Søren Johansen   Iain Johnstone

Some Uses of Spherical Geometry in Simultaneous Inference and Data Analysis

Søren Johansen   and   Iain Johnstone*

SOME USES OF SPHERICAL GEOMETRY IN

SIMULTANEOUS INFERENCE AND DATA ANALYSIS

SOME USES OF SPHERICAL GEOMETRY IN SIMULTANEOUS INFERENCE AND DATA ANALYSIS.

Søren Johansen
Iain Johnstone

University of Copenhagen
Stanford University

April 1985

Abstract

We illustrate by contemporary examples the continued value of Hotelling's (1939) geometric approach to simultaneous probability calculations. Hotelling shows how to reduce the calculation of certain normal theory significance probabilities to finding the volume of a tube about a curve in a d-sphere and shows that this is often exactly given by length times cross sectional area. We illustrate this point of view by construction of confidence and prediction bands in curvilinear regression, Andrews plots in multivariate data analysis, inference problems for projection pursuit regression and Tukey–Kramer confidence intervals in the analysis of variance.

§1. Introduction

The purpose of this largely theorem-free paper is to present some examples in simultaneous inference and data analysis for which the spherical geometric viewpoint of Hotelling (1939) continues to be of use. In this now somewhat neglected article (although see Diaconis and Efron (1985)) Hotelling shows that the volume of a tube lying about a curve in a hypersphere is exactly its length multiplied by its cross-sectional area (ignoring overlap) and applies this to the computation of significance probabilities for tests based on the multivariate normal distribution.

We consider principally four situations: simultaneous confidence and prediction bands in regression. Andrews' plots in multivariate data analysis, significance tests in projection pursuit regression, and Tukey-Kramer confidence intervals in pairwise comparisons of means.

Firstly, it is shown that the problem of simultaneous confidence bands for the general linear model with Gaussian errors, and the problem of simultaneous prediction regions in Gaussian random coefficient models can be given a common formulation which allows one to apply Hotelling's result (via an important complement due to Naiman (1985)) to get a conservative solution and via a conditioning argument to give an exact solution for sufficiently wide bands. A stimulus for this paper was work of Olshen (1985) on the application of trigonometric regression to the analysis of gait in children (Sutherland et al,(1980)), and for some trigonometric regression situations we investigate (in Section 4) the accuracy of the Hotelling prediction region solution. This approach may be used (though we do not do so here) to reproduce theoretically results derived empirically from bootstrap resampling by Olshen.

Secondly, we study Andrews' (1972) device for representing points in high dimensional space R^d by trigonometric polynomials whose coefficients are the coordinates of the corresponding original data points. Thought of as a projection pursuit method, the plot traces out a curve of projection directions on S^{d-1}. We compute the fraction of possible projections that are seen, assuming a given "squint angle" for the data analyst, and obtain serviceable bounds on the distance of the furthest (unseen) projection from the curve.

Thirdly, an outline of an approach to significance tests in an idealized projection pursuit regression setting is given. A thorny issue with projection pursuit methods is to assess the magnitude of the selection effect implied by the search over many directions. (See, for example, Miller's discussion to Huber (1985)). In our idealized model of orthogonal polynomial regression, computation of a significance probability for the fitted terms of a given degree can be expressed (approximately) in terms of the closest direction from a curve (or surface) in a hypersphere to a Gaussian vector, so that the Hotelling's method applies.

Finally, in an expository section, we consider the problem of assigning simultaneous confidence intervals to all pairwise differences of means in an unbalanced one way analysis of variance. The long open conjecture of Tukey (1953) and Kramer (1956), settled in a difficult proof by Hayter (1984), asserted that intervals based on the studentized range for the balanced

case yield a conservative solution. The problem is formulated here in terms of a (constrained) positioning of caps on the surface of a sphere so as to maximize the area of the union. This suggests quite directly and intuitively the optimal configuration, and for illustration, we complete the proof in the relatively simple case of three populations.

§2. Simultaneous confidence and prediction bands

Let X be distributed as $N_d(\xi, \Gamma)$ and let $K \subset R^d$ denote a set of vectors specifying linear combinations of interest to us. We want to make simultaneous confidence statements about $\{\lambda'\xi, \lambda \in K\}$ and form prediction sets for the random variables $\{\lambda'X, \lambda \in K\}$.

In either case we start from the random variable

$$(2.1) \qquad T = T(X, \xi) = \sup_{\lambda \in K} \frac{\lambda'(X - \xi)}{(\lambda'\Gamma\lambda)^{1/2}}.$$

If we can find the $P_{\xi,\Gamma}$ distribution of T we can construct a $1 - \varepsilon$ confidence set C_X as follows:

$$C_X = [\{\lambda'\xi\}_{\lambda \in K} | T(X, \xi) \leq z_{1-\varepsilon}]$$

where $z_{1-\varepsilon}$ is the $1 - \varepsilon$ quantile in the distribution of T.

It is easily seen that the random set C_X covers the point $\{\lambda'\xi\}_{\lambda \in K}$ with $P_{\xi,\Gamma}$ probability $1 - \varepsilon$. Similarly we can construct a $1 - \varepsilon$ prediction set R_ξ, by

$$R_\xi = [\{\lambda'X\}_{\lambda \in K} | T(X, \xi) \leq z_{1-\varepsilon}]$$

and it follows easily that the random point $\{\lambda'X\}_{\lambda \in K}$ is contained in the set R_ξ with $P_{\xi,\Gamma}$ probability $(1 - \varepsilon)$.

We shall now show how the result of Hotelling (1939) provides a way to discuss the distribution of T.

The variable T given by (2.1) can be decomposed as follows:

$$T = WS$$

where $S^2 = (X - \xi)'\Gamma^{-1}(X - \xi)$ and

$$\begin{aligned}
W &= \sup_{\lambda \in K} \frac{\lambda'(X - \xi)}{(\lambda'\Gamma\lambda)^{1/2}((X - \xi)'\Gamma^{-1}(X - \xi))^{1/2}} \\
&= \sup_{\lambda \in K} \frac{(\Gamma^{1/2}\lambda)'\Gamma^{-1/2}(X - \xi)}{|\Gamma^{1/2}\lambda||\Gamma^{-1/2}(X - \xi)|}.
\end{aligned}$$

If we define

$$\gamma(\lambda) = \Gamma^{1/2}\lambda/|\Gamma^{1/2}\lambda|$$

and

$$U = \Gamma^{-1/2}(X - \xi)/|\Gamma^{-1/2}(X - \xi)|,$$

then U is uniformly distributed on ∂S^d, the surface of the sphere of radius 1 in d dimensions.

