Model-Based Sufficient Dimension Reduction for Regression

R.D. Cook
School of Statistics
University of Minnesota

Isaac Newton Institute
January 2008

In collaboration with Francesca Chiaromonte, Liliana Forzani, Bing Li and Lexin Li.

Fisher (1924), "The Influence of Rainfall on the Yield of Wheat at Rothamsted"

Opening issue: $n < p$**.**

Standard methodology of the time: Convert $n < p$ **to** $n > p^*$ **by using marginal regressions.**

Fisher's conclusion: "The meteorological variables to be employed must be chosen without reference to the actual crop record."

Principal Components in Regression

Goal : Reduce the dimension of the predictor vector $\mathbf{X} \in \mathbb{R}^p$ prior to studying the regression of Y on $\mathbf{X}$.

PC's : Replace $\mathbf{X}$ with $\widehat{\gamma}_1^T\mathbf{X}, \ldots, \widehat{\gamma}_d^T\mathbf{X}$ for some $d < p$, $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p$ ordered eigenvectors of $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{var}}(\mathbf{X})$.

Historical reasons :

- **mitigate the impact of collinearity**
- **facilitate computation and exploratory analysis**
- **facilitate analysis when $n < p$ (Chiaromonte & Martinelli 2002; Li & Li, 2004; Bair, et al. 2006, supervised principal components)**

Cox (1968):

> **A difficulty seems to be that there is no logical reason why the dependent variable should not be closely tied to the least important principal component.**

Some agreed with Cox, providing examples of relevant “least important principal components.”

Mosteller and Tukey (1977) asked

> **. . . how can we find linear combinations of the [predictors] that will be likely, or unlikely, to pick up regression from some as yet unspecified y?**

and answered

> **A malicious person who knew our x's and our plan for them could always invent a y to make our choices look horrible. But we don't believe nature works that way – more nearly that nature is, as Einstein put it (in German), "tricky, but not downright mean."**

On balance, the role for PC's in reg. seems elusive: (a) they are computed from the marginal of X, and (b) are not usefully invariant or equivariant.

Today they are ubiquitous in the applied sciences . . . but not widely represented in Statistics – IR methods, lasso, MAVE, PLS, projection pursuit, contour regression, directional regression.

Sufficient Reductions

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 the dimension of $\mathbf{X}$ without loss of relevant information on $Y|\mathbf{X}$.**

Sufficient Reduction: **A reduction $R : \mathbb{R}^p \to \mathbb{R}^q$, $q \leq p$, is sufficient if any of the following hold:**

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

We will pursue reductions via #1.

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

"Model-free" inverse regression $\mathbf{X}|Y$, #1.

The central subspace $\mathcal{S}_{Y|\mathbf{X}}$**:** $\boldsymbol{\alpha} \in \mathbb{R}^{p \times d}$

$$R(\mathbf{X}) = \boldsymbol{\alpha}^T \mathbf{X} \to \text{span}(\boldsymbol{\alpha}) \to \mathcal{S}_{Y|\mathbf{X}} = \cap \text{span}(\boldsymbol{\alpha}).$$

SIR, SAVE, PHD, IHT, IR methods, contour reg., directional reg., Fourier methods, moment methods, partial methods, kernel methods . . .

Parametric inverse regression $\mathbf{X}|Y$, #1.

Going back. $\mathbf{X}|Y \to Y|\mathbf{X}$ **when** $(\mathbf{X}, Y)$ **has a density,**

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

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

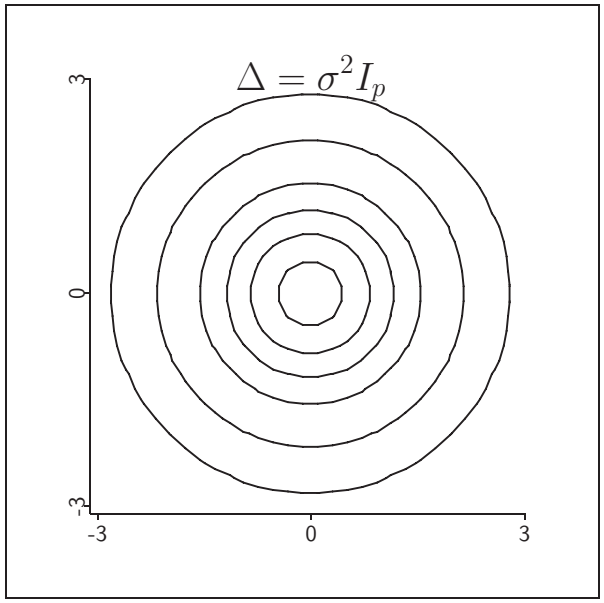
A First Nonlinear Inverse Model

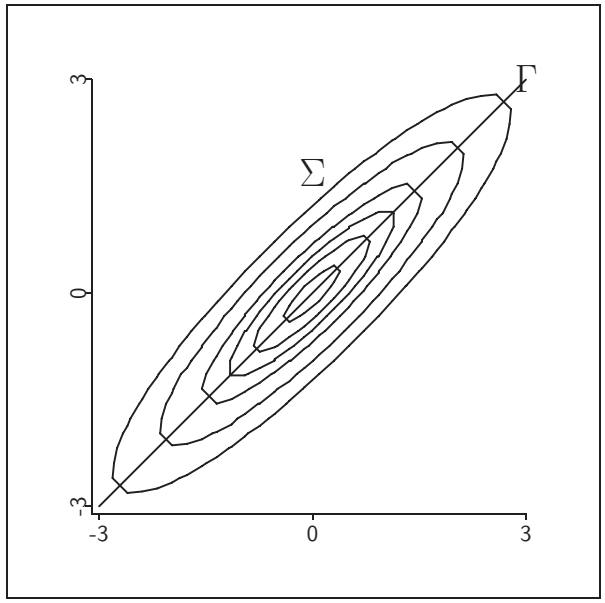
$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \sigma\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, I_p)$$

