ASYMPTOTICS OF GRAPHICAL PROJECTION PURSUIT

BY

PERSI DIACONIS and DAVID FREEDMAN

TECHNICAL REPORT NO. 195
NOVEMBER 1982

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ASYMPTOTICS OF GRAPHICAL PROJECTION PURSUIT

BY

PERSI DIACONIS and DAVID FREEDMAN

TECHNICAL REPORT NO. 195
November 1982

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Also issued as Technical Report No. 014 ORION, Department of Computer Science, Stanford University.
ASYMPTOTICS OF GRAPHICAL PROJECTION PURSUIT

Persi Diaconis\(^1\)
Department of Statistics
Stanford University
Stanford, California 94305

David Freedman\(^2\)
Department of Statistics
University of California, Berkeley
Berkeley, California 94720

Abstract

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.

Key Words: Projections, projection pursuit, random probabilities, empirical characteristic function.

AMS 1981 Classification Numbers:

\(^1\) Research partially supported by NSF Grant MCS-80-24649.

\(^2\) Research partially supported by NSF Grant MCS-80-02535.
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 entropy such as $-f \log f$ or Fisher information. Huber observes that the numerical criteria used by Friedman and Tukey essentially minimizes $-f^2$, another measure of entropy. 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. We also present classes of data sets where most projections are close to the same non-Gaussian distribution. For such data sets, 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).

The main results will now be stated. Let \( x_1, x_2, \ldots, x_n \) be (non-random) 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 \( v \). As \( v \) tends to infinity, so do \( n \) and \( p \). Suppose that for \( \sigma^2 \) positive and finite, for any positive \( \varepsilon \), as \( v \) tends to infinity,

1. \( \frac{1}{n} \text{card} \{ j \leq n : \|x_j\|^2 - \sigma^2 p \geq \varepsilon p \} \to 0 \)

2. \( \frac{1}{n^2} \text{card} \{ 1 \leq j, k \leq n : x_j \cdot x_k \geq \varepsilon p \} \to 0 \)

Condition (1) says that most vectors have length near \( \sigma^2 p \). Condition (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).

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 \( \gamma \) has coordinates

3. \( \gamma \cdot x_1, \gamma \cdot x_2, \ldots, \gamma \cdot x_n \).

Let \( \theta_v(\gamma) \) be the empirical distribution of this sequence, assigning mass \( 1/n \) to each \( \gamma \cdot x_j \). The first theorem says that \( \theta_v(\gamma) \) is close to \( N(0, \sigma^2) \) for most \( \gamma \), for large \( v \). 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 \( \theta_v(\cdot) \).

**Theorem 1.** Under conditions (1) and (2), as \( v \to \infty \), the empirical distribution \( \theta_v \) tends to \( N(0, \sigma^2) \) weakly in probability.
Theorem 1 is proved in Section 2. The approach is quite similar to the techniques in Freedman and Lane (1979, 1981). Section 3 gives examples where conditions (1) and (2) hold. Section 4 gives examples where most projections are not normal. The random measures $\mathbf{\theta}_V(\gamma)$ 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 continue to hold if the data is 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 is standardized using the mean and standard deviation. Some notation is needed. Let $a_V$, be the mean of the projected data, $s^2_V$ the variance, and $t^2_V$ the second moment. Thus

$$(4) \quad a_V = \frac{1}{n} \sum_{j=1}^{n} \gamma \cdot x_j, \quad s^2_V = \frac{1}{n} \sum_{j=1}^{n} (\gamma \cdot x_j - a_V)^2, \quad t^2_V = \frac{1}{n} \sum (\gamma \cdot x_j)^2$$

Let

$$(5) \quad \bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

The conditions required for Theorem 2 are

$$(6) \quad \frac{1}{np} \sum_{j=1}^{n} \|x_j\|^2 \to \sigma^2 \quad \text{where} \quad 0 < \sigma^2 < \infty$$

$$(7) \quad \frac{1}{n} \text{card} \{j \leq n : \|x_j\|^2 - p\sigma^2 > \epsilon p\} \to 0$$

$$(8) \quad \frac{1}{(np)^2} \sum_{j,k=1}^{n} (x_j \cdot x_k)^2 \to 0$$

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

**Theorem 2.**

(a) Under conditions (6-8), as $\nu \to \infty$,

- the empirical second moment $t_{\nu}^2$ converges to $\sigma^2$ in probability

- the scaled empirical $\theta_{\nu}^1(\gamma)$ converges to $N(0,1)$ weakly in probability.

(b) If conditions (6-8) 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 $\theta_{\nu}^0(\gamma)$ converges to $N(0,\sigma^2)$ weakly in probability

- the standardized empirical $\theta_{\nu}^2(\gamma)$ converges to $N(0,1)$ weakly in probability.

**Remarks.** Of course, part (b) of Theorem 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 $\frac{1}{n} \sum ||x_j - \bar{x}||^2 = p$. The conditions become

\begin{align*}
(1-\epsilon) \frac{1}{n} \sum_{j=1}^{n} ||x_j - \bar{x}||^2 < ||x_k - \bar{x}||^2 < (1+\epsilon) \frac{1}{n} \sum_{j=1}^{n} ||x_j - \bar{x}||^2 \\
\text{except for } o(n) \text{ indices } k = 1,2,\ldots,n.
\end{align*}

\begin{align*}
\frac{1}{n^2} \sum_{j,k=1}^{n} [(x_j - \bar{x}) \cdot (x_k - \bar{x})]^2 = o\left[\frac{1}{n} \sum_{j=1}^{n} ||x_j - \bar{x}||^2 \right]^2.
\end{align*}

**Acknowledgement.** This work began during a seminar at Harvard with David Donoho and Peter Huber. It owes much to their suggestions.
2. Proofs of Theorems 1 and 2.

