form is a constant is

$$\mathbf{0}=\lambda\mathbf{N}^*(\boldsymbol{\beta}-\mathbf{b}^*)+\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})$$

where λ is a Lagrange multiplier. Solving this for $\boldsymbol{\beta}$, we obtain

$$\boldsymbol{\beta}(\lambda)=(\lambda\mathbf{N}^*+\mathbf{X}'\mathbf{X})^{-1}(\lambda\mathbf{N}^*\mathbf{b}^*+\mathbf{X}'\mathbf{X}\mathbf{b}), \tag{5.14}$$

which should remind you of the curve décolletage defined in Section 3.3. This equation defines the locus of tangencies between the prior ellipsoidal surfaces and the likelihood ellipsoidal surfaces. In two dimensions the curve is a hyperbola (see Figure 5.8).[11]

Points off the information contract curve are inefficient in the sense that

[10]It is interesting to observe that if $\mathbf{N}^*=\mathbf{I}$, and if the coefficients are independent, then f_1 is necessarily a multivariate normal distribution.

[11]The relevant part of the curve consists of the segment between $\mathbf{b}$ and $\mathbf{b}^*$, since any other points on the curve are dominated by points on this segment. It can also be shown that the points on the hyperbola but not on the line segment between $\mathbf{b}$ and $\mathbf{b}^*$ involve negative values of the Lagrange multiplier λ. These are ruled out by the monotonicity assumption on the labeling functions.

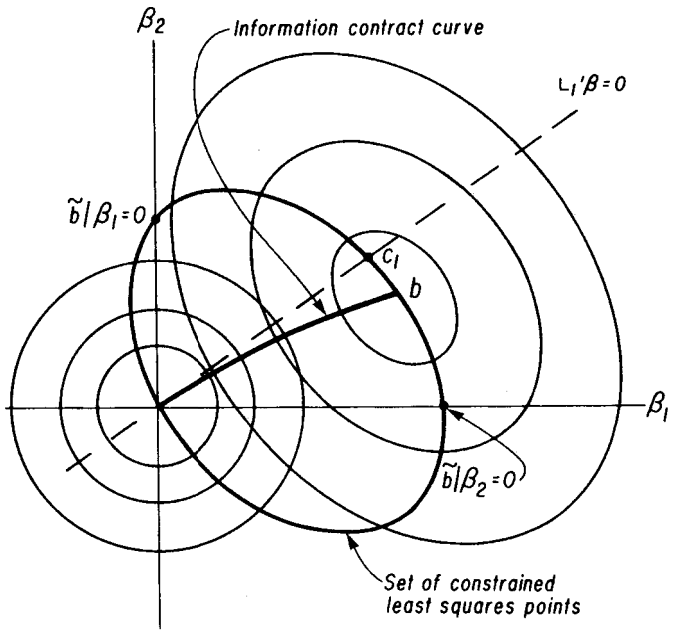

Fig. 5.8 The information contract curve with a spherical prior; $\mathbf{c}_1$= principal component regression.

both the researcher and the data may be made "happier" by moving to point on the contract curve. A researcher who imposes constraints on tl estimates is trying to make himself happier, but at least if he tests tl constraints, he is also keeping in mind the data preferences. He is thi apparently trying to describe the information contract curve, and tl question to be addressed now is how accurately a set of constraine estimates approximates the information contract curve.

The first result to be discussed is that the contract curve (5.14) can l written as a weighted average of 2^k constrained least-squares points. Tl contract curve is also shown to be a weighted average of $k+1$ princip component points. Without loss of generality, the prior precision matrix assumed to be diagonal $\mathbf{N}^*=\text{diag}\{d_1,d_2,\ldots,d_k\}$, and the prior location taken as the origin $\mathbf{b}^*=\mathbf{0}$.

THEOREM 5.5 (DIAGONAL PRIOR PRECISION). *A matrix-weighte average can be written as a weighted average of 2^k constrained leas squares points*:

$$\mathbf{b}^{**}=(\mathbf{H}+\mathbf{D}^*)^{-1}\mathbf{H}\mathbf{b}=\sum_I w_I\mathbf{b}_I, \tag{5.1}$$

where $\mathbf{D}^*$ is a diagonal matrix, $\mathbf{D}^* = \text{diag}\{d_1, d_2, \ldots, d_k\}$, $\mathbf{H} = \sigma^{-2}\mathbf{X}'\mathbf{X}$, I is a subset of the first k integers, $\mathbf{b}_I$ minimizes the quadratic form $(\boldsymbol{\beta} - \mathbf{b})'\mathbf{H}(\boldsymbol{\beta} - \mathbf{b})$ subject to the constraints $\beta_i = 0$ for $i \in I$. The weights are

$$w_I = \frac{\left(\prod_{i \in I} d_i\right)|\mathbf{H}_I|}{|\mathbf{H} + \mathbf{D}|} \geqslant 0$$

$$\sum w_I = 1,$$

where $\mathbf{H}_I$ is the square matrix formed by deleting all rows i and columns i of $\mathbf{H}$ for all $i \in I$. By definition, if I is the complete set of k integers, $|H_I| \equiv 1$, and if I is the null set, $\prod_{i \in I} d_i \equiv 1$.

Proof: See Appendix to Chapter 5.

The implication of this theorem is that a researcher who has elliptically uniform priors with major axes in the directions of the coordinate axes can find the points on his information contract curve by computing the 2^k regressions formed by omitting variables in all different combinations. A stronger result is that the contract curve can be written as a weighted average of $k+1$ principal component points. Before that result is presented, principal component regression is explained.

The intuitive foundation of principal component regression is the assertion that if the explanatory variables were orthogonal, that is, if $\mathbf{X}'\mathbf{X}$ were a diagonal matrix, then a "natural" interpretive search strategy would be to omit the variables with the smallest variance first. Some people have objected that, in fact, it is more "natural" to test the variables to see if they belong in the equation, that is, to omit the variables with the smallest t-values. The point that is being made in this section is that the prior distribution determines what is "natural," and arguments over features of search strategies can only be resolved by explicit reference to features of prior distributions. As it turns out, spherical priors make it "natural" to omit orthogonal variables in the order of their variance. There does not seem to be a prior that would lead one to test the restrictions implicit in principal component regression.

There is, of course, a transformation—in fact many—that make the explanatory variables orthogonal. Write the regression process as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}\mathbf{L}\mathbf{L}^{-1}\boldsymbol{\beta} + \mathbf{u} = \mathbf{Z}\boldsymbol{\theta} + \mathbf{u} \tag{5.16}$$

where $\mathbf{Z} = \mathbf{X}\mathbf{L}$, $\boldsymbol{\theta} = \mathbf{L}^{-1}\boldsymbol{\beta}$, and where $\mathbf{L}$ is a $(k \times k)$ matrix that diagonalizes $\mathbf{X}'\mathbf{X}$: $\mathbf{Z}'\mathbf{Z} = \mathbf{L}'\mathbf{X}'\mathbf{X}\mathbf{L} = \boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$. The transformed model $\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta} + \mathbf{u}$ thus has orthogonal explanatory variables, and principal component regression would omit these variables (impose the constraints $\theta_i = 0$ as ordered by the variances λ_i.

Among the many matrices that take $\mathbf{X}'\mathbf{X}$ into a diagonal matrix, it is necessary to select one. By Theorem 35 in Appendix 1, there is a (unique) matrix $\mathbf{L}$ such that $\mathbf{L}'\mathbf{X}'\mathbf{X}\mathbf{L}$ is a diagonal matrix and $\mathbf{L}'\mathbf{N}^*\mathbf{L}$ is an identity matrix, where $\mathbf{N}^*$ is an arbitrary $(k \times k)$ symmetric positive definite matrix. Thus choice of $\mathbf{N}^*$ with the restriction on $\mathbf{L}$, $\mathbf{L}'\mathbf{N}^*\mathbf{L} = \mathbf{I}_k$, determines $\mathbf{L}$ uniquely. Although it is common to choose $\mathbf{N}^* = \mathbf{I}_k$, perhaps with the original explanatory variables standardized to have equal variance, this decision is a critical step in the interpretive search strategy and should not be taken lightly. As is shown, it amounts to choosing a prior with elliptical isodensity surfaces $\boldsymbol{\beta}'\mathbf{N}^*\boldsymbol{\beta} = z$.

When $\mathbf{N}^* = \mathbf{I}_k$, the matrix $\mathbf{L}$ is a matrix of eigenvectors of $\mathbf{X}'\mathbf{X}$. The vector of parameters $\boldsymbol{\theta}$ is $\mathbf{L}^{-1}\boldsymbol{\beta} = \mathbf{L}'\boldsymbol{\beta}$, and the constraint $\theta_i = 0$ is equivalent to the constraint $\mathbf{L}_i'\boldsymbol{\beta} = 0$, where $\mathbf{L}_i$ is the ith eigenvector of $\mathbf{X}'\mathbf{X}$. The diagonal elements of $\mathbf{L}'\mathbf{X}'\mathbf{X}\mathbf{L}$ are the eigenvalues of $\mathbf{X}'\mathbf{X}$. Thus principal component regression sequentially imposes the constraints that the vector $\boldsymbol{\beta}$ is orthogonal to the eigenvectors of $\mathbf{X}'\mathbf{X}$, with the constraints ordered from smallest to largest eigenvalue.[12]

In Appendix 1, it is shown that the eigenvectors of $\mathbf{X}'\mathbf{X}$ are the principal axes of the ellipsoid $\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = r^2$, and the eigenvalues are ordered the reverse of the ordering of the lengths of the axes. The first restriction is that $\boldsymbol{\beta}$ is orthogonal to the largest axis of the ellipse. For a two-dimensional problem, this restriction is illustrated in Figure 5.8. Notice that $\mathbf{c}_1 - \mathbf{b}$ and $\mathbf{c}_1$ are orthogonal by construction.

The following theorem asserts that the convex hull of the three points $\mathbf{b}$, $\mathbf{c}_1$, and $\mathbf{0}$ contains the contract curve.

THEOREM 5.6 (SPHERICAL PRIOR). *The contract curve* (5.14) *with* $\mathbf{N}^* = \mathbf{I}$ *and* $\mathbf{b}^* = \mathbf{0}$ *can be written as*

$$\boldsymbol{\beta}(\lambda) = (\lambda\mathbf{I} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = \sum_{j=0}^{k} w_j(\mathbf{X}, \lambda)\mathbf{c}_j$$

where $\mathbf{c}_j$ *is the jth principal component point formed by "dropping" from the equation the j principal components of* $\mathbf{X}'\mathbf{X}$ *with the smallest roots.*

[12]Another way to describe this is that the linear combination of variables $\mathbf{X}\mathbf{L}_1$ is selected to have smallest variance $\mathbf{L}_1'\mathbf{X}'\mathbf{X}\mathbf{L}_1$, given the normalization $\mathbf{L}_1'\mathbf{L}_1 = 1$. The next linear combination, $\mathbf{X}\mathbf{L}_2$, is restricted to be orthogonal to $\mathbf{X}\mathbf{L}_1$, $0 = \mathbf{L}_2'\mathbf{X}'\mathbf{X}\mathbf{L}_1$, but is otherwise selected to minimize its variance.

Mathematically, the vectors $\mathbf{c}_j$ *are defined using the eigenvector coordinates*

$$\mathbf{L}'\mathbf{X}'\mathbf{X}\mathbf{L}=\mathbf{\Lambda}=\operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_k)$$
$$\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_k$$
$$\mathbf{L}\mathbf{L}'=\mathbf{I}$$
$$\mathbf{g}=\mathbf{L}'\mathbf{b}$$

with $\mathbf{h}_j'=(0,0,\ldots,0,g_{j+1},g_{j+2},\ldots,g_k)$

$$\mathbf{c}_j=\mathbf{L}\mathbf{h}_j.$$

The weights applying to the principal component points are

$$w_0=\frac{\lambda_1}{(\lambda_1+\lambda)}$$
$$w_j=\frac{\lambda(\lambda_{j+1}-\lambda_j)}{(\lambda_{j+1}+\lambda)(\lambda_j+\lambda)} \qquad 0<j<k$$
$$w_k=\frac{\lambda}{(\lambda_k+\lambda)}$$

The proof follows directly by writing $\boldsymbol{\beta}(\lambda)$ *in principal component coordinates,* $\boldsymbol{\beta}(\lambda)=(\lambda\mathbf{I}+\mathbf{L}\mathbf{L}'\mathbf{X}'\mathbf{X}\mathbf{L}\mathbf{L}')^{-1}\mathbf{X}'\mathbf{X}\mathbf{L}\mathbf{L}'\mathbf{b}=\mathbf{L}(\lambda\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}g$.

We can then write $(\lambda\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}\mathbf{g}=\sum_{j=0}^{k}w_jh_j$ where w_j is the solution to the triangular system $w_0=\lambda_1/(\lambda_1+\lambda), w_0+w_1=\lambda_2/(\lambda_2+\lambda),\ldots,\sum_{i=0}^{k-1}w_i=\lambda_k/(\lambda_k+\lambda),\sum_{i=0}^{k}w_i=1$.

These last two results determine a subset of constrained estimates that a Bayesian with an elliptical prior would be interested in. The principal component result is perhaps more useful, since it involves only $k+1$ regressions instead of 2^k and since by a linear transformation any ellipsoid can be taken into a sphere. The principal component result may be used to resolve several questions that trouble users of principal component regression. The first concerns the arbitrary order in which the principal component restrictions are imposed, and the second is the arbitrary normalization rule $\mathbf{l}'\mathbf{l}=1$. Some writers have suggested that one ought to "test" the restrictions, that is, to order the restrictions by their t values. Theorem 5.6 does not apply if the restrictions are so ordered. They must be ordered by their respective eigenvalues (variances). The arbitrariness in normalization is also resolved, since the researcher is required to use a coordinate system in which his prior is spherical. Equivalently, if his prior is uniform on the ellipsoids $\boldsymbol{\beta}'\mathbf{N}^*\boldsymbol{\beta}$, he should use the normalization $\mathbf{L}'\mathbf{N}^*\mathbf{L}=1$. The reverse

result is also true: if a Bayesian computes principal component points with respect to some normalization, he reveals that he has the prior implicit in that normalization in the sense that otherwise points on the contract curve might not be weighted averages of the points he computes (Leamer, 1977).

Another element of arbitrariness that is not widely recognized arises when the principal component analysis is applied to only a subset of the variables. In that case it is not clear at the outset whether one should attempt to minimize the marginal variance of the components, or the variance conditional on unaffected variables. By the logic in this section, it should be the conditional variance (Leamer, 1977).

HYPERBOLICALLY UNIFORM PRIORS

Although the principal component method of estimation is used occasionally, it is much more common to express restrictions in a predetermined coordinate system, that is, to drop particular variables. In Figure 5.8 the third point that would most often be reported is not $\mathbf{c}_1$ but $\tilde{\mathbf{b}}|\beta_1=0$ or $\tilde{\mathbf{b}}|\beta_2=0$. This is clearly undesirable with elliptically uniform priors, since the information contract curve may be very poorly represented by such a sequence of points. For example, in Figure 5.8 if the variable with the lower t value is dropped first, the convex hull of the three estimates $(\mathbf{b},[\tilde{\mathbf{b}}|\beta_2=0],\mathbf{0})$ does not contain the contract curve. (The t value of the first coefficient exceeds the t value of the second because $\tilde{\mathbf{b}}|\beta_2=0$ is closer in the data metric to $\mathbf{b}$ than is $\tilde{\mathbf{b}}|\beta_1=0$. In other words, the data are "happier" with $\tilde{\mathbf{b}}|\beta_2=0$ than $\tilde{\mathbf{b}}|\beta_1=0$.)

The common procedure of dropping variables with low t values may therefore, be highly undesirable with elliptically uniform priors. On the other hand, there may be perfectly reasonable priors that do lead to this kind of processing of the data. A completely trivial case in point illustrated in Figure 5.9 occurs when the prior isodensity surfaces are $z=\min(|\beta_1|,|\beta_2|)$. The solid line linking $\mathbf{b}$ to $\tilde{\mathbf{b}}|\beta_2=0$ contains all global modes, although local modes may lie on the dotted line linking $\mathbf{b}$ to $\tilde{\mathbf{b}}|\beta_1=0$. Thus a data analysis that reported $\mathbf{b}$ and $\tilde{\mathbf{b}}|\beta_2=0$ would convey part of the essential features of the data, since global modes are necessarily convex combinations of these two points. Local modes on the segment connecting $\mathbf{b}$ to $\tilde{\mathbf{b}}|\beta_1=0$ are also natural candidates to be reported.

