CHAPTER 4

Testing for Latent Structure

We consider two closely related hypothesis testing problems. In the test c *homogeneity* (against heterogeneity), we test

$$H\text{: one component} \quad \text{versus} \quad A_1\text{: any mixture.}$$

In testing for the *number of components*, the usual test compares

$$H\text{: one component} \quad \text{versus} \quad A_2\text{: two components.}$$

These two problems are clearly closely related, and it can be anticipated tha tests valid for one might also be applied to the other. As such, we consider th two problems together.

In a survey of the tests of homogeneity, DerSimonian (1989) found that mos of the available procedures were one of two types. The first, the $C(\alpha)$ test fo homogeneity, is simple and tractable, with a known limiting distribution an a local optimality property. Strictly speaking, it is designed for the alterna tive A_1.

The second, the likelihood ratio test for H versus A_2, has long been a enigma. In addition to the aforementioned problems with multiple likelihoo roots when the number of components is fixed, we have the problem that th standard likelihood ratio statistic has an unknown limiting distribution. In terest has persisted in this method, no doubt partially due to the mystery, bu also related to the phenomenal record of reliability for this testing procedur when used in other standard problems.

The goals of this chapter are to provide essential background on these pro cedures, plus provide some fundamentally new insights to both. We will star this chapter with the $C(\alpha)$ testing procedure. It is a simple and highly ef fective method of testing for overdispersion, and is the method the autho would recommend for most practical situations. A new result is given for it optimality, which had been clouded by work of Moran (1973).

The second part of this chapter is devoted to an extensive treatment of th asymptotic distribution of the likelihood ratio test for one versus two com ponents. A complete description is given for a class of multinomial mixtur

models, together with a simple approximation formula. As by-products of this analysis, we derive the appropriate score test for the two-component problem and give a new result for the one component versus nonparametric mixture model as well. These results require a careful, and somewhat difficult, geometric analysis related to the geometry given in Chapter 2.

4.1. Dispersion score tests. The most popular method for testing for overdispersion is Neyman's $C(\alpha)$ test for homogeneity, although it is not always identified by this name because the same test can be derived in a number of different ways. We will start by motivating the test statistic from the point of view of the gradient function used in nonparametric maximum likelihood.

4.1.1. *The dispersion score.* Recall from the mixture NPMLE theorem of Chapter 1 that the degenerate distribution $\Delta_{\hat{\phi}}$ is the NPMLE for Q if and only if

$$D_1(\phi) \leq 0 \quad \text{for all } \phi. \tag{4.1}$$

Here $D_1(\phi)$ is unicomponent gradient function $D_{\Delta_{\hat{\phi}}}(\phi)$. We note also that we have

$$D_1(\hat{\phi}) = 0,$$

$$D_1'(\hat{\phi}) = 0.$$

The first equation is by direct calculation and the second, where the prime is the derivative with respect to ϕ, holds because of its equivalence to the likelihood equation which defines $\hat{\phi}$. [*Exercise.*] Thus by Taylor expansion about $\hat{\phi}$ we have, locally,

$$D_1(\phi) \approx \tfrac{1}{2}(\phi - \hat{\phi})^2 D_1''(\hat{\phi}). \tag{4.2}$$

It follows that the unicomponent gradient inequality (4.1) is violated in a neighborhood of $\hat{\phi}$ if $D_1''(\hat{\phi}) > 0$, but if < 0, it must hold locally. That is,

$$D_1''(\hat{\phi}) > 0 \implies \Delta_{\hat{\phi}} \text{ is not NPMLE.} \tag{4.3}$$

Thus $D_1''(\hat{\phi})$ seems to be an important summary statistic for checking a unicomponent hypothesis against local violations of that hypothesis. We therefore investigate its form. We define the (*Neyman*) *dispersion score* function to be

$$v_2(\phi, x_i) := \frac{f''(x_i;\phi)}{f(x_i;\phi)} = v(\phi; x_i)^2 + v'(\phi; x_i).$$

By direct calculation,

$$D_1''(\hat{\phi}) = \sum_i v_2(\hat{\phi}; x_i).$$

If we further examine the form of v_2 in the exponential family, using the natural parameterization, we find that

$$v_2(\phi, x) = (x - \mu)^2 - \sigma^2,$$

where μ and σ^2 are the mean and variance of the statistic X. Since under the null hypothesis, $\hat{\mu}_{\text{mle}} = \bar{x}$, we can conclude that there are no local gradient violations of the unicomponent model if and only if the sample variance is smaller than the variance predicted under the one-component model:

$$D_1''(\hat{\phi}) = \sum_i (x_i - \bar{x})^2 - n\sigma^2(\hat{\phi}).$$

Moreover, $D_1''(\hat{\phi})/n$ can be seen to be a consistent estimator of the variance of the mean value parameter μ under the latent distribution Q. [*Exercise.*]

However, the relationship of this statistic to the gradient function also indicates it may have low power to detect the need for a second component ϕ that is *not near* to the unicomponent estimate $\hat{\phi}$, because the gradient at a distant ϕ is likely to be less well predicted by the second order Taylor series approximation (4.2). Indeed, as we will see, the likelihood ratio test does use the gradient information for ϕ away from $\hat{\phi}$.

4.1.2. *Neyman and Scott's $C(\alpha)$ test.* For a second development of this statistic, we turn to Neyman and Scott's original derivation (1966). Let G be a distribution for Θ that has mean 0 and variance 1 (we assume the parameterization has been chosen so as to make this possible). We construct a location-scale family of distributions for Φ, with parameters (a, b) by setting

$$\Phi =_{\text{dist}} a + b\Theta,$$

where Θ has distribution G. [The parameters (a, b) may need to be restricted to ensure that the distribution has all its mass in the parameter space for ϕ.] We note that as the parameter b goes to zero, the distribution converges to a degenerate distribution at parameter a. Thus testing

$$H\colon\ b = 0 \quad \text{versus} \quad A\colon\ b > 0$$

is a test of overdispersion.

We consider the construction of the locally most powerful score test for this problem, so as to obtain maximum local power. We treat a as fixed for the moment. We calculate the first derivative of the log likelihood with respect to b:

$$\frac{\partial}{\partial b} \ln \int f(x; a + bs)\, dG(s) := S_1(a, b) = \frac{\int s f'(x; a + bs)\, dG(s)}{\int f(x; a + bs)\, dG(s)}.$$

Since G has mean zero, as $b \to 0$ this score converges to

$$\left[\int s\, dG(s)\right] \cdot \frac{f'(x; a)}{f(x; a)} = 0$$

under appropriate regularity conditions.

Since the first derivative of the log likelihood becomes degenerate as we approach the null, the approximate form of the *Neyman–Pearson tests* in a

neighborhood of $b = 0$ must be determined by the second derivative of the log likelihood. Differentiating once more, we obtain

$$\frac{\partial^2}{\partial b^2} \ln \int f(x; a + bs)\, dG(s) = \frac{\int s^2 f''(x; a + bs)\, dG(s)}{\int f(x; a + bs)\, dG(s)} - S_1^2(a, b).$$

Since $\int s^2 dG(s) = 1$ by assumption, as $b \to 0$ this score converges to the dispersion score

$$v_2(a; x) = \frac{f''(x; a)}{f(x; a)}.$$

For a *fixed* value of a, the locally most powerful test of our hypotheses is therefore to reject for large positive values of $\sum v_2(a; x_i)$.

[*Note:* Earlier workers parameterized the family of latent distributions by (a, σ^2), with $\sigma^2 = b^2$. If one does this, one finds that the limit of the *first* derivative with respect to σ^2 is equal to (through the use of l'Hôpital's rule) the above limiting *second* derivative with respect to b. We have taken the scale parameter approach here because of an important point to be touched upon later.]

To complete construction of the $C(\alpha)$ procedure, as described by Neyman (1959), we must compute a $\sqrt{n}$ consistent estimator of a under the null hypothesis. Since the null hypothesis is the unicomponent model, we may use the one-component MLE $\hat{\phi}$. In this case, the $C(\alpha)$ test statistic will be the statistic we earlier derived from the gradient,

$$V_2 := \sum v_2(\hat{\phi}; x_i) = D_1''(\hat{\phi}),$$

normalized so as to have asymptotic variance 1. Note that except for the estimated parameter, V_2 is an i.i.d. sum. The asymptotic variance of V_2 is $n \cdot \tau(\phi)$, where

$$\tau(\phi) = E[v_2(X)]^2 - \frac{E^2[v_2(X)v_1(X)]}{E[v_1(X)]^2}.$$

Note that the second term in this variance formula is necessary to allow for the estimation of ϕ in V_2. Thus $V_2/\sqrt{n\tau(\hat{\phi})}$ is an asymptotically standard normal test statistic and has certain local optimality properties described by Neyman.

We note that the $C(\alpha)$ test can be constructed using a different point estimator than the MLE $\hat{\phi}$, but we must replace v_2 with a *corrected* dispersion score $\tilde{v}_2 := v_2 - \rho v_1$, where

$$\rho = \frac{E[v_2(X)v_1(X)]}{E[v_1(X)]^2}.$$

The adjustment is necessary to correct for the estimation of the nuisance parameter under the null hypothesis, but if one uses the MLE of the nuisance parameter, the correction term is zero. The variance term $\tau(\phi)$ is in fact equal

to $\mathrm{Var}(\tilde{v}_2(X))$. These corrections of scores for the effects of nuisance parameters will show up again in the derivation of properties of the likelihood ratio test.

For its practical value, we indicate how the variance of the test statistic can be calculated when there are other auxiliary parameters in the problem, say $\theta_1, \ldots, \theta_m$. If we let $u_1, \ldots, u_m$ be the score functions for the auxiliary parameters, then one constructs the $m+2$ by $m+2$ covariance matrix i^* of the extended set of scores $u_1, \ldots, u_m, v_1, v_2$. Other than the last row and column, this is the Fisher information matrix i for the unicomponent problem. We write it in the partitioned form

$$i^* = \begin{bmatrix} i & a \\ a' & b \end{bmatrix}.$$

The formula for the asymptotic variance of $v_2(\hat{\theta}_1, \ldots, \hat{\theta}_m, \hat{\phi}; x)$ is then $b - a' i^{-1} a$.

4.1.3. *Dispersion test optimality.* A remarkable and important feature of the $C(\alpha)$ test for heterogeneity is that the test statistic does not depend on the alternative distribution G that was used and so it suggests that the statistic has power over a wide range of nearby alternatives. (If we think about the statistic in terms of the gradient, this becomes quite plausible.) Indeed, the statistic is often derived through other means than presented here, such as constructing the locally most powerful test for degeneracy in a conjugate family of latent densities.

However, a curiosity exists in the literature. Moran (1973) found that he could prove that this test, in the case of the test for Poisson overdispersion, was asymptotically best in the sense of best power under local alternatives, *only* if he assumed in addition that

$$m_3(G) := \int s^3 \, dG(s) = 0.$$

It appears that this condition is not strictly necessary, as we argue now briefly.

If we carry out a Taylor series expansion of the sample log likelihood about $b = 0$, we obtain the formula

$$\begin{aligned} \sum \ln\left[\frac{f(x_i; a, b)}{f(x_i; a, 0)}\right] &= b \cdot 0 + \frac{b^2}{2!} \sum v_2(a; x_i) + \frac{b^3}{3!} m_3(G) \sum v_3(a; x_i) \\ &\quad + \frac{b^4}{4!} \left[m_4(G) \sum v_4(a; x_i) - 3 \sum v_2(a; x_i)^2\right] \\ &\quad + \text{remainder}. \end{aligned} \tag{4.4}$$

Here we have used the higher order versions of the dispersion score:

$$v_k(a; x) := \frac{f^{(k)}(x; a)}{f(x; a)}.$$

Noting that the functions $v_k(a; X)$ all have mean *zero* under the null hypothesis, we find that the log likelihood ratios *converge to normality* if we set

$b = b_n = cn^{-1/4}$. With such a scaling the quadratic term is asymptotically normal, the cubic term converges to *zero, for any finite* $m_3(G)$, and for finite $m_4(G)$ the quartic term converges to the constant,

$$\frac{3c^4 E[v_2(a;X)^2]}{4!}.$$

From this point it appears that by taking local alternatives that approach the null at the unusual rate of $n^{-1/4}$ in the parameter b, one can derive the necessary asymptotic optimality, as well as local power.

The key here seems to be that instead of following the above approach, Moran considered the $n^{-1/2}$ convergence of the parameter b^2. We note that if we use the scale parameter b, then the parametric family of likelihoods generated by $\Phi = a + b\Theta$ is in fact well defined for b both positive and negative, and generates a smooth two-sided alternative to the null hypothesis. If Θ has a symmetric distribution G, then $-b$ generates the same distribution as b, but otherwise not. If not, however, this means that the family of likelihoods is not symmetric as a function of b, and so is not a function of b^2 and so cannot have a Taylor expansion in the variable b^2.

Thus, although the $\sqrt{n}$ convergence of b^2 might make it seem to be the appropriate parameterization, switching to the more slowly converging scale parameter b makes the expansions work properly, because then they are valid for b both positive and negative.

The fact that one has asymptotic power, in the b scale, only at distances of order $n^{-1/4}$, has importance in the general nonparametric theory as well. Chen (1993) used a derivation like that above, with G being a two point distribution, to show that the optimal rate of convergence of a consistent G estimator to an unknown discrete latent distribution is $n^{-1/4}$ if the number of components is not known or has been overspecified.

4.1.4. *Auxiliary parameters.* As a final note, we make the observation that further problems may arise in certain mixture models with auxiliary parameters. To give a simple example, if we consider the normal mixture model $N(G, \sigma^2)$, it is easily seen that the dispersion score v_2 is *equivalent* to the nuisance parameter score for σ^2. Since the nuisance parameter scores must be regressed out of the locally most powerful test statistic, it follows that the locally most powerful tests for heterogeneity from (4.4) are no longer independent of the form of G. If $m_3(G)$ is *not zero*, then the locally most powerful test is based on the corrected version of the "skewness" score v_3. If $m_3(G)$ is *zero*, then the best test depends on the "kurtosis" score v_4. In either case, the rate of convergence of detectable alternatives to the null must be slower yet than $n^{-1/4}$.

