1

Introduction

1.1 PREAMBLE

Stochastics prevail in every walk of life and science, amidst uncertainty, unpredictability, unknowability, and improbability to shear diversity on one hand, and inequality, inexplainability, and incompatibility to impossibility on the other hand. Stochastics distorts transparency of inherent *determinacy* so that the Trojan horses of *statistical inference* are often invoked for drawing valid and interpretable conclusions from pertinent experimental or observational data. Fathoming the depth of stochastics along with the latent chaos or chance elements (i.e., uncontrollable variation) is a challenging task that statistical inference shares intricately with *statistical decision theory* (SDT), for theoretical foundation, and *statistical methodology* for effective applications. In real-life problems, often, there is an additional layer of complexity: *Constraints* (latent or not) of diverse types mar the underlying determinacy (even if that exists), and, hence, there may be a genuine need to sort out the underlying determinacy from its superimposed stochastics; we need to focus on such logical restraints, and incorporate them mathematically as well as statistically in contemplated decision making processes. *Constrained statistical inference* (CSI) has its roots in this complex.

With the ever-expanding horizons of the domain of statistical inference, from the tenuous simple parametric models (such as the binomial, Poisson, normal, Laplace, exponential laws) to more complex *functional parametric* models (such as in many *stochastic processes*), to *semiparametrics* to the omnibus *nonparametrics*, it became necessary to recharge its arsenals with the battery of compatible developments in SDT, and thereby regulate the needed methodology to meet the demand of the vast

interdisciplinary fields of applications. Within this diversity, often, constraints crop up either as (partial, (non-)linear, total, or even implicit) *ordering* on the so-called parameter space(s), or as (linear or not) *inequalities* on contemplated parameters or functionals; they may also show up on the observations or experimental outcomes, termed the *responses*. In this sense, statistical constraints have their abode in the *decision* as well as the *sample* spaces. Albeit, such constrained universes could be portrayed in matching abstractions from the mathematical universe with SDT tones added for stochastic resolutions, for a more down-to-earth, comprehensive treatise of CSI, we shall adopt a pedestrian's path through the alleys of simple parametrics to functional parametrics to the vast domain, *beyond parametrics*. We specifically focus here on the constraints in the form of ordering and inequality restraints of diverse types, and keep in mind the vast field of applications. Perhaps it would be more motivating, and convincing too, to portrait first some simple and yet genuine inequality and order constrained inference problems as illustrative examples, and then step out to a more general prescription for a broader class of problems encountered in CSI. This is advocated in Section 1.2. The final section in this chapter attempts to capture the basic organization and coverage of the present treatise of CSI.

1.2 EXAMPLES

To briefly indicate an important feature of constrained statistical inference, let us first consider the simple example of testing $\mu = 0$ based on a random sample from the univariate normal distribution, $N(\mu, 1)$. Suppose that μ is known to be nonnegative. Then, we know from elementary statistics that a one-sided test with the sample mean as the test statistic is more powerful than a two-sided one. Intuitively, we would expect to do better if we know that $\mu \geqslant 0$ and if such a constraint is incorporated in the statistical inference procedure. We would expect a similar trend in multiparameter situations as well. In other words, if we know that parameters are restricted by some constraints then it is reasonable to expect that we should be able to do better by incorporating such additional information than by ignoring them. While this appears reasonable, it may or may not be an easy task.

The main theme of this book is incorporation of inequality and order restrictions on the parameters of the statistical model, and our objective is to provide a systematic development of the relevant statistical theory, motivated and illustrated by examples. In the rest of this section, we provide a broad range of motivating examples.

Example 1.2.1 *Ordered treatment means in one-way layout*

An experiment was conducted to evaluate the effect of exercise on the age at which a child starts to walk. Let Y denote the age (in months) at which a child starts to walk; the data on Y are given in Table 1.1. (The original experiment consisted of another treatment; however, here we consider only three treatments for simplicity. The data and analyzes involving all four treatments will be given in a later chapter.)

The first treatment group received a special walking exercise for 12 minutes per day beginning at age 1 week and lasting 7 weeks. The second group received daily

Table 1.1 The age at which a child first walks

Treatment (i)	Age (in months)						n_i	$\bar{y}_i$	μ_i
1	9.00	9.50	9.75	10.00	13.00	9.50	6	10.125	μ_1
2	11.00	10.00	10.00	11.75	10.50	15.00	6	11.375	μ_2
3	13.25	11.50	12.00	13.50	11.50		5	12.35	μ_3

Reprinted from: *Science*, Volume 176, P. R. Zelazo, N. A. Zelazo, and S. Kolb, Walking in the newborn, Pages 314–315, Copyright (1972), with permission from AAAS.

exercises but not the special walking exercises. The third group is the control; they did not receive any exercises or other treatments. For Treatment i ($i = 1, 2, 3$), let

$$\mu_i = \text{ Mean age (in months) at which a child starts to walk.}$$

In the traditional analysis of variance (ANOVA), one would usually test

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad \text{against} \qquad H_2 : \mu_1, \mu_2, \textit{ and } \mu_3 \textit{ are not all equal.}$$

However, let us suppose that the researcher was prepared to assume that the walking exercises would not have the negative effect of increasing the mean age at which a child starts to walk, and it was desired that this additional information be incorporated to improve on the statistical analysis. For illustrative purposes, let us suppose that the researcher wishes to incorporate the information $\mu_1 \leqslant \mu_2 \leqslant \mu_3$. In this case, the testing problem is

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ vs } H_1 : \mu_1 \leqslant \mu_2 \leqslant \mu_3 \text{ and } \mu_1, \mu_2, \textit{ and } \mu_3 \textit{ are not all equal.}$$

Clearly, the usual ANOVA where one would test

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ against } H_2 : \mu_1, \mu_2, \textit{ and } \mu_3 \text{ are not all equal,}$$

fails to incorporate the additional information $\mu_1 \leqslant \mu_2 \leqslant \mu_3$. Therefore, one would expect that we should be able to do better than the traditional F-test. In the next chapter, we illustrate a method that is specifically designed for this testing problem; the method to be illustrated, which we shall call the $\bar{F}$-test, is easy to understand and use because it uses essentially the same idea as that which underlies the traditional F-test and modifies it to incorporate additional information such as $\mu_1 \leqslant \mu_2 \leqslant \mu_3$. ■

Let us make a few remarks about this example. If the objective of the experiment was to establish that the special walking exercise results in a reduction in the mean age at which a child starts to walk, the testing problem needs to be formulated differently; for example, it may be formulated as test of

$$H_a : \mu_1 < \mu_2 < \mu_3 \textit{ does not hold} \quad \text{vs} \quad H_b : \mu_1 < \mu_2 < \mu_3.$$

This is a different statistical inference problem and it will be discussed in later examples.

To find out which treatments were effective and which were not would involve multiple tests and require Bonferroni-type adjustment to control overall type-I error; this is also likely to require large samples to achieve moderate power.

A test of H_0 against H_2 in one-way ANOVA, for example the F-test, is also a *valid* procedure for testing H_0 against H_1 as well. Here *valid* means that a 5% level test of H_0 against H_2 is also a 5% level test of H_0 against H_1. This is because the critical value and the size of a test are computed under the null hypothesis, and the null hypothesis for H_0 vs H_1 and for H_0 vs H_2 are the same. Therefore, the validity of the test is not an issue in this example.

The ordering on the μ_is in H_1 is an *assumption*. Even if the true values of $\{\mu_1, \mu_2, \mu_3\}$ do not satisfy the order restrictions in H_1, it is still possible for some tests of H_0 against H_1 to reject H_0. In this case, while rejecting H_0 would be the correct statistical decision, accepting H_1 as true would be a wrong conclusion.

In general, when we apply a test of hypothesis, it is assumed that either H_0 or H_1 is true. For this reason, $H_0 \cup H_1$ can be called the *maintained hypothesis*; in this example, the maintained hypothesis is $\mu_1 \leqslant \mu_2 \leqslant \mu_3$. It will be seen in the next chapter, that tests of H_0 against H_1 can be carried out easily.

Example 1.2.2 *Relationship between El Niño and hurricanes*

El Niño refers to unusually warm ocean currents in the Pacific that appear around Christmas time and last for several months. Monsoon rains in the central Pacific and droughts and forest fires in Indonesia and Australia have been linked to *El Niño*. A hypothesis concerning *El Niño* is the following (Kitchens (1998) page 812):

H : Warm phase of *El Niño* suppresses hurricanes and cold phase encourages.

The data[1] in Table 1.2 provide information on the numbers of hurricanes from 1950 to 1995. In this context, different types of hypothesis testing problems may arise depending on how the question of interest is formulated.

Let us first formulate the basic model as a simple one-way classification. Let *El Niño* be the factor with three levels : Cold ($i = 1$), Neutral ($i = 2$), and Warm ($i = 3$). Let Y_{ij} denote the number of hurricanes and let

$$Y_{ij} = \mu_i + e_{ij},$$

where μ_i is the expected number of hurricanes ($i = 1, 2, 3$).

Let us suppose that the hypothesis H is based on various conjectures and scientific reasoning, and it is of interest to know whether or not there is any evidence against H. In this case, the null and alternative hypotheses can be stated as

$$H_0 : \mu_1 \geqslant \mu_2 \geqslant \mu_3 \quad \text{and} \quad H_1 : \mu_1 \geqslant \mu_2 \geqslant \mu_3 \textit{ does not hold.}$$

A feature of this inference problem is that inequality constraints are present in the null hypothesis; this is different from the usual ANOVA problem where the null hypothesis would typically be of the form $\mu_1 = \mu_2 = \mu_3$.

[1] The authors are grateful to the National Hurricane Center for help with these data.

Table 1.2 Effect of *El Niño* on hurricanes

Y	50	51	52	53	54	55	56	57	58	59	60	61
H	11	8	6	6	8	9	4	3	7	7	4	8
E	c	w	n	w	c	c	n	w	n	n	n	c
Y	62	63	64	65	66	67	68	69	70	71	72	73
H	3	7	6	4	7	6	4	12	5	6	3	4
E	n	n	c	w	n	c	n	n	c	c	w	c
Y	74	75	76	77	78	79	80	81	82	83	84	85
H	4	6	6	5	5	5	9	7	2	3	5	7
E	c	c	w	w	n	n	n	n	w	w	n	c
Y	86	87	88	89	90	91	92	93	94	95		
H	4	3	5	7	8	4	3	4	3	10		
E	w	w	c	c	n	w	w	w	w	c		

Y:year (64 refers to 1964); H: number of hurricanes; E: *El Niño*; c: cold; w: warm; n: neutral.
Source: The original source of these data is the National Hurricane Center. The data in this table were extracted from the table on page 812 in Kitchens (1998).

On the other hand, if the interest is to establish that there is sufficient evidence to support the hypothesis H, then the testing problem is

$$H_0 : \mu_1 > \mu_2 > \mu_3 \textit{ does not hold} \quad \text{vs} \quad H_1 : \mu_1 > \mu_2 > \mu_3 \textit{ holds.}$$

The one-way model $Y_{ij} = \mu_i + e_{ij}$ used in the foregoing two examples is a simple one with no other covariates. If $\boldsymbol{x}_{ij}$ is a vector of covariates, then the model could be of the form[2]

$$Y_{ij} = \mu_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + e_{ij};$$

the hypotheses of interest are still the same. These inference problems come under the general framework of the linear model

$$\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{E}.$$

Within the context of such a linear model, three testing problems of interest are:

1. $H_0 : \boldsymbol{R\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{R\theta} \geqslant \boldsymbol{0}, \boldsymbol{R\theta} \neq \boldsymbol{0}$
2. $H_0 : \boldsymbol{R\theta} \geqslant \boldsymbol{0}$ against $H_1 : \boldsymbol{R\theta} \ngeqslant \boldsymbol{0}$
3. $H_0 : \boldsymbol{R\theta} > \boldsymbol{0}$ *does not hold* against $H_1 : \boldsymbol{R\theta} > \boldsymbol{0}$

where $\boldsymbol{R}$ is a given fixed matrix that does not depend on $\boldsymbol{\theta}$. The first two types of problems will be discussed in Chapters 3 and 4, and the third type will be discussed in later chapters under the topic *sign testing*. ■

[2] Vectors and matrices are shown in bold and the superscript T denotes transpose.

Example 1.2.3 *Comparison of two treatments when the response is multivariate (Pocock et al. (1987))*

Seventeen patients with asthma or chronic obstructive airways disease entered a trial to evaluate an inhaled active drug versus placebo. The experiment was designed as a randomized, double-blind, cross-over trial. The main purpose was to study the possibly harmful effects of drugs on lung mucociliary clearance. Analysis of those results produced no evidence of harm. In addition, standard respiratory function measures were taken at the end of both treatment periods. These were peak respiratory flow rates ($PEFR$), forced expiratory volume (FEV), and forced vital capacity (FVC), the latter two being expressed as a percentage of the predicted value for that patient's age, gender, and height in the normal population. Let us define the following variables for the changes (i.e., drug - placebo) in the three measures:

$$\begin{aligned} X_1 &= \text{ change in FEV,} \\ X_2 &= \text{ change in FVC,} \\ \text{and } X_3 &= \text{ change in PEFR.} \end{aligned}$$

Let

$$\boldsymbol{X} = (X_1, X_2, X_3)^T \quad \text{and} \quad \boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$$

where $\mu_i = E(X_i)$ for $i = 1, 2, 3$. A question of interest was whether the addition of the inhaled drug could further improve respiratory function. For each measure, there was no sign of period or carry-over effects. Therefore, the univariate analyzes of drug versus placebo was performed using paired t-test (see Table 1.3).

The estimated correlation matrix of $(\bar{X}_1, \bar{X}_2, \bar{X}_3)$ reported in Pocock et al. (1987) is

$$\begin{pmatrix} 1 & 0.95 & 0.219 \\ 0.95 & 1 & 0.518 \\ 0.219 & 0.518 & 1 \end{pmatrix}.$$

All three measures showed a mean improvement on active drug but none achieved statistical significance at the 5% level. Thus, Bonferroni correction would lead to a conclusion of no evidence of improvement for the active drug. In any case, such a procedure is not the most powerful method in this setting. What is required is a

Table 1.3 Effect of drug on chronic respiratory disease

Variable	X_1	X_2	X_3
Sample Mean $(i.e. \bar{X}_i)$	7.56	4.81	2.29
$s.e.(\bar{X}_i)$	18.53	10.84	8.51
t-statistic	1.63	1.77	1.11

Reprinted from: *Biometrics*, Vol 43, S. J. Pocock, N. L. Geller, and A. A. Tsiatis, The analysis of multiple endpoints in clinical trials, Pages 487–498, Copyright (1987), with permission from Blackwell Publishing.

multivariate statistical procedure to evaluate the collective evidence of the overall benefit of the drug. Therefore, the null and alternative hypotheses take the form

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{and} \quad H_1 : \boldsymbol{\mu} \gtrsim \mathbf{0},$$

respectively, where

$$\boldsymbol{\mu} \gtrsim \mathbf{0} \text{ means that } \mu_i \geqslant 0 \text{ for every } i \text{ and } \boldsymbol{\mu} \neq \mathbf{0}.$$

It would be helpful to point out an important difference between testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against $H_1 : \boldsymbol{\mu} \gtrsim \mathbf{0}$ and performing the three tests,

$$\begin{aligned} & H_{01} : \mu_1 = 0 \text{ vs } H_{11} : \mu_1 > 0, \\ & H_{02} : \mu_2 = 0 \text{ vs } H_{12} : \mu_2 > 0, \\ \text{and} \quad & H_{03} : \mu_3 = 0 \text{ vs } H_{13} : \mu_3 > 0 \end{aligned}$$

with Bonferroni correction. For illustrative purposes let us suppose that the standard errors, $se(\bar{X}_i)$, in the three columns of Table 1.3 are 1.0, 10.0, and 0.2, respectively. In this case, the three Bonferroni corrected tests would conclude that the effects on FEV and PEFR are significant, but not that on FVC at 5% level of significance. A test of $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against $H_1 : \boldsymbol{\mu} \gtrsim \mathbf{0}$ would also reject H_0 and accept H_1; however, this does not say which of FEV, PEFR, and FVC is significant and which is not. Therefore, the two testing procedures answer different questions.

Note that the parameter space for $\boldsymbol{\mu}$ has $2^3 = 8$ quadrants, and the Hotelling's T^2 looks for departure from the null value $\boldsymbol{\mu} = \mathbf{0}$ into all of the 2^3 quadrants, while a test against $H_1 : \boldsymbol{\mu} \gtrsim \mathbf{0}$ looks for departure into just one quadrant and hence would be more targeted to the specific objective of this study.

In a later chapter we will develop the theory for testing

$$H_0 : \boldsymbol{R\mu} = \mathbf{0} \text{ against } H_1 : \boldsymbol{R\mu} \neq \mathbf{0}, \boldsymbol{R}_1\boldsymbol{\mu} \geqslant \mathbf{0}$$

based on a sample from $N(\boldsymbol{\mu}, \boldsymbol{V})$ where $\boldsymbol{R}_1$ is a submatrix of $\boldsymbol{R}$ and $\boldsymbol{V}$ is unknown. These results can be used to deal with problems similar to those in this example. ■

Example 1.2.4 *Testing the validity of Liquidity Preference Hypothesis [an ordered hypotheses on the mean of a multivariate normal] (Richardson et al. (1992))*

Let $H(r, t)$ denote the rate of return during the time period t to $t + 1$ from a *bill* with *maturity* r (i.e., at time t, the maturity date of the investment/bill is $t + r$). At time t, the *term premium* on a bill with maturity r is defined as

$$\theta_r = E\{H(r, t) - H(1, t)\}$$

where the expectation is taken with respect to the information available to the investors at time t. Thus, the term premium is the additional rate of return that the investors expect at time t for investing in a longer-term bill rather than in a short-term one that would mature at the end of next time point. The *Liquidity Preference Hypothesis*

Table 1.4 Term premia of T-bills with up to 11 months to maturity for Aug/1964-Nov/1990

Average premium		t-statistics for $\theta_r - \theta_{r-1}$		Average premium		t-statistics for $\theta_r - \theta_{r-1}$	
θ_2	0.030	$\theta_3 - \theta_2$	6.78	θ_7	0.070	$\theta_8 - \theta_7$	4.55
θ_3	0.056	$\theta_4 - \theta_3$	-0.12	θ_8	0.088	$\theta_9 - \theta_8$	1.13
θ_4	0.055	$\theta_5 - \theta_4$	3.77	θ_9	0.093	$\theta_{10} - \theta_9$	-3.97
θ_5	0.070	$\theta_6 - \theta_5$	0.80	θ_{10}	0.071	$\theta_{11} - \theta_{10}$	1.36
θ_6	0.073	$\theta_7 - \theta_6$	-0.80	θ_{11}	0.077	$\theta_{11} - \theta_9$	-2.00

(*LPH*), which plays an important role in economics, says that the term premium, θ_r, is a nonincreasing function of r. The basic argument supporting the hypothesis is that longer term bills are more risky than short term ones, and hence investors expect higher return for higher risk. There has been a series of publications in which the validity of this hypothesis was investigated (see Richardson et al. (1992) and the references therein). Table 1.4 provides summary data and some relevant statistics that have been used in the debates concerning the validity of *LPH*.

According to *LPH*, $\theta_r - \theta_s$ is nonnegative for $r > s$. Yet the estimates of $\theta_4 - \theta_3, \theta_7 - \theta_6, \theta_{10} - \theta_9$, and $\theta_{11} - \theta_9$ are negative and the last two are large compared with 5% level critical values from a standard normal distribution (see Table 1.4). This led to interesting debates about the validity of *LPH*. If we were to examine all possible pairwise differences then there is a good chance that some "large" negative estimates of $\theta_r - \theta_{r-1}$ would be observed. In any case, multiple tests of all possible pairwise differences in the θ_rs is not quite the appropriate statistical method for this problem, although the t-statistics in Table 1.4 are useful. The problem needs to be formulated as a test of

$$H_0 : \theta_{11} \geqslant \theta_{10} \geqslant \ldots \geqslant \theta_2 \quad \text{against} \quad H_1 : H_0 \text{ does not hold.} \tag{1.1}$$

As in Example 1.2.2, a feature of this testing problem is that inequalities are present in the null hypothesis.

In a later chapter, we will develop the theory for testing $\boldsymbol{\theta} \geqslant \mathbf{0}$ against $\boldsymbol{\theta} \not\geqslant \mathbf{0}$ based on observations from $\boldsymbol{X}$ where $\boldsymbol{X}$ has the multivariate normal distribution, $N(\boldsymbol{\theta}, \boldsymbol{V})$; $\boldsymbol{V}$ is either known or an estimate of it is available. These results can be used for testing (1.1). A test of (1.1) would say whether or not there is sufficient evidence against

$$H_0 : \theta_{11} \geqslant \theta_{10} \geqslant \ldots \geqslant \theta_2;$$

when there is such evidence against H_0, the test would *not* provide information as to which of the inequalities in the null hypothesis fails to hold. On the other hand, multiple tests on the pairwise differences with Bonferroni adjustment, attempt to identify which of the pairwise differences are positive; thus the later approach attempts to answer a much deeper question, one that is not really required for testing the validity of *LPH*. Therefore, in this example, it would be unnecessary to carry out multiple tests on pairwise differences; a set of multiple tests is likely to have low

power compared with one that is specifically designed for testing H_0 against H_1 in (1.1). If there are a large number of observations, then multiple tests can be used for the testing problem (1.1).

In the foregoing discussions, the problem was formulated as

$$\boldsymbol{X} \sim N(\boldsymbol{\beta}, \boldsymbol{V}) \text{ where } \boldsymbol{\beta} = (\theta_2, \ldots, \theta_{11})^T,$$

and the estimate of $\boldsymbol{\beta}$ was the sample mean of 313 observations on $\boldsymbol{X}$. It is quite possible that in a more elaborate model, $(\theta_2, \ldots, \theta_{11})$ may appear as some parameters in a regression type model. The theory for such problems will also be developed in a later chapter. ■

Example 1.2.5 *Test against a set of one-sided hypotheses in a linear regression model (Wolak (1989))*

Consider the following double-log demand function:

$$\log Q_t = \alpha + \beta_1 \log PE_t + \beta_2 \log PG_t + \beta_3 \log I_t + \gamma_1 D1_t + \gamma_2 D2_t + \gamma_3 D3_t + \epsilon_t,$$

where

Q_t = aggregate electricity demand,
PE_t = average price of electricity to the residential sector,
PG_t = price of natural gas to the residential sector,
I_t = income per capita,
and $D1_t, D2_t$, and $D3_t$ are seasonal dummy variables.

The error terms $\{\epsilon_t\}$ may or may not be independent; for example, they may satisfy an AR(1) process. Prior knowledge suggests that $(-\beta_1, \beta_2, \beta_3) \geqslant (0, 0, 0)$. A typical model selection question that arises is whether or not the foregoing model provides a better fit than the simpler model,

$$\log Q_t = \alpha + \gamma_1 D1_t + \gamma_2 D2_t + \gamma_3 D3_t + \epsilon_t.$$

This would require a test of

$$H_0 : (\beta_1, \beta_2, \beta_3) = (0, 0, 0) \text{ against } H_1 : (-\beta_1, \beta_2, \beta_3) \geqslant (0, 0, 0).$$

Strictly speaking, we should write the alternative hypothesis as $(-\beta_1, \beta_2, \beta_3) \gtrsim (0, 0, 0)$. However, in what follows, we shall always interpret H_0 vs H_1 as H_0 vs $H_1 - H_0$ (i.e., $H_1 \setminus H_0$), so that the two hypothesis are disjoint. A test of

$$H_0 : (\beta_1, \beta_2, \beta_3) = (0, 0, 0) \text{ vs } H_2 : (-\beta_1, \beta_2, \beta_3) \neq (0, 0, 0)$$

is not the best for this testing/model-selection problem, although this is the one used by most standard statistical software packages. ■

Example 1.2.6 *Multivariate ANOVA (Dietz (1989))*

Vinylidene fluoride is suspected of causing liver damage. An experiment was carried out to evaluate its effects. Four groups of 10 male Fischer-344 rats received, by inhalation exposure, one of several dosages of vinylidene fluoride. Among the response variables measured on the rats were three serum enzymes: SDH, SGOT, and SGPT. Increased levels of these serum levels are often associated with liver damage. It is of interest to test whether or not these enzyme levels are affected by vinylidene fluoride. The data are given in Table 1.5 [the authors gratefully acknowledge receiving generous help from the National Toxicology Program of NIEHS regarding these data].

Let $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})^T$ denote the observations on the three enzymes for ith subject ($i = 1, \ldots, 10$) in treatment $j (j = 1, \ldots, 4)$. Let θ_{jk} denote the mean (or median) response for j^{th} treatment (i.e., dose) and k^{th} variable and let $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \theta_{j3})^T$ for $j = 1, \ldots, 4$. Now, one formulation of the null and alternative hypotheses is

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}_3 = \boldsymbol{\theta}_4 \quad \text{and} \quad H_1 : \boldsymbol{\theta}_1 \leqslant \boldsymbol{\theta}_2 \leqslant \boldsymbol{\theta}_3 \leqslant \boldsymbol{\theta}_4.$$

For this type of problem, general procedures for multivariate linear models can be used; these include methods based on likelihood, M-estimators, and ranks. The latter two methods have some robustness properties against nonnormal error distributions compared with the normal theory likelihood approach, similar to the case in the location scale model. ■

Table 1.5 Serum enzyme levels in rats

		Rat within dosage									
Dosage	Enzyme	1	2	3	4	5	6	7	8	9	10
0	SDH	18	27	16	21	26	22	17	27	26	27
	SGPOT	101	103	90	98	101	92	123	105	92	88
	SGPT	65	67	52	58	64	60	66	63	68	56
1500	SDH	25	21	24	19	21	22	20	25	24	27
	SGOT	113	99	102	144	109	135	100	95	89	98
	SGPT	65	63	70	73	67	66	58	53	58	65
5000	SDH	22	21	22	30	25	21	29	22	24	21
	SGOT	88	95	104	92	103	96	100	122	102	107
	SGPT	54	56	71	59	61	57	61	59	63	61
15000	SDH	31	26	28	24	33	23	27	24	28	29
	SGOT	104	123	105	98	167	111	130	93	99	99
	SGPT	57	61	54	56	45	49	57	51	51	48

Serum enzyme levels are in international units/liter; dosage of vinylidene in parts/million.

Reprinted from: Thirteen-week study of Vinylidene fluoride in F344 male rats, with permission from The National Toxicology Program, NIEHS; the data appeared in a report prepared by Litton Bionetics Inc. (1984). These data also appeared in Dietz (1989), where they were used to illustrate constrained inference.

Just as in the linear regression model, inequality constrained hypotheses arise in more general models such as logistic regression models, proportional hazards models, loglinear models for categorical data, Generalized Linear Models, time-series models, and threshold models. Some examples of this type are provided below.

Example 1.2.7 *Test against an ordered hypothesis on the regression parameters of a binomial response model (Piegorsch (1990), Silvapulle (1994))*

An assay was carried out with the bacterium *E. coli* strain 343/358(+) to evaluate the genotoxic effects of 9-aminoacridine (9-AA) and potassium chromate (KCr). The data are presented in Table 1.6.

An objective of the study was to evaluate the synergistic effects of KCr when it is administered simultaneously with 9-AA. The null model of interest is known as the *Simple Independent Action* (SIA) model; the reasons for choosing this particular model for the null hypothesis are discussed in Piegorsch (1990) and the references therein. They show that the complementary loglinear model,

$$-\log(1 - \pi_{ij}) = \mu + \alpha_i + \beta_j + \eta_{ij}$$

with $\alpha_1 = \beta_1 = \eta_{1j} = \eta_{i1} = 0$ for every (i, j), provides an appropriate framework to test for/against *synergism.* When $\eta_{ij} = 0\ \forall (i, j)$, the SIA hypothesis holds. If η_{ij} are nonnegative and at least one of them is positive, then there is departure away from SIA towards synergism. Thus, the null and the alternative hypotheses take the forms

$$H_0 : \boldsymbol{\eta} = \mathbf{0} \quad \text{and} \quad H_1 : \boldsymbol{\eta} \geq \mathbf{0}, \tag{1.2}$$

respectively.

If we wish to test against any possible departure from the SIA model, we would test $H_0 : \boldsymbol{\eta} = \mathbf{0}$ against $H_2 : \boldsymbol{\eta} \neq \mathbf{0}$. Methods of testing H_0 against H_2 are available in standard text books on the analysis of binomial data. As was discussed in a previous example, a test of $H_0 : \boldsymbol{\eta} = \mathbf{0}$ against $H_2 : \boldsymbol{\eta} \neq \mathbf{0}$ is valid for testing $H_0 : \boldsymbol{\eta} = \mathbf{0}$ against $H_1 : \boldsymbol{\eta} \geq \mathbf{0}$, but is unlikely to be as powerful as one that is specifically developed for testing against $H_1 : \boldsymbol{\eta} \geq \mathbf{0}$.

Table 1.6 Genotoxicity of potassium chromate (KCr) and 9-aminoacridine (9-AA)

	KCr(μm)				
9-AA(μm)	0	1.4	2.9	4.3	5.7
0	0/192	9/192	3/192	1/192	3/192
40	49/102	92/192	19/192	77/192	60/192

Note: 9/192 means 9 positive responses out of 192 trials.

Reprinted from: *Mutation Research*, Vol 171, La Velle, Potasium chromate potentiates frameshift mutagenesis in *E. coli* and *S. Ttyphimurium*, Pages 1-10, Copyright (1986), with permission from Elsevier.

This type of issue arises in other areas. For example, noncarcinogenic toxic reactions to therapeutic agents have been caused by interactions of compounds that were normally safe when given alone (see Piegorsch (1990)). A formulation of this phenomenon also involves sign constraints on the interaction terms similar to (1.2).

In Chapter 4, we will consider general statistical models that include the generalized linear model in which the distribution of y_i belongs to an exponential family and that for some suitable link function, g, we have that

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\eta}$$

where μ_i is the mean of y_i. The loglikelihood functions of such generalized linear models (and some partial likelihoods) are well behaved, for example, they are concave and smooth. Consequently, we can (for example, see Silvapulle (1994)) obtain elegant results for testing

$$H_0 : \boldsymbol{R\eta} = \mathbf{0} \text{ against } H_1 : \boldsymbol{R\eta} \geq \mathbf{0}.$$

This collection of results can be used for the foregoing testing problem in the binomial response model. ■

Example 1.2.8 *Testing that a treatment is better than the control when the responses are ordinal (Grove (1980), Silvapulle (1994))*

The results of a trial involving two treatments for ulcer are given in Table 1.7. The objective is to test the hypothesis that Treatment B is better than Treatment A.

This testing problem can be formulated in several ways. One possible approach is to say that Treatment B is better than Treatment A, if the local odds ratios are all greater than or equal to 1 with at least one of them being greater than one; recall that the local odds ratios are defined as follows:

$$\phi_1 = \pi_{11}\pi_{22}/\{\pi_{21}\pi_{12}\}, \phi_2 = \pi_{12}\pi_{23}/\{\pi_{22}\pi_{13}\}, \text{ and } \phi_3 = \pi_{13}\pi_{24}/\{\pi_{23}\pi_{14}\}$$

where π_{ij} is the probability that the response of a given individual in the ith row falls in column j. Therefore, the null and alternative hypotheses are

$$H_0 : \boldsymbol{\phi} = \mathbf{1} \quad \text{and} \quad H_1 : \boldsymbol{\phi} \geq \mathbf{1}, \tag{1.3}$$

where $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)^T$ and $\mathbf{1} = (1,1,1)^T$. The restriction $\boldsymbol{\phi} \geqslant \mathbf{1}$ is known as the *likelihood ratio order*.

Table 1.7 Response of two groups to treatments to an ulcer

	Response				
Treatment group	Worse	Good	Very good	Excellent	Total
A (Control)	12	10	4	6	32
B (Treatment)	5	8	8	11	32

Reprinted from: *Lancet*, Vol 259, R. Doll and F. Pygott, Factors influencing the rate of healing of gastric ulcers, Pages 171-175, Copyright (1952), with permission from Elsevier.

Table 1.8 The 2 x 2 tables formed by merging adjacent columns of Table 1.7

	1		2		3	
Treatment group	Col 1	Col 2,3,4	Col 1,2	Col 3,4	Col 1,2,3	Col 4.
A	12	20	22	10	26	6
B	5	27	13	19	21	11
Odds ratio :	γ_1		γ_2		γ_3	

Reprinted from: *Biometrics*, Vol 50, M. J. Silvapulle, On tests against one-sided hypotheses in some generalized linear models, Pages 853-858, Copyright (1994), with permission from Blackwell Publishing.

Grove (1980) argues that (1.3) may be a rather stringent formulation of the hypothesis, and suggests the following formulation of the alternative hypothesis:

$$H_1 : (\sum_{j=1}^{q} \pi_{1j})(\sum_{j=q+1}^{C} \pi_{2j})\{(\sum_{j=1}^{q} \pi_{2j})(\sum_{j=q+1}^{C} \pi_{1j})\}^{-1} \geqslant 1, q = 1, \ldots, C-1 \tag{1.4}$$

where $C = 4$ is the number of columns. To interpret this, consider the collection of 2×2 tables (see Table 1.8) that may be formed by merging the adjacent columns of Table 1.7. Let $\gamma_1, \gamma_2, \gamma_3$ denote the odds ratios for the three 2×2 tables. Now the alternative hypothesis in (1.4) is the same as

$$H_1 : \gamma_1 \geqslant 1, \gamma_2 \geqslant 1, \gamma_3 \geqslant 1;$$

this is not as stringent as the H_1 in (1.3). A detailed discussion of inference when the response variable is ordinal is provided in Chapter 6 where other formulations of the alternative hypothesis such as *hazard rate order, continuation odds ratio order,* and *stochastic order* will also be studied. ■

Example 1.2.9 *A group sequential study (Lipids and cardiovascular diseases)*

The National Heart, Lung and Blood Institute (USA) planned a multicenter clinical trial in 1971 to investigate the impact of lowering the blood cholesterol level on the risk for cardiovascular diseases. The study involved 3952 males, between the ages 35 and 60, chosen from a large pool of people having no previous heart problems but having an elevated blood cholesterol level. They were randomly divided into a placebo and a treatment group, the treatment group receiving some drug to lower the cholesterol level to about 160 or less. Based on cost and other medical considerations, a 12-year (1972-1984) study was planned, and the data coordination task was given to the University of North Carolina, Chapel Hill. The primary response variable was the failure time, while there were a large number of explanatory and covariables.

Medical ethics prompted that if at any time during the study, the treatment is judged superior, the trial should be terminated, and all surviving persons should be switched to the treatment protocol. This clearly set the tone for *interim analysis,* looking at the accumulating data set regularly. It was decided to have a multiple one-sided

hypothesis testing setup, in a time-sequential framework: every quarter, there would be a statistical test, one-sided alternative, and if at any time a significance is observed, the trial would be concluded along with the rejection of the null hypothesis of no treatment effect (i.e., the cholesterol level in blood has no impact on heart problems). These test statistics were not independent, nor could they be based on simple normal or exponential laws. So, the first task was to develop a time-sequential hypothesis testing procedure for one-sided alternatives. For some of these details, we refer to Sen (1981, Ch.11) and Sen (1999a). Much of these technicalities have become a part of standard statistical interim analysis, and we shall refer to that again in Chapter 5. The restraints on the parameters (survival functions) were compatible with the medical and statistical considerations, and a conclusive result was announced in the 1984 issue of the *Journal of the American Medical Association.* ■

Example 1.2.10 *Testing for/against the presence of random effects (Stram and Lee (1994))*

Consider the well-known growth curve data set of Potthoff and Roy (1964). The data consists of measurements of the distance (in *mm*) from the center pituitary to the pteryomaxillary fissure for 27 children (11 girls and 16 boys). Every subject has four measurements, taken at ages 8, 10, 12, and 14 years. The object is to model the growth in distance as a function of age and sex. To allow for individual variability in the growth function, the model being considered is

$$y_i = (1, Age, Sex, Age.Sex)\alpha + (1, Age)\beta_i + \epsilon_i$$

where y_i is 4×1 for the 4 measurements of the i^{th} subject, $Age = (8, 10, 12, 14)^T$, $Sex = (a, a, a, a)^T$ where a is 0 for boys and 1 for girls. The parameter $\alpha(4 \times 1)$ captures the fixed effects and $\beta_i(2 \times 1)$ captures the random effect of subject i. Assume that

$$\epsilon_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}), \quad \beta_i \sim N(\mathbf{0}, \Psi)$$

for some 2×2 positive semi-definite matrix Ψ, and that ϵ_i and β_i are independent. More generally, the regression model can be written as

$$y_i = X_i\alpha + Z_i\beta_i + \epsilon_i.$$

In the process of analyzing these data, one would be interested to know whether or not some random effects are present; equivalently, would a smaller model fit as well as a larger one? In this case, the parameter of interest is $\Psi = \text{cov}(\beta_i)$, and the following testing problems are of interest:

1. $H_0 : \Psi = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ vs $H_1 : \Psi$ is positive semi-definite.

2. $H_0 : \Psi = \begin{pmatrix} \psi_{11} & 0 \\ 0 & 0 \end{pmatrix}$ and $\psi_{11} \geqslant 0$ *vs* $H_1 : \Psi$ is positive semi-definite.

If $\Psi = \mathbf{0}$ then there are no random effects. If $\Psi = (\psi_{11}, 0 \mid 0, 0)$ then the individual variability is captured by a random intercept, and hence the regression planes are parallel for different subjects. If Ψ is positive definite then individual variability results in not only a random intercept but also a random slope parameter for Age.

More generally, consider a generalized linear model with linear predictor of the form

$$X_i\alpha + U_i\gamma_i + Z_i\beta_i$$

where γ_i and β_i are random effects, for example

$$\gamma_i \sim N(\mathbf{0}, D_1) \quad \text{and} \quad \beta_i \sim N(\mathbf{0}, D_2).$$

Suppose that we wish to test

$$H_0 : D_1 = \mathbf{0} \quad \text{vs} \quad H_1 : D_1 \text{ is positive semi-definite.}$$

These hypotheses can be tested using Likelihood Ratio and Local Score tests, among others; the theory for this will be developed in Chapter 4. ■

Example 1.2.11 *Sign testing problem: testing whether an identified treatment is the best*

The problem here is to test whether or not a particular treatment, say A, is better than each of K other treatments. To illustrate an example where this arises, let A denote a drug formed by combining drugs B and C; in the medical/pharmaceutical literature, A is known as *combination drug/therapy*. Since A is a combination of drugs, it may have a higher level of risk of adverse effects. Therefore, a standard requirement of regulatory authorities to approve a combination drug, such as A, is that it must be shown that the combination drug is *better* than each of its components (see Laska and Meisner, 1989). Let X_0, X_1 and X_2 be three random variables that characterize the effects of the drugs A, B, and C, respectively. Suppose that the distribution of X_i is associated with a parameter, μ_i, and assume that larger values of μ_i are more desirable. Let

$$\theta_1 = \mu_0 - \mu_1 \text{ and } \theta_2 = \mu_0 - \mu_2.$$

Now, the null and alternative hypotheses are

$$H_0 : \theta_1 \leqslant 0 \textit{ or } \theta_2 \leqslant 0, \quad \text{and} \quad H_1 : \theta_1 > 0 \textit{ and } \theta_2 > 0$$

respectively. In each of the previous examples, the alternative parameter space included its boundary as well. By contrast, in this example, the alternative parameter space does not include the boundary, and the null parameter space is the union of the second, third, and fourth quadrants, and the boundary of the first quadrant. These types of testing problems can be solved using *intersection-union tests.*

In this example, we assumed that there is only one primary variable of interest and hence X_i was univariate. There are clinical trials in which the assessment involves several variables (called, *multiple end points*); for example, interest may be on several symptoms simultaneously. Therefore, X_i and μ_i may be vectors. ■

Example 1.2.12 *Combination drug for low-back pain and spasm (Snappin (1987))*

Dolobid(D) and Flexeril(F) are approved drugs for spasm and pain relief respectively. It was conjectured that the new drug DF, a combination of D and F, is likely to be better. As was indicated in Example 1.2.11, a requirement of the regulatory authorities for the approval of a new combination drug such as DF is that it must be better than its components D and F.

A study was conducted to evaluate the efficacy of DF relative to D only, F only, and Placebo (P). Patients entering the study were randomly assigned to one of the four treatments, DF, D, F, and P. At the end of the study period, each patient provided an evaluation of the treatment on a five-point ordinal scale ranging from Marked Improvement to Worsening of the condition (see Table 1.9). The objective is to test the claim that the combination drug is better than each of the other three treatments.

Let

$$DF \succ F \text{ denote } DF \text{is better than } F;$$

this can be formulated using different odds ratios as in Example 1.2.8. Now, the inference problem may be formulated as test of

$$H_0 : \text{Not } H_1 \quad \text{vs} \quad H_1 : DF \succ F, DF \succ D, \text{ and } DF \succ P.$$

The inference problem can also be formulated in the following simpler form. Let μ_0, μ_1, μ_2, and μ_3 denote parameters that capture the effects of the four treatments in Table 1.9; such parameters can be defined by introducing various parametric assumptions (see Snappin (1987)). Let $\theta_i = \mu_0 - \mu_i$ for $i = 1, 2, 3$. Now, one way of formulating the inference problem is test of

$$H_0 : \theta_1 \leqslant 0, \textit{ or } \theta_2 \leqslant 0, \textit{ or } \theta_3 \leqslant 0, \text{ vs } H_1 : \theta_1 > 0, \theta_2 > 0, \textit{ and } \theta_3 > 0.$$

This problem will be discussed under intersection union tests. ■

Table 1.9 Comparison of Dolobid/Flexeril with Dolobid, Flexeril, and placebo.

	Degree of improvement					
Treatment	Marked	Mod	Mild	No Change	Worse	Total
Dolobid/Flexeril	50	29	19	11	5	114
Dolobid alone	41	28	23	22	2	116
Flexeril alone	38	27	18	27	3	113
Placebo	35	20	23	37	4	119

Example 1.2.13 *Cross-over interaction in a multicenter clinical trial (Ciminera et al. (1993b))*

A multicenter clinical trial was conducted to compare a drug with a placebo; the drug was a treatment to reduce blood pressure (see Table 1.10). This is an unpaired design (i.e., parallel group) within each center, and there were 12 centers. In the context of this example, we say that there is *cross-over interaction* between the treatment and center if the drug is beneficial in at least one center and harmful in at least one other. The question of interest is whether or not there is any evidence of crossover interaction.

The problem may be formulated as follows. Let δ_i denote the effect of the treatment in Center i ($i = 1, \ldots, 12$). Then the null and the alternative hypotheses may take the form

$$H_0 : \delta \geqslant 0 \text{ or } \delta \leqslant 0, \quad \text{and} \quad H_1 : \delta \not\geqslant 0 \text{ and } \delta \not\leqslant 0,$$

respectively. In this formulation, the null hypothesis says that there is no cross-over interaction. It is also possible to interchange the roles of H_0 and H_1 in the testing problem so that the null and alternative hypotheses are H_1 and H_0, respectively. In all of these, it may be of interest to replace δ_i by $\delta_i - \epsilon_i$ for some given ϵ_i so that $\delta_i - \epsilon_i > 0$ corresponds to a treatment effect that is clinically significant; such a modification through the introduction of small ϵ_is helps to accommodate the fact that a cross-over interaction may not be important if it is not clinically significant. ■

Table 1.10 Comparison of a drug with a placebo in a multicenter trial [a]

i	n_{1i}	n_{2i}	df	$\bar{z}_{1i\cdot}$	$\bar{z}_{2i\cdot}$	$\hat{\delta}_i$	ss_i	$\hat{\delta}_i / se(\hat{\delta}_i)$
1	3	3	4	-7.0	-15.3	8.3	176.67	1.53
2	4	6	8	-13.0	-6.0	-7.0	180.00	-2.29
3	7	8	13	-3.0	-15.9	12.9	1492.87	2.33
4	7	8	13	-13.6	-11.0	-2.6	693.71	-0.69
5	6	8	12	-3.0	-11.8	8.8	345.50	3.04
6	7	6	11	2.7	-4.3	7.0	236.76	2.71
7	5	8	11	-2.2	-5.0	2.8	472.80	0.75
8	7	9	14	-8.3	-17.6	9.3	1080.93	2.10
9	8	8	14	-7.4	-12.2	4.8	507.38	1.59
10	8	9	15	-7.4	-6.7	-0.7	711.87	-0.21
11	4	4	6	2.0	-8.5	10.5	519.00	1.60
12	3	3	4	0.7	-12.0	12.7	116.67	2.88
Total	69	80	125				2067.7	

[a] The within center sample means for treatments 1 and 2 are denoted by $\bar{z}_{1i\cdot}$ and $\bar{z}_{2i\cdot}$ respectively; $\hat{\delta}_i = \bar{z}_{1i\cdot} - \bar{z}_{2i\cdot}$; ss_i = within group sum of squares for group i; $se(\cdot)$ = standard error.

Example 1.2.14 *Testing against the presence of ARCH effect in ARCH or ARCH-M models (Silvapulle and Silvapulle (1995) and Beg et al. (2001))*

The AutoRegressive Conditional Heteroscedasticity (ARCH) model, introduced by Engle (1982), is one of the widely used econometric models, particularly in Financial Economics. This example involves two models in this family.

Let y_t denote the rate of return from an investment (portfolio) at time t; here we are thinking of investment in stocks or bonds. The basic form of the ARCH model is

$$y_t = \mu + \beta^T x_t + \epsilon_t, \tag{1.5}$$

$$\epsilon_t | \mathcal{F}_{t-1} \sim N(0, h_t^2), \qquad h_t^2 = \alpha_0 + \psi_1 \epsilon_{t-1}^2 + \ldots + \psi_p \epsilon_{t-p}^2 \tag{1.6}$$

where $\mathcal{F}_{t-1}$ is the collection of information up to time $(t-1)$ and x_t is a vector of exogenous variables. A main feature of this model is that periods of small (respectively, large) random fluctuations in y tend to be clustered. This is a frequently observed phenomenon in many high-frequency financial time series.

Another frequently observed phenomenon is that large random fluctuations in stock price would be seen as periods of high risk and this would in turn affect the stock price as investors demand higher returns for higher risk. To incorporate this, Engle et al. (1987) suggested that the specification in (1.5) be modified as

$$y_t = \mu + \beta^T x_t + \phi h_t + \epsilon_t. \tag{1.7}$$

Note that according to (1.7), the conditional variance h_t^2 has a direct effect on the mean of y_t; in other words, the degree of risk, measured by h_t, affects the level of stock price and rate of return. The model (1.6)-(1.7) is known as the ARCH-in-Mean (ARCH-M) model.

In many empirical applications involving the ARCH and ARCH-M models, a question of basic interest is whether or not the conditional variance h_t^2 is a constant over time; or more generally, are some of the ψ parameters equal to zero. Since h_t^2 cannot be negative, $\psi_1, \ldots, \psi_p$ are all nonnegative. Therefore, if we were to test against the alternative that h_t^2 depends on time, then the testing problem takes the form

$$H_0 : \psi = 0 \text{ against } H_1 : \psi_1 \geqslant 0, \ldots, \psi_p \geqslant 0. \tag{1.8}$$

For (1.5), the method that is usually applied by econometric software packages is a score test that ignores the constraints $\psi \geqslant 0$ in the alternative hypothesis. However, it is possible to improve on this by applying a method that incorporates the constraints. The theory for this will be developed in Chapter 4; one of the methods that will be developed is a simple generalization of Rao's Score test, which we shall call a local score test (see Silvapulle and Silvapulle (1995)). For (1.7), there are some additional theoretical issues that need to be taken care of because a parameter becomes unidentified under the null hypothesis; this is a nonstandard problem (see Beg, Silvapulle and Silvapulle (2001), and Andrews (2001)). ■

Now, we provide a few more examples that fall into the categories of the previous ones.

Example 1.2.15 *Acceptance sampling (Berger (1982))*

Table 1.11 lists specifications for upholstery fabric. Similar specifications are available for other products.

It is required to develop an acceptance sampling scheme for a batch that consists of a large number of units. Suppose that a random sample of n units are to be chosen from a batch and tested. The objective is to decide whether or not to accept the batch after testing the sample of n units. Note that some of the tests would destroy or change the fabric, and therefore once a piece of fabric has been used for one test, it may not be possible to use it for another test. The inference problem can be formulated as test of

$$H_0 : \theta \notin \Theta \quad \text{vs} \quad H_1 : \theta \in \Theta$$

where

$$\Theta = \{\theta : \theta_1 \geqslant 50, \theta_2 \geqslant 6, \theta_3 \leqslant 5, \theta_4 \leqslant 2, \theta_5 \in A_1, \\ \theta_6 \in B_4, \theta_7 \in C_4, \theta_8 \in D_3, \theta_9 \in E_4, \theta_{10} \in P\}.$$

First, suppose that each unit in the sample can be subjected to every test in the table of specifications. Typically, this would be a reasonable assumption if the properties of the product are uniform within each sampling unit and it can be divided into subunits for different tests. In this case, the statistical inference problem reduces to elementary inference on a single population proportion.

A different approach is required if a sampling unit cannot be subjected to all the different tests in the table of specifications. In this case, the *intersection union test*

Table 1.11 Standard specification for woven upholstery fabric–plain, tufted, or flocked

Variable	Minimum standard Requirement	Formulation
Breaking strength	50 pounds	$\theta_1 \geqslant 50$
Tong tear strength	6 pounds	$\theta_2 \geqslant 6$
Dimensional change	5% shrinkage	$\theta_3 \leqslant 5$
	2% gain	$\theta_4 \leqslant 2$
Surface abrasion (heavy duty)	15000 cycles	$\theta_5 \in A_1$
Colorfastness to		
Water	class 4	$\theta_6 \in B_4$
Crocking		
Dry	class 4	$\theta_7 \in C_4$
Wet	class 3	$\theta_8 \in D_3$
Light-40 AATCCF (fading units)	class 4	$\theta_9 \in E_4$
Flammability	Pass	$\theta_{10} \in P$

Source: The original source of the information contained in this table is the Annual Book of ASTM Standards (1978), American Society for Testing and Materials, Philadelphia, part 32, page 717. This table is reproduced from *Technometrics*, Volume 24, R. L. Berger, Multiparameter hypothesis testing and acceptance sampling, pages 295–300.

provides a convenient solution to this problem. Essentially, a test at level-α accepts the batch if *each* specification is passed by a level-α test; there are no Bonferroni type adjustment because this does not come within the framework of what is usually known as multiple testing. ■

Example 1.2.16 *Analysis of* $2 \times 2 \times 2$ *tables (Cohen et al. (1983a))*

A random sample of 15 individuals who have been vaccinated against a certain disease, and another independent random sample of 15 individuals who have not been vaccinated against the same disease were chosen. The individuals were cross-classified with respect to the binary variables, Vaccination (yes, no), Immune Level (low, high), and Infection (well, sick). The data are given in Table 1.12.

Table 1.12 Effect of vaccination and immune level on infection

Infection	Not vaccinated & low immun.	Not vaccinated & high immun.	Vaccinated & low immun.	Vaccinated & high immun.
Well	4	3	3	7
Sick	7	1	3	2

Reprinted from: *Journal of the American Statistical Association*, Vol 78, A. Cohen, C. Gatsonis, and J. Marden, Hypothesis testing for marginal probabilities in a 2 × 2 × 2 contingency table with conditional independence, Pages 920–929, Copyright (1983), with permission from the American Statistical Association.

Assume that vaccination affects the risk of infection only through its effect on the immune level. It is required to test the validity of the claim

H : vaccination improves immune level, which in turn reduces the risk of infection;

the terms *improves* and *reduces* lead to multiparameter one-sided hypotheses. This type of inference problem can be formulated as test of

$$H_0 : \alpha \notin A \text{ or } \beta \notin B \qquad \text{vs} \qquad H_1 : \alpha \in A \text{ and } \beta \in B$$

for some parameters α and β and suitably chosen sets A and B. Cohen et al. (1983a) developed a methodology for this type of problem and illustrated their results using these data. This example will be studied in Chapter 6. ■

Example 1.2.17 *Meta-analysis (Cutler et al. (1991), Follmann (1996a))*

A recent meta analysis summarized the effect of sodium reduction on blood pressure from published clinical trials. An objective of the study was to provide an overall probability statement on the joint effect of sodium reduction on reducing both the systolic blood pressure (X_1) and diastolic blood pressure (X_2).

Table 1.13 provides the data for the meta analysis. In fact, the table provides considerable information that would lend to more extensive analyzes. Let $\boldsymbol{X}$ be

Table 1.13 Randomized trials of sodium reduction

Sample size	Study length (months)	Reduction in urinary sodium	Reduction in BP (mm Hg) SBP (se)	DBP (se)
Cross-over trials; hypertensive subjects				
15	1	98	6.7 (3.76)	3.2 (3.47)
19	1	76	10 (3.06)	5.0 (1.78)
18	1	56	0.5 (1.5)	0.3 (0.8)
12	1	105	5.2 (4.1)	1.8 (3.55)
40	1.5	72	0.8 (1.8)	0.8 (1.55)
20	1	82	8.0 (2.6)	5.0 (1.62)
9	1	76	9.7 (4.33)	5.1 (2.94)
88	2	67	3.6 (0.7)	2.1 (0.4)
Parallel studies with treatment and control groups; hypertensive subjects				
31/31	24	27	1.5 (4.6)	6.9 (2.12)
6/6	2	98	*	6.0 (4.08)
6/6	2	78	*	4.0 (3.30)
10/15	12	53	8.7 (10.23)	6.3 (4.41)
15/19	1.5	117	−1.8 (4.14)	−0.5 (2.37)
15/15	2.3	89	13.3 (5.46)	6.7 (3.07)
18/12	3	171	2.0 (4.96)	−2 (2.84)
44/50	6	58	2.7 (2.2)	3.4 (1.7)
48/52	3	54	5.1 (1.42)	4.2 (0.9)
37/38	6	32	1.1 (2.15)	0.2 (1.41)
17/17	3	59	13.0 (5.99)	1.8 (3.48)
50/53	2	71	5.5 (1.48)	2.8 (0.85)
20/21	12	*	18.3 (4.35)	5.9 (2.91)
Cross-over trials; normotensive subjects				
20	0.5	170	2.7 (2.36)	3.0 (2.03)
113	2	69	0.6 (0.84)	1.4 (1.07)
35	0.9	74	1.4 (0.74)	1.2 (0.93)
31	1	60	0.5 (0.82)	1.4 (0.90)
172	1	130	3.5 (1.25)	1.9 (0.9)
Parallel studies with treatment and control groups; normotensive subjects				
19/19	0.5	117	1.5 (3.32)	1.1 (2.03)
174/177	36	16	−0.1 (1.00)	−0.2 (0.8)

Reprinted from: *Hypertension*, Vol 17, Cutler, Follmann, Elliott, and Suh, An overview of randomized trials of sodium reduction and blood pressure, Pages 27–33, Copyright (1991), with permission from Wolters Kluwer Health.

the bivariate vector that consists of X_1 and X_2. Let the mean of (X_1, X_2) be denoted by (θ_1, θ_2). Now, the statistical inference problem can be formulated as test of $H_0 : (\theta_1, \theta_2) = (0, 0)$ against $H_1 : (\theta_1, \theta_2) \geq (0, 0)$. Different rows in Table 1.13 correspond to studies with different sample sizes. Therefore, we have independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with a common mean $\boldsymbol{\theta}$ but with different covariance matrices. ■

Example 1.2.18 *Distribution-free analysis of covariance (Boyd and Sen (1986))*

The nature of the problem in this example is that there are several treatments and some covariates. We are interested to test for equality against an ordered alternative of the treatment effects. However, there was adequate evidence to suggest that the error distribution is unlikely to be normal and, further, the effects of the covariates on the response variable are unlikely to be linear. Thus, some kind of distribution-free method is required.

A trial was conducted (by A.H. Robins Company) to compare the efficacy of three treatments for the reversal of anesthesia. For confidentiality purposes, the treatments were identified as 1, 2, and 3. The response variable, Y, is *time elapsed from administration of treatment to completion of reversal.* The covariates are *Depth* = "depth of neuromuscular block at time of reversal (time treatment administered)" and *Age* = "the age (in years) of the patient"; the data are given in Table 1.14.

Based on various diagnostics such as plot of the residuals, there is adequate evidence that the distribution of Y at a fixed value of (Depth, Age) is not normal. Further, within each treatment, the dependence of Y on (Depth, Age) is unlikely to be linear. Therefore, the usual ANCOVA based on normal theory linear model is inappropriate.

Table 1.14 Patient response to treatments for reversal of anesthesia

Treatment	*Y*	*Depth*	*Age*	*Treatment*	*Y*	*Depth*	*Age*
1	4.8	17	27				
1	13.2	6	41	2	5.2	61	45
1	5.8	30	27	2	6.6	26	34
1	4.6	40	33	2	2.7	45	41
1	6	30.5	34	2	5.4	45	29
1	2.9	48	21	2	8.2	21	25
1	5.2	50	43	3	16.4	15	41
1	5.6	20	26	3	6.7	6	32
1	3.9	40	36	3	6.7	6	32
1	5.6	24.4	23	3	7.9	15	28
2	6	30	25	3	6	32	46
2	9.6	25	58	3	19.4	0	45
2	15.5	25	55	3	19	15.7	24
2	8.7	56	43	3	2.8	59	23
2	7.9	51	36	3	6.6	26	29
				3	10.4	21	35

Reprinted from: *Journal of the American Statistical Association*, Vol 81, M. N. Boyd and P. K. Sen, Union-intersection rank tests for ordered alternatives in ANOCOVA, Pages 526–532, Copyright (1986), with permission from The American Statistical Association.

The problem may be formulated as follows. Let $F(y - \theta_k|x)$ denote the distribution of Y for Treatment k at *(Depth, Age)* = x. The functional form of F and the form of dependence of a location parameter of Y on x are unknown. The null and alternative hypotheses are $H_0 : \theta_1 = \theta_2 = \theta_3$ and $H_1 : \theta_1 \leqslant \theta_2 \leqslant \theta_3$. Some general results based on ranks can be applied for this type of problems. ■

1.3 COVERAGE AND ORGANIZATION OF THE BOOK

Statistical Decision Theory (SDT) attempts to bring both wings of statistical inference under a common shade. In that process, it often becomes more abstract than what could be more easily routed to fruitful applications. Faced with this dilemma, we plan to treat the estimation theory and hypothesis testing more or less on a complementary basis, although at the end, in Chapter 8, we would attempt to bind them together by SDT strings. Constrained statistical inference (CSI) arises in many areas of applications, and some of these have been illustrated in the preceding section. From pure theoretical perspectives, *optimal* statistical decision procedures generally exist only under very specific regularity assumptions that are not that likely to be tenable in constrained stochastic environments. As such, in the sequel, we shall first trace the rather tenuous domain of CSI problems where finite-sample optimal decisions exist. In a way to get out of this restricted scope of CSI, we would spring up in a beyond parametric scenario and illustrate the variety of competing methodologies that attempt to capture the glimpses of CSI with due emphasis on robustness, validity, and efficacy considerations in a far more general setup (where exact optimality may not percolate). In the rest of this section, we attempt to provide a motivating coverage of the book.

Ours is not the first book or monograph in this specified field. We would like to acknowledge with profound thanks the impetus we have had from the earlier two precursors: Barlow et al. (1972), to be referred to in the sequel as $B4$ (1972), and Robertson et al. (1988), those two being 16 years apart and otherwise sequentially very much related. We reverently regard $B4$ as the foundation of the CSI at a time when it was quite difficult to impart the much needed SDT to support the basic axioms. Sixteen years later, the foundation has been fortified further in Robertson et al. (1988) with annexation of some research published after 1970. In the past 16 years, there has been a phenomenal growth of research literature in CSI, not only covering the groundwork reported in the earlier two volumes, but much more so, in novel areas that have been annexed to statistical inference more intensively during the past two decades. As such, with enthusiasm, we have attempted to bring out an updated coverage of the entire integrated area of CSI, strangely after a lapse of another 16 years!

Given our motivation to integrate the CSI across the parametrics to all the way to beyond parametrics, our task has not been very easy or routine. Based on our contemplated intermediate level of presentation (for the convenience of upper-undergraduate- and lower-graduate-level students as well as researchers in applied areas), we have therefore recast the setup as follows.

In Chapters 2, 3, and 4, we consider the so called exact CSI in the parametric setting with due emphasis on the role of the classical *likelihood principle* (LP) in CSI. Chapter 2 provides an introduction to the traditional order-restricted tests in the one-way classification model. An objective of this chapter is to illustrate that, for comparing several populations in the usual one-way classification, it is easy to use a natural generalization of the traditional F-test to incorporate order restrictions on the population means. It is hoped that these methods will be incorporated into standard statistical software so that they could be used easily. Chapter 3 provides a rigorous theoretical development of the likelihood approach for testing hypotheses about the mean of a multivariate normal and the regression parameter in the normal theory linear model. These results lay the foundation on a firm footing for CSI in general statistical models that may be parametric, semiparametric, or nonparametric. In Chapter 4, we broaden the multivariate normal framework of the previous chapter and consider general parametric models which includes the generalized linear models and nonlinear time-series models. In this setting, very little is known about the exact distribution of statistics (estimators/test statistics); this is to be expected because even when there are no order restrictions on parameters, most of the available results are asymptotic in nature. Chapter 4 is devoted to developing the essentials for large sample statistical inference based mainly on the likelihood when there are constraints on parameters. Chapters 3 and 4 form the core of constrained statistical inference and open the door to a vast literature of current research interest. The remaining chapters assume familiarity with the results in these two chapters.

It would become more evident as we make progress in the subsequent chapters that in CSI, LP is often encountered with some challenges, particularly, in not so simple models, and there could be some alternative approaches that may have greater flexibility in this respect. Chapter 5 brings out this issue with the exploration of the Roy (1953) *union-intersection principle* (UIP), which has been more conveniently adapted in beyond parametrics scenarios, too. Chapters 5, 6, and 7 highlight these developments with genuine nonparametrics as well as some other semiparametric models. Functional parameters as well as more complex statistical functionals arising in survival and reliability analysis and multiple comparisons have been focused in this context. Chapter 8 deals with the essentials of SDT and attempts to tie-up some loose ends in theory that were relegated from earlier chapters to this one. In particular, shrinkage estimation and Bayes tests in CSI are unified in this coverage. Some miscellaneous CSI problems, we found hard to integrate with the presentations in earlier chapters, are relegated to the concluding one. In this way, our contention has been to present the theory and methodology at a fairly consistent and uniform (intermediate) level, albeit, for technical reasons, in latter chapters, we might have some occasional detours into higher level of abstractions, to a small extent. We therefore strive to provide more up-to-date and thorough coverage of the bibliography, particularly the post-1987 era, so that serious and advanced readers could find their desired way out through the long bibliography appended.

2

Comparison of Population Means and Isotonic Regression

The need to compare several treatments/groups/populations with respect to their means or medians or some other location parameters arises frequently in many areas of applications. Typically, we have k populations with means $\mu_1, \ldots, \mu_k$ and independent samples from the k populations. One of the statistical inference problems that often arises in this context is test of

$$H_0 : \mu_1 = \ldots = \mu_k \quad \text{vs} \quad H_2 : \mu_1, \ldots, \mu_k \text{ are not all equal.}$$

Standard methods for this include the F-test in analysis of variance (ANOVA) and nonparametric methods based on ranks. However, when several treatments are being compared, one is usually prepared to assume that certain treatments are not worse than another, for example, $\mu_1 \leq \mu_2 \leq \mu_3$. In this setup, it is of interest to incorporate the prior information $\mu_1 \leq \mu_2 \leq \mu_3$ for the purposes of improving the statistical analysis. In other words, one would like to target the analysis to detect a departure away from $\mu_1 = \mu_2 = \mu_3$ but in the direction of $\mu_1 \leq \mu_2 \leq \mu_3$. The usual unrestricted F-test in ANOVA is not particularly suitable for this because it does not incorporate prior knowledge such as $\mu_1 \leq \mu_2 \leq \mu_3$ and hence it is not specifically targeted to detect departures in any particular direction of (μ_1, μ_2, μ_3). This chapter provides an introduction to an approach for dealing with this type of statistical inference problems.

Most of the results used in this chapter are also discussed in the following chapters under more general settings. Therefore, advanced readers may start with the next chapter and refer back to this chapter when necessary, or read up to Section 2.3 and then turn to the next chapter.

In the statistics literature, a constraint/prior-knowledge of the form $\mu_i \leqslant \mu_j$ is called an *order restriction* or an *inequality constraint* on the parameters. When an

order restriction appears in a hypothesis, as in the foregoing example, it is called an *ordered hypothesis* or more generally a *constrained hypothesis.*

As an example, consider Example 1.2.1 (page 2) on the comparison of two exercise programs with a control. In this example, the response variable is the age at which a child starts to walk. Let μ_i denote the population mean of the response variable for Treatment i, $i = 1, 2, 3$. An inference problem that is of interest in this context is test of

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{vs} \quad H_2 : \mu_1, \mu_2, \text{ and } \mu_3 \text{ are not all equal.}$$

For illustrative purposes let us suppose that the exercise program was expected not to have detrimental effects, and that the researcher was interested to ensure that the prior knowledge, $\mu_1 \leqslant \mu_2 \leqslant \mu_3$, is incorporated in the statistical analysis when testing H_0 against H_2. This chapter introduces what we call the $\bar{F}$-test, a natural generalization of the usual F-test, to incorporate prior knowledge in the form of order restrictions. A classical test known as the $\bar{E}^2$-test is also introduced. For most practical purposes the difference between these two tests appear to be not that large. An objective of this chapter is to illustrate that it is fairly easy to apply these tests to incorporate prior information such as $\mu_1 \leqslant \mu_2 \leqslant \mu_3$. To this end, only the essential results to implement the test are presented here; a detailed study of the distribution theory in a general setting will be presented in the next chapter.

Section 2.1 provides an introduction to test of $H_0 : \mu_1 = \ldots = \mu_k$ against

$$H_1 : \mu_1, \ldots, \mu_k \text{ are not all equal and they satisfy several order restrictions.}$$

Example 1.2.2 indicated a scenario where order restrictions appear in the null hypothesis. These type of problems are discussed in Section 2.2. More specifically, this section develops a test of

$$H_1 : \{\mu_1, \ldots, \mu_k\} \textit{ satisfy several order restrictions} \quad \text{vs} \quad H_2 : \textit{not } H_1.$$

It will be shown that the tests developed in Section 2.1 for H_0 vs H_1 and those in Section 2.2 for H_1 vs H_2 can be implemented by using a simulation approach. Further, only a short computer program is required to implement this simulation approach. Thus, an important conclusion of the first two sections is that test of hypothesis involving $\mu_1, \ldots, \mu_k$ when there are order restrictions in the null or the alternative hypothesis can be implemented easily. Such a simulation approach overcomes the computational difficulties encountered in the classical approach of computing the exact critical/p-values and/or bounds for them. Consequently, the statistical results become clearer and easier to use.

Section 2.3 provides an introduction to *isotonic regression*. The foregoing example involving three treatments in which the treatment means are ordered is an example of topics studied under isotonic regression. The origin of the term isotonic regression is that the basic model used is the *regression* model $y_{ij} = \mu_i + \epsilon_{ij}$, and the parameters $\mu_1, \ldots, \mu_k$ are required to have the *same tone* (*iso tone*) as the order imposed on the treatments, for example,

$$\text{Treatment 1} \preceq \text{Treatment 2} \preceq \text{Treatment 3,}$$

where Treatment 1 $\preceq$ Treatment 2 may be read as "the mean for Treatment 2 is at least as large as that for Treatment 1"; more generally, it could also be read as "Treatment 2 is at least as good as Treatment 1."

The results in Sections 2.1 and 2.2 for testing $H_0 : \mu_1 = \ldots = \mu_k$ against an order restriction are based on the linear model $y_{ij} = \mu_i + \epsilon_{ij}$. Most of these results have natural extensions to the more general linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{E}$ where $\boldsymbol{E}$ has mean zero and covariance $\sigma^2\boldsymbol{U}$, σ is unknown and $\boldsymbol{U}$ is known. These will be discussed in the next chapter.

These have also been extended to incorporate robust tests based on M-estimators and those based on ranks; see Silvapulle (1992b, 1992c, 1985), Puri and Sen (1985), and Geyer (1994). The finite sample distribution theory for normal error distribution will be derived in the next chapter, and the large sample theory will be derived in the following chapter. The results have been extended to bounded influence tests in general parametric models; see Silvapulle (1997c). Section 2.3 provides an introduction to isotonic regression; most of this is related to the results in Section 2.1.

2.1 ORDERED ALTERNATIVE HYPOTHESES

In this section we adopt the same context as that is usually adopted for one-way ANOVA. Suppose that there are k treatments to be compared. Let y_{ij} denote the jth observation for Treatment i, and assume that $y_{ij} \sim N(\mu_i, \sigma^2)$ and that the observations are independent ($j = 1, \ldots, n_i, i = 1, \ldots, k$); see Table 2.1. Later we shall relax the requirement that y_{ij} be normally distributed. In the next subsection, we consider the walking exercise example (see Example 1.2.1) to motivate some of the ideas that underlie the $\bar{F}$-test and to illustrate its application when the number of order restrictions of the form $\mu_i \leqslant \mu_j$ is less than four. Then in the following subsection, we consider the more general case with an arbitrary number of treatments and test of

$$H_0 : \mu_1 = \ldots = \mu_k \quad \text{vs} \quad H_1 : \mu_i - \mu_j \geqslant 0 \text{ for } (i, j) \in B \tag{2.1}$$

where

$$B \subset \{(i, j) : i, j = 1, \ldots, k\}. \tag{2.2}$$

This particular form of the alternative hypothesis allows only pairwise contrasts; more general contrasts will be discussed in the next chapter.

Table 2.1 Normal theory one-way layout

Treatment	Independent observations	Sample mean	Population distribution
1	$y_{11}, \cdots, y_{1n_1}$	$\bar{y}_1$	$N(\mu_1, \sigma^2)$
.	.	.	.
.	.	.	.
k	$y_{k1}, \cdots, y_{kn_k}$	$\bar{y}_k$	$N(\mu_k, \sigma^2)$

2.1.1 Test of $H_0 : \mu_1 = \mu_2 = \mu_3$ Against an Order Restriction

In this subsection, we consider the special case when there are only three treatments and the null hypothesis is $H_0 : \mu_1 = \mu_2 = \mu_3$; this simple case is convenient to introduce the basic ideas. In this special case, there are three possible ordered alternatives that may be of interest in the context of one-way classification; they are

$$(1)\quad \mu_1 \leqslant \mu_2 \leqslant \mu_3, \quad (2)\quad \mu_1 \leqslant \mu_2 \text{ and } \mu_1 \leqslant \mu_3, \quad \text{and} \quad (3)\quad \mu_1 \leqslant \mu_2.$$

For illustrative purposes, we shall consider the first; the results for the second and third are similar and are also stated in this subsection.

Consider Example 1.2.1, and assume that the setting in Table 2.1 holds. If there is no prior knowledge in the form of an order among the treatment means, then the null and alternative hypotheses would be

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{and} \quad H_2 : \mu_1, \mu_2, \text{ and } \mu_3 \text{ are not all equal,}$$

respectively; the notation H_2, rather than H_1, is used here for the alternative hypothesis to suggest that it is "two-sided" or unrestricted. For testing H_0 against H_2, the classical F-test is usually preferred.

The purpose of the study was to test the claim that walking exercises are associated with a reduction in the mean age at which children start to walk. Let us suppose that the alternative hypothesis is

$$H_1 : \mu_1 \leqslant \mu_2 \leqslant \mu_3, \text{ and } \{\mu_1, \mu_2, \mu_3\} \text{ are not all equal.}$$

As indicated in the previous chapter, we shall abbreviate the statement of this hypothesis to

$$H_1 : \mu_1 \leqslant \mu_2 \leqslant \mu_3.$$

More generally, we shall adopt the following convention throughout this book:

Notation: *When we refer to test of "H_0 against H_1," it should be read as "H_0 against $H_1 \setminus H_0$"; in the literature, $H_1 \setminus H_0$ is also written as $H_1 - H_0$.*

In view of this notation, test of $\mu = 0$ vs $\mu \geqslant 0$ and test of $\mu = 0$ vs $\mu > 0$ are exactly the same testing problems, where μ is a scalar.

To test H_0 against H_1, it is perfectly valid to apply the usual F-test for H_0 against H_2 on $(2, \nu)$ degrees of freedom, where ν is the error degrees of freedom. The validity is justified by the fact that both testing problems have the same null hypothesis. However, one would expect that the standard F-test for H_0 against H_2 is unlikely to have good power properties for testing H_0 against H_1 because it does not make use of the additional information $\mu_1 \leqslant \mu_2 \leqslant \mu_3$ and hence is not specifically targeted to detect departures in the direction of H_1.

Recall that the standard F-statistic is defined as

$$F = \{RSS(H_0) - RSS(H_2)\}(k-1)^{-1}/S^2 \tag{2.3}$$

where S^2 is the error mean square, k is the number of treatments, $RSS(H)$ is the abbreviation for Residual Sum of Squares under the hypothesis H,

$$RSS(H_0) = \inf_{H_0} \sum\sum (y_{ij} - \mu_i)^2 = \sum\sum (y_{ij} - \bar{y})^2$$

$$RSS(H_2) = \inf_{H_2} \sum\sum (y_{ij} - \mu_i)^2 = \sum\sum (y_{ij} - \bar{y}_i)^2$$

and $\bar{y}$ is the grand mean. The term $\{RSS(H_0) - RSS(H_2)\}$ in the numerator of F (see (2.3)) is a measure of the discrepancy between the null H_0 and the alternative H_2. The denominator, S^2, acts as a scaling factor so that the null distribution of the test statistic does not depend on the unknown scale parameter σ for the error; S^2 does not contribute to quantify the discrepancy between H_0 and H_2.

Remark: Strictly speaking, $RSS(H_2)$ should have been written as $RSS(H_0 \cup H_2)$. Since H_0 is on the boundary of H_2, this does not affect the value of $RSS(H_2)$.

These observations suggest that to test H_0 against H_1, a reasonable test statistic is obtained by modifying the foregoing F-statistic as follows:

$$\bar{F} = \{RSS(H_0) - RSS(H_1)\}/S^2 \tag{2.4}$$

where

$$RSS(H_1) = \min_{H_1} \sum\sum (y_{ij} - \mu_i)^2 = \sum\sum (y_{ij} - \tilde{\mu}_i)^2$$

and $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3)$ is the point at which the sum of squares $\sum\sum (y_{ij} - \mu_i)^2$ is minimized subject to the constraint in H_1. Thus, $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3)$ is an estimate of (μ_1, μ_2, μ_3) under H_1 and $RSS(H_1)$ is the corresponding sum of squares of the residuals; in fact, since the errors are *iid* as $N(0, \sigma^2)$, it can be shown that $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3)$ is the *mle* of (μ_1, μ_2, μ_3) under H_1 [see Problem 2.3]. Thus, the numerator $\{RSS(H_0) - RSS(H_1)\}$ of $\bar{F}$ is a measure of the discrepancy between H_0 and the restricted alternative H_1, just as the numerator of F is a measure of the discrepancy between H_0 and the unrestricted alternative H_2. The constant $(k-1)^{-1}$ in the numerator of the usual F statistic was not included in the definition of $\bar{F}$, but it could have been included without affecting the essential nature of any results.

To implement the foregoing procedure, we need to compute the restricted estimator $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3)$ under $H_1 : \mu_1 \leq \mu_2 \leq \mu_3$. Since

$$\sum_i \sum_j (y_{ij} - \mu_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i (\bar{y}_i - \mu_i)^2 n_i \tag{2.5}$$

it suffices to minimize $\sum_i (\bar{y}_i - \mu_i)^2 n_i$ subject to $H_1 : \mu_1 \leq \mu_2 \leq \mu_3$ for the purposes of computing $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3)$.

If the alternative hypothesis is any other order restriction, say H_1^*, then the foregoing discussions hold with trivial modifications; the main modification is that $RSS(H_1^*) = \sum\sum (y_{ij} - \tilde{\mu}_i)^2$ where $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3)$ is the point at which $\sum\sum (y_{ij} - \mu_i)^2$ reaches its minimum subject to the constraints in H_1^*. Now we have the following result; this is a consequence of more general results in the next chapter (see (3.25)).

Proposition 2.1.1 *Let the setting be as in Table 2.1 with $k = 3$ and let the null hypothesis be $H_0 : \mu_1 = \mu_2 = \mu_3$. Then, for testing against any order restriction of the form* (2.1) *on page 27, the null distribution of $\bar{F}$ is*

$$pr(\bar{F} \leqslant c \mid H_0) = w_0 + w_1 pr(F_{1,\nu} \leqslant c) + w_2 pr(F_{2,\nu} \leqslant c/2), \quad (c > 0) \tag{2.6}$$

Table 2.2 Formulae for determining $\{w_1, w_2, w_3\}$ in (2.6)

H_1	ρ in (2.7)
$\mu_1 \leqslant \mu_2 \leqslant \mu_3$,	$\rho = -[n_1 n_3/\{(n_1+n_3)(n_2+n_3)\}]^{1/2}$
$\mu_1 \leqslant \mu_2$ *and* $\mu_1 \leqslant \mu_3$	$\rho = [n_2 n_3/\{(n_1+n_2)(n_1+n_3)\}]^{1/2}$.
$\mu_1 \leqslant \mu_2$	1.0 [i.e. $(w_0, w_1, w_2) = (0, 0.5, 0.5)$]

where

$$w_1 = 0.5,\ w_2 = (0.5 - q),\ w_0 + w_1 + w_2 = 1, q = (2\pi)^{-1}\cos^{-1}(\rho) \qquad (2.7)$$

for some ρ. The formulas for determining ρ are given in Table 2.2. ■

We chose the notation $\bar{F}$ for the statistic in (2.4) because it is related to the unrestricted F-ratio and its null distribution is a weighted average of probabilities associated with F distributions. Some authors have used $\bar{F}$ for a different statistic.

It follows from (2.6) that the p-value for the $\bar{F}$-test is given by

$$p\text{-value} = w_1 \text{pr}(F_{1,\nu} \geqslant \bar{f}_{obs}) + w_2 \text{pr}(2F_{2,\nu} \geqslant \bar{f}_{obs}) \qquad (2.8)$$

where $\bar{f}_{obs}$ is the sample value of $\bar{F}$.

Example 2.1.1

Now let us consider the exercise program example discussed at the beginning of this section. Suppose that we are interested to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad \text{vs} \qquad H_1 : \mu_1 \leqslant \mu_2 \leqslant \mu_3.$$

The unrestricted estimate of (μ_1, μ_2, μ_3), obtained by minimizing $\sum(\bar{y}_i - \mu_i)^2 n_i$ is $(\bar{y}_1, \bar{y}_2, \bar{y}_3) = (10.125, 11.375, 12.35)$. Since this estimate satisfies the constraints in H_1, it follows that the estimate of (μ_1, μ_2, μ_3) subject to the constraint in H_1 is also equal to the unconstrained estimate. Therefore, $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3) = (\bar{y}_1, \bar{y}_2, \bar{y}_3) = (10.125, 11.375, 12.35)$ and, by simple substitution, we have that $\bar{F} = 5.978$. Now, let us assume that the setting in Table 2.1 holds. Then, it follows from (2.8) that for testing $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_1 : \mu_1 \leqslant \mu_2 \leqslant \mu_3$, we have that

$$p\text{-value} = 0.5\text{pr}(F_{1,14} \geqslant 5.989) + \{(0.5 - 0.329)\}\text{pr}(F_{2,14} \geqslant 2.989) = 0.028.$$

By contrast, for the unrestricted F-statistic for testing

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ vs } H_2 : \{\mu_1, \mu_2, \mu_3\} \text{ are not all equal}$$

p-value $= \text{pr}(F_{2,14} \geqslant 2.989) = 0.083$. Note that the p-value for the F-test is larger than that for the $\bar{F}$-test. If the alternative hypothesis is $\mu_1 \leqslant \mu_2 \leqslant \mu_3$ and the sample means satisfy the corresponding inequality $\bar{y}_1 \leq \bar{y}_2 \leq \bar{y}_3$, then it can be shown using (2.6) that the p-value for $\bar{F}$ would be smaller than that for F. Of course, it is reasonable to expect that the $\bar{F}$-test, which is specifically targeted for testing against $\mu_1 \leqslant \mu_2 \leqslant \mu_3$, would provide stronger evidence to reject H_0 than the unrestricted F-test when the sample means satisfy $\bar{y}_1 \leq \bar{y}_2 \leq \bar{y}_3$; it is reassuring to note that this, in fact, is the case. ■

In this example, the calculations were simple because there were only two order restrictions and the sample means satisfied the constraints that correspond to those in the alternative hypothesis. If the alternative hypothesis has up to three order restrictions, then closed-form expressions similar to (2.6)-(2.7) are available for computing the p-value. If the number of order restrictions is four or more then the null distribution of $\bar{F}$ is a weighted sum similar to (2.6), but in general it is quite inconvenient to use it for computing the p-value exactly. A simulation approach offers a simple and practically feasible method of computing the p-value sufficiently precisely irrespective of the number of order restrictions, for any error distribution. These are discussed in the next subsection. The derivations of the null distribution of $\bar{F}$ and that of the likelihood ratio test statistic will be provided in the next chapter.

2.1.2 Test of $H_0 : \mu_1 = \ldots = \mu_k$ Against an Order Restriction

Suppose that there are k treatments. Let the independent observations $\{y_{ij}\}$ be as in Table 2.3, where F is a cumulative distribution function. Clearly, μ_i is a location parameter for treatment i, but it does not need to be the mean ($i = 1, \ldots, k$); for example, it could be the median. Let the null and alternative hypotheses be

$$H_0 : \mu_1 = \ldots = \mu_k \quad \text{and} \quad H_1 : \mu_i - \mu_j \geqslant 0 \text{ for } (i,j) \in B, \qquad (2.9)$$

respectively, for some $B \subset \{(i,j) : i, j = 1, \ldots, k\}$. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$ and H denote a hypothesis concerning $\boldsymbol{\mu}$, for example, H can be H_0 or H_1. Now, we define *the residual sum of squares under H* as

$$RSS(H) = \inf_{\mu \in H} \sum\sum(y_{ij} - \mu_i)^2. \qquad (2.10)$$

It is easily verified that this definition leads to the familiar expressions

$$RSS(H_0) = \sum\sum(y_{ij} - \bar{y})^2 \quad \text{and} \quad RSS(H_2) = \sum\sum(y_{ij} - \bar{y}_i)^2 \qquad (2.11)$$

for the residual sum of squares under $H_0 : \mu_1 = \ldots = \mu_k$ and under $H_2 : \mu_1, \ldots, \mu_k$ *are not restricted*, respectively. In this sense, (2.10) is a natural extension of the two

Table 2.3 Comparison of k means

Treatment	Independent observations	Sample mean	Population distribution (cdf)
1	$y_{11}, \ldots, y_{1n_1}$	$\bar{y}_1$	$F\{(t-\mu_1)/\sigma^2\}$
⋮	⋮	⋮	⋮
k	$y_{k1}, \ldots, y_{kn_k}$	$\bar{y}_k$	$F\{(t-\mu_k)/\sigma^2\}$

familiar expressions in (2.11) for residual sums of squares. Let

$$\tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_k)^T = \arg\min_{\mu \in H_1} \sum\sum (y_{ij} - \mu_i)^2.$$

Thus, by (2.10), we have that

$$RSS(H_1) = \sum\sum (y_{ij} - \tilde{\mu}_i)^2 \tag{2.12}$$

for the residual sum of squares under H_1. If the error distribution is normal then $(\tilde{\mu}_1, \ldots, \tilde{\mu}_k)$ is also the *mle* of $(\mu_1, \ldots, \mu_k)$ under H_1. By motivations similar to those leading to the $\bar{F}$ in (2.4), let us define

$$\bar{F} = \{RSS(H_0) - RSS(H_1)\}/S^2, \tag{2.13}$$

where $S^2 = \nu^{-1} \sum\sum (y_{ij} - \bar{y}_i)^2$ is the mean square for error and $\nu = (n_1 + \ldots + n_k - k)$ is the error degrees of freedom. As in the previous subsection, the $\bar{F}$ in (2.13) can be considered as a modification of the usual F-statistic to incorporate the order restrictions in H_1. The exact finite sample null distribution of $\bar{F}$ for normal errors will be discussed at the end of this subsection. The p-value of $\bar{F}$ can be estimated by simulation. To this end, the following result is important.

Theorem 2.1.2 *Consider the setting in Table 2.3 and the testing problem* (2.9). *Suppose that the null hypothesis, $H_0 : \mu_1 = \ldots = \mu_k$, holds; let μ denote the common value of $\mu_1, \ldots, \mu_k$. Then the exact finite sample null distribution of $\bar{F}$ depends on the functional form of F but not on (μ, σ). Further, the asymptotic null distribution of $\bar{F}$ does not depend on (F, μ, σ), where the limit is taken as $(n_1 + \ldots + n_k) \to \infty$ such that $0 < \lim_{n \to \infty} n_i/(n_1 + \ldots + n_k) < 1$ for $i = 1, \ldots, k$.*

Proof: Let $y^*_{ij} = (y_{ij} - \mu)/\sigma$ for every (i, j). Let $RSS^*(H_0), RSS^*(H_1), S^*$ and $\bar{F}^*$ denote the values corresponding to $\{y^*_{ij}\}$. Then $RSS^*(H_0) = \sigma^{-2} RSS(H_0)$, $RSS^*(H_1) = \sigma^{-2} RSS(H_1)$, $(S^*)^2 = \sigma^{-2} S^2$, and hence

$$\bar{F} = \{RSS(H_0) - RSS(H_1)\}/S^2 = \{RSS^*(H_0) - RSS^*(H_1)\}/(S^*)^2 = \bar{F}^*.$$

Under H_0, the distribution of y^*_{ij} is $F(t)$, which does not depend on (μ, σ). Now, since $\bar{F}^*$ is a function of $\{y^*_{ij}\}$ only and $\bar{F} = \bar{F}^*$, it follows that the distribution of $\bar{F}$ does not depend on (μ, σ). The proof of the second part will be discussed in the next chapter. ■

This result suggests the following simulation method for computing the p-value.

Computation of the exact p-value for the $\bar{F}$-test:

Suppose that H_0 and H_1 are as in (2.9), $\bar{F}$ is as in (2.13), the setting is as in Table 2.3 and the functional form F of the error distribution is known. Then the following steps are adequate to compute the p-value for $\bar{F}$.

1. Generate independent observations $\{y_{ij} : j = 1, \ldots, n_i, i = 1, \ldots, k\}$ from $F\{(t - \mu_0)/\sigma_0\}$ where (μ_0, σ_0) can have any arbitrary values, but must be held fixed for different (i, j).

2. Compute the $\bar{F}$ statistic in (2.13) with $RSS(H_0)$ and $RSS(H_1)$ as in (2.11) and (2.12), respectively.

3. Repeat the previous two steps N times (say N = 10000), and estimate the p-value by M/N where M is the number of times the $\bar{F}$ statistic in the second step exceeded its sample value. ■

Note that in the first step of the simulation, the observations may be generated from a distribution with any value for the common location and scale parameters because, in view of Theorem 2.1.2, the null distribution of $\bar{F}$ does not depend on the common value of $\mu_1, \ldots, \mu_k$ or σ. For example, if the errors are assumed to be $N(0, \sigma^2)$ where σ is unknown, then it suffices to generate the observations from $N(0, 1)$ in the first step of the simulation for computing the p-value. Similarly, if the error distribution is a logistic with unknown variance, then it suffices to generate the observations from a logistic distribution with any values for its mean and variance.

The simulation approach can also be used to implement a bootstrap-type procedure in which the observations may be generated from the empirical distribution of the within treatment residuals. To improve robustness of validity against violation of the assumption of normal error distribution, we may estimate the p-value corresponding to several forms of error distribution and then choose their maximum as the p-value.

The second part of Theorem 2.1.2 says that if the error distribution is $F(t/\sigma)$ for some (F, σ), which may be unknown, then the asymptotic distribution of $\bar{F}$ does not depend on (F, σ). Therefore, if n_i is large for $i = 1, \ldots, k$ then the foregoing simulation for normal error would estimate the asymptotic p-value, irrespective of whether or not the true error distribution is normal. Similarly, the simulation method with any error distribution would also estimate the asymptotic p-value if n_i is large for $i = 1, \ldots, k$, irrespective of the true error distribution. It is reasonable to conjecture that closer the $F\{(. - \mu)/\sigma\}$ to the true error distribution, for some (μ, σ), the better the approximation.

Suppose that the precise form of F is unknown but it is known that F is a member of the class of distributions $\mathcal{F}$ for some $\mathcal{F}$. Let p_F denote the p-value corresponding to F in Table 2.3. Then we have

$$p\text{-value} = \sup_{F \in \mathcal{F}} p_F \tag{2.14}$$

As an example, suppose that the error distribution is known to be normal, logistic, or a t-distribution with four degrees of freedom (T_4). Then, the foregoing p-value in (2.14) for $\bar{F}$ can be computed by applying the foregoing simulation procedure to compute the p-values corresponding to normal, logistic, and T_4 and then taking their maximum as the p-value in (2.14). In most practical applications, we would not know $\mathcal{F}$, although it would not be difficult to specify a small class of distributional forms that would be well spread out in $\mathcal{F}$. In this case, a reasonable procedure is to compute the p-values corresponding to a range of choices of F (such as normal, logistic, T_r, and χ^2_r) that are well spread out in $\mathcal{F}$ and take their maximum as a reasonable approximation to the p-value.

Example 2.1.2

The experiment in Example 1.2.1 also included a fourth treatment; it was not considered in the previous worked example for simplicity. The group of children receiving the fourth treatment were checked weakly for progress but they did not receive any special exercises.

Table 2.4 Effect of exercise on the age at which a child first walks

Treatment	Age						n	$\bar{y}$
Group 1:	9.00	9.50	9.75	10.00	13.00	9.50	6	10.125
Group 2:	11.00	10.00	10.00	11.75	10.50	15.00	6	11.375
Group 3:	13.25	11.50	12.00	13.50	11.50		5	12.35
Group 4:	11.50	12.00	9.00	11.50	13.25	13.00	6	11.7

Reprinted from: *Science*, Volume 176, P. R. Zelazo, N. A. Zelazo, and S. Kolb, Walking in the newborn, Pages 314–315, Copyright (1972), with permission from AAAS.

Since Groups 1 and 2 were also checked weekly, it may be argued that Group 4 is more appropriate as a control than Group 3; we shall not pursue this issue any further here. The data for all four groups are given in Table 2.4.

Suppose that we wish to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against an alternative that captures our *a priori* information. There is no unique way of formulating the alternative hypothesis. For illustrative purposes, let us suppose that we wish to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ against } H_1 : \mu_1 \leqslant \mu_3, \mu_2 \leqslant \mu_3, \mu_1 \leqslant \mu_4, \mu_2 \leqslant \mu_4.$$

This says that Treatments 1 and 2 are at least as good as Treatments 3 and 4, but no ordering is suggested between Treatments 1 and 2, or between Treatments 3 and 4. In this case, there is no convenient formula similar to (2.8) for computing the p-value for $\bar{F}$; this is a common feature of most of the problems that involve test against inequality constraints. However, in view of Theorem 2.1.2, the simulation method on page 32 is still applicable.

Since $(\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4) = (10.1, 11.4, 12.4, 11.7)$, the sample means satisfy the restrictions in H_1. Therefore, the constrained and unconstrained estimators of $(\mu_1, \mu_2, \mu_3, \mu_4)$ are the same; in particular, $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3, \tilde{\mu}_4) = (\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4) = (10.1, 11.4, 12.4, 11.7)$. Now substituting directly into the formula for $\bar{F}$ in (2.13), the sample value of $\bar{F}$ is 6.43. The p-values corresponding to a range of error distributions were computed using the simulation approach. The computed values are given in the first row of Table 2.5; the figures in the second row are also p-values for a different statistic, which will be discussed later.

It is also of interest to compute the p-value by resampling; to this end, we set the error distribution equal to the empirical distribution of the residuals about the treatment means. This can be implemented by resampling with replacement from the set of within treatment residuals. The computed p-value is also given in the last column of Table 2.5. This example illustrates that simulation is a very convenient

Table 2.5 The p-values for the $\bar{F}$- and $\bar{E}^2$-tests for different error distributions

	Error Distribution							
Test	$N(0,\sigma^2)$	T_4	T_{10}	χ^2_1	χ^2_2	χ^2_4	χ^2_7	RS
$\bar{F}$	0.052	0.051	0.058	0.050	0.048	0.049	0.051	0.048
$\bar{E}^2$	0.046	0.047	0.053	0.046	0.044	0.044	0.047	0.043

The symbol RS refers to resampling from the unrestricted within treatment residuals. To compute the p-value corresponding to $N(0,\sigma^2)$, we used $\sigma = 1$.

way of implementing tests against any order restriction in the one-way classification model when the errors are iid even if the common error distribution is not normal.

It is well known that the critical values from the F-tables are reasonably robust even if the error distribution is not normal for the unrestricted F-test of $H_0 : \mu_1 = \ldots = \mu_k$ against $H_2 : \mu_1, \ldots \mu_k$ *are unequal*. It will be argued later that a similar comment is likely to be applicable for the $\bar{F}$-test as well. Since the p-values in Table 2.5 are close for different error distributions, the results of this example are consistent with this conjecture.

The value of the standard unconstrained F-statistic for testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_2 : \{\mu_1, \mu_2, \mu_3, \mu_4\}$ *are not all equal* is 2.14 and its p-value is 0.129 based on F-statistic $\sim F_{3,19}$. The numerical results for this example again highlight the fact that if the sample means satisfy the constraints that correspond to those in the alternative hypothesis, then the estimate of μ under H_1 and H_2 are the same and the p-value for the constrained test would be smaller than that for the unconstrained F-test. ■

2.1.2.1 Exact finite sample results when the error distribution is $N(0, \sigma^2)$

The results for normal theory likelihood ratio test of H_0 against H_1 are very similar to those for the $\bar{F}$-test. *Assume that the errors are independent and distributed as* $N(0,\sigma^2)$. Let us define

$$\bar{E}^2 = \{RSS(H_0) - RSS(H_1)\}/RSS(H_0).$$

Then the $\bar{E}^2$-test rejects H_0 for large values of $\bar{E}^2$. If the error distribution is normal then, it may be verified that

$$\bar{E}^2 = \{1 - \exp(-LRT/n)\} \tag{2.15}$$

where LRT denotes the likelihood ratio statistic ($= -2\log\Lambda$) for testing H_0 against H_1. It follows from (2.15) that LRT is an increasing function of $\bar{E}^2$, and, therefore, two tests are equivalent.

It is worth noting that $\bar{F}$ and $\bar{E}^2$ have the same numerator, namely $\{RSS(H_0) - RSS(H_1)\}$, which measures the discrepancy between H_0 and H_1. The denominators of these statistics act as scaling factors so that the null distribution of the test statistics

do not depend on the unknown scale parameter of the error distribution; they do not make any contribution to quantifying the discrepancy between H_0 and H_1. Now, Theorem 2.1.2 holds with $\bar{F}$ replaced by $\bar{E}^2$ as well; simply replace $\bar{F}$ by $\bar{E}^2$ in the proof. If the error distribution is not normal then $\bar{E}^2$ is not necessarily a monotonic function of the likelihood ratio statistic, but it is still a suitable test statistic, and the three-step simulation for computing the p-value for $\bar{F}$ (see page 32) can also be applied for $\bar{E}^2$, with the latter replacing the former in the second step of the simulation.

For the Walking Exercise example discussed earlier in this section, the sample value of $\bar{E}^2$ is 0.253. The corresponding p-values for several error distributions are given in Table 2.5; they range from 0.043 for the resampling method to 0.053 for T_{10}. Note that the results for $\bar{F}$ and $\bar{E}^2$ are close; we conjecture that the difference between $\bar{F}$-test and $\bar{E}^2$-test in terms of p-value for a given set of data and in terms of power, would be small in most practical situations.

2.1.2.2 Computing $\min_{\mu\in H_1}\sum\sum(y_{ij}-\mu_i)^2$, $\bar{F}$, and $\bar{E}^2$.

The implementation of the $\bar{F}$- and $\bar{E}^2$-tests requires the computation of $RSS(H_1)$ that is equal to

$$\min_{\mu\in H_1}\sum\sum(y_{ij}-\mu_i)^2 \quad \text{where } H_1:\mu_i-\mu_j\geqslant 0 \text{ for } (i,j)\in B.$$

Let

$$q(\boldsymbol{\mu})=(\bar{\boldsymbol{y}}-\boldsymbol{\mu})^T\boldsymbol{W}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}) \tag{2.16}$$

where $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_k)^T$, $\bar{\boldsymbol{y}}=(\bar{y}_1,\ldots,\bar{y}_k)^T$ and $\boldsymbol{W}=\text{diag}\{n_1,\ldots,n_k\}$, and let $\boldsymbol{A}$ be a matrix such that

$$\{\boldsymbol{\mu}:\mu_i-\mu_j\geqslant 0 \text{ for } (i,j)\in B\}=\{\boldsymbol{\mu}:\boldsymbol{A}\boldsymbol{\mu}\geqslant\boldsymbol{0}\}. \tag{2.17}$$

In this particular case, each row of $\boldsymbol{A}$ is a permutation of the k-vector $(1,-1,0,\ldots,0)$. Since

$$\sum\sum(y_{ij}-\mu_i)^2=q(\boldsymbol{\mu})+C(\boldsymbol{y}),$$

where $C(\boldsymbol{y})=\sum\sum(y_{ij}-\bar{y}_i)^2$, which does not depend on $\boldsymbol{\mu}$, we have that

$$\bar{F}=\{\min_{H_0}q(\boldsymbol{\mu})-\min_{H_1}q(\boldsymbol{\mu})\}/S^2,$$

$$\text{and}\quad \bar{E}^2=\{\min_{H_0}q(\boldsymbol{\mu})-\min_{H_1}q(\boldsymbol{\mu})\}/\{\min_{H_0}q(\boldsymbol{\mu})+C(\boldsymbol{y})\}.$$

We can use the following general method for computing the minimum of $q(\boldsymbol{\mu})$ subject to $\boldsymbol{\mu}\in H_i$, $(i=1,2)$.

Let $\boldsymbol{x}\in\mathbb{R}^k$ be given, $\boldsymbol{B}$ be a given $k\times k$ positive definite matrix and

$$g(\boldsymbol{\mu})=(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{B}(\boldsymbol{x}-\boldsymbol{\mu}),\qquad \boldsymbol{\mu}\in\mathbb{R}^k. \tag{2.18}$$

Suppose that we wish to solve

$$\min g(\boldsymbol{\mu})\quad \text{subject to } \boldsymbol{A}_1\boldsymbol{\mu}\geqslant\boldsymbol{0} \text{ and } \boldsymbol{A}_2\boldsymbol{\mu}=\boldsymbol{0} \tag{2.19}$$

for some matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ that do not depend on $\boldsymbol{\mu}$. This constrained minimization problem in which the objective function is a quadratic in $\boldsymbol{\mu}$ and the constraints are linear equality and inequality constraints in $\boldsymbol{\mu}$ is called a *quadratic program*. There are efficient computer algorithms and software for this optimization problem. For example, we found that the subroutine QPROG in IMSL worked well; similar procedures in packages such as MATLAB, GAUSS, and SPLUS are likely to work just as well.

The foregoing quadratic programming problem is sometimes expressed in the following slightly different, but equivalent, form. Note that

$$g(\boldsymbol{\mu})=2f(\boldsymbol{\mu})+\text{constant},\quad \text{where } f(\boldsymbol{\mu})=\boldsymbol{a}^T\boldsymbol{\mu}+(1/2)\boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{\mu}$$

and $\boldsymbol{a}=-\boldsymbol{B}^T\boldsymbol{x}$. Therefore, the minimization of $g(\boldsymbol{\mu})$ subject to some constraints on $\boldsymbol{\mu}$ is equivalent to the minimization of $f(\boldsymbol{\mu})$ subject to the same constraints on $\boldsymbol{\mu}$. Therefore, $\tilde{\boldsymbol{\mu}}$, the solution to (2.19), is also the solution to

$$\min\{\boldsymbol{a}^T\boldsymbol{\mu}+(1/2)\boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{\mu}\} \text{ subject to } \boldsymbol{A}_1\boldsymbol{\mu}\geqslant\boldsymbol{0} \text{ and } \boldsymbol{A}_2\boldsymbol{\mu}=\boldsymbol{0}. \tag{2.20}$$

2.1.3 General Remarks

It will be shown in the next chapter that, if the errors are independent and distributed as $N(0,\sigma^2)$, then the p-value for $\bar{F}$ is

$$\sum_{i=0}^{k}w_i(H_0,H_1)\text{pr}(iF_{i,\nu}\geqslant\bar{f}_{obs}), \tag{2.21}$$

where $\{w_i(H_0,H_1)\}$ are some nonnegative weights, which depend on the null hypothesis H_0 and the alternative hypothesis H_1, and $\bar{f}_{obs}$ is the sample value of $\bar{F}$. The quantities $\{w_i(H_0,H_1)\}$ are known as *chi-bar-square weights* and also as *level probabilities*; the reasons for these terms will become clear in the next chapter. They appear in the null distribution of several test statistics when there are inequality constraints on parameters.

Often, when there are order restrictions, the null distributions of various test statistics, including $\bar{F}$ and $\bar{E}^2$, take the form $w_0\text{pr}(X_0\leqslant c)+\ldots+w_k\text{pr}(X_k\leqslant c)$ where $w_i=w_i(H_0,H_1)$ and $X_0,\ldots,X_k$ are some random variables. In this expression, computation of $\text{pr}(X_i\leqslant c)$ is usually not difficult $(i=0,\ldots,k)$. However, simple closed-forms are not available for computing $w_0,\ldots,w_k$ exactly, in general. Consequently, there have been various attempts to obtain approximations/bounds for these quantities, so that they can be used to obtain approximations/bounds for the p-value. The details of these are not easy and/or tedious. It appears that the complicated nature of these technical details and unavailability of closed-forms for computing the p-value have led to the opinion among some authors that tests of H_0 against an order restriction are difficult to apply. Fortunately, such an opinion is ill-founded, because there is no need to compute the weights $\{w_i\}$ exactly. In view of the fact that the p-value can be computed sufficiently precisely by simulation, the aforementioned

approximations/bounds on $\{w_1, \ldots, w_k\}$, and the related bounds on the exact null distributions of $\bar{F}$ and $\bar{E}^2$ are not that important for implementing tests of hypotheses involving order restrictions on $\mu_1, \ldots, \mu_k$. Even if $w_0, \ldots, w_k$ can be computed exactly, the simulation approach can be used to compute a p-value that would be more robust against violation of distributional assumptions about the error term, although for a one-way classification problem we believe that the p-values corresponding to normal and other error distributions are likely to be close. However, the bounds on $\{w_1, \ldots, w_k\}$ may be useful in nonlinear models; this will be discussed in Chapter 4.