These right-angled prior indifference curves represent a peculiar kind of ordering that seems difficult to approximate with any continuous probability functions. They, furthermore, reflect a peculiar willingness to ignore all but one of the constraints at any but special points in the parameter space. What seems to be a more reasonable family of curves are the hyperbolas

$$z=\prod_i|\beta_i|.$$

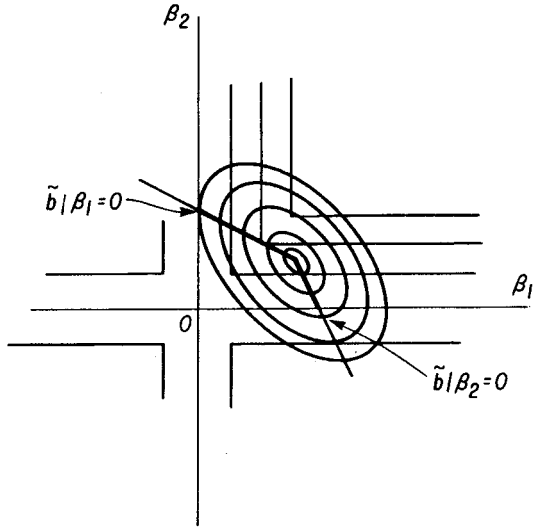

Fig. 5.9 The contract curve with rectangularly uniform priors.

As an example of such a density, products of independent Student functions

$$f(\boldsymbol{\beta}) \propto \prod_i (\nu + \beta_i^2)^{-(\nu+1)/2}$$

are hyperbolically uniform in the degenerate case $\nu = 0$

$$f(\boldsymbol{\beta}) \propto \prod_i |\beta_i|^{-1}.$$

The information contract curve Equation (5.14) with hyperbolically uniform priors becomes

$$\lambda\{\beta_i^{-1}\} + \mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) = \mathbf{0}, \tag{5.17}$$

where $\{\beta_i^{-1}\}$ indicates a vector with elements β_i^{-1}. Solving out the Lagrange multiplier λ yields the system of quadratics

$$\mathbf{n}_1'(\boldsymbol{\beta} - \mathbf{b})\beta_1 = \mathbf{n}_i'(\boldsymbol{\beta} - \mathbf{b})\beta_i, \qquad i > 1$$

where $\mathbf{n}_i'$ is the ith row of $\mathbf{X}'\mathbf{X}$. See Figure 5.10.

This system of equations can also be written as

$$\boldsymbol{\beta} = (\lambda\mathbf{D} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} \tag{5.18}$$

where $\mathbf{D} = \text{diag}\{\beta_1^{-2}, \beta_2^{-2}, \ldots, \beta_k^{-2}\}$. This equation appears to be a matrix-weighted average with the prior ellipsoids $\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}$ located at the origin, with

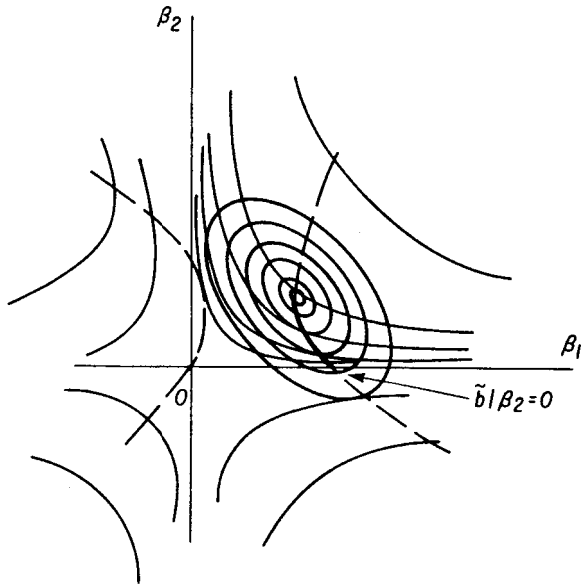

Fig. 5.10 The information contract curve with hyperbolically uniform priors.

principal axes equal to the coordinate axes. The relative lengths of these axes are, however, functions of $\boldsymbol{\beta}$. It is thus useful to think of hyperbolically uniform priors as if they were elliptically uniform priors located at the origin with principal axes equal to the coordinate axes but with the lengths of the principal axes "uncertain." In fact, the Student priors mentioned previously can be written as (elliptically uniform) normals with uncertain variance hyperparameters.

For any arbitrary $\mathbf{D}$ in Equation (5.18), Theorem 5.5 implies that $\boldsymbol{\beta}$ is contained in the hull of the 2^k constrained least-squares points formed by dropping variables. Since $\mathbf{D}$ is not free, it may be possible to further reduce the set of points. Paralleling the previous section, we would like to find a sequence of $k+1$ constrained estimates that contain the relevant curve segment. Although the prior determines the coordinate system, the sequence for imposing constraints is very ambiguous. Many different orderings seem possible. For example, (1) we might delete variables as ordered by the t coefficients in the original equation, or (2) we might recompute t values as constraints are imposed. (3) Proceeding in the other direction, variables may be added to the equation as ordered by their simple correlations with the dependent variable, or (4) as ordered by their partial correlations holding fixed the variables that are already included.

(5) Yet another alternative is to find the set of j variables that maximizes the multiple correlation coefficient, for $j=1,2,\ldots,k$.

It is obvious that the hyperbolic prior cannot induce a change in sign of any coefficient. The sequence of dropping variables as ordered by the sequentially computed t statistics necessarily preserves the orthant of the estimate (Leamer, 1975):

THEOREM 5.7 (ORTHANT PRESERVATION). *The least-squares estimate of β_j subject to a single linear constraint must lie in the interval $(b_j - V_{jj}^{1/2}|t|, b_j + V_{jj}^{1/2}|t|)$, where b_j is the unconstrained least-squares estimate of β_j, $V_{jj}^{1/2}$ is the jth diagonal element of $\mathbf{V}=s^2(\mathbf{X}'\mathbf{X})^{-1}$, and t is the t statistic for testing the restriction.*

Proof: The least-squares estimate subject to the constraint $\mathbf{R}\boldsymbol{\beta}=r$ (where $\mathbf{R}$ is a row vector) is $\hat{\boldsymbol{\beta}}=\mathbf{b}-\mathbf{V}\mathbf{R}'(\mathbf{R}\mathbf{V}\mathbf{R}')^{-1}(\mathbf{R}\mathbf{b}-\mathbf{r})$, with $\mathbf{V}=s^2(\mathbf{X}'\mathbf{X})^{-1}$. The t statistic for testing $\mathbf{R}\boldsymbol{\beta}=r$ is $t=(\mathbf{R}\mathbf{b}-r)/(\mathbf{R}\mathbf{V}\mathbf{R}')^{1/2}$. Thus we may write the constrained least-squares estimate of β_j as $\hat{\beta}_j=b_j-[\mathbf{V}_j'\mathbf{R}'(\mathbf{R}\mathbf{V}\mathbf{R}')^{-1/2}V_{jj}^{-1/2}]V_{jj}^{1/2}t$, where $\mathbf{V}_j$ is the jth column of $\mathbf{V}$. The term in square brackets is just the correlation between b_j and $\mathbf{R}\mathbf{b}$, which must be between -1 and $+1$. These two extreme values imply the bound in the statement of the theorem.

A consequence of this theorem is that there can be no change in sign of any coefficient that is more significant than the coefficient of an omitted variable. In particular, if the least significant variable is omitted, all the other coefficients will retain their signs. Thus the sequence of omitting variables in the order of their sequentially computed t statistics necessarily preserves the orthant of the estimate.

This theorem increases my probability of the truthfulness of the conjecture that this sequence of estimates contains the contract curve (5.17), but I have been unable to construct a proof of the proposition. The proposition is true in two dimensions, but in two dimensions all five sequences described in the foregoing paragraph imply estimates that contain the curve. I do have a tedious proof that the curve is contained in the hull of the unconstrained point and the k constrained points formed by omitting a single variable.

LEXICOGRAPHIC ORDERING

Spherically uniform priors have been seen to imply constraints in the coordinate system of the sample. The sequence of imposing these constraints depends on the eigenvalues of $\mathbf{X}'\mathbf{X}$ but not at all on the data $\mathbf{Y}$ or as a result on any test statistics. Hyperbolically uniform priors on the other

hand, imply a predetermined coordinate system but with a sequence o imposing constraints that is data dependent. Occasionally, both the coordi nate system and the sequence of imposing constraints are predetermine By this I mean that a researcher decides before looking at either $\mathbf{Y}$ or first to omit variable one, then variable two, and so on. As an example, i distributed lag analysis a general model involving many lagged explana tory variables is often simplified by omitting sequentially the variable wit the longest lag.

In consumer theory, a consumer who first satisfies his desires for goo A, then for good B, and so on, is said to have a lexicographic utilit ordering. Similarly, a researcher who proceeds this way has a lexicographi information ordering. A probabilistic structure that can effect such a ordering allocates positive probability to a nested sequence of subspace with a one-dimensional informative prior in each subspace.

CONICALLY UNIFORM PRIORS

A fairly common form of prior information about sets of parameters i expressed in the pair of sentences "I think these coefficients are the sam size and sign. I have very little information about their particular magni tudes." This could be translated into one of several families of isodensit surfaces. The traditional degenerate normal distributions would lead t cylindrical surfaces around the vector of ones. Such a density has bee used by Lindley and Smith (1972) to produce a modified Stein estimato and by Shiller (1973) to produce a distributed lag estimator. For reasons t be explained below, a better family of isodensity surfaces consists of cone from the origin also around the vector of ones.

Normal priors that reflect this information may be constructed a follows. Begin with a multivariate normal distribution for $\boldsymbol{\beta}$ located at th origin with covariance matrix $\sigma^2\mathbf{I}$. The joint distribution of $\boldsymbol{\beta}$ and the mea of the coefficients $\bar{\beta}$ (a scalar) then has covariance matrix

$$\operatorname{var}\begin{bmatrix}\boldsymbol{\beta}\\ \bar{\beta}\end{bmatrix}=\sigma^2\begin{bmatrix}\mathbf{I} & \mathbf{1}k^{-1}\\ k^{-1}\mathbf{1}' & k^{-1}\end{bmatrix}$$

where $\mathbf{1}$ is a vector of ones, and $\mathbf{1}'\mathbf{1}=k$. Conditional on $\bar{\beta}$ the moments $\boldsymbol{\beta}$, therefore, would be

$$E(\boldsymbol{\beta}|\bar{\beta})=\mathbf{1}\bar{\beta}$$

$$V(\boldsymbol{\beta}|\bar{\beta})=\sigma^2(\mathbf{I}-\mathbf{1}\mathbf{1}'/k).$$

Retain these conditionals but let $\bar{\beta}$ have variance $v\neq\sigma^2/k$. Marginally,

then has mean vector zero and covariance matrix

$$V(\beta)=\sigma^2\mathbf{I}+\mathbf{1}\left(v-\frac{\sigma^2}{k}\right)\mathbf{1}'.$$

Isodensity contours take the form

$$\begin{aligned}z&=\beta' V^{-1}(\beta)\beta\\&=\beta'\left(\mathbf{I}-\mathbf{1}(\mathbf{1}'\mathbf{1}+s^{-2})^{-1}\mathbf{1}'\right)\beta\end{aligned}$$

where

$$s^2=v-\frac{\sigma^2}{k}.$$

Last, let v get large to become diffuse on $\bar{\beta}$, and the isodensity contours become

$$z=\beta'(\mathbf{I}-\mathbf{1}k^{-1}\mathbf{1}')\beta.$$

These are cylinders around the vector $\mathbf{1}$, since by analogy to the error sum of squares in the least-squares algebra, z is the length of the difference between β and the projection of β onto $\mathbf{1}$.

In other words, a normal probability function for β with conditional distribution $f(\beta|\bar{\beta})$ as if β were spherical and with $\bar{\beta}$ diffuse implies cylindrical isodensity surfaces. In two dimensions, these are lines parallel to the vector $(1,1)$ in Figure 5.11. To this author, this is an exceedingly poor characterization of the statement "I think β_1 and β_2 have the same sign and magnitude," since it indicates indifference between, for example, the vector $(1,-1)$ and the vector $(100,98)$. The former fails the test of equality of coefficients miserably, and the latter passes it admirably.

A better family of isodensity surfaces to express this kind of prior information is the conically uniform family. This family indicates indifference between all vectors that make the same angle with the vector of ones. That is, isodensity surfaces depend on the cosine of the angle between β and $\mathbf{1}$

$$z=1-\frac{(\beta'\mathbf{1})^2}{\beta'\beta\mathbf{1}'\mathbf{1}}=1-\cos^2(\beta,\mathbf{1})$$

The information contract curve (5.12) can then be written as

$$\mathbf{0}=\left[\frac{(\beta'\mathbf{1})}{\beta'\beta}\mathbf{1}-\frac{(\beta'\mathbf{1})^2}{(\beta'\beta)^2}\beta\right]\lambda+\mathbf{X}'\mathbf{X}(\beta-\mathbf{b}) \tag{5.19}$$

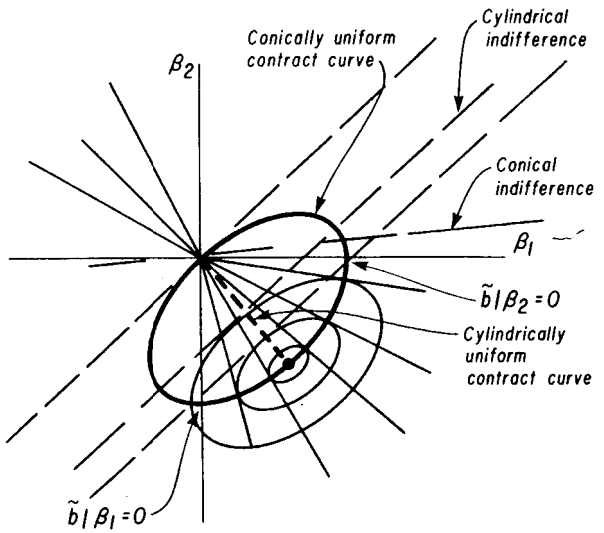

Fig. 5.11 Conically uniform and cylindrically uniform priors.

where

$$a^*(\beta)=\frac{(\beta'\beta)}{(\beta'\mathbf{1})}=\frac{\bar{\beta}}{\cos^2(\beta,\mathbf{1})}$$

$$\lambda^*=\frac{\lambda(\beta'\mathbf{1})^2}{(\beta'\beta)^2}.$$

Equation (5.20) appears to be the familiar matrix-weighted average of t
least-squares estimate **b** and the vector $a^*(\beta)\mathbf{1}$. Thus we can loosely thi
of the conically uniform prior as inducing a spherically uniform pri
located at $a^*(\beta)\mathbf{1}$. Note that the location of this vector involves
expansion of $\mathbf{1}\bar{\beta}$ by the amount $\cos^{-2}(\beta,\mathbf{1})$. Thus when β and $\mathbf{1}$ a
orthogonal the effective location of the prior is $\pm\mathbf{1}\infty$, and the estima
may be substantially pulled from its location.

A more accurate understanding of the behavior of this informati
contract curve can be obtained by premultiplying (5.19) by β':

$$\mathbf{0}=\left(\frac{(\beta'\mathbf{1})^2}{(\beta'\beta)}-\frac{(\beta'\mathbf{1})^2}{(\beta'\beta)}\right)\lambda+\beta'\mathbf{X}'\mathbf{X}(\beta-\mathbf{b})$$

which can be rewritten as

$$\left(\beta-\frac{\mathbf{b}}{2}\right)'\mathbf{X}'\mathbf{X}\left(\beta-\frac{\mathbf{b}}{2}\right)=\frac{\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}}{4}.$$

This is the same ellipsoid as the set of all constrained least-squares points (5.1). Since the contract curve is thus a continuous set of points lying on the feasible ellipsoid, it is obvious that the hull of no finite set of constrained least-squares points can contain the curve. In that sense it is undesirable to try to characterize the curve by a set of constrained least-squares points.

An example illustrating the difference between conically and cylindrically uniform priors is depicted in Figure 5.11. The information contract curve with cylindrically uniform priors is a straight line from the least-squares point to the origin, always in the fourth quadrant in this example. In contrast, the information contract curve with conically uniform priors is an ellipse connecting $\mathbf{b}$ to $\mathbf{b}_{\{1\}}$ (the variable with the lower t value is dropped) to the origin. In other words, although imposing the constraint $\beta_1=\beta_2$ leads to the estimate (0,0), this point and the least-squares point alone greatly distort the pooled information. It would be better to report also the estimate computed when the variable with the lower t value is dropped. Even this constitutes a distortion of the pooled information, since the contract curve travels very distinctly through the third quadrant. (The arm of the contract curve that travels into the first quadrant is dominated by the arm that travels into the third quadrant.)

To summarize this section, the vague notion that coefficients are equal in magnitude and sign is not well captured by elliptically uniform priors or their degenerate counterparts—cylindrically uniform priors. A better choice of isodensity surfaces is a family of cones. A family of cones leads to a very different kind of information contract curve than does a family of ellipsoids. Of the contract curves discussed in this section it is the one that is most poorly approximated by a sequence of constrained least-squares estimates.