- $\mathbf{X}_y \sim \mathbf{X}|(Y = y)$
- $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times d}$, $\boldsymbol{\nu}_y \in \mathbb{R}^d$, **and** d **all unknown.**
- $R(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$, $\boldsymbol{\alpha}$ **any basis for** $\mathcal{S}_{Y|\mathbf{X}} \equiv \text{span}(\boldsymbol{\Gamma})$ **is a** ***minimal sufficient reduction***

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \sigma\boldsymbol{\varepsilon}$$

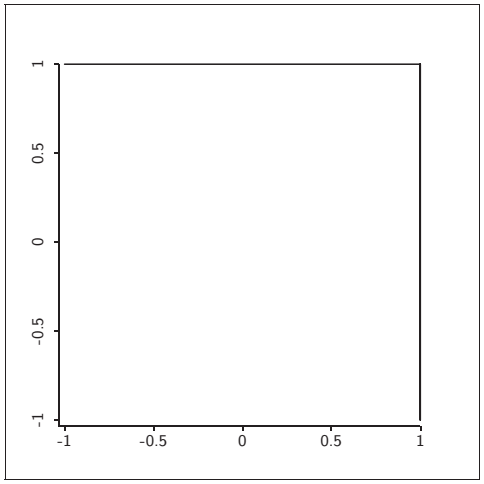
Using MLE, $\widehat{\mathcal{S}}_{Y|\mathbf{X}} = \mathrm{span}\{\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_d\}$**;** $R(\mathbf{X})$ **is estimated by 1st** d **sample PC's,** $\widehat{R}(\mathbf{X}) = (\widehat{\boldsymbol{\gamma}}_1^T\mathbf{X}, \ldots, \widehat{\boldsymbol{\gamma}}_d^T\mathbf{X})$**; role of** Y**.**





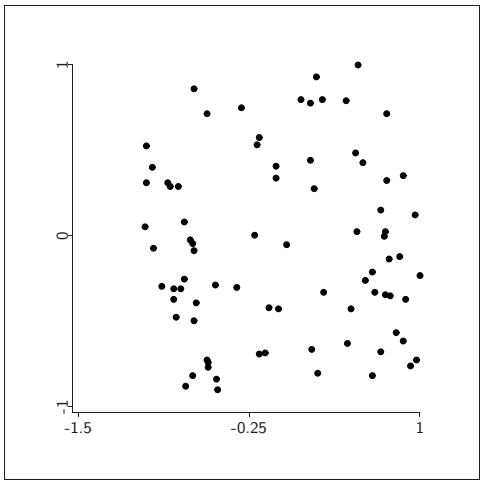
$$\boldsymbol{\Delta} = \mathbf{Var}\big(\mathbf{X}\big|Y\big) \qquad \boldsymbol{\Sigma} = \mathbf{Var}\big(\mathbf{X}\big)$$

$$\boldsymbol{\nu}_y, n = 80, d = 2$$

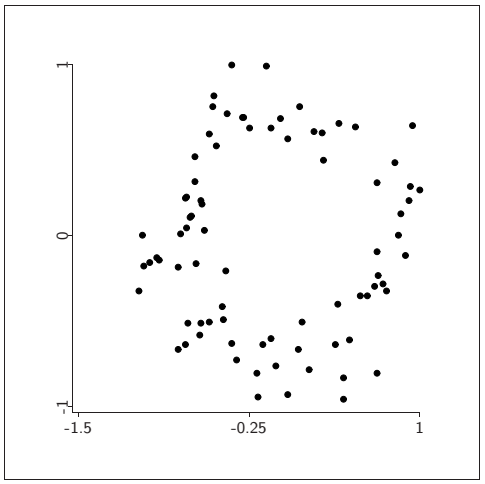


$$\begin{aligned} \mathbf{X}_y &= \boldsymbol{\Gamma}\boldsymbol{\nu}_y + .25N(0, I_p) \\ \text{vec}(\boldsymbol{\Gamma}) &\sim N(0, I_{2p}) \end{aligned}$$

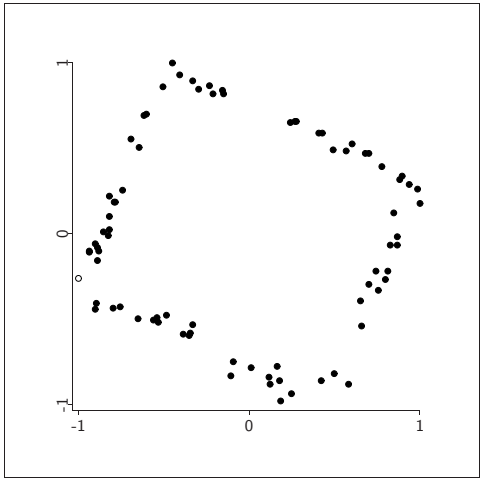
Next slide: Plots of $\widehat{R}(\mathbf{X}_i) : 2 \times 1$**,** $i = 1, \ldots, 80$**.**