Intuitively, this arises because of an identifiability issue. If the latent distribution G is $N(v, \tau^2)$, then X has a distribution that can be represented as either a mixture $N(G, \sigma^2)$ or a different normal $N(v, \sigma^2 + \tau^2)$. Thus any hope of detecting a mixture alternative will depend on the degree to which the G differs from the normal, as evidenced through its cumulants.

4.2. LRT for number of components. In this section we will give some background on the likelihood ratio test (LRT) for the number of components in a mixture model. The last section of this chapter will give some new results regarding the limiting distributions involved, both for the test of one versus two components, and the test of one component versus the nonparametric model. However, it is important to point out that from a practical point of view, the $C(\alpha)$ test is much simpler to calculate and has a much nicer distribution theory.

4.2.1. *The testing problem.* The nature of the limiting distribution for this likelihood ratio test is a long-standing mystery. If we consider just the simple model with no auxiliary parameters and where ϕ is unidimensional, then the usual distribution theory suggests that the likelihood ratio test has a chi-squared distribution with 2 degrees of freedom, corresponding to one free parameter under the null hypothesis H of one component, with density $f(x;\phi)$, and three parameters under the alternative A_2, with density

$$\pi f(x;\phi_1) + (1-\pi) f(x;\phi_2).$$

However, it is known that the usual regularity conditions are not satisfied.

To consider the difficulties, we consider the parameter space for the two-component model, where we can restrict attention to

$$\pi \in [0,1] \quad \text{and} \quad \phi_1 \leq \phi_2.$$

In this setting we can describe a single element of the null hypothesis, say unicomponent with parameter ϕ_0, with many elements of the alternative parameter space. We find three lines on the boundary of the parameter space all give the same null distribution:

- The line for $\phi_1 = \phi_2 = \phi_0$, $\pi =$ anything.
- The line where $\pi = 0$, $\phi_2 = \phi_0$ and $\phi_1 =$ anything.
- The line where $\pi = 1$, $\phi_1 = \phi_0$ and $\phi_2 =$ anything.

In Figure 4.1 we plot the alternative parameter values that correspond to a single null distribution, unicomponent with parameter ϕ_0.

The union of all such lines, over all values of ϕ_0, makes up the null hypothesis and so it corresponds to three entire boundary surfaces of the alternative parameter space. It is clear from this description that there are many points in the alternative that seem to be close to any one distribution in the null hypothesis.

The first fundamental difficulty in establishing the asymptotic structure of the likelihood ratio test lies in determining what happens to the score functions in the alternative hypothesis as we approach the null. Because one can approach the null from so many directions, we will find that the set of limit functions at the null hypothesis is an *infinite*-dimensional score space, even though there are only three scores in the alternative hypothesis, corresponding to three parameters. In turn, this structure implies that the problem generates what we will call a type III likelihood ratio test, a rather awkward beast

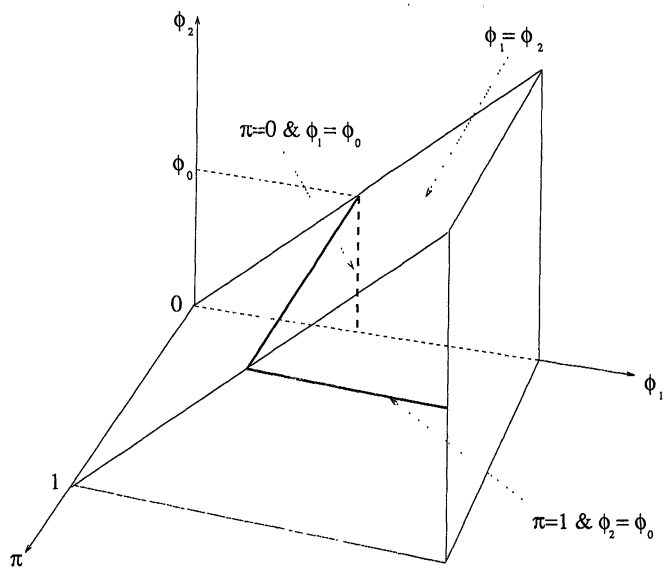

FIG. 4.1. *The parameter space for the two-component mixture.*

at best. (Type I is the standard testing problem with limiting chi-squared distributions; type II generates the chi-bar-squared distributions.)

4.2.2. *Historical perspective.* For background on the problem of likelihood ratio testing in mixture models, the books by Titterington, Smith and Makov (1985) and McLachlan and Basford (1988) provide extensive discussion. Over the years considerable simulation work has been done; see Böhning, Dietz, Schaub, Schlattman and Lindsay (1994) for a variety of exponential family simulations. In all, the limiting distributions are not completely apparent from such studies, but the normal theory distribution, in the unknown variance case, has appeared tantalizingly like the chi-squared with 2 degrees of freedom, and the Böhning study shows that one parameter exponential families appear to have tails very much like mixtures of chi-squares with differing degrees of freedom; these results are in accordance with the distributions to be derived in the next section.

Exact theoretical results have been obtained in a number of special cases. For example, if we have two known components, with unknown π only, and the null hypothesis is $\pi = 0$, then it will be shown in the next section that the limiting distribution is

$$0.5\chi_0^2 + 0.5\chi_1^2.$$

Here the symbol χ_0^2, the "chi-squared with 0 degrees of freedom," is a degenerate distribution with all its mass at 0. Such a mixture of chi-squared distributions of different degrees of freedom is called a *chi-bar-squared* distribution. Another interesting set of special cases was considered by Goffinet,

Loisel and Laurent (1992), who were interested in normal model hypotheses when the weight π was known a priori. The limiting distributions were again of the chi-bar-squared type.

The difficulties with the irregular parameter space in the general problem led Aitkin and Rubin (1985) to attempt to make the problem regular by using a prior distribution on π. However, Quinn, McLachlan and Hjort (1987) showed that even with this approach the usual regularity conditions were violated.

Ghosh and Sen (1985) developed a theory for models that satisfy a rather severe identifiability constraint. The most relevant and important insight into the general problem can be found in Hartigan (1985), who established that in the normal model, if one of the ϕ's is known and if σ^2 is known, then the likelihood ratio test statistic has no interesting limiting distribution, but rather goes to *infinity* with probability 1.

4.2.3. *Initial observations.* For the moment we assume the model has no auxiliary parameters. We first make some observations about the probability that the likelihood ratio statistic equals zero, corresponding to having a χ_0^2 component in the limiting distribution. We let $\hat{\phi}$ be the maximum likelihood estimator under the null hypothesis and make the following claim:

$$\text{LRT} = 0 \quad \Leftrightarrow \quad D_1(\phi) \leq 0 \qquad \forall \phi. \tag{4.5}$$

Check this, using the fact that the gradient function D_1 satisfies the right-hand inequality if and only if $\Delta_{\hat{\phi}}$ is the nonparametric maximum likelihood estimator of Q. Further, it follows from the asymptotic normality of the Neyman–Scott dispersion test statistic and relationship of the dispersion test to the gradient inequality as in (4.3) that the gradient inequality is violated in a neighborhood of $\hat{\phi}$ with an asymptotic null probability of at least 0.5. It follows that there is asymptotically *at least* probability 0.5 that the LRT is greater than zero.

The gradient equation (4.5) makes it straightforward to do simulations of the probability that the likelihood ratio statistic takes on the value 0. However, if the gradient inequality fails and we are testing one component versus two, then there are some fundamental difficulties with simulation studies due to the nonuniqueness of the solutions to the likelihood equations. Two simulation studies that use different algorithms, different starting values or different convergence criteria are studying different statistics. Unfortunately, we know very little about how much difference it might make.

From a practical point of view, one would like the simulation study to exactly mimic the procedure one uses on a data set, so that the results are appropriate to the estimator actually used. An approach used by Furman and Lindsay (1994b) was to use the unique and easy-to-calculate moment estimators as initial values, then iterate 100 times with the EM algorithm. Since this process is easy to replicate, it gives a fast and useful way to construct critical values by simulation.

In the face of this, it seems even more important to understand the asymptotic structure of the mixture problem better, especially the nature of the multimodality problem.

4.3. Asymptotic multinomial geometry. We now offer a side trip into a particular geometric formulation of asymptotics in the multinomial model, designed to be background for the analysis of likelihood ratio testing in the mixture model. The reader may find that his needs are suited by skimming this section and considering just the implications presented in the next section of this chapter. The results found in this section are not new, but the geometric formulation of the asymptotics is presented with more completeness than is to be found elsewhere, as far as I know.

We return to a general discrete density $f(t;\eta)$ for $t \in \{0, 1, \ldots, T\}$, written in vector form as $\mathbf{f}_\eta$. We suppose that we have an i.i.d. sample of size n and that $\mathbf{d}$ is the vector of sample proportions. We suppose for the probability calculations that $\eta = \eta_0$ is the true parameter value, with corresponding vector $\mathbf{f}_0 := \mathbf{f}_{\eta_0}$. The asymptotic result used to drive the analysis is the fact that as $n \to \infty$, we have

$$n^{1/2}(\mathbf{d} - \mathbf{f}_0) \Rightarrow \mathrm{MVN}(\mathbf{0}, \mathbf{V}).$$

Here the matrix $\mathbf{V} = \mathrm{diag}(\mathbf{f}_0) - \mathbf{f}_0\mathbf{f}_0'$. The covariance matrix is rank deficient since the probability vectors are constrained to lie in the T-dimensional probability simplex, now denoted $\mathbb{P}$.

4.3.1. *The dagger simplex.* To set up the geometry, it is useful to perform a transformation of the vectors in the simplex that will give our calculations a natural probabilistic interpretation. The transformed space arises from constructing ratios of densities with respect to the true density f_0. For any function $g(t)$ on the sample space, we define the *dagger operation* to be

$$g^\dagger(t) := \frac{g(t)}{f_0(t)} - 1.$$

Note that the inverse of the dagger operation is $g(t) = f_0(t)[g^\dagger(t) + 1]$ and that we are necessarily assuming that the true density f_0 is strictly positive on the sample space.

If we apply the dagger operation to all elements of the simplex, $\mathbf{p} \to \mathbf{p}^\dagger$, we obtain a transformed simplex $\mathbb{P} \to \mathbb{P}^\dagger$. The dagger simplex is the convex hull of the extreme points $\mathbf{e}_k^\dagger$. Most importantly for our calculations, the dagger simplex lies in the hyperplane of vectors orthogonal to $\mathbf{f}_0$:

$$\mathbb{P}^\dagger \subset \mathcal{M}_0 := \{\mathbf{v}\colon\ \mathbf{v} \cdot \mathbf{f}_0 = 0\}.$$

Note also that the stochastic interpretation of the equation $\mathbf{v} \cdot \mathbf{f}_0 = 0$ is that $v(X)$, if $X \sim f_0$, is a mean zero variable: $\mathbf{v} \cdot \mathbf{f}_0 = \sum v(t) f_0(t)$. Thus $\mathcal{M}_0$ is the linear space of mean zero variables under the null model. We note moreover that the dagger transformation took the true density and placed it at the

origin: $\mathbf{f}_0^\dagger = \mathbf{0}$. It is important to remember that the origin plays the role of a selected element of the null hypothesis.

On this space we will use the E_0 inner product defined by expectations under E_0 :

$$\langle \mathbf{g}^\dagger, \mathbf{h}^\dagger \rangle = E_0[g^\dagger(X)h^\dagger(X)] = \sum f_0(t)g^\dagger(t)h^\dagger(t).$$

In this space, the distance between two vectors is then

$$\|\mathbf{g}^\dagger - \mathbf{h}^\dagger\|^2 = E_0[g^\dagger(X) - h^\dagger(X)]^2 = \sum \frac{[g(t) - h(t)]^2}{f_0(t)},$$

a form of chi-squared distance on the undaggered functions.

If we let $\mathbf{F}$ be the surface of model densities in the simplex,

$$\mathbf{F} = \{\mathbf{f}_\eta\colon\ \eta \in \text{parameter space}\},$$

then there is a corresponding model surface $\mathbf{F}^\dagger$ in the dagger simplex, and it contains the origin (the selected null model). If η is p-dimensional, we let $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_p$ be the score function vectors at $\eta = \eta_0$; that is, the vectors with tth component defined by

$$s_j(t, \eta) = \frac{\partial \ln f(t; \eta)}{\partial \eta_j}.$$

The score functions give us a way to approximate the space of models $\mathbf{F}^\dagger$ with a linear manifold on which it is easier to do asymptotics. If we pick any *direction* $\mathbf{h}$ in the parameter space, then a first order Taylor expansion gives

$$\mathbf{f}^\dagger_{\eta_0+\mathbf{h}} = (\mathbf{f}^\dagger_{\eta_0+\mathbf{h}} - \mathbf{0}) \approx (\mathbf{s}_1, \ldots, \mathbf{s}_p)\mathbf{h} = \sum h_i \mathbf{s}_i. \tag{4.6}$$

That is, in this space, elements of the model surface $\mathbf{F}^\dagger$ can be approximated in a neighborhood of the null model by linear combinations of the score vectors, the *score tangent space*

$$\mathbf{S} = \left\{ \sum h_i \mathbf{s}_i\colon\ h_i \in \mathbf{R} \right\}. \tag{4.7}$$

This approximation is pictured in Figure 4.2. Note also that the E_0 squared length of a score vector is just the corresponding *Fisher information* $E_0[s_j(X)]^2$. Thus the lengths in this space have a natural statistical interpretation.

Further, and this is quite important for our investigation, if there are *model constraints* on η that limit the directions $\mathbf{h}$ one can move in the parameter space away from η_0, the approximating tangent surface $\mathbf{S}$ will have corresponding constraints on the coefficients h_i. For example, if $\eta_0 = (0, \ldots, 0)'$ and if within the model $\mathbf{F}$ some of the η_i are restricted to be nonnegative, then, from (4.6), the approximating surface (4.7) should have the same restriction on the h_i.