SUMMARY

The literature on biased estimation of the regression parameter vector can be thought to involve three choices: (a) the choice of a (prior) location, (b) the choice of a distance function for measuring closeness to that location, and (c) the choice of a particular value of the distance function, or in the other language, the shrinkage factor. The "Stein" estimators and both Bayesian and non-Bayesian variants of Hoerl and Kennard's ridge regression presume the existence of (1) the (prior) location and (2) the metric, and they argue over (3) the shrinkage factor. But once the location and metric have been selected, the set of potential estimates has been reduced

to a curve, or possibly mixtures of points along a curve. In this section the choice of metric is emphasized, and the choice of shrinkage factor is deemphasized. It is more important to know which curve is appropriate than to pick a particular point on the curve. Furthermore, the choice of a point on the curve depends significantly on features of the prior that are likely to be difficult to select; in that event a useful data analysis tool is a graphical or mathematical representation of the whole curve.

The choice of metric is important from the purist Bayesian viewpoint since it shrinks the set of potential posterior modes from the feasible ellipsoid to an information contract curve. The choice of metric is equivalently important when only constrained regressions are computable, since it determines the coordinate system, and also the order in which constraints are imposed. All that is left undetermined is the weighting function, which determines a single estimate as a weighted average of the set of constrained estimates.

The usual elliptical metrics are closely associated with principal component regression. The more common regression-selection strategies cannot be justified with elliptical metrics, and we have been forced to consider hyperbolic, lexicographic, and conical metrics. These have been associated respectively, with strategies that omit insignificant variables, that omit predetermined variables, and that impose equality constraints. Conical metrics might be more appropriate than cylindrical metrics when coefficients are thought to be similar in size, but no regression selection strategy is appropriate with conical metrics. A more careful description of the information contract curve is required.

5.6.2 *Choice of Weight Functions*

An interpretive search strategy involves, first, a choice of origin, second, a choice of coordinate system for imposing constraints, and third, a weighting function that selects among the set of constrained estimates. Prior isodensities imply a contract curve that can be approximated by a sequence of constrained estimates. In that sense the choice of origin and coordinate system corresponds to the choice of prior isodensities. Selection of one of the constrained estimates or more generally the specification of a weighting function over constrained estimates requires a fully specified prior. In this section we select labeling functions for the elliptically uniform priors and explore the resultant weight functions.

In discussing the weights it is notationally convenient to write the omitted variables as a matrix $\mathbf{Z}$. That is, the regression process may be written

$$\mathbf{Y}=\mathbf{W}\delta+\mathbf{Z}\gamma+\mathbf{u}. \qquad (5.21$$

where $\mathbf{W}$ and $\mathbf{Z}$ are $(T \times k_w)$ and $(T \times k_z)$ observable matrices, and $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are $(k_w \times 1)$ and $(k_z \times 1)$ unobservable vectors. *Letting the prior be normal* with mean vector $\mathbf{0}$ and variance $\mathbf{D}^* = \text{diag}(d_1, d_2, \ldots, d_k)$ the weight to be applied to the estimate $[(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}, \mathbf{0}]$ is from Theorem 5.5

$$w_x \propto \left(\prod_I d_i \right) |\mathbf{W}'\mathbf{W}| \sigma^{-2k_w} \tag{5.22}$$

where the nonempty set I contains the k_z indices subscripting the left-out variables. These are conditional weights applicable marginally when both the process variance σ^2 is known and when the prior for the coefficients is in the normal family with known variance. Note especially that these weights are independent of the sample result $\mathbf{Y}$ and thus do not depend on any test statistics. This straightforwardly parallels the result that under these assumptions a posterior mean is a fixed (independent of $\mathbf{Y}$) matrix weighted average of the sample mean and the prior mean.

Another result discussed in Chapter 3 is that a conjugate normal-gamma prior with σ^2 uncertain also leads to a posterior mean that is a fixed, weighted average of the sample point and prior mean. Similarly, the weights in (5.22) would be independent of the sample. That is, with $d_i = \sigma^{-2} n_i$ we have

$$w_z \propto \left(\prod_I n_i \right) |\mathbf{W}'\mathbf{W}|$$

since $\sigma^{k_w + k_z}$ is a constant.

The fixed weighted mixing of sample and prior implied by conjugate distributions has justifiably encountered the criticism of Dickey (1975) and others. A fairly tractable analysis that implies variable weighting results when the coefficient vector comes from a multivariate Student distribution. That is, let us write $d_i = \sigma_1^{-2} n_i$, $\mathbf{N}^* = \text{diag}(n_1, n_2, \ldots, n_k)$, $k_1 = \text{rank}(\mathbf{N}^*)$, $\sigma_1^{-2} \sim f_\gamma(\ |1, \nu_1^*)$ where $f_\gamma(\ |1, \nu_1)$ indicates a gamma distribution with location and scale parameters 1 and ν_1. Furthermore, let us employ a gamma prior for σ^{-2}, $\sigma^{-2} \sim f_\gamma(\ |s_1^2, \nu_1)$. The resulting posterior is proportional to the product of two multivariate Student distributions and is relatively intractable. The marginal mode, as described in Chapter 3, requires the iterative solution of

$$\mathbf{b}^{**}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{N}^*)^{-1} \mathbf{X}'\mathbf{Y}$$

$$\lambda = \frac{[\nu_1 s_1^2 + (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})]/[T + \nu_1]}{[\nu_1^* + \boldsymbol{\beta}'\mathbf{N}^*\boldsymbol{\beta}]/[\nu_1^* + k]}$$

where λ estimates the variance ratio σ^2/σ_1^2. The weights analogous to (5.22) are

$$w_z \propto \left(\prod n_i \right) |\ \mathbf{W}'\mathbf{W}| \lambda^{-k_w}. \tag{5.23}$$

These weights depend on the sample $\mathbf{Y}$ only through the variance rat λ. The sample-dependent factor λ^{-k_w} is a constant for all regressio involving the same number of restrictions. Thus *the sample influences on the choice of number of restrictions*. The critical value of λ is one. When λ less than one, equations with few restrictions are favored, conversely for greater than one. The following result makes this point explicit by e pressing the posterior mean as a weighted average of $k+1$ rotatic invariant average regressions, each of which is a *fixed* weighted average constrained least squares estimates involving exactly j restrictions (j $0, 1, \ldots, k$).

THEOREM 5.8 (ROTATION INVARIANT AVERAGE REGRESSIONS) *Tl posterior mean corresponding to a spherical prior can be written as*

$$\mathbf{b}^{**}(\lambda) = (\mathbf{N} + \lambda \mathbf{I})^{-1} \mathbf{N}\mathbf{b} = \sum_{j=0}^{k} w_j(\mathbf{X}, \lambda) \mathbf{a}_j \tag{5.2}$$

where

$$\mathbf{a}_j = \sum_{I \in C_j} |\mathbf{N}_I| \mathbf{b}_I p_j^{-1}$$

$$p_j = \sum_{I \in C_j} |\mathbf{N}_I|$$

$$w_j(\mathbf{X}, \lambda) = p_j \lambda^j \Big/ \sum_{j=0}^{k} p_j \lambda^j$$

with C_j the set of all subsets of the first k integers taken j at a time a with $\mathbf{N}_I$ a matrix formed by deleting rows i and columns i of $\mathbf{N} = \mathbf{X}'\mathbf{X}$ f all $i \in I$.

Proof: See Appendix 3.

Any prior that is uniform on ellipsoids can, by a linear transformatic be made uniform on spheres. By Theorem 5.8, posterior modes implied such distributions are weighted averages of $k+1$ rotation invariant avera regressions, $\mathbf{a}_j$. Each such point is a weighted average of constrain least-squares points involving exactly j restrictions with weights th cannot depend on the data $\mathbf{Y}$. Given an elliptical prior, the choice labeling function can thus influence *only* the number of restrictions th are imposed.

The rotation invariant average regressions derive their name from t surprising property that they are invariant to rotations of the parame space. This property is illustrated in the two-dimensional case in Figu

5.12. In the original coordinate system connect with a straight line the two constrained least-squares points given $\beta_1=0$ or $\beta_2=0$. Find the constrained least-squares points in any other (orthogonal) coordinate system and connect them with a straight line. The point of intersection is $\mathbf{a}_1$, because $\mathbf{a}_1$ is a weighted average of constrained least-squares points in any coordinate system.

The weights (5.23) are, in general, very complicated but in the special case when the contract curve is a straight line they do imply familiar test procedures:

ONE RESTRICTION

When prior information is diffuse in all directions but one, the posterior mean is a simple weighted average of the estimates resulting from dropping and not dropping the relevant variable. The weight to be applied to the restricted estimate is given in (5.23)

$$w_z \propto n_z\left(\mathbf{Z}'\mathbf{Z}-\mathbf{Z}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Z}\right)^{-1}\lambda^{-1}. \tag{5.25}$$

The weight given the unrestricted estimator is

$$w_x \propto 1.$$

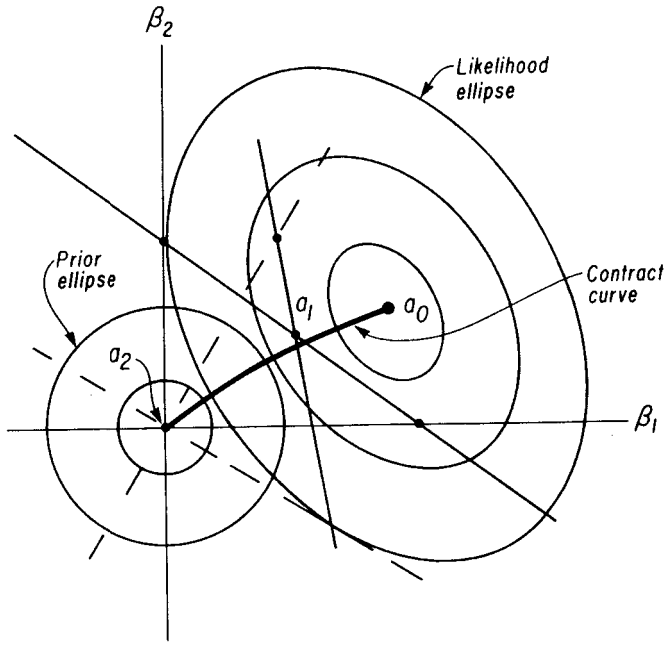

Fig. 5.12 Rotation invariant average regressions.

Note now that for ν_1^* and ν_1 small, the first iteration of the equation f computing λ has

$$\lambda=\frac{\hat{\sigma}^2}{n_z\hat{\gamma}^2}$$

where $\hat{\gamma}$ and $\hat{\sigma}^2$ are the maximum likelihood estimates. Thus the weig becomes approximately

$$w_z \propto t_\gamma^{-2}$$

where t_γ is the t value for testing the restriction $\gamma=0$. Perhaps mc informatively, we may write

$$w_z=\frac{1}{1+t_\gamma^2}, \qquad w_x=\frac{t_\gamma^2}{1+t_\gamma^2}.$$

The estimate of σ^2 in these results is uncorrected for degrees of freedo this is just the first iteration toward the mode, and the mode is only o aspect of the posterior distribution. Note especially that further iteratic to the mode move the estimate closer to the restricted least-squares poi Nonetheless, a somewhat distorted Bayesian analysis with information one dimension only results in an estimate that is a weighted average dropping and not dropping the variable with a weight on the unrestrict estimate proportional to the F statistic commonly used to test the restr tion.

MULTIPLE RESTRICTIONS: THE STEIN SOLUTION

When information is available in several directions, the weights implied (5.23) are more complicated. One special case is interesting, howev Suppose $\mathbf{X}'\mathbf{X}$ is proportional to $\mathbf{N}^*$.

$$\mathbf{N}^*=c\mathbf{X}'\mathbf{X}.$$

The posterior modal equation can then be written as a weighted average the origin and the unrestricted estimator

$$\boldsymbol{\beta}=(1+\lambda c)^{-1}[\mathbf{b}+\lambda c\mathbf{0}]$$

and the variance ratio λ becomes

$$\lambda=\frac{\left(\nu_1 s_1^2+(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})+(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})\right)/(T+\nu_1)}{(\nu_1^*+c\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta})/(\nu_1^*+k)}.$$

The first iteration of these two equations for ν_1^* and ν_1 small yields

$$\lambda=\frac{(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})/T}{c\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/k}\cong(cF)^{-1}$$

where F is the value of the F statistic for testing the restriction $\boldsymbol{\beta}=\mathbf{0}$. Using this in the formula for the mode of $\boldsymbol{\beta}$ and indicating the weight assigned to the zero vector by w_0 and the weight assigned to the unrestricted estimator by w, we obtain approximately

$$w_0=\frac{1}{1+F}, \qquad w=\frac{F}{1+F}.$$

In words, a somewhat distorted Bayesian analysis with prior information structurally equivalent to sample information ($\mathbf{X}'\mathbf{X}\propto\mathbf{N}^*$) results in a posterior mode that is approximately a weighted average of the zero vector and the least-squares vector with the weight on the least-squares vector proportional to the F statistic commonly used to test the restriction. The structural equivalence assumption thus appears to be implicit in "Stein estimators." This is a vindication to a Bayesian of such estimators only in the unlikely case that his prior is, in fact, structurally equivalent to the sample.[13]

5.7 Multicollinearity and Local Sensitivity Analysis[14]

An interpretive search that involved omitting variables would be worthless if the data were orthogonal, since the estimates and their standard errors[15] would not change and therefore would not "improve." Thus collinear data is a handmaiden of interpretative searches, and this is a useful point to discuss the "collinearity problem."

The principal claim of this section is that the most important aspects of the collinearity problem derive from the existence of potentially useful, uncertain, prior information, which causes major problems in interpreting the data evidence. It is claimed here that if our a priori knowledge of parameter values were either completely certain or "completely uncertain," the aspects of the collinearity problem that most of us worry about would disappear. As an empirical test of this proposition, consider the situations when collinearity is identified as a culprit. Usually, signs are wrong or point estimates are otherwise peculiar. Occasionally, confidence intervals overlap unlikely regions of the parameter space. Yet to say these things is to say there exists uncertain prior information.

[13]A more general result is that a simple weighted average of imposing and not imposing the restriction is valid (in the sense discussed) if and only if the curve décolletage or contract curve $\mathbf{b}^{**}(\lambda)=(\lambda\mathbf{N}^*+\mathbf{X}'\mathbf{X})^{-1}(\lambda\mathbf{N}^*\mathbf{b}^*+\mathbf{X}'\mathbf{Y})$ is a straight line. [See Chamberlain and Leamer (1976) for a discussion of the necessary and sufficient conditions.] In this case, the weights are approximated by $F/(1+F)$ and $1/(1+F)$ where F is the usual F statistic for testing the restriction. Weights of this form were originally suggested by Huntsberger (1955). They are also discussed in Alam and Thompson (1964) and Baranchik (1970). Feldstein (1973) uses t^2 weights in a Monte Carlo study.

[14]This section is taken from Leamer (1973).

[15]Conditional on σ^2. See comments at the end of this section.

Classical inference, with the possible exception of the pretesting liter ture, necessarily excludes undominated uncertain prior information.[16] As result, most discussions of the collinearity problem miss a critical poi The textbook discussions, including Theil (1971, p. 149), Malinvaud (197 p. 218), and Goldberger (1964, p. 192), observe that when the desi matrix $\mathbf{X}'\mathbf{X}$ becomes singular, the least-squares estimator is nonunique, a the sampling distribution has finite variance only for certain "estimabl functions. Thus extreme collinearity is implicitly defined as total lack sample information about some parameters.

The case of less extreme collinearity is not dealt with so trivially, sin there is nothing in the least-squares theorems that is obviously depend on the "near noninvertibility" of the design matrix. This fact has l Kmenta (1971, p. 391) to conclude "that a high degree of multicollinear is simply a feature of the sample that contributes to the unreliability of t estimated coefficients, but has no relevance for the conclusions drawn a result of this unreliability."

To put this another way, the problem of defining collinearity may solved by identifying a distance function for measuring the closeness of design matrix to some noninvertible matrix in which the collinear problem is unambiguously extreme. Since the extreme case is associa with infinite marginal variances on the parameters, authors such as Th (1971, p. 152), Malinvaud (1970, p. 218), and Goldberger (1964, p. 193) a distance function informally related to the sampling variance of coefficients. Collinearity is defined as large variances. The failure of t definition is that instead of defining a new *problem*, it identifies a n *cause* of an already well-understood problem—weak evidence. Altho collinearity as a cause of the weak-evidence problem can be distinguish from other causes, such as small samples or large residual error varian collinearity as a problem is by this definition indistinguishable from weak-data problem in general. Thus Kmenta's conclusion that there really nothing special about the collinearity problem is appropriate. Stil gnawing confusion remains. Goldberger (1964, p. 201) concludes discussion with accurate ambiguity, "...when orthogonality is absent concept of the contribution of an individual regressor remains inheren ambiguous."

The point of this section is that there *is* a special problem caused collinearity. This is the problem of *interpreting* multidimensional evide Briefly, collinear data provide relatively good information about lin combinations of coefficients. The interpretation problem is the problem deciding how to allocate that information to individual coefficients. T depends on prior information. A solution to the interpretation prob thus involves formalizing and utilizing effectively all prior informati

[16]Precise prior information is quite relative to the given data information.