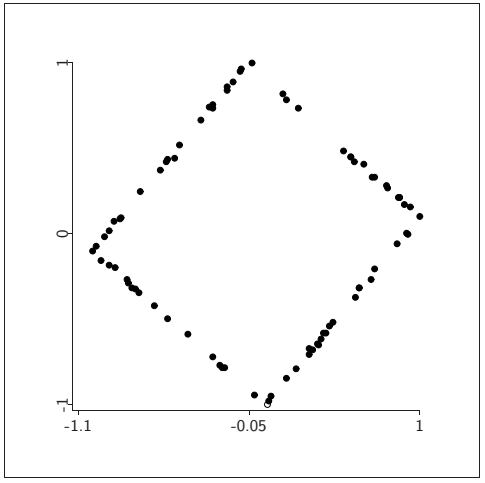
$p = 3$



$p = 5$



$p = 25$



$p = 500$

Extensions of $\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \sigma\boldsymbol{\varepsilon}$

- **Model** $\boldsymbol{\nu}_y = \boldsymbol{\beta}\mathbf{f}_y$**;** $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$**,** $d \leq r$**,** $\mathbf{f}_y \in \mathbb{R}^r$ **known.**
- **More flexibility in** $\boldsymbol{\Delta} = \mathbf{Var}(\mathbf{X}_y)$**.**

Principal Fitted Components

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon}$$

- $R(\mathbf{X}) = \alpha^T\mathbf{X}$ **still a minimal sufficient reduction,** α **any basis for** $\mathcal{S}_{Y|\mathbf{X}} = \mathrm{span}(\boldsymbol{\Gamma})$**.**
- **ML Estimation: Let** $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$ **sample covariance matrix of fitted values from the linear regression fit of** $\mathbf{X}_y$ **on** $\mathbf{f}_y$**.**

  Then $\widehat{\mathcal{S}}_{Y|\mathbf{X}}$ **is the span of the first** d **eigenvectors of** $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$**, leading to Principal Fitted Components.**
- **Robustness:** $\sqrt{n}$ **consistency under**
  - **Normal** $\boldsymbol{\varepsilon} \to t_5$**,** χ_5**,** $U(0,1)$**.**
  - **Approximation of** $\mathbf{f}_y \in \mathbb{R}^r$ **with** $\mathbf{g}_y$ **provided** $\mathrm{rank}\{\mathbf{cov}(\mathbf{f}_Y, \mathbf{g}_Y)\} = r$**.**

Application: Gene Expression (Leek & Storey, Sept. 2007):

$$
\begin{aligned}
\mathbf{X}_{y,\mathbf{u}} &= \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Gamma}_2\boldsymbol{\nu}_{\mathbf{u}} + \sigma\boldsymbol{\varepsilon} \\
\mathbf{X}_{y,\mathbf{c},\mathbf{u}} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Gamma}_1\boldsymbol{\beta}_1\mathbf{g}_{\mathbf{c}} + \boldsymbol{\Gamma}_2\boldsymbol{\nu}_{\mathbf{u}} + \sigma\boldsymbol{\varepsilon} \\
\mathbf{c} &= \text{known confounders} \\
\mathbf{u} &= \text{unknown confounders} \\
\mathbf{f}_y, \mathbf{g}_{\mathbf{c}} &= \text{known functions of } y, \mathbf{c}.
\end{aligned}
$$

Adding $\boldsymbol{\Delta} \equiv \mathbf{Var}(\mathbf{X}_y) > 0$

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon}, \ \boldsymbol{\Delta} > 0$$

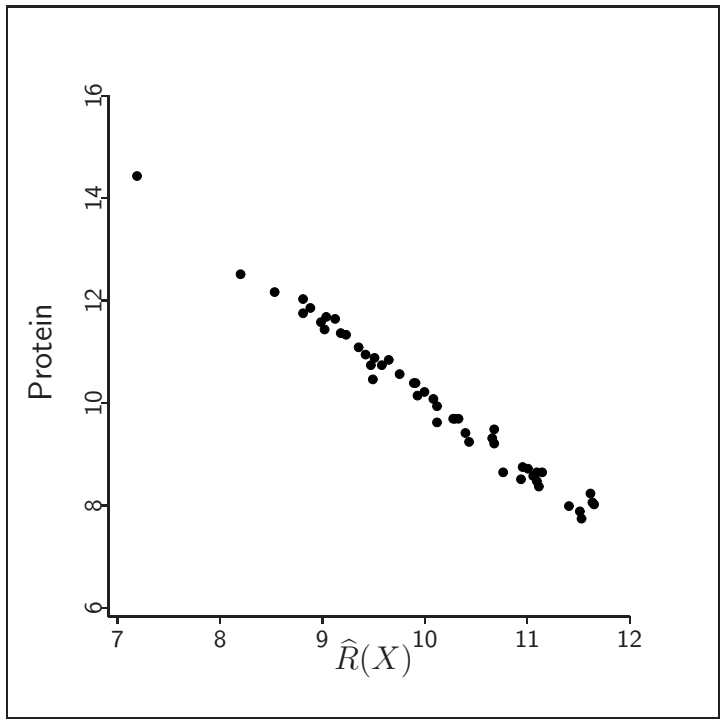
- $R(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$, $\boldsymbol{\alpha}$ **any basis for** $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Delta}^{-1}\mathrm{span}(\boldsymbol{\Gamma})$.
- $R(\mathbf{AX}) \equiv R(\mathbf{X})$ **and adapted to** Y.
- $\widehat{R}$ **computed as PFC's based on** $\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X}$**: Inner product of first** d **eigenvectors of** $\widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\widehat{\boldsymbol{\Sigma}}^{-1/2}$ **with** $\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X}$.
- **Robustness to non-normality and misspecificaton of** $\mathbf{f}_y$ **still hold,** $\mathrm{rank}\{\mathbf{cov}(\mathbf{f}_Y, \mathbf{g}_Y)\} = r$.
- **SIR**

