ROBUSTNESS AND HIGH DIMENSIONAL DATA

Peter J. Bickel

UC Berkeley

MSU 9/2012

(Joint with Boaz Nadler, Bin Yu, N. el Karoui, Derek Bean and Chingway Lim)

Outline

1. Robust M estimation in Linear Regression for fixed number of covariates p
2. What is known if, $\frac{p}{n} \to 0$, $p \to \infty$?
3. Least squares and Lasso: Some current results
4. Some curious simulations
5. Heuristics
6. Projection Pursuit
7. Discussion

The Regression Model and Least Squares Data

$$X_i = (\mathbf{Z}_i, Y_i) \quad i = 1, \ldots, n \quad \mathbf{Z}_i \quad p \times 1 \text{ iid}$$

Assumed model: $(n = 1)$

$$Y = \mathbf{Z}^T \beta_0 + \mathbf{e}$$

$e \perp \mathbf{Z}$

Used as an approximation to general model

$$Y = \mu(\mathbf{Z}) + e, \quad E(e|\mathbf{Z}) \equiv 0$$

Basic Theorem

If $\hat{\beta} = \arg\min\{\|Y - \mathbf{Z}^T\beta\|_n^2\}$ where $\|f(X)\|_n \equiv \frac{1}{n}\sum_{i=1}^n f^2(X_i)$

a) $\hat{\beta} = \left[\frac{1}{n}(\mathcal{Z}^{(n)}[\mathcal{Z}^{(n)}]^T)\right]^{-1}(\mathbf{Z}^{(n)}, Y)_{(n)}$
where $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)^T$, $(\mathbf{Z}^{(n)}, \mathbf{Y})_{(n)} \equiv \frac{1}{n}\sum_{i=1}^n Y_i\mathbf{Z}_i$
$\mathcal{Z}^{(n)}_{p\times n} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$
$\mathcal{Z}^{(n)}[\mathcal{Z}^{(n)}]^T = \sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^T$

b) If $\Sigma \equiv E(\mathbf{Z}\mathbf{Z}^T)$ is nonsingular, β_0 is TRUE

$$\sqrt{n}(\hat{\beta} - \beta_0) \Longrightarrow N(\mathbf{0}, \sigma^2\Sigma^{-1})$$
$$\beta_0 = \Sigma^{-1}\left[E(Y\mathbf{Z})\right]$$

Robust M Estimation in Regression

$$\boldsymbol{\beta}_0 = \arg\min E_0 \rho(Y - \mathbf{Z}^T \boldsymbol{\beta})$$

$\boxed{p \text{ fixed}}$

ρ convex, symmetric about 0.

$\hat{\boldsymbol{\beta}}_\rho \equiv \arg\min \frac{1}{n} \sum_{i=1}^n \rho(Y_i - \mathbf{Z}_i^T \boldsymbol{\beta})$

Thm (Huber) If $\psi \equiv \rho'$ is smooth, *p is fixed*, $n \to \infty$, $E_0 \psi^2(e) < \infty$, $E_0 \psi'(e) \neq 0$ and $\mathbf{Z}$ is full dimensional, $E\mathbf{Z}\mathbf{Z}^T$ non singular

Robust M Estimation in Regression

$$\hat{\boldsymbol{\beta}}_\rho = \boldsymbol{\beta}_0 - \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i \frac{\psi(e)}{E_0\psi'(e)} + o_P(n^{-\frac{1}{2}})$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \Longrightarrow N_p\left(\mathbf{0}, \left[E_0(\mathbf{Z}\mathbf{Z}^T)\right]^{-1}\sigma^2(\rho)\right)$$

$$\sigma^2(\rho) = \frac{E_0\psi^2(e)}{[E_0\psi'(e)]^2}$$

E.g.: $\psi(t) = t$ LSE not robust against heavy tails

$$\begin{aligned} \psi(t) = h_k(t) &= t, \ |t| \le k \ \text{(Huber)} \\ &= k\,\mathrm{sgn}(t), |t| > k \\ \psi(t) = \mathrm{sgn}(t) & \qquad\qquad (L1) \end{aligned}$$

The curent focus of interest: p, n both large

What if $p \to \infty$?

