*The Annals of Statistics*
1984, Vol. 12, No. 3, 793–815

# ASYMPTOTICS OF GRAPHICAL PROJECTION PURSUIT

BY PERSI DIACONIS[1] AND DAVID FREEDMAN[2]

*Stanford University and University of California, Berkeley*

Mathematical tools are developed for describing low-dimensional projections of high-dimensional data. Theorems are given to show that under suitable conditions, most projections are approximately Gaussian.

**1. Introduction.** One mainstay of data-analysis is the use of low-dimensional projections to study high-dimensional data sets. One-dimensional projections may be represented by histograms; two-dimensional ones, by scatter diagrams. A number of interactive data-analysis programs allow projection of a high-dimensional data set into a low-dimensional subspace selected by the user, who can search for interesting projections. See Fisherkeller, Friedman and Tukey (1974) or Donoho, Huber and Thoma (1981) for details. In recent years, Kruskal (1969, 1972) and Friedman and Tukey (1974) have suggested various algorithms for finding interesting projections. Heuristically, a projection will be uninteresting if it is random or unstructured. One standard measure of randomness is entropy. This gives a numerical criteria suggested by Huber (1981): a projection is interesting if it has small entropy relative to other projections, using a measure of randomness such as entropy ($-\int f \log f$) or Fisher information. Huber observes that the numerical criteria used by Friedman and Tukey essentially minimizes $-\int f^2$, another measure of randomness. If the scale is fixed, maximum entropy is attained by the Gaussian distribution. This suggests another heuristic: a projection is interesting if it is far from Gaussian. The data-analytic conclusion is to look at only a few of the projections which are close to Gaussian, and to look at more of the ones which are far from Gaussian.

This paper presents a different rationale for looking at non-Gaussian projections. For many data sets, we show that most projections are nearly the same and approximately Gaussian. Thus, if a data set is being inspected by projections, the non-Gaussian projections are the ones that are special. On the other hand, we also present classes of data sets where most projections are close to the same non-Gaussian distribution. For such a data set, a different criterion seems in order—the interesting projections may even be the ones which are close to Gaussian.

This paper introduces mathematical machinery for describing the distribution of projections. Most of the results are stated for one-dimensional projections, although the results generalize (Section 5).

---

Received January 1983; revised March 1984.
[1] Research partially supported by NSF Grant MCS-80-24649.
[2] Research partially supported by NSF Grant MCS-80-02535.
*AMS* 1980 *subject classifications.* Primary 60F99, 62-07
*Key words and phrases.* Projections, projection pursuit, random probabilities, empirical characteristic function, graphical methods.

The main results will now be stated. Let $x_1, x_2, \cdots, x_n$ be (nonrandom) vectors in $\mathbb{R}^p$. This is the data set. For mathematical convenience, suppose that $n, p$, and $x_i$ depend on a hidden index $\nu$. As $\nu$ tends to infinity, so do $n$ and $p$. Suppose that for $\sigma^2$ positive and finite, for any positive $\varepsilon$, as $\nu$ tends to infinity,

$$(1.1) \qquad (1/n)\text{card}\{j \leq n: \mid \| x_j \|^2 - \sigma^2 p \mid > \varepsilon p\} \to 0$$

$$(1.2) \qquad (1/n^2)\text{card}\{1 \leq j, k \leq n: \mid x_j \cdot x_k \mid > \varepsilon p\} \to 0.$$

Condition (1.1) says that most vectors have length near $\sigma^2 p$. Condition (1.2) says that most vectors are nearly orthogonal. The word "nearly" is important: of course, only $p$ vectors can be exactly orthogonal. The conditions are satisfied if e.g. the $x_i$ are observed values of independent identically distributed vectors with independent identically distributed $L_4$ coordinates (Section 3), or the $n = 2^p$ vertices of a unit cube centered at the origin.

Turn now to projections. Let $S_{p-1}$ be the unit sphere in $\mathbb{R}^p$. Put the uniform distribution on $S_{p-1}$. Let $\gamma$ be a typical element of $S_{p-1}$. The projected data in direction $\gamma$ have coordinates

$$(1.3) \qquad \gamma \cdot x_1, \gamma \cdot x_2, \cdots, \gamma \cdot x_n.$$

Let $\theta_\nu(\gamma)$ be the empirical distribution of this sequence, assigning mass $1/n$ to each $\gamma \cdot x_j$. The first theorem says that $\theta_\nu(\gamma)$ is close to $N(0, \sigma^2)$ for most $\gamma$, for large $\nu$. Here "close" is in the sense of the weak topology; "most" is relative to the uniform distribution on $S_{p-1}$. A technical description involves convergence in probability of the random measures $\theta_\nu(\cdot)$.

THEOREM 1.1. *Under conditions* (1.1) *and* (1.2), *as* $\nu \to \infty$, *the empirical distribution* $\theta_\nu$ *tends to* $N(0, \sigma^2)$ *weakly in probability.*

Theorem 1.1 is proved in Section 2. The approach is quite similar to the techniques in Freedman and Lane (1980, 1981). Section 3 gives examples where conditions (1.1) and (1.2) hold. Section 4 gives examples where most projections are not normal. The random measures $\theta_\nu(\cdot)$ may converge in probability to nonnormal limits; or in distribution but not in probability to random limits. Examples include the case of strongly correlated coordinates, and clusters.

The results in Theorem 1.1 continue to hold if the data are standardized using robust (i.e., weakly continuous) measures of location and scale such as the median and interquartile range. Consider next the case where the data are standardized, using the mean and standard deviation. Some notation is needed. Let $a_\nu$ be the mean of the projected data, $s_\nu^2$ the variance, and $t_\nu^2$ the second moment. Thus

$$(1.4) \qquad a_\nu = (1/n) \sum_{j=1}^n \gamma \cdot x_j, \quad s_\nu^2 = (1/n) \sum_{j=1}^n (\gamma \cdot x_j - a_\nu)^2,$$
$$t_\nu^2 = (1/n) \sum (\gamma \cdot x_j)^2.$$

Let

$$(1.5) \qquad \bar{x} = (1/n) \sum_{j=1}^n x_j.$$

The conditions required for Theorem 1.2 are

$$(1.6) \qquad (1/np)\sum_{j=1}^{n} \| x_j \|^2 \to \sigma^2 \quad \text{where} \quad 0 < \sigma^2 < \infty$$

$$(1.7) \qquad (1/n)\text{card}\{j \le n: \, | \, \| x_j \|^2 - p\sigma^2 \, | > \varepsilon p\} \to 0$$

$$(1.8) \qquad (1/(np)^2)\sum_{j,k=1}^{n} (x_j \cdot x_k)^2 \to 0.$$

These conditions imply (1.1–1.2), by Chebychev's inequality. Let $\theta_\nu^0(\gamma)$ be the centered empirical measure, assigning mass $1/n$ to $\gamma \cdot x_j - a_\nu$. Let $\theta_\nu^1(\gamma)$ be the scaled empirical measure, assigning mass $1/n$ to $\gamma \cdot x_j/t_\nu$. Let $\theta_\nu^2(\gamma)$ be the standardized empirical measure, assigning mass $1/n$ to $(\gamma \cdot x_j - a_\nu)/s_\nu$.

THEOREM 1.2.

(a) *Under conditions* (1.6–1.8), *as* $\nu \to \infty$,
- *the empirical second moment* $t_\nu^2$ *converges to* $\sigma^2$ *in probability*
- *the scaled empirical* $\theta_\nu^1$ *converges to* $N(0, 1)$ *weakly in probability.*

(b) *If conditions* (1.6–1.8) *hold for the centered data* $x_j - \bar{x}$, *then*
- *the empirical variance* $s_\nu^2$ *converges to* $\sigma^2$ *in probability*
- *the centered empirical* $\theta_\nu^0$ *converges to* $N(0, \sigma^2)$ *weakly in probability*
- *the standardized empirical* $\theta_\nu^2$ *converges to* $N(0, 1)$ *weakly in probability.*

REMARKS. Of course, part (b) of Theorem 1.2 follows from part (a). If the focus is on the standardized empirical, it is harmless to center the data and scale it so that $(1/n)\sum \| x_j - \bar{x} \|^2 = p$. The conditions become

$$(1 - \varepsilon)(1/n)\sum_{j=1}^{n} \| x_j - \bar{x} \|^2$$

$$(1.9) \qquad < \| x_k - \bar{x} \|^2 < (1 + \varepsilon)(1/n)\sum_{j=1}^{n} \| x_j - \bar{x} \|^2$$

$$\text{except for} \quad o(n) \quad \text{indices} \quad k = 1, 2, \cdots, n$$

$$(1.10) \qquad (1/n^2)\sum_{j,k=1}^{n} [(x_j - \bar{x}) \cdot (x_k - \bar{x})]^2 = o[(1/n)\sum_{j=1}^{n} \| x_j - \bar{x} \|^2]^2.$$

Theorem 2 will be proved in Section 2.