Thus T is decomposed as a product of two independent variables, and we only have to find the distribution of W, since that of S^2 is just a χ^2 distribution with d degrees of freedom. Now

$$W = \sup_{\lambda \in K} \gamma(\lambda)'U = \sup_{\gamma \in \gamma K} \gamma'U$$

and the distance $d(\gamma, U)$ satisfies

$$d^2(\gamma, U) = |\gamma|^2 + |U|^2 - 2\gamma'U = 2(1 - \gamma'U)$$

which shows that

$$\inf_{\gamma \in \gamma(K)} d(\gamma, U)^2 = 2(1 - W).$$

Thus W measures the cosine of the angle between U and the point $\hat{\gamma} \in \gamma(K)$ which is closest to U. On the other hand the set $\{W \geq w\}$ is the set of points on ∂S^d which are within a distance $(2(1-w))^{1/2}$ of $\gamma(K)$. This set is called the tube around $\gamma(K)$ of radius $(2(1-w))^{1/2}$ or angular radius θ where $\cos(\theta) = w$.
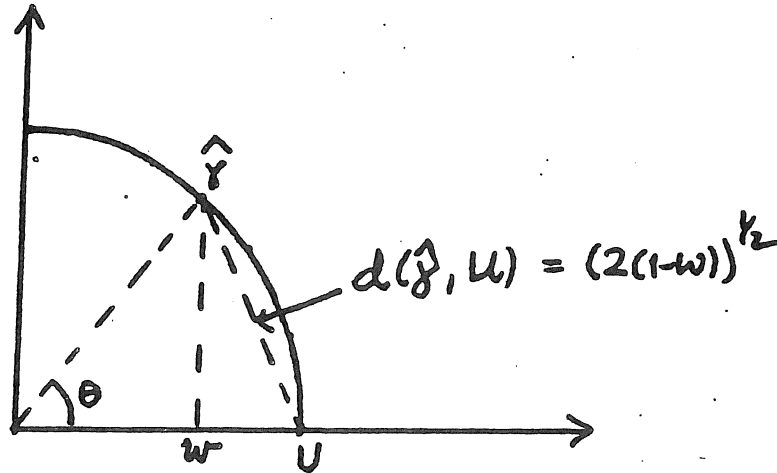

**Figure 2.1. Illustration of the relation
between θ, w and $d(\hat{\gamma}, U)$.**

We can then state the main result of this section

THEOREM 2.1 (Hotelling). Let $\gamma(K)$ be a twice continuously differentiable and closed curve in ∂S^d with length $|\gamma(K)|$. Then

(2.2) $$P\{W \geq w\} = \frac{|\gamma(K)|}{2\pi}(1 - w^2)^{(d-1)/2}$$

if there is no local self overlap of the tube, i.e. if $\sin(\theta) = (1 - w^2)^{1/2} < \rho$, where ρ is the minimal radius of curvature of $\gamma(K)$ considered as a subset of R^d, and if there is no global

overlap of the tube.

<u>Proof</u> Hotelling proved that under the conditions stated the volume of the tube around $\gamma(K)$ of radius $(2(1-w))^{1/2}$ is given by

$$|\gamma(K)| \ \mathrm{Vol}[S^{d-2}(\{1-w^2\}^{1/2})]$$

corresponding to the length of the curve times the volume of a sphere of dimension $d-2$ and radius $(1-w^2)^{1/2} = \sin(\theta)$. If we divide by $\mathrm{Vol}(\partial S^d(1))$ we obtain (2.2). Note that $\mathrm{Vol}(S^d(r)) = \pi^{d/2}r^d/\Gamma(1+d/2)$ and $\mathrm{Vol}(\partial S^d(r)) = 2\pi^{d/2}r^{d-1}/\Gamma(d/2)$. Note also that the expression (2.2) is exact for all w if $\gamma(K)$ is a great circle.

For the applications that we are interested in it is clear that one cannot in general avoid local and global overlap of the tube, since we want w to range from 1 to 0, thereby expanding the tube to cover all of ∂S^d. Thus the right hand side can be used as an approximation to the required probability, and more detailed numerical or analytic studies are required to check the quality of the approximation. Since the curve γ corresponding to trigonometric regression is relevant to Olshen's example (discussed below) and to the discussion of Andrews' plots, we analyse the quality of the approximation for this case in Section 4.

If the curve is not closed the tube will contain a half sphere of dimension $d-1$ and radius $(2(1-w^2))^{1/2}$ at each end and the right hand side needs an extra term

$$(2.3) \qquad \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_w^1 (1-z^2)^{(d-2)/2} dz.$$

Naiman (1985) proved that with this extra term added the right hand side of (2.2) is an upper bound on the probability.

The above general formulation clearly also covers the case of multivariate regression, but in this case $\gamma(K)$ is not a curve but a surface with dimension equal to the number of regressors. Thus the volume we need to find is the volume of all points within a certain distance of a submanifold of ∂S^d. An expression for this has been given by Weyl (1939), and this allows one to derive similar but more complicated results for multivariate regression.

If the tube is sufficiently narrow, i.e. $W \geq w_0$ say, then the result (2.2) holds exactly. The problem of simultaneous confidence or prediction bands require the distribution of $T = WS$. Thus we cannot in general find exact expressions for the distribution of T. What can be done, however, is to change the procedure slightly as follows: Instead of just predicting $\{\lambda'X\}_{\lambda \in K}$ let us first predict the length of X, as measured by $S^2 = (X-\xi)'\Gamma^{-1}(X-\xi)$ and then conditionally on this outcome let us construct the prediction set for $\{\lambda'X\}_{\lambda \in K}$ using T. For fixed S, the distribution of T is just that of a scale transform of W. Thus we need the tail probabilities of W, which are precisely those that can be found from the result (2.2).

Remark: The C_X, R_ξ can also be considered in parameter space and observation space R^d as follows: For each $\lambda \in K$ define the half space $H_\lambda \subset R^d$ by

$$z \in H_\lambda \leftrightarrow \frac{\lambda' z}{(\lambda' \Gamma \lambda)^{1/2}} \leq z_{1-\varepsilon}$$

and let

$$M = \bigcap_{\lambda \in K} H_\lambda.$$

Then M is a convex and closed set, and

$$\lambda' \xi\}_{\lambda \in K} \in C_X \leftrightarrow \frac{\lambda'(X - \xi)}{(\lambda' \Gamma \lambda)^{1/2}} \leq z_{1-\varepsilon}, \quad \forall \lambda \in K$$
$$\leftrightarrow \xi \in X - M.$$

Similarly

$$\{\lambda' X\}_{\lambda \in K} \in R_\xi \leftrightarrow X \in \xi + M.$$

The convex region M clearly contains the ellipsoid

$$E = \bigcap_{\lambda \in R^d} H_\lambda = \{z | z' \Gamma^{-1} z \leq z_{1-\varepsilon}\}$$

which is proportional to the set one would get by considering all linear combinations in R^d.