The weak-evidence problem, however, remains, even when the interpretation problem is solved. The solution to the weak-evidence problem is more and better data. Within the confines of the given data set there is nothing that can be done about weak evidence.

A Bayesian with a well-defined prior distribution can, of course, have no problem interpreting the sample evidence, since he merely computes his posterior distribution. A Bayesian with poorly defined priors or a wide readership may have extreme difficulties in reporting and interpreting evidence. This suggests the following definition:

Definition. The *collinearity problem* is said to affect a parameter β_i if the apparent sample evidence about β_i depends on ambiguous uncertain prior information about other parameters, where ambiguous means that readers differ in their judgments or that they are not too sure how to select features of their prior. This is made more precise subsequently.

Since classical inference provides no assistance in using uncertain prior information, this definition does not apply directly to everyday "classical" inferences. An easy, ad hoc procedure used when analyzing data is to neglect the off-diagonal terms of $(\mathbf{X}'\mathbf{X})^{-1}$ and to proceed as if the sample evidence were generated by an orthogonal experiment. This may lead to significant misinterpretations of the data and suggests an alternative definition:

Definition. The *collinearity problem* is said to affect β_i if the sample evidence about β_i is distorted by an analysis that proceeds as if the data were orthogonal. This is also made more clear shortly.

MULTICOLLINEARITY: THE WEAK-DATA PROBLEM

The unique problem associated with collinear data is the problem of interpreting multidimensional evidence. Collinear data is also a *cause* of the weak-data problem. In this section we show how collinearity causes weak evidence, where weak evidence is defined as the necessary coincidence of the prior and posterior distribution for some parameter $g(\boldsymbol{\beta})$.

In particular, suppose there is an extreme collinearity problem with the columns of $\mathbf{X}$ being perfectly collinear. Then there exists a vector $\boldsymbol{\psi}$ such that $\mathbf{X}\boldsymbol{\psi}=\mathbf{0}$. We wish to show that there is a function $g(\boldsymbol{\beta})$ that necessarily has the same prior and posterior distribution regardless of the sample outcome $\mathbf{Y}$. This is true, in particular, for $g(\boldsymbol{\beta})=\boldsymbol{\psi}'\mathbf{N}^*\boldsymbol{\beta}$, Malinvaud (1970, pp. 246–249). The prior moments of $\boldsymbol{\psi}'\mathbf{N}^*\boldsymbol{\beta}$ are

$$E(\boldsymbol{\psi}'\mathbf{N}^*\boldsymbol{\beta})=\boldsymbol{\psi}'\mathbf{N}^*\mathbf{b}^*$$

where the prior variance matrix has been set to $(h_1\mathbf{N}^*)^{-1}$. The posteric moments are

$$\begin{aligned}E(\boldsymbol{\psi}'\mathbf{N}^*\boldsymbol{\beta}|\mathbf{Y})&=\boldsymbol{\psi}'\mathbf{N}^*\mathbf{b}^{**}\\&=\boldsymbol{\psi}'(\mathbf{N}+\mathbf{N}^*)\mathbf{b}^{**}\qquad(\text{because }\boldsymbol{\psi}'\mathbf{N}=\mathbf{0})\\&=\boldsymbol{\psi}'(\mathbf{N}\mathbf{b}+\mathbf{N}^*\mathbf{b}^*)=\boldsymbol{\psi}'\mathbf{N}^*\mathbf{b}^*\\V(\boldsymbol{\psi}'\mathbf{N}^*\boldsymbol{\beta}|\mathbf{Y})&=\boldsymbol{\psi}'\mathbf{N}^*V(\boldsymbol{\beta}|\mathbf{Y})\mathbf{N}^*\boldsymbol{\psi}\\&=h_1^{-1}\boldsymbol{\psi}'(h\mathbf{N}+h_1\mathbf{N}^*)V(\boldsymbol{\beta}|\mathbf{Y})\mathbf{N}^*\boldsymbol{\psi}\\&=h_1^{-1}\boldsymbol{\psi}'\mathbf{N}^*\boldsymbol{\psi}\end{aligned}$$

where $h=\sigma^{-2}$. These moments are seen to be the same as the pri moments, and, conditional on h and h_1, there can be no learning abo $\boldsymbol{\psi}'\mathbf{N}^*\boldsymbol{\beta}$. The evidence is thus necessarily weak about this linear combin tion of parameters. (Note, by the way, that this linear combination d pends on the prior through $\mathbf{N}^*$. See Section 5.9 also.)

Weaker forms of collinearity imply that this result is almost true. The are functions of the parameters about which we can learn very little. The is, of course, no cure for weak data, except more and better data. collinearity were only a cause of the weak-data problem it would me very little mention. The more interesting and more difficult aspect collinearity is the interpretation problem.

INTERPRETING COLLINEAR DATA: A BAYESIAN ANALYSIS OF AN AD HOC PROCEDURE

Although it is possible to make enlightened use of prior informati through interpretive searches, perhaps as suggested by the pretesting lite ture, we assume that a researcher has before him only the sufficie statistics and no computer, as would be the case of a reader of a techni report. Off-diagonal terms of $(\mathbf{X}'\mathbf{X})^{-1}$ may not be reported and, even they are, classical inference provides no very clear way of interpreti them. Instead, many of us in this situation would proceed as if $(\mathbf{X}'\mathbf{X})$ were diagonal. Furthermore, when prior information on the coefficients available, we may choose to ignore the a priori covariance terms.

An example can usefully illustrate what I have in mind. A logarithr regression of a volume index of purchases of meat C_m on money income price of meat P_m, and a general consumer price index P_Y yields regression that is underreported as

$$\begin{aligned}\log C_m&=\alpha+\beta_1\log Y+\beta_2\log P_m+\beta_3\log P_Y\\&=5.0+\underset{(.2)}{.9}\log Y-\underset{(.2)}{.2}\log P_m-\underset{(.2)}{.1}\log P_Y\end{aligned}$$

where standard errors are indicated in parentheses. The researcher th

against his marginal prior for each coefficient. He finds the money-income elasticity β_1 to be a "bit high," the direct-price elasticity β_2 to be "about right," and the cross-price elasticity β_3 "to have the wrong sign but not significantly so." The error that is being made here is, first, to ignore the data tradeoffs implied by the off-diagonal terms of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix and, second, to ignore the prior tradeoffs implied by a nondiagonal $\mathbf{N}^*$ matrix. For example, the researcher may have independent information about a homogeneity parameter $\beta_1+\beta_2+\beta_3$, a real income elasticity $\beta_1-\beta_3$, and the price elasticity β_2, and he is unlikely to regard β_1, β_2, and β_3 to be a priori independent.

Proceeding as he did, the researcher has made an error in pooling the prior information and the sample information. He has treated a k-dimensional problem as if it were k one-dimensional problems; he will be making misinterpretations of the data evidence unless his prior and data fit together in a special way. Thus the collinearity problem creates a situation in which it is necessary to process prior information carefully.

More formally, inferences about the coefficient vector often proceed as if the posterior mean were on the *diagonalized contract curve*

$$\mathbf{d}(\lambda)=(\mathbf{D}+\lambda\mathbf{D}^*)^{-1}(\mathbf{D}\mathbf{b}+\lambda\mathbf{D}^*\mathbf{b}^*) \tag{5.26}$$

where $\mathbf{D}^{-1}$ and $\mathbf{D}^{*-1}$ are diagonal matrices formed by setting the off-diagonal elements of $\mathbf{N}^{-1}$ and $\mathbf{N}^{*-1}$ to zero. If $\mathbf{N}$ and $\mathbf{N}^*$ are diagonal, Bayes rule may be applied coefficient by coefficient, and the resultant conditional posterior mean is given by (5.26). For $\mathbf{N}$ or $\mathbf{N}^*$ nondiagonal, the true contract curve

$$\mathbf{b}^{**}(\lambda)=(\mathbf{N}+\lambda\mathbf{N}^*)^{-1}(\mathbf{N}\mathbf{b}+\lambda\mathbf{N}^*\mathbf{b}^*) \tag{5.27}$$

may deviate substantially from $\mathbf{d}(\lambda)$ and the ad hoc use of prior information may cause major data misinterpretations and ultimately unnecessary expected losses (see Figure 5.13). Collinearity thus creates an incentive to use prior information more carefully:

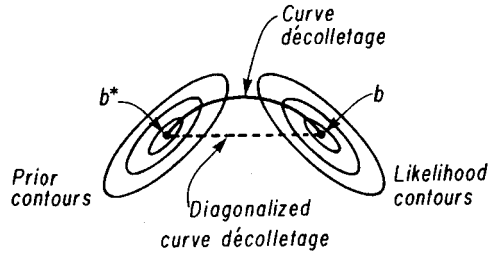

Fig. 5.13 Diagonalized contract curve.

Definition. A coefficient β_i is said to suffer from the collinearity problem if the ith component of the true contract curve $b_i^{**}(\lambda)$ differs substantially from the ith component of the diagonalized contract curve $d_i(\lambda)$.

The difference between $\mathbf{d}(\lambda)$ and $\mathbf{b}^{**}(\lambda)$ may be assessed in several ways. One suggestion of Leamer (1973) is the following:

Rectangle Test. One aspect of a univariate problem is that the posterior mean is necessarily between the prior mean and the sample mean. Thus the location change induced by the sample evidence has an unambiguous sign and is limited in distance by the sample point. The diagonalized contract curve also has this property, coefficient by coefficient; that is, it lies in the rectangular solid with diagonal $[\mathbf{b}^*, \mathbf{b}]$. The true contract curve need not have this property, and the sign of the elements of $\mathbf{b}^{**}(\lambda)-\mathbf{b}$ and $\mathbf{b}^{**}(\lambda)-\mathbf{b}$ may be ambiguous. Thus, it may not be possible to say whether the data suggest positive or negative revisions to opinions about some coefficient or to limit the distance of the revision, until a full prior distribution is specified. This suggests the following definition.

Definition. The collinearity problem in the *rectangle sense* is said to affect a coefficient β_j if the contract curve $\mathbf{b}^{**}(\lambda)$ travels outside the slab $b_j \leqslant b_j^{**}(\lambda) \leqslant b_j^*$. The collinearity problem in the rectangle sense affects no parameters if the contract curve lies everywhere in the rectangular solid with diagonal $[\mathbf{b}, \mathbf{b}^*]$.

The contract curve may, in general, lie anywhere, Theorem (5.10). Even with orthogonal data, that is, with $\mathbf{X}'\mathbf{X}$ diagonal, it need not be restricted to the appropriate rectangular solid. *Thus orthogonal data is not sufficient to prevent the collinearity problem. We must also restrict the class of priors.* If $\mathbf{N}$ and $\mathbf{N}^*$ are both diagonal, there is no collinearity problem in this sense. If $\mathbf{N}$ is proportional to $\mathbf{N}^*$, the contract curve is a straight line, and there is, again, no collinearity problem.[17]

INTERPRETING COLLINEAR DATA: THE BAYESIAN PROBLEMS

Given a prior distribution, the posterior distribution is fully defined, and there is no ambiguity about measures of location and thus no collinearity problem in the sense of the previous section. A Bayesian, therefore apparently has no special difficulty working with collinear data. This, however, ignores difficulties in selecting an acceptable prior distribution.

[17]Note that even if $\mathbf{N}$ and $\mathbf{N}^*$ are both diagonal, collinearity may be said to affect certain linear combinations of parameters. Collinearity affects no linear combinations only if $\mathbf{N}$ proportional to $\mathbf{N}^*$.

When collinearity is present, the posterior distribution may be highly sensitive to changes in the prior, and apparently innocuous differences in the prior may be amplified into significant differences in the posterior distribution. Thus the collinearity problem is transformed from a problem of characterizing and interpreting a multidimensional likelihood function into a problem of characterizing and interpreting a multidimensional prior distribution.

In the clearly collinearity-free case with $\mathbf{N}$ and $\mathbf{N}^*$ diagonal, the posterior distribution of any coefficient is conditionally independent of the prior distributions of the other coefficients. This suggests the following slightly ambiguous definition of the collinearity problem:

Definition. The collinearity problem is said to affect parameter β_i if the interpretation of the sample evidence about β_i depends meaningfully on uncertain prior information about the other parameters. The *interpretation of the sample* evidence about β_i is a mapping of marginal prior distributions for β_i into marginal posterior distributions. The phrase "depends meaningfully" can be interpreted in terms of both the location change and the scale change induced by the sample evidence. As in the previous section, we restrict ourselves to the location change.

In general, the interpretation of the sample evidence about one coefficient is sensitive to the prior about others, because of prior correlations. To make sense out of this definition, we thus have to define meaningful classes of priors within which to perform the sensitivity analysis.

The sensitivity of the posterior mean to variations in the prior mean holding other things constant is indicated by the matrix of derivatives

$$\frac{\partial \mathbf{b}^{**}(\lambda)}{\partial \mathbf{b}^*} = \lambda(\mathbf{N}+\lambda\mathbf{N}^*)^{-1}\mathbf{N}^*.$$

The off-diagonal elements of this matrix indicate the extent to which the conditional posterior mean of one coefficient depends on the prior mean of the others. These are zero for $\mathbf{N}$ and $\mathbf{N}^*$ diagonal, for $\mathbf{N}$ proportional to $\mathbf{N}^*$, and for $\mathbf{N}^*$ or λ equal to zero.

We may also be interested in the sensitivity of the posterior mean to changes in the prior variances. Let us write

$$(\mathbf{N}^*)^{-1} = \mathbf{DRD}$$

where $\mathbf{D}$ is a diagonal matrix with $\sqrt{V(\beta_i)}$ on the diagonal and $\mathbf{R}$ is the matrix of correlation coefficients. A change in the prior variances induces a change in the prior precision matrix according to the formula

$$d\mathbf{N}^* = -(\mathbf{N}^*\mathbf{D}^{-1}d\mathbf{D}+\mathbf{D}^{-1}d\mathbf{D}\mathbf{N}^*),$$

and a change in $\mathbf{N}^*$ induces a change in $\mathbf{b}^{**}$ according to the formula

$$d\mathbf{b}^{**}(\lambda) = (\mathbf{N}+\lambda\mathbf{N}^*)^{-1}(\lambda d\mathbf{N}^*)(\mathbf{N}+\lambda\mathbf{N}^*)^{-1}\mathbf{N}(\mathbf{b}^*-\mathbf{b}).$$

In the diagonal case, the ith element of $d\mathbf{b}^{**}(\lambda)$ depends only on the ith differential $\{d\mathbf{D}\}_{ii}$. Otherwise, changes in the prior variance of one coefficient induce changes in the posterior means of other coefficients.

Sensitivity analysis can also be performed with respect to the variance ratio λ. In the orthogonal case, the posterior mean (the curve decolletage) lies everywhere in the rectangular solid with diagonal $[\mathbf{b}, \mathbf{b}^*]$. In that case the sign and maximum distance of the mapping from prior to posterior mean are unambiguous, and we could say that the sample evidence does not depend meaningfully on prior information about λ. When the curve decolletage travels outside this region, the sample evidence does become ambiguous, and collinearity is the culprit. Note, by the way, that this is mathematically the same as the collinearity problem in the rectangle sense discussed in the previous section, although the interpretations are quite different. In that section the curve decolletage was assumed to lie in the relevant rectangular solid, and when it did not, a major data misinterpretation occurred. Here, we know where the curve décolletage lies, but we are uncertain whether particular points on the curve outside the rectangular solid are relevant, since our prior information about the variance ratio is ambiguous.

The derivatives just reported imply a local sensitivity analysis in which the consequences of small perturbations in the prior are analyzed. Global sensitivity analyses are discussed in the next section, but one result given there yields an especially interesting measure of collinearity. It is shown that if only the prior location is known (taken here to be the origin), then any posterior mean must lie within the feasible ellipsoid (5.1), and any point in the ellipsoid is a posterior mean for some prior. Projection of the ellipsoid on the ith axis yields the set of feasible estimates of β_i. In contrast, if β_i were the only parameter, then the posterior mean would necessarily lie between zero and b_i, the least-squares estimate. The ratio of the lengths of these two intervals is shown in the next theorem to be

$$c_{1i} = \left(\frac{\chi^2}{Z_i^2}\right)^{1/2} \geq 1,$$

where χ^2 is the chi-square statistic for testing the multivariate restriction $\boldsymbol{\beta}=\mathbf{0}$ and Z_i is the normal statistic for testing $\beta_i=0$.[18] (See Figure 5.14.) When c_{1i} is one, inferences about β_i are unaffected by the fact that the

[18]Notice that c_{1i} is proportional to the inverse of the square of the t statistic for testing $\beta_i=0$. Thus the ranking of coefficients by t statistics is equivalent to a ranking by the collinearity measure.