Theorem (Huber) (1973) (Negative)
If $\frac{p}{n} \to c > 0$
$\exists$ contrast $\mathbf{t}^T \boldsymbol{\beta}_0$
$\mathbf{t}^T(\hat{\boldsymbol{\beta}}_{\text{LSE}} - \boldsymbol{\beta}_0)$ is not asymptotically Gaussian

Note: $E\left[X^T(\hat{\boldsymbol{\beta}}_{\text{LSE}} - \hat{\boldsymbol{\beta}}_0)\right]^2 = \sigma^2 \frac{p}{n}$
$\Longrightarrow$ Data picked contrast is inconsistent

(Huber, Portnoy) (Positive)

(Huber) If the projection matrix $[\mathcal{Z}^{(n)}]^T\{\mathcal{Z}^{(n)}[\mathcal{Z}^{(n)}]^T\}^{-1}\mathcal{Z}^{(n)}$ has diagonal $\pi_{ii} \equiv \frac{p}{n}$ and

$$\boxed{\frac{p^3}{n} \to 0}$$

$a^T(\hat{\beta} - \beta) \sim N\big(0, \sigma^2(a, \psi)\big)$

$$\sigma^2(a,\psi) = \frac{E\psi^2(e)}{\big(E\psi'(e)\big)^2} a^T \big[(\mathcal{Z}^{(n)}[\mathcal{Z}^{(n)}]^T\big]^{-1} a$$

Improved Conditions: Portnoy (1985) AS

Gaussian Linear Regression Model

$$\mathbf{Y}_{n\times 1} = Z_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}$$

$$\mathbf{e} = (e_1, \dots, e_n)^T \text{ iid } N(0, \sigma^2)$$

$$\mathbf{Z}^{(j)} \equiv \begin{pmatrix} Z_{ij} \\ \vdots \\ Z_{nj} \end{pmatrix}, \qquad j = 1, \dots, p$$

$$Z \equiv (\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(p)})_{n\times p} = [\mathcal{Z}^{(n)}]^T$$

Suppose $|\mathbf{Z}^{(j)}|^2 = n$, $\mathbf{Z}^{(a)} \perp \mathbf{Z}^{(b)}$ $a \neq b$
(Canonical Gaussian Model)
Equivalent to:

Gaussian White Noise Model (Donoho, Johnstone, Kerkyacharian, Picard (1995))

$$X_j = \beta_j + \varepsilon_j,\ j = 1, \dots, p,\ \varepsilon_j \sim N\big(0, \frac{\sigma^2}{n}\big)\, iid$$

$$X_j = \frac{[\mathbf{Z}^{(j)}]^T \mathbf{Y}}{n}$$

Assume

i) $\boldsymbol{\beta}$ sparse: If $S = \{\beta_j; \beta_j \neq 0\}$, $|S| = s << p$.

ii) Signal strong: $j \in S \Longrightarrow |\beta_j| \geq \delta_n > 0$

Let,

$$\hat{X}_j \equiv X_j - h_K(X_j)$$

$$h_K \equiv \text{ Huber function, } K = \sigma\sqrt{\frac{2\log p}{n}}$$

GWN Result: If $\delta_n\sqrt{\frac{n}{\log p}} \to \infty$,

$$\sum_{j=1}^{p} E(\hat{X}_j - \beta_j)^2 = \frac{s\sigma^2}{n}(1 + 0(1))$$

(Best possible if S is known)

If $\delta_n = \Omega\sqrt{\frac{\log p}{n}}$, $s \to s\log p$.

The Lasso: Donoho, Saunders, Chen (1998), Tibshirani (1996)

$$\hat{\beta}_L \equiv \arg\min \left\{ |\mathbf{Y} - Z\beta|^2 + \lambda|\beta|_1 \right\}$$

For canonical model

$$\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(p)} \text{ orthonormal } |\mathbf{Z}^{(j)}|^2 = n, \; j = 1, \ldots, p \; .$$

Then, for suitable $\lambda(K)$

$$\hat{\beta}_{jL} = \hat{X}_j \; .$$

Conclusion

a) If $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(p)}$ are nearly orthogonal
b) $\boldsymbol{\beta}$ is sparse
c) The signal is strong

$\hat{\boldsymbol{\beta}}_L$ behaves as LS when we know S.