2.2 ORDERED NULL HYPOTHESES

Now we consider testing problems in which there are inequality constraints in the null hypothesis. To introduce the general form of the testing problems, let μ_i denote the mean of population i, and let y_{ij} denote the jth observation from population i $(i = 1, \ldots, k, j = 1, \ldots, n_i)$. Assume that the observations are independent; thus the setting is the same as that in Table 2.3 for one-way classification. Suppose that we wish to test

$$H_1 : \mu_i - \mu_j \geqslant 0 \text{ for } (i,j) \in B \quad \text{vs} \quad H_2 : \textit{No restriction on } \mu_1, \ldots, \mu_k \tag{2.22}$$

where B is a given subset of $\{(i,j) : i,j = 1, \ldots, k\}$. Thus the hypothesis H_1 in (2.22) is the same as that in (2.9) although in the present context it is the null hypothesis; further, let us also note that in view of the convention introduced earlier (see page 28), the alternative parameter space is the set of all values $\{\mu_1, \ldots, \mu_k\}$ that do not satisfy the constraints H_1 in (2.22). This section provides an introduction and the essentials for carrying out a test of H_1 vs H_2. The distribution theory will be presented in the next chapter.

First, let us consider a general testing problem of the form H_a vs H_b, where H_a and H_b are given. By motivations similar to those leading to (2.13), we define

$$\bar{F} = \{RSS(H_a) - RSS(H_a \cup H_b)\}/S^2, \tag{2.23}$$

for testing H_a against H_b, where RSS is defined as in (2.10). This is a natural generalization of the usual unrestricted F-statistic for H_0 vs H_2 to a test of H_a vs H_b; clearly, the $\bar{F}$ in (2.13) is a special case of (2.23). Thus, for testing H_1 vs H_2, we have

$$\bar{F} = \{RSS(H_1) - RSS(H_1 \cup H_2)\}/S^2, \tag{2.24}$$

where, by the definition of $RSS(H)$ in (2.10), we have that

$$\begin{aligned} RSS(H_1) &= \min_{\mu} \sum\sum (y_{ij} - \mu_i)^2, \text{subject to } (\mu_i - \mu_j) \geqslant 0 \text{ for } (i,j) \in B \\ RSS(H_1 \cup H_2) &= \min_{\mu} \sum\sum (y_{ij} - \mu_i)^2, \text{subject to no restriction on } \mu. \end{aligned} \tag{2.25}$$

This leads to

$$RSS(H_1) = \sum\sum (y_{ij} - \tilde{\mu}_i)^2, \quad RSS(H_1 \cup H_2) = \sum\sum (y_{ij} - \bar{y}_i)^2. \tag{2.26}$$

Now, with $q(\mu)$ as in (2.16), the $\bar{F}$ in (2.24) for the testing problem in (2.22) takes the form,

$$\bar{F} = \min_{\mu \in H_1} q(\mu)/S^2, \tag{2.27}$$

and therefore the computational methods of subsection 2.1.2.2 are also applicable and adequate.

Let $\bar{f}_{obs}$ denote the sample value of $\bar{F}$. We need to take some care in defining the p-value and critical value for this $\bar{F}$-test of H_1 vs H_2 because $\text{pr}(\bar{F} \geqslant \bar{f}_{obs} \mid \textit{the null hypothesis, } H_1)$ is a function of the particular value of $(\mu_1, \ldots, \mu_k)$ in the null parameter space $\{(\mu_1, \ldots, \mu_k) : \mu_i - \mu_j \geqslant 0 \text{ for } (i,j) \in B\}$, for any $\bar{f}_{obs} > 0$. This is in sharp contrast to the testing problem studied in the previous subsection where $\text{pr}(\bar{F} \geqslant \bar{f}_{obs} \mid \textit{the null hypothesis, } H_0)$ does not depend on the particular value of $(\mu_1, \ldots, \mu_k)$ in the null parameter space $\{(\mu_1, \ldots, \mu_k) : \mu_1 = \ldots = \mu_k\}$.

For testing H_1 vs H_2 we define

$$p\text{-value} = \sup_{\mu \in H_1} \text{pr}_{\mu}(\bar{F} \geqslant \bar{f}_{obs}), \tag{2.28}$$

and a level-α test of H_1 vs H_2 rejects H_1 if this p-value $\leqslant \alpha$. An interpretation of this standard approach is that if the data are consistent with at least one point in the null parameter space then we do not reject the null hypothesis. Let $\boldsymbol{\mu}^*$ denote the value of $\boldsymbol{\mu}$ at which the supremum in (2.28) is reached. Then, $\boldsymbol{\mu}^*$ is the point in the null parameter space with which the data are most consistent, and the right-hand side of (2.28) is a measure of the degree of this consistency. If this is large, then the data are consistent with the value $\boldsymbol{\mu}^*$ in the null parameter space and hence we would not reject the null hypothesis. If, on the other hand, the right-hand side of (2.28) is small then the data are not consistent with any point in the null parameter space and hence we would reject the null hypothesis, H_1.

At first, it may seem that computing the supremum in (2.28) is a formidable task; fortunately, it is not. It can be shown that for the particular null hypothesis (i.e., H_1) in (2.22), the supremum in (2.28) is reached where $\mu_1 = \ldots = \mu_k$; further, the right-hand side of (2.28) does not depend on the common value of $\mu_1, \ldots, \mu_k$ (see the theorem below). Hence, (2.28) can be written as

$$p\text{-value} = \text{pr}(\bar{F} \geqslant \bar{f}_{obs} \mid \mu_1 = \ldots = \mu_k). \tag{2.29}$$

The main result concerning the $\bar{F}$-test of H_1 vs H_2 is given in the next result.

Theorem 2.2.1 *Suppose that the setting is as in Table 2.3, the testing problem is (2.22), and $\bar{F}$ is as in (2.24). Let μ denote the common value of $\mu_1, \ldots, \mu_k$ when they are all equal. Then we have the following:*

1. $\sup_{\mu \in H_1} pr_{\mu}(\bar{F} \geqslant c) = pr(\bar{F} \geqslant c \mid \mu_1 = \ldots = \mu_k)$, *for any* $c > 0$.

2. *$pr(\bar{F} \geqslant c \mid \mu_1 = \ldots = \mu_k = \mu)$, depends on the functional form F of the error distribution but not on (μ, σ).*

3. *The asymptotic distribution of $\bar{F}$ at $\mu_1 = \ldots = \mu_k (= \mu, \text{ say})$, and hence the value of $\lim pr(\bar{F} \geqslant c \mid \mu_1 = \ldots = \mu_k = \mu)$ as $n \to \infty$, do not depend on (F, μ, σ) where $c > 0$, $n = (n_1, \ldots, n_k)$ and $0 < \lim(n_i/n) < 1$.* ■

The proof of the first part will be discussed in the next chapter; see Theorem 3.8.1 and Section 3.9. The proof of the second part follows by arguments very similar to those for the corresponding result in the previous subsection, and hence is omitted; the proof of the last part will be provided in Chapter 4. It follows from the foregoing result that the p-value and critical value for testing H_1 vs H_2 can be estimated very precisely by simulation; the main steps are given below.

Computing the exact p-value for the $\bar{F}$-test of H_1 vs H_2

Suppose that the setting is as in Table 2.3, the testing problem is (2.22) and the functional form F of the error distribution is known. Thus, $y_{ij} \sim F\{(t - \mu_i)/\sigma\}$ where F is known but σ may be unknown.

1. Generate independent observations $\{y_{ij} : j = 1, \ldots, n_i, i = 1, \ldots, k\}$ from $F\{(t - \mu_0)/\sigma_0\}$ where we may choose (μ_0, σ_0) to have any convenient value but must be the same for different values of (i, j).

2. Compute the $\bar{F}$ statistic; to this end we may either use (2.24) with RSS as in (2.26), or use (2.27).

3. Repeat the previous two steps N times (say N=10000), and estimate the p-value by M/N where M is the number of times the simulated value of $\bar{F}$ in the second step was not less than its sample value. ■

Note that in the first step of the simulation, it suffices to generate the observations from a distribution with any values for the mean and variance because, in view of Theorem 2.1.2, the null distribution of $\bar{F}$ does not depend on the common location or scale of the error distribution.

Now, suppose that the functional form F of the error distribution is unknown, but is known to be member of the class $\mathcal{F}$. Let p_F denote the p-value when the error distribution is $F(t/\sigma)$ for some σ as in Table 2.3; by Theorem 2.2.1, p_F does not depend on σ. Therefore, we have

$$p\text{-value} = \sup_{F \in \mathcal{F}} p_F. \tag{2.30}$$

As in the arguments that followed (2.14), a procedure to compute a reasonable approximation to the foregoing p-value is to compute p_F corresponding to a range of choices of F (such as normal, logistic, t_r and χ^2_r) that are well spread out in $\mathcal{F}$ and take their maximum as an approximation to the p-value in (2.30).

The results for the normal theory likelihood ratio test of H_1 against H_2 are closely related to those for the $\bar{F}$ test. For testing H_1 vs H_2, let us define

$$\bar{E}^2 = \{RSS(H_1) - RSS(H_1 \cup H_2)\}/RSS(H_1).$$

Then, it may be verified that

$$\bar{E}^2 = \{1 - \exp(-LRT/n)\}. \tag{2.31}$$

where LRT is the normal theory likelihood ratio statistic for testing H_1 vs H_2. Therefore, $\bar{E}^2$ is a monotonically increasing function of the LRT, and hence the LRT and $\bar{E}^2$-tests of H_1 vs H_2 are equivalent. Further, $\bar{E}^2$ $[= \bar{F}/(1 + \bar{F})]$ is also a monotonic function of $\bar{F}$ for testing H_1 against H_2. Therefore, for this testing problem $\bar{E}^2$-test and the $\bar{F}$-test are equivalent.

The exact finite sample null distributions of $\bar{F}$ and $\bar{E}^2$ when the errors are normally distributed will be obtained in the next chapter. It is worth emphasizing that these exact results are for the specific case when the errors are normal; the exact null distributions of $\bar{F}$ and $\bar{E}^2$ are unknown for other error distributions. Hence the p-value computed using these exact results correspond to normal error only. By contrast the simulation approach to computing the p-value does not require the errors to be normally distributed.

In general, hypothesis testing problems with order restrictions in the null hypothesis arise in contexts that are more complicated than the simple one-way layout studied in this section. Nevertheless, the ideas developed in this section provide a valuable introduction and will be helpful in later sections where these are explored further. The following numerical example illustrates the main results of this subsection on test of hypothesis when there are order restrictions in the null hypothesis.

Example 2.2.1 *An example with order restrictions in the null hypothesis*

The sample mean $\bar{y}_i$ and the sample size n_i for population i are given in Table 2.6 for $i = 1, \ldots, 5$; the error mean square, $s^2 = 0.817$.

Table 2.6 Summary statistics for comparing $\mu_1, \ldots, \mu_5$

i	1	2	3	4	5
n_i	6	5	6	6	7
$\bar{y}_i$	11.825	10.811	11.751	11.837	11.231

Let the null and alternative hypothesis be

$$H_1 : \mu_1 \leqslant \mu_2, \mu_1 \leqslant \mu_3, \mu_4 \leqslant \mu_2, \mu_4 \leqslant \mu_3, \mu_1 \leqslant \mu_5, \quad \text{and} \quad H_2 : \text{not } H_1,$$

respectively. Assume that the setting in Table 2.3 holds; in particular, the observations are assumed to be *iid*. Let us apply the $\bar{F}$-test. It follows from the definition of $\bar{F}$,

Table 2.7 Estimated p-values of $\bar{F}$ for different error distributions

Dist:	$N(0,\sigma^2)$	T_4	T_{10}	χ^2_1	χ^2_4	χ^2_8	RS
p-value:	0.091	0.091	0.092	0.074	0.086	0.085	0.091

RS : resample with replacement from the empirical distribution of the within treatment residuals, which is the same as bootstrap.

(2.27), (2.24), and (2.16) that

$$\bar{F} = \{RSS(H_1) - RSS(H_1 \cup H_2)\}/S^2 = \min_{\mu \in H_1} q(\mu)/S^2$$
$$= [\min_{A\mu \geqslant 0} (\bar{y} - \mu)^T W (\bar{y} - \mu)]/S^2$$

where A is the 5×5 matrix $(-1,1,0,0,0 \mid -1,0,1,0,0 \mid 0,1,0,-1,0 \mid 0,0,1,-1,0 \mid -1,0,0,0,1)$; the five rows correspond to the five order restrictions in H_1. We used the subroutine QPROG in IMSL for solving this constrained minimization problem. The sample value of $\bar{F}$ is 5.04. If the error distribution is $F(t/\sigma)$, then p-value = $pr_F(\bar{F} \geqslant 5.04 \mid \mu_1 = \ldots = \mu_5)$, where the probability is computed when the error distribution is $F(t)$. In view of Theorem 2.2.1, the unknown σ can be set equal to one or any other convenient value when computing the p-value. Because the distribution function F is unknown, we approximate the p-value by the maximum of the p-values corresponding to several choices of F (see Table 2.7). In view of the results in Table 2.7, we conclude that the p-value $\geqslant$ 0.074. The results in Table 2.7 show that the p-values are not that sensitive to the form of the error distribution in this particular example; we conjecture that this is likely to be the case more generally. ■

The foregoing results and simulation approaches to computing p-values extend to tests involving order restrictions on the regression parameter $\boldsymbol{\beta}$ in the linear model, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{E}$ where $\boldsymbol{E}$ has mean zero and covariance $\sigma^2\boldsymbol{U}$, σ is unknown and $\boldsymbol{U}$ is known; the inequality constraints may be in the null or alternative hypothesis. This includes (1) test of $H_0 : \boldsymbol{R\beta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{R}_1\boldsymbol{\beta} \geqslant \boldsymbol{0}$, where $\boldsymbol{R}_1$ is a submatrix of $\boldsymbol{R}$, and (2) test of $H_1 : \boldsymbol{R}_1\boldsymbol{\beta} \geqslant \boldsymbol{0}, \boldsymbol{R}_2\boldsymbol{\beta} = \boldsymbol{0}$ against $H_2 : \boldsymbol{\beta}$ *is unrestricted.* These will be discussed in the next chapter.

2.3 ISOTONIC REGRESSION

Consider k populations with μ_i denoting a scalar parameter of interest for group i, $i = 1,\ldots,k$. Suppose that the μ_i s are known to satisfy the *pairwise constraints,*

$$\mu_i - \mu_j \geqslant 0 \text{ for } (i,j) \in B \tag{2.32}$$

where B is a given subset of $\{(i,j) : i,j = 1,\ldots,k\}$. The objective in *isotonic regression* is to address statistical inference on $\mu_1,\ldots,\mu_k$ when they are constrained by (2.32).

Suppose that the random variable Y_i is an estimate of μ_i for $i = 1,\ldots,k$; for example, Y_i could be an observation from a population with mean μ_i or it could be the mean of several *iid* observations from a population with mean μ_i. A natural method of estimating $\{\mu_1,\ldots,\mu_k\}$ when they are known to satisfy (2.32) is to

$$\text{minimize } \sum_{i=1}^{k} (Y_i - \mu_i)^2 w_i \quad \text{subject to (2.32)} \tag{2.33}$$

where $\{w_1,\ldots,w_k\}$ are given or suitably chosen weights. For example, if $Y_i \sim N(\mu_i, \sigma^2/w_i)$ for $i = 1,\ldots,k$ and $Y_1,\ldots,Y_k$ are independent then (2.33) leads to the maximum likelihood estimate of $(\mu_1,\ldots,\mu_k)$ under (2.32). On the other hand, if $Y_1,\ldots,Y_k$ are not necessarily normal but the variance of Y_i is σ^2/w_i, where σ is unknown and $w_1,\ldots,w_k$ are known positive numbers, then (2.33) is the method of weighted least squares with weights, w_i $(i = 1,\ldots,k)$. For the time being, we shall not assume that $Y_1,\ldots,Y_k$ are normal.

Let $(\tilde{\mu}_1,\ldots,\tilde{\mu}_k)$ denote the value of $(\mu_1,\ldots,\mu_k)$, which solves the constrained minimization problem (2.33). Then we say that $(\tilde{\mu}_1,\ldots,\tilde{\mu}_k)$ is the *isotonic regression* of $(Y_1,\ldots,Y_k)$ with respect to the weights $(w_1,\ldots,w_k)$ and the order restriction (2.32). Thus, an *isotonic regression* is a weighted least squares estimate subject to a set of pairwise constraints; different weights and/or pairwise constraints correspond to different isotonic regressions.

As an example, consider the one-way ANOVA setting with k treatments: $E(X_{ij}) = \mu_i$, $\text{var}(X_{ij}) = \sigma^2$ for $j = 1,\ldots,n_i$ and $i = 1,\ldots,k$, and the X_{ij} s are independent. Let Y_i denote the sample mean $\bar{X}_i$ for $i = 1,\ldots,k$; then $E(Y_i) = \mu_i$ and $\text{var}(Y_i) = \sigma^2/n_i$. Suppose that $\mu_1,\ldots,\mu_k$ are known to satisfy

$$\textit{simple order}: \ \mu_1 \leqslant \ldots \leqslant \mu_k.$$

Clearly this order restriction is of the form (2.32). Then the isotonic regression of $(Y_1,\ldots,Y_k)$ with respect to the weights $(n_1,\ldots,n_k)$ and the simple order is the value of $(\mu_1,\ldots,\mu_k)$, which minimizes $\sum(Y_i - \mu_i)^2 n_i$ subject to $\mu_1 \leqslant \ldots \leqslant \mu_k$.

Two other important examples of pairwise constraints are:

1. *Tree order*: $\mu_1 \leqslant \mu_2,\ldots,\mu_1 \leqslant \mu_k$, also written as $\mu_1 \leqslant [\mu_2,\ldots,\mu_k]$; and
2. *Umbrella order*: $\mu_1 \leqslant \ldots \leqslant \mu_m \geqslant \ldots \geqslant \mu_k$, where m is known.

Usually, the tree order arises when a control (= group 1) is compared with several treatments. It is also worth noting that $\mu_1 \leqslant \mu_2 = \mu_3$ is also of the type in (2.32), because the constraint $\mu_2 = \mu_3$ is equivalent to $\mu_2 \leqslant \mu_3$ and $\mu_3 \leqslant \mu_2$.

As was noted in (2.17), the set of constraints in (2.32) can also be expressed as

$$\boldsymbol{A\mu} \geqslant 0 \tag{2.34}$$

for some matrix $\boldsymbol{A}$ in which each row is a permutation of the k-vector, $(-1,1,0,\ldots,0)$, and $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_k)^T$. For example, for the simple and tree orders, the matrix $\boldsymbol{A}$

in (2.34) of order $(k-1)\times k$ takes the forms

$$\begin{bmatrix} -1 & 1 & 0 & \dots & & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \dots & & & & & \\ 0 & \dots & & 0 & -1 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} -1 & 1 & 0 & \dots & & 0 \\ -1 & 0 & 1 & 0 & \dots & 0 \\ \dots & & & & & \\ -1 & 0 & \dots & 0 & 0 & 1 \end{bmatrix}, \tag{2.35}$$

respectively. Since

$$\sum (Y_i - \mu_i)^2 w_i = (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{\mu}),$$

where $\boldsymbol{Y} = (Y_1, \dots, Y_k)^T$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^T$, and $\boldsymbol{W} = \text{diag}\{w_1, \dots, w_k\}$, the minimization problem (2.33) can be expressed as

$$\text{minimize } (\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{\mu}) \quad \text{subject to } \boldsymbol{A}\boldsymbol{\mu} \geqslant \boldsymbol{0}.$$

This is a standard quadratic program; see subsection 2.1.2.2 for a general method of solving it.

There are two types of inference problems that arise in isotonic regression: (i) k is fixed, and (ii) k increases with the sample size. In most parts of this text, only the first case is considered. The second case arises when a real function is to be estimated subject to shape constraints, for example, the nonparametric estimation of (i) a monotone density function, (ii) unimodal density function, (iii) monotone regression function, and (iv) convex regression function; the statistical inference problems arising in such cases are more challenging and they will be discussed later.

2.3.1 Quasi-order and Isotonic Regression

Let the labels $x_1, \dots, x_k$ denote the k groups and let $X = \{x_1, \dots, x_k\}$. Let x be a qualitative explanatory variable taking values in X; for group i, the explanatory variable x takes the value x_i. Note that $x_1, \dots, x_k$ are not necessarily real numbers. In this section, the following notation is adopted to identify a real function on X with a point in $\mathbb{R}^k$. To any given real function g on X, we associate the point $\boldsymbol{g}$ in $\mathbb{R}^k$ defined by $\boldsymbol{g} = (g_1, \dots, g_k)^T$ where $g_i = g(x_i)$ for $i = 1, \dots, k$. Conversely, to any given point $\boldsymbol{g} = (g_1, \dots, g_k)^T$ in $\mathbb{R}^k$ we associate the real function g on X defined by $g(x_i) = g_i$ for $i = 1, \dots, k$. Therefore, we shall freely interchange the roles of functions like g on X and the corresponding vector, $\boldsymbol{g} = (g_1, \dots, g_k)^T$.

Let us now consider our original problem of estimating $\mu_1, \dots, \mu_k$ subject to (2.32). Let y_i denote an estimator of μ_i, $i = 1, \dots, k$, and let

$$y(x_i) = y_i, \quad \boldsymbol{y} = (y_1, \dots, y_k)^T, \quad \mu(x_i) = \mu_i, \text{ and } \boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^T.$$

Consider the regression model

$$y(x_i) = \mu(x_i) + \epsilon_i, \tag{2.36}$$

where ϵ_i is the error term, $i = 1, \dots, k$. Thus, we have a regression model in which the domain of the regression function $\mu(x)$ is the set of labels X, and the function

$\mu(x)$ satisfies the restrictions (2.32). An example of the foregoing setting is when y_i is the sample mean response of several observations from group i and μ_i is the population mean. Another example is where μ_i is the probability of success and y_i is the sample proportion of successes in group i, $i = 1, \dots, k$.

First, let us recall some definitions on binary relations. A binary relation $\preccurlyeq$ on the set $X = \{x_1, \dots, x_k\}$ is said to be (a) *reflexive* if $x \preccurlyeq x$ for every $x \in X$; (b) *transitive*, if $x \preccurlyeq y$ and $y \preccurlyeq z$ imply $x \preccurlyeq z$; (c) *antisymmetric* if $x \preccurlyeq y$ and $y \preccurlyeq x$ imply $x = y$; (d) a *partial order* if it is reflexive, transitive, and antisymmetric; and (e) a *quasi-order* if it is reflexive and transitive. If $\preccurlyeq$ is a quasi-order and if $x_i \preccurlyeq x_j$, then we may interpret it as "group i is not higher than group j."

Let $\preccurlyeq$ be a given quasi-order on the set of labels $X = \{x_1, \dots, x_k\}$. A real function, μ, on X is said to be *isotonic* with respect to $\preccurlyeq$ if

$$x_i \preccurlyeq x_j \text{ implies that } \mu(x_i) \leqslant \mu(x_j), \text{ for every } x_i \text{ and } x_j \text{ in } X.$$

Thus, an isotonic function has the *same tone* as the quasi-order, and hence the term *iso tone*.

Suppose that $\{\mu_1, \dots, \mu_k\}$ satisfy the constraints (2.32). Then, these constraints induce the order $\preccurlyeq$ on X, defined by

$$x_i \preccurlyeq x_j \quad \text{if and only if} \quad (i, j) \in B.$$

It is easily seen that the induced relation $\preccurlyeq$ is a quasi-order. Conversely, suppose that a quasi-order $\preccurlyeq$ on X is given. Then, define the corresponding set B in (2.32) as $\{(i, j) : x_i \preccurlyeq x_j\}$. Therefore, the pairwise constraints (2.32) may be referred to as a quasi-order.

It follows that estimation of $\boldsymbol{\mu}$ subject to (2.32), and estimation of the function $\mu(x)$ subject to it being isotonic with respect to the quasi-order induced by (2.32) are equivalent problems. A formal statement is given in the next proposition.

Proposition 2.3.1 *For a given a quasi-order $\preccurlyeq$ on X there exists a matrix $\boldsymbol{A}$ such that each row of $\boldsymbol{A}$ is a permutation of the k-vector (1, -1, 0, ..., 0), and*

$$\boldsymbol{A}\boldsymbol{\mu} \geqslant \boldsymbol{0} \text{ is equivalent to } \mu \text{ is isotonic with respect to } \preccurlyeq. \tag{2.37}$$

Conversely, given a matrix $\boldsymbol{A}$ in which each row is a permutation of the k-vector (1, -1, 0, ..., 0), there exists a quasi-order $\preccurlyeq$ on X such that (2.37) *holds.* ■

Let y be a given real function on X and w be a given nonnegative function on X. Let

$$\mathcal{F} = \{\mu : \mu \text{ is an isotonic function on } X\} \text{ and } \mathcal{Q} = \{\mu : \boldsymbol{A}\boldsymbol{\mu} \geqslant 0\}$$

where $\boldsymbol{A}$ is as in Proposition 2.3.1. It follows from the foregoing proposition that $\mathcal{F}$ and $\mathcal{Q}$ are equivalent, in the sense that there is a one–one correspondence between $\mathcal{F}$ and $\mathcal{Q}$.

An isotonic function y^* on X is said to be *the least squares isotonic regression of* y *with respect to the weight* w if

$$\sum_{x\in X}\{y(x)-\mu(x)\}^2 w(x) \tag{2.38}$$

reaches its minimum over $\mu \in \mathcal{F}$ at $\mu = y^*$. It is also possible to consider L^p-isotonic regression with respect to w by minimizing $\sum |y(x)-\mu(x)|^p w(x)$; we shall not consider them for the time being. If there is no possibility of ambiguity, then we shall refer to y^* simply as the *isotonic regression* of y without referring to least squares explicitly.

Let

$$\boldsymbol{W} = \text{diag}\{w_1, \ldots, w_k\}.$$

Then

$$\sum_{x\in X}\{y(x)-\mu(x)\}^2 w(x) = (\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}).$$

Now, since $\mathcal{F}$ and $\mathcal{Q}$ are equivalent sets, the following two minimization problems are also equivalent:

$$(A)\quad \textit{minimize} \sum_{x\in X}\{y(x)-\mu(x)\}^2 w(x) \qquad \text{subject to } \mu \in \mathcal{F}, \tag{2.39}$$

$$(B)\quad \textit{minimize}\ (\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}) \qquad \text{subject to } \boldsymbol{A\mu} \geqslant \boldsymbol{0}. \tag{2.40}$$

Let y^* denote the isotonic regression of y in the foregoing minimization problem (A). Then $\boldsymbol{y}^*$ is the *least square projection* of $\boldsymbol{y}$ onto $\mathcal{Q}$ with respect to the inner product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^T\boldsymbol{W}\boldsymbol{y}$. Such projections will be studied in more detail in a later chapter; see Section 2.1.2.2 for some comments on computations relating to the problem (B).

2.4 ISOTONIC REGRESSION: RESULTS RELATED TO COMPUTATIONAL FORMULAS

In this section, several results relating to the computation of the isotonic regression are discussed. They are included here because the details are instructive and/or the results may be useful for studying the properties of the estimator. No attempt is made to provide a comprehensive discussion of the numerical algorithms that may be used for computing the isotonic regression; a discussion of such computational issues is outside the scope of this book. Since the computation of the isotonic regression is a problem in quadratic programming, interested readers may wish to consult the relevant literature in optimization. In this chapter, we will provide only an introduction.

2.4.1 Isotonic Regression Under Simple Order

Let y_i and μ_i denote an observation and a parameter of interest, respectively, for group i, $i = 1, \ldots, k$. Assume that $y_i = \mu_i + \epsilon_i$ and that the parameters, $\mu_1, \ldots, \mu_k$ are unknown but known to satisfy the simpler order $\mu_1 \leqslant \ldots \leqslant \mu_k$. Suppose that it is desired to estimate these parameters by the method of weighted least squares:

$$\min_{\mu_1\leqslant\ldots\leqslant\mu_k} \sum_{i=1}^{k}(y_i-\mu_i)^2 w_i \tag{2.41}$$

where $w_1, \ldots, w_k$ are given positive numbers. This problem has been studied extensively in the early literature, and detailed discussions may be found in Barlow et al. (1972) and Robertson et al. (1988).

Let $\boldsymbol{y}^*$ denote the constrained least squares estimate of $\boldsymbol{\mu}$ resulting from (2.41); $\boldsymbol{y}^*$ is also the isotonic regression of $\boldsymbol{y}$ with respect to weights $\{w_i\}$ and the simple order. It turns out that $\boldsymbol{y}^*$ has simple and instructive explicit forms. It can be computed by a very simple algorithm known as the *Pool Adjacent Violators Algorithm*; this is described below.

Pool Adjacent Violators Algorithm (PAVA)

A set of consecutive elements of X will be called a *block*. For example, $\{x_1, x_2\}$ and $\{x_3, x_4, x_5, x_6\}$ are blocks, but $\{x_1, x_2, x_6\}$ is not a block because 1, 2, and 6 are not consecutive integers. Initially, X is partitioned into the k blocks $\{x_1\}, \ldots, \{x_k\}$ with one label in each block. Now, PAVA pools consecutive blocks as follows:
(i) If $y_1 \leqslant \ldots \leqslant y_k$ then stop the iteration and the solution is $y_j^* = y_j$ for $j = 1, \ldots, k$.
(ii) Otherwise, let i be the smallest index such that $y_i > y_{i+1}$. Now, pool the categories x_i and x_{i+1} to form a single block. Let the response variable for this block be the weighted response, $(w_i y_i + w_{i+1}y_{i+1})/(w_i + w_{i+1})$ with weight $(w_i + w_{i+1})$.
(iii) Now repeat the process for the new blocks with their weighted responses and corresponding weights until the responses are nondecreasing.
(iv) Once this process is completed, y_i^* is equal to the weighted response of the block which contains x_i.

The following example illustrates the main steps of PAVA.

Example: Let $k = 7$ and the values of $\{y_i\}$ and $\{w_i\}$ be as in the top three lines of Table 2.8. The values of y_i^* are given in the last line. The calculations in the table show that the estimator of μ subject to the simple order $\mu_1 \leqslant \cdots \leqslant \mu_7$ is (-5/2, -5/2, 2/3, 2/3, 3, 5, 6). ■

Although the foregoing discussions are limited to the case when the objective function is a weighted sum of squares, the result has been extended to include the case when it is of the form $(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{G}(\boldsymbol{y}-\boldsymbol{\mu})$ where $\boldsymbol{G}$ is a positive definite matrix (Díaz and Salvador (1988)) and the simple order set is replaced by an *acute* cone. The PAVA has also been extended to include more general quadratic programming problems and concave regression problems (Tang and Lin (1991), Qian (1994b))

The isotonic regression under simple order has another form that is instructive. Although the variable x is treated as qualitative in isotonic regression, let us consider

Table 2.8 Illustration of PAVA

i	1	2	3	4	5	6	7
w_i	1	1	2	1	2	2	1
y_i	-1	-4	1	0	3	5	6
First pool:							
Blocks	{1,2}		{3}	{4}	{5}	{6}	{7}
Pooled w	2		2	1	2	2	1
Pooled y	-5/2		1	0	3	5	6
Second pool:							
Blocks	{1,2}		{3,4}		{5}	{6}	{7}
Pooled w	2		3		2	2	1
Pooled y	-5/2		2/3		3	5	6
y_i^*	-5/2	-5/2	2/3	2/3	3	5	6

the special case when it also has a quantitative interpretation. Let $\mu(x)$ be a nondecreasing function of x and let $F(x) = \int^x \mu(t)dt$. Then $F(x)$ is a convex function and $\mu(x)$ is the left slope of $F(x)$. This suggests that an alternative approach to estimating $\mu(x)$ subject to the constraint that μ is nondecreasing, is to estimate F subject to the constraint that it is convex and then estimate $\mu(x_i)$ as the left slope of the estimate of F at x_i. In fact a version of this approach and PAVA are equivalent. Such an equivalence holds, with suitable modifications, even when $x_1, \ldots, x_k$ are not quantitative but k categories. These are discussed below.

Let P_0 denote the origin in two-dimensions, and let the coordinates of P_i be defined by

$$P_i \equiv (w_1 + \ldots + w_i, w_1y_1 + \ldots + w_iy_i), \qquad i = 1, \ldots, k.$$

Let the *Cumulative Sum Diagram* (*CSD*) be the curve which joins the consecutive points, $P_0, P_1, \ldots, P_k$ by straight line segments. The *Greatest Convex Minorant* (*GCM*) of *CSD* is defined as

$$GCM(t) = \max\{g(t) : g \text{ is convex and } g(t) \leqslant CSD(t) \text{ for every } t\}. \quad (2.42)$$

Therefore, *GCM* of *CSD* is the maximum of the convex functions that do not lie above $CSD(t)$ at any t. Alternatively, $GCM(t)$ can be seen as the path of a taut string that joins P_0 and P_k such that $P_0, \ldots, P_k$ are on or above the string (see Figures 2.1 and 2.2). Now we have the following; see pages 9-10 in Barlow et al. (1972) or Theorem 1.2.1 in Robertson et al. (1988) for a proof.

Proposition 2.4.1 *For the simple increasing order, $\mu_1 \leqslant \ldots \leqslant \mu_k$, we have that*

$$y_i^* = \textit{left slope of } GCM(t) \textit{ at } t = (w_1 + \ldots + w_i),\ i = 1, \ldots, k. \quad \blacksquare$$

The solution based on *GCM* can be given an explicit form using min-max and max-min functions. To this end, let

$$Av(s,t) = (w_sy_s + \ldots + w_ty_t)/(w_s + \ldots + w_t) \quad (2.43)$$

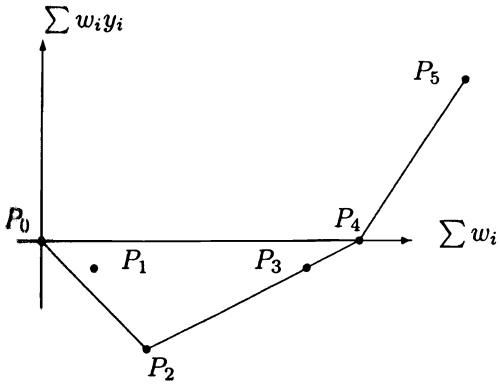

Fig. 2.1 Greatest convex minorant of $P_0P_1P_2P_3P_4P_5$

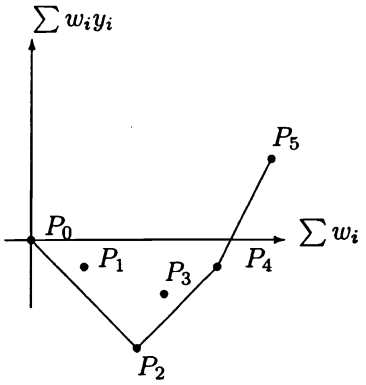

Fig. 2.2 Greatest convex minorant of $P_0P_1P_2P_3P_4P_5$

for $t \geqslant s$. Thus, $Av(s,t)$ is the weighted mean of the observations $y_s, \ldots, y_t$ and it is also the left slope of the line segment $P_{s-1}P_t$. Then equivalent closed-forms of y_i^* are given in the next result, and it is followed by an illustrative example; for more details see Barlow et al. (1972), Robertson et al. (1988), and Qian (1992).

Proposition 2.4.2 *For the simple increasing order, $\mu_1 \leqslant \ldots \leqslant \mu_k$, we have that*

$$\begin{aligned} y_i^* &= \min_{t \geqslant i} \max_{s \leqslant i} Av(s,t) = \min_{t \geqslant i} \max_{s \leqslant t} Av(s,t) \\ &= \max_{s \leqslant i} \min_{t \geqslant i} Av(s,t) = \max_{s \leqslant i} \min_{t \geqslant s} Av(s,t) \end{aligned} \quad \blacksquare$$

Example: Let $k = 6$. The values of (w_i, y_i) and the coordinates of the points P_i are given in Table 2.9, and Fig. 2.1 shows the points. Let us denote the slope of P_iP_j by $[P_iP_j]$. Now, by the foregoing Proposition, we have that

$$\begin{aligned} y_2^* &= \max_{s \leqslant 2} \min_{t \geqslant 2} Av(s,t) = \max_{s \leqslant 2} \min_{t \geqslant 2} [P_{s-1}P_t] \\ &= \max\{\min_{t \geqslant 2}[P_0P_t], \min_{t \geqslant 2}[P_1P_t]\} \\ &= \max\{\ \min\{[P_0P_2], [P_0P_3], [P_0P_4], [P_0P_5], \\ &\qquad \min\{[P_1P_2], [P_1P_3], [P_1P_4], [P_1P_5]\}\ \} \\ &= \max\{[P_0P_2], [P_1P_2]\} = [P_0P_2]. \end{aligned}$$

This simple example illustrates how the max-min formula is related to the left slope of the *GCM(t)* at $t = w_1 + w_2$. Similarly, it would be instructive to verify that y_3^* is equal to the left slope of the *GCM(t)* at $t = (w_1 + \ldots + w_3)$. Fig. 2.2 shows another example of a greatest convex minorant of $P_0P_1P_2P_3P_4P_5$. $\blacksquare$

There are more general versions of the *min max* and *max min* formulas; see Section 1.4 in Robertson et al. (1988).

Table 2.9 Greatest convex minorant

i	1	2	3	4	5	6
w_i	1	1	3	1	2	2
y_i	-1	-3	1	1	3	4
Pooled blocks	$\{1,2\}$		$\{3\}$	$\{4\}$	$\{5\}$	$\{6\}$
Pooled w	2		3	1	2	2
Pooled y	-2		1	1	3	4
Final Estimate						
y_i^*	-2	-2	1	1	3	4
Greatest convex minorant :						
$\sum_{j=1}^{i} w_j$	1	2	5	6	8	10
$\sum_{j=1}^{i} w_j y_j$	-1	-4	- 1	0	6	14
P_i	(1,-1)	(2,-4)	(5,-1)	(6,0)	(8,6)	(10,14)

2.4.2 Ordered Means of Exponential Family

This section contains more advanced topics and it can be ignored at first reading. Further, most of the results in this section will not be required until Chapter 7.

There are different forms of what is known as the exponential family. First, we shall consider the form that is used in Generalized Linear Models where there may be an unknown scale parameter (for example, see McCullagh and Nelder (1989)); another form in which there are no unknown scale parameters will be considered at the end of this subsection. Suppose that the variable y has probability function,

$$f(y;\theta,\phi) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\}, \tag{2.44}$$

where θ and ϕ are scalars and $a(\cdot) > 0, b(\cdot)$ and $c(\cdot)$ are some functions. The function f may be the probability density of a continuous random variable such as normal or gamma, or it may the probability function of a discrete random variable such as poisson or binomial. Let $\ell(\theta,\phi)$ denote $\log f(y;\theta,\phi)$. Then, it is well known that (for example, see McCullagh and Nelder (1989), p 22)

$$(\partial/\partial\theta)\ell(\theta,\phi) = \{y - \dot{b}(\theta)\}/a(\phi), \quad (\partial^2/\partial\theta^2)\ell(\theta,\phi) = -\ddot{b}(\theta)/a(\phi),$$
$$E(y) = \mu = \dot{b}(\theta) \text{ and } \mathrm{var}(y) = \ddot{b}(\theta)a(\phi).$$

Let $\hat{\theta}$ and $\hat{\mu}$ denote the unconstrained *mle*s of θ and μ based on a single observation of y. Then, we have $\hat{\mu} = \dot{b}(\hat{\theta}) = y$.

Now, suppose that we have k groups and assume that the variable y_i in group i has probability function,

$$f(y_i;\theta_i,\phi_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i;\phi_i)\}, \tag{2.45}$$

where θ_i and ϕ_i are scalars and $a(\cdot) > 0, b(\cdot)$ and $c(\cdot)$ are some functions. Let μ_i denote the population mean of y_i. The function $a(\phi_i)$ usually takes the form ϕ/m_i;

ϕ and m_i are known as the dispersion parameter and prior weight, respectively. As an example, if y_i is the mean of m_i independent observations from $N(\mu_i,\sigma^2)$ then $a(\phi_i) = \sigma^2/m_i$. Let $y_{i1},\ldots,y_{in_i}$ denote n_i independent observations from group $i, i = 1,\ldots,k$. Let $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)^T$, $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_k)^T$, $\bar{y}_i = n_i^{-1}(y_{i1} + \ldots + y_{in_i})$, $\bar{\boldsymbol{y}} = (\bar{y}_1,\ldots,\bar{y}_k)^T$. We are interested to obtain the *mle* of $\boldsymbol{\mu}$ subject to

$$\mu_i - \mu_j \geqslant 0 \text{ for } (i,j) \in B \tag{2.46}$$

where B is a given subset of $\{(i,j) : i,j = 1,\ldots,k\}$. Let $\boldsymbol{A}$ be defined by

$$\{\boldsymbol{\mu} : \boldsymbol{A\mu} \geqslant 0\} = \{\boldsymbol{\mu} : \mu_i - \mu_j \geqslant 0 \text{ for } (i,j) \in B\}.$$

It follows from (2.45) that the loglikelihood $L(\boldsymbol{\theta},\boldsymbol{\phi})$ can be expressed as

$$L(\boldsymbol{\theta},\boldsymbol{\phi}) = \sum_{i=1}^{k} \frac{n_i\{\bar{y}_i\theta_i - b(\theta_i)\}}{a(\phi_i)} + H(\boldsymbol{y};\boldsymbol{\phi}) \tag{2.47}$$

where H is some function that does not depend on $\boldsymbol{\theta}$. Let $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\phi}})$ denote the unconstrained *mle* of $(\boldsymbol{\theta},\boldsymbol{\phi})$; thus $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\phi}})$ is the unconstrained maximizer of $L(\boldsymbol{\theta},\boldsymbol{\phi})$. Let the corresponding estimator of $\boldsymbol{\mu}$ be denoted by $\hat{\boldsymbol{\mu}}$. If $\boldsymbol{\mu}(\boldsymbol{\theta},\boldsymbol{\phi})$ expresses $\boldsymbol{\mu}$ as a function of $(\boldsymbol{\theta},\boldsymbol{\phi})$ then by the invariance of *mle*, we have that $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\phi}})$. Since the exponential family was expressed in the canonical form, $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$. Similarly, let $(\tilde{\boldsymbol{\phi}},\tilde{\boldsymbol{\mu}})$ denote the *mle* of $(\boldsymbol{\phi},\boldsymbol{\mu})$ subject to $\boldsymbol{\mu} \in \mathcal{Q}$ where $\mathcal{Q} = \{\boldsymbol{\mu} : \boldsymbol{A\mu} \geqslant \boldsymbol{0}\}$. Since the *mle* is invariant under transformations, the constrained *mle* of $\boldsymbol{\theta}$ is $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\tilde{\boldsymbol{\mu}})$. It turns out that, under some conditions concerning the dispersion parameter $\boldsymbol{\phi}$, $\tilde{\boldsymbol{\mu}}$ is equal to the isotonic regression of $\hat{\boldsymbol{\mu}}$ with respect to some weights, and, therefore, $\tilde{\boldsymbol{\mu}}$ can be obtained by minimizing a quadratic function of $\bar{\boldsymbol{y}}$. This is stated in the next result; the proof is given in the Appendix.

Proposition 2.4.3 *Suppose that the parameter $\boldsymbol{\theta}$ is to be estimated by maximizing the function $L(\boldsymbol{\theta},\boldsymbol{\phi})$ in (2.47) subject to* (2.46)*, which we write as $\boldsymbol{A\mu} \geqslant \boldsymbol{0}$ where $\mu_i = E(\bar{y}_i)$. Let $(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\mu}})$ denote the resulting estimator of $(\boldsymbol{\theta},\boldsymbol{\mu})$.*
(a) Assume that $\phi_1,\ldots,\phi_k$ are known constants. Then $\tilde{\boldsymbol{\mu}}$ is also the value of $\boldsymbol{\mu}$ at which $\sum(\bar{y}_i - \mu_i)^2\{n_i/a(\phi_i)\}$ reaches its minimum subject to $\boldsymbol{A\mu} \geqslant \boldsymbol{0}$.
(b) Assume that $\phi_1 = \ldots = \phi_k (= \phi$, say), and ϕ may be an unknown nuisance parameter. Then $\tilde{\boldsymbol{\mu}}$ is also the value of $\boldsymbol{\mu}$ at which $\sum(\bar{y}_i - \mu_i)^2 n_i$ reaches its minimum subject to $\boldsymbol{A\mu} \geqslant \boldsymbol{0}$. ■

To consider important special cases of this result, let $y_{i1},\ldots,y_{in_i}$ denote n_i independent observations from any one of the following (here we use the same notation as in McCullagh and Nelder (1989) page 30):

1. Normal with mean μ_i and unknown variance σ^2.
2. Poisson with mean μ_i.
3. Binary with $\mathrm{pr}(y_{ij} = 1) = \mu_i$ and $\mathrm{pr}(y_{ij} = 0) = 1 - \mu_i$.

4. Gamma with parameters (μ_i, ν), where $\text{var}(y_i) = \mu_i^2/\nu$ and ν is unknown.

5. Inverse gaussian with parameter (μ_i, σ^2) where σ is unknown.

All of these cases belong to the canonical exponential family in (2.44). Therefore, for each of the five cases just listed, it follows from Proposition 2.4.3 that

$$\tilde{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{A\mu} \geqslant \mathbf{0}} \sum (\bar{y}_i - \mu_i)^2 n_i.$$

If the constraint $\boldsymbol{A\mu} \geqslant \mathbf{0}$ is the same as $\mu_1 \leqslant \ldots \leqslant \mu_k$ then it follows from Proposition 2.4.2 that

$$\tilde{\mu}_i = \min_{t \geqslant i} \max_{s \leqslant i} (n_s \bar{y}_s + \ldots + n_t \bar{y}_t)/(n_s + \ldots + n_t); \tag{2.48}$$

further, $\tilde{\mu}_i$ can also be expressed as the left slope of a greatest convex minorant of a particular piecewise linear function as shown in Proposition 2.4.1.

Now, let us consider the case when $\phi_1, \ldots, \phi_k$ are unknown nuisance parameters; this was not considered in Proposition 2.4.3. Let $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\phi}})$ denote the solution of

$$\min L(\boldsymbol{\theta}, \boldsymbol{\phi}) \text{ subject to } \boldsymbol{A\mu} \geqslant \mathbf{0}.$$

It can be shown using Lemmas 2.5.3 and 2.5.2 in the appendix to this chapter that $\tilde{\boldsymbol{\mu}}$ is still the isotonic regression of $\bar{\boldsymbol{y}}$ with respect to the weights $\tilde{w}_i = n_i/a(\tilde{\phi}_i)$ as in part (a) of Proposition 2.4.3. However, since the weights depend on $\tilde{\boldsymbol{\phi}}$, which in turn may depend on $\tilde{\boldsymbol{\mu}}$, it follows that $\tilde{\boldsymbol{\mu}}$ is not just a function of $\bar{\boldsymbol{y}}$ alone, as is the case in parts (a) and (b) of Proposition 2.4.3. Therefore, the foregoing results do not provide a simple closed-form for computing $\tilde{\boldsymbol{\mu}}$ when $\phi_1, \ldots, \phi_k$ are unknown.

If the conditions of Proposition 2.4.3 are satisfied, then it provides a convenient way of computing $\tilde{\boldsymbol{\mu}}$ starting from the unconstrained estimator $\hat{\boldsymbol{\mu}}$. Further, it provides analytically simpler forms for the solution. This may be helpful in the study of the properties of the solutions. For example, to study the asymptotic properties of $\tilde{\boldsymbol{\mu}}$, it may be convenient to use the fact that $\tilde{\boldsymbol{\mu}}$ is obtained by minimizing a quadratic function of $\hat{\boldsymbol{\mu}}$; since properties of $\hat{\boldsymbol{\mu}}$ are easier to obtain, and minimizers of sum of squares are more tractable, this approach may turn out to be easier.

If there are no unknown scale parameters, then a result similar to the foregoing proposition holds even when the exponential family is not in the canonical form (2.44). Suppose that the density function of y_{ij} is

$$f(y; \theta_i, \phi_i) = \exp\{p_1(\theta_i) p_2(\phi_i) K(y; \phi_i) + S(y; \phi_i) + q(\theta_i, \phi_i)\},$$

where $(\partial/\partial\theta_i) q(\theta_i, \phi_i) = -\theta_i (\partial/\partial\theta_i) p_1(\theta_i) p_2(\phi_i)$ and ϕ_i is known $(i = 1, \ldots, k)$. Then, we have $E\{K(y_i; \phi_i)\} = \theta_i$ and $var\{K(Y_i; \phi_i)\} = \{(\partial/\partial\theta_i) p_1(\theta_i) p_2(\phi_i)\}^{-1}$. Further, the unconstrained *mle* $\hat{\theta}_i$ of θ_i is $(1/n_i)\{K(y_{i1}; \phi_i) + \ldots + K(y_{in_i}; \phi_i)\}$; let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^T$. Now, the *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{A\theta} \geqslant \mathbf{0}$ is the isotonic regression of $\hat{\boldsymbol{\theta}}$ with weights $w_i = n_i p_2(\phi_i)$ (see Theorem 1.5.2 in Robertson et al. (1988); they attribute this result to the student, Zehua Chen).

The next result (proof is given in the appendix to this chapter) appears on page 45 in Barlow et al. (1972); also on page 38 in Robertson et al. (1988). This result will be used for obtaining simpler formulas for the constrained *mle* of a multinomial parameter, and also the constrained nonparametric *mle* of a monotonic density function in a later chapter.

Corollary 2.4.4 *(Maximum of a product). Let c_i and y_i be given positive numbers, s be a given real number,*

$$L(\boldsymbol{\theta}) = \theta_1^{y_1} \ldots \theta_k^{y_k},$$

$\boldsymbol{A}$ be a given matrix in which every row is a permutation of the k-vector $(-1, 1, 0, \ldots, 0)$, and $z_i = (y_i/\sum y_i)(s/c_i)$ $(i = 1, \ldots, k)$. Then

$$\textit{the maximum of } L(\boldsymbol{\theta}) \textit{ subject to } \boldsymbol{A\theta} \geqslant \mathbf{0} \textit{ and } \sum c_i \theta_i = s$$

and

$$\textit{the minimum of } \sum (z_i - \theta_i)^2 c_i \textit{ subject to } \boldsymbol{A\theta} \geqslant \mathbf{0}$$

are reached at the same value of $\boldsymbol{\theta}$. ■

Example 2.4.1 *Multinomial Parameters*

Consider a multinomial distribution with k categories, and probability function

$$n!(y_1! \ldots y_k!)^{-1} \theta_1^{y_1} \ldots \theta_k^{y_k},$$

where $n = (y_1 + \ldots + y_k)$ is the total number of observations. Let $\tilde{\boldsymbol{\theta}}$ denote the *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{A\theta} \geqslant \mathbf{0}$ and $\sum \theta_i = 1$. Then, we have that

$$\tilde{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{A\theta} \geqslant \mathbf{0}} \sum_{i=1}^{k} (\hat{\theta}_i - \theta_i)^2 \text{ where } \hat{\theta}_i = y_i/n \qquad \text{for } i = 1, \ldots, k.$$

This can be deduced from the previous corollary as follows.

Since $n!(y_1! \ldots y_k!)^{-1}$ does not depend on $\boldsymbol{\theta}$ and $\sum \theta_i = 1$ is equivalent to $\sum c_i \theta_i = s$ where $c_i = 1 = s$, the results of the previous corollary on maximization of a product are applicable. Note that $z_i = \{y_i/(\sum y_i)\}(s/c_i) = (y_i/n) = \hat{\theta}_i$, which is the unrestricted *mle* of θ_i $(i = 1, \ldots, k)$. Further, $w_i = (y_i/z_i) = n$. Now, the desired result follows from the previous corollary. ■

2.5 APPENDIX: PROOFS

The proof of Proposition 2.4.3 involves several intermediate results that are of independent interest. These are stated below as lemmas. The proofs of these lemmas, given below, use various mathematical results on projections onto polyhedrals. A detailed account of the relevant results and definitions are given in the next chapter. Therefore, these proofs are meant to be read only if the reader is familiar with the

basic results on projections on to polyhedrals discussed in the Appendix to the next chapter. Proofs of the following lemmas are also given from first principles in Barlow et al. (1972); see also Section 1.7 in Robertson et al. (1988).

Lemma 2.5.1 *Let $\boldsymbol{A}$ be a matrix in which every row is a permutation of* $(-1, 1, 0, 0, \ldots, 0)$, $\mathcal{Q} = \{\boldsymbol{\mu} : \boldsymbol{A\mu} \geqslant \boldsymbol{0}\}$, *and let $\boldsymbol{y}^*$ denote the value of $\boldsymbol{\mu}$ at which $\sum\{y_i - \mu_i\}^2 w_i$ reaches its minimum over $\boldsymbol{A\mu} \geqslant \boldsymbol{0}$. For any real function ϕ on $\mathbb{R}$, define $\boldsymbol{\phi}^* = (\phi(y_1^*), \ldots, \phi(y_k^*))^T$. Then we have the following:*

1. *Let F be the unique face of $\mathcal{Q}$ such that $\boldsymbol{y}^*$ is in the relative interior of F. Then, $\boldsymbol{\phi}^*$ is in the linear space spanned by F.*
2. $(\boldsymbol{y} - \boldsymbol{y}^*)^T \boldsymbol{W} \boldsymbol{\phi}^* = 0$ *where* $\boldsymbol{W} = \operatorname{diag}\{w_1, \ldots, w_k\}$.

Proof: Let $\boldsymbol{a}_i$ denote a typical row of $\boldsymbol{A}$. Let I and J be such that

$$F = \{\boldsymbol{\mu} : \boldsymbol{a}_i^T \boldsymbol{\mu} = 0 \text{ for } i \in I, \text{ and } \boldsymbol{a}_j^T \boldsymbol{\mu} \geq 0 \text{ for } j \in J\}.$$

Then

$$ri(F) = \{\boldsymbol{\mu} : \boldsymbol{a}_i^T \boldsymbol{\mu} = 0 \text{ for } i \in I, \text{ and } \boldsymbol{a}_j^T \boldsymbol{\mu} > 0 \text{ for } j \in J\},$$

where $ri()$ denotes the relative interior. Let $\langle F \rangle = \{\boldsymbol{\mu} : \boldsymbol{a}_i^T \boldsymbol{\mu} = 0 \text{ for } i \in I\}$, the linear space spanned by F. Then $(\boldsymbol{y} - \boldsymbol{y}^*) \perp_{\boldsymbol{W}^{-1}} \langle F \rangle$. Suppose that $\boldsymbol{a}_i^T \boldsymbol{\mu} = 0$ for some $i \in I$ and that the nonzero elements of $\boldsymbol{a}_i$ occur at positions p and q. Then $\mu_p = \mu_q$. Consequently, $\phi(\mu_p) = \phi(\mu_q)$ and $\boldsymbol{a}_i^T \boldsymbol{\phi}_0 = 0$ where $\boldsymbol{\phi}_0 = (\phi(\mu_1), \ldots, \phi(\mu_k))^T$. Since $\boldsymbol{y}^* \in F \subset \langle F \rangle$, it follows that $\boldsymbol{a}_i^T \boldsymbol{y}^* = 0$ and hence $\boldsymbol{a}_i^T \boldsymbol{\phi}^* = 0$ for every $i \in I$; hence, $\boldsymbol{\phi}^*$ also belongs to $\langle F \rangle$. This completes the proof of the first part. The proof of the second part follows from $(\boldsymbol{y} - \boldsymbol{y}^*) \perp_{\boldsymbol{W}^{-1}} \langle F \rangle$. ■

The first part of the lemma provides a geometric picture and the second part follows from the first. The next lemma says that the minimum of a weighted sum of squares and that of a more general function, both subject to the same constraints, are reached at the same point.

Lemma 2.5.2 *Let $y_1, \ldots, y_k$ be given real numbers, and let I be an interval containing $\{y_1, \ldots, y_k\}$. Let Φ be a convex function that is finite on I and ∞ elsewhere. Let ϕ be the derivative of Φ (at a corner, choose a value between the left and right derivatives); assume that ϕ is finite on I. Let*

$$\Delta(u, v) = \begin{cases} \Phi(u) - \Phi(v) - (u - v)\phi(v), & \text{if } u, v \in I \\ \infty, & \text{if } u \notin I \text{ or } v \notin I. \end{cases}$$

For any two points, $\boldsymbol{y}$ and $\boldsymbol{\mu}$ in $\mathbb{R}^k$, let

$$\Delta[\boldsymbol{y}, \boldsymbol{\mu}] = \sum_{i=1}^{k} \Delta(y_i, \mu_i) w_i.$$

Let $\boldsymbol{A}$ be a matrix in which every row is a permutation of $(-1, 1, 0, 0 \ldots, 0)$, $\mathcal{Q} = \{\boldsymbol{\mu} : \boldsymbol{A\mu} \geqslant \boldsymbol{0}\}$, *and let $\boldsymbol{y}^*$ denote the value of $\boldsymbol{\mu}$ at which $\sum\{y_i - \mu_i\}^2 w_i$ reaches its minimum over $\boldsymbol{A\mu} \geqslant \boldsymbol{0}$. Then,*

(i) $\Delta[\boldsymbol{y}, \boldsymbol{\mu}] \geqslant \Delta[\boldsymbol{y}, \boldsymbol{y}^*] + \Delta[\boldsymbol{y}^*, \boldsymbol{\mu}]$, *for any* $\boldsymbol{\mu} \in \mathcal{Q}$.

(ii) $\boldsymbol{y}^*$ *solves* $\min_{\boldsymbol{\mu} \in \mathcal{Q}} \Delta[\boldsymbol{y}, \boldsymbol{\mu}]$.

(iii) $\boldsymbol{y}^*$ *also solves* $\max_{\boldsymbol{\mu} \in \mathcal{Q}} \sum[\Phi(\mu_i) + (y_i - \mu_i)\phi(\mu_i)]w_i$.

(iv) If Φ is strictly convex then $\Delta[\boldsymbol{y}, \boldsymbol{\mu}]$ has a unique minimum over $\boldsymbol{\mu} \in \mathcal{Q}$.

Proof:
[This proof is essentially the same as that in Barlow et al. (1972), page 42]. It follows from the definition of Δ that, for real numbers r, s and t,

$$\Delta(r, t) = \Delta(r, s) + \Delta(s, t) - (r - s)\{\phi(s) - \phi(t)\}.$$

Let $\boldsymbol{\mu}$ be a point in $\mathcal{Q}$, $\boldsymbol{\phi}_0 = (\phi(\mu_1), \ldots, \phi(\mu_k))^T$, and $\boldsymbol{\phi}^* = (\phi(y_1^*), \ldots, \phi(y_k^*))^T$. It follows from (2.49) that

$$\Delta[\boldsymbol{y}, \boldsymbol{\mu}] = \Delta[\boldsymbol{y}, \boldsymbol{t}^*] + \Delta[\boldsymbol{y}^*, \boldsymbol{\mu}] - (\boldsymbol{y} - \boldsymbol{y}^*)^T \boldsymbol{W}(\boldsymbol{\phi}^* - \boldsymbol{\phi}_0). \tag{2.49}$$

Since $\boldsymbol{A\mu} \geqslant \boldsymbol{0}$ and ϕ is nondecreasing, it follows that $\boldsymbol{A\phi}_0 \geqslant \boldsymbol{0}$ and hence $\boldsymbol{\phi}_0 \in \mathcal{Q}$. Therefore, it follows from Lemma 3.12.3 (see page 114) that

$$(\boldsymbol{y} - \boldsymbol{y}^*)^T \boldsymbol{W} \boldsymbol{\phi}_0 \leqslant 0. \tag{2.50}$$

Further, by Lemma 2.5.1, we have that

$$(\boldsymbol{y} - \boldsymbol{y}^*)^T \boldsymbol{W} \boldsymbol{\phi}^* = 0. \tag{2.51}$$

Now, part (i) of the theorem follows from (2.49), (2.50), and (2.51). Part (ii) follows from part (i) because $\Delta[\cdot, \cdot] \geqslant 0$. Part (iii) follows from

$$\Delta[\boldsymbol{y}, \boldsymbol{\mu}] = \sum_{i=1}^{k} \Phi(y_i) w_i - \sum_{i=1}^{k} \{\Phi(\mu_i) + (y_i - \mu_i)\phi(\mu_i)\} w_i]$$

and the fact that the first term, $\sum \Phi(y_i) w_i$, does not depend on $\boldsymbol{\mu}$. Part (iv) follows from $\Delta[\boldsymbol{y}^*, \boldsymbol{\mu}] > 0$ if $\boldsymbol{y}^* \neq \boldsymbol{\mu}$. ■

Lemma 2.5.3 *Let $L(\boldsymbol{\theta}, \boldsymbol{\phi})$ denote the loss function in (2.47). Then,*

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \sum_{i=1}^{k} \{\bar{y}_i \theta_i - b(\theta_i)\}\{n_i / a(\phi_i)\} + \sum_{i,j} c(y_{ij}; \phi_i) \\ &= \sum_{i=1}^{k} \{\Phi(\mu_i) + \theta(\mu_i)(\bar{y}_i - \mu_i)\}\{n_i / a(\phi_i)\} + H(\boldsymbol{y}; \boldsymbol{\phi}, \mu_0), \end{aligned}$$

for some function H, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)^T$, and $\Phi(t) = \int_{\mu_0}^{t} \theta(\mu) d\mu$.

Proof:

Let θ, μ and ϕ denote scalar parameters. Let $\dot{b}(\theta)$ and $\ddot{b}(\theta)$ denote the first and second derivatives of $b(\theta)$. For the exponential family in (2.47), we have that

$$\mu = \dot{b}(\theta) \text{ and } var(Y) = \ddot{b}(\theta)a(\phi);$$

for example, see McCullagh and Nelder (1989), page 29). Since $\ddot{b}(\theta) > 0$, it follows that $\dot{b}(\theta)$ is a strictly increasing function. Therefore, $\mu = \dot{b}(\theta)$ can be inverted, and hence we have that

$$\theta(\mu) = \dot{b}^{-1}(\mu), \mu = \dot{b}\{\theta(\mu)\}, \text{ and } (d/d\mu)b\{\theta(\mu)\} = \dot{b}(\theta)\dot{\theta}(\mu) = \mu\dot{\theta}(\mu).$$

Now, to obtain the unconstrained *mle,* $\hat{\theta}$, we solve $(\partial/\partial\theta)L(\theta, \phi) = 0$. This leads to

$$\hat{\mu}_i = \bar{y}_i = \dot{b}(\hat{\theta}_i) = \dot{b}\{\theta(\hat{\mu}_i)\}. \tag{2.52}$$

Let μ_0 be a value in the allowable range of μ. Then

$$\begin{aligned} b\{\theta(\mu)\} &= \int_{\mu_0}^{\mu} (d/dt)b\{\theta(t)\}dt + C(\mu_0) \\ &= \int_{\mu_0}^{\mu} t\dot{\theta}(t)dt + C(\mu_0) \\ &= \mu\theta(\mu) - \int_{\mu_0}^{\mu} \theta(t)dt + C(\mu_0). \end{aligned} \tag{2.53}$$

Substituting (2.52) and (2.53) in (2.47), we have

$$Loglikelihood = \sum\{n_i/a(\phi_i)\}\{\Phi(\mu_i) + \theta(\mu_i)(\hat{\mu}_i - \mu_i)\} + H(\boldsymbol{y}, \phi)$$

where $\Phi(x) = \int_{\mu_0}^{x} \theta(t)dt$. ■

Proof of Proposition 2.4.3:

Follows from the previous two lemmas.

Proof of Corollary 2.4.4:

Let $w_i = (y_i/z_i)$ for $i = 1, \ldots, k$, and $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. Then

$$\ell(\boldsymbol{\theta}) = \sum y_i \log\theta_i = \sum(z_i \log\theta_i)w_i. \tag{2.54}$$

Since $\sum(z_i - \theta_i)w_i = s^{-1}(\sum y_i)(s - \sum c_i\theta_i)$, it follows that

$$\sum c_i\theta_i = s \text{ if and only if } \sum(z_i - \theta_i)w_i = 0. \tag{2.55}$$

Let $\Phi(t) = t\log(t)$ and let Δ be defined as in Lemma 2.5.2. Then

$$\begin{aligned} \Delta[a, b] &= \Phi(a) - \Phi(b) - (a - b)\phi(b) \\ &= a\log a - a\log b - (a - b) \end{aligned}$$

and hence

$$\Delta[z_i, \theta_i]w_i = \sum(z_i \log z_i)w_i - \{\ell(\boldsymbol{\theta}) + r(\boldsymbol{\theta})\} \tag{2.56}$$

where $r(\boldsymbol{\theta}) = \sum(z_i - \theta_i)w_i$. Since $\sum(z_i \log z_i)w_i$ does not depend on $\boldsymbol{\theta}$, we can restrict our attention to $\{\ell(\boldsymbol{\theta}) + r(\boldsymbol{\theta})\}$ for the purposes of finding the minimum of $\sum\Delta[z_i, \theta_i]w_i$. Let $\mathcal{A} = \{\boldsymbol{\theta} : \boldsymbol{A\theta} \geqslant \boldsymbol{0}\}$ and $\mathcal{R} = \{\boldsymbol{\theta} : r(\boldsymbol{\theta}) = 0\}$. Let

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathcal{A}} \sum(z_i - \theta_i)^2 w_i.$$

Let $\phi(t) = 1$ for every $t \in \mathbb{R}$. Then, by Lemma 2.5.1, we have that $\boldsymbol{\phi}^* = (\phi(\tilde{\theta}_1), \ldots, \phi(\tilde{\theta}_k))^T = \boldsymbol{1}^T$, and $(\boldsymbol{z} - \tilde{\boldsymbol{\theta}})^T\boldsymbol{W}\boldsymbol{1} = 0$; hence $r(\tilde{\boldsymbol{\theta}}) = \sum(z_i - \tilde{\theta}_i)w_i = 0$. Further,

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}\in\mathcal{A}} \sum\Delta[z_i, \theta_i]w_i \quad \text{by Lemma 2.5.2, part(ii)} && (2.57) \\ &= \arg\max_{\boldsymbol{\theta}\in\mathcal{A}}\{\ell(\boldsymbol{\theta}) + r(\boldsymbol{\theta})\} \quad \text{by (2.56)} && (2.58) \end{aligned}$$

Now,

$$\begin{aligned} \max_{\boldsymbol{\theta}\in\mathcal{A}\cap\mathcal{R}} \ell(\boldsymbol{\theta}) &= \max_{\boldsymbol{\theta}\in\mathcal{A}\cap\mathcal{R}}\{\ell(\boldsymbol{\theta}) + r(\boldsymbol{\theta})\}, \quad \text{because } r(\boldsymbol{\theta}) = 0 \text{ for } \boldsymbol{\theta} \in \mathcal{R}. \\ &\leqslant \max_{\boldsymbol{\theta}\in\mathcal{A}}\{\ell(\boldsymbol{\theta}) + r(\boldsymbol{\theta})\}, \quad \text{because } \mathcal{A}\cap\mathcal{R} \subset \mathcal{A} \\ &= \ell(\tilde{\boldsymbol{\theta}}) + r(\tilde{\boldsymbol{\theta}}) \quad \text{by (2.57) and (2.58)} \\ &= \ell(\tilde{\boldsymbol{\theta}}) \quad \text{because } r(\tilde{\boldsymbol{\theta}}) = 0. \end{aligned}$$

Since $r(\tilde{\boldsymbol{\theta}}) = 0$, we have that $\tilde{\boldsymbol{\theta}} \in \mathcal{R}$, and hence $\max_{\boldsymbol{\theta}\in\mathcal{A}\cap\mathcal{R}} \ell(\boldsymbol{\theta}) = \ell(\tilde{\boldsymbol{\theta}})$. Since Φ is strictly convex, $\tilde{\boldsymbol{\theta}}$ is uniquely determined. ■

Problems

2.1

In Example 1.2.2, assume that the error distribution is normal. Now, test the following hypotheses.

1. $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_1 : \mu_1 \geqslant \mu_2 \geqslant \mu_3$.
2. $H_0 : \mu_1 \geqslant \mu_2 \geqslant \mu_3$ vs $H_1 : \mu_1 \geqslant \mu_2 \geqslant \mu_3$ *does not hold.*

[Hint: Apply the $\bar{F}$-test and follow the steps as in the worked example on walking exercise].

2.2

For the Example 2.1.2 on page 33, compute the p-values when the error distribution is (a) Normal, (b) Logistic, and (c) χ^2_5 for each of the following:

1. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs $H_1 : \mu_1 \leqslant \mu_2 \leqslant \mu_3, \mu_2 \leqslant \mu_4$.
2. $H_1 : \mu_1 \leqslant \mu_2 \leqslant \mu_3, \mu_2 \leqslant \mu_4$ vs H_2 : *No restriction on* $(\mu_1, \ldots, \mu_4)$.

2.3 Consider the normal theory one-way ANOVA context in Table 2.1. Let H_1 be as in (2.9), and $(\tilde{\mu}_1, \ldots, \tilde{\mu}_k)$ be the solution of

$$\min \sum\sum (y_{ij} - \mu_i)^2 \text{ subject to } H_1.$$

Show that $\tilde{\mu}_1, \ldots, \tilde{\mu}_k)$ is the constrained *mle* of $(\mu_1, \ldots, \mu_k)$ under H_1.

[Hint: Let $L(\mu_1, \ldots, \mu_k, \sigma)$ denote the log likelihood corresponding to $y_{ij} \sim N(\mu_i, \sigma^2)$. Then $L(\mu_1, \ldots, \mu_k, \sigma) = -(2\sigma^2)^{-1} \sum\sum (y_{ij} - \mu_i)^2 - N \log \sigma^2$. For any given, $(\mu_1, \ldots, \mu_k)$, L reaches its minimum at $\tilde{\sigma}^2 = N^{-1} \sum\sum (y_{ij} - \mu_i)^2$. Therefore, the *concentrated* loglikelihood obtained by eliminating σ is

$$L^*(\mu_1, \ldots, \mu_k) = L(\mu_1, \ldots, \mu_k, \tilde{\sigma}) = -\sum\sum (y_{ij} - \mu_i)^2 - 2N^{-1}.$$

The *mle* of $(\mu_1, \ldots, \mu_k)$ under H_1 is its value at which $\sum\sum (y_{ij} - \mu_i)^2$ reaches its minimum.]

2.4 Let y be a binary random variable taking the values 1 and 0 with probability μ and $(1 - \mu)$, respectively, $(0 < \mu < 1)$. Let $f(y)$ denote the probability function of y. Verify that $f(y) = \mu^y (1 - \mu)^{1-y} = \exp[y \log\{\mu/(1 - \mu)\} + \log(1 - \mu)]$, and deduce that the probability function of y is of the form (2.45) with $\theta(\mu) = \log\{\mu/(1 - \mu)\}$, $a(\phi) = 1$ and $b(\theta) = -\log(1 - \mu)$.

Suppose that there are k groups and for group i, let $y_{i1}, \ldots, y_{in_i}$ denote n_i independent binary random variables with $\text{pr}(y_{ij} = 1) = \mu_i (i = 1, \ldots, k)$. Verify that

$$\text{likelihood} = \Pi_{i=1}^{k} \Pi_{j=1}^{n_i} \mu_i^{y_{ij}} (1 - \mu_i)^{1 - y_{ij}}.$$

Let $\boldsymbol{A}$ be a matrix with k columns in which every row is a permutation of $(1, -1, 0, \ldots, 0)$. Use Proposition 2.4.3 and show that the *mle* $\tilde{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ subject to $\boldsymbol{A\mu} \geqslant \boldsymbol{0}$ is also the value of $\boldsymbol{\mu}$ which

$$\text{maximizes } \sum (\bar{y}_i - \mu_i)^2 n_i \text{ subject to } \boldsymbol{A\mu} \geqslant \boldsymbol{0}. \tag{2.59}$$

Now, suppose that the constraint $\boldsymbol{A\mu} \geqslant \boldsymbol{0}$ is the same as $\mu_1 \leqslant \ldots \leqslant \mu_k$. Show that

$$\tilde{\mu}_i = \min_{t \geqslant i} \max_{s \leqslant i} (n_s \bar{y}_s + \ldots + n_t \bar{y}_t)/(n_s + \ldots + n_t).$$

Suppose that $k = 5, (n_1, \ldots, n_5) = (10, 15, 25, 12, 16)$ and $\bar{\boldsymbol{y}} = (3/10, 5/15, 8/25, 8/12, 14/16)$. Use a diagram to illustrate that $\tilde{\mu}_i$ can also be expressed as the left slope of a greatest convex minorant of a particular piece-wise linear function. [Hint: see Proposition 2.4.1].

3

Tests on Multivariate Normal Mean

3.1 INTRODUCTION

In this chapter we consider two broad classes of hypothesis testing problems when the population distribution is normal. For example, the observations may be independently and identically distributed from the multivariate normal distribution, $N(\boldsymbol{\theta}, \boldsymbol{V})$, or they may satisfy a linear model of the form $\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{E}$ with normal errors. For the standard problem of testing $H_0 : \boldsymbol{R\theta} = \boldsymbol{0}$ against $H_2 : \boldsymbol{R\theta} \neq \boldsymbol{0}$, where $\boldsymbol{R}$ is a given fixed matrix, it is easy to apply the likelihood ratio test because the (likelihood ratio) test statistic can be computed easily, and the statistical tables for its null distribution, χ_q^2 where $q = rank(\boldsymbol{R})$, are easily available. In what follows, we shall use the acronym LRT for the likelihood ratio statistic and the test as well.

If the hypotheses involve inequalities in $\boldsymbol{\theta}$, then the theory becomes more complicated, and several difficulties in applying the results are also encountered. One such difficulty is caused by the fact that the null distribution of LRT depends on the particular inequalities. For example, when $\boldsymbol{V}$ is known, the null distribution of LRT for testing

$$H_0 : \boldsymbol{R\theta} = \boldsymbol{0} \text{ against } H_1 : \boldsymbol{R\theta} \geqslant \boldsymbol{0}, \boldsymbol{R\theta} \neq \boldsymbol{0} \tag{3.1}$$

depends on the matrix $\boldsymbol{R}$ through $\boldsymbol{RVR}^T$, not just on $rank(\boldsymbol{R})$, as is the case for testing $H_0 : \boldsymbol{R\theta} = \boldsymbol{0}$ against $H_2 : \boldsymbol{R\theta} \neq \boldsymbol{0}$. Consequently, it is not practicable to have statistical tables of critical values because we would need one table for each $\boldsymbol{RVR}^T$ that may arise in practice, and there are infinite number of them. Another difficulty is that, except in some very special cases, it is difficult to compute the critical values exactly.

Fortunately, the p-values and critical values of the tests discussed in this chapter can be computed easily by simulation. A simple example illustrated this in the previous chapter. Its extension to the setting of this chapter is easy. The required computer programs can be developed so that they are applicable to most practical situations in which the constraints on θ are linear. The main requirement for this is a computer program for solving a quadratic program; such programs are available in IMSL, NAG and several other software packages. See Bazaraa et al. (1993) for relevant theory and references.

Sometimes, constrained inference problems may lead to unexpected results; the simple example in Silvapulle (1997a) illustrated this. For example, the sample means of a bivariate response may appear to suggest that a treatment performed substantially worse than the control, but a likelihood ratio test may conclude to accept that the treatment is better than the control. The crux of these issues are also discussed.

We adopt the general approach of computing the p-values approximately by simulation. For most practical purposes, the simulation approach is simple, adequate, and easy to implement.

The next section provides a discussion of the two types of testing problems that are discussed in this chapter, which we call Type A and Type B problems. Subsection 3.2.1 shows that for a large class of problems considered in this chapter, considerations may be restricted to a single observation from a normal distribution; the reasoning is based on sufficiency of the sample mean for a random sample from a multivariate normal. Some examples in two dimensions are discussed in detail in Section 3.3 to introduce the essential basics. A distribution, called chi-bar-square, arises naturally in Type A and Type B testing problems. This class of distributions is defined in Section 3.4. Computing the critical values of a chi-bar-square distribution is not as easy as reading a number from a table of critical values. In fact, except in some very special cases, it is virtually impossible to compute the tail probability of a chi-bar-square distribution using a hand calculator. Various results relating to the tail probability of a chi-bar-square distribution are discussed in Section 3.5. Then, in Sections 3.7 and 3.8 we consider the two types of testing problems in multivariate analysis with known covariance matrices. These results are extended in the following section to deal with the same type of testing problems in the linear model when the covariance matrix is $\sigma^2 U$ where σ^2 is unknown but U is known. Then, we consider the same two types of testing problems when the observations are from a multivariate normal and the covariance matrix is completely unknown. The proofs of several results are demanding, and therefore they are given in an Appendix to ensure that the main ideas and results are presented without the interruption of lengthy technical details.

3.2 STATEMENT OF TWO GENERAL TESTING PROBLEMS

Let us first define some terms. A set $\mathcal{A} \subset \mathbb{R}^p$ is said to be *convex* if $\{\lambda x + (1-\lambda)y\} \in \mathcal{A}$ whenever $x, y \in \mathcal{A}$ and $0 < \lambda < 1$. Therefore $\mathcal{A}$ is a convex set if the line segment joining x and y is in $\mathcal{A}$ whenever the points x and y are in $\mathcal{A}$. A set $\mathcal{A}$ is said to be a *cone with vertex* x_0 if $x_0 + k(x - x_0) \in \mathcal{A}$ for every $x \in \mathcal{A}$ and $k \geqslant 0$; if the vertex is the origin, then we shall simply refer to it as a *cone*. Therefore, a cone is a set that consists of infinite straight lines starting from the origin. A *polyhedral* $\mathcal{P}$ in $\mathbb{R}^p$ is a set of the form $\{\theta \in \mathbb{R}^p : a_1^T\theta \geqslant 0, \ldots, a_k^T\theta \geqslant 0\}$ where $a_1, \ldots, a_k$ are given elements of $\mathbb{R}^p$. Thus, $\mathcal{P}$ is the intersection of the *half-spaces*, $\{a_1^T\theta \geqslant 0\}, \ldots \{a_k^T\theta \geqslant 0\}$. It is easily seen that a polyhedral is a closed convex cone. A more detailed discussion of polyhedrals is given in the first Appendix to this chapter.

Notation: Throughout this chapter, we shall adopt the following notation without any further comment:
(a) C_a, C_b, C, and $\mathcal{M}$ are subsets of an Euclidean space, for example $\mathbb{R}^p$.
(b) C is a closed convex cone, $\mathcal{M}$ is a linear space, $\mathcal{M} \subset C$, and $C_a \subset C_b$.

We shall consider two types of inequality constrained testing problems. These are stated below in their most general forms, where $\theta \in \mathbb{R}^p$.

Type A : Test $H_0 : \theta \in \mathcal{M}$ against $H_1 : \theta \in C$,
Type B : Test $H_1 : \theta \in C$ against $H_2 : \theta \notin C$.

Clearly, the null hypothesis in Type A is of the form $R\theta = 0$, for some matrix R. Strictly speaking, the alternative hypothesis in the foregoing Type A problem should be stated as $H_1 : \theta \in C, \theta \notin \mathcal{M}$. However, we shall continue to adopt the convention introduced in the previous chapter (see page 28): *"Test of H_0 against H_1" is to be interpreted as "test of H_0 against $H_1 - H_0$."* Therefore, the foregoing Type A and Type B testing problems may also be stated in the following equivalent forms:

Type A : Test $H_0 : \theta \in \mathcal{M}$ against $H_1 : \theta \in C$ and $\theta \notin \mathcal{M}$.
Type B : Test $H_1 : \theta \in C$ against $H_2 : \theta \in \mathbb{R}^p$.

Some examples of Type A problems:

1. $H_0 : \theta = 0$ against $H_1 : \theta \geqslant 0$.
2. $H_0 : (\theta_a^T, \theta_b^T)^T = 0$ against $H_1 : \theta_a \geqslant 0$.
3. $H_0 : R\theta = 0$ against $H_1 : R_1\theta \geqslant 0$, where R_1 is a submatrix of R.
4. $H_0 : A(\theta) = 0$ against $H_1 : A(\theta)$ *is negative semi-definite*, where $\theta = (\theta_1, \theta_2, \theta_3)^T$ and $A(\theta)$ is the 2×2 matrix, $(\theta_1, \theta_3 \mid \theta_3, \theta_2)$; this is equivalent to $H_0 : (\theta_1, \theta_2, \theta_3) = 0$ against $H_1 : \theta_1 \leq 0, \theta_2 \leq 0, \theta_1\theta_2 - \theta_3^2 \geqslant 0$.

In the last example, although the parameter space $\{\theta : \theta_1 \leq 0, \theta_2 \leq 0, \theta_1\theta_2 - \theta_3^2 \geqslant 0\}$ involves nonlinear inequalities, it turns out to be a closed convex cone. In view of the convention stated above, $\theta \neq 0, (\theta_a^T, \theta_b^T)^T \neq 0$, $R\theta \neq 0$, and $A(\theta) \neq 0$ are assumed to be implicit in the alternative hypothesis of the foregoing examples (1), (2), (3), and (4), respectively.

Some examples of Type B problems:

1. $H_1 : \theta \geqslant 0$ against $H_2 : \theta \ngeqslant 0$.
2. $H_1 : \theta_a \geqslant 0$ and $\theta_b = 0$ against $H_2 : \theta \in \mathbb{R}^p$, where $\theta = (\theta_a^T, \theta_b^T)^T$.

3. $H_1 : \boldsymbol{R}_1\boldsymbol{\theta} \geqslant \mathbf{0}$ and $\boldsymbol{R}_2\boldsymbol{\theta} = \mathbf{0}$ against $H_2 : H_1$ *is not true*, where $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are some known fixed matrices.

4. $H_1 : \boldsymbol{A}(\boldsymbol{\theta})$ *is negative semi-definite* against $H_2 : H_1$ *is not true*, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ and $\boldsymbol{A}(\boldsymbol{\theta})$ is the 2×2 matrix $(\theta_1, \theta_3 \mid \theta_3, \theta_2)$; this is equivalent to $H_1 : \theta_1 \leq 0, \theta_2 \leq 0, \theta_1\theta_2 - \theta_3^2 \geqslant 0$, and H_2 : *no restriction on* $\boldsymbol{\theta}$.

These examples provide some idea of the type of testing problems studied in this chapter. The main difference between Type A and Type B problems is that inequalities may be present in the alternative hypothesis of Type A and the null hypothesis of Type B problems.

Here are some examples that are not Type A or Type B testing problems:

1. $H_0 : \theta_1 \leq 0$ or $\theta_2 \leq 0$ against $H_1 : \theta_1 > 0$ and $\theta_2 > 0$.

2. $H_0 : \boldsymbol{\theta} \geqslant \mathbf{0}$ or $\boldsymbol{\theta} \leq \mathbf{0}$ against $H_1 : \boldsymbol{\theta} \not\geqslant \mathbf{0}$ and $\boldsymbol{\theta} \not\leq \mathbf{0}$.

3. $H_0 : \theta_1 = \theta_2 = 0$ against $H_1 : \theta_2^2 \geqslant |\theta_1|$.

The first of these arises in clinical trials involving *combination drugs*; this testing problem is known as the *sign testing* problem. The second one arises in tests against *cross-over interactions* between treatments and groups. For the first two examples, some exact finite sample results are available. These will be studied in Chapter 9. The third is an artificial example; this example resembles a Type A problem but the alternative parameter space is not a closed convex cone. For such problems, large sample results will be studied in Chapter 4.

When the population distribution is $N(\boldsymbol{\theta}, \boldsymbol{V})$, the nature of the solutions to inequality constrained testing problems also depend on what is known about the covariance matrix $\boldsymbol{V}$ in addition to the structure of the null and alternative parameter spaces. We shall consider three cases: (i) $\boldsymbol{V}$ is a known positive definite matrix, (ii) $\boldsymbol{V} = \sigma^2\boldsymbol{U}$, where $\boldsymbol{U}$ is a known positive definite matrix and σ is unknown, and (iii) $\boldsymbol{V}$ is unknown. In the next chapter, we shall study asymptotic results when $\boldsymbol{V}$ is a function of some nuisance parameters.

Almost all of the results in this chapter are for Type A or Type B problems. Occasionally we shall consider test of $H_a : \boldsymbol{\theta} \in \mathcal{C}_a$ against $H_b : \boldsymbol{\theta} \in \mathcal{C}_b$ because it incorporates Type A and Type B problems and it is instructive to introduce some results within this general context.

3.2.1 Reduction by Sufficiency

Often, we shall consider test of hypotheses based on a single observation from a multivariate normal distribution. Although this case is considered for simplicity, it can also be justified using the sufficiency of sample mean. To illustrate this, let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n independent and identically distributed observations from $N(\boldsymbol{\theta}, \boldsymbol{V})$, where $\boldsymbol{V}$ is a known positive definite matrix and $\boldsymbol{\theta} \in \mathbb{R}^p$. Suppose that we wish to test

$$H_a : \boldsymbol{\theta} \in \mathcal{C}_a \text{ against } H_b : \boldsymbol{\theta} \in \mathcal{C}_b,$$

where, as was indicated earlier, $\mathcal{C}_a \subset \mathcal{C}_b$. In this case, the likelihood ratio test can be formulated in terms of a single observation of $\bar{\boldsymbol{X}}$, where $\bar{\boldsymbol{X}}$ is the sample mean. As is shown below, this is a consequence of the fact that $\bar{\boldsymbol{X}}$ is sufficient for $\boldsymbol{\theta}$.

Let $L(\boldsymbol{\theta})$ denote the loglikelihood for the n observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Since

$$L(\boldsymbol{\theta}) = -(n/2)^{-1} \log|\boldsymbol{V}| - (np/2)\log(2\pi) - 2^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{X}_i - \boldsymbol{\theta}),$$

we have that

$$L(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + g(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{V}),$$

where

$$\ell(\boldsymbol{\theta}) = -2^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\theta})^T(n^{-1}\boldsymbol{V})^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\theta})$$

and $g(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{V})$ does not depend on $\boldsymbol{\theta}$. Therefore, $\ell(\boldsymbol{\theta})$ is the *kernel* of the loglikelihood for $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. Now, the *LRT* for testing $H_a : \boldsymbol{\theta} \in \mathcal{C}_a$ against $H_b : \boldsymbol{\theta} \in \mathcal{C}_b$, based on $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ is

$$2[\max\{L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}_b\} - \max\{L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}_a\}].$$

Therefore, it follows that

$$LRT = 2[\max\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}_b\} - \max\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}_a\}]. \tag{3.2}$$

On the other hand, since $\bar{\boldsymbol{X}} \sim N(\boldsymbol{\theta}, n^{-1}\boldsymbol{V})$, the kernel of the loglikelihood for the single observation $\bar{\boldsymbol{X}}$ is $\ell(\boldsymbol{\theta})$. Therefore, the maximum likelihood estimator (*mle*) and the *LRT* based on $\bar{\boldsymbol{X}}$ and those based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are identical. Hence, in much of the developments in this chapter where the covariance matrix $\boldsymbol{V}$ is known, we will consider statistical inference based on a single observation that may be thought of as $\boldsymbol{X}$.

3.3 THEORY: THE BASICS IN TWO DIMENSIONS

Before we study general testing problems, it would be instructive to consider some simple special cases. We shall illustrate the derivation of the null distribution of the LRT for some special cases and then use the ideas developed to introduce the general results.

Example 3.3.1 *Maximum Likelihood Estimation*

Let $\boldsymbol{X} = (X_1, X_2)^T \sim N(\boldsymbol{\theta}, \boldsymbol{I})$, where $\boldsymbol{I}$ is the 2×2 identity matrix and $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$. Consider the maximum likelihood estimation of $\boldsymbol{\theta}$ based on one observation on $\boldsymbol{X}$ and subject to the constraint $\boldsymbol{\theta} \in \mathcal{C}$ where $\mathcal{C}$ is the nonnegative orthant $\{(\theta_1, \theta_2)^T : \theta_1 \geqslant 0, \theta_2 \geqslant 0\}$. For the single observation $\boldsymbol{X}$, the kernel $\ell(\boldsymbol{\theta})$ of the loglikelihood is given by

$$-2\ell(\boldsymbol{\theta}) = \{(X_1 - \theta_1)^2 + (X_2 - \theta_2)^2\} = \|\boldsymbol{X} - \boldsymbol{\theta}\|^2$$

where $\|\cdot\|$ is the Euclidean norm defined by $\|x\|^2 = (x_1^2 + x_2^2)$. Let $\tilde{\theta}$ be the *mle* of θ subject to $\theta \geqslant 0$. Since $-2\ell(\theta)$ is equal to the squared distance between X and θ, $\tilde{\theta}$ is the point in C that is closest to X; in other words, $\tilde{\theta}$ is the *projection* of X onto C (see Fig. 3.1). With Π denoting the projection function, we can write

$$\tilde{\theta} = \Pi(X \mid \mathbb{R}^{+2}).$$

Let Q_1, Q_2, Q_3, Q_4 denote the four quadrants in the two-dimensional plane as in Fig. 3.1. Then the *mle* $\tilde{\theta}$ is given by

$$(\tilde{\theta}_1, \tilde{\theta}_2) = \begin{cases} (X_1, X_2) & \text{if } X \in Q_1 \\ (0, X_2) & \text{if } X \in Q_2 \\ (0,0) & \text{if } X \in Q_3 \\ (X_1, 0) & \text{if } X \in Q_4. \end{cases} \tag{3.3}$$

■

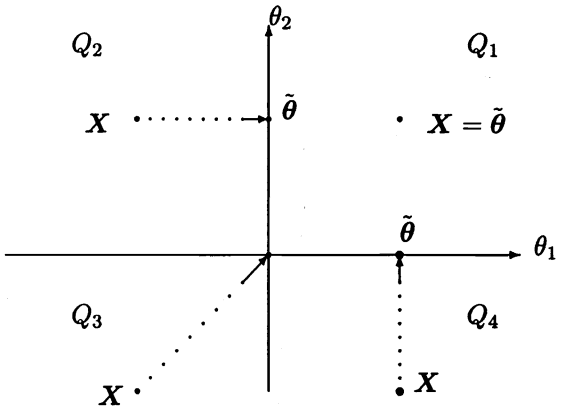

Fig. 3.1 The maximum likelihood estimator $\tilde{\theta}$ of θ subject to $\theta \geqslant 0$, based on a single observation of X, where $X \sim N(\theta, I)$.

Since the distribution function of X is known and $\tilde{\theta}$ is a function of X only, we can write down explicit expressions for the distribution of $\tilde{\theta}$. In contrast to the familiar situation where the parameter space for θ is $\mathbb{R}^2$, here we observe that the distribution of the constrained *mle*, $\tilde{\theta}$, is not normal. Further, the distribution of $(\tilde{\theta} - \theta)$ and that of any scaled form of it depend on θ. Therefore, we cannot construct a confidence region for θ in the usual way based on the distribution of $(\tilde{\theta} - \theta)$. Consequently, the explicit form of the distribution of $(\tilde{\theta} - \theta)$ has not found much use in statistical inference.

Example 3.3.2 *Type A testing problem (Example 3.3.1 continued).*

Let $X = (X_1, X_2)^T \sim N(\theta, I)$, where I is the 2×2 identity matrix and $\theta = (\theta_1, \theta_2)^T$. Consider the likelihood ratio test of

$$H_0 : \theta_1 = \theta_2 = 0 \text{ against } H_1 : \theta_1 \geqslant 0, \theta_2 \geqslant 0,$$

based on a single observation of X. Since $-2\ell(\theta) = \|X - \theta\|^2$ and $LRT = 2[\max\{\ell(\theta) : \theta \in H_1\} - \max\{\ell(\theta) : \theta \in H_0\}]$ we have that

$$LRT = \|X\|^2 - \|X - \tilde{\theta}\|^2 = \|\tilde{\theta}\|^2.$$

The last step follows since $(X - \tilde{\theta})^T \tilde{\theta} = 0$ as is clear from Figure 3.1; it may also be verified by considering each of the four cases in (3.3) (see Fig. 3.1). Since we are interested in the distribution of LRT under the null hypothesis, *let us suppose that the null hypothesis is true* for the rest of these derivations. To obtain an expression for $\text{pr}(LRT \leq c)$, note that

$$\begin{aligned} \text{pr}(LRT \leq c) &= \sum_{i=1}^{4} \text{pr}(LRT \leq c \text{ and } X \in Q_i) \\ &= \sum_{i=1}^{4} \text{pr}(LRT \leq c \mid X \in Q_i)\text{pr}(X \in Q_i). \end{aligned} \tag{3.4}$$

Let us evaluate each of the four summands in the last expression.

Using the circular symmetry of $N(0, I)$, it can be shown that the direction $(= X/\|X\|)$ and the length $(= \|X\|)$ of X are statistically independent (see Exercise 3.1). Consequently, the conditional distribution of $(X_1^2 + X_2^2)$, given that the direction of X is in Q_1, is the same as that of its unconditional distribution. Therefore,

$$\begin{aligned} \text{pr}(LRT \leqslant c \mid X \in Q_1) &= \text{pr}(X_1^2 + X_2^2 \leqslant c \mid X \in Q_1) \\ &= \text{pr}(X_1^2 + X_2^2 \leqslant c) = \text{pr}(\chi_2^2 \leqslant c). \end{aligned} \tag{3.5}$$

Now, consider the second summand in (3.4) :

$$\begin{aligned} \text{pr}(LRT \leq c \mid X \in Q_2) &= \text{pr}(X_2^2 \leq c \mid X_2 \geqslant 0, X_1 \leq 0) \\ &= \text{pr}(X_2^2 \leq c \mid X_2 \geqslant 0) \quad \text{since } X_1 \text{ and } X_2^2 \text{ are independent} \\ &= \text{pr}(X_2^2 \leq c) \quad \text{since } X_2 \text{ is symmetric} \\ &= \text{pr}(\chi_1^2 \leq c) \quad \text{since } X_2 \sim N(0,1). \end{aligned} \tag{3.6}$$

Similarly, $\text{pr}(LRT \leq c \mid X \in Q_4) = \text{pr}(\chi_1^2 \leq c)$. Therefore, we have that

$$LRT = \begin{cases} X_1^2 + X_2^2 & \text{given } X \in Q_1 \\ X_2^2 & \text{given } X \in Q_2 \\ 0 & \text{given } X \in Q_3 \\ X_1^2 & \text{given } X \in Q_4 \end{cases} \sim \begin{cases} \chi_2^2 \\ \chi_1^2 \\ 0 \\ \chi_1^2. \end{cases} \tag{3.7}$$

Now, it follows from (3.4)-(3.7) that the null distribution of the *LRT* is the following weighted sum of chi-square distributions:

$$\text{pr}(LRT \leq c) = (1/4) + (1/2)\text{pr}(\chi_1^2 \leq c) + (1/4)\text{pr}(\chi_2^2 \leq c) \tag{3.8}$$

for $c > 0$. For notational convenience, we define χ_0^2, the *chi-square distribution with zero degrees of freedom*, to be the distribution that takes the value zero with probability

one. With this notation, the foregoing result is usually written as

$$\text{pr}(LRT \leq c \mid H_0) = \sum_{i=0}^{2} w_i \text{pr}(\chi_i^2 \leq c),$$

where $(w_0, w_1, w_2) = (0.25, 0.5, 0.25)$. For later use, let us note that w_0, w_1, and w_3 are the probabilities that $\boldsymbol{X}$ falls in the cones Q_3, $Q_2 \cup Q_4$, and Q_1 respectively. It follows from (3.7) that the critical region, $\{LRT \geqslant c\}$, is the region to the upper-right of the curve $PQRS$ in Fig. 3.2; PQ is parallel to the x_1-axis, RS is parallel to x_2-axis and QR is a circular arc of radius $\sqrt{c}$. ■

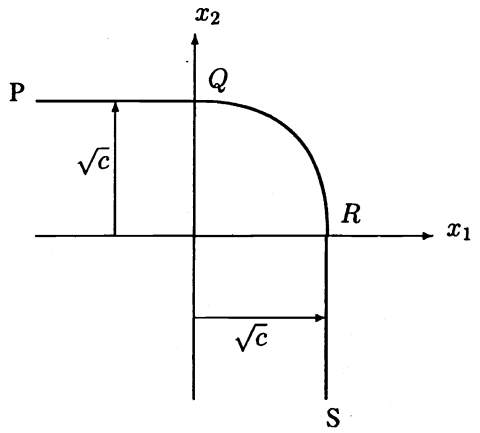

Fig. 3.2 The likelihood ratio test of $\boldsymbol{\theta} = \mathbf{0}$ against $\boldsymbol{\theta} \geqslant \mathbf{0}$ based on a single observation of $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$; the critical region, $\{\boldsymbol{x} : LRT \geqslant c\}$ is the upper-right region bounded by the curve $PQRS$.

The foregoing two examples show, in a very simple situation, how the constrained *MLE* over C behaves like a projection of a normal random variate onto C (see Fig. 3.1), and how a mixture of chi-square distributions arise when the alternative hypothesis involves inequality constraints. In fact, more general mixtures of chi-square distributions arise when the hypotheses involve other forms of inequalities. The ideas underlying such general results are introduced below using simple bivariate examples.

Example 3.3.3 *Likelihood ratio test of* $H_0 : \boldsymbol{\theta} = \mathbf{0}$ *against* $H_1 : \boldsymbol{R\theta} \geqslant \mathbf{0}$.

Let $\boldsymbol{X} = (X_1, X_2)^T \sim N(\boldsymbol{\theta}, \boldsymbol{I})$. Consider the LRT of

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \quad \text{against} \quad H_1 : \boldsymbol{R\theta} \geqslant \mathbf{0}$$

where $\boldsymbol{R}$ is the 2×2 nonsingular matrix $(1, 4 \mid 1, -2)$ (i.e., the two rows of $\boldsymbol{R}$ are $(1, 4)$ and $(1, -2)$ respectively). Let $C = \{\boldsymbol{\theta} : \boldsymbol{R\theta} \geqslant \mathbf{0}\}$. Then, C is a closed convex cone (see Fig. 3.3).

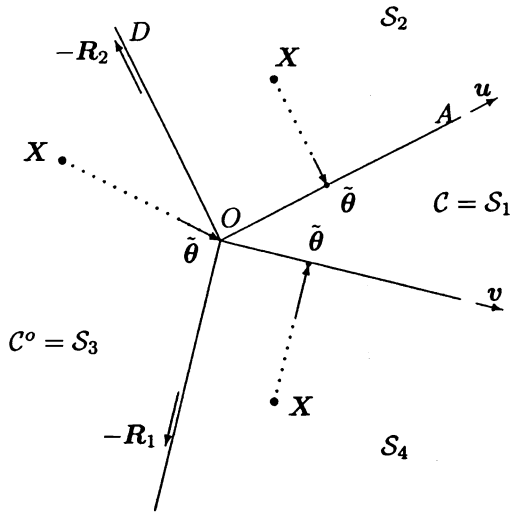

Fig. 3.3 The *mle* $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ based on a single observation of $\boldsymbol{X}$ when $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$ and $\boldsymbol{\theta}$ is constrained to lie in the cone C.

Let $C^o = \{\boldsymbol{\alpha} : \boldsymbol{\alpha}^T\boldsymbol{\theta} \leq 0 \text{ for every } \boldsymbol{\theta} \in C\}$; C^o is called the *negative dual* or *polar cone* of C with respect to the inner product, $\langle \boldsymbol{\alpha}, \boldsymbol{\theta} \rangle = \boldsymbol{\alpha}^T\boldsymbol{\theta} = \alpha_1\theta_1 + \alpha_2\theta_2$. Thus, C^o is the collection of vectors which do not form an acute angle with any vector in C. It may be verified that the boundaries of C^o are the perpendiculars to the boundaries of C, and that C^o is the closed convex cone formed by these perpendiculars (see Fig.3.3). Let $\boldsymbol{R}_1^T$ and $\boldsymbol{R}_2^T$ denote the two rows of $\boldsymbol{R}$. It may also be verified that $-\boldsymbol{R}_1$ and $-\boldsymbol{R}_2$ are parallel to the boundaries of C^o as shown in Fig. 3.3. Clearly, C and C^o partition the plane into 4 cones; let us denote these by $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4$ with $\mathcal{S}_1 = C$ and $\mathcal{S}_3 = C^o$ (see Fig. 3.3). The partition $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$ in this example corresponds to $\{\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3, \mathcal{Q}_4\}$ in Example 3.3.2.