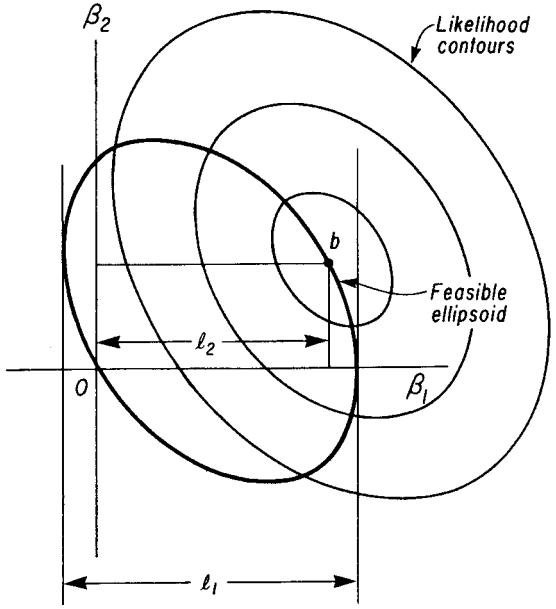

Fig. 5.14 A measure of collinearity $c_1 = l_1/l_2 = \sqrt{\chi^2/Z_1^2}$.

are other parameters, since the posterior location is constrained to lie in the interval $(0, b_i)$. When c_{1i} is greater than one, the existence of other parameters does cause difficulties for interpreting the evidence about β_i in the sense that the set of feasible estimates is enlarged.

THEOREM 5.9 (PROJECTION OF THE FEASIBLE ELLIPSOID). *Given* $\boldsymbol{\beta}$ *constrained to the ellipsoid*

$$(\boldsymbol{\beta}-\mathbf{f})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{f}) = c$$

where $\mathbf{f} = (\mathbf{b}+\mathbf{b}^*)/2$ *and* $c = (\mathbf{b}-\mathbf{b}^*)'\mathbf{X}'\mathbf{X}(\mathbf{b}-\mathbf{b}^*)/4$, *the extreme values of the linear combination* $\psi'\boldsymbol{\beta}$ *occur at the points*

$$\boldsymbol{\beta}^* = \mathbf{f} \pm (\mathbf{X}'\mathbf{X})^{-1}\psi\left(\frac{c}{\psi'(\mathbf{X}'\mathbf{X})^{-1}\psi}\right)^{1/2}.$$

The linear function $\psi'\boldsymbol{\beta}$ *at these points takes on the values*

$$\psi'\boldsymbol{\beta}^* = \psi'\mathbf{f} \pm \left(c\psi'(\mathbf{X}'\mathbf{X})^{-1}\psi\right)^{1/2}.$$

Proof: Setting the derivatives of the Lagrangian to zero yields the vector of equations $\mathbf{0} = \psi + \mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{f})\lambda$, which implies $(\boldsymbol{\beta}-\mathbf{f}) = -(\mathbf{X}'\mathbf{X})^{-1}\psi\lambda^{-1}$. Thus $c = (\boldsymbol{\beta}-\mathbf{f})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{f}) = \lambda^{-2}\psi'(\mathbf{X}'\mathbf{X})^{-1}\psi$, and $\lambda^2 = \psi'(\mathbf{X}'\mathbf{X})^{-1}\psi/c$.

This interval for $\psi'\boldsymbol{\beta}$ *has length equal to* $2(c\psi'(\mathbf{X}'\mathbf{X})^{-1}\psi)^{1/2} = (\chi^2 \mathrm{Var}(\psi'\mathbf{b}))^{1/2}$ *where* χ^2 *is the chi-square statistic for testing* $\boldsymbol{\beta} = \mathbf{b}^*$, $\chi^2 = \sigma^2(\mathbf{b}-\mathbf{b}^*)'\mathbf{X}'\mathbf{X}(\mathbf{b}-\mathbf{b}^*)$, *and* $\mathrm{Var}(\psi'\mathbf{b})$ *is the sampling variance of* $\psi'\mathbf{b}$, $\sigma^2\psi'(\mathbf{X}'\mathbf{X})^{-1}\psi$.

These measures of collinearity have described how the posterior location of one parameter β_i depends on the prior information about the other parameters. One measure mentioned in Section 5.5 describes the relationship between the (posterior) variance of β_i and the prior variance of the other parameters. If the prior for the vector $\boldsymbol{\beta}$ is diffuse, the posterior variance of β_i is proportional to $[(\mathbf{X}'\mathbf{X})^{-1}]_{ii}$, whereas if all the parameters except β_i were known exactly, then the variance of β_i would be proportional to $([\mathbf{X}'\mathbf{X}]_{ii})^{-1}$. The square root of the ratio of these two numbers thus measures the incentive to formulate prior information

$$c_2(\beta_i) = \left(\frac{\left[(\mathbf{X}'\mathbf{X})_{ii}\right]^{-1}}{\left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{ii}}\right)^{1/2}.$$

A value of $c_2(\beta_i)$ equal to one occurs when the ith row of $\mathbf{X}'\mathbf{X}$ has zeroes except in its ith element. A value of one indicates that there can be no gain in information about β_i by gathering more information about the other coefficients. A value of $c_2(\beta_i) = \frac{1}{2}$ indicates that if the other coefficients could be specified exactly, the confidence interval for β_i would be cut in half.

More generally, if we are interested in estimating the linear combination $\psi'\boldsymbol{\beta}$, the incentive to gather information about $\mathbf{R}\boldsymbol{\beta}$ can be measured by

$$c_2^2(\psi, \mathbf{R}) = \frac{\psi' V(\boldsymbol{\beta}|\mathbf{R}\boldsymbol{\beta})\psi}{\psi' V(\boldsymbol{\beta})\psi}.$$

where $V(\boldsymbol{\beta}|\mathbf{R}\boldsymbol{\beta})$ and $V(\boldsymbol{\beta})$ are the conditional and unconditional variances of $\boldsymbol{\beta}$.

These measures of the incentive to gather other information are similar to measures of the intercorrelation of the explanatory variables. Other measures of intercorrelation between a subset $\mathbf{X}_J$, and its complement $\mathbf{X}_{\bar{J}}$ are Hotelling's (1936) "coefficient of alienation"

$$\rho_a^2 = \frac{\det\left[\mathbf{X}_J'\left(\mathbf{I}-\mathbf{X}_{\bar{J}}(\mathbf{X}_{\bar{J}}'\mathbf{X}_{\bar{J}})^{-1}\mathbf{X}_{\bar{J}}'\right)\mathbf{X}_J\right]}{\det(\mathbf{X}_J'\mathbf{X}_J)}$$

and Hooper's (1962) "trace correlation coefficient"

$$\rho_t^2 = J^{-1}\mathrm{tr}\left[(\mathbf{X}_J'\mathbf{X}_J)^{-1}\left(\mathbf{X}_J'\left(\mathbf{I}-\mathbf{X}_{\bar{J}}(\mathbf{X}_{\bar{J}}'\mathbf{X}_{\bar{J}})^{-1}\mathbf{X}_{\bar{J}}'\right)\mathbf{X}_J\right)\right].$$

The analogous measure $c^2(\psi, \mathbf{R})$, with $\psi' = (\psi_J, \mathbf{0})$ and $\mathbf{R} = (\mathbf{0}, \mathbf{I}_{\bar{J}})$ is

$$c^2 = \frac{\psi_J'(\mathbf{X}_J'\mathbf{X}_J)^{-1}\psi_J}{\psi_J'\left[\mathbf{X}_J\left(\mathbf{I} - \mathbf{X}_{\bar{J}}(\mathbf{X}_{\bar{J}}'\mathbf{X}_{\bar{J}})^{-1}\mathbf{X}_{\bar{J}}'\right)\mathbf{X}_J\right]^{-1}\psi_J}$$

Notice that the difference between c^2 and ρ_a^2 is only that c^2 involves variances of a specific linear combination, whereas ρ_a^2 uses the generalized variances, $\det(V(\boldsymbol{\beta}_J))$. They are identical if the set J has only one element.

The discussion to this point has made use of the assumption that the residual variance σ^2 is known. In that event, the conditional confidence interval for $\psi'\boldsymbol{\beta}$ given $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ can never be larger than the confidence interval computed without benefit of the restriction. But if σ^2 is unknown, imposition of the constraint may, in fact, lead to a larger confidence interval, since the estimate of σ^2 necessarily changes. The multicollinearity measure, which by definition is the ratio of the length of a conditional to the length of an unconditional interval, may exceed one. This seems to be saying paradoxically that more information is less information; the more information corresponding to knowledge that $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, and the less information corresponding to the fact that the interval for $\psi'\boldsymbol{\beta}$ increases in length.

It is, in fact, *not* paradoxical that specific information may make you less certain, especially if that information is greatly at odds with what you currently believe. Knowledge of $\mathbf{R}\boldsymbol{\beta}$ will sometimes increase and sometimes decrease the confidence interval for $\psi'\boldsymbol{\beta}$. But the *expected* variance of $\psi'\boldsymbol{\beta}$ given $\mathbf{R}\boldsymbol{\beta}$, expected with respect to the distribution of $\mathbf{R}\boldsymbol{\beta}$, will always be less than the unconditional variance of $\psi'\boldsymbol{\beta}$, since $V(\psi'\boldsymbol{\beta}) = V(E(\psi'\boldsymbol{\beta}|\mathbf{R}\boldsymbol{\beta})) + E(V(\psi'\boldsymbol{\beta}|\mathbf{R}\boldsymbol{\beta}))$.

The incentive to obtain prior information about $\mathbf{R}\boldsymbol{\beta}$ should be measured in terms of the *expected* variance of $\psi'\boldsymbol{\beta}$ given $\mathbf{R}\boldsymbol{\beta}$:

$$c^2(\psi', \mathbf{R}) = \frac{E\left[V(\psi'\boldsymbol{\beta}|\mathbf{R}\boldsymbol{\beta})\right]}{V(\psi'\boldsymbol{\beta})}.$$

It is easy to show that this measure is precisely the same as the measures previously suggested when σ^2 was assumed to be known. Thus the results heretofore discussed continue to apply. The proof of this proposition could appeal to properties of multivariate Student distributions, but it is easier merely to observe that, conditional on σ^2, both $V(\psi'\boldsymbol{\beta})$ and $E[V(\psi'\boldsymbol{\beta}|\mathbf{R}\boldsymbol{\beta})]$ are proportional to σ^2. Integrating each with respect to the (posterior) distribution of σ^2, $f(\sigma^2|\mathbf{Y}, \mathbf{X})$, will thus multiply each by the *same* constant. Thus the ratio is unaffected by uncertainty in σ^2.

SUMMARY

The most discussed aspect of collinearity is the weak-data problem associated with large standard errors of estimated coefficients and, in a

Bayesian analysis, the coincidence of prior and posterior distributions on certain subspaces. As Kmenta suggests, there is nothing special about this problem in the collinearity context. What is special in the collinearity context is the major problems of interpreting the evidence.

When prior information is fully specified and unique, both personally and publicly, the posterior mean and hence the interpretation of the evidence are unambiguous. The diagonalizations of the data matrix $\mathbf{X}'\mathbf{X}$ and the prior covariance matrix that some of us implicitly perform may, however, lead to very poor approximations to the posterior mean. Qualitative and quantitative summaries of the error of approximation provide one way of assessing the collinearity problem.

The principal implication of collinearity is that data evidence cannot be interpreted in a parameter-by-parameter fashion. The informal use of nondata-based prior information by practicing classical statisticians almost necessarily implies a parameter-by-parameter analysis, and consequently, the data misinterpretation just described. The great benefit of a Bayesian approach is that it provides instruction on how to deal with prior information in a multiparameter problem. For example, the posterior mean is, under suitable assumptions, a *matrix*-weighted average of the prior mean and the sample estimate, not a simple average.

Although the Bayesian approach appropriately spotlights the fundamental source of the collinearity problem—personal prior information—it necessarily leaves the resolution of the problem to the individual. He must "merely" construct his prior distribution. Difficulties in constructing a personal prior and/or variation in opinions among intended readers may cause major difficulties in analyzing and reporting collinear evidence. Thus the problem of collinearity from a Bayesian viewpoint concerns the sensitivity of the posterior distribution to changes in the prior distribution, and quantitative measures of that sensitivity may be used to describe the degree of the problem.

The principal claim of this section is that the collinearity problem concerns the way in which sample evidence fits together with prior information. If prior information dominated sample evidence in all directions, there would be no collinearity problem. When there is a collinearity problem, classical inference, which excludes undominated uncertain prior information, fails as a method of interpreting evidence. Peculiarities in the likelihood surface make the BLUE (least-squares) estimate almost irrelevant. A fuller exploration of the likelihood contours informally directed by prior information is difficult and rarely convincing, especially when the number of dimensions of prior information is more than one. Although a Bayesian approach cannot provide a complete cure, it does indicate the source of the disease.

5.8 Global Sensitivity Analysis: Properties of Matrix-Weighted Averages

The posterior mean of a normal, linear-regression model with normal priors is a matrix-weighted average of a prior location vector and a sample location vector. The prior weight matrix is arbitrary, either because prior distributions are impossible to measure without error or because intended readers may differ in their prior judgements. A Bayesian analysis based on any particular prior distribution, as a result, is of little interest. Practical users of the Bayesian tools necessarily face the difficult reporting problem of characterizing economically the mapping implied by the given data from interesting prior distributions into their respective posterior distributions, thereby servicing a wide readership as well as identifying those features of the prior that critically determine the posterior.

One way of characterizing the mapping from priors into posteriors is a local sensitivity analysis discussed in the previous section that identified the relative sensitivity of aspects of the posterior distribution to infinitesimal changes in the prior. The usefulness of a local sensitivity analysis is somewhat limited, since to have great content it must be performed for many different prior distributions. An alternative is a global sensitivity analysis that constructs a correspondence between classes of priors and classes of posteriors. A correspondence can be constructed by answering questions of the form: "if my prior is a member of this class of priors, what can I say about my posterior?"

Although both the location and the dispersion of the posterior are of interest, we consider here only the location parameter. The location of the prior is taken as given, and a correspondence is developed between classes of prior covariance matrices and regions in the space of the posterior location vector. A great deal can be said about the posterior location without precisely specifying the prior covariance matrix. What is not true, except under special and unlikely circumstances, is that, element by element, the posterior location lies algebraically between the prior location and the sample location. The inappropriateness of this bound is an important reason why a multiparameter problem is fundamentally different from a uniparameter problem.

The three most interesting results of Chamberlain and Leamer (1976) are reported here. The posterior mean, as usual, is written as

$$\mathbf{b}^{**} = E(\boldsymbol{\beta}|\mathbf{Y},\mathbf{H},\mathbf{H}^*) = (\mathbf{H}+\mathbf{H}^*)^{-1}(\mathbf{H}\mathbf{b}+\mathbf{H}^*\mathbf{b}^*) \tag{5.28}$$

where $\mathbf{H}=\sigma^{-2}\mathbf{X}'\mathbf{X}$ and $\mathbf{H}\mathbf{b}=\sigma^{-2}\mathbf{X}'\mathbf{Y}$. The first result is that if only the location vectors $\mathbf{b}$ and $\mathbf{b}^*$ are given, then $\mathbf{b}^{**}$ may lie essentially anywhere. This contrasts with the analogous one-dimensional result, which constrains the scalar b^{**} to lie algebraically between the scalars b^* and b. Next, it is shown that if $\mathbf{X}'\mathbf{X}$ is known as well as $\mathbf{b}$ and $\mathbf{b}^*$, then the posterior location $\mathbf{b}^{**}$ must lie in the feasible ellipsoid (5.1). The third result has already been reported in Theorem 5.5: if $\mathbf{b}$, $\mathbf{b}^*$ and $\mathbf{X}'\mathbf{X}$ are given and if $\mathbf{H}^*$ is a diagonal matrix, then $\mathbf{b}^{**}$ is a weighted average of 2^k constrained least-squares points.

The first result of practical interest is that if only the locations $\mathbf{b}$ and are known, the posterior mean may lie essentially anywhere. The proof requires the following lemma.