Adding $\boldsymbol{\Delta} \equiv \mathbf{var}(\mathbf{X}_y) > 0$

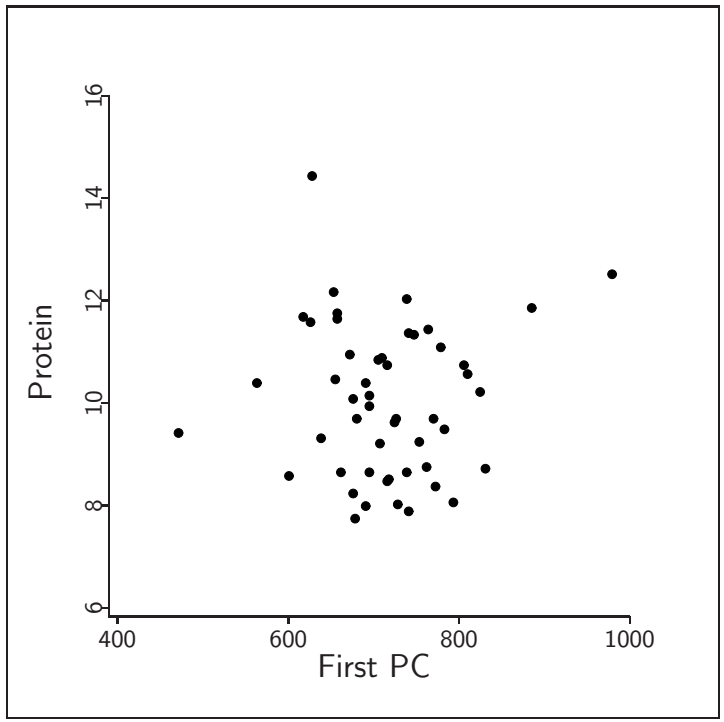
$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon}, \ \boldsymbol{\Delta} > 0$$

- $R(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$, $\boldsymbol{\alpha}$ **any basis for** $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Delta}^{-1}\mathrm{span}(\boldsymbol{\Gamma})$.
- $R(\mathbf{AX}) \equiv R(\mathbf{X})$ **and adapted to** Y.
- $\widehat{R}$ **computed as PFC's based on** $\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X}$**: Inner product of first** d **eigenvectors of** $\widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}\widehat{\boldsymbol{\Sigma}}^{-1/2}$ **with** $\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X}$.
- **Robustness to non-normality and misspecificaton of** $\mathbf{f}_y$ **still hold,** $\mathrm{rank}\{\mathbf{cov}(\mathbf{f}_Y, \mathbf{g}_Y)\} = r$.
- **SIR gives the MLE when** Y **is categorical.**

Illustration: NIR for Protein Content of Wheat. $p = 6$**,** $n = 50$**.** $\mathbf{f}_y^T = (y, y^2, y^3)$**;** $.9 < \widehat{\text{corr}}(X_i, X_k) < .999$**.** $\widehat{d} = 1$.



Sufficient reduction



First PC

Where can we go from here?

1. PC: $\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \sigma\boldsymbol{\varepsilon}$

2. PFC: $\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon}$

Envelope models for increased efficiency

3. PFC: $\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon}$

Models for increased scope

$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Delta}_y{}^{1/2}\boldsymbol{\varepsilon}$

Increasing Scope

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Delta}_y^{1/2}\boldsymbol{\varepsilon}$$

$R(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$ **is a sufficient reduction iff**

1. $\text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}) \subseteq \text{span}(\boldsymbol{\alpha})$, **where** $\boldsymbol{\Delta} = \mathbf{E}(\boldsymbol{\Delta}_Y)$.
2. $\boldsymbol{\alpha}_0^T\boldsymbol{\Delta}_y^{-1}$ **is constant in** y, **where** $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0)$ **is orthogonal.**