4.3.2. *Maximum likelihood and projections.* The next goal is to reexpress the asymptotics of maximum likelihood estimation in the language of this geometry via studying the properties of the *projections* of the sample proportions onto the tangent score space, which might be called *data-to-model projections*, or more simply, *data projections*.

To do so, we consider the geometric relationship between the daggered data vector $\mathbf{d}^\dagger$ and maximum likelihood estimation. First we note that the E_0 squared distance between $\mathbf{d}^\dagger$ and true density $\mathbf{f}_0^\dagger = \mathbf{0}$ is

$$\sum \frac{[d(t) - f_0(t)]^2}{f_0(t)},$$

which is n^{-1} times the Pearson chi-squared statistic for testing

$$H\colon\ f(t) = f_0(t)$$

against a general multinomial alternative. It therefore has a limiting chi-squared distribution with T degrees of freedom.

Let $\mathscr{P}_{\mathbf{S}}\mathbf{v}$, for $\mathbf{v} \in \mathbf{P}^\dagger$, be the E_0 *projection* of $\mathbf{v}$ onto the linear space $\mathbf{S}$ of the scores. That is,

$$\mathscr{P}_{\mathbf{S}}\mathbf{v} = \sum b_i \mathbf{s}_i,$$

where the coefficients b_i are chosen to minimize the E_0 distance

$$\left\| \mathbf{v} - \sum b_i \mathbf{s}_i \right\|.$$

Let $\hat{\eta}$ be the maximum likelihood estimator of the parameter η, with $\hat{\mathbf{f}} := \mathbf{f}_{\hat{\eta}}$ thereby being the maximum likelihood estimator of the density vector. Our *claim* is that the data projection $\mathscr{P}_{\mathbf{S}}\mathbf{d}^\dagger$ is asymptotically equivalent to the MLE $\hat{\mathbf{f}}^\dagger$ in the sense that

$$n\|\hat{\mathbf{f}}^\dagger - \mathscr{P}_{\mathbf{S}}\mathbf{d}^\dagger\|^2 \to 0 \quad \text{in probability.}$$

This is pictured in Figure 4.2.

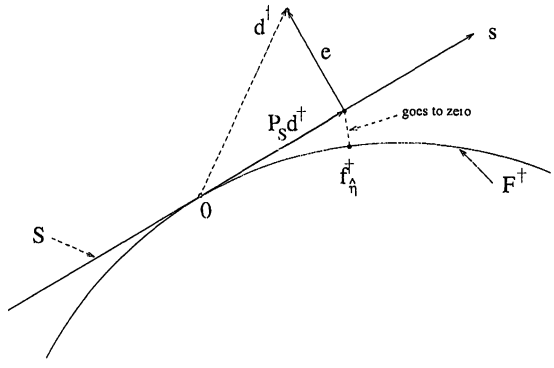

FIG. 4.2. *The score tangent space approximates the daggered model.*

For simplicity's sake, we consider the proof of the claim in the ($p = 1$)-dimensional case. First, a Taylor expansion of $\hat{\mathbf{f}}^\dagger$ in η about η_0 gives us

$$\hat{\mathbf{f}}^\dagger = \mathbf{0} + (\hat{\eta} - \eta_0)\mathbf{s} + O_p(n^{-1}).$$

[We note, as an important aside for novices in asymptotics, that this implies that "all the action" is taking place near the origin because $(\hat{\eta} - \eta_0) = O_p(n^{-1/2})$. As $n \to \infty$, provided the null model is true, we can count on the statistical objects of interest becoming closer and closer to the origin, which is where the tangent score approximation works best.]

Second, we can explicitly calculate the projection of $\mathbf{d}^\dagger$ on the score space to be

$$\mathscr{P}_{\mathbf{S}}\mathbf{d}^\dagger = \hat{b}\mathbf{s} \quad \text{where } \hat{b} = \frac{\langle \mathbf{s}, \mathbf{d}^\dagger \rangle}{\|\mathbf{s}\|^2} = \sum d(t)s(t;\eta_0)i^{-1}, \tag{4.8}$$

where i is the Fisher information about η at η_0. [*Exercise.*] However, examination of $\hat{b}$ in the last equation shows it is equal to the first order influence function expansion of the maximum likelihood functional $\hat{\eta}$ about η_0, so that $\hat{\eta} - \eta_0 = \hat{b} + O_p(n^{-1})$. Hence

$$n\|\hat{\mathbf{f}}^\dagger - \mathscr{P}_{\mathbf{S}}\mathbf{d}^\dagger\|^2 = n\|(\hat{\eta} - \eta_0 - \hat{b})\mathbf{s} + O_p(n^{-1})\|^2 = O_p(n^{-1}).$$

This establishes our claim.

Thus finding the maximum likelihood estimator of the density is, to the appropriate statistical order, equivalent to finding a projection of the daggered data onto the tangent score space.

Finally, we establish a simple asymptotic distribution theory for data projections. Note that taking inner products with the daggered data vector gives us sample averages:

$$\langle \mathbf{v}, \mathbf{d}^\dagger \rangle = \sum_t f_0(t)v(t)\left(\frac{d(t)}{f_0(t)} - 1\right) = n^{-1}\sum_i v(X_i),$$

provided $v(t) \in \mathscr{M}_0$. Thus for elements $\mathbf{v} \in \mathscr{M}_0$ there exists a very simple rule for calculation of limiting distributions for pairs (and vectors) of inner products with the daggered data vector. It is a multivariate normal distribution, with E_0 inner products determining the covariance matrix:

$$n^{1/2}\begin{bmatrix} \langle \mathbf{v}, \mathbf{d}^\dagger \rangle \\ \langle \mathbf{w}, \mathbf{d}^\dagger \rangle \end{bmatrix} \to N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \|\mathbf{v}\|^2 & \langle \mathbf{w}, \mathbf{v} \rangle \\ \langle \mathbf{w}, \mathbf{v} \rangle & \|\mathbf{w}\|^2 \end{pmatrix}\right]. \tag{4.9}$$

This is a simple *exercise* for the reader.

This leads to an elegant expression for the limiting distribution for a data projection. Suppose we have a linear space $\mathbf{T}$ with an *orthonormal basis* (E_0 inner product) $\mathbf{t}_1, \ldots, \mathbf{t}_k$, so that $\|\mathbf{t}_j\| = 1$ and $\langle \mathbf{t}_i, \mathbf{t}_j \rangle = 0$. The use of such a basis is desirable, because then the projection of $\mathbf{d}^\dagger$ onto $\mathbf{T}$ can be represented very simply as the vector

$$\langle \mathbf{t}_1, \mathbf{d}^\dagger \rangle \mathbf{t}_1 + \cdots + \langle \mathbf{t}_k, \mathbf{d}^\dagger \rangle \mathbf{t}_k.$$

This leads to a *fundamental result* that will simplify many of the calculations that we will undertake in the following. From (4.9), the asymptotic distribution of the data projection is therefore

$$n^{1/2}\mathscr{P}_{\mathbf{T}}(\mathbf{d}^{\dagger}) \to Z_1\mathbf{t}_1 + \cdots + Z_k\mathbf{t}_k, \tag{4.10}$$

where $(Z_1, \ldots Z_k)$ are i.i.d. standard normal variates.

4.3.3. *Type I likelihood ratio testing.* We now turn to the geometric interpretation of the likelihood ratio tests. Our claim is that the likelihood ratio tests for model hypotheses are asymptotically equivalent to the lengths of certain E_0 projections in the tangent score space.

First, we consider the simple versus composite hypotheses $\eta = \eta_0$ versus $\neq$. Here we assume knowledge from standard asymptotic results available elsewhere that the quadratic form of the score test statistic is asymptotically equivalent to the likelihood ratio test statistic. It is easily checked that the normalized quadratic form score test statistic

$$n\left[\sum d(t)\mathbf{s}(\eta_0, t)\right] i^{-1}\left[\sum d(t)\mathbf{s}(\eta_0, t)\right]$$

for these hypotheses is $n\|\mathscr{P}_{\mathbf{S}}(\mathbf{d}^{\dagger})\|^2$. [*Exercise*: See (4.8).] Moreover, we note that, in general, n times the squared length of the projection of $\mathbf{d}^{\dagger}$ onto a linear space such as $\mathbf{S}$ will result in a variate that has a limiting chi-squared distribution, with degrees of freedom equal to the dimension of $\mathbf{S}$, which is the same as the number of parameters in the model. [*Exercise*: Work this out using (4.10).]

The next level of difficulty is to incorporate nuisance parameters into the testing problem. We partition $\eta = (\theta, \gamma)$ and consider the hypotheses $\theta = \theta_0$ versus $\neq$. We now have two sets of score functions corresponding to the two sets of parameters. We will denote the θ scores by $\mathbf{u}$'s and the γ scores by $\mathbf{v}$'s. We then decompose the score tangent space $\mathbf{S}$ into two parts: the *nuisance score space* $\mathbf{V}$, generated by linear combinations of the $\mathbf{v}_j$, and the *corrected score space* $\tilde{\mathbf{U}}$, which will be the orthogonal complement of $\mathbf{V}$ within $\mathbf{S}$. Note that if we construct *corrected scores* via

$$\tilde{\mathbf{u}}_k := \mathbf{u}_k - \mathscr{P}_{\mathbf{V}}(\mathbf{u}_k),$$

then $\tilde{\mathbf{U}}$ is generated by linear combinations of the $\tilde{\mathbf{u}}_k$. (These are also known as the *efficient* scores.)

In this case, the likelihood ratio statistic for the composite hypotheses $\theta = \theta_0$ versus $\neq$ corresponds asymptotically to the difference in two score test statistics. That is, we can write

$$\begin{aligned} &2[\ln f(x; \hat{\theta}, \hat{\gamma}) - \ln f(x; \theta_0, \hat{\gamma}_{\theta_0})] \\ &\quad = 2[\ln f(x; \hat{\theta}, \hat{\gamma}) - \ln f(x; \theta_0, \gamma_0)] \\ &\quad\quad - 2[\ln f(x; \theta_0, \hat{\gamma}_{\theta_0}) - \ln f(x; \theta_0, \gamma_0)], \end{aligned}$$

which is the difference between the test statistics for two simple versus composite hypotheses and so is asymptotically equivalent to

$$n\|\mathscr{P}_{\mathbf{S}}(\mathbf{d}^{\dagger})\|^2 - n\|\mathscr{P}_{\mathbf{V}}(\mathbf{d}^{\dagger})\|^2 = n\|\mathscr{P}_{\tilde{\mathbf{U}}}(\mathbf{d}^{\dagger})\|^2.$$

The last equality above derives from orthogonality of $\mathbf{V}$ and $\tilde{\mathbf{U}}$ and the fact that they span $\mathbf{S}$. The test statistic has, therefore, from the above remarks, an asymptotic chi-squared distribution with degrees of freedom equal to the dimension of θ. A testing situation in which the above standard likelihood ratio asymptotic theory applies will be called a *type I* problem.

4.4. The type II likelihood ratio problem. The next step in our analysis is to take account of *restrictions* on the parameter space that will alter the preceding distribution theory. This section increases the level of difficulty to the next degree. Although it is not sufficient to carry us all the way through to the solution of the mixture likelihood ratio solution, it is essential background.

4.4.1. *Parameter constraints.* Inasmuch as this section is designed to lead to the next, we will simplify the general context somewhat. In the general context, there are *focal* parameters θ, parameters of interest that are restricted by the null hypothesis and *nuisance* parameters γ that are not specified by the hypotheses. In general, both the focal and nuisance parameters will be divisible into two types: those constrained by the model and those not. (This description is a bit vague, but we hope that further considerations will clarify it.) In order to be true to our objective—the mixture model LRT—and keep things as simple as possible, we will reduce to the case where the focal parameters *all* have constraints put upon them in the neighborhood of the null hypothesis, but that *none* of the nuisance parameters do. To further simplify, we will assume that the null hypothesis is $\boldsymbol{\theta} = \mathbf{0}$ and that the alternative is $\boldsymbol{\theta} \geq \mathbf{0}$.

As they will be useful to us the next section, we also will follow throughout this section the following two examples.

EXAMPLE 11. We have a single focal parameter θ, with the restriction that $\theta \geq 0$. The null hypothesis is $\theta = 0$ and the alternative is $\theta \geq 0$. All nuisance parameters are unconstrained in the neighborhood of the true model (θ_0, γ_0).

EXAMPLE 12. In addition to an arbitrary set of nuisance parameters, we have two focal parameters (θ_1, θ_2), with the constraint under the alternative:

$$\theta_1 \geq 0, \qquad \theta_2 \geq 0.$$

4.4.2. *Convex cones.* As we noticed earlier, under a set of restrictions the model surface $\mathbf{F}^{\dagger}$ is no longer approximated by $\mathbf{S}$, the entire linear space generated by the scores $\mathbf{s}_i$, but rather by a restricted set, in our case the *score tangent cone:*

$$\mathbf{S}^* = \left\{ \sum a_i \mathbf{u}_i + \sum b_j \mathbf{v}_j \colon\ a_i \geq 0,\ b_j \in \mathscr{R} \right\}.$$

It is useful to note that we can also write $\mathbf{S}^*$ in terms of the corrected focal scores:

$$\mathbf{S}^* = \left\{ \sum a_i \tilde{\mathbf{u}}_i + \sum b_j \mathbf{v}_j \colon\ a_i \geq 0,\ b_j \in \mathscr{R} \right\}.$$

At this point we need some further terminology to describe the kind of set we are dealing with. First, a set of vectors $\mathbf{W}$ is a *cone* if it contains all *rays* through points in $\mathbf{W}$: that is, if $\mathbf{w} \in \mathbf{W}$, then $c\mathbf{w} \in \mathbf{W}$ for all $c > 0$. The set $\mathbf{W}$ is a *convex cone* if it is a cone that is also convex. Check that the restricted tangent space $\mathbf{S}^*$ is a convex cone.