It is certainly possible for the convex region M to be substantially larger than the ellipsoid E. Thus, for example, the confidence band for a simple linear regression line over a small interval in R will contain lines whose coefficients are substantially different from those of the estimated line and which would certainly be rejected on the basis of, say, a Hotelling T^2 test on the coefficients. If we are interested in diagnosing departure from a given model via the use of prediction bands, it may be reasonable to first conduct a comparison on the coefficients, using the T^2-like statistic S, before examining whether the observed curve lies within the prediction band. This leads to the procedure suggested above.

We shall illustrate the above results with a random coefficient regression model where the purpose is to compare a new individual with the prediction region derived from a population of control individuals.

Thus let us consider the following trigonometric regression model as suggested by Olshen (1985).

Let the observations $Y_1, \ldots, Y_n$ be independent and distributed as the vector with components

$$Y_i = A_0 + \sum_{j=1}^{k} [A_j \cos(j\theta_i) + B_j \sin(j\theta_i)] + V_i, \quad i = 1, \ldots, m$$

where $\theta_i = 2\pi(i - 1)/m$, and V_i are independent and distributed as $N(0, \sigma^2)$, whereas the vector $(A_0, A_1, \ldots, A_k, B_1, \ldots, B_k)$ is distributed as $N_{2k+1}(\xi, \Gamma)$ with $\xi' = (\alpha_0, \alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k)$.

If for each individual we fit the trigonometric regression curve by least squares, we get estimates $X_j = (\hat{A}_{0j}, \hat{A}_{1j}, \ldots \hat{A}_{kj}, \hat{B}_{1j}, \ldots, \hat{B}_{kj})'$ with distribution $N_{2k+1}(\xi, \Gamma + \sigma^2 M^{-1})$ where M is a diagonal matrix.

The prediction region for the estimated coefficients for a new individual will be derived from

$$S^2 = (X - \xi)'(\Gamma + \sigma^2 M^{-1})^{-1}(X - \xi)$$

and if we are interested in the values of the curve

$$A_0 + \sum_{j=1}^{k}[A_j \cos(j\theta) + B_j \sin(j\theta)]$$

then we define the set of combination vectors

$$K_0 = [1, \{\cos(j\theta), \sin(j\theta)\}_{j=1}^{k}, 0 \le \theta < 2\pi]$$

and

$$K = K_0 \cup (-K_0).$$

Then the statistic

$$T(X, \xi) = \sup_{\lambda \in K} \frac{\lambda'(X - \xi)}{(\lambda'(\Gamma + \sigma^2 M^{-1})\lambda)^{1/2}}$$

becomes

$$\sup_{\theta} \left| \frac{A_0 - \alpha_0 + \sum_{j=1}^{k}[(A_j - \alpha_j)\cos(j\theta) + (B_j - \beta_j)\sin(j\theta)]}{(\text{Var (Numerator)})^{1/2}} \right|.$$

In practice we shall have to estimate (ξ, Γ, σ^2) from the control population, and insert the estimates into S and T, but we shall not pursue this point here. Rather, we emphasize that if each new individual is characterised by its Fourier coefficients $(A_0, A_1, \ldots, A_k, B_1, \ldots, B_k)$, it seems reasonable to include in the diagnostic of the new individual a comparison of these coefficients with those of the control group, as well as a comparison of the estimated curve with a prediction band for this curve, as derived from the control group. The conditional approach proposed above provides one convenient and distributionally tractable way to handle this.

It follows from the above analysis that this two stage procedure can be solved by the result of Hotelling. The calculations performed in Section 3 for the Andrews' plot are relevant here too, since the curve is the same. However, when the inner product, as given by $(\Gamma + \sigma^2 M^{-1})^{-1}$, is not a multiple of the identity matrix (as in Olshen's application), a separate analysis of when the formula (2.2) is exact is needed.

§3. Andrews' plots

Andrews (1972) proposed and discussed an interesting method of plotting high dimensional data. Each data point $x = (x_1, \ldots, x_d)$ is mapped into a trigonometric polynomial

$$f_x(\theta) = x_1/\sqrt{2} + x_2 \sin\theta + x_3 \cos\theta + x_4 \sin 2\theta + \cdots$$

containing d terms, and this polynomial is then plotted for $\theta \in C = [0, 2\pi)$. The mapping is an isometry of R^d onto a subspace of $L^2[0, 2\pi)$ spanned by $\{\theta \to \frac{1}{\sqrt{2}}, \theta \to \sin\theta, \theta \to \cos\theta, \ldots\}$, and so preserves means and interpoint distances if the squared distance between two points x and y on the plot is measured by

$$\frac{1}{\pi} \int_0^{2\pi} (f_x(\theta) - f_y(\theta))^2 \, d\theta.$$

For a fixed θ, Andrews notes that the collection of values $\{f_x(\theta)\}$ as x runs through the data is a projection of the original data onto the unit vector $\gamma(\theta) = w_d(\theta)/ \mid w_d(\theta) \mid$, where

$$w_d(\theta) = (\frac{1}{\sqrt{2}}, \sin\theta, \cos\theta, \cdots, \sin k\theta) \qquad \text{if } d = 2k \text{ is even}$$

$$= (\frac{1}{\sqrt{2}}, \sin\theta, \cos\theta, \ldots, \sin k\theta, \cos k\theta) \quad \text{if } d = 2k + 1 \text{ is odd.}$$

The Andrews' plot can be seen therefore as a projection pursuit method (e.g. Friedman and Tukey, (1974); Huber (1985)) as it is used to look for multivariate structure by searching amongst a (subset) of one-dimensional projections.

Projections of data onto unit vectors u and v that are close will produce similar results – thus it is not necessary (or possible) to look at all projections. Suppose that we deem it unnecessary to use projection directions v which make an angle less than φ with a chosen direction u (Huber (1985) terms this the "squint angle"). If we employ an Andrews' plot, the percentage of possible projections that we see for a squint angle φ is just the ratio of the volume of the tubes of radius $\sqrt{2(1 - \cos\varphi)}$ about $\pm\gamma(C)$ to the volume of ∂S^d. (Both $+u$ and $-u$ are counted since a projection in direction $-u$ is just a reflection of that in direction u).

In this section we study the curves $\gamma(C)$ (or a slight modification given in (3.2) for the case of d even) and the overlap properties of tubes surrounding them. Particular questions addressed include

(i) the tube radii at which global and local self-overlap occurs

(ii) the radii at which overlap with the antipodal tube about $-\gamma(C)$ occurs, and bounds for the overlap, and

(iii) determining bounds on the distance of the furthest projection from the curve corresponding to the Andrews' plot.