Let $\boldsymbol{u}$ and $\boldsymbol{v}$ denote unit vectors parallel to the upper and lower boundaries of C. Since the kernel of the loglikelihood is equal to $-0.5\|\boldsymbol{X} - \boldsymbol{\theta}\|^2$, the *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \in C$ is the point in C closest to $\boldsymbol{X}$. In other words, the *mle* is the projection of $\boldsymbol{X}$ onto C, which we write as $\Pi(\boldsymbol{X} \mid C)$. As in Examples 3.3.1 and 3.3.2, we obtain the following for the *mle*:

$$\tilde{\boldsymbol{\theta}} = \begin{cases} \boldsymbol{X} & \text{given } \boldsymbol{X} \in \mathcal{S}_1 \\ (\boldsymbol{u}^T\boldsymbol{X})\boldsymbol{u} & \text{given } \boldsymbol{X} \in \mathcal{S}_2 \\ \mathbf{0} & \text{given } \boldsymbol{X} \in \mathcal{S}_3 \\ (\boldsymbol{v}^T\boldsymbol{X})\boldsymbol{v} & \text{given } \boldsymbol{X} \in \mathcal{S}_4. \end{cases} \tag{3.9}$$

By considering each of the cases, $\boldsymbol{X} \in \mathcal{S}_i, i = 1, 2, 3, 4$ separately, it may be verified that $(\boldsymbol{X} - \tilde{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}}$ are orthogonal when $\tilde{\boldsymbol{\theta}}$ is not zero; Fig. 3.3 shows this clearly. Therefore, it follows from (3.2) that

$$LRT = \|\boldsymbol{X}\|^2 - \|\boldsymbol{X} - \tilde{\boldsymbol{\theta}}\|^2 = \|\tilde{\boldsymbol{\theta}}\|^2.$$

To derive the null distribution of *LRT*, first note that

$$\text{pr}(LRT \leq c) = \sum_{i=1}^{4} \text{pr}(LRT \leq c \mid \boldsymbol{X} \in \mathcal{S}_i)\text{pr}(\boldsymbol{X} \in \mathcal{S}_i). \tag{3.10}$$

Suppose that H_0 holds. Then $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Therefore, the length and direction of $\boldsymbol{X}$ are independent (see Exercise 3.1). Hence $\|\boldsymbol{X}\|^2 \leqslant c$ and $\boldsymbol{X} \in \mathcal{S}_1$ are independent, and

$$\begin{aligned}\text{pr}(LRT \leq c \mid \boldsymbol{X} \in \mathcal{S}_1) &= \text{pr}(X_1^2 + X_2^2 \leq c \mid \boldsymbol{X} \in \mathcal{S}_1) \\ &= \text{pr}(X_1^2 + X_2^2 \leq c) = \text{pr}(\chi_2^2 \leq c).\end{aligned}$$

To obtain an expression for $\text{pr}(LRT \leq c \mid \boldsymbol{X} \in \mathcal{S}_2)$, note that we can think of OA and OD as the first and second axes. In this new orthogonal coordinate system, the *LRT* given $\boldsymbol{X} \in \mathcal{S}_2$ is the squared length of the first coordinate and hence is distributed as χ_1^2; essentially the same argument was used in the previous example leading to (3.6). Similarly, the *LRT* given $\boldsymbol{X} \in \mathcal{S}_4$ is distributed as χ_1^2. Thus, as in Examples 3.3.1 and 3.3.2, we obtain the following for the conditional null distribution of *LRT*:

$$LRT = \begin{cases} X_1^2 + X_2^2 & \text{given } \boldsymbol{X} \in \mathcal{S}_1 \\ (\boldsymbol{u}^T\boldsymbol{X})^2 & \text{given } \boldsymbol{X} \in \mathcal{S}_2 \\ 0 & \text{given } \boldsymbol{X} \in \mathcal{S}_3 \\ (\boldsymbol{v}^T\boldsymbol{X})^2 & \text{given } \boldsymbol{X} \in \mathcal{S}_4. \end{cases} \sim \begin{cases} \chi_2^2 \\ \chi_1^2 \\ \chi_0^2 \\ \chi_1^2. \end{cases} \tag{3.11}$$

Collecting these results and substituting in (3.10), we have that

$$\text{pr}(LRT \leq c \mid H_0) = q\text{pr}(\chi_0^2 \leq c) + 0.5\text{pr}(\chi_1^2 \leq c) + (0.5 - q)\text{pr}(\chi_2^2 \leq c), \tag{3.12}$$

where $q = \text{pr}(\boldsymbol{X} \in \mathcal{S}_3 \mid H_0) = \text{pr}(\boldsymbol{X} \in \mathcal{C}^o \mid H_0) = (2\pi)^{-1}\gamma$ and γ is the angle (in radians) of the cone $\mathcal{C}^o$ at its vertex; therefore,

$$q = (2\pi)^{-1}\cos^{-1}[\boldsymbol{R}_1^T\boldsymbol{R}_2/\{\boldsymbol{R}_1^T\boldsymbol{R}_1)(\boldsymbol{R}_2^T\boldsymbol{R}_2)\}^{1/2}], \tag{3.13}$$

where $\boldsymbol{R}_1^T$ and $\boldsymbol{R}_2^T$ are the two rows of $\boldsymbol{R}$. The critical region $\{LRT \geqslant c\}$ is the region to the right of the curve $PQRS$ in Figure 3.4; PQ is orthogonal to the upper boundary of $\mathcal{C}$, RS is orthogonal to the lower boundary of $\mathcal{C}$ and QR is a circular arc of radius $\sqrt{c}$. ■

Example 3.3.4 *Likelihood ratio test of* $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ *against* $H_1 : \boldsymbol{R\theta} \geqslant \boldsymbol{0}$ *when* $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ *and* $\boldsymbol{V}$ *is arbitrary.*

Let $\boldsymbol{X} = (X_1, X_2)^T \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V}$ is an arbitrary positive definite matrix. Consider the LRT of $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ where $\mathcal{C} = \{\boldsymbol{\theta} : \boldsymbol{R\theta} \geqslant \boldsymbol{0}\}$ is a closed convex cone and $\boldsymbol{R}$ is a 2×2 nonsingular matrix. We can transform this problem to the set up in Example 3.3.3. Let $\boldsymbol{V}^{-1}$ be factorized as $\boldsymbol{V}^{-1} = \boldsymbol{A}^T\boldsymbol{A}$, for some $\boldsymbol{A}$ (for example, use Cholesky decomposition; see Anderson (1984) page 586);

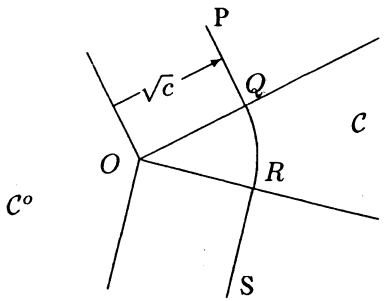

Fig. 3.4 The region to the right of $PQRS$ is the critical region $\{\boldsymbol{x} : \|\tilde{\boldsymbol{\theta}}(\boldsymbol{x})\|^2 \geqslant c\}$ for the *LRT* of $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ based on a single observation of $\boldsymbol{X}$ where $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$; $\tilde{\boldsymbol{\theta}}(\boldsymbol{x})$ is the *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \in \mathcal{C}$ when $\boldsymbol{X} = \boldsymbol{x}$, as shown in Figure 3.3

if $\boldsymbol{V} = [v_{ij}]$ then $\boldsymbol{A} = (v_{11}v_{22} - v_{12}^2)^{-1/2}[\sqrt{v_{22}}, -v_{12}/\sqrt{v_{22}} \mid 0, v_{11} - (v_{12}^2/v_{22})]$. Now, transform $\boldsymbol{X}$ and the associated quantities, $\boldsymbol{\theta}$ and $\mathcal{C}$, by $\boldsymbol{A}$:

$$\boldsymbol{Y} = \boldsymbol{AX}, \boldsymbol{\alpha} = \boldsymbol{A\theta}, \mathcal{P} = \boldsymbol{A}\mathcal{C}$$

where $\boldsymbol{A}\mathcal{C} = \{\boldsymbol{A\theta} : \boldsymbol{\theta} \in \mathcal{C}\}$. Then $\mathcal{P} = \{\boldsymbol{\alpha} : \boldsymbol{RA}^{-1}\boldsymbol{\alpha} \geqslant \boldsymbol{0}\}$, $\boldsymbol{Y} \sim N(\boldsymbol{\alpha}, \boldsymbol{I})$, and the testing problem is equivalent to the test of $H_0 : \boldsymbol{\alpha} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\alpha} \in \mathcal{P}$ based on $\boldsymbol{Y}$. This is precisely the set-up in the last example with $\{\boldsymbol{Y}, \mathcal{P}\}$ in place of $\{\boldsymbol{X}, \mathcal{C}\}$. Since the *mle* and *LRT* are invariant under nonsingular transformations, we have that

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{A}^{-1}\tilde{\boldsymbol{\alpha}} \quad \text{and} \quad LRT = \|\tilde{\boldsymbol{\alpha}}\|^2 = \tilde{\boldsymbol{\theta}}^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{\theta}}.$$

Now, by (3.13) and (3.12), we have that

$$\text{pr}(LRT \leq c \mid H_0) = q\text{pr}(\chi_0^2 \leq c) + 0.5\text{pr}(\chi_1^2 \leq c) + (0.5 - q)\text{pr}(\chi_2^2 \leq c), \tag{3.14}$$

where

$$q = (2\pi)^{-1}\cos^{-1}[(\boldsymbol{R}_1^T\boldsymbol{VR}_2)/\{(\boldsymbol{R}_1^T\boldsymbol{VR}_1)(\boldsymbol{R}_2^T\boldsymbol{VR}_2)\}^{1/2}], \tag{3.15}$$

and $\boldsymbol{R}_1^T$ and $\boldsymbol{R}_2^T$ are the two rows of $\boldsymbol{R}$. ■

This example shows that we do not have to consider the case for a general $\boldsymbol{V}$ separately because the inference problem can be restated in terms of the identity covariance matrix and therefore can be handled by the method for the previous example. However, it is instructive to review the details in terms of the original variables and parameters without using such a transformation because some important features tend to be masked by the transformation.

Geometry of *mle* and LRT when $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V}), \boldsymbol{V} \neq \boldsymbol{I}$.

Since $\boldsymbol{V}^{-1}$ is positive definite, it induces the inner product defined by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{V}} = \boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{y}$. This further induces the distance function defined by $\|\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{V}} = \{(\boldsymbol{x} -$

$y)^T V^{-1}(x-y)\}^{1/2}$. Two vectors x and y are said to be *orthogonal* with respect to $\langle\cdot\rangle_V$ if $\langle x, y\rangle_V = 0$ and denote it by $x \perp_V y$. If necessary, we shall use the abbreviation V-orthogonal, V-distance, etc. to indicate that the relevant inner product is $\langle\cdot\rangle_V$.

Let $\mathcal{B}$ be a given subset of $\mathbb{R}^2$ and let $\tilde{\theta}$ denote the *mle* of θ over $\theta \in \mathcal{B}$ based on a single observation of X where $X \sim N(\theta, V)$. Then, since $\ell(\theta) = (-1/2)(X-\theta)^T V^{-1}(X-\theta) = (-1/2)\|X-\theta\|_V^2$, it follows that $\tilde{\theta} = \arg\min_{\theta\in\mathcal{B}} \|X-\theta\|_V^2$ and hence $\tilde{\theta}$ is the point in $\mathcal{B}$ that is V-closest to X. In this case, we say that $\tilde{\theta}$ is the V-projection of X onto $\mathcal{B}$ and denote it by $\Pi_V(X \mid \mathcal{B})$ or equivalently by $\Pi_V(X, \mathcal{B})$. More details about such projections are given in the appendix to this chapter. The following result relating to the contours of $N(0, V)$ is useful.

Proposition 3.3.1 *The ellipse $x^T V^{-1} x = Const$ and any line through the center of the ellipse are V-orthogonal at their points of intersection.*

Proof: Let us write

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \text{ and hence } V^{-1} = \Delta^{-1}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

where $\Delta = \{\sigma_1^2\sigma_2^2(1-\rho^2)\}$. Now, the ellipse $x^T V^{-1} x = Constant$ can be expressed as

$$\sigma_2^2 x_1^2 - 2\rho\sigma_1\sigma_2 x_1 x_2 + \sigma_1^2 x_2^2 = Constant. \tag{3.16}$$

By implicit differentiation, we have that

$$dx_2/dx_1 = (\sigma_2/\sigma_1)(\rho\sigma_1 x_2 - \sigma_2 x_1)/(\sigma_1 x_2 - \rho\sigma_2 x_1); \tag{3.17}$$

this is the slope of the tangent to the contour (3.16) at (x_1, x_2). Now consider the line ℓ defined by $x_2 = m x_1$ (see Figure 3.5). For the time being, assume that $\sigma_1 m \neq \sigma_2\rho$ to avoid division by zero; this can be relaxed by taking the limit $\sigma_1 m \to \sigma_2\rho$.

Note that the slope, a, of the tangent to the contour (3.16) at the point of intersection with the line ℓ is given by

$$a = (\sigma_2/\sigma_1)(\rho\sigma_1 m - \sigma_2)/(\sigma_1 m - \rho\sigma_2). \tag{3.18}$$

Further, $(1, m)$, a vector parallel to the line ℓ, and $(1, a)$, a vector parallel to the tangent of the ellipse at the point of intersection with ℓ, are V-orthogonal. Therefore, the line ℓ and any contour (3.16) are V-orthogonal at their points of intersections. ■

To illustrate this further, let C_1 and C_2 be two contours defined by $x^T V^{-1} x = c_1^2$ and $x^T V^{-1} x = c_2^2$ respectively, for some c_1 and c_2 (see Fig. 3.5). Let O denote the origin and, as shown in Fig. 3.5, let F and Q be the points at which ℓ intersects C_1 and C_2, respectively. Let T_Q be the tangent to C_2 at Q; similarly, let T_F be the tangent to C_1 at F. Let P and E be points on T_Q and T_F respectively. Now, we have the following:

1. For a given i, $(i = 1, 2)$, every point on C_i is at V-equidistant from O.

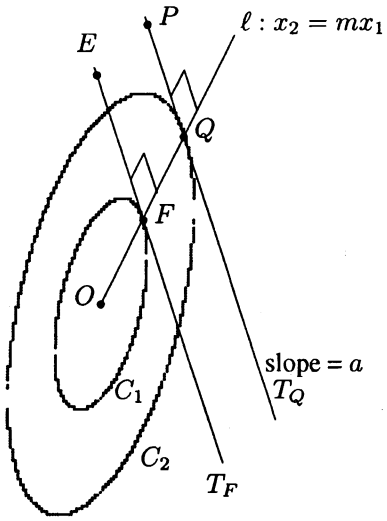

Fig. 3.5 V-projection onto a radial line ℓ through the center O of the ellipses C_1 and C_2; the point on ℓ that is V-closest to P is Q; the point on T_F that is V-closest to O is F.

2. T_Q and T_F are parallel and their common slope a is given by (3.18).

3. T_Q and T_F are V-orthogonal to ℓ.

By Pythagoras theorem, the point on a line that is V-closest to a point, say P, is obtained by dropping a V-perpendicular to the line from P. Therefore, in Fig. 3.5 the point on ℓ that is V-closest to P is Q; the point on T_Q that is V-closest to O is also Q. If $X \sim N(\theta, V)$, the observed value of X is at O, and θ is restricted to T_Q then the *mle* of θ is at Q; if the observation is at P and θ is restricted to ℓ then the *mle* of θ is at Q.

The foregoing are also helpful to illustrate the form of the *mle* of θ when it is constrained to lie in a cone. Let $X \sim N(\theta, V)$ and let us consider the *mle* of θ based on a single observation of X when θ is constrained to lie on a convex cone. Let C be the convex cone AOB in Fig. 3.6 and suppose that θ is constrained to lie in it. Fig. 3.6 shows the *mle* when θ is restricted to C and Fig. 3.7 shows the critical region for testing $H_0 : \theta = 0$ against $H_1 : \theta \in C$. Let $OC \perp_V OA$ and $OD \perp_V OB$. Then, COD is the polar cone of C with respect to $\langle,\rangle_V$. Let Q and R be the points of intersection of an arbitrary contour ($x^T V^{-1} x = constant$) with OA and OB, respectively. Let PQ and SR be the tangents to the contour at Q and R, respectively. Thus, $PQRS$ is a smooth curve with continuous slope everywhere. If $X \in AOB$ then $\tilde{\theta} = X$; if $X \in COA$, say $X = \overline{OP}$, then $\tilde{\theta} = \overline{OQ}$; if $X \in DOC$ then $\tilde{\theta} = 0$; and if $X \in DOB$, say $X = \overline{OS}$, then $\tilde{\theta} = \overline{OR}$. Now, the boundary of a typical critical region is $PQRS$, where the QR is the segment of the contour C (ie. $x^T V^{-1} x = constant$) that lies in AOB.

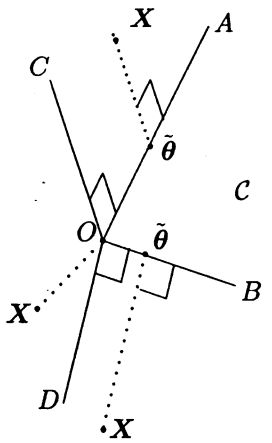

Fig. 3.6 The constrained *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \in C$ when $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$.

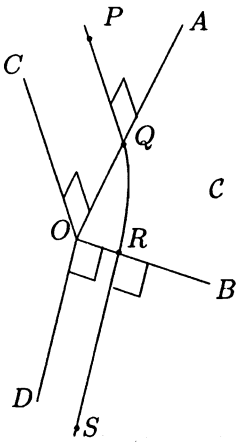

Fig. 3.7 The *LRT* of $H_0 : \boldsymbol{\theta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\theta} \in C$ when $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ and $\boldsymbol{V} \neq \boldsymbol{I}$; a typical boundary of the critical region is $PQRS$

Example 3.3.5 *A numerical example: likelihood ratio test against inequality constraints*

Let $\boldsymbol{X}$ be a bivariate random variable distributed as $N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V} = (1, \rho; \rho, 1)$ with $\rho = 0.9$. Now, consider the estimation of $\boldsymbol{\theta}$ and test of $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$ based on the following five observations on $\boldsymbol{X}$: (-4,-3), (-4, -3), (-3, -2), (-2, -1), and (-2, -1). For these data, the mean $\bar{\boldsymbol{x}}$ is (-3, -2); as was pointed out in Section 3.2.1, the estimation and tests depend on the data only through this mean because $\boldsymbol{V}$ is known. The kernel, $\ell(\boldsymbol{\theta})$, of the loglikelihood is $\{-0.5n(\bar{\boldsymbol{X}} - \boldsymbol{\theta})^T \boldsymbol{V}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\theta})\}$. To estimate $\boldsymbol{\theta}$, we minimize $(\bar{\boldsymbol{x}} - \boldsymbol{\theta})^T \boldsymbol{V}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\theta})$ over $\{\theta_1 \geqslant 0, \theta_2 \geqslant 0\}$; this achieves its minimum at $(0, 0.7)$. This may be verified by using the geometric method in the previous example. Alternatively, since the dimension of $\boldsymbol{\theta}$ is only two, the minimum can be computed as explained in the next paragraph; while we would not use this method in high dimensions, it is instructive in two dimensions.

Note that $(\bar{\boldsymbol{x}} - \boldsymbol{\theta})^T \boldsymbol{V}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\theta})$ is proportional to $f(\theta_1, \theta_2)$ defined by

$$f(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2 - 1.8\theta_1\theta_2 + 2.4\theta_1 - 1.4\theta_2 + 2.2.$$

This function is convex because its second derivative is proportional to $\boldsymbol{V}^{-1}$, which is positive definite. The minimum of $f(\boldsymbol{\theta})$, without any restrictions on (θ_1, θ_2), is achieved at $\boldsymbol{\theta} = \bar{\boldsymbol{x}} = (-3, -2)$, which is outside the region $\{\boldsymbol{\theta} \geq \mathbf{0}\}$. Because $f(\boldsymbol{\theta})$ is convex, the minimum of $f(\boldsymbol{\theta})$ subject to $\{\boldsymbol{\theta} \geq \mathbf{0}\}$ is achieved on the boundary of $\{\boldsymbol{\theta} \geq \mathbf{0}\}$, which consists of the nonnegative θ_1- and θ_2-axes. Now, to find the absolute minimum of $f(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \geq \mathbf{0}$, it suffices to find the minima of $f(\boldsymbol{\theta})$ on the nonnegative segments of the θ_1- and θ_2-axes separately; the lowest of the two minima is the absolute minimum of $f(\boldsymbol{\theta})$ subject to $\boldsymbol{\theta} \geqslant \mathbf{0}$.

The null distribution of LRT is given by (see (3.14) and (3.15)):

$$pr(LRT \geq c \mid H_0) = 0.5pr(\chi_1^2 \geq c) + 0.5(1 - \pi^{-1}\cos^{-1}\rho)pr(\chi_2^2 \geq c).$$

Substituting $\rho = 0.9$, the sample value of the likelihood ratio statistic, $n\tilde{\boldsymbol{\theta}}^T \boldsymbol{V}^{-1}\tilde{\boldsymbol{\theta}}$, is 12.9 for which the p-value is

$$0.5\text{pr}(\chi_1^2 \geqslant 12.9) + 0.5(1 - \pi^{-1}\cos^{-1} 0.9)\text{pr}(\chi_2^2 \geqslant 12.9),$$

which is less than 0.05. ■

Now, let us consider a curious phenomenon of the LRT based on the foregoing example.

Example 3.3.6 *A curious example involving the LRT; the data are inconsistent with the the null and the alternative hypotheses (Silvapulle (1997a))*

Consider an experiment to establish that a new treatment for pain relief is better than a placebo. For each patient, let X_1 and X_2 denote the reductions in pain at two locations (say neck and back) due to the treatment. For illustrative purposes, suppose that the distribution of $\boldsymbol{X}$ and the data are the same as in the previous example. Suppose also that the new treatment is expected to be at least as good as the placebo in controlling pain. Thus, it is assumed *a priori* that $\boldsymbol{\theta} \geqslant \mathbf{0}$. Suppose that the inference problem for establishing that the new treatment is better than the placebo be formulated as test of $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$. Since every observed value is negative, the pain has increased under the new treatment for each patient in the sample; clearly, the new treatment has performed worse than the placebo. For these data, the *mle* of (θ_1, θ_2) under $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$ is $(0, 0.7)$ and the LRT of H_0 vs H_1 would reject H_0 in favor of H_1 at 5% level (LRT=12.9, p-value < 0.05).

At first, it would appear that there is something wrong because there is no way that we could claim that the new treatment is better than the placebo when, in fact, the pain under the new treatment turned out to be worse for each patient. In fact, there is nothing wrong as far as the statistical calculations are concerned. However, several issues arise in the interpretation of the results.

To provide an insight into why the *mle* of θ_2 is positive and the LRT rejects H_0, first recall that the maximum likelihood estimator and the likelihood ratio statistic are based on the distribution of $\bar{\boldsymbol{X}}$ being $N(\boldsymbol{\theta}, n^{-1}\boldsymbol{V})$. The contours of $N(\boldsymbol{\theta}, n^{-1}\boldsymbol{V})$ passing through the observed value $\bar{\boldsymbol{x}} = (-3, -2)$ for $\boldsymbol{\theta} = (0, 0)$ and $\boldsymbol{\theta} = (0, 0.7)$ are shown in Fig. 3.8; these are indicated by C_2 and C_1, respectively. Note that C_1 is smaller than C_2. Therefore, if $g(\boldsymbol{x}|\boldsymbol{\theta})$ denotes the density function of $\bar{\boldsymbol{X}}$ then $g((-3, -2)|(0, 0.7)) > g((-3, -2)|(0, 0))$. Thus, the likelihood at $\boldsymbol{\theta} = (0, 0.7)$ is larger than that at $\boldsymbol{\theta} = (0, 0)$. Therefore, the data are more likely to have arisen from $N(\boldsymbol{\theta}, \boldsymbol{V})$ with $\boldsymbol{\theta} = (0, 0.7)$ than $\boldsymbol{\theta} = \mathbf{0}$.

A small p-value for $H_0 : \boldsymbol{\theta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$ says that the data are considerably more inconsistent with $\boldsymbol{\theta} = \mathbf{0}$ than with $\boldsymbol{\theta} \gtrsim \mathbf{0}$. If we accept the formulation of the problem as given above, then there are no contradictions; the LRT does exactly what we expect it to do.

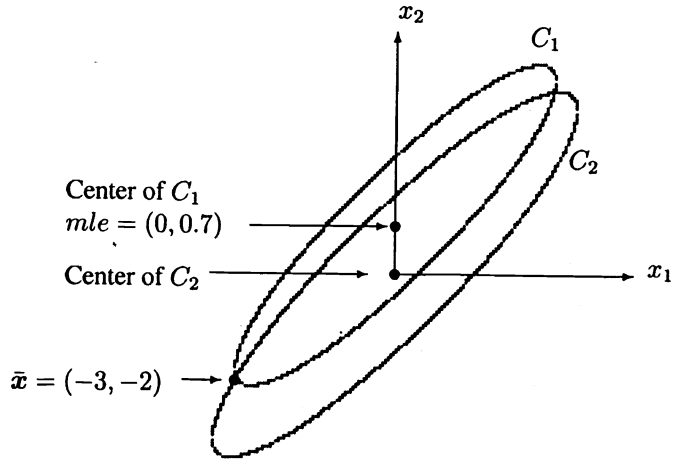

Fig. 3.8 The *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \geqslant \mathbf{0}$ when $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$, $\boldsymbol{V} = (1, 0.9 \mid 0.9, 1)$ and the sample mean is $(-3, -2)$.

Up to this point there are no difficulties in the statistical calculations. However, recently, there have been debates about the interpretation of these results. If the assumption $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V} = (1, 0.9|0.9, 1)$ is correct and $\bar{\boldsymbol{x}} = (-3, -2)^T$, should we accept that the new treatment is better? Some authors argue that *LRT* is deficient, while others argue there is nothing wrong with *LRT*. It is not possible to discuss all the relevant issues concerning this. Our objective here has been to draw attention to the phenomenon. We refer the readers to Silvapulle (1997a), Cohen et al. (2000), Cohen and Sackrowitz (2004), Perlman and Wu (2002b), and Chaudhuri and Perlman (2004). Although these authors discuss different examples, the statistical issues therein are the same as those in the foregoing example.

If we wish to establish that the new treatment is better than the placebo, then the testing problem could have been formulated as $H_0 : \boldsymbol{\theta} \notin A$ vs $H_1 : \boldsymbol{\theta} \in A$ where A is the set of all values of $\boldsymbol{\theta}$ that correspond to the treatment being better than the placebo. For example, one possible choice is $A = \{\theta_1 > 0, \theta_2 > 0\}$; for this particular choice of A, if the test rejects H_0, then clearly there is statistical evidence not only to reject H_0 but also to support H_1. This particular testing problems will be discussed in later chapters under the topic, ***sign testing***. More generally, in a test of H_a vs H_b for some H_a and H_b, if we wish to interpret the rejection of H_a as statistical evidence in support of H_b, it may be necessary to ensure that $H_a \cup H_b$ is the full parameter space. ■

The phenomenon in the simple example just discussed manifests in other settings as well; see Problem 3.9 for an example. In the analysis of ordinal response data, this

property of LRT translates to it being not *concordant monotone*; for a discussion of this see Cohen et al. (2000). For further discussions and different points, see Cohen and Sackrowitz (2004), Chaudhuri and Perlman (2004), and Perlman and Wu (1999, 2002b).

It is possible to construct tests of $\boldsymbol{\theta} = \mathbf{0}$ vs $\boldsymbol{\theta} \geqslant \mathbf{0}$ in which points such as (-3,-2) do not lie in the critical region. Cohen and Sackrowitz (1998) introduced a class of tests that they called the *Directed Tests* for testing this type of hypotheses. They developed their ideas on precise mathematical formulations of the problem; this will be introduced in Chapter 9. However, at the time of writing this book, there is a continuing debate concerning the advantages and disadvantages of various tests for constrained inference problems, in particular, as to whether or not *LRT* is deficient, and as to whether or not the directed tests are themselves deficient (see the references in the previous paragraph).

3.4 CHI-BAR-SQUARE DISTRIBUTION

The family of chi-bar-square distributions, to be introduced in this section, plays an important role when the null and/or alternative hypothesis involve parameter inequalities. The null distributions of likelihood ratio statistics for Type A and Type B problems with multivariate normal data turn out to be chi-bar-square. The role of chi-bar-square distributions in Type A and Type B testing problems is similar to that of chi-square distributions in the application of likelihood ratio for testing $\boldsymbol{\theta} = \mathbf{0}$ against $\boldsymbol{\theta} \neq \mathbf{0}$ in small and in large samples. In this section, we introduce the general form of chi-bar-square distribution and show how it relates to an important fundamental result on likelihood ratio test.

Let $C \subset \mathbb{R}^p$ (C is a closed convex cone) and let $\boldsymbol{Z}_{p\times 1} \sim N(\mathbf{0}, \boldsymbol{V})$, where $\boldsymbol{V}$ is a positive definite matrix. We define $\bar{\chi}^2(\boldsymbol{V}, C)$ to be the random variable, which has the same distribution as $[\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z} - \min_{\boldsymbol{\theta}\in C}(\boldsymbol{Z}-\boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}-\boldsymbol{\theta})]$. If there is no possibility of any ambiguity, we shall also write

$$\bar{\chi}^2(\boldsymbol{V}, C) = \boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z} - \min_{\boldsymbol{\theta}\in C}(\boldsymbol{Z}-\boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}-\boldsymbol{\theta}). \tag{3.19}$$

It would be instructive to provide a geometric interpretation of $\bar{\chi}^2$ and its related quantities. Fig. 3.9 provides a schematic diagram of C and a typical value of $\boldsymbol{Z}$ represented by OA. Let the inner product be defined as $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\boldsymbol{V}} = \boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{y}$, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$; this induces the norm $\|\boldsymbol{x}\|_{\boldsymbol{V}} = \{\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{x}\}^{1/2}$ and the distance $\|\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{V}} = \{(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{y})\}^{1/2}$ between $\boldsymbol{x}$ and $\boldsymbol{y}$. Let B be the point in C that is $\boldsymbol{V}$-closest to A, and let $\tilde{\boldsymbol{Z}}$ denote the vector OB. Therefore, $\tilde{\boldsymbol{Z}}$ is the value of $\boldsymbol{x}$ at which

$$(\boldsymbol{Z}-\boldsymbol{x})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}-\boldsymbol{x}), \quad \boldsymbol{x} \in C$$

is the minimum. In other words, $\tilde{\boldsymbol{Z}}$ is the $\boldsymbol{V}$-projection of $\boldsymbol{Z}$ onto C; let us denote it by $\Pi_{\boldsymbol{V}}(\boldsymbol{Z}|C)$ or equivalently $\Pi_{\boldsymbol{V}}(\boldsymbol{Z}, C)$. It is shown in the Appendix that $\boldsymbol{Z} - \tilde{\boldsymbol{Z}}$ is $\boldsymbol{V}$-orthogonal to $\tilde{\boldsymbol{Z}}$; in other words, OB is $\boldsymbol{V}$-orthogonal to AB. Therefore, by

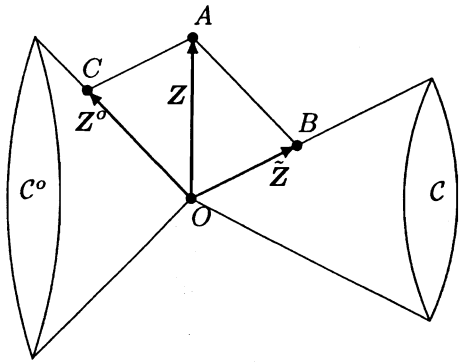

Fig. 3.9 V-projection: OB and OC are the V-projections of OA onto C and C^o respectively.

applying the Pythagoras theorem to OAB, we have that

$$\|OA\|_V^2 = \|OB\|_V^2 + \|BA\|_V^2,$$

and hence

$$Z^T V^{-1} Z = \tilde{Z}^T V^{-1} \tilde{Z} + \min_{x \in C} (Z - x)^T V^{-1} (Z - x).$$

Therefore,

$$\bar{\chi}^2(V, C) = \|OB\|_V^2.$$

We denote the V-distance between the point Z and the set C by $\|Z - C\|_V$ and it is defined by

$$\|Z - C\|_V^2 = \inf\{(Z - x)^T V^{-1} (Z - x) : \quad x \in C\};$$

in Fig. 3.9, $\|Z - C\|_V$ is equal to $\|AB\|_V$.

The *polar cone* C^o of C with respect to the inner product $\langle x, y \rangle_V = x^T V^{-1} y$ is defined by

$$C^o = \{x : x^T V^{-1} y \leq 0 \text{ for every } y \in C\}.$$

Because C is a closed convex cone, it may be verified easily that C^o is also a closed convex cone. It consists of the vectors that form obtuse angles with every vector in C; here obtuse means that $x^T V^{-1} y \leq 0$.

Let Z^o denote the point in C^o that is closest to Z. In other words, Z^o is the projection of Z onto C^o and is defined as the value of x at which

$$(Z - x)^T V^{-1} (Z - x), \quad x \in C^o$$

is a minimum. Let C in C^o be the point such that $Z^o = OC$. Again, the Pythagoras theorem is applicable to OAC. In fact, $OBAC$ is a rectangle, and hence

$$OC = BA, CA = OB, \|OA\|_V^2 = \|OC\|_V^2 + \|CA\|_V^2 = \|OB\|_V^2 + \|BA\|_V^2.$$

It would also be useful to note that

$$\bar{\chi}^2(V, C^o) = \|OC\|_V^2 = \|Z - C\|^2.$$

These results are summarized below.

Proposition 3.4.1 *Let V be a $p \times p$ positive definite matrix. Then*

1. $\|Z\|_V^2 = \|\Pi_V(Z \mid C)\|_V^2 + \|Z - \Pi_V(Z \mid C)\|_V^2$.
2. $\Pi_V(Z \mid C^o) = Z - \Pi_V(Z \mid C)$.
3. *If $Z \sim N(0, V)$ then*

$$\|\Pi_V(Z \mid C)\|_V^2 \sim \bar{\chi}^2(V, C) \quad \text{and} \quad \|Z - \Pi_V(Z \mid C)\|_V^2 \sim \bar{\chi}^2(V, C^o). \blacksquare$$

Now we have the following fundamental result; proof is given in the Appendix.

Theorem 3.4.2 *Let C be a close convex cone in $\mathbb{R}^p$ and V be a $p \times p$ positive definite matrix. Then the distribution of $\bar{\chi}^2(V, C)$ is given by*

$$pr\{\bar{\chi}^2(V, C) \leq c\} = \sum_{i=0}^{p} w_i(p, V, C) pr(\chi_i^2 \leq c) \tag{3.20}$$

where $w_i(p, V, C), i = 0, \ldots, p$ are some nonnegative numbers and $\sum_{i=0}^{p} w_i(p, V, C) = 1$. $\blacksquare$

More details about the quantities $w_i(p, V, C)\}$ and their computation are discussed in the next section. Several proofs of Theorem 3.4.2 have appeared in the literature for progressively more general cases. The proofs in Gourieroux et al. (1982) and Shapiro (1985) are noteworthy. The proof in Gourieroux et al. (1982) is for the special case when $C = \{\theta : R\theta \geqslant 0\}$ and R is a full row-rank matrix. This proof is instructive because it is a direct extension of the ideas presented earlier. The proof in Shapiro (1985) is for general closed convex cones and is the most general known; further, Shapiro's proof is based on results relating to projections on to polyhedrals. The proof given in the Appendix to this chapter is based on the proof due to Shapiro (1985). The special case of the theorem for $p = 2$ was discussed earlier in Section 3.3. The main idea underlying the proof for the general case is similar although the technical details are more involved.

In the special case of two dimensions, we saw that the sum of the three weights is 1. Theorem 3.4.2 shows that this generalizes to the case in (3.20). Thus the right-hand side of (3.20) is a weighted mean of several tail probabilities of χ^2-distributions and hence is known as a *chi-bar-square distribution.* We shall refer to $\{w_i(p, V, C)\}$ as *chi-bar-square weights* or simply as *weights.* Another term used for these weights is *level probabilities.*

Now, suppose that we wish to test $H_0 : \theta = 0$ against $H_1 : \theta \in C$ based on a single observation of the p-dimensional vector X where $X \sim N(\theta, V)$ and V is a

known positive definite matrix. It follows from the definition, that the LRT takes the form

$$\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} - \min\{(\boldsymbol{X}-\boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{X}-\boldsymbol{\theta}) : \boldsymbol{\theta} \in C\}, \tag{3.21}$$

which is the expression on the right-hand side of (3.19) with $\boldsymbol{X}$ in place of $\boldsymbol{Z}$. Therefore, the distribution of LRT under $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ is $\bar{\chi}^2(\boldsymbol{V}, C)$. In view of this, (3.21) is sometimes called a $\bar{\chi}^2$-statistic. It is worth noting that the $\bar{\chi}^2$-statistic is based on principles of generalized least squares, and therefore it is a reasonable test statistic even if the distribution of $\boldsymbol{X}$ is not normal. The null distribution of the $\bar{\chi}^2$-statistic in (3.21) is a $\bar{\chi}^2$-distribution if $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{V})$.

3.5 COMPUTING THE TAIL PROBABILITIES OF CHI-BAR-SQUARE DISTRIBUTIONS

As was indicated earlier, the null distribution of several test statistics for or against inequality constraints turn out to be $\bar{\chi}^2$. Therefore, we need to be able to compute its tail probability to obtain the p-value and/or critical value. This would be easy if the chi-bar-square weights $\{w_i\}$ are known. Unfortunately, the *exact* computation of $\{w_i\}$ is quite difficult in general. The difficulties in computing $\{w_i\}$ may give the false impression that likelihood ratio tests of hypotheses for Type A and Type B testing problems are too complicated to implement. However, we can compute the tail probability of a chi-bar-square distribution by simulation; this is easy to use and fast. Therefore, for most practical needs, the fact that w_i has a complicated form does not cause any serious difficulties. Later, we will see situations where the chi-bar-square weights depend on unknown nuisance parameters; in those cases our inability to compute the weights accurately and fast may cause difficulties.

Simulation 1: To compute pr$\{\bar{\chi}^2(\boldsymbol{V}, C) \geqslant c\}$

(1) Generate $\boldsymbol{Z}$ from $N(\boldsymbol{0}, \boldsymbol{V})$. (2) Compute $\bar{\chi}^2(\boldsymbol{V}, C)$ in (3.19). (3) Repeat the first two steps N times (say, $N = 10000$). (4) Estimate pr$\{\bar{\chi}^2(\boldsymbol{V}, C) \geqslant c\}$ by (n/N) where n is the number of times $\bar{\chi}^2(\boldsymbol{V}, C)$ in the second step turned to be greater than or equal to c. ■

If C involves only linear constraints, then a quadratic program can be used for computing $\bar{\chi}^2$ (see Section 2.1.2.2). We can also try to compute the tail probability of a $\bar{\chi}^2$ distribution by computing the weights first. However, the weights of a chi-bar-square distribution are likely to be required only for computing the tail probability; if this is the case, then it is likely to be just as easy to compute the tail probability directly by the foregoing simulation method.

Now, let us consider the case when

$$C = \{\boldsymbol{\theta} \in \mathbb{R}^p : \boldsymbol{a}_i^T\boldsymbol{\theta} \geqslant 0, i = 1, \ldots, k\}. \tag{3.22}$$

If the weights $w_i(p, \boldsymbol{V}, C\}$ are of interest, then formula (3.29) given in the next section can be used for a simulation-based method for computing them. The simulation steps are easy to implement. It requires only a quadratic program. Such programs are

available in several software packages; a simple and easy to implement algorithm for solving quadratic programs is also given in Wollan and Dykstra (1987).

Simulation 2: To compute $w_i(p, \boldsymbol{V}, C)$ when C is the polyhedral in (3.22)

(1) Generate $\boldsymbol{Z}$ from $N(\boldsymbol{0}, \boldsymbol{V})$. (2) Compute $\tilde{\boldsymbol{Z}}$, the point at which $(\boldsymbol{Z}-\boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}-\boldsymbol{\theta})$ is a minimum over $\boldsymbol{\theta} \in C$. (3) Let $J = \{j : \boldsymbol{a}_j^T\tilde{\boldsymbol{Z}} = 0\}$ and $\phi = \{\boldsymbol{\theta} : \boldsymbol{a}_j^T\boldsymbol{\theta} = 0$ for every $j \in J\}$ and s be the dimension of ϕ; compute s. (4) Repeat the previous steps N times (say $N = 10000$). (5) Estimate $w_i(p, \boldsymbol{V}, C)$ by the proportion of times s turned out to be equal to i, $(i = 1, \ldots, p)$. ■

For more details and proofs relating to the foregoing simulation method see Silvapulle (1996) and Shapiro (1985).

The case when C is the nonnegative orthant $\mathbb{R}^{+p}$ arises frequently in inequality constrained statistical inference. When the $\boldsymbol{a}_j$'s defining the polyhedral $C = \{\boldsymbol{x} : \boldsymbol{a}_j^T\boldsymbol{x} \geqslant 0\}$ are linearly independent, a result in the next section shows that $\{w_i(p, \boldsymbol{V}, C)\}$ can be obtained from $\{w_i(k, \boldsymbol{W}, \mathbb{R}^{+k})\}$ for some $k \times k$ positive definite matrix $\boldsymbol{W}$. Thus, the case when the constrained set is the nonnegative orthant is of particular interest. If the weights take the form $w_i(k, \boldsymbol{W}, \mathbb{R}^{+k})$, then the foregoing Simulation 2 reduces to the following simpler form:

Simulation 3: To compute $w_i(p, \boldsymbol{W}, \mathbb{R}^{+p}), i = 1, \ldots, p$

(1) Generate $\boldsymbol{Z}$ from $N(\boldsymbol{0}, \boldsymbol{W})$. (2) Compute $\tilde{\boldsymbol{Z}}$, the point at which $(\boldsymbol{Z}-\boldsymbol{\theta})^T\boldsymbol{W}^{-1}(\boldsymbol{Z}-\boldsymbol{\theta})$ is the minimum over $\boldsymbol{\theta} \geqslant \boldsymbol{0}$. (3) Count the number of positive components of $\tilde{\boldsymbol{Z}}$ (this is equal to s in Step 3 of Simulation 2). (4) Repeat the previous three steps N times (say $N = 10000$). (5) Estimate $w_i(p, \boldsymbol{W}, \mathbb{R}^{+p})$ by the proportion of times $\tilde{\boldsymbol{Z}}$ turned out to have exactly i positive components, $i = 1, \ldots, p$. ■

This approach for the case when the polyhedral is $\mathbb{R}^{+k}$ appears in Wolak (1987). The general idea of using simulation to compute the weights has been around for sometime, for example, see Perlman (1969) and Gourieroux et al. (1982). The authors have a FORTRAN program that implements the foregoing procedure using subroutines from IMSL; this works quite fast.

If $pr\{\bar{\chi}^2(\boldsymbol{V}, C) \geqslant c\}$ is very small, then Simulation 1 would require a large number of simulations. In this case, we can overcome the problem by computing the weights of the chi-bar-square distribution separately, for example, by simulation. Once this is done, the tail probabilities of the chi-square distributions in $\sum w_i(p, \boldsymbol{V}, C)\text{pr}(\chi_i^2 \geqslant c)$ may be computed using methods that are already available. The advantage of this method is that the value of c has no bearing on $w_i(p, \boldsymbol{V}, C)$ and standard statistical software packages have programs for computing $\text{pr}(\chi_i^2 \geqslant c)$ very accurately for large values of c.

In most practical applications, C is a polyhedral. However, for completeness, let us also consider the case when C is not a polyhedral. In this general case, closed-form expressions for $w_i(p, \boldsymbol{V}, C)$ are quite complicated. They can be expressed as some integrals (see Kuriki (1993), and Lin and Lindsay (1997)). At this stage, it is unclear whether or not such integral representations can be used for computing them. However, a simulation approach can be used for computing the weights even in this general case.

Simulation 4: To compute $w_i(p, V, C)$ when C is not necessarily a polyhedral

(1) Let $0 < c_1 < \ldots < c_{p+1} < \infty$ and $a_{ij} = \text{pr}(\chi_i^2 \leq c_{j+1})$ for $i, j = 0, \ldots, p$. (2) Generate Z from $N(\mathbf{0}, V)$, and compute $\bar{\chi}^2(V, C)$ in (3.19). (3) Repeat the previous step a large number of times and let p_j be the proportion of times $\bar{\chi}^2(V, C)$ computed in the previous step fell in the interval $(0, c_{j+1}), j = 0, \ldots, p$. (4) Estimate $w_j(p, V, C)$ by solving the system of $(p+1)$ equations $p_j = a_{0j}w_0 + \ldots + a_{pj}w_p$, $(j = 0, \ldots, p)$. ■

For most of the Type A and Type B testing problems, the tail probability P of the likelihood ratio statistic takes the form

$$P = \sum_{i=0}^{p} w_i(p, V, C)\text{pr}(\chi_{i+k}^2 \geqslant c),$$

for some k. Since $\text{pr}(\chi_i^2 \geqslant c) \leq \text{pr}(\chi_j^2 \geqslant c)$ for $i < j$, $\sum_0^p w_i(p, V, C) = 1$ and $0 \leq w_i(p, V, C) \leq (1/2)$ by part 4 in Proposition 3.6.1, we have the following lower and upper bounds for P :

$$(1/2)\{\text{pr}(\chi_k^2 \geqslant c) + \text{pr}(\chi_{k+1}^2 \geqslant c)\} \leq P \leq (1/2)\{\text{pr}(\chi_{p+k-1}^2 \geqslant c) + \text{pr}(\chi_{p+k}^2 \geqslant c)\}. \tag{3.23}$$

These bounds are easy to compute and may turn out to be adequate for practical purposes in some cases; for example, if the upper bound is small then there may not be any need to compute the exact p-value to reject the null hypothesis. It is worth pointing out that these may be crude bounds if p is large.

The proof of Theorem 3.4.2 shows that $\text{pr}\{\bar{\chi}^2(V, C) \geqslant c\}$ and $w_i(p, V, C)$ are continuous in (V, C). Therefore, if C is close to $\mathbb{R}^{+p}$ and V is close to I, then $w_i(p, V, C)$ is close to $w_i(p, I, \mathbb{R}^{+p})$, and an approximation to $\text{pr}(\bar{\chi}^2(V, C) \geqslant c)$ is $\sum_{i=0}^p 2^{-p}p!/\{i!(p-i)!\}\text{pr}(\chi_i^2 \geqslant c)$ because $w_i(p, I, \mathbb{R}^{+p}) = 2^{-p}p!/\{i!(p-i)!\}$ by part 2 of Proposition 3.6.1. It may be verified that this approximation lies between the lower and the upper bounds given in (3.23). This provides a quick and rough estimate of the tail probability $\text{pr}(\bar{\chi}^2(V, C) \geqslant c)$. The closer the (V, C) to $(I, \mathbb{R}^{+p})$, the better the approximation; for some comparisons on this, see Piegorsch (1990).

The following closed-form expressions for $w_i(p, V, \mathbb{R}^{+p})$ are useful.

1. Let $p = 1$. Then

$$w_0(1, V) = w_1(1, V) = 0.5. \tag{3.24}$$

Therefore, $\text{pr}(\bar{\chi}^2 \geqslant t) = 0.5\text{pr}(\chi_1^2 \geqslant t)$ for $t > 0$. This is related to the familiar result where the p-value for a two-sided alternative concerning a scalar parameter is equal to twice the p-value for the corresponding one-sided alternative.

2. Let $p = 2$. Then (see Example 3.3.4)

$$\begin{aligned} &w_0(2, V) = (1/2)\pi^{-1}\cos^{-1}(\rho_{12}), \qquad w_1(2, V) = (1/2), \text{ and} \\ &w_2(2, V) = (1/2) - (1/2)\pi^{-1}\cos^{-1}(\rho_{12}) \end{aligned} \tag{3.25}$$

where ρ_{12} is the correlation coefficient $v_{12}\{v_{11}v_{22}\}^{-1/2}$.

3. Let $p = 3$. Then, (for example, see Kudo, 1963, pp 414 - 415)

$$\begin{aligned} w_3(3, V) &= (4\pi)^{-1}(2\pi - \cos^{-1}\rho_{12} - \cos^{-1}\rho_{13} - \cos^{-1}\rho_{23}), \\ w_2(3, V) &= (4\pi)^{-1}(3\pi - \cos^{-1}\rho_{12.3} - \cos^{-1}\rho_{13.2} - \cos^{-1}\rho_{23.1}), \\ w_1(3, V) &= (1/2) - w_3(3, V), \\ w_0(3, V) &= (1/2) - w_2(3, V), \end{aligned} \tag{3.26}$$

where ρ_{ij} is the correlation coefficient $v_{ij}\{v_{ii}v_{jj}\}^{-1/2}$, and $\rho_{ij.k}$ is the partial correlation coefficient $(\rho_{ij} - \rho_{ik}\rho_{jk})/\{(1-\rho_{ik}^2)(1-\rho_{jk}^2)\}^{1/2}$.

4. Let $V = I$. Then, by part 2 of Proposition 3.6.1,

$$w_i(p, I) = 2^{-p}p!/\{i!(p-i)!\} \text{ for every } p \text{ and } i. \tag{3.27}$$

When $p = 4$, all the weights can be expressed as functions of $w_4(4, V)$ and V (for example, see Wolak (1987)). However, numerical integration or Monte Carlo simulation needs to be used to compute $w_4(4, V)$. Therefore, these formulas may not be that helpful, because it may be just as easy to compute all the weights or the tail probability of the $\bar{\chi}^2$ distribution by simulation.

The simulation approaches discussed in this section require a computer program for computing $\min\{(Z-\theta)^T V^{-1}(Z-\theta) : \theta \in C\}$. If C involves only linear inequalities, we can use a standard quadratic program (see Section 2.1.2.2). If C involves nonlinear inequalities in θ, then a general-purpose nonlinear optimization program may be used; such programs are available in IMSL, MATLAB, and NAG. If C involves nonlinear inequality constraints and minimization of $(Z-\theta)^T V^{-1}(Z-\theta)$ subject to the constraint $\theta \in C$ is difficult to be incorporated in a simulation program, then another approach worth considering is to compute bounds for the critical values and/or p-values, for example, by the following method. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two polyhedrals, such that $\mathcal{P}_1 \subset C \subset \mathcal{P}_2$. Since $\bar{\chi}^2(V, \mathcal{P}_1) \leqslant \bar{\chi}^2(V, C) \leqslant \bar{\chi}^2(V, \mathcal{P}_2)$ we have that

$$\text{pr}\{\bar{\chi}^2(V, \mathcal{P}_1) \geqslant c\} \leqslant \text{pr}\{\bar{\chi}^2(V, C) \geqslant c\} \leqslant \text{pr}\{\bar{\chi}^2(V, \mathcal{P}_2) \geqslant c\}.$$

This approach is useful for obtaining lower and upper bounds for the tail probability of a chi-bar-square distribution; this method was used in Silvapulle (1992b).

3.8 RESULTS ON CHI-BAR-SQUARE WEIGHTS

Often, the constraints defining C are linear and independent. In this case, $w_i(p, V, C)$ can be expressed as $w_i(k, W, \mathbb{R}^{+k})$ for some W and k. Therefore, to simplify notation, we denote $w_i(p, V, C)$ by $w_i(p, V)$ when C is the positive orthant:

$$w_i(p, V) = w_i(p, V, \mathbb{R}^{+p}). \tag{3.28}$$

The main results concerning $w_i(p, V, C)$ are stated in this section; the proofs are not given here but may be found in Kudo (1963), Perlman (1969), Wolak (1987b), Shapiro (1985, 1988), and Silvapulle (1996a).

Proposition 3.6.1 *Let C be a closed convex cone in $\mathbb{R}^p$ and V be a $p \times p$ nonsingular covariance matrix. Then we have the following:*

1. *Let $C = \{x : a_i^T x \geqslant 0, i = 1, \ldots, k\}$ for some $a_1, \ldots, a_k$, $J = \{j : a_j^T \Pi_V(Z|C) = 0\}$, and $\phi = \{\theta : a_j^T \theta = 0, \forall j \in J\}$. Then*

$$w_i(p, V, C) = pr(\text{The linear space } \phi \text{ is of dimension } i). \tag{3.29}$$

This can be restated as follows: Suppose that C is a polyhedral, $Z \sim N(0, V)$ and let $F(Z, V, C)$ denote the face of C such that $\Pi_V(Z|C)$ lies in the relative interior of $F(Z, V, C)$; for a definition of a face of a cone and relative interior of a face see the Appendix to this chapter (page 123). Then

$$w_i(p, V, C) = pr\{\text{dimension of the linear space spanned by } F(Z, V, C) \text{ is } i\}.$$

2. *Let $Z \sim N(0, V)$ and C be the nonnegative orthant. Then, $w_i(p, V) = pr\{\Pi_V(Z|C)$ has exactly i positive components$\}$.*
3. $\sum_0^p (-1)^i w_i(p, V, C) = 0$.
4. $0 \leq w_i(p, V, C) \leq 0.5$.
5. *Let C^o denote the polar cone, $\{x \in \mathbb{R}^p : x^T V^{-1} y \leq 0$ for every $y \in C\}$, of C with respect to the inner product $\langle x, y \rangle = x^T V^{-1} y$. Then $w_i(p, V, C^o) = w_{p-i}(p, V, C)$.*
6. *Let $C = \{\theta \in \mathbb{R}^p : R\theta \geqslant 0\}$ where R is a $p \times p$ nonsingular matrix. Then $\bar{\chi}^2(V, C) = \bar{\chi}^2(RVR^T, \mathbb{R}^{+p})$ and $w_i(p, V, C) = w_i(p, RVR^T)$.*
7. $w_i(p, V) = w_{p-i}(p, V^{-1})$.
8. *Let $C = \{\theta \in \mathbb{R}^p : R\theta \geqslant 0\}$ where R is a $k \times p$ matrix of rank $k (\leq p)$. Then*

$$w_{p-k+i}(p, V, C) = \begin{cases} w_i(k, RVR^T) & \text{for } i = 0, \ldots, k \\ 0 & \text{otherwise} \end{cases}$$

9. *Let $C = \{\theta \in \mathbb{R}^p : R_1\theta \geqslant 0, R_2\theta = 0\}$ where R_1 is $s \times p$, R_2 is $t \times p$, $s + t \leq p$, $[R_1^T, R_2^T]$ is of full rank, and*

$$A = R_1 V R_1^T - (R_1 V R_2^T)(R_2 V R_2^T)^{-1}(R_2 V R_1^T).$$

Then

$$w_{p-s-t+j}(p, V, C) = \begin{cases} w_j(s, A) & \text{for } j = 0, \ldots, s \\ 0 & \text{otherwise.} \end{cases}$$

10. *Let R be a $r \times p$ matrix of rank r, R_1 be a $q \times p$ submatrix of R, $\mathcal{M} = \{\beta : R\beta = 0\}$, $C = \{\beta : R_1\beta \geqslant 0\}$, and $\mathcal{M}^\perp$ be the orthogonal complement of $\mathcal{M}$ with respect to $\langle x, y \rangle_V = x^T V^{-1} y$. Then*

$$w_{r-q+j}(p, V, C \cap \mathcal{M}^\perp) = \begin{cases} w_j(q, R_1 V R_1^T)) & \text{for } j = 0, \ldots, q \\ 0 & \text{otherwise} \end{cases}$$

11. *Let C denote the correlation matrix corresponding to V. Then $\bar{\chi}^2(V, C) = \bar{\chi}^2(C, C)$ and $w_i(p, V, C) = w_i(p, C, C)$ for every i.*

Proof: For (1) see Shapiro (1985); (2) follows from (1) but it has also been known for quite some time (for example, see Perlman (1969)). See (5.3), (5.4), (5.5), (5.8) and (5.10) in Shapiro (1988) for (6), (7), (8), (9), and (10) respectively. ■

Kudo (1963) showed that $w_i(p, V) = \sum p\{(\Lambda_N)^{-1}\} p(\Lambda_{M:N})$ where $M \subset \{1, \ldots, p\}$, the sum extends over all M with exactly i elements, $N = \{1, \ldots, p\} \setminus M$, Λ_M is the covariance matrix of the vector $\{Z_i : i \in M\}$, $\Lambda_{M:N}$ is the same under $Z_j = 0$ for $j \notin M$, and $p(A)$ is $\text{pr}(X \geqslant 0)$ when $X \sim N(0, A)$. Thus, $w_i(p, V)$ can be expressed in terms of orthant probabilities. Schervish (1983) provided a program for computing probabilities of rectangles corresponding to multivariate normal; this can also be used for computing the orthant probabilities. Bohrer and Chow (1978) provided a computer program based on this result for computing $w_i(p, V)$, for $p \leq 10$. Another FORTRAN program is given in Sun (1988a). Kudo (1963) also discussed its use for computations. Formulas for the chi-bar square weights when V is a diagonal matrix are given in Section 2.4 of Robertson et al. (1988) for some special cases.

The following useful corollary to part 5 of the foregoing proposition follows from Theorem 3.4.2; it says that when C is replaced by its polar cone the weights appear in the reverse order.

Corollary 3.6.2 *Let C and V be as in Proposition 3.6.1. Then*

$$pr\{\bar{\chi}^2(V, C^o) \leq c\} = \sum_{i=0}^{p} w_{p-i}(p, V, C) pr(\chi_i^2 \leq c). \quad ■$$

3.7 LRT FOR TYPE A PROBLEMS: V IS KNOWN

Section 3.3 provided detailed discussions on the distribution of *LRT* in two dimensions, and the discussions therein showed how chi-bar-square distributions arise. The main result in Section 3.4 showed that the *LRT* for testing $\theta = 0$ against $\theta \in C$ based on a single observation of X, where $X \sim N(\theta, V)$, is the $\bar{\chi}^2$ statistic

$$X^T V^{-1} X - \min_{\theta \in C} (X - \theta)^T V^{-1} (X - \theta)$$

and that its null distribution is the chi-bar-square distribution

$$\sum_{i=0}^{p} w_i(p, V, C) \text{pr}(\chi_i^2 \leq c).$$

A similar result holds even if the null parameter space, $\{0\}$, is replaced by a linear space; this section provides a discussion of such results.

Let $\mathcal{M}$ be a linear space contained in C. Suppose that we are interested to test

$$H_0 : \theta \in \mathcal{M} \text{ against } H_1 : \theta \in C$$

based on a single observation of $\boldsymbol{X}$, where $\boldsymbol{\theta}$ is a location parameter of $\boldsymbol{X}$, for example, it could be the mean of $\boldsymbol{X}$. A least squares statistic for testing H_0 against H_1 is

$$L_A = \min\{q(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{M}\} - \min\{q(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{C}\}, \tag{3.30}$$

where

$$q(\boldsymbol{a}) = (\boldsymbol{X} - \boldsymbol{a})^T \boldsymbol{V}^{-1} (\boldsymbol{X} - \boldsymbol{a})$$

and $\boldsymbol{V}$ is a known positive definite matrix. If $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ then L_A is the *LRT* for testing H_0 against H_1. A general result on the distribution of *LRT* for a Type A testing problem is given in the next result.

Theorem 3.7.1 *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V}$ is a positive definite matrix. Then, the LRT for the Type A testing problem, $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, is similar*[1] *and its null distribution is given by*

$$pr(LRT \le c \mid H_0) = \sum_{i=0}^{p} w_i(p, \boldsymbol{V}, \mathcal{C} \cap \mathcal{M}^{\perp}) pr(\chi_i^2 \le c) \tag{3.31}$$

where $\mathcal{M}^{\perp} = \{\boldsymbol{x} : \boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{y} = 0$ for every $\boldsymbol{y} \in \mathcal{M}\}$ is the orthogonal complement of $\mathcal{M}$ with respect to the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{V}} = \boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{y}$.

Proof: The proof of this theorem is based on deeper results about projections onto to convex cones. These are discussed in the Appendix to this chapter. Here, we shall provide an intuitive explanation of the main ideas after a short proof.

It follows from parts (h) and (i) of Proposition 3.12.6 in the Appendix that

$$\begin{aligned} LRT &= \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|_{\boldsymbol{V}}^2 : \boldsymbol{\theta} \in \mathcal{M}\} - \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|_{\boldsymbol{V}}^2 : \boldsymbol{\theta} \in \mathcal{C}\} \\ &= \|\boldsymbol{X}\|_{\boldsymbol{V}}^2 - \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|_{\boldsymbol{V}}^2 : \boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{M}^{\perp}\}. \end{aligned}$$

The last expression on the right-hand side is also the *LRT* for testing $\boldsymbol{\theta} = \boldsymbol{0}$ against $\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{M}^{\perp}$. Therefore, the proof follows from Theorem 3.4.2. ■

Figure 3.10 provides an intuitive picture of the ideas underlying the foregoing proof in three-dimensions. Let O be the origin and OR be the θ_1 axis; θ_2 and θ_3 axes are not shown. Let $\mathcal{M}$ be the linear space spanned by the line OR; thus, $\mathcal{M}$ is simply the θ_1 axis. The rectangle $OPQR$ spans a plane; similarly, the rectangle $OSTR$ spans another plane. Let $\mathcal{C}$ be the convex cone generated by these two planes as shown in the figure; thus, the cone $\mathcal{C}$ has a wedge shape with its *spine* being OR. Clearly, $\mathcal{M}^{\perp}$ is the plane perpendicular to the θ_1 axis and passes through O. Therefore, $\mathcal{M}^{\perp}$ is spanned by the θ_2-axis and the θ_3-axis, and $\mathcal{C} \cap \mathcal{M}^{\perp}$ is the cone SOP. Let OA represent the vector $\boldsymbol{X}$. Let B be the projection of A onto $\mathcal{C}$; more precisely, let OB be the projection of OA onto $\mathcal{C}$. Suppose that B is an interior point of OPQR. Then B is also the projection of A onto the plane spanned by $OPQR$. Therefore,

[1]Here, similar means that the null distribution of the test statistic is the same at every point in the null parameter space.

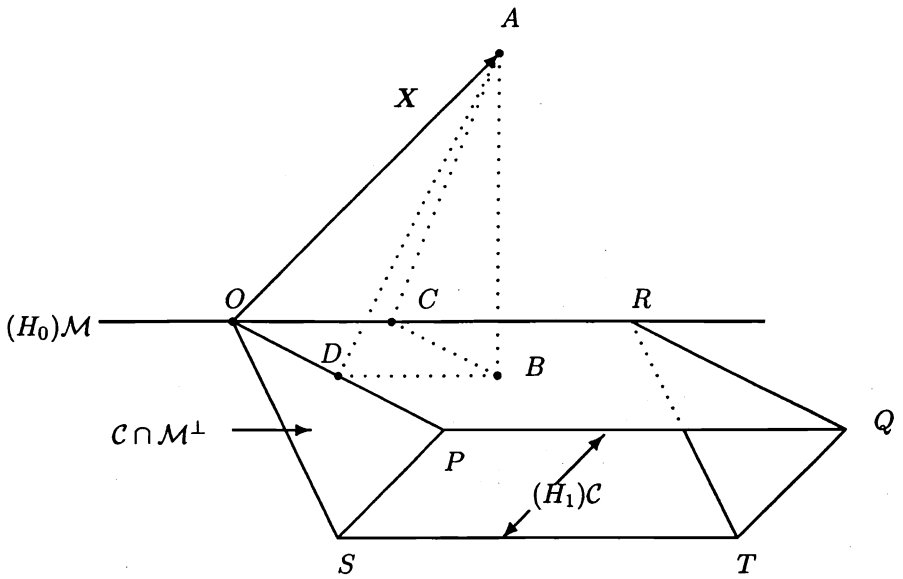

Fig. 3.10 Projections onto $\mathcal{C}$, $\mathcal{M}$, and $\mathcal{C} \cap \mathcal{M}^{\perp}$ where $\mathcal{M} \subset \mathcal{C}$.

AB is orthogonal to $OPQR$. Let D and C be the projections of A onto $\mathcal{C} \cap \mathcal{M}^{\perp}$ and $\mathcal{M}$ respectively. Then D and C are also the projections of B onto $\mathcal{C} \cap \mathcal{M}^{\perp}$ and $\mathcal{M}$ respectively. Now, we have the following :

$$AB \perp BD, AB \perp BC, AC \perp OC, AD \perp OD.$$

It follows that

$$\begin{aligned} &LRT \text{ of } \boldsymbol{\theta} \in \mathcal{M} \text{ vs } \boldsymbol{\theta} \in \mathcal{C} \\ &\quad = \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|_{\boldsymbol{V}}^2 : \boldsymbol{\theta} \in \mathcal{M}\} - \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|_{\boldsymbol{V}}^2 : \boldsymbol{\theta} \in \mathcal{C}\} \\ &\quad = \|AC\|^2 - \|AB\|^2 = \|BC\|^2 = \|DO\|^2 = \|OA\|^2 - \|AD\|^2 \\ &\quad = \|\boldsymbol{X}\|_{\boldsymbol{V}}^2 - \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|_{\boldsymbol{V}}^2 : \boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{M}^{\perp}\}. \end{aligned}$$

This is also the *LRT* for testing $\boldsymbol{\theta} = \boldsymbol{0}$ against $\boldsymbol{\theta} \in \mathcal{C} \cap \mathcal{M}^{\perp}$. Since $\mathcal{C} \cap \mathcal{M}^{\perp}$ is a closed convex cone, its null distribution is a chi-bar-square and is given by (3.31).

In general, the weights of the chi-bar-square distribution in (3.31) depend on the parameter spaces $\mathcal{M}$ and $\mathcal{C}$. There is no easy way to compute these weights for an arbitrary $\{\mathcal{C}, \mathcal{M}, \boldsymbol{V}\}$. The simulation procedure indicated in Section 3.5 is available as a general purpose procedure for computing $\text{pr}(LRT \le c \mid H_0)$. Fortunately, often in practice, the constraints defining $\mathcal{C}$ turn out to be linear and independent. In this case, we have the following simpler result, which follows from Proposition 3.6.1.

Corollary 3.7.2 *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V}$ is a positive definite matrix, $\boldsymbol{R}$ be a matrix of order $r \times p$, $rank(\boldsymbol{R}) = r \le p$, and let $\boldsymbol{R}_1$ be a submatrix of $\boldsymbol{R}$ of order*

$q \times p$. *Let the null and alternative hypothesis be* $H_0 : \boldsymbol{R\theta} = \mathbf{0}$ *and* $H_1 : \boldsymbol{R}_1\boldsymbol{\theta} \geqslant \mathbf{0}$ *respectively. Then, the LRT is similar and its null distribution is given by*

$$pr(LRT \leq c|H_0) = \sum_{i=0}^{q} w_i(q, \boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_1^T) pr(\chi^2_{r-q+i} \leq c) \qquad \blacksquare$$

Note that the chi-bar-square weights appearing in this corollary take the simpler form $w_i(q, \boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_1^T)$, which by definition is equal to $w_i(q, \boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_1^T, \mathbb{R}^{+q})$. Further, the number of terms in the foregoing $\bar{\chi}^2$ distribution depends on the number of inequalities in H_1 only, not on the dimension p of $\boldsymbol{\theta}$. If the alternative hypothesis has only independent linear inequalities, and the number of them does not exceed 3, then we can use the explicit formulas for weights given in (3.24), (3.25), and (3.26). If the alternative hypothesis does not have any inequality constraints then $q = 0$ and hence we recover the classical result, $\text{pr}(LRT \leq c|H_0) = \text{pr}(\chi^2_r \leq c)$.

If $\boldsymbol{X}$ is not normal but the distribution of $\boldsymbol{X} - \boldsymbol{\theta}$ does not depend on any unknown parameters, then L_A in (3.30) is still a suitable statistic for testing $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$; it is a generalized least squares based statistic. For this we have the following:

Proposition 3.7.3 *Suppose that the distribution of* $\boldsymbol{X} - \boldsymbol{\theta}$ *does not depend on any unknown parameters. Then the null distribution of* L_A *in* (3.30) *does not depend on the particular value of* $\boldsymbol{\theta}$ *in the null parameter space; ie. the test is similar.*

Proof: It may be verified easily that $\mathcal{C} + \boldsymbol{\alpha} = \mathcal{C}$ for any $\boldsymbol{\alpha} \in \mathcal{M}$, where $\mathcal{C} + \boldsymbol{\alpha}$ is defined as $\{\boldsymbol{x}+\boldsymbol{\alpha} : \boldsymbol{x} \in \mathcal{C}\}$; and $\mathcal{M}+\boldsymbol{\alpha} = \mathcal{M}$ for any $\boldsymbol{\alpha} \in \mathcal{M}$. Let $\boldsymbol{\theta}$ be an element of $\mathcal{M}$ and let $\boldsymbol{Y} = (\boldsymbol{X} - \boldsymbol{\theta})$. Then $L_A = \min\{q^*(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{M}\} - \min\{q^*(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{C}\}$, where $q^*(\boldsymbol{a}) = (\boldsymbol{Y} - \boldsymbol{a})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{a})$, and the distribution of $\boldsymbol{Y}$ does not depend on any unknown parameters. Therefore, if H_0 is true then the distribution of L_A is the same at every value in the null parameter space, $\mathcal{M}$. $\blacksquare$

In view of the foregoing Proposition, even if $\boldsymbol{X}$ is not normal, we can talk about the null distribution of L_A without specifying the null value of $\boldsymbol{\theta}$, and for the purposes of computing the p-value/critical value of L_A, we may assume that $\boldsymbol{\theta} = \mathbf{0}$. With t denoting the observed value of L_A, we have that

$$p\text{-value} = \text{pr}(L_A \geqslant t \mid H_0) = \text{pr}(L_A \geqslant t \mid \boldsymbol{\theta} = \mathbf{0}). \tag{3.32}$$

A consequence of this is that, even if $\boldsymbol{X}$ is not normal but the distribution of $\boldsymbol{X} - \boldsymbol{\theta}$ is known and does not depend on any unknown parameters, then exact finite sample p-value corresponding to L_A can be computed by simulation with pseudo-random observations generated at $\boldsymbol{\theta} = \mathbf{0}$.

The foregoing results are useful in large samples even if the population distribution is not normal. Let us suppose that $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independently and identically distributed as $\boldsymbol{Z}$, with $E(\boldsymbol{Z}) = \boldsymbol{\theta}$ and $\text{cov}(\boldsymbol{Z}) = \boldsymbol{W}$; it is assumed that the distribution of $\boldsymbol{Z}$ may not be normal, $\boldsymbol{W}$ is known, and that the distribution of $(\boldsymbol{Z} - \boldsymbol{\theta})$ does not depend on any unknown parameters. Suppose that we are interested to test $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$. Then, with $\boldsymbol{X}$ denoting $n^{-1}\sum \boldsymbol{Z}_i$ and assuming

that n is large, we have that $\boldsymbol{X}$ is approximately $N(\boldsymbol{\theta}, \boldsymbol{V})$, where $\boldsymbol{V} = n^{-1}\boldsymbol{W}$. Therefore, L_A in (3.30) is still a suitable statistic for testing H_0 against H_1, although it may not be the likelihood ratio statistic. Further, for large samples, its null distribution is approximately $\bar{\chi}^2$. In particular, the results in Theorem 3.7.1 and Corollary 3.7.2 are applicable as large sample approximations.

Another large sample application of the foregoing may take the following form. Let $\boldsymbol{\theta}$ denote a parameter associated with a general statistical model, and let $\boldsymbol{\theta}_0$ denote its true value. For large samples, we typically have $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx N(\mathbf{0}, \boldsymbol{V})$. Assume that $\hat{\boldsymbol{V}}$ is a consistent estimator of $\boldsymbol{V}$. Let

$$\hat{T} = \min_{\boldsymbol{\theta}\in\mathcal{M}} n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\hat{\boldsymbol{V}}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \min_{\boldsymbol{\theta}\in\mathcal{C}} n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\hat{\boldsymbol{V}}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

If the $\hat{\boldsymbol{V}}$ in the foregoing definition of $\hat{T}$ is replaced by its probability limit, $\boldsymbol{V}$, then the resulting change on $\hat{T}$ is of order $o_p(1)$. Let $\boldsymbol{Z}_n = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $\boldsymbol{Z} \sim N(\mathbf{0}, \boldsymbol{V})$. Suppose that $\boldsymbol{\theta}_0 \in \mathcal{M}$ and $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z}$. Then, we have that

$$\begin{aligned}\hat{T} &= \min_{\boldsymbol{\theta}\in\mathcal{M}}(\boldsymbol{Z}_n - \boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}_n - \boldsymbol{\theta}) - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\boldsymbol{Z}_n - \boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}_n - \boldsymbol{\theta}) + o_p(1) \\ &\xrightarrow{d} \min_{\boldsymbol{\theta}\in\mathcal{M}}(\boldsymbol{Z} - \boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z}_n - \boldsymbol{\theta}) - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\boldsymbol{Z} - \boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Z} - \boldsymbol{\theta}). \end{aligned} \tag{3.33}$$

The distribution of this limit is given by (3.31); if the constraints in H_1 are linear and independent then the limiting distribution is given by Corollary 3.7.2. A difficulty in applying this general result will arise if the weights, $w_i(p, \boldsymbol{V}, \mathcal{M}^\perp \cap \mathcal{C})$ in (3.31) depend on the unknown parameters defining $\boldsymbol{V}$. Methods of dealing with such nuisance parameters will be discussed in more detail in the next chapter. For now, we note that if $\boldsymbol{V}$ depends on unknown nuisance parameters then a standard procedure in statistical inference is to define the p-value as the supremum of the tail probability over the nuisance parameter:

$$p\text{-value} = \sup_{\boldsymbol{V}} \text{pr}(\hat{T} \geqslant t_{obs} \mid H_0)$$

where t_{obs} is the observed value of $\hat{T}$; this is discussed at the beginning of the next section. Now, if adopt the large sample arguments leading to (3.33), we have

$$p\text{-value} \simeq \sup_{\boldsymbol{V}} \lim \text{pr}(\hat{T} \geqslant t_{obs} \mid H_0) = \sup_{\boldsymbol{V}} \text{pr}\{\bar{\chi}^2(\boldsymbol{V}, \mathcal{C} \cap \mathcal{M}^\perp) \geqslant t_{obs}\}.$$

Example 3.7.1 *Effect of in-breeding on growth (Kudo (1963))*

In a large scale survey of child health carried out in Hiroshima and Nagasaki (see Neel and Schull (1954)), approximately 10000 children born of related as well as unrelated parents were studied. One facet of the study involved obtaining four anthropometric measurements as indices of growth and development. These were weight (y_1 gm), height (y_2 mm), head girth (y_3 mm), and chest girth (y_4 mm). The multivariate linear regression model, $E(\boldsymbol{Y}) = \boldsymbol{\mu} + t\boldsymbol{\alpha} - c\boldsymbol{\beta}$ was estimated where $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^T$, t is the age in months at examination, c is the in-breeding

coefficient (see Neel and Schull (1954), p71) of the individual, and $\{\mu, \alpha, \beta\}$ are unknown parameters. This analysis does not allow for difference between the two cities, Hiroshima and Nagasaki; preliminary analysis failed to detect any significant difference between the two cities. For now, we assume that there is no difference between the two cities; the method of analysis given below would not be affected even if there is a difference. The null and alternative hypotheses of interest may be stated as

$$H_0: \text{ In-breeding has no effect on growth,}$$
$$\textit{and} \quad H_1: \text{ In-breeding has a negative effect on growth.}$$

respectively. The foregoing statement of the testing problem is broad, and it needs to be formulated in terms of some parameters so that it can be tested statistically. We formulate the problem as

$$H_0: (\beta_1, \beta_2, \beta_3, \beta_4) = (0,0,0,0) \text{ and } H_1: (\beta_1, \beta_2, \beta_3, \beta_4) \geqslant (0,0,0,0). \quad (3.34)$$

The null hypothesis will be rejected if there is evidence that at least one β_i is positive. The least squares estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (2.710, 0.775.0.068, 0.329)^T$; the upper triangle of an estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is [1.199, 0.2164, 0.0839, 0.1841 | 0.0850, 0.0179, 0.0338 | 0.0232, 0.0137 | 0.0532], where "|" is used to separate different rows. The estimated covariance matrix is based on 10483 degrees of freedom. Now, since $\hat{\boldsymbol{\beta}}$ is approximately $N(\boldsymbol{\beta}, \boldsymbol{V})$, we wish to apply Corollary 3.7.2 with $\boldsymbol{X}$ replaced by $\hat{\boldsymbol{\beta}}$. Note that we have only an estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$. However, because the estimate is based on very large number of observations (more than 10000 degrees of freedom), it is reasonable to apply the results of the earlier sections with the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ being treated as the exact covariance matrix of $\hat{\boldsymbol{\beta}}$; essentially we are applying (3.33). Since each component of $\hat{\boldsymbol{\beta}}$ is positive, the estimate of $\boldsymbol{\beta}$ based on the single observation of $\hat{\boldsymbol{\beta}}$ subject to $\boldsymbol{\beta} \geqslant \mathbf{0}$ is also $\hat{\boldsymbol{\beta}}$. The sample value of the test statistic is

$$L_A = \|\hat{\boldsymbol{\beta}}\|_{\boldsymbol{V}}^2 - \min\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) : \boldsymbol{\beta} \geqslant \mathbf{0}\} = \|\hat{\boldsymbol{\beta}}\|_{\boldsymbol{V}}^2 = 9.3088.$$

Therefore, the p-value, $\text{pr}(L_A \geqslant 9.3088 \mid \boldsymbol{\beta} = \mathbf{0})$, is equal to

$$w_0(4, \boldsymbol{V})\text{pr}(\chi_0^2 \geqslant 9.3088) + \ldots + w_4(4, \boldsymbol{V})\text{pr}(\chi_4^2 \geqslant 9.3088).$$

Now, we compute the weights $w_0(4, \boldsymbol{V}), \ldots, w_4(4, \boldsymbol{V})$ by simulation (see page 79). The computed values are $w_0 = 0.0062$, $w_1 = 0.0682$, $w_2 = 0.2700$, $w_3 = 0.4355$, and $w_4 = 0.2201$; a FORTRAN computer program took a few seconds to compute these weights by simulation using 50000 samples. Now,

$$\begin{aligned} p\text{-value} &= 0.0682 \times \text{pr}(\chi_1^2 \geqslant 9.3088) + 0.27 \times \text{pr}(\chi_2^2 \geqslant 9.3088) \\ &\quad + 0.4355 \times \text{pr}(\chi_3^2 \geqslant 9.3088) + 0.2201 \times \text{pr}(\chi_4^2 \geqslant 9.3088) = 0.026. \end{aligned}$$

Therefore, it appears that there is sufficient evidence to accept that in-breeding has a negative effect on growth. Let us note that the p-value could have been computed using Simulation 1 without computing the weights first.

Another approach to formulating and testing the foregoing hypothesis is to carry out the multiple hypotheses tests of $\beta_1 = 0$ vs $\beta_1 > 0$, $\beta_2 = 0$ vs $\beta_2 > 0$, $\beta_3 = 0$ vs $\beta_3 > 0$, and $\beta_4 = 0$ vs $\beta_4 > 0$. There are advantages and disadvantages of this approach. The main advantage is that the multiple tests attempt to identify which of $\beta_1, \ldots, \beta_4$ are positive, and hence which aspects of growth are affected and which are not by in-breeding. However, a disadvantage of the multiple testing formulation is that they require an adjustment to the individual levels of significance, such as Bonferroni-type adjustment, which is known to result in low power for the combined tests. ■

Example 3.7.2 *Meta analysis: Effect of sodium on blood pressure (Follmann (1996b)).*

A large number of studies have been carried out to evaluate the effect of sodium reduction in reducing blood pressure. Let $\boldsymbol{X}^T = (X_1, X_2)$ = (systolic blood pressure, diastolic blood pressure). The extent of evidence in favor of the positive effect of sodium reduction in reducing the blood pressure vary considerably across various studies. What is desired is an overall probability statement that summarizes the strength of evidence in favor of the hypothesis that sodium reduction reduces blood pressure. To this end, Cutler et al. (1991) collected summary information available in a large number of published studies. It is worth pointing out that meta analysis requires considerable care in selecting the sample and ensuring that the observations are independent and that they are measuring the same parameter; we shall not consider such issues here. The relevant data are given in Table 1.13 (Cutler et al. (1991)).

Let us consider the data for the normotensive subjects. There are $n = 7$ observations. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ denote the mean reduction in blood pressure. Let us formulate the null and alternative hypotheses as $H_0: \boldsymbol{\theta} = \mathbf{0}$ and $H_1: \boldsymbol{\theta} \geqslant \mathbf{0}$, respectively. This is a Type A testing problem with two inequality constraints in H_1. Because the sample size in each of the seven studies with normotensive subjects is reasonably large, let us assume that the estimated standard error of $\hat{\theta}_{ij}$ is its true standard deviation. From ith published study we obtain an estimate $\hat{\boldsymbol{\theta}}_i$ and an estimated standard error for each component of $\hat{\boldsymbol{\theta}}_i$. Because these standard errors are not the same across different studies, we apply weighted least squares to estimate $\boldsymbol{\theta}$; each $\hat{\theta}_{ij}$ is weighted by the inverse of its variance. This leads to the following estimates:

$$\hat{\boldsymbol{\theta}} = (0.17, 1.04)^T; \hat{\boldsymbol{V}} = (0.1524, 0.0974 \mid 0.0974, 0.1510).$$

Now, the projection of $\hat{\boldsymbol{\theta}}$, with respect to $\|.\|_{\hat{\boldsymbol{V}}}$, onto H_0 is $\mathbf{0}$ and that onto the positive orthant is $\hat{\boldsymbol{\theta}}$ itself because both components of $\hat{\boldsymbol{\theta}}$ are positive. Direct substitution provides $\hat{\boldsymbol{\theta}}^T \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{\theta}} = 9.97$. Let ρ denote the correlation coefficient corresponding to $\boldsymbol{V}$; thus it is the correlation coefficient between systolic and diastolic blood pressures. If ρ were known, a large sample p-value would be

$$[0.5\text{pr}\{\chi_1^2 \geqslant 9.97\} + (2\pi)^{-1} \cos^{-1}(\rho)\text{pr}\{\chi_2^2 \geqslant 9.97\}].$$

Since ρ is unknown,

$$p\text{-value} \simeq \sup_{\rho}[0.5\text{pr}\{\chi_1^2 \geqslant 9.97\} + (2\pi)^{-1} \cos^{-1}(\rho)\text{pr}\{\chi_2^2 \geqslant 9.97\}].$$

Let us make the reasonable assumption that the reductions in systolic and diastolic blood pressures are nonnegatively correlated; therefore, $\cos^{-1}\rho \leq \pi/2$. Hence

$$\sup_{\rho>0}[(1/2)\mathrm{pr}\{\chi_1^2 \geqslant 9.97\} + \{(2\pi)^{-1}\cos^{-1}\rho\}\mathrm{pr}\{\chi_1^2 \geqslant 9.97\}] \leq 0.001.$$

Therefore, a large sample estimate of the p-value does not exceed 0.001; there is sufficient evidence to reject H_0.

The results presented so far for testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$ say that there is sufficient statistical evidence to reject $\boldsymbol{\theta} = \mathbf{0}$ but this alone does not mean that there is sufficient statistical evidence to accept $\boldsymbol{\theta} \gtrsim \mathbf{0}$. Suppose that we are prepared to assume *a priori* that a reduction in sodium intake cannot increase mean systolic or diastolic blood pressure, which translates to $\boldsymbol{\theta} \geqslant \mathbf{0}$. Now, since we have already noted that there is sufficient statistical evidence to reject $\boldsymbol{\theta} = \mathbf{0}$ we can conclude that there is sufficient statistical evidence to accept $\boldsymbol{\theta} \gtrsim \mathbf{0}$.

Now, let us suppose that we do not wish to assume *a priori* that $\boldsymbol{\theta} \geqslant \mathbf{0}$ but would like to establish that $\boldsymbol{\theta} \gtrsim \mathbf{0}$. This type of problems will be studied in a later chapter under the topic *sign testing*; see also *combination therapy*. It will be seen that we can formulate the inference problem as test of

$$H_0 : \theta_1 \leq 0 \text{ or } \theta_2 \leq 0 \text{ vs } H_1 : \theta_1 > 0 \text{ and } \theta_2 > 0.$$

Once the problem is formulated in this form, the so-called *min test* can be applied. In this formulation if the p-value turns out to be small then we can conclude that there is sufficient statistical evidence not only to reject $H_0 : \theta_1 \leq 0$ or $\theta_2 \leq 0$ but also to accept $H_1 : \theta_1 > 0$ and $\theta_2 > 0$. ■

3.8 LRT FOR TYPE B PROBLEMS: V IS KNOWN

Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, V)$, where V is a given positive definite matrix. We consider the Type B problem of testing

$$H_1 : \boldsymbol{\theta} \in C \text{ against } H_2 : \boldsymbol{\theta} \notin C. \tag{3.35}$$

In contrast to Type A problems, a main feature of Type B problems is that the null hypothesis involves inequalities. In this case, some important issues arise. To illustrate some of these, we first consider a simple bivariate case.

Let $\boldsymbol{X} = (X_1, X_2)^T$ be a bivariate random variable with density function $f(\boldsymbol{x}-\boldsymbol{\theta})$ for some f and $\boldsymbol{\theta}$, and let the null and alternative hypotheses be

$$H_1 : (\theta_1, \theta_2) \geqslant (0,0) \quad \text{and} \quad H_2 : (\theta_1, \theta_2) \not\geqslant (0,0),$$

respectively. This is a special case of (3.35) with $C = \{\boldsymbol{\theta} : \boldsymbol{\theta} \geqslant \mathbf{0}\}$. Let

$$L_B = \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|^2 : \boldsymbol{\theta} \geqslant \mathbf{0}\} - \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|^2 : \boldsymbol{\theta} \in \mathbb{R}^2\}, \tag{3.36}$$

where $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x}$. Let $\tilde{\boldsymbol{\theta}}$ be the projection of $\boldsymbol{X}$ onto the positive orthant. Then

$$L_B = \|\boldsymbol{X} - \tilde{\boldsymbol{\theta}}\|^2.$$

If $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$ then $\tilde{\boldsymbol{\theta}}$ is the *mle* of $\boldsymbol{\theta}$ and L_B is the *LRT*; however, for now we shall not assume that $\boldsymbol{X}$ is $N(\boldsymbol{\theta}, \boldsymbol{I})$. With ℓ_B denoting the sample value of L_B, one is tempted to define the p-value as $\mathrm{pr}(L_B \geqslant \ell_B \mid H_0)$. A difficulty that arises with this is that this probability depends on the particular null value, $\boldsymbol{\theta}$, which may be anywhere in the null parameter space $\{\boldsymbol{\theta} : \boldsymbol{\theta} \geqslant \mathbf{0}\}$. Thus, $\mathrm{pr}_{\boldsymbol{\theta}}(L_B \geqslant \ell_B \mid \boldsymbol{\theta} \in H_0)$ is not just a fixed number on the null parameter space, but a function of $\boldsymbol{\theta}$, and hence does not define a p-value. In this case, a reasonable approach to overcome the difficulty appears to be not to reject null hypothesis if there is at least one value in the null parameter space with which the data are consistent; or equivalently reject H_0 if the data are inconsistent with every value of $\boldsymbol{\theta}$ in the null parameter space. In fact, the usual procedure is to define the p-value as $\sup_{\boldsymbol{\theta}\in H_0}\mathrm{pr}_{\boldsymbol{\theta}}(L_B \geqslant \ell_B)$. At first, it may seem to be a formidable task to evaluate this supremum. Fortunately, it is reached at $\boldsymbol{\theta} = \mathbf{0}$, and therefore

$$\text{p-value} = \sup_{\boldsymbol{\theta}\in H_0}\mathrm{pr}_{\boldsymbol{\theta}}(L_B \geqslant \ell_B) = \mathrm{pr}_{\mathbf{0}}(L_B \geqslant \ell_B). \tag{3.37}$$

The proof of this is easy, and is given below in Theorem 3.8.1.

It is implicit in the statement (3.37) that the strength of evidence against H_0 and in favor of H_1 based on $L_B = \ell_B$ depends of the assumed true value of $\boldsymbol{\theta}$ in the null parameter space and that it is *least* when $\boldsymbol{\theta} = \mathbf{0}$. For this reason, $\boldsymbol{\theta} = \mathbf{0}$ is called the *least favorable null value* of L_B (also, the *least favorable null configuration of* L_B), and the distribution of L_B at this value is called the *least favorable null distribution* of L_B. The foregoing result holds in higher dimensions as well for Type B testing problems even if the distribution of $\boldsymbol{X}$ is not normal; this is stated below.

Theorem 3.8.1 *Assume that the distribution of $\boldsymbol{X} - \boldsymbol{\theta}$ does not depend on any unknown parameters. Let the null and alternative hypotheses be $H_1 : \boldsymbol{\theta} \in C$ and $H_2 : \boldsymbol{\theta} \notin C$, respectively, where C is a closed convex cone. Let V be a given positive definite matrix, and let $L_B = \min\{(\boldsymbol{X} - \boldsymbol{a})^T V^{-1}(\boldsymbol{X} - \boldsymbol{a}) : \boldsymbol{a} \in C\}$. Then, $\mathrm{pr}_{\boldsymbol{\theta}}(L_B \geqslant c \mid \boldsymbol{\theta} \in C) \leq \mathrm{pr}_{\mathbf{0}}(L_B \geqslant c)$, and therefore the least favorable null value of L_B is $\mathbf{0}$ and*

$$p\text{-value} = \sup_{\boldsymbol{\theta}\in C}\mathrm{pr}_{\boldsymbol{\theta}}(L_B \geqslant \ell_B) = \mathrm{pr}_{\mathbf{0}}(L_B \geqslant \ell_B) \tag{3.38}$$

where ℓ_B is the sample value of L_B. If C contains a linear space $\mathcal{M}$ then every point in $\mathcal{M}$ is also a least favorable null value for L_B.

Proof : The proof follows a set containment argument. The following proof is illustrated in Fig. 3.6 for the special case, $p = 2$, $V = \boldsymbol{I}$ and $C = \mathbb{R}^{+2}$. Suppose that $\boldsymbol{\theta} \in C$ and let $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{\theta}$. Then, for $c > 0$,

$$\begin{aligned}\mathrm{pr}_{\boldsymbol{\theta}}(L_B \geqslant c) &= \mathrm{pr}_{\boldsymbol{\theta}}[\|\boldsymbol{X} - C\|_V^2 \geqslant c] \\ &= \mathrm{pr}[\|(\boldsymbol{Z} + \boldsymbol{\theta}) - C\|_V^2 \geqslant c] = \mathrm{pr}[\|\boldsymbol{Z} - (C - \boldsymbol{\theta})\|_V^2 \geqslant c] \\ &\leq \mathrm{pr}[\|\boldsymbol{Z} - C\|_V^2 \geqslant c];\end{aligned}$$

the last step follows because $C \subset C - \boldsymbol{\theta}$ and hence the distance from $\boldsymbol{Z}$ to C is at least as large as that to $C - \boldsymbol{\theta}$. Therefore, we have that $\mathrm{pr}_{\boldsymbol{\theta}}(L_B \geqslant c|\boldsymbol{\theta} \in C) \leq \mathrm{pr}_{\mathbf{0}}(L_B \geqslant c)$. If $\boldsymbol{\theta} \in \mathcal{M}$ then $C = C - \boldsymbol{\theta}$ and hence the last part of the theorem follows. ■.

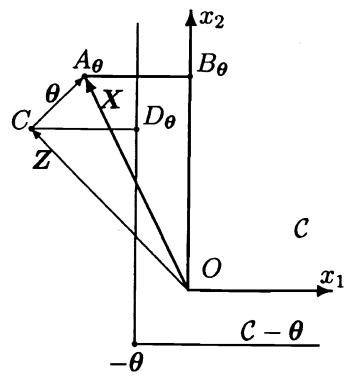

Fig. 3.11 Least favorable null value for $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$ vs $H_2 : \boldsymbol{\theta} \not\geqslant \mathbf{0}$ is the origin

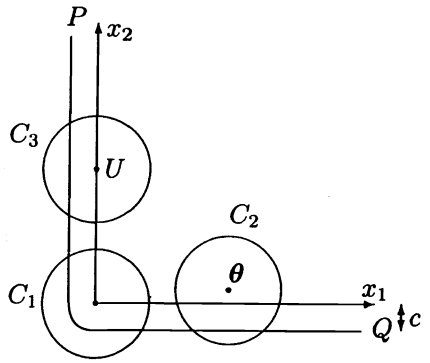

Fig. 3.12 Least favorable null value for $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$ vs $H_2 : \boldsymbol{\theta} \not\geqslant \mathbf{0}$ is the origin

Least favorable null value when $p = 2$

The foregoing proof is illustrated in Fig. 3.11 for the special case $p = 2, \boldsymbol{V} = \boldsymbol{I}$ and $\mathcal{C} = \mathbb{R}^{+2}$. Let $\boldsymbol{Z}$ be a random variable and let it be represented by OC. Let us think of $\boldsymbol{X}$ as the sum of the deterministic part $\boldsymbol{\theta}$ in $\mathbb{R}^{+2}$ and the random part $\boldsymbol{Z}$ that is unrelated to $\boldsymbol{\theta}$. Thus, the position of the point C is not related to $\boldsymbol{\theta}$. The vector $\boldsymbol{X}$ is represented by $OA_{\boldsymbol{\theta}}$ where the position of $A_{\boldsymbol{\theta}}$ depends on $\boldsymbol{\theta}$. Let $B_{\boldsymbol{\theta}}$ and $D_{\boldsymbol{\theta}}$ be the points in $\mathbb{R}^{+2}$ and $\mathbb{R}^{+2} - \boldsymbol{\theta}$ that are closest to $A_{\boldsymbol{\theta}}$ and C, respectively. Then $L_B = \|A_{\boldsymbol{\theta}}B_{\boldsymbol{\theta}}\|^2 = \|CD_{\boldsymbol{\theta}}\|^2$; because the position of C is unrelated $\boldsymbol{\theta}$, it follows that $\|CD_{\boldsymbol{\theta}}\|^2$ is a maximum when $\boldsymbol{\theta} = \mathbf{0}$. Therefore, $\boldsymbol{\theta} = \mathbf{0}$ is the least favorable null value.

It would be instructive to provide a different geometric interpretation of the foregoing proof. To this end, let us rewrite the proof of the previous theorem slightly differently. Let c and d be positive numbers. Suppose that $\boldsymbol{\theta} \in \mathcal{C}$ and let $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{\theta}$. Now, the critical region is of the form $\mathcal{A} = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x} - \mathcal{C}\| \geqslant c\}$. Again, $\mathcal{C} \subset \mathcal{C} - \boldsymbol{\theta}$ for any $\boldsymbol{\theta} \in \mathcal{C}$, and we have that

$$\text{pr}_{\boldsymbol{\theta}}(\|\boldsymbol{X} - \mathcal{C}\|_{\boldsymbol{V}} \geqslant c \text{ and } \|\boldsymbol{Z}\|_{\boldsymbol{V}} \leqslant d) = \text{pr}(\|\boldsymbol{Z} - (\mathcal{C} - \boldsymbol{\theta})\|_{\boldsymbol{V}} \geqslant c \text{ and } \|\boldsymbol{Z}\|_{\boldsymbol{V}} \leqslant d)$$

$$\leqslant \text{pr}(\|\boldsymbol{Z} - \mathcal{C}\|_{\boldsymbol{V}} \geqslant c \text{ and } \|\boldsymbol{Z}\|_{\boldsymbol{V}} \leqslant d) = \text{pr}_{\mathbf{0}}(\|\boldsymbol{X} - \mathcal{C}\|_{\boldsymbol{V}} \geqslant c \text{ and } \|\boldsymbol{Z}\|_{\boldsymbol{V}} \leqslant d).$$

Now, the proof follows by taking the limit $d \to \infty$.

This provides a different geometric interpretation of the main idea that underlies Theorem 3.8.1. Let us illustrate this using the same simple bivariate example. Since the test statistic is $L_B = \|\boldsymbol{X} - \mathbb{R}^{+2}\|^2$, the critical region is of the form $\mathcal{A} = \{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x} - \mathbb{R}^{+2}\| \geqslant c\}$ for some $c > 0$; this is the region to the left/bottom of the curve PQ in Fig. 3.12. Let C_1 be a circle of radius d with center at the origin. Let $\boldsymbol{\theta}$ be the point in $\mathcal{C}$ as shown in Figure 3.12. Let the circle C_2 be obtained by shifting the center of C_1 to $\boldsymbol{\theta}$. Note that the part of the interior of C_1 in the critical region is larger than that corresponding to C_2, for any $\boldsymbol{\theta}$ in $\mathcal{C}$. Now, it is easily seen

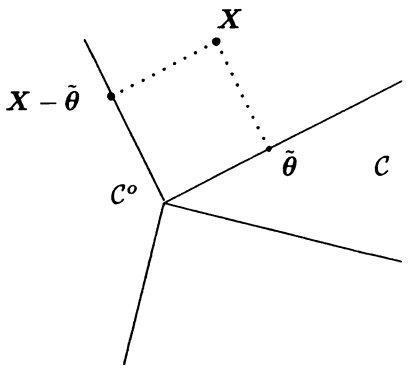

Fig. 3.13 Likelihood ratio test of $\boldsymbol{\theta} \in \mathcal{C}$ against $\boldsymbol{\theta} \notin \mathcal{C}$.

that the probability of $\boldsymbol{X}$ falling in the critical region and $\|\boldsymbol{Z}\| \leq d$ decreases as $\boldsymbol{\theta}$ moves along any straight line from the origin into the nonnegative orthant. This is true for any $d > 0$ and therefore the probability of $\boldsymbol{X}$ falling in the critical region is a maximum when $\boldsymbol{\theta} = \mathbf{0}$, and hence the least favorable null value of L_B is $\mathbf{0}$.

Least favorable null distribution of LRT when $p = 2$

Now, let us derive the least favorable null distribution of the LRT in two dimensions. Some important results do not become clear when $\mathcal{C}$ is the nonnegative orthant, $\mathbb{R}^{+2}$. Therefore, we consider an arbitrary $\mathcal{C}$. Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$, $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ and $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$; this set-up is the same as that in Example 3.3.3 in Section 3.3 and the details therein would be helpful here as well. For testing H_1 vs H_2, we have that

$$LRT = \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|^2 : \boldsymbol{\theta} \in \mathcal{C}\} - \min\{\|\boldsymbol{X} - \boldsymbol{\theta}\|^2 : \boldsymbol{\theta} \in \mathbb{R}^2\} = \|\boldsymbol{X} - \tilde{\boldsymbol{\theta}}\|^2,$$

where $\tilde{\boldsymbol{\theta}}$ is the *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \in \mathcal{C}$. Note that $(\boldsymbol{X} - \tilde{\boldsymbol{\theta}})$ is also the point in $\mathcal{C}^o$ that is closest to $\boldsymbol{X}$. Fig. 3.13 shows this for the present two dimensional case; this also holds in $\mathbb{R}^p$ (see Proposition 3.12.4). Since $\mathcal{C}^o$ is a closed convex cone in $\mathbb{R}^2$, $\|\boldsymbol{X} - \tilde{\boldsymbol{\theta}}\|^2$ is the LRT for testing

$$H_0^* : \boldsymbol{\theta} = \mathbf{0} \text{ against } H_1^* : \boldsymbol{\theta} \in \mathcal{C}^o.$$

We just noted that $\|\boldsymbol{X} - \tilde{\boldsymbol{\theta}}\|^2$ is also the LRT for testing $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$ based on a single observation of $\boldsymbol{X}$. Therefore, the least favorable null distribution of the LRT for testing $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$, and the null distribution of the LRT for testing $H_0^* : \boldsymbol{\theta} = \mathbf{0}$ against $H_1^* : \boldsymbol{\theta} \in \mathcal{C}^o$ are the same.

Since the angles (in radians) of $\mathcal{C}$ and $\mathcal{C}^o$ at their vertices sum to π, we conclude from Example 3.3.3 in Section 3.3 that

$$\begin{aligned} \text{pr}(LRT \geqslant c \,|\boldsymbol{\theta} \in \mathcal{C}) &\leq \text{pr}(LRT \geqslant c \mid \boldsymbol{\theta} = \mathbf{0}) \\ &= (0.5 - q)\text{pr}(\chi_0^2 \geqslant c) + 0.5\text{pr}(\chi_1^2 \geqslant c) + q\text{pr}(\chi_2^2 \geqslant c) \end{aligned} \tag{3.39}$$

where $q = \gamma/(2\pi)$, and γ is the angle (in radians) of the cone $\mathcal{C}^o$ at the vertex. Note that for this Type B testing problem, the weights in the relevant chi-bar-square

distribution appear in the reverse order to that of the corresponding Type A problem of testing $\theta = 0$ against $\theta \in C$ (see Example 3.3.3).

In the foregoing example, if we were to assume that the covariance matrix of X is V rather than I, then the details of Example 3.3.4 are relevant and (3.39) would still hold with q as in (3.15).

This example in two-dimensions illustrates some important results that hold in higher dimensions as well; this is stated in the next theorem.

Theorem 3.8.2 *Let $X \sim N(\theta, V)$ where V is a $p \times p$ positive definite matrix, and let the null and alternative hypotheses be*

$$H_1 : \theta \in C \quad \text{and} \quad H_2 : \theta \notin C,$$

respectively. Then the least favorable null value of LRT is $\theta = 0$. The least favorable null distribution of LRT is $\bar{\chi}^2(V, C^o)$ and

$$pr(LRT \leq c \mid \theta = 0) = \sum_{i=0}^{p} w_{p-i}(p, V, C) pr(\chi_i^2 \leq c).$$

Proof: It follows from Propositions 3.4.1 (page 77) and 3.12.4 (page 116) that

$$\begin{aligned} LRT &= \min\{(X-a)^T V^{-1}(X-a) : a \in C\} = \|\Pi_V(X \mid C^o)\|_V^2 \\ &= X^T V^{-1} X - \min\{(X-\theta)^T V^{-1}(X-\theta) : \theta \in C^o\} \\ &= \bar{\chi}^2(V, C^o). \end{aligned}$$

Now, the proof follows from Corollary 3.6.2 on page 83. ∎

As in Type A testing problems, if the inequalities defining C are all linear and independent, then the weights of the $\bar{\chi}^2$ distribution in the previous theorem take the simpler forms corresponding to the nonnegative orthant. The following corollary can be deduced from Proposition 3.6.1 (page 82).

Corollary 3.8.3 *Let $X \sim N(\theta, V)$. Let the null and alternative hypotheses be*

$$H_1 : R_1\theta \geqslant 0, R_2\theta = 0 \quad \text{and} \quad H_2 : \theta \text{ is not restricted},$$

where R_1 is $s \times m$, R_2 is $t \times m$, and the rank of $[R_1^T, R_2^T]$ is $(s+t)$. Then

$$LRT = \min\{(X-a)^T V^{-1}(X-a) : R_1 a \geqslant 0, R_2 a = 0\},$$

and the least favorable null distribution of LRT is

$$pr(LRT \leq c \mid \theta = 0) = \sum_{0}^{s} w_{s-i}(s, A) pr(\chi_{t+i}^2 \leq c),$$

where $A = R_1 V R_1^T - (R_1 V R_2^T)(R_2 V R_2^T)^{-1}(R_2 V R_1^T)$. ∎

In some situations, it is possible that linear equality constraints may be present in both the null and the alternative hypotheses as well. For example, suppose that the testing problem is

$$H_1 : \theta \in C \quad \text{vs} \quad H_2 : \theta \in \mathcal{L}$$

where $\mathcal{L}$ is a linear subspace of $\mathbb{R}^p$ and $C \subset \mathcal{L}$. In this case, the testing problem can be reformulated in the form discussed in this section. To this end, let $k = dim(\mathcal{L})$ and A be a matrix such that $rank(A) = k (\leq p)$ and $\mathcal{L} = \{A\beta : \beta \in \mathbb{R}^k\}$. Then the hypotheses take the form $H_1 : \beta \in \mathcal{P}$ and $H_2 : \beta \notin \mathcal{P}$, where $\mathcal{P}$ is the closed convex cone, $\{\beta : A\beta \in C\}$. Now, with L denoting $-2\log(likelihood)$, we have the following for $\theta \in \mathcal{L}$:

$$\begin{aligned} L &= (X-\theta)^T V^{-1}(X-\theta) = (X - A\beta)^T V^{-1}(X - A\beta) \\ &= (X - A\hat{\beta})^T V^{-1}(X - A\hat{\beta}) + (\hat{\beta} - \beta)^T W^{-1}(\hat{\beta} - \beta) \end{aligned}$$

where $\hat{\beta} = (A^T V^{-1} A)^{-1} A^T V^{-1} X$ and $W = (A^T V^{-1} A)^{-1}$. Therefore, the problem is equivalent to testing $H_1 : \beta \in \mathcal{P}$ against $H_2 : \beta \notin \mathcal{P}$ based on a single observation of $\hat{\beta}$ where $\hat{\beta} \sim N(\beta, W)$. This is of the same form as that considered in the foregoing theorem and the corollary.