LEMMA 5.1. *Given the prior and sample locations,* $\mathbf{b}^*$ *and* $\mathbf{b}$, *and an invertible matrix of common conjugate axes* $\mathbf{B}$ *such that* $\mathbf{B}'\mathbf{H}\mathbf{B}=\rho\mathbf{I}$ *and* $\mathbf{B}'\mathbf{H}^*\mathbf{B}=\mathbf{D}^*$, *with* ρ *an arbitrary positive scalar and* $\mathbf{D}^*$ *an arbitrary positive diagonal matrix, the transformed posterior mean*

$$\mathbf{a}^{**}=\mathbf{B}^{-1}\mathbf{b}^{**}=\mathbf{B}^{-1}(\mathbf{H}+\mathbf{H}^*)^{-1}(\mathbf{H}\mathbf{b}+\mathbf{H}^*\mathbf{b}^*)$$

lies everywhere in the orthotope

$$\left|a_i^{**}-\frac{(a_i+a_i^*)}{2}\right|<\frac{|a_i-a_i^*|}{2}$$

where $\mathbf{a}=\mathbf{B}^{-1}\mathbf{b}$ *and* $\mathbf{a}^*=\mathbf{B}^{-1}\mathbf{b}^*$. *The transformation of this region back into natural coordinates takes the axes into the columns of* $\mathbf{B}$. *Thus the edges of the bound that are axes in transformed coordinates are columns of* $\mathbf{B}$ *in natural coordinates. The resultant bound is a parallelotope with* $\mathbf{b}^*$ *and* $\mathbf{b}$ *at opposite vertices and with edges parallel to the common conjugate axes.*[19]

Conversely, given any point $\mathbf{b}^{**}$ *in this bound, there is a unique (up to a factor of proportionality) diagonal matrix* $\mathbf{D}^*$ *such that* $\mathbf{b}^{**}$ *is a posterior mean. Thus the bound is minimal.*

Proof: This lemma follows trivially by writing

$$\mathbf{a}^{**}=(\mathbf{B}'(\mathbf{H}+\mathbf{H}^*)\mathbf{B})^{-1}(\mathbf{B}'\mathbf{H}\mathbf{B}\mathbf{B}^{-1}\mathbf{b}+\mathbf{B}'\mathbf{H}^*\mathbf{B}\mathbf{B}^{-1}\mathbf{b}^*)$$

$$=(\rho\mathbf{I}+\mathbf{D}^*)^{-1}(\mathbf{a}\rho+\mathbf{D}^*\mathbf{a}^*)$$

where $\mathbf{a}=\mathbf{B}^{-1}\mathbf{b}$, and $\mathbf{a}^*=\mathbf{B}^{-1}\mathbf{b}^*$. Thus, element by element a_i^{**} is a simple weighted average of a_i and a_i^*

$$a_i^{**}=(\rho+d_i)^{-1}(\rho a_i+d_i a_i^*) \tag{5.29}$$

and is constrained to lie in the orthotope described previously.

[19] A parallelotope is an n-dimensional generalization of a parallelogram. It is generated by a pair of points (opposite vertices) and n vectors that define the faces at each of the points. The parallelotope is an orthotope if the vectors are orthogonal to each other.

The converse of this theorem is also true: any point in the orthotope is a posterior mean for some set of eigenvalues d_i. This follows simply by picking ρ and inverting (5.29) to write d_i as a function of a_i^{**}.

THEOREM 5.10 (MATRIX-WEIGHTED AVERAGES CAN LIE ANYWHERE). *Given only the prior and sample locations,* $\mathbf{b}$ *and* $\mathbf{b}^*$, *with* $\mathbf{H}$ *and* $\mathbf{H}^*$ *any symmetric positive definite matrices, the posterior mean may lie on the open-line segment* $(\mathbf{b},\mathbf{b}^*)$ *and anywhere off the line through it. That is, only points on the line through* $\mathbf{b}$ *and* $\mathbf{b}^*$ *exterior to the open line segment* $(\mathbf{b},\mathbf{b}^*)$ *are excepted.*

Proof: Choose any point $\mathbf{b}^{**}$satisfying the foregoing bound, and form a parallelogram in the plane of $\mathbf{b}^{**}$, $\mathbf{b}^*$, and $\mathbf{b}$ that contains $\mathbf{b}^{**}$ and has $\mathbf{b}^*$ and $\mathbf{b}$ at opposite vertices. If we simply choose the edges of this parallelogram as the first two common conjugate axes described in Lemma 1 and further choose $k-2$ additional linearly independent vectors to complete the selection of conjugate axes, by the converse of Lemma 1 there exists a set of d_i such that $\mathbf{b}^{**}$ is a posterior mean. The exception derives from the impossibility of forming such a parallelogram if $\mathbf{b}^{**}$ is on the line through $\mathbf{b}$ and $\mathbf{b}^*$ exterior to the segment $(\mathbf{b},\mathbf{b}^*)$. A tedious algebraic proof is in Leamer (1971).

The following result used to prove Theorem 5.11 is essentially the same as Pratt's (1970); the proof parallels his proof.

LEMMA 2. *Given the sample and prior locations* $\mathbf{b}$ *and* $\mathbf{b}^*$ *and the information that the prior and sample ellipsoids have a complete set of common principal axes (i.e., that the positive definite matrices* $\mathbf{H}$ *and* $\mathbf{H}^*$ *commute,* $\mathbf{H}\mathbf{H}^*=\mathbf{H}^*\mathbf{H}$*), the posterior mean is constrained to lie in the hypersphere*

$$(\mathbf{b}^{**}-\mathbf{c})'(\mathbf{b}^{**}-\mathbf{c})<\frac{(\mathbf{b}-\mathbf{b}^*)'(\mathbf{b}-\mathbf{b}^*)}{4} \tag{5.30}$$

where

$$\mathbf{c}=\frac{(\mathbf{b}+\mathbf{b}^*)}{2}.$$

In words, the posterior mean is constrained to lie in the hypersphere with diameter $[\mathbf{b},\mathbf{b}^*]$.

Conversely, any point in this hypersphere is a posterior mean for some choice of $\mathbf{H}$ *and* $\mathbf{H}^*$ *with* $\mathbf{H}\mathbf{H}^*=\mathbf{H}^*\mathbf{H}$. *Thus the bound is minimal.*

Proof: We may write

$$\mathbf{b}^{**}-\mathbf{b}^*=(\mathbf{H}+\mathbf{H}^*)^{-1}\mathbf{H}(\mathbf{b}-\mathbf{b}^*)$$

$$\mathbf{b}^{**}-\mathbf{b}=(\mathbf{H}+\mathbf{H}^*)^{-1}\mathbf{H}^*(\mathbf{b}^*-\mathbf{b})$$

and

$$(\mathbf{b}^{**}-\mathbf{b}^*)'(\mathbf{b}^{**}-\mathbf{b})=-(\mathbf{b}-\mathbf{b}^*)'\mathbf{H}(\mathbf{H}+\mathbf{H}^*)^{-2}\mathbf{H}^*(\mathbf{b}-\mathbf{b}^*)<0,$$

since the matrix of this quadratic form is positive definite when $\mathbf{H}$ and $\mathbf{H}^*$ commute. After some straightforward rearrangements, this inequality is the same as inequality (5.30).

Conversely, any point in the hypersphere lies in a rectangular solid with vertices at $\mathbf{b}$ and $\mathbf{b}^*$. The edges of such a rectangle can be taken as common conjugate axes, and the converse of this lemma follows trivially from the converse of Lemma 1. For an algebraic proof see Pratt (1970).

The following theorem is of considerable practical interest, since it deals with a typical case when the sample precision matrix is known and the prior precision is completely arbitrary.

THEOREM 5.11 (ELLIPSOID BOUND). *Given the sample and prior locations,* $\mathbf{b}$ *and* $\mathbf{b}^*$, *and the sample precision* $\mathbf{H}$ *up to a scale factor, the posterior mean is constrained to lie in the ellipsoid*

$$(\mathbf{b}^{**}-\mathbf{c})'\mathbf{H}(\mathbf{b}^{**}-\mathbf{c})<\frac{1}{4}(\mathbf{b}-\mathbf{b}^*)'\mathbf{H}(\mathbf{b}-\mathbf{b}^*)$$

where $\mathbf{c}=(\mathbf{b}^*+\mathbf{b})/2$. *In words, the posterior mean must lie everywhere within an ellipsoid from the sample family of ellipsoids with center at the midpoint of the line segment joining* $\mathbf{b}$ *to* $\mathbf{b}^*$ *and with boundary including* $\mathbf{b}$ *and* $\mathbf{b}^*$.

Conversely, any point in this ellipsoid is a posterior mean for some $\mathbf{H}^*$. *Thus the bound is minimal.*

Incidentally, the boundary of this ellipsoid is the set of contrained least-squares points described in Section 5.1 and illustrated in Figure 5.1

Proof: Find the coordinate system that transforms the sample ellipsoid into concentric spheres, $\mathbf{B}'\mathbf{H}\mathbf{B}=\mathbf{I}$. In terms of these coordinates

$$\mathbf{a}^{**}=\mathbf{B}^{-1}\mathbf{b}^{**}=(\mathbf{I}+\mathbf{A})^{-1}(\mathbf{a}+\mathbf{A}^*\mathbf{a}^*)$$

where $\mathbf{A}^* = \mathbf{B}'\mathbf{H}^*\mathbf{B}$, $\mathbf{a} = \mathbf{B}^{-1}\mathbf{b}$, and $\mathbf{a}^* = \mathbf{B}^{-1}\mathbf{b}^*$. By Lemma 2, since $\mathbf{I}$ and $\mathbf{A}^*$ commute, $\mathbf{a}^{**}$ is constrained to the hypersphere

$$(\mathbf{a}^{**} - \mathbf{g})'(\mathbf{a}^{**} - \mathbf{g}) < \frac{1}{4}(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*)$$

where $\mathbf{g} = (\mathbf{a} + \mathbf{a}^*)/2$. Transforming back into natural coordinates we obtain

$$(\mathbf{B}^{-1}\mathbf{b}^{**} - \mathbf{B}^{-1}\mathbf{c})'(\mathbf{B}^{-1}\mathbf{b}^{**} - \mathbf{B}^{-1}\mathbf{c})$$

$$= (\mathbf{b}^{**} - \mathbf{c})'\mathbf{H}(\mathbf{b}^{**} - \mathbf{c}) < \frac{1}{4}(\mathbf{b} - \mathbf{b}^*)'\mathbf{H}(\mathbf{b} - \mathbf{b}^*)$$

where $\mathbf{c} = (\mathbf{b} + \mathbf{b}^*)/2$.

The converse is also true and follows straightforwardly from the converse of Lemma 2 by noting that $\mathbf{A}^*$ is arbitrary.

Knowledge of the sample moment matrix is enough to shrink the bound from essentially complete freedom to a well-defined ellipsoid. We may now consider how various items of information about the prior precision matrix further shrink the bound. One fact we may know is that certain linear combinations of parameters are independent of each other, or equivalently, $\mathbf{H}^*$ may be diagonalized by a known transformation. Theorem 5.5, discussed in Section 5.6.1, implies that the posterior mean is then a weighted average of the 2^k constrained regressions formed by omitting variables in all different combinations.[20] The converse is not true. All weighted averages of the 2^k constrained least-squares points are not necessarily feasible; see Chamberlain and Leamer (1976).

[20]Note, incidentally, that the weighting function w_I in Theorem 5.5 allows us to collapse these 2^k points into a smaller number of points whenever any of the diagonal elements d_i are constrained to be equal. In particular, given k_1 of the diagonals equal to one number, k_2 equal to another..., the 2^k points can be collapsed into $\Pi_j(k_j+1)$ points. The extreme case with all the diagonal elements equal is equivalent to knowing $\mathbf{H}^*$ up to a scale factor, and the resulting minimal bound is the curve decolletage. Theorem 5.5 describes that curve as a convex combination of $k+1$ points.

Theorem 5.5 applies to several familiar models. The exchangeable model of Lindley and Smith (1972) with $\beta_j \sim N(\xi, \sigma_\beta^2)$ and $\xi \sim N(0, \sigma_\xi^2)$ implies a variance matrix for the vector $\boldsymbol{\beta}$ with $\sigma_\xi^2 + \sigma_\beta^2$ on the diagonal and σ_ξ^2 on the off-diagonal. The eigenvalues of this matrix (Rao, 1965, p. 54), are $d_1^{-1} = \sigma_\beta^2 + k\sigma_\xi^2$ with eigenvector $(1, 1, \ldots, 1)$ and $d_j^{-1} = \sigma_\beta^2$ of multiplicity $k-1$ with any set of $(k-1)$ eigenvectors orthogonal to $(1, 1, \ldots, 1)$. Since the eigenvectors are independent of the uncertain parameters (σ_β^2 and σ_ξ^2), there is a known linear transformation that takes the prior variance into a diagonal matrix, and theorem 5.5 applies. The multiplicity of the second eigenvalue implies that the 2^k points can be collapsed into $2k$ points. The constrained least-squares estimates involve one constraint $\Sigma_i \beta_i = 0$ and $k-1$ constraints of the form $\mathbf{l}_j'\boldsymbol{\beta} = 0$ with $\mathbf{l}_j$ a set of eigenvectors orthogonal to the vector of ones. Each of the $2k$ points is a weighted average of constrained least-squares points with or without $\Sigma_i \beta_i = 0$ and with exactly m of the other $k-1$ constraints for $m = 0, 1, \ldots, k-1$. The "ridge regression" special case with $\sigma_\xi^2 = 0$ has only a single distinct eigenvalue, and the number of points is reduced to $k+1$. The limiting degenerate case $\sigma_\xi^2 \to \infty$ has one zero eigenvalue d_1, and all constrained least-squares estimates involving the constraint $\Sigma_i \beta_i = 0$ have zero weight.

Other bounds are discussed in Chamberlain and Leamer (1976). Th point of this section is that it is possible to say interesting things about th posterior distribution when the prior is not fully specified. If only the prio location and sample location are known, then the posterior modes may li essentially anywhere. Knowledge of the sample ellipsoid shrinks the set c feasible points to an ellipsoid; knowledge of the major axes of the prio further restricts the feasible region to the hull of the 2^k restricted leas squares points.

5.9 Identification

If two models imply the same distribution of the data, no observed dat can be said to favor one or the other. The posterior odds, defined as th ratio of the posterior probability of one to the posterior probability of th other, necessarily equal the prior odds ratio. A mathematical translation c this statement is, essentially, the classical definition of the identificatio problem used by Koopman and Riersol (1950). The shortcoming of th definition is that a "model" usually determines a *family* of data distribu tions indexed by some uncertain parameter vector. The generalization c the definition to deal with families of data distributions is not obvious.

The probability model we use to illustrate the concepts is the usu normal linear-regression model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, with $\mathbf{u}$ normally distribute with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$ with σ^2 assumed known. It assumed that the model suffers from the extreme multicollinearity pro lem, $\mathbf{X}\boldsymbol{\eta}_i = \mathbf{0}$, for some set of p linearly independent vectors, $\boldsymbol{\eta}_i$, $i = 1, \ldots,$ Where required, we use a prior for $\boldsymbol{\beta}$ that is normal with mean $\mathbf{b}^*$ ar variance matrix $(\mathbf{H}^*)^{-1}$. It should be emphasized that the following discu sion applies to this linear-regression model, and the definitions and resul do not necessarily generalize to other statistical models.

Example. It is useful to have in mind a more specific example. Suppo that the model included only two explanatory variables, $\mathbf{Y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$ $\mathbf{u}$, with the two variables identical, $\mathbf{x}_1 = \mathbf{x}_2$. In this case, there is only one vector, $\boldsymbol{\eta}' = (1, -1)$, (or any vector proportional to it).

A special case of the model used here has an $\mathbf{X}$ matrix with zero vecto as the first p columns and $k-p$ linearly independent vectors as i remaining columns. Data generated by such a model pretty clearly pr vides evidence only about the last $k-p$ parameters. By a linear transfc mation, any model with p linear dependencies can be taken into this for Let $\mathbf{C}$ be a $k \times k$ invertible matrix with the p vectors $\boldsymbol{\eta}_i$ as its first columns, and write the regression process as

$$\mathbf{Y} = (\mathbf{XC})(\mathbf{C}^{-1}\boldsymbol{\beta}) + \mathbf{u}.$$

Let the vector $\boldsymbol{\theta}=\mathbf{C}^{-1}\boldsymbol{\beta}$ be partitioned $\boldsymbol{\theta}'=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_1$ has p elements. By construction, the first p columns of $\mathbf{XC}$ are zero, and the process can be written

$$\mathbf{Y}=\mathbf{X}_2^*\boldsymbol{\theta}_2+\mathbf{u}, \tag{5.31}$$

where $\mathbf{X}_2^*$ is a $T\times(k-p)$ matrix with linearly independent columns.

Example. The two variable regression model with $\mathbf{x}_1=\mathbf{x}_2$ can be written $\mathbf{Y}=\mathbf{x}_1\beta_1+\mathbf{x}_2\beta_2+\mathbf{u}=\mathbf{x}_1(\beta_1+\beta_2)+\mathbf{u}$, and we are led to conclude that the process produces evidence about $\theta=\beta_1+\beta_2$.

The first pair of definitions are the ones most commonly used in the econometric literature.

Definition. A parameter value $\boldsymbol{\beta}^a$ is *observationally equivalent* to a parameter value $\boldsymbol{\beta}^b$ if the data distributions $f(\mathbf{Y}|\boldsymbol{\beta}=\boldsymbol{\beta}^a)$ and $f(\mathbf{Y}|\boldsymbol{\beta}=\boldsymbol{\beta}^b)$ are identical.

Discussion. Let $\boldsymbol{\beta}^b=\boldsymbol{\beta}^a+\boldsymbol{\eta}$ with $\boldsymbol{\eta}$ chosen such that $\mathbf{X}\boldsymbol{\eta}=\mathbf{0}$. Then $\mathbf{X}\boldsymbol{\beta}^b=\mathbf{X}\boldsymbol{\beta}^a$, and both parameter values determine the same data distribution.

Definition. The regression vector $\boldsymbol{\beta}$ is *identifiable* if there exist no two (distinct) observationally equivalent values of $\boldsymbol{\beta}$. The model is *identified* if its parameter vector is identifiable.