**2. Proofs of Theorems 1.1 and 1.2.** Let $\zeta$ be $N(0, I_p)$, i.e., a vector of $p$ independent $N(0, 1)$ variables. Then $\zeta/\| \zeta \|$ is uniformly distributed over $S_{p-1}$. On the other hand, $\| \zeta \|/\sqrt{p} \to 1$ almost surely as $p \to \infty$. Hence, it is enough to prove the theorems with $\gamma$ replaced by $\zeta/\sqrt{p}$; a variant of Slutzky's lemma is involved in this step. The advantage is that normal theory can be used. To economize on indices, we use $\sqrt{-1}$ instead of $i$. The first two lemmas are standard.

LEMMA 2.1. *Let* $U$ *be a random variable with characteristic function* $\phi$. *Then*

$$P\left\{ | \, U \, | \ge \frac{2}{\varepsilon} \right\} \le \frac{1}{\varepsilon} \int_{-\varepsilon}^{\varepsilon} [1 - \text{Re } \phi(t)] \, dt.$$

For the next lemma, let $\theta_\nu$ be a random probability on the line, with random

characteristic function $\phi_\nu$. Let $\theta_0$ be a deterministic probability on the line, with deterministic characteristic function $\phi_0$.

LEMMA 2.2. *$\theta_\nu \to \theta_0$ weakly in probability if and only if the random characteristic functions $\phi_\nu(t)$ converge to $\phi_0(t)$ in probability for each $t$.*

PROOF. "Only if" is clear. In the other direction, let $T_\delta = (-\infty, 2/\delta] \cup [2/\delta, \infty)$. Then

$$E\{\theta_\nu(T_\delta)\} \leq \frac{1}{\delta}\int_{-\delta}^{\delta} [1 - \operatorname{Re} E\{\phi_\nu(t)\}]\, dt$$

and $\limsup_\nu E\{\theta_\nu(T_\delta)\} \leq (1/\delta)\int_{-\delta}^{\delta}[1 - \phi_0(t)]\, dt$ is small for $\delta$ small. Given $\varepsilon$ positive, there is a positive $\delta$ so small that $P\{\theta_\nu(T_\delta) < \varepsilon\} > 1 - \varepsilon$ for all $\nu$. Thus $\{\theta_\nu\}$ is tight. □

Now let $\theta_\nu(\zeta)$ be the empirical measure of

$$(\zeta \cdot x_1)/\sqrt{p}, \cdots, (\zeta \cdot x_n)/\sqrt{p}.$$

PROPOSITION 2.1. *Under conditions* (1.1–1.2), *$\theta_\nu \to N(0, \sigma^2)$ weakly in probability.*

PROOF. The characteristic function of $\theta_\nu(\zeta)$ is

$$\phi_\nu(\zeta, t) = (1/n)\sum_{j=1}^{n} \exp\{\sqrt{-1}\, t(\zeta \cdot x_j)/\sqrt{p}\}. \tag{2.1}$$

Clearly,

$$E\{\phi_\nu(\cdot, t)\} = (1/n)\sum_{j=1}^{n} \exp\{-\tfrac{1}{2}t^2 \| x_j \|^2/p\} \to \exp\{-\tfrac{1}{2}t^2\sigma^2\} \tag{2.2}$$

by condition (1.1). Likewise

$$\begin{aligned} E\{|\phi_\nu(\cdot, t)|^2\} &= E\{\phi_\nu(\cdot, t)\bar{\phi}_\nu(\cdot, t)\} \\ &= (1/n^2)\textstyle\sum_{jk} E\{\exp[\sqrt{-1}\, t(\zeta \cdot (x_j - x_k))/\sqrt{p}]\} \\ &= (1/n^2)\textstyle\sum_{jk} \exp\{-\tfrac{1}{2}t^2 \| x_j - x_k \|^2/p\}. \end{aligned} \tag{2.3}$$

Of course,

$$\| x_j - x_k \|^2 = \| x_j \|^2 + \| x_k \|^2 - 2x_j \cdot x_k.$$

The summands in (2.3) are between 0 and 1. By conditions (1.1) and (1.2), except for a set of pairs $(j, k)$ of cardinality $o(n^2)$, we have simultaneously

$$\sigma^2 p - \varepsilon p < \| x_j \|^2 < \sigma^2 p + \varepsilon p,$$

$$\sigma^2 p - \varepsilon p < \| x_k \|^2 < \sigma^2 p + \varepsilon p$$

$$-\varepsilon p < x_j \cdot x_k < \varepsilon p$$

and hence

$$2\sigma^2 - 4\varepsilon < \| x_j - x_k \|^2/p < 2\sigma^2 + 4\varepsilon.$$

Hence

$$E\{|\phi_\nu(\cdot, t)|^2\} \to \exp\{-t^2\sigma^2\}.$$

By Chebychev's inequality, $\phi_\nu(\cdot, t) \to \exp\{-\frac{1}{2}t^2\sigma^2\}$ in probability. □

To prove Theorem 1.2, another estimate is needed.

LEMMA 2.3. *$E\{(\zeta \cdot x_j)^2(\zeta \cdot x_k)^2\}$ equals $3\|x_j\|^4$ if $j = k$, and $2(x_j \cdot x_k)^2 + \|x_j\|^2\|x_k\|^2$ if $j \neq k$.*

PROOF. Clearly, $\zeta \cdot x_j$ is normal with mean 0 and variance $\|x_j\|^2$, so the case $j = k$ is trivial. If $j \neq k$, consider the regression of $x_j$ on $x_k$, viz., $x_j = \alpha x_k + \delta$, where $\alpha$ is a scalar, $\delta \in R^p$, and $\delta \perp x_k$. Now $\zeta \cdot x_j = (\alpha\zeta \cdot x_k) + (\zeta \cdot \delta)$, and the terms on the right are independent. The rest is routine. □

For the scaling, now let $t_\nu(\zeta)^2$ be the empirical second moment:

$$t_\nu(\zeta)^2 = (1/n)\sum_{j=1}^n (\zeta \cdot x_j)^2/p.$$

The empirical variance $s_\nu^2$ is $t_\nu^2$ applied to the centered data $x_j - \bar{x}$.

LEMMA 2.4.

(a) $npE\{t_\nu(\zeta)^2\} = \sum_{j=1}^n \|x_j\|^2$

(b) $n^2p^2E\{t_\nu(\zeta)^4\} = 2\sum_{jk}(x_j \cdot x_k)^2 + (\sum_j \|x_j\|^2)^2$.

PROOF. Claim (a) is easy. For (b),

$$E\{t_\nu(\zeta)^4\} = (1/n^2p^2)\sum_{jk} E\{(\zeta \cdot x_j)^2(\zeta \cdot x_k)^2\}.$$

Using Lemma 2.3, the double sum can be evaluated as

$$3\sum_j \|x_j\|^4 + 2\sum_{j\neq k}(x_j \cdot x_k)^2 + \sum_{j\neq k}\|x_j\|^2\|x_k\|^2$$

which can be rewritten as

$$2\sum_{jk}(x_j \cdot x_k)^2 + (\sum_j \|x_j\|^2)^2. \quad \square$$

As before, let $\theta_\nu^1(\zeta)$ be the scaled empirical.

PROPOSITION 2.2. *Under conditions* (1.6–1.8), *the empirical second moment $t_\nu^2$ converges to $\sigma^2$ in probability, and the scaled empirical distribution $\theta_\nu^1(\zeta)$ converges to $N(0, \sigma^2)$ weakly in probability.*

PROOF. That $t_\nu^2 \to \sigma^2$ follows from Lemma 2.4, and $\theta_\nu \to N(0, \sigma^2)$ by Proposition 2.1; then $\theta_\nu^1$ can be handled, in effect by Slutzky's lemma. □

**3. Examples with most projections Gaussian.** This section presents examples of data sets that satisfy conditions (1, 2) or (6, 7, 8). The examples are

made up of independent and identically distributed random vectors

$$X_j = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{pj} \end{pmatrix}, \quad 1 \le j \le n.$$

The first example shows that conditions (1.1) and (1.2) hold almost surely for independent identically distributed (iid) coordinates.

EXAMPLE 3.1. *iid coordinates.* Let $X_{ij}$ be iid for $i = 1, 2, \cdots$, and $j = 1, 2, \cdots$. Suppose

$$E(X_{ij}) = 0, \quad \sigma^2 = E\{X_{ij}^2\} > 0 \quad \text{and} \quad E\{|X_{ij}|^{2+\eta}\} < \infty \quad \text{for some} \quad \eta > 0. \tag{3.1}$$

Then, for almost all realizations of the array $\{X_{ij}\}$, conditions (1.6–1.8) and so conditions (1.1–1.2) are satisfied, no matter how $n$ and $p$ tend to infinity.