Let \( \zeta \) be \( N(0,I_p) \), i.e., a vector of \( p \) independent \( N(0,1) \) variables. Then \( \| \zeta \| / \sqrt{p} \) is uniformly distributed over \( S_{p-1} \). On the other hand, \( \| \zeta \| / \sqrt{p} \to 1 \). Hence, it is enough to prove the theorems with \( \gamma \) replaced by \( \zeta / \sqrt{p} \); a variant of Slutsky'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

\[
P\{ |U| \geq \frac{2}{\varepsilon} \} \leq \frac{1}{\varepsilon} \int_{-\varepsilon}^{\varepsilon} [1 - \text{Re} \phi(t)]dt.
\]

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

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

and

\[
\limsup \nu E\{\theta_\nu(T_\delta)\} \leq \frac{1}{\delta} \int_{-\delta}^{\delta} [1 - \phi_0(t)]dt
\]

which is small for \( \delta \) small. Given \( \varepsilon \) positive, there is a positive \( \delta \) so small that
Pr{θν(Tδ) < ε} > 1 - ε for all ν.

Thus, \{θν\} is tight. □

Now let θν(ζ) be the empirical measure of

\((ζ \cdot x_1)/\sqrt{p}, \ldots, (ζ \cdot x_n)/\sqrt{p}\).

**Proposition 2.1.** Under conditions (1-2), θν → N(0,σ²) weakly in probability.

**Proof.** The characteristic function of θν(ζ) is

\[
φ_ν(ζ,t) = \frac{1}{n} \sum_{j=1}^{n} \exp\{ i \cdot t \cdot (ζ \cdot x_j)/\sqrt{p}\}.
\]

Clearly,

\[
E[φ_ν(ζ,t)] = \frac{1}{n} \sum_{j=1}^{n} \exp\{- \frac{1}{2} t^2 \cdot ||x_j||^2/p\}
\]

\[\rightarrow \exp\{- \frac{1}{2} t^2 \cdot σ^2\}\]

by condition (1). Likewise

\[
E[|φ_ν(ζ,t)|^2] = E[φ_ν(ζ,t)\bar{φ}_ν(ζ,t)]
\]

\[
= \frac{1}{n^2} \sum_{j,k} E[\exp\{- i \cdot t \cdot (ζ \cdot (x_j-x_k))/\sqrt{p}\}]
\]

\[= \frac{1}{n^2} \sum_{j,k} \exp\{- \frac{1}{2} t^2 \cdot ||x_j-x_k||^2/p\} .
\]

Of course,

\[||x_j-x_k||^2 = ||x_j||^2 + ||x_k||^2 - 2 x_j \cdot x_k .
\]

The summands in (2.3) are between 0 and 1. By conditions (1) and (2), except for a set of pairs (j,k) of cardinality o(n²), 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
\[
E[|\phi_v(\cdot,t)|^2] \to \exp[-t^2 \sigma^2]
\]

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

To prove Theorem 2, another estimate is needed.

**Lemma 2.3.** \( E[(\xi \cdot x_j)^2 (\xi \cdot x_k)^2] \) 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, \( \xi \cdot x_j \) 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.,

\[
x_j = \alpha x_k + \delta, \text{ where } \delta \perp x_k
\]

As usual,
\[
\|x_k\|^2 \alpha = x_j \cdot x_k
\]
\[
\|x_k\|^2 \|\delta\|^2 = \|x_j\|^2 \|x_k\|^2 - (x_j \cdot x_k)^2
\]

Now
\[ \zeta \cdot x_j = \alpha \zeta \cdot x_k + \zeta \cdot \delta, \]

so

\[
(\zeta \cdot x_j)^2 (\zeta \cdot x_k)^2 = \alpha^2 (\zeta \cdot x_k)^4 + 2\alpha (\zeta \cdot x_k)^3 (\zeta \cdot \delta) + (\zeta \cdot x_k)^2 (\zeta \cdot \delta)^2. 
\]

But \( \zeta \cdot x_k \) and \( \zeta \cdot \delta \) are independent, so

\[
E[(\zeta \cdot x_j)^2 (\zeta \cdot x_k)^2] = 3\alpha^2 \| x_k \|^4 + 2\alpha \| x_k \| \| \delta \|^2 
= 2(x_j \cdot x_k)^2 + \| x_j \|^2 \| x_k \|^2. \quad \Box
\]

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

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

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

**Lemma 2.4.**

a) \( \nu p E[t_{\nu}(\zeta)^2] = \sum_{j=1}^{n} \| x_j \|^2 \)

b) \( \nu^2 p^2 E[t_{\nu}(\zeta)^4] = 2\sum_j (x_j \cdot x_k)^2 + \sum_j \| x_j \|^2 \| x_k \|^2 \)

**Proof.** Claim a) is easy. For b),

\[
E[t_{\nu}(\zeta)^4] = \frac{1}{\nu^2 p^2} \sum_{jk} E[(\zeta \cdot x_j)^2 (\zeta \cdot x_k)^2].
\]

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
\[ 2 \sum_{jk} (x_j \cdot x_k)^2 + \left( \sum_j \|x_j\|^2 \right)^2 \]

As before, let \( \theta_1^{\nu}(\zeta) \) be the scaled empirical.

