This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'avenir

Dimension Reduction via Sliced Inverse Regression:

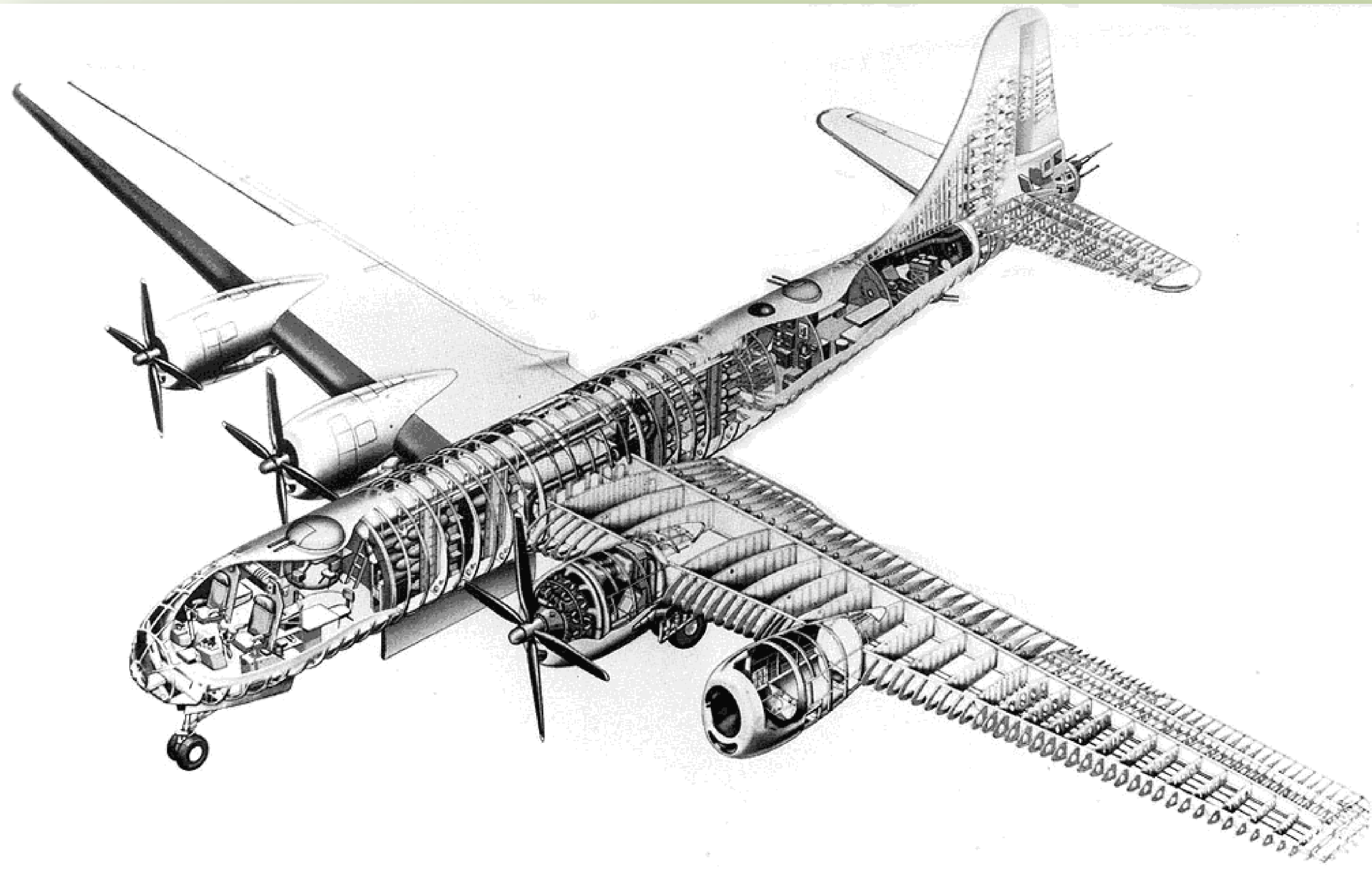
Ideas and Extensions

Alessandro Chiancone

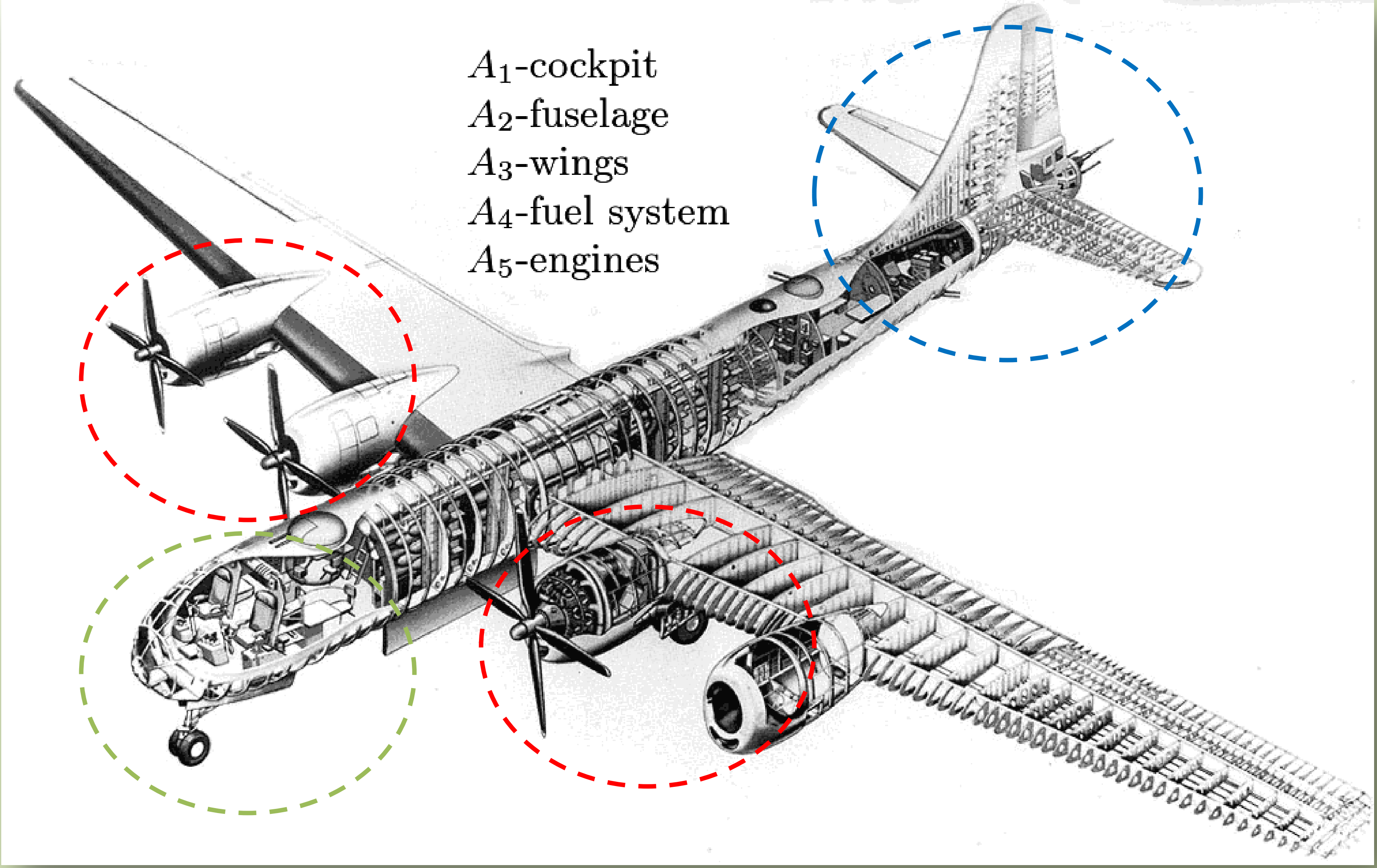
Jocelyn Chanussot

Stéphane Girard

A PROBLEM FROM WWII



A PROBLEM FROM WWII



A PROBLEM FROM WWII

$$\mathbf{X} = (x_1, x_2, x_3, x_4, x_5) \in \mathbb{R}^5,$$

where x_i is the area of A_i hit by bullets.

$$Y = f(\mathbf{X}) : \mathbb{R}^5 \longrightarrow [0, 1]$$

f-overall damage of the aircraft

A PROBLEM FROM WWII

$$\mathbf{X} = (x_1, x_2, x_3, x_4, x_5) \in \mathbb{R}^5,$$
where x_i is the area of A_i hit by bullets.

$$Y = f(\mathbf{X}) : \mathbb{R}^5 \longrightarrow [0, 1]$$
f-overall damage of the aircraft

Y depends only on the total area hit by bullets:
$$Y = f(x_1 + x_2 + x_3 + x_4 + x_5) = f(\beta^T \mathbf{X}),\ \beta = (1, 1, ..., 1)$$

$$\mathbf{X} = (x_1, x_2, x_3, x_4, x_5) \in \mathbb{R}^5,$$
where x_i is the area of A_i hit by bullets.

$$Y = f(\mathbf{X}) : \mathbb{R}^5 \longrightarrow [0, 1]$$
f-overall damage of the aircraft

Y depends only on the total area hit by bullets:
$$Y = f(x_1 + x_2 + x_3 + x_4 + x_5) = f(\beta^T \mathbf{X}),\ \beta = (1, 1, ..., 1)$$

There are more and less vulnerable parts of the plane that have a different impact:
$$Y = f(\beta^T \mathbf{X}),\ \beta =?$$

Fundamental Problem

Given the n-sample $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$ generated from the model

$$Y = f(\beta^T \mathbf{X})$$

is there a way to recover β if there are no assumption on f?

Fundamental Problem

Given the n-sample $(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n)$ generated from the model

$$Y = f(\beta^T \mathbf{X})$$

is there a way to recover β if there are no assumption on f?

YES! Sliced Inverse Regression solves this problem

K.-C. Li, Sliced inverse regression for dimension reduction,
Journal of the American Statistical Association 86 (414) (1991) 316–327.

Let us consider a r.v. $\mathbf{X} \in \mathbb{R}^p$ and the model

$$Y = g(\mathbf{X}, \epsilon) : \mathbb{R}^{p+1} \longrightarrow \mathbb{R}$$

where ϵ is an error independent of $\mathbf{X}$.

Let us consider a r.v. $\mathbf{X} \in \mathbb{R}^p$ and the model

$$Y = g(\mathbf{X}, \epsilon) : \mathbb{R}^{p+1} \longrightarrow \mathbb{R}$$

where ϵ is an error independent of $\mathbf{X}$.

$$g(\mathbf{X}, \epsilon) = f(\beta_1^T \mathbf{X}, ..., \beta_k^T \mathbf{X}, \epsilon) : \mathbb{R}^{k+1} \longrightarrow \mathbb{R}$$

when $k = p$, is trivial $(f = g)$. If $k \ll p$ we can solve the same regression problem in lower dimension if the vectors $\beta_1, ..., \beta_k \in \mathbb{R}^p$ are available.