Finally, let L_A denote the LRT for testing $\theta = 0$ vs $\theta \in C$ and L_B denote the LRT for testing $\theta \in C$ vs $\theta \notin C$ based on a single observation of X from $N(\theta, V)$. Then, by using arguments similar to the proof of Lemma 3.13.6 (page 129), it can be deduced that

$$\text{pr}(L_A \in A \text{ and } L_B \in B \mid \theta = 0) = \sum_{i=0}^{p} w_i(p, V, C)\text{pr}(\chi_i^2 \in A)\text{pr}(\chi_{p-i}^2 \in B).$$

This was obtained by Raubertas et al. (1986). There may be some scenarios in which this may be useful. For example, if one wishes to reject $\theta = 0$ and accept $\theta \in C \setminus \{0\}$ if L_A is large and L_B is small, this result may have some use. However, a sketch of the critical region would show that it has an unappealing shape when $V \neq I$.

3.9 TESTS ON THE LINEAR REGRESSION PARAMETER

In this section it will be shown that most of the normal theory results in the previous sections extend to the normal theory linear model as well. It will be seen that most of the results concerning the LRT on the comparison of normal means studied in the previous chapter are special cases of the results presented in this section.

Let us consider the linear model,

$$Y = X\theta + E \tag{3.40}$$

where Y is an $N \times 1$ vector of observations, X is a known $N \times p$ matrix of rank p with $p < N$, θ is a $p \times 1$ vector of unknown parameters and E has mean 0 and

covariance matrix $\sigma^2 U$, U is known and σ is unknown. Assume that the distribution of E does not depend any unknown parameters. Let

$$\hat{\theta} = (X^T U^{-1} X)^{-1} X^T U^{-1} Y, \qquad W = (X^T U^{-1} X)^{-1},$$
$$Q(\theta) = (Y - X\theta)^T U^{-1} (Y - X\theta), \quad q(\theta) = (\hat{\theta} - \theta)^T W^{-1} (\hat{\theta} - \theta). \tag{3.41}$$

Using the normal equation, it is easily seen that

$$Q(\theta) = (Y - X\hat{\theta})^T U^{-1} (Y - X\hat{\theta}) + q(\theta). \tag{3.42}$$

Let the null and alternative hypotheses be

$$H_a : \theta \in C_a \text{ and } H_b : \theta \in C_b,$$

respectively, where $C_a \subset C_b$. Let θ^a and θ^b be the points at which $Q(\theta)$ is minimized over C_a and C_b, respectively. To define a test statistic based on generalized least squares without assuming normality, first note that $Q(\theta)$ is a generalized sum of squares. Therefore, $\{Q(\theta^a) - Q(\theta^b)\}$ is the reduction in this generalized sum of squares; it is a measure of the discrepancy between H_a and H_b. However, as it stands, this reduction cannot be used as a statistic for testing H_a against H_b because its null distribution depends on the unknown scale parameter σ^2; but, the null distribution of $\{Q(\theta^a) - Q(\theta^b)\}/\sigma^2$ does not depend on σ^2. Therefore, it provides a good starting point for constructing a test statistic. As in the classical least squares theory, the idea is to replace the σ^2 in the denominator of $\{Q(\theta^a) - Q(\theta^b)\}/\sigma^2$ by a suitable estimate of it so that the distribution of the resulting statistic is independent of σ^2. One such estimator turns out to be $Q(\theta^a)$, and we define the following statistic:

$$\bar{E}^2 = \{Q(\theta^a) - Q(\theta^b)\}/Q(\theta^a). \tag{3.43}$$

In the analysis of the classical normal theory linear models with no inequality constraints, tests on regression parameters are usually carried out using F-tests. A similar test statistic can also be developed when the hypotheses involve inequality constraints. Let

$$S^2 = \nu^{-1} (Y - X\hat{\theta})^{-1} U^{-1} (Y - X\hat{\theta})$$

with $\nu = (N - p)$, the usual error mean square. Motivated by $\{Q(\theta^a) - Q(\theta^b)\}/\sigma^2$ as a measure of discrepancy between H_a and H_b, we define

$$\bar{F} = \{Q(\theta^a) - Q(\theta^b)\}/S^2. \tag{3.44}$$

In the definition of the traditional unrestricted/two-sided F-statistic for testing equality constraints, the numerator of the F-statistic is $\{Q(\theta^a) - Q(\theta^b)\}/k$ where k is the number of equality constraints imposed by the null hypothesis. However, we did not introduce the divisor k in the definition of $\bar{F}$ because it does not really simplify anything.

The difference between the definitions of $\bar{E}^2$ and $\bar{F}$ is that $\bar{E}^2$ uses the dispersion of the *restricted* residuals, $Y - X\theta^a$, to estimate σ^2 while $\bar{F}$ uses the dispersion of the *unrestricted* residuals, $Y - X\hat{\theta}$. Because of the close similarity between these statistics, we would not expect the differences between them to be substantial, at least in large samples. Wright (1988) provided some calculations to compare the powers of $\bar{E}^2$ with $\bar{F}$ in one-way lay-out models.

The basic idea of the $\bar{F}$ statistic appears to be due to Kudo (1963) who suggested it when the observations are *iid* from $N(\theta, \sigma^2 U)$ and stated its null distribution. Wolak (1987a) applied it for inference with linearly independent constraints in the linear model, Wright (1988) considered it for ANOVA models, and Silvapulle (1996a) obtained results for the case of testing for or against inequality constraints in a more general context. Dufour (1989) obtained bounds when the parameter spaces may not be cones. Wolak (1987a) also provided a detailed discussion of $\bar{F}$ and how it relates to Kuhn-Tucker multipliers. Large sample results relating to $\bar{F}$ will be obtained in the next chapter.

So far, the discussions have been motivated by generalized least squares, without assuming that E has any known distributional form. Now, suppose that $E \sim N(0, \sigma^2 U)$. Then the loglikelihood $L(\theta, \sigma^2)$ takes the form

$$L(\theta, \sigma^2) = -(1/2)\sigma^{-2} Q(\theta) - (1/2) np \log(\sigma^2) + Const. \tag{3.45}$$

For any given θ, $L(\theta, \sigma^2)$ is maximized at $\sigma^2 = (np)^{-1} Q(\theta)$. Therefore, the profile (or concentrated) loglikelihood $\tilde{L}(\theta)$ is obtained by substituting $\sigma^2 = (np)^{-1} Q(\theta)$ in $L(\theta, \sigma^2)$. This leads to

$$\tilde{L}(\theta) = \max_{\sigma} L(\theta, \sigma^2) = -(1/2) np \log Q(\theta) + Const.$$

Therefore, θ^a and θ^b, which were introduced as generalized least squares estimates, are also the *mle*'s of θ under H_a and H_b, respectively. Now, the likelihood ratio statistic for testing H_a against H_b is given by

$$LRT = np \log\{Q(\theta^a)/Q(\theta^b)\}. \tag{3.46}$$

Since

$$\bar{E}^2 = [1 - \exp\{-LRT/(np)\}],$$

LRT is an increasing function of $\bar{E}^2$. Therefore, the $\bar{E}^2$-test, which rejects H_a for large values of $\bar{E}^2$, is equivalent to the likelihood ratio test when $E \sim N(0, \sigma^2 U)$.

3.9.1 Null Distributions

Many of the results that we established for $\bar{\chi}^2$ statistics also have corresponding results for $\bar{E}^2$ and $\bar{F}$. In what follows, the suffices A and B indicate that the test statistics correspond to Type A and Type B problems, respectively. The proof of the next result is given in the Appendix (page 130).

Theorem 3.9.1 *Let the linear model be as in (3.40), and assume that the distribution of E/σ does not depend on any unknown parameters. Then we have the following:*
(a) For the Type A testing problem,

$$H_0 : \theta \in \mathcal{M} \quad \textit{against} \quad H_1 : \theta \in C,$$

the null distributions of $\bar{E}^2$ and $\bar{F}$ do not depend on σ or on the value of $\boldsymbol{\theta}$ in the null parameter space; in other words, the tests are similar.
(b) For the Type B testing problem,

$$H_0 : \boldsymbol{\theta} \in \mathcal{C} \quad \textit{against} \quad H_1 : \boldsymbol{\theta} \notin \mathcal{C},$$

a least favorable null value of $\bar{E}^2$ and of $\bar{F}$ is $\boldsymbol{\theta} = \mathbf{0}$; if the null hypothesis also includes the linear equality constraint, $\boldsymbol{R\theta} = \mathbf{0}$, then any $\boldsymbol{\theta}$ satisfying $\boldsymbol{R\theta} = \mathbf{0}$ is also a least favorable null value. ■

Suppose that the testing problem is as in part (a) or part (b) of the foregoing Theorem. An important consequence of this result is that the critical values and p-values can be computed by simulation for a large class of error distributions that includes the normal distribution; the simulations can be carried out at any fixed value of $(\boldsymbol{\theta}, \sigma)$. The simulation steps for computing the p-values of $\bar{F}$ and $\bar{E}^2$ corresponding to any error distribution are given below.

Computation of the p-values for $\bar{F}$ and $\bar{E}^2$ by simulation
Suppose that the conditions of Theorem 3.9.1 are satisfied. Let $F(\epsilon/\sigma)$ denote the *cdf* of the error $\boldsymbol{E}$ where σ may be unknown, but F is assumed known; for example, the distribution of $\boldsymbol{E}$ may be $N(\mathbf{0}, \sigma^2\boldsymbol{U})$. Now, the following steps would compute the p-value for $\bar{F}$ (respectively, for $\bar{E}^2$).

1. Generate one observation of $\boldsymbol{Y}$ from $F(\epsilon)$; this is same as generating one observation $\boldsymbol{E}$ from $F(\epsilon)$ and then computing $\boldsymbol{Y}$ as in (3.40) with $\boldsymbol{\theta} = \mathbf{0}$.
2. Compute the $\bar{F}$ (respectively, $\bar{E}^2$) statistic .
3. Repeat the previous two steps N times (say N = 10000), and estimate the p-value by M/N where M is the number of times the $\bar{F}$ (respectively, $\bar{E}^2$) statistic in the second step exceeded its sample value. ■

Note that in the first step of the simulation, the observations may be generated from a distribution with any value for the common scale parameter because, in view of Theorem 3.9.1, the null distributions of $\bar{F}$ and $\bar{E}^2$ do not depend on σ. Thus, if $\boldsymbol{E} \sim N(\mathbf{0}, \sigma^2\boldsymbol{U})$ where σ is unknown, then it suffices to generate $\boldsymbol{Y}$ from $N(\mathbf{0}, \boldsymbol{U})$ in the first step of the simulation for computing the p-value. The level-α critical value can also be computed by ordering the N pseudo-random values of the statistic computed in step 2, and then finding upper level-α quantile of the empirical distribution of the N values.

The forgoing results would be particularly useful in model selection. For example, suppose that we wish to test H_0 vs H_1 where

$$H_0 : y = \theta_0 + \theta_1 x_1 + \ldots + \theta_q x_q + e,$$
$$\textit{and} \quad H_1 : y = \theta_0 + \theta_1 x_1 + \ldots + \theta_q x_q + \theta_{q+1} x_{q+1} + \ldots + \theta_p x_p + e.$$

Suppose also that some of $\{\theta_{q+1}, \ldots, \theta_p\}$ are known to be nonnegative. Then the task of comparing the smaller model (H_0) to the larger one (H_1) reduces to a testing problem of the type just discussed.

If we assume that the error distribution is normal, then we can derive the null distributions of $\bar{E}_A^2$ and $\bar{F}_A$. The main result is given in the next theorem, and the proof is given in the Appendix (see page 131). The derivations use the same basic approach and constructions as for deriving the $\bar{\chi}^2$ distribution.

Theorem 3.9.2 *Let the linear model be as in (3.40), and assume that $\boldsymbol{E} \sim N(\mathbf{0}, \sigma^2\boldsymbol{U})$.*
(a) Let the null and alternative hypotheses be

$$H_0 : \boldsymbol{\theta} \in \mathcal{M} \quad \textit{and} \quad H_1 : \boldsymbol{\theta} \in \mathcal{C}$$

respectively, where $dim(\mathcal{M}) = q$. Then the null distributions of $\bar{E}^2$ and $\bar{F}$ are given by

$$pr(\bar{E}_A^2 \le c \mid H_0) = \textstyle\sum_{i=0}^{p} w_i(p, \boldsymbol{W}, \mathcal{C} \cap \mathcal{M}^{\perp}) pr[\beta\{i/2, (N-q-i)/2\} \le c], \tag{3.47}$$

$$pr(\bar{F}_A \le c \mid H_0) = \textstyle\sum_{i=0}^{p} w_i(p, \boldsymbol{W}, \mathcal{C} \cap \mathcal{M}^{\perp}) pr(iF_{i,\nu} \le c) \tag{3.48}$$

where $\beta(a, b)$ is the beta distribution with parameters a and b.
(b) Let the null and alternative hypotheses be

$$H_0 : \boldsymbol{\theta} \in \mathcal{C} \quad \textit{and} \quad H_1 : \boldsymbol{\theta} \notin \mathcal{C}$$

respectively. Then $\boldsymbol{\theta} = \mathbf{0}$ is a least favorable null value for $\bar{E}_B^2$ and $\bar{F}_B$. Further, the least favorable null distributions of $\bar{E}_B^2$ and $\bar{F}_B$ are:

$$pr(\bar{E}_B^2 \le c \mid \boldsymbol{\theta} = \mathbf{0}) = \textstyle\sum_{i=0}^{p} w_{p-i}(p, \boldsymbol{W}, \mathcal{C}) pr[\beta\{i/2, (N-p)/2\} \le c],$$

$$pr(\bar{F}_B \le c \mid \boldsymbol{\theta} = \mathbf{0}) = \textstyle\sum_{i=0}^{p} w_{p-i}(p, \boldsymbol{W}, \mathcal{C}) pr(iF_{i,\nu} \le c).$$ ■

As in Sections 3.7 and 3.8 on results concerning $\bar{\chi}^2$ statistics, if the constraints are linear and independent then the foregoing results take simpler forms because the weights can be expressed in terms of those corresponding to the nonnegative orthant using the results in Proposition 3.6.1; this is stated in the next corollary.

Corollary 3.9.3 *Let the linear model be as in (3.40), and assume that $\boldsymbol{E} \sim N(\mathbf{0}, \sigma^2\boldsymbol{U})$.*
(a) Let $\boldsymbol{R}$ be a matrix of order $r \times p$, $r \le p$, $rank(\boldsymbol{R}) = r$, $\boldsymbol{R}_1$ be a submatrix of $\boldsymbol{R}$ of order $q \times p$. Let the null and alternative hypotheses be

$$H_0 : \boldsymbol{R\theta} = \mathbf{0} \textit{ and } H_1 : \boldsymbol{R}_1\boldsymbol{\theta} \geqslant \mathbf{0}.$$

Then we have the following for the null distributions of $\bar{E}_A^2$ and $\bar{F}_A$:

$$pr(\bar{E}_A^2 \le c | \boldsymbol{R\theta} = \mathbf{0}) = \textstyle\sum_{i=0}^{q} w_i(q, \boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_1^T) pr[\beta\{(r-q+i)/2, (N-p+q-i)/2\} \le c]$$

$$pr(\bar{F}_A \le c | \boldsymbol{R\theta} = \mathbf{0}) = \textstyle\sum_{i=0}^{q} w_i(q, \boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_1^T) pr[(r-q+i)F_{r-q+i,\nu} \le c].$$

(b) Now, let the null and alternative hypotheses be

$$H_1 : \boldsymbol{R}_1\boldsymbol{\theta} \geqslant \mathbf{0}, \boldsymbol{R}_2\boldsymbol{\theta} = \mathbf{0} \textit{ and } H_2 : \boldsymbol{\theta} \textit{ is not restricted},$$

where $\boldsymbol{R}_1$ is $s \times p$, $\boldsymbol{R}_2$ is $t \times p$, and $[\boldsymbol{R}_1^T, \boldsymbol{R}_2^T]$ has full rank. Then, any $\boldsymbol{\theta}$ satisfying $\boldsymbol{R}_1\boldsymbol{\theta} = \boldsymbol{0}$ and $\boldsymbol{R}_2\boldsymbol{\theta} = \boldsymbol{0}$ is a least favorable null value for $\bar{E}_B^2$ and $\bar{F}_B$. Further,

$$pr(\bar{E}_B^2 \leq c|\boldsymbol{\theta} = \boldsymbol{0}) = \sum_{i=0}^{s} w_{s-i}(s, \boldsymbol{A}) pr[\beta\{(t+i)/2, (N-k)/2\} \leq c],$$

$$pr(\bar{F}_B \leq c|\boldsymbol{\theta} = \boldsymbol{0}) = \sum_{i=0}^{s} w_{s-i}(s, \boldsymbol{A}) pr[(t+i)F_{t+i,\nu} \leq c]. \quad \blacksquare$$

The foregoing results for Type B testing problems are useful even if some equality constraints on the regression parameters are present under both the null and alternative hypotheses. For example, consider the Type B testing problem and suppose that some linear equality constraints are present in the null and alternative hypotheses. Then the hypotheses take the form $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ and $H_2 : \boldsymbol{\theta} \in \mathcal{L}$, where $\mathcal{L} = \{\boldsymbol{A}\beta : \beta \in \mathbb{R}^k\}$ for some matrix $\boldsymbol{A}$ of order $p \times k$ and $rank(\boldsymbol{A}) = k$. Let $\mathcal{P} = \{\beta : \boldsymbol{A}\beta \in \mathcal{C}\}$. Then the linear model takes the form $\boldsymbol{Y} = \boldsymbol{B}\beta + \boldsymbol{E}$, where $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{A}$, and the null and alternative hypotheses take the standard form, $H_0 : \beta \in \mathcal{P}$ and $H_1 : \beta \notin \mathcal{P}$. Now we can apply the foregoing results.

3.10 TESTS WHEN V IS UNKNOWN (PERLMAN'S TEST AND ALTERNATIVES)

In the foregoing sections we considered the cases when the covariance matrix $\boldsymbol{V}$ is either known or of the form $\sigma^2\boldsymbol{U}$ where $\boldsymbol{U}$ is known and σ is unknown. Now we consider the case when $\boldsymbol{V}$ is completely unknown.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n independently and identically distributed observations from $N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V}$ is unknown. Exact results for the likelihood ratio test were obtained by Perlman in his seminal paper, Perlman (1969). Corresponding results for a closely related test were obtained by Silvapulle (1995); conditional version of this test was introduced by Perlman and Wu (2002a). Other procedures include Tang (1994), Wang and McDermott (1998a, 1998b), Tang et al. (1989a), O'Brien (1984), and Follmann (1996b).

Let us consider the general problem of testing

$$H_a : \boldsymbol{\theta} \in \mathcal{C}_a \text{ against } H_b : \boldsymbol{\theta} \in \mathcal{C}_b \tag{3.49}$$

where $\mathcal{C}_a$ and $\mathcal{C}_b$ are *one-sided* closed cones; a set A is said to be *one-sided* if there exists $\boldsymbol{a}$ such that $A \subset \{\boldsymbol{x} : \boldsymbol{a}^T\boldsymbol{x} \geqslant 0\}$. In this section, the cones considered are all *one-sided*. Since $\boldsymbol{V}$ is unknown, the loglikelihood up to a constant is

$$\ell(\boldsymbol{\theta}, \boldsymbol{V}) = \sum_{i=1}^{n} (-1/2)\{\log|\boldsymbol{V}| + (\boldsymbol{X}_i - \boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{X}_i - \boldsymbol{\theta})\}. \tag{3.50}$$

Further, we have that

$$LRT = 2[\max\{\ell(\boldsymbol{\theta}, \boldsymbol{V}) : \boldsymbol{\theta} \in \mathcal{C}_b, \boldsymbol{V} > 0\} - \max\{\ell(\boldsymbol{\theta}, \boldsymbol{V}) : \boldsymbol{\theta} \in \mathcal{C}_a, \boldsymbol{V} > 0\}],$$

where $\boldsymbol{V} > 0$ means that $\boldsymbol{V}$ is positive definite. Note that the maximization needs to be carried out over $\boldsymbol{\theta}$ and over all the positive definite matrices, $\boldsymbol{V} > 0$, because now

$\boldsymbol{V}$ is an unknown parameter; therefore, this *LRT* is not the same as that for the case when $\boldsymbol{V}$ is known. Let

$$\boldsymbol{S} = n^{-1}\sum(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T$$

denote the sample covariance matrix, and let

$$\mathcal{U}(\mathcal{C}_a, \mathcal{C}_b) = \{\|\Pi_{\boldsymbol{S}}(\bar{\boldsymbol{X}}|\mathcal{C}_b)\|_{\boldsymbol{S}}^2 - \|\Pi_{\boldsymbol{S}}(\bar{\boldsymbol{X}}|\mathcal{C}_a)\|_{\boldsymbol{S}}^2\}\{1 + \|\bar{\boldsymbol{X}} - \|\Pi_{\boldsymbol{S}}(\bar{\boldsymbol{X}}|\mathcal{C}_b)\|_{\boldsymbol{S}}^2\}^{-1}. \tag{3.51}$$

Now, we have the following that simplifies the *LRT*.

Proposition 3.10.1 *(Perlman (1969), Theorem 5.1). $\mathcal{U}(\mathcal{C}_a, \mathcal{C}_b)$ is a strictly increasing function of LRT for testing $H_a : \boldsymbol{\theta} \in \mathcal{C}_a$ against $H_b : \boldsymbol{\theta} \in \mathcal{C}_b$.* $\blacksquare$

It follows that the likelihood ratio test is equivalent to rejecting H_a for large values of $\mathcal{U}(\mathcal{C}_a, \mathcal{C}_b)$. Therefore, to apply the *LRT*, we do not have to compute the maximum of the function, $\ell(\boldsymbol{\theta}, \boldsymbol{V})$, or any other function of $\boldsymbol{V}$ over the complicated parameter space, $\{\boldsymbol{V} > 0\}$. If the constraints in $\mathcal{C}_a$ and $\mathcal{C}_b$ are linear, then $\mathcal{U}(\mathcal{C}_a, \mathcal{C}_b)$ can be computed using only a quadratic program. The foregoing result does not require observations to be *iid* but it does require an observation of $\boldsymbol{Y}$ from $N(\sqrt{n}\boldsymbol{\theta}, \boldsymbol{V})$, and an observation $\boldsymbol{S}$ from the Wishart distribution $W(n-1, \boldsymbol{V})$ where $\boldsymbol{Y}$ and $\boldsymbol{S}$ are independent.

Now, let us introduce *Perlman's test* as follows:

Reject $H_0 : \boldsymbol{\theta} \in \mathcal{C}_a$ in favor of $H_1 : \boldsymbol{\theta} \in \mathcal{C}_b$ for large values of $\mathcal{U}(\mathcal{C}_a, \mathcal{C}_b)$.

Before we study properties of this test, let us introduce another class of tests. In classical multivariate analysis, tests of $\boldsymbol{\theta} = \boldsymbol{0}$ against $\boldsymbol{\theta} \neq \boldsymbol{0}$ are carried out using the well-known Hotelling's T^2 statistic, $\bar{\boldsymbol{X}}^T\boldsymbol{S}^{-1}\bar{\boldsymbol{X}}$. For this unrestricted testing problem, the LRT reduces to $\bar{\boldsymbol{X}}^T\boldsymbol{S}^{-1}\bar{\boldsymbol{X}}$, the Hotelling's T^2. To motivate another test, let us note that, since $\boldsymbol{S}$ is positive definite, $\|.\|_{\boldsymbol{S}}^2 = \bar{\boldsymbol{X}}^T\boldsymbol{S}^{-1}\bar{\boldsymbol{X}}$ is the squared distance between $\bar{\boldsymbol{X}}$ and the null parameter space. A similar statistic is available for the two-sample problem. Based on an idea of Shorack (1967), generalizations of the statistic $\bar{\boldsymbol{X}}^T\boldsymbol{S}^{-1}\bar{\boldsymbol{X}}$ have been proposed when there are inequalities in $\boldsymbol{\theta}$ (see Silvapulle (1995)). The basic form of the statistic is quite simple:

$$T^{*2} = \|\bar{\boldsymbol{X}} - \mathcal{C}_a\|_{\boldsymbol{S}}^2 - \|\bar{\boldsymbol{X}} - \mathcal{C}_b\|_{\boldsymbol{S}}^2. \tag{3.52}$$

In other words, T^{*2} is the difference between "the squared distance from $\bar{\boldsymbol{X}}$ to the null parameter space" and "the squared distance from $\bar{\boldsymbol{X}}$ to the alternative parameter space." Let us remark that the foregoing T^{*2} and the Hotelling T^2 have similarities with respect to their algebraic forms; but, Hotelling's T^2 is based on likelihood ratio and the T^{*2} in (3.52) is not.

For the special case of testing $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta} \geqslant \boldsymbol{0}$, Sen and Tsai (1999), obtained a union-intersection test. Their test statistic, which they denote by T_n^*, is the same as T^{*2}; hence the latter can also be motivated from the union-intersection principle.

Perlman (1969) obtained the null distribution of $\mathcal{U}(C_a, C_b)$ for some important testing problems. Several corresponding results for the T^{*2} were obtained by Silvapulle (1995). The important ones are discussed below. Let U denote either the $\mathcal{U}(C_a, C_b)$ in (3.51) or the T^{*2} in (3.52). Let u denote the sample value of U. In general, the null distribution of U depends on $(\boldsymbol{\theta}, \boldsymbol{V})$. Therefore, $\text{pr}(U \geqslant u \mid H_0, \boldsymbol{V})$ is not an operational p-value. A suitable test procedure is,

$$\text{reject } H_0 \text{ if } \sup\{\text{pr}(U \geqslant u \mid \boldsymbol{\theta}, \boldsymbol{V}) : \boldsymbol{\theta} \in H_0, \boldsymbol{V} > 0\} \text{ is small,} \tag{3.53}$$

where the supremum is taken over $\boldsymbol{\theta}$ in H_0 and over all positive definite matrices $\boldsymbol{V}$; for example, see Bickel and Doksum (1977) page 170. Similarly, if $\inf\{\text{pr}(U \geqslant u|\boldsymbol{\theta}, \boldsymbol{V}) : \boldsymbol{\theta} \in H_0, \boldsymbol{V} > 0\}$ is large then do not reject H_0. Otherwise, the test is not conclusive. In some cases, for example, in Type A testing problems, $\text{pr}(U \geqslant u|H_0, \boldsymbol{V})$ does not depend on the value of $\boldsymbol{\theta}$ in the null parameter space but it depends on $\boldsymbol{V}$. In such cases, we shall also consider the possibility of estimating $\text{pr}(U \geqslant u|H_0, \boldsymbol{V})$ by substituting $\hat{\boldsymbol{V}}$ for $\boldsymbol{V}$, but it does require caution (see Section 4.3.2).

The null distributions of *LRT* and T^{*2} for inequality constrained problems, involve the following two random variables:

$$G(i,j) = \chi_i^2/\chi_j^2 \quad \text{and} \quad H(k,r,n) = (\chi_k^2/\chi_{n-r}^2)(1 + \chi_{r-k}^2/\chi_{n-r+k}^2),$$

where the different χ^2 variates are independent. For any given c, the tail probabilities $\text{pr}\{G(i,j) \geqslant c\}$ and $\text{pr}\{H(k,r,n) \geqslant c\}$ can be computed easily by simulation because generation of random numbers from independent χ^2 variates is straight forward. Consequently, it will be seen that, in terms of computational demands, the difference between T^{*2} and *LRT* is small; the choice between the two would need to be based on their statistical properties.

It is also worth noting that $(j/i)G(i,j)$ is the usual F distribution with (i,j) degrees of freedom which we denote by $F_{i,j}$. Therefore,

$$\text{pr}\{G(i,j) \geqslant c\} = \text{pr}((i/j)F_{i,j} \geqslant c),$$

and hence the F-distribution can be used to compute the tail probabilities of $G(i,j)$.

3.10.1 Type A Testing Problem

To consider a special Type A problem, let us partition $\boldsymbol{\theta}$ as $(\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ where $\boldsymbol{\theta}_1$ is $q \times 1$ and $\boldsymbol{\theta}_2$ is $r \times 1$; let $\boldsymbol{V}, \boldsymbol{S}$ and $\boldsymbol{X}$ be partitioned conformably. Now, consider the test of

$$H_0 : \boldsymbol{\theta}_2 = \mathbf{0} \text{ against } H_1 : \boldsymbol{\theta}_2 \in \mathcal{P} \tag{3.54}$$

where $\mathcal{P}$ is a one-sided closed convex cone. Then, it is easily shown that the null distributions of *LRT* and T^{*2} do not depend on $\boldsymbol{\theta}$ in the null parameter space; the argument is virtually the same as for Proposition 3.7.3. A bounds test can be carried out using the bounds given in the next result.

Proposition 3.10.2 *(Perlman (1969) Theorem 6.3, and Silvapulle (1995) Theorem 3.) Let the testing problem be as in (3.54), and let $\mathcal{U}$ denote $\mathcal{U}(C_a, C_b)$ for this testing problem. Then under H_0,*

$$\begin{aligned} \inf_{\boldsymbol{\theta},\boldsymbol{V}} pr(\mathcal{U} \geqslant u \mid H_0, \boldsymbol{V}) &= (1/2)pr\{G(1, n-r) \geqslant u\}, \\ \inf_{\boldsymbol{\theta},\boldsymbol{V}} pr(T^{*2} \geqslant u \mid H_0, \boldsymbol{V}) &= (1/2)pr\{H(1,r,n) \geqslant u\}. \end{aligned} \tag{3.55}$$

If $\mathcal{P}$ contains an open set of dimension r then

$$\begin{aligned} &\sup_{\boldsymbol{\theta},\boldsymbol{V}} pr(\mathcal{U} \geqslant u \mid H_0, \boldsymbol{V}) = \\ &\quad (1/2)[pr\{G(r-1, n-r) \geqslant u\} + pr\{G(r, n-r) \geqslant u\}], \\ &\sup_{\boldsymbol{\theta},\boldsymbol{V}} pr(T^{*2} \geqslant u \mid H_0, \boldsymbol{V}) = \\ &\quad (1/2)[pr\{H(r-1, r, n) \geqslant u\} + pr\{G(r, n-r) \geqslant u\}]. \end{aligned} \tag{3.56}$$

■

The results in (3.55) and (3.56) are adequate to carry out the test as in (3.53). These two results are still quite general because the parameter space $\mathcal{P}$ in (3.54) may include nonlinear inequalities in $\boldsymbol{\theta}$. In most practical situations, the inequalities in $\boldsymbol{\theta}$ are linear and independent. In this case we can obtain closed forms for the null distributions of $\mathcal{U}$ and T^{*2}; these are given in the next result.

Proposition 3.10.3 *(Perlman (1969) Corollary 7.6, and Silvapulle (1995)). Let the null and alternative hypotheses be*

$$H_0 : \boldsymbol{\theta}_2 = \mathbf{0} \text{ and } H_1 : \boldsymbol{A}\boldsymbol{\theta}_2 \geqslant \mathbf{0} \tag{3.57}$$

respectively, where $\boldsymbol{A}$ is a given nonsingular matrix. Let $\mathcal{U}$ denote the $\mathcal{U}(C_a, C_b)$ for this testing problem. Then

$$\begin{aligned} pr(\mathcal{U} \geqslant u \mid H_0, \boldsymbol{V}) &= \textstyle\sum_{i=1}^{r} pr\{G(i, n-r) \geqslant u\} w_i(r, \boldsymbol{W}), \quad (3.58) \\ pr(T^{*2} \geqslant u \mid H_0, \boldsymbol{V}) &= \textstyle\sum_{i=1}^{r} pr\{H(k, r, n) \geqslant u\} w_i(r, \boldsymbol{W}). \quad (3.59) \end{aligned}$$

where $\boldsymbol{W} = \boldsymbol{A}^{-1}\boldsymbol{V}_{22}(\boldsymbol{A}^{-1})^T$, and $\boldsymbol{V}_{22}$ is the bottom-right submatrix of $\boldsymbol{V}$ obtained by partitioning $\boldsymbol{V}$ to conform with the partitioning $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$. ■

Note that the expressions on the right-hand side of (3.58) depend on $\boldsymbol{W}$ but not on $\boldsymbol{\theta}$ in the null parameter space. Therefore, we have to use its infimum and supremum over $\boldsymbol{V} > 0$, given in (3.55) and (3.56), to obtain bounds on the p-value and perform the test as in (3.53). These bounds can be quite far from those corresponding to the true $\boldsymbol{W}$ if the sample size is small. This is the price that we pay for not knowing anything about the covariance matrix, $\boldsymbol{V}$. One approach to improve the situation is to narrow down the class of matrices allowed for $\boldsymbol{V}$ based on nonsample information, and then take the supremum of (3.58) over that smaller class. Finding and evaluating the *exact* supremum over such a restricted set of matrices would typically be difficult, but it should be possible to generate a well-spread finite number of matrices within that smaller class, evaluate (3.58) for each $\boldsymbol{V}$ and take the maximum as an estimate of the p-value. If the sample size is large, an estimate of (3.58) is obtained by substituting an estimate $\hat{\boldsymbol{W}}$ of $\boldsymbol{W}$ for $\boldsymbol{W}$ in $w_i(r, \boldsymbol{W})$. One needs to be cautious in using this as a p-value because if $\hat{\boldsymbol{W}}$ is far from $\boldsymbol{W}$, the resulting estimate of the p-value could be a

poor estimate. A more desirable approach would be to compute the expressions on the right-hand side of (3.58) for a range of possible values of W and take the maximum of these values. Even this has a certain degree of subjective element, but it is still better than simply substituting an estimate for W. A formal way of doing this is to take the supremum over a confidence region for W and make some adjustments for using only a confidence region rather than all possible W. A rigorous development of this procedure is provided in the next chapter (see Silvapulle (1996b) and Berger and Boos (1994)).

The *LRT* is biased when $\mathcal{P}$ contains a non-empty open set; see Perlman (1969), p 558. There are points in the alternative parameter space that are close to the null space at which the restricted test, (3.53), has less power than the unrestricted one. However, for the restricted testing problem, we would still prefer the restricted tests to the unrestricted one because the former would have better overall performance (see Silvapulle (1995) and Perlman and Wu (2002a)). The results of a simulation study shows that *LRT* is neither dominated by nor dominates T^{*2}. For *Type A* problems, some calculations show that the T^{*2} is likely to be more powerful than the *LRT* for values for θ near the boundary; for values of θ in the central direction, *LRT* is likely to be more powerful than the T^{*2} (see Silvapulle (1995)). At this stage we do not have sufficient results to recommend one over the other.

3.10.2 Type B Testing Problem

Now, consider the Type B problem of testing

$$H_1 : \theta_2 \in \mathcal{P} \text{ against } H_2 : \theta_2 \notin \mathcal{P}, \tag{3.60}$$

where as in the previous subsection we assume that θ is partitioned as $(\theta_1^T, \theta_2^T)^T$ and $\mathcal{P}$ is assumed to be a one-sided closed cone. In the notation introduced at the beginning of this section, $\mathcal{C}_a = \mathcal{P} \subset \mathcal{C}_b = \mathbb{R}^r$. Let $\mathcal{U}$ denote $\mathcal{U}(\mathcal{C}_a, \mathcal{C}_b)$ for this testing problem and let u denote the sample value of $\mathcal{U}$. It follows from (3.51) and (3.52) that T^{*2} and $\mathcal{U}$ are equal. The null distribution of $\mathcal{U}$ depends on (θ_2, V) and therefore, the test is

$$\text{reject } H_0 \text{ if } \sup_{\theta_2 \in \mathcal{P}, V>0} \text{pr}(\mathcal{U} \geq u \mid \theta, V) \text{ is small.} \tag{3.61}$$

The main result on the null distribution of $\mathcal{U}$ is the following:

Proposition 3.10.4 *(Perlman (1969), section 8). Let $\mathcal{U}$ denote $\mathcal{U}(\mathcal{C}_a, \mathcal{C}_b)$ for the testing problem (3.60). Then we have that,*

1. $\mathcal{U} = min\{(\bar{X}_2 - a)^T S_{22}^{-1} (\bar{X}_2 - a) : a \in \mathcal{P}\}$.
2. $pr(U \geq u \mid H_0, V) \leq pr(U \geq u \mid \theta = 0, V)$, *for any* V.
3. $\sup_{\theta_2 \in C, V>0} pr(U \geq u \mid \theta, V) \leq$

$$(1/2)[pr\{G(r-1, n-r-1) \geq u\} + pr\{G(r, n-r) \geq u\}].$$

4. *If $\mathcal{P}$ contains an open set of dimension r, then*

$$\inf_{\theta_2 \in \mathcal{P}, V>0} pr(LRT \geq t \mid H_0, V) = 0. \quad \blacksquare$$

There are other procedures for testing when V is unknown. None of the tests is uniformly best. A conditional version of the T^{*2} was suggested by Perlman and Wu (2002a); they also compared the performance of several statistics for this testing problem. Likelihood based conditional tests have been proposed by Wang and McDermott(1998a,b) ; the implementation of this procedure is difficult and there appears to be some issues that need to be resolved (see Perlman and Wu (1999)). Other procedures have also been suggested, such as the simple test in Tang (1994). Although the test in Tang (1994) is uniformly more powerful than the likelihood ratio test, it does not follow that this test is better than the likelihood ratio test or the T^{*2} test. These will be discussed later.

3.10.3 Conditional Tests of $H_0 : \theta = 0$ vs $H_1 : \theta \geq 0$

An approach to dealing with the unknown nuisance parameter V is to eliminate it from the null distribution by conditioning on a sufficient statistic for V or modify the test so that the probability of Type I error, as a function of V, is as flat as possible. Tests of the former type were developed by Wang and McDermott(1998a,b), and those of the latter type were developed by Perlman and Wu (2002a). These two types of tests are discussed below in turn.

Wang-McDermott test:

Let $X_1, \ldots, X_n$ be *iid* as X where $X \sim N(\theta, V)$. Let the null and alternative hypotheses be

$$H_0 : \theta = 0 \qquad \text{and} \qquad H_1 : \theta \geq 0$$

respectively. Let $W = \sum X_i X_i^T, S = W - n\bar{X}\bar{X}^T$, $\tilde{\theta}$ be the *mle* of θ subject to $\theta \geq 0$,

$$\mathcal{U} = n\tilde{\theta}^T S^{-1} \tilde{\theta} \{1 + (\bar{X} - \tilde{\theta})^T S^{-1} (\bar{X} - \tilde{\theta})\}^{-1},$$

and u_{obs} be the observed value of $\mathcal{U}$. It was shown earlier that the *unconditional* LRT of H_0 against H_1 is, *reject H_0 for large values of $\mathcal{U}$*. For this test, the p-value is $\sup_{V>0} \text{pr}_V(\mathcal{U} \geq u_{obs})$.

Under H_0, W is a complete sufficient statistic for V, and hence the distribution of $\mathcal{U}$ conditional on $W = w$ does not depend on V or any other unknown parameters. Now, the level-α *conditional* LRT of Wang and McDermott is the following:

$$\text{reject } H_0 \text{ if } \mathcal{U} \geq c_\alpha(w) \text{ where } \text{pr}_0\{\mathcal{U} \geq c_\alpha(w) \mid W = w\} = \alpha.$$

Note that the probability on the LHS does not depend on V. Therefore, the p-value for this conditional test is

$$\text{pr}_0(\mathcal{U} \geq u_{obs} \mid W = w). \tag{3.62}$$

It turns out that the density of $\bar{X}$ under H_0, conditional on $W = w$, has a simple closed form. Further, $\mathcal{U}$ has a simple closed-form for the purposes of computing it. Therefore, the p-value in (3.62) can be computed by using a simple brute force numerical integration or Monte Carlo; an easy-to-implement algorithm for computing the foregoing p-value is given in Wang and McDermott (1998a).

The power function of this conditional test approaches 1 uniformly in (θ, V) as $\theta^T V^{-1}\theta \to \infty$; further, the test is consistent. However, it is biased; see Sen and Tsai (1999).

It also turns out that $c_\alpha^* \geqslant c_\alpha(w)$ where c_α^* and $c_\alpha(w)$ are the level-α critical values of the unconditional and conditional LRTs, respectively. Since the test statistics for the two tests are the same, it follow that the conditional LRT is uniformly more powerful than the unconditional test. It does not automatically follow that the unconditional test is better because the probability of Type I error is higher for the unconditional test although both tests have the same size.

Conditional versions of the T^{*2} discussed in the previous subsections were studied by Wang and McDermott (1998b). Essentially the same ideas are applicable. Briefly the test is

reject H_0 if $T^{*2} > c'_\alpha$ where c'_α satisfies $pr_{0,V}\{T^{*2} > c'_\alpha \mid W = w\} = \alpha$.

Perlman-Wu test:

Now, let us consider the class of conditional tests proposed by Perlman and Wu (2002a). These authors proposed a class of tests in which they condition on K, where K is the number of positive components of $\tilde{\theta}$. In contrast to the statistic W in the foregoing discussions, the statistic K is not sufficient for V, and the resulting conditional test turns out to be not similar. However, Perlman and Wu (2002a) identify a particular member of this class of conditional tests that is nearly similar, in a sense to be made precise in the theorems stated below.

Let C_α denote the class of tests of H_0 against H_1 for which the test statistic is $\mathcal{U}$ and they are conditional on K. Let $c = (c_1, \ldots, c_k)$ where $c_i > 0$ for $i = 1, \ldots, p$. Now, define the conditional test of Perlman and Wu as follows:

Do not reject H_0 if $K = 0$, and reject H_0 if $K = k$ and $\mathcal{U} > c_k$ for $k = 1, \ldots, p$.

Let $T(c)$ denote this test. The unconditional LRT is a special case of this corresponding to $c_1 = \ldots = c_k$. Each choice of c defines a conditional test and it is not similar. One way of choosing an optimal c is to find that value for which the minimum probability of Type I error over the null parameter space is a maximum over all possible values of $\{c\}$. Perlman and Wu (2002a) obtained results in this direction.

Theorem 3.10.5 *(Perlman and Wu (2002a)). For any α, $0 < \alpha < 1$, we have that*

$$\max_{T(c)\in C_\alpha} \inf_{V>0} pr_{0,V}\{T(c) \text{ rejects } H_0\} = \alpha/2.$$

Further, this maximum is attained only by the test $T(c_\alpha)$ where $c_\alpha = (c_{1,\alpha}, \ldots, c_{p,\alpha})^T$ and $pr(\chi_k^2/\chi_{n-p}^2 \geqslant c_{k,\alpha}) = \alpha$ where χ_k^2 and χ_{n-p}^2 are independent ($k = 1, \ldots, p$). ■

The same conditioning approach is applicable to the statistic T^{*2}. The Perlman and Wu (2002a) type conditional version of T^{*2} is the following:

Do not reject H_0 if $K = 0$, and reject H_0 if $K = k$ and $T^{*2}(c) > c_k$

for $k = 1, \ldots, p$. The optimal choice is given in the next theorem.

Theorem 3.10.6 *(Perlman and Wu (2002a)). For any α, $0 < \alpha < 1$, we have that*

$$\max_{T^{*2}(c)\in C_\alpha} \inf_{V>0} pr_{0,V}\{T^{*2}(c) \text{ rejects } H_0\} = \alpha/2.$$

Further, this maximum is attained only by the test $T^{*2}(c_\alpha^*)$ where $c_\alpha^* = (c_{1,\alpha}^*, \ldots, c_{p,\alpha}^*)^T$ and $pr[(\chi_k^2/\chi_{n-p}^2)\{1+(\chi_{p-k}^2/\chi_{n-p-k}^2)\} \geqslant c_{k,\alpha}^*] = \alpha$ where the various χ^2 variates are independent ($k = 1, \ldots, p$). ■

The forgoing conditional tests are unlikely to be admissible. The acceptance regions of $T(c)$ and $T^{*2}(c)$ are neither monotonic nor convex. In fact, Perlman and Wu (2002a) conjecture that the conditional tests of Wang and McDermott (1998a,b) and the unconditional tests, T^{*2} and LRT, are also inadmissible. A simulation study in Perlman and Wu (2002a) compared the performance of the foregoing conditional and unconditional tests; they observed that overall, $T(c)$ and $T^{*2}(c)$ performed better.

3.11 OPTIMALITY PROPERTIES

In the classical unrestricted setting, the parameter spaces tend to be open and several tests tend to have various optimality properties such as UMP, UMPI, unbiasedness, consistency, monotonic power function, etc. However, these optimality properties do not automatically carry over when the parameter space is not open. Further, often different optimality/desirable properties cannot be achieved without sacrificing some other desirable properties. In this section we shall consider consistency and monotonicity of power function. A thorough investigation of all these relevant issues is outside the scope of this book.

3.11.1 Consistency of Tests

Let n denote the sample size and let a level-α test of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \notin \Theta_0$ be " reject H_0 if $T_n \geqslant c_n$" where T_n is a test statistic and c_n is the critical value. As the sample size increases, one would expect that the information in the sample about θ would also increase, and, hence, if T_n is a good test statistic then the probability of rejecting H_0 would tend 1 as $n \to \infty$ for $\theta \notin \Theta_0$. This leads to the following definition. A test of $H_0 : \theta \in \Theta_0$ is said to be *consistent* at θ ($\notin \Theta_0$) if the probability of rejecting the null hypothesis tends to 1 as $n \to \infty$ when the true value of the parameter is θ. Tests under inequality constraints are usually consistent at points in the alternative parameter space, but not necessarily at every point in the parameter space. Let us consider a simple example in two-dimensions to illustrate the main idea.

Example 3.11.1 *Consistency of the LRT in 2-dimensions.*

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be *iid* as $N(\boldsymbol{\theta}, \boldsymbol{I})$, and let the null and alternative hypotheses be defined as

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \text{ and } H_1 : \boldsymbol{\theta} \geqslant \mathbf{0},$$

respectively, where $\boldsymbol{\theta} \in \mathbb{R}^2$. Then, $LRT = n\{\|\bar{\boldsymbol{X}}\|^2 - \|\bar{\boldsymbol{X}} - \mathbb{R}^{+2}\|^2\}$. First, let us consider the case when the true value of $\boldsymbol{\theta}$ satisfies $\theta_1 > 0$ and $\theta_2 > 0$. Since $\bar{\boldsymbol{X}} \xrightarrow{p} \boldsymbol{\theta}$, it follows that, with probability tending to 1, $\bar{\boldsymbol{X}}$ lies in an arbitrarily small neighborhood of $\boldsymbol{\theta}$ contained in $\mathbb{R}^{+2}$ and $LRT = n\|\bar{\boldsymbol{X}}\|^2$; therefore, $LRT \xrightarrow{p} \infty$ as $n \to \infty$. Hence, LRT is consistent at $\boldsymbol{\theta}$ when $\theta_1 > 0$ and $\theta_2 > 0$. Similarly, by considering the interior and the boundary of each of the other three quadrants separately, it is easily seen that

$$\text{LRT is consistent at } \boldsymbol{\theta} \text{ if and only if } \boldsymbol{\theta} \notin \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_1 \leq 0, \theta_2 \leq 0\}.$$

In fact, it may be verified that $lim_{n\to\infty} pr\{LRT \text{ rejects } H_0 | \boldsymbol{\theta}\}$ is 1, less than 1, or zero according as $\boldsymbol{\theta}$ is not in Q_3, lies on the boundary of Q_3, or lies in the interior of Q_3, where $Q_3 = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_1 \leq 0, \theta_2 \leq 0\}$.

Now consider the LRT of $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$ against $H_2 : \boldsymbol{\theta} \ngeqslant \mathbf{0}$. Again, by considering the interior and boundary of each quadrant separately, we have that

$$\text{LRT is consistent at any } \boldsymbol{\theta} \notin \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_1 \geqslant 0, \theta_2 \geqslant 0\};$$

further, $lim_{n\to\infty} pr\{LRT \text{ rejects } H_1 | \boldsymbol{\theta}\}$ is 1, less than 1, or zero according as $\boldsymbol{\theta}$ is not in $\mathbb{R}^{+2}$, lies on the boundary of $\mathbb{R}^{+2}$, or lies in the interior of $\mathbb{R}^{+2}$. ■

Essentially the same arguments can be used to study the consistency properties of more general tests of inequality constrained hypotheses (see Problem 3.29). For the rest of this section all inner products will be assumed to be with respect to a given positive definite matrix, $\boldsymbol{V}$. Let $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n$ be *iid* as $\boldsymbol{X}$ where $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ and $\boldsymbol{V}$ is known. Then we have the following:

(1) Let the null and alternative hypotheses be $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ and $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, respectively. Then $LRT = n\|\bar{\boldsymbol{X}} - \mathcal{P}\|^2$ where $\mathcal{P} = (\mathcal{C}^o \oplus \mathcal{M}) = (\mathcal{C} \cap \mathcal{M}^\perp)^o$; see Theorem 3.7.1 on page 84 and Proposition 3.12.6 on page 118. Since $(\bar{\boldsymbol{X}} - \boldsymbol{\theta}) = o_p(1)$ and $\mathcal{P}$ is a closed set, it follows that LRT is consistent if and only if $\boldsymbol{\theta} \notin \mathcal{P}$. As a special case, let $\mathcal{M} = \{\mathbf{0}\}$. Then, since $LRT = n\|\bar{\boldsymbol{X}} - \mathcal{C}^o\|^2$, $(\bar{\boldsymbol{X}} - \boldsymbol{\theta}) = o_p(1)$ and $\mathcal{C}^o$ is a closed set, it follows that LRT is consistent if and only if $\boldsymbol{\theta} \notin \mathcal{C}^o$.
(2) Let the null and alternative hypotheses be $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ and $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$, respectively. Then $LRT = n\|\bar{\boldsymbol{X}} - \mathcal{C}\|^2$, $(\bar{\boldsymbol{X}} - \boldsymbol{\theta}) = o_p(1)$ and $\mathcal{C}$ is a closed set. Therefore, LRT is consistent at any $\boldsymbol{\theta} \notin \mathcal{C}$.

Now, let us consider the case when $\boldsymbol{V}$ is unknown, and the null and alternative hypotheses are $H_0 : \boldsymbol{\theta} = \mathbf{0}$ and $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, respectively. In this case, the *LRT* statistic is consistent (see Wang and McDermott (1998a) p 383). Sen and Tsai (1999) note that essentially the same arguments are applicable to T^{*2} because $T^{*2} \geqslant LRT$; hence T^{*2} is also consistent. Wang and McDermott (1998) also showed that their conditional *LRT* is consistent.

It may also be of interest to study the behavior of the power when the effect of the sample size is absorbed into the value of $\boldsymbol{\theta}$. Problem 3.30 is a result along this line.

3.11.2 Monotonicity of the Power Function

For the simple univariate case of testing $H_0 : \mu = 0$ against $H_1 : \mu > 0$ based on a single observation of X where $X \sim N(\mu, 1)$, the power of a test with critical region $\{x \in \mathbb{R} : x \geqslant c\}$, for any fixed $c > 0$, increases to 1 as μ increases to ∞. Similarly, it is also known that for testing $\boldsymbol{\theta} = \mathbf{0}$ against $\boldsymbol{\theta} \neq \mathbf{0}$ based on a single observation of $\boldsymbol{X}$ from $N(\boldsymbol{\theta}, \boldsymbol{V})$, where $\boldsymbol{V}$ is a given positive definite matrix, the power of the LRT ($= \|\boldsymbol{X}\|_{\boldsymbol{V}}^2$) at $k\boldsymbol{\Delta}$ increases to 1 as k increases from zero to ∞, for any fixed $\boldsymbol{\Delta} \neq \mathbf{0}$ (see Anderson (1955)); this result holds even when the distribution of $\boldsymbol{X}$ is not normal but symmetric and unimodal (see Theorem 3.11.4).

The result of Anderson (1955) (stated as Theorem 3.11.4 on page 111) is applicable when the acceptance region is convex and symmetric. Therefore, it is usually not applicable when there are order restrictions on the parameters because the acceptance region is usually not symmetric and further it may not be even convex. However, it turns out that the power function of several tests when there are order restrictions on $\boldsymbol{\theta}$ are monotonic along certain straight lines. The main results concerning the monotonicity of power are stated below, and the proofs are given for some of them; for the others, references are given where detailed proofs may be found.

There are essentially two approaches to study the monotonicity of the power function. The simplest one is called the *set containment argument*, similar to that in Lemma 8.2 of Perlman (1969). This uses the fact $\mathcal{C} + \boldsymbol{\theta} \subset \mathcal{C}$ when $\boldsymbol{\theta} \in \mathcal{C}$ and hence the distance from any given point to $\mathcal{C} + \boldsymbol{\theta}$ is not more than the distance from the same point to $\mathcal{C}$. The second method is *geometric*. This approach starts with the fact that the power of a test is equal to the volume of a region under the probability density function and above the critical region and uses geometric properties of this volume. So far, general results using such a geometric approach have assumed that the density function is elliptically symmetric and unimodal; the set containment argument does not require the density function to be elliptically symmetric or unimodal. Therefore, the results based on set containment arguments are applicable to a broader class of distributions. However, there are cases in which the geometric approach works but not the set containment argument; in some cases the converse is true. Therefore, both approaches are useful and essential.

For the rest of this section, the inner product (and hence $\|.\|, \Pi()$, etc) will be assumed to be with respect to a given positive definite matrix $\boldsymbol{V}$, for tests based on a single observation.

Proposition 3.11.1 *Let $\boldsymbol{X} = \boldsymbol{\theta} + \boldsymbol{E}$ where the distribution of $\boldsymbol{E}$ does not depend on any unknown parameters. Let $\boldsymbol{V}$ be a given positive definite matrix, $\pi_{01}(\boldsymbol{\theta})$ denote the power function of $\|\boldsymbol{X} - (\mathcal{C} \cap \mathcal{M}^\perp)^o\|_{\boldsymbol{V}}^2$ for testing $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$; similarly, let $\pi_{12}(\boldsymbol{\theta})$ denote the power function of $\|\boldsymbol{X} - \mathcal{C}\|_{\boldsymbol{V}}^2$ for testing $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$. Then, $\pi_{01}(\boldsymbol{\theta}) \leq \pi_{01}(\boldsymbol{\theta}')$ for $\boldsymbol{\theta} - \boldsymbol{\theta}' \in \mathcal{C}^o$; and $\pi_{12}(\boldsymbol{\theta}) \geqslant \pi_{12}(\boldsymbol{\theta}')$ for $\boldsymbol{\theta}' - \boldsymbol{\theta} \in \mathcal{C}$.*

Proof: The proof follows a set containment argument. To prove the first, let $\boldsymbol{\theta} - \boldsymbol{\theta}' \in \mathcal{C}^o$. Then $\mathcal{C}^o + (\boldsymbol{\theta} - \boldsymbol{\theta}') \subset \mathcal{C}^o$, and hence $\mathcal{C}^o - \boldsymbol{\theta}' \subset \mathcal{C}^o - \boldsymbol{\theta}$. Now, $\pi_{01}(\boldsymbol{\theta}) =$

$\text{pr}\{\|\boldsymbol{X}-\mathcal{C}^o\|_V^2 \geqslant c|\boldsymbol{\theta}\} = \text{pr}\{\|\boldsymbol{E}+\boldsymbol{\theta}-\mathcal{C}^o\|_V^2 \geqslant c\} = \text{pr}\{\|\boldsymbol{E}-(\mathcal{C}^o-\boldsymbol{\theta})\|_V^2 \geqslant c\} \leq \text{pr}\{\|\boldsymbol{E}-(\mathcal{C}^o-\boldsymbol{\theta}')\|_V^2 \geqslant c\} = \pi_{01}(\boldsymbol{\theta}')$.

To prove the second, let $\boldsymbol{\theta}'-\boldsymbol{\theta} \in \mathcal{C}$. Then $\mathcal{C}+(\boldsymbol{\theta}'-\boldsymbol{\theta}) \subset \mathcal{C}$, and hence $\mathcal{C}-\boldsymbol{\theta} \subset \mathcal{C}-\boldsymbol{\theta}'$. Now, $\pi_{12}(\boldsymbol{\theta}) = \text{pr}\{\|\boldsymbol{X}-\mathcal{C}\|_V^2 \geqslant c|\boldsymbol{\theta}\} = \text{pr}\{\|\boldsymbol{E}-(\mathcal{C}-\boldsymbol{\theta})\|_V^2 \geqslant c\} \geqslant \text{pr}\{\|\boldsymbol{E}-(\mathcal{C}-\boldsymbol{\theta}')\|_V^2 \geqslant c\} = \pi_{12}(\boldsymbol{\theta}')$. ■

A simple diagram can be used to illustrate this result. The first part of the result says that for testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, the power of the test at any point, say $\boldsymbol{\theta}_0$, increases in any direction of $-\mathcal{C}^o$ and decreases in any direction of $\mathcal{C}^o$. Similarly, the second part of the result says that for testing $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$, the power of the test at any point, say $\boldsymbol{\theta}_0$, increases in any direction of $-\mathcal{C}$ and decreases in any direction of $\mathcal{C}$.

Now, we state the main results that have been obtained using the geometric approach. Let h be a nonincreasing function on $[0,\infty)$, $\boldsymbol{V}$ be a positive definite matrix of order $p \times p$, and $\boldsymbol{X}$ be a random variable with a density function $|\boldsymbol{V}|^{-1/2}h(\|\boldsymbol{x}-\boldsymbol{\theta}\|^2)$; a density function of this form is said to be *elliptically symmetric* and *unimodal* about $\boldsymbol{\theta}$. Let

$$p(\boldsymbol{\theta}) = 1 - \int_A |\boldsymbol{V}|^{-1/2}h(\|\boldsymbol{x}-\boldsymbol{\theta}\|_V^2)d\boldsymbol{x}, \qquad A \subset \mathbb{R}^p, \boldsymbol{\theta} \in \mathbb{R}^p.$$

For several tests, including the LRT, of $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ and of $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$, the power of the test at $\boldsymbol{\theta}$ takes the form $p(\boldsymbol{\theta})$, where A is the acceptance region. Therefore, monotonicity properties of functions of the form $p(\boldsymbol{\theta})$ are of interest. Anderson's theorem is applicable to establish the monotonicity of $p(\boldsymbol{\theta})$ when A is symmetric and convex. When there are inequality constraints, the acceptance region A is usually not symmetric and it may not be even convex. Mukerjee et al. (1986) studied the monotonicity properties of $p(\boldsymbol{\theta})$ when A satisfies certain conditions that include that A be convex but not necessarily symmetric. They showed that $p(k\boldsymbol{\theta}_0)$ is monotonic in $|k|$ for $\boldsymbol{\theta}_0 \in \mathbb{R}^p$; their main result (stated as Theorem 3.11.5 on page 111) leads to the following useful corollary.

Corollary 3.11.2 *(Mukerjee et al. (1986)). Let the distribution of $\boldsymbol{X}$ be elliptically symmetric and unimodal about $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_1 \in \mathbb{R}^p$ and $\boldsymbol{\theta}_2 \in \mathcal{C}$. Then as a function of k, $\pi_{12}(\boldsymbol{\theta}_1 + k\boldsymbol{\theta}_2)$ is nondecreasing for $k \in (-\infty,\infty)$ and $\pi_{12}(\boldsymbol{\theta}_1 + k\tilde{\boldsymbol{\theta}}_1)$ is nondecreasing for $k \geqslant -1$ where $\tilde{\boldsymbol{\theta}}_1 = \Pi(\boldsymbol{\theta}_1|\mathcal{C}^o)$.* ■

It is also shown in Robertson et al. (1988), page 106, that (a) $\pi_{01}(k\boldsymbol{\Delta})$ is nondecreasing in k, where $(k > 0)$ and $\boldsymbol{\Delta} \in \mathcal{C}$, and (b) $\pi_{12}(k\boldsymbol{\Delta})$ and $\pi_{12}^*(k\boldsymbol{\Delta})$ are nondecreasing in k, where $k > 0$, $\boldsymbol{\Delta} \in \mathcal{C}^o \oplus \mathcal{M}$ and π_{12}^* is the power function of the $\bar{E}^2$-test.

For some $\bar{E}^2$-tests, the acceptance region is not convex and hence the results of Mukerjee et al. (1986) are not applicable. Hu and Wright (1994b) studied the monotonicity of $p(\boldsymbol{\theta}_0 + t\boldsymbol{\theta}_1)$ in t for some given $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ and when A takes a special form; their regularity conditions do not require A to be convex or symmetric. Their result extends some of the results in Mukerjee et al. (1986), and is applicable for the $\bar{E}^2$-test of H_0 against H_1. A useful corollary of their main result (see Theorem 3.11.6) is the following:

Corollary 3.11.3 *(Hu and Wright (1994b)). Suppose that $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are iid as $N(\boldsymbol{\theta},\sigma^2\boldsymbol{U})$ where $\boldsymbol{U}$ is known and σ is unknown. Then, the power function of the LRT (namely, $\bar{E}^2$-test) of $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ is nondecreasing on each line segment starting at a point in $\mathcal{M}$ and continuing in the direction of a vector in $\mathcal{C}$. Consequently, this test is unbiased as well.* ■

Since $\bar{F} = \bar{\chi}^2/S^2$ and S^2 is independent of $\boldsymbol{\theta}$, monotonicity of $\bar{\chi}^2$ carry over to $\bar{F}$ as well; hence, monotonicity of $\bar{F}$ is not studied separately.

The main results concerning the monotonicity of $p(\boldsymbol{\theta})$ are stated below for completeness.

Theorem 3.11.4 *(Anderson (1955)). Let A be a convex and symmetric set in $\mathbb{R}^p$. Let f be a function such that (i) $f(\boldsymbol{x}) = f(-\boldsymbol{x})$, (ii) $\{\boldsymbol{x} : f(\boldsymbol{x}) \geqslant u\}$ is convex for every $u > 0$, and (iii) $\int_A f(\boldsymbol{x})d\boldsymbol{x} < \infty$ (in the Lebesgue sense). Then $\int_A f(\boldsymbol{x}+k\boldsymbol{y}) \geqslant \int_A f(\boldsymbol{x}+\boldsymbol{y})d\boldsymbol{y}$ for $0 \leq k \leq 1$.* ■

Theorem 3.11.5 *(Mukerjee et al. (1986)). Let $\boldsymbol{\theta} \in \mathbb{R}^p$ and $S_{\boldsymbol{\theta}} = \{b\boldsymbol{\theta} : -\infty < b < \infty\}$ be the linear space spanned by $\boldsymbol{\theta}$; for a given convex set $A \subset \mathbb{R}^p$ let the positive part of A in the direction of $\boldsymbol{\theta}$ be defined by*

$$A^+ = \{\boldsymbol{x} \in A : \Pi_{\boldsymbol{V}}(\boldsymbol{x}|S_{\boldsymbol{\theta}}) = b\boldsymbol{\theta} \text{ where } b > 0\}.$$

Let f be an elliptically symmetric and unimodal density function. Suppose that $\boldsymbol{x} - 2\Pi_{\boldsymbol{V}}(\boldsymbol{x}|S_{\boldsymbol{\theta}}) \in A$ for each $\boldsymbol{x} \in A^+$. Then, $\int_{A-k\boldsymbol{\theta}} f(\boldsymbol{x})d\boldsymbol{x}$ is a nonincreasing function of k for $k > 0$. ■

Theorem 3.11.6 *(Hu and Wright (1994b)). Suppose that $\boldsymbol{X}$ has the unimodal and elliptically symmetric density $|\boldsymbol{V}|^{-1/2}h(\|\boldsymbol{x}\|_{\boldsymbol{V}}^2)$ where h is nonincreasing on $[0,\infty)$. Let $\mathcal{C} \subset \mathbb{R}^p$ be a closed convex cone, $\mathcal{M}$ be a linear subspace such that $\mathcal{M} \subset \mathcal{C}$, and*

$$A = \{\boldsymbol{x} \in \mathbb{R}^p : \|\Pi_{\boldsymbol{V}}(\boldsymbol{x}|\mathcal{M}) - \Pi_{\boldsymbol{V}}(\boldsymbol{x}|\mathcal{C})\|_{\boldsymbol{V}}^2 \leq a + b\|\boldsymbol{x} - \Pi_{\boldsymbol{V}}(\boldsymbol{x}|\mathcal{C})\|_{\boldsymbol{V}}^2\},$$

where $a > 0$ and $b > 0$. If $\boldsymbol{\mu}_0 \in \mathcal{M}$ and $\boldsymbol{\nu}_0 \in \mathcal{C}$ then $pr(\boldsymbol{X} + (\boldsymbol{\mu}_0 + t\boldsymbol{\nu}_0) \in A)$, is nonincreasing in t for $t > 0$. ■

Now let us consider the case when the covariance matrix $\boldsymbol{V}$ is completely unknown, and the null and alternative hypotheses are $H_0 : \boldsymbol{\theta} = \mathbf{0}$ and $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, respectively. The power function of *LRT* and T^{*2} are monotonic. In particular, the set containment argument can be used to establish that the power functions of *LRT* and T^{*2} at $\boldsymbol{\theta}_0$ increase in any direction of $\mathcal{C}$ (see Sen and Tsai (1999)).

3.12 APPENDIX 1: CONVEX CONES, POLYHEDRALS, AND PROJECTIONS

3.12.1 Introduction

This Appendix provides a brief account of the main results on projections onto convex cones in finite dimensional spaces as they relate to the statistical inference problems

discussed in this book. An excellent account of projections onto convex cones is given in Stoer and Witzgall (1970)(SW); other references include Bazaraa et al. (1993) and Hiriart-Urruty and Lemaréchal (1993). Zarantonello (1998) provides an extensive account of projections onto convex sets in Hilbert spaces; while some of the results and ideas therein are relevant, such general results are not required at this stage. In this Appendix, we provide a reasonably self-contained discussion that would be adequate for this book.

Let $\mathbb{R}^p$ denote the p-dimensional Euclidean space, $\boldsymbol{V}$ be a $p \times p$ symmetric positive definite matrix, $\boldsymbol{x} \in \mathbb{R}^p$, and $\boldsymbol{y} \in \mathbb{R}^p$. Then, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{V}} = \boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{y}$ defines an *inner product* on $\mathbb{R}^p$. This induces the corresponding *norm* $\|\boldsymbol{x}\|_{\boldsymbol{V}} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{V}}^{1/2}$; the corresponding *distance* between $\boldsymbol{x}$ and $\boldsymbol{y}$ is $\|\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{V}}$. If $\boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{y} = 0$ then we say that $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* with respect to $\boldsymbol{V}$, which we denote by $\boldsymbol{x} \perp_{\boldsymbol{V}} \boldsymbol{y}$. We may abbreviate "orthogonal with respect to $\boldsymbol{V}$", "distance with respect to $\boldsymbol{V}$" to $\boldsymbol{V}$-orthogonal, $\boldsymbol{V}$-distance, etc. However, if $\boldsymbol{V}$ is obvious we will drop the reference to $\boldsymbol{V}$.

If $\boldsymbol{x} \perp_{\boldsymbol{V}} \boldsymbol{y}$ then a version of the Pythagoras theorem holds: namely, $\|\boldsymbol{x} + \boldsymbol{y}\|_{\boldsymbol{V}}^2 = \|\boldsymbol{x}\|_{\boldsymbol{V}}^2 + \|\boldsymbol{y}\|_{\boldsymbol{V}}^2$. A consequence of this is that the shortest distance between a point and a plane is the distance from the point to the plane along a line that is orthogonal to the plane; here, distance and orthogonality are with respect to $\langle , \rangle_{\boldsymbol{V}}$.

Let C be a closed convex set in $\mathbb{R}^p$ and $\boldsymbol{x} \in \mathbb{R}^p$. Let $\tilde{\boldsymbol{x}}$ in C be the point in C that is closest to $\boldsymbol{x}$ with respect to the distance $\|\cdot\|_{\boldsymbol{V}}$; thus,

$$\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| = \min_{\boldsymbol{\theta} \in C} (\boldsymbol{x} - \boldsymbol{\theta})^T \boldsymbol{V}^{-1} (\boldsymbol{x} - \boldsymbol{\theta}).$$

The vector $\tilde{\boldsymbol{x}}$ is called the *projection* of $\boldsymbol{x}$ onto C, and is denoted by $\Pi_{\boldsymbol{V}}(\boldsymbol{x} \mid C)$ (SW p 47); thus

$$\tilde{\boldsymbol{x}} = \Pi_{\boldsymbol{V}}(\boldsymbol{x} \mid C) = \arg \min_{\boldsymbol{\theta} \in C} (\boldsymbol{x} - \boldsymbol{\theta})^T \boldsymbol{V}^{-1} (\boldsymbol{x} - \boldsymbol{\theta}).$$

Almost all the results on projections onto convex cones discussed here are generalizations of the corresponding results for projections onto linear spaces. Therefore, the basic results on the latter are essential for the rest of this appendix. For detailed discussions of relevant results on projections onto linear spaces, see textbooks on linear models, such as Arnold (1981), Chapter 2, or Rao (1972).

We may assume, without loss of generality, that $\boldsymbol{V} = \boldsymbol{I}$ when considerations are restricted to $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{V}}$ where $\boldsymbol{x}$ and $\boldsymbol{y}$ are elements of $\mathbb{R}^p$. To show this, let $\boldsymbol{V}^{-1} = \boldsymbol{U}^T \boldsymbol{U}$ be a factorization of $\boldsymbol{V}^{-1}$, for example, the Cholesky factorization. Now apply the invertible linear transformation $\boldsymbol{U}$ on $\mathbb{R}^p$. Then, any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^p$ are mapped to $\boldsymbol{p}$ and $\boldsymbol{q}$ in $\mathbb{R}^p$ where $\boldsymbol{p} = \boldsymbol{U}\boldsymbol{x}$ and $\boldsymbol{q} = \boldsymbol{U}\boldsymbol{y}$. Since $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{V}} = \boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{y} = (\boldsymbol{U}\boldsymbol{x})^T (\boldsymbol{U}\boldsymbol{y}) = \boldsymbol{p}^T \boldsymbol{q} = \langle \boldsymbol{p}, \boldsymbol{q} \rangle_{\boldsymbol{I}}$, the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{V}}\}$ is equal to the distance between their images $\boldsymbol{p}$ and $\boldsymbol{q}$ in $\{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{I}}\}$. Therefore, for the purposes of studying properties that depend only on distances between points, we may identify $\{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{V}}\}$ with $\{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{I}}\}$ and, hence, without loss of generality, assume that $\boldsymbol{V}$ is the identity matrix. Instead of writing $\langle \boldsymbol{p}, \boldsymbol{q} \rangle_{\boldsymbol{I}}$ and $\|\boldsymbol{p} - \boldsymbol{q}\|_{\boldsymbol{I}}$ we shall use the simpler notation, $\boldsymbol{p}^T \boldsymbol{q}$ and $\|\boldsymbol{p} - \boldsymbol{q}\|$, respectively.

Remark: Since $\boldsymbol{U}$ and $\boldsymbol{U}^{-1}$ are one-to-one, onto and, continuous, $\boldsymbol{U} : \{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{V}}\} \to \{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{I}}\}$ is a *homeomorphism* and hence $\{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{V}}\}$ and $\{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{I}}\}$ are *homeomorphic*. Thus, these two spaces have the same topological structure. Since $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{V}} = \langle \boldsymbol{U}\boldsymbol{x}, \boldsymbol{U}\boldsymbol{y} \rangle_{\boldsymbol{I}}$, $\boldsymbol{U}$ is an *isomorphism* and $\{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{V}}\}$ and $\{\mathbb{R}^p, \langle \cdot \rangle_{\boldsymbol{I}}\}$ are *isomorphic*.

A set S is said to be the *orthogonal sum* of S_1 and S_2 if $S = S_1 + S_2 (= \{s_1 + s_2 : s_1 \in S_1, s_2 \in S_2\})$ and every element $\boldsymbol{x} \in S$ admits a unique *orthogonal decomposition* of the form $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ with $\boldsymbol{y} \in S_1, \boldsymbol{z} \in S_2$ and $\boldsymbol{y}^T \boldsymbol{z} = 0$. Such an orthogonal sum will be denoted by $S_1 \oplus S_2$. As an example, the *direct sum* of two orthogonal linear subspaces is an orthogonal sum.

For any set $S \subset \mathbb{R}^p$, its *orthogonal complement* $S^{\perp}$ is defined as $\{\boldsymbol{y} \in \mathbb{R}^p : \boldsymbol{y}^T \boldsymbol{x} = 0, \text{ for all } \boldsymbol{x} \in S\}$. Clearly, $S^{\perp}$ is a linear space. For a set $S \subset \mathbb{R}^p$, we define the *polar cone* S^o as

$$S^o = \{\boldsymbol{x} : \boldsymbol{x}^T \boldsymbol{y} \leq 0 \text{ for all } \boldsymbol{y} \in S\}.$$

The next result states a few important but elementary properties of cones and polar cones; the proofs are easy and hence omitted.

Proposition 3.12.1 *Let C be a closed convex cone and P and Q be subsets of $\mathbb{R}^p$. Then we have the following:*

(1) C is closed under addition: $\boldsymbol{x} \in C$ and $\boldsymbol{y} \in C \Rightarrow \boldsymbol{x} + \boldsymbol{y} \in C$.

(2) $C + \boldsymbol{x} \subset C \subset C - \boldsymbol{x}, \quad \forall \boldsymbol{x} \in C$.

(3) P^o is a closed convex cone.

(4) If $P \subset Q$ then $Q^o \subset P^o$.

(5) If P is a linear subspace then $P^{\perp} = P^o$. ■

Clearly, we can also define orthogonal sum, orthogonal complement, polar cone, etc. with respect to any symmetric and positive definite matrix, $\boldsymbol{V}$; all that it requires is to replace $\boldsymbol{x}^T \boldsymbol{y}$ by $\boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{y}$ in the foregoing definitions. For example, the *polar cone of S with respect to $\boldsymbol{V}$* is $\{\boldsymbol{x} : \boldsymbol{x}^T \boldsymbol{V}^{-1} \boldsymbol{y} \leq 0 \text{ for all } \boldsymbol{y} \in S\}$. This can be denoted by $S^o_{\boldsymbol{V}}$; however, we will write S^o instead if the suffix $\boldsymbol{V}$ is obvious. Similarly, the *orthogonal complement of S with respect to $\boldsymbol{V}$* is defined as $\{\boldsymbol{y} \in \mathbb{R}^p : \boldsymbol{y}^T \boldsymbol{V}^{-1} \boldsymbol{x} = 0, \text{ for all } \boldsymbol{x} \in S\}$. This can be denoted by $S^{\perp}_{\boldsymbol{V}}$, although we would typically write $S^{\perp}$ if the suffix $\boldsymbol{V}$ is obvious. In what follows, the main results are presented for the case $\boldsymbol{V} = \boldsymbol{I}$; however, the corresponding results for a general $\boldsymbol{V}$ would be obvious.

For statistical inference based on the linear model, projections onto linear spaces play an important role. The corresponding results on projections onto convex cones play a similar important role in constrained statistical inference in the exact normal and large sample theory.

3.12.2 Projections onto Convex Cones

Before we consider projections onto convex cones, let us recall some results on projections onto linear spaces. Let $\boldsymbol{P}$ be a $p \times p$ matrix. We say that $\boldsymbol{P}$ is a *projection matrix* if $\boldsymbol{P} = \boldsymbol{P}^T$ and $\boldsymbol{P}^2 = \boldsymbol{P}$.

Proposition 3.12.2 *Let P be $p \times p$ projection matrix, $Q = I - P$, $\mathcal{L}_P = \{Px : x \in \mathbb{R}^p\}$ and $\mathcal{L}_Q = \{Qx : x \in \mathbb{R}^p\}$. Then we have the following: (1) P is positive semi-definite. (2) Q is a projection matrix. (3) $\mathcal{L}_P$ and $\mathcal{L}_Q$ are orthogonal linear subspaces. (4) Any $x \in \mathbb{R}$ can be represented uniquely as $x = p + q$ where $p \in \mathcal{L}_P$ and $q \in \mathcal{L}_Q$.*

Proof: See Bazaraa et al. (1993), page 448, for example. ■

Let $\mathcal{L}$ be a linear subspace of $\mathbb{R}^p$. Some results on projections onto linear spaces that are important for inference in linear models, include the following:

1. The projection $\Pi(x|\mathcal{L})$ of x onto $\mathcal{L}$ exists and is unique.

2. Let $\mathcal{L} = \{X\beta : \beta \in \mathbb{R}^k\}$, and $P = X(X^TX)^{-1}X^T$, where $k \leqslant p$, X is $p \times k$ and $\text{rank}(X) = k$. Then, $\Pi(x|\mathcal{L}) = Px$ for every x and P is called the *projection matrix* onto $\mathcal{L}$.

3. $x - \Pi(x|\mathcal{L})$ is the projection of x onto $\mathcal{L}^\perp$, and the projection matrix onto $\mathcal{L}^\perp$ is $(I - P)$.

4. If x is projected onto $\mathcal{L}$ and then the projection itself is projected onto a subspace $\mathcal{M}$ of $\mathcal{L}$, the resulting projection is the projection of x onto $\mathcal{M}$ [i.e., $\Pi\{\Pi(x|\mathcal{L})|\mathcal{M}\} = \Pi(x|\mathcal{M})$].

The essence of many of these results carries over to projections onto convex cones and plays similar important roles in constrained statistical inference. The main difference that causes technical difficulties is that the projection of x onto a convex cone is not a linear function of x and hence it cannot be expressed as Qx where x is arbitrary and Q does not depend on x. The geometric nature of projections onto linear spaces and their relevance in the context of statistical inference in linear models appear in standard textbooks on linear models. Therefore, we do not discuss such results here, but provide a discussion of the corresponding results for closed convex cones and schematic diagrams to help interpret the results.

Just as in the case for linear spaces, projection of a point onto a closed convex set exists uniquely, and it can be characterized by the angle between the line of projection and the set. These results are given below.

Proposition 3.12.3 *Let C be a nonempty closed convex set in $\mathbb{R}^p$ and $a \in \mathbb{R}^p \setminus C$. Then we have the following:*

(a) *There exists a unique $a^* \in C$ that is closest to a in the sense that, for $x \in C$, $\|a - x\|$ is a minimum at $x = a^*$.*

(b) *The point $a^* \in C$ is the unique point in C closest to a if and only if $(a - a^*)^T(x - a^*) \leq 0$ for all $x \in C$.*

(c) *There exists a vector p and a scalar α such that $p^Ta > \alpha$ and $p^Tx \leq \alpha$ for every $x \in C$.*

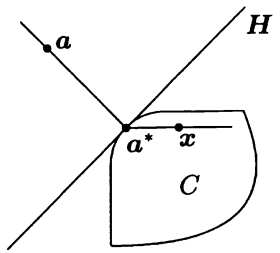

Fig. 3.14 The projection a^* of a onto the convex set C; the plane H, which is orthogonal to $a - a^*$, separates C and the point a

[For (a), see SW 3.3.1 or Bazaraa et al. (1993), page 43; for (b) see Bazaraa et al. (1993), page 43 or Hiriart-Urutty and Lemaréchal (1993), page 117]

Proof of (a): Let $f : C \to \mathbb{R}$ be defined by $f(x) = \|a - x\|$. Let $x_0 \in C$ and $S = \{y \in \mathbb{R}^p : f(y) \leq f(x_0)\}$. Then $S \cap C$ is nonempty and compact. Since f is continuous, it has a minimum over $S \cap C$, say at a^*. Clearly, $f(a^*) < f(x)$ for $x \notin S$. Therefore, a^* is also the global minimum of $f(x)$ over C. Now, to show that a^* is unique, suppose that $\tilde{a} \in C, \tilde{a} \neq a^*$, and $\|a - \tilde{a}\| = \|a - a^*\|$. Let $\bar{a} = 0.5(a^* + \tilde{a})$. Clearly, $\bar{a} \in C$ because C is convex. Further, we have that $\|a - \bar{a}\|^2 = 0.5\|a - a^*\|^2 + 0.5\|a - \tilde{a}\|^2 - 0.25\|\tilde{a} - a^*\|^2 < \|a - a^*\|^2$. Thus, $\bar{a}$ is closer to a than a^*; this is a contradiction. Therefore, a^* is unique.

Proof of (b): Assume that $a^* \in C$ and $(a - a^*)^T(x - a^*) \leq 0$ for every $x \in C$. Let $x \in C$. Then $\|a - x\|^2 = \|a - a^*\|^2 + \|a^* - x\|^2 + 2(a - a^*)^T(a^* - x)$. Since $(a - a^*)^T(a^* - x) \geqslant 0$ by assumption, we have that $\|a - x\|^2 \geqslant \|a - a^*\|^2$. Therefore, a^* is the point in C closest to a. To prove the converse, assume that $a^* \in C$ is the point in C closest to a. Let $x \in C$. Then, $a^* + \lambda(x - a^*) \in C$ for $0 < \lambda \leq 1$ because C is convex. Now, since a^* is closer to a than $a^* + \lambda(x - a^*)$, we have $\|a - a^*\|^2 \leq \|a - \{a^* + \lambda(x - a^*)\}\|^2 = \|a - a^*\|^2 + \lambda^2\|x - a^*\|^2 - 2\lambda(a - a^*)^T(x - a^*)$. Therefore, $2\lambda(a - a^*)^T(x - a^*) \leq \lambda^2\|a - a^*\|^2$. Now, divide by $\lambda(> 0)$ and take the limit as $\lambda \to 0^+$. This leads to $(a - a^*)^T(x - a^*) \leq 0$ for every $x \in C$.

Proof of (c): Let $a^* = \Pi(a \mid C)$, p be a unit vector parallel to $(a - a^*)$ and $\alpha = p^Ta^*$. Now, the proof follows from part (b). ■

Fig. 3.14 illustrates the nature of the foregoing results. It shows that if $\Pi(a|C)$ is treated as the origin then the angle between a and x must be obtuse for any $x \in C$ (the *angle* θ between the vectors x and y in $\mathbb{R}^p$ is defined by $\cos\theta = x^Ty\{\|x\|\|y\|\}^{-1/2}$ and this angle is said to be *obtuse* if $\cos\theta < 0$). Further, the hyperplane $H = \{x \in \mathbb{R}^p : x^Tp = \alpha\}$ is orthogonal to $(a - a^*)$, passes through a^*, and *separates* a and C; here separates means that the point a and the set C lie on the opposite sides of the hyperplane.

An important result on projections onto linear spaces is that, for any $x \in \mathbb{R}^p$ and linear space $\mathcal{L}$, we have a unique orthogonal decomposition of the form $x =$

$\Pi(\boldsymbol{x}|\mathcal{L}) + \Pi(\boldsymbol{x}|\mathcal{L}^{\perp})$ that corresponds to the orthogonal sum $\mathbb{R}^p = \mathcal{L} \oplus \mathcal{L}^{\perp}$. For a closed convex cone C, essentially the same results hold with the polar cone C^o playing the role of $\mathcal{L}^{\perp}$. These are given in the next proposition

Proposition 3.12.4 *Let C be a closed convex cone and $\boldsymbol{x} \in \mathbb{R}^p$.*

(a) Assume that $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ with $\boldsymbol{y} \in C$, $\boldsymbol{z} \in C^o$ and $\boldsymbol{y}^T\boldsymbol{z} = 0$. Then $\boldsymbol{y} = \Pi(\boldsymbol{x}|C)$ and $\boldsymbol{z} = \Pi(\boldsymbol{x}|C^o)$.

(b) Conversely, $\boldsymbol{x} = \Pi(\boldsymbol{x}|C) + \Pi(\boldsymbol{x}|C^o)$ and $\Pi(\boldsymbol{x}|C)^T\Pi(\boldsymbol{x}|C^o) = 0$.

(c) $\mathbb{R}^p = C \oplus C^0$.

(d) $C^{oo} = C$.

(For (a) and (b) see Hiriart-Urruty (1993), pp120-121 or SW 2.7.5.)

Proof of (a): Let $\boldsymbol{a}$ be an arbitrary point in C. Then, $\boldsymbol{a}^T\boldsymbol{z} \leq 0$, and $\|\boldsymbol{x} - \boldsymbol{a}\|^2 = \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \|\boldsymbol{y} - \boldsymbol{a}\|^2 - 2\boldsymbol{a}^T\boldsymbol{z} \geqslant \|\boldsymbol{x} - \boldsymbol{y}\|^2$. Therefore, $\boldsymbol{y}$ is the point in C closest to $\boldsymbol{x}$ and hence $\boldsymbol{y} = \Pi(\boldsymbol{x}|C)$. Similarly, let $\boldsymbol{b} \in C^o$. Then $\boldsymbol{b}^T\boldsymbol{y} \leq 0$, and $\|\boldsymbol{x} - \boldsymbol{b}\|^2 = \|\boldsymbol{x} - \boldsymbol{z}\|^2 + \|\boldsymbol{z} - \boldsymbol{b}\|^2 - 2\boldsymbol{b}^T\boldsymbol{y} \geqslant \|\boldsymbol{x} - \boldsymbol{z}\|^2$. Therefore, $\boldsymbol{z}$ is the point in C^o closest to $\boldsymbol{x}$ and hence $\boldsymbol{z} = \Pi(\boldsymbol{x}|C^o)$.

Proof of (b): Let $\boldsymbol{x}^*$ denote $\Pi(\boldsymbol{x}|C)$ and $\boldsymbol{y} = (1 + \delta)\boldsymbol{x}^*$ for $|\delta| < 1$. Then, by Proposition 3.12.3 (on page 114), $(\boldsymbol{x} - \boldsymbol{x}^*)^T(\boldsymbol{y} - \boldsymbol{x}^*) \leq 0$ for $|\delta| < 1$. Therefore, $\delta(\boldsymbol{x} - \boldsymbol{x}^*)^T\boldsymbol{x}^* \leq 0$ for $|\delta| < 1$. Since δ may be positive or negative, we have that $(\boldsymbol{x} - \boldsymbol{x}^*)^T\boldsymbol{x}^* = 0$. Let $\boldsymbol{u}$ be an arbitrary point in C and $\boldsymbol{y} = \boldsymbol{x}^* + \boldsymbol{u}$. Then, $\boldsymbol{y} \in C$ because C is a convex cone. Now, $0 \geqslant (\boldsymbol{x} - \boldsymbol{x}^*)^T(\boldsymbol{y} - \boldsymbol{x}^*) = (\boldsymbol{x} - \boldsymbol{x}^*)^T\boldsymbol{u}$. Since $\boldsymbol{u}$ is arbitrary, we have that $\boldsymbol{x} - \boldsymbol{x}^* \in C^o$. Thus, we have shown that $\boldsymbol{x} = \boldsymbol{x}^* + (\boldsymbol{x} - \boldsymbol{x}^*), \boldsymbol{x}^* \in C, (\boldsymbol{x} - \boldsymbol{x}^*) \in C^o$ and $(\boldsymbol{x} - \boldsymbol{x}^*)^T\boldsymbol{x}^* = 0$. Therefore, by part (a), $\boldsymbol{x} - \boldsymbol{x}^* = \Pi(\boldsymbol{x}|C^o)$, and hence $\boldsymbol{x} = \Pi(\boldsymbol{x}|C) + \Pi(\boldsymbol{x}|C^o)$ and $\Pi(\boldsymbol{x}|C)^T\Pi(\boldsymbol{x}|C^o) = 0$.

Proof of (c); Follows directly from parts (a) and (b).

Proof of (d): It follows from the definition of polar cones that $C \subset C^{oo}$. To prove the inclusion in the other direction, let $\boldsymbol{x} \in C^{oo}$ and let $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ be the orthogonal decomposition corresponding to $\mathbb{R}^p = C \oplus C^o$. Now, $\boldsymbol{z}^T\boldsymbol{z} = \boldsymbol{z}^T\boldsymbol{z} + \boldsymbol{y}^T\boldsymbol{z} = (\boldsymbol{y} + \boldsymbol{z})^T\boldsymbol{z} = \boldsymbol{x}^T\boldsymbol{z} \leq 0$, since $\boldsymbol{x} \in C^{oo}$ and $\boldsymbol{z} \in C^o$. It follows that $\boldsymbol{z}^T\boldsymbol{z} = 0$ and hence $\boldsymbol{x} = \boldsymbol{y} \in C$. Therefore, $C^{oo} \subset C$. ■

This result is illustrated in Figures 3.15 and 3.16; the first one shows the projections with respect to an arbitrary matrix $\boldsymbol{V}$ and the second shows those for the identity matrix. Let A be an arbitrary point, and $\boldsymbol{x}$ denote OA. Let $\boldsymbol{y} = OB$ and $\boldsymbol{z} = OC$ be the projections of OA onto C and C^o, respectively. Now, the foregoing proposition says that OC and OB are orthogonal, $OBAC$ is a rectangle (with respect to $\langle . \rangle_V$), and $OA = OB + OC$. Further, the orthogonal decomposition of OA into OB and OC where B and C lie C and C^o, respectively, is unique.

The following result appears in Rockafellar (1970) on page 146:

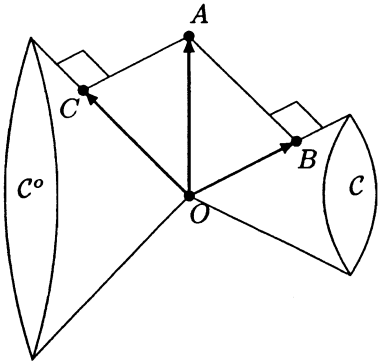

Fig. 3.15 The orthogonal projections of OA onto C and C^o when $\boldsymbol{V} \neq \boldsymbol{I}$

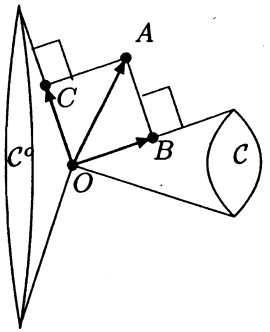

Fig. 3.16 The orthogonal projections of OA onto C and C^o when $\boldsymbol{V} = \boldsymbol{I}$.

Proposition 3.12.5 . *Let $C_1, \ldots, C_k$ be k convex cones in $\mathbb{R}^p$. Then*

$$(C_1 + \ldots + C_k)^o = C_1^o + \ldots + C_k^o,$$

$$\textit{and} \quad (C_1 \cap \ldots \cap C_k)^o = \textit{closure of } (C_1^o + \ldots + C_k^o). \qquad \blacksquare$$

Example: Let $X, P, Q \subset \mathbb{R}^p$, $P \subset Q^o$, and $X = P \oplus Q$. Show that for any $\boldsymbol{x} \in X$, there exist unique $\boldsymbol{p} \in P$ and $\boldsymbol{q} \in Q$ such that $\boldsymbol{x} = \boldsymbol{p} + \boldsymbol{q}$ and $\boldsymbol{p}^T\boldsymbol{q} = 0$; further, $\boldsymbol{p} = \Pi(\boldsymbol{x}|P)$ and $\boldsymbol{q} = \Pi(\boldsymbol{x}|Q)$.
Solution: Let $\boldsymbol{a} \in P$. Then

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{a}\|^2 &= \|\boldsymbol{p} + \boldsymbol{q} - \boldsymbol{a}\|^2 = \|\boldsymbol{p} - \boldsymbol{a}\|^2 + \|\boldsymbol{q}\|^2 + 2(\boldsymbol{p} - \boldsymbol{a})^T\boldsymbol{q} \\ &= \|\boldsymbol{p} - \boldsymbol{a}\|^2 + \|\boldsymbol{q}\|^2 - 2\boldsymbol{a}^T\boldsymbol{q}, \qquad \text{because } \boldsymbol{p}^T\boldsymbol{q} = 0 \\ &\geqslant \|\boldsymbol{p} - \boldsymbol{a}\|^2 + \|\boldsymbol{q}\|^2; \end{aligned} \tag{3.63}$$

the last step follows because $\boldsymbol{a} \in P$ and $\boldsymbol{q} \in Q \subset P^o$ and hence $\boldsymbol{a}^T\boldsymbol{q} \leqslant 0$. Therefore, $\|\boldsymbol{x} - \boldsymbol{a}\|^2$ reaches its minimum over $\boldsymbol{a} \in P$ when $\boldsymbol{a} = \boldsymbol{p}$. Hence, $\boldsymbol{p} = \Pi(\boldsymbol{x}|P)$. Similarly, $\boldsymbol{q} = \Pi(\boldsymbol{x}|Q)$. ■

Another important feature of projections onto linear spaces is that if $\mathcal{M}$ and $\mathcal{L}$ are linear subspaces and $\mathcal{M} \subset \mathcal{L}$ then projecting $\boldsymbol{x}$ onto $\mathcal{L}$ first and then onto $\mathcal{M}$ leads to the same point as projecting $\boldsymbol{x}$ directly onto $\mathcal{M}$; further, this two-stage projection leads to the orthogonal decomposition, $\Pi(\boldsymbol{x}|\mathcal{L}) = \Pi(\boldsymbol{x}|\mathcal{M}) + \Pi(\boldsymbol{x}|\mathcal{L} \cap \mathcal{M}^{\perp})$, and hence $\|\Pi(\boldsymbol{x}|\mathcal{L})\|^2 = \|\Pi(\boldsymbol{x}|\mathcal{M})\|^2 + \|\Pi(\boldsymbol{x}|\mathcal{L} \cap \mathcal{M}^{\perp})\|^2$. This plays an important role for statistical inference in normal theory linear models. For example, consider the likelihood ratio statistic for testing $\boldsymbol{\theta} \in \mathcal{M}$ against $\boldsymbol{\theta} \in \mathcal{L}$ based on $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$. This LRT statistic is equal to the LRT statistic for testing $\boldsymbol{\theta} = \boldsymbol{0}$ against $\boldsymbol{\theta} \in \mathcal{L} \cap \mathcal{M}^{\perp}$, and hence its null distribution is χ^2_q where $q = dim(\mathcal{L} \cap \mathcal{M}^{\perp}) = dim(\mathcal{L}) - dim(\mathcal{M})$. Similar results hold for projections onto closed convex cones, C_1 and C_2, as well provided that $C_1 \subset C_2$ and at least one of them is a linear space. These results are stated below. It is worth noting that if neither C_1 nor C_2 is a linear space then we

cannot obtain orthogonal decompositions in general; it will be seen that this is the root of the difficulties in obtaining least favorable null distributions when the null and alternative parameter spaces are C_1 and C_2 with neither of them being a linear space.

Proposition 3.12.6 *Let $\mathcal{L}$ and $\mathcal{M}$ be linear spaces, $\mathcal{C}$ be a closed convex cone, $\mathcal{M} \subset \mathcal{C} \subset \mathcal{L} \subset \mathbb{R}^p$, and $\boldsymbol{y} \in \mathbb{R}^p$. Then*

(a) $\Pi\{\Pi(\boldsymbol{y}|\mathcal{L})|\mathcal{C}\} = \Pi(\boldsymbol{y}|\mathcal{C})$.

(b) $\Pi\{\Pi(\boldsymbol{y}|\mathcal{C})|\mathcal{M}\} = \Pi(\boldsymbol{y}|\mathcal{M})$.

(c) $\mathcal{C} = \mathcal{M} \oplus (\mathcal{C} \cap \mathcal{M}^\perp)$.

(d-1) Given $\boldsymbol{x} \in \mathbb{R}^p$, there exist unique $\boldsymbol{m} \in \mathcal{M}, \boldsymbol{m}_o \in \mathcal{M} \cap \mathcal{C}^\perp$, and $\boldsymbol{c}_o \in \mathcal{C}^o$ such that $\boldsymbol{x} = \boldsymbol{m} + \boldsymbol{m}_o + \boldsymbol{c}_o$ and the components are pairwise orthogonal, namely $\boldsymbol{m}^T\boldsymbol{m}_o = \boldsymbol{m}^T\boldsymbol{c}_o = \boldsymbol{m}_o^T\boldsymbol{c}_o = 0$. Further, the components $\boldsymbol{m}, \boldsymbol{m}_o$ and $\boldsymbol{c}_0$ are the projections of $\boldsymbol{x}$ onto $\mathcal{M}, \mathcal{C} \cap M^\perp$ and $\mathcal{C}^o$, respectively.

(d-2) $\Pi\{\Pi(\boldsymbol{x} \mid \mathcal{C}) \mid \mathcal{C} \cap \mathcal{M}^\perp\} = \Pi(\boldsymbol{x} \mid \mathcal{C} \cap \mathcal{M}^\perp)$.

(d-3) For any $\boldsymbol{x} \in \mathcal{C}$, we have

$$\boldsymbol{x} = \Pi(\boldsymbol{x} \mid \mathcal{M}) + \Pi(\boldsymbol{x} \mid \mathcal{C} \cap \mathcal{M}^\perp) + \Pi(\boldsymbol{x} \mid \mathcal{C}^o)$$

and the three projections in this decomposition are pairwise orthogonal.

(d-4) $\mathbb{R}^p = \mathcal{M} \oplus (\mathcal{C} \cap \mathcal{M}^\perp) \oplus \mathcal{C}^o$

(d-5) $(\mathcal{C} \cap \mathcal{M}^\perp)^o = \mathcal{C}^o \oplus \mathcal{M}$.

(e) $\mathcal{L} = \mathcal{C} \oplus (\mathcal{L} \cap \mathcal{C}^o)$.

(f) $\mathbb{R}^p = \mathcal{C} \oplus (\mathcal{C}^o \cap \mathcal{L}) \oplus \mathcal{L}^\perp$.

(g) $\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{C}) \perp \Pi(\boldsymbol{y}|\mathcal{M}) - \Pi(\boldsymbol{y}|\mathcal{C})$.

(h) $\|\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{M})\|^2 - \|\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{C})\|^2 = \|\Pi(\boldsymbol{y}|\mathcal{M}) - \Pi(\boldsymbol{y}|\mathcal{C})\|^2$.

(i) $\Pi(\boldsymbol{y}|\mathcal{C}) - \Pi(\boldsymbol{y}|\mathcal{M}) = \Pi(\boldsymbol{y}|\mathcal{C} \cap \mathcal{M}^\perp)$.

Proof of (a): Let $\mathcal{L} = \{\boldsymbol{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^k\}$ where $\boldsymbol{X}$ is a $p \times k$ matrix of rank k, $\mathcal{K} \subset \mathbb{R}^k$ be a closed convex cone such that $\mathcal{C} = \boldsymbol{X}\mathcal{K}$, and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Then, by using the normal equation, we have that

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

and $\Pi(\boldsymbol{X}|\mathcal{L}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Therefore, the value of $\boldsymbol{\beta} \in \mathcal{K}$, say $\tilde{\boldsymbol{\beta}}$, at which $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ is a minimum is also the same as that at which $\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ is a minimum. Since $\Pi(\boldsymbol{y}|\mathcal{L}) = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \Pi(\boldsymbol{X}\hat{\boldsymbol{\beta}}|\mathcal{C}) = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ and $\Pi(\boldsymbol{y}|\mathcal{C}) = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$, the proof follows. One could also show this directly using $\|\boldsymbol{y} - \boldsymbol{x}\|^2 = \|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\|^2 + \|\boldsymbol{P}\boldsymbol{y} - \boldsymbol{x}\|^2$ where $\boldsymbol{P}$ is the projection matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ onto $\mathcal{L}$, $\boldsymbol{x} \in \mathcal{L}$, and $\boldsymbol{y} \in \mathbb{R}^p$.

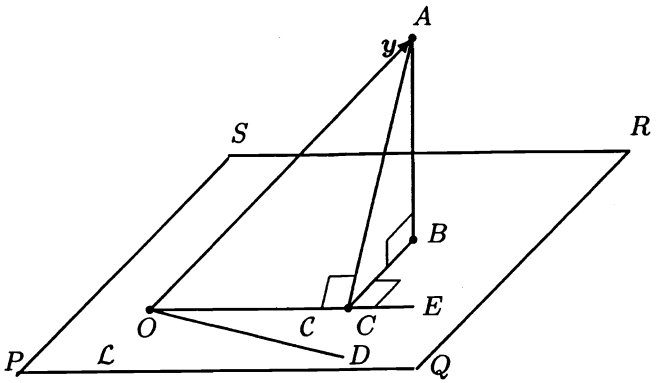

Fig. 3.17 Projections of OA onto the cone, $\mathcal{C}$, and the linear space, $\mathcal{L}$, where $\mathcal{C} \subset \mathcal{L}$.

Fig. 3.17 illustrates this result; $\mathcal{L}$ is the two-dimensional plane containing $PQRS$; O is the origin; $\mathcal{C}$ is the cone EOD contained in $\mathcal{L}$; B is the projection of A onto $\mathcal{L}$; and C is the projection of B onto $\mathcal{C}$. Then C is also the projection of A onto $\mathcal{C}$.
Proof of (b) (Raubertas et al. (1986), p 2815) : Since $\Pi(\boldsymbol{y}|\mathcal{C}) = \boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{C}^o)$, we have that $\Pi\{\Pi(\boldsymbol{y}|\mathcal{C})|\mathcal{M}\} = \Pi\{\boldsymbol{y} - \Pi(\boldsymbol{y}|\mathcal{C}^o)|\mathcal{M}\} = \Pi(\boldsymbol{y}|\mathcal{M}) - \Pi\{\Pi(\boldsymbol{y}|\mathcal{C}^o)|\mathcal{M}\}$; the last equality follows since any projection onto any linear space is a linear function. Since $\mathcal{M} \subset \mathcal{C}$, we have that $\mathcal{C}^o \subset \mathcal{M}^\perp$. Therefore, $\Pi\{\Pi(\boldsymbol{y}|\mathcal{C}^o)|\mathcal{M}\} = \mathbf{0}$, and we conclude that $\Pi\{\Pi(\boldsymbol{y}|\mathcal{C})|\mathcal{M}\} = \Pi(\boldsymbol{y}|\mathcal{M})$ (see Fig. 3.10, page 85).

Proof of (c): Let $\mathcal{C}^* = \mathcal{M}^\perp \cap \mathcal{C}$. First let us show that $\mathcal{C} = \{\boldsymbol{y}+\boldsymbol{z} : \boldsymbol{y} \in \mathcal{M}, \boldsymbol{z} \in \mathcal{C}^*\}$. Recall that $\mathbb{R}^p = \mathcal{M} \oplus \mathcal{M}^\perp$. Given $\boldsymbol{x} \in \mathcal{C}$, we can write $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$, where $\boldsymbol{y} \in \mathcal{M}, \boldsymbol{z} \in \mathcal{M}^\perp$ and $\boldsymbol{y}^T\boldsymbol{z} = 0$. Since $-\boldsymbol{y} \in \mathcal{M} \subset \mathcal{C}$ and $\mathcal{C}$ is a closed convex cone, it follows that $\boldsymbol{z} = \boldsymbol{x} + (-\boldsymbol{y}) \in \mathcal{C}$. Therefore, $\boldsymbol{z} \in \mathcal{C}^*$ and hence $\mathcal{C} \subset \mathcal{M} + \mathcal{C}^*$. Since $\mathcal{M} \subset \mathcal{C}$, $\mathcal{C}^* = \mathcal{M}^\perp \cap \mathcal{C} \subset \mathcal{C}$ and $\mathcal{C}$ is a closed convex cone it follows that $\mathcal{M} + \mathcal{C}^* \subset \mathcal{C}$. Therefore, $\mathcal{C} = \mathcal{M} + \mathcal{C}^*$.