Many Results: Buhlmann, van de Geer, Tsybakov, Meinshausen, Yu, Fan and collaborators, have found minimal versions of a)-c) extended GWN result.

Robust Case:

Bradic, Fan, Wang, JRSS(B) (2011)
Variable selection including robust objective functions give results of this type with

$$\sigma^2 \frac{s}{n} \to \frac{E\psi^2(e)}{[E\psi'(e)]^2} \frac{s}{n}$$

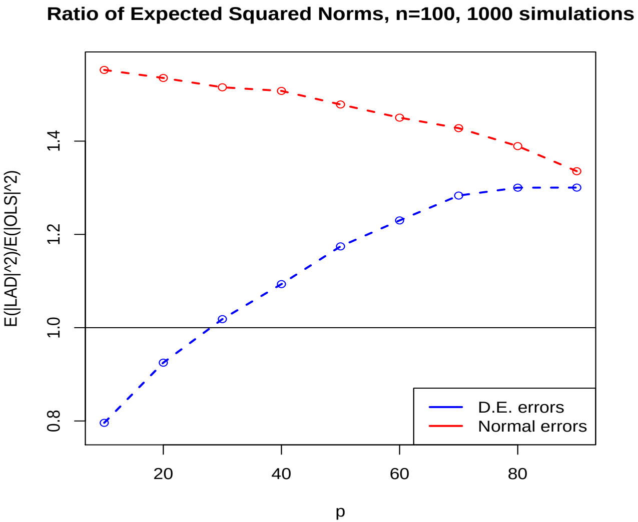
Open problems in paralleling the work done for LS + Lasso but see GLM results of van de Geer and others (Buhlmann, van de Geer(to appear)) Statistics for High Dimensional Data

Behavior of $||\hat{\beta} - \beta_0||^2$

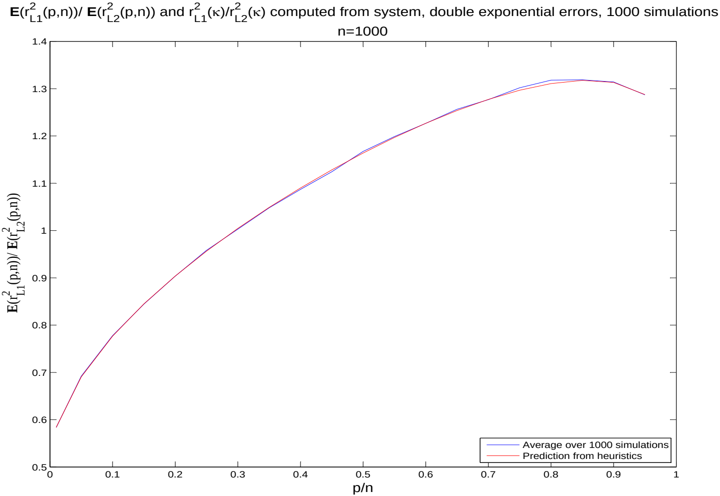
A. What if a) or b) or c) conditions don't hold:
Another Lecture

B. What if $\frac{p}{n} \to 0 < c < 1$ and robust and least squares are compared *without penalization*?

Surprising simulations



Surprising simulations



(Semi-heuristic) Results of el Karoui, Bean, Bickel, Lim and Yu

$$Y_i = X_i^T \beta_0 + \epsilon_i \quad i = 1, \dots, n$$

- ϵ_i i.i.d. $g \perp\!\!\!\perp \{X_i : i = 1, \dots, n\}$
- X_i i.i.d. $\mathcal{N}(0, \Sigma)$

Define $\hat{\beta}(\rho; \beta_0, \Sigma) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(Y_i - X_i^T \beta)$

- ρ convex
- $n \to \infty$, $p/n \to \kappa < 1$.

Key Lemma

$$\hat{\beta}(\rho; \beta_0, \Sigma) \stackrel{\mathcal{L}}{=} \beta_0 + \|\hat{\beta}(\rho; 0, I_p)\| \Sigma^{1/2} u,$$

where u is uniform on the p sphere of radius 1.

$\therefore$ Can assume $\beta_0 = 0$, $\Sigma = I_p$.