**Proposition 2.2.** Under conditions (6-7-8), the empirical second moment \( t^2_{\nu} \) converges to \( \sigma^2 \) in probability, and the scaled empirical distribution \( \theta_1^{\nu}(\zeta) \) converges to \( N(0,\sigma^2) \) weakly in probability.

**Proof.** That \( t^2_{\nu} \to \sigma^2 \) follow from Lemma 2.4, and \( \theta_1^{\nu} \to N(0,\sigma^2) \) by Proposition 2.1; then \( \theta_1^{\nu} \) can be handled, in effect by Slutsky'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

\[
X_j = \begin{pmatrix} X_{ij} \\ X_{2j} \\ \vdots \\ X_{nj} \end{pmatrix}, \quad 1 \leq j \leq n.
\]

The first example shows that conditions (1) and (2) hold almost surely for independent identically distributed (iid) coordinates.

**Example 3.1: iid coordinates.** Let \( X_{ij} \) be iid for \( i = 1, 2, \ldots \), and \( j = 1, 2, \ldots \). Suppose

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

Then, for almost all realizations of the array \( \{X_{ij}\} \), conditions (6,7,8) and so conditions (1,2) are satisfied, no matter how \( n \) and \( p \) tend to infinity.
Proof. Condition (6) is easy

\[(3.2) \quad \frac{1}{np} \sum_{i=1}^{p} \sum_{j=1}^{p} x_{ij}^2 - \sigma^2 \rightarrow a.e.\]

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 (7), fix \(\varepsilon > 0\). Let \(\xi_{pj}\) be 1 if

\[\left| \frac{1}{p} \sum_{i=1}^{p} x_{ij}^2 - \sigma^2 \right| < \varepsilon\]

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

\[(3.3) \quad \lim_{n,p \to \infty} \frac{1}{n} \sum_{j=1}^{n} \xi_{pj} = 0 \quad a.e.\]

Suppose first \(E(x_{ij}^4) < \infty\). Fix \(\delta\) positive but small. Let \(A_{pn}\) be the event \(\sum_{j=1}^{n} \xi_{pj} \geq \delta n\). 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, \(\pi_p \leq A/p\) where \(A = \text{Var}(X_{ij}^2) \leq E(X_{ij}^4)\). By a version of Bernstein's inequality,

\[P(A_{pn}) \leq (e \pi_p / \delta)^{\delta n} \leq (Ae / p \delta)^{\delta n}\]

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

\[2(Ae / p \delta)^{\delta N}\]

If \(N > 1/\delta\), 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

$$Y_{ij} = X_{ij} \quad \text{when} \quad |X_{ij}| \leq L$$

$$= 0 \quad \text{when} \quad |X_{ij}| > L$$

$$Z_{ij} = X_{ij} - Y_{ij}$$

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, $\frac{1}{p} \sum_{i=1}^{p} Y_{ij}^2$ can be dealt with under the fourth-moment condition. On the other hand,

$$\sup \frac{1}{p} \sum_{i=1}^{p} Z_{ij}^2 = V_j$$

are independent and identically distributed in $j$; while $E(V_j) \leq 4E(Z_{ij}^2)$ is small so $\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} V_j = E(V_j)$ is small. See [Doob, 1953, Theorem 3.4 on p. 317].

This completes the argument for condition (7).

We turn now to condition (8). It is convenient to deal first with the term $j = k$. We claim

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

Let $V_j = \sup_p \frac{1}{p} \sum_{j=1}^{n} X_{ij}^2$. Then $V_j \in L_{1+\delta}$, and the $V_j$ are independent and identically distributed. Even if the $V_j$ were just $L_1$ and identically distributed,

$$\frac{1}{n} \max_{j=1, \ldots, n} V_j \to 0 \quad \text{a.e.}$$

So
\[ \frac{1}{n^2} \sum_{j=1}^{n} \left( \frac{1}{p} \sum_{i=1}^{p} x_{ij}^2 \right)^2 \leq \left( \frac{1}{n} \max_{j=1, \ldots, n} V_j \right) \times \left( \frac{1}{np} \sum_{i=1}^{p} \sum_{j=1}^{n} x_{ij}^2 \right). \]

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

We now take up the terms \( j \neq k \) in (8). We claim

\[
\lim_{n,p \to \infty} \frac{1}{n^2 p^2} \sum_{1 \leq j < k \leq n} \left( \sum_{i=1}^{p} x_{ij} x_{ik} \right)^2 = 0 \text{ a.e.}
\]

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

\[ T_{np} = \frac{1}{n(n-1)} \sum_{1 \leq j < k \leq n} h_p(x_j, x_k) \]

where

\[ X_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{pj} \end{bmatrix} \]

and

\[ h_p(x,y) = \frac{1}{p^2} (x \cdot y)^2 - \frac{\sigma^4}{p} \]

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

\[ \mathbb{E}\{h_p(X_j, X_k)\} = 0, \]

\[ \text{Var}\{h_p(X_j, X_k)\} = \frac{5\sigma^8}{p^2} + \frac{\mu_4}{p^3} \]