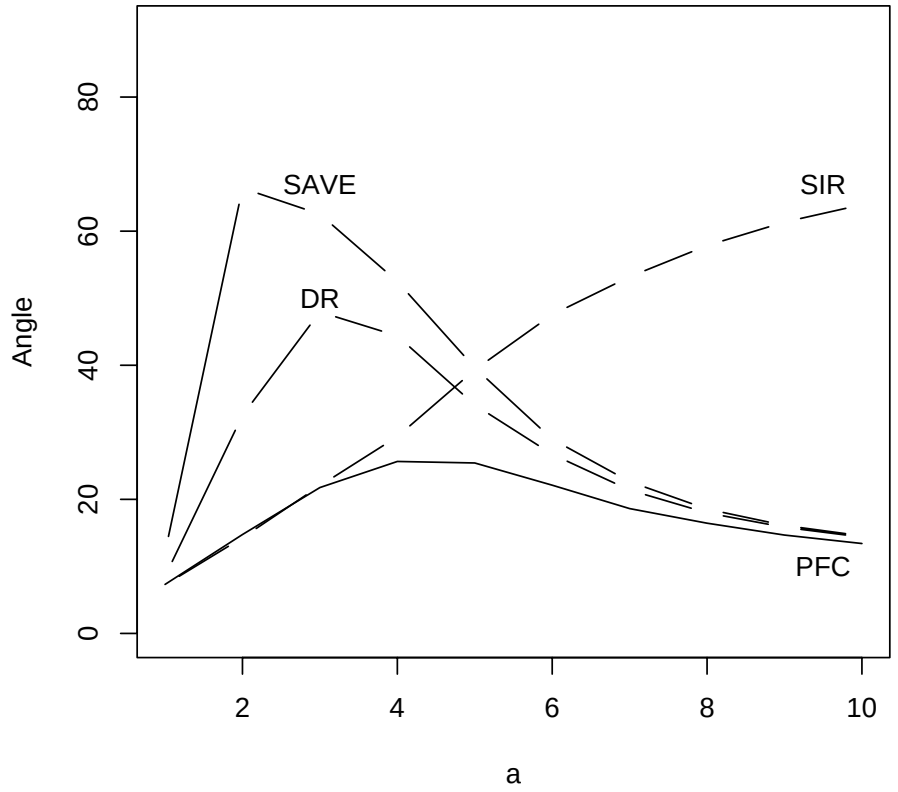
For y **categorical,** $\widehat{R}(\mathbf{X}) = \widehat{\boldsymbol{\alpha}}^T\mathbf{X}$, $\widehat{\boldsymbol{\alpha}}$ **is a basis for**

$$\widehat{\mathcal{S}}_{Y|\mathbf{X}} = \arg\max_{\mathcal{S}\in\mathcal{G}_{d,p}} n\log|P_{\mathcal{S}}\widehat{\boldsymbol{\Sigma}}P_{\mathcal{S}}|_0 - \sum_y n_y \log|P_{\mathcal{S}}\widehat{\boldsymbol{\Delta}}_y P_{\mathcal{S}}|_0$$

$|\mathbf{A}|_0 =$ **product of non-zero eigenvalues of** $\mathbf{A}$.

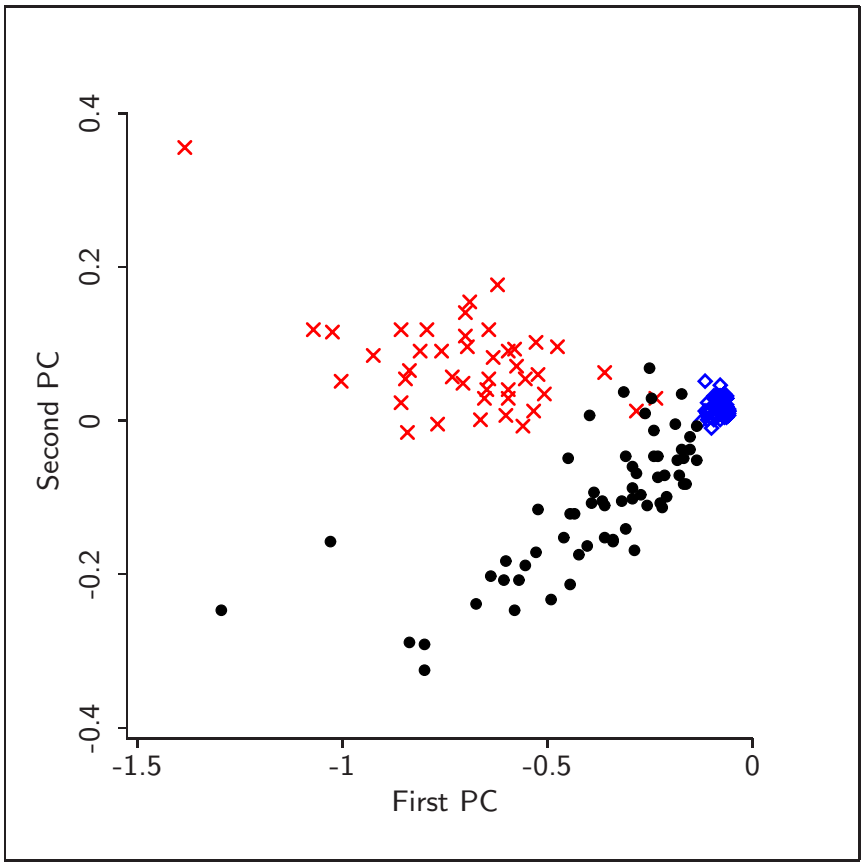
Simulation Model, $p = 8$

$$Y = X_1/a + aX_1^2/10 + \varepsilon, \mathcal{S}_{Y|\mathbf{X}} = \text{span}(1, 0 \ldots, 0)^T$$