The claim that $\mathcal{M}+\mathcal{C}^*$ is an orthogonal sum follows directly from $\mathbb{R}^p = \mathcal{M} \oplus \mathcal{M}^\perp$. Let $\boldsymbol{x} \in \mathcal{C}$ and let us start with the decomposition, $\boldsymbol{x} = \boldsymbol{y}+\boldsymbol{z}$ in the previous paragraph where $\boldsymbol{y} \in \mathcal{M}, \boldsymbol{z} \in \mathcal{M}^\perp$ and $\boldsymbol{y}^T\boldsymbol{z} = 0$; the pair $\{\boldsymbol{y}, \boldsymbol{z}\}$ is uniquely defined. Since we just proved that $\boldsymbol{z}$ also belongs to $\mathcal{C} \cap \mathcal{M}^\perp$, we have shown the following: $\mathcal{C} = \mathcal{M} + \mathcal{C}^*$ and given $\boldsymbol{x} \in \mathcal{C}$ there exist unique $\boldsymbol{y} \in \mathcal{M}$ and $\boldsymbol{z} \in \mathcal{C}^*$ such that $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ and $\boldsymbol{y}^T\boldsymbol{z} = 0$. Therefore, $\mathcal{C} = \mathcal{M} \oplus \mathcal{C}^*$.
Remark: Note that $\boldsymbol{z}$ is the point in $\mathcal{M}^\perp$ closest to $\boldsymbol{x}$, and further since $\boldsymbol{z} \in \mathcal{C}^* \subset \mathcal{M}^\perp$ it is also the point in $\mathcal{C}^*$ closest to $\boldsymbol{x}$ as well. Therefore, $\boldsymbol{z} = \Pi(\boldsymbol{x} \mid \mathcal{C}^*)$ and $\boldsymbol{y} = \Pi(\boldsymbol{x}|\mathcal{M})$.

Proof of (d-1): Let $\boldsymbol{x} \in \mathbb{R}^p$. Since $\mathbb{R}^p = \mathcal{C} \oplus \mathcal{C}^o$, there exist unique $\boldsymbol{c} \in \mathcal{C}$ and $\boldsymbol{c}_0 \in \mathcal{C}^o$ such that $\boldsymbol{x} = \boldsymbol{c} + \boldsymbol{c}_o$ and $\boldsymbol{c}^T\boldsymbol{c}_o = 0$. We also have $\boldsymbol{c} = \Pi(\boldsymbol{x}|\mathcal{C})$ and $\boldsymbol{c}_0 = \Pi(\boldsymbol{x}|\mathcal{C}^o)$. Since $\mathcal{C} = \mathcal{M} \oplus (\mathcal{C} \cap \mathcal{M}^\perp)$, there exist unique $\boldsymbol{m} \in \mathcal{M}$ and $\boldsymbol{m}_0 \in \mathcal{C} \cap \mathcal{M}^\perp$ such that $\boldsymbol{c} = \boldsymbol{m} + \boldsymbol{m}_0$ and $\boldsymbol{m}^T\boldsymbol{m}_0 = 0$. We also have

$$\boldsymbol{m} = \Pi(\boldsymbol{c} \mid \mathcal{M}) = \Pi\{\Pi(\boldsymbol{x} \mid \mathcal{C}) \mid \mathcal{M}\} = \Pi(\boldsymbol{x} \mid \mathcal{M}) \qquad \text{by part (b)}, \qquad (3.64)$$

$$m_0 = \Pi(c \mid \mathcal{C} \cap \mathcal{M}^\perp) = \Pi\{\Pi(x \mid \mathcal{C}) \mid \mathcal{C} \cap \mathcal{M}^\perp\}. \tag{3.65}$$

Since $\mathcal{M} \subset \mathcal{C}$, we have that $\mathcal{C}^o \subset \mathcal{M}^\perp$ and hence $c_o^T m = 0$. Now

$$0 = c_o^T c = c_0^T(m + m_o) = c_o^T m + c_o^T m_o = c_o^T m_o.$$

Therefore, given $x \in \mathbb{R}^p$, we have shown that there exist $m \in \mathcal{M}, m_o \in \mathcal{M} \cap \mathcal{C}^\perp$, and $c_o \in \mathcal{C}^o$ such that $x = m + m_o + c_o$ and the three components are pairwise orthogonal, (i.e., $m^T m_o = m^T c_o = m_o^T c_o = 0$).

To prove uniqueness of the decomposition let $x \in \mathbb{R}^p$. Suppose that $x = m' + m'_o + c'_o$ where $m' \in \mathcal{M}, m'_o \in \mathcal{C} \cap \mathcal{M}^\perp, c'_o \in \mathcal{C}^o$, and $\{m', m'_o, c'_o\}$ are pairwise orthogonal. Then, since $m' \in \mathcal{C}$ and $m'_o \in \mathcal{C}$ it follows that $m' + m'_o \in \mathcal{C}$. It also follows from the pairwise orthogonality of the components that $(m' + m'_o)^T c'_o = 0$. Since $\mathbb{R}^p = \mathcal{C} \oplus \mathcal{C}^o$ it follows from the orthogonal decomposition $x = (m' + m'_0) + c'_0$ that $c'_o = \Pi(x \mid \mathcal{C}^o) = c_0$ and $(m' + m'_o) = \Pi(x \mid \mathcal{C})$. Since $\mathcal{C} = \mathcal{M} \oplus (\mathcal{C} \cap \mathcal{M}^\perp)$, $m' + m'_o \in \mathcal{C}, m' \in \mathcal{M}, m'_o \in \mathcal{C} \cap \mathcal{M}^\perp$, and $m'^T m'_o = 0$ it follows that $m' = \Pi\{m' + m'_o \mid \mathcal{M}) = \Pi\{\Pi(x \mid \mathcal{C}) \mid \mathcal{M}\} = \Pi(x \mid \mathcal{M})$ and $m'_o = \Pi\{m' + m'_o \mid \mathcal{C} \cap \mathcal{M}^\perp) = \Pi\{\Pi(x \mid \mathcal{C}) \mid \mathcal{C} \cap \mathcal{M}^\perp\} = m_0$. Therefore, the decomposition $x = m + m_o + c_o$ is unique.

We already noted that $c_o = \Pi(x \mid \mathcal{C}^o)$ and $m = \Pi\{\Pi(x \mid \mathcal{C}) \mid \mathcal{M}\} = \Pi(x \mid \mathcal{M})$. Therefore, to complete the proof of (d-1) it suffices to prove that $m_o = \Pi(x \mid \mathcal{C} \cap \mathcal{M}^\perp)$. To this end, let $b \in \mathcal{C} \cap \mathcal{M}^\perp$. Now,

$$\begin{aligned}
(m + c_o)^T(m_o - b) &= m^T(m_0 - b) + c_o^T(m_0 - b) \\
&= 0 + c_o^T(m_0 - b), \qquad \text{because } m_o - b \in \mathcal{M}^\perp \text{ and } m \in \mathcal{M} \\
&= c_o^T m - c_o^T b \\
&= 0 - c_o^T b \qquad \text{because } c_o \in \mathcal{C}^o \subset \mathcal{M}^\perp \text{ and } m \in \mathcal{M}
\end{aligned}$$

The last term is nonnegative because $c_o \in \mathcal{C}^o$ and $b \in \mathcal{C}$. Therefore, $(m+c_o)^T(m_o - b)$ is nonnegative and reaches its minimum of zero when $b = m_0$. Now,

$$\begin{aligned}
\|x - b\|^2 &= \|m + m_o + c_o - b\|^2 \\
&= \|m + c_o\|^2 + \|m_o - b\|^2 + 2(m + c_o)^T(m_o - b). \\
&\geqslant \|m + c_o\|^2 + \|m - b\|^2 \qquad \forall b \in \mathcal{C} \cap \mathcal{M}^\perp,
\end{aligned}$$

equality holding when $b = m_0$. Hence, $\|x - b\|^2$ reaches its minimum over $b \in \mathcal{C} \cap \mathcal{M}^\perp$ when $b = m_o$. Therefore, $\Pi(x \mid \mathcal{C} \cap \mathcal{M}^\perp) = m_o = \Pi\{\Pi(x \mid \mathcal{C}) \mid \mathcal{C} \cap \mathcal{M}^\perp\}$.

Proofs of part (d-2), (d-3), and (d-4): Contained in the proof of (d-1).

Proof of part (d-5): Since $\mathcal{C} \cap \mathcal{M}^\perp \subset \mathcal{C}$ we have $\mathcal{C}^o \subset (\mathcal{C} \cap \mathcal{M}^\perp)^o$. Since $\mathcal{C} \cap \mathcal{M}^\perp \subset \mathcal{M}^\perp$ it follows that $\mathcal{M} = (\mathcal{M}^\perp)^\perp = (\mathcal{M}^\perp)^o \subset (\mathcal{C} \cap \mathcal{M}^\perp)^o$. Since $(\mathcal{C} \cap \mathcal{M}^\perp)^o$ is a closed convex cone, it is closed under addition, and hence $\mathcal{C}^o + \mathcal{M} \subset (\mathcal{C} \cap \mathcal{M}^\perp)^o$.

To prove the set containment in the reverse order, first note that $\mathcal{C}^o \subset \mathcal{C}^o + \mathcal{M}$, and hence $(\mathcal{C}^o + \mathcal{M})^o \subset \mathcal{C}$. Since $\mathcal{M} \subset \mathcal{C}^o + \mathcal{M}$, it follows that $(\mathcal{C}^o + \mathcal{M})^o \subset \mathcal{M}^\perp$. Therefore, $(\mathcal{C}^o + \mathcal{M})^o \subset \mathcal{C} \cap \mathcal{M}^\perp$ and $(\mathcal{C} \cap \mathcal{M}^\perp)^o \subset \mathcal{C}^o + \mathcal{M}$.

Therefore, we have proved $(\mathcal{C} \cap \mathcal{M}^\perp)^o = \mathcal{C}^o + \mathcal{M}$. The proof that this is an orthogonal sum follows from part (d-4).

Proof of (e): The proof is similar to that for part (c) and is based on $\mathbb{R}^p = \mathcal{C} \oplus \mathcal{C}^o$.

Proof of (f): Given $y \in \mathbb{R}^p$ we have the unique orthogonal decomposition, $y = \Pi(y|\mathcal{L}) + \Pi(y|\mathcal{L}^\perp)$. Since $\mathcal{L} = \mathcal{C} \oplus (\mathcal{C}^o \cap \mathcal{L})$, we also have the unique orthogonal decomposition

$$\begin{aligned}
\Pi(y|\mathcal{L}) &= \Pi\{\Pi(y|\mathcal{L})|\mathcal{C})\} + \Pi\{\Pi(y|\mathcal{L})|\mathcal{C}^o \cap \mathcal{L})\} \\
&= \Pi(y|\mathcal{C}) + \Pi(y|\mathcal{C}^o \cap \mathcal{L}) \qquad \text{by part (a)}.
\end{aligned} \tag{3.66}$$

Therefore, we have the orthogonal decomposition,

$$y = \Pi(y|\mathcal{C}) + \Pi(y|\mathcal{C}^o \cap \mathcal{L}) + \Pi(y|\mathcal{L}^\perp).$$

Uniqueness of the decomposition can also be established as for part (d-1). Hence, $\mathbb{R}^p = \mathcal{C} \oplus (\mathcal{C}^o \cap \mathcal{L}) \oplus \mathcal{L}^\perp$.

Proofs of (g) and (h): It is easily verified that $\mathcal{M}^\perp = \mathcal{M}^o$, and if $\mathcal{C}_1 \subset \mathcal{C}_2$ then $\mathcal{C}_2^o \subset \mathcal{C}_1^o$. Since $\mathcal{M} \subset \mathcal{C}$ we have that $\mathcal{C}^o \subset \mathcal{M}^\perp$ and $\Pi(y|\mathcal{C}^o)^T \Pi(y|\mathcal{M}) = 0$. Note that $\Pi(y|\mathcal{C}^o) \in \mathcal{C}^o \subset \mathcal{M}^\perp$. Therefore, $\Pi(y|\mathcal{C}^o) \perp \mathcal{M}$. Now,

$$\begin{aligned}
&\{y - \Pi(y|\mathcal{C})\}^T\{\Pi(y|\mathcal{C}) - \Pi(y|\mathcal{M})\} \\
&\quad = \Pi(y|\mathcal{C}^o)^T \Pi(y|\mathcal{C}) - \Pi(y|\mathcal{C}^o)^T \Pi(y|\mathcal{M}) = \mathbf{0}.
\end{aligned} \tag{3.67}$$

Therefore, $\|y - \Pi(y|\mathcal{M})\|^2 = \|y - \Pi(y|\mathcal{C})\|^2 + \|\Pi(y|\mathcal{C}) - \Pi(y|\mathcal{M})\|^2$.

Proof of (i):

$$\begin{aligned}
\Pi(x \mid \mathcal{C}) &= \Pi\{\Pi(x \mid \mathcal{C}) \mid \mathcal{M}\} + \Pi\{\Pi(x \mid \mathcal{C}) \mid \mathcal{C} \cap \mathcal{M}^\perp\} \\
&= \Pi(x \mid \mathcal{M}) + \Pi(x \mid \mathcal{C} \cap \mathcal{M}^\perp) \qquad \text{by parts (b) and (c)}
\end{aligned} \tag{3.68}$$

The Fig. 3.17 illustrates the main ideas of this proof in three-dimensions. ■

A consequence of the orthogonal sum $\mathcal{C} = \mathcal{M} \oplus (\mathcal{C} \cap \mathcal{M}^\perp)$ in part(c) of the foregoing proposition is that any $y \in \mathcal{C}$ has a unique decomposition of the form $y = a + b$, where $a \in \mathcal{M}, b \in \mathcal{C} \cap \mathcal{M}^\perp, a^T b = 0$; further for such a decomposition, $a = \Pi(y \mid \mathcal{M})$, and $b = \Pi(y \mid \mathcal{C} \cap \mathcal{M}^\perp)$. Similarly, a consequence of the orthogonal sum $\mathbb{R}^p = \mathcal{M} \oplus (\mathcal{C} \cap \mathcal{M}^\perp) \oplus \mathcal{C}^o$ is that any $y \in \mathbb{R}^p$ has a unique decomposition of the form $y = a + b + c$, where $a \in \mathcal{M}, b \in \mathcal{C} \cap \mathcal{M}^\perp, c \in \mathcal{C}^o$, $a^T b = 0, a^T c = 0, b^T c = 0$; and, further, for such a decomposition we have that $a = \Pi(y \mid \mathcal{M})$, $b = \Pi(y \mid \mathcal{C} \cap \mathcal{M}^\perp)$ and $c = \Pi(y \mid \mathcal{C}^o)$.

In contrast to the decompositions in linear subspaces, the decomposition $x = x_1 + x_2$, with $x_1 \in \mathcal{C}$ and $x_2 \in \mathcal{C}^o$ is not unique when $\mathcal{C}$ is a closed convex cone unless we impose the requirement $x_1^T x_2 = 0$; however, the decomposition is optimal in the sense stated below (Hiriart-Urruty and Lemaréchal (1993), pp 118, 121).

Proposition 3.12.7 *Let C be a closed convex cone $x \in \mathbb{R}^p$. Then we have the following:*
(a) $x = x_1 + x_2, x_1 \in C, x_2 \in C^o \Rightarrow \|x_1\| \geqslant \|\Pi(x|C)\|$ *and* $\|x_2\| \geqslant \|\Pi(x|C^o)\|$.
(b) $\|\Pi(x_1|C) - \Pi(x_2|C)\|^2 \leq \{\Pi(x_1|C) - \Pi(x_2|C)\}^T(x_1 - x_2)$, *for* $x_1, x_2 \in \mathbb{R}^p$.
(c) $\|\Pi(x_1|C) - \Pi(x_2|C)\| \leq \|(x_1 - x_2)\|$, *for* $x_1, x_2 \in \mathbb{R}^p$. ■

3.12.3 Polyhedral Cones

Let $a_1, \ldots, a_q$ be q points in $\mathbb{R}^p$ and $\mathcal{P} = \{x \in \mathbb{R}^p : a_i^T x \geqslant 0 \text{ for } i = 1, \ldots, q\}$. Then $\mathcal{P}$ is a closed convex cone and it is called a *polyhedral cone.* Note that $\mathcal{P}$ is the intersection of the half-spaces, $\{x : a_1^T x \geqslant 0\}, \ldots,$ and $\{x : a_q^T x \geqslant 0\}$. With the $p \times q$ matrix A defined as $[a_1, \ldots, a_q]$, we may express $\mathcal{P}$ as $\{x \in \mathbb{R}^p : A^T x \geqslant 0\}$. Because the equality constraint $a^T x = 0$ is equivalent to the two inequality constraints $a^T x \geqslant 0$ and $(-a)^T x \geqslant 0$, the set of constraints $A^T x \geqslant 0$ may contain equality constraints as well. We will say that the set of constraints $A^T x \geqslant 0$ defining $\mathcal{P}$ is *tight* if $\mathcal{P}$ cannot be defined using a submatrix of A with fewer columns than A; in other words, A has no redundant columns, and hence no redundant constraints. Without loss of generality, we shall always assume that the set of constraints defining $\mathcal{P}$ is tight.

Let S be a subset of $\mathbb{R}^p$. Then the *cone generated* by S consists of elements of the form $w_1 b_1 + \ldots + w_m b_m$ where $\{b_1, \ldots, b_m\} \subset S$, $w_i \geqslant 0$, for $i = 1, \ldots, m$; it is also called the *conical hull* or *positive hull* of S. The elements of S are called *generators* of the cone generated by S. The conical hull is the smallest cone that contains S. The set consisting of elements of the form $w_1 b_1 + \ldots + w_m b_m$ where $\{b_1, \ldots, b_m\} \subset S$, $\sum w_i = 1$, and $w_i \geqslant 0$, for $i = 1, \ldots, m$, is called the *convex hull* of S. The convex hull is the smallest convex set containing S.

An explicit formula for the relationship between a polyhedral cone and its polar cone is given in the next result.

Proposition 3.12.8 *Let $a_1, \ldots, a_q$ be q points in $\mathbb{R}^p$ and let $C = \{w_1 a_1 + \ldots + w_q a_q : w_i \geqslant 0, i = 1, \ldots, q\}$, the cone generated by $\{a_1, \ldots, a_q\}$. Then $C^o = \{x \in \mathbb{R}^p : a_i^T x \leq 0$ for $i = 1, \ldots, q\}$. Consequently, if $\mathcal{P} = \{x : a_i^T x \geqslant 0$, for $i = 1, \ldots, q\}$ then $\mathcal{P}^o = \{w_1 a_1 + \ldots + w_q a_q : w_i \leq 0, i = 1, \ldots, q\}$.*
[For example see, Hiriart-Urutty and Lemaréchal (1993), p 119].

Proof: Let $x \in C^o$. Then $x^T(w_1 a_1 + \ldots + w_q a_q) \leq 0$ for $(w_1, \ldots, w_q) \geqslant 0$. Now, by choosing $w_i > 0$ and the $w_j = 0$ for $j \neq i$, we have that $a_i^T x \leq 0$ for $i = 1, \ldots, q$. Therefore, $x \in \{x : a_i^T x \leq 0, \text{ for } i = 1, \ldots, q\}$. To prove the set containment in the opposite direction, let $x \in \{x : a_i^T x \leq 0, \text{ for } i = 1, \ldots, q\}$ and $z \in C$. Then $z = (w_1 a_1 + \ldots + w_q a_q)$ for some $w_i \geqslant 0$ for $i = 1, \ldots, q$. Now, $x^T z = w_1 a_1^T x + \ldots + w_q a_q^T x \leq 0$. Therefore, $x \in C^o$. The second part of the proposition follows from the first by substituting $\mathcal{P} = -C^o$ and noting that $\mathcal{P}^o = -C^{oo} = -C$. ■

The following restatement of the foregoing proposition in matrix notation is useful.

Corollary 3.12.9 *Let $a_1, \ldots, a_q$ be q points in $\mathbb{R}^p$ and let $A = [a_1, \ldots, a_q]$.*

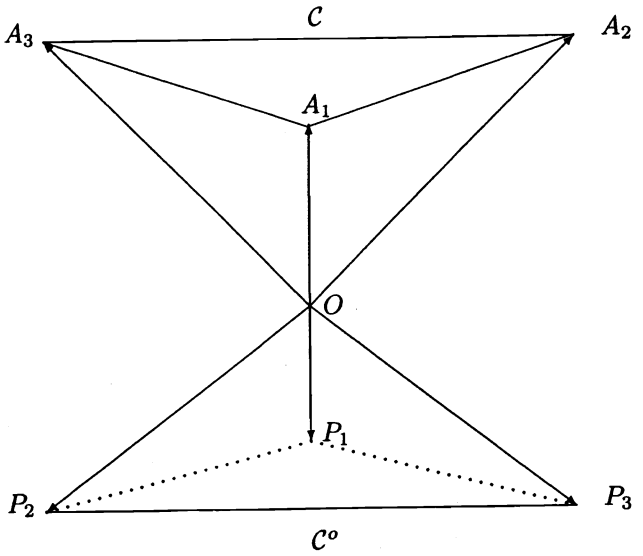

Fig. 3.18 A polyhedral, C, and its polar cone C^o

1. *If* $\mathcal{P} = \{x : A^T x \geqslant 0\}$ *then* $\mathcal{P}^o = \{Ay : y \leq 0\}$.
2. *If* $C = \{Ax : x \geqslant 0\}$ *then* $C^o = \{y : A^T y \leq 0\}$. ■

Fig. 3.18 provides a representation of this relationship between a polyhedral and its polar in three-dimensions. Let OA_1, OA_2, and OA_3 correspond to a_1, a_2, and a_3, respectively, and $C = \{Ax : x \geqslant 0\}$. Let $OA_1 \perp OP_2P_3$, $OA_2 \perp OP_1P_3$, and $OA_3 \perp OP_1P_2$. Then $OP_1 \perp OA_2A_3$, $OP_2 \perp OA_1A_3$, and $OP_3 \perp OA_1A_2$. The polyhedrals $OA_1A_2A_3$ and $OP_1P_2P_3$ are C and C^o, respectively. Note that every boundary plane of C corresponds to an edge of C^o, the two being orthogonal to each other; similarly, every edge of C corresponds to a boundary plane of C^o, again the two being orthogonal to each other. This correspondence defines the function Φ in Proposition 3.12.11.

Now we state two useful results on polyhedral cones; for example see SW 2.8.6 and 2.8.8.

Proposition 3.12.10 *(a) Minkowski's theorem: For every polyhedral cone $\mathcal{P}$ there exist $a_1, \ldots, a_q$ in $\mathbb{R}^p$ such that $\mathcal{P} = \{w_1 a_1 + \ldots + w_q a_q : w_i \geqslant 0 \, i = 1, \ldots, q\}$.*
(b) Weyl's theorem: Every closed convex cone with finitely many generators is a polyhedral cone. ■

Let $A = (a_1, \ldots, a_q)$ be a $p \times q$ matrix, and $\mathcal{P} = \{x : A^T x \geqslant 0\}$. Let $J = \{1, \ldots, m\}$ be a subset of $\{1, \ldots, q\}$; J may be empty. Let $I = \{1, \ldots, q\} \setminus J$. Let A_J and A_I denote the matrices with their columns being $\{a_j : j \in J\}$ and $\{a_i : i \in I\}$ respectively. The set $F_J = \{x : A_J^T x = 0, A_I^T x \geqslant 0\}$ is called a *face* of $\mathcal{P}$; without loss of generality, it is assumed that the set of constraints defining F_J

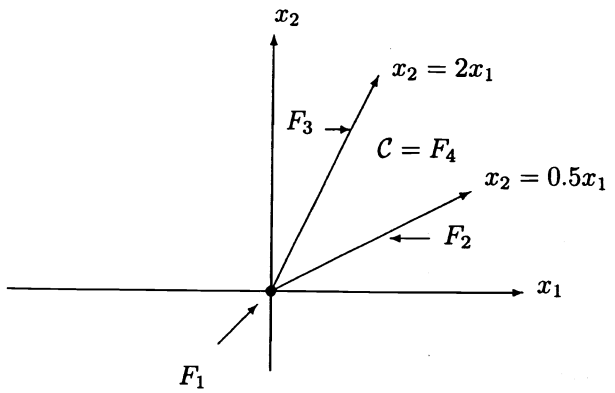

Fig. 3.19 The faces of the polyhedral $\{x : 2x_1 - \frac{1}{2}x_2 \geqslant 0, x_2 - \frac{1}{2}x_1 \geqslant 0\}$.

is tight. Now, the *relative interior* of F_J, denoted, $ri(F_J)$, is defined as the interior of F_J with respect to the relative topology induced in the linear space spanned by F_J; it may be verified that the relative interior of a face is obtained by replacing the inequality constraints defining F_J by their corresponding strict inequalities; thus, $ri(F_J) = \{x : A_J^T x = 0, A_I^T x > 0\}$. The next example illustrates these definitions.

Example: Let $C = \{x : 2x_1 - x_2 \geqslant 0, x_2 - \frac{1}{2}x_1 \geqslant 0\}$.

Then C is a polyhedral cone and its faces are (see Fig. 3.19)

$$\begin{aligned} F_1 = &\quad \{x : 2x_1 - x_2 = 0, x_2 - 2^{-1}x_1 = 0\}, \\ F_2 = &\quad \{x : 2x_1 - x_2 \geqslant 0, x_2 - 2^{-1}x_1 = 0\}, \\ F_3 = &\quad \{x : 2x_1 - x_2 = 0, x_2 - 2^{-1}x_1 \geqslant 0\}, \\ F_4 = &\quad \{x : 2x_1 - x_2 \geqslant 0, x_2 - 2^{-1}x_1 \geqslant 0\}, \end{aligned}$$

Descriptions of the faces of C and their relative interiors are:

$$\begin{aligned} F_1 &= \text{the vertex}, & ri(F_1) &= \text{the vertex}, \\ F_2 &= \text{the lower boundary of } C & ri(F_2) &= F_2 \text{ excluding the vertex} \\ F_3 &= \text{the upper boundary of } C & ri(F_3) &= F_3 \text{ excluding the vertex} \\ F_4 &= C & ri(F_4) &= C \text{ excluding its boundary, } F_2 \cup F_3. \blacksquare \end{aligned}$$

Note that the polyhedral $\mathcal{P}$ is the union of sets of the form $ri(F_J)$, each of which is an open set in a linear subspace, and hence does not include its boundary. A main feature of statistical inference problems under inequality constraints is that the parameter space includes the boundary. An advantage of $ri(F_J)$ is that the parameter space $\mathcal{P}$ that includes its boundary is partitioned into sets of the form $ri(F_J)$, each of which is open with respect to the linear space spanned by $ri(F_J)$ and has a linear structure. Consequently, we can apply well-known results for open parameter spaces to a polyhedral $\mathcal{P}$ as well by considering each member of the partition of $\mathcal{P}$ separately.

There is an important relationship between the faces of a polyhedral cone and those of its polar cone. This was useful for deriving the chi-bar-square distribution for a general closed convex cone; however, these relationships are not required to understand the proofs in the form they are presented in this monograph. For completeness, we state these results below; for example, see SW 2.13.2, 2.13.3.

Proposition 3.12.11 *Let $\mathcal{F}(\mathcal{P}) = \{F : F \text{ is a face of } \mathcal{P} \text{ and } F \text{ is nonempty}\}$; it is the collection of faces of $\mathcal{P}$. Then*

1. $F^\perp \cap \mathcal{P}^o$ is a face of $\mathcal{P}^o$.

2. For any given face F of $\mathcal{P}$, let Φ denote the operation of choosing $F^\perp \cap \mathcal{P}^o$. Then $\mathcal{F}(\mathcal{P}) \xrightarrow{\Phi} \mathcal{F}(\mathcal{P}^o) \xrightarrow{\Phi} \mathcal{F}(\mathcal{P}^{oo}) = \mathcal{F}(\mathcal{P})$ and $\Phi\Phi(F) = F$.

3. $< F^\perp \cap \mathcal{P}^o >=< F >^\perp$ where $< A >$ denotes the linear space spanned by A. Further, every face of $\mathcal{P}$ is uniquely associated with the face $F^\perp \cap \mathcal{P}^o$ of $\mathcal{P}^o$, and any face of $\mathcal{P}^o$ is of the form $F^\perp \cap \mathcal{P}^o$ for some face F of $\mathcal{P}$. ■

3.13 APPENDIX 2: PROOFS

Proof of Theorem 3.4.2: To prove this theorem it is convenient to establish several lemmas that are of independent interest.

Lemma 3.13.1 .

1. *Suppose that $Z \sim N(0, I)$ and let $\mathcal{P} \subset \mathbb{R}^p$ be a cone. Then the direction and length of Z are independent. Consequently, $\|Z\|$ and $Z/\|Z\|$ are independent random variables and $\{\|Z\| \geqslant c\}$ and $\{Z \in \mathcal{P}\}$ are independent events.*

2. *More generally, suppose that the distribution of $\boldsymbol{X}$ is orthogonally invariant (i.e., $\boldsymbol{X}$ and $\boldsymbol{PX}$ have the same distribution for any orthogonal matrix $\boldsymbol{P}$). Then $\{\|Z\| \geqslant c\}$ and $\{Z \in \mathcal{P}\}$ are independent events and $Z/\|Z\|$ is distributed uniformly on the unit sphere.*

Proof: First, let us transform Z into polar coordinates: $Z_1 = R\sin\theta_1$, $Z_2 = R\cos\theta_1 \sin\theta_2, Z_3 = R\cos\theta_1\cos\theta_2\sin\theta_3, \ldots, Z_{p-1} = R\cos\theta_1\cos\theta_2\ldots\cos\theta_{p-2}\sin\theta_{p-1}$, $Z_p = R\cos\theta_1\cos\theta_2\ldots\cos\theta_{p-2}\cos\theta_{p-1}$, where $-(\pi/2) < \theta_i \leqslant (\pi/2)$ for $i = 1, \ldots, p-2$ and $-\pi < \theta_{p-1} \leqslant \pi$. Then $R^2 = \|Z\|^2$ and the Jacobian of the transformation is $R^{p-1}\cos^{p-2}\theta_1\cos^{p-3}\theta_2\ldots\cos\theta_{p-2}$ (see Anderson (1984), page 280). Now, since the density of Z at z is a function of $\|z\|$ only, it follows that the density of $(R, \theta_1, \ldots, \theta_{p-1})$ factors into two components, one is a function of R only and the other is a function of $(\theta_1, \ldots, \theta_{p-1})$. Now the proof follows because $\{Z/\|Z\|$ depends on $(\theta_1, \ldots, \theta_{p-1})$ only. A similar result also appeared in Kudo (1963).

The second result is stated in Perlman (1969) and is essentially the same as the first. ■

Lemma 3.13.2 *Let $\mathcal{P} \subset \mathbb{R}^p$ be a polyhedral cone, $y \in \mathbb{R}^p$, F be a face of $\mathcal{P}$, $\mathcal{L}$ denote the linear space spanned by F, and $\boldsymbol{P}$ denote the projection matrix onto $\mathcal{L}$. Let the statements (A), (B) and (C) be defined as follows:*

(A) $\Pi(\boldsymbol{y}|\mathcal{P}) \in ri(F)$.

(B) $\boldsymbol{Py} \in ri(F)$ *and* $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y} \in F^{\perp} \cap \mathcal{P}^o$.

(C) $\boldsymbol{Py} = \Pi(\boldsymbol{y}|\mathcal{P})$ *and* $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y} = \Pi(\boldsymbol{y}|\mathcal{P}^o)$.

Then, (A) is true if and only if (B) is true. If (B) is true then (C) is true.

Proof: Let us recall that $\mathbb{R}^p = \mathcal{L} \oplus \mathcal{L}^{\perp}$ where $\oplus$ is an orthogonal sum. Consequently, $\boldsymbol{y} \in \mathbb{R}^p$ has a unique orthogonal decomposition of the form $\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{b}$, where $\boldsymbol{a} \in \mathcal{L}, \boldsymbol{b} \in \mathcal{L}^{\perp}$ and $\boldsymbol{a}^T\boldsymbol{b} = 0$; further, $\boldsymbol{a} = \Pi(\boldsymbol{y}|\mathcal{L}) = \boldsymbol{Py}$ and $\boldsymbol{b} = \Pi(\boldsymbol{y}|\mathcal{L}^{\perp}) = (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}$. A similar result holds for closed convex cones as well: $\mathbb{R}^p = \mathcal{P} \oplus \mathcal{P}^o$ is an orthogonal sum. Consequently, $\boldsymbol{y} \in \mathbb{R}^p$ has a unique orthogonal decomposition of the form $\boldsymbol{y} = \boldsymbol{p} + \boldsymbol{q}$, where $\boldsymbol{p} \in \mathcal{P}, \boldsymbol{q} \in \mathcal{P}^o$, and $\boldsymbol{p}^T\boldsymbol{q} = 0$; further, $\boldsymbol{p} = \Pi(\boldsymbol{y}|\mathcal{P})$ and $\boldsymbol{q} = \Pi(\boldsymbol{y}|\mathcal{P}^o)$. Consequently, if $\boldsymbol{y} = \boldsymbol{p}+\boldsymbol{q}$ where $\boldsymbol{p} \in \mathcal{P}, \boldsymbol{q} \in \mathcal{P}^o$, and $\boldsymbol{p}^T\boldsymbol{q} = 0$, then $\boldsymbol{p} = \Pi(\boldsymbol{y}|\mathcal{P})$ and $\boldsymbol{q} = \Pi(\boldsymbol{y}|\mathcal{P}^o)$. Let us write down the two unique orthogonal decompositions of $\boldsymbol{y}$:

$$\boldsymbol{y} = \Pi(\boldsymbol{y}|\mathcal{L}) + \Pi(\boldsymbol{y}|\mathcal{L}^{\perp}) = \Pi(\boldsymbol{y}|\mathcal{P}) + \Pi(\boldsymbol{y}|\mathcal{P}^o) \tag{3.69}$$

Proof of (B) $\Rightarrow$ *(C)* : Assume that (B) is true. Since $\Pi(\boldsymbol{y}|\mathcal{L}) = \boldsymbol{Py} \in ri(F) \subset \mathcal{P}$, $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y} = \Pi(\boldsymbol{y}|\mathcal{L}^{\perp}) = F^{\perp} \cap \mathcal{P}^o \subset \mathcal{P}^o$ and $\Pi(\boldsymbol{y}|\mathcal{L})^T\Pi(\boldsymbol{y}|\mathcal{L}^{\perp}) = 0$ it follows that the two orthogonal decompositions in (3.69) are identical, and hence $\boldsymbol{Py} = \Pi(\boldsymbol{y}|\mathcal{L}) = \Pi(\boldsymbol{y}|\mathcal{P})$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y} = \Pi(\boldsymbol{y}|\mathcal{L}^{\perp}) = \Pi(\boldsymbol{y}|\mathcal{P}^o)$; this completes the proof of (C).
Proof of (B) $\Rightarrow$ *(A)*: Follows since $\Pi(\boldsymbol{y}|\mathcal{P}) = \boldsymbol{Py} \in ri(F)$. (see Fig. 3.20).
Proof of (A) $\Rightarrow$ *(B)*: Assume that (A) is true. Let $\boldsymbol{y}^*$ denote $\Pi(\boldsymbol{y}|\mathcal{P})$. Then $(\boldsymbol{y}-\boldsymbol{y}^*)^T(\boldsymbol{x}-\boldsymbol{y}^*) \leq 0$ for every $\boldsymbol{x} \in \mathcal{P}$ (by Proposition 3.12.3 on page 114). Suppose that $(\boldsymbol{y}-\boldsymbol{y}^*)^T(\boldsymbol{x}-\boldsymbol{y}^*)$ is negative for some $\boldsymbol{x} \in ri(F)$. Then $g(t) = (\boldsymbol{y}-\boldsymbol{y}^*)^T(t-\boldsymbol{y}^*)$ is negative for values of t in a small neighborhood of $\boldsymbol{x}$ contained in $ri(F)$. Therefore, there is an $\boldsymbol{x}_1 \in ri(F)$ such that $(\boldsymbol{y}-\boldsymbol{y}^*)^T(\boldsymbol{x}_1-\boldsymbol{y}^*) < 0$, and hence $\boldsymbol{y}$ is closer to $\boldsymbol{x}_1$ in $\mathcal{P}$ than to $\boldsymbol{y}^*$; this is a contradiction because $\boldsymbol{y}^*$ is the closest point in $\mathcal{P}$. Therefore, $(\boldsymbol{y}-\boldsymbol{y}^*)^T(\boldsymbol{x}-\boldsymbol{y}^*) = 0$ for every $\boldsymbol{x}$ in $ri(F)$ and $\boldsymbol{y}-\boldsymbol{y}^*$ is orthogonal to $\mathcal{L}$. Since $\boldsymbol{y}^* \in ri(F)$ it follows that $\boldsymbol{y}^* \in \mathcal{L}$. Therefore, $\boldsymbol{Py} = \Pi(\boldsymbol{y}|\mathcal{L}) = \boldsymbol{y}^* \in ri(F)$. Further, $\boldsymbol{y}-\boldsymbol{y}^* = \Pi(\boldsymbol{y}|\mathcal{P}^o) \in \mathcal{P}^o$ by (3.69) and $\boldsymbol{y}-\boldsymbol{y}^* = (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y} = \Pi(\boldsymbol{y}|\mathcal{L}^{\perp}) \in \mathcal{L}^{\perp} \subset F^{\perp}$. Therefore, $\boldsymbol{y}-\boldsymbol{y}^* \in F^{\perp} \cap \mathcal{P}^o$. This completes the proof of (A) $\Rightarrow$ (B). ■

Lemma 3.13.3 *Let* $\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{P}$ *be a projection matrix,* $r = rank(\boldsymbol{P})$ *and* $\mathcal{C} \subset \mathbb{R}^p$ *be a nonempty cone. Then the distribution of* $\|\boldsymbol{PY}\|^2$, *conditional on* $\boldsymbol{PY} \in \mathcal{C}$, *is* χ^2_r.

Proof : Let $\boldsymbol{P} = \boldsymbol{UDU}^T$ where $\boldsymbol{U}$ is an orthogonal matrix and $\boldsymbol{D}$ is a diagonal matrix that we assume, without loss of generality, to be $diag(1, \ldots, 1, 0, \ldots, 0)$ with each of the first r diagonal elements equal to 1 and the rest being zeros. Let $\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{Y}$, and $\boldsymbol{Z}_a = (Z_1, \ldots, Z_r)^T$. Then $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{Z}_a \sim N_r(\boldsymbol{0}, \boldsymbol{I})$. Further,

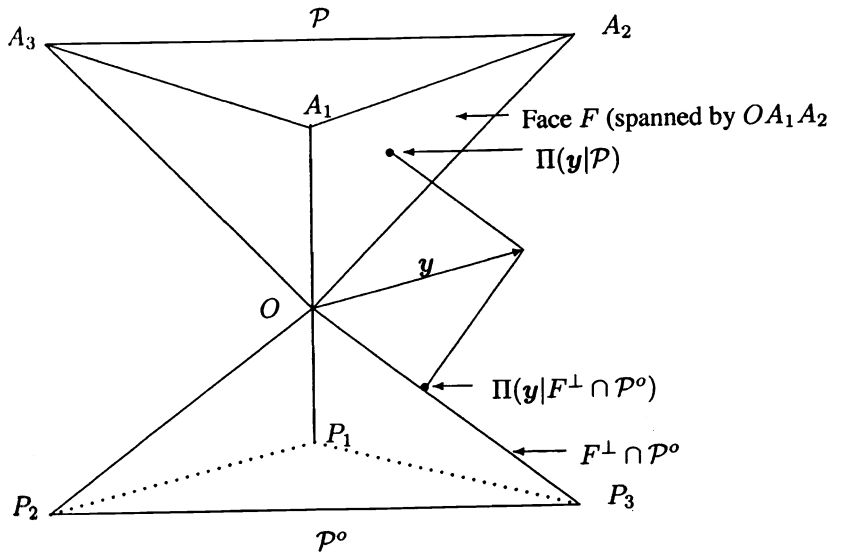

Fig. 3.20 Projections onto F and $F^{\perp} \cap \mathcal{P}^o$.

since $\boldsymbol{U}$ is a rotation matrix, we have that $\|\boldsymbol{PY}\|^2 = \|\boldsymbol{UDU}^T\boldsymbol{Y}\|^2 = \|\boldsymbol{UDZ}\|^2 = \|\boldsymbol{DZ}\|^2 = Z_1^2 + \ldots + Z_r^2$. Now, with $\{\boldsymbol{PY} \in \mathcal{C}\}$ denoting the event that $\boldsymbol{PY} \in \mathcal{C}$, we have that $\{\boldsymbol{PY} \in \mathcal{C}\} = \{\boldsymbol{UDU}^T\boldsymbol{Y} \in \mathcal{C}\} = \{\boldsymbol{U}(\boldsymbol{Z}_a : \boldsymbol{0})^T \in \mathcal{C}\} = \{\boldsymbol{Z}_a \in \mathcal{K}\}$, where $\mathcal{K}$ is a cone in $\mathbb{R}^r$. Since the length and direction of $N(\boldsymbol{0}, \boldsymbol{I})$ are independent, we have that $\text{pr}(\|\boldsymbol{PY}\|^2 \geqslant c \mid \boldsymbol{PY} \in \mathcal{C}) = \text{pr}(Z_1^2 + \ldots + Z_r^2 \geqslant c \mid \boldsymbol{Z}_a \in \mathcal{K}) = \text{pr}(\chi^2_r \geqslant c)$. ■

Lemma 3.13.4 *Let* $\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\mathcal{P} \subset \mathbb{R}^p$ *be a polyhedral cone,* F *be a face of* $\mathcal{P}$, $\boldsymbol{P}$ *be the projection matrix onto the linear space spanned by* F, *and* $r = rank(\boldsymbol{P})$. *Then we have the following.*
(i) $pr\{\|\Pi(\boldsymbol{Y}|\mathcal{P})\|^2 \geqslant c \mid \Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F)\} = pr(\chi^2_r \geqslant c)$.
(ii) Conditional on $\Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F)$, $\|\Pi(\boldsymbol{y}|\mathcal{P})\|^2$ *and* $\|\Pi(\boldsymbol{y}|\mathcal{P}^o)\|^2$, *are independent and are distributed as* χ^2_r *and* χ^2_{p-r}, *respectively. Hence*

$$\begin{aligned} &pr\{\|\Pi(\boldsymbol{Y}|\mathcal{P})\|^2 \geqslant c_1, \|\Pi(\boldsymbol{Y}|\mathcal{P}^o)\|^2 \geqslant c_2 \mid \Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F)\} \\ &\quad = pr(\chi^2_r \geqslant c_1) pr(\chi^2_{p-r} \geqslant c_2). \end{aligned} \tag{3.70}$$

Proof:
(i) Since $\boldsymbol{PY}$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ are independent and $ri(F)$ is a cone, it follows from the previous two lemmas that $\text{pr}\{\|\Pi(\boldsymbol{y}|\mathcal{P})\|^2 \geqslant c \mid \Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F)\} = \text{pr}\{\|\boldsymbol{PY}\|^2 \geqslant c \mid \Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F)\} = \text{pr}\{\|\boldsymbol{PY}\|^2 \geqslant c \mid \boldsymbol{PY} \in ri(F), (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} \in F^{\perp} \cap \mathcal{P}^o\} = \text{pr}\{\|\boldsymbol{PY}\|^2 \geqslant c \mid \boldsymbol{PY} \in ri(F)\} = \text{pr}(\chi^2_r \geqslant c)$.
(ii) By arguments similar to those for the previous part and the fact that $F^{\perp} \cap \mathcal{P}^o$ is

a cone, we have the following:

$$\begin{aligned}
&\text{pr}\{\|\Pi(\boldsymbol{Y}|\mathcal{P})\|^2 \geqslant c_1, \|\Pi(\boldsymbol{Y}|\mathcal{P}^o)\|^2 \geqslant c_2 \mid \Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F)\} \\
&= \text{pr}\{\|\boldsymbol{PY}\|^2 \geqslant c_1, \|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}\|^2 \geqslant c_2 \mid \boldsymbol{PY} \in ri(F), (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} \in F^{\perp} \cap \mathcal{P}^o\} \\
&= \text{pr}\{\|\boldsymbol{PY}\|^2 \geqslant c_1 \mid \boldsymbol{PY} \in ri(F)\} \times \\
&\qquad \text{pr}\{\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}\|^2 \geqslant c_2 \mid (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} \in F^{\perp} \cap \mathcal{P}^o\} \\
&= \text{pr}(\chi^2_r \geqslant c_1)\text{pr}(\chi^2_{p-r} \geqslant c_2). \qquad \blacksquare
\end{aligned}$$

Lemma 3.13.5 *Let $\mathcal{P}$ be a polyhedral cone in $\mathbb{R}^p$. Then there exist a collection of faces of $\mathcal{P}$, say $\{F_1, \ldots, F_K\}$, such that the collection of their relative interiors, $\{ri(F_1), \ldots, ri(F_K)\}$, forms a partition of $\mathcal{P}$. Further,*

$$\|\Pi(\boldsymbol{y}|\mathcal{P})\|^2 = \sum_{i=0}^{K} I\{\Pi(\boldsymbol{y}|\mathcal{P}) \in ri(F_i)\}\|\boldsymbol{P}_i\boldsymbol{y}\|^2$$

where $I(.)$ is the indicator function and $\boldsymbol{P}_i$ is the projection matrix onto the linear space spanned by F_i.

Proof: Since $\mathcal{P}$ is a closed convex set, $\Pi(\boldsymbol{Y}|\mathcal{P})$ exists and is unique (see Proposition 3.12.3 on page 114), or Stoer and Witzgall (1970), p 51). If F and G are two different faces of $\mathcal{P}$, then $ri(F) \cap ri(G)$ is the empty set (see Stoer and Witzgall (1970), p 43, 2.4.14). The polyhedral cone $\mathcal{P}$, a typical face F, and its relative interior $ri(F)$ can be expressed as follows :

$$\begin{aligned}
&\mathcal{P} = \{\boldsymbol{x} : \boldsymbol{a}_i^T\boldsymbol{x} \leq 0 \text{ for } i = 1, \ldots, m\}, \\
&F = \{\boldsymbol{x} : \boldsymbol{a}_i^T\boldsymbol{x} = 0 \text{ for } i \in I, \boldsymbol{a}_j^T\boldsymbol{x} \leq 0 \text{ for } j \in J\}, \\
&ri(F) = \{\boldsymbol{x} : \boldsymbol{a}_i^T\boldsymbol{x} = 0 \text{ for } i \in I, \boldsymbol{a}_j^T\boldsymbol{x} < 0 \text{ for } j \in J\}
\end{aligned} \tag{3.71}$$

where $\{I, J\}$ is a partition of $\{1, \ldots, m\}$; it is assumed that the equality constraints defining F are contained in $\{\boldsymbol{a}_i^T\boldsymbol{x} = 0, i \in I\}$ and hence the set of constraints $\{\boldsymbol{a}_j^T\boldsymbol{x} \leq 0, j \in J\}$ does not include a constraint of the form $\boldsymbol{a}^T\boldsymbol{x} = 0$ for some $\boldsymbol{a} \neq \boldsymbol{0}$. Thus, it is clear that the collection of all relative interiors of faces of $\mathcal{P}$ covers $\mathcal{P}$. Let $\{F_1, \ldots, F_K\}$ denote a collection of faces of $\mathcal{P}$ such that $\{ri(F_1), \ldots, ri(F_K)\}$ forms a partition of $\mathcal{P}$. Let $\boldsymbol{P}_i$ denote the projection matrix onto the linear space spanned by F_i, $i = 1, \ldots, K$. Now, if $\Pi(\boldsymbol{y}|\mathcal{P}) \in ri(F_i)$ then $\Pi(\boldsymbol{y}|\mathcal{P}) = \boldsymbol{P}_i\boldsymbol{y}$, by Lemma 3.13.2. This completes the proof. ■

The next lemma is really a restatement of the results in the previous lemmas, but it is stated here because this alternative statement is helpful.

Lemma 3.13.6 *Let $\mathcal{P}$ be a convex polyhedral in $\mathbb{R}^p$, and $\boldsymbol{X}_{p\times 1} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Then there exists a collection of faces of $\mathcal{P}$, say $\{F_1, \ldots, F_k\}$, such that their relative interiors, $\{ri(F_1), \ldots, ri(F_k)\}$, are a partition of $\mathcal{P}$. Let $\mathcal{S}_i = \{\boldsymbol{x} : \Pi(\boldsymbol{x}|\mathcal{P}) \in ri(F_i)\}$, $\boldsymbol{P}_i$ denote the projection matrix onto the linear space spanned by F_i, and $\bar{\chi}^2(\boldsymbol{I}, \mathcal{P}) = \|\boldsymbol{X}\|^2 - \min\{\|\boldsymbol{X} - \boldsymbol{a}\|^2 : \boldsymbol{a} \in \mathcal{P}\}$. Then, we have the following :*

(i) $\bar{\chi}^2(\boldsymbol{I}, \mathcal{P}) = \sum_{i=1}^{k} I(\boldsymbol{X} \in \mathcal{S}_i)\|\boldsymbol{P}_i\boldsymbol{X}\|^2$.
(ii) If $\boldsymbol{X} \in \mathcal{S}_i$, then $\boldsymbol{P}_i\boldsymbol{X}$ and $(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{X}$ are the projections of $\boldsymbol{X}$ onto $\mathcal{P}$ and $\mathcal{P}^o$, respectively; hence $\bar{\chi}^2(\boldsymbol{I}, \mathcal{P}) = \|\boldsymbol{P}_i\boldsymbol{X}\|^2$ and $\bar{\chi}^2(\boldsymbol{I}, \mathcal{P}^o) = \|(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{X}\|^2$.
(iii) If $\boldsymbol{X} \in \mathcal{S}_i$, then $\|\boldsymbol{P}_i\boldsymbol{X}\|^2$ and $\|(\boldsymbol{I}-\boldsymbol{P}_i)\boldsymbol{X}\|^2$ are independent and are distributed as χ^2_ν and $\chi^2_{p-\nu}$, respectively, where $\nu = rank(\boldsymbol{P}_i)$. ■

Lemma 3.13.7 *Let $\boldsymbol{Y} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$, $\mathcal{P}$ be a polyhedral cone, and let $\bar{\chi}^2$ denote $\|\Pi(\boldsymbol{Y}|\mathcal{P})\|^2$. Then*

$$pr(\bar{\chi}^2 \geqslant c) = \sum_{i=0}^{p} w_i pr(\chi^2_i \geqslant c), \text{ where } w_i = \sum_{rank(\boldsymbol{P}_j)=i} pr\{\Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F_j)\}.$$

Thus, $w_i = pr\{\Pi(\boldsymbol{Y}|\mathcal{P})$ lies in the relative interior of a face of $\mathcal{P}$ of dimension $i\}$, for $i = 0, \ldots, p$. Further, $w_0 + \ldots + w_p = 1$.

Proof: In view of the previous lemma, we have the following:

$$\begin{aligned}
&\text{pr}(\bar{\chi}^2 \geqslant c) \\
&= \sum_{j=0}^{k} \text{pr}\{\|\Pi(\boldsymbol{Y}|\mathcal{P})\|^2 \geqslant c \mid \Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F_j)\} \times \text{pr}\{\Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F_j)\} \qquad (3.72) \\
&= \sum_{i=0}^{p} w_i \text{pr}(\chi^2_i \geqslant c), \text{ where } w_i = \sum_{rank(\boldsymbol{P}_j)=i} \text{pr}\{\Pi(\boldsymbol{Y}|\mathcal{P}) \in ri(F_j)\}.
\end{aligned}$$

Thus, $w_i = \text{pr}\{\Pi(\boldsymbol{Y}|\mathcal{P})$ *lies on the relative interior of a face of dimension i* $\}$, the dimension of a face being defined as the dimension of the linear space spanned by the face. ■

Proof of Theorem 3.4.2 continued: We will prove the theorem for a polyhedral cone first, and then deduce the result for a general cone. Assume that $\mathcal{C}$ is a polyhedral cone. Let $\boldsymbol{V}^{-1} = \boldsymbol{A}^T\boldsymbol{A}$ be a factorization of $\boldsymbol{V}^{-1}$, $\boldsymbol{X} = \boldsymbol{AY}$, and $\mathcal{P}$ be the polyhedral cone $\boldsymbol{A}\mathcal{C} = \{\boldsymbol{Ax} : \boldsymbol{x} \in \mathcal{C}\}$. Then the faces of $\mathcal{C}$ take the form $\boldsymbol{AF}$ for some face F of $\mathcal{P}$ and $dim(\boldsymbol{AF}) = dim(F)$ = dimension of the linear space spanned by F. Now,

$$\begin{aligned}
LRT &= \boldsymbol{Y}^T\boldsymbol{V}^{-1}\boldsymbol{Y} - \min\{(\boldsymbol{Y} - \boldsymbol{\alpha})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathcal{C}\} \\
&= \boldsymbol{X}^T\boldsymbol{X} - \min\{(\boldsymbol{X} - \boldsymbol{\beta})^T(\boldsymbol{X} - \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{P}\}.
\end{aligned} \tag{3.73}$$

It follows from Lemma 3.13.7 that $\text{pr}(LRT \leq c \mid H_0) = \sum_{i=0}^{p} w_i(p, \boldsymbol{V}, \mathcal{P})\text{pr}(\chi^2_i \leq c)$, where $w_i(p, \boldsymbol{V}, \mathcal{P}) = \text{pr}\{\Pi(\boldsymbol{X}||\mathcal{P})$ lies on the relative interior of a face G of $\mathcal{P}$ with $dim(G) = i\} = \text{pr}\{\Pi_{\boldsymbol{V}}(\boldsymbol{Y}|\mathcal{C})$ lies on the relative interior of a face F of $\mathcal{C}$ with $dim(F) = i\}$.

Now let us relax the assumption that $\mathcal{C}$ is a polyhedral and assume only that it is a closed convex cone. Let $\mathcal{C}_1 \subset \mathcal{C}_2 \subset \ldots$ be a sequence of polyhedral cones such that it converges to $\mathcal{C}$. Then, for any $c > 0$, $\text{pr}(\|\Pi_{\boldsymbol{V}}(\boldsymbol{Y}, \mathcal{C}_n)\|^2_{\boldsymbol{V}} \leq c)$ converges to $\text{pr}(\|\Pi_{\boldsymbol{V}}(\boldsymbol{Y}, \mathcal{C})\|^2_{\boldsymbol{V}} \leq c)$ as $n \to \infty$. For any fixed $(i, p, \boldsymbol{V})$, the sequence

$\{w_i(p, \boldsymbol{V}, \mathcal{C}_n)\}$ is bounded, and hence has limit points. Therefore, $\text{pr}(\|\Pi_{\boldsymbol{V}}(\boldsymbol{Y}, \mathcal{C}_n)\|_{\boldsymbol{V}}^2 \leq c)$ has limit points of the form $\sum_{i=0}^{p} w_i \text{pr}(\chi_i^2 \leq c)$ for some $\{w_i\}$. Since all such limit points are equal to $\text{pr}(\|\Pi_{\boldsymbol{V}}(\boldsymbol{Y}, \mathcal{C})\|_{\boldsymbol{V}}^2 \leq c)$, for any c, it follows that limit points of $\{w_i(p, \boldsymbol{V}, \mathcal{C}_n)\}$ are unique. Therefore, we define $w_i(p, \boldsymbol{V}, \mathcal{C}) = \lim_{n\to\infty} w_i(p, \boldsymbol{V}, \mathcal{C}_n)$. ■

Proof of Proposition 3.6.1: The proofs of virtually all of these are based on the details of the proof of the previous theorem. The proof of part 1 is contained in the proof of Theorem 3.4.2 just given. Part 2 is a corollary to part 1. It was also proved by Nuesch (1964); see also Perlman (1969) and Wolak (1987). To our knowledge, part 3 has been a known result but it is not clear whether or not a proof has been published. This was posed as a conjecture (see Shapiro (1987)); Professor Alexander Shapiro informed us that he has received more than one proof of the conjecture, one by Professor J. Kinoses and another by Professor J. Lawrence; Professor Tim Robertson also informed us that he is aware of other proofs. Part 4 follows from part 3. Part 5 is shown in Shapiro (1985); see also Shapiro (1988, equation 3.4). Part 6 can be established easily by applying the transformation $\boldsymbol{Y} = \boldsymbol{RZ}$ where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{V})$; see also Shapiro (1988, equation 5.3). See equations (5.4), (5.5), (5.8), and (5.10) in Shapiro (1988) for parts 7, 8, 9, and 10, respectively. ■

Proof of Theorem 3.9.1: (a) Let $\boldsymbol{Z} = \boldsymbol{E}/\sigma$ and assume that the null hypothesis holds. Then, we have that

$$\bar{E}_A^2 = \{\min_{\boldsymbol{a}\in\mathcal{M}} Q^*(\boldsymbol{a}) - \min_{\boldsymbol{a}\in\mathcal{C}} Q^*(\boldsymbol{a})\} / \min_{\boldsymbol{a}\in\mathcal{M}} Q^*(\boldsymbol{a}) \tag{3.74}$$

where $Q^*(\boldsymbol{a}) = (\boldsymbol{AZ} - \boldsymbol{a})^T \boldsymbol{W}^{-1} (\boldsymbol{AZ} - \boldsymbol{a})$ and $\boldsymbol{A} = \boldsymbol{W}^{-1}\boldsymbol{X}^T\boldsymbol{U}^{-1}$. The distribution of $\boldsymbol{Z}$ and hence that of (3.74) does not depend on any unknown parameters. Consequently, the null distribution of $\bar{E}_A^2$ is the same for any value of $\boldsymbol{\theta}$ in $\mathcal{M}$ and any value of $\sigma > 0$; by similar arguments, the corresponding result holds for $\bar{F}_A$ as well.

(b) Let $\boldsymbol{\theta} \in \mathcal{C}$. Let

$$\begin{aligned} g_1(\boldsymbol{\theta}, \boldsymbol{E}) &= \min\{\|\boldsymbol{X\theta} + \boldsymbol{E} - \boldsymbol{Xa}\|_{\boldsymbol{U}}^2 : \boldsymbol{a} \in \mathcal{C}\} \\ \text{and} \quad g_2(\boldsymbol{\theta}, \boldsymbol{E}) &= \min\{\|\boldsymbol{X\theta} + \boldsymbol{E} - \boldsymbol{Xa}\|_{\boldsymbol{U}}^2 : \boldsymbol{a} \in \mathbb{R}^p\}. \end{aligned} \tag{3.75}$$

To show the dependence of $\bar{E}_B^2$ on $\boldsymbol{\theta}$, let us write

$$\bar{E}_B^2(\boldsymbol{\theta}, \boldsymbol{E}) = \{g_1(\boldsymbol{\theta}, \boldsymbol{E}) - g_2(\boldsymbol{\theta}, \boldsymbol{E})\}/g_1(\boldsymbol{\theta}, \boldsymbol{E}).$$

Now,

$$\begin{aligned} &\min\{\|\boldsymbol{Y} - \boldsymbol{Xa}\|_{\boldsymbol{U}}^2 : \boldsymbol{a} \in \mathcal{C}\} = \min\{\|\boldsymbol{X\theta} + \boldsymbol{E} - \boldsymbol{Xa}\|_{\boldsymbol{U}}^2 : \boldsymbol{a} \in \mathcal{C}\} \\ &= \min\{\|\boldsymbol{E} - \boldsymbol{X}(\boldsymbol{a} - \boldsymbol{\theta})\|_{\boldsymbol{U}}^2 : \boldsymbol{a} \in \mathcal{C}\} = \min\{\|\boldsymbol{E} - \boldsymbol{Xb}\|_{\boldsymbol{U}}^2 : \boldsymbol{b} \in \mathcal{C} - \boldsymbol{\theta}\} \\ &\leq \min\{\|\boldsymbol{E} - \boldsymbol{Xb}\|_{\boldsymbol{U}}^2 : \boldsymbol{b} \in \mathcal{C}\} \end{aligned}$$

the last step follows since $\mathcal{C} \subset \mathcal{C} - \boldsymbol{\theta}$. Similarly,

$$\min\{\|\boldsymbol{Y} - \boldsymbol{Xa}\|_{\boldsymbol{U}}^2 : \boldsymbol{a} \in \mathbb{R}^p\} = \min\{\|\boldsymbol{E} - \boldsymbol{Xb}\|_{\boldsymbol{U}}^2 : \boldsymbol{b} \in \mathbb{R}^p\}.$$

Thus, $g_1(\boldsymbol{\theta}, \boldsymbol{E}) \leq g_1(\boldsymbol{0}, \boldsymbol{E})$ and $g_2(\boldsymbol{\theta}, \boldsymbol{E}) = g_2(\boldsymbol{0}, \boldsymbol{E})$. Now,

$$\begin{aligned} \bar{E}_B^2(\boldsymbol{\theta}, \boldsymbol{E}) &= 1 - \{g_2(\boldsymbol{\theta}, \boldsymbol{E})/g_1(\boldsymbol{\theta}, \boldsymbol{E})\} = 1 - \{g_2(\boldsymbol{0}, \boldsymbol{E})/g_1(\boldsymbol{\theta}, \boldsymbol{E})\} \\ &\leqslant 1 - \{g_2(\boldsymbol{0}, \boldsymbol{E})/g_1(\boldsymbol{0}, \boldsymbol{E})\} = \bar{E}_B^2(\boldsymbol{0}, \boldsymbol{E}). \end{aligned}$$

Therefore, $\bar{E}_B^2(\boldsymbol{\theta}, \boldsymbol{E}) \leq \bar{E}_B^2(\boldsymbol{0}, \boldsymbol{E})$ and

$$\text{pr}\{\bar{E}_B^2(\boldsymbol{\theta}, \boldsymbol{E}) \geqslant c\} \leq \text{pr}\{\bar{E}_B^2(\boldsymbol{0}, \boldsymbol{E}) \geqslant c\}, \quad \boldsymbol{\theta} \in \mathcal{C}.$$

Hence the least favorable null value for $\bar{E}_B^2$ is $\boldsymbol{0}$.

To prove the corresponding result for $\bar{F}$, note that $S^2(\boldsymbol{\theta}, \boldsymbol{E}) = S^2(\boldsymbol{0}, \boldsymbol{E})$. Now,

$$\begin{aligned} \bar{F}(\boldsymbol{\theta}, \boldsymbol{E}) &= \{g_1(\boldsymbol{\theta}, \boldsymbol{E}) - g_2(\boldsymbol{\theta}, \boldsymbol{E})\}/S^2(\boldsymbol{\theta}, \boldsymbol{E}) \\ &= \{g_1(\boldsymbol{\theta}, \boldsymbol{E}) - g_2(\boldsymbol{0}, \boldsymbol{E})\}/S^2(\boldsymbol{0}, \boldsymbol{E}) \\ &\leqslant \{g_1(\boldsymbol{0}, \boldsymbol{E}) - g_2(\boldsymbol{0}, \boldsymbol{E})\}/S^2(\boldsymbol{0}, \boldsymbol{E}) = \bar{F}(\boldsymbol{0}, \boldsymbol{E}). \end{aligned}$$

Therefore, $\text{pr}\{\bar{F}_B(\boldsymbol{\theta}, \boldsymbol{E}) \geqslant c\} \leq \text{pr}\{\bar{F}_B(\boldsymbol{0}, \boldsymbol{E}) \geqslant c\}$ for $\boldsymbol{\theta} \in \mathcal{C}$. ■

Proof of Theorem 3.9.2 : (a) Without loss of generality assume that $\sigma = 1$, $\boldsymbol{U}$ is the identity matrix, and the null hypothesis is $\boldsymbol{R\theta} = \boldsymbol{0}$ where $\boldsymbol{R}$ is a $r \times p$ matrix of rank r. Now

$$\bar{E}_B^2 = \{q(\boldsymbol{\theta}^0) - q(\boldsymbol{\theta}^1)\}/Q(\boldsymbol{\theta}^0).$$

Note that

$$Q(\boldsymbol{\theta}^0) = [\{q(\boldsymbol{\theta}^0) - q(\boldsymbol{\theta}^1)\} + \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}\|^2 + q(\boldsymbol{\theta}^1)].$$

Since $\hat{\boldsymbol{\theta}}$ is independent of $(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})$, we have that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}$ is also independent of $\{q(\boldsymbol{\theta}^0) - q(\boldsymbol{\theta}^1), q(\boldsymbol{\theta}^1)\}$. Let $\{F_1, \ldots, F_k\}$ be a collection of faces of $\mathcal{C} \cap \mathcal{M}^\perp$ as in Lemma 3.13.4.

Now, suppose that $\Pi_{\boldsymbol{W}}(\hat{\boldsymbol{\theta}}|\mathcal{C} \cap \mathcal{M}^\perp) \in ri(F)$ for some $F \in \{F_1, \ldots, F_k\}$. Let $dim\{ri(F)\} = j$. Then $q(\boldsymbol{\theta}^0) - q(\boldsymbol{\theta}^1) = \|\boldsymbol{P}\hat{\boldsymbol{\theta}}\|^2$ where $\boldsymbol{P}$ is the projection matrix onto the linear space spanned by F. Now, since $q(\boldsymbol{\theta}^0) \sim \chi_r^2, q(\boldsymbol{\theta}^0) - q(\boldsymbol{\theta}^1) = \|\boldsymbol{P}\hat{\boldsymbol{\theta}}\|^2 \sim \chi_j^2$, and $q(\boldsymbol{\theta}^0) - q(\boldsymbol{\theta}^1)$ and $q(\boldsymbol{\theta}^1)$ are independent by Proposition 3.12.6 part (g), we have that $q(\boldsymbol{\theta}^1) \sim \chi_{r-j}^2$. Therefore,

$$\begin{aligned} \bar{E}_A^2 &= \{q(\boldsymbol{\theta}^0) - q(\boldsymbol{\theta}^1)\}/[\{q(\boldsymbol{\theta}^0) - q(\boldsymbol{\theta}^1)\} + \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}\|^2 + q(\boldsymbol{\theta}^1)] \\ &\sim \chi_j^2/(\chi_j^2 + \chi_{N-p+r-j}^2), \end{aligned} \tag{3.76}$$

where the two terms in the denominator are independent. Therefore,

$$\bar{E}_A^2 \sim \beta\{2^{-1}j, 2^{-1}(N - p + r - j)\}.$$

Let B_i denote the event $\Pi(\hat{\theta}|\mathcal{C}\cap\mathcal{M}^{\perp}) \in ri(F_i)\}$. Now, by adopting an argument similar to the derivation of $\bar{\chi}^2$ we have the following:

$$\begin{aligned}\mathrm{pr}(\bar{E}_A^2 \le c|H_0) &= \sum_i \mathrm{pr}\{\bar{E}_A^2 \le c \mid B_i\}\mathrm{pr}\{B_i\}.\\ &= \sum_{j=0}^{p}\sum_{dim(F_i)=j} \mathrm{pr}\{\bar{E}_A^2 \le c \mid B_i\}\mathrm{pr}(B_i)\\ &= \sum_{j=0}^{p} w_j(p, \boldsymbol{W}, \mathcal{C}\cap\mathcal{M}^{\perp})\mathrm{pr}[\beta\{2^{-1}j, 2^{-1}(N-p+r-j)\} \le c].\end{aligned}$$

The distribution of $\bar{F}$ is obtained similarly.

$$\begin{aligned}\mathrm{pr}(\bar{F}_A \le c|H_0) &= \sum_{i=1}^{K}\mathrm{pr}\{\|\Pi(\hat{\theta}|\mathcal{C}\cap\mathcal{M}^{\perp})\|^2 S^{-2} \le c \mid B_i\}\mathrm{pr}(B_i)\\ &= \sum_{j=0}^{p}\sum_{dim(F_i)=j} \mathrm{pr}\{\|\boldsymbol{P}_i\hat{\theta}\|^2 S^{-2} \le c \mid B_i)\mathrm{pr}(B_i)\\ &= \sum_{j=0}^{p} w_j(p, \boldsymbol{W}, \mathcal{C}\cap\mathcal{M}^{\perp})\mathrm{pr}(jF_{j,\nu} \le c).\end{aligned}$$

■

Proof of part (b):

$$\bar{E}_B^2 = \{Q(\theta^1) - Q(\theta^2)\}/Q(\theta^1) = q(\theta^1)/[q(\theta^1) + \|\boldsymbol{Y} - \boldsymbol{X}\hat{\theta}\|_U^2].$$

As in Lemma 3.13.4, let $\{F_1, \ldots, F_k\}$ be a collection of faces of $\mathcal{C}$ so that their relative interiors are a partition of $\mathcal{C}$. Let A_i be the event $\Pi_{\boldsymbol{W}}(\hat{\theta}|\mathcal{C}) \in ri(F_i)$. Now, we have

$$\begin{aligned}\mathrm{pr}(\bar{E}_B^2 \le c \mid \theta = \mathbf{0}) &= \sum_{i=0}^{K}\mathrm{pr}[q(\theta^1)/\{q(\theta^1) + \|\boldsymbol{Y} - \boldsymbol{X}\hat{\theta}\|_U^2\} \le c \mid A_i]\mathrm{pr}(A_i)\\ &= \sum_{j=0}^{p}\sum_{dim(F_i)=j} \mathrm{pr}[\chi^2_{p-j}/\{\chi^2_{p-j} + \chi^2_{N-p}\} \le c]\mathrm{pr}(A_i)\\ &= \sum_{j=0}^{p} w_{p-j}(p, \boldsymbol{W}, \mathcal{C})\mathrm{pr}[\beta\{2^{-1}j, 2^{-1}(N-p)\} \le c].\end{aligned}$$

The proof for the $\bar{F}$ is similar.

$$\begin{aligned}\mathrm{pr}(\bar{F}_B \le c|\theta = \mathbf{0}) &= \sum_{i=0}^{K}\mathrm{pr}[q(\theta^1)/S^2 \le c \mid A_i]\mathrm{pr}(A_i)\\ &= \sum_{i=0}^{k}\sum_{j=0}^{p}\mathrm{pr}[\chi^2_{p-j}/S^2 \le c]\mathrm{pr}(A_i) = \sum_{j=0}^{p} w_{p-j}(p, \boldsymbol{W}, \mathcal{C})\mathrm{pr}[jF_{j,\nu} \le c].\end{aligned}$$

■

Problems

Section 3.3

3.1 *The length and direction of $N(\mathbf{0}, \boldsymbol{I})$ are independent; two-dimensional case:* Let $\boldsymbol{X} = (X_1, X_2)^T \sim N(\mathbf{0}, \boldsymbol{I})$, where $\boldsymbol{I}$ is the 2×2 identity matrix. Let (r, α) be the polar coordinates of $\boldsymbol{X}$ given by $(X_1, X_2) = r(\sin\alpha, \cos\alpha)$. Obtain the joint density of (r, α), and deduce that r and α are independent and that $\|\boldsymbol{X}\|$ and $\boldsymbol{X}/\|\boldsymbol{X}\|$ are independent [see Anderson (1984), p 279].

3.2 *Derivation of a $\bar{\chi}^2$ distribution in three dimensions:* Let $\boldsymbol{X} = (X_1, X_2, X_3)^T \sim N(\boldsymbol{\theta}, \boldsymbol{I})$ and let the null and alternative hypotheses be $H_0 : \boldsymbol{\theta} = \mathbf{0}$ and $H_1 : \boldsymbol{\theta} \geqslant \mathbf{0}$. Let $\tilde{\boldsymbol{\theta}}$ be the *mle* of $\boldsymbol{\theta}$ under H_1 based on a single observation of $\boldsymbol{X}$; the test of H_0 against H_1 is also to be based on a single observation of $\boldsymbol{X}$. Show that $\tilde{\boldsymbol{\theta}} = (X_1 I\{X_1 > 0\}, X_2 I\{X_2 > 0\}, X_3 I\{X_3 > 0\}$, where I is the indicator function and $LRT = X_1^2 I(X_1 > 0) + X_2^2 I(X_2 > 0) + X_3^2 I(X_3 > 0)$. Show that $\boldsymbol{X}$ has exactly 3, 2, 1, or 0 positive components with probabilities (1/8), (3/8), (3/8), or (1/8), respectively. Deduce that, for $c > 0$, the null distribution of LRT is given by, $\mathrm{pr}(LRT \le c) = \sum_{i=0}^{3} w_i \mathrm{pr}(\chi_i^2 \le c)$, where $(w_0, w_1, w_2, w_3) = (1/8, 3/8, 3/8, 1/8)$. Verify that $w_i = \mathrm{pr}$(exactly i components of $\boldsymbol{Z}$ are positive), where $\boldsymbol{Z}_{3\times 1} \sim N(\mathbf{0}, \boldsymbol{I})$, and $i = 0, \ldots, 3$.

3.3 *Derivation of a $\bar{\chi}^2$ distribution in three dimensions:* Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$, where $\boldsymbol{X}$ is 3×1, and let $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^3 : \theta_2 \geqslant 0, \theta_3 \geqslant 0\}$. Show that the *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \in \mathcal{C}$ and the LRT of $H_0 : \boldsymbol{\theta} = \mathbf{0}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ are given by

$$(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3) = \begin{cases} (X_1, X_2, X_3) & \text{if } X_2 \geqslant 0 \text{ and } X_3 \geqslant 0\\ (X_1, X_2, 0) & \text{if } X_3 < 0 \text{ and } X_2 > 0\\ (X_1, 0, X_3) & \text{if } X_2 < 0 \text{ and } X_3 > 0\\ (X_1, 0, 0) & \text{if } X_2 < 0 \text{ and } X_3 < 0 \end{cases}$$

and

$$LRT = \begin{cases} X_1^2 + X_2^2 + X_3^2 & \sim \chi_3^2 & \text{if } X_2 \geqslant 0 \text{ and } X_3 \geqslant 0\\ X_1^2 + X_2^2 & \sim \chi_2^2 & \text{if } X_3 < 0 \text{ and } X_2 > 0\\ X_1^2 + X_3^2 & \sim \chi_2^2 & \text{if } X_2 < 0 \text{ and } X_3 > 0\\ X_1^2 & \sim \chi_1^2 & \text{if } X_2 < 0 \text{ and } X_3 < 0, \end{cases}$$

respectively. Deduce that the null distribution of LRT is given by

$$\mathrm{pr}(LRT \le c|H_0) = (1/4)\mathrm{pr}(\chi_1^2 \le c) + (1/2)\mathrm{pr}(\chi_2^2 \le c) + (1/4)\mathrm{pr}(\chi_3^2 \le c).$$

3.4 *Derivation of a $\bar{\chi}^2$ distribution in four dimensions:* Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$ where $\boldsymbol{X}$ is 4×1. Let the null and alternative hypothesis be $H_0 : \boldsymbol{\theta} = \mathbf{0}$ and $H_1 : \theta_3 \geqslant 0, \theta_4 \geqslant 0$. Show that

$$\mathrm{pr}(LRT \le c|H_0) = (1/4)\mathrm{pr}(\chi_2^2 \le c) + (1/2)\mathrm{pr}(\chi_3^2 \le c) + (1/4)\mathrm{pr}(\chi_4^2 \le c).$$

3.5 *Derivation of a $\bar{\chi}^2$ distribution in p dimensions:* Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$ where $\boldsymbol{X}$ is $p \times 1$. Let the null and alternative hypothesis be $H_0 : \boldsymbol{\theta} = \mathbf{0}$ and $H_1 : \boldsymbol{\theta}_2 \geqslant 0$

where θ_2 is a subvector of θ consisting of k elements. Show that $\text{pr}(LRT \leq c|H_0) = w_{p-k}\text{pr}(\chi^2_{p-k} \leq c) + \ldots + w_p\text{pr}(\chi^2_p \leq c)$ where $w_{p-k+i} = 2^{-k}k!/\{(k-i)!i!\}$.

3.6 *Application of LRT in two dimensions:* Let $\boldsymbol{X} = (X_1, X_2)^T \sim N(\boldsymbol{\theta}, \boldsymbol{I})$. The mean of a sample of five observations on $\boldsymbol{X}$ is $(-1, 2)$. Compute the *mle* under the constraint $\theta_1 \geqslant 0, \theta_2 \geqslant 0$ and the p-value for testing $H_0 : \theta_1 = \theta_2 = 0$ against $H_1 : \theta_1 \geqslant 0, \theta_2 \geqslant 0$.

3.7 *Derivation of a $\bar{\chi}^2$ distribution in two dimensions:* Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}), H_0 : \theta_1 = \theta_2 = 0, H_1 : \theta_1 \geqslant 0, \theta_2 - m\theta_1 \geqslant 0$, where m is known. Show from first principles that the distribution of the LRT under H_0 is given by $\text{pr}(LRT \leq c|H_0) = \sum_{i=0}^{i=2} w_i\text{pr}(\chi^2_i \leq c)$ where $w_0 = q, w_1 = 0.5, w_2 = 0.5 - q$, and $q = (2\pi)^{-1}cos^{-1}\{-m/(1+m^2)^{1/2}\}$. Indicate the critical region corresponding to the critical value 4. [The details of the derivations are also given in (Gourieroux et al., 1982, section 4.2).]

3.8 *Derivation of the critical region of LRT in two dimensions:* For Example 3.3.5, prove the following: (i) $(1, \rho) \perp_V (0, 1)$ and $(1, \rho^{-1}) \perp_V (1, 0)$. (ii) The boundaries of the polar cone of the nonnegative orthant are $\theta_2 = \rho^{-1}\theta_1$ and $\theta_2 = \rho\theta_1$. (iii) For $\rho = 0.9$ and $\bar{\boldsymbol{x}} = (-3, 2)$, the *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \geqslant \boldsymbol{0}$ is the point of intersection of the θ_2 axis and the line of slope ρ through $\bar{\boldsymbol{x}}$. (iv) For $\rho = 0.9$ and $\bar{\boldsymbol{x}} = (3, -2)$, the *mle* of $\boldsymbol{\theta}$ subject to $\boldsymbol{\theta} \geqslant \boldsymbol{0}$ is the point of intersection of the θ_1 axis and the line of slope ρ^{-1} through $\bar{\boldsymbol{x}}$. Indicate the critical region $\{LRT \geqslant 6\}$.

3.9 Let X_i be an observation from the ith treatment and assume that $X_i \sim N(\mu_i, 1)$ for $i = 1, 2, 3$. Let $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 \geq \mu_2$ and $\mu_1 \geq \mu_3$. (i) Verify that the LRT of H_0 against H_1 rejects H_0 at 5% level when $(X_1, X_2, X_3) = (50, 150, 70)$. (ii) Verify that the phenomenon in part (i) is essentially the same the same as that in Example 3.3.6.
[Hint: Let $Y_1 = X_1 - X_2, Y_2 = X_1 - X_3, \boldsymbol{Y} = (Y_1, Y_2)^T$ and $\boldsymbol{X} = (X_1, X_2)^T$. Then $\boldsymbol{Y} = \boldsymbol{AX}, \boldsymbol{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$, and $\boldsymbol{\theta} = \boldsymbol{A\mu}$ where $\boldsymbol{V} = \boldsymbol{AA}^T, \boldsymbol{A} = (1, -1, 0; 1, 0, -1)$, and $\boldsymbol{V} = 2(1, 0.5; 0.5, 1)$. Verify that the LRT of H_0 against H_1 is equivalent to the LRT of $H_0^* : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1^* : \boldsymbol{\theta} \geq \boldsymbol{0}$ based on a single observation of $\boldsymbol{Y}$. Verify (by using the result in Problem 8) that $\boldsymbol{Y} = (-100, -20)^T$, which corresponds to $\boldsymbol{X} = (50, 150, 70)^T$, is in the critical region.]

Section 3.4

3.10 *Pythagoras theorem in $\mathbb{R}^p$:* Let $\boldsymbol{V}$ be a $p \times p$ positive definite matrix. Define the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_V = \boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{y}$, and the distance function $\|\boldsymbol{x} - \boldsymbol{y}\|_V = \{\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{y}\}^{1/2}$. We say that $\boldsymbol{x}$ is *orthogonal* to $\boldsymbol{y}$ with respect to $\boldsymbol{V}$, denoted $\boldsymbol{x} \perp_V \boldsymbol{y}$, if $\boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{y} \neq \boldsymbol{0}$, and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_V = 0$. If A, B, C are three points in $\mathbb{R}^p$ such that $AB \perp_V AC$, show that $\|AB\|_V^2 + \|AC\|_V^2 = \|BC\|_V^2$.

3.11 Let $\mathcal{L}$ be a subspace of $\mathbb{R}^p, r = dim(\mathcal{L})$, and $\boldsymbol{x} \in \mathbb{R}^p$. Let O be the origin and A be the point such $\boldsymbol{x} = OA$. Let $\boldsymbol{V}$ be a $p \times p$ positive definite matrix. Show that the point in $\mathcal{L}$ that is $\boldsymbol{V}$-closest to A is obtained by dropping a $\boldsymbol{V}$-perpendicular to $\mathcal{L}$ from A. [Hint: Use the Pythagoras theorem.]

3.12 Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V}), H_0 : \boldsymbol{\theta} = \boldsymbol{0}$, and $H_1 : \boldsymbol{\theta} \in \mathcal{C}$. Suppose that $\boldsymbol{V}^{-1} = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ for some $\boldsymbol{A}$ and $\boldsymbol{B}$. Show that $\boldsymbol{BA}^{-1}$ is an orthogonal matrix. Now, transform the setting from $\{\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{V}, \mathcal{C}\}$ by $\boldsymbol{A}$ as in Example 3.3.4 on page 68; also, consider the transformation by $\boldsymbol{B}$. Show that the geometry of the image under transformation A is the same as that under transformation $\boldsymbol{B}$ except for a rotation defined by the orthogonal matrix $\boldsymbol{BA}^{-1}$.

3.13 Let $Y_1, \ldots, Y_n$ be independent standard normal rv's, and let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. Let $\boldsymbol{B}$ be a $k \times n$ matrix of given constants such that $\{\boldsymbol{y} : \boldsymbol{By} \geqslant \boldsymbol{0}\}$ is nonempty. Show that the conditional distribution of $\|\boldsymbol{Y}\|^2$ given $\boldsymbol{BY} \geqslant \boldsymbol{0}$ is χ^2_n.

3.14 A function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be Lipchitz continuous if $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leqslant k\|\boldsymbol{x} - \boldsymbol{y}\|$ for some $k > 0$. Let $\mathcal{A} \subset \mathbb{R}^p$ and $d_{\mathcal{A}}$ be the distance function $d_{\mathcal{A}} : \mathbb{R}^p \to \mathbb{R}$ defined by $d_{\mathcal{A}}(\boldsymbol{x}) = \|\boldsymbol{x} - \mathcal{A}\|$. Show that $d_{\mathcal{A}}$ is Lipchitz continuous with $k = 1$.

3.15 Let $\boldsymbol{A}$ be an $m \times n$ matrix and $rank(\boldsymbol{A}) = m$. Let $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{A}^T(\boldsymbol{AA}^T)^{-1}\boldsymbol{A}$. Then, show that $\boldsymbol{P}$ is the projection matrix onto the null space of $\boldsymbol{A}$.

3.16 Let $\boldsymbol{V}$ be a positive definite matrix, and let $\boldsymbol{V}^{-1} = \boldsymbol{P}^T\boldsymbol{D}^{-1}\boldsymbol{P}$ for some $\boldsymbol{D}$ and $\boldsymbol{P}$, where $\boldsymbol{P}$ is orthogonal. Consider the problem of testing $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ based on a single observation of $\boldsymbol{X}$ where $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$. Show that the LRT of H_0 against H_1 based on a single observation of $\boldsymbol{X}$ is equivalent to testing $H_0^* : \boldsymbol{\alpha} = \boldsymbol{0}$ against $H_1^*; \boldsymbol{\alpha} \in \mathcal{P}$ based on a single observation of $\boldsymbol{Y}$, where $\boldsymbol{Y} = \boldsymbol{PX}$ and $\mathcal{P} = \boldsymbol{P}\mathcal{C} = \{\boldsymbol{Px} : \boldsymbol{x} \in \mathcal{C}\}$ which is also a closed convex cone. Deduce that $\bar{\chi}^2(\boldsymbol{V}, \mathcal{C}) \sim \bar{\chi}^2(\boldsymbol{D}, \boldsymbol{P}\mathcal{C})$.

Similarly, let $\boldsymbol{V}^{-1} = \boldsymbol{A}^T\boldsymbol{A}$; thus $\boldsymbol{A}$ can be the Cholesky factor or $\boldsymbol{D}^{-1/2}\boldsymbol{P}$. Let $\boldsymbol{Z} = \boldsymbol{AX}$ and $\mathcal{Q} = \boldsymbol{A}\mathcal{C} = \{\boldsymbol{Ax} : \boldsymbol{x} \in \mathcal{C}\}$. Show that $\mathcal{Q}$ is also a closed convex cone, the testing problem is equivalent to test of $H_0' : \boldsymbol{\gamma} = \boldsymbol{0}$ vs $H_1' : \boldsymbol{\gamma} \in \mathcal{Q}$ based on a single observation of $\boldsymbol{Z} \sim N(\boldsymbol{\gamma}, \boldsymbol{I})$ and $\bar{\chi}^2(\boldsymbol{V}, \mathcal{C}) \sim \bar{\chi}^2(\boldsymbol{I}, \mathcal{Q})$.

3.17 *The $\bar{\chi}^2(\boldsymbol{V}, \mathcal{C})$ distribution and the weights depend on $\boldsymbol{V}$ only through the correlation matrix corresponding to $\boldsymbol{V}$:* Let $\boldsymbol{V} = (v_{ij})$ be a positive definite matrix of order $p \times p$. Consider the problem of testing $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ based on a single observation of $\boldsymbol{X}$ where $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$. Let Ω denote the correlation matrix corresponding to $\boldsymbol{V}$. Let $\boldsymbol{A}$ be the diagonal matrix $diag(a_1, \ldots, a_p)$ where $a_i = v_{ii}^{1/2}$. Verify that $\boldsymbol{V} = \boldsymbol{A\Omega A}$. Show that $\boldsymbol{A}\mathcal{C} = \mathcal{C}$. Deduce that $\bar{\chi}^2(\boldsymbol{V}, \mathcal{C}) \sim \bar{\chi}^2(\Omega, \mathcal{C})$ and hence the $\bar{\chi}^2$-weights depend on $\boldsymbol{V}$ only through its correlation matrix.

3.18 Let $\boldsymbol{R}$ be a $k \times p$ matrix of rank $k, k \leq p$, and $\boldsymbol{y} = \boldsymbol{Ra}$. Verify that

$$\min\{(\boldsymbol{x}-\boldsymbol{a})^T\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{a}) : \boldsymbol{Ra} \geqslant \boldsymbol{0}\} = \min\{(\boldsymbol{y}-\boldsymbol{b})^T(\boldsymbol{RVR}^T)^{-1}(\boldsymbol{y}-\boldsymbol{b}) : \boldsymbol{b} \geqslant \boldsymbol{0}\}.$$

Deduce that

$$\bar{\chi}^2(\boldsymbol{V}, \mathcal{C}) \sim \bar{\chi}^2(\boldsymbol{RVR}^T, \mathbb{R}^{+k})$$

where $\mathcal{C} = \{\boldsymbol{\theta} : \boldsymbol{R\theta} \geqslant \boldsymbol{0}\}$.
[Hint: Find a matrix $\boldsymbol{Q}$ such that $\boldsymbol{RVQ}^T = \boldsymbol{0}$ and $\boldsymbol{C}^T = [\boldsymbol{R}^T, \boldsymbol{Q}^T]$ is non-singular; then show that $\min\{(\boldsymbol{x} - \boldsymbol{a})^T\boldsymbol{V}^{-1}(\boldsymbol{x} - \boldsymbol{a}) : \boldsymbol{Ra} \geqslant \boldsymbol{0}\} = \min\{(\boldsymbol{z} - \boldsymbol{c})^T(\boldsymbol{CVC}^T)^{-1}(\boldsymbol{z} - \boldsymbol{c}) : \boldsymbol{c} \geqslant \boldsymbol{0}\}$. Now use the fact that $\boldsymbol{CVC}^T$ is block diagonal.]

Section 3.8

3.19 *Derivation of the null distribution of LRT for a Type B problem in two dimensions:* Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}), H_1 : \theta_1 \geqslant 0, \theta_2 - m\theta_1 \geqslant 0$ and H_2 : *not* H_1, where m is known. Verify the following: (a) $LRT = \min\{(\boldsymbol{X}-\boldsymbol{a})^T(\boldsymbol{X}-\boldsymbol{a}) : a_1 \geqslant 0, a_2 - ma_1 \geqslant 0\}$. (b) $\text{pr}(LRT \leq c|\boldsymbol{\theta}=\boldsymbol{0}) = (0.5-q)\text{pr}(\chi_0^2 \leq c) + 0.5\text{pr}(\chi_1^2 \leq c) + q\text{pr}(\chi_2^2 \leq c)$, where $q = (2\pi)^{-1}\cos^{-1}(-m(1+m^2)^{-1/2})$.

3.20 Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I})$, where $\boldsymbol{X}$ is 2×1, $H_0 : \boldsymbol{\theta} \in \mathcal{C}$ and $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$. Let $\mathcal{C}^o$ denote the polar cone of $\mathcal{C}$ with respect to $\boldsymbol{I}$. Let $\mathcal{S}_1, \ldots, \mathcal{S}_4$ denote the four cones as in Example 3 in Section 3.3. (a) Show that, conditional on $\boldsymbol{X} \in \mathcal{S}_i$, $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{X} - \tilde{\boldsymbol{\theta}}$ are independent, $i = 1, \ldots, 4$. (b) Let L_A and L_B denote the LRT for the Type A testing problem, $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta} \geqslant \boldsymbol{0}$, and for the Type B testing problem, $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$, respectively. Show that the joint distribution of L_A and L_B at $\boldsymbol{\theta} = \boldsymbol{0}$ is given by

$$\text{pr}(L_A \geqslant c_1, L_B \geqslant c_2|\boldsymbol{\theta}=\boldsymbol{0}) = (0.5-q)\text{pr}(\chi_0^2 \geqslant c_1)\text{pr}(\chi_2^2 \geqslant c_2) + 0.5\text{pr}(\chi_1^2 \geqslant c_1)\text{pr}(\chi_1^2 \geqslant c_2) + q\text{pr}(\chi_2^2 \geqslant c_1)\text{pr}(\chi_0^2 \geqslant c_2).$$

3.21

3.21.1. Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$. Let $\tilde{\boldsymbol{\theta}}$ denote the *mle* subject to $\boldsymbol{\theta} \in \mathcal{C}$, L_A and L_B denote the *LRT* s for $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ and $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ against $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$, respectively. Then the joint distribution of L_A and L_B at $\boldsymbol{\theta} = \boldsymbol{0}$ is given by the following:

$$\text{pr}(L_A \leq c_1, L_B \leq c_2|\boldsymbol{\theta}=\boldsymbol{0}) = \sum_{i=0}^{p} w_i(p, \boldsymbol{V}, \mathcal{C})\text{pr}(\chi_i^2 \leq c_1)\text{pr}(\chi_{p-i}^2 \leq c_2).$$

3.21.2. Let the hypotheses H_0, H_1, and H_2 be defined as $H_0 : \boldsymbol{\theta} \in \mathcal{M}$, $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, and $H_2 : \boldsymbol{\theta} \in \mathcal{L}$ where $\mathcal{M}$ and $\mathcal{L}$ are linear spaces, $\mathcal{C}$ is a closed convex cone, $\mathcal{M} \subset \mathcal{C} \subset \mathcal{L}$, and $dim(\mathcal{L}) = k \leq p$. Obtain an expression for $\text{pr}(L_A \leq c_1, L_B \leq c_2)$.

3.21.3. Suppose that $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \sigma^2\boldsymbol{U})$ where σ is unknown and $\boldsymbol{U}$ is known. Using similar arguments obtain expressions for $\text{pr}_0(\bar{E}_A^2 \geqslant c, \bar{E}_B^2 \geqslant d)$ and $\text{pr}_0(\bar{F}_A \geqslant c, \bar{F}_B \geqslant d)$.

3.21.4. Consider the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{E}$ where $\boldsymbol{E} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{U})$, σ is unknown and $\boldsymbol{U}$ is known. Using similar arguments obtain expressions for $\text{pr}_0(\bar{E}_A^2 \geqslant c, \bar{E}_B^2 \geqslant d)$ and $\text{pr}_0(\bar{F}_A \geqslant c, \bar{F}_B \geqslant d)$.
[Hint:

$$\begin{aligned}&\text{pr}_0(L_A \leq c_1, L_B \leq c_2)\\ &\quad = \sum \text{pr}_0(L_A \leq c_1, L_B \leq c_2 \mid \boldsymbol{X} \in \mathcal{S}_i)\text{pr}(\boldsymbol{X} \in \mathcal{S}_i)\\ &\quad = \sum \text{pr}_0(\|\boldsymbol{P}_i\boldsymbol{X}\|^2 \leq c_1, \|(\boldsymbol{I}-\boldsymbol{P})_i\boldsymbol{X}\|^2 \leq c_2 \mid \boldsymbol{X} \in \mathcal{S}_i)\text{pr}(\boldsymbol{X} \in \mathcal{S}_i).\end{aligned}$$

Now use Lemma 3.13.6.]

Section 3.9

3.22 For Example 1.2.2 in Chapter 1 on *El Niño*, recall that the hypothesis of interest is that

the warm phase of El Niño tends to suppress hurricanes whereas cold phase encourages hurricanes.

(a) Test the foregoing hypothesis as a null hypothesis.
(b) Test the same hypothesis as the alternative.

3.23 *Connections among $\bar{E}^2, \bar{F}$ and LRT:* Let $\boldsymbol{X} = (X_1, X_2)^T$ and assume that $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \sigma^2\boldsymbol{U})$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n independently and identically distributed observations on $\boldsymbol{X}$. Show that this set up is a special case of the linear model in 3.40. Let the null and alternative hypotheses be $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ and $H_1 : \boldsymbol{\theta} \geqslant \boldsymbol{0}$. Let $Q_1(\boldsymbol{\theta}) = \sum(\boldsymbol{X}_i - \boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\boldsymbol{X}_i - \boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta}) = n(\bar{\boldsymbol{X}} - \boldsymbol{\theta})^T\boldsymbol{U}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\theta})$. Show that $Q_1(\boldsymbol{\theta}) = \sum(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T\boldsymbol{U}^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) + q_1(\boldsymbol{\theta})$ and that the loglikelihood $L(\boldsymbol{\theta}, \sigma^2)$ takes the form $L(\boldsymbol{\theta}, \sigma^2) = -(1/2)\sigma^{-2}Q_1(\boldsymbol{\theta}) - \frac{1}{2}np\log(\sigma^2) + Const.$ Let the null and alternative hypotheses be $H_a : \boldsymbol{\theta} \in \mathcal{C}_a$ and $H_b : \boldsymbol{\theta} \in \mathcal{C}_b$, respectively, where $\mathcal{C}_a \subset \mathcal{C}_b$. Let $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$ be the points at which $Q_1(\boldsymbol{\theta})$ is minimized over $\mathcal{C}_a$ and $\mathcal{C}_b$, respectively. Show that, the likelihood ratio statistic for testing H_a against H_b is given by $LRT = np\log\{Q_1(\boldsymbol{\theta}^a)/Q_1(\boldsymbol{\theta}^b)\}$. Let $\bar{E}^2 = \{Q_1(\boldsymbol{\theta}^a) - Q_1(\boldsymbol{\theta}^b)\}/Q_1(\boldsymbol{\theta}^a)$ and $\bar{F} = \{Q_1(\boldsymbol{\theta}^a) - Q_1(\boldsymbol{\theta}^b)\}/S^2$, where $S^2 = \nu^{-1}\sum(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T\boldsymbol{U}^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$ and $\nu = 2(n-1)$. Show that $\bar{E}^2 = [1 - \exp\{-LRT/(np)\}]$. Deduce that the $\bar{E}^2$-test that rejects H_a for large values of $\bar{E}^2$, is equivalent to the likelihood ratio test.
(a) Now let the null and alternative hypotheses be $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ and $H_1 : \boldsymbol{\theta} \geqslant \boldsymbol{0}$. Let the four quadrants $Q_1, \ldots, Q_4$ of $\mathbb{R}^2$ be defined as in Examples 1 and 2 of Section 3.3. Show that $\bar{X}_1, \bar{X}_2, \sum\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}\|^2$ are mutually independent. By arguments similar to those for (3.7) show that

$$\bar{E}_A^2 = \begin{cases} T/\{T + \sum\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}\|^2\} & \sim \beta(\frac{2}{2}, \frac{n-2}{2}) \quad \text{given } \bar{\boldsymbol{X}} \in Q_1 \\ (n\bar{X}_1^2)/\{T + \sum\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}\|^2\} & \sim \beta(\frac{1}{2}, \frac{n-1}{2}) \quad \text{given } \bar{\boldsymbol{X}} \in Q_4 \\ (n\bar{X}_2^2)/\{T + \sum\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}\|^2\} & \sim \beta(\frac{1}{2}, \frac{n-1}{2}) \quad \text{given } \bar{\boldsymbol{X}} \in Q_2 \\ 0/\{0 + T + \sum\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}\|^2\} & \sim \beta(\frac{0}{2}, \frac{n-0}{2}) \quad \text{given } \bar{\boldsymbol{X}} \in Q_3 \end{cases}$$

where $T = n\bar{X}_1^2 + n\bar{X}_2^2$.

By applying $\text{pr}(\bar{E}_A^2 \leq c|H_0) = \sum_1^4 \text{pr}(\bar{E}_A^2 \leq c|\bar{\boldsymbol{X}} \in Q_i)\text{pr}(\bar{\boldsymbol{X}} \in Q_i)$ or otherwise, deduce that $\text{pr}(\bar{E}_A^2 \leq c|H_0)$ is equal to

$$\frac{1}{4}\text{pr}\{\beta(\frac{0}{2}, \frac{n-0}{2}) \leq c\} + \frac{1}{2}\text{pr}\{\beta(\frac{1}{2}, \frac{n-1}{2}) \leq c) + \frac{1}{4}\text{pr}\{\beta(\frac{2}{2}, \frac{n-2}{2}) \leq c\}.$$

(b) Show that

$$\bar{F}_A = \begin{cases} (n\bar{X}_1^2 + n\bar{X}_2^2)/S^2 & \sim 2F_{2,\nu} \quad \text{conditional on } \bar{\boldsymbol{X}} \in Q_1 \\ n\bar{X}_1^2/S^2 & \sim 1F_{1,\nu} \quad \text{conditional on } \bar{\boldsymbol{X}} \in Q_4 \\ n\bar{X}_2^2/S^2 & \sim 1F_{1,\nu} \quad \text{conditional on } \bar{\boldsymbol{X}} \in Q_2 \\ 0/S^2 & \sim 0F_{0,\nu} \quad \text{conditional on } \bar{\boldsymbol{X}} \in Q_3. \end{cases}$$

Show that, under the null hypothesis,

$$\text{pr}(\bar{F}_A \leq c) = (1/4)\text{pr}(2F_{2,\nu} \leq c) + (1/2)\text{pr}(1F_{1,\nu} \leq c) + (1/4)\text{pr}(0F_{0,\nu} \leq c).\blacksquare$$

3.24 Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \sigma^2 \boldsymbol{U})$, where σ is unknown. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independently and identically distributed as $\boldsymbol{X}$. Let the null and alternative hypotheses be $H_1 : \boldsymbol{R}_1\boldsymbol{\theta} \geqslant \boldsymbol{0}, \boldsymbol{R}_2\boldsymbol{\theta} = \boldsymbol{0}$ and $H_2 : \boldsymbol{\theta}$ *is not restricted*, where $\boldsymbol{R}_1^T$ is $s \times m$ and $\boldsymbol{R}_2^T$ is $t \times m$, and $[\boldsymbol{R}_1^T, \boldsymbol{R}_2^T]$ has full rank. Show that the least favorable null value of $\bar{E}^2$ and of $\bar{F}$ is $\boldsymbol{\theta} = \boldsymbol{0}$, and their least favorable null distributions are

$$\mathrm{pr}_0(\bar{E}^2 \leq c | \boldsymbol{\theta} = \boldsymbol{0}) = \textstyle\sum_{i=0}^{s} w_{s-i}(s, \boldsymbol{A}) \mathrm{pr}[\beta\{\frac{1}{2}(t+i), \frac{1}{2}(np-p)\}\} \leq c].$$

$$\mathrm{pr}_0(\bar{F} \leq c | \boldsymbol{\theta} = \boldsymbol{0}) = \textstyle\sum_{i=0}^{s} w_{s-i}(s, \boldsymbol{A}) \mathrm{pr}\{(t+i)F_{t+i,\nu} \leq c\}$$

where $\boldsymbol{A} = \boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_1^T - (\boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_2^T)(\boldsymbol{R}_2\boldsymbol{V}\boldsymbol{R}_2^T)^{-1}(\boldsymbol{R}_2\boldsymbol{V}\boldsymbol{R}_1^T)$.

3.25 Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independently and identically distributed as $\boldsymbol{X}$ where $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \sigma^2 \boldsymbol{U})$. Let the null and alternative hypotheses be $H_0 : \boldsymbol{R}\boldsymbol{\theta} = \boldsymbol{0}$ and $H_1 : \boldsymbol{R}_1\boldsymbol{\theta} \geqslant \boldsymbol{0}$, where $\boldsymbol{R}$ is a full row-rank matrix of order $r \times p$ and $\boldsymbol{R}_1$ is a submatrix of $\boldsymbol{R}$ of order $q \times p$. Show that $\bar{E}^2$ and $\bar{F}$ tests are similar and that their null distributions are given by

$$\mathrm{pr}(\bar{E}^2 \leq c | H_0) = \textstyle\sum_{j=0}^{q} w_j(q, \boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_1^T) \mathrm{pr}\{\beta(2^{-1}(r-q+j), 2^{-1}(np-p+q-j) \leq c\},$$

$$\mathrm{pr}(\bar{F} \leq c) = \textstyle\sum_{j=0}^{q} w_j(q, \boldsymbol{R}_1\boldsymbol{V}\boldsymbol{R}_1^T) \mathrm{pr}\{(r-q+j)F_{r-q+j,\nu} \leq c\}.$$

Section 3.10

3.26 For the hypertensive patients in the sodium reduction example (see Example 1.2.17 in Chapter 1), obtain an overall p-value for testing the hypothesis that a reduction in sodium intake leads to a reduction in blood pressure. State the assumptions that you made and discuss their limitations.