In addition to the information gained about Andrews' plots, the results are relevant to the trigonometric regression situation discussed in Section 2 in the case that $\Gamma = M = I$. Some of the methods can be extended to cover other curves, such as the correlated trigonometric regression case considered by Olshen.

Consider, for example, the curves $\gamma^{(1)}(\theta) = \frac{1}{\sqrt{k}}(\cos\theta, \sin\theta, \ldots, \cos k\theta, \sin k\theta)$ in S^{2k-1} and $\gamma^{(2)}(\theta) = \sqrt{\frac{2}{2k+1}}(\frac{1}{\sqrt{2}}, \cos\theta, \sin\theta, \ldots, \cos k\theta, \sin k\theta)$ in S^{2k}. (Our reasons for using $\gamma^{(1)}$ rather than $w_d(\theta)/\mid w_d(\theta)\mid$ for $d = 2k$ are discussed below.) Ignoring overlap (which we may for the angles listed, as is shown below) the percentage of projections seen for a given squint angle φ is given by

$$(3.1) \qquad 2P(W_0 \geq \cos\varphi) = 2[(2k^2 + 3k + 1)/6]^{\frac{1}{2}} \sin^{2k-2}\varphi \quad \text{and} \quad 2[(k^2 + k)/3]^{\frac{1}{2}} \sin^{2k-1}\varphi$$

respectively for $\gamma^{(1)}$ and $\gamma^{(2)}$.

	k	2	3	4	5	6
$2P(W_0 \geq \cos\varphi)$	5^0	.024	.0002		0	
	10^0	.095	.0039	.0002		
(for $\gamma^{(1)}$)	15^0	.2118	.0194	.0016	.0001	
	20^0	.3699	.0591	.0088	.0012	.0002
$2P(W_0 \geq \cos\varphi)$	5^0	.0019			0	
	10^0	.0148	.0006			
(for $\gamma^{(2)}$)	15^0	.0490	.0046	.0004		
	20^0	.1132	.0187	.0028	.0004	.0001

Table 3.1 Percentage of projections 'seen' using an Andrews' plot for the curves $\gamma^{(1)}(\theta), \gamma^{(2)}(\theta)$ in S^{2k-1}, S^{2k} respectively for squint angles $5°(5)20°$.

Remark: In even dimensions, the Andrews' plot based on $\gamma(\theta) = w_d(\theta)/\mid w_d(\theta)\mid$ suffers from the slight quirk that $\mid\dot\gamma(\theta)\mid, \mid\ddot\gamma(\theta)\mid$ and hence the curvature of the curve all depend somewhat on θ. An alternative and computationally simpler curve which we shall use when $d = 2k$ is

$$(3.2) \qquad\qquad \tilde{w}_d(\theta) = (\cos\theta, \sin\theta, \ldots, \cos k\theta, \sin k\theta)$$

which has constant length, speed, curvature, etc. As Andrews illustrates (with a more extreme example), use of projection directions based on higher frequency sinusoids allows the curve to

fill out more of the unit sphere, but also forces the picture to become more oscillatory and hard to interpret. For the case $k = 2$ we ran a crude simulation and found that as expected, $\tilde{S}$ is stochastically larger than S, with the Kolmogorov-Smirnov distance between the two empirical c.d.f.'s being as high as .26.

To begin the study of questions (i)–(iii), consider first a general smooth closed curve $\Gamma = \gamma(C)$ curve lying in S^{d-1}. In order to determine the range of squint angles for which Hotelling's result is exact, we wish to find the size of the tube about Γ for which global overlap first occurs. We say that global overlap occurs at a point ξ in the tube section about $\gamma_0 = \gamma(\theta_0)$ if there exists another value $\theta \in [0, 2\pi)$ for which $\gamma(\theta)'\xi > \gamma(\theta_o)'\xi$ or equivalently, $|\gamma(\theta) - \xi| < |\gamma(\theta_0) - \xi|$. To determine when this occurs, write $\xi = w\gamma_o + \sqrt{1 - w^2}v$ for the generic point in the $(d - 3)$ dimensional tube section about γ_o whose inner product with γ_0 is $w = \cos\phi$ (see Fig. 3.1 below). Note that v satisfies the constraints

$$|v| = 1 \quad v'\gamma_o = 0 \quad v'\dot{\gamma}_o = 0$$

where $\dot{\gamma}_o = \dot{\gamma}(\theta_0)$. Denote by P_M the projection onto $M = \text{span}\{\gamma_o, \dot{\gamma}_o\}$: since $\gamma \perp \dot{\gamma}_0$, $P_M u = (u'\gamma_0)\gamma_0 + (u'\dot{\gamma}_0)\dot{\gamma}_0 / |\dot{\gamma}_0|^2$.
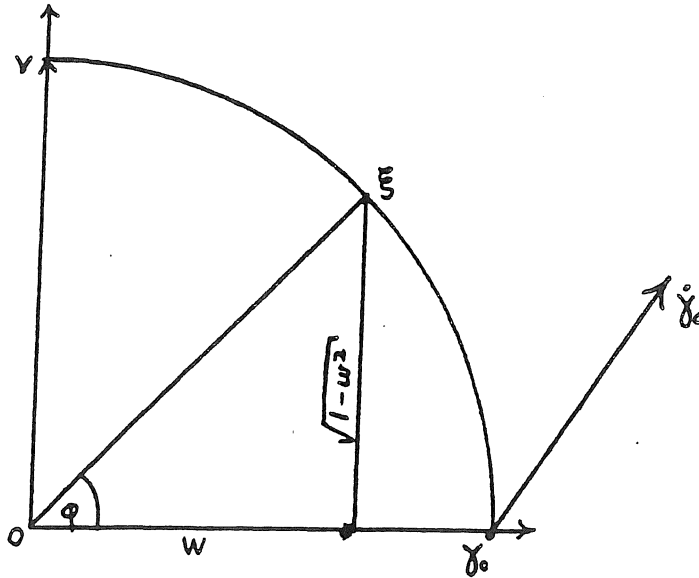

Figure 3.1 Parametrization of a point ξ in the tube section about γ_0.

Consider a fixed point $\gamma = \gamma(\theta)$: the point ξ^o of the above form which is closest to γ corresponds to

$$v = u^o/|u^o|, \quad u^o = (I - P_M)(\gamma - w\gamma_o).$$

Now from the definition of M, $\gamma' u^* = |u^*|^2 = 1 - \gamma' P_M \gamma$, so that

$$\gamma' \xi^* = w \gamma' \gamma_0 + \sqrt{1 - w^2} \sqrt{1 - \gamma' P_M \gamma}.$$

Thus the condition for overlap at θ_0 with a tube of angle φ reduces to the existence of some $\gamma = \gamma(\theta)$ for which

$$w(1 - \gamma' \gamma_0) < \sqrt{1 - w^2} \sqrt{1 - \gamma' P_M \gamma}.$$

Suppose that $\gamma' \gamma_0$ and $\gamma' P_M \gamma$ are functions of $\theta - \theta_0$, then the angle φ at which global overlap first occurs is obtained from the solution $w = \cos \varphi$, of

(3.3)
$$\frac{w^2}{1 - w^2} = \sup_\theta \frac{1 - \gamma' P_M \gamma}{(1 - \gamma' \gamma_0)^2} = \sup_\theta V_\theta = V,$$

where $\gamma' P_M \gamma = (\gamma' \gamma_0)^2 + (\gamma' \dot{\gamma}_0)^2 / |\dot{\gamma}_0|^2$.