The shortcoming of this definition, as stressed by Kadane (1975), is that it tends to give up too soon. Although a model may fail to be identified, it may, nonetheless, provide a great deal of information about some functions of the parameter. The following is a special case of a definition due to Kadane (1975).

Definition. The linear *function* $\boldsymbol{\psi}'\boldsymbol{\beta}$ is *identified* if, whenever a parameter value $\boldsymbol{\beta}^a$ is observationally equivalent to another parameter value $\boldsymbol{\beta}^b$, it is also true that $\boldsymbol{\psi}'\boldsymbol{\beta}^a=\boldsymbol{\psi}'\boldsymbol{\beta}^b$.

Discussion. The logic of this definition is that although two parameter points may be indistinguishable, we do not, in fact, need to distinguish them, if we are interested in functions that assign the same value to both points.

THEOREM 5.12 (IDENTIFICATION OF $\boldsymbol{\psi}'\boldsymbol{\beta}$). *Given the full set of exact linear dependencies among the columns of* $\mathbf{X}$, $\mathbf{X}\boldsymbol{\eta}_i=\mathbf{0}$, $i=1,\ldots,p$, *the function* $\boldsymbol{\psi}'\boldsymbol{\beta}$ *is identified if and only if* $\boldsymbol{\psi}'\boldsymbol{\eta}_i=0$ *for* $i=1,\ldots,p$.

Proof: It is easily seen that the complete set of values observationally equivalent to $\boldsymbol{\beta}^a$ is $\boldsymbol{\beta}^b=\boldsymbol{\beta}^a+\sum_{i=1}^{p}\boldsymbol{\eta}_i w_i$ for any values of w_i. It is also clear that $\boldsymbol{\psi}'\boldsymbol{\beta}^a=\boldsymbol{\psi}'\boldsymbol{\beta}^b$ for all pairs of vectors in this class if and only if $\boldsymbol{\psi}'\boldsymbol{\eta}_i=0$, $i=1,\ldots,p$.

A word that is very close in spirit to identifiable is the word estimable. The concept of estimable functions is due to Bose (1944) and is discussed in Scheffé (1959). As is now shown, estimable is mathematically equivalent to identifiable.

Definition. The linear combination of parameters $\boldsymbol{\psi}'\boldsymbol{\beta}$ is *estimable* if there exists an unbiased linear estimator of $\boldsymbol{\psi}'\boldsymbol{\beta}$.

Discussion. A linear estimator is $\mathbf{w}'\mathbf{Y}$, where $\mathbf{w}$ is a vector of constants. It is an unbiased estimator of $\boldsymbol{\psi}'\boldsymbol{\beta}$ if, for all $\boldsymbol{\beta}$, $E(\mathbf{w}'\mathbf{Y}|\boldsymbol{\beta})=\boldsymbol{\psi}'\boldsymbol{\beta}$, that is, if $\mathbf{w}'\mathbf{X}\boldsymbol{\beta}=\boldsymbol{\psi}'\boldsymbol{\beta}$. This condition is satisfied for all $\boldsymbol{\beta}$ if and only if $\mathbf{w}'\mathbf{X}=\boldsymbol{\psi}'$.

THEOREM 5.13. *Given the full set of exact linear dependencies among the columns of* $\mathbf{X}$, $\mathbf{X}\boldsymbol{\eta}_i=\mathbf{0}$, $i=1,\ldots,p$, *a necessary and sufficient condition for* $\boldsymbol{\psi}'\boldsymbol{\beta}$ *to be estimable is that* $\boldsymbol{\psi}$ *and* $\boldsymbol{\eta}_i$ *be orthogonal,* $\boldsymbol{\psi}'\boldsymbol{\eta}_i=0$, *for all* i.

Proof: Post-multiplying the condition $\mathbf{w}'\mathbf{X}=\boldsymbol{\psi}'$ by $\boldsymbol{\eta}_i$ produces the equality $0=\boldsymbol{\psi}'\boldsymbol{\eta}_i$. This establishes the necessity of $\boldsymbol{\psi}'\boldsymbol{\eta}_i=0$. A constructive proof of the sufficiency of the condition is useful. We can make use of the same notation as used in Equation (5.31). Observe that $\hat{\boldsymbol{\theta}}_2=(\mathbf{X}_2^{*\prime}\mathbf{X}_2^*)^{-1}\mathbf{X}_2^{*\prime}\mathbf{Y}$ is an unbiased estimator of $\boldsymbol{\theta}_2$. Let $\hat{\boldsymbol{\theta}}$ have arbitrarily chosen values for its first p elements and have $\hat{\boldsymbol{\theta}}_2$ for its last $k-p$ elements, $\hat{\boldsymbol{\theta}}'=(\mathbf{a}',\hat{\boldsymbol{\theta}}_2')$. An estimator of $\boldsymbol{\psi}'\boldsymbol{\beta}$ is $\boldsymbol{\psi}'\mathbf{C}\hat{\boldsymbol{\theta}}$ with expected value

$$E\boldsymbol{\psi}'\mathbf{C}\hat{\boldsymbol{\theta}}=\boldsymbol{\psi}'\mathbf{C}\begin{bmatrix}\mathbf{a}\\ \boldsymbol{\theta}_2\end{bmatrix}=\boldsymbol{\psi}'\mathbf{C}\boldsymbol{\theta}+\boldsymbol{\psi}'\mathbf{C}\begin{bmatrix}\mathbf{a}-\boldsymbol{\theta}_1\\ \mathbf{0}\end{bmatrix}$$

$$=\boldsymbol{\psi}'\boldsymbol{\beta}+\boldsymbol{\psi}'\mathbf{C}\begin{bmatrix}\mathbf{a}-\boldsymbol{\theta}_1\\ \mathbf{0}\end{bmatrix}.$$

The last term in this expression is zero if the first p elements of $\boldsymbol{\psi}'\mathbf{C}$ are zero, that is, if $\boldsymbol{\psi}'\boldsymbol{\eta}_i=0$; and then $\boldsymbol{\psi}'\mathbf{C}\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\psi}'\boldsymbol{\beta}$.

Example. In the two-variable model with $\boldsymbol{\eta}=[1,-1]$, the function $\beta_1+\beta_2$ is estimable.

A feature of a set of observationally equivalent parameter values is that they are all assigned the same value by the likelihood function regardless

of the data $\mathbf{Y}$. The likelihood function is

$$L(\boldsymbol{\beta};\mathbf{Y}) \propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right].$$

Given two observationally equivalent values, $\boldsymbol{\beta}^a$ and $\boldsymbol{\beta}^b$, it is easy to see that $L(\boldsymbol{\beta}^a;\mathbf{Y}) = L(\boldsymbol{\beta}^b;\mathbf{Y})$ regardless of the value of $\mathbf{Y}$. This suggests some alternative definitions.

Definition. A parameter value $\boldsymbol{\beta}^a$ is *observationally equivalent* to a parameter value $\boldsymbol{\beta}^b$ if the likelihood function $L(\boldsymbol{\beta};\mathbf{Y})$ satisfies $L(\boldsymbol{\beta}^a;\mathbf{Y}) = L(\boldsymbol{\beta}^b;\mathbf{Y})$ for all $\mathbf{Y}$.

Definition. The model is said to be *identified* if the likelihood function attains its maximum at a single point.

Example. Given the two-variable regression model with $\mathbf{x}_1 = \mathbf{x}_2$, the likelihood function is maximized along the line $(\beta_1+\beta_2) = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{Y}$ (see Figure 5.15).

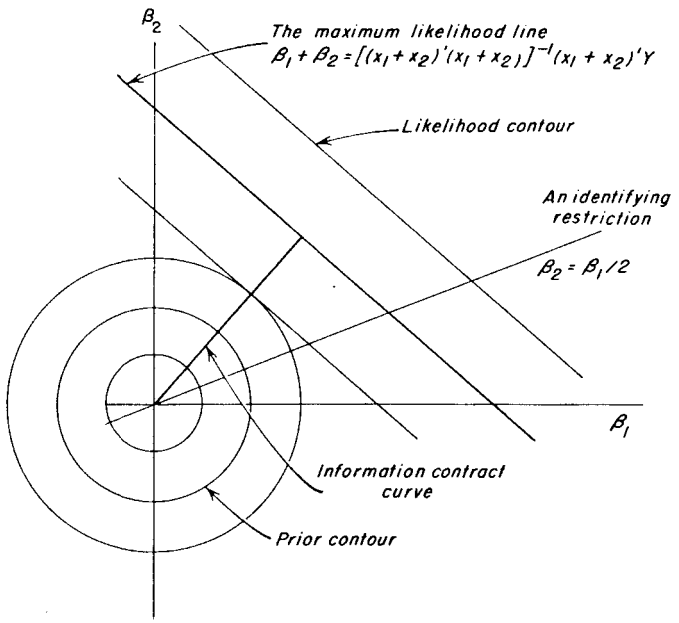

Fig. 5.15 The likelihood function of an unidentified model.

We now consider the role that prior information plays in identifying a model. We first consider exact linear restrictions and then probabilistic prior information.

Definition. A set of q restrictions $\mathbf{R}\boldsymbol{\beta}=\mathbf{r}$ where $\mathbf{R}$ is a $q\times k$ matrix of rank q is set of *identifying restrictions* if on the set $\{\boldsymbol{\beta}|\mathbf{R}\boldsymbol{\beta}=\mathbf{r}\}$ the likelihood function attains its maximum at a single point. The set of restrictions is said to *overidentify* the model if a subset of the restrictions identifies the model.

Discussion. Using the notation of Equation (5.31), the set of restrictions $\boldsymbol{\theta}_1=\mathbf{a}$ is a set of identifying restrictions (see Figure 5.15).

THEOREM 5.14. *Given a set of identifying restrictions any function $\psi'\boldsymbol{\beta}$ is identified and (equivalently) estimable.*

Proof: Left to reader.

Dreze (1962) and (1975) comments that exact restrictions are unlikely to hold with probability one and suggests using probabilistic prior information. Whereas a model is defined before to include the restrictions, a model is now defined to include any probabilistic prior information:

Definition. A model is said to be *identified in probability* if the posterior distribution for $\boldsymbol{\beta}$ is proper regardless of the data $\mathbf{Y}$.

Discussion. It is enough to have a proper prior distribution for $\boldsymbol{\theta}_1$ to have a proper posterior distribution for $\boldsymbol{\theta}$ and hence for $\boldsymbol{\beta}$. If the prior is proper for $\boldsymbol{\beta}$, any model is identified in probability. This leads one to conclude erroneously, that the concept of identification is uninteresting to a Bayesian. More discussion follows.

Another observation due to Dreze (1962) is that if it is known in advance of the data $\mathbf{Y}$ that the likelihood function will necessarily be uniform on some subspace, then, conditional on this subspace, the prior and posterior distributions will coincide:

Definition. A model is not identified if there exist a set of restriction $\mathbf{A}\boldsymbol{\beta}=\mathbf{a}$ such that $f(\boldsymbol{\beta}|\mathbf{A}\boldsymbol{\beta}=\mathbf{a}) = f(\boldsymbol{\beta}|\mathbf{A}\boldsymbol{\beta}=\mathbf{a},\mathbf{Y})$ for all $\mathbf{Y}$.

Discussion. The reader may verify that $f(\boldsymbol{\beta}|\boldsymbol{\theta}_2) = f(\boldsymbol{\beta}|\boldsymbol{\theta}_2,\mathbf{Y})$ regardless of the prior $f(\boldsymbol{\beta})$.

Without saying so explicitly, these definitions of the identification problem describe deficiencies in the information afforded by the experiment. The concept of estimability, however, does make explicit reference to the sample information. An alternative is to compare the prior distribution with the posterior distribution:

Definition. Given a particular prior distribution, the experiment is *personally uninformative* about a linear function $\psi'\beta$ if the posterior distribution of $\psi'\beta$ is equal to the given prior distribution for all data values $\mathbf{Y}$.

Example. If $\mathbf{H}^*$ is the identity matrix and $\eta=(-1,1)$, then from Section 5.7 the data is personally uninformative about $\eta'\mathbf{H}^*\beta=\beta_2-\beta_1$. Notice in Figure 5.15 that the information contract curve is $\beta_2-\beta_1=0$, regardless of the data $\mathbf{Y}$. Notice also that this curve changes if you choose a different prior metric, $\mathbf{H}^*$.

Discussion. With $\mathbf{H}^*$ as the prior precision matrix, the experiment is personally uninformative about $\eta_i'\mathbf{H}^*\beta$, $i=1,\ldots,p$ and any linear combination of these (see Section 5.7), but the experiment is personally informative (i.e., contains information) about all other linear combinations. Notice the sharp contrast between this notion and the notion of identifiability. Given some linear dependencies among the columns of $\mathbf{X}$, almost all functions are unidentified. Nonetheless, the data are personally informative about almost all functions. The following definition *is* the analogue of identifiable.

Definition. The experiment is *publicly informative* about the linear combination $\psi'\beta$ if there exists no positive definite, prior precision matrix $\mathbf{H}^*$ such that the experiment is personally uninformative about $\psi'\beta$.

THEOREM 5.15. *Given the full set of exact linear dependencies among the columns of* $\mathbf{X}$, $\mathbf{X}\eta_i=0$, $i=1,\ldots,p$, *a necessary and sufficient condition for the experiment to be publicly informative about* $\psi'\beta$ *is* $\psi'\eta_i=0$ *for all* i.

Proof: The experiment is personally uninformative about the function $\sum_i w_i\eta_i'\mathbf{H}^*\beta$, for arbitrary w_i. The experiment is publicly informative about $\psi'\beta$ if we cannot find a positive definite $\mathbf{H}^*$ and constants w_i such that $\sum_i w_i\eta_i'\mathbf{H}^*=\psi'$. Write this equation as $(\mathbf{H}^*)^{-1}\psi=\sum w_i\eta_i$, and postmultiply it by ψ', $\psi'\mathbf{H}^{*-1}\psi=\sum w_i\psi'\eta_i$. Given the conditions $\psi'\eta_i=0$, this equation amounts to $\psi'(\mathbf{H}^*)^{-1}\psi=0$ which cannot hold for positive definite $\mathbf{H}^*$. Thus $\psi'\eta_i=0$, $i=1,\ldots,$p is a sufficient condition. Conversely, suppose there is a η such that $\mathbf{X}\eta=0$ and $\psi'\eta\neq 0$. Then let c_j, $j=2,\ldots,k$ be a set of orthonormal vectors orthogonal to ψ, and let $\mathbf{H}^*=\psi\psi'(\psi'\eta)^{-1}+\sum_j c_jc_j'$.

Observation of an experiment is personally valuable in the sense that i changes the observer's opinions. Observation of an experiment is sociall valuable in the sense that it leads to a consensus, thereby eliminating th need for the unresolvable argument over whose prior is best. The last set c definitions are aimed at this social-learning phenomenon.

Definition. An experiment *leads to a consensus* if, given a sufficie number of independent replications of the experiment, all observers wi nondogmatic priors have essentially the same posterior distribution, r gardless of their prior.

Definition. An experiment *leads to consensus* about the linear functi $\psi'\beta$ if, given a sufficient number of independent replications of t experiment, all observers with nondogmatic priors have essentially t same posterior distribution for $\psi'\beta$.

THEOREM 5.16. *An experiment leads to consensus if and only if the mo is identified.*

THEOREM 5.17. *An experiment leads to consensus about* $\psi'\beta$ *if and o if the function* $\psi'\beta$ *is identified.*

Proof: Left to reader.

An experiment cannot lead to consensus if it is impossible to distingu one prior from another. The following definition is due to Zellner (197

Definition. A prior distribution $f^a(\beta)$ is observationally equivalent another prior distribution $f^b(\beta)$ if the marginal data distributions $f^a(\mathbf{Y}$ $\int f(\mathbf{Y}|\beta)f^a(\beta)d\beta$ and $f^b(\mathbf{Y})=\int f(\mathbf{Y}|\beta)f^b(\beta)d\beta$ are identical.

Observation. Translate the location of a prior $\mathbf{b}^*$ to $\mathbf{b}^*+\eta$ where satisfies $\mathbf{X}\eta=0$ to construct an observationally equivalent prior.

In summary, the words identifiable, estimable, and publicly informat and the phrase "leads to a consensus" are interchangeable. If a mo implies a likelihood function that attains its maximum on a (linear) se points, the model is not identifiable, and, conditional on that set of po (or certain other sets), the prior and posterior distribution coincide.

The concept of personal informativeness is quite different from concept of public informativeness. No individual would want to disc the information generated by a model just because the model is identified, or even because the linear function of interest is not identif Given this prior information, the model may, nonetheless, imply us information.

5.10 Examples

Two examples of a sensible Bayesian analysis of data are reported in this section. The adjective "Bayesian" refers to the admission that subjective nonsample information is used to interpret the data. The modifier "sensible" refers to the fact that it is unlikely that anyone could with confidence select a particular prior, and as a result we explore the implications of many different prior distributions. These results were computed by a program entitles SEARCH: *S*eeking *E*xtreme and *A*verage *R*egression *C*oefficient *H*ypotheses, which is available on request.

