Likelihood-Based Sufficient Dimension Reduction for Regression

R.D. Cook
School of Statistics
University of Minnesota

JSM 2009

Recent work in collaboration with Liliana Forzani

Sufficient Reductions for Regression

Variables: $Y \in \mathbb{R}^1$**,** $\mathbf{X} \in \mathbb{R}^p$**,** $(Y, \mathbf{X}) \sim F$

Data: $(Y_i, \mathbf{X}_i)$ iid F**,** $i = 1, \ldots, n$**.**

Goal: **Reduce** $\dim(\mathbf{X})$ **without loss of information on** $Y|\mathbf{X}$**.**

Sufficient Reduction: $R : \mathbb{R}^p \to \mathbb{R}^q$**,** $q \leq p$**,**

1. $\mathbf{X}|(Y, R(\mathbf{X})) \sim \mathbf{X}|R(\mathbf{X})$ **(Inverse regression)**
2. $Y|\mathbf{X} \sim Y|R(\mathbf{X})$ **(Forward regression)**
3. $Y \perp\!\!\!\perp \mathbf{X}|R(\mathbf{X})$ **(Joint reduction)**

Bijective trans. of R **are also sufficient.**

Routes to methodology: **Model-free & Model-based.**

Model-free reduction: **restrict** $R(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$**,** $\boldsymbol{\alpha} \in \mathbb{R}^{p\times q}$**, so** $\mathbf{X}|(Y, \boldsymbol{\alpha}^T\mathbf{X}) \sim \mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}$**.**

Central subspace: $\mathcal{S}_{Y|\mathbf{X}} = \cap\,\mathrm{span}(\boldsymbol{\alpha})$**.**

Goal: Exhaustive estimation of $\mathcal{S}_{Y|\mathbf{X}}$ **under minimal assumptions, guided by consistency.**

SIR, SAVE, PHD, IHT, IR methods, contour reg., directional reg. , 3rd and 4th moment methods, partial methods, kernel methods . . . mostly moment based.

Issues: Exhaustive estimation of $\mathcal{S}_{Y|\mathbf{X}}$**. Efficiency. Predictor types mixed. Post reduction prediction of** $Y|\widehat{R}(\mathbf{X}_{\mathrm{new}})$**.**

Model-based reduction: **Model $\mathbf{X}|Y$, derive and est. R via maximum likelihood.**

- **Restriction to linear R's unnecessary. Mixed predictor types OK.**
- **Efficiency: Methods inherit optimal properties from likelihood theory.**
- **Prediction: $\mathbf{X}|Y \to Y|\mathbf{X}$ (Adragni & Cook '09)**

$$\widehat{\mathbf{E}}(Y|\mathbf{X}_{\mathbf{new}}) = \frac{\sum_{i=1}^{n} y_i \widehat{f}\{R(\mathbf{X}_{\mathbf{new}})|Y = y_i\}}{\sum_{i=1}^{n} \widehat{f}\{R(\mathbf{X}_{\mathbf{new}})|Y = y_i\}}$$

- **Issue: Robustness**

Nonlinear Inverse Normal Model

$$\mathbf{X}|(Y=y) \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}(y) + N_p(0, \boldsymbol{\Delta})$$

- $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, $d < p$, $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, $\boldsymbol{\Delta}$ **and** d **are unknown.**
- $\mathbf{f}(y) \in \mathbb{R}^r$ **known vector of basis functions, like polynomials, piecewise polynomials, Fourier series, or indicator functions.**
- $R(\mathbf{X}) = (\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}(\mathbf{X} - \mathbf{E}(\mathbf{X})) \in \mathbb{R}^d$ **is a** ***minimal sufficient reduction.***

$$\mathbf{X}|(Y = y) \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}_{p\times d}\boldsymbol{\beta}_{d\times r}\mathbf{f}(y) + N_p(0, \boldsymbol{\Delta})$$