PROOF. Condition (1.6) is easy

$$(1/np) \sum_{i=1}^{p} \sum_{j=1}^{n} X_{ij}^2 \to \sigma^2 \quad \text{a.e.} \tag{3.2}$$

Convergence in (3.2) is as $n$ and $p$ tend to infinity in any arbitrary way: the null set does not depend on the path. This strong result fails if it is only assumed that $E(X_{ij}^2) < \infty$. See Smythe (1973) for details.

For condition (1.7), fix $\varepsilon > 0$. Let $\xi_{pj}$ be 1 if

$$|(1/p) \sum_{i=1}^{p} X_{ij}^2 - \sigma^2| > \varepsilon;$$

otherwise let $\xi_{pj}$ be 0. We claim

$$\lim_{n,p\to\infty} (1/n) \sum_{j=1}^{n} \xi_{pj} = 0 \quad \text{a.e.} \tag{3.3}$$

Suppose first $E\{X_{ij}^4\} < \infty$. Fix $\delta$ positive but small. Let $A_{pn}$ be the event $\sum_{j=1}^{n} \xi_{pj} \ge \delta n$. We will show that $P\{A_{pn}\}$ sums over $n > n_0$ and $p > p_0$ when $n_0$ and $p_0$ are large; Borel–Cantelli completes the proof in the $L_4$ case.

Let $\pi_p = P\{\xi_{pj} = 1\}$. By Chebychev's inequality, $\pi_p \le A/\varepsilon^2 p$ where $A = \mathrm{Var}\{X_{ij}^2\} \le E\{X_{ij}^4\}$. By a version of Bernstein's inequality,

$$P(A_{pn}) \le (e\pi_p/\delta)^{\delta n} \le (Ae/\varepsilon^2 p\delta)^{\delta n}.$$

See Freedman (1973, Theorem 4b). Fix $p$ so large that $(Ae/\varepsilon^2 p\delta)^{\delta} < \frac{1}{2}$. The sum on $n$ of $(Ae/\varepsilon^2 p\delta)^{\delta n}$ from $n = n_0$ to $\infty$ is at most

$$2(Ae/\varepsilon^2 p\delta)^{\delta n_0}.$$

If $n_0 > 1/\delta$, this sums in $p$, completing the proof of (3.3) under the assumption $E\{X_{ij}^4\} < \infty$.

The fourth moment condition is eliminated by truncation. Fix $L$ large but

finite. Let

$$Y_{ij} = X_{ij} \quad \text{when} \quad |X_{ij}| \le L$$
$$= 0 \quad \text{when} \quad |X_{ij}| > L$$
$$Z_{ij} = X_{ij} - Y_{ij}.$$

Now

$$X_{ij}^2 = Y_{ij}^2 + Z_{ij}^2$$

because $Y_{ij}Z_{ij} = 0$. So, $Y_{ij}$ is uniformly bounded, and $E(Y_{ij}^2)$ is almost $\sigma^2$, while $E(Z_{ij}^2)$ is small. Now, $(1/p)\sum_{i=1}^p Y_{ij}^2$ can be dealt with under the fourth-moment condition. On the other hand,

$$\sup_p(1/p)\sum_{i=1}^p Z_{ij}^2 = V_j$$

are independent and identically distributed in $j$. The averages over $i = 1, \cdots, p$ of $Z_{ij}^2$ form a backwards martingale in $p$. Fix $\alpha$ a little bit larger than 1. Then $E(V_j^\alpha) \le (\alpha/(\alpha - 1))^\alpha E(Z_{ij}^{2\alpha})$ (see Doob, 1953, Theorem 3.4 on page 317). Now

$$E(V_j) \le (\alpha/(\alpha - 1))E(Z_{ij}^{2\alpha})^{1/\alpha}$$

is small for $L$ large and

$$\lim_{n\to\infty}(1/n)\sum_{j=1}^n V_j = E(V_j) \quad \text{a.e.}$$

is small. This completes the argument for condition (1.7).

We turn now to condition (1.8). It is convenient to deal first with the term $j = k$. We claim

$$(1/n^2p^2)\sum_{j=1}^n \left(\sum_{i=1}^p X_{ij}^2\right)^2 \to 0 \quad \text{a.e.} \tag{3.4}$$

Let $V_j = \sup_p(1/p)\sum_{i=1}^p X_{ij}^2$. Using Doob's inequality again, $V_j \in L_{1+\eta/2}$, and the $V_j$ are independent and identically distributed. Even if the $V_j$ were just in $L_1$ and identically distributed

$$(1/n)\max_{j=1,\cdots,n} V_j \to 0 \quad \text{a.e.}$$

This last follows from $V_n/n \to 0$ a.e. which in turn follows from the Borel-Cantelli lemma. It now follows that

$$\frac{1}{n^2}\sum_{j=1}^n \left(\frac{1}{p}\sum_{i=1}^p X_{ij}^2\right)^2 \le \left(\frac{1}{n}\max_{j=1,\cdots,n} V_j\right) \times \left(\frac{1}{np}\sum_{i=1}^p \sum_{j=1}^n X_{ij}^2\right).$$

The first factor goes to 0 a.e., and the second to $\sigma^2$. Thus, (3.4) holds.

We now take up the terms $j \ne k$ in (1.8). We claim

$$\lim_{n,p\to\infty}(1/n^2p^2)\sum_{1\le j<k\le n}\left(\sum_{i=1}^p X_{ij}X_{ik}\right)^2 = 0 \quad \text{a.e.} \tag{3.5}$$

Suppose first $\mu_4 = E\{X_{ij}^4\} < \infty$. Then the idea is to use Hoeffding's $U$-statistic argument. Let

$$T_{np} = (1/n(n-1))\sum_{1\le j<k\le n} h_p(X_j, X_k)$$

where

$$X_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{pj} \end{bmatrix}$$

and

$$h_p(x, y) = (1/p^2)(x \cdot y)^2 - (\sigma^4/p)$$

for column $p$-vectors $x$ and $y$. It is enough to show that $T_{np} \to 0$ a.e. By a slightly tedious calculation,

$$E\{h_p(X_j, X_k)\} = 0$$

$$\mathrm{Var}\{h_p(X_j, X_k)\} = 3(p-1)\sigma^4/p^3 + \mu_4/p^3 - \sigma^8/p^2.$$

Let

$$\alpha_p(x) = E\{h_p(x, X_k)\} = \sigma^2 \| x \|^2/p^2 - \sigma^4/p$$

so

$$\mathrm{Var}\ \alpha_p(X_j) = (\sigma^4/p^3)(\mu_4 - \sigma^4).$$

Let

$$h_p^*(x, y) = h_p(x, y) - \alpha_p(x) - \alpha_p(y).$$

Then

$$\alpha_p^*(x) = E\{h_p^*(x, X_k)\} = E\{h_p^*(X_j, y)\} = \alpha_p^*(y) = 0$$

and

$$\mathrm{Var}\{h_p^*(X_j, X_k)\} \le A/p^2$$

for some constant $A$. Now

$$T_{np} = (1/n) \sum_{j=1}^n \alpha_p(X_j) + T_{np}^*$$

where

$$(2/n) \sum_{j=1}^n \alpha_p(X_j) = (\sigma^2/p^2 n) \sum_{i=1}^p \sum_{j=1}^n (X_{ij}^2 - \sigma^2) \to 0 \quad \text{a.e.}$$

and

$$T_{np}^* = (1/n(n-1)) \sum_{1 \le j < k \le n} h_p^*(X_j, X_k)$$

has mean 0 and variance

$$(2/n(n-1))\mathrm{Var}\{h_p^*(X_j, X_k)\} \le B/n^2p^2$$

for some constant $B$. This is the key point; the reason is that $\alpha_p^* = 0$. The upshot is that

$$\sum_{np} P\{|\, T_{np}^* \,| > \varepsilon\} < \infty, \quad \text{so} \quad T_{np}^* \to 0 \quad \text{a.e.}$$

We now eliminate the fourth moment condition by truncation, and show that under condition (3.1) only,

$$(1/n^2p^2) \sum_{j,k=1}^n (X_j \cdot X_k)^2 \to 0 \quad \text{a.e.}$$

Let

$$Y_{ij} = X_{ij} \quad \text{when} \quad |X_{ij}| \le L$$
$$= 0 \quad \text{otherwise},$$
$$Z_{ij} = X_{ij} - Y_{ij},$$
$$A_{jk} = (X_j \cdot X_k)/p = (1/p) \sum_{i=1}^{p} X_{ij}X_{ik} = B_{jk} + C_{jk} + D_{jk} + F_{jk},$$

where

$$B_{jk} = (1/p) \sum_{i=1}^{p} Y_{ij}Y_{ik}, \quad C_{jk} = (1/p) \sum_{i=1}^{p} Y_{ij}Z_{ik},$$
$$D_{jk} = (1/p) \sum_{i=1}^{p} Z_{ij}Y_{ik}, \quad F_{jk} = (1/p) \sum_{i=1}^{p} Z_{ij}Z_{ik}.$$