3.27 Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V}$ is unknown. Let $\boldsymbol{\theta}_2$ and $\boldsymbol{S}$ be as in Section 3.10. Consider the test of $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta}_2 \in \mathcal{P}$, where $\mathcal{P}$ is a one-sided closed cone. Show that the *LRT* takes the simpler form (see Perlman (1969), equation (6.13))

$$LRT = \{\|\boldsymbol{Y}\|_{\boldsymbol{T}}^2 - \|\boldsymbol{Y} - \Pi_{\boldsymbol{T}}(\boldsymbol{Y}|\mathcal{P})\|_{\boldsymbol{T}}^2\}\{1 + \|\boldsymbol{Y} - \Pi_{\boldsymbol{T}}(\boldsymbol{Y}|\mathcal{P})\|_{\boldsymbol{T}}^2\}^{-1}$$

where $\boldsymbol{Y} = \bar{\boldsymbol{X}}_{22}$ and $\boldsymbol{T} = \boldsymbol{S}_{22}$.

3.28 Let $\boldsymbol{A}$ be the $k \times (k-1)$ matrix such that $\boldsymbol{A}\boldsymbol{\mu} \geqslant \boldsymbol{0}$ is equivalent to $\mu_1 \leqslant \ldots \leqslant \mu_k$. Let $\boldsymbol{y}^*$ be the point at which $(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\mu})$ reaches its minimum over $\boldsymbol{A}\boldsymbol{\mu} \geqslant \boldsymbol{0}$, where $\boldsymbol{W}$ is a diagonal matrix with positive diagonal elements. Let $\mathcal{Q} = \{\boldsymbol{\mu} : \boldsymbol{A}\boldsymbol{\mu} \geqslant \boldsymbol{0}\}$. Let $\mathcal{M}$ be the linear space spanned by $(1, \ldots, 1)$. Deduce the following [Hint: Use the results relating to Fig. 3.10, in particular those in Section 3.12.2. Proofs of these from first principles are provided in Barlow et al. (1972); see also Robertson et al. (1988)].

1. $\mathcal{M}$ is the largest linear space contained in $\mathcal{Q}$.
2. $(\boldsymbol{y} - \boldsymbol{y}^*) \perp_{\boldsymbol{W}^{-1}} \mathcal{M}$.

3. $(\boldsymbol{y} - \boldsymbol{y}^*)\boldsymbol{W}\boldsymbol{1} = 0$. (Hint: $\boldsymbol{1} \in \mathcal{M}$).
4. $(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\mu}) \geqslant (\boldsymbol{y} - \boldsymbol{y}^*)^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{y}^*) + (\boldsymbol{y}^* - \boldsymbol{\mu})^T\boldsymbol{W}(\boldsymbol{y}^* - \boldsymbol{\mu})$ for every $\boldsymbol{\mu} \in \mathcal{Q}$.
5. $(\boldsymbol{y} - \boldsymbol{y}^*)^T\boldsymbol{W}\boldsymbol{y}^* = 0$.
6. $(\boldsymbol{y} - \boldsymbol{y}^*)^T\boldsymbol{W}\boldsymbol{\mu} \leqslant 0$ for every $\boldsymbol{\mu} \in \mathcal{Q}$.
7. If $\tilde{\boldsymbol{y}} \in \mathcal{Q}$ and $(\boldsymbol{y} - \tilde{\boldsymbol{y}})^T\boldsymbol{W}(\tilde{\boldsymbol{y}} - \boldsymbol{\mu}) \geqslant 0$ for every $\boldsymbol{\mu} \in \mathcal{Q}$ then $\tilde{\boldsymbol{y}} = \boldsymbol{y}^*$.
8. Let $\tilde{\boldsymbol{y}} \in \mathcal{Q}$. Then $\tilde{\boldsymbol{y}} = \boldsymbol{y}^*$ if and only if $(\boldsymbol{y} - \tilde{\boldsymbol{y}})^T\boldsymbol{W}\tilde{\boldsymbol{y}} = 0$ and $(\boldsymbol{y} - \tilde{\boldsymbol{y}})^T\boldsymbol{W}\boldsymbol{\mu} \leqslant 0$ for every $\boldsymbol{\mu} \in \mathcal{Q}$.

3.29 Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}$ where $\boldsymbol{\theta} \in \mathbb{R}^p$. Suppose that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to $N(\boldsymbol{0}, \boldsymbol{V})$ where $\boldsymbol{V}$ is positive definite. Let $\hat{\boldsymbol{V}}$ be a consistent estimator of $\boldsymbol{V}$ (i.e., $\hat{\boldsymbol{V}} \xrightarrow{p} \boldsymbol{V}$), $H_0 : \boldsymbol{\theta} \in \mathcal{M}$, $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, and $H_2 : \boldsymbol{\theta} \notin \mathcal{C}$. $L_A = n\{\|\hat{\boldsymbol{\theta}} - \mathcal{M}\|^2 - \|\hat{\boldsymbol{\theta}} - \mathcal{C}\|^2\}$ and $L_B = n\|\hat{\boldsymbol{\theta}} - \mathcal{C}\|^2$, where $\|\cdot\|$ is defined with respect to $\langle\cdot\rangle_{\hat{\boldsymbol{V}}}$. Following the notation introduced in this chapter, it is implicitly assumed that $\mathcal{C}$ is a closed convex cone, $\mathcal{M}$ is a linear space, and $\mathcal{M} \subset \mathcal{C}$. Show that L_A is consistent at any $\boldsymbol{\theta} \notin (\mathcal{C} \cap \mathcal{M}^{\perp})^o$ and that L_B is consistent at any $\boldsymbol{\theta} \notin \mathcal{C}$.

Now assume that $\mathcal{C}$ is a polyhedral. Explain why the limiting probability of rejecting the null hypothesis may depend on the particular face of the polyhedral on which the true value lies.

3.30

3.30.1. Let $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ and $H_1 : \boldsymbol{\theta} \in \mathcal{C}$ be the null and alternative hypotheses where $\boldsymbol{\theta} \in \mathbb{R}^p$. Assume that $\boldsymbol{X} = \boldsymbol{\theta} + \boldsymbol{E}$, where the distribution of $\boldsymbol{E}$ does not depend on any unknown parameters. Let $\boldsymbol{V}$ be a positive definite matrix, the inner product be defined with respect to $\langle\cdot\rangle_{\boldsymbol{V}}$, and let

$$L_A = \|\boldsymbol{X}\|^2 - \|\boldsymbol{X} - \mathcal{C}\|^2.$$

Then prove the following: (a) $L_A = \|\boldsymbol{X} - \mathcal{C}^o\|^2 = \|\boldsymbol{E} - \mathcal{P}_{\boldsymbol{\theta}}\|^2$, where $\mathcal{P}_{\boldsymbol{\theta}} = \mathcal{C}^o - \boldsymbol{\theta}$. (b) $\|\boldsymbol{\theta} - \mathcal{C}^o\| = \|\boldsymbol{0} - \mathcal{P}_{\boldsymbol{\theta}}\|$; in other words, the distance between $\boldsymbol{\theta}$ and $\mathcal{C}^o$ is equal to the distance between $\boldsymbol{0}$ and $\mathcal{P}_{\boldsymbol{\theta}}$. (c) For any given $K > 0$, if the distance between $\boldsymbol{\theta}$ and $\mathcal{C}^o$ is more than K then $B_K(\boldsymbol{0}) \cap \mathcal{P}_{\boldsymbol{\theta}}$ is the empty set, where $B_K(\boldsymbol{0})$ is the ball of radius K centered at $\boldsymbol{0}$. (d) $L_A \xrightarrow{p} \infty$ as $\|\boldsymbol{\theta} - \mathcal{C}^o\| \to \infty$. (e) For the test that rejects the H_0 for large values of L_A, show that the power tends to 1 as $\|\boldsymbol{\theta} - \mathcal{C}^o\| \to \infty$.

3.30.2. Let $\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{E}, \boldsymbol{V}$ and the inner product be as in the previous part. Let $H_0 : \boldsymbol{\theta} \in \mathcal{M}$ and $H_1 : \boldsymbol{\theta} \in \mathcal{C}$, and let

$$L_A = \|\boldsymbol{X} - \mathcal{M}\|^2 - \|\boldsymbol{X} - \mathcal{C}\|^2;$$

$\mathcal{M}$ is a linear space, $\mathcal{C}$ is a closed convex cone, and $\mathcal{M} \subset \mathcal{C}$. Consider the test that rejects the H_0 for large values of L_A. Prove that

$$\text{the power of } L_A \text{ tends to 1 as } \|\boldsymbol{\theta} - (\mathcal{C} \cap \mathcal{M}^{\perp})^o\| \to \infty.$$

[Hint: Recall that $L_A = \|X - (\mathcal{C} \cap \mathcal{M}^{\perp})^o\|^2$ and hence L_A has the same numerical value as the L_A in the previous part with $\mathcal{C}$ replaced by $\mathcal{C} \cap \mathcal{M}^{\perp}$. Now apply the previous part.]

3.30.3. Let X, θ, E, V and the inner product be as in the previous part. Let $H_1 : \theta \in \mathcal{C}$ and $H_2 : \theta \notin \mathcal{C}$, and let $L_B = \|X - \mathcal{C}\|^2$. Show that $L_B \xrightarrow{p} \infty$ and that the power of the test that rejects H_1 for large values of L_B tends to 1 as $\|\theta - \mathcal{C}\| \to \infty$. [Hint: Apply a method similar to that for the first part].

3.31 Let A be a nonsingular matrix of order $p \times p$, $x \in \mathbb{R}^p$ and $\mathcal{P}$ be a cone. Then prove the following:

1. $A\Pi(x|\mathcal{P}) = \Pi_{AVA^T}(Ax|A\mathcal{P})$.
2. $\|\Pi_V(x|\mathcal{P})\|_V = \|\Pi_{AVA^T}(Ax|A\mathcal{P})\|_{AVA^T}$.
3. $\|x - \Pi_V(x|\mathcal{P})\|_V = \|Ax - \Pi_{AVA^T}(Ax|A\mathcal{P})\|_{AVA^T}$.
4. Let $\mathcal{P}_1 \supset \mathcal{P}_2 \ldots \supset$ be s sequence of decreasing cones in $\mathbb{R}^p$. Then $\|x - \Pi_V(x|\mathcal{P}_n)\|_V$ is a nondecreasing sequence and converges to $\|x - \Pi_V(x|\mathcal{P})\|_V$ where $\mathcal{P} = \lim \mathcal{P}_n$; a similar result holds for increasing sequences.

[Perlman (1969)]

3.32 (a) Let $x, y \in \mathbb{R}^p$, $\mathcal{C}$ be a closed convex cone, V be a positive definite matrix of order $p \times p$, Π denote Π_V, $\|\cdot\|$ denote $\|\cdot\|_V$, and $\mathcal{C}^o$ denote the polar cone of $\mathcal{C}$ with respect to $\langle\cdot\rangle_V$. Show that [see Lemma 2.2 in Mukerjee et al. (1986)]

$$\Pi(x + y \mid \mathcal{C}) \leq \|\Pi(x \mid \mathcal{C}) + \Pi(y \mid \mathcal{C})\| \leq \|\Pi(x \mid \mathcal{C})\| + \|\Pi(y \mid \mathcal{C})\|.$$

(b) $\|\Pi(y + x \mid \mathcal{C})\| \leq \|\Pi(y \mid \mathcal{C})\|, \quad x \in \mathcal{C}^o$.
(c) $\|y - x\|^2 \geq \|\Pi(y \mid \mathcal{C}) - \Pi(x \mid \mathcal{C})\|^2, \quad x, y \in \mathbb{R}^p$.

3.33 Let $Y = (Y_1, \ldots, Y_n)^T$ where $Y_1, \ldots, Y_n$ are independent standard normal random variables. Let $\mathcal{L}$ be a linear subspace of $\mathbb{R}^n$, $dim(\mathcal{L}) = r$, A be an $m \times n$ matrix such that each row of A is orthogonal to $\mathcal{L}$, and $\tilde{Y}$ denote the projection of Y onto $\mathcal{L}$. Assume that $\{y : Ay \geqslant 0\}$ is nonempty. Then, show that the conditional distribution of $\|\tilde{Y}\|^2$, given $AY \geqslant 0$, is χ^2_r. [Meyer (2003b)]
[Hint: Use Lemma 3.13.3. Let P be the symmetric projection matrix onto $\mathcal{L}$. Then $\mathcal{L} = \{Py : y \in \mathbb{R}^n\}$. Verify that if a_i^T is the i^{th} row of A then $a_i = Py_i$ for some y_i $(i = 1, \ldots, m)$. Deduce that $A = BP$ where $B^T = [y_1, \ldots, y_m]$, and hence $AY \geq 0$ is equivalent to $PY \in \mathcal{C} = \{x : Bx \geqslant 0\}$. For a different proof, see Meyer (2003b)].

3.34 Let $Y = (Y_1, \ldots, Y_n)^T$ where $Y_1, \ldots, Y_n$ are independent standard normal random variables. Let $\mathcal{L}$ be a linear subspace of $\mathbb{R}^n$, $dim(\mathcal{L}) = r$. Let A be an $m \times n$ matrix such that each row of A is a vector in $\mathcal{L}$. Let $\tilde{Y}$ denote the projection of Y onto $\mathcal{L}$. Assume that $\{Ay \geqslant 0\}$ is non-empty. Then, show that the conditional distribution of $\|\tilde{Y}\|^2$, given $AY \geqslant 0$, is χ^2_r. [Meyer (2003b)]

[Hint: Use Lemma 3.13.3. Let P be a projection matrix onto $\mathcal{L}$. Let the i^{th} row of A be a_i^T. Then $a_i = (I - P)y_i$ for some y_i $(i = 1, \ldots, m)$. Deduce that $A = Y_0^T(I - P)$, where $Y_0^T = [y_1, \ldots, y_m]$, and hence $AZ \geqslant 0$ is equivalent to $(I - P)Z \in \mathcal{P}$ where $\mathcal{P} = \{x : Y_0^T x \geqslant 0\}$. Now, use the fact that PZ and $(I - P)Z$ are independent to verify that $\text{pr}(\|PZ\|^2 \leqslant c \mid BZ \geqslant 0) = \text{pr}(\|PZ\|^2 \leqslant c \mid (I - P)Z \in \mathcal{P}) = \text{pr}(\|PZ\|^2 \leqslant c)$.]

4

Tests in General Parametric Models

4.1 INTRODUCTION

In Chapter 3, tests of hypotheses concerning the mean of a multivariate normal were considered. Often, the parameter of interest in a hypothesis is more than just the mean of a multivariate normal. Some examples of this type were discussed in Chapter 1. In this chapter, the relevant theory is developed in a general context. To illustrate the nature of the problem, consider Example 1.2.8. It describes the results of a study to establish that a new treatment is better than an old one when the response is an ordinal variable. In this example, the inference problem can be formulated as test of $g(\pi) = \mathbf{0}$ against $g(\pi) \geqslant \mathbf{0}$ where π is the vector consisting of all the $\{\pi_{ij}\}$ and π_{ij} is the probability that the response for Treatment i is the j^{th} ordinal category of the response variable, and g is a nonlinear function; in fact, every component of g is an odds ratio. This example does not fit into the setting in Chapter 3. The theory developed in this chapter is applicable to this and much more general settings. A sample of models to which the theory of this chapter is applicable include the following:

(1) nonlinear regression
(2) generalized linear models (logistic regression, log-linear models, etc.)
(3) proportional hazards model, Cox's regression model
(4) time series model (for example, ARCH models used in finance)
(5) mixed effects models (positive semi-definite covariance matrices)
(6) quasi-likelihood
(7) generalized estimating equations (GEE)

(8) generalized method of moments (GMM).

It is thus clear that the results of this chapter are applicable to a broad range of contexts.

Let $\boldsymbol{\theta}$ denote the parameter of a statistical model and Θ denote the parameter space. A common feature of most of the inference problems with no inequality constraints on $\boldsymbol{\theta}$ is that Θ is open and the likelihood function is smooth over Θ. This feature plays a crucial role in ensuring that (i) $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx N\{\mathbf{0}, \mathcal{I}(\boldsymbol{\theta})^{-1}\}$, and (ii) the asymptotic null distribution of the likelihood ratio statistic for testing $\boldsymbol{\theta}_a = \mathbf{0}$ against $\boldsymbol{\theta}_a \neq \mathbf{0}$ is a chi-square where $\boldsymbol{\theta}_a$ is a subvector of $\boldsymbol{\theta}$. These two results form the basis of statistical inference in these models when there are no inequality constraints on $\boldsymbol{\theta}$.

In this chapter, we will build on these and obtain corresponding results when there are inequality constraints on $\boldsymbol{\theta}$ or the parameter space for $\boldsymbol{\theta}$ is constrained so that the constrained space is not open (topologically). These results will form the basis of inference in this setting. Most of the details in this chapter are given for the case when the observations are independently and identically distributed. However, corresponding results hold for non-*iid* cases, which include regression type models, time series models, and stochastic processes. As a general guide, with $\hat{\boldsymbol{\theta}}$ denoting an unconstrained estimator of $\boldsymbol{\theta}$, if $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal and the asymptotic null distribution of the likelihood ratio statistic for testing $\boldsymbol{\theta}_a = \mathbf{0}$ against $\boldsymbol{\theta}_a \neq \mathbf{0}$ is a chi-square, where $\boldsymbol{\theta}_a$ is a subvector of $\boldsymbol{\theta}$, then it is very likely that most of the results in this chapter for likelihood ratio tests would be applicable for constrained statistical inference on $\boldsymbol{\theta}$ as well.

The plan of this chapter is as follows. Section 4.2 provides a brief discussion of regularity conditions relating to the likelihood function that would be relevant to the rest of the chapter. Section 4.3 considers statistical models, which are usually referred to as *regular models,* and develops the likelihood ratio, Wald and score-type tests of $\boldsymbol{R\theta} = \mathbf{0}$ against $\boldsymbol{R\theta} \geqslant \mathbf{0}$ where $\boldsymbol{R}$ is a given matrix of constants. These results are of particular importance because a large number of real-life examples with inequality constraints on $\boldsymbol{\theta}$ are of this type. In any case, a detailed study of the case when the constraints are linear is essential before we consider nonlinear constraints. Section 4.4 extends the results of the previous section to the case when the constraints are nonlinear in $\boldsymbol{\theta}$. In particular, tests of $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$ against $\boldsymbol{h}(\boldsymbol{\theta}) \geqslant \mathbf{0}$ where $\boldsymbol{h}$ is a vector function will be discussed; it will be seen that there are several tests that are asymptotically equivalent to the LRT. Subsection 4.4.2 provides a statistic for testing $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$ against $\boldsymbol{h}_2(\boldsymbol{\theta}) \geqslant \mathbf{0}$, where $\boldsymbol{h}(\boldsymbol{\theta})$ is a smooth vector function of $\boldsymbol{\theta}$ and $\boldsymbol{h}_2(\boldsymbol{\theta})$ is a subvector of $\boldsymbol{h}(\boldsymbol{\theta})$. The main requirement for this is that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ be asymptotically normal; thus $\hat{\boldsymbol{\theta}}$ may not be the *mle*. In Section 4.5, a brief overview of likelihood ratio, Wald and score tests is provided with particular reference to the results that would be required for developing a class of score tests against inequality constraints. In Section 4.6, the Rao's score and Neyman's C(α) tests are generalized to tests against inequality constraints. These are further extended to develop tests against inequality constraints when the model is estimated by solving a set of estimating equations; for example, the class of *Generalized Estimating Equation*(GEE) methods and the *Generalized Method of Moments* (GMM) that is widely used in economics, fall into this category.

The results in Sections 4.3–4.6 provide a good coverage of the standard results required for testing $\boldsymbol{R\theta} = \mathbf{0}$ against $\boldsymbol{R\theta} \geqslant \mathbf{0}$. These results are extended further in the next two sections where the null and alternative hypotheses are stated in the general form, $H_0 : \boldsymbol{\theta} \in \Theta_0$ and $H_1 : \boldsymbol{\theta} \in \Theta_1$, respectively, where $\Theta_0 \subset \Theta_1 \subset \Theta$. In this case, the asymptotic null distributions of test statistics such as the likelihood ratio/Wald/ score, depend on the local shapes of Θ_0 and Θ_1 at the assumed true value in the null parameter space. The relevant local shapes of Θ_0 and Θ_1 are characterized by cones that approximate them at $\boldsymbol{\theta}_0$; these are called *approximating cones*. These will be discussed in detail later in this chapter.

General results for testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$ are discussed in Section 4.8. A topic of particular interest in economics involves test of $\boldsymbol{h}(\boldsymbol{\theta}) \geqslant \mathbf{0}$ against $\boldsymbol{h}(\boldsymbol{\theta}) \ngeqslant \mathbf{0}$ where $\boldsymbol{h}$ is a given nonlinear function. An example of this type is test of the hypothesis that a certain function (for example, a production function or a cost function) is concave against that it is not concave. The general theory for this is discussed in Section 4.8. The asymptotic distribution of the constrained local and global *mle* are studied in the last section. Some of the proofs are relegated to the Appendix at the end of this chapter.

4.2 PRELIMINARIES

Let $Y_1, \ldots, Y_n$ be independently and identically distributed with common density function $f(y; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ and y can be univariate or multivariate. Let

$$\ell(\boldsymbol{\theta}) = \sum \log f(Y_i; \boldsymbol{\theta}), \qquad \boldsymbol{S}(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})\ell(\boldsymbol{\theta})$$
$$\text{and} \quad \mathcal{I}_{\boldsymbol{\theta}} = \mathcal{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{(\partial/\partial\boldsymbol{\theta}) \log f(Y; \boldsymbol{\theta})(\partial/\partial\boldsymbol{\theta}^T) \log f(Y; \boldsymbol{\theta})\} \tag{4.1}$$

denote the loglikelihood, score function, and the information matrix for one observation, respectively. Let $\boldsymbol{\theta}_0$ denote the *true value* and $\hat{\boldsymbol{\theta}}$ denote the global *mle* over Θ.

Let A be a subset of Θ containing the true value of $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_A$ denote the *mle* over A defined by

$$\hat{\boldsymbol{\theta}}_A = \arg\max_{\boldsymbol{\theta} \in A} \ell(\boldsymbol{\theta});$$

equivalently, $\ell(\hat{\boldsymbol{\theta}}_A) = \max_{\boldsymbol{\theta} \in A} \ell(\boldsymbol{\theta})$. Strictly speaking, it is better to define the *mle* over A by $\ell(\hat{\boldsymbol{\theta}}_A) = \sup_{\boldsymbol{\theta} \in A} \ell(\boldsymbol{\theta}) + o_p(1)$; if we do so, $\hat{\boldsymbol{\theta}}_A$ would be well defined even if $\sup_{\boldsymbol{\theta} \in A} \ell(\boldsymbol{\theta})$ is reached at a point on the boundary of A that is not in A. However, we shall not be concerned with this technical detail. Every result in this chapter requires $\hat{\boldsymbol{\theta}}_A$ to be consistent, where A may the natural parameter space Θ or the ones defined by null/alternative hypotheses. For the case when the true value of $\boldsymbol{\theta}$ is an interior point of A, consistency of the *mle* has been discussed in great detail in the literature; for example, see Lehmann (1983, section 6.3). For a discussion of the case when the true parameter may be on the boundary of the parameter space, see Self and Liang (1987, Theorem 1) and Andrews (1999, Section 3.1); van der Vaart and Wellner (1996, Section 3.2) consider the more general case, which includes M-estimation.

Consistency of $\hat{\theta}_A$ would follow from the following two, albeit strong, conditions (see Andrews (1999)): There exists a function $\ell_0(\theta)$ such that
(i) $\sup_{\theta \in A} |n^{-1}\ell(\theta) - \ell_0(\theta)| \xrightarrow{p} 0$, and
(ii) $\sup_{\theta \in A, \|\theta - \theta_0\| > \epsilon} \ell_0(\theta_0) < \ell_0(\theta_0), \quad \epsilon > 0.$
The first condition says that $n^{-1}\ell(\theta)$ converges uniformly over the parameter space, and the second says that θ_0 is the global maximum of $\ell_0(\theta)$ and it is separated from all the other maxima. The uniform law of large numbers of Andrews (1992) may be helpful in establishing the uniform convergence of $n^{-1}\ell(\theta)$. In any case, let us note that if the estimator is the global maximizer of a function, such as the likelihood, then constraints on θ do not generally introduce additional difficulties for establishing consistency. Throughout this chapter, we shall assume the following:

Assumption A1: *The mle, $\hat{\theta}_A$, is consistent where $\theta_0 \in A \subset \Theta$.* ■

Further, we shall also typically assume that a set of regularity conditions are satisfied to ensure that the asymptotic null distribution of the *LRT* of $\theta = \theta_0$ against $\theta \neq \theta_0$ is a chi-square, and $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically normal. If this were not true, it is quite unlikely that the results stated in this section about the asymptotic null distribution of LRT for test against inequality restrictions would hold. It is also of interest to note that such a set of regularity conditions is sufficient to derive the asymptotic distributions of LRT under inequality constraints. With this in mind, let us define the following set of regularity conditions:

Condition Q:

1. Distinct values of θ correspond to distinct distributions.
2. The first three partial derivatives of $\log f(y;\theta)$ with respect to θ exist almost everywhere.
3. There exists a $G(y)$ such that $\int G(y)dy < \infty$ and the absolute values of the first three partial derivatives of $\log f(y;\theta)$ with respect to θ are bounded by $G(y)$ in a neighborhood of θ_0.
4. The Fisher information matrix, $\mathcal{I}(\theta)$, is finite and positive definite. ■

Remark: In most cases, it would be possible to replace the condition involving the third derivative of $\log f(y;\theta)$ by one that involves only the second derivative. However, we shall not concern ourselves with this technical detail.

In what follows we shall assume that Condition Q is satisfied. Under this condition, we have the following standard results; for example, see Cox and Hinkley (1974), Sen and Singer (1993), Lehmann (1991), and Ferguson (1996).

Proposition 4.2.1 *Suppose that Condition Q is satisfied. Then we have the following:*

1. $n^{-1/2}S(\theta_0) \xrightarrow{d} N\{0, \mathcal{I}(\theta_0)\}$.
2. $n^{-1/2}\mathcal{I}(\theta_0)^{-1}S(\theta_0) = n^{1/2}(\hat{\theta} - \theta_0) + o_p(1)$.
3. $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\{0, \mathcal{I}(\theta_0)^{-1}\}$.

4. *The LRT for testing $R\theta = 0$ against $R\theta \neq 0$ is asymptotically χ^2_r under H_0 where $r = rank(R)$.* ■

This proposition states important first order results that are useful for inference when there are no inequality constraints on θ. Typically, these results are proved using a quadratic approximation of $\ell(\theta)$ and a linear approximation of $\nabla\ell(\theta)$ near θ_0. Since we will be making use of such quadratic approximations, they are stated in the next proposition.

Proposition 4.2.2 *Assume that Condition Q is satisfied. Let $u = \sqrt{n}(\theta - \theta_0)$ and $K > 0$ be given. Then we have the following:*

$$(A) \quad \ell(\theta) = \ell(\theta_0) + n^{-1/2}u^T S(\theta_0) - (1/2)u^T \mathcal{I}_{\theta_0} u + r_n(u), \tag{4.2}$$
where $\sup_{\|u\| < K} |r_n(u)| = o_p(1)$.

$$(B) \quad \ell(\theta) = \ell(\theta_0) + (1/2)n^{-1}S(\theta_0)^T \mathcal{I}_{\theta_0}^{-1} S(\theta_0) - (1/2)(Z_n - u)^T \mathcal{I}_{\theta_0}(Z_n - u) + \delta_n(u), \tag{4.3}$$
where $Z_n = n^{-1/2}\mathcal{I}_{\theta_0}^{-1}S(\theta_0)$ *and* $\sup_{\|u\| < K} |\delta_n(u)| = o_p(1)$.

$$(C) \quad \ell(\theta) = \ell(\hat{\theta}) - (1/2)(Z_n - u)^T \mathcal{I}_{\theta_0}(Z_n - u) + \epsilon_n(u), \tag{4.4}$$
where $Z_n = \sqrt{n}(\hat{\theta} - \theta_0)$, $\hat{\theta}$ *is* $\sqrt{n}$*-consistent, and* $\sup_{\|u\| < K} |\epsilon_n(u)| = o_p(1)$.

Further, $\delta_n(u) = n^{-1/2}\|u\|^3 O_p(1)$. ■

These three quadratic approximations are very important in constrained inference, just as in unconstrained inference. A detailed proof of the first approximation is given in Sen and Singer (1993, Section 5.3); the second approximation is used in Self and Liang (1987), and the third one is used in Silvapulle (1994). For a detailed account see Andrews (1999).

Practically all the asymptotic results of this chapter are based on the assumption that the *mle* is $\sqrt{n}$-consistent–i.e., $\sqrt{n}(\hat{\theta}_A - \theta_0) = O_p(1)$, where $\theta_0 \in A \subset \Theta$. Fortunately, $\sqrt{n}$-consistency of *mle* follows from the consistency of the *mle* and the quadratic approximations in the previous proposition; this is stated in the next result.

Lemma 4.2.3 *Suppose that Condition Q is satisfied and $\theta_0 \in A \subset \Theta$. Then, the consistency of $\hat{\theta}_A$ ensures that $\sqrt{n}(\hat{\theta}_A - \theta_0) = O_p(1)$.*

Proof: For the case when the observations are *iid*, see Lemma 1 in Chernoff (1954). For the non-*iid* case, see Andrews (1999). For a proof of a similar result, see page 141 in Pollard (1984); see also Lemma 4.3 in Geyer (1994). ■

The $\sqrt{n}$-consistency of $\hat{\theta}_A$ enables us to restrict our attention to neighborhoods of the form $\{\theta : \sqrt{n}\|\theta - \theta_0\| < K\}$, where K is arbitrary but fixed, for the purposes of evaluating the global maximum of $\ell(\theta)$. Once we restrict attention to such

a neighborhood, the remainder terms in Proposition 4.2.2 become negligible because they would converge to zero uniformly over the neighborhood. This is the approach that will be used throughout.

A result similar to Lemma 4.2.3 is also available for local *mle*. Except for the very last section of this chapter, the results in this chapter relate to test of hypotheses and hence local *mle* would not play a role. However, for completeness let us state the result; the following is Lemma 1 in Self and Liang (1987)

Proposition 4.2.4 *Suppose that Condition Q is satisfied and that, for some* $\delta_0 > 0$, $B = A \cap \{\theta : \|\theta - \theta_0\| \leq \delta\}$ *is a closed set for* $0 < \delta < \delta_0$. *Then, there exists a sequence of points* $\theta^\dagger$ *in* B *at which* $\ell(\theta)$ *has a local maximum and* $\theta^\dagger \xrightarrow{p} \theta_0$. *Further,* $\sqrt{n}(\theta^\dagger - \theta_0) = O_p(1)$. ■

4.3 TESTS OF $R\theta = 0$ AGAINST $R\theta \geqslant 0$

Let the setting be as in the previous section; in particular, $\theta \in \mathbb{R}^p$ denotes the parameter associated with a statistical model, Θ denotes the parameter space for θ, θ_0 denotes the true parameter, and $\{\ell(\theta), S(\theta), \mathcal{I}(\theta)\}$ be as in (4.1). Let $\boldsymbol{R}$ be a known full-rank matrix of constants of order $r \times p$ ($r \leq p$). In this section, we consider test of

$$H_0 : \boldsymbol{R\theta} = \mathbf{0} \quad \text{vs} \quad H_1 : \boldsymbol{R\theta} \geqslant \mathbf{0}. \tag{4.5}$$

In this section, we will assume that Θ *is open.* The results for the case when Θ is open also hold for many cases when it is not open, but the technical details are more involved. The main advantage of considering the case when Θ is open is that the classical arguments based on Taylor series, for example, as in Cox and Hinkley (1974) and Lehmann (1991), can also be used. For most of this section, *the observations will be assumed to be independently and identically distributed.*

For the time being, let us note that practically every result in this section for testing (4.5) also holds, with minor modifications, for testing

$$H_0^* : \theta \in \mathcal{M} \quad \text{vs} \quad H_1^* : \theta \in \mathcal{C} \tag{4.6}$$

as well where $\mathcal{M}$ is a linear space, $\mathcal{C}$ is a closed convex cone, and $\mathcal{M} \subset \mathcal{C}$. These results also have generalizations when the $\mathcal{C}$ in H_1^* is replaced by a set that may be neither concave nor conical; such generalizations will be discussed in a later section.

Notation: *Unless the contrary is clear, the following notation will be adopted:* $\hat{\theta}$ *is the unconstrained estimator over* Θ, *and* $\tilde{\theta}$ *and* $\bar{\theta}$ *denote the estimators under* H_1 *and* H_0, *respectively.*

4.3.1 Likelihood Ratio Test with *iid* Observations

For testing $H_0 : \boldsymbol{R\theta} = \mathbf{0}$ against $H_1 : \boldsymbol{R\theta} \geqslant \mathbf{0}$, we have

$$LRT = 2[\sup_{\boldsymbol{R\theta} \geqslant \mathbf{0}} \ell(\theta) - \sup_{\boldsymbol{R\theta} = \mathbf{0}} \ell(\theta)]; \tag{4.7}$$

its asymptotic null distribution is given in the next result.

Proposition 4.3.1 *Let* $Y_1, \ldots, Y_n$ *be independently and identically distributed with common density function* $f(y;\theta)$. *Assume that Condition Q is satisfied. Let* H_0 *and* H_1 *be as in* (4.5). *Suppose that* H_0 *is satisfied and let* θ_0 *denote the true value in* H_0. *Let the function* $D(z)$, $z \in \mathbb{R}^p$, *be defined by*

$$D(z) = \min_{\boldsymbol{R\theta}=\mathbf{0}} (z-\theta)^T \mathcal{I}_{\theta_0} (z-\theta) - \min_{\boldsymbol{R\theta} \geqslant \mathbf{0}} (z-\theta)^T \mathcal{I}_{\theta_0} (z-\theta). \tag{4.8}$$

Let $Z_n = n^{-1/2} \mathcal{I}_{\theta_0}^{-1} S(\theta_0)$ *and* $Z \sim N(\mathbf{0}, \mathcal{I}_{\theta_0}^{-1})$. *Then, we have the following:*

1. $LRT = D(Z_n) + o_p(1)$, *and hence* $LRT \xrightarrow{d} D(Z)$.
2. *The asymptotic null distribution of the LRT is the chi-bar square given by*

$$\lim_{n\to\infty} pr_{\theta_0}(LRT \geqslant c | H_0) = \sum_{i=0}^{r} w_i(r, R\mathcal{I}_{\theta_0}^{-1} R^T) pr(\chi_i^2 \geqslant c). \tag{4.9}$$

Proof: The large sample arguments used here are similar to those used for the case when there are no inequality constraints. Let the notation be as in (4.3). By the same result, we have that

$$\ell(\theta) = A_n - (1/2) q(Z_n - u) + \delta_n(u), \tag{4.10}$$

where $Z_n = n^{-1/2} \mathcal{I}_{\theta_0}^{-1} S(\theta_0)$, $q(Z_n - u) = (Z_n - u)^T \mathcal{I}_{\theta_0} (Z_n - u)$, A_n does not depend on u, and $\sup\{|\delta_n(u)| : \|u\| < K\} = o_p(1)$. Since $\boldsymbol{R\theta}_0 = \mathbf{0}$, it follows that $\boldsymbol{R\theta} \geq \mathbf{0}$ if and only if $\boldsymbol{Ru} \geq \mathbf{0}$, and $\boldsymbol{R\theta} = \mathbf{0}$ if and only if $\boldsymbol{Ru} = \mathbf{0}$. Now, it follows from the foregoing quadratic approximation that

$$\begin{aligned} LRT &= 2[\sup\{\ell(\theta) : \boldsymbol{R\theta} \geqslant \mathbf{0}\} - \sup\{\ell(\theta) : \boldsymbol{R\theta} = \mathbf{0}\}] \\ &= \inf\{q(Z_n - u) : \boldsymbol{Ru} = \mathbf{0}\} - \inf\{q(Z_n - u) : \boldsymbol{Ru} \geq \mathbf{0}\} + o_p(1) \\ &= D(Z_n) + o_p(1). \end{aligned}$$

Note that $Z_n \xrightarrow{d} Z$, where $Z \sim N(\mathbf{0}, \mathcal{I}_{\theta_0}^{-1})$, and $D(z)$ is continuous in z (in fact, it can be shown, using the triangular inequality, that $|D(z) - D(x)| \leq \|z - x\|$ and hence $D(z)$ is Lipchitz-continuous). It follows that $D(Z_n)$ and the LRT have a common limiting distribution and it is equal to the distribution of $D(Z)$. By Corollary 3.7.2 (on page 85), the distribution of $D(Z)$ is the chi-bar-square in (4.9). ■

This result is a direct generalization of the classical result that the asymptotic distribution of the likelihood ratio statistic for testing $\boldsymbol{R\theta} = \mathbf{0}$ against $\boldsymbol{R\theta} \neq \mathbf{0}$ is χ_r^2. In fact, for testing (4.5) and (4.6) the family of chi-bar-square distributions plays a role similar to that of the chi-squared distributions in inference under no inequality constraints.

It would be helpful to note that the main steps in the proof of Proposition 4.3.1 are very similar to those for proving that the *LRT* of $R\theta = 0$ against $R\theta \neq 0$ is asymptotically χ^2_r under H_0. In fact, the main idea of the proof carries over to much more general contexts, although some of the technical details would be different.

Let t_{obs} denote the observed sample value of *LRT*. If

$$\sum_{i=0}^{r} w_i(r, R\mathcal{I}_{\theta_0}^{-1}R^T)\text{pr}(\chi_i^2 \geqslant t_{obs}) \tag{4.11}$$

does not depend on the assumed true value θ_0 in the null parameter space, then it would be the large sample p-value. However, in general, it depends on θ_0 through $\mathcal{I}_{\theta_0}^{-1}$ in $w_i(q, R\mathcal{I}_{\theta_0}^{-1}R^T)$. Thus, θ_0 is a nuisance parameter and hence (4.11) is not a usable definition of a p-value. Some methods of dealing with this are discussed in the next subsection.

4.3.2 Tests in the Presence of Nuisance Parameters

Tests of statistical hypotheses in the presence of nuisance parameters are encountered frequently in statistical analysis. The presence of such parameters usually causes difficulties for statistical inference. Let θ be partitioned as $(\lambda : \psi)$ where $\psi \in \mathbb{R}^r$; here, and in what follows, we shall use the following notation for partitioning a vector such as θ to avoid writing the superscript "T" several times:

$$(\lambda : \psi) \quad \text{denotes} \quad (\lambda^T, \psi^T)^T.$$

For simplicity, we shall consider the special case,

$$H_0 : \psi = 0 \text{ and } H_1 : \psi \geqslant 0; \tag{4.12}$$

it will be clear that the main procedures discussed below are applicable for testing $\psi = 0$ against $\psi \in C$ where $C \subset \mathbb{R}^r$ and for more general testing problems as well.

Let T be a test statistic such that large values of T provide evidence against H_0, t denote the sample value of T, and λ_0 denote the true value of λ. Since the discussion of this subsection is centered around p-value and probability of Type I error, let us suppose that $H_0 : \psi = 0$ holds; therefore, we let $\theta_0 = (\lambda_0 : 0)$ denote the true value. Suppose that $\text{pr}(T \geqslant t \mid \psi = 0, \lambda)$ depends on λ. If λ_0 were known, then the p-value would be $\text{pr}(T \geqslant t \mid \psi = 0, \lambda = \lambda_0)$. Because λ_0 is unknown, it is not usable as a p-value.

Under the null hypothesis, $H_0 : \psi = 0$, the only unknown parameter is λ, and therefore the notation $\text{pr}_\lambda(\cdot|H_0)$ will be used for the probability evaluated under H_0 at $\theta = (\lambda : 0)$. Let

$$p_n(c; \lambda) = \text{pr}_\lambda(LRT \geqslant c|H_0) \quad \text{and} \quad p_\infty(c; \lambda) = \lim_{n\to\infty} p_n(c; \lambda).$$

In many cases, closed-form expressions are available for $p_\infty(t; \lambda)$, but not for $p_n(t; \lambda)$. In this case, our main interest would be on large sample tests based on $p_\infty(t, \lambda)$.

By Proposition 4.3.1,

$$p_\infty(t; \lambda) = \sum_{i=0}^{r} w_i\{r, \mathcal{I}^{\psi\psi}(\lambda)\}\text{pr}(\chi_i^2 \geqslant t) \tag{4.13}$$

where $\mathcal{I}^{\psi\psi}(\lambda)$ is the (ψ, ψ) block of $\mathcal{I}^{-1}(\theta)$ evaluated at $\theta = (\lambda : 0)$. If $p_\infty(t; \lambda)$ does not depend on λ then the large sample p-value corresponding to $LRT = t$ is $p_\infty(t; \lambda)$. Therefore, for the following discussions, we shall consider only the case when $p_\infty(c; \lambda)$ does depend on λ.

Bounds test:

Let $p_{n,L}(t)$ and $p_{n,U}(t)$ be positive numbers such that

$$p_{n,L}(t) \leq p_n(t; \lambda) \leq p_{n,U}(t) \qquad \text{for any } \lambda.$$

Then, the *bounds test* at level α is the following: reject H_0 if $p_{n,U}(t) \leq \alpha$, do not reject H_0 if $p_{n,L}(t) > \alpha$ and the test is inconclusive if $p_{n,L}(t) \leq \alpha < p_{n,U}(t)$. Usually, it is difficult to obtain sharp bounds for $p_n(t; \lambda)$. If the sample size is large then $p_n(t; \lambda)$ can be approximated by the tail probability of the chi-bar-square distribution in (4.13), and hence the bounds in (3.23) can be used for a large sample bounds test.

Supremum of pr(Type I Error) over the null parameter space (p_{sup}-test):

A test, which we call the p_{sup}-test, is based on

$$p\text{-value} = \sup_{\lambda\in\Lambda} p_n(t; \lambda) \tag{4.14}$$

where t is the sample value of LRT and Λ is the parameter space for λ, for example, see Lehmann (1983). This is the generally accepted procedure for exact inference. The argument underlying this procedure is that if the value of t for LRT is consistent with at least one point in the null parameter space then we should not reject H_0. A large sample approximation of (4.14) is

$$\text{reject } H_0 \text{ if } \sup_{\lambda\in\Lambda} p_\infty(t; \lambda) \leq \alpha. \tag{4.15}$$

Ideally, it would be better to approximate $lim_n sup_\lambda p_n(t; \lambda)$ than $sup_\lambda lim_n p_n(t; \lambda)$. It does not appear that this has been investigated for tests against inequality constraints.

In some cases, $p_\infty(t; \lambda)$ may vary substantially as a function of λ in Λ and hence this test may be too conservative. Let us illustrate this using an example. Consider the Perlman's test of $\psi = 0$ against $\psi \geqslant 0$ based on a random sample from $N(\psi, \lambda)$, where λ is an unknown $q \times q$ covariance matrix. For Perlman's statistic, $\mathcal{U}$, which is equivalent to the *LRT*, the finite sample null distribution is the distribution given by (see Proposition 3.10.3 on page 103)

$$p_n(t; \lambda) = \text{pr}_\lambda(\mathcal{U} \geqslant t|H_0) = \sum_{i=1}^{r} w_i(r, \lambda)\text{pr}(\chi_i^2/\chi_{n-r}^2 \geqslant t); \tag{4.16}$$

it was shown in (3.56) that

$$p_{sup} = \sup_{\lambda\in\Lambda} p_n(t; \lambda) = 0.5\{\text{pr}(\chi_{r-1}^2/\chi_{n-r}^2 \geqslant t) + \text{pr}(\chi_r^2/\chi_{n-r}^2 \geqslant t)\}. \tag{4.17}$$

It can be shown that the supremum of $p_n(t;\lambda)$ over λ is achieved when the correlation matrix corresponding to λ is equal to J, where J is the matrix in which every element is equal to 1 [Silvapulle (1996b), Perlman (1969)]. Clearly, J is an extreme form for a correlation matrix. The true correlation matrix in most practical situations is unlikely to be close to J. If we restrict the parameter space for λ away from a neighborhood of J, then the supremum of $p_n(t;\lambda)$ would be smaller than the expression on the right-hand side of (4.17). Consequently, the test in (4.14) with p_{sup} as in (4.17) appears to be conservative for most practical applications. For a discussion of this example, see Silvapulle (1996b).

There are some values of λ that one would be able to eliminate without looking at the data. After looking at the data, one may be tempted to eliminate values of λ that are far from an estimate of λ; however, this needs to be done carefully to ensure that the probability of Type I error does not exceed the nominal level. One approach that may overcome the conservative nature of p_{sup}-test is to approximate the p-value by the supremum of $\text{pr}(LRT \geqslant t|\lambda)$ over a confidence region for λ. The p^*-test discussed below provides such an approach.

Supremum of pr(Type I Error) over a confidence region(p^*-test):

In this method, the test is carried out in two stages. For illustrative purposes, suppose that we wish to test H_0 vs H_1 at level 0.05. In the first stage, we construct a confidence region for λ, say a 99 % confidence region; let this be denoted by $\mathcal{A}$. Let $p^* = \{(1-0.99) + \sup_{\lambda\in\mathcal{A}} p_n(t;\lambda)\}$. Now, in the second stage, a test at level 0.05 is

$$\text{reject } H_0 \text{ if } p^* \leq 0.05.$$

An interpretation of this is the following: In the first stage, when we replace the full parameter space for λ by a 99% confidence region for it, the "probability" of an error is $(1-0.99)$. In the second stage, we reject H_0 if $\sup_{\lambda\in\mathcal{A}} p_n(t;\lambda) \leq 0.04$, so that the combined probability of Type I error is not more than $(1-0.99)+0.04 = 0.05$. We call this, the *p^*-test,* and the foregoing p^* is an upper bound for the p-value of this test. Obviously, it is also possible to use a 98% confidence region for Stage 1 and carry out the test in Stage 2 with $p^* = \{(1-0.98) + \sup_{\lambda\in\mathcal{A}} p_n(t;\lambda)\}$. A rigorous statement of this is given below in Proposition 4.3.2.

Suppose that we wish to test

$$H_0 : \psi = 0 \text{ against } H_1 : \psi \in C$$

for some C. Let T_λ be a test statistic for a given λ, and let t_λ denote the sample value of T_λ; note that the test statistic may depend on λ. For simplicity, assume that large values of T_λ favor H_1. Let α_1 be a given small number in the range $0 \leq \alpha_1 < 1$. Let $\mathcal{A}$ be a random subset of Λ such that

$$\text{pr}(\lambda_0 \in \mathcal{A} \mid H_0) \geqslant 1-\alpha_1.$$

Thus, once the sample values have been substituted, $\mathcal{A}$ becomes a $100(1-\alpha_1)\%$ confidence region for λ_0 under H_0. Let

$$p^* = \alpha_1 + \sup_{\lambda\in\mathcal{A}} \text{pr}\{T_\lambda \geqslant t_\lambda \mid \theta = (\lambda : 0)\}.$$

Now, a level-α *p^*-test* is

$$\text{reject } H_0 \text{ if } p^* \leq \alpha. \tag{4.18}$$

This is stated in the next result, and the proof is given in the appendix to this chapter (see page 215).

Proposition 4.3.2 *The size of the p^*-test in* (4.18) *does not exceed α. Consequently, an upper bound for the p-value of this test is p^*.* ■

Although the p^*-test was motivated by the need for a test with better power than the p_{sup}-test, there is no guarantee that the former would have such a power advantage. However, it is worth noting that the power of p^*-test at level α is no less than that of the p_{sup}-test at level $(\alpha-\alpha_1)$. This suggests that the power of the p^*-test cannot be much worse than that of the p_{sup} test provided α_1 is small. Simulations in Silvapulle (1996b) suggest that the p^*-test has better power than the p_{sup}-test in the specific cases considered; for illustrative examples of this method and further discussions, see Silvapulle (1996b) and Berger and Boos (1994).

Point estimate of the p-value ($\hat{p}$-test):

Because λ_0 is unknown, one procedure is to use an estimated value of $p_n(t;\lambda_0)$ under H_0. Let $\hat{p}$ be a consistent estimate of $p_n(t;\lambda_0)$ under H_0; for example, it could be $p_n(t;\hat{\lambda})$ or $p_\infty(t;\hat{\lambda})$ where $\hat{\lambda}$ is a consistent estimate of λ under H_0. Now the large sample test, which we call the $\hat{p}$-test, is

$$\text{reject } H_0 \text{ if } \hat{p} \leq \alpha. \tag{4.19}$$

In small samples, this procedure must be used with caution because there is a possibility that the size of the test may exceed the nominal level α substantially. For example, for testing $\psi = 0$ against $\psi \geqslant 0$ based on a sample from $N(\psi,\lambda)$, the tail probability

$$\sum w_i(r,\hat{\lambda})\text{pr}(\chi_i^2/\chi_{n-r}^2 \geqslant t)$$

with $\hat{\lambda}$ being a consistent estimator of λ may turn out to be a poor estimate of

$$\sum w_i(r,\lambda_0)\text{pr}(\chi_i^2/\chi_{n-r}^2 \geqslant t)$$

if the chi-bar-square weights, $w_i(r,\lambda)$, are sensitive to the form of the covariance matrix λ and $\hat{\lambda}$ is not close to λ_0; this was illustrated in Silvapulle (1996b).

4.3.3 Wald- and Score-type Tests with *iid* Observations

In the standard inference literature, Wald and score tests of $R\theta = 0$ against $R\theta \neq 0$ are known to be asymptotically equivalent to the likelihood ratio test. These tests may be generalized in different directions when there are inequality constraints. In this section, we will introduce Wald- and score-type tests that are asymptotically equivalent to the likelihood ratio test for testing against inequality constraints.

Wald-type tests

The usual Wald statistic for testing $R\theta = 0$ against $R\theta \neq 0$ is

$$n(R\hat{\theta})^T(R\hat{\mathcal{I}}^{-1}R^T)^{-1}(R\hat{\theta}) \tag{4.20}$$

where $\hat{\mathcal{I}}$ is a consistent estimator of $\mathcal{I}(\theta)$. The following interpretations of this statistic are helpful; they are based on $\sqrt{n}(\hat{\theta} - \theta) \approx N\{0, \mathcal{I}(\theta)^{-1}\}$ and $\sqrt{n}(R\hat{\theta} - R\theta) \approx N(0, A)$ where $A = \{R\mathcal{I}(\theta)^{-1}R^T\}$:

1. It [i.e., (4.20)] is a measure of $\|R\hat{\theta}\|^2$. Since $R\hat{\theta}$ is an estimate of $R\theta$ obtained without imposing the constraints in the null hypothesis, it is reasonable to expect that $\|R\hat{\theta}\|^2$ would be small if $R\theta = 0$; and it would be large if $R\theta \neq 0$.

2. First, construct the LRT (=$n(R\hat{\theta})^T(RVR^T)^{-1}(R\hat{\theta})$) for testing $R\theta = 0$ against $R\theta \neq 0$ based on a single observation of $\hat{\theta}$ under the assumption that $\hat{\theta}$ is exactly $N(\theta, V)$, where V is assumed known. Now, (4.20) is a large sample approximation to this statistic obtained by substituting $V = n^{-1}\hat{\mathcal{I}}^{-1}$.

3. It is a measure of $\{d(\hat{\theta}, H_0) - d(\hat{\theta}, H_2)\}$, where $d(\hat{\theta}, H_i)$ is a measure of the distance between $\hat{\theta}$ and the parameter space defined by the hypothesis H_i $(i = 0, 1)$; (4.20) is obtained by choosing

$$d(\hat{\theta}, H) = \min\{n(\hat{\theta} - \theta)^T\hat{\mathcal{I}}(\hat{\theta} - \theta) : \theta \in H\}.$$

For testing $R\theta = 0$ against $R\theta \geqslant 0$, a statistic that resembles (4.20) and is along the lines of the first interpretation (of the foregoing three) is

$$W = n(R\tilde{\theta})^T(R\tilde{\mathcal{I}}^{-1}R^T)^{-1}(R\tilde{\theta}) \tag{4.21}$$

where $\tilde{\mathcal{I}}$ is an estimator of $\mathcal{I}$ under $H_1 : R\theta \geqslant 0$; thus, $\tilde{\mathcal{I}}$ can be $\mathcal{I}(\tilde{\theta})$ or $\mathcal{I}(\hat{\theta})$. We shall call this a Wald-type statistic. Note that the definition in (4.21) does not follow from the asymptotic distribution of $R\tilde{\theta}$ in the same way as (4.20) does from that of $R\hat{\theta}$ because $R\tilde{\theta}$ is not asymptotically normal. It will be seen that the W in (4.21) is asymptotically equivalent to *LRT*.

A statistic along the lines of the third interpretation is

$$D = \inf_{R\theta=0}\{n(\hat{\theta} - \theta)^T\hat{\mathcal{I}}(\hat{\theta} - \theta)\} - \inf_{R\theta\geqslant 0}\{n(\hat{\theta} - \theta)^T\hat{\mathcal{I}}(\hat{\theta} - \theta)\}. \tag{4.22}$$

We shall call this a *Distance Statistic*.

Because the parameter θ enters the hypotheses only through $R\theta$, we can think of $R\theta$ as the parameter of interest and define a statistic along the lines of the second interpretation with $R\theta$ playing the role of the parameter of interest. This leads to a statistic which is equal to D in (4.22) although its functional form is different:

$$\begin{aligned} D = \inf_{R\theta=0}\{&n(R\hat{\theta} - R\theta)^T(R\hat{\mathcal{I}}^{-1}R^T)^{-1}(R\hat{\theta} - R\theta) - \\ &\inf_{R\theta\geqslant 0}\{n(R\hat{\theta} - R\theta)^T(R\hat{\mathcal{I}}^{-1}R^T)^{-1}(R\hat{\theta} - R\theta). \end{aligned} \tag{4.23}$$

It will be seen later that D is also asymptotically equivalent to the *LRT*.

A Global score statistic

Recall that the score statistic for testing $\theta = \theta_0$ against $\theta \neq \theta_0$ is

$$n^{-1}S(\theta_0)^T\mathcal{I}_{\theta_0}^{-1}S(\theta_0) \tag{4.24}$$

where $S(\theta) = (\partial/\partial\theta)\ell(\theta)$ is the score function. The motivation for this is that if θ_0 is the true value then $E\{\ell(\theta)\}$ has a global maximum and a stationary value at θ_0, $n^{-1/2}S(\theta_0) \xrightarrow{d} N\{0, \mathcal{I}_{\theta_0}\}$, and $S(\theta_0)$ tends to be close to (respectively, far from) 0 when θ_0 is the true (respectively, not true) value. One possible way of extending this to the inequality constrained testing problem is to assess the extent to which $\{S(\bar{\theta}) - S(\tilde{\theta})\}$ is different from zero, where $\bar{\theta}$ and $\tilde{\theta}$ are estimates of θ under the null and alternative hypotheses, respectively. If H_0 is true, then $\bar{\theta}$ and $\tilde{\theta}$ are expected to be close and hence $\{S(\bar{\theta}) - S(\tilde{\theta})\}$ is expected to be close to zero; on the other hand, if H_0 is not true then $S(\tilde{\theta})$ is expected to be close to zero but not $S(\bar{\theta})$ and therefore, $\{S(\bar{\theta}) - S(\tilde{\theta})\}$ is expected to be away from zero. This suggests that a test can be based on the difference $\{S(\bar{\theta}) - S(\tilde{\theta})\}$. Instead of the score function $S(\theta)$ we may use the effective score, $R\mathcal{I}(\theta_0)^{-1}S(\theta_0)$. This suggests the following:

$$S_G = [R\hat{\mathcal{I}}^{-1}\{S(\bar{\theta}) - S(\tilde{\theta})\}]^T(R\hat{\mathcal{I}}^{-1}R^T)^{-1}[R\hat{\mathcal{I}}^{-1}\{S(\bar{\theta}) - S(\tilde{\theta})\}]. \tag{4.25}$$

The following form has also been suggested (see Robertson et al. (1988)):

$$\{S(\bar{\theta}) - S(\tilde{\theta})\}^T\mathcal{I}(\bar{\theta})^{-1}\{S(\bar{\theta}) - S(\tilde{\theta})\}.$$

The reason for referring to these as "Global" score statistics is that they use information about the shape of the score function over the whole parameter space. This is in contrast to the Rao score statistic and Neyman's $C(\alpha)$ statistic, which are based only on the properties of the score function in a *local neighborhood* of the assumed true null value.

It is well known that the likelihood ratio, Wald, and score tests are asymptotically equivalent for testing $R\theta = 0$ against $R\theta \neq 0$. The next proposition says that a corresponding result also holds for the foregoing Wald-type statistic in (4.21), the distance statistic (4.22), and the score statistic in (4.25). In fact, this result holds even in some regression-type models such as the generalized linear models (see Silvapulle (1994)); this will be discussed in the next subsection.

Proposition 4.3.3 *Suppose that Condition Q is satisfied. Then for testing* $H_0 : R\theta = 0$ *vs* $H_1 : R\theta \geq 0$, *we have the following:*

$$LRT = W + o_p(1) = D + o_p(1) = S_G + o_p(1) \text{ under } H_0. \tag{4.26}$$

Consequently, the common asymptotic null distribution of LRT, W, D, and S_G *is the chi-bar square in* (4.9).

Proof: Follows from the local quadratic approximations of loglikelihood in Proposition 4.2.2 (see Gourieroux and Monfort (1995) and Silvapulle (1994)). ■

It follows that, under some reasonable regularity conditions, the test statistics LRT, W, D, and S_G are asymptotically equivalent in the sense that they have the same asymptotic local power, more precisely Pitman asymptotic efficiency against local alternatives. Although the four statistics in (4.26) are asymptotically equivalent, very little is known about their relative advantages. From a computational point of view, D is easier to use than the others because it requires only the unconstrained estimator, $\hat{\boldsymbol{\theta}}$, which is usually easier to compute than the constrained estimators. The statistics S_G and W do not have any significant computational advantage over LRT because all of them require the constrained estimator. In contrast to S_G, W, and D, *LRT* has the advantage that it is invariant under a reparameterization. If a test that is computationally simpler than the LRT is required, then D is worth considering. In view of these, at this stage, S_G and W do not appear to be serious competitors to *LRT* or D.

For testing $\boldsymbol{R\theta} = \mathbf{0}$ against $\boldsymbol{R\theta} \neq \mathbf{0}$, the classical score statistic, also known as Rao's score statistic and the Lagrange multiplier statistic, provides a simple way of testing when the full model is complicated; the implementation requires estimation of the model only under H_0. This is important when the full model is significantly more complicated than the null model; this particular feature of the score statistic has been crucial for its popularity. Unfortunately, the global score statistic, S_G, does not have this simplicity. A natural generalization of the Rao's score statistic, which preserves the attractive feature (i.e., requires only the null model to be estimated) of the classical Rao's score statistic will be introduced in a later section.

4.3.4 Independent Observations with Covariates

Most of the foregoing results for independently and identically distributed observations extend in a natural way to other more general cases with non-*iid* observation, for example, linear and nonlinear regression models, generalized linear models, and some stochastic processes.

Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be independent, $f_i(\boldsymbol{y}; \boldsymbol{\theta})$ denote the density function of $\boldsymbol{Y}_i$ and $\ell(\boldsymbol{\theta})$ denote the loglikelihood, $\sum \log f_i(\boldsymbol{Y}_i; \boldsymbol{\theta})$. This includes the setting where the observations are $(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1, \ldots, n$, with $\boldsymbol{x}$ denoting a covariate; thus, regression-type models are included. Recall that in the previous subsections, we defined $\mathcal{I}(\boldsymbol{\theta})$ as the information per observation; it was the same for every observation because they were identically distributed. Now, since the observations are not identically distributed, we need to make some modifications. Let

$$\nabla = (\partial/\partial\boldsymbol{\theta}) \text{ and } \nabla^2 = (\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T).$$

Suppose that

$$n^{-1/2}\nabla\ell(\boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{V}_{\boldsymbol{\theta}}), \qquad \text{and} \qquad n^{-1}\nabla^2\ell(\boldsymbol{\theta}) \xrightarrow{a.s.} -\mathcal{V}_{\boldsymbol{\theta}} \tag{4.27}$$

for some positive definite matrix $\mathcal{V}_{\boldsymbol{\theta}}$. As was introduced at the beginning of this section, let $\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ denote the maximizers of $\ell(\boldsymbol{\theta})$ over Θ, $\{\boldsymbol{R\theta} \geqslant \mathbf{0}\}$ and $\{\boldsymbol{R\theta} = \mathbf{0}\}$, respectively. For testing

$$H_0 : \boldsymbol{R\theta} = \mathbf{0} \text{ against } H_1 : \boldsymbol{R\theta} \geqslant \mathbf{0}, \tag{4.28}$$

the likelihood ratio statistic is

$$LRT = 2\{\ell(\tilde{\boldsymbol{\theta}}) - \ell(\bar{\boldsymbol{\theta}})\}. \tag{4.29}$$

We shall continue to assume that Assumption A1 on page 146 holds with appropriate modifications. Assume that (a) the quadratic approximation (4.3) in Proposition 4.2.2 (page 147) holds with $\mathcal{I}(\boldsymbol{\theta})$ replaced by $\mathcal{V}(\boldsymbol{\theta})$, (b) $\mathcal{V}(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and (c) $\tilde{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ are $n^{1/2}$-consistent. The proof of the quadratic approximation may involve lengthy technical details; for example, for the case of generalized linear models, see Fahrmeir and Kaufmann (1985). The $\sqrt{n}$-consistency of $\tilde{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ would follow from the consistency of the global *mle* and the quadratic approximation, in much the same way as in the *iid* case. Thus, conceptually, the conditions required for the regression type setting is not that different from that for the *iid* setting but the technical details could be complicated.

Now, the proofs of Propositions 4.3.1 and 4.3.3 still hold with appropriate modifications; see Silvapulle (1994) for details.

Proposition 4.3.4 *The asymptotic null distribution of LRT for testing*

$$H_0 : \boldsymbol{R\theta} = \mathbf{0} \text{ against } H_1 : \boldsymbol{R\theta} \geqslant \mathbf{0}$$

is equal to the distribution of

$$\inf_{\boldsymbol{R\theta}=\mathbf{0}}\{(\boldsymbol{Z}-\boldsymbol{\theta})^T\mathcal{V}_{\boldsymbol{\theta}_0}(\boldsymbol{Z}-\boldsymbol{\theta})\} - \inf_{\boldsymbol{R\theta}\geqslant\mathbf{0}}\{(\boldsymbol{Z}-\boldsymbol{\theta})^T\mathcal{V}_{\boldsymbol{\theta}_0}(\boldsymbol{Z}-\boldsymbol{\theta})\}$$

where $\boldsymbol{Z} \sim N(\mathbf{0}, \mathcal{V}_{\boldsymbol{\theta}_0}^{-1})$. *Consequently, as* $n \to \infty$,

$$pr_{\boldsymbol{\theta}_0}(LRT \geqslant c \mid H_0) \to \sum_{i=0}^{r} w_i(r, \boldsymbol{R}\mathcal{V}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{R}^T)pr(\chi_i^2 \geqslant c). \qquad \blacksquare$$

It appears that, if

1. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges to $N\{\mathbf{0}, \mathcal{V}_{\boldsymbol{\theta}_0}^{-1}\}$, and
2. the asymptotic null distribution of the likelihood ratio statistic, $2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\bar{\boldsymbol{\theta}})\}$, for testing $\boldsymbol{R\theta} = \mathbf{0}$ against $\boldsymbol{R\theta} \neq \mathbf{0}$ is a chi-square,

then it is almost certain that the conclusions of the foregoing proposition also hold. If the parameter space for $\boldsymbol{\theta}$ is open and the asymptotic null distribution of $2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\bar{\boldsymbol{\theta}})\}$ for testing $H_0 : \boldsymbol{R\theta} = \mathbf{0}$ against $H_2 : \boldsymbol{R\theta} \neq \mathbf{0}$ is not a chi-square, then the results of Proposition 4.3.4 are unlikely to hold.

It is also possible to construct Wald-type, score-type, and distance-based tests as follows:

$$W = n(\boldsymbol{R}\tilde{\boldsymbol{\theta}})^T(\boldsymbol{R}\hat{\boldsymbol{V}}^{-1}\boldsymbol{R}^T)^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\theta}})$$
$$S_G = n^{-1}\boldsymbol{U}^T(\boldsymbol{R}\hat{\boldsymbol{V}}^{-1}\boldsymbol{R}^T)^{-1}\boldsymbol{U}, \quad \text{where } \boldsymbol{U} = \boldsymbol{R}\hat{\boldsymbol{V}}^{-1}\{\boldsymbol{S}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{S}(\bar{\boldsymbol{\theta}})\}$$
$$D = \{d(\hat{\boldsymbol{\theta}}, H_0) - d(\hat{\boldsymbol{\theta}}, H_1)\}, \quad \text{where } d(\hat{\boldsymbol{\theta}}, H) = \inf_{\boldsymbol{\theta}\in H} n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{4.30}$$

Now, it can be shown that (see Silvapulle (1994), Gourieroux and Monfort (1995))

$$LRT = W + o_p(1) = S_G + o_p(1) = D + o_p(1). \tag{4.31}$$

Therefore, these tests have the same asymptotic null distribution. Further, under some weak conditions, (4.31) also hold under Pitman-type local alternatives and hence the tests in (4.31) also have the same asymptotic local power.

Least squares in linear regression

The foregoing results extend in a natural way to some quasi-likelihood models as well. To illustrate the main ideas, let us first consider the linear regression model,

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\theta} + \epsilon_i,$$

where the error terms are *iid* with common variance σ^2. Let

$$R_n(\boldsymbol{\theta}) = (-1/2)\sum(y_i - \boldsymbol{x}_i^T\boldsymbol{\theta})^2 \qquad \text{and} \qquad \boldsymbol{W} = \lim(n^{-1}\boldsymbol{X}^T\boldsymbol{X}),$$

where $\boldsymbol{X}$ is the design matrix $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$. Let $\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}$, and $\bar{\boldsymbol{\theta}}$ be the estimators corresponding to the objective function $R_n(\boldsymbol{\theta})$; thus $\hat{\boldsymbol{\theta}}$ is simply the unconstrained ordinary least squares estimator. Let $\boldsymbol{S}(\boldsymbol{\theta}) = \nabla R_n(\boldsymbol{\theta})$. Then we have that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, \sigma^2\boldsymbol{W}^{-1})$,

$$n^{-1/2}\boldsymbol{S}(\boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, \sigma^2\boldsymbol{W}) \quad \text{and} \quad -n^{-1}\nabla^2 R_n(\boldsymbol{\theta}) \xrightarrow{p} \boldsymbol{W}.$$

Note that the asymptotic covariance of $\boldsymbol{S}(\boldsymbol{\theta}_0)$ and the limit of $-n^{-1}\nabla^2 R_n(\boldsymbol{\theta}_0)$ are proportional, but not equal; if R_n were the loglikelihood then they would have been equal. Now, for testing $H_0 : \boldsymbol{R\theta} = \boldsymbol{0}$ against $H_2 : \boldsymbol{R\theta} \neq \boldsymbol{0}$, an LR-type statistic and its asymptotic null distribution are given by

$$T_n = 2\{R_n(\hat{\boldsymbol{\theta}}) - R_n(\bar{\boldsymbol{\theta}})\}/\hat{\sigma}^2 \xrightarrow{d} \chi^2_r \quad \text{under } H_0,$$

where $\hat{\sigma}$ is a consistent estimator of σ.

Note that

$$T_n = \{\|(\boldsymbol{Y} - \boldsymbol{X}\bar{\boldsymbol{\theta}})/\sigma\|^2 - \|(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})/\sigma\|^2\}(\sigma^2/\hat{\sigma}^2),$$

$(\sigma^2/\hat{\sigma}^2) \xrightarrow{p} 1$, and the distribution of the term in $\{\ldots\}$ does not depend on σ. Thus, essentially σ^2 factors out and cancels with $\hat{\sigma}^{-2}$ in the limit. This illustrates that in the linear model, we do not need to know the exact population distribution to construct tests of $H_0 : \boldsymbol{R\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{R\theta} \neq \boldsymbol{0}$.

In what follows we extend Proposition 4.3.4 to incorporate this type of setting.

A quasi-likelihood method

Consider a general parametric model with unknown parameter $\boldsymbol{\theta}$. Let $R_n(\boldsymbol{\theta})$ denote an objective function which will be maximized for estimating $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}$, and $\bar{\boldsymbol{\theta}}$ be the estimators corresponding to this objective function. Let $\boldsymbol{S}_n(\boldsymbol{\theta}) = \nabla R_n(\boldsymbol{\theta})$. Suppose that

$$n^{-1/2}\boldsymbol{S}(\boldsymbol{\theta}_0) \xrightarrow{d} N\{\boldsymbol{0}, \sigma^2\boldsymbol{W}(\boldsymbol{\theta}_0)\} \quad \text{and} \quad n^{-1}\nabla^2 R_n(\boldsymbol{\theta}) \xrightarrow{p} -\boldsymbol{W}(\boldsymbol{\theta}).$$

Let $\boldsymbol{Z}_n(\boldsymbol{\theta}_0) = n^{-1/2}\boldsymbol{W}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{S}(\boldsymbol{\theta}_0)$. Assume that R_n satisfies regularity conditions similar to those required for the *iid* setting. Therefore, we have $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\boldsymbol{0}, \sigma^2\boldsymbol{W}(\boldsymbol{\theta}_0)^{-1}\}$, $\boldsymbol{Z}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N\{\boldsymbol{0}, \sigma^2\boldsymbol{W}(\boldsymbol{\theta}_0)^{-1}\}$, and a quadratic approximation similar to (4.3) holds for $R_n(\boldsymbol{\theta})$ with the information matrix $\mathcal{I}$ replaced by $\boldsymbol{W}$. Now consider test of $H_0 : \boldsymbol{R\theta} = \boldsymbol{0}$ vs $H_1 : \boldsymbol{R\theta} \geqslant \boldsymbol{0}$. A suitable test statistic for this is

$$T_n = 2\{R_n(\tilde{\boldsymbol{\theta}}) - R_n(\bar{\boldsymbol{\theta}})\}/\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is a consistent estimator of σ. Then we have

$$\begin{aligned} T_n &= 2\{R_n(\tilde{\boldsymbol{\theta}}) - R_n(\bar{\boldsymbol{\theta}})\}/\sigma^2 + o_p(1) \\ &= D(\boldsymbol{Z}_n)/\sigma^2 + o_p(1) = D(\boldsymbol{Z}_n/\sigma) + o_p(1) \\ &\xrightarrow{d} D(\boldsymbol{U}) \quad \text{under } H_0 \end{aligned} \tag{4.32}$$

where $D(\boldsymbol{z})$ is as in (4.8) with $\mathcal{I}_{\boldsymbol{\theta}_0}$ replaced by $\boldsymbol{W}(\boldsymbol{\theta}_0)$, and $\boldsymbol{U} \sim N\{\boldsymbol{0}, \boldsymbol{W}(\boldsymbol{\theta}_0)^{-1}\}$. Therefore,

$$T_n \xrightarrow{d} \bar{\chi}^2\{\boldsymbol{RW}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{R}^T, \mathbb{R}^{+r}\} \quad \text{under } H_0.$$

We can also define Wald, score, and distance type test statistics as in (4.30). All that we need to do is to introduce $\hat{\sigma}^{-2}$ on the RHS of the definitions of W, S_G, and D in (4.30) and replace $\hat{\boldsymbol{V}}$ by $\hat{\boldsymbol{W}}$, a consistent estimator of $\boldsymbol{W}(\boldsymbol{\theta}_0)$:

$$W = \hat{\sigma}^{-2}n(\boldsymbol{R}\tilde{\boldsymbol{\theta}})^T(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^T)^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\theta}})$$
$$S_G = \hat{\sigma}^{-2}n^{-1}\boldsymbol{U}^T(\boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\boldsymbol{R}^T)^{-1}\boldsymbol{U}, \quad \text{where } \boldsymbol{U} = \boldsymbol{R}\hat{\boldsymbol{W}}^{-1}\{\boldsymbol{S}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{S}(\bar{\boldsymbol{\theta}})\}$$
$$D = \hat{\sigma}^{-2}\{d(\hat{\boldsymbol{\theta}}, H_0) - d(\hat{\boldsymbol{\theta}}, H_1)\}, \quad \text{where } d(\hat{\boldsymbol{\theta}}, H) = \inf_{\boldsymbol{\theta}\in H} n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\hat{\boldsymbol{W}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{4.33}$$

The rest of the details for the asymptotic equivalence of T_n, W, S_G, and D remain essentially the same; see Silvapulle (1994) and Gourieroux and Monfort (1995) for details. Finally, let us note that, for a generalized linear model with canonical link, the foregoing results are adequate to apply quasi-likelihood ratio test in the presence of an overdispersion parameter, σ.

Robust test in the linear model

The foregoing results can be used for developing *robust tests* of $\boldsymbol{R\theta} = \mathbf{0}$ vs $\boldsymbol{R\theta} \geqslant \mathbf{0}$ in the linear regression model $\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{E}$. Suppose that the objective function is

$$R_n(\theta) = -\sum \rho\{(y_i - x_i^T\theta)/s\}s,$$

where ρ is a suitably chosen convex function and s is an estimate of scale, for example, a multiple of the median of the absolute values of the least square residuals. Now, it may be verified that the foregoing tests of $\boldsymbol{R\theta} = \mathbf{0}$ vs $\boldsymbol{R\theta} \geqslant \mathbf{0}$ are applicable. See Silvapulle (1992b) and Silvapulle (1992c) for details where proofs of the asymptotic equivalence of the tests are also given; these provide details for the statistics of the form T_n and D, respectively. It is shown that the robustness properties of the M-estimator carry over to the corresponding test statistics T_n and D. These results also hold in *group sequential analysis*; see Silvapulle and Sen (1993).

4.3.5 Examples

Example 4.3.1

An assay was carried out with the bacterium *E. coli* to evaluate the genotoxic effects of 9-AA and potassium chromate. Further background and the data are given in Example 1.2.7 of Chapter 1 (page 11). The objective is to test whether potassium chromate and 9-AA have a synergistic effect. The *simple independent action* (SIA) model has attracted interest for studies of the type in this example. To test SIA against synergism, let $\pi_{ij} = \text{pr}\{$ positive response for a test unit in cell $(i,j)\}$, and

$$\log(1 - \pi_{ij}) = \mu + \alpha_i + \tau_j + \eta_{ij} \tag{4.34}$$

where $i = 1, 2$ and $j = 1, \ldots, 5$. To ensure that the parameters in (4.34) are identified, let us impose the constraints, $\alpha_1 = \tau_1 = \eta_{i1} = \eta_{1j} = 0$ for all (i,j). Let $\boldsymbol{\theta} = (\mu, \alpha_2, \tau_2, \tau_3, \tau_4, \tau_5, \eta_{22}, \eta_{23}, \eta_{24}, \eta_{25})^T$, $\boldsymbol{\lambda} = (\mu, \alpha_2, \tau_2, \tau_3, \tau_4, \tau_5)^T$ and $\boldsymbol{\psi} = (\eta_{22}, \eta_{23}, \eta_{24}, \eta_{25})^T$. It is easily seen that the number of distinct π_{ij} parameters, namely ten, is also the number of distinct parameters on the right-hand side of (4.34), and that (4.34) is a one-to-one transformation from $\boldsymbol{\pi}$ to $\boldsymbol{\theta}$, where $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}, \pi_{15}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{24}, \pi_{25})^T$. The SIA model is equivalent to $\boldsymbol{\psi} = \mathbf{0}$. If $\boldsymbol{\psi} \geqslant \mathbf{0}$ and $\boldsymbol{\psi} \neq \mathbf{0}$, then there is synergism; if $\boldsymbol{\psi} \leq \mathbf{0}$ and $\boldsymbol{\psi} \neq \mathbf{0}$, then there is antagonism.

In this example, the question of interest is whether or not there is any synergism. Therefore, the statistical inference problem may be formulated as a test of

$$H_0 : \boldsymbol{\psi} = \mathbf{0} \text{ against } H_1 : \boldsymbol{\psi} \geqslant \mathbf{0}. \tag{4.35}$$

This is within the general framework considered in this section; it is a special case of (4.28) with $\boldsymbol{R} = [\mathbf{0} : \boldsymbol{I}_4]$, where $\boldsymbol{I}_4$ is the 4×4 identity matrix and $\mathbf{0}$ is the 4×6 matrix of zeroes. Now, the null and alternative hypotheses are $H_0 : \boldsymbol{R\theta} = \mathbf{0}$ and $H_1 : \boldsymbol{R\theta} \geqslant \mathbf{0}$, respectively.

For cell (i,j), let n_{ij} and y_{ij} denote the number of units tested and the number of positive responses, respectively. Then the likelihood is the product of all the binomial probabilities, and therefore

$$\ell(\theta) = \sum_{i=1}^{2}\sum_{j=1}^{5}[y_{ij}\log\pi_{ij} + (n_{ij} - y_{ij})\log(1 - \pi_{ij}) + c_{ij}]$$

where π_{ij} as a function of $\boldsymbol{\theta}$ is given by (4.34) and the constant c_{ij} is the logarithm of the binomial coefficient, $n_{ij}!\{y_{ij}!(n_{ij} - y_{ij})!\}^{-1}$. By definition,

$$\begin{aligned} LRT &= 2\{\max_{\theta\in H_1}\ell(\theta) - \max_{\theta\in H_0}\ell(\theta)\} \\ &= 2\{\max_{\psi\geqslant 0}\ell(\theta) - \max_{\psi=0}\ell(\theta)\}. \end{aligned} \tag{4.36}$$

In this model, the loglikelihood is concave, and by modifying the arguments in Silvapulle (1981), it can be verified that the *mle* exists at a finite point with probability one. The value of $\max\{\ell(\theta) : \psi \geqslant 0\}$ under H_1 can be computed using a nonlinear inequality constrained optimization program in software libraries, such as *IMSL*, *MATLAB*, or *NAG*. In this example, we did not need to use such a constrained optimization program, for reasons explained below.

Note that the equation (4.34) is a reparameterization of $\boldsymbol{\pi}$ in terms of $\boldsymbol{\theta}$. Therefore, to compute the global *mle*, $\hat{\boldsymbol{\theta}}$, it suffices to solve (4.34) with $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}$ where $\hat{\boldsymbol{\pi}}$, the vector of sample proportions of successes, is the global *mle* of $\boldsymbol{\pi}$. The computed value of $\hat{\boldsymbol{\psi}}$ turned to be nonnegative for each component. Therefore, the *mle* of $\boldsymbol{\theta}$ under the constraint $\boldsymbol{\psi} \geqslant \mathbf{0}$ is also the unconstrained estimator, $\hat{\boldsymbol{\theta}}$. Therefore, the sample value of the *LRT* for testing $\boldsymbol{\psi} = \mathbf{0}$ against $\boldsymbol{\psi} \geqslant \mathbf{0}$ is equal to that for testing $\boldsymbol{\psi} = \mathbf{0}$ against $\boldsymbol{\psi} \neq \mathbf{0}$. Although the *LRT* for the two tests have the same numerical value for these data, they have different p-values, because their null distributions are different.

The computed value of *LRT* is 60.5. A large sample approximation of the p-value for testing $\boldsymbol{\psi} = \mathbf{0}$ against $\boldsymbol{\psi} \geqslant \mathbf{0}$ is

$$\sup_{\boldsymbol{\lambda}}\sum_{i=1}^{4} w_i\{4, \mathcal{I}^{\psi\psi}(\boldsymbol{\lambda})\}\text{pr}(\chi_i^2 \geqslant 60.5).$$

Because this supremum is not easy to compute, it is worth applying a bounds test first. An upper bound for the foregoing large sample p-value is (see, (3.23))

$$0.5[\text{pr}(\chi_3^2 \geqslant 60.5) + \text{pr}(\chi_4^2 \geqslant 60.5)]; \tag{4.37}$$

this is less than 0.001. Therefore, the large sample p-value corresponding to a sample value of 60.5 for the *LRT* is less than 0.001. Therefore, there is sufficient evidence to reject H_0 and accept H_1; in other words, reject the SIA model and accept that there is synergism between potassium chromate and 9-AA.

For illustrative purposes, let us apply the *LRT* for testing $\boldsymbol{\psi} = \mathbf{0}$ against $\boldsymbol{\psi} \neq \mathbf{0}$ as well. It was noted that the *LRT* for this is also 60.5. The large sample p-value for this unrestricted test is $\text{pr}(\chi_4^2 \geqslant 60.5)$, which is also less than 0.001. Therefore,

assuming that the nominal level is 0.05, we would reject $\psi = 0$ and accept $\psi \neq 0$. It is valid to apply this unrestricted test for testing $\psi = 0$ against $\psi \geqslant 0$ as well. Recall that the reason that we prefer to apply a one-sided/restricted test as opposed to a two-sided/unrestricted test for testing $\psi = 0$ against $\psi \geqslant 0$ is for power advantage; validity is not the issue because a 5% level test of $\psi = 0$ against $\psi \neq 0$ is also a 5% level test of $\psi = 0$ against $\psi \geqslant 0$. Since the p-value for $\psi = 0$ against $\psi \neq 0$ is small, we would reject $\psi = 0$ and accept $\psi \geqslant 0$. Therefore, in this example, the evidence against $\psi = 0$ is sufficiently strong and we need not have carried out a one-sided test to decide whether or not to reject H_0. However, if it is known that $\psi \geqslant 0$ then a test of $H_0 : \psi = 0$ against $H_1 : \psi \geqslant 0$ would in general summarize the statistical evidence better than a test of $H_0 : \psi = 0$ against $H_2 : \psi \neq 0$.

The foregoing computations can be modified to incorporate an overdispersion parameter as well. ■

Example 4.3.2 *Comparison of treatments with ordinal response data.*

Example 1.2.8 of Chapter 1 (page 12) provides details of a trial to compare two treatments for ulcer; the data are given in Table 1.7. The objective of the study is to show that Treatment B is better than Treatment A. The hypotheses can be formulated in different ways. One possibility is to say that Treatment B is better than Treatment A if the *local odds ratios* are all greater than or equal to 1 with at least one of them being greater than 1; for interpretations and discussions about odds ratios, see Chapter 6. However, this may be an unnecessarily stringent requirement. A less stringent, and hence perhaps more desirable, formulation of the problem is

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 1 \quad \text{and} \quad H_1 : \gamma_1 \geqslant 1, \gamma_2 \geqslant 1, \gamma_3 \geqslant 1, \tag{4.38}$$

$$\text{where } \gamma_q = \Big(\sum_{j=1}^{q} \pi_{1j}\Big)\Big(\sum_{j=q+1}^{C} \pi_{2j}\Big)\Big\{\Big(\sum_{j=1}^{q} \pi_{2j}\Big)\Big(\sum_{j=q+1}^{C} \pi_{1j}\Big)\Big\}^{-1} \tag{4.39}$$

and $\pi_{ij} = \text{pr}(\text{the response of a subject in row } i \text{ falls in column } j)$.

The γ_i s are also certain odds ratios of interest as discussed in Example 1.2.8 of Chapter 1. The parameter $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23})^T$ defines the unrestricted model completely. The likelihood is simply the product of the two multinomial probabilities for Treatment A and Treatment B:

$$\text{Likelihood at } \boldsymbol{\pi} \propto \Pi_{i=1,j=1}^{i=2,j=4} \pi_{ij}^{y_{ij}}. \tag{4.40}$$

To apply the results of this section, it is important that the hypotheses involve only linear functions of the parameter that appears in the likelihood. However, in the foregoing formulation, this is not the case because the likelihood is expressed as a function of $\boldsymbol{\pi}$ (see 4.40) and the hypothesis involve nonlinear functions of $\boldsymbol{\pi}$. General methods of testing hypotheses involving nonlinear functions of parameters will be discussed later. Such methods can be applied for testing (4.38). However, in this particular example, there is a simple reparameterization such that the hypotheses involve only linear functions and hence the results of this section are applicable.

Let us parameterize the π_{ij} s as

$$\log(\pi_{i1} + \ldots + \pi_{iq}) = \lambda_q - \psi_{iq} \tag{4.41}$$

where $i = 1, 2$ and $q = 1, 2, 3$. Let $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_3, \psi_{21}, \psi_{22}, \psi_{23})^T$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$ and $\boldsymbol{\psi} = (\psi_{21}, \psi_{22}, \psi_{23})^T$. Then $\gamma_1 \geqslant 1, \gamma_2 \geqslant 1, \gamma_3 \geqslant 1$ is equivalent to $\boldsymbol{\psi} \geqslant \mathbf{0}$. Now,

$$\text{Likelihood at } \boldsymbol{\theta} = L(\boldsymbol{\theta}) = \text{ Const } \Pi_{i=1,j=1}^{i=2,j=4} \{\pi_{ij}(\boldsymbol{\theta})\}^{y_{ij}}$$

where the value of $\pi_{ij}(\boldsymbol{\theta})$ in terms of $\boldsymbol{\theta}$ is given by (4.41). For the data in Table 1.7, $\hat{\gamma}_1 = (12 \times 27)/(5 \times 20) = 3.24$, $\hat{\gamma}_2 = (22 \times 19)/(13 \times 10) = 3.22,4$ and $\hat{\gamma}_3 = (26 \times 11)/(21 \times 6) = 2.27$. Thus, each $\hat{\gamma}_i$ is greater than 1 and hence $\hat{\boldsymbol{\psi}} \geqslant \mathbf{0}$. Again, as in the previous example, the unrestricted estimator of $\boldsymbol{\theta}$ satisfies the inequalities in H_1. Consequently, the estimator of $\boldsymbol{\theta}$ under H_1 is also the unrestricted estimator of $\boldsymbol{\theta}$. Now, the *LRT* for testing $\boldsymbol{\psi} = \mathbf{0}$ against $\boldsymbol{\psi} \geqslant \mathbf{0}$ is 6.0. Under H_0, we have that $\pi_{1j} = \pi_{2j}$ for $j = 1, \ldots, 4$. Therefore, the *mle* of $(\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})$ under H_0 is the vector of sample proportions for column totals ($i = 1, 2$). Hence, the *mle* of $\boldsymbol{\pi}$ under H_0 is $\bar{\boldsymbol{\pi}} = (\bar{\boldsymbol{\pi}}_1^T, \bar{\boldsymbol{\pi}}_2^T)^T = (17/64, 18/64, 12/64, 17/64, 17/64, 18/64, 12/64, 17/64)^T$. Because the *mle* is invariant under transformations, the *mle* of $\boldsymbol{\lambda}$ under H_0 is obtained by solving (4.41) with $\boldsymbol{\pi} = \bar{\boldsymbol{\pi}}$ and $\boldsymbol{\psi} = \mathbf{0}$; this leads to $\bar{\boldsymbol{\lambda}} = \{\log(17/64), \log(35/64), \log(47/64)\}$. Now, an estimate of the large sample p-value for *LRT* is (see $\hat{p}$-test)

$$\hat{p} = \sum_{i=1}^{3} w_i\{3, \mathcal{I}^{\psi\psi}(\bar{\boldsymbol{\lambda}})\}\text{pr}(\chi_i^2 \geqslant 6.0) = 0.054;$$

the formulas in (3.26) were used for computing $w_i\{3, \mathcal{I}^{\psi\psi}(\bar{\boldsymbol{\lambda}})\}$. Using the same closed-form for the weights, the supremum of the tail probability over $\boldsymbol{\lambda}$ was computed as

$$\sup_{\boldsymbol{\lambda}} \sum_{i=1}^{3} w_i\{3, \mathcal{I}^{\psi\psi}(\boldsymbol{\lambda})\}\text{pr}(\chi_i^2 \geqslant 6.0) = 0.083.$$

Therefore, a large sample approximation of the p-value is 0.083; as expected, this is larger than $\hat{p}$ (= 0.054) and the difference between the two is not negligible. In this example, some of the observed cell frequencies (in Table 1.7) are not that large, and, therefore, it may be prudent to be conservative and use $0.083(= p_{sup})$ rather than $0.054(= \hat{p})$ as a large sample p-value.

It would be of interest to compare these results with a test against the unrestricted alternative that Treatments A and B are different for which the null and alternative hypotheses take the form

$$H_0 : \boldsymbol{\psi} = \mathbf{0} \text{ and } H_2 : \boldsymbol{\psi} \neq \mathbf{0},$$

respectively. Because the unrestricted *mle* of $\boldsymbol{\psi}$ satisfies the constraint $\boldsymbol{\psi} \geqslant \mathbf{0}$, the *LRT* of $\boldsymbol{\psi} = \mathbf{0}$ against $\boldsymbol{\psi} \neq \mathbf{0}$ and that of $\boldsymbol{\psi} = \mathbf{0}$ against $\boldsymbol{\psi} \geqslant \mathbf{0}$ have the same numerical value, namely 6.0. However, the p-values are different because the null

distributions are different. For $H_0 : \psi = 0$ against $H_2 : \psi \neq 0$, the large sample p-value is $\text{pr}(\chi^2_3 \geqslant 6.0) = 0.112$. As expected (because the components of U are all positive), this p-value is larger than those obtained earlier for testing against $R\theta \geq 0$.

As has been noted earlier, if the unrestricted estimate satisfies the inequalities in H_1, then the *LRT* against $H_1 : \psi \geqslant 0$ and that against $H_2 : \psi \neq 0$ would have the same numerical value but the latter would have a larger p-value. This follows easily because the chi-bar-square distribution corresponding to the test against $\psi \geqslant 0$ has shorter tails than the χ^2 distribution corresponding to the test against $\psi \neq 0$. This is consistent with our intuition: if the unrestricted estimate, $\hat{\psi}$ is nonnegative, then the data are pointing us in the direction of $\psi \geqslant 0$ and the evidence against $\psi = 0$ from a test against $\psi \geqslant 0$ is likely to be stronger than that from a global test against $\psi \neq 0$. The corresponding phenomenon is well-known in the classical one-sided t-test on the mean of a normal distribution. ■

4.4 TESTS OF $h(\theta) = 0$

The results in Section 4.1 for linear constraints of θ can be extended to nonlinear constraints as well. For simplicity we shall first consider the case when the observations are *iid*. Throughout this section we shall assume that the natural parameter space Θ is open and that Condition Q is satisfied. Most of the results of this section can be extended to the setting when the observations are not *iid*. Some of these will be discussed later in this chapter.

Let $h(\theta) = (h_1(\theta), \ldots, h_r(\theta))^T$ be a continuously differentiable vector function. In the next subsection, we will define several tests of $h(\theta) = 0$ vs $h(\theta) \geqslant 0$. Then in subsection 4.4.2 we will consider a test against $h_2(\theta) \geqslant 0$ where h_2 is a subvector of h.

4.4.1 Test of $h(\theta) = 0$ Against $h(\theta) \geqslant 0$

Let us consider the hypotheses

$$H_0 : h(\theta) = 0 \qquad \text{vs} \qquad H_1 : h(\theta) \geqslant 0. \tag{4.42}$$

Let $H(\theta) = (\partial/\partial\theta^T)h(\theta)$. Assume that

$$H(\theta_0) \text{ has full row rank.} \tag{4.43}$$

As usual, let $\hat{\theta}$ denote the unconstrained *mle*, and $\tilde{\theta}$ and $\bar{\theta}$ denote the *mle*'s under H_1 and H_0, respectively. The definition of *LRT* does not require any new ideas: $LRT = 2\{\ell(\tilde{\theta}) - \ell(\bar{\theta})\}$. The asymptotic null distribution of this turns out to be a chi-bar-square. Another convenient statistic for testing (4.42) is the distance statistic based on $\hat{\theta}$. Since $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\{0, \mathcal{I}(\theta_0)^{-1}\}$, we define

$$D = d(\hat{\theta}, H_0) - d(\hat{\theta}, H_1), \text{ where } d(\hat{\theta}, H) = \inf_{\theta \in H} n(\hat{\theta} - \theta)^T \hat{\mathcal{I}} (\hat{\theta} - \theta)$$

where $\hat{\mathcal{I}}$ is a consistent estimator of $\mathcal{I}(\theta_0)$. It will be seen that the *LRT* and D are asymptotically equivalent and have the common asymptotic null distribution, $\bar{\chi}^2\{H(\theta_0)\mathcal{I}(\theta_0)^{-1}H(\theta_0)^T, \mathbb{R}^{+r}\}$. Even if $H(\theta_0)$ has rank less than the dimension of h, the asymptotic null distribution of *LRT* is still a chi-bar-square, but it has a different form.

Much of the existing literature concerning tests of $h(\theta) = 0$ usually assumes that $H(\theta_0)$ is of full row-rank. It is important to note that this may fail to be satisfied even in what appears to be well-behaved models.

An example where $H(\theta_0)$ is not of full rank:

Let $A(\theta)$ be the 2×2 matrix $(\theta_1, \theta_3 \mid \theta_3, \theta_2)$,

$$H_0 : A(\theta) = 0 \quad \text{and} \quad H_1 : A(\theta) \text{ is positive semi-definite.}$$

This may be restated in the equivalent form $H_0 : h(\theta) = 0$ and $H_1 : h(\theta) \geqslant 0$, where $h(\theta) = (-\theta_1, \theta_1\theta_2 - \theta_3^2)^T$. Now, $(\partial/\partial\theta^T)h(\theta) = [1, 0, 0 | \theta_2, \theta_1, -2\theta_3]$; the last row becomes a zero vector at the null value, $\theta = 0$, and hence $(\partial/\partial\theta^T)h(\theta)$ is not of full rank. Therefore, condition (4.43) is not satisfied.

To provide a more concrete example, we consider the regression model

$$y = g(x;\theta) + \epsilon, \text{ where } g(x;\theta) = \theta_1 x_1^2 + 2\theta_3 x_1 x_2 + \theta_2 x_2^2 + \theta_4 x_1 + \theta_5 x_2 + \theta_6.$$

Let the hypotheses be

$$H_0 : g(x;\theta) = \text{linear in } x \qquad \text{and} \qquad H_1 : g(x;\theta) \text{ is concave in } x.$$

First note that the Hessian matrix of $g(x;\theta)$ with respect to x is $2[\theta_1, \theta_3 | \theta_3, \theta_2]$. Now, concavity of $g(x;\theta)$ is equivalent to the Hessian matrix, $(1/2)A(\theta)$, of $g(x;\theta)$ being negative semi-definite. Therefore, the hypotheses take the form $H_0 : A(\theta) = 0$ and $H_1 : A(\theta)$ *is positive semi-definite.*

Another example where this arises is when A is the covariance matrix of a collection of random effects; see Example 1.2.10 on page 14. ■

We can also define tests that resemble the W and S_G introduced earlier for linear constraints of θ. All of these test statistics differ only by a $o_p(1)$ term under the null hypothesis. Thus, they have the same asymptotic null distribution. Further, they also have the same asymptotic local power against Pitman-type local alternatives; this would require some conditions, for example, contiguity. Other types of asymptotic efficiencies of these tests have not been studied in any detail. As noted earlier, not much is known about their relative advantages and disadvantages. For completeness, they are introduced here.

For simplicity, we shall adopt the following notation. A function, say $h(\theta)$, evaluated at $\tilde{\theta}$ and $\bar{\theta}$ will be denoted by $\tilde{h}$ and $\bar{h}$, respectively; thus $\tilde{h} = h(\tilde{\theta})$ and $\bar{h} = h(\bar{\theta})$. Let $\check{H}$ and $\check{\mathcal{I}}$ denote consistent estimators of $H(\theta)$ and $\mathcal{I}(\theta)$ respectively, under the null hypothesis; therefore, they could be obtained by substituting $\hat{\theta}$, $\tilde{\theta}$ or $\bar{\theta}$ for θ.

(1) Wald-type test

This is very similar to the Wald-type test introduced earlier for linear constraints. Recall that the Wald-type statistic for testing $h(\theta) = 0$ against $h(\theta) \neq 0$ is

$n\hat{h}^T(\check{H}\check{\mathcal{I}}^{-1}\check{H}^T)^{-1}\hat{h}$. This suggests the statistic

$$W = n\tilde{h}^T(\check{H}\check{\mathcal{I}}^{-1}\check{H}^T)^{-1}\tilde{h}.$$

(2) Hausman-Wald type test

This simply evaluates the distance between the *mle*'s under H_0 and H_1.

$$W_H = n(\tilde{\theta} - \bar{\theta})^T\check{\mathcal{I}}(\tilde{\theta} - \bar{\theta}).$$

(3) Global score test:

Motivated by arguments given for the linear constraints setting, we define

$$S_G = n^{-1}U^T(\check{H}\check{\mathcal{I}}^{-1}\check{H}^T)^{-1}U, \quad \text{where } U = \check{H}\check{\mathcal{I}}^{-1}(\tilde{S} - \bar{S}).$$

Another, possible global score statistic is (see Robertson et al. (1988)),

$$n^{-1}(\tilde{S} - \bar{S})^T\check{\mathcal{I}}^{-1}(\tilde{S} - \bar{S}).$$

(4) Kühn-Tücker multiplier test

First, it would be instructive to provide an insight into the so called *Lagrange multiplier test*. Let us consider the optimization problem

$$\text{Problem P1:} \qquad \max \ell(\boldsymbol{\theta}) \qquad \text{subject to } \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{b}$$

where $\boldsymbol{b}$ is a given set of constraints. Let the Lagrangian function be defined as $\{\ell(\boldsymbol{\theta}) + \boldsymbol{\lambda}^T(\boldsymbol{h}(\boldsymbol{\theta}) - \boldsymbol{b})\}$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_r)^T$. Let the optimal solution be $(\bar{\boldsymbol{\theta}}_{\boldsymbol{b}} : \bar{\boldsymbol{\lambda}}_{\boldsymbol{b}})$; $\bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}$ is the Lagrange multiplier. Now, we have (for example, see Intriligator (1971) page 36)

$$\bar{\boldsymbol{\lambda}}_{\boldsymbol{b}} = -(\partial/\partial \boldsymbol{b})\ell(\bar{\boldsymbol{\theta}}_{\boldsymbol{b}}).$$

Thus, $\bar{\boldsymbol{\lambda}}_{\boldsymbol{b}}$ is a measure of the sensitivity of the constrained maximum of $\ell(\boldsymbol{\theta})$ over $\{\boldsymbol{\theta} : \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{b}\}$ to the value $\boldsymbol{b}$. If the unconstrained maximum of $\ell(\boldsymbol{\theta})$ is achieved over $\{\boldsymbol{\theta} : \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{b}\}$ then $\bar{\boldsymbol{\lambda}}_{\boldsymbol{b}} = \boldsymbol{0}$.

For brevity of notation, let us assume that $\boldsymbol{b} = \boldsymbol{0}$ and suppress the suffix $\boldsymbol{b}$ in $\boldsymbol{\lambda}_{\boldsymbol{b}}$. Aichison and Silvey (1958) showed that, under $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$, $\bar{\boldsymbol{\lambda}}/\sqrt{n}$ is asymptotically $N\{\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\theta}_0)\}$ where $\boldsymbol{V}(\boldsymbol{\theta}) = [\boldsymbol{H}\mathcal{I}^{-1}\boldsymbol{H}^T]_{\boldsymbol{\theta}}^{-1}$; it is asymptotically normal with nonzero mean under local alternatives. Therefore, it is reasonable to expect that the Lagrange multipliers have the potential to be building blocks for constructing a test of $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$ against $\boldsymbol{h}(\boldsymbol{\theta}) \neq \boldsymbol{0}$. The so called *Lagrange multiplier* statistic for testing $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$ vs $\boldsymbol{h}(\boldsymbol{\theta}) \neq \boldsymbol{0}$ is $n^{-1}\bar{\boldsymbol{\lambda}}^T\boldsymbol{V}(\bar{\boldsymbol{\theta}})^{-1}\bar{\boldsymbol{\lambda}}$. Since $\bar{\boldsymbol{\lambda}}$ satisfies $[(\partial/\partial\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \boldsymbol{H}(\boldsymbol{\theta})^T\bar{\boldsymbol{\lambda}}]_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \boldsymbol{0}$ we have

$$n^{-1}\bar{\boldsymbol{\lambda}}^T\boldsymbol{V}(\bar{\boldsymbol{\theta}})^{-1}\bar{\boldsymbol{\lambda}} = n^{-1}[\boldsymbol{S}(\boldsymbol{\theta})^T\mathcal{I}(\boldsymbol{\theta})^{-1}\boldsymbol{S}(\boldsymbol{\theta})]_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}};$$

the expression on the right-hand side is Rao's score statistic.

The same idea can be extended to construct tests when there are inequality constraints. For example, consider the optimization problem

$$\text{Problem P2:} \qquad \max \ell(\boldsymbol{\theta}) \qquad \text{subject to } \boldsymbol{h}(\boldsymbol{\theta}) \geqslant \boldsymbol{b}$$

Let the Lagrangian function be defined as above. The first-order Kühn-Tücker conditions (also known as Karush-Kühn-Tücker conditions) are

$$\nabla\ell(\tilde{\boldsymbol{\theta}}) + \nabla\boldsymbol{h}(\tilde{\boldsymbol{\theta}})^T\tilde{\boldsymbol{\lambda}} = \boldsymbol{0}, \boldsymbol{h}(\tilde{\boldsymbol{\theta}}) \geqslant \boldsymbol{0}, \tilde{\boldsymbol{\lambda}} \geqslant \boldsymbol{0}, \lambda_j h_j(\tilde{\boldsymbol{\theta}}) = 0 \text{ for } j = 1, \ldots, r.$$

Let the solution be $(\tilde{\boldsymbol{\theta}}_{\boldsymbol{b}}, \tilde{\boldsymbol{\lambda}}_{\boldsymbol{b}})$; $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{b}}$ are known as Kühn-Tücker multipliers. Again $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{b}}$ can also be interpreted as a measure of sensitivity of the constrained maximum of $\ell(\boldsymbol{\theta})$ to $\boldsymbol{b}$, in much the same way as for the foregoing equality constrained optimization problem (for details about this interpretation and its relevance, see Intriligator (1971) page 60).

Let $\bar{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\lambda}}$ denote the Kühn-Tücker multipliers associated with the maximization of $\ell(\boldsymbol{\theta})$ subject to $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$ and $\boldsymbol{h}(\boldsymbol{\theta}) \geqslant \boldsymbol{0}$, respectively. Now, the KT-statistic is defined as

$$KT = n(\tilde{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}})^T(\check{\boldsymbol{H}}\check{\mathcal{I}}^{-1}\check{\boldsymbol{H}})(\tilde{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}).$$

Now, the next result says that the foregoing tests are asymptotically equivalent.

Proposition 4.4.1 *Under the null hypothesis, $H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$, we have*

$$LRT = W + o_p(1) = W_H + o_p(1) = S_G + o_p(1) = D + o_p(1) = KT + o_p(1).$$

Proof: The proofs of these, except that for D, are given in Gourieroux and Monfort (1995, Chapter, 21); the equivalence of D to the *LRT* follows from the quadratic approximation (C) in Proposition 4.2.2. ■

An unattractive feature of W, W_H, S_G, and KT is that they depend on $\boldsymbol{H}(\boldsymbol{\theta})$ explicitly; further, some of the technical arguments also assume that $\boldsymbol{H}$ is of full rank. By contrast, *LRT* and distance-based tests do not depend on the choice of the function $\boldsymbol{h}$ to define a given set of constraints.