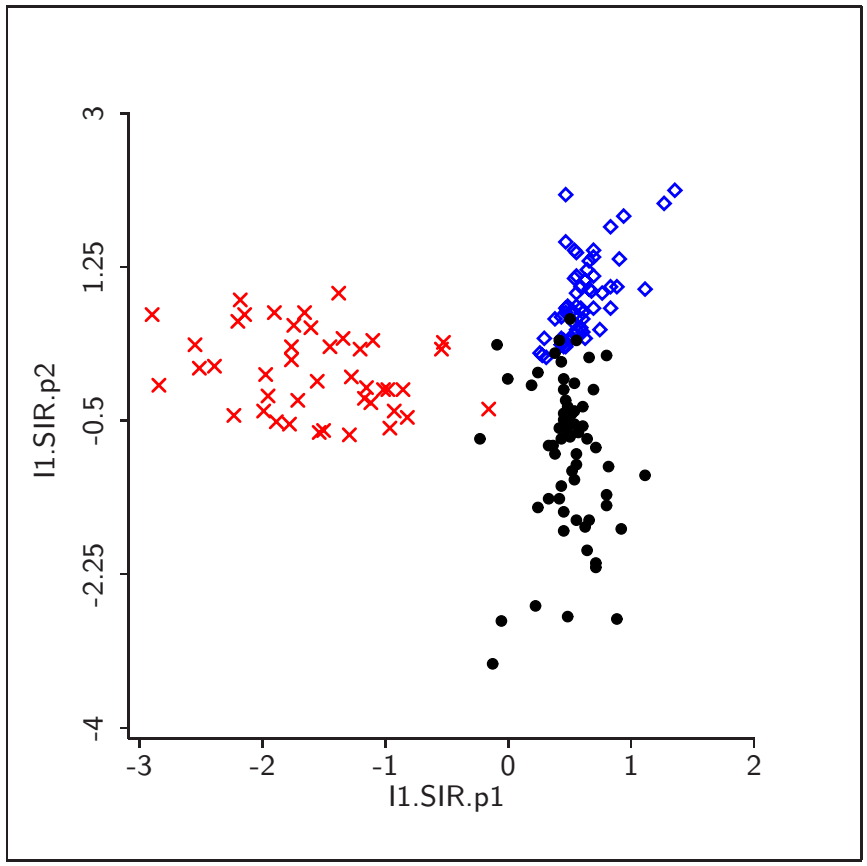


angle vs. a

Birds and Planes and Cars

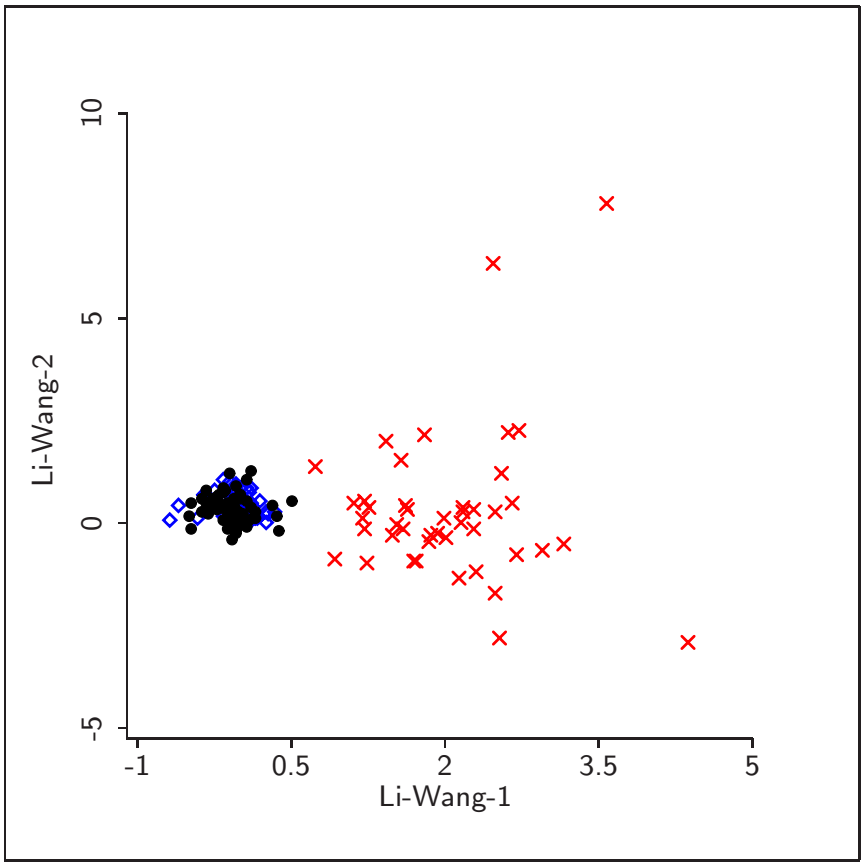


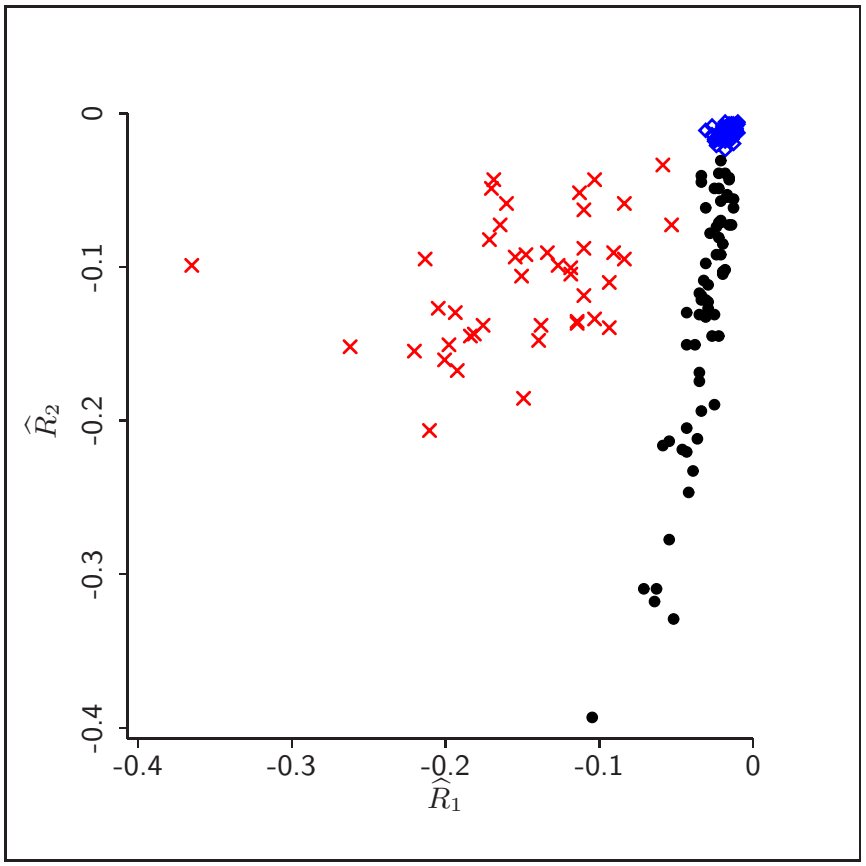
PC



LDA

Birds and Planes and Cars





Li & Wang, JASA, 2007

Sufficient reductions: $\widehat{R}_1$ & $\widehat{R}_2$

Envelope Models for Increasing Efficiency

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon}$$