Let us now consider two examples relevant to Andrews' plots. For S^{2k-1}, let $\gamma = \gamma^{(1)}(\theta) = \frac{1}{\sqrt{k}}(\cos \theta, \sin \theta, \ldots, \cos k\theta, \sin k\theta)$. The derivatives $\dot{\gamma}(\theta)$ and $\ddot{\gamma}(\theta)$ have constant length, namely $|\dot{\gamma}(\theta)|^2 = (k+1)(2k+1)/6$ and $|\ddot{\gamma}(\theta)|^2 = (k+1)(2k+1)(3k^2 + 3k - 1)/30$. The tube with central angle φ does not have any local self overlap if $\sin \varphi < \rho = |\dot{\gamma}|^2 / |\ddot{\gamma}|$. This leads to the critical values φ_L listed in Table 3.2 below. By using a grid on $[0, 2\pi)$ with $30k$ points, the values for V, w and $\varphi_G = \cos^{-1} w$ in (3.3) were obtained numerically. The values of φ_G hover close to $45°$: This is not unexpected since $\gamma' \gamma_0 = \frac{1}{k} \sum_{j=1}^{k} \cos[j(\theta - \theta_0)]$ is close to zero for most values of θ and hence the half-angle between $\gamma(\theta)$ and $\gamma(\theta_0)$ is about $45°$. Since global overlap first occurs before local overlap (as φ increases), we calculated the volume of the largest tube for which Hotelling's formula (2.2) is exact: This is listed as $P_0(W \geq w_G)$.

For S^{2k}, let $\gamma = \gamma^{(2)}(\theta) = \sqrt{\frac{2}{2k+1}}(\frac{1}{\sqrt{2}}, \cos \theta, \sin \theta, \ldots, \cos k\theta, \sin k\theta)$. Again $\dot{\gamma}(\theta)$ and $\ddot{\gamma}(\theta)$ have constant length and the condition for absence of local self-overlap and the rest of Table 3.3 is easily computed as for Table 3.2.

Let us return to the question of the percentage of possible projections that are 'seen' with an Andrews' plot and a squint angle φ. Consider now the maximum tube radius that can occur before the tubes about Γ and $-\Gamma$ overlap. By tracing through the preceding argument, it is evident that this angle $\varphi_B = \cos^{-1} w_B$ is found from the solution of

$$\frac{w^2}{1 - w^2} = \sup_\theta \frac{1 - \gamma' P_M \gamma}{(1 + \gamma' \gamma_0)^2},$$

that is, by changing the sign of $\gamma' \gamma_0$ in (3.3). Use of the same $30k$ point grid as above led to the numerical estimates of φ_B tabulated in Tables 3.2 and 3.3. Numerical methods may be used to obtain lower bounds on the total volume covered by tubes about Γ and $-\Gamma$ in the case of overlap. These are constructed using the shift invariant relationship between Γ and $-\Gamma$, but we omit further details.

k	ϕ_L	V	w	ϕ_G	$P_0(W \geq w_G)$	ϕ_B	$2P_0(W \geq w_B)$
2	59.0	1.0	.707	45.0	.79	28.0	.697
3	54.7	1.04	.714	44.4	.52	32.1	.344
4	52.9	1.08	.720	43.9	.30	33.8	.163
5	51.8	1.11	.725	43.5	.167	35.0	.077
6	51.2	1.14	.730	43.1	.0868	35.6	.0346
7	50.7	1.15	.731	43.0	.0453	36.0	.0154
8	50.4	1.17	.734	42.8	.0223	36.4	.0068
9	50.1	1.18	.736	42.6	.0110	36.7	.0030
10	49.9	1.19	.737	42.5	.0054	36.9	.0013
11	49.7	1.30	.752	41.2	.0016	37.1	.0006
12	49.6	1.21	.740	42.3	.0012	37.2	.0002
13	49.5	1.21	.740	42.3	.0006	37.3	.0001
14	49.4	1.22	.741	42.2	.0003	37.4	~0
15	49.3	1.22	.741	42.2	.0001	37.5	~0
16	49.2	1.23	.743	42.0	.0001	37.6	~0

Table 3.2 For curve $\Gamma = \gamma^{(1)}(C)$ in S^{2k-1} φ_L = angle (°) of first local overlap, V = maximal value of (3.3), $w = \cos \varphi_G$ defines angle of first global overlap. $P_0(W \geq w_G)$ = percentage of S^{2k-1} in largest non-self overlapping tube, φ_B = angle of first intersection of tubes about Γ and $-\Gamma$, $2P_0(W \geq w_B)$ = percentage of S^{2k-1} in largest non-intersecting tubes.

Another question of interest is "what projection of the data lies furthest from the Andrews' plot and how closely does the plot approach it?". If the curve on S^{d-1} traced out by the projection vectors is $\gamma(C)$, then we wish to find

$$M_1 = \min_{c \in S^{d-1}} \max_{\theta} c'\gamma(\theta)$$

or, since the projection on $\gamma(\theta)$ is equivalent to that on $-\gamma(\theta)$,

$$M_2 = \min_{c \in S^{d-1}} \max_{\theta} | c'\gamma(\theta) |,$$

and, if possible, the minimax value(s) of c.

To begin with, consider the curve $\gamma(\theta) = \frac{1}{\sqrt{k}}(\cos\theta, \sin\theta, \ldots, \cos k\theta, \sin k\theta)$ in S^{2k-1}, $\theta \in (0, 2\pi]$. It is easy to establish that $M_2 \in [\frac{1}{\sqrt{2k}}, \frac{1}{\sqrt{k}}]$. The upper bound follows from the choice $c = (1, 0, \ldots, 0)$. For the lower bound, consider $(c'\gamma(\theta))^2 / | c |^2$ and note that if

k	ϕ_L	V	w	ϕ_G	$P_0(W \geq w_G)$	ϕ_B	$2P_0(W \geq w_B)$
2	50.1	1.5	.776	39.2	.358	37.8	.653
3	49.1	1.38	.761	40.4	.230	38.3	.367
4	48.7	1.34	.757	40.8	.132	38.5	.186
5	48.5	1.32	.754	41.0	.071	38.6	.091
6	48.4	1.31	.754	41.1	.037	38.6	.042
7	48.4	1.31	.753	41.2	.0189	38.6	.0189
8	48.3	1.31	.753	41.2	.0093	38.7	.0084
9	48.3	1.30	.752	41.2	.0045	38.7	.0037
10	48.3	1.30	.752	41.2	.0022	38.7	.0016
11	48.3	1.40	.764	40.2	.0007	38.7	.0007
12	48.3	1.30	.752	41.3	.0005	38.7	.0003
13	48.2	1.30	.751	41.3	.0002	38.7	.0001
14	48.2	1.30	.752	41.3	.0001	38.7	.0001
15	48.2	1.30	.751	41.3	.0001	38.7	~0
16	48.2	1.30	.751	41.3	~0	38.7	~0

Table 3.3   As for Table 3.2, but refers to $\Gamma = \gamma^{(2)}(C)$ in S^{2k}.