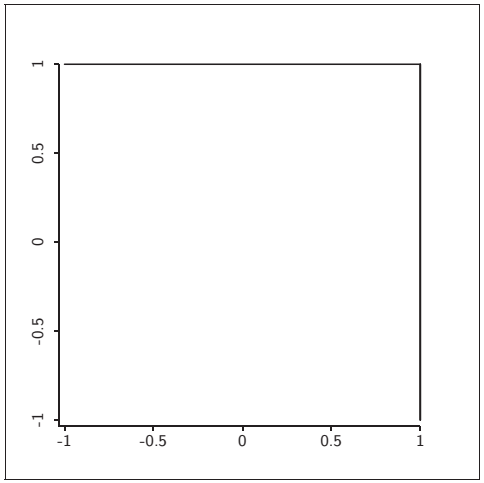
Principal Components: $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$**,** $\mathbf{f}(y_i)$ **indicator vector** $\mathbf{f}_i$ **for** i**-th case, so** $r = n$**.**

Then MLE $\widehat{R}(\mathbf{X}_i)$ **of** $R(\mathbf{X}_i)$**,** $i = 1, \ldots, n$**, is the first** d **principal components (PC's) of X.**

No requirement that $n > p$**.**

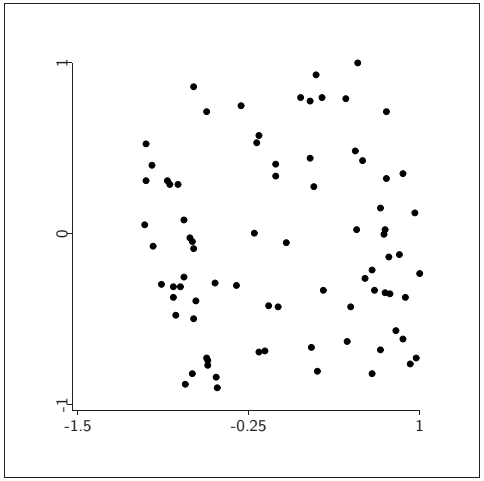
$\boldsymbol{\beta}\mathbf{f}_i$**,** $n = 80$**,** $d = 2$

(Adragni & Cook '09)

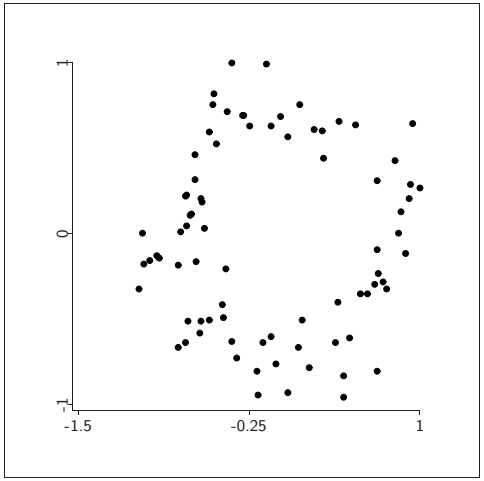


$$\begin{aligned} \mathbf{X}_i &= \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_i + N(0, I_p) \\ \mathrm{vec}(\boldsymbol{\Gamma}) &\sim N(0, I_{2p}) \end{aligned}$$

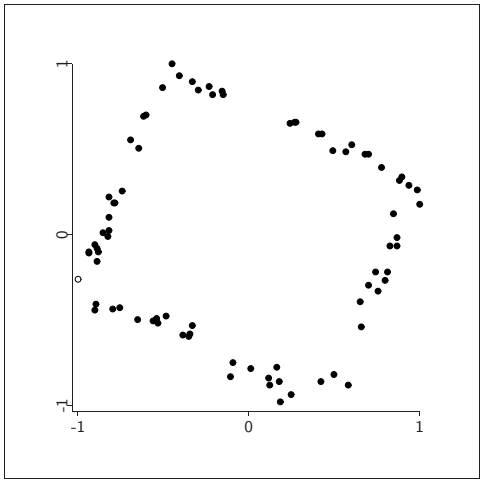
Next slide: Plots of $\widehat{R}(\mathbf{X}_i) : 2 \times 1$**,** $i = 1, \ldots, 80$**.**



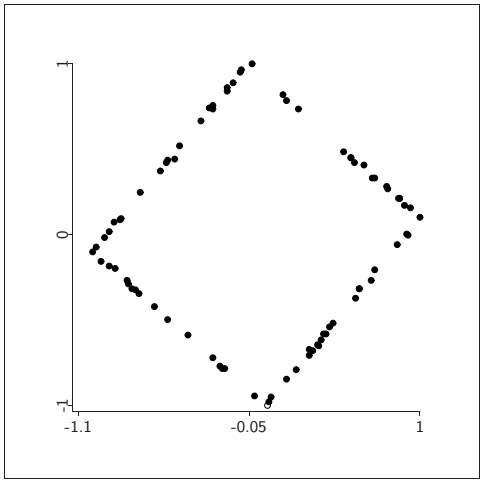
$p = 3$



$p = 5$



$p = 25$



$p = 500$

$$\mathbf{X}|(Y=y) \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}_{p\times d}\boldsymbol{\beta}_{d\times r}\mathbf{f}(y) + N_p(0, \boldsymbol{\Delta})$$