Let
\[ \alpha_p(x) = E_h(x, x_k) = \frac{\sigma^2 \|X\|_2}{p^2} - \frac{\sigma^4}{p} \]

so

\[ \text{Var} \alpha_p(x_j) = \frac{\sigma^4}{p^3} (\mu_4 - \sigma^4) \]

Let

\[ h_p^*(x, y) = h_p(x, y) - \alpha_p(x) - \alpha_p(y) \]

Then

\[ \alpha_p^*(x) = E(h_p^*(x, x_j)) = E(h_p^*(x_j, y)) = 0 \]

\[ \text{Var}(h_p^*(x_j, x_k)) \leq A/p^2 \]

for some constant A. Now

\[ T_{np} = \frac{2}{n} \sum_{j=1}^{n} \alpha_p(x_j) + T_{np}^* \]

where

\[ \frac{2}{n} \sum_{j=1}^{n} \alpha_p(x_j) = \frac{2\sigma^2}{p^2} \sum_{i=1}^{p} \sum_{j=1}^{n} (x_{ij}^2 - \sigma^2) \to 0 \text{ a.e.} \]

\[ T_{np}^* = \frac{1}{n(n-1)} \sum_{1 \leq j < k \leq n} h_p^*(x_j, x_k) \]

has mean 0 and variance

\[ \frac{2}{n(n-1)} \text{Var}(h_p^*(x_j, x_k)) \leq \frac{B}{n^2 p^2} \]

for some constant B. This is the key point; the reason is that \( \alpha_p^* = 0 \). The upshot is that
\[ \sum_{np} \mathbb{P}\{|T_{np}^*| > \varepsilon\} < \infty, \text{ so } T_{np}^* \to 0 \text{ a.e.} \]

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

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

Let

\[ Y_{ij} = X_{ij} \quad \text{when} \quad |X_{ij}| \leq L \]

\[ = 0 \quad \text{otherwise} \]

\[ Z_{ij} = X_{ij} - Y_{ij} \]

\[ A_{jk} = \frac{X_j \cdot X_k}{p} \]

\[ = \frac{1}{p} \sum_{i=1}^{p} X_{ij} X_{ik} \]

\[ = B_{jk} + C_{jk} + D_{jk} + F_{jk} \]

where

\[ B_{jk} = \frac{1}{p} \sum_{i=1}^{p} Y_{ij} Y_{ik} \]

\[ C_{jk} = \frac{1}{p} \sum_{i=1}^{p} Y_{ij} Z_{ik} \]

\[ D_{jk} = \frac{1}{p} \sum_{i=1}^{p} Z_{ij} Y_{ik} \]

\[ F_{jk} = \frac{1}{p} \sum_{i=1}^{p} Z_{ij} Z_{ik} \]

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
\[
\limsup_{n,p \to \infty} \frac{1}{n^2} \sum_{jk=1}^{n} C_{jk}^2 \leq \varepsilon_L \text{ a.e.}
\]

where \( \varepsilon_L \to 0 \) as \( L \to \infty \). Indeed,

\[
C_{jk}^2 \leq \left( \frac{1}{p} \sum_{i=1}^{p} Y_{ij}^2 \right) \times \left( \frac{1}{p} \sum_{i=1}^{p} Z_{ik}^2 \right)
\]

So

\[
\frac{1}{n^2} \sum_{jk=1}^{n} C_{jk}^2 \leq \left( \frac{1}{np} \sum_{i=1}^{p} \sum_{j=1}^{n} Y_{ij}^2 \right) \times \left( \frac{1}{np} \sum_{i=1}^{p} \sum_{k=1}^{n} Z_{ik}^2 \right)
\]

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

\[
\limsup_{n,p \to \infty} \frac{1}{n^2} \sum_{jk=1}^{n} B_{jk}C_{jk} \leq \delta_L \text{ a.e.}
\]

where \( \delta_L \to 0 \) as \( L \to \infty \). Indeed,

\[
\left( \frac{1}{n^2} \sum_{jk=1}^{n} B_{jk}C_{jk} \right)^2 \leq \left( \frac{1}{n^2} \sum_{jk=1}^{n} B_{jk}^2 \right) \times \left( \frac{1}{n^2} \sum_{jk=1}^{n} C_{jk}^2 \right)
\]

The first factor on the right goes to 0 a.e. by (3.4-5); the second factor is under control by (3.6). This proves (3.7). Likewise for \( D \) and \( F \). This completes the verification of condition (8).

Conditions (6,7,8) hold for the centered data \( X_j - \bar{X} \), where \( \bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j \), under the assumptions (3.1). This is comparatively easy to deduce from example 3.1. One useful fact:
\[
\frac{1}{p} \sum_{i=1}^{P} \left( \frac{1}{n} \sum_{j=1}^{n} X_{ij} \right)^2 \rightarrow 0 \text{ a.e.}
\]

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

\[
S_i = \sup_{n>n_0} \left( \frac{1}{n} \sum_{j=1}^{n} X_{ij} \right)^2
\]

Then the \( S_i \) are iid and

\[
E(S_i) \leq 4E\left(\left( \frac{1}{n_0} \sum_{j=1}^{n_0} X_{ij} \right)^2\right)
\]

\[
= 4 \sigma^2/n_0
\]

So for \( n > n_0 \),

\[
\limsup \frac{1}{p} \sum_{i=1}^{P} \left( \frac{1}{n} \sum_{j=1}^{n} X_{ij} \right)^2 \leq 4\sigma^2/n_0 \text{ a.e.}
\]