	2	3	4	5	6	7	8	9
$\sup_\theta \sqrt{k} f_k$	1.0	.942	.943	.932	.933	.930	.930	.928
$\sup_\theta \sqrt{k} \tilde{f}_k$	.707	.793	.834	.857	.872	.882	.889	.893

	10	11	12	13	14	15	16
$\sup_\theta \sqrt{k} f_k$	.929	.928	.928	.927	.927	.927	.927
$\sup_\theta \sqrt{k} \tilde{f}_k$	.897	.900	.902	.904	.906	.907	.908

Table 3.4   Upper bound for M_2 is $k^{-1/2} \times$ indicated value
in table. The two bounds correspond to f_k and $\tilde{f}_k$ defined
in (3.4) and (3.5) respectively.

$c = (a_1, b_1, \ldots, a_k, b_k)$, then

$$| c |^2 = \frac{1}{\pi} \int_{-\pi}^{\pi} (\sum_{j=1}^{k} a_j \cos j\theta + b_j \sin j\theta)^2 d\theta$$

$$= \frac{k}{\pi} \int_{-\pi}^{\pi} (c'\gamma(\theta))^2 d\theta \leq 2k \max_{\theta} (c'\gamma(\theta))^2.$$

Equality would occur in the last expression if $f_c(\theta) = c'\gamma(\theta)$ were constant in absolute value. Of course, this is not possible since f_c is a degree k trigonometric polynomial, which in particular integrates to 0. The simplest function $f(\theta)$ with $| f(\theta) | = 1$ and $\int f(\theta) d\theta = 0$, seems to be $f(\theta) = \text{sign } (\theta)$ on $[-\pi, \pi)$. Thus a naive attempt to improve the upper bound for M_2 would be to use for f_c the (normalized) k term Fourier series approximation to $f(\theta) = \text{sign } (\theta)$ on $[-\pi, \pi)]$, namely

(3.4)
$$f_k(\theta) = \sum_{\substack{1 \leq j \leq k \\ j \text{ odd}}} \frac{\sin j\theta}{j} / \sqrt{\sum_{\substack{1 \leq j \leq k \\ j \text{ odd}}} 1/j^2}.$$

Another possibility would be to emply the approximation used by Jackson to study degree of approximation of continuous functions by trigonometric polynomials (see Davis (1963) §13.3, especially Lemma 13.3.5). In our present case, this sets

(3.5)
$$\tilde{f}_k(\theta) = \sum_{\substack{1 \leq j \leq k \\ j \text{ odd}}} \rho_{k,j} \frac{\sin j\theta}{j} / \sqrt{\sum_{\substack{1 \leq j \leq k \\ j \text{ odd}}} \rho_{k,j}^2/j^2}$$

where

$$\rho_{k,j} = \sum_{s=0}^{k-j} a_s a_{s+j} / \sum_{s=0}^{k} a_s^2$$

and

$$a_s = \sin[(s+1)\pi/(k+2)].$$

Numerical maximization of f_k and $\tilde{f}_k$ over a mesh of $30k$ points in $[-\pi, \pi]$ gave the results in Table 3.4.

Although $\tilde{f}_k$ consistently has smaller supremum than f_k, it seems that at least for larger k, a further substantial reduction (from 1.0 in the direction of .5) ought to be possible.

§4. Significance tests in projection pursuit regression

We indicate briefly how the tubes method can be used to derive approximate tests of significance appropriate for projection pursuit regression in an idealized setting. It is hoped that a more complete discussion will be given separately in Donoho and Johnstone (1985).

Consider a model in which the data consists of n i.i.d. observations $(X_j, Y_i) \in R^d \times R$ from the regression

$$(4.1) \qquad Y_j = f(\mathbf{X}_j) + \epsilon_j$$

in which X_j has a standard Gaussian distribution on R^d, $E(\epsilon_i \mid X_i) = 0$ and $E(\epsilon_j^2 \mid X_j) = 1$. Suppose that a significance test of the null hypothesis that $f = 0$ is desired. To motivate the test, consider the problem of fitting the model $Y = g(u'X) + e$, where $\mid u \mid = 1$ and $g : R \to R$ is a one dimensional "ridge" function of the linear combination $u'x$. It is, therefore, necessary to fit both u and $g(\cdot)$. For fixed u, we fit an orthogonal (with respect to Gaussian measure) polynomial of degree m using normalized Hermite polynomials. Specifically, if $H_m(t) = e^{t^2/2}(-d/dt)^m e^{-t^2/2}$ is the mth univariate Hermite polynomial, then set $e_m(t) = H_m(t)/\sqrt{t!}$. A method of moments (or bootstrap) estimator of g given u would be

$$(4.2) \qquad \hat{g}_u(u'x) = \widehat{P_u f}(u'x) = \sum_{r=1}^{m} \hat{c}_r(u) e_r(u'x),$$

where

$$(4.3) \qquad \hat{c}_r(u) = \frac{1}{n} \sum_{1}^{n} Y_i e_r(u'X_i).$$

Note that $P_u f(u'X) = E(f \mid u'X)$ and that $\hat{c}_r(u) \longrightarrow E\, f(X)e_r(u'X) = E\, P_u f(u'X)e_r(u'X)$ as $n \to \infty$ if (X_i, Y_i) are i.i.d. from the model (4.1). The regression function $f(x)$ could be of the form $g(u_0'x)$ for some u_0, but need not be. It will, therefore, be seen that $\widehat{P_u f}$ is measuring the best degree m approximation to $P_u f$, which in turn is the best L^2 approximation to f amongst ridge functions in direction u.