Principal Fitted Components: $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$**.**

MLE $\widehat{R}(\mathbf{X}_i)$**,** $i = 1, \ldots, n$**, consists of the first** d **Principal Fitted Components (PFC), PC's of the fitted vectors from the linear regression of** $\mathbf{X}$ **on** $\mathbf{f}(y)$**, including an intercept.**

$n > p$ **is not required .**

$$\mathbf{X}|(Y=y) \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}_{p\times d}\boldsymbol{\beta}_{d\times r}\mathbf{f}(y) + N_p(0, \boldsymbol{\Delta})$$

PFC$_{\boldsymbol{\Delta}}$: $\widehat{R}$ computed as PFC's with $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$, but using the standardized predictors $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1/2}\mathbf{X}$.

$n > p$ is required.

New Results: Large p regressions

1. PFC: $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}(y) + N_p(0, \sigma^2\mathbf{I})$**,** $n < p$ **permitted.**

2. PFC$_{\boldsymbol{\Delta}}$: $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}(y) + N_p(0, \boldsymbol{\Delta})$**,** $n > p$ **required.**

How do likelihood methods under 1 perform as $p \to \infty$ with n fixed when, in fact, 2 is the true model?

How do PC (indicator $\mathbf{f}_i$) and PFC compare?

Limiting reductions as $p \to \infty$, with n fixed

Recall,

$$R_p(\mathbf{X}) = (\mathbf{\Gamma}^T \mathbf{\Delta}^{-1} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{\Delta}^{-1} (\mathbf{X} - \mathbf{E}(\mathbf{X})) \in \mathbb{R}^d$$

What happens to R_p as $p \to \infty$? Do PC's and PFC's estimate the limiting reduction?

Assume that

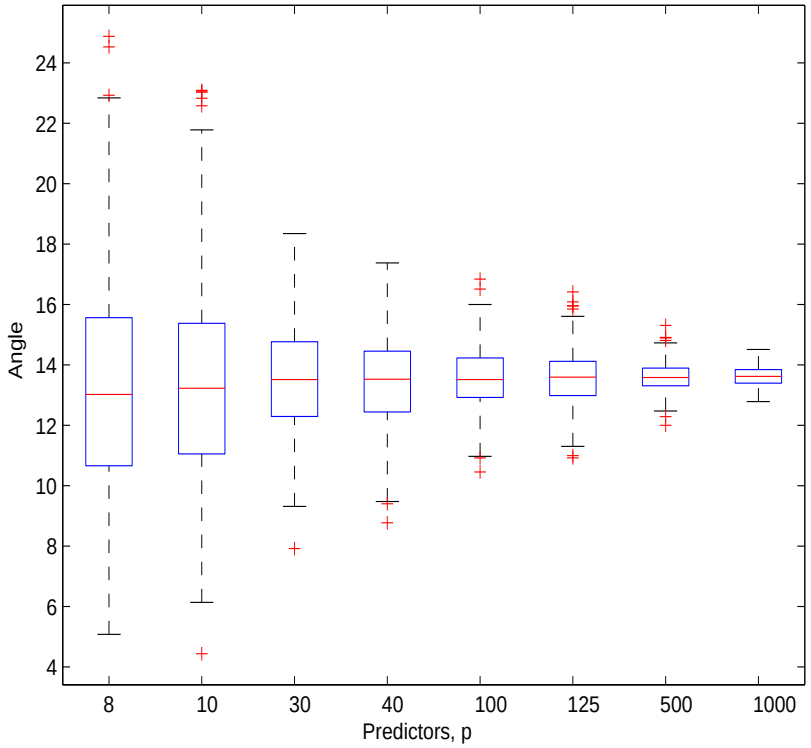
1. $\lim_{p\to\infty} p^{-1}\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{G} > 0$
2. $\lim_{p\to\infty} p^{-1}\text{trace}(\mathbf{\Delta}) < \infty$
3. $\lim_{p\to\infty} p^{-2}\text{trace}(\mathbf{\Delta}^2) = 0$