Model assumption

$$Y = f(\beta_1^T \mathbf{X}, ..., \beta_k^T \mathbf{X}, \epsilon)$$

Linearity Design Condition (LDC)

$$\forall b \in \mathbb{R}^p \;\; \mathbb{E}(b^t \mathbf{X} | \beta_1^T \mathbf{X}, ..., \beta_k^T \mathbf{X})$$

is linear in $\beta_1^T \mathbf{X}, ..., \beta_k^T \mathbf{X}$

Model assumption

$$Y = f(\beta_1^T \mathbf{X}, \ldots, \beta_k^T \mathbf{X}, \epsilon)$$

Linearity Design Condition (LDC)

$$\forall b \in \mathbb{R}^p \;\; \mathbb{E}(b^t \mathbf{X} | \beta_1^T \mathbf{X}, \ldots, \beta_k^T \mathbf{X})$$

is linear in $\beta_1^T \mathbf{X}, \ldots, \beta_k^T \mathbf{X}$

Theorem (Li, 1991). The centered inverse regression curve $\mathbb{E}(\mathbf{X}|Y) - \mathbb{E}(\mathbf{X})$ is contained in the space spanned by the $\Sigma^{-1}\beta_i, i = 1, \ldots, k$, where $\Sigma = cov(\mathbf{X})$.

Corollary. The matrix $\Sigma^{-1}\Gamma$, $\Gamma = cov(\mathbb{E}(\mathbf{X}|Y))$ is degenerated in any direction orthogonal to the effective dimension reduction (e.d.r.) directions $\beta_i, i = 1, \ldots, k$.

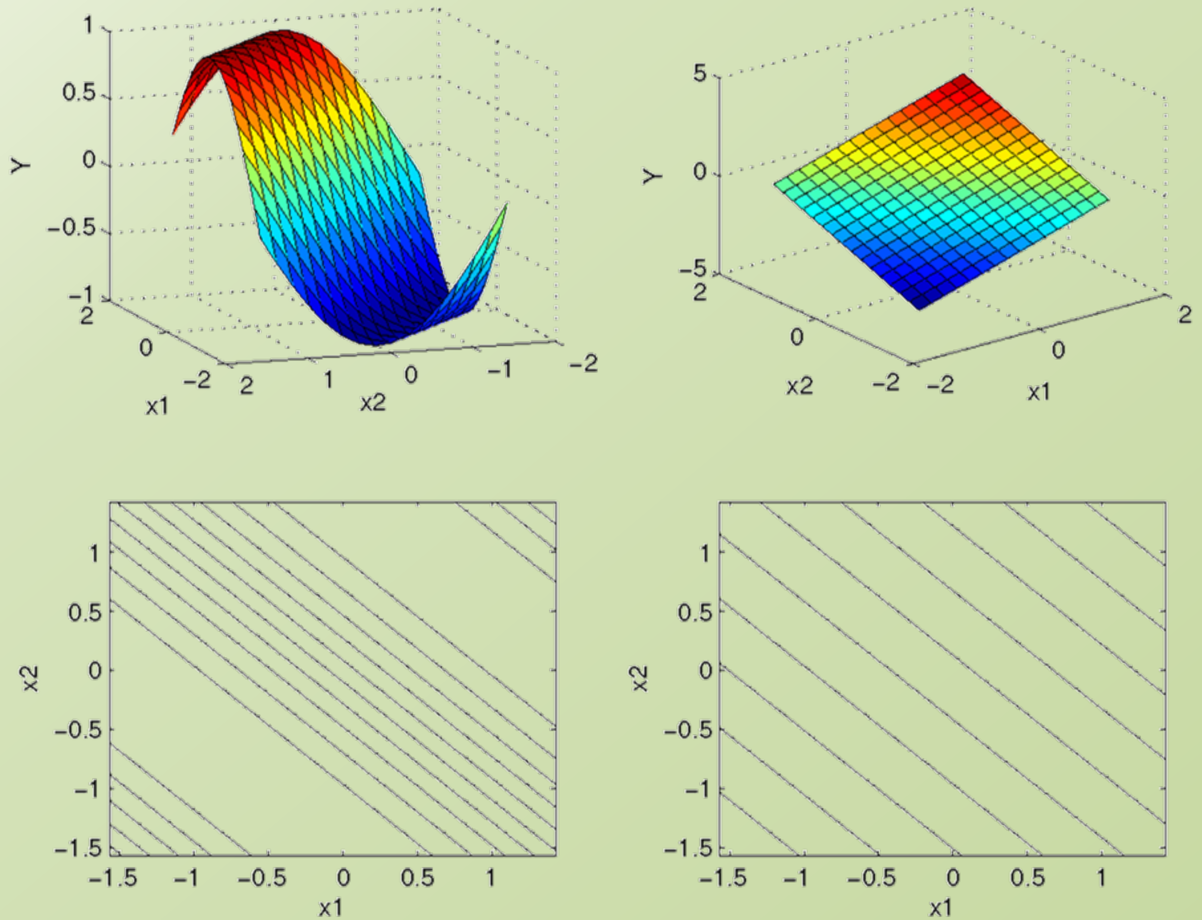
K.-C. Li, Sliced inverse regression for dimension reduction, Journal of the American Statistical Association 86 (414) (1991) 316–327.

- Split the range of Y in h slices
- Estimate Γ using the slices.
  Γ is the between-slice covariance matrix
- Compute the eigendecomposition of the matrix $\Sigma^{-1}\Gamma$
- Select the eigenvectors corresponding to the k-highest eigenvalues

$\rightarrow$ $\beta_1, \ldots, \beta_k$

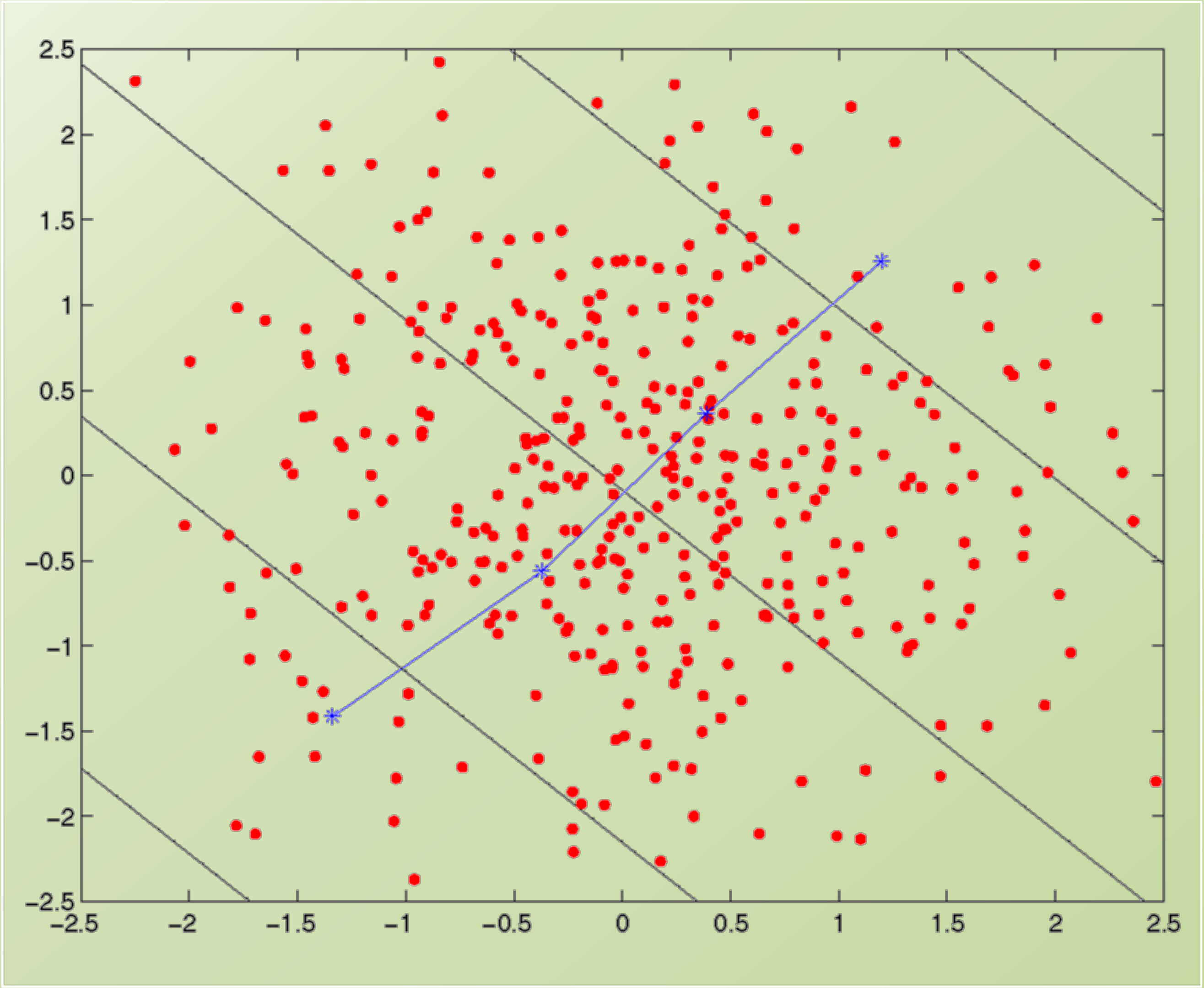
COUNTOUR LINES

$$\mathbf{X} \in \mathbb{R}^2,\ Y = g(x_1 + x_2),\ \beta = (1, 1)$$



WHY SIR WORKS?