To choose u, simply find the direction $\hat{u}$ for which $\widehat{P_u f}$ has maximum variance (i.e. "explains" the data best), namely, $E[\widehat{P_u f}]^2$, which we shall write as $\| \widehat{P_u f} \|^2$. The fitted model at this point would be $Y = (\widehat{P_{\hat{u}} f})(\hat{u}'X) + e$. The algorithm of Friedman and Stuetzle (1981) would now iterate, applying analogs of the above steps to the residuals e to obtain a new ridge function in direction u_2, and so on till no substantial improvement in fit results. For our purposes here, it seems convenient and tractable to consider only the first iteration.

To derive a significance test of the null hypothesis that the regression function (namely f) is zero, we consider, in view of the preceding discussion

$$T = \sup_u \| \widehat{P_u f} \|^2 = \sup_u \| \widehat{T_u f} \|^2$$

where $\| \widehat{T_u f} \|^2 = \sum_{r=1}^{m} \hat{c}_r^2(u)$.

To approximate the null distribution of T, we need to isolate the effects of u as much as

possible. This is done via the Hermite polynomial identity.

$$(4.4) \qquad e_r(u'x) = \sum_{|k|=r} \sqrt{\binom{r}{k}} u^k e_k(x),$$

where $k = (k_1, \ldots, k_p)$ and $u^k = u^{k_1} \cdots u_p^{k_p}$ are multi-indices and $e_k(x) = \prod_i e_{k_i}(x_i)$. Substituting (4.4) into (4.3), we may write

$$\hat{c}_r(u) = \sum_{|k|=r} \sqrt{\binom{r}{k}} u^k \hat{c}_k,$$

where $\hat{c}_k = \frac{1}{n} \Sigma_1^n Y_i e_k(X_i)$. Now set $\gamma_k = \sqrt{\binom{r}{k}} u^k$, $\gamma^{(r)} = (\gamma_k; | k | = r)$, and $Z^{(r)} = (\hat{c}_k; \ | k | = r)$: we have

$$T = \sup_\gamma \sum_{r \leq m} (\gamma^{(r)'} Z^{(r)})^2 \leq \sum_{r \leq m} \sup_\gamma (\gamma^{(r)'} Z^{(r)}) = \sum_{r \leq m} T_r$$

where $| \gamma^{(r)} | = 1$ for each r.

Each $\hat{c}_k$ is a normalized sum of i.i.d. random variables, has mean zero, variance $\frac{1}{\sqrt{n}}$ and is uncorrelated with all other $\hat{c}_k$. We, therefore, approximate $\sqrt{n}Z^{(r)}$ by a standard normal vector of the appropriate dimension.

Thus, if we consider only the component at degree r, $T_r = \sup_\theta (\gamma^{(r)'} Z^{(r)})^2$, we are in the situation described in Section 2. If $d = 2$, then $\gamma_j^{(r)}(\theta) = \cos^j \theta \sin^{r-j} \theta$, so that $\theta \to \gamma^{(r)}(\theta)$ is a curve on S^{r+1}, so that the methods of Section 2 and 3 can be applied to give approximately conservative P-values for T_r.

The null distribution of T cannot be immediately obtained by the same method since it does not have the form of a single inner product. A simple alternative approach if n is not too large would be just to combine the independent P values associated with the individual components T_r.

§5. Pairwise comparisons of means

Let $\overline{X}_i, i = 1, \ldots, d$ be the average of n_i independent $N(\mu_i, \sigma^2)$ random variables. Consider the problem of providing confidence intervals for all pairwise differences of means $\{\mu_i - \mu_j, i \neq j\}$ with simultaneous coverage probability $100(1 - \alpha)\%$. Tukey (1953) and Kramer (1956) proposed the intervals

$$\overline{X}_i - \overline{X}_j \pm q_{d,N-d}^\alpha s \sqrt{\frac{1}{2}(\frac{1}{n_i} + \frac{1}{n_j})}, \qquad N = \sum_1^d n_i$$

where $q_{d,N-d}^{\alpha}$ is the upper 100α percentile of the studentized range distribution and would yield an exact coverage probability $1 - \alpha$ if all $n_i = N/d$ (Miller (1981)). Here, of course, s^2 is the usual pooled unbiased estimate of variance and is independent of all $\{\overline{X}_i\}$. It was long conjectured that these intervals were conservative, but Hayter (1984), building on the work of Brown (1979) was the first to give a general proof. The geometric approach, also discussed by Brown (1984), offers a useful perspective on this result, and we use it to give, as an example, a proof in the relatively simple case $d = 3$.

As a preliminary reduction, we condition on s and write $Z_i = \sqrt{n_i}(\overline{X}_i - \mu_i)/\sigma$, and $\sigma_i = \sigma/\sqrt{n_i}$. It then suffices to show that if $Z = (Z_1,\ldots,Z_d) \sim N(0,I)$, then

$$(5.1) \qquad P\big(\max_{i\neq j} \frac{|\sigma_i Z_i - \sigma_j Z_j|}{\sqrt{\sigma_i^2 + \sigma_j^2}} > c\big)$$

is maximized when all σ_i^2 are equal.

Now write $Z = RU$, where U is uniformly distributed on S^{d-1} and $R^2 \sim \chi_{(d)}^2$. Define also $\sigma_{ij} = (\sigma_i e_i - \sigma_j e_j)/\sqrt{\sigma_i^2 + \sigma_j^2}$ and $\overline{\sigma}_{ij} = -\sigma_{ij}$. Conditioning now on R, we have to show for $0 < \varphi < \pi/2$ that

$$(5.2) \qquad P\{\max_{i\neq j} |\sigma_{ij}'U| > \cos\varphi\}$$

is maximized when all σ_i^2 are equal. In other words, we consider the total surface area of the union of the spherical caps $C_{ij}(\varphi)$ and $\overline{C}_{ij}(\varphi)$ of angle φ centered on $\Sigma = \{\sigma_{ij}\} \cup \{\overline{\sigma}_{ij}\}$ and have to show that this is maximized when

$$(5.3) \qquad \sigma_{ij} = (e_i - e_j)/\sqrt{2} \qquad \text{all } i \neq j.$$

Two caps $C_{ij}(\varphi)$ and $C_{kl}(\varphi)$ will intersect if and only if the angle between σ_{ij} and σ_{kl} is less than 2φ. Thus to reduce the amount of intersection amongst the caps and thus maximize (5.2), it seems that we must push their centers apart so as to maximize the minimum angle between the elements of Σ. It is a simple matter to show that this maximum angle is $60°$ and is uniquely attained by the configuration (5.3). Of course, this heuristic completely solves the problem associated with (5.2) only for $\varphi \leq 30°$, but it simultaneously helps in understanding Hayter's result and, by the link to sphere packing problems, may indicate why his results appear to lie deep.