See [Doob, 1953, Theorem 3.4 on p. 317]. \( \square \)

**Remark.** Let \( \theta(np\zeta) \) be the empirical distribution of

\[
\zeta \cdot X_1/\sqrt{p}, \ldots, \zeta \cdot X_n/\sqrt{p}
\]

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

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

Under condition (3.1), \( \theta(np\zeta) \rightarrow N(0,\sigma^2) \) 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 $(n_\nu, p_\nu): \nu = 1, 2, \ldots$. 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 $\|z\|$. 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 $j = 1, 2, \ldots$. In fact,
the processes $\{S_{pj}: p = 1, 2, \ldots\}$ are independent in $j$, but this is immaterial
here. For each $j$, the variables $S_{pj}: p = 1, 2, \ldots$, are dependent, but nearly
independent for widely separated $p$'s. Let

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

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 $\Theta(npXz)$ 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 $(n_\nu, p_\nu)$? We can find $\nu_n$ so
large that
\[ P\{A_{p,n} \text{ for at least one } p \leq v_n\} \geq 1 - \frac{1}{n^2} \]

Now consider the path \((n, p, v)\) that results from stacking the indices in the following order:

\[
\begin{array}{ccccccccc}
1 & 1 & & & & & & & \\
1 & 2 & & & & & & & \\
& & & \vdots & & & & & \\
& & 1 & \uparrow v_1 & & & & & \\
2 & 1 & & & & & & & \\
2 & 2 & & & & & & & \\
& & & \vdots & & & & & \\
& & 2 & \uparrow v_2 & & & & & \\
& & & & & \vdots & & & \\
\end{array}
\]

It follows that \( P(A_{n, i, 0}) = 1 \), defeating almost sure convergence.

In principle, it is possible to get bounds on the rates of convergence in Theorems 1 and 2, using Chebychev's inequality and Esseen's smoothing lemma (Feller, 1971, p. 536). If

\[
\frac{1}{n} \text{card } \{j : 1 \leq j \leq n \text{ and } \|x_j\|^2 - \sigma^2 p > \varepsilon p\} < \varepsilon
\]

and

\[
\frac{1}{n^2} \text{card } \{j, k : 1 \leq j, k \leq n \text{ and } |x_j \cdot x_k| > \varepsilon p\} < \varepsilon
\]

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

\[ \gamma \cdot x_1, \ldots, \gamma \cdot x_n \]
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 \( \|F-G\| = \sup_J |F(J) - G(J)| \) 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

\[
\gamma X_1, \gamma X_2, \ldots, \gamma X_n
\]

where

\[
X_j = \begin{bmatrix}
X_{ij} \\
\vdots \\
X_{pj}
\end{bmatrix}
\]

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 \( \sqrt{p} V_p(\gamma) \) 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 $\gamma^*X_j$. Clearly,

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

where

$$U_{np\gamma} = \|\theta(\gamma) - F(\gamma)\|$$

$$W_{np\gamma} = \|F(\gamma) - \phi\|$$

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, $W_{np} \leq V_p(\gamma)$ where

$$V_p(\gamma) = K_3 \sum_{i=1}^p |\gamma_i|^3$$

and $K$ is a universal positive constant: see (Petrov, 1972, p. 111). We must now demonstrate the limiting behavior of $\sqrt{p} V_p(\gamma)$. Let $Z_1, Z_2, \ldots$, be independent standard normal variables. Then $\sqrt{p} V_p(\gamma)$ is distributed as $K_3$ times

$$\frac{1}{p} \sum_{i=1}^p |Z_1|^3 / \frac{1}{p} \sum_{i=1}^p |Z_i|^2$$

which converges a.e. to $E(|Z_1|^3) = 4/\sqrt{2\pi}$. \qed

4. Examples of Non-Gaussian Projections.

Theorems 1 and 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 $X_i$ be independent, with common symmetric stable density of index $\alpha < 2$, having characteristic function $\exp(-|t|^{\alpha})$. Let $\Theta(npXY)$ be the empirical distribution of

$$p^{1/2} \gamma \cdot X_i/p^{1/\alpha}, \ldots, p^{1/2} \gamma \cdot X_n/p^{1/\alpha}$$

where

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

and $\gamma$ is uniform on the unit sphere $S_{p-1}$ in $\mathbb{R}^p$, 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, $\Theta(npXY)$ converges weakly in probability to a symmetric stable law of index $\alpha$, having characteristic function $\psi(t) = \exp(-C_{\alpha}|t|^{\alpha})$.

**Proof.** Let $\phi_{npXY}(t)$ be the empirical characteristic function

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

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

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

As is easily seen,

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

Hence, $\phi_{npXY}(t) \rightarrow \psi(t)$ in probability. Likewise, $|\phi_{npXY}(t)|^2 \rightarrow \psi(t)^2$. □
Remarks. Replace $\sqrt{p} \gamma$ by $\zeta$ and take expectations to get $\phi_{npX}$. This will not converge a.e.: see the corresponding remark in Section 3. 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 projections converge in probability to a scale mixture of normals. To be more precise, let $F_\nu$ 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

$$F_\nu \Rightarrow F \quad \text{weakly}$$

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 $Y$ 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 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 $\sigma_1, \sigma_2, \ldots$, 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) $\frac{1}{np} \sum_{j=1}^{n} \|X_j\|^2 \to \sigma^2$

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

(c) $\frac{1}{np^2} \sum_{j=1}^{n} (X_j \cdot X_k)^2 \to 0$.