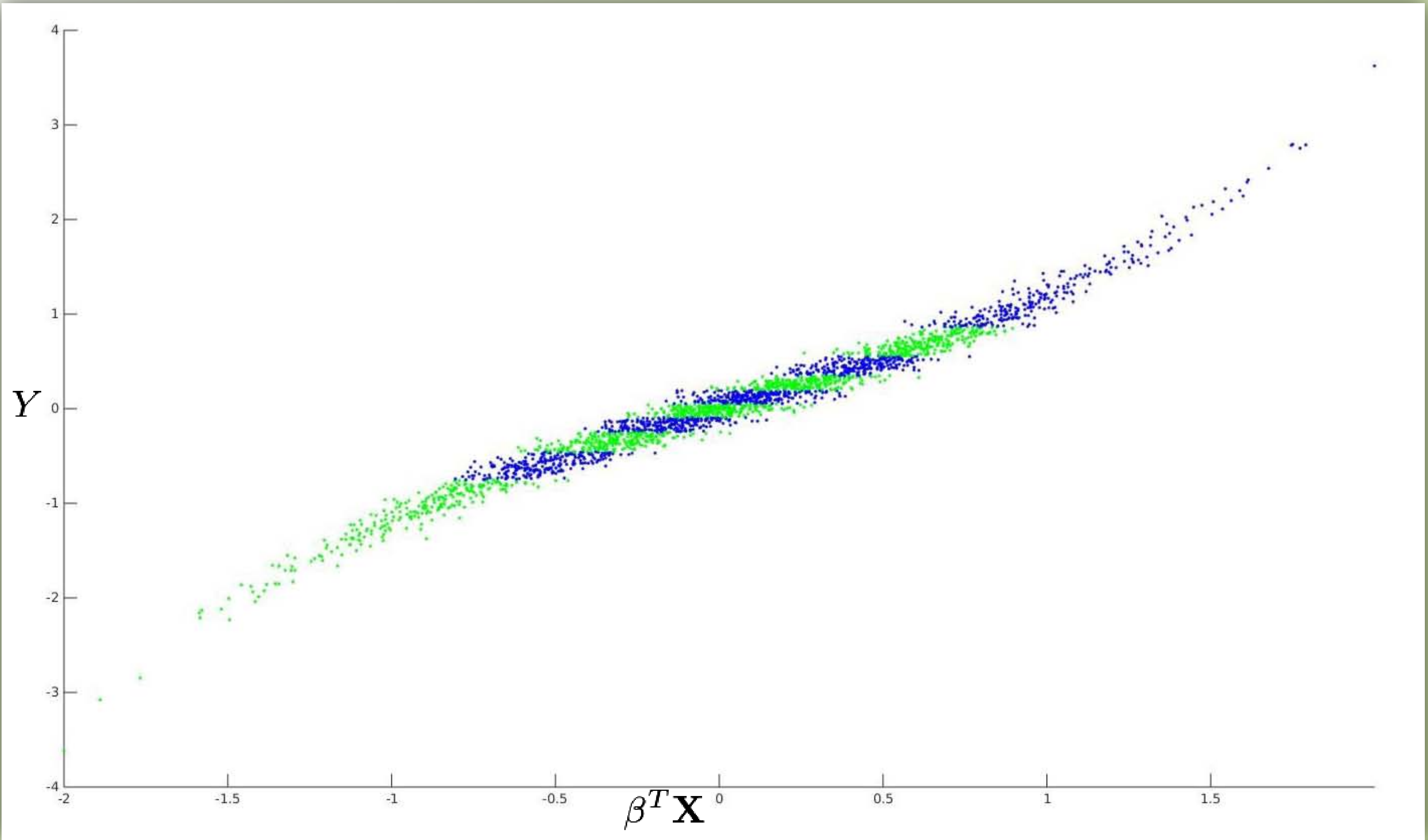
$$\mathbf{X} \in \mathbb{R}^2,\ Y = g(x_1 + x_2),\ \beta = (1, 1)$$



SIR IN ACTION

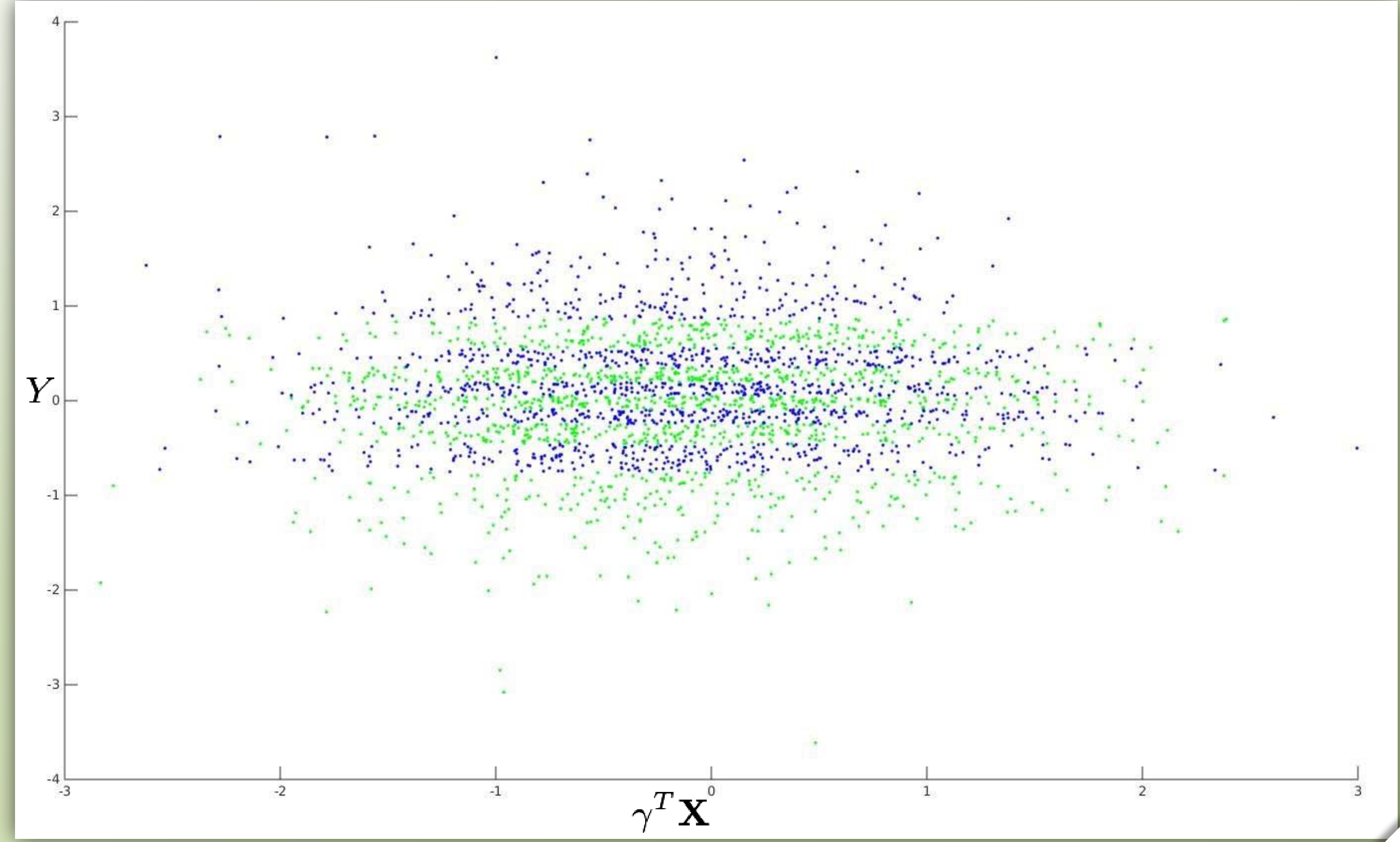
Li (1991)

$$\mathbf{X} \in \mathbb{R}^2,\ Y = sin(x_1 + x_2) + \epsilon,\ \beta = (1,1),\ \epsilon \sim N(0,V)$$



SIR IN ACTION

$$\mathbf{X} \in \mathbb{R}^2,\ Y = sin(x_1 + x_2) + \epsilon,\ \beta = (1, 1),\ \epsilon \sim N(0, V)$$



WEAK POINTS OF SIR AND EXTENSIONS

- The Linearity Design Condition (LDC)
- SIR is not robust to outliers
- Sparsity

- The Linearity Design Condition (LDC)

R. D. COOK AND C. J. NACHTSHEIM, Reweighting to achieve elliptically contoured covariates in regression, Journal of the American Statistical Association, 89 (1994), pp. 592–599.

L. LI, R. D. COOK, AND C. J. NACHTSHEIM, Cluster-based estimation for sufficient dimension reduction, Computational Statistics & Data Analysis, 47 (2004), pp. 175–193.

V. KUENTZ AND J. SARACCO, Cluster-based sliced inverse regression, Journal of the Korean Statistical Society, 39 (2010), pp. 251–267.

Linearity Design Condition (LDC)

$$\forall b \in \mathbb{R}^p \ \ \mathbb{E}(b^t\mathbf{X}|\beta_1^T\mathbf{X},...,\beta_k^T\mathbf{X})$$

is linear in $\beta_1^T\mathbf{X},...,\beta_k^T\mathbf{X}$

Linearity Design Condition (LDC)

$$\forall b \in \mathbb{R}^p \;\; \mathbb{E}(b^t\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})$$

is linear in $\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X}$

The (LDC) holds if $\mathbf{X}$ follows an elliptically symmetric distribution

Linearity Design Condition (LDC)

$$\forall b \in \mathbb{R}^p \;\; \mathbb{E}(b^t\mathbf{X}|\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X})$$

is linear in $\beta_1^T\mathbf{X}, ..., \beta_k^T\mathbf{X}$

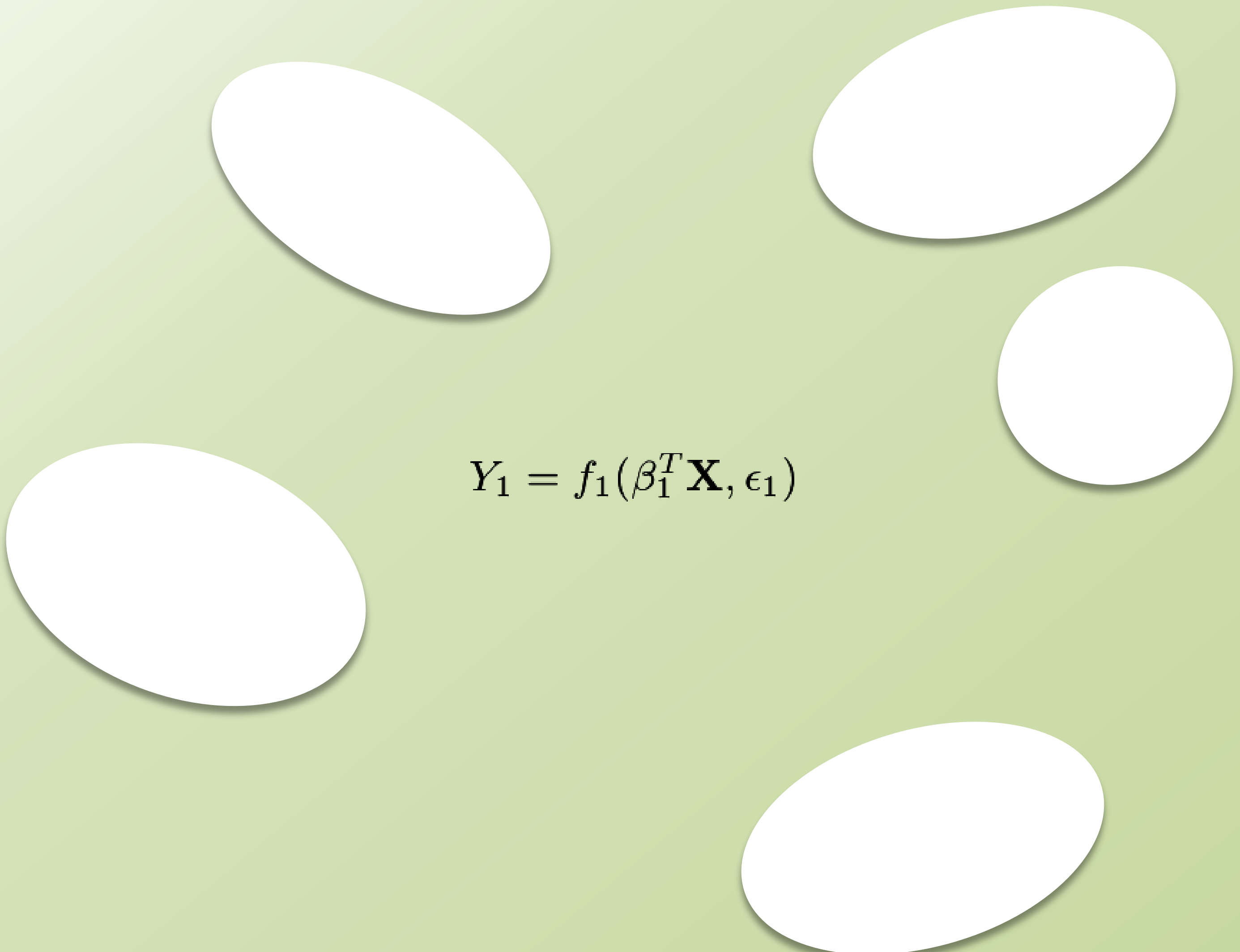
The (LDC) holds if $\mathbf{X}$ follows an elliptically symmetric distribution