$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon}$$

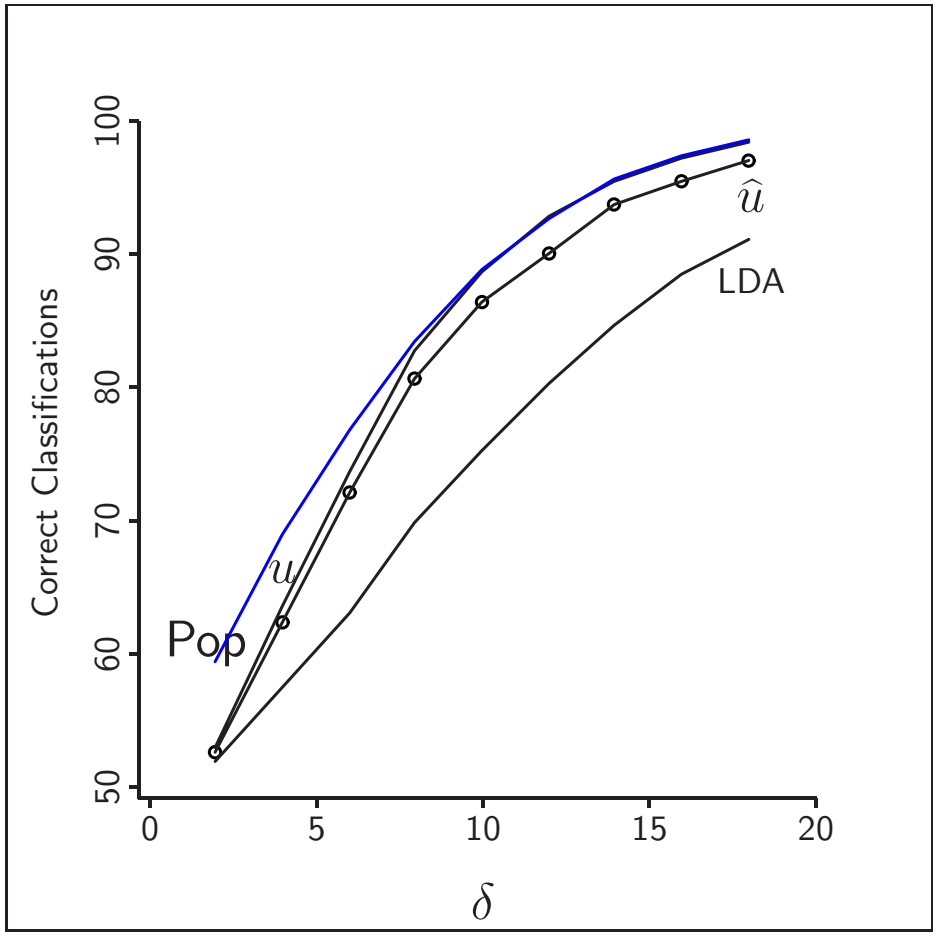
$\mathcal{E}_{\boldsymbol{\Delta}}(\mathcal{S}_{\boldsymbol{\Gamma}})$ = **smallest reducing subspace of $\boldsymbol{\Delta}$ that contains $\mathcal{S}_{\boldsymbol{\Gamma}}$, the $\boldsymbol{\Delta}$ envelope of $\mathcal{S}_{\boldsymbol{\Gamma}}$.** $u = \dim\{\mathcal{E}_{\boldsymbol{\Delta}}(\mathcal{S}_{\boldsymbol{\Gamma}})\}$**.**