We will call $\mathbf{W}$ the *positive cone* generated by the vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ if

$$\mathbf{W} = \left\{ \sum\nolimits_i a_i \mathbf{w}_i \colon\ a_i \geq 0 \right\}.$$

A positive cone is clearly a convex cone.

Although convex cones do not have all the desirable features of a linear space, if they are *closed*, they do share with a linear space the uniqueness of projections. That is, given any vector $\mathbf{v}$ and a closed convex cone $\mathbf{W}$, there is a unique vector $\mathbf{w} \in \mathbf{W}$, *the projection of* $\mathbf{v}$ *onto* $\mathbf{W}$, that minimizes $\|\mathbf{v} - \mathbf{w}\|$. We will study some further properties of these projections later.

Chernoff (1954) proved that the limiting distribution theory of the likelihood ratio test can be generated by doing data projections onto the score tangent cones, provided that these cones approximate the model surface. In our example, this means that the likelihood ratio statistic is asymptotically equivalent to

$$n\|\mathscr{P}_{\mathbf{S}^*}(\mathbf{d}^\dagger)\|^2 - n\|\mathscr{P}_{\mathbf{V}}(\mathbf{d}^\dagger)\|^2,$$

where $\mathbf{V}$ is the tangent space of the nuisance parameters. Check that this is also the squared length of the data projection $\mathscr{P}_{\tilde{\mathbf{U}}^*}(\mathbf{d}^\dagger)$, where

$$\tilde{\mathbf{U}}^* := \left\{ \sum a_i \tilde{\mathbf{u}}_i \colon\ a_i \geq 0 \right\}$$

is the positive cone generated by the corrected score functions. (See Figure 4.3.) That is, we may without loss of generality restrict our attention to the linear space $\tilde{\mathbf{U}}$ that contains $\tilde{\mathbf{U}}^*$ and ignore the orthogonal directions corresponding to the space $\mathbf{V}$.

4.4.3. *The z-coordinate system.* It is useful to define a new coordinate system for the linear space $\tilde{\mathbf{U}}$. If we do so appropriately, then we can reduce the general problem to a standardized form involving projections of i.i.d. standard normal random variables, so that we can directly deduce the limiting distributions of the projections. Returning to the original coordinate system then will give us the implications for a general model.

We taken any E_0 orthonormal basis for $\tilde{\mathbf{U}}$, say $\mathbf{b}_1, \ldots, \mathbf{b}_d$, and represent points $\mathbf{w}$ in this space by their orthonormal coordinates $\mathbf{z} = \mathbf{z}(\mathbf{w})$ defined by

$$z_1 = \langle \mathbf{b}_1, \mathbf{w} \rangle, \ldots, z_d = \langle \mathbf{b}_d, \mathbf{w} \rangle.$$

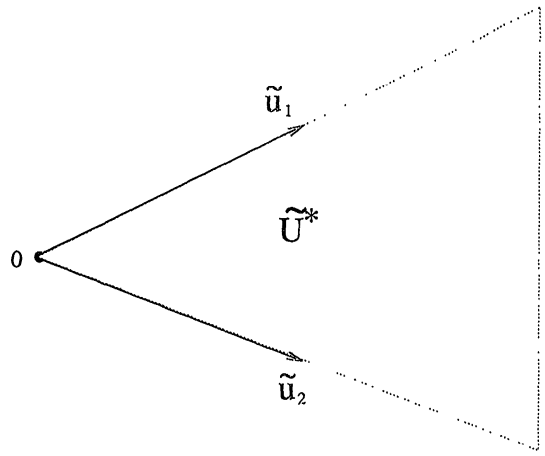

FIG. 4.3. *The positive cone generated by two corrected score functions.*

[Of course, there exists an appropriate $d \times (T+1)$ matrix $\mathbf{B}$ such that $\mathbf{Bw} = \mathbf{z}$.] We will let the z-coordinate representation of the normalized data projection

$$n^{1/2}\mathscr{P}_{\tilde{\mathbf{U}}}\mathbf{d}^\dagger = \mathscr{P}_{\tilde{\mathbf{U}}}(n^{1/2}\mathbf{d}^\dagger)$$

be denoted by $\mathbf{Z}$. Recall from our original considerations (4.10) that $\mathbf{Z}$ is asymptotically standard normal.

We have thus replaced our original data with a set of standard normal variables. We next need to understand how the restrictions in the original score space show up in this transformed version and how the original E_0 geometry is transformed.

The first simplification comes because when working with the z-coordinates we replace the E_0 geometry with the ordinary Euclidean inner product and distance. This arises because the E_0 inner product between any two points in $\tilde{\mathbf{U}}$ equals the ordinary Euclidean inner product for their $\mathbf{z}$-coordinate representations:

$$\begin{aligned}\left\langle \sum z_{i1}\mathbf{b}_i, \sum z_{i2}\mathbf{b}_i \right\rangle &= E_0\left[\left(\sum z_{i1} b_i(X)\right)\left(\sum z_{i2} b_i(X)\right)\right] \\ &= \sum z_{i1} z_{i2} \\ &= \mathbf{z}_1 \cdot \mathbf{z}_2.\end{aligned}$$

In this new coordinate system, we have a transformed version $\mathbf{z}(\tilde{\mathbf{U}}^*)$ of the original cone $\tilde{\mathbf{U}}^*$, corresponding to the set of coordinates $\mathbf{z}$ of all the points in $\tilde{\mathbf{U}}^*$. As an *exercise*, show that it is generated as the positive cone of the normalized extremal vectors

$$\mathbf{p}_i := \mathbf{z}(\tilde{\mathbf{u}}_i)/\|\mathbf{z}(\tilde{\mathbf{u}}_i)\|.$$

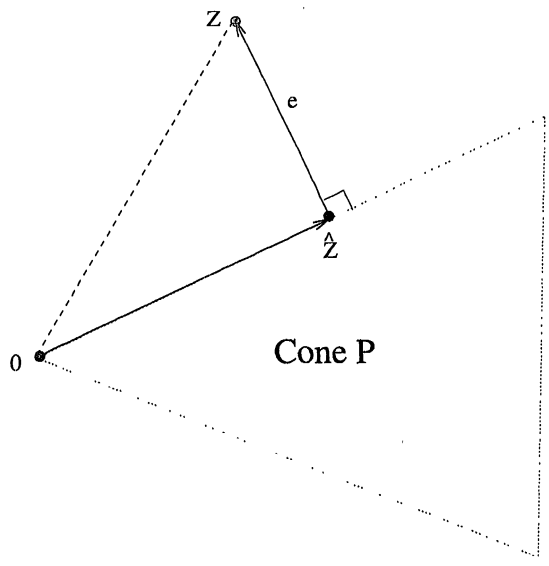

FIG. 4.4. *The projection of* **z** *onto the cone* **P**.

We will call this cone in z-space

$$\mathbb{P} = \left\{ \sum a_i \mathbf{p_i} \colon a_i \geq 0 \right\}$$

the *primal cone*.

Suppose we wish to find the z coordinates of the E_0 projection of $\mathbf{d}^\dagger$ onto $\tilde{\mathbf{U}}^*$ so that we can determine the limiting distribution of our data projection. The projection can be carried out in two steps, by first projecting onto $\tilde{\mathbf{U}}$, then onto $\tilde{\mathbf{U}}^*$. Because of the equivalence above for inner products, and therefore distances, we can *instead* work in the z-coordinate space and use Euclidean distances. That is, we take the z-coordinates of $\mathscr{P}_{\tilde{\mathbf{U}}}\mathbf{d}^\dagger$ and do an ordinary Euclidean projection of them onto $\mathbb{P}$.

Thus we have transformed our asymptotic problem into the problem of determining the distribution of the squared length of the projection of a vector of *standard normal variables* **Z** onto a convex cone **P**. The situation is pictured in Figure 4.4, where the projection is denoted $\hat{\mathbf{Z}}$.

4.4.4. *Projections onto convex cones.* We now study the properties of projections onto convex cones because then we can determine the limiting distributions of the data projections. A useful concept, akin to the notion of the orthogonal subspace, is the idea of the dual convex cone.

The *dual (or polar) cone* $\mathbf{P}^o$ to a convex cone **P** is defined to be the set of all vectors that are negatively correlated with all the vectors in **P**, namely,

$$\mathbf{P}^o = \{\mathbf{y} \colon \mathbf{y} \cdot \mathbf{m} \leq 0 \text{ for all } \mathbf{m} \in \mathbf{P}\}.$$

If the cone $\mathbf{P}$ is closed and convex, then $\mathbf{P}^o$ is closed and convex and $(\mathbf{P}^o)^o = \mathbf{P}$. To simplify the notation, let

$$\hat{\mathbf{z}} := \mathscr{P}_{\mathbf{P}}\mathbf{z} \quad \text{and} \quad \mathbf{e} := \mathscr{P}_{\mathbf{P}^o}\mathbf{z}$$

be the *primal* and *dual projections* of $\mathbf{z}$. The terminology *fitted value vector* and *residual vector* would also be appropriate, as a reminder of the similarity of this problem to the regression problem, in which the vector of observations $\mathbf{y}$ is decomposed into $\hat{\mathbf{y}} + \mathbf{e}$, where $\hat{\mathbf{y}}$ is the vector of fitted values, determined by linear projection of $\mathbf{y}$ onto the model space, and $\mathbf{e}$ are the residuals, determined by the projection onto the subspace orthogonal to the model. The second part of the following proposition indicates that the parallel to the regression problem extends to the orthogonality of the primal and dual projections.

PROPOSITION 13. *Let* $\mathbf{P}$ *be a closed convex cone and let* $\mathbf{D}$ *be its dual cone. The projection* $\hat{\mathbf{z}}$ *is the unique element of* $\mathbf{P}$ *that satisfies the gradient inequality*

$$(\mathbf{z} - \hat{\mathbf{z}}) \cdot \mathbf{m} \leq 0 \quad \textit{for all } \mathbf{m} \in \mathbf{P}.$$

Further, there exists an orthogonal decomposition of $\mathbf{z}$ *into its primal projection and dual projection. That is,*

$$\mathbf{z} = \hat{\mathbf{z}} + \mathbf{e} \quad \textit{with } \hat{\mathbf{z}} \cdot \mathbf{e} = 0.$$

PROOF. [*Exercise.*] First show that $(\mathbf{z} - \hat{\mathbf{z}}) \cdot \hat{\mathbf{z}} = 0$ using the fact that cones contain all rays and that therefore $\|\mathbf{z} - \alpha\hat{\mathbf{z}}\|$ is minimized at $\alpha = 1$. For the gradient inequality, use convexity of the cone and the fact that $\|\mathbf{z} - (1-\varepsilon)\hat{\mathbf{z}} - \varepsilon\mathbf{m}\|$ is therefore minimized at $\varepsilon = 0$. To show that $(\mathbf{z} - \hat{\mathbf{z}})$ is the dual projection $\mathbf{e}$, show that it satisfies the gradient inequality on the dual cone. □

To derive the limiting distribution theory, we will use the elegant geometric description of the conal projection problem given by Fraser and Massam (1989), who apply it to construct an algorithm for the projection. The discussion is therefore limited to convex cones with a certain simple structure. However, the loss in generality is balanced by a set of important insights.

4.4.5. *The dual basis.* Suppose that our convex cone $\mathbf{P}$ is generated as the positive cone induced by a linearly independent set of unit vectors $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_d$ in d-dimensional space, the primal basis. We can construct a *dual basis* for the dual cone by sequentially finding *unit* vectors $\mathbf{d}_1, \ldots, \mathbf{d}_d$ such that

$$\mathbf{d}_i \cdot \mathbf{p}_j = 0 \quad \text{for } j \neq i.$$

This orthogonality determines $\mathbf{d}_i$ up to its sign, which we determine by making it negatively correlated with its primal partner:

$$\mathbf{d}_i \cdot \mathbf{p}_i \leq 0.$$

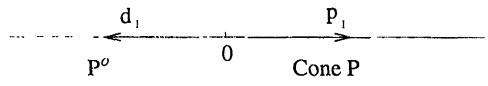

FIG. 4.5. *The dual and primal cones for Example 11.*

It is easy to check that every $\mathbf{d}_i$ is an element of the dual cone and that every positively weighted linear combination of the $\mathbf{d}_i$ is an element of the dual cone. In fact, the dual cone is just the positive cone generated by the dual basis.

Moreover, we can partition the space into 2^d positive cones by using as a basis for each cone $\mathbf{b}_1, \ldots, \mathbf{b}_d$, where each $\mathbf{b}_i$ equals either $\mathbf{p}_i$ or $\mathbf{d}_i$. Check that the set of vectors $\mathbf{b}_1, \ldots, \mathbf{b}_d$ is also linearly independent. We will call each such a cone a *sector*. We can index the sectors by $\delta = (\delta_1, \ldots, \delta_d)$, where δ_i is 1 if $\mathbf{b}_i = \mathbf{p}_i$ and is 0 if $\mathbf{b}_i = \mathbf{d}_i$. The basis vectors used in the sector will be called the *active* basis for that sector. We will denote such a sector by

$$\mathscr{S}(\delta) = \left\{ \sum a_i [\delta_i \mathbf{p}_i + (1 - \delta_i)\mathbf{d}_i] \colon a_i \geq 0 \right\}.$$

In each such sector, the active primal vectors $\mathbf{p}$ are orthogonal to the active dual basis vectors $\mathbf{d}$, a fact that is important in the projection and, later, the distribution theory.

We illustrate these notions with our two simple examples. In Example 11, the cone $\tilde{\mathbf{U}}^*$ is simply a single vector, so we have only a single coordinate z_1 in z space. See Figure 4.5. The induced primal cone is then simply $\{z_1 \geq 0\}$. The projection of v onto this cone is therefore either v itself, if v is nonnegative, or 0, if v is negative. The dual cone is $\{z_1 \leq 0\}$. The reader should check the validity of the theorem in this simple case.