Thus, condition (1) fails, but (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 $\gamma X_j$ is close to the $F$-scale mixture of normals. Further, the empirical mean of the projections $\gamma X_j$ 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) and (2) fail, and so do the conclusion of Theorem 1; indeed, the empirical distribution $\Theta(\gamma)$ depends strongly on $\gamma$. To see this, consider nonrandom $p$-vectors $x_1, x_2, \ldots, x_n$. Define

$$f_j = \frac{1}{p} \sum_{i=1}^{p} x_{ij} \quad \text{and} \quad \varepsilon_{ij} = x_{ij} - f_j$$

The following conditions are assumed:

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

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

$$\varepsilon_j = \begin{bmatrix} \varepsilon_{ij} \\ \vdots \\ \varepsilon_{pj} \end{bmatrix}$$

For distribution functions $G$ and $H$, recall $\|G - H\| = \sup_J |G(J) - H(J)|$ where $J$ ranges over intervals.
Proposition 4.4. Assume (4.2-3). Let $\theta_\nu(\gamma)$ be the empirical distribution of $\gamma^*x_1, \ldots, \gamma^*x_n$. 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

$$(f_j, \gamma^*e_j): j = 1, \ldots, n$$

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

(4.4) $$\theta_\nu^{(2)} + \psi^{(2)} \text{ weakly in probability}$$

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

$$\frac{1}{n} \sum_{j=1}^n \exp[itf_j + iu\gamma^*e_j / \sqrt{p}]$$

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

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

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

24
Clearly,
\[ \gamma \mathcal{X}_j = \Gamma_p f_j + \gamma \varepsilon_j \]

Let \( J \) be a linear interval. So \( \gamma \mathcal{X}_j \in J \) iff \( (f_j, \gamma \varepsilon_j) \) falls in the convex set \( \{(u,v): \Gamma_p u + v \in J\} \)

Thus
\[ \|\theta_{\psi_1}(\gamma) - \psi_1(\gamma)\| \leq \|\theta_{\psi_1}(\gamma) - \psi_1(\gamma)\| . \]

\[ \square \]

\textbf{Remarks.} a) 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 early verified, the law of \( \Gamma_p U + Z \) determines \( \Gamma_p \).) Fix a sequence \( \zeta_1, \zeta_2, \ldots \) of independent standard normals; realize \( \gamma_i \) as \( \zeta_i/\|\zeta\| \). Even so, \( \theta_\psi \) converges in law but not in probability: because the same is true of \( \Gamma_p \).

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

\[ \|x_j\|^2 = f_j^2 + \|\varepsilon_j\|^2 = f_j^2 + \sigma^2 \]

is strongly dependent on \( j \); and

\[ \frac{1}{p} x_j \cdot x_k = f_j f_k + \frac{1}{p} \varepsilon_j \varepsilon_k \]

Both (1) and (2) fail.

c) Consider the one-factor model
\[ X_{ij} = U_j + V_{ij} . \]
Suppose the $U_i$ are independent with common distribution $F$; the $V_{ij}$ are iid with mean $U$, variance $\sigma^2$ and finite $2+\delta$th moment. Then conditions (4.2-3) hold for almost all realizations of $X$; independence of $U$ and $V$ is not required, nor any moment condition on $U$. Indeed, let

$$V_{ij} = \frac{1}{p} \sum_{i=1}^{p} V_{ij}$$

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

$$f_j = \frac{1}{p} \sum_{i=1}^{p} X_{ij} = U_j + V_{ij}$$

$$\varepsilon_{ij} = X_{ij} - f_j = V_{ij} - V_{ij}.$$ 

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

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

$$\frac{(\Gamma \ U + Z) / \sqrt{\frac{\gamma^2}{P} \tau^2 + \sigma^2}}{}.$$

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

e) What happens if we scale the vectors separately? The idea is to make condition (1) hold by brute force, replacing $x_j$ by

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

We assume conditions (4.2-3) hold for $x_j$. Recall (4.6-7):
\[ \frac{1}{p} \hat{x}_j \hat{x}_k = \frac{f_j}{\sqrt{f_j^2 + \sigma^2}} \times \frac{f_k}{\sqrt{f_k^2 + \sigma^2}} \]

and condition (2) fails. We turn to the asymptotic behavior of \( \hat{\psi}_\nu(\gamma) \), the empirical distribution for \( j = 1, \ldots, n \) of

\[ \gamma \hat{x}_j = \frac{(\Gamma f_j + \gamma \varepsilon_j)}{\sqrt{f_j^2 + \varepsilon_j^2}} \]

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

\[ \frac{(\Gamma^j U + Z)}{\sqrt{U^2 + \sigma^2}} \]

To make this precise, introduce a metric for weak convergence like 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

\[ \mu(K) \leq \mu'(e^\varepsilon K) + \varepsilon \]

\[ \mu'(K) \leq \mu(e^\varepsilon K) + \varepsilon \]

Clearly, \( d(\mu, \mu') \leq \|\mu - \mu'\| \).