Special case of basic result

If $r_\rho(p,n) \equiv \|\hat{\beta}(\rho; 0, I_p)\|$
Under suitable regularity conditions,

$r_\rho(p,n) \xrightarrow{P} r_\rho(\kappa)$ solving:

$$\mathbb{E}\left[\mathrm{prox}_c(\rho)\right]'(\hat{z}_\epsilon) = 1 - \kappa$$
$$\mathbb{E}\left(\hat{z}_\epsilon - [\mathrm{prox}_c(\rho)](\hat{z}_\epsilon)\right)^2 = \kappa r_\rho^2(\kappa),$$

$\hat{z}_\epsilon \stackrel{\mathcal{L}}{=} \epsilon + r_\rho(\kappa) Z$, $\epsilon \perp\!\!\!\perp Z$, $Z \sim \mathcal{N}(0,1)$.

Remarks

$$\mathrm{prox}_c(\rho)(x) = \operatorname*{argmin}_y \left(\rho(y) + \frac{(x-y)^2}{2c} \right)$$

Solves, if ρ is differentiable, strictly convex:

$$y + c\rho'(y) = x.$$

Key ideas:

I. Leave out 1 predictor $\{X_{pi} : i = 1, \dots, n\}$.
$r_{i,[p]} \equiv \epsilon_i - V_i^T \hat{\gamma}$
where $V_i \equiv (X_{1i}, \dots, X_{p-1,i})^T$, $\hat{\gamma} \equiv$ estimate of $(\beta_{01}, \dots, \beta_{0,p-1})^T$ without $X^{(p)} \equiv (X_{p1}, \dots, X_{pn})^T$. Then:

$$(*) \qquad \hat{\beta}_p = \frac{\sum_i X_{pi} \psi(r_{i,[p]})}{\sum_i X_{pi}^2 \psi'(r_{i,[p]}) - v_p^T \mathcal{S}_p^{-1} v_p} + o_p(n^{-1/2})$$

$v_p = \sum_i \psi'(r_{i,[p]}) V_i X_{pi}$
$v_p^T \mathcal{S}_p^{-1} v_p = \left[X^{(p)}\right]^T D^{1/2} \Pi_V D^{1/2} \left[X^{(p)}\right]$, $D_{ii} = \psi'(r_{i,[p]})$.
Π_V is a projection matrix of rank $p-1$.
$X^{(p)} \perp\!\!\!\perp r_{i,[p]}, V_i$

Key ideas:

For LS $\psi(x) = x$ and $(*)$ holds exactly.
Asymptotically, $r_{i,[p]} \sim g * \text{Gaussian}$
$\sqrt{n}\hat{\beta}_p \Rightarrow \mathcal{N}(0, \sigma^2(\rho, g, \kappa))$.
$\sigma^2(\rho, g, \kappa) = \frac{\sigma^2}{1-\kappa}$ for LS.

II. Analysis requires leave out one (X_i, Y_i) as well.

Projection Pursuit

(J) Kruskal (1969),(1972), Switzer (1970), Switzer and Wright (1971), Friedman Tukey (1974), *Huber* (1985), Diaconis and Freedman (1985)

Given:

$$X_1, \ldots, X_n \qquad p \times 1 \quad \text{iid}$$

Find "interesting" projections i.e.

$$a, |a| = 1 \ni$$

$$P_{n,a} \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_{a^T X_i} \text{ is as non-normal as possible}$$

In *expectation* $P_{n,a} \approx P_a \leftrightarrow f_a \equiv$ density of $a^T X$

Measures of Nonnormality

$$SK(P_{n,a}) \leftrightarrow \frac{E_a(X - E_a X)^3}{\left[E_a(X - E_a X)^2\right]^{\frac{3}{2}}} \equiv SK(P_a)$$

$$KURT(P_{n,a}) \leftrightarrow \frac{E_a(X - E_a X)^4}{\left[E_a(X - E_a X)^2\right]^2} - 3 \equiv K(P_a)$$

These are highly nonrobust to outliers.

Robust and "efficient" measures

Alternatives: Robust Measures of Skewness , Kurtosis by Trimming

Efficient Measure (Estimate)

$$\int \log f_a f_a(x) dx + \log(2\pi e)^{\frac{1}{2}} \left[E_a(X - E_a X)^2 \right]^{\frac{1}{2}}$$

Procedure: Maximize over a

A Rationale

Diaconis, Freedman (1985)
If $p \to \infty$ with $n \to \infty$ under weak conditions e.g.

$$\text{If } X_j \text{ iid } F,\ \ EX^2 < \infty,\ \ X = (X_1, \ldots, X_p)^T,$$

Then, almost all $P_{n,a}$ are asymptotically Gaussian, where "almost" all is with respect to Lebesgue measure on surface of unit sphere in R^p.