Then

$$A_{jk}^2 = B_{jk}^2 + 2B_{jk}(C_{jk} + D_{jk} + F_{jk}) + (C_{jk} + D_{jk} + F_{jk})^2.$$

We claim that

$$\limsup_{n,p\to\infty} (1/n^2) \sum_{jk=1}^{n} C_{jk}^2 \le \varepsilon_L \quad \text{a.e.} \tag{3.6}$$

where $\varepsilon_L \to 0$ as $L \to \infty$. Indeed,

$$C_{jk}^2 \le ((1/p) \sum_{i=1}^{p} Y_{ij}^2) \times ((1/p) \sum_{i=1}^{p} Z_{ik}^2).$$

So

$$(1/n^2) \sum_{jk=1}^{n} C_{jk}^2 \le ((1/np) \sum_{i=1}^{p} \sum_{j=1}^{n} Y_{ij}^2) \times ((1/np) \sum_{i=1}^{p} \sum_{k=1}^{n} Z_{jk}^2).$$

As $n, p \to \infty$, the first factor on the right converges a.e. to $E\{Y_{ij}^2\}$, which is nearly $E\{X_{ij}^2\}$ for $L$ large; the second factor converges a.e. to $E\{Z_{ik}^2\}$, which is nearly 0 for $L$ large. This proves (3.6); analogous results for $D$ and $F$ may be obtained by the same argument.

Next, we claim that

$$\limsup_{np\to\infty} (1/n^2) \sum_{jk=1}^{n} B_{jk}C_{jk} \le \delta_L \quad \text{a.e.}, \tag{3.7}$$

where $\delta_L \to 0$ as $L \to \infty$. Indeed,

$$((1/n^2) \sum_{jk=1}^{n} B_{jk}C_{jk})^2 \le ((1/n^2) \sum_{jk=1}^{n} B_{jk}^2) \times ((1/n^2) \sum_{jk=1}^{n} C_{jk}^2).$$

The first factor on the right goes to 0 a.e. by (3.4–3.5); the second factor is under control by (3.6). This proves (3.7). Likewise for $D$ and $F$. This completes the verification of condition (1.8). □

Conditions (1.6–1.8) hold for the centered data $X_j - \bar{X}$, where $\bar{X} = (1/n) \sum_{j=1}^{n} X_j$, under the assumptions (3.1). This is comparatively easy to deduce from example 3.1. One useful fact:

$$(1/p) \sum_{i=1}^{p} ((1/n) \sum_{j=1}^{n} X_{ij})^2 \to 0 \quad \text{a.e.} \tag{3.8}$$

For the proof, observe that $(1/n) \sum_{j=1}^{n} X_{ij}$ is a backwards martingale. Fix $n_0$ and

let

$$S_i = \sup_{n \geq n_0}((1/n) \sum_{j=1}^{n} X_{ij})^2.$$

Then the $S_i$ are iid and

$$E(S_i) \leq 4E\{((1/n_0) \sum_{j=1}^{n_0} X_{ij})^2\} = 4\sigma^2/n_0.$$

So for $n > n_0$,

$$\lim \sup_{p \to \infty}(1/p) \sum_{i=1}^{p} ((1/n) \sum_{j=1}^{n} X_{ij})^2 \leq 4\sigma^2/n_0 \quad \text{a.e.}$$

See Doob (1953, Theorem 3.4 on page 317). □

REMARK 1. We have been assuming a $2 + \eta$th moment. We believe the argument goes through if $X_{ij}^2$ is in $L \log L$, by using a more sophisticated truncation.

REMARK 2. Let $\theta(npX\zeta)$ be the empirical distribution of

$$(\zeta \cdot X_1)/\sqrt{p}, \cdots, (\zeta \cdot X_n)/\sqrt{p}$$

where $\zeta = (\zeta_1, \zeta_2, \cdots, \zeta_p)$ is the first $p$ of a sequence of iid $N(0, 1)$ variables, independent of $\{X_{ij}\}$; and

$$X_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{pj} \end{bmatrix}.$$

Condition (3.1) implies $\theta(npX\zeta) \to N(0, \sigma^2)$ weakly in probability as $n$ and $p$ tend to infinity given $X$, for almost all realizations of $X$. Is the convergence a.e. in $\zeta$? The answer is negative, even if the $X_{ij}$ are $N(0, 1)$. Here, we are asking about free convergence of $n$ and $p$ to infinity; however, the answer is still negative for sufficiently peculiar fixed paths $(n_\nu, p_\nu)$: $\nu = 1, 2, \cdots$. We do get a positive answer by letting $n \to \infty$ and having only one $p_n$ for each $n$.

For results a.e., there is no point in conditioning on $X$, so leave $X$ free. Also, it is harmless to replace $\sqrt{p}$ by $\| \zeta \|$. Let

$$S_{pj} = \sum_{i=1}^{p} \zeta_i X_{ij} / \sqrt{\sum_{i=1}^{p} \zeta_i^2}.$$

Each $S_{pj}$ is $N(0, 1)$, and the $S_{pj}$ are independent for $j = 1, 2, \cdots$. In fact, the processes $\{S_{pj}: p = 1, 2, \cdots\}$ are independent in $j$, but this is immaterial here. For each $j$, the variables $S_{pj}: p = 1, 2, \cdots$, are dependent, but nearly independent for widely separated $p$'s. Let

$$A_{pn} = \{S_{pj} > 0 \text{ for all } j = 1, \cdots, n\}.$$

Then $P\{A_{pn}\} = 1/2^n$ for each $n$, and the $A_{pn}$ are nearly independent for widely separated $p$'s. So

$$P\{A_{pn} \text{ for infinitely many } p\} = 1.$$

Thus

$$P\{A_{pn} \text{ for infinitely many } p \text{ for all } n\} = 1.$$

In short, for any $n$, no matter how large, there are infinitely many $p$'s such that the empirical measure $\theta(npX\zeta)$ sits on the positive halfline $(0, \infty)$. This defeats convergence to $N(0, 1)$ weakly a.e. as $n, p \to \infty$ freely.

How about convergence along a peculiar path $(n_\nu, p_\nu)$? Fix any function $f$ from the positive integers into the positive integers, with $f(n)$ strictly increasing. We can find $g(n) > f(n)$ so large that

$$P\{A_{pn} \text{ for at least one } p \text{ with } f(n) \le p \le g(n)\} \ge 1 - 1/n^2.$$

Now consider the path $[n_\nu, p_\nu]$ that results from stacking the indices in the following order:

$$[1, f(1)]$$
$$[1, f(1) + 1]$$
$$\vdots$$
$$[1, g(1)]$$
$$[2, f(2)]$$
$$[2, f(2) + 1]$$
$$\vdots$$
$$[2, g(2)]$$
$$\vdots$$

That is, $[n_1, p_1] = [1, f(1)]$, and $[n_2, p_2] = [1, f(1) + 1]$, and so forth. As is easily seen, $P(A_{n_\nu p_\nu} \text{ i.o.}) = 1$, defeating almost sure convergence.

In principle, it is possible to get bounds on the rates of convergence in Theorems 1.1 and 1.2 using Chebychev's inequality and Esseen's smoothing lemma (Feller, 1971, page 536). If

$$(1/n)\text{card}\{j: 1 \le j \le n \text{ and } | \, \| x_j \|^2 - \sigma^2 p \, | > \varepsilon p\} < \varepsilon$$

and

$$(1/n^2)\text{card}\{j, k: 1 \le j, k \le n \text{ and } | \, x_j \cdot x_k \, | > \varepsilon p\} < \varepsilon$$

then, except for a set of $\gamma$'s of measure at most $f(\varepsilon)$, the empirical distribution of

$$\gamma \cdot x_1, \cdots, \gamma \cdot x_n$$

is within $f(\varepsilon)$ of $N(0, \sigma^2)$. The function $f$ may be estimated by the argument indicated above, but so far we have only very crude results; we hope to return to this issue later.

Here is a somewhat different argument, with a similar conclusion: for random data of the type considered in this section, the projections are normal up to a random error of size

$$O_p(1/\sqrt{n}) + O_p(1/\sqrt{p}).$$

To be more specific, for distribution functions $F$ and $G$ on the line, let $\| F - G \| = \sup_J | F(J) - G(J) |$, where $J$ is an interval. Let the $X_{ij}$ be independent with continuous distributions which may depend on $i$ but not on $j$. Suppose these distributions all have mean 0, variance 1, and a finite absolute third moment bounded by $\alpha_3 < \infty$. For $\gamma \in S_{p-1}$, let $\theta(\gamma)$ be the empirical distribution of

$$\gamma \cdot X_1, \gamma \cdot X_2, \cdots, \gamma \cdot X_n$$

where

$$X_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{pj} \end{bmatrix}.$$

Let $\Phi$ be the standard normal distribution function.