We claim

\[ (4.8) \quad d(\hat{\psi}_\nu, \bar{\psi}_\nu) \to 0 \text{ in probability as } \nu \to \infty. \]

Indeed,

\[ d(\hat{\psi}_\nu, \bar{\psi}_\nu) \leq d(\hat{\psi}_\nu, \hat{\psi}_\nu) + d(\hat{\psi}_\nu, \bar{\psi}_\nu) \]

where \( \hat{\psi}_\nu \) is the empirical distribution of

\[ \frac{(\Gamma f_j + \gamma \varepsilon_j)}{\sqrt{f_j^2 + \sigma^2}} \]

27
Fix real \( a, b, c \). Now the planer set

\[
\{(x,y) : (ax+y)/\sqrt{a^2+\sigma^2} \leq c \}
\]

either is convex or has a convex complement. So

\[
d(\tilde{\Theta}_v, \tilde{\Psi}_v) \leq \| \tilde{\Theta}_v - \tilde{\Psi}_v \|
\]

\[
\leq \| \Theta_v^{(2)} - \Psi_v^{(2)} \|
\]

\[\rightarrow 0 \text{ in probability}\]

by (4.5). But \( d(\tilde{\Theta}_v, \tilde{\Theta}_v) \rightarrow 0 \) too, because for large \( n \) except for \( o(n) \) indices \( j = 1, \ldots, n \), by (4.3)

\[
e^{-2\varepsilon \sigma^2} < \frac{1}{p} \| \Xi_j \|^2 < e^{2\varepsilon \sigma^2}
\]

so

\[
e^{-\sqrt{f_j^2 + \sigma^2}} < \sqrt{f_j^2 + \frac{1}{p} \| \Xi_j \|^2} < e^{\sqrt{f_j^2 + \sigma^2}}
\]

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

**Example 4.7.** Clustered data. This example determines the behavior of projections of data clustered about \( k \) centers. The following assumptions will be made:

(4.4) Let \( c_1, c_2, \ldots, c_k \) be distinct p-vectors.

(4.5) Let \( V_{ij} \) be iid with mean zero, variance \( \sigma^2 \) and a finite 3rd moment. Let \( V_1 = (V_{i1}, \ldots, V_{ip})^T \).

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

\[
n_0 = 0 < n_1 < n_2 < \ldots < n_k = n
\]
with

\[ \frac{n_i}{n} \rightarrow \lambda_i, \quad \lambda_0 = 0 < \lambda_1 < \lambda_2 < \ldots < \lambda_{k-1} < 1 = \lambda_k \quad \text{as} \quad n \rightarrow \infty. \]

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

**Proposition 4.5.** Assume (4.4-4.6). For \( i = 1, \ldots, k \) let \( \psi_i' \) be the law of a normal variable having mean \( \gamma \cdot c_i \) and variance \( \sigma^2 \). Let \( \psi_i' \) be the mixture of \( \psi_i' \) with weights \( \lambda_i - \lambda_{i-1} \). Let \( \theta_{\gamma} \) be the empirical measure of \( \gamma \cdot X_1, \ldots, \gamma \cdot X_n \).

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 \( \theta_{\gamma}^j \) 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 \( \theta_{\gamma}^j \) and a normal \((0, \sigma^2)\) variable tends to zero almost surely, in probability. The empirical \( \theta_{\gamma} \) is a mixture of \( \theta_{\gamma}^j \) 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 zero's elsewhere. Thus the distance between the vertices was 10. The 65 vectors from the \( i \)-th cluster thus project to points of the form

\[ \gamma_i \cdot \frac{10}{\sqrt{2}} + Z \]

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

29
will look like the result of choosing 15 independent centers \( \gamma_1 10/\sqrt{2} \) 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 non-normal 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 below shows three 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, based on work of J. M. Steele implies that if \( n \) and \( p \) tend to infinity in such a way that \( p/\sqrt{n}/\log \log n \to 0 \), then the least normal projection is close to normal.

**Proposition 5.1.** Let \( F \) be a probability distribution on \( \mathbb{R}^p \) with a bounded density. Let \( X_1, X_2, \ldots, X_n \) be a sample from \( F \) with empirical measure \( F_n \). Let \( F^Y \) denote the law of \( \gamma X_1 \). Let

\[
D = \sup_{\gamma} \sup_{t} |F^Y - F_n^Y|.
\]

If \( p < \delta \sqrt{n}/\log \log n \), then

\[
P(D > \epsilon) < ce^{-An}
\]

where \( c \) is a constant depending only on the density of \( F \) and \( A = \frac{e^2}{8} - \delta \).
Figure 1. Histograms of a highly non-normal projection from three samples of 50 from a 10 dimensional spherically symmetric normal distribution.

Proof. Theorem 2.3 of Steele (1975) gives an explicit bound on the maximum difference between $F_n$ and $F$ taken over all half spaces. In the notation above, Steele shows that there is a constant $c$ depending only on the density of $F$ such that for any $\Lambda > 0$,

$$P\{D > \Lambda/\sqrt{n}\} \leq c(\log n)^p e^{-\Lambda^2/8}.$$  

Choose $\Lambda = \epsilon \sqrt{n}$, the given bound follows.