In some cases, it maybe difficult to compute the *mle* and/or the information matrix. In such cases, it may be of interest to construct a test based on an estimator that may be easier to obtain than the *mle*. Such a method is given in the next subsection.

4.4.2 Test of $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$ Against $\boldsymbol{h}_2(\boldsymbol{\theta}) \geqslant \boldsymbol{0}$

Let us consider the setting in Section 4.2. In particular, the parameter space Θ is open and Condition Q is satisfied. As in the previous subsection, let $\boldsymbol{h}$ be an $r \times 1$ vector function of $\boldsymbol{\theta}$; it is assumed that $\boldsymbol{h}$ is continuously differentiable. Let $\boldsymbol{H}(\boldsymbol{\theta})_{r\times p} = (\partial/\partial\boldsymbol{\theta}^T)\boldsymbol{h}(\boldsymbol{\theta})$ and assume that its rank is r for values of $\boldsymbol{\theta}$ in the null parameter space.

Let $\boldsymbol{h}$ be partitioned as $(\boldsymbol{h}_1 : \boldsymbol{h}_2)$ where $\boldsymbol{h}_2$ is $q \times 1$; recall that $(\boldsymbol{h}_1 : \boldsymbol{h}_2)$ denotes the partitioned vector $(\boldsymbol{h}_1^T, \boldsymbol{h}_2^T)^T$. In this section, we shall introduce a test of

$$H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0} \text{ against } H_1 : \boldsymbol{h}_2(\boldsymbol{\theta}) \geqslant \boldsymbol{0} \tag{4.44}$$

based on the basic idea that underlies the Wald test.

Let $\hat{\boldsymbol{\theta}}$ denote an estimator of $\boldsymbol{\theta}$ such that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, \boldsymbol{W}(\boldsymbol{\theta}_0)\}$$

for some $\boldsymbol{W}(\boldsymbol{\theta})$. For example, $\hat{\boldsymbol{\theta}}$ can be the unrestricted maximum likelihood estimator. It follows that

$$n^{1/2}\{\boldsymbol{h}(\hat{\boldsymbol{\theta}}) - \boldsymbol{h}(\boldsymbol{\theta}_0)\} \xrightarrow{d} N\{\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta}_0)\}, \tag{4.45}$$

where $\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta})^T$. Now, we can use some of the arguments used in subsection 4.3.3 for defining Wald-type and distance statistics. Let $\hat{\boldsymbol{V}}$ be a consistent estimator of $\boldsymbol{V}(\boldsymbol{\theta})$ under the null hypothesis, for example, $\hat{\boldsymbol{V}} = \boldsymbol{H}(\hat{\boldsymbol{\theta}})\boldsymbol{W}(\hat{\boldsymbol{\theta}})\boldsymbol{H}(\hat{\boldsymbol{\theta}})^T$.

A suitable statistic for testing H_0 against H_1 is

$$D = n[\boldsymbol{h}(\hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{h}(\hat{\boldsymbol{\theta}}) - \min_{\boldsymbol{a}}\{(\boldsymbol{h}(\hat{\boldsymbol{\theta}}) - \boldsymbol{a})^T \hat{\boldsymbol{V}}^{-1} (\boldsymbol{h}(\hat{\boldsymbol{\theta}}) - \boldsymbol{a}) : \boldsymbol{a}_2 \geqslant \mathbf{0}\}] \tag{4.46}$$

where $\boldsymbol{a}$ in $\mathbb{R}^r$ is partitioned as $(\boldsymbol{a}_1 : \boldsymbol{a}_2)$ to conform with $\boldsymbol{h} = (\boldsymbol{h}_1 : \boldsymbol{h}_2)$; see Kodde and Palm (1986). Note that $D = d\{\boldsymbol{h}(\hat{\boldsymbol{\theta}}), H_0\} - d\{\boldsymbol{h}(\hat{\boldsymbol{\theta}}), H_1\}$, where $d\{\boldsymbol{h}(\hat{\boldsymbol{\theta}}), H\}$ is a measure of the distance between $\boldsymbol{h}(\hat{\boldsymbol{\theta}})$ and H. The test rejects H_0 if D is large.

Suppose that H_0 is satisfied. It follows from $\{\boldsymbol{V}(\boldsymbol{\theta}_0) - \hat{\boldsymbol{V}})\} = o_p(1)$ that, under H_0, the value of D would change by $o_p(1)$ if $\hat{\boldsymbol{V}}$ is replaced by $\boldsymbol{V}(\boldsymbol{\theta}_0)$ (see Lemma 4.10.2 in the Appendix to this chapter). Therefore, the limiting distribution of D in (4.46) is equal to the distribution of

$$T = \boldsymbol{X}^T \boldsymbol{V}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{X} - \min_{\boldsymbol{a}}\{(\boldsymbol{X} - \boldsymbol{a})^T \boldsymbol{V}(\boldsymbol{\theta}_0)^{-1} (\boldsymbol{X} - \boldsymbol{a}) : \boldsymbol{a}_2 \geqslant \mathbf{0}\}$$

where $\boldsymbol{X} \sim N\{\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta}_0)\}$. Consequently, the asymptotic null distribution of D is given by

$$\text{pr}_{\boldsymbol{\theta}_0}\{D \geqslant c \mid H_0\} \to \sum_{i=0}^{q} w_i\{q, \boldsymbol{V}_{22}(\boldsymbol{\theta}_0)\}\text{pr}(\chi^2_{r-q+i} \geqslant c), \tag{4.47}$$

where $\boldsymbol{V}_{(22)}$ is the diagonal block of $\boldsymbol{V}$ corresponding to $\boldsymbol{h}_2$. Note that the large sample null distribution of D depends on the assumed true value $\boldsymbol{\theta}_0$ of the parameter $\boldsymbol{\theta}$ in the null parameter space. Approaches to dealing with such nuisance parameter problems were discussed in an earlier section.

If $\hat{\boldsymbol{\theta}}$ is the *mle*, then it follows from the quadratic approximations of the loglikelihood, that *LRT* and the other tests introduced in this subsection are asymptotically equivalent and hence their asymptotic null distribution is given by (4.47) with $\mathcal{I}(\boldsymbol{\theta})^{-1}$ substituted for $\boldsymbol{W}(\boldsymbol{\theta})$ in the formula for $\boldsymbol{V}(\boldsymbol{\theta})$.

4.5 AN OVERVIEW OF SCORE TESTS WITH NO INEQUALITY CONSTRAINTS

In this section, a brief summary of the relevant asymptotic results on likelihood ratio, Wald- and score-type tests is provided when there are no inequality constraints.

Although formal proofs are not given, the essential arguments are indicated. Details of the arguments may be found in Basawa (1985), Cox and Hinkley (1974), Sen and Singer (1993), and van der Vaart (2000). Hall and Mathiason (1990) provide a thorough account of many of these results at an intermediate level. These will be used in Section 4.6 to introduce local score and score-type tests against inequality constraints.

Let $f(y;\boldsymbol{\theta})$ denote the density of Y where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, and $Y_1, \ldots, Y_n$ denote n *iid* observations on the random variable/vector Y. It is assumed that Θ is open in $\mathbb{R}^p$. First, we shall consider the case of testing $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and then consider the more general case when there are nuisance parameters. In this section we will be studying the distribution of statistics when H_0 is not true. Therefore, the true value may not be $\boldsymbol{\theta}_0$.

4.5.1 Test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ Against $H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$

Let

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(Y_i;\boldsymbol{\theta}), \qquad \boldsymbol{S}(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})\ell(\boldsymbol{\theta})$$
$$\text{and} \quad \mathcal{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{(\partial/\partial\boldsymbol{\theta}) \log f(Y;\boldsymbol{\theta})(\partial/\partial\boldsymbol{\theta}^T) \log f(Y;\boldsymbol{\theta})\} \tag{4.48}$$

denote the loglikelihood, score function, and the information matrix, respectively. Assume that Condition Q in subsection 4.2 is satisfied. Let the null and alternative hypotheses be

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{and} \quad H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \tag{4.49}$$

respectively. Then, because $\log x \leq (x - 1)$ (or by Jensen's inequality), it follows that

$$E_{\boldsymbol{\theta}} \log\{f(Y;\boldsymbol{\theta}^*)/f(Y;\boldsymbol{\theta})\} \leq E_{\boldsymbol{\theta}}\{f(Y;\boldsymbol{\theta}^*)/f(Y;\boldsymbol{\theta})\} - 1 = 0,$$

and hence

$$\text{plim}_{\boldsymbol{\theta}} n^{-1}\ell(\boldsymbol{\theta}^*) = E_{\boldsymbol{\theta}} \log f(Y;\boldsymbol{\theta}^*) \leq E_{\boldsymbol{\theta}} \log f(Y;\boldsymbol{\theta}) = \text{plim}_{\boldsymbol{\theta}} n^{-1}\ell(\boldsymbol{\theta})$$

where $\text{plim}_{\boldsymbol{\theta}}$ is the probability limit when $\boldsymbol{\theta}$ is the true value. Therefore, $\text{plim}_{\boldsymbol{\theta}} n^{-1}\ell(\boldsymbol{\theta}^*)$ achieves its global maximum at $\boldsymbol{\theta}^* = \boldsymbol{\theta}$. Further, $\boldsymbol{\theta}^* = \boldsymbol{\theta}$ is a stationary point of $\text{plim}_{\boldsymbol{\theta}} n^{-1}\ell(\boldsymbol{\theta}^*)$. This suggests that it may be possible to use the slope of loglikelihood at $\boldsymbol{\theta}_0$ to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. To this end, note that

$$n^{-1}\boldsymbol{S}(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0} \text{ under } H_0, \text{ and } n^{-1}\boldsymbol{S}(\boldsymbol{\theta}_0) \xrightarrow{p} \boldsymbol{a}(\boldsymbol{\theta}) \text{ under } H_2 \tag{4.50}$$

where $\boldsymbol{a}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{(\partial/\partial\boldsymbol{\theta}) \log f(Y;\boldsymbol{\theta}_0)\}$; assume that $\boldsymbol{a}(\boldsymbol{\theta}) \neq \mathbf{0}$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. We also have $n^{-1/2}\boldsymbol{S}(\boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)\}$ under H_0. This suggests the *score statistic*

$$S_L = n^{-1}\boldsymbol{S}(\boldsymbol{\theta}_0)^T \mathcal{I}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{S}(\boldsymbol{\theta}_0) \tag{4.51}$$

for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs $H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Because $S_L \xrightarrow{d} \chi^2_p$ under H_0 and $S_L \xrightarrow{p} \infty$ under H_2, the score test rejects H_0 if S_L is large.

The *Wald statistic* for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_2 : \boldsymbol{\theta} \neq \mathbf{0}$ is defined as

$$T_W = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \hat{\mathcal{I}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{4.52}$$

where $\hat{\mathcal{I}}$ is a consistent estimator of $\mathcal{I}(\boldsymbol{\theta}_0)$. In the usual definition of a Wald statistic, the normalizing matrix $\hat{\mathcal{I}}$ is chosen so that it would be consistent irrespective of whether the null hypothesis is true or not; however, for the validity of the test, it suffices $\hat{\mathcal{I}}$ to be consistent under the null hypothesis only. Clearly, $T_W \xrightarrow{d} \chi^2_p$ under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $T_W \xrightarrow{p} \infty$ under $H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, and therefore the Wald test rejects H_0 if T_W is large.

Under $H_2 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, the probability of rejecting H_0 tends to 1 as $n \to \infty$ for the foregoing score and Wald tests at a fixed level. Similarly, at any fixed level, the power of many other tests at $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ also tend to 1. Therefore, to compare the performance of different tests, the limiting power of a test at a fixed level and at a fixed point in the alternative parameter space is not particularly helpful. The procedure adopted here to compare the asymptotic power of tests is the limiting power at *local alternatives* of the form $H_n : \boldsymbol{\theta} = (\boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{\delta})$ where $\boldsymbol{\delta}$ is fixed and the level of the test is also fixed. The sequence of hypotheses $H_n : \boldsymbol{\theta} = \boldsymbol{\theta}_o + n^{-1/2}\boldsymbol{\delta}$ are also known as Pitman-type local alternatives; for a thorough discussion of these see Serfling (1980) or Hájek et al. (1999). Here, we shall indicate the main arguments and results.

Let $\boldsymbol{\theta}_n = (\boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{\delta})$ where $\boldsymbol{\delta}$ is fixed. Let a sequence of local hypotheses be defined by $H_n : \boldsymbol{\theta} = \boldsymbol{\theta}_n$. Then, by a one-term Taylor expansion of $\boldsymbol{S}(\boldsymbol{\theta}_0)$ about $\boldsymbol{\theta}_n$, we have that

$$n^{-1/2}\boldsymbol{S}(\boldsymbol{\theta}_0) = n^{-1/2}\boldsymbol{S}(\boldsymbol{\theta}_n) + \mathcal{I}(\boldsymbol{\theta}_n)\boldsymbol{\delta} + o_p(1) \quad \text{under } H_n : \boldsymbol{\theta} = \boldsymbol{\theta}_n \tag{4.53}$$

Since $\mathcal{I}(\boldsymbol{\theta}_n)\boldsymbol{\delta} \xrightarrow{p} \mathcal{I}(\boldsymbol{\theta}_0)\boldsymbol{\delta}$ and $\boldsymbol{S}(\boldsymbol{\theta}_n)$ is a sum of *iid* random variables, we have that

$$n^{-1/2}\boldsymbol{S}(\boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathcal{I}(\boldsymbol{\theta}_0)\boldsymbol{\delta}, \mathcal{I}(\boldsymbol{\theta}_0)\} \quad \text{and} \quad S_L \xrightarrow{d} \chi^2_p(\Delta) \quad \text{under } H_n, \tag{4.54}$$

where $\Delta = \boldsymbol{\delta}^T \mathcal{I}(\boldsymbol{\theta}_0)\boldsymbol{\delta}$ is the noncentrality parameter. Further, by modifying the arguments to accommodate that $\boldsymbol{\theta}_n$ is the true value that may not be a fixed point, it can be shown that

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\boldsymbol{\delta}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}\} \text{ and } T_W \xrightarrow{d} \chi^2_p(\Delta) \quad \text{under } H_n \tag{4.55}$$

where $\Delta = \boldsymbol{\delta}^T \mathcal{I}(\boldsymbol{\theta}_0)\boldsymbol{\delta}$. Therefore, by (4.54) and (4.55), the score and Wald tests have the same asymptotic distribution under $H_n : \boldsymbol{\theta} = \boldsymbol{\theta}_n$ and hence the same limiting power against the sequences of local alternatives, H_n.

Now, *assume that $\boldsymbol{\theta}_0$ is the true value.* Then expansion of $n^{-1/2}\boldsymbol{S}(\hat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}_0$ leads to

$$n^{-1/2}\mathcal{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{S}(\boldsymbol{\theta}_0) = n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1). \tag{4.56}$$

Consequently, we can interchange $n^{-1/2}\mathcal{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{S}(\boldsymbol{\theta}_0)$ and $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ in many cases without affecting the first-oder asymptotic properties of the resulting test. Thus, we have that $T_W = S_L + o_p(1)$. An expansion of $\ell(\boldsymbol{\theta}_0)$ about $\hat{\boldsymbol{\theta}}$ leads to (see (4.4))

$$\ell(\boldsymbol{\theta}_0) = \ell(\hat{\boldsymbol{\theta}}) - (1/2)n(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) + o_p(1). \tag{4.57}$$

It follows from (4.51), (4.52), (4.56), and (4.57) that

$$LRT = T_W + o_p(1) = S_L + o_p(1) \qquad \text{under } H_0. \tag{4.58}$$

Consequently, LRT, T_W, and S_L, have the same asymptotic null distribution. It can also be shown that (for example, see Basawa (1985))

$$LRT = T_W + o_p(1) = S_L + o_p(1) \qquad \text{under } H_n. \tag{4.59}$$

This follows easily if H_n is contiguous to H_0; otherwise quadratic approximations can be used. Consequently, we have that

$$LRT, T_W, S_L \xrightarrow{d} \chi^2(\Delta) \qquad \text{under } H_n$$

where $\Delta = \boldsymbol{\delta}^T \mathcal{I}_{\boldsymbol{\theta}_0}\boldsymbol{\delta}$, and hence the three tests have the same local power. Usually, once we have shown that the difference between two test statistics is $o_p(1)$ at $\boldsymbol{\theta}_0$ in the null parameter space, then the same would also hold under $H_n : \boldsymbol{\theta} = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{\delta}$ as well. The proof of this may or may not be simple. For the *iid* setting, we can adopt an approach based on a quadratic approximation of $\ell(\boldsymbol{\theta})$ in $n^{-1/2}$-neighborhoods. A different approach would be to use methods based on contiguity of hypotheses.

Essentially the same types of arguments are used when the null hypothesis is not simple. These are discussed in the next subsection.

4.5.2 Test of $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ Against $H_2 : \boldsymbol{\psi} \neq \boldsymbol{\psi}_0$

The discussions in the previous subsection are particularly useful because the simplicity of the null hypothesis with no nuisance parameters enabled us to bring out some important features without delicate technical details. In many practical situations, tests of hypotheses usually involve only some components of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}$ be partitioned as $(\boldsymbol{\lambda} : \boldsymbol{\psi})$; recall that when $\boldsymbol{\theta}$ is partitioned as $(\boldsymbol{\lambda}^T, \boldsymbol{\psi}^T)^T$ it would be denoted by $(\boldsymbol{\lambda} : \boldsymbol{\psi})$. Let the null and alternative hypotheses be

$$H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0 \text{ and } H_2 : \boldsymbol{\psi} \neq \boldsymbol{\psi}_0. \tag{4.60}$$

For this testing problem, *LRT*, Wald, and score tests are known to be asymptotically equivalent.

Let $\boldsymbol{\lambda}_0$ be the true value of $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}_0 = (\boldsymbol{\lambda}_0 : \boldsymbol{\psi}_0)$ be the true value of $\boldsymbol{\theta}$ when $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ holds. Different extensions of the score statistic in the previous section are available for this testing problem. Let the score function $\boldsymbol{S}(\boldsymbol{\theta})$ in (4.48) be partitioned into $(\boldsymbol{S}_\lambda : \boldsymbol{S}_\psi)$ to conform with $(\boldsymbol{\lambda} : \boldsymbol{\psi})$. Let $\{\mathcal{I}_{\lambda\lambda}, \mathcal{I}_{\lambda\psi} \mid \mathcal{I}_{\psi\lambda}, \mathcal{I}_{\psi\psi}\}$ be obtained by partitioning the information matrix $\mathcal{I}$ to conform with $(\boldsymbol{\lambda} : \boldsymbol{\psi})$; similarly, let $\{\mathcal{I}^{\lambda\lambda}, \mathcal{I}^{\lambda\psi} \mid \mathcal{I}^{\psi\lambda}, \mathcal{I}^{\psi\psi}\}$ be obtained by partitioning $\mathcal{I}^{-1}$. Let

$$\mathcal{I}_{\psi\psi.\lambda} = (\mathcal{I}_{\psi\psi} - \mathcal{I}_{\psi\lambda}\mathcal{I}_{\lambda\lambda}^{-1}\mathcal{I}_{\lambda\psi});$$

then $\mathcal{I}^{\psi\psi} = \mathcal{I}^{-1}_{\psi\psi.\lambda}$.

Let $\bar{\lambda}$ denote the *mle* of λ under H_0, defined by $S_\lambda(\bar{\theta}) = 0$ where $\bar{\theta} = (\bar{\lambda} : \psi_0)$; thus $\bar{\theta}$ is the *mle* of θ under H_0. Let a sequence of local hypotheses be defined by

$$H_n : \psi = \psi_0 + n^{-1/2}\delta,$$

where $\delta \in \mathbb{R}^k$ is fixed. By employing Taylor series expansions similar to those leading to (4.53), we have

$$n^{-1/2}S_\psi(\bar{\theta}) \xrightarrow{d} N\{\mathcal{I}_{\psi\psi.\lambda}(\theta_0)\delta, \mathcal{I}_{\psi\psi.\lambda}(\theta_0)\} \quad \text{under } H_n. \tag{4.61}$$

Therefore, a score statistic for testing $H_0 : \psi = \psi_0$ vs $H_2 : \psi \neq \psi_0$ is

$$S_L = n^{-1}[S_\psi^T \mathcal{I}^{-1}_{\psi\psi.\lambda} S_\psi]_{\theta=\bar{\theta}} \tag{4.62}$$

where the suffix $\theta = \bar{\theta}$ for [.] indicates that [.] is evaluated at $\theta = \bar{\theta}$ (see Rao (1973), p 418). It follows from (4.61) that

$$S_L \xrightarrow{d} \chi^2_k(\Delta) \text{ under } H_n, \tag{4.63}$$

where $\Delta = \delta^T \mathcal{I}_{\psi\psi.\lambda}(\theta_0)\delta$. Because $\mathcal{I}(\theta_0)$ is positive definite, it follows that $\mathcal{I}(\theta_0)^{-1}$ and $\mathcal{I}_{\psi\psi.\lambda}(\theta_0)$ are positive definite, and hence $\Delta > 0$ for $\delta \neq 0$. Therefore, a test of $H_0 : \psi = \psi_0$ against $H_2 : \psi \neq \psi_0$ based on S_L in (4.62) rejects H_0 when it is large. This is the simple form of the score test of $\psi = \psi_0$ against $\psi \neq \psi_0$.

Because $n^{1/2}(\hat{\psi} - \psi_0) \xrightarrow{d} N\{0, \mathcal{I}^{-1}_{\psi\psi.\lambda}(\theta_0)\}$, the Wald statistic for testing $H_0 : \psi = \psi_0$ vs $H_2 : \psi \neq \psi_0$ is defined as

$$T_W = n(\hat{\psi} - \psi_0)^T \mathcal{I}_{\psi\psi.\lambda}(\hat{\theta})(\hat{\psi} - \psi_0). \tag{4.64}$$

Since $n^{1/2}(\hat{\psi} - \psi_0) \xrightarrow{d} N\{\delta, \mathcal{I}^{-1}_{\psi\psi.\lambda}(\theta_0)\}$ under H_n, it follows that

$$T_W \xrightarrow{d} \chi^2_k(\Delta) \text{ under } H_n.$$

Therefore, T_W in (4.64) and S_L in (4.62) have the same asymptotic power against H_n. In fact, by arguments essentially similar to those in the previous subsection, we have

$$LRT = S_L + o_p(1) = T_W + o_p(1) \text{ under } H_0 \text{ and under } H_n.$$

Thus, the three tests have the same asymptotic distribution under the null hypothesis, and also under the sequence of local alternatives, H_n; consequently, they have the same local power as well.

4.5.3 Tests Based on Estimating Equations

As in the previous subsection, let θ be partitioned as $(\lambda : \psi)$, and let the null and alternative hypotheses be the same as in the previous subsection, namely

$$H_0 : \psi = \psi_0 \quad \text{and} \quad H_2 : \psi \neq \psi_0,$$

respectively. Instead of defining $S(\theta) = (\partial/\partial\theta)\ell(\theta)$, where $\ell(\theta)$ is the loglikelihood, a much larger class of tests can be developed by allowing $S(\theta)$ to be a suitable estimating function. For example, for large sample inference on β in the linear regression model, $y = x^T\beta + \epsilon$, there is no need to assume that the errors are normal; the estimator of β defined as the solution of the normal equations can be used for inference on β without assuming normal error distribution. Other important special cases of the estimating equation approach includes (i) the class of quasi-likelihood equations, which are particularly useful in generalized linear models (see McCullagh and Nelder (1998, Chapter 9), and (ii) the Generalized Estimating Equation (GEE) approach, which has found use in many areas of applications. The essential results are mentioned below; they are based on Basawa (1985); for a brief summary of these see Silvapulle and Silvapulle (1995) and for an in-depth and detailed coverage of quasi-likelihood see Heyde (1997).

Let $S(\theta) = 0$ be a $p \times 1$ *estimating equation* for θ; $S(\theta)$ is called an *estimating function*. First, we shall state a set of conditions that are used in the literature to ensure that the choice of the estimating function is appropriate. However, later we shall relax this and state a simpler set of high-level sufficient conditions.

Assume that the following condition is satisfied

Condition A:
There exist nonsingular matrices $G(\theta)$ and $V(\theta)$ such that, for $a > 0$, as $n \to \infty$,

$$n^{-1/2}S(\theta_0) \xrightarrow{d} N\{0, V(\theta_0)\}, \tag{4.65}$$

$$\sup_{\|u\| \leq a} \|n^{-1/2}\{S(\theta_0 + n^{-1/2}u) - S(\theta_0)\} + G(\theta_0)u\| = o_p(1) \tag{4.66}$$

and the convergences in (4.65) and (4.66) are uniform in θ_0. ■

As in the previous subsection, let $\theta_0 = (\lambda_0 : \psi_0)$ denote the true value of θ when $H_0 : \psi = \psi_0$ holds. Let $\hat{\theta}$ be defined by $S(\hat{\theta}) = 0$. Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\{0, A(\theta_0)\}, \quad \text{where } A = G^{-1}VG^{-T} \text{ and } G^{-T} = (G^{-1})^T.$$

We can use $\hat{\theta}$ or $S(\theta)$ for testing hypotheses about θ. Let $(S_\lambda : S_\psi)$ be the partition of S to conform with $\theta = (\lambda : \psi)$. Similarly, let the partition of G and V be $\{G_{\lambda\lambda}, G_{\lambda\psi} \mid G_{\psi\lambda}, G_{\psi\psi}\}$ and $\{V_{\lambda\lambda}, V_{\lambda\psi}, \mid V_{\psi\lambda}, V_{\psi\psi}\}$, respectively. Define

$$C(\theta) = [(V_{\psi\psi} - G_{\psi\lambda}G^{-1}_{\lambda\lambda}V_{\lambda\psi}) - (V^T_{\lambda\psi} - G_{\psi\lambda}G^{-1}_{\lambda\lambda}V_{\lambda\lambda})G^{-T}_{\lambda\lambda}G^T_{\psi\lambda}]_\theta. \tag{4.67}$$

$$\text{and} \qquad Z(\theta) = n^{-1/2}[S_\psi - G_{\psi\lambda}G^{-1}_{\lambda\lambda}S_\lambda]_\theta. \tag{4.68}$$

The crucial result that enables us to construct local score-type tests is the following:

Proposition 4.5.1 *Let $\bar{\lambda}$ be the estimator of λ under H_0 defined by $S_\lambda(\bar{\lambda} : \psi_0) = 0$. Further, let $\bar{\theta} = (\bar{\lambda} : \psi_0)$ and let the sequence of local hypotheses be defined by $H_n : \psi = \psi_0 + n^{-1/2}\delta$ where δ is fixed. Then,*

1. *$Z(\bar{\theta}) = Z(\theta_0) + o_p(1)$ under H_0.*

2. $Z(\theta_0)$ *and* $Z(\bar{\theta}) \xrightarrow{d} N\{G_{\psi\psi.\lambda}(\theta_0)\delta, C(\theta_0)\}$ *under* H_n.

3. *For testing* $\psi = \psi_0$ *against* $\psi \neq \psi_0$ *a (local) score-type statistic is*

$$S_L = [Z^T C^{-1} Z]_{\bar{\theta}}. \tag{4.69}$$

The asymptotic distribution of S_L *in (4.69), under* H_n, *is* $\chi^2_k(\Delta)$, *where* $\Delta = \delta^T [G^T_{\psi\psi.\lambda} C^{-1} G_{\psi\psi.\lambda}]_{\theta_0} \delta$.

Proof: Let $B(\theta) = [-G_{\psi\lambda} G^{-1}_{\lambda\lambda}, I]_\theta$. Then $Z(\theta) = n^{-1/2} B(\theta) S(\theta)$ is a linear function of the statistic $S(\theta)$. It follows from (4.65) that $Z(\theta_0) \xrightarrow{d} N\{0, C(\theta_0)\}$ under H_0. One term Taylor expansion of $Z(\bar{\theta})$ about θ_0 leads to

$$Z(\bar{\theta}) - Z(\theta_0) \xrightarrow{p} 0 \quad \text{under } H_0.$$

It follows from (4.65) and (4.66) that $n^{-1/2} S(\theta_0) \xrightarrow{d} N\{G(\theta_0)\Delta, V(\theta_0)\}$ under $H_n : \psi = \psi_0 + n^{-1/2}\delta$, where $\Delta = (0 : \delta)$ and therefore,

$$Z(\bar{\theta}) \xrightarrow{d} N\{G_{\psi\psi.\lambda}(\theta_0)\delta, C(\theta_0)\} \text{ under } H_n : \psi = \psi_0 + n^{-1/2}\delta. \tag{4.70}$$

■

Now, consider the special case when $S(\theta)$ is the score function corresponding to loglikelihood. Let $\bar{S} = S(\bar{\theta})$ and $\bar{\mathcal{I}} = \mathcal{I}(\bar{\theta})$. Then $G(\theta) = V(\theta) = \mathcal{I}(\theta)$, and

$$Z(\bar{\theta}) = n^{-1/2}(\bar{S}_\psi - \bar{\mathcal{I}}_{\psi\lambda} \bar{\mathcal{I}}^{-1}_{\lambda\lambda} \bar{S}_\lambda) = n^{-1/2} \bar{S}_{\psi.}$$

In this case, the S_L in (4.69) reduces to Rao's score statistic (4.62) (see Rao (1973), p 418),

$$n^{-1} \bar{S}^T_\psi \bar{\mathcal{I}}^{-1}_{\psi\psi.\lambda} \bar{S}_\psi.$$

It follows from *Condition A* given at the beginning of this subsection that

$$n^{-1/2} G^{-1}(\theta_0) S(\theta_0) = n^{1/2}(\hat{\theta} - \theta_0) + o_p(1).$$

Because of this asymptotic linear relationship between $S(\theta_0)$ and $(\hat{\theta} - \theta_0)$ one would conjecture that a test based on $S(\theta_0)$ and one based on $\hat{\theta}$ are likely to be equivalent; this in fact is the case, as will be seen. It follows from *Condition A* that

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N\{0, A(\theta_0)\}, \quad \text{where } A = [G^{-1} V G^{-T}].$$

Therefore,

$$\sqrt{n}(\hat{\psi} - \psi_0) \xrightarrow{d} N\{0, A_{\psi\psi}(\theta_0)\},$$

and hence a Wald-type statistic for testing $H_0 : \psi = \psi_0$ against $H_1 : \psi \neq \psi_0$ is

$$T_W = n(\hat{\psi} - \psi_0)^T (\hat{A}_{\psi\psi})^{-1} (\hat{\psi} - \psi_0). \tag{4.71}$$

where $\hat{A}_{\psi\psi}$ is a consistent estimator of $A_{\psi\psi}$, for example, $A_{\psi\psi}(\hat{\theta})$. Now we have the following:

Proposition 4.5.2 *Under* $H_n : \psi = \psi_0 + n^{-1/2}\delta$, S_L *in (4.69) and* T_W *in (4.71) converge to* $\chi^2_k(\Delta)$ *where* $\Delta = \delta^T G^T_{\psi\psi.\lambda} C^{-1} G_{\psi\psi.\lambda} \delta$. ■

The definition of S_L in (4.69) depends on $\bar{\theta}$, which needs to be obtained by solving $S_\lambda(\theta) = 0$ under H_0. Similarly, T_W depends on $\hat{\theta}$, which is defined by $S(\hat{\theta}) = 0$. In large samples, the $\bar{\theta}$ and $\hat{\theta}$ in these definitions can be replaced by one-step estimators, which may be significantly easier to compute in complicated models. These one-step estimators are obtained by applying Newton-Raphson iteration once to a suitably chosen starting value. Usually they are asymptotically as efficient as the fully iterated solution and the difference between them is of order $o_p(n^{-1/2})$. The main steps are indicated below.

Let $t_{n\lambda}$ be a preliminary estimator of λ such that $n^{1/2}(t_{n\lambda} - \lambda_0) = O_p(1)$ and let $\bar{\lambda}^* = t_{n\lambda} + n^{-1} G^{-1}_{\lambda\lambda}(\theta^\dagger) S_\lambda(\theta^\dagger)$, where $\theta^\dagger = (t_{n\lambda} : \psi_0)$, and $\bar{\theta}^* = (\bar{\lambda}^* : \psi_0)$. Then, under H_0, $Z(\bar{\theta}^*) - Z(\bar{\theta}) = o_p(1)$. Consequently, the $\bar{\theta}$ in (4.69) can be replaced by $\bar{\theta}^*$ without affecting the large sample distribution of S_L under H_0. More details and references to the results of this subsection may be found in Basawa (1985).

4.6 LOCAL SCORE-TYPE TESTS OF $H_0 : \psi = 0$ AGAINST $H_1 : \psi \in \Psi$

In this section, the main results mentioned in section 4.5 will be used to develop a local score-type test of $\psi = 0$ against $\psi \in \Psi$ where Ψ is a given subset of $\mathbb{R}^k$ and $\theta = (\lambda : \psi)$. Initially, we shall consider the case when $\Psi = \mathcal{C}$, where $\mathcal{C}$ is a closed convex cone. In the last subsection, we shall consider the case when Ψ is not necessarily a cone. We shall consider the settings in subsections 4.5.1, 4.5.2, and 4.5.3 in turn. To introduce the ideas, we will first consider the special case of testing $\theta = 0$ against $\theta \in \mathcal{C}$ where $\mathcal{C}$ is a cone, and provide sufficient motivation for the definition of a suitable statistic. This is then extended in subsection 4.6.2 to the case when nuisance parameters are present. Then in subsection 4.6.3 we extend these and define a statistic based on estimating equations. In all these, the score-type statistic will be denoted by S_L, but the setting will become progressively more general. In the final subsection, it will be general enough to incorporate quasi-likelihood, partial likelihood, and estimating equations. This section is based on Silvapulle and Silvapulle (1995) with particular attention to the local score type tests introduced there.

4.6.1 Local Score Test of $H_0 : \theta = 0$ Against $H_1 : \theta \in \mathcal{C}$

Let $f(y; \theta)$ be a density function and let $Y_1, \ldots, Y_n$ be independently and identically distributed with common density function $f(y; \theta)$ where $\theta \in \Theta \subset \mathbb{R}^p$. Assume that $\mathbf{0}$ is an interior point of Θ. Let $\mathcal{I}_0$ denote $\mathcal{I}(\mathbf{0})$. Assume that Condition Q in subsection

4.2 is satisfied. Let the null and alternative hypotheses be

$$H_0 : \theta = 0 \text{ and } H_1 : \theta \in C, \tag{4.72}$$

respectively, where C is a closed convex cone. Let $S(\theta)$ denote the score function, $(\partial/\partial\theta)\ell(\theta)$, where $\ell(\theta)$ is the loglikelihood, and let

$$U = n^{-1/2}\mathcal{I}_0^{-1}S(0). \tag{4.73}$$

Let $\delta \in \mathbb{R}^p$ be a fixed point and $\theta_n = n^{-1/2}\delta$. It follows from (4.53) that

$$U \xrightarrow{d} N(\delta, \mathcal{I}_0^{-1}) \text{ under } H_n : \theta = \theta_n. \tag{4.74}$$

This can be used as a guide to constructing a test of $\theta = 0$ against $H_1 : \theta \in C$. To this end, suppose that the true value of θ is $n^{-1/2}\delta$. Because $\theta \in C$ is equivalent to $\delta \in C$, it may be adequate to construct a test of $\delta = 0$ against $\delta \in C$ for testing $\theta = 0$ against $\theta \in C$. In view of (4.74), U can be regarded as an estimator of δ. If the limiting distribution of U in (4.74) were the exact distribution of U for every n under H_n, then the likelihood ratio statistic for testing $\delta = 0$ against $\delta \in C$ based on a single observation of U would be

$$S_L = U^T\mathcal{I}_0U - \min\{(U-a)^T\mathcal{I}_0(U-a) : a \in C\}. \tag{4.75}$$

The arguments leading to this definition of S_L are based on the local properties of U and $S(0)$. Therefore, these motivations are closely related to those leading to the classical Rao's score statistic. We shall refer to S_L as the *local score statistic.*

The foregoing local arguments were used only as a motivation and guide to constructing S_L. Whether or not it is a good statistic would depend on its properties and performance. Therefore, it would be of interest to compare the asymptotic performance of S_L with other statistics. Because $n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N\{0, \mathcal{I}(\theta_0)^{-1}\}$, where θ_0 is the true value, a distance statistic for testing $\theta = 0$ against $\theta \in C$ is

$$D = n[\hat{\theta}^T\hat{\mathcal{I}}\hat{\theta} - \min\{(\hat{\theta}-a)^T\hat{\mathcal{I}}(\hat{\theta}-a) : a \in C\}], \tag{4.76}$$

where $\hat{\mathcal{I}}$ is a consistent estimator of $\mathcal{I}_0$. It follows from (4.56), (4.75), and (4.76) that $S_L = D + o_p(1)$ under H_0. Therefore, S_L and D are asymptotically equivalent for testing $\theta = 0$ against $\theta \in C$. These are summarized in the following proposition.

Proposition 4.6.1 *Assume that Condition Q is satisfied. Let U, S_L, D, and the hypotheses be as in (4.73), (4.75), (4.76), and (4.72), respectively. Then, under H_0,*

$$LRT = S_L + o_p(1) = D + o_p(1). \tag{4.77}$$

Further, the common asymptotic null distribution of LRT, S_L and D is $\bar{\chi}^2(\mathcal{I}_0^{-1}, C)$.

A consequence of this result is that the local score statistic, the likelihood ratio statistic, and the distance statistic have the same asymptotic null distribution. Further, they also have the same local power if (4.77) holds under $H_n : \theta = n^{-1/2}\delta$, which would be the case under fairly general conditions. In this sense all three tests are asymptotically equivalent. Since there are no nuisance parameters under H_0, S_L may be computed without estimating the model. By contrast, the *LRT* requires the model to be estimated subject to the inequality constraints of H_1, and D requires estimation of the full model without the constraints.

Since the Rao score statistic is based on the score $S(0)$ as the basic building block and thinking of it as the slope of $\ell(\theta)$ at the null value, one might be tempted to define a score statistic for testing $\theta = 0$ vs $\theta \in C$ by replacing the role of U in (4.75) by $S(0)$. This leads to the statistic

$$R^* = S_0^T\mathcal{I}_0^{-1}S_0 - \min\{(S_0-a)^T\mathcal{I}_0^{-1}(S_0-a) : a \in C\}.$$

Since $n^{-1/2}S(0) \xrightarrow{d} N(\mathcal{I}_0\delta, \mathcal{I}_0)$ under H_n, this essentially assumes that a test of $\theta = 0$ vs $\mathcal{I}_0\theta \in C$ would also be suitable for testing $\theta = 0$ vs $\theta \in C$. Since this is not a reasonable assumption, S_L appears to be better than R^*.

4.6.2 Local Score Test of $H_0 : \psi = 0$ Against $H_1 : \psi \in C$

It is very rarely that a null hypothesis would specify a value for every component of θ. Therefore, the testing problem considered in the previous subsection is very rare, although the discussions therein were meant to be instructive. In this subsection, we consider the more realistic setting in which the hypotheses impose constraints on some components only. The main ideas leading to the foregoing local score statistic, S_L, can also be applied when there are nuisance parameters.

Let $\theta = (\lambda : \psi)$, where ψ is $k \times 1$, and let the null and alternative hypotheses be

$$H_0 : \psi = 0 \quad \text{and} \quad H_1 : \psi \in C, \tag{4.78}$$

respectively, where C is a closed convex cone in $\mathbb{R}^k$. It is convenient to adopt the notation in subsection 4.5.2. Let $S(\theta)$ denote the score function $(\partial/\partial\theta)\ell(\theta)$ and let it also be partitioned as $(S_\lambda : S_\psi)$. Let $\bar{\lambda}$ denote the *mle* of λ under H_0 defined by $S_\lambda(\bar{\lambda} : 0) = 0$ and let $\bar{\theta} = (\bar{\lambda} : 0)$ be the *mle* of θ under H_0. Define

$$U = n^{-1/2}\mathcal{I}^{\psi\psi}(\bar{\theta})S_\psi(\bar{\theta}). \tag{4.79}$$

Let δ be a fixed point in $\mathbb{R}^k$. Then, since $\mathcal{I}_{\psi\psi.\lambda} = (\mathcal{I}_{\psi\psi} - \mathcal{I}_{\psi\lambda}\mathcal{I}_{\lambda\lambda}^{-1}\mathcal{I}_{\lambda\psi})$ and $(\mathcal{I}_{\psi\psi})^{-1} = \mathcal{I}_{\psi\psi.\lambda}$, we have that (see (4.61))

$$U \xrightarrow{d} N\{\delta, \mathcal{I}_{\psi\psi.\lambda}^{-1}(\theta_0)\}, \text{ under } H_n : \psi = n^{-1/2}\delta, \tag{4.80}$$

where $\theta_0 = (\lambda_0 : 0)$, the true value of θ under H_0. Now, by arguments similar to those leading to (4.75), we define the *local score statistic* as

$$S_L = U^T\bar{\mathcal{I}}_{\psi\psi.\lambda}U - \inf\{(U-a)^T\bar{\mathcal{I}}_{\psi\psi.\lambda}(U-a) : a \in C\} \tag{4.81}$$

for testing (4.78), where $\bar{\mathcal{I}} = \mathcal{I}(\bar{\theta})$.

Because $n^{1/2}(\hat{\psi}-\psi_0) \xrightarrow{d} N\{0, \mathcal{I}^{-1}_{\psi\psi\cdot\lambda}(\theta_0)\}$, a distance statistic for testing H_0 : $\psi = 0$ vs $H_1 : \psi \in C$ is

$$D = n[\hat{\psi}^T \hat{\mathcal{I}}_{\psi\psi\cdot\lambda}\hat{\psi} - \min\{(\hat{\psi}-a)^T\hat{\mathcal{I}}_{\psi\psi\cdot\lambda}(\hat{\psi}-a) : a \in C\}] \tag{4.82}$$

where $\hat{\mathcal{I}}_{\psi\psi\cdot\lambda} = \mathcal{I}_{\psi\psi\cdot\lambda}(\hat{\theta})$ which is a consistent estimator of $\mathcal{I}_{\psi\psi\cdot\lambda}(\theta_0)$. Now, we have the following result, which says that the local score statistic S_L in (4.81) is asymptotically equivalent to the likelihood ratio and distance statistics; see Silvapulle and Silvapulle (1995) for a proof.

Proposition 4.6.2 *Let U, S_L, D, and the hypotheses be as in (4.79), (4.81), (4.82), and (4.78), respectively. Then, under H_0,*

$$LRT = S_L + o_p(1) = D + o_p(1). \tag{4.83}$$

Further, the common asymptotic null distribution of LRT, S_L, and D is $\bar{\chi}^2(\mathcal{I}^{-1}_{\psi\psi\cdot\lambda}(\theta_0), C)$.

If the alternative hypothesis in (4.78) is $H_1 : \psi \neq 0$ then S_L in (4.81) reduces to the Rao score statistic; further, since S_L is asymptotically equivalent to the *LRT* and they both are based on the same building block, the S_L in (4.81) can be considered as a natural generalization of Rao's score statistic.

4.6.3 Local Sore-Type Test Based on Estimating Equations

By employing arguments similar to those in the previous subsection, the main results in subsection 4.5.3 can be extended to introduce a test based on estimating equations. Let $\theta = (\lambda : \psi)$ where ψ is $k \times 1$ and $S(\theta) = 0$ be an estimating equation for θ. Initially, we shall assume that Condition A on page 173 is satisfied. In the next subsection, we will relax this further to develop a large class of tests. Consequently, the procedure becomes more flexible; for example, quasi-likelihood, partial likelihood, and Generalized Estimating Equations (GEE) are incorporated.

Let the null and alternative hypotheses be

$$H_0 : \psi = 0 \text{ and } H_1 : \psi \in C, \tag{4.84}$$

respectively, where C is a closed convex cone and $\theta_0 = (\lambda_0 : 0)$, the true value of θ when H_0 is true. Let $\bar{\theta} = (\bar{\lambda} : 0)$ denote the estimator obtained by solving $S_\lambda(\lambda : 0) = 0$; i.e., solve $S_\lambda(\theta) = 0$ under H_0. Let

$$U = \bar{G}^{-1}_{\psi\psi\cdot\lambda}\bar{Z}, \tag{4.85}$$

where $\bar{G} = G(\bar{\theta})$, $\bar{Z} = Z(\bar{\theta})$, and $G(.)$ and $Z(.)$ are as in subsection 4.5.3. Then, by Proposition 4.5.1,

$$U \xrightarrow{d} N\{\delta, A_{\psi\psi}(\theta_0)\} \quad \text{under } H_n : \psi = n^{-1/2}\delta, \tag{4.86}$$

where $A_{\psi\psi} = G^{-1}_{\psi\psi\cdot\lambda}C(G^{-1}_{\psi\psi\cdot\lambda})^T$. Therefore, we define the local score-type test statistic for testing $H_0 : \psi = 0$ against $H_1 : \psi \in C$ as

$$S_L = U^T\bar{A}^{-1}_{\psi\psi}U - \min\{(U-a)^T\bar{A}^{-1}_{\psi\psi}(U-a) : a \in C\}, \tag{4.87}$$

where $\bar{A}_{\psi\psi} = A_{\psi\psi}(\bar{\theta})$. Further, the limiting distribution of S_L is $\bar{\chi}^2\{A_{\psi\psi}(\theta_0), C\}$ under H_0.

If $S(\theta)$ is the likelihood score function $(\partial/\partial\theta)\ell(\theta)$, then $G(\theta) = V(\theta) = \mathcal{I}(\theta)$ and

$$U = n^{1/2}\bar{\mathcal{I}}^{\psi\psi}(\bar{S}_\psi - \bar{\mathcal{I}}_{\psi\lambda}\bar{\mathcal{I}}^{-1}_{\lambda\lambda}\bar{S}_\lambda),$$

where $\bar{S} = S(\bar{\theta})$ and $\bar{\mathcal{I}} = \mathcal{I}(\bar{\theta})$; consequently, the S_L in (4.87) reduces to the local score statistic in the previous subsection (see 4.81). Further, if the alternative hypothesis is $\psi \neq 0$ then S_L reduces to

$$n^{-1}\{\bar{\mathcal{I}}^{\psi\psi}(\bar{S}_\psi - \bar{\mathcal{I}}_{\psi\lambda}\bar{\mathcal{I}}^{-1}_{\lambda\lambda}\bar{S}_\lambda)\}^T\bar{\mathcal{I}}^{-1}_{\psi\psi\cdot\lambda}\{\bar{\mathcal{I}}^{\psi\psi}(\bar{S}_\psi - \bar{\mathcal{I}}_{\psi\lambda}\bar{\mathcal{I}}^{-1}_{\lambda\lambda}\bar{S}_\lambda)\},$$

which is Neyman's $C(\alpha)$-statistic. In view of these observations, the local score-type statistic, S_L, in (4.87) may be seen as a generalization of the $C(\alpha)$-statistic. Since $S(\theta)$ is not necessarily the likelihood score function, the S_L in (4.87) and the *LRT* may not be equivalent.

4.6.4 A General Local Score-Type Test of $H_0 : \psi = 0$

This section assumes familiarity with the definition and properties of *approximating cone* introduced later in this chapter. However, the topic of this subsection fits in better in this section, because it is a direct extension of the results in the previous subsections. An approximating cone is purely a mathematical object. The introduction to this topic in Section 4.7 can be read independently; readers who are not familiar with it (i.e., approximating cone) may wish to read about it before reading the rest of this section.

So far we have assumed that the alternative hypothesis is of the form $\psi \in C$ where C is a closed convex cone. The results in the previous subsection extend in a natural way to provide a large class of tests of $\psi = 0$ against quite general alternatives based on quantities that are not necessarily the likelihood score function. Let $0 \in \Psi \subset \mathbb{R}^k$, and let the testing problem be

$$H_0 : \psi = 0 \quad \text{vs} \quad H_1 : \psi \in \Psi.$$

Although this formulation allows 0 to be an interior point of Ψ, our main focus will be on the case when it is on the boundary of Ψ. Let $\mathcal{A}$ denote the *approximating cone* of Ψ at the null value 0. If the null hypothesis is formulated as $H_0 : \psi = \psi_0$, then $\mathcal{A}$ should be the approximating cone of Ψ at ψ_0. Although we will continue to use the notation $\theta = (\lambda : \psi)$, here we allow λ to be an infinite dimensional nuisance parameter. Thus, *semiparametric* models are also incorporated in this subsection. Let $\theta_0 = (\lambda_0 : 0)$ and let $\delta \in \mathcal{A}$ be arbitrary but fixed. Now, let us introduce the following condition:

Condition LS:

1. There exists a statistic U (i.e., a function of the data only) and a positive definite matrix $D(\theta_0)$ such that

$$U \xrightarrow{d} N\{\delta, D(\theta_0)\} \text{ under } H_n : \psi = n^{-1/2}\delta.$$

2. There exists a consistent estimator, $\bar{D}$ of $D(\theta_0)$. ■

Assume that this condition is satisfied. Then, a local score-type statistic for testing $H_0 : \psi = 0$ against $H_1 : \psi \in \Psi$ is

$$S_L = U^T \bar{D}^{-1} U - \inf_{b \in \mathcal{A}} (U - b)^T \bar{D}^{-1} (U - b). \tag{4.88}$$

This is the general form of the local score-type statistic introduced in Silvapulle and Silvapulle (1995) and further developed in Silvapulle et al. (2002) for the semiparametric model. If $\mathcal{A}$ is a closed convex cone then the limiting distribution of S_L is given by

$$S_L \xrightarrow{d} \bar{\chi}^2\{D(\theta_0), \mathcal{A}\} \text{ under } H_0.$$

The asymptotic distribution of S_L under H_n is equal to the expression on the right-hand side (4.88) with $U \sim N\{\delta, D(\theta_0)\}$ and $\bar{D}$ replaced by $D(\theta_0)$. If λ is finite dimensional and U is the efficient score then S_L is asymptotically equivalent to the likelihood ratio test [see Theorem 1 in Silvapulle and Silvapulle (1995)].

The test statistic in (4.88) provides a very flexible class of simple tests because its implementation requires only the null model to be estimated and it is based on essentially the same building blocks as those used for the traditional score tests. The applicability of the foregoing local score test in several semiparametric models (including linear regression, AR(p), ARCH, and nonlinear regression models) in which the error distribution is treated as an unknown infinite dimensional nuisance parameter, is discussed in Silvapulle et al. (2002) where *adaptive tests* and *efficient semiparametric tests* are developed. The results in Silvapulle et al. (2002) suggest that once an adaptive test is obtained for testing against an unconstrained hypothesis, it is likely that the same test can be extended to incorporate restricted alternatives as well. The reason for this is that a semiparametric test against an unrestricted alternative is likely to be based on a U satisfying Condition LS.

4.6.5 A Simple One-Dimensional Test of $\psi = 0$ Against $\psi \in C$

The local score-type statistic, S_L, in (4.87) is based on the distribution of the score vector, S, and its asymptotic null distribution would typically be a chi-bar-square. Although the computation of this may be considerably easier than that of the *LRT*, it is not as easy as the usual t-ratio for a single parameter. There is a need for a test that is simpler than S_L even if it is less powerful. To this end, we introduce the following simple test.

Let the testing problem be $H_0 : \psi = 0$ vs $H_1 : \psi \in C$, where C is a closed convex cone. Assume that there exists a vector c in C such that $c^T x > 0$ for every $x \in C \setminus \{0\}$. For example, if C is the positive orthant, then c could be the center direction $(1, \ldots, 1)^T$; if C is a polyhedral generated by $a_1, \ldots, a_k$ then it suffices to find a c such that $a_i^T c > 0$ for $i = 1, \ldots, k$. Now, the basic idea is to adopt the working assumption that $\psi = \eta c$, and test $\eta = 0$ against $\eta > 0$. It appears reasonable to expect that a test that can detect a departure in the direction of c would also be able to detect departures in other directions of C that are close to c as well.

Let U be as in (4.85). Then, it follows from (4.86) that

$$c^T U \xrightarrow{d} N\{c^T \delta, c^T A_{\psi\psi}(\theta_0) c\} \quad \text{under } H_n : \psi = n^{-1/2}\delta. \tag{4.89}$$

Because c was chosen such that $c^T \delta \geqslant 0$ for $\delta \in C$, a simple test of $H_0 : \psi = 0$ against $H_1 : \psi \in C$ rejects H_0 for large values of T_c where

$$T_c = c^T U / \{c^T \bar{A}_{\psi\psi} c\}^{1/2}, \tag{4.90}$$

and $\bar{A}_{\psi\psi}$ is a consistent estimator of $A_{\psi\psi}(\theta_0)$. Clearly, this test is considerably simpler than the ones in the previous subsections. The cost of simplicity is that this test is unlikely to be as powerful as the local score test, S_L, on average, although T_c is likely to be more powerful than S_L in the direction of c.

A closely related procedure was suggested by King and Wu (1997); this test has some locally optimal properties. For this test, let $S(\theta)$ be the likelihood score function. It follows from the foregoing results that

$$n^{-1/2} c^T S_\psi(\bar{\theta}) \xrightarrow{d} N\{c^T \mathcal{I}_{\psi\psi.\lambda}(\theta_0)\delta, c^T \mathcal{I}_{\psi\psi.\lambda}(\theta_0) c\} \quad \text{under } H_n : \psi = n^{-1/2}\delta.$$

Now, suppose that the following is satisfied: $c^T \mathcal{I}_{\psi\psi.\lambda}(\theta_0)\delta > 0$ for $\delta \in C \setminus \{0\}$. This condition would be satisfied if the elements of $\mathcal{I}_{\psi\psi.\lambda}(\theta_0)$ are nonnegative and $c = (1, \ldots, 1)^T$. Then a test of H_0 against H_1 rejects H_0 if

$$n^{-1/2} c^T S_\psi(\bar{\theta}) / \{c^T \bar{\mathcal{I}}_{\psi\psi.\lambda} c\}^{1/2}$$

is large compared with the critical value from the standard normal distribution. For an example to illustrate the application of this to ARCH/GARCH models in time series, see Lee and King (1993); see also King and Wu (1997).

4.6.6 A Data Example: Test Against ARCH Effect in an ARCH Model.

A common feature of many financial time series such as commodity prices, exchange rates, and stock returns is that periods of large changes tend to cluster together, and similarly periods of small changes tend to cluster together. Because high variability is associated with high risk, studying the variability of such time series is important. The AutoRegressive Conditional Heteroscedasticity (ARCH) model and its various generalizations/modifications are now widely used in modeling such volatility clustering patterns. For a comprehensive account of ARCH modeling see Gourieroux (1997).

Let $\mathcal{F}_t$ denote the information available up to time t, Y_t denote the dependent variable, and $\boldsymbol{x}_t$ be a column vector of explanatory variables. Typically, $\boldsymbol{x}_t$ includes lagged values of the dependent variable. The pth order ARCH model of Engle (1982) can be stated as

$$Y_t = \boldsymbol{x}_t^T\boldsymbol{\beta} + \epsilon_t, E(\epsilon_t|\mathcal{F}_{t-1}) = 0, \text{ and } h_t = \alpha_0 + \psi_1\epsilon_{t-1}^2 + \ldots + \psi_p\epsilon_{t-p}^2, \quad (4.91)$$

where $h_t = var(Y_t|\mathcal{F}_{t-1})$. Note that (4.91) provides a specification not only for the mean but also for the variance. If ψ_j in (4.91) is negative for some j ($j = 1, \ldots, p$), then a large value for ϵ_{t-j} would result in a negative variance for $Y_t|\mathcal{F}_{t-1}$. Therefore, the admissible range for $\psi_1, \ldots, \psi_p$ is $\{\psi_1 \geqslant 0, \ldots, \psi_p \geqslant 0\}$. In the context of this example, the following question is important: "is the variance, h_t, constant over time?" As a general rule, this testing problem is formulated as

$$H_0 : \psi_1 = \ldots = \psi_p = 0 \quad \text{against} \quad H_1 : \psi_1 \geqslant 0, \ldots, \psi_p \geqslant 0. \quad (4.92)$$

In this data example, we consider an ARCH model for an overall rate of return from stocks. Let $Y_t = \log(I_t/I_{t-1})$ where I_t is the *All Ordinaries Index (Australia)* on day t. This index is a weighted average of the prices of selected shares in Australia; it corresponds to the *Dow Jones Index* in the United States. The data that we used are for the period 5/January/1984 to 29/November/1985. This provides a total of 484 observations; see Silvapulle and Silvapulle (1995) for the data. For illustrative purposes, we chose $p = 3$; the method would be applicable for other values of p. The specific model that we consider is

$$\boldsymbol{x}_t^T\boldsymbol{\beta} = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_4 Y_{t-4} \text{ and } h_t = \alpha_0 + \psi_1\epsilon_{t-1}^2 + \psi_2\epsilon_{t-2}^2 + \psi_3\epsilon_{t-3}^2. \quad (4.93)$$

Let $\boldsymbol{\theta} = (\boldsymbol{\lambda} : \boldsymbol{\psi})$ where $\boldsymbol{\lambda} = (\beta_0, \ldots, \beta_4, \alpha_0)^T$ and $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3)^T$. The null and alternative hypotheses are

$$H_0 : \boldsymbol{\psi} = \mathbf{0} \text{ and } H_1 : \boldsymbol{\psi} \geqslant \mathbf{0}.$$

Let $\boldsymbol{S}(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})L(\boldsymbol{\theta})$ where $L(\boldsymbol{\theta}) = -(1/2)\sum\{\log(2\pi h_t) + \epsilon_t^2/h_t\}$; thus $L(\boldsymbol{\theta})$ is the loglikelihood if we were to assume that $Y_t|\mathcal{F}_{t-1} \sim N(\boldsymbol{x}_t^T\boldsymbol{\beta}, h_t)$. Now, let us compute the local score-type test statistic S_L in (4.87). First, we need to estimate the nuisance parameter under the null hypothesis. Therefore, we solve the equation $\boldsymbol{S}_{n\lambda}(\bar{\boldsymbol{\theta}}) = \mathbf{0}$ where $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\lambda}} : \mathbf{0})$. Under the null hypothesis, the error variance is constant, and therefore the null model is estimated by ordinary least squares. Thus we have that $\bar{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and $\bar{\alpha}_0 = n^{-1}\sum(Y_t - \boldsymbol{x}_t^T\bar{\boldsymbol{\beta}})^2$. Now, $\bar{\boldsymbol{\lambda}} = (\bar{\boldsymbol{\beta}} : \bar{\alpha}_0)$. Note that to apply the local score-type test statistic S_L, this is the only estimation required despite the fact that the ARCH model itself is rather involved. The next step is to compute $\bar{\boldsymbol{G}}$, $\bar{\boldsymbol{V}}$, $\bar{\boldsymbol{Z}}$, $\boldsymbol{U}$, and $\bar{\boldsymbol{A}}_{\psi\psi}$. Closed form expressions for these quantities are obtained directly from the closed-form available for $\boldsymbol{S}(\boldsymbol{\theta})$; see Silvapulle and Silvapulle (1995) for these formulas.

The computed value of $\boldsymbol{U}$ is $(3.56, 1.30, 2.20)^T$. Because every component of $\boldsymbol{U}$ is positive, $\min\{(\boldsymbol{U} - \boldsymbol{a})^T\bar{\boldsymbol{A}}_{\psi\psi}^{-1}(\boldsymbol{U} - \boldsymbol{a}) : \boldsymbol{a} \geqslant \mathbf{0}\} = 0$ and hence

$$S_L = \boldsymbol{U}^T\bar{\boldsymbol{A}}_{\psi\psi}^{-1}\boldsymbol{U} - \min\{(\boldsymbol{U} - \boldsymbol{a})^T\bar{\boldsymbol{A}}_{\psi\psi}^{-1}(\boldsymbol{U} - \boldsymbol{a}) : \boldsymbol{a} \geqslant \mathbf{0}\} = \boldsymbol{U}^T\bar{\boldsymbol{A}}_{\psi\psi}^{-1}\boldsymbol{U} = 4.36.$$

Let

$$\zeta(\boldsymbol{\lambda}) = [\sum_{i=1}^{3} w_i\{3, \boldsymbol{A}_{\psi\psi}(\boldsymbol{\theta})\}\text{pr}(\chi_j^2 \geqslant 4.36)]_{\boldsymbol{\theta}=(\boldsymbol{\lambda}:\mathbf{0})}.$$

Using the closed forms given in the last chapter for the chi-bar square weights in this expression (see (3.25)), we can compute $sup_\lambda\zeta(\boldsymbol{\lambda})$. However, before we do so, let us apply a bounds test. It follows from (3.23) that

$$\zeta(\boldsymbol{\lambda}) \leq 0.5\{\text{pr}(\chi_2^2 \geqslant 4.36) + \text{pr}(\chi_3^2 \geqslant 4.36)\} = 0.17.$$

This upper bound is too high to reject H_0. Therefore, we computed $p_{sup} = sup_\lambda\zeta(\boldsymbol{\lambda})$ and $\hat{p} = \zeta(\bar{\boldsymbol{\lambda}})$. The computed values are 0.10 and 0.085, respectively. Thus, it appears that there may be some evidence in favor of ARCH effect, but it is weak.

For testing $H_0 : \boldsymbol{\psi} = \mathbf{0}$ against $H_1 : \boldsymbol{\psi} \geqslant \mathbf{0}$, it is also valid to apply an unrestricted test ignoring the inequality constraints in H_1. To this end, let the testing problem be

$$H_0 : \boldsymbol{\psi} = \mathbf{0} \quad \text{vs} \quad H_2 : \boldsymbol{\psi} \neq \mathbf{0}.$$

Even though the ARCH model would not be well-defined with a negative value for any of the components of $\boldsymbol{\psi}$, it is valid to apply a score-type test of $H_0 : \boldsymbol{\psi} = \mathbf{0}$ against $H_2 : \boldsymbol{\psi} \neq \mathbf{0}$; in fact, this is the procedure that is found in much of the econometrics literature. The only information the score-type statistic $\boldsymbol{U}^T\bar{\boldsymbol{A}}_{\psi\psi}^{-1}\boldsymbol{U}$ uses is that it is approximately $N(0, 1)$ under H_0; if H_0 is not true then it is expected to be large. The sample value of this is

$$\boldsymbol{U}^T\bar{\boldsymbol{A}}_{\psi\psi}^{-1}\boldsymbol{U} = 4.36;$$

and its large sample p-value is $\text{pr}(\chi_3^2 \geqslant 4.36) = 0.22$. As expected, because the components of $\boldsymbol{U}$ turned out to be all positive, the statistics for the one-sided and two-sided tests have the same numerical value and the latter has a larger p-value.

4.7 APPROXIMATING CONES AND TANGENT CONES

Most of the results in the previous sections have generalizations to the case when the parameter space takes more general forms. Typically, such parameter spaces arise because the constraints are nonlinear in the parameter. It turns out that the main first-order results, namely the asymptotic distribution of an estimator, the asymptotic null distribution of a test statistic, and the asymptotic local power of a test are unchanged when the null and alternative parameter spaces are replaced by cones that approximate them at the true value of the parameter provided that the boundaries of the parameter spaces are sufficiently smooth. The objective of this section is to introduce the basics that are required to study the asymptotics when the parameter spaces are neither linear spaces nor cones; for example, the parameter space may have boundaries that are curved or the slope of the boundary may be discontinuous (i.e., kinked).

As an example, let $\Omega = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 \geqslant g_1(\theta_1) \text{ and } \theta_2 \geqslant g_2(\theta_1)\}$ where $g_1(\theta_1) = \theta_1^2 + \theta_1$ and $g_2(\theta_1) = \theta_1^2 - \theta_1$. In Fig. 4.1, O is the origin; the two

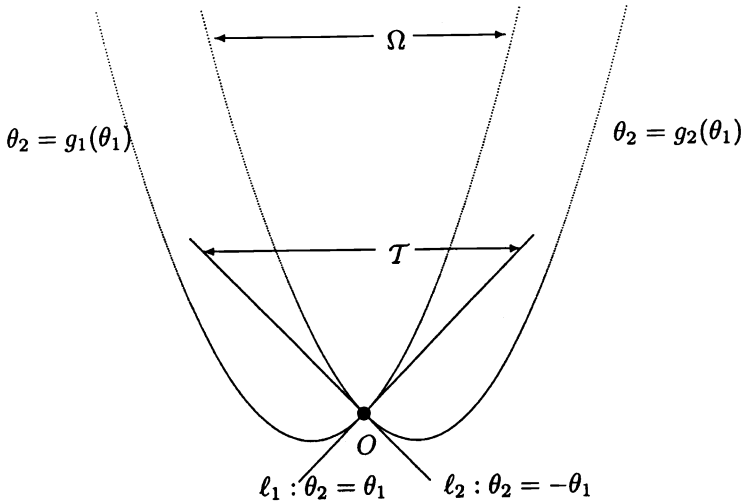

Fig. 4.1 The *cone of tangents*, $\mathcal{T}$, of Ω at O.

parabolas, $\theta_2 = g_1(\theta_1)$ and $\theta_2 = g_2(\theta_1)$, are also shown. Further, Ω is the region bounded by these two parabolas, as shown.

Let ℓ_1 and ℓ_2 be the tangents at O to the two smooth curves that form the boundary of Ω; thus the equations for ℓ_1 and ℓ_2 are $\theta_2 = \theta_1$ and $\theta_2 = -\theta_1$, respectively. Let $\mathcal{T} = \{\theta : \theta_2 \geqslant \theta_1 \text{ and } \theta_2 \geqslant -\theta_1\}$. Locally at the origin, Ω is close to the cone $\mathcal{T}$. In this example, $\mathcal{T}$ is known as the *linearizing cone* of Ω at O, because it is obtained by linearizing the boundaries defining the set Ω at O. The cone is also called the *contingent cone* and *cone of tangents* of Ω at O. The formal definitions of these terms will be introduced in the next subsection.

Let $\boldsymbol{\theta}_0 (\neq \mathbf{0})$ be a given fixed point in $\mathbb{R}^p$. Now, let us move $\{\Omega, \ell_1, \ell_2, \mathcal{T}\}$ by $\boldsymbol{\theta}_0$, and let the new objects be denoted by $\{\Theta, \ell_1', \ell_2', \mathcal{A}\}$ (see Fig. 4.2). Therefore, $\Theta = \boldsymbol{\theta}_0 + \Omega, \mathcal{A} = \boldsymbol{\theta}_0 + \mathcal{T}$, ℓ_1' and ℓ_2' are the tangents to the boundaries of Θ at $\boldsymbol{\theta}_0$, and $\mathcal{A}$ is the cone formed by ℓ_1' and ℓ_2'. Then $\mathcal{A}$ is close to Θ at $\boldsymbol{\theta}_0$, in the same way as $\mathcal{T}$ is close to Ω at O. The cone $\mathcal{A}$ with vertex at $\boldsymbol{\theta}_0$ is called the *approximating cone* of Θ at $\boldsymbol{\theta}_0$. An important result relating to the approximating cone is that if Θ is the parameter space and $\boldsymbol{\theta}_0$ is the true value then Θ can be replaced by its approximating cone $\mathcal{A}$ at $\boldsymbol{\theta}_0$ for obtaining the asymptotic null distribution of *LRT* and other first order asymptotic results when $\boldsymbol{\theta}_0$ is the true value.

Example 4.7.1 *Inference on normal mean when the parameter space is not a cone.*

Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be *iid* as the bivariate normal distribution, $N(\boldsymbol{\theta}, \boldsymbol{I})$, where $\boldsymbol{\theta} \in \mathbb{R}^2$ and let $\ell(\boldsymbol{\theta})$ denote the kernel of the loglikelihood $-(n/2)\|\bar{\boldsymbol{Y}} - \boldsymbol{\theta}\|^2$. Let the null and alternative hypotheses be

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \text{and} \qquad H_1 : \boldsymbol{\theta} \in \Theta,$$

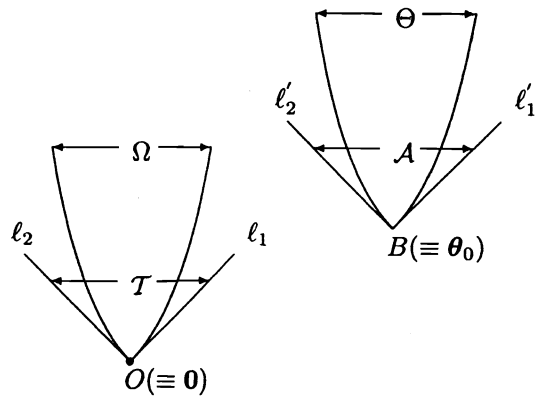

Fig. 4.2 The *cone of tangents*, $\mathcal{T}$, and the *approximating cone*, $\mathcal{A}$, of Θ at B.

respectively, where $\boldsymbol{\theta}_0 = (1,1)^T$, $\Theta = \boldsymbol{\theta}_0 + \Omega$ and Ω is as in Fig. 4.2. Let $\hat{\boldsymbol{\theta}}_\Theta$, $\hat{\boldsymbol{\theta}}_\Omega$ and $\hat{\boldsymbol{\theta}}_\mathcal{T}$ denote the *mle* over Θ, Ω, and $\mathcal{T}$, respectively. Then

$$LRT = 2\{\ell(\hat{\boldsymbol{\theta}}_\Theta) - \ell(\boldsymbol{\theta}_0)\} = n\{\|\bar{\boldsymbol{Y}} - \boldsymbol{\theta}_0\|^2 - \|\bar{\boldsymbol{Y}} - \Theta\|^2\}$$

where $\|\bar{\boldsymbol{Y}} - \Theta\|^2 = \inf_{\boldsymbol{\theta}\in\Theta}(\bar{\boldsymbol{Y}} - \boldsymbol{\theta})^T(\bar{\boldsymbol{Y}} - \boldsymbol{\theta})$ is the squared distance between $\bar{\boldsymbol{Y}}$ and Θ. Suppose that the true value of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$. Now, the results to be presented later in this section show that the asymptotic distribution of the constrained *mle* $\hat{\boldsymbol{\theta}}_\Theta$ and that of the *LRT* of $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $\boldsymbol{\theta} \in \Theta$ remain the same even if the constrained parameter space Θ is replaced by $\mathcal{A}$, the approximating cone of Θ at $\boldsymbol{\theta}_0$. More specifically,

$$\begin{aligned} LRT &= n\{\|\bar{\boldsymbol{Y}} - \boldsymbol{\theta}_0\|^2 - \|\bar{\boldsymbol{Y}} - \Theta\|^2\} \\ &= n\{\|\bar{\boldsymbol{Y}} - \boldsymbol{\theta}_0\|^2 - \|\bar{\boldsymbol{Y}} - \mathcal{A}\|^2\} + o_p(1), \qquad (4.94) \\ \text{and} \quad & n^{1/2}(\hat{\boldsymbol{\theta}}_\Theta - \hat{\boldsymbol{\theta}}_\mathcal{A}) = o_p(1). \end{aligned}$$

Further, the constrained estimator $\hat{\boldsymbol{\theta}}_\Theta$ and the projection of the unconstrained estimator $\hat{\boldsymbol{\theta}}$ onto the approximating cone have the same asymptotic distributions in the following sense:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_\Theta - \boldsymbol{\theta}_0) \xrightarrow{d} \Pi(\boldsymbol{Z}, \mathcal{T}) \quad \text{and} \quad \sqrt{n}\{\Pi(\hat{\boldsymbol{\theta}}, \mathcal{A}) - \boldsymbol{\theta}_0\} \xrightarrow{d} \Pi(\boldsymbol{Z}, \mathcal{T}),$$

where $\boldsymbol{Z} \sim N(\mathbf{0}, \boldsymbol{I})$. ■

It is worth noting that the process of replacing Θ by $\mathcal{A}$ is equivalent to replacing the nonlinear constraints defining Θ by their linear approximations at the null value. In the following subsections, we will extend these ideas to more general settings. To do so, one important requirement is that the parameter space must satisfy certain regularity conditions. These conditions must be satisfied irrespective of the properties of the likelihood function and other stochastic conditions. These are discussed in the next subsection.

Before we introduce regularity conditions relating to parameter spaces, let us make a remark regarding our terminology. Recall that a cone is defined to have vertex at the

origin unless the contrary is clear (see Appendix 1). According to our definition, the approximating cone does not necessarily have its vertex at the origin. Some authors refer to the $\mathcal{T}$ in Fig. 4.2 as the approximating cone of Θ at $\boldsymbol{\theta}_0$. We shall deviate from this terminology because it would be helpful to identify $\mathcal{T}$ and $\mathcal{A}$ separately using different terms; we shall refer to them as tangent cone of Θ at $\boldsymbol{\theta}_0$ and approximating cone of Θ at $\boldsymbol{\theta}_0$, respectively.

4.7.1 Chernoff Regularity

Let $\Theta \subset \mathbb{R}^p$ and $\boldsymbol{\theta}_0 \in \Theta$. A vector $\boldsymbol{w} \in \mathbb{R}^p$ is said to be a *tangent* to Θ at θ_o if there exists a sequence $\{\boldsymbol{\theta}_n\}$ in Θ and a sequence of positive numbers $\{t_n\}$ converging to zero such that $t_n^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \to \boldsymbol{w}$ as $n \to \infty$. This essentially means that $\boldsymbol{w} \in \mathbb{R}^p$ is a tangent to Θ at $\boldsymbol{\theta}_0$ if either $\boldsymbol{w} = \mathbf{0}$ or there exists a sequence of points in Θ converging to $\boldsymbol{\theta}_0$ such that the direction of the vector from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_n$ converges to the direction of $\boldsymbol{w}$. Therefore, a tangent at $\boldsymbol{\theta}_o$ is a vector rather than the set of points on a line passing through $\boldsymbol{\theta}_0$. This definition of a tangent is more general than the usual one where it is seen as a line that touches a smooth curve at a point. The following examples illustrate this generalized notion of a tangent.

Example 4.7.2 *Illustration of a generalized notion of tangents*

1. For $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 = \theta_1^2\}$, the tangents at $\mathbf{0}$ are $k(1,0)$ and $k(-1,0)$ where $k \geqslant 0$ (see Fig. 4.3). We do not view the line $\theta_2 = 0$ as a tangent.

2. For $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 = \theta_1^2 \text{ and } \theta_1 \geqslant 0\}$ the only tangent at $\mathbf{0}$ is $k(1,0)$ where $k \geqslant 0$; note that $(-1,0)$ is not a tangent at $\mathbf{0}$.

3. For $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 \geqslant \theta_1^2\}$, the tangents at $\mathbf{0}$ are (θ_1, θ_2) where $\theta_2 \geqslant 0$ and $\theta_1 \in \mathbb{R}$ (see Fig. 4.4). Thus, the tangents to the curve $\theta_2 = \theta_1^2$, and those to the area bounded by the curve are different.

4. Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 = \sin(1/\theta_1) \text{ and } \theta_1 > 0\} \cup \{\mathbf{0}\}$. The tangents to Θ at $\mathbf{0}$ are $k\boldsymbol{\theta}$ where $\boldsymbol{\theta} = (\theta_1, \theta_2), \theta_1 > 0$ and $k \geqslant 0$. Thus, even though $(d\theta_2/d\theta_1)$ does not exist at the origin, the set $\{\boldsymbol{\theta} : \theta_2 = \sin(1/\theta_1)\}$ has an infinite number of tangents at the same point. ■

The *cone of tangents,* denoted $\mathcal{T}(\Theta; \boldsymbol{\theta}_0)$, of Θ at $\boldsymbol{\theta}_0$ is defined as the set of all tangents to Θ at $\boldsymbol{\theta}_0$. This cone is also called the *contingent cone, Boulingard tangent cone,* and the *ordinary tangent cone.* Note that $\mathcal{T}(\Theta; \boldsymbol{\theta}_0)$ has vertex at the origin irrespective of whether or not $\boldsymbol{\theta}_0$ is the origin. A tangent $\boldsymbol{v}$ is said to be *derivable* if there exists a function $\boldsymbol{f} : [0, \epsilon] \to \Theta$ for some $\epsilon > 0$ such that $\boldsymbol{f}(0) = \boldsymbol{\theta}_0$ and $(d/dt+)\boldsymbol{f}(0) = \boldsymbol{v}$. Intuitively, this means that there is a smooth curve in Θ with one end at $\boldsymbol{\theta}_0$ such that the slope of the curve at $\boldsymbol{\theta}_0$ is parallel to $\boldsymbol{v}$. The set of all derivable tangents to Θ at $\boldsymbol{\theta}_0$ is called the *derived tangent cone* of Θ at $\boldsymbol{\theta}_0$. It is also called the *cone of attainable directions*; see Bazaraa et al (1993).

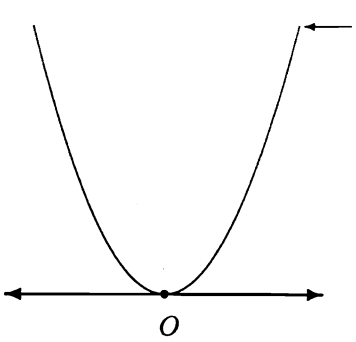

Fig. 4.3 The tangents to $\{\theta_2 = \theta_1^2\}$ at O.

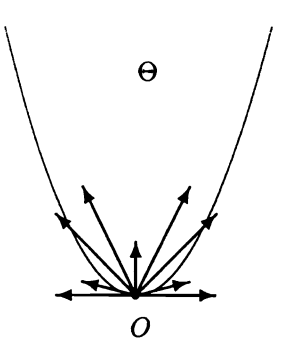

Fig. 4.4 Tangents to $\{\theta_2 \geqslant \theta_1^2\}$ at O.

While it is not essential for us at this stage, let us also state equivalent definitions of the cones just introduced. The derived tangent cone is equal to

$$\{\boldsymbol{v} : \forall t_n \downarrow 0, \exists \boldsymbol{\theta}_n \in \Theta \text{ such that } \boldsymbol{\theta}_n \to \boldsymbol{\theta}_0 \text{ and } t_n^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \to \boldsymbol{v}\},$$

and the cone of tangents is equal to

$$\{\boldsymbol{v} : \exists t_n \downarrow 0, \exists \boldsymbol{\theta}_n \in \Theta \text{ such that } \boldsymbol{\theta}_n \to \boldsymbol{\theta}_0 \text{ and } t_n^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \to \boldsymbol{v}\}.$$

It is clear from these definitions that the derived tangent cone is contained in the cone of tangents, because the former must satisfy a condition for all $t_n \downarrow 0$ while the cone of tangents must satisfy the same condition for some $t_n \downarrow 0$.

The following example of (Rockafellar and Wets, 1998, p 199) is helpful to illustrate the difference between the cone of tangents and the derived tangent cone.

Example 4.7.3 *Illustration of a difference between cone of tangents and the derived tangent cone*

Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 = \theta_1 \sin(1/\theta_1), \theta_1 > 0\} \cup \{\mathbf{0}\}$. Note that there is no curve that lies in Θ and is smooth at $\mathbf{0}$. Therefore, the only member of the derived tangent cone is $\{\mathbf{0}\}$. By contrast, the straight line $\theta_2 = k\theta_1$ where $|k| \leq 1$, intersects with Θ at a sequence of points $\{\boldsymbol{\theta}_n\}$ and $\|\boldsymbol{\theta}_n\|^{-1}\boldsymbol{\theta}_n \to (1, k)$. Therefore, the cone of tangents is $\{\boldsymbol{\theta} \in \mathbb{R}^2 : |\theta_2| \leq \theta_1, \theta_1 > 0\}$. ■

The foregoing definitions and the example suggest that if the cone of tangents and the derived tangent cone of Θ at $\boldsymbol{\theta}_0$ are not equal, then Θ is likely to be quite irregular near $\boldsymbol{\theta}_0$. A set $\Theta \subset \mathbb{R}^p$ is said to be *Chernoff regular* at $\boldsymbol{\theta}_0$ if the cone of tangents and the derived tangent cone of Θ at $\boldsymbol{\theta}_0$ are equal. If Θ is Chernoff regular at $\boldsymbol{\theta}_0$ then we shall refer to the *cone of tangents* and the *derived tangent cone* as *tangent cone.*

An intuitively simple way of verifying Chernoff regularity of Θ at $\boldsymbol{\theta}_0$ is the following (this is due to Rockafellar and Wets (1998), page 198). Consider the set $\mathcal{A}_t = t(\Theta - \boldsymbol{\theta}_0)$ where $t > 0$. Now take the limit as t increases to ∞ (see Fig. 4.5). Note that $\mathcal{A}_t$ provides a magnified view of Θ near $\boldsymbol{\theta}_0$ because it is like viewing Θ

through a magnifying glass focused at θ_0. As t increases to ∞, the magnification increases to ∞ and if the magnified view of Θ at θ_0 converges to a limit (i.e., if $\mathcal{A}_t$ has a limit) then Θ is Chernoff regular at θ_0, and the limit is the tangent cone of Θ at θ_0.

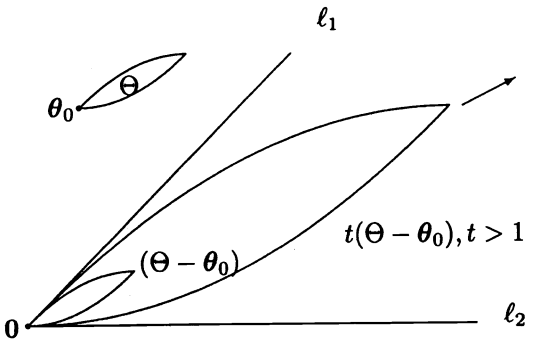

Fig. 4.5 The limit of $t(\Theta-\theta_0)$ as $t\to\infty$ is the cone between the lines ℓ_1 and ℓ_2, and it is equal to the tangent cone, $\mathcal{T}(\Theta;\theta_0)$, of Θ at θ_0.
Reproduced from: *Variational Analysis*, Rockafellar and Wets, Copyright (1998), with permission from Springer, New York.

Let $\Theta\subset\mathbb{R}^p$ and $\theta_0\in\Theta$. The set Θ is said to be *approximated by a cone $\mathcal{A}$ at θ_0* if

$$\begin{aligned}&(a)\quad d(\theta,\mathcal{A})=o(\|\theta-\theta_0\|)\text{ for }\theta\in\Theta,\\ \text{and }&(b)\quad d(x,\Theta)=o(\|x-\theta_0\|)\text{ for }x\in\mathcal{A},\end{aligned}\tag{4.95}$$

where d is the distance between a point and a set defined by

$$d(\theta,A)=\inf_{x\in A}\|\theta-x\|.$$

In this case, $\mathcal{A}$ is called an *approximating cone of Θ at θ_0* and will be denoted by $\mathcal{A}(\Theta_0;\theta_0)$.

This definition was introduced by Chernoff (1954) and it is illustrated in Fig. 4.6 using two different shapes for Θ. In each of the two cases, let Θ be the region bounded by the two curves as shown and O be the point corresponding to θ_0. In the diagram on the right-hand side, A is a point in Θ. Let $\theta=OA$ and B be the projection of A onto the approximating cone $\mathcal{A}$; therefore, $\|AB\|=d(\theta,\mathcal{A})$. Condition (a) of the definition (4.95) says that $\|AB\|/\|OA\|\to 0$ as A converges to O while staying in Θ. In the diagram on the left-hand side, A is a point in $\mathcal{A}$ and let $x=OA$. Let B be the projection of A on to Θ. Condition (b) of the definition (4.95) says that $\|AB\|/\|OA\|\to 0$ as A converges to O while staying in $\mathcal{A}$.

Andrews (1999, page 1358) noted that the foregoing definition of approximating cone can be stated using Hausdorff distance; see also Le Cam (1970, p 819). For two sets A and B, define the *Hausdorff distance* between A and B by

$$h(A,B)=\max\{\sup_{b\in B}d(b,A),\sup_{a\in A}d(a,B)\}.$$

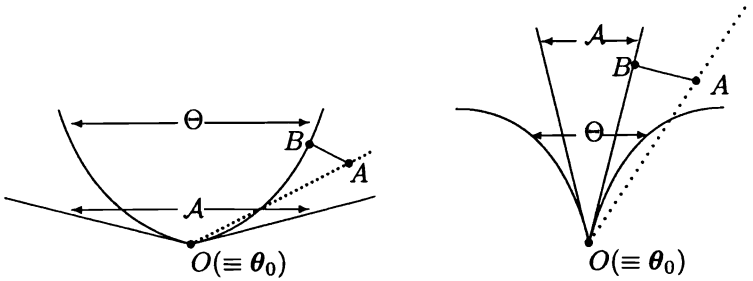

Fig. 4.6 Chernoff's definition of approximating cone $\mathcal{A}$ of Θ at θ_0 : $\|AB\|=o(\|OA\|$

Now, the next result provides an equivalent form of definition (4.95).

Proposition 4.7.1 *Let the setting be as in* (4.95) *and* $B_\epsilon=\{\theta:\|\theta-\theta_0\|<\epsilon\}$, *where* $\epsilon>0$. *Let $\mathcal{A}$ be a cone with vertex at θ_0. Then $\mathcal{A}$ is an approximating cone of Θ at θ_0 as defined by* (4.95) *if and only if*

$$h\{\Theta\cap B_\epsilon,\mathcal{A}\cap B_\epsilon\}=o(\epsilon)\quad\text{as }\epsilon\searrow 0.\tag{4.96}$$

Proof: For a given $\epsilon>0$ let $\Theta_\epsilon=\Theta\cap B_\epsilon,\mathcal{A}_\epsilon=\mathcal{A}\cap B_\epsilon,\theta_\epsilon=\arg\max_{\theta\in\Theta_\epsilon}d(\theta,\mathcal{A}_\epsilon)$, and $x_\epsilon=\arg\max_{x\in\mathcal{A}_\epsilon}d(x,\Theta_\epsilon)$. Then, the Hausdorff distance between Θ_ϵ and $\mathcal{A}_\epsilon$ takes the form

$$h(\Theta_\epsilon,\mathcal{A}_\epsilon)=\max\{d(\theta_\epsilon,\mathcal{A}_\epsilon),d(x_\epsilon,\Theta_\epsilon)\}.$$

Now suppose that $\mathcal{A}$ is an approximating cone according to Chernoff's definition (4.95). Then

$$d(\theta_\epsilon,\mathcal{A}_\epsilon)=o(\|\theta_\epsilon-\theta_0\|)=o(\epsilon);$$

the first equality follows by (4.95) and the second follows since $\|\theta_\epsilon-\theta_0\|\le\epsilon$. Similarly, $d(x_\epsilon,\Theta_\epsilon)=o(\epsilon)$. Therefore, $h(\Theta_\epsilon,\mathcal{A}_\epsilon)=o(\epsilon)$.

Now to prove the converse, suppose that (4.96) is satisfied. Let $\theta\in\Theta$ and $\delta=\|\theta-\theta_0\|$. Then,

$$d(\theta,\mathcal{A})=d(\theta,\mathcal{A}_\delta)\le d(\theta_\delta,\mathcal{A}_\delta)=o(\delta);$$

the first equality follows because $\|\Pi(\theta,\mathcal{A})-\theta_0\|\le\|\theta-\theta_0\|$ and hence $\Pi(\theta,\mathcal{A})=\Pi(\theta,\mathcal{A})$, the middle inequality follows by the definition of θ_δ, and the third equality follows by (4.96). ■

For a discussion about Hausdorff distance in statistics, see van der Vaart and Wellner (1996, p 162). Andrews (1999, p 1358) extended the definition of Chernoff (1954) to define a sequence of sets being approximated by a cone; such a generalization may be required/helpful in more general settings. We say that a sequence of sets $\{\Theta_n\}$ is locally approximated at θ_0 by a cone $\mathcal{A}$ if

$$\begin{aligned}&(a)\quad d(\theta_n,\mathcal{A})=o(\|\theta_n-\theta_0\|)\text{ for }\theta_n\in\Theta_n\text{ and }\|\theta_n-\theta_0\|\to 0\\ \text{and }&(b)\quad d(x_n,\Theta_n)=o(\|x_n-\theta_0\|)\text{ for }x_n\in\mathcal{A}\text{ and }\|x_n-\theta_0\|\to 0.\end{aligned}\tag{4.97}$$

Now the foregoing results for Θ also hold for Θ_n with appropriate modifications.

Throughout this book, parameter spaces will be assumed to be Chernoff regular unless the contrary is made clear; this is a very mild condition as was indicated earlier. The following important result, due to Geyer (1994, Theorem 2.1), establishes the link between the notion of approximating cone introduced by Chernoff and several tangent cones that have been known in the mathematics literature.

Proposition 4.7.2 *(Geyer 1994). The cone $\mathcal{K}$ is an approximating cone of Θ at $\boldsymbol{\theta}_0$ if and only if the derived tangent cone and the cone of tangents are equal and*

$$\textit{Closure of } \mathcal{K} = \boldsymbol{\theta}_0 + \mathcal{T}(\Theta; \boldsymbol{\theta}_0) = \boldsymbol{\theta}_0 + \textit{ Derived tangent cone of } \Theta \textit{ at } \boldsymbol{\theta}_0.$$

■

This essentially says that Θ has an approximating cone at $\boldsymbol{\theta}_0$ if an only if Θ is Chernoff regular at $\boldsymbol{\theta}_0$.

The following result provides a convenient way of obtaining the approximating cone of Θ at $\boldsymbol{\theta}_0$ when Θ is defined by a set of nonlinear equality and inequality constraints, in many cases; the proof is given in the appendix to this chapter.

Proposition 4.7.3 *Suppose that Ω is open and let*

$$\Theta = \{\boldsymbol{\theta} \in \Omega : h_1(\boldsymbol{\theta}) = \ldots = h_\ell(\boldsymbol{\theta}) = 0, h_{\ell+1}(\boldsymbol{\theta}) \geqslant 0, \ldots, h_k(\boldsymbol{\theta}) \geqslant 0\},$$

where $h_1, \ldots, h_k$ are continuously differentiable. Let $\boldsymbol{\theta}_0$ be a point in Θ, and let $\boldsymbol{a}_i = (\partial/\partial\boldsymbol{\theta})h_i(\boldsymbol{\theta}_0)$ for $i = 1, \ldots, k$, and $J(\boldsymbol{\theta}_0) = \{i : h_i(\boldsymbol{\theta}_0) = 0$ and $\ell + 1 \leq i \leq k\}$. Assume that the following condition, known as the Mangasarian-Fromowitz constraint qualification (MF-CQ), is satisfied at $\boldsymbol{\theta}_0$: There exists a nonzero $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{a}_1^T\boldsymbol{b} = \ldots = \boldsymbol{a}_\ell^T\boldsymbol{b} = 0$, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_\ell$ are linearly independent and $\boldsymbol{a}_i^T\boldsymbol{b} > 0$ for $i \in J(\boldsymbol{\theta}_0)$. Then $\mathcal{A}(\Theta; \boldsymbol{\theta}_0)$ is equal to

$$\{\boldsymbol{\theta} \in \mathbb{R}^p : \boldsymbol{a}_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = 0 \textit{ for } i = 1, \ldots, \ell; \boldsymbol{a}_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \geqslant 0 \textit{ for } i \in J(\boldsymbol{\theta}_0)\}.$$

The foregoing result shows that if MF-CQ is satisfied then the approximating cone at $\boldsymbol{\theta}_0$ is obtained by substituting the first order approximation (i.e., linear approximation at $\boldsymbol{\theta}_0$), $\{h_i(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T(\partial/\partial\boldsymbol{\theta})h_i(\boldsymbol{\theta}_0)\}$, of $h_i(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$ if $h_i(\boldsymbol{\theta}_0) = 0$; if $h_s(\boldsymbol{\theta}_0) > 0$ (i.e., $h_s(\boldsymbol{\theta}_0) \geq 0$ is inactive) then the constraint $h_s(\boldsymbol{\theta}_0) \geqslant 0$ does not play a part in determining the approximating cone at $\boldsymbol{\theta}_0$ ($s = \ell + 1, \ldots, k$).

The approximating cone obtained by applying the result in the previous proposition is a polyhedral. Thus, if the approximating cone is not a polyhedral then this method of obtaining the approximating cone would not be applicable in its current form. For example, in Fig. 4.1 if $\Theta = A \cup B$ where A is the region bounded by $\theta_2 = g_2(\theta_1)$ and $\theta_2 = \theta_1$, and B is the region bounded by $\theta_2 = g_1(\theta_1)$ and $\theta_2 = -\theta_1$, then the approximating cone of Θ at O is not convex. Therefore, for this Θ, the method in Proposition 4.7.3 is not directly applicable; in any case, this would be obvious because Θ is not of the form required by Proposition 4.7.3. It can be seen that A and B are

of the form required by Proposition 4.7.3 and $\mathcal{A}(A \cup B; \mathbf{0}) = \mathcal{A}(A; \mathbf{0}) \cup \mathcal{A}(B; \mathbf{0})$; now the method based on MF-CQ can be applied to A and B separately to obtain $\mathcal{A}(A \cup B; \mathbf{0})$. More generally, the foregoing method can be applied to a parameter space that is equal to the union of sets each of which satisfies the conditions of Proposition 4.7.3.

Example 4.7.4 *Tangent Cones*

(1) Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 \geqslant (1/2)\theta_1, \theta_2 \leq 2\theta_1\}$ (see Fig. 4.7). Then, we have the following:
$\mathcal{T}(\Theta_0; \boldsymbol{\theta}_0) = \Theta$, where $\boldsymbol{\theta}_0 = (0, 0)^T$.
$\mathcal{T}(\Theta; \boldsymbol{\theta}_a) = \{\boldsymbol{x} \in \mathbb{R}^2 : x_2 \geqslant (1/2)x_1\}$, where $\boldsymbol{\theta}_a = (2, 1)^T$.
$\mathcal{T}(\Theta; \boldsymbol{\theta}_b) = \mathbb{R}^2$, where $\boldsymbol{\theta}_b = (2, 2)^T$.
$\mathcal{T}(\Theta; \boldsymbol{\theta}_c) = \{\boldsymbol{x} \in \mathbb{R}^2 : x_2 \leq 2x_1\}$, where $\boldsymbol{\theta}_c = (1, 2)^T$.

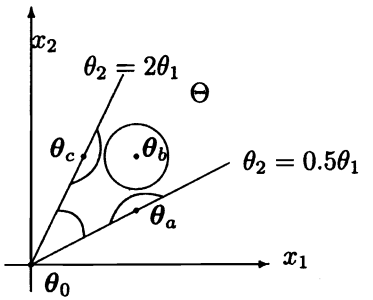

Fig. 4.7 The tangent cones of $\Theta = \{\boldsymbol{\theta} : \theta_2 \leq 2\theta_1, \theta_2 \geqslant 0.5\theta_1\}$ at $\boldsymbol{\theta}_0, \boldsymbol{\theta}_a, \boldsymbol{\theta}_b$ and $\boldsymbol{\theta}_c$

Note that at $\boldsymbol{\theta}_a$, the inequality $\theta_2 \leq 2\theta_1$ holds strictly (i.e., it is not active) and therefore this inequality constraint can be ignored for determining the tangent cone at $\boldsymbol{\theta}_a$. At $\boldsymbol{\theta}_b$, both inequalities are satisfied strictly, and therefore both constraints can be ignored for determining the tangent cone. Since there are no other active constraints, the tangent cone is the full space $\mathbb{R}^2$. At $\boldsymbol{\theta}_c$, the inequality $\theta_2 \geqslant (1/2)\theta_1$ holds strictly and hence it can be ignored for determining the tangent cone; the rest of the arguments are similar to those for $\boldsymbol{\theta}_a$. ■

(2) Let the matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ be given and let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : \boldsymbol{R}_1\boldsymbol{\theta} \geqslant \mathbf{0}, \boldsymbol{R}_2\boldsymbol{\theta} \geqslant \mathbf{0}\}$. Then, by Proposition 4.7.3 we have the following:

$$\begin{aligned}
&\boldsymbol{R}_1\boldsymbol{\theta}_a > \mathbf{0}, \boldsymbol{R}_2\boldsymbol{\theta}_a = \mathbf{0} \Rightarrow \mathcal{T}(\Theta; \boldsymbol{\theta}_a) = \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{R}_2\boldsymbol{x} \geqslant \mathbf{0}\}. \\
&\boldsymbol{R}_1\boldsymbol{\theta}_0 = \mathbf{0}, \boldsymbol{R}_2\boldsymbol{\theta}_0 = \mathbf{0} \Rightarrow \mathcal{T}(\Theta; \boldsymbol{\theta}_0) = \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{R}_1\boldsymbol{x} \geqslant \mathbf{0}, \boldsymbol{R}_2\boldsymbol{x} \geqslant \mathbf{0}\}. \quad (4.98) \\
&\boldsymbol{R}_1\boldsymbol{\theta}_b > \mathbf{0}, \boldsymbol{R}_2\boldsymbol{\theta}_b > \mathbf{0} \Rightarrow \mathcal{T}(\Theta; \boldsymbol{\theta}_b) = \mathbb{R}^p.
\end{aligned}$$

(3) Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 \geqslant -\theta_1^2 + 2\theta_1, \theta_2 \geqslant -\theta_1^2 - 2\theta_1\}$ and $\boldsymbol{\theta}_0 = \mathbf{0}$.

It is easily verified that the MF-CQ is satisfied at $\boldsymbol{\theta}_0$ and therefore the tangent cone is determined by linearizing the constraints at $\boldsymbol{\theta}_0$ as follows. Let the two constraints

be expressed as $g_1(\boldsymbol{\theta}) \geqslant 0$ and $g_2(\boldsymbol{\theta}) \geqslant 0$. Then the linearized versions of these constraints at $\boldsymbol{\theta}_0$ are

$$g_1(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T(\partial/\partial\boldsymbol{\theta})g_1(\boldsymbol{\theta}_0) \geqslant 0 \text{ and } g_2(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T(\partial/\partial\boldsymbol{\theta})g_2(\boldsymbol{\theta}_0) \geqslant 0,$$

respectively. These linearized versions are precisely $\theta_2 \geqslant 2\theta_1$ and $\theta_2 \geqslant -2\theta_1$, respectively. Since the two constraints defining Θ are active at $\boldsymbol{\theta}_0$, we have $\mathcal{T}(\Theta; \boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \theta_2 \geqslant 2\theta_1, \theta_2 \geqslant -2\theta_1\}$. In this particular example, we can sketch the shape of Θ easily and the tangent cone can be written down quickly; the above analytical approach is meant to be instructive.

(4) Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_2 \geqslant \theta_1^2, \theta_2 \leq \theta_1\}$. Let $g_1(\boldsymbol{\theta}) = \theta_2 - \theta_1^2$ and $g_2(\boldsymbol{\theta}) = \theta_1 - \theta_2$. Then $\nabla g_1(\boldsymbol{\theta}) = (-2\theta_1, 1)^T$, $\nabla g_2(\boldsymbol{\theta}) = (1, -1)^T$. Now, applying the same technique as in the previous example, we have $\mathcal{T}(\Theta; \boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \theta_1 \geqslant \theta_2 \geqslant 0\}$, and Θ is Chernoff regular at 0 (see Fig. 4.8). ■

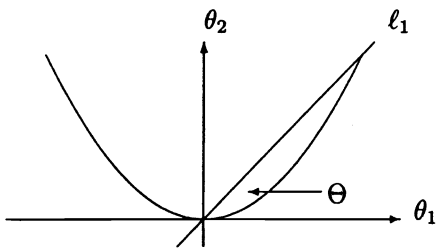

Fig. 4.8 The tangent cone of $\Theta = \{\theta_2 \geqslant \theta_1^2, \theta_2 \leq \theta_1\}$ at the origin is the cone between ℓ_1 and the θ_1 axis.

The next result is useful in asymptotic derivations because it enables us to substitute the approximating cone for the parameter space at the true value even if it (i.e., the true parameter value) is on the boundary which may or may not be smooth.

Proposition 4.7.4 *Let $\Theta \subset \mathbb{R}^p$, V be a positive definite matrix of order $p \times p$ and $\boldsymbol{\theta}_0 \in \Theta$ and $\mathcal{A}$ be an approximating cone of Θ at $\boldsymbol{\theta}_0$. Then*

$$\|\boldsymbol{y} - \Theta\|_V^2 - \|\boldsymbol{y} - \mathcal{A}(\Theta; \boldsymbol{\theta}_0\|_V^2 = o(\|\boldsymbol{y} - \boldsymbol{\theta}_0\|_V^2) \text{ as } \boldsymbol{y} \to \boldsymbol{\theta}_0.$$

The proof is given in the Appendix; in fact, two different proofs are given, one uses Chernoff's definition of an approximating cone and simple geometry and the other uses Hausdorff distance. Fig. 4.9 illustrates this result and it shows two possible shapes of Θ bounded by two curves intersecting at O. Let A be an arbitrary point, and C and D be the points in $\mathcal{A}(\Theta; \boldsymbol{\theta}_0)$ and Θ, respectively, that are closest to A. Now Proposition 4.7.4 says that $|AD^2 - AC^2| = o(OA^2)$.

A useful corollary to the foregoing proposition is the following.

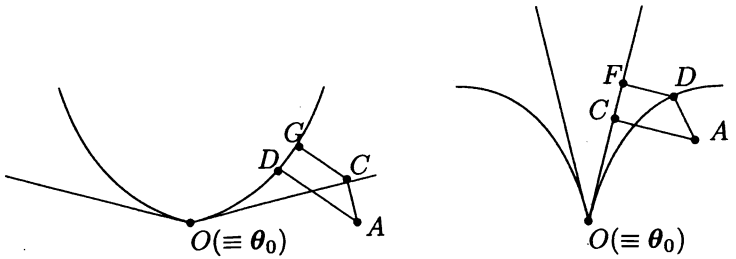

Fig. 4.9 $\|\boldsymbol{y} - \Theta\|^2 - \|\boldsymbol{y} - \mathcal{A}(\Theta; \boldsymbol{\theta}_0)\|^2 = o(\|\boldsymbol{y} - \boldsymbol{\theta}_0\|^2)$ [i.e., $|AD^2 - AC^2| = o(OA^2)$]

Corollary 4.7.5 *Let $\Theta \subset \mathbb{R}^p$, $\boldsymbol{\theta}_0 \in \Theta$, V be a $p \times p$ positive definite matrix and Θ be Chernoff regular at $\boldsymbol{\theta}_0$. Then we have the following:*

1. *If $n^{1/2}(\boldsymbol{T}_n - \boldsymbol{\theta}_0) = O_p(1)$ then*

$$\|\boldsymbol{T}_n - \Theta\|_V^2 - \|\boldsymbol{T}_n - \mathcal{A}(\Theta; \boldsymbol{\theta}_0)\|_V^2 = o_p(n^{-1}).$$

2. *If $n^{1/2}(\boldsymbol{T}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \boldsymbol{Z}$ then $n\|\boldsymbol{T}_n - \Theta\|_V^2 \xrightarrow{d} \|\boldsymbol{Z} - \mathcal{T}(\Theta; \boldsymbol{\theta}_0)\|_V^2$*

Proof: Let us first state a lemma.