Example 12 has more geometric content. There are now two primal vectors $\mathbf{p}_1$ and $\mathbf{p}_2$. In Figure 4.6, we show the location of the dual basis vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, which are orthogonal to $\mathbf{p}_2$ and $\mathbf{p}_1$, respectively, with direction chosen so as to preserve the negative correlation with the other $\mathbf{p}$ vector. These two bases generate a division of the plane into four conal sectors, corresponding to the primal and dual cones, $\mathbf{P} = \mathbf{P}(1,1)$ and $\mathbf{P}^o = \mathbf{P}(0,0)$, and two regions $\mathbf{P}(0,1)$ and $\mathbf{P}(1,0)$, where we use 0 and 1 to indicate whether the cone was generated with dual or primal basis element in that position.

4.4.6. *Sector decomposition and projection.* The beauty of the sector decomposition is that in each sector there is a simple expression for the projection of $\mathbf{z}$ onto $\mathbf{P}$. Check that in the preceding example,

$$\begin{aligned}
\mathbf{z} \in \mathbf{P} &\Rightarrow \hat{\mathbf{z}} = \mathscr{P}_{\mathbf{P}}\mathbf{z} = \mathbf{z}, \\
\mathbf{z} \in \mathbf{P}^o &\Rightarrow \hat{\mathbf{z}} = \mathbf{0}, \\
\mathbf{z} \in \mathbf{P}(1,0) &\Rightarrow \hat{\mathbf{z}} = (\mathbf{p}_1 \cdot \mathbf{z})\mathbf{p}_1, \\
\mathbf{z} \in \mathbf{P}(0,1) &\Rightarrow \hat{\mathbf{z}} = (\mathbf{p}_2 \cdot \mathbf{z})\mathbf{p}_2.
\end{aligned}$$

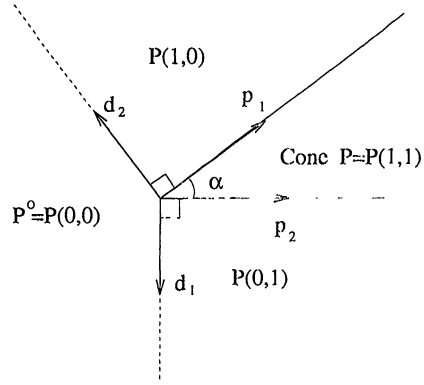

FIG. 4.6. *The cones and sectors for Example 12.*

That is, in this case, and more generally, when $\mathbf{z}$ is in a conal sector, the projection $\hat{\mathbf{z}}$ is formed by *projecting onto the linear space of the active primal vectors*. It is also clear that $\mathbf{e}$ is found in each case by projecting onto the active dual basis for the sector.

It will be useful to characterize the relationships between these projections. Define the linear space $\mathscr{L}_1(\delta)$ to be the linear space spanned by the active primal vectors and let

$$\mathscr{C}_1(\delta) = \left\{ \sum a_i \delta_i \mathbf{p}_i \colon a_i \geq 0 \right\}$$

be the corresponding positive cone. Let $\mathscr{L}_2(\delta)$ and $\mathscr{C}_2(\delta)$ be the corresponding linear and convex cones for the active dual vectors. Let $\mathbf{z}_1$ and $\mathbf{z}_2$ be the linear projections of $\mathbf{z}$ onto $\mathscr{L}_1(\delta)$ and $\mathscr{L}_2(\delta)$, respectively. We then have the following relationships:

PROPOSITION 14. *The following statements are equivalent:* $\mathbf{z} \in \mathscr{S}(\delta)$ *if and only if* $\mathbf{z}_1 \in \mathscr{C}_1(\delta)$ *and* $\mathbf{z}_2 \in \mathscr{C}_2(\delta)$ *if and only if* $\hat{\mathbf{z}} = \mathbf{z}_1$ *and* $\mathbf{e} = \mathbf{z}_2$.

PROOF. [*Exercise.*] It will be useful to note that $\mathscr{L}_1$ and $\mathscr{L}_2$ are orthogonal spaces and that $\mathbf{z}$ has a unique representation in terms of the sector's basis vectors $\mathbf{b}$. □

4.4.7. *The type II LRT.* To do a limiting distribution calculation for the asymptotic version of the likelihood ratio test statistic, namely, $d^2 = \|\mathscr{P}_{\mathbf{P}}\mathbf{Z}\|^2$, we can calculate probabilities via the law of total probability,

$$\Pr(d^2 > t) = \sum_{\text{sectors}} \Pr(d^2 > t | \mathbf{Z} \in \text{sector}) \Pr(\mathbf{Z} \in \text{sector}). \tag{4.11}$$

We tackle this calculation in two steps. Part of the calculation of the limiting distribution turns out to be quite straightforward.

PROPOSITION 15. $\Pr(d^2 > t \mid \mathbf{Z} \in \text{sector}) = \Pr(\chi_k^2 > t)$, *where* k *is the number of active primal constraints in the sector.*

PROOF. From the preceding proposition, we need to calculate

$$\Pr[\|\hat{Z}\|^2 > \mathbf{t} \mid \mathbf{Z}_1 \in \mathscr{C}_1(\delta) \text{ and } \mathbf{Z}_2 \in \mathscr{C}_2(\delta)],$$

which in turn equals

$$\Pr[\|\mathbf{Z}_1\|^2 > \mathbf{t} \mid \mathbf{Z}_1 \in \mathscr{C}_1(\delta) \text{ and } \mathbf{Z}_2 \in \mathscr{C}_2(\delta)].$$

However, by the independence of $\mathbf{Z}_1$ and $\mathbf{Z}_2$, since they arise from orthogonal projections, the condition on $\mathbf{Z}_2$ is irrelevant in this calculation. Further, the conal structure implies that we can replace the condition $\mathbf{Z}_1 \in \mathscr{C}_1$ with its equivalent condition $\mathbf{Z}_1/\|\mathbf{Z}_1\| \in \mathscr{C}_1$. However, this unit vector has a distribution independent of its length $\|\mathbf{Z}_1\|$ (e.g., by Basu's theorem). Hence our calculation boils down to $\Pr(\|\mathbf{Z}_1\|^2 > \mathbf{t})$, which has the given chi-squared property. □

The second part of the calculation, $\Pr(\mathbf{Z} \in \text{sector})$, is more difficult, and the answer will vary from problem to problem. The problem can be turned into the problem of calculating the surface area of a region on the unit sphere as follows. Let $R = \sqrt{Z_1^2 + \cdots + Z_d^2}$. Because of normality, the conditional distribution of the vector $\mathbf{Z}$ given R is uniform on the sphere of radius $R = r$. It follows that $\mathbf{U} := \mathbf{Z}/R$ is uniformly distributed on the unit sphere. Now $\mathbf{Z}$ is in a positive cone, say $\mathbf{P}$, if and only if $\mathbf{U}$ is in the intersection $\mathbf{I}$ of the sphere with $\mathbf{P}$. The probability of $\mathbf{Z}$ being in the cone is therefore equal to the spherical surface area of $\mathbf{I}$ divided by the total surface area of the sphere. (I will herein refer to "area" rather than "volume," even though the sphere is of arbitrary dimension.) We will illustrate such a calculation shortly.

It follows from (4.11) that the distribution of d^2 has the form of a mixture of chi-squared distributions,

$$\sum_{i=0}^{\dim(\theta)} w_i \chi_i^2,$$

with the weights w_i determined by the probabilities of the vector $\mathbf{z}$ following into various sectors. Such a mixture of chi-squared distributions is called a *chi-bar-squared distribution* and is written $\bar{\chi}^2$.

The chi-bar-squared distributional result holds more generally than for the cones considered here. It can be extended to the positive cones generated by vector sets having an arbitrary number of elements by extending the above ideas. Shapiro (1985) used this fact to show that the chi-bar-squared distribution holds for projections onto an *arbitrary* convex cone. (Essentially, we can approximate arbitrary convex cones with the positive cones associated with a set of primal vectors and then take limits on the number of primal vectors used.)

When a likelihood ratio test has the structure of a convex cone, with the resulting distribution being chi-bar-squared, we will refer to it as a *type II problem.*

4.4.8. *Applications.* Example 11 illustrates the simplest such conal projection result. We have but a single parameter of interest, with inequality constraint on it. The theory above leads us to conclude that the likelihood ratio statistic has the limiting distribution $0.5\chi_0^2 + 0.5\chi_1^2$, where 0.5 is the probability of z_1 being positive, the projection therefore equaling z_1, and the resulting sector distribution being χ_1^2. The second component $0.5\chi_0^2$ arises from z_1 being negative and the projection therefore being $\hat{z}_1 = 0$.

An example of this type from the mixture problem is the following. Suppose wish to test H: $\pi = 0$ versus A: $\pi > 0$ in the mixture model $(1 - \pi)f + \pi g$, where f and g are both known. That is, we ask if the distribution f has been contaminated by observations from the second distribution g.

In Example 12, it is clear now that the distribution of d^2 is a mixture of χ_k^2, for $k = 0, 1, 2$. To determine the weights, note first that the weight for the χ_1^2 component is 1/2, because $\mathbf{P}(0, 1)$ and $\mathbf{P}(0, 1)$ each contribute probability 1/4. (This is the probability of falling in an arc of 90° on the unit circle under the uniform measure on the circle.)

The probability of falling in the other two sectors is proportional to the angles they subtend. It suffices to determine the sector probability for $\mathbf{P}(1, 1)$. We need the *angle* α between the primal basis vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, but we need to express it in terms of the original variables that were projected into the $\mathbf{z}$ coordinate system. It can be argued that

$$\cos(\alpha) = \frac{E[\tilde{u}_1(X)\tilde{u}_2(X)]}{\sqrt{E[\tilde{u}_1(X)^2]E[\tilde{u}_2(X)^2]}}, \tag{4.12}$$

in which case the sector probability is $\alpha/2\pi$. (This is the angle between the generating vectors, as expressed using the E_0 inner product. As we have already seen, this agrees with the angle in the z-coordinate system of the transformed vectors.) All the terms on the right hand side of (4.12) can be calculated from the elements of the Fisher information matrix for the parameters at the null hypothesis.

However, there is an important issue here if there are nuisance parameters in the model. In the type I likelihood ratio test theory, there is no possible dependence of the limiting distribution on the nuisance parameters in the null hypothesis. Here, however, there is nothing to prevent the angle α from depending on the value of the nuisance parameter γ under the null hypothesis, in which case the limiting distribution varies over the elements of the null hypothesis.

When this problem arises, we will say that we have a *parameter dependent limiting null distribution.* An example will be given in the mixture problem in the next section.

If parameter dependence occurs, then one must develop a secondary strategy for conducting the test. If one desires the test to have the desired proba-

bility of type I error, a conservative strategy would be to use the critical value, say $c(\gamma)$ from the *least favorable* null distribution, with parameter γ. With a little further care as to the asymptotics, one can estimate the nuisance parameter under the null hypothesis, and use the critical value $c(\hat{\gamma})$ from the estimated distribution. In the latter case, one might for the sake of conservativeness employ the least favorable critical value within a confidence interval for the nuisance parameter.

4.5. Asymptotic mixture geometry. Our first problem is to come to an understanding of how we can put the mixture model into the geometric framework of the preceding section. The problem will be that of testing the one-component model against two components. Thus the null hypothesis is described by one-component density $f(t;\phi_0)$. The nuisance parameter in the null hypothesis is the parameter ϕ, which gives us a single nuisance score function $v(\phi,t)$, and this gives us the nuisance score space $\mathbf{V}$.

4.5.1. *Directional score functions.* A *fundamental* difficulty arises in ascertaining the nature of the tangent space $\mathbf{S}$ of score functions at the null hypothesis. In the previous discussion, it was assumed that the score functions under the alternative were also well defined under the null hypothesis. However, this is not true for the usual mixture parameterization. We have three score functions at each alternative parameter value, corresponding to one weight and two component parameters. Their limits as the null hypothesis is approached are problematical, however, because there are problems relating the parameters in the alternative hypothesis to those in the null model.

To illustrate, if we treat the null hypothesis as specifying the equality of the two component parameters, $\phi_1 = \phi_2$, then the score functions for the parameters ϕ_1 and ϕ_2 are both equal to the nuisance score function when evaluated at the null hypothesis. Moreover, the score function for the weights is identically zero there. Thus the score space appears to degenerate to just the score space for the nuisance parameter ϕ.

This analysis is misleading; the easiest way to deal with the problem of determining the score space $\mathbf{S}^*$ is to return to the geometric considerations of Chapter 2. Recall our plots of the structure of the binomial mixture model. The two-component mixture model generates a smooth family of models $\mathbf{F}$ in the probability simplex, a three-dimensional manifold corresponding to the three free parameters. The one-component models are a one-dimensional curve along the edge of this smooth manifold. Recall that our objective is to construct a score manifold that approximates the daggered model surface $\mathbf{F}^\dagger$ in the neighborhood of the null model. We can do this directly as follows. The *directional scores* at a particular null hypothesis point are geometrically constructed by taking limits of the form

$$\varepsilon^{-1}\mathbf{f}_{2\varepsilon}^{\dagger} \longrightarrow \mathbf{s}$$

as $\varepsilon \to 0$, where $\mathbf{f}_{2\varepsilon}$ is a family in $\varepsilon > 0$ of elements of the alternative hypothesis manifold of two-component mixtures that is approaching the null

hypothesis point $\mathbf{f}_0$, a one-component density, with sufficient smoothness that the limit exists. We will let the set of all scores created thusly be the *directional score cone* $\mathbf{S}^*$. If indeed we have the simple model structure of the preceding section, where the focal parameter scores are meaningfully defined in the null hypothesis, then this gives the score cone $\mathbf{S}^*$ of that section, which was there defined in terms of the focal and nuisance parameter score functions.