A complete proof in this vein for $d = 3$ is immediate, however, and we sketch this informally. (It appears from Brown (1984) that a proof for $d = 3$ along these lines was given by Kurtz (1956).) Note that the vectors σ_{ij} are all coplanar, lying in the subspace $M = M_\sigma$ perpendicular to $(1/\sigma_1, 1/\sigma_2, 1/\sigma_3)$. For fixed $\sigma = (\sigma_1, \sigma_2, \sigma_3)$, decompose Z into its projections $W = P_M Z$ and $W^\perp$. Clearly $W \sim N_2(0, P_M)$ and from the independence of W and $W^\perp$,

$$\mathcal{L}(\max_{i\neq j} |\sigma_{ij}'Z|) = \mathcal{L}(\max_{i\neq j} |\sigma_{ij}'W|).$$

Decomposing $W = \tilde{R}\tilde{U}$, with $\tilde{R}^2 \sim X^2_{(2)}$ and $\tilde{U}$ independently uniformly distributed on the copy of S^1 lying in M, we consider

$$P(\max_{i \neq j} | \sigma'_{ij}\tilde{U} | > \cos\varphi).$$

This reduces to considering the total measure (=length) of the union of six circular arcs centered on σ_{ij}, each arc being of equal length. Over all possible configurations of six points on the circle, this total length is maximized when the six points are spaced at angles of 60° around the circle. But this occurs when the σ_{ij} have the configuration (5.3)!

Acknowledgments

Figure 5.1   Configuration of spherical caps of common radius centered about σ_{ij} when $d = 3$. Centers of caps are constrained to lie on a great circle.
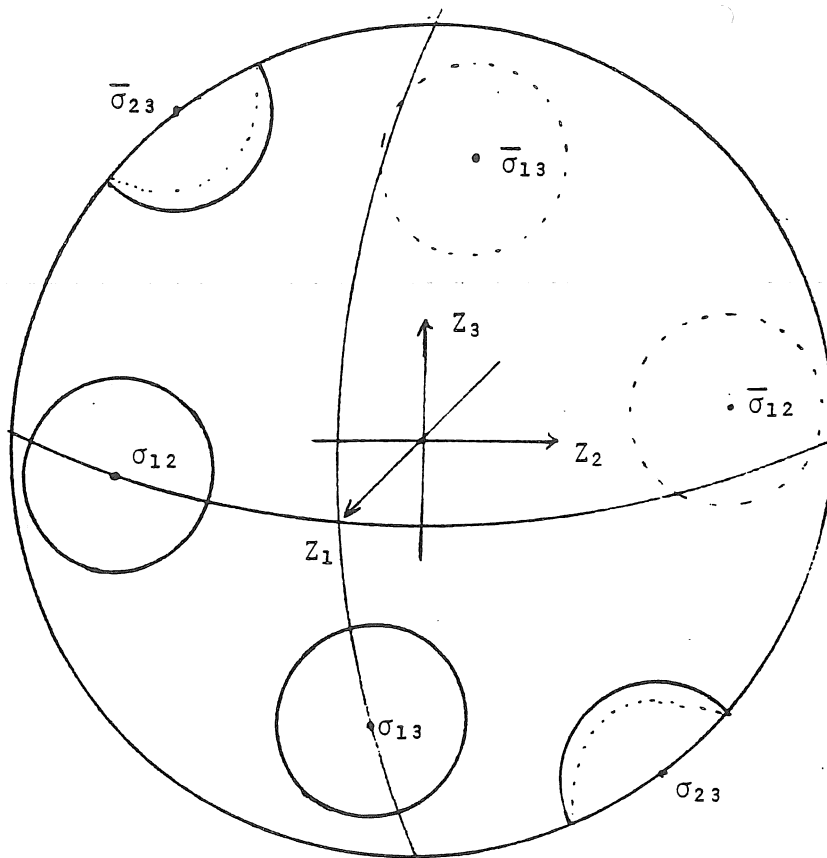
References

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics* **28** 137–156.

Brown, L. D. (1979). A proof that the Tukey-Kramer multiple comparison procedure for differences between treatment means is level-α for 3, 4, or 5 treatments. Unpublished technical report, Department of Mathematics, Cornell University, Ithaca, N. Y.

Brown, L. D. (1984). A note on the Tukey-Kramer procedure for pairwise comparisons of correlated means. In *Design of Experiments: ranking and selection: essays in honor of Robert E. Bechhofer*, eds. T.J. Santner and A.C. Tamhane, Dekker, New York.

Davis, P. J. (1963). *Interpolation and Approximation*, Blaisdell, New York.

Diaconis, P. and Efron, B. (1985). Testing for Independence in a two-way table: New interpretations of the chi-square statistic (with discussion). To appear, *Ann. Statist*

Donoho, D. and Johnstone, I. M. (1985). Manuscripts in preparation.

Friedman, J. H., Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions Computers.* **C-23** 881–889.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Ann. Statist.* **12** 61–75.

Hotelling, H. (1939). Tubes and spheres in n-spaces, and a class of statistical problems. *Amer. J. Math.* **61** 440–460.

Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* to appear.

Kramer, C. W. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* **12** 307–310.

Kurtz, T. E. (1956). An extension of a multiple comparison procedure. Unpublished Ph.D. thesis, Princeton University, Princeton, NJ.

Miller, R. G., Jr. (1981). *Simultaneous Statistial Inference*. Springer, New York.

Naiman, D. (1984). Conservative confidence bands in curvilinear regression. Technical Report, Department of Mathematics, Johns Hopkins University.

Olshen, R. A. (1985). Manuscript in preparation.

Sutherland, D.H., Olshen, R.A. et al., (1980). The development of mature gait. *J. Bone and Joint Surgery* **62A**, 336-353.

Tukey, J. W. (1953). The problem of multiple comparisons, unpublished report, Princeton, N. J.

Weyl, H. (1939). On the volume of tubes. *Amer. J. Math.* **61** 461–472.

Working, H. and Hotelling, H. (1929). Application of the theory of error to the interpretation of trends. *J. Amer. Statist. Assoc., Suppl. (Proc.)* **24** 73–85.

PREPRINTS 1984

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF
MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

PREPRINTS 1985

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF
MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1   Cohn, Harry: Almost Sure Convergence for Stochastically Monotone
          Temporally Homogeneous Markov Processes and Applications.

No. 2   Rootzén, Holger: Maxima and Exceedances of Stationary Markov Chains.

No. 3   Asmussen, Søren und Johansen, Helle: Über eine Stetigkeitsfrage
          betreffend das Bedienungssystem GI/GI/s.

No. 4   Asmussen, Søren and Thorisson, Hermann: A Markov Chain Approach to
          Periodic Queues.

No. 5   Hald, Anders: Galileo as Statistician.

No. 6   Johansen, Søren: The Mathematical Structure of Error Correction
          Models.

No. 7   Johansen, Søren and Johnstone, Iain: Some Uses of Spherical Geometry
          in Simultaneous Inference and Data Analysis.