In case of mixture of elliptically symmetric distributions the (LDC) holds locally but not globally

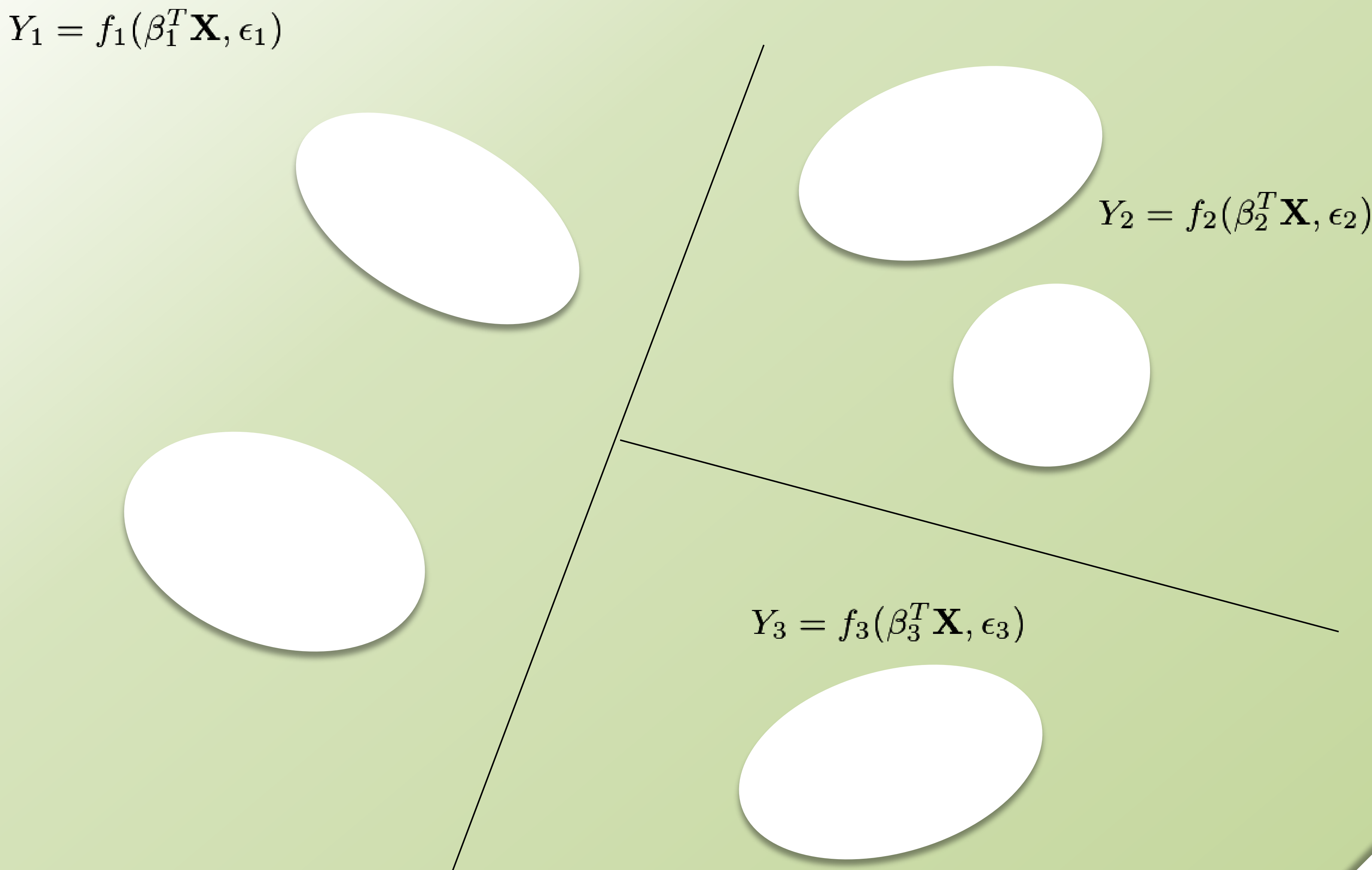
Kuentz&Saracco clusterized the predictor space to force the condition to hold locally

CLUESTER-BASED SIR

Kuentz&Saracco (2010)

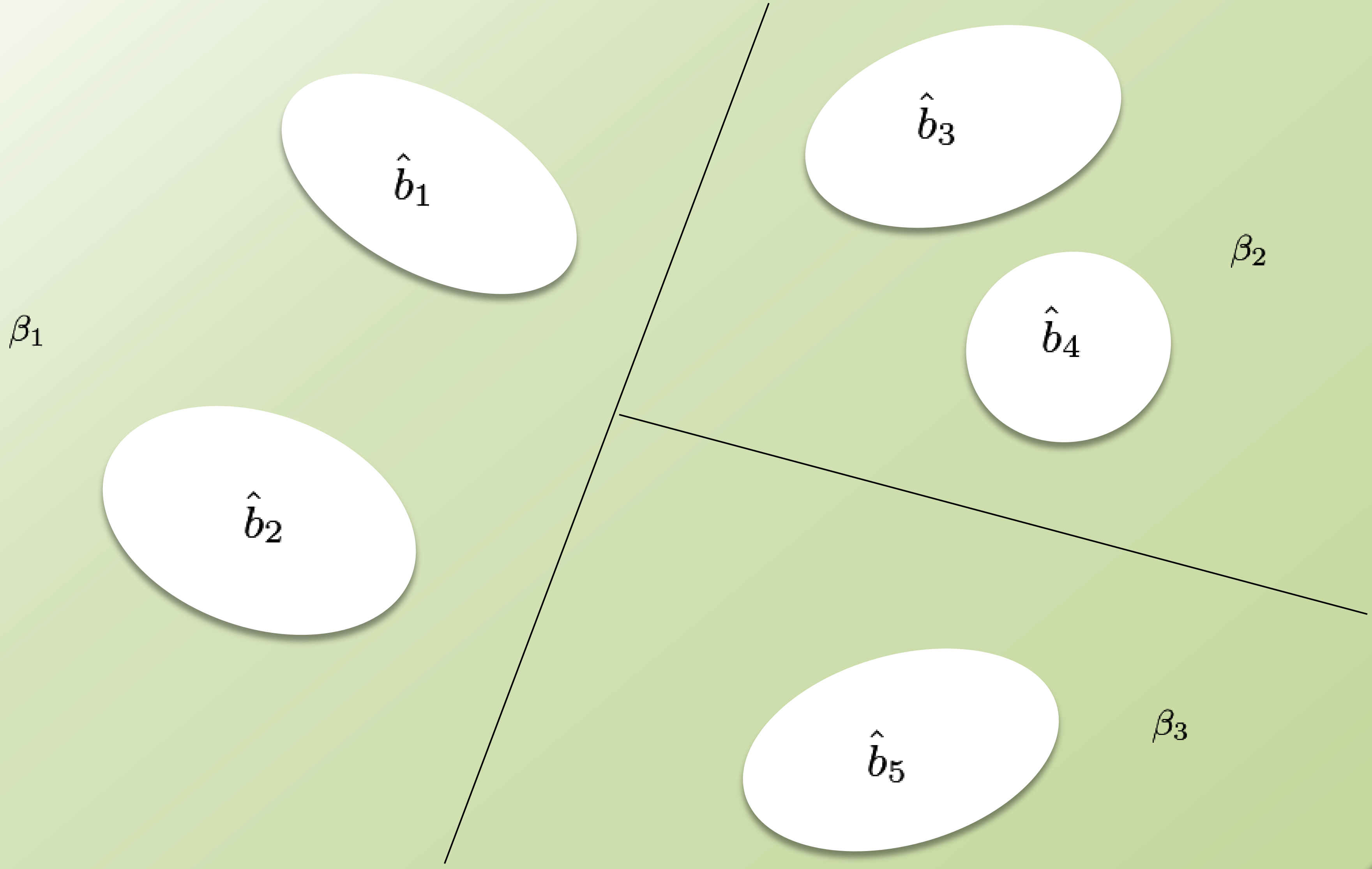


$$Y_1 = f_1(\beta_1^T \mathbf{X}, \epsilon_1)$$



$Y_1 = f_1(\beta_1^T \mathbf{X}, \epsilon_1)$
$Y_2 = f_2(\beta_2^T \mathbf{X}, \epsilon_2)$
$Y_3 = f_3(\beta_3^T \mathbf{X}, \epsilon_3)$

COLLABORATIVE SIR



Let $V = \{v_1, v_2, ..., v_{|V|}\}$ be a set of unit vectors in dimension p with associated weights w_i. We define the quantity $\lambda(V)$:

$$\lambda(V) = \max_{v \in \mathbb{R}^p} \frac{1}{\sum w_i} \sum_{i=1}^{|V|} w_i m(v_i, v) \text{ s.t. } ||v|| = 1$$

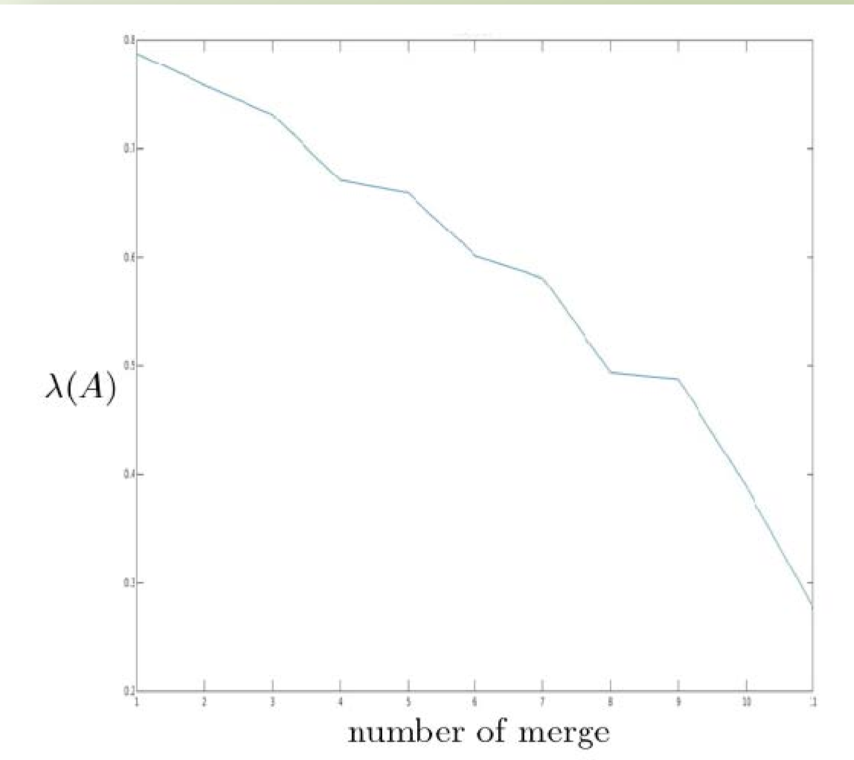
Let $V = \{v_1, v_2, ..., v_{|V|}\}$ be a set of unit vectors in dimension p with associated weights w_i. We define the quantity $\lambda(V)$:

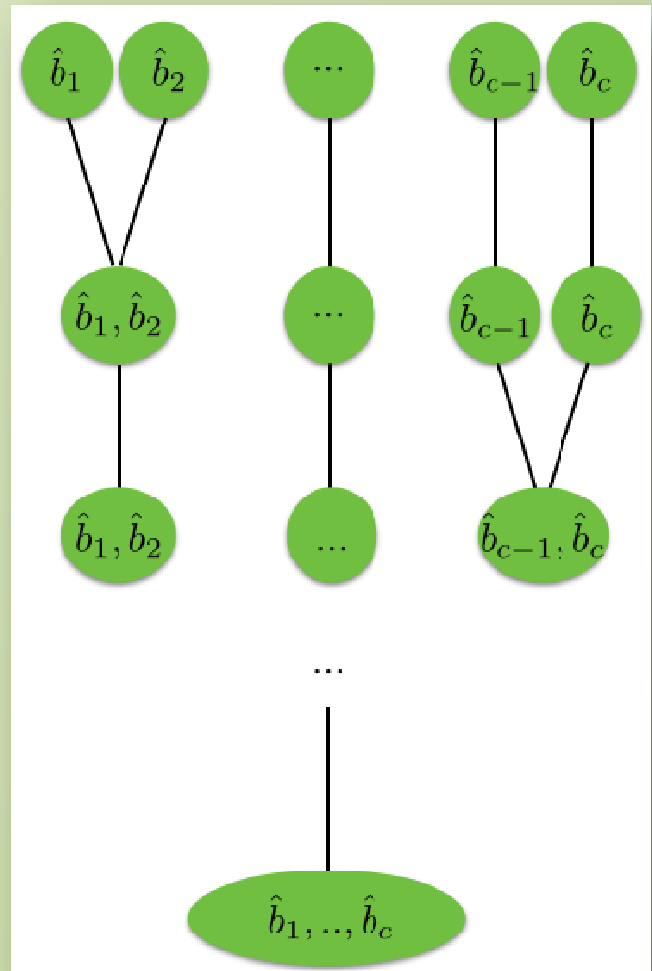
$$\lambda(V) = \max_{v \in \mathbb{R}^p} \frac{1}{\sum w_i} \sum_{i=1}^{|V|} w_i m(v_i, v) \text{ s.t. } ||v|| = 1$$

setting $m(v_i, v) = cos^2(v_i, v) = (v_i^T v)^2$

$$\lambda(V) = \text{largest eigenvalue of } \frac{1}{\sum w_i} \sum_{i=1}^{|V|} w_i v_i v_i^T$$

COLLABORATIVE SIR





Galaxy dataset

$n = 362887$

Y stellar formation rate

$\mathbf{X} \in \mathbb{R}^p$ $p = 46$-spectral characteristics of the galaxies

Aberrant points have been removed in accordance with experts

↓

$n = 292766$

A special thanks goes to Didier Fraix-Burnet, IPAG

Galaxy dataset

$$n = 362887$$

Y stellar formation rate

$\mathbf{X} \in \mathbb{R}^p$ $p = 46$-spectral characteristics of the galaxies

Aberrant points have been removed in accordance with experts

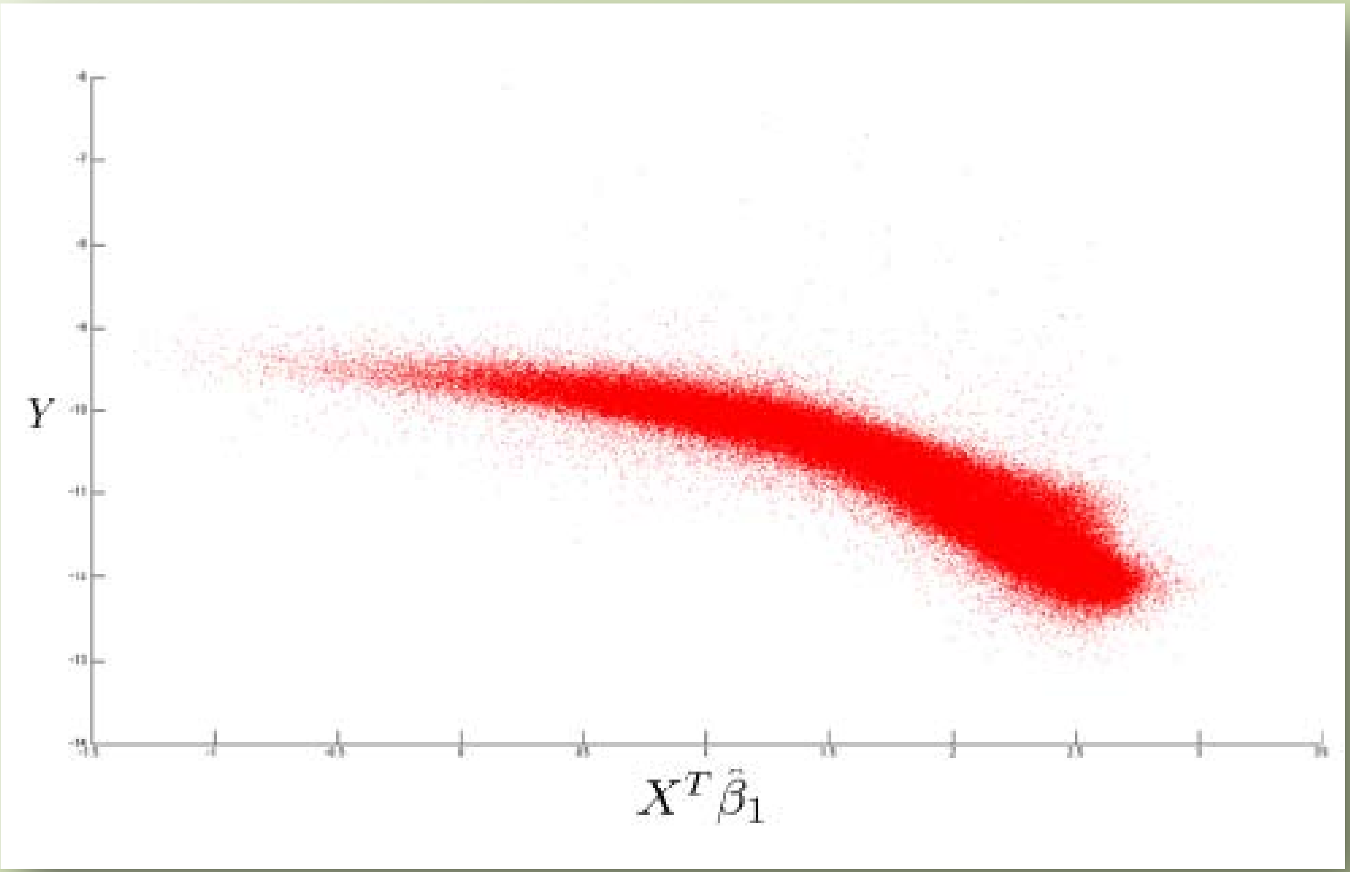
↓

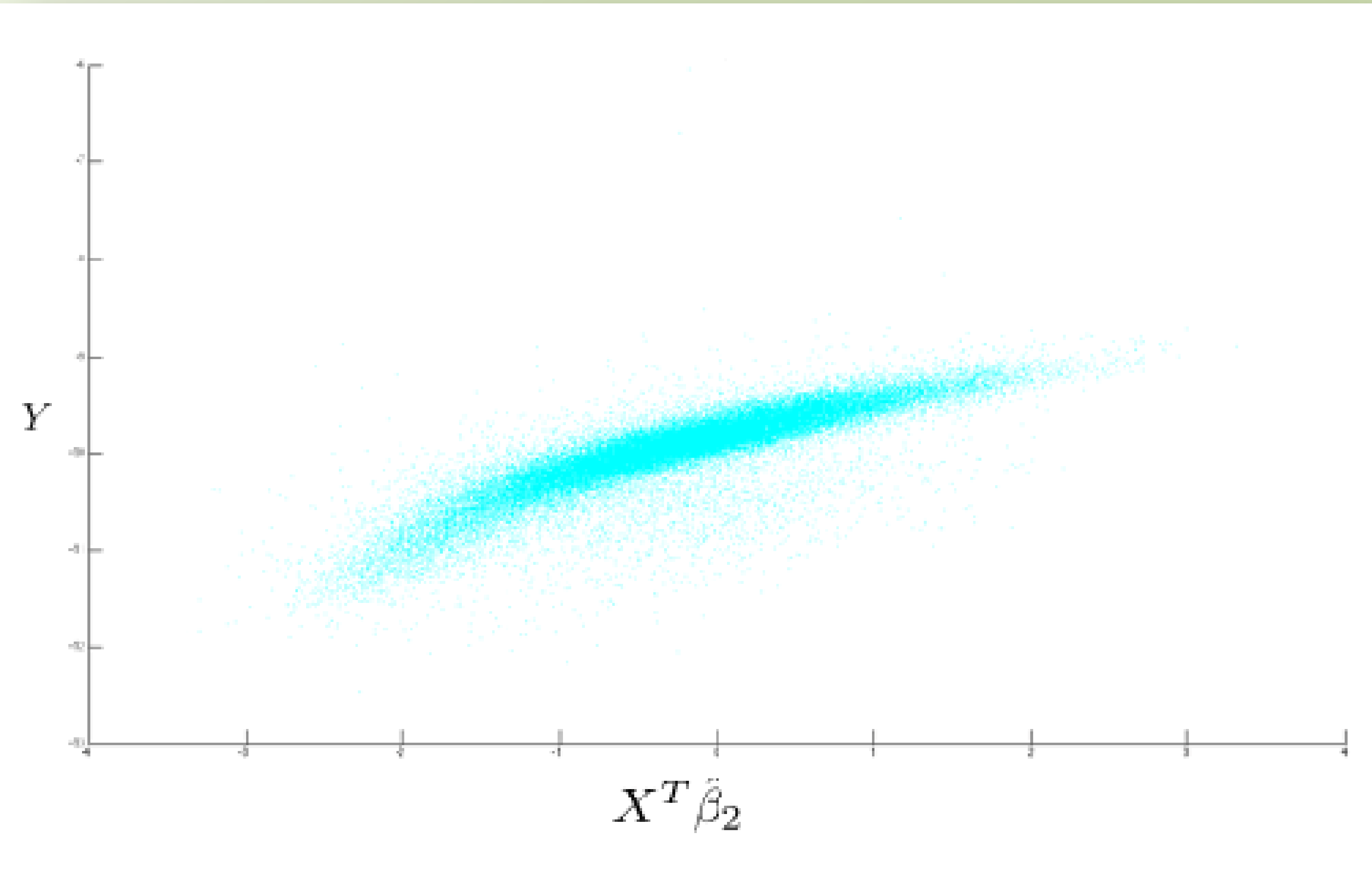
$$n = 292766$$

A special thanks goes to Didier Fraix-Burnet, IPAG

Collaborative SIR found two distinct subrgoups and directions $\hat{\beta}_1, \hat{\beta}_2$