SEARCH assumes that the prior distribution is built in three steps. The prior location is first selected, then the prior "metric" (isodensities), and lastly a particular density value is assigned to each of the isodensity surfaces. If only the prior location is known, the data support an ellipsoid of estimates described alternatively by Theorem 5.11 as a hull of contract curves or by Theorem 5.1 as a set of constrained least-squares estimates. The choice of prior isodensity surfaces further limits the set of estimates to a curve within this feasible ellipsoid. Lastly, the labeling of the prior isodensities selects from this curve a point or a set of points as posterior modes.

SEARCH describes the ellipsoid of constrained estimates in terms of extreme values of coefficients of interest. The information contract curve is described in terms of the "rotation invariant average regressions" (Theorem 5.8) and also in terms of a set of points on the curve.

The priors we are about to discuss are not informative on all the coefficients, and the feasible ellipsoid is suitably adjusted. The prior has implicit in it a set of q uncertain constraints, $\mathbf{R}\boldsymbol{\beta}=\mathbf{r}$. The feasible set of estimates can be described as the set of constrained regressions subject to constraints $\mathbf{MR}\boldsymbol{\beta}=\mathbf{Mr}$, with $\mathbf{R}$ and $\mathbf{r}$ taken as given and $\mathbf{M}$ free to vary. If $\mathbf{R}$ is invertible, then as $\mathbf{M}$ is varied any point on the "feasible" ellipsoid of constrained estimates is attainable. But when $\mathbf{R}$ is not invertible, only a subset of constrained estimates is attainable. This is made more clear in the following examples.

DOUBTFUL VARIABLES

It is very common to have a model with a few explanatory variables that are known to belong in the equation and a longer list of "doubtful" explanatory variables. The first set of variables is likely to be the focus of the analysis, and the second set is used to "control" for other influences. If the list of doubtful variables is long, estimation with all the doubtful variables included in the equation will produce large standard errors on the coefficients of the "focus" variables. In this situation, it is typical to try different subsets of the doubtful variables, and it is hoped that the

coefficients of the focus variables will not change much as the list c doubtful variables is changed. But this search is both haphazard an nonexhaustive. Furthermore, if the coefficients of the focus variable change very much, this ad hoc search does not suggest how to average th many computed estimates into a single number.

SEARCH is ideally suited to deal with this problem. The interestin bounds that the program can report are the extreme estimates of the focu coefficients with ideally chosen doubtful variables included in the equa tion. There is no way of "fiddling" with the doubtful variables to get a estimate outside the reported range. The points on the contract curv reported by the program are mixtures of the 2^q regressions that could b computed using subsets of the q doubtful variables. Thus the program bot searches exhaustively the set of possible regressions and also suggest weighted averages of the regressions, the latter being important when th bounds are wide.

The following example has eight "doubtful" regional dummy variables The dependent variable is the wage rate, and the focus variables are th education of the wage earner, his age, and the square of his age. A dumm variable for a region is necessary if the labor market in the given region i "separated" from the markets in other regions. To say that the dumm variables are doubtful is to say that in the absence of evidence to th contrary, we should view the labor market as a national market.

The estimated model with all the dummy variables included is (standarc errors in parentheses):

$$
\begin{array}{llll}
W = .041\,D_1 & +.098\,D_2 & +.051\,D_3 & -.019\,D_4 \\
\quad (.34) & \quad (.32) & \quad (.46) & \quad (.34) \\
+.004\,D_5 & -.178\,D_6 & +.086\,D_7 & +.060\,D_8 \\
\quad (.46) & \quad (.43) & \quad (.50) & \quad (.35) \\
+.05\ \mathrm{EDUC} & +.137\ \mathrm{AGE} & -.0015\ (\mathrm{AGE})^2 & +5.737 \\
\quad (.030) & \quad (.047) & \quad (.0006) & \quad (.96)
\end{array}
$$

where

D_1 = Mid-Atlantic
D_2 = East North Central
D_3 = West North Central
D_4 = South Atlantic
D_5 = East South Central
D_6 = West South Central
D_7 = Mountain
D_8 = Pacific

(New England omitted)

The bounds for the coefficients of the three focus variables are reported in the table below. The numbers in parentheses are the standard errors of these coefficients if the model that implied the estimate could be taken as given. (Remember that these bounds include regressions subject to constraints such as $\beta_1 = \beta_2$, which says the Mid-Atlantic and East North Central regions can be aggregated. They also include constraints of the form $\beta_i = 0$.)

Bounds for the Focus Coefficients

EDUC	AGE	$(AGE)^2$
.0577 (.0177)	.139 (.029)	−.00147 (.00035)
.0446 (.0178)	.131 (.029)	−.00155 (.00035)

Each of these coefficients is quite insensitive to the choice of regional dummy variables.

Choice of points within these (narrow) bounds requires a more completely specified prior. Suppose that the coefficients of the doubtful variables are thought to be small in the sense that $\sum_{i=1}^{8} \beta_i^2$ is likely to be small. This prior "metric" implies the contract curve incompletely reported in Table 5.1. On the contract curve the extremes of all coefficients occur at the end points. One end point is least squares with all the dummies included; the other is least squares with all the dummies excluded. The extremes for the focus variables are:

EDUC	AGE	$(AGE)^2$
.0521	.1332	−.001489
.0502	.1336	−.001535

These bounds are almost points and it hardly seems necessary to select a particular point on the contract curve. But notice from Table 5.1 that the equation with the dummies omitted has a low likelihood ratio (equivalently a large F) and the data have a distinct preference for an estimate close to the unconstrained least-squares points.

Table 5.1

Points on Contract Curve

Likelihood Ratio	EDUC	AGE	$-(AGE)^2$
.14	.0521	1.33	.00148
.31	.0517	1.34	.00150
.48	.0514	1.34	.00150
.66	.0511	1.35	.00151
.83	.0507	1.35	.00152
1.0	.0502	1.37	.00153

To conclude, for this particular problem the ambiguity in the specifi tion does not translate into substantial ambiguity in the focus coefficien The specification error implies, for example, an interval of estimates the education coefficient from .0446 to .0577. But the sampling standa error of this coefficient in the unconstrained model is .03, which is la compared to the specification range .0577 − .0446 = .0131. To put it brie the sampling error is more important than the specification error.

DISTRIBUTED LAG ESTIMATION

Another common problem in economics is the estimation of distributed processes. Consider the import demand function estimated by ordin least squares

$$\begin{aligned} M_t = {} & .13 Y_t + 2.0 Y_{t-1} - .91 Y_{t-2} \\ & (.42) \quad (.48) \quad (.48) \\ & + .56 Y_{t-3} - .33 Y_{t-4} - .42 P_t - .53 P_{t-1} \\ & (.50) \quad (.39) \quad (.50) \quad (.55) \\ & + .33 P_{t-2} - .72 P_{t-3} + .23 P_{t-4} - .15 + .96 e_{t-1} \\ & (.50) \quad (.51) \quad (.48) \end{aligned}$$

where standard errors are in parenthesis and where

M_t = logarithm (United States imports in the tth quarter divided by a pri index of imports)
Y_t = logarithm (United States GNP in quarter t divided by the GNP pri index)
P_t = logarithm (import price index divided by GNP price index)
t = 1951 first quarter to 1967 fourth quarter

Economists would generally expect to see the coefficients on the inc variables positive and the coefficients on the price variables negative. peculiar saw-tooth pattern of coefficients would be regarded as hig unlikely, and some constraint on the coefficients would undoubtably used to "improve" or to smooth the estimates. One possibility is constrain the coefficients of each of the distributed lag patterns to lie o line. The resulting estimates are

$$\begin{aligned} M_t = \alpha + {} & .83 Y_t + .55 Y_{t-1} + .27 Y_{t-2} - .01 Y_{t-3} \\ & - .29 Y_{t-4} - .56 P_{t-1} - .38 P_{t-2} - .19 P_{t-3} \\ & - 0 P_{t-4} + .19 P_{t-5}. \end{aligned}$$

Although this constraint does eliminate the wild pattern of coefficient does not produce coefficients that are all the same sign for each varia Perhaps we should constrain them all to be equal, yielding the estima

equation

$$M_t = \alpha + .28 \sum_{\tau=0}^{4} Y_{t-\tau} - .31 \sum_{\tau=0}^{4} P_{t-\tau}.$$

Each of these three estimated equations is appropriate for one extreme form of prior information about the coefficients. Since the researcher in fact holds as his opinions neither one of these three forms of prior information, he may informally mix together the three results. He notes that the sum of the coefficients on the income variable is either 1.29, 1.35, or 1.21, and the sum of the coefficients on the price variable is either -1.04, $-.94$, or -1.29. Neither of these estimates is particularly sensitive to the form of prior information. The shape of the lag distribution does seem to be highly sensitive to the form of the prior, but it does seem that the biggest coefficient is "probably" either the first or second.

The point of much of the discussion in Section 5.6, especially Theorem 5.5, is that a Bayesian can do nothing more than compute sensible weighted averages of constrained estimates. The value of the Bayesian approach is that it provides instruction concerning both the choice of constraints and the choice of sensible weight functions.

The analysis now to be discussed makes use of three different prior distributions. These priors make use of the assumption that $\mathbf{R}\beta$ has spherical normal distribution with $\mathbf{R}$ defined below:

PRIOR 1 SMALL DIFFERENCES

$$\mathbf{R} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}$$

This prior reflects the fact that the first five coefficients are likely to be similar, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$, and the next five coefficients are likely to be similar, $\beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10}$. (This is one of Shiller's (1973) proposals, first proposed by Whitakker and Robinson, 1940).

PRIOR 2

$$\mathbf{R} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

This prior will smooth only the income coefficient pattern.

PRIOR 3

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}$$

This prior will smooth only the price coefficients.

The bounds over all regressions which make use of the prior constraint are reported in Table 5.2. Note, for example, that the prior 1 bound for th sum of the income coefficients is the fairly small interval [1.2, 1.65] whereas the bounds for the individual coefficients are rather wide. Thi indicates that the choice of smoothness prior does not have much impac on the estimated long-run effect but does have a substantial effect on hov that impact is allocated over the individual coefficients.

The set of constrained regressions just discussed includes those that us constraints of the form $\beta_1 = \beta_2$; but the set of constraints also includes th unlikely constraint $\beta_1 - \beta_2 = \beta_7 - \beta_6$. Recall that the set of constraints ar of the form $\mathbf{MR}\beta = 0$ for any $\mathbf{M}$. To avoid constraints that involve jointl the price coefficients and the income coefficients we would have to restric $\mathbf{M}$ to a block diagonal. As it turns out, this is a difficult computationa task, and instead we use priors 2 and 3, which are diffuse, respectively, o the price coefficients and the income coefficients. The prior 2 bound fc the sum of the income coefficients is [1.24, 1.61] which indicates the set c estimates if only the income coefficients are smoothed. But the set c

Table 5.2

Bounds for Coefficients

β_1	β_2	β_3	β_4	β_5	$\Sigma_{i=1}^{5}\beta_i$	β_6	β_7	β_8	β_9	β_{10}	$\Sigma_{i=}^{10}$
Prior 1											
1.15	2.22	.80	1.55	.89	1.65	.81	.69	.92	.63	1.10	–
−.74	.02	−1.42	−.71	−.93	1.2	−1.53	−1.52	−.90	−1.66	−1.17	−2
Prior 2											
1.04	2.09	.66	1.41	.78	1.61	−.13	.44	.45	−.25	.37	−1
−.62	.156	−1.28	−.57	−.82	1.24	−1.22	−.63	−.31	−.33	−.14	−2
Prior 3											
.27	2.12	−.62	.68	−.03	1.46	.12	.04	.39	−.04	.43	–
−.08	1.70	−1.05	.21	−.38	1.35	−.78	−.81	−.31	−.92	−.44	−1

Table 5.3

Ideal Points for Prior 1 (Rotation Invariant Average Regressions)

β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
.28	.28	.28	.28	.28	−.31	−.31	−.31	−.31	−.31	−.15
.50	.45	.27	.14	.07	−.28	−.25	−.26	−.29	−.30	−.15
.61	.59	.23	.55	−.02	−.41	−.24	−.19	−.23	−.19	−.16
.65	.72	.16	−.00	−.05	−.55	−.22	−.12	−.21	−.12	−.16
.65	.86	.06	−.04	−.04	−.68	−.18	−.06	−.22	−.07	−.16
.60	1.03	−.06	−.04	−.03	−.77	−.14	−.02	−.27	−.03	−.16
.50	1.23	−.21	.01	−.04	−.80	−.12	.01	−.34	.01	−.16
.36	1.50	−.45	.15	−.10	−.74	−.18	.09	−.44	.06	−.16
.13	1.96	−.91	.56	−.33	−.42	−.53	.33	−.72	.23	−.15

possible constraints is still too large. It includes the higher order polynomial constraint $(\beta_1-\beta_2)=(\beta_2-\beta_3)$, but it also includes the unlikely constraint $(\beta_1-\beta_2)=-(\beta_2-\beta_3)$. Ideally, we would require the matrix $(\mathbf{MM}')^{-1}$ to be positive, but this too creates computational burdens.

The rotation invariant average regressions for prior 1 are given in Table 5.3. The first point is constrained least squares given all the constraints implicit in the prior. The last point is constrained least squares given *none* of the constraints; that is, it is just the unconstrained least-squares estimate. The intermediate points are weighted averages of constrained least-squares points. The next to last point is a weighted average of all regressions that involve *one* constraint. The next point uses constraints two at a time....

Any point on the contract curve is a weighted average of these average regressions points. A trace of the contract curve risks missing important features of highly variable curves, but the rotation invariant average regressions cannot.

Selected points on the contract curve implied by prior 1 are reported in Table 5.4. There are several observations that can now be made. First of all, observe that the long-run coefficients (the sums) are relatively insensitive to this form of prior information. It simply makes little difference how confident you are that the differences of the coefficients are small. Individual coefficients are in contrast quite sensitive to the precision of the prior. Next, observe that whereas both end points of the contract curve are peculiar, intermediate points have attractively smooth coefficient patterns. Thus although neither extreme form of prior information—diffuse prior or zero difference prior—implies acceptable estimates, "partial" imposition of the prior constraints does yield sensible estimates. Another thing to notice is that the income coefficients are smoothed easily, but the price coefficients resist smoothing. There is pretty clear evidence that the re-

Table 5.4

Selected Points on the Contract Curve, Prior 1

Rel. Like.[a]	β_1	β_2	β_3	β_4	β_5	$\sum_{i=1}^{5}\beta_i$	β_6	β_7	β_8	β_9	β_{10}	$\sum_{i=6}^{10}\beta_i$
.94	.28	.28	.28	.28	.28	1.4	−.31	−.31	−.31	−.31	−.31	−1.5
.96	.50	.48	.25	.14	.08	1.45	−.35	−.26	−.24	−.28	−.26	−1.3
.97	.60	.65	.18	.04	.00	1.47	−.48	−.23	−.16	−.23	−.17	−1.2
.99	.49	1.24	−.24	.05	−.07	1.47	−.74	−.17	.03	−.35	.02	−1.2
1.0	.13	1.95	−.91	.56	−.33	1.4	−.42	−.53	.33	−.72	.23	−1.1

[a]Relative likelihood of the reported point to the maximum likelihood value computed with σ^2 set to s^2.

sponse to the income stimulus is more rapid than the response to the pric stimulus. In fact, the data seem to suggest that other lagged price variable might be added to the equation. Finally, observe that whereas the max mum of the first coefficient reported in Table 5.4 is .60, the maximum ide point in Table 5.3 is .65. Although the value .65 is not attainable, numbe above .60 are attainable, and to some extent Table 5.4 is misleading.

A comparison of Table 5.2 with Tables 5.3 and 5.4 reveals the impo tance of the choice of "metric." Tables 5.3 and 5.4 make use of th assumption that $\boldsymbol{\beta}'\mathbf{R}'\mathbf{R}\boldsymbol{\beta}$ is small, whereas Table 5.2 uses only the assum tion that $\boldsymbol{\beta}'\mathbf{R}'\mathbf{H}^*\mathbf{R}\boldsymbol{\beta}$ is small where $\mathbf{H}^*$ may be any symmetric positiv definite matrix. With a suitable choice of $\mathbf{H}^*$, β_1 may be as large as 1.15 as small as −.74. But Tables 5.3 and 5.4 reveal that if you are willing t restrict $\mathbf{H}^*$ to be proportional to the identity matrix, then β_1 cannot excee .65 nor fall short of .13.

The next step in the analysis is to select a particular point or set points on the contract curve. Formally, this can be done by specifyin completely the prior distribution which has to this point been defined onl in terms of the surfaces on which the density is constant. This is a step tend to resist, since I have a very difficult time finding sensible questior that would reveal with any accuracy my opinions about the density value I do think it makes sense to examine the contract curve in several ways. I this case I note that, with a relative likelihood deterioration only to .97, can get a pattern of coefficients that makes me happy. Incidentally, yo may infer from this last sentence that there are features of my prior I hav not formally used, namely, that the coefficients should not change sign an should decay in absolute value. For this reason, too, I resist form methods for selecting points on the contract curve.