Then

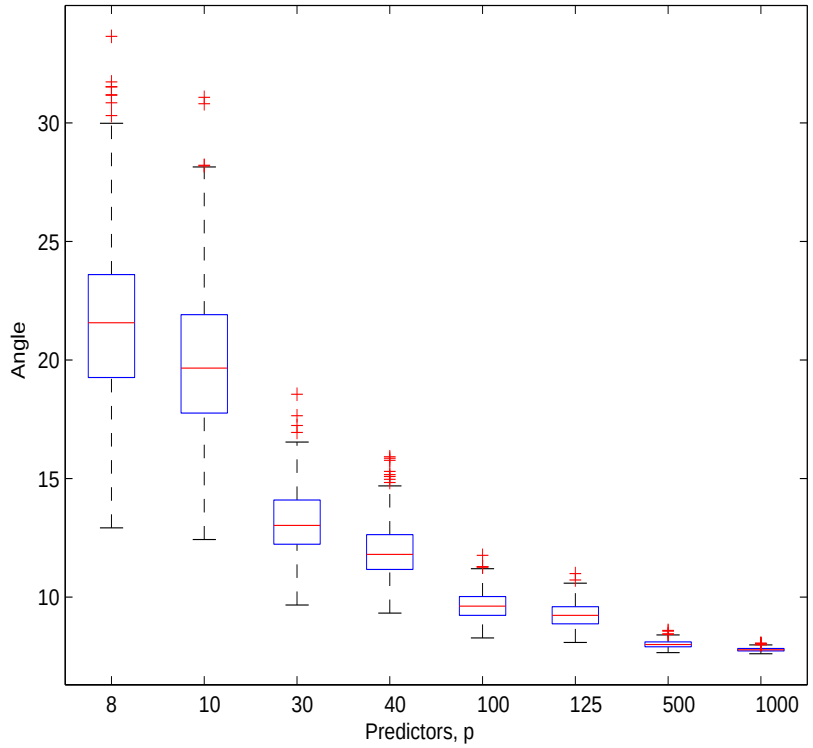
- $R_p(\mathbf{X}) \xrightarrow{p} R_\infty(\mathbf{X}) = \boldsymbol{\beta}\mathbf{f}(y)$**.**
- **The first** d **PC's and PFC's are both consistent estimators of** $R_\infty(\mathbf{X}_i)$**,** $i = 1, \ldots, n$**.**

Inconsistency of EV's does not imply inconsistency of PC's.

$$n = 20, d = 1, \mathbf{X}_i = \mathbf{1}_p y_i + N(0, \mathbf{I}_{20}), Y_i \sim N(0, 1).$$