We note first that $\mathbf{S}^*$ is a cone, because one can simply change the definition of the smooth family to alter the speed of the approach to the null hypothesis to get $c\mathbf{s}$. Note, however, that it is not generally true that such a local alternative family can be extended smoothly from positive ε through 0 into negative ε, so that $-\mathbf{s}$ is *not* necessarily in the directional score cone. This is very important in the mixture model, due to the way that the null lies at the boundary of the alternative. It follows that we must pay attention to the sign of the score function if we wish the score surface to approximate the model surface.

4.5.2. *The gradient scores.* This can be further illustrated by considering the scores generated by the weight parameter. Let us consider the boundaries of the parameter space in Figure 4.1 corresponding to $\pi = 0$ or 1. If for each fixed ϕ we construct the score function for the parameter π in the one parameter density

$$(1 - \pi)f(t; \phi_0) + \pi f(t; \phi),$$

then taking its limit as $\pi \to 0$, we obtain a score function of the form

$$s_\phi(\phi_0; t) = \frac{f(t; \phi)}{f(t; \phi_0)} - 1.$$

This leads to a number of important observations:

1. This generates an *infinite* family of directional score functions $\mathbf{C} = \{c\mathbf{s}_\phi\}$ corresponding to the infinitely many possible values of ϕ. By considering this family, and certain limits, we will be able to determine all the local corrected score functions. However, the dimension of the score space is no longer equal to the number of parameters in the alternative, and this will prevent us from using the simple geometry of the preceding section to derive the limiting distributions.
2. The sample version of the above score is

$$n \sum d(t) s_\phi(\phi_0; t) = D_{\phi_0}(\phi). \tag{4.13}$$

   That is to say, the unicomponent gradient function represents the collective sample values of these scores. For this reason, we will call $\mathbf{s}$ a *gradient score.*
3. These scores are one directional in that the one parameter family used to generate them does not extend beyond $\pi = 0$, so the score is not two-sided. That is, the model surface in the simplex has an edge at the null hypothesis.

4. The gradient scores are related very simply to the dagger operation. That is, we have

$$\mathbf{s}_\phi = \mathbf{f}_\phi^\dagger.$$

Thus if we wish to picture the cone generated by these scores, we plot the one-component model in the dagger space and the cone $\mathbf{C}$ is all rays from the origin through other points in the model surface.

Before considering further issues, we ask the reader to solidify his or her understanding by considering how these facts relate to the pictures of the binomial mixture models found in Chapter 2. The pictures are roughly the same, with the coordinates undergoing some stretching by the daggering operation. Thus, using the Bin(2, p) as an example, Figure 4.7, we see that the gradient score vectors $\mathbf{s}_\phi$ correspond to all the vectors from the origin (the null model) to points on the curve $\{\mathbf{f}_\phi^\dagger\}$ generated by the one-component model.

From the figure we also obtain an important insight. We see that there are sequences of two-component models that approach the null model in such a way that the directional score is not in $\mathbf{C} = \{c\mathbf{s}_\phi\}$. Thus we must create other score functions if we are to generate the entire directional score tangent cone $\mathbf{S}^*$.

4.5.3. *Other directional scores.* We return to our derivation of the $C(\alpha)$ test in order to generate some further score functions. We follow the Neyman–Scott derivation, letting the distribution G be a two point distribution, in which case the location-scale family $\Phi = a + b\Theta$ is also entirely two point distributions. If we let $b \searrow 0$, we should expect to find that the directional score for any such two-component alternative results in the dispersion score $v_2(\phi_0, t)$. However, we note that because the first derivative in b is zero, we must let $b = \sqrt{\varepsilon}$ to get this score. Moreover, we get exactly the same score (no change in sign)

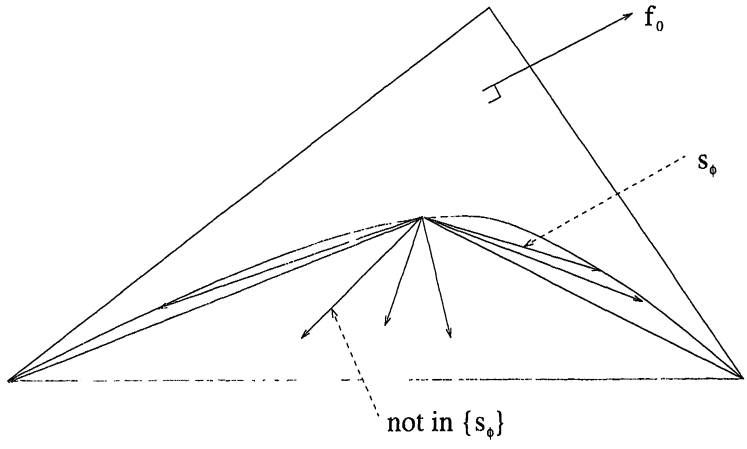

FIG. 4.7. *The directional scores for the* Bin(2, p) *model.*

if $b = -\sqrt{\varepsilon}$. That is, v_2 generates a ray in only one direction in the tangent cone. In the plot for the model Bin$(2, \phi)$, it corresponds to a ray down into the two-component models.

A further set of scores can be generated if we consider all approaches to the null model in which $a \to \phi_0$ at the same time as $b \to 0$, with appropriate rates, in which case we get scores of the form $c_2\mathbf{v}_2 + c_1\mathbf{v}_1$, where $c_2 \geq 0$.

At this point, a rather delicate analysis is needed to verify that we have found *all* possible directional score functions. We offer some heuristics for this: If the local family of alternatives is such that π does not go to 0 or 1, we must have ϕ_1 and ϕ_2 converging to ϕ_0. In this case, we will get a limit involving the dispersion score and the nuisance score. If π does converge to 0, then ϕ_1 must converge to ϕ_0 and the resulting score will be a gradient score provided ϕ_2 converges to something other than ϕ_0.

As in the preceding section, the relevant portion of the directional score cone for determining the asymptotic distributions is orthogonal to the nuisance parameter score space. This leads to a further simplification in the analysis, because, fortunately, the dispersion score does not need to be handled separately from the gradient scores $\mathbf{s}_\phi$ in the geometric analysis. That is, we claim that

$$\tilde{\mathbf{S}}^* = \tilde{\mathbf{C}}.$$

This is because the relevant scores for asymptotic analysis are the corrected scores obtained by the E_0 regression residuals $\tilde{\mathbf{s}}_\phi = \mathbf{s}_\phi - \rho\mathbf{v}_1$. Noting that $\mathbf{s}_\phi/(\phi - \phi_0) \to \mathbf{v}_1$ as $\phi \to \phi_0$, the interested reader should check that the normalized dispersion score $\tilde{\mathbf{v}}_2/\|\tilde{\mathbf{v}}_2\|$ is the limiting vector of the normalized corrected gradient scores $\tilde{\mathbf{s}}_\phi/\|\tilde{\mathbf{s}}_\phi\|$ as $\phi \to \phi_0$.

Thus we only need consider the corrected score cone $\tilde{\mathbf{C}}$ generated by the closure of the cone of the corrected gradient scores, because the other scores are limit points thereof.

4.5.4. *Simple binomial examples.* In the Bin$(2, p)$ plot, there are only two dimensions, and the nuisance parameter score is tangent to the unicomponent model, so the corrected gradient score space $\tilde{\mathbf{C}}$ must lie in the one-dimensional subspace orthogonal to it. It is a directional cone, pointing down into the model. This direction corresponds to the corrected dispersion score $\tilde{\mathbf{v}}_2 = \mathbf{v}_2 - \rho\mathbf{v}_1$. Thus, after this reduction of dimensionality, we are in the setting of Example 11 of the previous section and have the corresponding chi-bar-squared distribution.

In the Bin$(3, p)$ model, Figure 4.8 will enable us to visualize the relevant geometric constructs. The cross section we are viewing corresponds to the plane orthogonal to the nuisance score and so will contain the corrected gradient scores. The corrected gradient scores all lie *between* the two extremal scores corresponding to the mixtures with latent support at $\theta = 0$ and $\theta = 1$, respectively. Thus in this case the corrected tangent cone $\tilde{\mathbf{C}}$ for the model is generated as the positive cone of these two extremal scores. Thus, because of the limited number of dimensions in the simplex, the corrected tangent

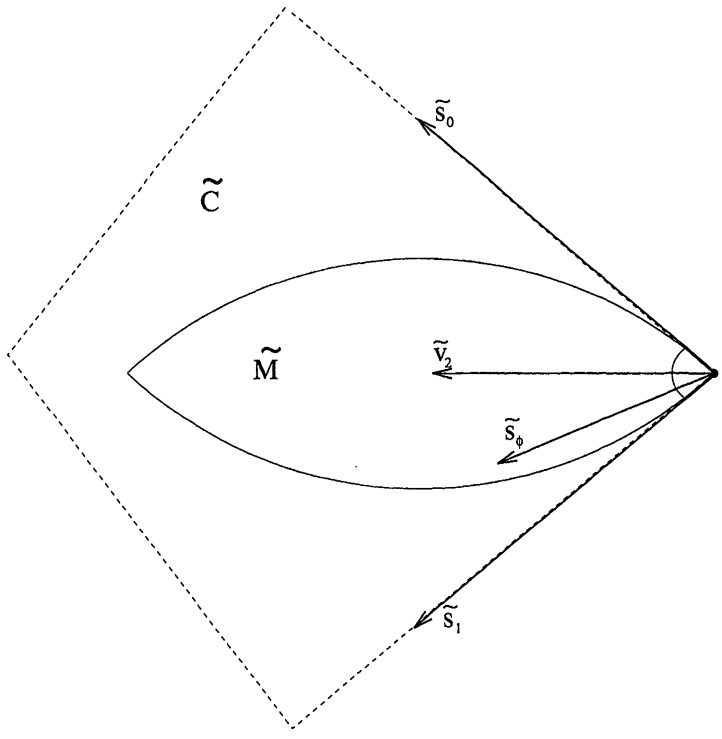

FIG. 4.8. *The corrected gradient score cone of the* Bin(3, p) *model.*

score space is just two dimensional, despite the fact that it was generated by infinitely many gradient score functions. The corresponding picture from our conal analysis is Figure 4.6.

Thus in both these binomial models, the geometry coincides exactly with the type II LRT theory of the previous section. We can conclude that in a Bin(2, p) model, the test of one versus two components has the chi-bar-squared distribution $0.5\chi_0^2 + 0.5\chi_1^2$. In the Bin(3, p) distribution, we have a mixture of chi-squared $(0.5-\alpha)\chi_0^2+0.5\chi_1^2+\alpha\chi_2^2$, where the component weights must be determined by calculation of α in (4.12). This calculation can readily be carried out and we find that the angle is not constant. For example, $\alpha = 0.167$, 0.193 and 0.226 when the log odds parameter equals 0, 2 and 4, respectively.

This proves that in this binomial case the LRT distribution is *parameter dependent* in the null hypothesis, and suggests that it is unlikely to be distribution constant in many other examples.

4.5.5. *The nonparametric LRT.* Before proceeding to the next level of difficulty, where the chi-bar-squared distributions fail, we can use what we already know to make an observation about the nonparametric likelihood ratio test, because the necessary geometric background has been laid. The result is that

the nonparametric likelihood ratio test has a chi-bar-squared distribution as well.

PROPOSITION 16. *In the multinomial model with $T + 1$ cells, under regularity the nonparametric mixture likelihood ratio test statistic has a limiting distribution of the form*

$$w_0\chi_0^2 + \cdots + w_T\chi_T^2 \quad \textit{with} \ \sum w_j = 1.$$

PROOF. The key here is that the tangent cone generated by the mixture models, viewed as the limiting directions in the direction of the mixture models from the null model, is clearly a convex cone and so we can apply the result of Shapiro (1985). □

Although this result is a start on the nonparametric distribution theory, it is useless without the ability to calculate the weights, and we anticipate that is a difficult issue. Additionally, it seems likely that there will be parameter dependence in the weights w, which leads to further difficulties in constructing critical values.

We note that the geometric analysis shows the relationship between the $C(\alpha)$ test for homogeneity and the nonparametric LRT. The Neyman dispersion score measures the tendency of the data to lie in the central direction $\mathbf{v}_2$ of the cone, but ignores the more subtle features of the conal structure. However, it has the clear advantage of an easy distribution theory and simple calculation. It seems likely that one can develop extensions of this test that measure some additional departure in the direction of heterogeneity, such as skewness features, yet still retain manageable limiting distributions.

4.5.6. *A nonconvex score cone.* Once we leave the three-dimensional simplex, where the corrected score space is forced to live in two dimensions, the distribution of the likelihood ratio test for one versus two components becomes significantly more challenging.

Suppose the one-component model is $\text{Bin}(4, p)$. From our earlier geometric analysis we know that the set of mixture probability vectors is a four-dimensional set and that if we constrain ourselves to the plane in the simplex orthogonal to the nuisance score function, then the two-component models generate a two-dimensional surface in the three-dimensional space: see Figure 4.9. In fact, there are two boundary surfaces to the mixture set consisting of the two types of mixtures with index 2: those that mix two values of θ that are in $(0, 1)$ and those that mix $\theta = 0$, $\theta = 1$ and one $\theta \in (0, 1)$. Between these two surfaces are two seams, one on each side, consisting of the mixtures of index 1.5. Only the first surface of index 2 concerns us, because the other would be considered, in a statistical sense, to consist of three-component models.

We can imagine the plot of the surface as being like the $\text{Bin}(3, p)$ plot, only with an extension into the third dimension; something like an American

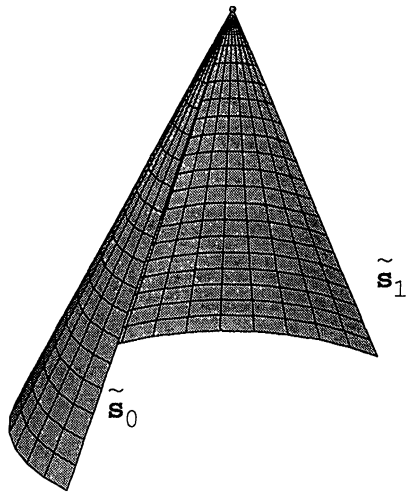

FIG. 4.9. *The corrected gradient score cone of the* Bin(4, p) *model.*

football, but with only two seams. The null model sits at the top of the football, so we can picture those tangent vectors to the surface at the null model that correspond to the directions toward two-component models as creating a cone that has the general shape of a half tepee, sliced in two from its apex down.