Let $\boldsymbol{\Phi}$: $p \times u$ **orthonormal basis for $\mathcal{E}_{\boldsymbol{\Delta}}(\mathcal{S}_{\boldsymbol{\Gamma}})$.** $d \leq u \leq p$**.**
$\boldsymbol{\Theta}$: $u \times d$ **semi-orthogonal matrix.**

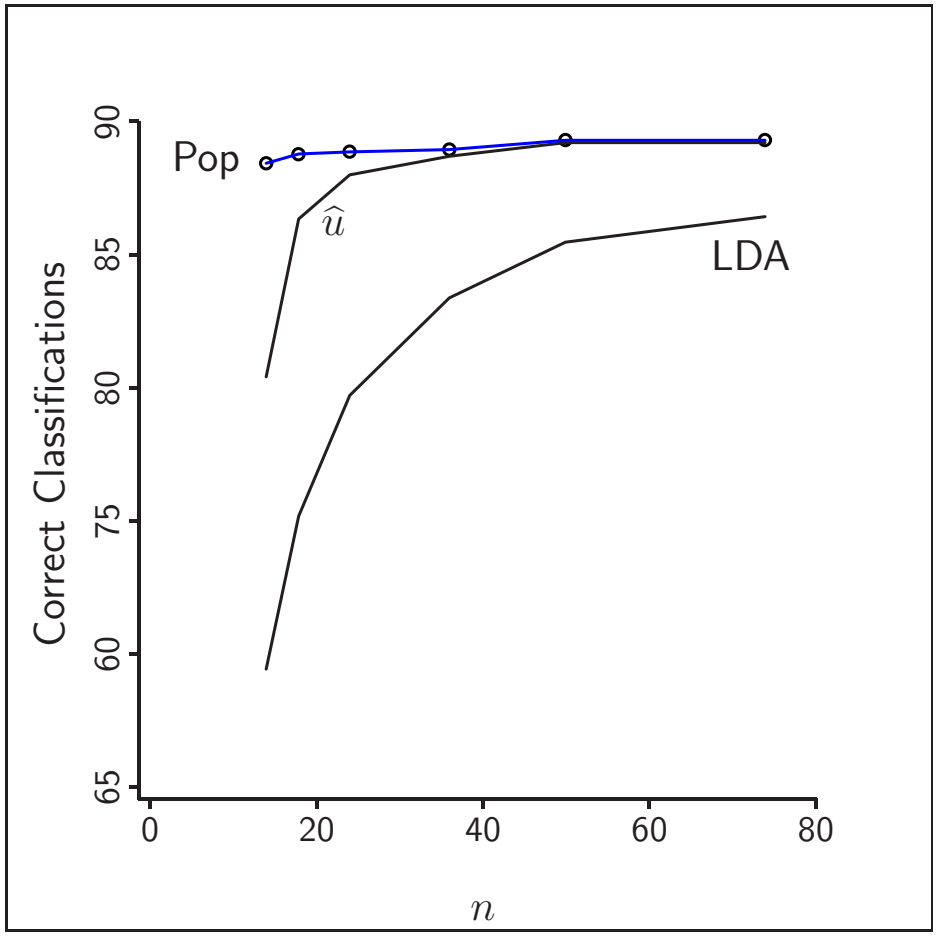
$$\begin{aligned} \mathbf{X}_y &= \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\Theta}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon} \\ \boldsymbol{\Delta} &= \boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^T + \boldsymbol{\Phi}_0\mathbf{M}_0\boldsymbol{\Phi}_0^T \\ R(\mathbf{X}) &= \boldsymbol{\alpha}^T\mathbf{X} \end{aligned}$$

$\boldsymbol{\alpha}$ **is a basis for** $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\Phi}\mathbf{M}^{-1}\boldsymbol{\Theta})$**.**

Discriminant analysis, two equally likely spherical normal populations, $p = 10$**,** $n = 18$**,** $\delta = \|\mu_1 - \mu_2\|$**.** 100 **obs/pop for classification,** 400 **reps.**



Same setting but varying n with $\delta = 10$.



Comments

- **Other distributions for $\mathbf{X}_y$, like $\mathbf{X}_y \sim QEF$. (Genotype data; SNPs)**
- **Screening: Regression of each elements of $\mathbf{X}_y$ on $\mathbf{f}_y$, looking for any dependence.**
- **Covariance reducing methods with application in evolutionary biology.**
  - **With y categorical, characterize $\{\mathbf{\Sigma}_y\}$**
  - **Flury's spectral models**
  - **Instead, base methodology on finding $\alpha \in \mathbb{R}^{p \times d}$ so that the distn of $\widehat{\mathbf{\Sigma}}_y | (\alpha^T \widehat{\mathbf{\Sigma}}_y \alpha, n_y)$ is constant in y.**

- **Multivariate normal linear model**

$$\mathbf{Y}_i = \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{X}_i + \mathbf{V}^{1/2} N(0, I_r), \quad \mathbf{V} \geq 0, \ i = 1, n$$

  with inference on β and prediction as the goals.

  - **Reparameterize in terms of $\mathcal{E}_{\mathbf{V}}(\mathrm{span}(\boldsymbol{\beta}))$**
  - **Use MLE and find the var in the limiting distn of $\widehat{\boldsymbol{\beta}}$.**

- **Partial least squares**

- $\mathbf{X}|Y$ **vs** $Y|\mathbf{X}$**.**