PROPOSITION 3.1. $\| \theta(\gamma) - \Phi \| \leq U_{np\gamma} + V_p(\gamma)$

*where*:

- $\sqrt{n}U_{np\gamma}$ *is a random variable with a Kolmogorov-Smirnov distribution which does not depend on* $p$ *or* $\gamma$, *or the laws of* $X_{ij}$, *and converges weakly to a limiting distribution as* $n \to \infty$.
- $V_p(\gamma)$ *is a function of* $p$ *and* $\gamma$ *only; and* $\sqrt{p}V_p(\cdot)$ *tends to* $4K\alpha_3/\sqrt{2\pi}$ *in probability as* $p \to \infty$, *where* $K$ *is the universal positive constant in the Berry-Esseen bound.*

PROOF. Let $F(\gamma)$ be the common theoretical law of $\gamma \cdot X_j$. Clearly,

$$\| \theta(\gamma) - \Phi \| \leq U_{np\gamma} + W_{np\gamma}$$

where

$$U_{np\gamma} = \| \theta(\gamma) - F(\gamma) \|, \quad W_{np\gamma} = \| F(\gamma) - \Phi \|.$$

Now the law of $\sqrt{n}U_{np\gamma}$ has the usual Kolmogorov-Smirnov distribution, whatever $p$ or $\gamma$ or $F_\gamma$ may be, and this converges as $n \to \infty$.

On the other hand, by the Berry-Esseen bound, $W_{np\gamma} \leq V_p(\gamma)$, where

$$V_p(\gamma) = K\alpha_3 \sum_{i=1}^p | \gamma_i |^3$$

and $K$ is a universal positive constant: see Petrov (1972, page 111). We must now demonstrate the limiting behavior of $\sqrt{p}V_p(\cdot)$. Let $Z_1, Z_2, \cdots$, be independent standard normal variables. Then $\sqrt{p}V_p(\cdot)$ is distributed as $K\alpha_3$ times

$$(1/p) \sum_{i=1}^p | Z_i |^3 / ((1/p) \sum_{i=1}^p | Z_i^2 |)^{3/2}$$

which converges a.e. to $E\{ | Z_i |^3 \} = 4/\sqrt{2\pi}$. □

**4. Examples of non-Gaussian projections.** Theorems 1.1 and 1.2 break down if the conditions are violated. In some cases, it is still possible to describe the asymptotic distribution of most projections. The examples presented here include cases in which most projections have the same non-Gaussian distribution and cases in which the projection depends on the direction $\gamma$.

With long-tailed data, asymptotic normality can fail. For instance, with Cauchy data, most projections (suitably scaled) are Cauchy. A bit more generally, let $X_{ij}$ be independent, with common symmetric stable density of index $\alpha < 2$, having characteristic function $\exp(-|t|^\alpha)$. Let $\theta(npX\gamma)$ be the empirical distribution of

$$p^{1/2}(\gamma \cdot X_1)/p^{1/\alpha}, \cdots, p^{1/2}(\gamma \cdot X_n)/p^{1/\alpha}$$

where

$$X_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{pj} \end{bmatrix}$$

and $\gamma$ is uniform on the unit sphere $S_{p-1}$ in $R^p$, independent of $X$. Let $Z$ be a standard normal variable, and $C_\alpha = E\{|Z|^\alpha\}$.

PROPOSITION 4.1. *As $n$ and $p$ tend to infinity, $\theta(npX\gamma)$ converges weakly in probability to a symmetric stable law of index $\alpha$, having characteristic function $\psi(t) = \exp(-C_\alpha|t|^\alpha)$.*

PROOF. Let $\phi_{npX\gamma}(t)$ be the empirical characteristic function

$$(1/n)\sum_{j=1}^n \exp\{\sqrt{-1}tp^{1/2}(\gamma \cdot X_j)/p^{1/\alpha}\}.$$

Take the expectation over $X$, holding $\gamma$ fixed, to get

$$\exp\{-|t|^\alpha(1/p)\sum_{i=1}^p (p^{1/2}|\gamma_i|)^\alpha\}.$$

As is easily seen,

$$(1/p)\sum_{i=1}^p (p^{1/2}|\gamma_i|)^\alpha \to C_\alpha \quad \text{in probability.}$$

Hence, $E\{\phi_{npX\gamma}(t)\} \to \psi(t)$. Likewise, $E\{|\phi_{npX\gamma}(t)|^2\} \to \psi(t)^2$. □

REMARK 1. Replace $\sqrt{p}\gamma$ by $\zeta$ and take expectations over $\zeta$ to get $\phi_{npX}$. This will not converge a.e.: see the corresponding remark in Section 3.

REMARK 2. Proposition 4.1 remains valid if $X_{ij}$ are in the domain of attraction of a symmetric stable law.

If the lengths of the vectors $x_j$ depend strongly on $j$, so condition (1.1) fails, but the inner products are negligible in the sense of condition (1.2), then the empiricals of the projections converge in probability to scale mixtures of normals. To be more precise, let $F_v$ be the empirical measure of the $n$ numbers $\|x_j\|/\sqrt{p}$.

Let $F$ be a distribution function on $(0, \infty)$. A condition generalizing (1.1) is

$$F_\nu \to F \quad \text{weakly.} \tag{4.1}$$

If $F$ is nondegenerate, this captures the idea that $\| x_j \| / \sqrt{p}$ depends strongly on $j$. If $Z$ and $Y$ are independent, with $Z$ being standard normal and $Y \geq 0$ having law $F$, the variable $ZY$ is said to be an $F$-scale mixture of normals.

PROPOSITION 4.2. *Suppose* (4.1) *and* (1.2). *As $\nu$ tends to infinity, the empirical distribution $\theta_\nu$ tends to the $F$-scale mixture of normals weakly in probability.*

The proof is just like that of Theorem 1.1 and is omitted. For a discussion of scale mixtures of normals, see Efron and Olshen (1979). Here is an example of data satisfying the conditions (4.1) and (1.2). Let $W_{ij}$ be iid with mean zero, variance 1, and finite $2 + \delta$th moment. Let $\sigma_1, \sigma_2, \cdots$, be iid with a common distribution $F$ on $(0, \infty)$. Suppose that $F$ has a finite fourth moment. Let $\sigma^2 = E(\sigma_j^2)$. Let $X_{ij} = \sigma_j W_{ij}$ and

$$X_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{pj} \end{bmatrix}.$$

PROPOSITION 4.3. *For almost all realizations of $W_{ij}$ and $\sigma_j$, the array $X_{ij}$ satisfies* (4.1) *and* (1.2). *Further*

(a) $(1/np) \sum_{j=1}^n \| X_j \|^2 \to \sigma^2$

(b) $\text{card}\{j: j = 1, \cdots, n \text{ and } | \| X_j \|^2 - \sigma_j^2 p | > \varepsilon p\} < \varepsilon n$, *for all large $n$ and $p$, for any positive $\varepsilon$.*

(c) $(1/n^2p^2) \sum_{jk=1}^n (X_j \cdot X_k)^2 \to 0$.

Thus, condition (1.1) fails, but (1.2) holds. The proof of Proposition 4.3 is omitted, being quite similar to the arguments in Section 3. Together with Proposition 4.2, it implies that for most $\gamma$, the empirical distribution of $\gamma \cdot X_j$ is close to the $F$-scale mixture of normals. Further, the empirical mean of the projections $\gamma \cdot X_j$ is for most $\gamma$ nearly 0 and the variance is nearly $\sigma^2$, so standardizing still results in a scale mixture of normals.

We turn next to models suggested by factor analysis. In these models, condition (1.1) and (1.2) fail, and so does the conclusion of Theorem 1.1; indeed, the empirical distribution $\theta(\gamma)$ depends strongly on $\gamma$. We consider nonrandom $p$-vectors $x_1, x_2, \cdots, x_n$. Define

$$f_j = (1/p) \sum_{i=1}^p x_{ij} \quad \text{and} \quad \varepsilon_{ij} = x_{ij} - f_j.$$

The following conditions are assumed:

(4.2) The empirical distribution of $f_1, \cdots, f_n$ converges to a continuous distribution function $F$.

(4.3) The vectors $\varepsilon_j$ satisfy conditions (1.1) and (1.2), where

$$\varepsilon_j = \begin{bmatrix} \varepsilon_{1j} \\ \vdots \\ \varepsilon_{pj} \end{bmatrix}.$$

For distribution functions $G$ and $H$, recall $\| G - H \| = \sup_J | G(J) - H(J) |$, where $J$ ranges over intervals.