That is, this is a situation where the tangent cone $\mathbf{S}$ is *not* convex, and so we cannot use the theory of the type II likelihood ratio test. In essence, the two-dimensional score manifold does not lie in a two-dimensional plane.

Of importance in our further discussion will be the two extremal vectors, corresponding to the two edges of the tepee surface. These correspond, in the Bin(4, p) case, to the two seams with index equal to 1.5, and so arise from a mixture with $p = 0$ on one edge and $p = 1$ on the other.

4.6. The LRT on nonconvex cones. As part of our analysis, we must therefore gain some further understanding of projections onto curved surfaces. In addition, the limiting distribution theory will be closely related to results for normal theory nonlinear regression, where exact results are unusual.

4.6.1. *Projections onto nonconvex cones.* We first consider the issue of the nonuniqueness of projections. If we are in $\mathbf{R}^2$—the Euclidean plane—and we wish to find the projection of a point $\mathbf{z}$ onto a curve $\mathscr{C}$, such as a parabola or hyperbola or circle, then there may be multiple points of minimum distance from $\mathbf{z}$ on the given curve. To take an extreme, but instructive, case, if the curve $\mathscr{C}$ is the unit circle and $\mathbf{z}$ is the center of the circle, then it is equidistant from *all* the points of the circle. However, in every other case in this same example, there is a unique projection of $\mathbf{z}$. For any point $\mathbf{z}$ that is outside $\mathscr{C}$, this is clear. On the other hand, if $\mathbf{z}$ is inside the circle, the set of points that have a fixed distance c from $\mathbf{z}$ lie on a circle $\mathscr{S}_c$ of radius c. See Figure 4.10. As c grows, we

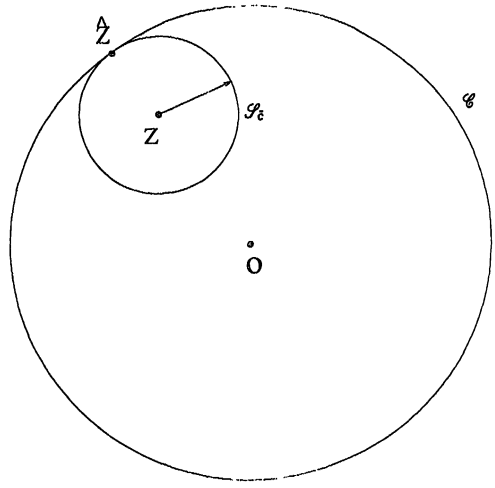

FIG. 4.10. *Projections onto a circle.*

can visualize that the minimum distance point on $\mathscr{C}$ corresponds to the first intersection point of $\mathscr{C}$ and $\mathscr{S}_{\tilde{c}}$, where $\tilde{c}$ is the minimum possible distance. This intersection point is unique due to the greater curvature of $\mathscr{S}_{\tilde{c}}$ than $\mathscr{C}$. Thus the curvature of the surfaces involved play a critical role in uniqueness considerations, as well as the multimodality of the distance function.

Although the projections $\hat{\mathbf{z}}$ on $\mathscr{C}$, may not be unique, there are still some useful facts available if $\mathscr{C}$ is a closed cone.

1. The distance function is continuous, so for closed surfaces (containing their limit points), there does exist a well defined minimum distance to the cone $\|\mathbf{z} - \hat{\mathbf{z}}\|$ that is attained for some point $\hat{\mathbf{z}}$.
2. Since we are projecting onto a *cone*, whatever solution we find, say $\hat{\mathbf{z}}$, is orthogonal to $\mathbf{z} - \hat{\mathbf{z}} := \mathbf{e}$, so we still have the basic orthogonal decomposition into fitted values and residuals,

$$\mathbf{z} = \hat{\mathbf{z}} + \mathbf{e},$$

with the accompanying sums of squares decomposition, even if we no longer have uniqueness for the vectors in this decomposition or a dual projection interpretation of the residual $\mathbf{e}$.

To visualize the concepts in the second point, imagine that the cone consists of two rays from the origin in $\mathbf{R}^2$. Then when $\mathbf{z}$ is between the two rays, it can be projected onto either ray to find a point of local minimum distance, and when $\mathbf{z}$ is exactly midway between, the two local minimum distances are equal. See Figure 4.11. However, in both cases $\hat{\mathbf{z}}$ is orthogonal to $\mathbf{e}$. Because of this, we may continue to use the relationship

$$\|\mathbf{z}\|^2 = \|\hat{\mathbf{z}}\|^2 + \|\mathbf{e}\|^2. \tag{4.14}$$

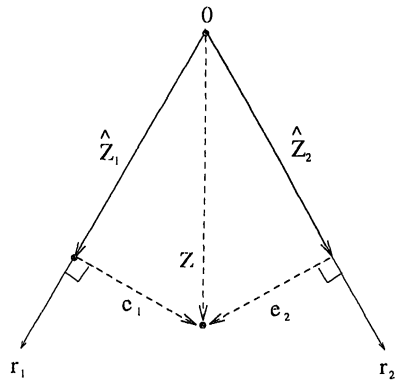

FIG. 4.11. *Nonunique projections onto two rays, with unique lengths.*

Another useful fact is that we can write an explicit formula for $\|\hat{\mathbf{z}}\|^2$ based on the elements of $\mathscr{C}$. We claim that

$$\|\hat{\mathbf{z}}\| = \sup\left\{\frac{(\mathbf{z}\cdot\mathbf{m})^+}{\|\mathbf{m}\|}:\ \mathbf{m}\in\mathscr{C}\right\}. \tag{4.15}$$

The argument goes as follows: View each ray $\{c\mathbf{m}:\ c\geq 0\}$ as a cone on which we do a projection, arriving at $\hat{\mathbf{z}}$ and $\mathbf{e}$ that depend on $\mathbf{m}$. If we do so, we find that the length of "$\hat{\mathbf{z}}$" is $(\mathbf{z}\cdot\mathbf{m})^+/\|\mathbf{m}\|$. Now, for each ray $\mathbf{m}$ the decomposition (4.14) holds, and our goal in projection is to minimize the residual term $\|\mathbf{e}\|^2$. However, because $\|\mathbf{z}\|^2$ is fixed, we can equivalently maximize $\|\hat{\mathbf{z}}\|^2$. The result (4.15) now follows.

4.6.2. *Measuring distances.* Another important issue arises when the model's corrected directional score cone is not convex. We will need to determine statistically appropriate ways to measure distances along the cone. An important parameter in our statistical analysis will be an arc length distance along the unit sphere in the appropriate metric. We will derive it here, in advance of the main result.

We set up the appropriate geometry. In the preceding section, we transformed from the E_0 geometry appropriate to the scores into a z-coordinate system, because in this space everything could be recognized as the projection of standard normal variables onto convex cones. The arguments could have been worked out directly in the dagger space, but the sense of simplification would have been lost. We presume we are working in such a transformed space, and if we refer to a particular score functions, we are referring to their coordinate representation.

If we imagine our (transformed) corrected score cone, with the half-tepee shape, with apex at the origin, then it intersects the unit sphere in a one-dimensional curve Γ on the surface of the sphere. (The unit sphere in the E_0 geometry consists of mean zero variables of variance 1.)

If the cone is actually *flat*, so that it is the positive cone of its two extremal vectors, then Γ is a *great circle*, the shortest path on the sphere between its two endpoints. Moreover, in this case the *length of this path* between the two endpoints equals the angle α between the two extremal rays at the origin, because we are merely walking along the rim of a circle that connects these rays.

However, if the tangent cone is actually the positive cone generated by the two extremal scores, then we are in the setting of the type II likelihood ratio test and the results of the preceding section. Our problem arises due to the fact that Γ is not a great circle, reflecting the curvature of the tangent surface generated by the gradient scores. The points of Γ are the coordinates of the normalized (to variance 1) corrected scores

$$g_\phi(t) := \tilde{s}_\phi(t)/\sqrt{E[\tilde{s}_\phi^2(X)]}.$$

As ϕ varies, $\mathbf{g}_\phi$ traces out the curve Γ on the unit sphere. See Figure 4.12.

We will need the length of this curve in the z-coordinate sense. To solve this, we transform back into the original coordinates and find the E_0 length around the sphere. To do this, one can break the parameter space into a grid of intervals, say $\phi_i < \phi_{i+1}$, and sum the secant distances

$$d_i = \sqrt{E[g_{\phi_{i+1}}(X) - g_{\phi_i}(X)]^2}$$

between neighbors $\mathbf{g}_{\phi_{i+1}}$ and $\mathbf{g}_{\phi_i}$ to arrive at the appropriate approximating sum

$$\text{arc length} \approx \sum_i \sqrt{E[g_{\phi_{i+1}}(X) - g_{\phi_i}(X)]^2}.$$

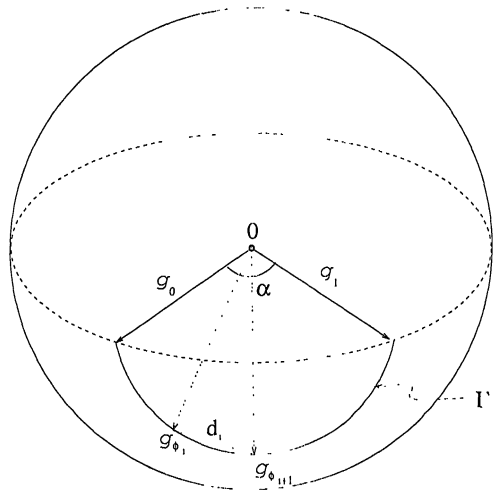

FIG. 4.12. *The trace of the curve* Γ *on the unit sphere.*

If we let the partition of the parameter space grow finer, we see that this approximation converges to

$$\text{Arc}(\Gamma) = \text{arc length} = \int \sqrt{E\left[\frac{\partial}{\partial\phi} g_\phi(X)\right]^2}\, d\phi. \tag{4.16}$$

[If we calculate the arc length using a partition approximation, somewhat greater accuracy can be obtained by summing the arc lengths α_i on the sphere rather than secant distances d_i, using the relationship $d_i^2 = 2 - 2\cos(\alpha_i)$.]

We have carried out these calculations, using the grid approximation, for several binomial models. In Table 4.1, we show these calculations. The parameter ϕ is the natural parameter; that is, the log odds. For comparison, we have shown the great circle distance between the two extremal rays of the cone so as to show that the statistical curvature in the problem can significantly increase the surface of the tangent score cone. In the binomial model, the arc length goes to infinity as $N \to \infty$. In the next section we will find how the arc length shows up in the limiting distribution.

4.6.3. *Tubes and distributions.* Now we tackle the rather severe distributional problem. Our first reduction is to turn this into a problem involving the calculation of surface areas on the unit sphere. The random variable $U = \|\mathbf{e}\|/R = \|\mathbf{e}\|/\|\mathbf{z}\|$ is scale invariant and so is statistically independent of R by Basu's theorem. Thus we can find its distribution by calculation of its conditional distribution given any fixed value of R, which we will take to be $R = 1$. However, the statistic whose distribution we desire is $\|\hat{\mathbf{z}}\|^2 = (1 - U^2)R^2$, so if we calculate the distribution of U, then we can use independence and the known chi-squared distribution of R^2 to find the desired distribution.

Now we use the fact that $\mathbf{Z}$ is, conditionally on $R = 1$, uniformly distributed on the unit sphere to note that the probability $\Pr[U \leq u]$ is the volume of a *tube* about the curve Γ. In Figure 4.13 we have attempted to recreate the geometry of the situation. If we let θ be in $[0, \pi/2]$ and consider all points that are within arc length θ of a point $\mathbf{g}_\phi$ on Γ, then we have a spherical cap at that point. Points $\mathbf{z}$ that are in that cap have a projection distance $\|\mathbf{e}\|$ to the cone no larger than $\sin\theta$.

TABLE 4.1

N	ϕ	Arc/2π	$\alpha/2\pi$
3	0	0.167	0.167
3	2	0.193	0.193
3	4	0.226	0.226
4	0	0.240	0.206
6	0	0.363	0.236
20	0	0.887	0.250

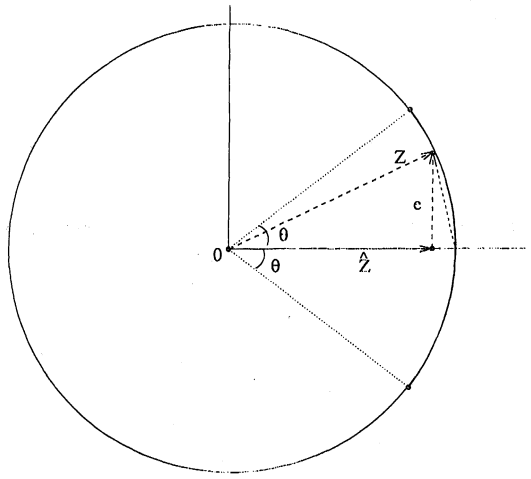

FIG. 4.13. *A cross section of the spherical cap with angular radius* θ.

Thus if we find the spherical volume of the set of all points within θ in arc length from Γ, and divide it by the total surface "volume" of the sphere, we will have the probability that $U \leq \sin(\theta)$.

The set $\mathcal{T}_\theta := \{\mathbf{z}\colon U \leq \sin\theta, R^2 = 1\}$ is the *tube of radius* θ *about the curve* Γ. It is sketched in Figure 4.14, showing in particular that the two endpoints of Γ generate two semispherical caps.