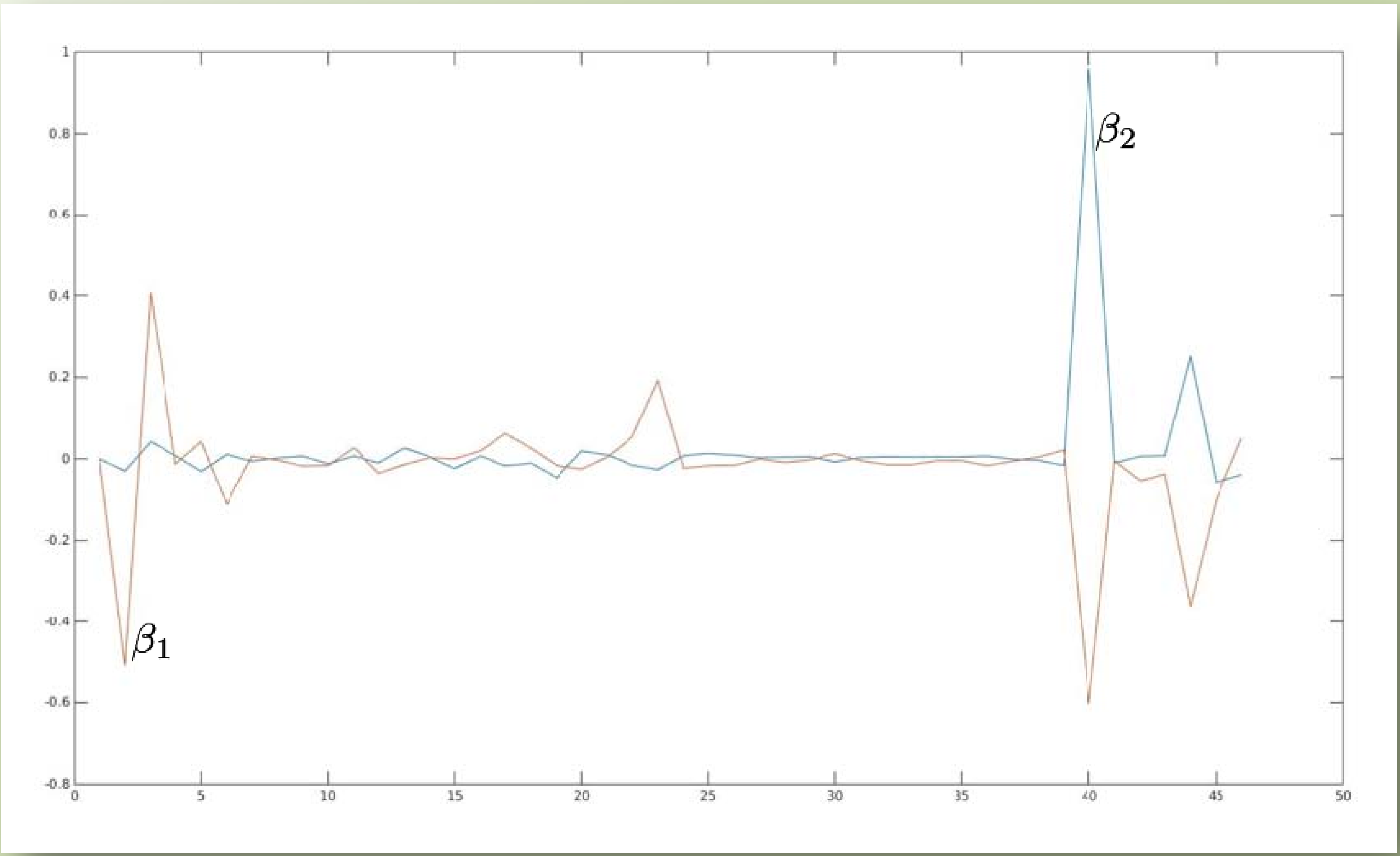
COLLABORATIVE SIR




Y
$X^T\hat{\beta}_2$

COLLABORATIVE SIR

Chiancone, Girard &Chanussot (2015)



- The Linearity Design Condition (LDC)
- SIR is not robust to outliers
- Sparsity

TRIGGERING PAPERS

- SIR is not robust to outliers

U. GATHER, T. HILKER, C. BECKER, A note on outlier sensitivity of sliced inverse regression. Statistics: A Journal of Theoretical and Applied Statistics 36.4 (2002): 271-281.

U. GATHER, T. HILKER, C. BECKER, A robustified version of sliced inverse regression, in: Statistics in Genetics and in the Environmental Sciences, Trends in Mathematics, Springer, 2001, Ch. 2, pp. 147–157.

B. LI AND Y. DONG, Dimension reduction for nonelliptically distributed predictors, The Annals of Statistics, 37 (2009), pp. 1272–1298.

STUDENT SIR

- Remove outliers before the analysis
- Adopt a more flexible model

Dimension reduction subspace (d.r.s) S is a space such that:

Y is independent of $\mathbf{X}$ given $P_S\mathbf{X}$, P_S-orthogonal projection on S

All information necessary to regress Y is contained in $P_S\mathbf{X}$.

Dimension reduction subspace (d.r.s) S is a space such that:

Y is independent of $\mathbf{X}$ given $P_S\mathbf{X}$, P_S-orthogonal projection on S

All information necessary to regress Y is contained in $P_S\mathbf{X}$.

The smallest d.r.s is called central subspace $S_{Y|X}$

SIR estimates a basis of this space

R. D. COOK (1996), "Graphics for Regressions With Binary Response," Journal of the American Statistical Association, 91, 983–992.

R. D. COOK, Fisher lecture: Dimension reduction in regression, Statistical Science, 22 (2007), pp. 1–26.

$$\mathbf{X} = \mu + VBc(Y) + \epsilon$$

where $\mu \in \mathbb{R}^p$ is a non random vector, B is a non random $p \times d$ matrix with $B^T B = I_d$ and c is a nonrandom function.

ϵ centered random error independent of Y with covariance matrix V

$$\mathbf{X} = \mu + VBc(Y) + \epsilon$$

where $\mu \in \mathbb{R}^p$ is a non random vector, B is a non random $p \times d$ matrix with $B^T B = I_d$ and c is a nonrandom function.

ϵ centered random error independent of Y with covariance matrix V

$\mathbb{E}(X|Y = y) = \mu + VBc(y)$ is contained in the subspace spanned by VB

$$\mathbf{X} = \mu + VBc(Y) + \epsilon$$

where $\mu \in \mathbb{R}^p$ is a non random vector, B is a non random $p \times d$ matrix with $B^T B = I_d$ and c is a nonrandom function.

ϵ centered random error independent of Y with covariance matrix V

$\mathbb{E}(X|Y=y) = \mu + VBc(y)$ is contained in the subspace spanned by VB

Proposition. If $\epsilon \sim N(0, V)$ then the distribution of $Y|\mathbf{X}$ is the same as $Y|B^T\mathbf{X}$

R. D. COOK, Fisher lecture: Dimension reduction in regression, Statistical Science, 22 (2007), pp. 1–26.

Theorem. The maximum likelihood estimator of B and the SIR estimator of the directions B span the same space

C. BERNARD-MICHEL, L. GARDES, AND S. GIRARD, Gaussian regularized sliced inverse regression, Statistics and Computing, 19 (2009), pp. 85–98.

Theorem. The maximum likelihood estimator of B and the SIR estimator of the directions B span the same space

C. BERNARD-MICHEL, L. GARDES, AND S. GIRARD, Gaussian regularized sliced inverse regression, Statistics and Computing, 19 (2009), pp. 85–98.

$$\mathbf{X} = \mu + VBc(Y) + \epsilon$$

Can we use an elliptically contoured distribution with heavy tails in order to better accommodate outliers?

Arellano Valle Generalized t-distribution - $\gamma = 1$, $\mathcal{S}_p(\mu, \Sigma, \alpha)$

$$\mathbf{X} = \mu + VBc(y) + \epsilon$$

Proposition. If $\epsilon \sim \mathcal{S}_p(0, V, \alpha)$ then the distribution of $Y|\mathbf{X}$ is the same as $Y|B^T\mathbf{X}$

$(X_i, Y_i)_{i=1,\dots,n}$ we want to find $\{\mu, V, B, \alpha\}$ and $c(\cdot)$

$$\mathbf{X} = \mu + VBC^T s(Y) + \epsilon,\ \theta = \{\mu, V, B, \alpha, C\}$$

No closed form solutions exist for ML estimators

STUDENT SIR

$(X_i, Y_i)_{i=1,\dots,n}$ we want to find $\{\mu, V, B, \alpha\}$ and $c(\cdot)$

$$\mathbf{X} = \mu + VBC^T s(Y) + \epsilon,\ \theta = \{\mu, V, B, \alpha, C\}$$

No closed form solutions exist for ML estimators

$$\mathcal{S}_p(x; \mu, \Sigma, \alpha) = \int_0^\infty \mathcal{N}_p(x; \mu, \Sigma/u)\ \mathcal{G}(u; \alpha, 1)\ \mathrm{d}u$$

STUDENT SIR

Chiancone, Forbes & Girard (2016)

$(X_i, Y_i)_{i=1,\ldots,n}$ we want to find $\{\mu, V, B, \alpha\}$ and $c(\cdot)$

$$\mathbf{X} = \mu + VBC^T s(Y) + \epsilon,\ \theta = \{\mu, V, B, \alpha, C\}$$

No closed form solutions exist for ML estimators

$$\mathcal{S}_p(x; \mu, \Sigma, \alpha) = \int_0^\infty \mathcal{N}_p(x; \mu, \Sigma/u)\ \mathcal{G}(u; \alpha, 1)\ \mathrm{d}u$$

Expectation Maximization algorithm

Latent variables $\mathbf{U} = \{U_1, \ldots, U_n\}$ with U_i independent of Y_i:

$$\begin{aligned} X_i | U_i = u_i, Y_i = y_i &\sim \mathcal{N}_p(\mu + VBC^T s(Y_i), V/u_i), \\ U_i | Y_i = y_i &\sim \mathcal{G}(\alpha, 1). \end{aligned}$$