PROPOSITION 4.4. *Assume* (4.2–4.3). *Let* $\theta_\nu(\gamma)$ *be the empirical distribution of* $\gamma \cdot x_1, \cdots, \gamma \cdot x_n$. *Let* $\Gamma_p = \sum_{i=1}^p \gamma_i$. *Let* $U$ *and* $Z$ *be independent, with* $U$ *having distribution* $F$ *and* $Z$ *being normal with mean* 0 *and variance* $\sigma^2$. *Let* $\psi_\nu(\gamma)$ *be the distribution of*

$$\Gamma_p U + Z.$$

*Then* $\| \theta_\nu - \psi_\nu \| \to 0$ *in probability as* $\nu$ *tends to infinity.*

PROOF. Let $\theta_\nu^{(2)}(\gamma)$ be the joint empirical distribution of

$$(f_j, \gamma \cdot \varepsilon_j): \quad j = 1, \cdots, n.$$

Let $\psi^{(2)} = F \times N(0, \sigma^2)$, another probability on the plane. We claim

$$\theta_\nu^{(2)} \to \psi^{(2)} \quad \text{weakly in probability.} \tag{4.4}$$

For this purpose, it is harmless to replace $\gamma_i$ by $\zeta_i/\sqrt{p}$, the $\zeta$'s being independent standard normals. Let $\phi_\nu^{(2)}(t, u)$ be the empirical characteristic function

$$(1/n) \sum_{j=1}^n \exp[\sqrt{-1}tf_j + \sqrt{-1}u\zeta \cdot \varepsilon_j/\sqrt{p}]$$

where $\zeta$ is the column $p$-vector $(\zeta_1, \cdots, \zeta_p)$. As usual

$$E\{\phi_\nu^{(2)}(t, u)\} \to \hat{F}(t)\exp[-\tfrac{1}{2}\sigma^2 u^2]$$

where $\hat{F}$ is the characteristic function of $F$. Likewise,

$$E\{| \phi_\nu^{(2)}(t, u) |^2\} \to | \hat{F}(t) |^2 \exp[-\sigma^2 u^2].$$

This proves (4.4), see Lemma 2.2.

For probabilities $\alpha$ and $\beta$ on $\mathbb{R}^2$, let

$$\| \alpha - \beta \| = \sup\{| \alpha(K) - \beta(K) |: K \text{ is Borel and convex.}\}.$$

Because $\psi^{(2)}$ assigns measure 0 to the boundary of each $K$, a theorem of Ranga Rao (1962) entails

$$\| \theta_\nu^{(2)} - \psi^{(2)} \| \to 0 \quad \text{in probability.} \tag{4.5}$$

Clearly,

$$\gamma \cdot x_j = \Gamma_p f_j + (\gamma \cdot \varepsilon_j).$$

Let $J$ be a linear interval. So $\gamma \cdot x_j \in J$ iff $(f_j, \gamma \cdot \varepsilon_j)$ falls in the convex set

$$\{(u, v): \Gamma_p u + v \in J\}.$$

Thus

$$\| \theta_\nu(\gamma) - \psi_\nu(\gamma) \| \leq \| \theta_\nu^{(2)}(\gamma) - \psi^{(2)}(\gamma) \|. \quad \square$$

REMARK 1. Suppose $F$ is non-Gaussian. Then the limiting distribution $\Gamma_p U + Z$ is non-Gaussian too. Also, this limit depends strongly on $\gamma$. Indeed, $\Gamma_p$ is nearly $N(0, 1)$ and therefore must vary with $\gamma$. As is easily verified, the law of $\Gamma_p U + Z$ determines $\Gamma_p$.

REMARK 2. In the theorems above we have used the uniform distribution on an $i$ dimensional sphere. It is possible to realize all of these uniform distributions on a common probability space and then ask about almost sure convergence of $\theta_\nu(\gamma_i)$: fix a sequence $\zeta_1, \zeta_2, \cdots$, of independent standard normals; realize $\gamma_i$ as $\zeta_i / \| \zeta \|$. Even in this restricted model, $\theta_\nu$ converges in law but not in probability, because the same is true of $\Gamma_p$.

REMARK 3. The conclusions of Theorem 1 fail here; what of the hypotheses? Suppose $\tau^2 = \int x^2 F(dx)$ is positive and finite; and $(1/n) \sum_{j=1}^n f_j^2 \to \tau^2$. By orthogonality,

$$(1/p) \| x_j \|^2 = f_j^2 + (1/p) \| \varepsilon_j \|^2 \doteq f_j^2 + \sigma^2 \tag{4.6}$$

is strongly dependent on $j$; and

$$(1/p)(x_j \cdot x_k) = f_j f_k + (1/p)(\varepsilon_j \cdot \varepsilon_k) \doteq f_j f_k. \tag{4.7}$$

Both (1.1) and (1.2) fail.

REMARK 4. Consider the one-factor model

$$X_{ij} = U_j + V_{ij}.$$

Suppose the $U_j$ are independent, with common distribution $F$; the $V_{ij}$ are iid with mean 0, variance $\sigma^2$, and finite $2 + \delta$th moment. Then conditions (4.2–4.3) hold for almost all realizations of $X$; independence of $U$ and $V$ is not required, nor any moment condition on $U$. Indeed, let

$$V_{\cdot j} = (1/p) \sum_{i=1}^p V_{ij}$$

which is negligible for most $j$ by previous arguments. Then

$$f_j = (1/p) \sum_{i=1}^p X_{ij} = U_j + V_{\cdot j}, \quad \varepsilon_{ij} = X_{ij} - f_j = V_{ij} - V_{\cdot j}.$$

Lemma 4.1 below is useful in verifying condition (4.2).

REMARK 5. What happens to the scaled empirical? Assume that $U$ with law $F$ has mean 0 and finite variance $\tau^2$. Then $\Gamma_p U + Z$ has mean 0 and a variance given $\gamma$ of $\Gamma_p^2 \tau^2 + \sigma^2$, suggesting that the scale of $\theta_\nu(\gamma)$ depends strongly on $\gamma$. To pin this down, assume the stronger conditions (6 − 7 − 8) on $\varepsilon_{ij}$. Then, as is easily verified, the mean of $\theta_\nu(\gamma)$ does tend to 0 in probability, and the variance to $\Gamma_p^2 \tau^2 + \sigma^2$. The standardized empirical $\theta_\nu(\gamma)$ will therefore look, for most $\gamma$,

like the distribution of

$$(\Gamma_p U + Z)/\sqrt{\Gamma_p^2 \tau^2 + \sigma^2}.$$

Again, this is non-Gaussian and strongly dependent on $\gamma$, for non-Gaussian $U$.

REMARK 6. What happens if we scale the vectors separately? The idea is to make condition (1.1) hold by brute force, replacing $x_j$ by

$$\hat{x}_j = \sqrt{p} x_j / \| x_j \|.$$

We assume conditions (4.2–4.3) hold for $x_j$. Recall (4.6–4.7):

$$\frac{1}{p} (\hat{x}_j \cdot \hat{x}_k) \doteq \frac{f_j}{\sqrt{f_j^2 + \sigma^2}} \times \frac{f_k}{\sqrt{f_k^2 + \sigma^2}}$$

and condition (1.2) fails.

We turn to the asymptotic behavior of $\hat{\theta}_\nu(\gamma)$, the empirical distribution for $j = 1, \cdots, n$ of

$$\gamma \cdot \hat{x}_j = (\Gamma_p f_j + (\gamma \cdot \varepsilon_j))/\sqrt{f_j^2 + (1/p) \| \varepsilon_j \|^2}.$$

This $\hat{\theta}_\nu(\gamma)$ merges with $\hat{\psi}_\nu(\gamma)$, the theoretical distribution of

$$(\Gamma_p U + Z)/\sqrt{U^2 + \sigma^2}.$$

To make the idea of merging precise, we introduce a metric for weak convergence, similar to Prokhorov's. If $\mu$ and $\mu'$ are two probabilities on the line, let $d(\mu, \mu')$ be the inf of $\varepsilon$ positive such that for all intervals $K$ containing the origin

$$\mu(K) \le \mu'(e^\varepsilon K) + \varepsilon, \quad \mu'(K) \le \mu(e^\varepsilon K) + \varepsilon$$

where if $\lambda$ is real and $K = (a, b)$, then $\lambda K = (\lambda a, \lambda b)$. Clearly, $d(\mu, \mu') \le \| \mu - \mu' \|$.