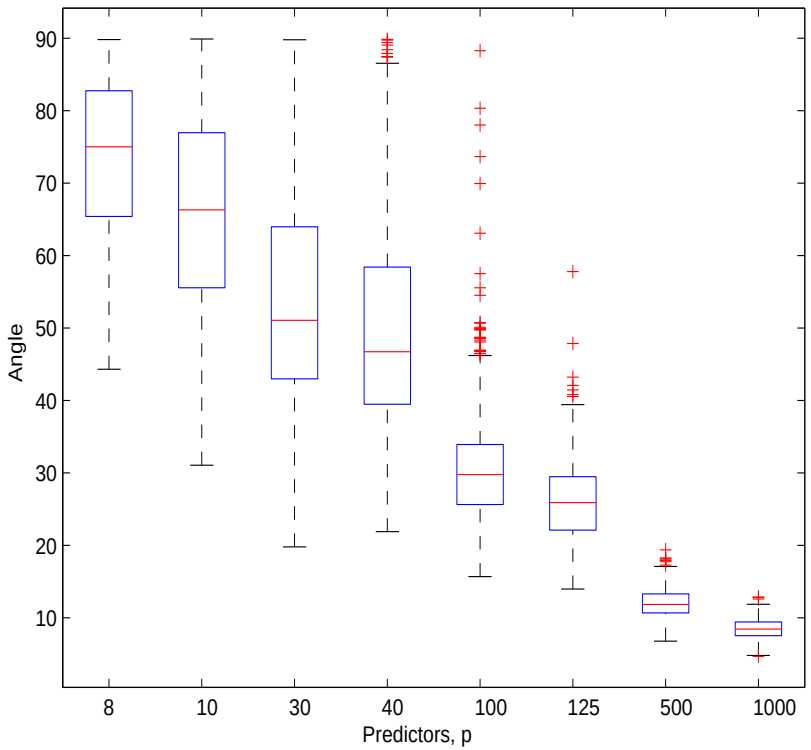
Eigenvectors



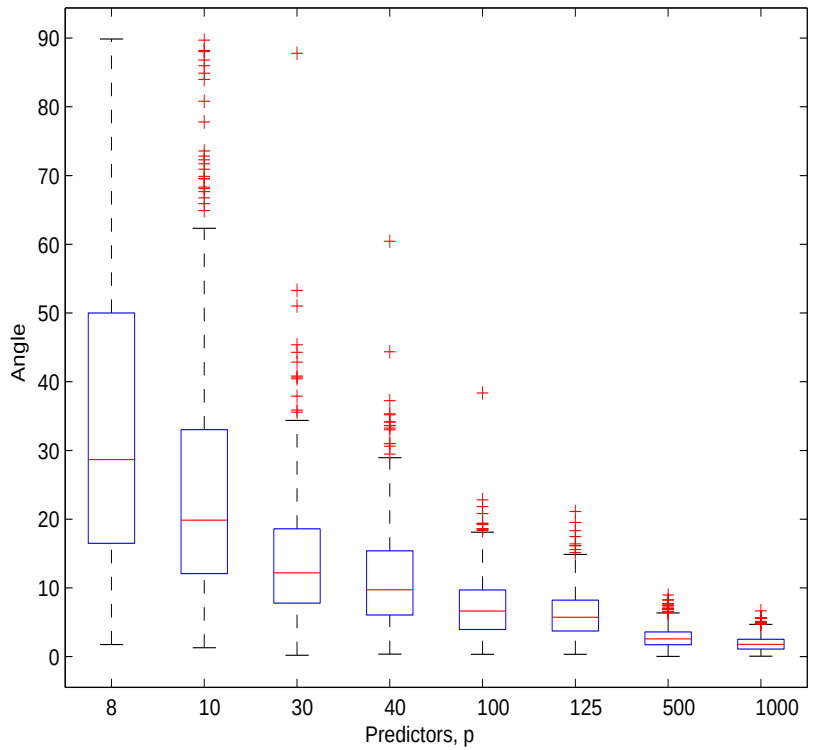
PC's

PFC dominates PC.

$$n = 20, d = 2, \mathbf{X}_i = \mathbf{\Gamma}\mathbf{f}_{\text{true}}(y) + N(0, \text{diag}(\chi_2^2)),$$
$$\mathbf{\Gamma}_{ij} \sim N(0,1), \mathbf{f}_{\text{true}}(y) = (y, |y|)^T, \mathbf{f}(y) = (y, |y|, y^3)^T.$$