We first carry out the calculation under the simplifying assumption that the curve Γ lies on a great circle (geodesic). We assume that the angle α it subtends is less than π, so that curve wraps no more than half way about the sphere. We let $V_1, \ldots, V_d$ be the uniformly distributed coordinates of the sphere and we suppose that we have rotated the sphere about so that the great circle is

$$\{(v_1, v_2, 0, \ldots, 0)\colon v_1^2 + v_2^2 = 1\}.$$

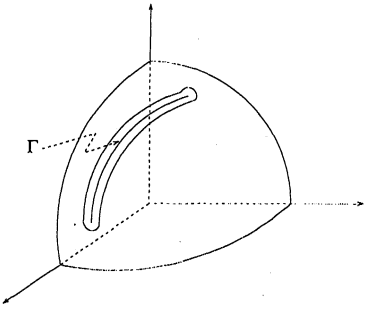

FIG. 4.14. *A tube of radius* θ *about the curve* Γ.

In this situation, we have already calculated the probability distribution for the likelihood ratio statistic, because we are in the setting of the type II likelihood ratio test, with the geometric structure of Example 12. The corrected score tangent cone is generated as the positive cone of the two extremal rays of the curve Γ, so it lies in the space $\{(z_1, z_2, 0, \ldots, 0)\colon\ z_1 \in R,\ z_2 \in R\}$ and so we can ignore the orthogonal variables $Z_3, \ldots, Z_d$, here corresponding to the spherical variables $V_3, \ldots, V_d$.

We rederive the distribution for Example 12 by starting with the probability that the uniform sphere variable $\mathbf{V}$ falls in the tube. Provided that the angle θ is in the range $[0, \pi/2)$, so that the tube does not overlap itself, an elementary geometric argument shows that

$$\Pr[U < \sin\theta] = \Pr[V_1 > \cos\theta] + \frac{\alpha}{2\pi}\Pr\left[\sqrt{V_1^2 + V_2^2} > \cos\theta\right]. \tag{4.17}$$

The first term in this expression is the probability of falling in the two endcaps, which together make one complete semispherical cap. [Note that here V_1 refers not to the first coordinate, but the distribution of the first coordinate]. The second term comes from the body of the tube, excluding the endcaps. By substituting $w = 1 - \sin^2\theta = \cos^2\theta$, this last formula becomes

$$\Pr[1 - U^2 > w] = \frac{1}{2}\Pr[V_1^2 > w] + \frac{\alpha}{2\pi}\Pr[V_1^2 + V_2^2 > w], \tag{4.18}$$

where the factor $1/2$ arises from the symmetry of V_1's distribution about zero. From here, we may finish the calculation of the distribution by noting that $R^2 \cdot (V_1^2 + V_2^2)$ has a χ_2^2 distribution and $R^2 \cdot V_1^2$ has a χ_1^2 distribution, so an easy *exercise* shows

$$\Pr[\|\hat{\mathbf{Z}}\|^2 > t] = \frac{1}{2}\Pr[\chi_1^2 > t] + \frac{\alpha}{2\pi}\Pr[\chi_2^2 > t], \tag{4.19}$$

exactly in accordance with our earlier derivation of the type II likelihood ratio test in Example 12.

4.6.4. *Approximations for tubes.* Hotelling (1939) showed that the tube formula in (4.17) is still an equality if the curve is not a great circle, provided that the curve is regular, and the chosen angle θ is sufficiently small and one replaces α with the arc length $A(\Gamma)$ of the curve. However, equality does fail for large values of θ when the tube displays curvature, that is, when Γ is not a great circle curve, and this failure occurs when θ exceeds the smallest value of the spherical radius of curvature of Γ.

However, Naiman (1986) has shown that even when equality fails to hold, the two sides of (4.17) are always related by an inequality of the form $\leq$. Heuristically, the inequality arises because when θ grows sufficiently large, the curvature of Γ causes a kink in the tube, in which case it has *less* volume than predicted by the formula on the right-hand side.

This said, we will say that we have a *type III likelihood ratio problem* if the tangent cone of the corrected scores is two dimensional and the corresponding curve Γ is smooth, with no points of self-intersection.

PROPOSITION 17. *In any type III likelihood ratio problem with finite arc length* $A = A(\Gamma)$, *the limiting distribution of* $\|\hat{\mathbf{Z}}\|^2$ *satisfies*

$$\Pr[\|\hat{\mathbf{Z}}\|^2 > t] \leq 0.5 \Pr[\chi_1^2 > t] + \frac{A}{2\pi} \Pr[\chi_2^2 > t].$$

Moreover, the ratio of the two sides goes to 1 *as* $t \to \infty$, *provided that the curve* Γ *has spherical curvature that is bounded above.*

PROOF. To prove the inequality, we apply Naiman's result to (4.17) and follow its consequences through (4.18) and (4.19). For the second part, we note that as long as θ is sufficiently small that the tube inequality is an equality, the first bound is an equality; hence, so is the second upper bound for w sufficiently close to 1. This implies that the upper tail of $\|\hat{\mathbf{Z}}\|^2 = R^2(1 - U^2)$, which is determined by large values of R and $W := (1 - U^2)$, is asymptotic to the bound, by the following line of argument. Let G be the distribution function for W and let a be such that the bound is exact for $W \geq a$. We can show that the tail bound is exact if the following ratio converges to zero as $t \to \infty$:

$$\frac{\int_0^a \Pr\{R^2 > t/w\}\, dG(w)}{\int_a^1 \Pr\{R^2 > t/w\}\, dG(w)} \leq \frac{\int_0^a \Pr\{R^2 > t/w\}\, dG(w)}{[1 - G(a)] \Pr\{R^2 > t/a\}}.$$

Thus it suffices to show that

$$\frac{\Pr\{R^2 > t/w\}}{\Pr\{R^2 > t/a\}} \to 0,$$

for $w < a$, since the ratio being less than 1 implies dominated convergence. However, this last statement is true by an application of l'Hôpital's rule and the use of the appropriate χ^2 density. □

Although it is perhaps possible to get more accurate descriptions of the limiting distribution theory, it is unlikely that it can be done without considerably more effort. However, such a geometric study will no doubt have further payoffs in understanding the nature of the problem of multimodal likelihoods. Note also that we can accommodate multinomial models with auxiliary parameters in this analysis; they will show up as nuisance scores in the calculation of the corrected mixture scores. Finally, we note that the limiting distribution may not be operative if the sample size is small relative to the parameter space, as we discuss in the next section.

4.6.5. *The arc length problem.* Despite the fact that there is much work to be done to turn the preceding theory into a viable strategy in the wide range of strictly multinomial problems that it could be applied to, it is hard to avoid the temptation to turn it to use in understanding the much studied mixture problem of testing two normal components. We will consider here just

the case of two normal components with common known variance, because the unknown variance case considerably changes the geometry.

To enhance our understanding, we imagine that the problem has been discretized, so that we are in a situation similar to a Bin(N, p) with N large. There are several important features of this problem that we can now see more clearly.

First, as the binomial parameter N increases, the arc length parameter grows without bound, increasing the weight applied to the χ^2_2 component in the approximation. Thus the inequality says that the likelihood ratio test is gaining heavier and heavier tails. If one carries out the formal calculation of the arc length using (4.16) in the normal model, one finds in fact that it becomes infinite when integrated over the parameter space.

This calculation confirms Hartigan's conclusion that the likelihood ratio test statistic diverges to infinity. However, in addition to not being useful, this result appears in contradiction to the great stability that has been found in many simulation studies. Indeed, the simulation study of Böhning, Dietz, Schaub, Schlattman and Lindsay (1994) shown in Figure 4.15 shows that for a sample of size 10,000, the distribution appears very much like the chi-bar-squared approximation, with some small finite arc length. (*Note:* The plot is of the conditional CDF of the nonzero statistic values only, because our right tail approximation gives no information about the probability of a zero.)

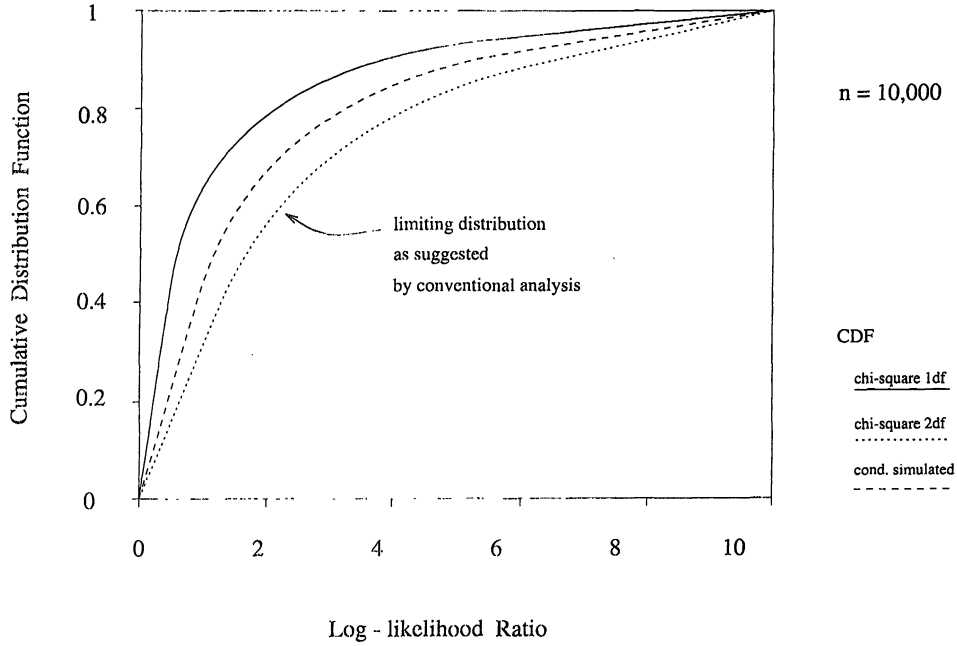

FIG. 4.15. *A simulation study of the likelihood ratio test distribution.*

To explain this phenomenon, suppose we were to restrict the search for a two-component alternative to a finite interval of parameter values, say $\phi \in [L, R]$. Then the theory of the preceding section would lead us to identical conclusions, but the arc length would be calculated over the restricted interval; call this length Arc$[L, R]$.

The importance of this is that in practice, for any finite sample size, we do not need to search for the two-component solution over an infinite range. For example, in a two-component normal mixture problem, there can be no likelihood solutions for which the support points are not between the smallest and largest order statistic of the data, $y_{1:n}$ and $y_{n:n}$. Thus a more relevant calculation would seem to be the arc length $A_n = A[L_n, U_n]$ corresponding to $L_n = E[Y_{1:n}]$ and $U_n = E[Y_{n:n}]$. In the normal model, this *effective arc length* grows very slowly in n. Some values are given in Table 4.2.

Although this gives us some further guidelines as to what to expect in a distributional theory, there are also some technical problems with the asymptotics that one should consider. The distributional theory relies on the corrected scores being asymptotically normally distributed. However, in the mixture case, the gradient scores $\tilde{s}_\phi$ can be a long way from achieving their limiting distribution for any finite sample size. Consider the normal case, where the selected null model is $N(0, 1)$:

$$s_\phi(t) = \frac{\exp(-0.5(t-\phi)^2)}{\exp(-0.5t^2)} - 1 = \exp(\phi t - 0.5\phi^2) - 1.$$

If ϕ is very large, say 100,000, then we can expect $s_\phi(X_i)$ to be nearly -1 across all observations in any reasonable sample size from a standard normal, even though the statistic is nominally mean zero. (The explanation arises from the enormous variance.) Thus the sum really behaves more like a Poisson variate than a normal because the finite sample behavior is determined by rare events, namely, very large observations from the normal distribution.

4.6.6. *Final comments.* As we indicated earlier, a pragmatic strategy for many problems is simply to use the Neyman dispersion score approach and use the resulting asymptotic normality to generate a simple procedure. If one desires more power against a wider class of alternatives, one must consider the likelihood ratio test as the method of choice. However, we do note that in the process of our analysis, we also derived the appropriate score function test for testing one versus two components, as we now argue.

TABLE 4.2

n	$A_n/2\pi$
44	0.44
740	0.75
31,000	1.07
3,500,000	1.40

Recall from (4.13) that inner product of the directional scores $\mathbf{s}_\phi$ from null value ϕ_0 with the data vector resulted in the normalized gradient $n^{-1}D_{\phi_0}(\phi)$. If we transform the projection statistic $\|\hat{\mathbf{z}}\|^2$ from (4.15) back into the original model geometry, we see that the projection statistic, and therefore the likelihood ratio statistic, is asymptotically equivalent to the *generalized score statistic:*

$$U^2(\phi_0) := n^{-1/2} \sup_\phi \left\{ \frac{D^+_{\phi_0}(\phi)}{\|\tilde{\mathbf{s}}_\phi\|} \right\}^2.$$

This statistic may be estimated under the null hypothesis by $U^2(\hat{\phi})$. This statistic will have the same asymptotic distribution as the likelihood ratio test and is considerably easier to compute provided that the corrected score variances can be calculated explicitly. We note that the argument of the supremum

$$\frac{D^+_{\phi_0}(\phi)}{\|\tilde{\mathbf{s}}_\phi\|}$$

approaches the positive part of the Neyman dispersion test statistic as $\phi \to \phi_0$, so the generalized score test clearly uses wider properties of the gradient than the Neyman test.

There is also a score test corresponding to the likelihood ratio test for one component against an arbitrary number of components, but it is considerably more complicated to compute.

Finally, we note that for both the likelihood ratio test and the generalized score test, the arc length problem goes away provided that one is willing to group data in the tails of the distribution. In a standard contingency table analysis one can group cells together so as to improve asymptotic approximations, and one can do so here. One can either group the data throughout its range, making sure the bins in the tails have sufficient observations, or one can construct a likelihood from the densities in the middle, but use the appropriate distribution functions in the tails. In either case, the arc length calculations will no longer be infinite, at any finite sample size, and the tail scores functions will be more nearly normal, justifying the normal approximations used in constructing the approximate distribution.