EXPECTATION - STEP

$$\bar{u_i}^{(t)} = E_{U_i}[U_i | x_i, y_i; \theta^{(t-1)}]$$

$$\tilde{u_i}^{(t)} = E_{U_i}[\log U_i | x_i, y_i; \theta^{(t-1)}]$$

EXPECTATION - STEP

$$\bar{u_i}^{(t)} = E_{U_i}[U_i|x_i, y_i; \theta^{(t-1)}]$$
$$\tilde{u_i}^{(t)} = E_{U_i}[\log U_i|x_i, y_i; \theta^{(t-1)}]$$

$$\bar{u}_i^{(t)} = \frac{\alpha^{(t-1)} + \frac{p}{2}}{1 + \frac{1}{2}\delta(x_i, \mu^{(t-1)} + V^{(t-1)}B^{(t-1)}C^{(t-1)T}s_i, V^{(t-1)})}$$
$$\tilde{u}_i^{(t)} = \Psi(\alpha^{(t-1)} + \frac{p}{2}) - \log(1 + \frac{1}{2}\delta(x_i, \mu^{(t-1)} + V^{(t-1)}B^{(t-1)}C^{(t-1)T}s_i, V^{(t-1)}))$$

where Ψ is the digamma function and δ is the Mahalanobis distance:

$$\delta(x_i, \mu + VBC^T s_i, V) = (\mu + VBC^T s_i - x_i)^T V^{-1}(\mu + VBC^T s_i - x_i)$$

MAXIMIZATION - STEP

$h \times h$ weighted covariance matrix W of $s(Y)$ defined by:

$$W = \frac{1}{n} \sum_{i=1}^{n} \bar{u_i} \, (s_i - \bar{s})(s_i - \bar{s})^T,$$

the $h \times p$ weighted covariance matrix M of $(s, \mathbf{X})$ defined by

$$M = \frac{1}{n} \sum_{i=1}^{n} \bar{u_i} \, (s_i - \bar{s})(x_i - \bar{x})^T,$$

and Σ the $p \times p$ weighted covariance matrix of $\mathbf{X}$

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} \bar{u_i} \, (x_i - \bar{x})(x_i - \bar{x})^T$$

where

$$\bar{x} = \frac{1}{\sum_{i=1}^{n} \bar{u_i}} \sum_{i=1}^{n} \bar{u_i} x_i \quad \text{and}$$

$$\bar{s} = \frac{1}{\sum_{i=1}^{n} \bar{u_i}} \sum_{i=1}^{n} \bar{u_i} s_i.$$

Proposition. Considering the inverse model, if W and Σ are regular, then the M-step for (μ, V, B, C) leads to the updated estimations $(\hat{\mu}, \hat{V}, \hat{B}, \hat{C})$ given below

- $\hat{B}$ is made of the eigenvectors associated to the largest eigenvalues of $\Sigma^{-1} M^T W^{-1} M$,
- $\hat{V} = \Sigma - (M^T W^{-1} M B)(B^T M^T W^{-1} M B)^{-1} (M^T W^{-1} M B)^T$.
- $\hat{C} = W^{-1} M \hat{B} (\hat{B}^T \hat{V} \hat{B})^{-1}$ and
- $\hat{\mu} = \bar{x} - \hat{V} \hat{B} \hat{C}^T \bar{s}$.

This proposition generalizes to multivariate t-Student error the result obtained for Gaussian error by BERNARD-MICHEL et al (2009)

For α is easy to see that:

$$\hat{\alpha} = \Psi^{-1}(\tilde{u})$$

Three different regression models are considered:

$$\mathbf{I}: \; Y = 1 + 0.6X_1 - 0.4X_2 + 0.8X_3 + 0.2\varepsilon,$$
$$\mathbf{II}: \; Y = (1 + 0.1\varepsilon)X_1,$$
$$\mathbf{III}: \; Y = X_1/(0.5 + (X_2 + 1.5)^2) + 0.2\varepsilon,$$

where ε follows a standard normal distribution. The three models are combined with three possible distributions for the predictors $\mathbf{X}$:

(i) $\mathbf{X}$ is multivariate normal distributed with mean vector 0 and covariance matrix defined by its entries as $\sigma_{ij} = 0.5^{|i-j|}$;

(ii) $\mathbf{X}$ is standard multivariate Cauchy distributed;

(iii) $\mathbf{X} = (X_1, \ldots, X_p)$, where each X_i is generated independently from a mixture of normal and uniform distributions denoted by $0.8\mathcal{N}(0,1) + 0.2\mathcal{U}(-\nu, \nu)$ where ν is a positive scalar value.

U. GATHER, T. HILKER, C. BECKER, A robustified version of sliced inverse regression, in: Statistics in Genetics and in the Environmental Sciences,Trends in Mathematics, Springer, 2001, Ch. 2, pp. 147–157.
B. LI AND Y. DONG, Dimension reduction for non elliptically distributed predictors, The Annals of Statistics, 37 (2009), pp. 1272–1298.

STUDENT SIR – SIMULATION STUDY

Model	X	Method					
		SIR	CP-SIR	WCAN	WIRE	SIME	st-SIR
I	(i)	**.99(.01)**	.99(.01)	.98(.01)	.98(.01)	**.99(.01)**	**.99(.01)**
	(ii)	.63(.18)	.92(.04)	.88(.06)	.87(.07)	.91(.04)	**.98(.01)**
	(iii)	**.99(.01)**	.86(.12)	.72(.27)	.98(.01)	.97(.01)	**.99(.01)**
II	(i)	**.99(.01)**	.98(.01)	.98(.01)	.98(.01)	98(.01)	**.99(.01)**
	(ii)	.61(.18)	.92(.04)	.89(.06)	.87(.08)	.91(.05)	**.98(.01)**
	(iii)	**.99(.01)**	.67(.25)	.69(.28)	.98(.01)	.97(.02)	**.99(.01)**
III	(i)	.88(.06)	.87(.06)	**.89(.05)**	.86(.06)	.87(.06)	.87(.06)
	(ii)	.40(.13)	.78(.10)	.78(.11)	.76(.11)	.78(.10)	**.85(.06)**
	(iii)	.84(.07)	.63(.12)	.67(.13)	**.85(.07)**	**.85(.07)**	.84(.07)

Proximity measure $r(B, \hat{B}) = \dfrac{trace(BB^T\hat{B}\hat{B}^T)}{d}$

200 repetitions, number of slices $h = 5$, predictor dimension $p = 10$ and sample size $n = 200$

CP-SIR: Contour projection for SIR

WCAN: Weighted canonical correlation

WIRE: Weighted sliced inverse regression estimation

SIME: Sliced inverse multivariate median estimation

Galaxy dataset

$n = 362887$

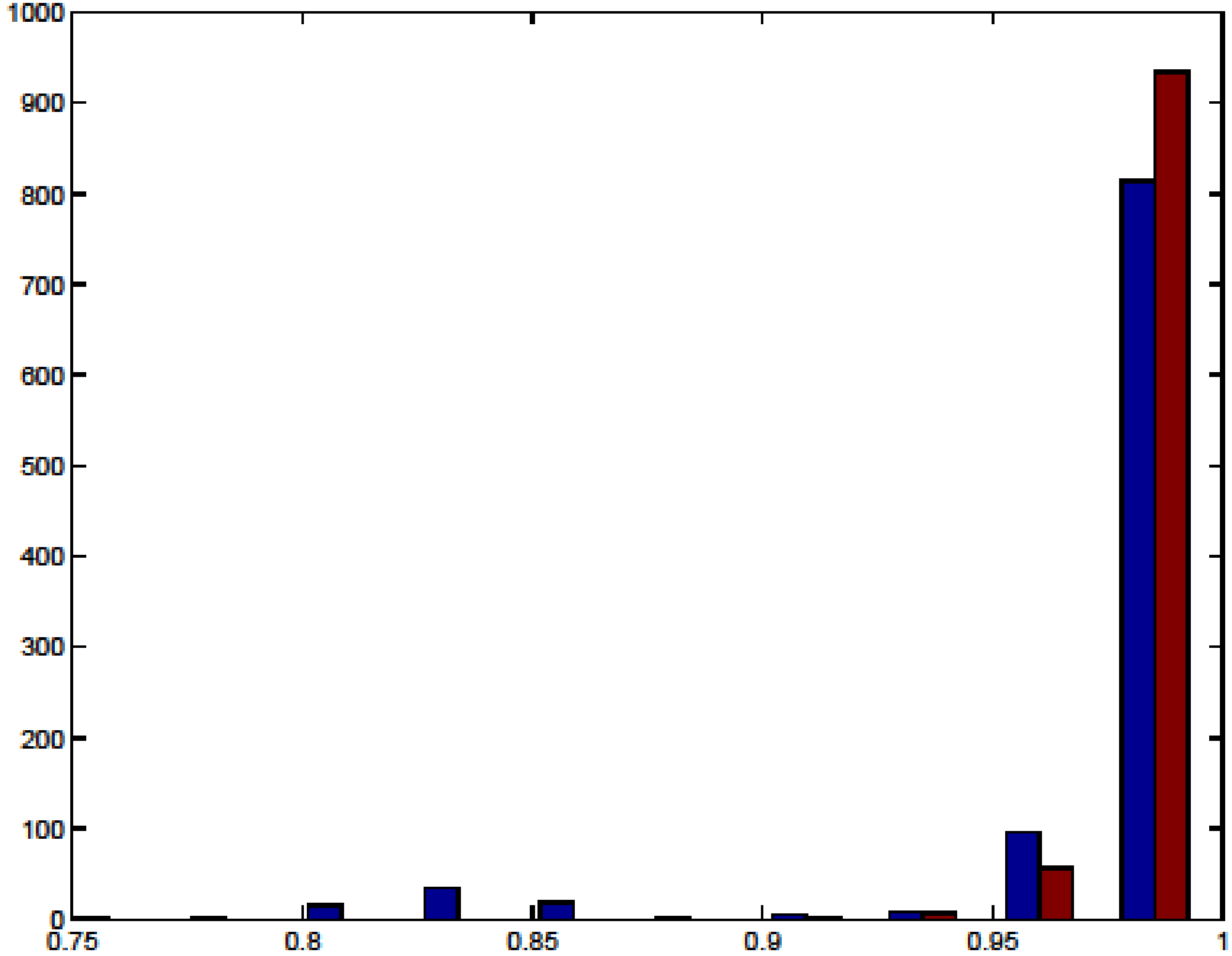
Y stellar formation rate

$\mathbf{X} \in \mathbb{R}^p$ $p = 46$-spectral characteristics of the galaxies

1000 random subsets of $\mathbf{X}$ of size $n_b = 30,000$

$d = 3$ has been selected via BIC computed for $d = 1, \ldots, 20$

STUDENT SIR



WEAK POINTS OF SIR AND EXTENSIONS

- The Linearity Design Condition (LDC)
- SIR is not robust to outliers
- Sparsity