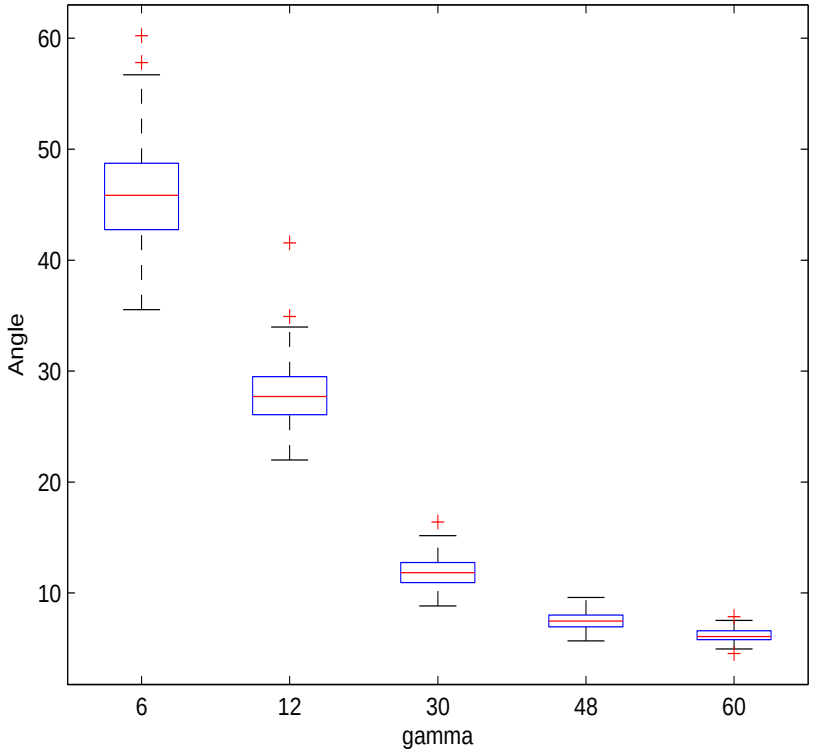
PC's



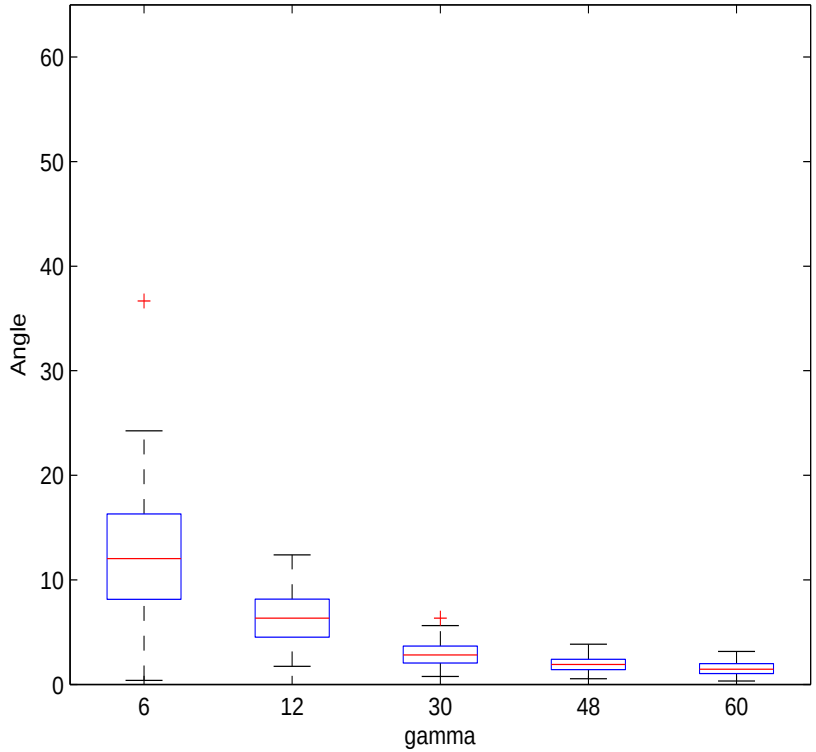
PFC's

PFC's & PC's can be useful even when inconsistent.

$p = 250$, $n = 50$, $d = 1$, $\mathbf{X}_i = \mathbf{\Gamma} y_i + N(0, \mathbf{\Delta})$, $\mathbf{\Gamma} = \boldsymbol{\gamma} \mathbf{1}_p$, $\mathbf{\Delta}$ **is a corr. matrix with** $\rho = .8$, **for fitting** $\mathbf{f}(y) = (y, \ldots y^4)^T$.



PC's



PFC's

Closing comments

- **Likelihood-based SDR can be quite useful in many regressions.**
- **PC's and PFC's can be useful in regressions where most predictors furnish some information on the response and the conditional predictor correlations are not "too large" or the signal/noise ratio is "large".**

- **Extensions: More flexibility in $\boldsymbol{\Delta} = \mathbf{Var}(\mathbf{X}|Y = y)$:**
  - $\boldsymbol{\Delta} = \sigma^2\mathbf{I}$ **(Cook '07)**
  - $\boldsymbol{\Delta} = \mathrm{diag}(\sigma_j^2)$ **(Cook & Forzani, Stat. Sci., '09)**
  - $\boldsymbol{\Delta} > 0$ **(Cook & Forzani, Stat. Sci., '09)**
  - $\boldsymbol{\Delta}(y) > 0$ **(Cook & Forzani, JASA, '09)**
- **Extension: $\mathbf{X}|Y \sim QEF$ (Cook & L. Li '09)**