We claim

$$d(\hat{\theta}_\nu, \hat{\psi}_\nu) \to 0 \quad \text{in probability as} \quad \nu \to \infty. \tag{4.8}$$

Indeed,

$$d(\hat{\theta}_\nu, \hat{\psi}_\nu) \le d(\hat{\theta}_\nu, \tilde{\theta}_\nu) + d(\tilde{\theta}_\nu, \hat{\psi}_\nu)$$

where $\tilde{\theta}_\nu$ is the empirical distribution of

$$(\Gamma_p f_j + (\gamma \cdot \varepsilon_j))/\sqrt{f_j^2 + \sigma^2}.$$

Fix real $a, b, c$. Now the planar set

$$\{(x, y): (ax + y)/\sqrt{x^2 + b^2} \le c\}$$

either is convex or has a convex complement. So

$$d(\tilde{\theta}_\nu, \hat{\psi}_\nu) \le \| \tilde{\theta}_\nu - \hat{\psi}_\nu \| \le \| \theta_\nu^{(2)} - \psi_\nu^{(2)} \| \to 0 \quad \text{in probability}$$

by (4.5). But $d(\hat{\theta}_\nu, \tilde{\theta}_\nu) \to 0$ too, because for large $n$, except for $o(n)$ indices $j = 1, \cdots, n$, by (4.3),

$$e^{-2\varepsilon} \sigma^2 < (1/p) \| \varepsilon_j \|^2 < e^{2\varepsilon} \sigma^2$$

so

$$e^{-\epsilon}\sqrt{f_j^2 + \sigma^2} < \sqrt{f_j^2 + (1/p)\|\varepsilon_j\|^2} < e^{\epsilon}\sqrt{f_j^2 + \sigma^2}.$$

The following lemma shows that small perturbations of empirical distributions do not change them much in the metric $\rho$. This was used in Remark 4 above.

LEMMA 4.1. *Let $\mu$ be the empirical distribution of the $n$ numbers $\xi_1, \cdots, \xi_n$; and $\mu'$ the empirical distribution of $\xi_1 + \eta_1, \cdots, \xi_n + \eta_n$. Let $\varepsilon > 0$. Suppose that $|\eta_j| \leq \varepsilon$ except for $\varepsilon n$ indices $j = 1, \cdots, n$. Let $\rho$ be Prokhorov's metric. Then $\rho(\mu, \mu') \leq \varepsilon$.*

The next example determines the behavior of projections of data clustered about $k$ centers. The following assumptions will be made:

(4.9) Let $c_1, c_2, \cdots, c_k$ be distinct $p$-vectors.

(4.10) Let $V_{ij}$ be iid with mean zero, variance $\sigma^2$ and a finite 3rd moment. Let $V_i = (V_{i1}, \cdots, V_{ip})^T$.

(4.11) For each $n$ there is a sequence $\{n_i\}$ of integers satisfying

$$n_0 = 0 < n_1 < n_2 < \cdots < n_k = n$$

with

$$n_i/n \to \lambda_i, \quad \lambda_0 = 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{k-1} < 1 = \lambda_k \quad \text{as} \quad n \to \infty.$$

Define $X_i = c_j + V_i$ for $n_{j-1} < i \leq n_j$, $j = 1, 2, \cdots, k$.

PROPOSITION 4.5. *Assume* (4.9–4.11). *For $i = 1, \cdots, k$ let $\psi'_\gamma$ be the law of a normal variable having mean $\gamma \cdot c_i$ and variance $\sigma^2$. Let $\psi_\gamma$ be the mixture of $\psi'_\gamma$ with weights $\lambda_i - \lambda_{i-1}$. Let $\theta_\gamma$ be the empirical measure of $\gamma \cdot X_1, \cdots, \gamma \cdot X_n$. Let*

$$D_\gamma = \sup_t |\theta_\gamma(t) - \psi_\gamma(t)|.$$

*Then, for almost all realizations of the array $V_{ij}$, $D_\gamma$ tends to zero in probability.*

PROOF. Let $\theta^j_\gamma$ be the empirical measure of the points in the $j$th cluster—that is, of the points $\gamma \cdot c_j + \gamma \cdot V_i$, $n_{j-1} < i \leq n_j$. Proposition 3.1 implies that the sup norm between $\theta^j_\gamma$ and a normal $(0, \sigma^2)$ variable tends to zero almost surely, in probability. The empirical $\theta_\gamma$ is a mixture of $\theta^j_\gamma$ with mixing weights that tend to $\lambda_j - \lambda_{j-1}$. □

Data generated from a model like the one just described is the base of Example B in Friedman and Tukey (1974). In that example, 65 points were centered at each of the 15 corners of a simplex in 15-dimensions. The coordinates of the points were independent standard normal. The simplex was scaled so that the $i$th vector $c_i$ was a vector with $10/\sqrt{2}$ in the $i$th coordinate and zeros elsewhere. Thus the distance between the vertices was 10. The 65 vectors from the $i$th

cluster thus project to points of the form

$$\gamma_i(10/\sqrt{2}) + Z$$

where $\gamma_i$ is, approximately, normal with mean zero and variance $1/15$, and $Z$ is standard normal. The data has 15 clusters, and a plot in a typical direction will look like the result of choosing 15 independent centers $\gamma_i(10/\sqrt{2})$ and putting a normal histogram based on a sample of size 65 about each center. As Friedman and Tukey demonstrate empirically, such a display will not be structured; it is not particularly normal either. Their projection pursuit algorithm found projections that clearly separate each cluster from the rest of the data.

**5. Final remarks.** This section treats normal data and higher dimensional projections.

Projection pursuit algorithms try to find nonnormal projections. One natural question is: suppose $X_i$ are iid $p$-dimensional vectors with independent standard normal coordinates. How much "structure" can be found? Figure 1 shows three

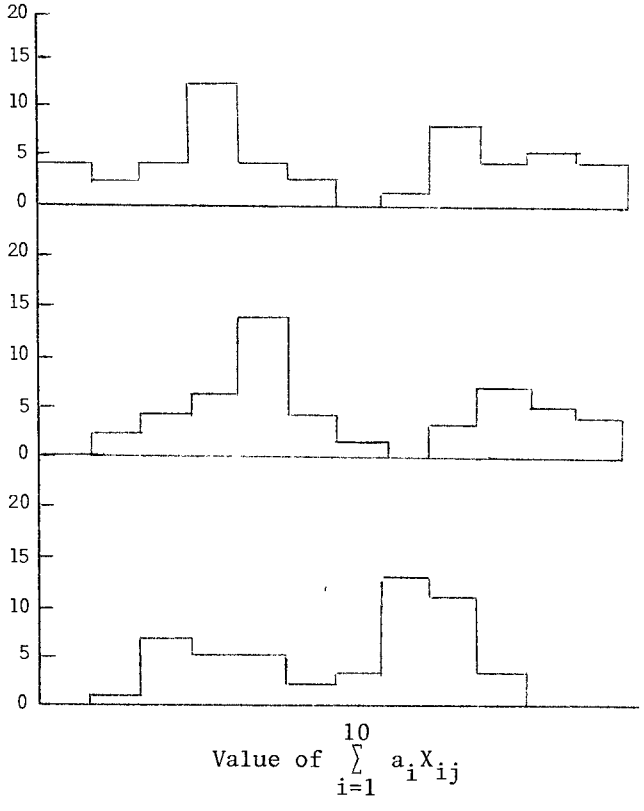


FIG. 1. *Histograms of a highly nonnormal projection from three samples of* 50 *from a* 10*-dimensional spherically symmetric normal distribution.*

clustered projections based on normal samples of 50 points in 10 dimensions. The data appear quite structured. These figures are based on simulations reported in Day (1969).

The following result shows that if $n$ and $p$ tend to infinity in such a way that $p/n \to 0$, then the least normal projection is close to normal. We would like to thank Ken Alexander for showing us how to improve an earlier result by making careful use of the results of Vapnik and Cervonenkis (1971).

PROPOSITION 5.1. *Let $F$ be a continuous probability on $\mathbb{R}^p$. Let $X_1, \cdots, X_n$ be a sample from $F$ with empirical measure $F_n$. Let $F^\gamma$ denote the law of $\gamma \cdot X_1$. Let*

$$D = \sup_\gamma \sup_t |F^\gamma - F_n^\gamma|.$$

*If $n$ and $p$ tend to infinity in such a way that $p/n \to 0$, then for any fixed $\varepsilon > 0$, $P\{D > \varepsilon\} \to 0$.*

PROOF. Theorem 2 of Vapnik and Cervonenkis (1971) implies that the probability that $D$ is larger than $\varepsilon$ is bounded above by

$$4m(p, 2n)\exp(-\varepsilon^2 n/8),$$

where $m(p, 2n) = \sum_{k=0}^{p} \binom{2n}{k}$ for $2n > p$. This is $2^{2n}$ times the lower tail of a binomial distribution. Feller (1968, VI.3) gives

$$m(p, 2n) \leq \binom{2n}{p} \frac{2n - p}{n - p}.$$

Now routine use of Stirling's formula shows that for universal positive constants $c_i$,

$$m(p, 2n) \leq c_1 \exp(c_2 p \log(n/p)).$$

It follows that $m(p, 2n)\exp(-\varepsilon^2 n/8) \to 0$. □

REMARK 1. If $F$ is $p$-dimensional standard normal, then $F^\gamma$ is standard normal for any $\gamma$, so this result says that even the least normal projection of normal data is close to normal.