Some Comfort

Theorem If F is Gaussian iid and $\frac{p}{n} \to 0$, then

$$\sup_a \sup_x \left| P_a(-\infty, x] - P_{na}(-\infty, x] \right| \xrightarrow{P} 0$$

Some Caution

But, even if F is Gaussian iid, if $\frac{p}{n} \to c > 0$,

$$\max_a \int a^T \left(\mathbf{x} - \mu(P_{n,a})\right)^2 dP_{n,a} \not\to 1$$

(Wigner, Geman)

How bad can things get?

Suppose $\frac{p}{n} \longrightarrow \infty$

Theorem (B, Nadler)
Let $X_1, \ldots, X_n$ i.i.d. $N(\mathbf{0}, I_p)$. Let G be *any* cdf such that $G - \Phi$ doesn't change sign.. Let $\hat{F}_a$ denote the empirical cdf of $a^T X_1, \ldots, a^T X_n$, $|a| = 1$. Then:

$$P\left[\inf_a \|\hat{F}_a - G\|_\infty \to 0\right] = 1$$

where $\|f\|_\infty = \sup_x |f(x)|$.

Idea of proof:

a) Let $\psi : R \to R$ monotone increasing bounded.

Let $\Psi_a \equiv \frac{1}{n} \sum_{i=1}^n \psi(a^T X_i)$

$N = \lambda^p$ for λ to be chosen, $\lambda > 1$.

$A = \{a_j : 1 \le j \le N\}$ points in $\mathcal{S}_p$

Then, for any $\varepsilon > 0$,

$$P_\Phi[K_0 - \varepsilon \le \Psi_{a_j} \le K_0 + \varepsilon \text{ for some } j,\ 1 \le j \le N] \to 1$$

b) Let $\psi^{(1)}, \dots, \psi^{(m)}$ be as above, $\varepsilon > 0$. For,

$$K_j \text{ arbitrary}, \;\; j = 1, \dots, m,$$
$$\text{sgn}\left(K_j - \int \psi^{(j)}(\xi)\phi(\xi)\mathrm{d}\xi\right) \text{ constant}$$

Then,

$$P_\Phi\left[K_j - \varepsilon \le \Psi_a^{(j)} \le K_j,\; 1 \le j \le m \text{ for some } a \in \mathcal{A}\right] \to 1$$

c) Let $\psi^{(j)}(u) = 1(x_j, \infty)$

$K_j = \overline{G}(x_j)$

By b)

$$P\big[\exists\, j \in \mathcal{A} \ni |\hat{\overline{F}}_{a_j}(x_k) - \overline{G}(x_k)| \le \varepsilon \text{ for all } 1 \le k \le m\big] \to 1$$

Lemma

$\exists\ \lambda > 1,\ \varepsilon > 0,\ a_1, \ldots, a_N \in \mathcal{S}_p$, such that, for $N = \lambda^p$,

$\exists\ a_1, \ldots, a_N$ so that $|(a_j, a_{j'})| \le 1 - \varepsilon$ for all $1 \le j \ne j' \le N$,

Then,

$$(a_1^T X_1, \ldots, a_N^T X_1)^T \sim \mathcal{N}_N(\mathbf{0}, R)$$

$$R = \|\rho_{ij}\|_{N \times N},\ \text{where } |\rho_{ij}| \le 1 - \varepsilon,\ \ i \ne j$$

$$
\begin{aligned}
\text{If } X &\sim N(\mathbf{0}, R_0) \\
R_0 &\equiv (1-\varepsilon)\mathbf{1}\mathbf{1}^T + \varepsilon I_d \\
\mathbf{1} &\equiv (1, \dots, 1)^T, \\
X &= (1-\varepsilon)\mathbf{z}_0\mathbf{1} + \left(1-(1-\varepsilon)^2\right)^{\frac{1}{2}}\mathbf{Z} \\
\mathbf{z}_0 &\sim \mathcal{N}_1(0,1) \perp \mathbf{Z} \sim \mathcal{N}_N(\mathbf{0}, I_d)
\end{aligned}
$$

Slepian's inequality

Extended by Joag-dev et al (1983) Ann. Prob.