Lemma 4.7.6 *Let $g : \mathbb{R}^p \to \mathbb{R}$ be such that $g(\boldsymbol{y}) = o(\|\boldsymbol{y} - \boldsymbol{\theta}_0\|^\beta)$ as $\|\boldsymbol{y} - \boldsymbol{\theta}_0\| \to 0$, where $\beta > 0$. Let $\boldsymbol{T}_n$ be a sequence of random p-vectors such that $\boldsymbol{T}_n - \boldsymbol{\theta}_0 = O_p(n^{-\gamma})$ where $\gamma > 0$. Then $g(\boldsymbol{T}_n) = o_p(n^{-\beta\gamma})$.*

Proof: Given $\epsilon > 0$, there exists a δ such that $|g(\boldsymbol{y})/\|\boldsymbol{y} - \boldsymbol{\theta}_0\|^\beta| < \epsilon$ for $\|\boldsymbol{y} - \boldsymbol{\theta}_0\| < \delta$. Therefore,

$$\text{pr}\{|g(\boldsymbol{T}_n)/\|\boldsymbol{T}_n - \boldsymbol{\theta}_0\|^\beta < \epsilon) \geqslant \text{pr}(\|\boldsymbol{T}_n - \boldsymbol{\theta}_0\| < \delta) \to 1 \text{ as } n \to \infty.$$

Thus, $g(\boldsymbol{T}_n)/\|\boldsymbol{T}_n - \boldsymbol{\theta}_0\|^\beta = o_p(1)$. Now,

$$g(\boldsymbol{T}_n) = \|\boldsymbol{T}_n - \boldsymbol{\theta}_0\|^\beta o_p(1) = |n^{-\gamma}O_p(1)|^\beta o_p(1) = o_p(n^{-\beta\gamma}).$$

This completes the proof of the lemma.

Now, the proof of the first part of the corollary follows from the foregoing Lemma and Proposition 4.7.4 by substituting $g(\boldsymbol{y}) = \|\boldsymbol{y} - \Theta\|_V^2 - \|\boldsymbol{y} - \mathcal{A}(\Theta; \boldsymbol{\theta}_0)\|_V^2$, $\beta = 2$ and $\gamma = 1/2$.

To prove the second part, let $\boldsymbol{Z}_n = n^{1/2}(\boldsymbol{T}_n - \boldsymbol{\theta}_0)$ and note that

$$\begin{aligned} n\|\boldsymbol{T}_n - \Theta\|_V^2 &= n\|\boldsymbol{T}_n - \mathcal{A}(\Theta; \boldsymbol{\theta}_0)\|_V^2 + o_p(n^{-1}) \\ &= \|\boldsymbol{Z}_n - \mathcal{T}(\Theta; \boldsymbol{\theta}_0)\|_V^2 + o_p(1) \xrightarrow{d} \|\boldsymbol{Z} - \mathcal{T}(\Theta; \boldsymbol{\theta}_0)\|_V^2 \end{aligned}$$

■

The corollary essentially says that $\mathcal{A}(\Theta;\boldsymbol{\theta}_0)$ can be substituted for Θ in most of the expressions; in particular, if $\boldsymbol{T}_n$ is a $\sqrt{n}$-consistent estimator, $\boldsymbol{\theta}_0$ is the true value and $\boldsymbol{\theta}_0 \in \Theta$ then the squared distance between $\boldsymbol{T}_n$ and Θ is equal to that between $\boldsymbol{T}_n$ and the approximating cone of Θ at $\boldsymbol{\theta}_0$ except for a $o_p(n^{-1})$ term. The corollary can also be extended to the case when Θ_n is a sequence of sets such that $\boldsymbol{\theta}_0 \in \Theta_n$ and $\mathcal{A}(\Theta_n;\boldsymbol{\theta}_0)$ converges to a cone, or Θ_n converges to a cone.

4.8 GENERAL TESTING PROBLEMS

Let $f(y;\boldsymbol{\theta})$ be a density function where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ and $\Theta_0 \subset \Theta_1 \subset \Theta$. Let Y be a random variable with density function $f(y;\boldsymbol{\theta})$. In this section, test of

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ against } H_1 : \boldsymbol{\theta} \in \Theta_1 \tag{4.99}$$

is studied when Θ_0, Θ_1 and Θ may have boundary points and the boundaries of these sets may or may not be smooth; for example, the boundary of a cone at its vertex is not smooth. The special case when Θ_0 is a linear space and Θ_1 is a polyhedral was studied in Section 4.3, where the special structure on the parameter spaces enabled us to obtain explicit expressions for the asymptotic null distributions of the test statistics. Now, we consider more general cases and obtain general results. It will be assumed that Θ_0 and Θ_1 have approximating cones at every point in Θ_0; in the terminology introduced in Section 4.7, this is same as saying that Θ_0 and Θ_1 are Chernoff regular at every $\boldsymbol{\theta} \in \Theta_0$. For simplicity, it is assumed that the observations on Y are independently and identically distributed; however, the main results of this section hold, under appropriate regularity conditions, for much more general stochastic processes although such generalizations are not considered.

Let $Y_1, \ldots, Y_n$ be independently and identically distributed as Y. As in the earlier sections, let

$$\ell(\boldsymbol{\theta}) = \sum \log f(Y_i;\boldsymbol{\theta}), \qquad \boldsymbol{S}(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta})\ell(\boldsymbol{\theta}), \tag{4.100}$$

$$\text{and} \quad \mathcal{I}(\boldsymbol{\theta}) = E\{(\partial/\partial\boldsymbol{\theta}) \log f(Y;\boldsymbol{\theta})(\partial/\partial\boldsymbol{\theta}^T) \log f(Y;\boldsymbol{\theta})\} \tag{4.101}$$

denote the loglikelihood, the score function, and the information matrix, respectively. Some important fundamental results on the *LRT* for testing problems of the form (4.99) were obtained by Chernoff (1954). The technical details in Chernoff (1954) assumed that Θ was open and that inference was based on likelihood. Chernoff's (1954) results have natural extensions to more general cases although modifications would be required to the regularity conditions and technical details; for example, see Self and Liang (1987), Andrews (1998), Vu and Zhou (1997), Shapiro (1985, 1989, 2000a), Geyer (1994), Silvapulle (1992a, 1992b, 1992c, 1994), and El Barmi (1996) among others. The areas covered include the cases when (i) Θ is not necessarily open (ii) observations are not necessarily independently and identically distributed, and (iii) inference is based on quasi-likelihood, partial likelihood, or general objective functions such as those used in, for example, M-estimation, minimum distance methods, and empirical likelihood.

In the next subsection, we consider the simple case of likelihood ratio test when Θ is open. The simplicity of this case makes it easier to convey the essentials. On the other hand, this case is also of independent interest because many practical problems fall into this category. Then, in subsection 4.8.2, we consider likelihood inference when Θ is not necessarily open and the true value of $\boldsymbol{\theta}$ can be a boundary point of Θ, Θ_0, and Θ_1 simultaneously. The distribution and implementation of likelihood ratio test are virtually the same whether or not $\boldsymbol{\theta}_0$ is a boundary point of Θ provided some conditions are satisfied; essentially all that we require is that it should be possible to approximate the objective function by a quadratic similar to that in (4.2) and (4.3). In this sense, the results in subsection 4.8.1 are not very different from those in subsection 4.8.2. In subsection 4.8.3, we consider inference based on general objective functions rather than just likelihood. Some results obtained in subsections 4.8.1 and 4.8.2 do not necessarily remain the same in this general setting. In subsection 4.8.4, some simple examples in two dimensions are considered to illustrate the theory. In subsection 4.8.5, a general form of the Type B testing problem, test of $\boldsymbol{h}(\boldsymbol{\theta}) \geqslant \boldsymbol{0}$ against $\boldsymbol{h}(\boldsymbol{\theta}) \ngeqslant \boldsymbol{0}$, is studied; this particular type of problems arises in economics, for example, the interest may be to test whether or not a regression function deviates from concavity.

As usual, we adopt the notation $\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}$, and $\bar{\boldsymbol{\theta}}$ for the estimators over Θ, Θ_1, and Θ_0 respectively.

4.8.1 Likelihood Approach When Θ is Open

Assume that Θ is open and that $H_0 : \boldsymbol{\theta} \in \Theta_0$ holds. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$ in Θ_0. The set of regularity conditions required for this setting are virtually the same as those usually required for establishing that $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal and the asymptotic null distribution of the *LRT* of $\boldsymbol{R\theta} = \boldsymbol{0}$ against $\boldsymbol{R\theta} \neq \boldsymbol{0}$ is chi-squared. There are different sets of conditions available for these results; some involve third-order partial derivatives of the log density, some involve only the second-order partial derivatives, and some replace the differentiability by other special types of differentiability (quadratic mean, stochastic, etc.). For simplicity, we shall work with conditions on the third-order derivatives of the log density, but we recognize that our technical conditions can be improved. Let

$$\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T \mathcal{I}_{\boldsymbol{\theta}_0} \boldsymbol{x}. \tag{4.102}$$

Define,

$$W_D = \min_{\boldsymbol{\theta}\in\Theta_0} n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \hat{\mathcal{I}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \min_{\boldsymbol{\theta}\in\Theta_1} n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \hat{\mathcal{I}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \tag{4.103}$$

where $\hat{\mathcal{I}} = \mathcal{I}(\hat{\boldsymbol{\theta}})$; however, in general $\hat{\mathcal{I}}$ can be any consistent estimator of $\mathcal{I}(\boldsymbol{\theta}_0)$. Note that W_D is a measure of $\{dist(\hat{\boldsymbol{\theta}}, H_0) - dist(\hat{\boldsymbol{\theta}}, H_1)\}$, where $dist(\hat{\boldsymbol{\theta}}, H)$ is a measure of distance between $\hat{\boldsymbol{\theta}}$ and the parameter space defined by H. Hence it is a suitable statistic for testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$. As in the case of test against the unrestricted alternative, *LRT* and W_D are asymptotically equivalent. A simple method of obtaining their common asymptotic distribution is given in the

next result; essentially, it says that the value of *LRT* changes by a term of order $o_p(1)$ if $\mathcal{A}(\Theta_i;\boldsymbol{\theta}_0)$ is substituted for Θ_i $(i=0,1)$.

Proposition 4.8.1 *Assume that Θ is open, Condition Q is satisfied, and $H_0:\boldsymbol{\theta}\in\Theta_0$ holds. Then, with the notation in (4.102), we have the following:*

1. *$LRT = W_D + o_p(1)$, where W_D is given in (4.103).*
2. *The common asymptotic null distribution of LRT and W_D is equal to the distribution of*

$$\|\boldsymbol{X}-\mathcal{A}(\Theta_0;\boldsymbol{\theta}_0)\|^2-\|\boldsymbol{X}-\mathcal{A}(\Theta_1;\boldsymbol{\theta}_0)\|^2, \tag{4.104}$$

which is also equal to the distribution of

$$\|\boldsymbol{Z}-\mathcal{T}(\Theta_0;\boldsymbol{\theta}_0)\|^2-\|\boldsymbol{Z}-\mathcal{T}(\Theta_1;\boldsymbol{\theta}_0)\|^2 \tag{4.105}$$

where $\boldsymbol{X}\sim N(\boldsymbol{\theta}_0,\mathcal{I}(\boldsymbol{\theta}_0)^{-1})$, $\boldsymbol{Z}\sim N(\mathbf{0},\mathcal{I}(\boldsymbol{\theta}_0)^{-1})$ and $\|\cdot\|$ is as in (4.102).

Proof: Let $\mathcal{A}_i=\mathcal{A}(\Theta_i;\boldsymbol{\theta}_0)$ and $\mathcal{T}_i=\mathcal{T}(\Theta_i;\boldsymbol{\theta}_0)$ for $i=0,1$. Now, by using the quadratic approximation (4.4), we have that

$$\begin{aligned} LRT &= 2[\sup_{\boldsymbol{\theta}\in\Theta_1}\ell(\boldsymbol{\theta})-\sup_{\boldsymbol{\theta}\in\Theta_0}\ell(\boldsymbol{\theta})] \\ &= n\|\hat{\boldsymbol{\theta}}-\Theta_0\|^2-n\|\hat{\boldsymbol{\theta}}-\Theta_1\|^2+o_p(1) \\ &= W_D+o_p(1); \end{aligned} \tag{4.106}$$

the last step, which says that $\|\cdot\|_{\mathcal{I}(\boldsymbol{\theta}_0)}$ can be replaced by $\|\cdot\|_{\hat{\mathcal{I}}}$, follows from Lemma 4.10.2 in the Appendix.

Now, let us prove the second part. By Corollary 4.7.5, Θ_i in (4.106) can be replaced by its approximating cone $(i=0,1)$. Let $\boldsymbol{Z}_n=\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$. Now, by applying essentially the same arguments as those used in the proof of the first part, we have

$$\begin{aligned} LRT &= n\{\|\hat{\boldsymbol{\theta}}-\mathcal{A}_0\|^2-\|\hat{\boldsymbol{\theta}}-\mathcal{A}_1\|^2\}+o_p(1), \\ &= \|\boldsymbol{Z}_n-\mathcal{T}_0\|^2-\|\boldsymbol{Z}_n-\mathcal{T}_1\|^2+o_p(1), \\ &\xrightarrow{d}\|\boldsymbol{Z}-\mathcal{T}_0\|^2-\|\boldsymbol{Z}-\mathcal{T}_1\|^2; \end{aligned} \tag{4.107}$$

the last step follows since $\|\boldsymbol{z}-\mathcal{T}_i\|^2$ is continuous in $\boldsymbol{z}$ and $\boldsymbol{Z}_n\xrightarrow{d}\boldsymbol{Z}$. ■

The foregoing results show that, if the null hypothesis is true and $\boldsymbol{\theta}_0$ denotes the true value, then the asymptotic distribution of the *LRT* of $\boldsymbol{\theta}\in\Theta_0$ against $\boldsymbol{\theta}\in\Theta_1$ is equal to the distribution of

1. The *LRT* of $\boldsymbol{\theta}\in\mathcal{A}(\Theta_0;\boldsymbol{\theta}_0)$ against $\boldsymbol{\theta}\in\mathcal{A}(\Theta_1;\boldsymbol{\theta}_0)$ based on a single observation of $\boldsymbol{X}$, where $\boldsymbol{X}\sim N\{\boldsymbol{\theta},\mathcal{I}(\boldsymbol{\theta}_0)^{-1}\}$ and the true value of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$.
2. The *LRT* of $\boldsymbol{\alpha}\in\mathcal{T}(\Theta_0;\boldsymbol{\theta}_0)$ against $\boldsymbol{\alpha}\in\mathcal{T}(\Theta_1;\boldsymbol{\theta}_0)$ based on a single observation of $\boldsymbol{Z}$ where $\boldsymbol{Z}\sim N\{\boldsymbol{\alpha},\mathcal{I}(\boldsymbol{\theta}_0)^{-1}\}$ and the true value of $\boldsymbol{\alpha}$ is $\mathbf{0}$.

Thus, as in classical inference and in Section 4.3, the asymptotic distribution of the test statistics can be expressed in terms of a single observation from the normal distribution. However, this apparently simpler problem may be still quite complicated, as was noted in Section 4.3, mainly due to the presence of the nuisance parameter in the null distribution; the nuisance parameter appears because the asymptotic null distribution of the *LRT* depends on $\mathcal{I}(\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is the assumed true value in the null parameter space. This is in sharp contrast to the case where the asymptotic null distribution of *LRT* for testing $\boldsymbol{R\theta}=\mathbf{0}$ against $\boldsymbol{R\theta}\neq\mathbf{0}$ is a chi-squared and hence does not depend on $\boldsymbol{\theta}_0$ in the null parameter space.

If $\mathcal{T}(\Theta_0;\boldsymbol{\theta}_0)$ is a linear space and $\mathcal{T}(\Theta_1;\boldsymbol{\theta}_0)$ is a closed convex cone, then it follows from the foregoing proposition that the asymptotic null distribution of the *LRT* is a chi-bar-square. For example, if the null and alternative hypotheses are $\boldsymbol{R\theta}=\mathbf{0}$ and $\boldsymbol{R\theta}\geqslant\mathbf{0}$, respectively, then the asymptotic null distribution of the *LRT* is chi-bar-square; this was also obtained in Proposition 4.3.1. In fact, Proposition 4.3.1 on the asymptotic null distribution of *LRT* is a special case of the foregoing proposition. However, in general, the asymptotic null distribution of the *LRT* is not a chi-bar-square.

The results in the foregoing proposition provide elegant asymptotic representations of the *LRT* and they lead to its asymptotic distribution, which has a simpler structure because the original parameter spaces have been replaced by their approximating cones and the limiting distribution is a function of $\boldsymbol{Z}$ where $\boldsymbol{Z}\sim N(\mathbf{0},\mathcal{I}_{\boldsymbol{\theta}_0}^{-1})$. Although Proposition 4.8.1 was established for the case when Θ is open, similar results are available even when Θ is not open and $\boldsymbol{\theta}_0$ is a boundary point of Θ.

4.8.2 Likelihood Approach When Θ is not Open

Now assume that Θ is not necessarily open and that $\boldsymbol{\theta}_0$ may not be an interior point of Θ. Then $(\partial/\partial\boldsymbol{\theta})\ell(\hat{\boldsymbol{\theta}})$ may not be zero and $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ may not be asymptotically normal. Consequently, the arguments in the previous subsection that depend on the asymptotic normality of $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ are not valid; also, arguments based on Taylor series may not be valid. However, most of the results still hold. For example, the asymptotic null distribution of *LRT* remains unchanged even if $\boldsymbol{\theta}_0$ is a boundary point Θ, but the technical details leading to it are different from those in the previous subsection. *Assume that the null hypothesis holds.* In most practical situations, $\ell(\boldsymbol{\theta})$ would have directional derivatives at $\boldsymbol{\theta}_0$, and hence would admit a quadratic approximation in a neighborhood of $\boldsymbol{\theta}_0$. Further, the maximizers of $\ell(\boldsymbol{\theta})$ over Θ_0 and over Θ_1 are likely to be $n^{1/2}$-consistent. This is because consistency of the estimator is not affected by the true parameter being on the boundary, and $\sqrt{n}$-consistency follows once we have a quadratic approximation of $\ell(\boldsymbol{\theta})$ similar to (4.2) by arguments indicated at the beginning of this chapter. If the objective function $\ell(\boldsymbol{\theta})$ does not have a quadratic approximation similar to (4.2) or the estimators are not $\sqrt{n}$-consistent, then the general approach of this section will need modifications to obtain the asymptotic distribution of *LRT*.

The Condition $Q2$ defined below is a slight modification of Condition Q; the modification is made in view of the fact that $\boldsymbol{\theta}_0$ may be a boundary point of Θ.

Condition Q2. There exist $S_*(\theta)$ and $\mathcal{I}_*(\theta)$ such that (4.3) holds with $S(\theta)$ and $\mathcal{I}(\theta)$ replaced by $S_*(\theta)$ and $\mathcal{I}_*(\theta)$ respectively. Further,

1. $n^{-1/2}S_*(\theta_0) \xrightarrow{d} N\{0, \mathcal{I}_*(\theta_0)\}$.
2. The maximizers of $\ell(\theta)$ over Θ_0 and over Θ_1 are $\sqrt{n}$-consistent for θ_0. ■

There is room for fine tuning this condition; for example, $\sqrt{n}$-consistency would follow from (4.3) and consistency. However, it is convenient for us to state them explicitly.

If θ_0 is an interior point of Θ, then S_* and $\mathcal{I}_*$ would be the S and $\mathcal{I}$ in the previous subsection. If Θ is a rectangle in the Euclidean space, θ_0 is a boundary point of Θ, and $\ell(\theta)$ has partial derivatives at θ_0 from the appropriate directions then Condition $Q2$ could be established using Taylor series expansions (for example, see Self and Liang (1987) and Andrews (1999)).

If Θ does not have the shape of a rectangle and θ_0 is a boundary point of Θ, then it is likely that Condition $Q2$ can be established using directional derivatives (see Shapiro (2000b)). Directional derivatives of $\ell(\theta)$ at θ_0 would be required only in directions that take us into the approximating cone of Θ at θ_0. The proof of the next result is essentially the same as that for the previous one, hence omitted.

Proposition 4.8.2 *Assume that Condition Q2 is satisfied and that $H_0 : \theta \in \Theta_0$ holds. Let θ_0 denote the true value of θ in Θ_0, $\|x\|^2 = x^T\mathcal{I}_*(\theta_0)x$, $Z \sim N\{0, \mathcal{I}_*(\theta_0)^{-1}\}$, $X = \theta_0 + Z$, and $U_n = n^{-1/2}\mathcal{I}_*(\theta_0)^{-1}S_*(\theta_0)$. Then,*

1. $LRT = \|U_n - \mathcal{T}(\Theta_0; \theta_0)\|^2 - \|U_n - \mathcal{T}(\Theta_1; \theta_0)\|^2 + o_p(1)$.
2. *The asymptotic null distribution of LRT is equal to the distribution of*

$$\|X - \mathcal{A}(\Theta_0; \theta_0)\|^2 - \|X - \mathcal{A}(\Theta_1; \theta_0)\|^2, \tag{4.108}$$

which is also equal to the distribution of

$$\|Z - \mathcal{T}(\Theta_0; \theta_0)\|^2 - \|Z - \mathcal{T}(\Theta_1; \theta_0)\|^2. \tag{4.109}$$

■

The conditions under which Condition $Q2$ holds when θ_0 is a boundary point of Θ, appear to give the impression that the likelihood must behave at θ_0 almost as if it is an interior point of an extended parameter space over which Condition $Q2$ is satisfied, and $E\{\ell(\theta)\}$ has a maximum at θ_0 that is also a stationary point of $E\{\ell(\theta)\}$. Thus, although Condition Q for the case when θ_0 is an interior point of Θ and Condition $Q2$ when θ_0 is a boundary point of Θ appear to be different, the requirements on the statistical model appear to be not very different. A question of interest is, if Condition $Q2$ is satisfied, perhaps with S_* and $\mathcal{I}_*$ defined using directional derivatives, then is it possible to extend the parameter space such that it includes an open neighborhood of θ_0 and the derivations in the previous subsection still hold? Although we believe that the answer is "yes," the technical details do not appear to be available yet.

If θ_0 is an interior point of Θ, then it follows from

$$n^{1/2}(\hat{\theta} - \theta_0) = n^{-1/2}\mathcal{I}_{\theta_0}^{-1}S(\theta_0) + O_p(n^{-1/2})$$

that $n^{1/2}(\hat{\theta} - \theta_0)$ and $n^{-1/2}\mathcal{I}(\theta_0)^{-1}S(\theta_0)$ are essentially interchangeable in the proofs of the asymptotic results (the Z_n in the proof of Proposition 4.8.1 and U_n in the proof Proposition 4.8.2 are equivalent, and hence the two derivations differ by an $o_p(1)$ term only).

4.8.3 Tests with General Objective Functions

For the case when there are no inequality constraints, the classical inference procedures include M-estimation, minimum distance, partial likelihood, quasi-likelihood, and empirical likelihood and more generally, pseudo-likelihoods (see Pace and Salvan (1997)). The principles underlying these also lend themselves for inference when there are inequality constraints on parameters. In this subsection, we will discuss some of the results that are available in these directions.

Let the objective function be denoted by $R_n(\theta)$; this replaces the role of loglikelihood. The estimator of θ is the maximizer of $R_n(\theta)$. Assume that the following condition, which is a more generalized version of those used in the previous subsections, is satisfied.

Condition $Q3$:
There exist a vector function $S_*(\theta)$ and symmetric and positive definite matrices, $G_*(\theta)$ and $\mathcal{I}_*(\theta)$, such that the following are satisfied:

1. (4.3) holds with $\{\ell(\theta), S(\theta), \mathcal{I}(\theta)\}$ replaced by $\{R_n(\theta), S_*(\theta), G_*(\theta)\}$.
2. $n^{-1/2}S_*(\theta_0) \xrightarrow{d} N\{0, \mathcal{I}_*(\theta_0)\}$.
3. The maximizer of $R_n(\theta)$ over Θ_i is $n^{1/2}$-consistent for θ_0 when $\theta_0 \in \Theta_i$ $(i = 0, 1)$. ■

Now, a likelihood ratio type statistic for testing $\theta \in \Theta_0$ against $\theta \in \Theta_1$ is

$$T_n = 2\{\sup_{\theta \in \Theta_1} R_n(\theta) - \sup_{\theta \in \Theta_0} R_n(\theta)\}. \tag{4.110}$$

The next proposition shows that the asymptotic null distribution of T_n is obtained by an argument similar to those in the previous subsection.

Proposition 4.8.3 *Assume that Condition Q3 is satisfied and that $H_0 : \theta \in \Theta_0$ holds. Let θ_0 denote the true value of θ in Θ_0, $\|x\|^2 = x^T G_*(\theta_0)x$, and $U_n = n^{-1/2}G_*(\theta_0)^{-1}S_*(\theta_0)$. Then*

1. $T_n = \|U_n - \mathcal{T}(\Theta_0; \theta_0)\|^2 - \|U_n - \mathcal{T}(\Theta_1; \theta_0)\|^2 + o_p(1)$.
2. *The asymptotic null distribution of T_n is equal to that of*

$$\|U - \mathcal{T}(\Theta_0; \theta_0)\|^2 - \|U - \mathcal{T}(\Theta_1; \theta_0)\|^2 \tag{4.111}$$

where $U \sim N\{0, G_*(\theta_0)^{-1}\mathcal{I}_*(\theta_0)G_*(\theta_0)^{-1}\}$.

Proof : It follows from Condition $Q3(1)$ that, for $\theta - \theta_0 = O(n^{-1/2})$,

$$R_n(\theta) = R_n(\theta_0) - (n/2)\|n^{-1/2}U_n - (\theta - \theta_0)\|^2 + (1/2)n^{-1}\|U(\theta_0)\|^2 + o_p(1).$$

Now,

$$\begin{aligned} T_n &= 2[\sup\{R_n(\theta) : \theta \in \Theta_1\} - \sup\{R_n(\theta) : \theta \in \Theta_0\}] \\ &= n\|n^{-1/2}U_n - (\Theta_0 - \theta_0)\|^2 - n\|n^{-1/2}U_n - (\Theta_1 - \theta_0)\|^2 + o_p(1) \\ &= \|U_n - \mathcal{T}(\Theta_0;\theta_0)\|^2 - \|U_n - \mathcal{T}(\Theta_1;\theta_0)\|^2 + o_p(1) \\ &\xrightarrow{d} \|U - \mathcal{T}(\Theta_0;\theta_0)\|^2 - \|U - \mathcal{T}(\Theta_1;\theta_0)\|^2. \quad \blacksquare \end{aligned}$$

Note that the quadratic form $\|U - \alpha\|^2$ in (4.111), which is equal to $(U - \alpha)^T G_*(\theta_0)(U - \alpha)$, is with respect to the matrix $G_*(\theta_0)^{-1}$ but the covariance matrix of U is $G_*(\theta_0)^{-1}\mathcal{I}_*(\theta_0)G_*(\theta_0)^{-1}$. Therefore, if $\mathcal{I}_*(\theta_0) \neq G_*(\theta_0)$, then the quadratic form $\|U - \alpha\|^2$ may not be with respect to the covariance matrix of U and hence the results in Chapter 2 for normal distributions are not applicable for obtaining the asymptotic distribution of T_n. In particular, (i) for testing $\theta = 0$ against $\theta \neq 0$, the asymptotic null distribution of T_n may not be a chi-square, and (ii) for testing $R\theta = 0$ against $R\theta \geqslant 0$, the asymptotic null distribution of T_n in (4.111) may not be a chi-bar-square.

Suppose that Θ is open and that $H_0 : \theta \in \Theta_0$ holds. As usual, let $\hat{\theta}$ denote the unrestricted estimator of θ. Then by standard arguments, $n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N\{0, V(\theta_0)\}$, where $V(\theta_0) = G_*(\theta_0)^{-1}\mathcal{I}_*(\theta_0)G_*(\theta_0)^{-1}$. Therefore, a W_D-type statistic for testing $\theta \in \Theta_0$ against $\theta \in \Theta_1$ is

$$W_D = \min_{\theta\in\Theta_0} n(\hat{\theta} - \theta)^T\hat{V}^{-1}(\hat{\theta} - \theta) - \min_{\theta\in\Theta_1} n(\hat{\theta} - \theta)^T\hat{V}^{-1}(\hat{\theta} - \theta), \quad (4.112)$$

where $\hat{V}$ is a consistent estimator of $V(\theta_0)$, for example, $V(\hat{\theta})$. The asymptotic distribution of the foregoing W_D is equal to that of

$$\min_{\theta\in\mathcal{T}(\Theta_0;\theta_0)} (Z - \theta)^T V(\theta_0)^{-1}(Z - \theta) - \min_{\theta\in\mathcal{T}(\Theta_1;\theta_0)} (Z - \theta)^T V(\theta_0)^{-1}(Z - \theta)$$

where $Z \sim N\{0, V(\theta_0)\}$.

Note that the quadratic form $(Z - \theta)^T V(\theta_0)^{-1}(Z - \theta)$, which appears in the asymptotic null distribution of W_D is with respect to the covariance matrix $V(\theta_0)$ of Z. Consequently, it follows that for testing $R\theta = 0$ against $R_1\theta \geqslant 0$, the asymptotic null distribution of W_D is a chi-bar-square. Therefore, the asymptotic null distribution of W_D in (4.112) and that of the likelihood ratio type statistic T_n in (4.110) are not necessarily the same, and hence are unlikely to be asymptotically equivalent in this type of setting.

For some partial-likelihood and quasi-likelihood, $G_*(\theta)$ and $\mathcal{I}_*(\theta)$ are equal except for a scalar multiplicative constant. Since the multiplicative constant factors out from matrix products, it (i.e., multiplicative constant) can be replaced by a consistent estimator of it and the required results can be obtained (see quasi-likelihood method on page 159 and Silvapulle (1994)).

4.8.4 Examples

Example 4.8.1 *Asymptotic distribution of the LRT of* $g(\theta) = 0$ *vs* $g(\theta) \geqslant 0$.

Let $Y = (Y_1, Y_2)^T \sim N(\theta, V)$ where $\theta \in \Theta = \mathbb{R}^2$ and V is known and does not depend on θ. Let the null and alternative hypotheses be

$$H_0 : g(\theta) = 0 \text{ and } H_1 : g(\theta) \geqslant 0,$$

respectively, where $g = (g_1, g_2)^T, g_1(\theta) = \theta_2 - \theta_1^2 - \theta_1$ and $g_2(\theta) = \theta_2 - \theta_1^2 + \theta_1$; see Fig. 4.1. Let us obtain the asymptotic null distribution of the *LRT* given by Proposition 4.8.1. Let $\bar{Y}$ denote the mean of n independently and identically distributed observations on Y; then $\bar{Y} \sim N(\theta, n^{-1}V)$. Now, $\Theta_0 = \{\theta \in \mathbb{R}^2 : g(\theta) = 0\}$ and $\Theta_1 = \{\theta \in \mathbb{R}^2 : g(\theta) \geqslant 0\}$, Θ_1 is shown as Ω in Fig. 4.1 on page 184. Clearly, $\Theta_0 = \{0\}$. Assume that H_0 is true; then the true value θ_0 is $0, \mathcal{T}(\Theta_0;\theta_0) = \{0\}$, and $\mathcal{T}(\Theta_1;\theta_0) = \{\theta : \theta_2 \geqslant \theta_1, \theta_2 \geqslant -\theta_1\} = \{\theta : R\theta \geqslant 0\}$ where R is the 2×2 matrix $[-1, 1|1, 1]$; $\mathcal{T}(\Theta_1; 0)$ is shown as $\mathcal{T}$ in Fig. 4.1. Let $Z \sim N(0, V)$. Now, since $\ell(\theta) = (-1/2)n(\bar{Y} - \theta)^T V^{-1}(\bar{Y} - \theta) + const$ and $\mathcal{I}(\theta) = V^{-1}$, we have the following:

$$\begin{aligned} LRT &= n\|\bar{Y}\|_V^2 - n\|\bar{Y} - \Theta_1\|_V^2 \\ &= n\|\bar{Y}\|_V^2 - n\|\bar{Y} - \mathcal{A}(\Theta_1;0)\|_V^2 + o_p(1) \\ &\xrightarrow{d} \|Z\|_V^2 - \|Z - \mathcal{T}(\Theta_1;0)\|_V^2 \sim \bar{\chi}^2\{V, \mathcal{T}(\Theta_1;0)\}. \end{aligned}$$

Therefore, by (3.8)

$$\text{pr}(LRT \geqslant c|H_0) \to 0.5\text{pr}(\chi_1^2 \geqslant c) + (0.5 - q)\text{pr}(\chi_2^2 \geqslant c) \quad c > 0$$

where $(0.5 - q)$ is as in (3.8). $\blacksquare$

Example 4.8.2 *Asymptotic null distribution of the LRT of* $g(\theta) \geqslant 0$ *vs* $g(\theta) \not\geqslant 0$.

Let $Y, \bar{Y}, V$, and g be as in the previous example. Let us obtain the asymptotic null distribution of the *LRT* for testing

$$H_0 : g(\theta) \geqslant 0 \text{ against } H_1 : g(\theta) \not\geqslant 0.$$

The relevant asymptotic result is given in Proposition 4.8.1. By definition, $\Theta_0 = \{\theta \in \mathbb{R}^2 : g(\theta) \geqslant 0\}$ and $\Theta_1 = \mathbb{R}^2$. Assume that the null hypothesis holds and let θ_0 denote the true value of θ. Clearly, $\mathcal{T}(\Theta_1;\theta_0) = \mathbb{R}^2$. However, $\mathcal{T}(\Theta_0;\theta_0)$ depends on exactly which of the inequalities defining Θ_0 are active at θ_0 (recall that $g_i(\theta) \geqslant 0$ is said to be *active* at θ_0 if $g_i(\theta_0) = 0$). We shall consider four different cases corresponding to different tangent cones of Θ_0 at θ_0.

Case 1: $g_1(\theta_0) > 0$ and $g_2(\theta_0) > 0$.

The condition "$g_1(\theta_0) > 0$ and $g_2(\theta_0) > 0$" is equivalent to saying that θ_0 is an interior point of Θ_0. Therefore, $\mathcal{T}(\Theta_0;\theta_0) = \mathbb{R}^2$ and hence $\mathcal{T}(\Theta_0;\theta_0) = \mathcal{T}(\Theta_1;\theta_0)$.

Now, it follows from (4.105) that $LRT \xrightarrow{p} 0$. Therefore, the probability of rejecting H_0 when the true value is an interior point of Θ_0 converges to zero. Intuitively, this is to be expected, because it follows from $\hat{\theta} \xrightarrow{p} \theta_0$ that $LRT = 0$ and hence the *LRT* will not reject the null hypothesis with probability going to one.

Case 2: $g_1(\theta_0) = 0, g_2(\theta_0) > 0$.

The tangent cone at θ_0 is obtained by replacing the curved boundaries at θ_0 by their linear approximations. Equivalently, by Proposition 4.7.3, we have that $\mathcal{A}(\Theta_0; \theta_0) = \{\theta : (\theta - \theta_0)^T(\partial/\partial\theta)g_1(\theta_0) \geqslant 0\}$ and $\mathcal{T}(\Theta_0; \theta_0) = \mathcal{A}(\Theta_0; \theta_0) - \theta_0$. Therefore, $\mathcal{T}(\Theta_0; \theta_0) = \{x : x^T(\partial/\partial\theta)g_1(\theta_0) \geqslant 0\} = \{x : x_2 \geqslant x_1(1 + 2\theta_{01})\}$. Now, with $Z \sim N(0, V)$ we have that

$$\text{pr}_{\theta_0}(LRT \geqslant c \mid H_0) \to \text{pr}_{\theta_0}[\|Z - \mathcal{T}(\Theta_0; \theta_0)\|_V^2 \geqslant c] = (1/2)\text{pr}(\chi_1^2 \geqslant c).$$

Case 3: $g_1(\theta_0) = g_2(\theta_0) = 0$.

As was shown in the previous example, $\theta_0 = 0$ and $\mathcal{T}(\Theta_0; \theta_0) = \{d \in \mathbb{R}^2 : d_2 \geqslant d_1, d_2 \geqslant -d_1\} = \{d : Rd \geqslant 0\}$, where R is the 2×2 nonsingular matrix $[-1, 1|1, 1]$. Therefore, by (4.105), $LRT \xrightarrow{d} \|Z - \mathcal{T}(\Theta_0; \theta_0)\|_V^2$ where $Z \sim N(0, V)$. Now by (3.39) it follows that

$$\text{pr}_{\theta_0}(LRT \geqslant c) \to 0.5\text{pr}(\chi_1^2 \geqslant c) + q\text{pr}(\chi_2^2 \geqslant c), \quad c > 0$$

where

$$q = (2\pi)^{-1}\cos^{-1}[(R_1VR_2^T)/\{(R_1VR_1^T)(R_2VR_2^T)\}^{1/2}].$$

Case 4: $g_1(\theta_0) > 0$ and $g_2(\theta_0) = 0$.

This is similar to Case 2 and hence $\text{pr}(LRT \geqslant c|H_0) \to (1/2)\text{pr}(\chi_1^2 \geqslant c)$.

It follows from the different cases considered above that $\text{pr}(LRT \geqslant c \mid H_0)$ is a maximum when $g_1(\theta_0) = g_2(\theta_0) = 0$. Therefore, Case 3 is the least favorable null configuration, and hence $\theta_0 = 0$ is the least favorable null value, and

$$\lim \text{pr}(LRT \geqslant c|H_0) \leq 0.5\text{pr}(\chi_1^2 \geqslant c) + q\text{pr}(\chi_2^2 \geqslant c).$$

Since q is known, we can apply the *LRT* easily in this case. It will be seen later that if the null hypothesis involves several inequality constraints, for example, say $g_1(\theta) \geqslant 0, \ldots, g_k(\theta) \geqslant 0$, $(k \geq 3)$, then a least favorable null value may or may not be a point in $\{\theta : g_1(\theta) = \ldots = g_k(\theta) = 0\}$. The foregoing example illustrates how the asymptotic null distribution of the *LRT* depends on the shape of the parameter space at the null value. ■

4.8.5 Test of $h(\theta) \geqslant 0$ Against $h(\theta) \ngeqslant 0$

Let $h(\theta)$ be a continuously differentiable $k \times 1$ vector function. Let the null and alternative hypotheses be

$$H_0 : h(\theta) \geqslant 0 \text{ and } H_1 : h(\theta) \ngeqslant 0, \tag{4.113}$$

respectively. The formulation allows for the number of inequalities in H_0 to be larger than the number of components of θ. Consequently, the rank of $\nabla h(\theta)$ may

be less than k. Therefore, even if $\sqrt{n}(\hat{\theta} - \theta_0)$ converges to $N(0, V)$ where V is positive definite, the asymptotic covariance of $h(\hat{\theta})$ may not be positive definite. The foregoing null and alternative hypotheses may also be stated as

$$H_0 : \theta \in \Theta_0 \qquad \text{and} \qquad H_0 : \theta \in \Theta_1,$$

respectively, where $\Theta_0 = \{\theta : h(\theta) \geqslant 0\}$ and $\Theta_1 = \mathbb{R}^p$. Therefore, $LRT = \sup_{h(\theta) \geqslant 0} \ell(\theta)$. As usual, assume that Θ_0 is Chernoff regular. Further, assume that Condition Q2 in Section 4.8.2 is satisfied.

To study the asymptotic null distribution of the *LRT*, assume that the null hypothesis holds and let θ_0 denote the true value. Wolak (1987, 1989, 1991) studied this testing problem and obtained important results; he also highlighted some of the difficulties that arise in this type of problems. Wolak (1988) pointed out that there is no guarantee that $h(\theta) = 0$ would correspond to the least favorable null value. It follows from Section 4.8.2 that $LRT \xrightarrow{d} \|Z - \mathcal{T}(\Theta_0; \theta_0)\|^2$ where $Z \sim N\{0, \mathcal{I}(\theta_0)^{-1}\}$ and $\|x\|^2 = x^T\mathcal{I}(\theta_0)x$.

For simplicity, let us consider the case when $\mathcal{T}(\Theta_0; \theta_0)$ is a closed convex cone. Now, from Theorem 3.4.2 and the fact that

$$\|Z - \mathcal{T}(\Theta_0; \theta_0)\|^2 = \|Z\|^2 - \|Z - \mathcal{T}^o(\Theta_0; \theta_0)\|^2,$$

it follows that

$$\text{pr}_{\theta_0}(LRT \geqslant c \mid H_0) \to \sum_{i=0}^{p} w_i\text{pr}(\chi_i^2 \geqslant c) \tag{4.114}$$

where $w_i = w_i\{p, \mathcal{I}(\theta_0)^{-1}, \mathcal{T}^o(\Theta_0; \theta_0)\}$ is the usual chi-bar square weight and $\mathcal{T}^o$ denotes the polar cone of $\mathcal{T}$ with respect to $\mathcal{I}(\theta_0)^{-1}$. In general, the $\bar{\chi}^2$-distribution in (4.114) depends on θ_0 through the weights, $w_i\{p, \mathcal{I}(\theta_0)^{-1}, \mathcal{T}^o(\Theta_0; \theta_0)\}$. Therefore, to implement a large sample test, we need to find

$$\sup_{\theta \in \Theta_0} \sum_{i=0}^{p} w_i\{p, \mathcal{I}(\theta_0)^{-1}, \mathcal{T}^o(\Theta_0; \theta_0)\}\text{pr}(\chi_i^2 \geqslant c), \quad c > 0. \tag{4.115}$$

Computation of this supremum can be a difficult task; however, we can always obtain bounds for it.

If θ_0 is an interior point of Θ_0 then $\mathcal{T}(\Theta_0; \theta_0) = \mathbb{R}^p$ and hence $\text{pr}(LRT \geqslant c|\theta = \theta_0) \to 0$ for $c > 0$. Therefore, the supremum in (4.115) cannot be achieved at an interior point of Θ_0, and hence for the purposes of computing (4.115), we may restrict the attention to values of θ_0 on the boundary of Θ_0.

It can be shown that at the least favorable null value, at least two of the inequality constraints in $h(\theta_0) \geqslant 0$ must be active; this was shown by Wolak (1991). Thus, it follows that if there are only two inequality constraints, say $h_1(\theta) \geqslant 0$ and $h_2(\theta) \geqslant 0$, then they both must be active at the least favorable null value.

In what follows, we shall assume that θ_0 is a boundary point of Θ_0. Let us consider a setting that is simpler, yet general enough for many practical situations. Assume that $h(\theta)$ is continuously differentiable and that the Mangasarian-Fromowitz condition

(see Proposition 4.7.3) is satisfied. Because θ_0 is a boundary point of Θ_0, it follows that $h_s(\theta_0) = 0$ for at least one s in $\{1, \ldots, k\}$. Let a_i denote $(\partial/\partial\theta)h_i(\theta_0)$ and let $\{j_1, \ldots, j_m\}$ denote $\{i : h_i(\theta_0) = 0\}$, the set of all indices of the active constraints at θ_0; note that m, the number of active constraints at θ_0, is itself a function of θ_0. Then

$$\mathcal{T}(\Theta_0; \theta_0) = \{\alpha : a_{j_1}^T \alpha \geqslant 0, \ldots, a_{j_m}^T \alpha \geqslant 0\}.$$

Because $\mathcal{T}(\Theta_0; \theta_0)$ is a polyhedral, it is a closed convex cone, and hence it follows that $\|Z - \mathcal{T}(\Theta_0; \theta_0)\|^2$, the limiting distribution of *LRT*, is a chi-bar-square (see Theorem 3.8.2).

Now, let $\boldsymbol{H}(\theta_0) = [a_{j_1}, \ldots, a_{j_m}]^T$. Let us consider the case when $rank\{\boldsymbol{H}(\theta_0)\} = m(\theta_0)$. Since $LRT \xrightarrow{d} \|Z - \mathcal{T}(\Theta_0; \theta_0)\|^2$, we have that

$$\text{pr}(LRT \geqslant c | \boldsymbol{\theta} = \boldsymbol{\theta}_0) \to \sum_{i=0}^{m} w_{m-i}\{m, \boldsymbol{V}(\boldsymbol{\theta}_0)\}\text{pr}(\chi_i^2 \geqslant c), \tag{4.116}$$

where $\boldsymbol{V}(\boldsymbol{\theta}_0) = \boldsymbol{H}(\boldsymbol{\theta}_0)\mathcal{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{\theta}_0)^T$. With t denoting the sample value of *LRT*, the large sample p-value, p_{sup}, is

$$\sup_{\boldsymbol{\theta}_0 \in \Theta_0} \sum_{i=0}^{m} w_{m-i}\{m, \boldsymbol{V}(\boldsymbol{\theta}_0)\}\text{pr}(\chi_i^2 \geqslant t).$$

In general, the value of $\boldsymbol{\theta}_0$ at which this supremum is achieved may also depend on t. Consequently, there may not be a single least favorable null value for all possible values of t. It is not easy to maximize the tail probability of a chi-bar square distribution numerically using Newton-type iteration, because analytical expressions are not available for the derivatives of the tail probability of (4.115). Recent developments using simulation approaches may be helpful (see Shapiro, 2000); the utility of this approach remains to be investigated. In general, there is no easy way to compute the supremum in (4.115).

If computation of the supremum in (4.115) does not appear feasible, we can apply a bounds test using,

$$\sup_{\boldsymbol{\theta}_0 \in \Theta_0} \sum_{i=0}^{p} w_i\{p, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}, \mathcal{T}^o(\Theta_0; \boldsymbol{\theta}_0)\}\text{pr}(\chi_i^2 \geqslant c)$$
$$\leq 0.5\{\text{pr}(\chi_{p-1}^2 \geqslant t) + \text{pr}(\chi_p^2 \geqslant t)\}.$$

If the maximum number of inequalities in the null hypothesis that can be satisfied simultaneously as equalities is, say q, and $q \leq p$, then the foregoing upper bound could be sharpened to $0.5\{\text{pr}(\chi_{q-1}^2 \geqslant t) + \text{pr}(\chi_q^2 \geqslant t)\}$. Thus, while we can write down the asymptotic distribution of statistics for testing $\boldsymbol{h}(\boldsymbol{\theta}) \geqslant \boldsymbol{0}$ vs $\boldsymbol{h}(\boldsymbol{\theta}) \not\geqslant \boldsymbol{0}$, there is scope for developing practically feasible computational procedures to implement them.

The foregoing result extends in a natural way for testing

$$H_0 : \boldsymbol{h}_1(\boldsymbol{\theta}) = \boldsymbol{0}, \boldsymbol{h}_2(\boldsymbol{\theta}) \geqslant \boldsymbol{0} \quad \text{vs} \quad H_1 : \text{No restrictions on } \boldsymbol{\theta}$$

where $\boldsymbol{h} = (\boldsymbol{h}_1 : \boldsymbol{h}_2)$ is a partition of $\boldsymbol{h}$ into two subvectors; only minor changes are required for this extension (see Kodde and Palm (1986) and Wolak (1989a)).

Example 1. Let $\boldsymbol{\theta} \in \mathbb{R}^2$. Consider a statistical model satisfying Condition Q2 in Section 4.8.2. Let the null and alternative hypotheses be

$$H_0 : \boldsymbol{\theta} \geqslant \boldsymbol{0} \text{ and } H_1 : \boldsymbol{\theta} \not\geqslant \boldsymbol{0},$$

respectively, where $\boldsymbol{\theta} \in \mathbb{R}^2$. Then $LRT = \sup\{\ell(\boldsymbol{\theta}) : \boldsymbol{\theta} \geqslant \boldsymbol{0}\}$. Suppose that the null hypothesis holds and let $\boldsymbol{\theta}_0$ denote the true null value. Then, the asymptotic null distribution of *LRT* is equal to that of $\|Z - \mathcal{T}(\Theta_0; \boldsymbol{\theta}_0)\|^2$ where Θ_0 is the positive orthant, $Z \sim N\{\boldsymbol{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}\}$ and $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\mathcal{I}(\boldsymbol{\theta}_0)\boldsymbol{x}$. Clearly, $\mathcal{T}(\Theta_0; \boldsymbol{\theta}_0)$, and hence the foregoing null asymptotic distribution of *LRT*, depend on $\boldsymbol{\theta}_0$. We shall consider four different cases.

Case 1: $\boldsymbol{\theta}_0 = (0, 0)^T$. It is easily seen that $\mathcal{T}(\Theta_0; \boldsymbol{\theta}_0)$ is the positive orthant and therefore,

$$\text{pr}(LRT \geqslant c \mid \boldsymbol{\theta} = \boldsymbol{0}) \to (1/2)\text{pr}(\chi_1^2 \geqslant c) + q\text{pr}(\chi_2^2 \geqslant c),$$

where $q = (2\pi)^{-1}\cos^{-1}\{\mathcal{I}^{12}/(\mathcal{I}^{11}\mathcal{I}^{22})^{1/2}\}$ and $\mathcal{I}(\boldsymbol{0})^{-1} = (\mathcal{I}^{ij})_{2\times 2}$.

Case 2: $\boldsymbol{\theta}_0 = (0, a)^T$, where $a > 0$. Because $\mathcal{T}(\Theta_0; \boldsymbol{\theta}_0)$ is the half-space $\{\alpha \in \mathbb{R}^2 : \alpha_1 \geqslant 0\}$,

$$\text{pr}\{LRT \geqslant c \mid \boldsymbol{\theta} = (0, a)\} \to (1/2)\text{pr}(\chi_1^2 \geqslant c).$$

Case 3: $\boldsymbol{\theta}_0 = (a, 0)^T$, where $a > 0$. The asymptotic null distribution is the same as that in Case 2.

Case 4: $\boldsymbol{\theta}_0 = (a, b^T)$, where $a > 0$ and $b > 0$. Now, $\mathcal{T}(\Theta_0; \boldsymbol{\theta}_0) = \mathbb{R}^2$, and therefore, $\|Z - \mathcal{T}(\Theta_0; \boldsymbol{\theta}_0)\| = 0$ and $\text{pr}\{LRT = 0 \mid \boldsymbol{\theta} = (a, b)\} \to 1$.

It follows from the foregoing four cases that $\boldsymbol{\theta} = \boldsymbol{0}$ is the least favorable null value. Therefore, the large sample p-value corresponding to $LRT = t$ is $(1/2)\text{pr}(\chi_1^2 \geqslant t) + q\text{pr}(\chi_2^2 \geqslant t)$. ■

Example 2. Let the testing problem be

$$H_0 : \boldsymbol{h}(\boldsymbol{\theta}) \geqslant \boldsymbol{0} \text{ vs } H_1 : \boldsymbol{h}(\boldsymbol{\theta}) \not\geqslant \boldsymbol{0}$$

where $\boldsymbol{h} = (h_1 : h_2)$. Suppose that $h_1(\boldsymbol{\theta}) = h_2(\boldsymbol{\theta}) = 0$ has a solution, say at $\boldsymbol{\theta}_0$ and $rank[\boldsymbol{H}(\boldsymbol{\theta}_0)] = 2$. As in the previous example, it would be convenient to consider four different cases.

Case 1: $h_1(\boldsymbol{\theta}_0) = 0$ and $h_2(\boldsymbol{\theta}_0) = 0$

$$\text{pr}\{LRT \geqslant c \mid h_1(\boldsymbol{\theta}_0) = 0, h_2(\boldsymbol{\theta}_0) = 0\}$$
$$\to (1/2)\text{pr}(\chi_1^2 \geqslant c) + q(\boldsymbol{\theta}_0)\text{pr}(\chi_2^2 \geqslant c),$$

where

$$q(\boldsymbol{\theta}) = (2\pi)^{-1}\cos^{-1}\{V_{12}/(V_{11}V_{22})^{1/2}\}, \text{ and } \boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta}_0).$$

Case 2: $h_1(\boldsymbol{\theta}_0) > 0$ and $h_2(\boldsymbol{\theta}_0) = 0$

$$\text{pr}\{LRT \geqslant c \mid h_1(\boldsymbol{\theta}_0) > 0, h_2(\boldsymbol{\theta}_0) = 0\} \to (1/2)\text{pr}(\chi_1^2 \geqslant c).$$

Case 3: $h_1(\boldsymbol{\theta}_0) = 0$ and $h_2(\boldsymbol{\theta}_0) > 0$

$$\text{pr}\{LRT \geqslant c \mid h_1(\boldsymbol{\theta}_0) = 0, h_2(\boldsymbol{\theta}_0) > 0\} \to (1/2)\text{pr}(\chi_1^2 \geqslant c);$$

Case 4: $h_1(\boldsymbol{\theta}_0) > 0$ and $h_2(\boldsymbol{\theta}_0) > 0$

$$\text{pr}\{LRT = 0 \mid h_1(\boldsymbol{\theta}_0) > 0, h_2(\boldsymbol{\theta}_0) > 0\} \to 1.$$

Therefore, the least favorable null value is a point at which $h_1(\boldsymbol{\theta}) = h_2(\boldsymbol{\theta}) = 0$. If $h_1(\boldsymbol{\theta}) = h_2(\boldsymbol{\theta}) = 0$ has more than one solution, then the solution at which $q(\boldsymbol{\theta})$ is the maximum is the least favorable null value. ■

Example 3 (Wolak (1991)). Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and

$$H_0 : h_1(\boldsymbol{\theta}) \geqslant 0 \text{ and } h_2(\boldsymbol{\theta}) \geqslant 0,$$

where $h_1(\boldsymbol{\theta}) = \theta_2 - \theta_1^2$ and $h_2(\boldsymbol{\theta}) = \theta_1 - \theta_2$; the parameter space is shown in Fig. 4.8. As in the previous example, there are four cases to be considered. It follows that the least favorable null value is a $\boldsymbol{\theta}_0$ satisfying $h_1(\boldsymbol{\theta}_0) = 0$ and $h_2(\boldsymbol{\theta}_0) = 0$. Let t denote the sample value of *LRT*. Because there are two values of $\boldsymbol{\theta}_0$, namely $(0,0)$ and $(1,1)$, satisfying these equations, and

$$\text{pr}(LRT \geqslant t | \boldsymbol{h}(\boldsymbol{\theta}_0) = \mathbf{0}) \to (1/2)\text{pr}(\chi_1^2 \geqslant t) + q(\boldsymbol{\theta}_0)\text{pr}(\chi_2^2 \geqslant t),$$

the least favorable null value is the $\boldsymbol{\theta}_0$ in $\{(0,0),(1,1)\}$ for which

$$(1/2)\text{pr}(\chi_1^2 \geqslant t) + q(\boldsymbol{\theta}_0)\text{pr}(\chi_2^2 \geqslant t)$$

is a maximum. If $\mathcal{I}(\boldsymbol{\theta}_0)$ is known for $\boldsymbol{\theta}_0 = (0,0)$ and $(1,1)$, then we can compute the foregoing tail probability for each of the two values of $\boldsymbol{\theta}_0$ and choose the largest as the p-value. ■

In the foregoing examples with only two inequality constraints in the null hypothesis, the least favorable null value turned out to be a point at which all the inequalities in the null hypothesis are active. We also observed in Chapter 2 that for testing $H_0 : \boldsymbol{\theta} \geqslant \mathbf{0}$ against $H_1 : \boldsymbol{\theta} \ngeqslant \mathbf{0}$, based on observations from $N(\boldsymbol{\theta}, \boldsymbol{V})$ where $\boldsymbol{V}$ does not depend on $\boldsymbol{\theta}$, the least favorable null value, namely $\mathbf{0}$, is the point at which every inequality constraint in the null hypothesis is active. However, in general, if the null hypothesis is $H_0 : h_1(\boldsymbol{\theta}) \geqslant 0, \ldots, h_k(\boldsymbol{\theta}) \geqslant 0$, where $k > 2$, then the least favorable null value may not be a point at which all the inequalities are active. The following example from Wolak (1991) illustrates this.

Example 4 (Wolak (1991)). Let $\boldsymbol{X} = (X_1, X_2)^T \sim N(\mathbf{0}, \boldsymbol{V})$ where

$$\boldsymbol{V} = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}.$$

Let $\boldsymbol{\theta} = (\sigma^2, \tau^2, \rho)^T$,

$$\Theta = \{\boldsymbol{\theta} : \theta_1 > 0, \theta_2 > 0, -1 < \theta_3 < 1\},$$

$\boldsymbol{\theta}^A = (1, 1, 0.95)^T$ and $\boldsymbol{\theta}^B = (1, 1, 0)^T$. Let the null hypothesis be $H_0 : \boldsymbol{\theta} \leq \boldsymbol{\theta}^A$. To express this testing problem in the notation of (4.113), let $\boldsymbol{h}(\boldsymbol{\theta}) = (1 - \theta_1, 1 - \theta_2, 0.95 - \theta_3)$; then $\boldsymbol{H}(\boldsymbol{\theta}_0) = -\boldsymbol{I}$. The inverse of the information matrix and the corresponding correlation matrix are given below (see Lehmann (1983), p 441).

$$\mathcal{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} 2\sigma^4 & 2\rho^2\sigma^2\tau^2 & \rho(1-\rho^2)\sigma^2 \\ 2\rho^2\sigma^2\tau^2 & 2\tau^4 & \rho(1-\rho^2)\tau^2 \\ \rho(1-\rho^2)\sigma^2 & \rho(1-\rho^2)\tau^2 & (1-\rho^2)^2 \end{bmatrix}$$

$$corr\{\mathcal{I}(\boldsymbol{\theta})^{-1}\} = \begin{bmatrix} 1 & \rho^2 & \rho/\sqrt{2} \\ \rho^2 & 1 & \rho/\sqrt{2} \\ \rho/\sqrt{2} & \rho/\sqrt{2} & 1 \end{bmatrix}.$$

Note that Θ is open, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}\}$, and Condition $Q2$ in subsection 4.8.2 is satisfied. Now,

$$\begin{aligned} \lim_{n\to\infty} \text{pr}(LRT \geqslant c \mid \boldsymbol{\theta} = \boldsymbol{\theta}^A) = w_1\{3, \mathcal{I}(\boldsymbol{\theta}_0)\}\text{pr}(\chi_3^2 \geqslant c) + \\ w_2\{3, \mathcal{I}(\boldsymbol{\theta}_0)\}\text{pr}(\chi_2^2 \geqslant c) + w_3\{3, \mathcal{I}(\boldsymbol{\theta}_0)\}\text{pr}(\chi_1^2 \geqslant c), \end{aligned} \tag{4.117}$$

where w_1, w_2, and w_3 are determined by the formulas in (3.26). By direct substitution into these formulas, we have $w_1 = 0.0153, w_2 = 0.1682, w_3 = 0.4847$,

$$\lim_{n\to\infty} \text{pr}\{LRT \geqslant c \mid \boldsymbol{\theta} = \boldsymbol{\theta}^B\} = 0.5\{\text{pr}(\chi_1^2 \geqslant c) + 0.5\text{pr}(\chi_2^2 \geqslant c)\}, \tag{4.118}$$

$$\text{and} \quad \lim_{n\to\infty} \text{pr}(LRT \geqslant c \mid \boldsymbol{\theta} = \boldsymbol{\theta}^A) < \text{pr}\{LRT \geqslant c \mid \boldsymbol{\theta} = \boldsymbol{\theta}^B\}.$$

Therefore, the least favorable null value is not $\boldsymbol{\theta}^A$, the only point at which all the inequalities in the null hypothesis are active.

Let $\boldsymbol{\theta}^*$ be a point at which exactly two of those inequalities are active. Then,

$$\lim \text{pr}\{LRT \geqslant c \mid \boldsymbol{\theta} = \boldsymbol{\theta}^*\} = 0.5\text{pr}(\chi_1^2 \geqslant c) + q\text{pr}(\chi_2^2 \geqslant c),$$

for some q in the range, $0 \leq q \leq 0.5$. Let $\boldsymbol{\theta}^+$ be a point at which exactly one of the inequalities in H_0 is active. Then,

$$\lim \text{pr}\{LRT \geqslant c \mid \boldsymbol{\theta} = \boldsymbol{\theta}^+\} = 0.5\text{pr}(\chi_1^2 \geqslant c).$$

Since this limit is smaller than the RHS of (4.118), it follows that $\boldsymbol{\theta} = \boldsymbol{\theta}^B$ is the least favorable null value. ■

Example 4.8.3 (Stram and Lee, 1994) *Testing for the presence of random effects.*

Consider the growth curve model in Example 1.2.10 (p 14). The data consist of measurements of the distance (in *mm*) from the center pituitary to the pteryomaxillary fissure for 27 children (11 girls and 16 boys). Every subject has four measurements, taken at ages 8, 10, 12 and 14 years. The object is to model the growth in the distance as a function of age and sex. To allow for individual variability in the growth function, the model being considered is

$$\boldsymbol{y}_i = (1, Age, Sex, Age.Sex)\boldsymbol{\alpha} + (1, Age)\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{y}_i$ is 4×1 for the four measurements of the ith subject, $Age = (8,10,12,14)^T$, $Sex = (a,a,a,a)^T$ where a is 0 for boys and 1 for girls. The parameter $\boldsymbol{\alpha}(4\times 1)$ captures the fixed effects and $\boldsymbol{\beta}_i(2\times 1)$ captures the random effect of subject i. Assume that $\boldsymbol{\epsilon}_i \sim N(0,\sigma^2\boldsymbol{I})$, $\boldsymbol{\beta}_i \sim N(\boldsymbol{0},\Psi)$ for some 2×2 positive semi-definite matrix Ψ and that $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\beta}_i$ are independent. The following models are of interest:

1. Model $M1 : \Psi = \boldsymbol{0}$. There are no random effects.
2. Model $M2 : \Psi = (\psi_{11}, 0 \mid 0, 0)$. The individual variability is captured by a random intercept; therefore, the regression planes are parallel for different subjects.
3. Model $M_3 : \Psi = (\psi_{11}, \psi_{12} \mid \psi_{21}, \psi_{22})$. Individual variability results in not only a random intercept but also a different slope parameter for age.

To simplify notation, let us write $\Psi = (\psi_1, \psi_2 \mid \psi_2, \psi_3)$. In the process of modeling these data, it is possible that we may be interested to know whether or not the fit for Model M3 is significantly better than that for the simpler Model M2. Therefore, we formulate the testing problem as

$$H_0 : \psi_1 \geqslant 0, \psi_2 = \psi_3 = 0 \quad \text{vs} \quad H_1 : \psi_1 \geqslant 0, \psi_3 \geqslant 0, \psi_1\psi_3 - \psi_2^2 \geqslant 0.$$

The regression model can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i.$$

With the assumptions as introduced earlier for the distributions of $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\beta}_i$, we have

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}(-n_i/2)\log(2\pi) - (1/2)\log\{\det(\boldsymbol{\Sigma}_i^{-1})\} - (1/2)\text{trace}(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{S}_i)$$

where $\boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{I} + \boldsymbol{Z}_i\Psi\boldsymbol{Z}_i^T$ and $\boldsymbol{S}_i = (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha})^T(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha})$. Now, under the null,

$$LRT \xrightarrow{d} \|U - \mathcal{T}(\Theta_0;\boldsymbol{\theta}_0)\|^2 - \|U - \mathcal{T}(\Theta_1;\boldsymbol{\theta}_0)\|^2$$

where $\boldsymbol{U} \sim N(\boldsymbol{0}, \mathcal{I}_{\boldsymbol{\theta}_0}^{-1})$,

$$\Theta_0 = \{\boldsymbol{\psi} \in \mathbb{R}^3 : \psi_2 = \psi_3 = 0, \psi_1 \geqslant 0\},$$
$$\text{and } \Theta_1 = \{\boldsymbol{\psi} \in \mathbb{R}^3 : \psi_1 \geqslant 0, \psi_3 \geqslant 0, \psi_1\psi_3 - \psi_2^2 \geqslant 0\}.$$

There are two possible values of $\boldsymbol{\theta}_0$ in Θ_0 that we need to consider separately: $(0,0,0)^T$ and $(a,0,0)^T$ where $a > 0$.

Case 1: $\boldsymbol{\theta}_0 = (0,0,0)^T$.

It may be verified that

$$\mathcal{T}(\Theta_0;\boldsymbol{\theta}_0) = \{\boldsymbol{\psi} \in \mathbb{R}^3 : \psi_2 = \psi_3 = 0, \psi_1 \geqslant 0\}.$$

To determine the tangent cone of Θ_1, note that

$$h_3(\boldsymbol{\theta}) = \psi_1\psi_3 - \psi_2^2, \qquad \text{and} \qquad (\partial/\partial\boldsymbol{\theta})h_3(\boldsymbol{\theta}) = (\psi_3, -2\psi_2, \psi_1)^T.$$

At $\boldsymbol{\theta}_0 = (0,0,0)^T$ we have $h_3(\boldsymbol{\theta}_0) = 0$ and $(\partial/\partial\boldsymbol{\theta})h_3(\boldsymbol{\theta}_0) = \boldsymbol{0}$. Therefore, MF-CQ is not satisfied at $\boldsymbol{\theta}_0$, and hence we cannot use Proposition 4.7.3 to write down the tangent cone; thus we should not simply replace the nonlinear constraints by their linear approximations at $\boldsymbol{\theta}_0$.

Since Θ_1 is a closed convex cone with vertex at the origin it follows that $\mathcal{T}(\Theta_1;\boldsymbol{\theta}_0) = \Theta_1$. Since neither $\mathcal{T}(\Theta_1;\boldsymbol{\theta}_0)$ nor $\mathcal{T}(\Theta_0;\boldsymbol{\theta}_0)$ is a linear space, the asymptotic null distribution of *LRT* may not be a chi-bar square. Exact distribution of *LRT* for this case is nonstandard. A bound can be obtained by replacing $\mathcal{T}(\Theta_1;\boldsymbol{\theta}_0)$ by the larger set $\{\psi_1 \geqslant 0, \psi_3 \geqslant 0\}$. In this case, the distribution is a mixture of chi-square distributions.

Case 2: $\boldsymbol{\theta}_0 = (a,0,0)^T$ where $a > 0$.

It may be verified that

$$\mathcal{T}(\Theta_0;\boldsymbol{\theta}_0) = \{\boldsymbol{\psi} \in \mathbb{R}^3 : \psi_2 = \psi_3 = 0\},$$
$$\text{and} \quad \mathcal{T}(\Theta_1;\boldsymbol{\theta}_0) = \{\boldsymbol{\psi} \in \mathbb{R}^3 : \psi_3 \geqslant 0\},$$

the Mangasarian-Fromowitz constraint qualification is satisfied at $\boldsymbol{\theta}_0$ and therefore we can use Proposition 4.7.3 to write down the tangent cones. Since $\mathcal{T}(\Theta_0;\boldsymbol{\theta}_0)$ is a linear space and $\mathcal{T}(\Theta_1;\boldsymbol{\theta}_0)$ is a closed convex cone it follows that the asymptotic null distribution of *LRT* is a chi-bar square. ■

4.9 PROPERTIES OF THE MLE WHEN THE TRUE VALUE IS ON THE BOUNDARY

We continue to consider the *iid* setting for simplicity although most of the results in this section would hold in more general settings as will be indicated later. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ denote *iid* observations from a population with density function $f(\boldsymbol{x};\boldsymbol{\theta})$. Let $\ell(\boldsymbol{\theta})$ denote the loglikelihood, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Let $\Omega \subset \Theta$ where Ω is not necessarily open. In this section we study large sample properties of local and global *mle*'s of $\boldsymbol{\theta}$. Let us denote an *mle*, global or local, by $\tilde{\boldsymbol{\theta}}$. The true value, denoted by $\boldsymbol{\theta}_0$, will be assumed to be a boundary point of Ω unless the contrary is made clear. *Assume that Condition Q is satisfied.* For simplicity, we shall restrict to maximum likelihood. However, most of the results of this section hold for M-estimators as well; for details of this case see Geyer (1994) and Shapiro (2000a).

Distribution of the *mle* over a restricted parameter space that is not open has not attracted as much attention as the corresponding hypothesis testing problems. If θ_0 were an interior point then $\sqrt{n}\hat{\mathcal{I}}^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I)$. Since this limiting distribution does not depend on any unknown parameters (i.e., $\sqrt{n}\hat{\mathcal{I}}^{-1/2}(\hat{\theta} - \theta_0)$ is a pivotal quantity), we can use this result to construct a confidence region for θ. Unfortunately, the asymptotic distribution of $\sqrt{n}\hat{\mathcal{I}}^{-1/2}(\tilde{\theta} - \theta_0)$ is a discontinuous function of θ_0. Further, it is unknown whether or not there is a a pivotal quantity based on $\tilde{\theta}$ that would lend itself for constructing a confidence region. Consequently, statistical inference based on the constrained estimator $\tilde{\theta}$ has not attracted much attention. Nevertheless, it is of interest to study properties of the constrained *mle* because it has an important role to play in the theory of statistical inference. In this section, we shall consider some examples to illustrate the main results. Then we shall state the main results and provide references to sources where detailed statements of the regularity conditions and proofs may be found.

At this stage it would be helpful to reconsider Example 3.3.1.

Example 4.9.1 *Asymptotic distribution of mle when* $\Omega = \mathbb{R}^{+2}$ *and* $X \sim N(\theta, I)$.

Let $X \sim N(\theta, I)$ and $\Omega = \mathbb{R}^{+2}$. Let $X_1, \ldots, X_n$ be *iid* as X. Then,

$$\ell(\theta) = (-n/2)\|\bar{X} - \theta\|^2$$

and $\tilde{\theta}$ is the point in $\mathbb{R}^{+2}$ that is closest to $\bar{X}$. Thus,

$$\tilde{\theta} = \Pi(\bar{X}, \mathbb{R}^{+2}) \text{ and } \tilde{\theta} \xrightarrow{p} \theta_0.$$

If θ_0 is an interior point of $\mathbb{R}^{+2}$ then $\text{pr}(\bar{X} = \tilde{\theta}) \to 1$ and hence $\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, I)$. Let $\mathcal{T} = \mathcal{T}(\mathbb{R}^{+2}; \theta_0)$, $Z_n = \sqrt{n}(\bar{X} - \theta_0)$ and $Z \sim N(0, I)$. Suppose that θ_0 lies on the boundary of $\mathbb{R}^{+2}$. Then it is easily seen that

$$(\tilde{\theta} - \theta_0) = \Pi(\bar{X} - \theta_0, \mathcal{T}) \text{ with probability approaching 1.}$$

Therefore,

$$\sqrt{n}(\tilde{\theta} - \theta_0) = \Pi(Z_n, \mathcal{T}) \text{ with prob approaching 1.}$$

Since $\Pi(z, \mathcal{T})$ is continuous in z and $Z_n \xrightarrow{d} Z$ it follows that $\Pi(Z_n, \mathcal{T}) \xrightarrow{d} \Pi(Z, \mathcal{T})$. Therefore, we have

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} \Pi(Z, \mathcal{T}).$$

This may be verified directly (in fact, it would be instructive to do so) by considering the following three cases separately: (i) θ_0 is on the positive θ_1-axis (this is the case shown in Fig. 4.10), (ii) θ_0 is on the positive θ_2-axis and (iii) θ_0 is at the origin. It is perhaps easier to visualize this result as

$$(\tilde{\theta} - \theta_0) \approx \Pi(Z/\sqrt{n}; \mathcal{T}),$$

for large n as shown in Fig. 4.10. ■

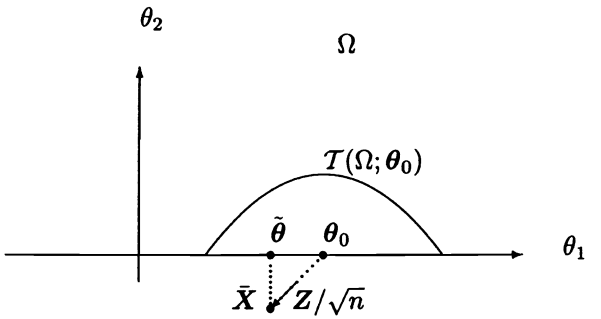

Fig. 4.10 $(\tilde{\theta} - \theta_0) \approx \Pi\{Z/\sqrt{n}; \mathcal{T}(\Omega; \theta_0)\}$,where $Z \sim N(0, I)$.

When θ_0 is an interior point, a consistent *mle* $\hat{\theta}$, whether it is a global or local *mle*, satisfies $n^{-1/2}\hat{\mathcal{I}}^{-1/2}S(\theta_0) = \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1)$. This is a neat representation of $\hat{\theta}$ as a sum of independently and identically distributed variables ($S(\theta_0)$ is a sum of *iid* variables). This is a direct result of applying a one-term Taylor expansion on the first-order condition $\nabla\ell(\hat{\theta}) = 0$. If θ_0 is not an interior point, then the first-order conditions, such as the Kuhn-Tucker conditions, involve inequalities. Consequently, $\sqrt{n}(\tilde{\theta} - \theta_0)$ does not have a simple representation as for $\sqrt{n}(\hat{\theta} - \theta_0)$. In fact, the properties of local *mle*'s are very subtle. Let us consider a simple example to illustrate this.

Example 4.9.2 *Nonconvex parameter space*

Let X have the bivariate normal distribution $N(0, I)$ and let $X_1, \ldots, X_n$ be *iid* as X. Then $\ell(\theta) = (-n/2)\|\bar{X} - \theta\|^2$. Let Ω be the union of the θ_1-axis and the θ_2-axis; thus $\Omega = \{(\theta_1, \theta_2) : \theta_1 = 0 \text{ or } \theta_2 = 0\}$. In this case, there are two local *mle*'s: $\tilde{\theta}^a = (\bar{X}_1, 0)$ and $\tilde{\theta}^b = (0, \bar{X}_2)$. It is clear that their large sample distributions are different. Further, when θ_0 is on the θ_1-axis, we have the following with probability $\to 1$:

$$\tilde{\theta}^a - \theta_0 = \text{ projection of } (\bar{X} - \theta_0) \text{ onto the } \theta_1\text{-axis,}$$

and $\sqrt{n}(\tilde{\theta}^a - \theta_0)$ converges to $\Pi(Z, \mathcal{P})$ where $Z \sim N(0, I)$ and $\mathcal{P} = \{\theta \in \mathbb{R}^2 : \theta_2 = 0\}$. In fact, we have

$$\sqrt{n}(\tilde{\theta}^a - \theta_0) \begin{cases} \xrightarrow{d} \Pi(Z, \mathcal{P}) & \text{if } \theta_0 \in \theta_1\text{-axis} \\ = \begin{bmatrix} \sqrt{n}\bar{X}_1 \\ -\sqrt{n}\theta_{02} \end{bmatrix} & \text{if } \theta_0 \notin \theta_1\text{-axis.} \end{cases}$$

A similar comment applies to $\tilde{\theta}^b$ by symmetry.

By contrast, with $\tilde{\theta}$ denoting the global *mle*, it is easily seen that

$$(\tilde{\theta} - \theta_0) = \Pi(\bar{X} - \theta_0, \mathcal{T}(\Omega; \theta_0)) \text{ with prob approaching 1.}$$

Therefore,

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \Pi(\boldsymbol{Z}, \mathcal{T}(\Omega; \boldsymbol{\theta}_0)), \text{ where } \boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}).$$

Thus, local and global *mle*'s have different asymptotic distributions. If the parameter space is $\mathbb{R}^{+2}$ then there is only one maximum, hence local and global *mle*'s are the same. This example illustrates that if the parameter space is not convex, or at least nearly convex in a way to be explained later, the global and local *mle*'s may have different properties. A consequence of such a difference between local *mle*'s is that if an iterative algorithm with Newton-Raphson type steps were to be used to find the global maximum, it could be trapped around a local maximum. ■

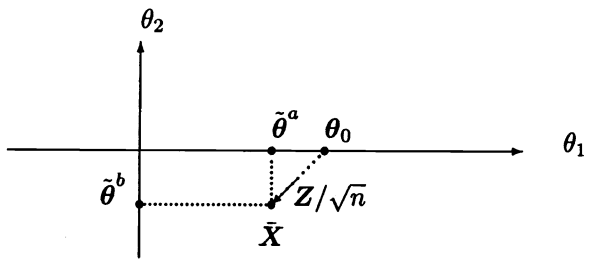

Fig. 4.11 Two local *mle*s when Ω is the union of θ_1-axis and the θ_2-axis, and $\boldsymbol{\theta}_0$ is not on the θ_2-axis; $\sqrt{n}(\tilde{\boldsymbol{\theta}}^a - \boldsymbol{\theta}_0)$ converges in distribution but $\|\sqrt{n}(\tilde{\boldsymbol{\theta}}^b - \boldsymbol{\theta}_0)\| \to \infty$.

The standard method that is adopted in the literature to study large sample properties of consistent *mle*'s is very similar to that was adopted so far to study the properties of *LRT*. To illustrate the main idea, let us first consider the case when the parameter space Ω is a closed convex cone with its vertex at $\boldsymbol{\theta}_0$. Let $\mathcal{T} = \Omega - \boldsymbol{\theta}_0$; thus $\mathcal{T}$ is the tangent cone of Ω at $\boldsymbol{\theta}_0$. Suppose that Condition Q is satisfied. Let us write the quadratic approximation of $\ell(\boldsymbol{\theta})$ as (see (4.3))

$$\ell(\boldsymbol{\theta}) = K_n - 2^{-1}(\boldsymbol{Z}_n - \boldsymbol{u})^T \boldsymbol{I}_{\boldsymbol{\theta}_0}(\boldsymbol{Z}_n - \boldsymbol{u}) + \delta_n(\boldsymbol{u}),$$

where $\boldsymbol{u} = \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and K_n does not depend on $\boldsymbol{\theta}$. Let

$$\boldsymbol{\theta}^\dagger = \text{argmin}_{\boldsymbol{\theta} \in \Omega}(\boldsymbol{Z}_n - \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0))^T \boldsymbol{I}_{\boldsymbol{\theta}_0}(\boldsymbol{Z}_n - \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)).$$

A consequence of $\delta_n(\boldsymbol{u}) = n^{-1/2}\|\boldsymbol{u}\|^3 O_p(1)$ is that

$$n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\dagger) = o_p(1)$$

(see Self and Liang (1987, Lemma 2)). Therefore, $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $n^{1/2}(\boldsymbol{\theta}^\dagger - \boldsymbol{\theta}_0)$ have the same asymptotic distribution. Now, recall the following notation for projection:

$$\Pi_{\boldsymbol{W}}(\boldsymbol{z}, \mathcal{T}) = \text{argmin}_{\boldsymbol{u} \in \mathcal{T}}(\boldsymbol{z} - \boldsymbol{u})^T \boldsymbol{W}^{-1}(\boldsymbol{z} - \boldsymbol{u}).$$

Since $\mathcal{T}$ is convex and projection onto a convex set is distance reducing, it follows that $\Pi_{\boldsymbol{W}}(\boldsymbol{z}, \mathcal{T})$ is continuous in $\boldsymbol{z}$ for every fixed $\boldsymbol{W}$; in fact, it follows from

$$\|\boldsymbol{z}^t(\boldsymbol{W}_1 - \boldsymbol{W}_2)\boldsymbol{z}\| \le \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|\|\boldsymbol{z}\|^2$$

that

$$\Pi_{\boldsymbol{W}}(\boldsymbol{z}, \mathcal{T}) \text{ is continuous in } (\boldsymbol{z}, \boldsymbol{W}).$$

Since $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})$ and $\sqrt{n}(\boldsymbol{\theta}^\dagger - \boldsymbol{\theta}_0) = \Pi(\boldsymbol{Z}_n, \mathcal{T})$, it follows that

$$\sqrt{n}(\boldsymbol{\theta}^\dagger - \boldsymbol{\theta}_0) \xrightarrow{d} \Pi(\boldsymbol{Z}, \mathcal{T}).$$

Therefore, the asymptotic distribution of the global *mle* is given by

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \Pi(\boldsymbol{Z}, \mathcal{T}).$$

These arguments can also be modified when Ω is not necessarily a cone, but provided $\mathcal{T}(\Omega, \boldsymbol{\theta}_0)$ is convex; see Andrews (1999) and Le Cam (1970, p 820) for details. This result has also been extended to the case when $\mathcal{T}(\Omega, \boldsymbol{\theta}_0)$ is not convex, but Ω is Chernoff regular at $\boldsymbol{\theta}_0$. The next result states the most general form known for the *iid* setting.

Proposition 4.9.1 *Suppose that Ω is Chernoff regular at $\boldsymbol{\theta}_0$, and let $\tilde{\boldsymbol{\theta}}$ denote the global mle. Suppose that Condition Q is also satisfied. Then*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \Pi_{\mathcal{I}(\boldsymbol{\theta}_0)^{-1}}\{\boldsymbol{Z}, \mathcal{T}(\Omega; \boldsymbol{\theta}_0)\}, \quad \textit{where } \boldsymbol{Z} \sim N(\boldsymbol{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}).$$

Proof: See Self and Liang (1987, Theorem 2) and Geyer (1994, Theorem 4.4); the proof in Geyer (1994) is applicable for M-estimators as well. ■

To state the foregoing result differently, let $\boldsymbol{Y} \sim N(\boldsymbol{\theta}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1})$ where $\boldsymbol{\theta} \in \Omega$. Let G denote the distribution of the *mle* of $\boldsymbol{\theta}$ based on a single observation of $\boldsymbol{Y}$ when $\boldsymbol{\theta} = \boldsymbol{0}$. Then the asymptotic distribution of $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is G. This is illustrated in Fig. 4.12. It shows that the first order asymptotic behavior of $(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is the same as that of $\Pi\{\boldsymbol{Z}/\sqrt{n}, \mathcal{T}(\Omega; \boldsymbol{\theta}_0)\}$,where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1})$.

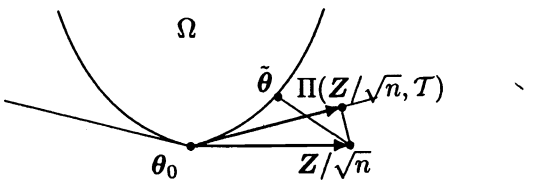

Fig. 4.12 $(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx \Pi\{\boldsymbol{Z}/\sqrt{n}, \mathcal{T}\}$,where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \mathcal{I}(\boldsymbol{\theta}_0))$ and $\mathcal{T} = \mathcal{T}(\Omega; \boldsymbol{\theta}_0)$.

To study properties of local *mle*'s, let us introduce the terms *near convexity of a set* and *pro-regular*.

Definition: We say that the set Ω is *nearly convex* at θ_0 if there exists a neighborhood V of θ_0 and a function $k(\theta, \theta^*)$ tending to zero as $\theta \to \theta_0, \theta^* \to \theta_0$, such that

$$dist\{\theta^* - \theta, T(\Omega; \theta)\} \leq k(\theta, \theta^*)\|\theta^* - \theta\|, \quad \forall \theta, \theta^* \in \Omega \cap V.$$

We say that the set Ω is *pro-regular* at θ_0 if there exists a neighborhood V of θ_0 and a constant K such that

$$dist\{\theta^* - \theta, T(\Omega; \theta)\} \leq K\|\theta^* - \theta\|^2, \quad \forall \theta, \theta^* \in \Omega \cap V.$$

If Ω is convex at θ_0 then it is nearly convex and pro-regular at θ_0. The concept of *near convexity* was introduced by Shapiro and Al-Khayyal (1993). *Pro-regularity* appears in Shapiro (1994) under the name "O(2)-Convexity"; for related details on this topic see Rockafellar and Wets (1998).

Another concept that is related to this is *monotonicity of normals*. This is closely related to what is known as *Clark regularity,* convexity of tangent cones and near convexity. See Shapiro (2000a) for details about these concepts. Geyer (1994) provides an excellent discussion of *Clark regularity*; for more detailed discussions see Rockafellar and Wets (1998).

If Ω satisfies the Mangasarian and Fromowitz constraint qualification at θ_0, then Ω is nearly convex at θ_0. This is particularly relevant because many of the constrained inference problems in statistics are likely to be specified by equality and inequality constraints on functions of θ, and the Mangasarian and Fromowitz condition is likely to be satisfied in many cases. Further, if the constraining functions $h_i(\theta)$ are Lipchitz continuous in a neighborhood of θ_0 then Ω is pro-regular at θ_0.

Proposition 4.9.2 *(Shapiro (2000a)). Suppose that Ω is nearly convex at θ_0 and Condition Q is satisfied. Let $\tilde{\theta}^a$ and $\tilde{\theta}^b$ be two local mle's. Then,*

$$\sqrt{n}(\tilde{\theta}^a - \tilde{\theta}^b) = o_p(1). \quad \blacksquare$$

Shapiro (2000a) showed, by constructing a counter example, that even if the tangent cone of Ω at θ_0 is convex, two local *mle*'s may fail to be asymptotically equivalent [i.e., $\sqrt{n}(\tilde{\theta}^a - \tilde{\theta}^b)$ may not be $o_p(1)$]. In fact, Shapiro (2000a) showed that even if Ω is *Clark regular*, $\sqrt{n}(\tilde{\theta}^a - \tilde{\theta}^b)$ may not be $o_p(1)$. Clark regularity is weaker than near convexity, but it is a considerably stronger than Chernoff regularity; for example, it ensures that the tangent cone is convex.

Suppose that Ω is pro-regular at θ_0 and $\tilde{\theta}^a$ and $\tilde{\theta}^b$ are strongly consistent. Then, under some regularity conditions on the smoothness of and uniformity of convergence of $n^{-1}\ell(\theta)$ we have $\tilde{\theta}^a = \tilde{\theta}^b$ with probability approaching one (see Shapiro (2000a) for details).

Throughout this section we restricted our discussions to the simple *iid* setting. However, these results hold under much more general conditions. The developments in Geyer (1994) and Shapiro (2000a) include M-estimation; Andrews (1999) studied properties of the global *mle* under conditions that are more general than the usual

iid setting. For example, his framework includes many time series models used in econometric modeling. He considered several important econometric models and illustrated the relevance of the theoretical results. In these papers quadratic approximations are used even if the objective function is not differentiable; see also Shapiro (1989).

4.10 APPENDIX: PROOFS

Proof of Proposition 4.3.2: Assume that the null hypothesis is true. The test statistic T_λ is a function of the sample. Let $T_\lambda(\mathcal{X})$ denote the value of the test statistic T_λ for the sample $\mathcal{X}$. Let $\mathcal{Y}$ be *iid* as $\mathcal{X}$. Then

$$p^*(\mathcal{X}) = \alpha_1 + \sup_{\lambda \in \mathcal{A}(\mathcal{X})} \text{pr}_{\mathcal{Y}|\mathcal{X}}\{T_\lambda(\mathcal{Y}) \geqslant T_\lambda(\mathcal{X}) \mid \theta = (\mathbf{0} : \lambda)\}.$$

Now,

$$\begin{aligned} \text{pr}_{\mathcal{X}}\{p^*(\mathcal{X}) \leq \alpha\} &= \text{pr}_{\mathcal{X}}\{p^*(\mathcal{X}) \leq \alpha \text{ and } \lambda_0 \in \mathcal{A}(\mathcal{X})\} + \\ &\quad \text{pr}\{p^*(\mathcal{X}) \leq \alpha \text{ and } \lambda_0 \notin \mathcal{A}(\mathcal{X})\} \\ &\leq \text{pr}\{p^*(\mathcal{X}) \leq \alpha \text{ and } \lambda_0 \in \mathcal{A}(\mathcal{X})\} + \text{pr}(\lambda_0 \notin \mathcal{A}(\mathcal{X})\}. \end{aligned} \tag{4.119}$$

Let F denote the cumulative distribution function of $T_{\lambda_0}(\mathcal{X})$. Then $F\{T_{\lambda_0}(\mathcal{X})\}$ has the uniform distribution on $(0, 1)$, and hence $\text{pr}_{\mathcal{Y}|\mathcal{X}}\{T_{\lambda_0}(\mathcal{Y}) \geqslant T_{\lambda_0}(\mathcal{X})\}$, which is equal to $[1 - F\{T_{\lambda_0}(\mathcal{X})\}]$, has the uniform distribution on $(0, 1)$. Now,

$$\begin{aligned} &\text{pr}_{\mathcal{X}}\{p^*(\mathcal{X}) \leq \alpha \text{ and } \lambda_0 \in \mathcal{A}(\mathcal{X})\} \\ &\quad = \text{pr}_{\mathcal{X}}[\alpha_1 + \\ &\qquad \sup_{\lambda \in \mathcal{A}(\mathcal{X})} \text{pr}_{\mathcal{Y}|\mathcal{X}}\{T_\lambda(\mathcal{Y}) \geqslant T_\lambda(\mathcal{X}) | \theta = (\mathbf{0} : \lambda)\} \leq \alpha \text{ and } \lambda_0 \in \mathcal{A}(\mathcal{X})] \\ &\quad \leq \text{pr}_{\mathcal{X}}[\alpha_1 + \text{pr}_{\mathcal{Y}|\mathcal{X}}\{T_{\lambda_0}(\mathcal{Y}) \geqslant T_{\lambda_0}(\mathcal{X}) | \theta = (\mathbf{0} : \lambda_0)\} \leq \alpha] \\ &\quad = \alpha - \alpha_1. \end{aligned}$$

Now pr(Type I error) $= \text{pr}\{p^*(\mathcal{X}) \leq \alpha\} \leq \alpha$, by (4.119). $\blacksquare$

In the proof of Proposition 4.8.1 and that of other similar asymptotic results, it is often sufficient to restrict attention to neighborhoods of the true value that shrink at the rate of $n^{-1/2}$, more precisely, neighborhoods of the form $\{\theta : \sqrt{n}\|\theta - \theta_0\| \leq K\}$ for large K. The next lemma provides a justification.

Lemma 4.10.1 *(1) Let $\Theta \subset \mathbb{R}^p$, $\theta_0 \in \Theta$, $n^{1/2}(T_n - \theta_0) = O_p(1)$ and $\Theta_{n,K} = \Theta \cap \{\theta : n^{1/2}\|\theta - \theta_0\| < K\}$. Then $\|T_n - \Theta\| = \|T_n - \Theta_{n,K}\|$ with arbitrarily large probability for sufficiently large K and n. More precisely, given ϵ there exist $n_0(\epsilon)$ and $K_0(\epsilon)$ such that pr$\{\Pi(T_n, \Theta) = \Pi(T_n, \Theta_{n,K})\} > 1 - \epsilon$ for $n > n_0$ and $K > K_0$.*

(2) Suppose that $\tilde{\theta}$ is the maximizer of the objective function $R_n(\theta)$ over Θ. Suppose also that $\sqrt{n}(\tilde{\theta} - \theta_0) = O_p(1)$. Then, given $\epsilon > 0$ there exists $n_0 > 0$ and $K_0 > 0$ such that

$$\sup_{\theta \in \Theta} R_n(\theta) = \sup_{\theta \in \Theta_{n,K}} R_n(\theta)$$

with arbitrarily large probability for $n > n_0$ and $K > K_0$.

Proof: Given $\epsilon > 0$ there exist $n_0(\epsilon)$ and $K_0(\epsilon)$ such that

$$\text{pr}(n^{1/2}\|\boldsymbol{T}_n - \boldsymbol{\theta}_0\| < K/2) > 1 - \epsilon, \text{ for } n > n_0 \text{ and } K > K_0.$$

Since $\boldsymbol{\theta}_0 \in \Theta$, we have that

$$\|\boldsymbol{T}_n - \boldsymbol{\theta}_0\| \geqslant \|\boldsymbol{T}_n - \Pi(\boldsymbol{T}_n, \Theta)\|.$$

Therefore,

$$\|\Pi(\boldsymbol{T}_n, \Theta) - \boldsymbol{\theta}_0\| \leq \|\Pi(\boldsymbol{T}_n, \Theta) - \boldsymbol{T}_n\| + \|\boldsymbol{T}_n - \boldsymbol{\theta}_0\| \leq 2\|\boldsymbol{T}_n - \boldsymbol{\theta}_0\|.$$

Now, if $n^{1/2}\|\boldsymbol{T}_n - \boldsymbol{\theta}_0\| < K/2$ then $n^{1/2}\|\Pi(\boldsymbol{T}_n, \Theta) - \boldsymbol{\theta}_0\| < K$ and hence $\Pi(\boldsymbol{T}_n, \Theta) = \Pi(\boldsymbol{T}_n, \Theta_{n,K})$. Therefore,

$$\text{pr}\{\Pi(\boldsymbol{T}_n, \Theta) = \Pi(\boldsymbol{T}_n, \Theta_{n,K})\} \geqslant \text{pr}(n^{1/2}\|\boldsymbol{T}_n - \boldsymbol{\theta}_0\| < K/2) > 1 - \epsilon$$

for $n > n_0$ and $K > K_0$.

The proof of the second part is similar. ■

Proof of Proposition 4.8.1: First let us first establish a lemma that would be of independent interest.

Lemma 4.10.2 *Suppose that $\boldsymbol{Y}_n = O_p(1), A \subset \mathbb{R}^p$, $\boldsymbol{V}$ is a positive definite matrix of order $p \times p$ and $\hat{\boldsymbol{V}}$ is a consistent estimator of $\boldsymbol{V}$. Then*

1. $\Pi_{\boldsymbol{V}}(\boldsymbol{Y}_n, A) = O_p(1)$.
2. $\|\boldsymbol{Y}_n - A\|_{\hat{\boldsymbol{V}}} = \|\boldsymbol{Y}_n - A\|_{\boldsymbol{V}} + o_p(1)$.
3. $\|\Pi_{\hat{\boldsymbol{V}}}(\boldsymbol{Y}_n, A)\|_{\hat{\boldsymbol{V}}} = O_p(1)$.

Proof:
(1) The first part follows by triangular inequality. Let P be a fixed point in A, O be the origin, $\boldsymbol{Y}_n = OQ$, and R be the point in A that is $\boldsymbol{V}$-closest to Q. Then

$$\|\boldsymbol{Y}_n - A\|_{\boldsymbol{V}} = QR \leq QP \leq QO + OP = O_p(1)$$

$$\|\Pi(\boldsymbol{Y}_n, A)\|_{\boldsymbol{V}} = OR \leq OQ + QR = O_p(1).$$

(2) For a given matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\|$ denote the matrix norm that is equal to the largest

eigenvalue of $\boldsymbol{A}\boldsymbol{A}^T$; then an upper bound for $\|\boldsymbol{A}\|$ is $\sum\sum|a_{ij}|$. Here we shall use the inequality

$$\|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}} \leq \|\boldsymbol{A}\|\|\boldsymbol{x}\|_{\boldsymbol{W}} \leq \sum\sum|a_{ij}|\|\boldsymbol{x}\|_{\boldsymbol{W}}$$

for any $\boldsymbol{x}$ and any positive definite matrix $\boldsymbol{W}$ (for example, see Pryce (1973, p 101)). From this and the Cauchy-Schwartz inequality, we have $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \leq \sum\sum|a_{ij}|\|\boldsymbol{x}\|^2$. Since $\hat{\boldsymbol{V}}$ is consistent, it follows that $\|\boldsymbol{V} - \hat{\boldsymbol{V}}\| = o_p(1)$. Let

$$f_n(\boldsymbol{\theta}) = (\boldsymbol{Y}_n - \boldsymbol{\theta})^T\hat{\boldsymbol{V}}^{-1}(\boldsymbol{Y}_n - \boldsymbol{\theta}) \quad \text{and} \quad g_n(\boldsymbol{\theta}) = (\boldsymbol{Y}_n - \boldsymbol{\theta})^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_n - \boldsymbol{\theta}).$$

Then

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in A}|f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})| &\leq \sup_{\boldsymbol{\theta} \in A}(\|\boldsymbol{Y}_n - \boldsymbol{\theta}\|^2)(\|\hat{\boldsymbol{V}}^{-1} - \boldsymbol{V}^{-1}\|) \\ &\leq (\|\boldsymbol{Y}_n - A\|_I^2)(\|\hat{\boldsymbol{V}}^{-1} - \boldsymbol{V}^{-1}\|) = O_p(1)o_p(1) = o_p(1)\end{aligned}$$

Note that, since f_n and g_n are nonnegative, we have

$$\inf f_n(\boldsymbol{\theta}) \leq |f_n(\boldsymbol{\theta})| \leq |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta}) + g_n(\boldsymbol{\theta})| \leq |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})| + g_n(\boldsymbol{\theta}).$$

Therefore, $\inf f_n(\boldsymbol{\theta}) \leq \sup|f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})| + g_n(\boldsymbol{\theta})$, and hence

$$\inf f_n(\boldsymbol{\theta}) \leq \sup|f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})| + \inf g_n(\boldsymbol{\theta}).$$

Thus, $\inf f_n(\boldsymbol{\theta}) - \inf g_n(\boldsymbol{\theta}) \leq \sup|f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})|$. By symmetry

$$\inf g_n(\boldsymbol{\theta}) - \inf f_n(\boldsymbol{\theta}) \leq \sup|f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})|.$$

Therefore, we have

$$|\inf_{\boldsymbol{\theta} \in A} f_n(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta} \in A} g_n(\boldsymbol{\theta})| \leq \sup_{\boldsymbol{\theta} \in A}|f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})| = o_p(1).$$

This completes the proof of part (2). Part (3) follows from the first two. ■

Proof of Proposition 4.7.3: The proof of this uses several results concerning Constraint Qualification (CQ) in nonlinear programming and optimization. The main steps are indicated here without attempting to define terms and concepts in nonlinear optimization that are used in this proof. We shall use the same notation as in Bazaraa et al. (1993, section 5.3); for convenience, the relevant notations therein are stated below:

$$\begin{aligned}T &= \{\boldsymbol{d} : \boldsymbol{d} = \lim_{k \to \infty} \lambda_k(\boldsymbol{\theta}_k - \boldsymbol{\theta}_0), \boldsymbol{\theta}_k \in \Theta_1, \boldsymbol{\theta}_k \to \boldsymbol{\theta}_0\}; \\ G' &= \{\boldsymbol{d} : \boldsymbol{d}^T\nabla \boldsymbol{h}_i(\boldsymbol{\theta}_0) \geqslant 0 \quad \text{for } i \in J(\boldsymbol{\theta}_0)\}, \\ H_0 &= \{\boldsymbol{d} : \boldsymbol{d}^T\nabla \boldsymbol{h}_i(\boldsymbol{\theta}_0) = 0 \quad \text{for } i = 1, \ldots, \ell\}, \\ A &= \{\boldsymbol{d} : \exists \delta > 0, \exists \boldsymbol{f} : \mathbb{R} \to \mathbb{R}^p \text{ such that } \boldsymbol{f}(\lambda) \in \Theta_1 \text{ for } \lambda \in (0, \delta), \\ &\qquad \boldsymbol{f}(\boldsymbol{0}) = \boldsymbol{\theta}_0, \text{ and } \lim_{\lambda \to 0} \lambda^{-1}[\boldsymbol{f}(\lambda) - \boldsymbol{f}(\boldsymbol{0})] = \boldsymbol{d}.\}\end{aligned}$$

Thus, T is the cone of tangents and A is the derived tangent cone. For these cones the following holds (Bazarra et al. (1993, p 193):

1. Closure of $A \subset T \subset G' \cap H_0$.
2. Kuhn-Tucker CQ : Closure of $A = G' \cap H_0$.
3. Cottle CQ $\Rightarrow$ Kuhn-Tucker (KT) CQ
4. Cottle CQ is equivalent to Mangasarian-Fromowitz (MF) CQ provided Θ is open.

Since Θ is open and the MF-CQ is satisfied, it follows that Cottle-CQ is satisfied, and hence KT-CQ is satisfied. Therefore, it follows that Closure of $A = T = G' \cap H_0$, and hence Θ_1 is Chernoff regular and the tangent cone is $G' \cap H_0$. ■

Proof of Proposition 4.7.4 using Hausdorff distance definition

Let $\epsilon > 0$ be given and $\|y - \theta_0\| < \epsilon/2$. We shall refer to Figure 4.9 for this proof.

Let $y = OA$ and $\Pi(A, \Theta)$ denote the point in Θ that is V-closest (i.e., closest with respect to $\|.\|_V$) to A. Let $C = \Pi(A, \mathcal{A}), D = \Pi(A, \Theta)$, $G = \Pi(C, \Theta)$ and $F = \Pi(D, \mathcal{A})$. First, note that since $C \in \mathcal{A}$ and $D \in \Theta$,

$$CD \leq h\{\Theta \cap B_\epsilon, \mathcal{A} \cap B_\epsilon\} = o(\epsilon).$$

By triangle inequality, $AC \leq AD + CD$; squaring both sides and subtracting AD^2, we have

$$AC^2 - AD^2 \leq CD^2 + 2AD.CD \leq o(\epsilon^2) + 2.OA.CD = o(\epsilon^2) + \epsilon o(\epsilon) = o(\epsilon^2).$$

Similarly, by considering $AD \leq AC + CD$ we have $AD^2 - AC^2 \leq o(\epsilon^2)$. Therefore, $AD^2 - AC^2 = o(\epsilon^2)$. ■

Proof of Proposition 4.7.4 without using Hausdorff distance

Let $y = OA$ and $\Pi(A, \Theta)$ denote the point in Θ that is V-closest (i.e., closest with respect to $\|.\|_V$) to A. Let $C = \Pi(A, \mathcal{A}), D = \Pi(A, \Theta)$, $G = \Pi(C, \Theta)$ and $F = \Pi(D, \mathcal{A})$.

Case 1: $AC > AD$; see the diagram on the right in Fig.4.9. Since $AD \leq OA$, we have that $OD/OA \leq (OA + AD)/OA \leq (OA + OA)/OA = 2$. Now by making use of the definition of "closest point", we have that $0 \leq (AC - AD)/OA \leq (AF - AD)/OA \leq (AD + DF - AD)/OA = DF/OA = (DF/OD)(OD/OA) \leq 2(DF/OD)$.

Case 2: $AC < AD$; see the diagram on the left in Fig.4.9. Since OAC is a right-angle triangle, we have $OC \leq OA$. Now, $0 \leq (AD - AC)/OA \leq (AG - AC)/OA \leq (AC + CG - AC)/OA \leq CG/OA \leq CG/OC$.

From cases (1) and (2), we have that

$$|(AC - AD)|/OA \leq \max\{2(DF/OD), CG/OC\}.$$

Since $AC \leq OA$ and $AD \leq OA$, it follows that $(AC + AD)/OA \leq 2$. Now, $|(AC^2 - AD^2)|/OA^2 = \{|(AC - AD)|/OA\}\{(AC + AD)/OA\} \leq 2|(AC - AD)|/OA \leq 2\max\{2DF/OD, CG/OC\}$.

Now, as $OA \to 0$, we have $OD \to 0$ and $OC \to 0$ because $OD \leq OA + AD \leq 2OA$. As $OD \to 0$ we have $DF/OD \to 0$ by condition (a) of the definition of approximating cone; as $OC \to 0$ we have $CG/OC \to 0$ by condition (b) of the definition of approximating cone. Therefore, as $OA \to 0$, we have that $\max\{2DF/OD, CG/OC\} \to 0$, $(|AC^2 - AD^2)|/OA^2 \to 0$ and this establishes the claim of the proposition. ■