REMARK 2. There is an evident discrepancy between Proposition 5.1 and the example in Figure 1. Just how large $p/n$ may be for practical values of $n$ and $p$ requires further simulation and theory.

REMARK 3. Work of Geman (1980) implies that if $n$ and $p = p_n$ tend to infinity in such a way that $p/n \to \eta > 0$, then the least normal projection of normal data will deviate from normality in some aspects. Indeed, if $X_j$ are iid $p$-dimensional standard normal for $1 \leq j \leq n$, the maximum variance of $X_1, \cdots, X_n$ is almost surely larger than $(1 + 2\sqrt{\eta})$ instead of 1. More specifically, let $M$ be the $n \times p$ matrix whose $ji$ element is $X_{ij}$. Let $L$ be the largest eigenvalue of

$M^TM$. Then Geman showed that $(1/n)L = \sup_{\|\gamma\|=1}(1/n)\ \sum_{j=1}^n\ (\gamma \cdot X_j)^2 \to (1 + \sqrt{\eta})^2$ a.s. (As usual, $\gamma$ is a $p$-vector.) But

$$\sup_{\|\gamma\|=1}\text{variance}(\gamma \cdot X_1, \cdots, \gamma \cdot X_n)$$

$$\geq \sup_{\|\gamma\|=1} \frac{1}{n} \sum_{j=1}^n (\gamma \cdot X_j)^2 - \sup_{\|\gamma\|=1}\left(\frac{1}{n} \sum_{j=1}^n \gamma \cdot X_j\right)^2$$

and

$$\sup_{\|\gamma\|}\left(\frac{1}{n} \sum_{j=1}^n \gamma \cdot X_j\right)^2 = \left\| \frac{1}{n} \sum_{j=1}^n X_j \right\|^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{\sqrt{n}} \sum_{j=1}^n X_{ij}\right)^2$$

is distributed as $\chi_p^2 \approx p/n \to \eta$ a.s., completing the argument.

De Wet, Venter, and van Wyck (1979) give some results on the maximum third and fourth moments in connection with a projection pursuit test for normality.

Thus far we have been working with 1-dimensional projections. These determine the behavior of most 2 or 3 dimensional projections. Consider the case where most projections are normal.

PROPOSITION 5.2. *Suppose conditions* (1.1) *and* (1.2) *are satisfied. For* $\beta$ *and* $\gamma$ *in* $S_{p-1}$ *let* $\theta_{\beta\gamma}$ *be the empirical distribution of* $(\beta \cdot X_1, \gamma \cdot X_1), \cdots, (\beta \cdot X_n, \gamma \cdot X_n)$. *Choose* $\gamma$ *uniformly on* $S_{p-1}$ *and* $\beta$ *uniformly among vectors orthogonal to* $\beta$. *As* $\nu \to \infty$, $\theta_{\beta\gamma}$ *tends to a standard bivariate normal measure, weakly in probability.*

PROOF. This can be proved directly via the argument for Theorem 1, using bivariate characteristic functions; further details are omitted. □

Similar results can be given for scale mixtures of normals. Under the conditions of Theorem 1.3, for most pairs $\gamma$, $\beta$ with $\gamma \perp \beta$, the empirical $\theta_{\beta\gamma}$ converges to the bivariate law of $Z\sigma$ where $Z$ is a standard bivariate normal and $\sigma$ is independent of $Z$ with law $F$. For the factor analysis situation, as in Proposition 4.4, the limit of $\theta_{\beta\gamma}$ tends to the law of

$$B_p f + Z_1, \quad \Gamma_p f + Z_2,$$

where $Z_1$ and $Z_2$ are independent normal variables, $f$ has law $F$, independent of $(Z_1, Z_2)$ and

$$B_p = \sum \beta_i, \quad \Gamma_p = \sum \gamma_i.$$

Further details are omitted.

Recall that $\theta_\nu(\gamma)$ is the empirical distribution of the data projected in direction $\gamma$. We view $\theta_\nu(\gamma)$ as a random probability: random because it depends on $\gamma$, which is uniformly distributed over $S_{p-1}$. In particular, $\theta_\nu$ itself has a distribution $\pi_\nu$ that is a probability on the probabilities on $\mathbb{R}^1$. When does $\pi_\nu$ converge? Arguing

as in Proposition 2.1, we can prove the following sufficient conditions:

$$\begin{aligned} \mu_\nu(s) &\quad \text{converges weakly for each } s \\ \mu_\nu(s, t) &\quad \text{converges weakly for each pair } (s, t) \\ \mu_\nu(s, t, u) &\quad \text{converges weakly for each triple } (s, t, u) \\ &\vdots \end{aligned}$$

where

$$\begin{aligned} \mu_\nu(s) &\quad \text{is the empirical of} \quad \| sx_j \|^2/p\text{:} \quad j = 1, \cdots, n \\ \mu_\nu(s, t) &\quad \text{is the empirical of} \quad \| sx_j + tx_k \|^2/p\text{:} \quad j, k = 1, \cdots, n \\ \mu_\nu(s, t, u) &\quad \text{is the empirical of} \quad \| sx_j + tx_k + ux_\ell \|^2/p\text{:} \quad j, k, \ell = 1, \cdots, n \\ &\vdots \end{aligned}$$

Let $\alpha_\nu$ be the three-dimensional empirical distribution of

$$\| x_j \|^2/p, \quad (x_j \cdot x_k)/p, \quad \| x_k \|^2/p. \tag{5.1}$$

At one time, we thought that the weak convergence of $\alpha_\nu$ might suffice for the weak convergence of $\pi_\nu$. This, however, turns out to be false in general, although there may be some germ of truth in it. A counterexample is given in Diaconis and Freedman (1982).

**Acknowledgement.** This work began during a seminar at Harvard with David Donoho and Peter Huber. It owes much to their suggestions. We thank Michael Steele and Ken Alexander for helping with Proposition 5.1.

## REFERENCES

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474.

DIACONIS, P. and FREEDMAN, D. (1982). Asymptotics of graphical projection pursuit. Technical Report No. 195, Department of Statistics, Stanford University.

DONOHO, D., HUBER, P. and THOMA, M. (1981). The use of kinematic displays to represent high dimensional data. In *Computer Science and Statistics, Proceedings of the 14th Annual Symposium on the Interface.* W. Eddy (ed.).

DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

EFRON, B. and OLSHEN, R. (1979). How broad is the class of normal scale mixtures? *Ann. Statist.* **6** 1159–1164.

FELLER, W. (1968). *An Introdution to Probability Theory and Its Applications*, Vol. I. Wiley, New York.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II. Wiley, New York.

FISHERKELLER, M. A., FRIEDMAN, J. H. and TUKEY, J. W. T. (1974). *PRIM*-9 An interactive multidimensional data display system. Stanford Linear Accelerator Pub—1408.

FREEDMAN, D. (1973). Another note on the Borel-Cantelli lemma and the strong law, with the Poisson approximation as a by-product. *Ann. Probab.* **1** 910–925.

FREEDMAN, D. and LANE D. (1980). The empirical distribution of Fourier coefficients. *Ann. Statist.* **8** 1244–1251.

FREEDMAN, D. and LANE, D. (1981). The empirical distribution of the Fourier coefficients of a sequence of independent, identically distributed long-tailed random variables. *Z. Wahrsch. verw. Gebiete* **58** 21–39.

FRIEDMAN, J. and TUKEY, J. W. T. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **9** 881–890.

GEMAN, S. (1980). A limit theorem for the norm of random matrices. *Ann. Probab.* **8** 252–261.

HUBER, P. (1984). Projection pursuit. Technical Report PJH-4, Department of Statistics, Harvard University. *Ann. Statist.*, to appear.

KRUSKAL, J. B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation that optimizes a new index of condensation. In *Statistical Computation.* R. C. Milton and J. A. Nelder (eds.). Academic, New York.

KRUSKAL, J. B. (1972). Linear transformation of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Vol. 1, *Theory.* Seminar Press.

PETROV, V. V. (1975). *Sums of Independent Random Variables.* Springer-Verlag, New York.

RAO, R. RANGA (1962). Relations between weak and uniform convegence of measures with applications. *Ann. Math. Statist.* **33** 659–680.

SMYTHE, R. T. (1973). Strong laws of large numbers for $r$-dimensional arrays of random variables. *Ann. Probab.* **1** 164–170.

VAPNIK, V. N. and CERVONENKIS, A. Y. A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

DE WET, T., VENTER, J. H., and VAN WYCK, J. W. J. (1979). The null distributions of some test criteria of multivariate normality. *South African Statist. J.* **13** 153–176.

WOLFOWITZ, J. (1960). Convergence of the empirical distribution function on half spaces. *Contributions to Probability and Statistics.* Stanford University Press, California.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720