We do not know how large $p$ can get and still have even the worst deviation from normal be small. Arguments of Ken Alexander suggest that $p$ can be as large as $n/\log n$. It follows from work of Gehman (1980) that for normal $X_i$, the maximum variance of $\gamma_X^1, \ldots, \gamma_X^N$ is almost surely $(1 + \sqrt{\gamma})$ if $p = \gamma n.$
de Vet, 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 $(\beta \cdot x_1, \gamma \cdot x_1, \ldots, \beta \cdot x_n, \gamma \cdot x_n)$. 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 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

$$B_p = \Sigma_{B_1}, \quad \Gamma_p = \Sigma_{Y_1}.$$ 

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:

\[ \mu_v(s) \text{ converges weakly for each } s \]
\[ \mu_v(s,t) \text{ converges weakly for each pair } (s,t) \]
\[ \mu_v(s,t,u) \text{ converges weakly for each triple } (s,t,u) \]
\[ \vdots \]

where

\[ \mu_v(s) \text{ is the empirical of } \| tx_j \|^2/p : j = 1, \ldots, n \]
\[ \mu_v(s,t) \text{ is the empirical of } \| sx_j + tx_k \|^2/p : j,k = 1, \ldots, n \]
\[ \mu_v(s,t,u) \text{ is the empirical of } \| sx_j + tx_k + nx_\ell \|^2/p : j,k,\ell = 1, \ldots, n \]
\[ \vdots \]

Let \( \alpha_v \) be the three-dimensional empirical distribution of

\[(5.1) \quad \| x_j \|^2/p, \quad x_j \cdot x_k / p, \quad \| x_k \|^2 / p . \]

At one time, we thought that the weak convergence of \( \alpha_v \) might suffice for the weak convergence of \( \pi_v \). This, however, turns out to be false in general, although there may be some germ of truth in it.

We now present a counterexample; here is the key preliminary:

If \( \beta \) is a probability on \( \mathbb{R}^3 \), let \( \hat{\beta} \) be the distribution of \( U \cdot V \), where \( U \) and \( V \) are independent with common distribution \( \beta \). Thus, \( \hat{\beta} \) is a probability on \( \mathbb{R}^1 \). Unfortunately, in general, \( \hat{\beta} \) does not determine \( \beta \), even up to rotations.

**Proposition 5.3.** There are two probabilities \( \beta_1 \) and \( \beta_2 \) on the unit sphere \( S_2 \) in \( \mathbb{R}^3 \) such that \( \hat{\beta}_1 = \hat{\beta}_2 \), but \( \beta_1 \) cannot be rotated into \( \beta_2 \). These probabilities have infinitely differentiable densities relative to surface area.

**Proof.** Choose three points \( c_1, c_2, c_3 \) on \( S_2 \) such that each pair subtends the same angle \( \theta \) at the origin, with \( 0 < \theta \leq \pi/2 \). Let \( C_i \) be a spherical cap
centered at $c_i$, but $C_i$ is not all of $S_2$; make all these $C_i$'s of the same radius. Let $f_i$ be a positive infinitely-differentiable function vanishing off $C_i$, circularly symmetric about $c_i$, integrating to 1 with respect to surface area. Suppose the three $f_i$'s can be rotated one onto the other. Let $\beta_1$ have density $\frac{1}{2} f_1 + \frac{1}{2} f_2$, while $\beta_2$ has density $\frac{2}{3} f_1 + \frac{1}{6} f_2 + \frac{1}{6} f_3$.

To verify that $\hat{\beta}_1 = \hat{\beta}_2$, let $\lambda_0$ be the law of $X \cdot Y$ when $X$ and $Y$ are independent, with common density $f_1$; let $\lambda_1$ be the law of $X \cdot Y$ when $X$ and $Y$ are independent, $X$ having density $f_1$ and $Y$ having density $f_2$. By symmetry,

$$\hat{\beta}_1 = \frac{1}{2} \hat{\lambda}_0 + \frac{1}{2} \lambda_1.$$ 

With respect to $\beta_2 \times \beta_2$, the chance that $X$ and $Y$ are both chosen from the same cap is

$$\left(\frac{2}{3}\right)^2 + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^2 = \frac{1}{2}.$$

Thus,

$$\hat{\beta}_2 = \frac{1}{2} \hat{\lambda}_0 + \frac{1}{2} \lambda_1 = \hat{\beta}_1.$$ 

Proposition 5.4. Let $\alpha_\nu$ be defined by (5.1). There are data sets consisting of $n_\nu$ vectors in $p_\nu$-dimensional space, all having the same length $\sqrt{p_\nu}$, such that $\alpha_\nu$ converges weakly but $\pi_\nu$ does not.

Proof. We use the $\beta_1$ and $\beta_2$ constructed in Proposition 5.3, but scaled up so they sit on vectors of length $\sqrt{3}$. Start with $p_1 = 3$ and $n_1$ large; the data vectors are sampled iid from $\beta_1$. The empirical distribution of $x_j \cdot x_k / 3$ is essentially $\hat{\beta}_1$. Now switch to $p_2 = 3$; again, $n_2$ is large; the data vectors have

coordinate $i = \text{coordinate i-3 for } i = 4, 5, 6$ 
coordinates $(1, 2, 3)$ sampled iid from $\beta_2$. 

34
In particular, the empirical distribution of $x_j^*x_k/6$ is essentially $\hat{\beta}_2 = \hat{\beta}_1$. And so on. Thus, $\alpha_\nu$ converges; but $\pi_\nu$ oscillates between a random projection of $\beta_1$ and a random projection of $\beta_2$. If the circular caps are small in diameter relative to the angle $\theta$ between their centers, a typical projection of $\beta_1$ is bimodal; of $\beta_2$, trimodal.

References