Let $\mathbf{Z}^{(j)} \sim \mathcal{N}(0, R^{(j)})\ \ j = 0, 1,$

$$R^{(j)}_{N\times N} \equiv \rho^{(j)}_{ab} = \delta_{ab} + (1 - \delta_{ab})\rho^{(j)}_{ab},$$
$$\text{and } \rho^{(0)}_{ab} \le \rho^{(1)}_{ab} \text{ for all } a, b$$

Let $\Psi : R^N \longrightarrow R$, bounded,
$\frac{\partial^2\psi}{\partial x_a \partial x_b} \ge 0$ all $a \ne b$. Then,

$$E\Psi(\mathbf{Z}^{(0)}) \le E\Psi(\mathbf{Z}^{(1)})$$

(Valid if $\Delta^2_{a,b}\psi \ge 0$, where
$\Delta^2_{a,b}\psi = \psi(x_a + h_a, x_b + h_b, x_c, c \ne a, b) - \psi(x_a + h_a, x_b, x_c, c \ne a, b) - \psi(x_a, x_b + h_b, x_c, c \ne a, b) + \psi(x_c, c = 1, \ldots, N)$.)

Let ψ_j, $j = 1, \ldots, m$ be bounded non-decreasing function

Consider $a^T X_1, \ldots, a^T X_n$, $a \in \mathcal{A}$, $a_1, \ldots, a_N$ as in Lemma.

Consider $\mathcal{Y}_{m \times n}$ where $\mathcal{Y}_{ij} = a_i^T \vec{X}_j$ and $\mathcal{X}_c(\vec{\mathcal{Y}})$ where

$$\mathcal{X}_c\left(u_{ik} : i = 1, \ldots, n,\ k = 1, \ldots, N\right) \equiv \prod_{k=1}^{N} \left[1 - \prod_{\ell=1}^{m} 1\left(\mathcal{X}^{(\ell)}(u_{1k}, \ldots, u_{nk}) \geq c_\ell\right)\right],$$

and $\mathcal{X}^{(\ell)}(v_1, \ldots, v_n) = \frac{1}{n} \sum_{\ell=1}^{n} \psi_\ell(u_i)$.

$\mathcal{X}$ satisfies our hypotheses.

f) Apply large deviation theory to

$$\frac{1}{n}\sum_{i=1}^{n}\psi(Z_{ij}^{(0)})$$

where $\mathbf{Z}_1^0, \dots, \mathbf{Z}_n^0$

$$\mathbf{Z}_i^{(0)} \equiv (Z_{i1}^{(0)}, \dots, Z_{iN}^{(0)})^T$$

are iid $N_N(\mathbf{0}, R_{N\times N}^{(0)})$

$$R_{N\times N}^{(0)} = \|\delta_{ab} + (1-\delta_{ab})(1-\varepsilon)\|_{N\times N}$$

to obtain

$$P\left[\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^{(0)} \notin [K_0-\delta, K_0+\delta] \text{ for any } 1 \le j \le N\right] \longrightarrow 0$$

g) Apply Slepian's inequality to get a).
Generalize to b), c) using Joag-dev's inequality.
Choose the $\{x_j\}$ to be dense to get $(*)$

Discussion

II: Huber (1985)

1) "Perhaps the practical conclusion to be drawn is that we shall have to acquiesce to the fact that PP will in practice reveal not only true but also spurious structure and that we must weed out the latter by other methods."
2) What structures survive if we consider a random set of m projections of the data?
   E.g. Suppose the true population is
   $$(1-\varepsilon)N(\mathbf{0}, I_p) + \varepsilon N(\mathbf{0}, \Sigma)$$
   Σ of rank $<< p$
   If we take $m = o(n)$ projections, what chance do we have of finding $N(0, \Sigma)$ structure?
3) Conjecture: Result holds for all G.

Discussion

I. Proof now available

Questions:

1) "Optimal" ρ, g, κ known: DONE.
2) Robust ρ, optimizing on g given g in a small neighborhood of ϕ: OPEN.