TRIGGERING PAPERS

- Sparsity

L. LI AND C. J. NACHTSHEIM, Sparse sliced inverse regression, Technometrics, 48 (2006), pp. 503–510.

L. LI, Sparse sufficient dimension reduction, Biometrika, 94 (2007), pp. 603–613.

R. F. BARBER AND E. J. CANDÈS, Controlling the false discovery rate via knockoffs, The Annals of Statistics, 43 (2015), pp. 2055–2085.

$$\mathbf{X} = \{x_1, ..., x_n\} \in \mathbb{R}^{n \times p} \text{ predictors}$$

$$\hat{\Sigma} = \mathbf{X}^T\mathbf{X}, \mathbb{E}(\mathbf{X}) = 0 \text{ and } diag(\hat{\Sigma}) = 1$$

KNOCKOFF FILTER

Barber&Candès (2014)

$$\mathbf{X} = \{x_1, \ldots, x_n\} \in \mathbb{R}^{n \times p} \text{ predictors}$$

$$\hat{\Sigma} = \mathbf{X}^T\mathbf{X},\ \mathbb{E}(\mathbf{X}) = 0 \text{ and } diag(\hat{\Sigma}) = 1$$

$$\text{Knockoff filter } \tilde{\mathbf{X}} = \{\tilde{x}_1, \ldots, \tilde{x}_n\} \in \mathbb{R}^{n \times p}$$

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \hat{\Sigma}$$

$$\mathbf{X}^T\tilde{\mathbf{X}} = \hat{\Sigma} - diag\{s\},\ s \in \mathbb{R}^p \text{ nonnegative}$$

$$(X_j, X_k) \text{ and } (X_j, \tilde{X}_k) \text{ have the same correlation for } k \neq j$$

$$X_j^T\tilde{X}_j = 1 - s_j$$

KNOCKOFF FILTER

Barber&Candès (2014)

$$\mathbf{X} = \{x_1, \ldots, x_n\} \in \mathbb{R}^{n \times p} \text{ predictors}$$

$$\hat{\Sigma} = \mathbf{X}^T\mathbf{X},\ \mathbb{E}(\mathbf{X}) = 0 \text{ and } diag(\hat{\Sigma}) = 1$$

$$\text{Knockoff filter } \tilde{\mathbf{X}} = \{\tilde{x}_1, \ldots, \tilde{x}_n\} \in \mathbb{R}^{n \times p}$$

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \hat{\Sigma}$$

$$\mathbf{X}^T\tilde{\mathbf{X}} = \hat{\Sigma} - diag\{s\},\ s \in \mathbb{R}^p \text{ nonnegative}$$

$$(X_j, X_k) \text{ and } (X_j, \tilde{X}_k) \text{ have the same correlation for } k \neq j$$

$$X_j^T\tilde{X}_j = 1 - s_j$$

$$n \geq 2p$$

R. F. BARBER AND E. J. CANDÈS, Controlling the false discovery rate via knockoffs, The Annals of Statistics, 43 (2015), pp. 2055–2085.

Concatenation $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$

Is there a link between the behavior of SIR if applied on $\mathbf{X}$ and the results obtained on the concatenation $[\mathbf{X}, \tilde{\mathbf{X}}]$?

$$\text{Concatenation } [\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n\times 2p}$$

Is there a link between the behavior of SIR if applied on $\mathbf{X}$ and the results obtained on the concatenation $[\mathbf{X}, \tilde{\mathbf{X}}]$?

Theorem. Let us assume the model $Y = f(\beta^T X, \epsilon)$, where $\beta \in \mathbb{R}^p$ spans the central subspace and ϵ is a random error idependent of X.

Given the predictors $\mathbf{X} = \{x_1, ..., x_n\} \in \mathbb{R}^{n\times p}$ and the response variable $Y = \{y_1, ..., y_n\} \in \mathbb{R}^{n\times 1}$, let us consider a knockoff filter $\hat{\mathbf{X}} \in \mathbb{R}^{n\times p}$ and the concatenation $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n\times 2p}$.

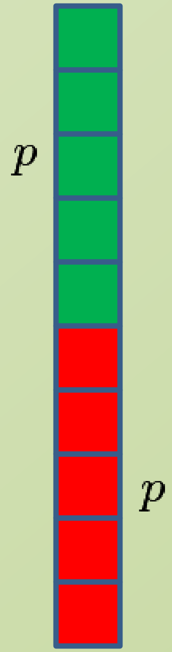
Then the SIR estimator $\hat{B} \in \mathbb{R}^{2p}$ for the concatenation $[\mathbf{X}, \tilde{\mathbf{X}}]$ has the form:

$$\hat{B} = [\beta, 0]$$

where 0 is a p-vector of all zeros.

IN PRACTICE

- $\mathbf{X} \in \mathbb{R}^{n \times p}$
- Build N different knockoff sets
- Gather the N solutions of SIR applied to the concatenation
- $\tilde{B}_1, \ldots, \tilde{B}_N$
- A Wilcoxon-Mann-Whitney test is used to distiguish active and inactive variables



SIMULATED DATA

$$Y = (x_1 + x_2 + x_3 - 10)^2 + \epsilon$$

where $\mathbf{X} \in \mathbb{R}^{10}$ follows a standard normal distribution and ϵ is a standard normal error independent of $\mathbf{X}$

n	TIR	FIR	#-slices
25	.81(.25)	.48(.20)	2
50	1(.0)	.16(.16)	5
75	1(.0)	.09(.12)	7
100	1(.0)	.08(.10)	10
150	1(.0)	.08(.11)	15
200	1(.0)	.06(.11)	20
250	1(.0)	.05(.08)	25
300	1(.0)	.04(.08)	30
400	1(.0)	.04(.06)	30

TABLE 4.1. Study on the sensitivity to the number of sample n, averages (and standard deviation in brackets) are obtained over 100 iterations. True Inclusion Rate (TIR) and False Inclusion Rate (FIR) are shown. The number of slices has been selected such that at least 10 samples are contained in each slice.

DIMENSION OF THE CENTRAL SUBSPACE

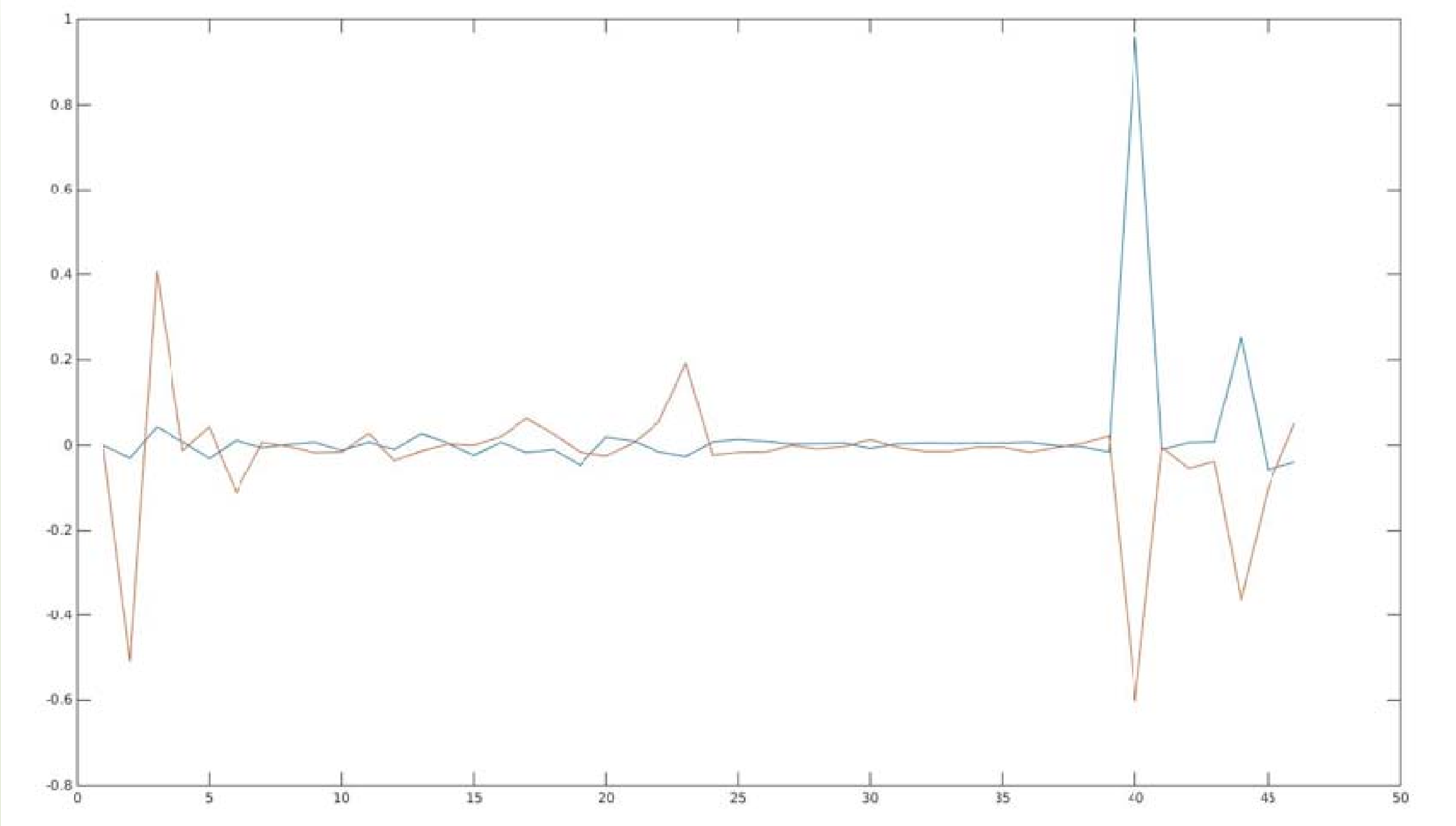
Galaxy dataset

$n = 362887$

Y stellar formation rate

$\mathbf{X} \in \mathbb{R}^p$ $p = 46$-spectral characteristics of the galaxies

DIMENSION OF THE CENTRAL SUBSPACE



The analysis via Knockoff SIR confirms the results of Collaborative SIR

A difference is that the 6th variable is not selected

The dimension of the central subspace has been identified as d=3 in accordance to Student SIR

PROBLEMS

- The procedure can be applied only when $n \geq 2p$
- Troubles in identifying variables with a small relative importance (e.g. $f(0.1x_1 + x_2)$)
- Wilcoxon-Mann-Whitney test is not always reliable

- M. Chini, A. Chiancone and S. Stramondo. Scale Object Selection (SOS) through a hierarchical segmentation by a multi-spectral per-pixel classification. Pattern Recognition Letters 49 (2014): 214-223.

- A. Chiancone, S. Girard, J. Chanussot. Collaborative Sliced Inverse Regression. Communication in Statistics - Theory and Methods, Taylor & Francis, 2016

- A. Chiancone, F. Forbes, S. Girard. Student Sliced Inverse Regression. Computational Statistics and Data Analysis, Elsevier, 2016

Personal projects

- A. Chiancone. What's wrong with classes? The theory of Knowledge

- Student SIR: when $p > n$ a regularization to overcome the problem of inversing the matrix Σ is necessary. The framework proposed in [BERNARD-MICHEL et al (2009)] is promising

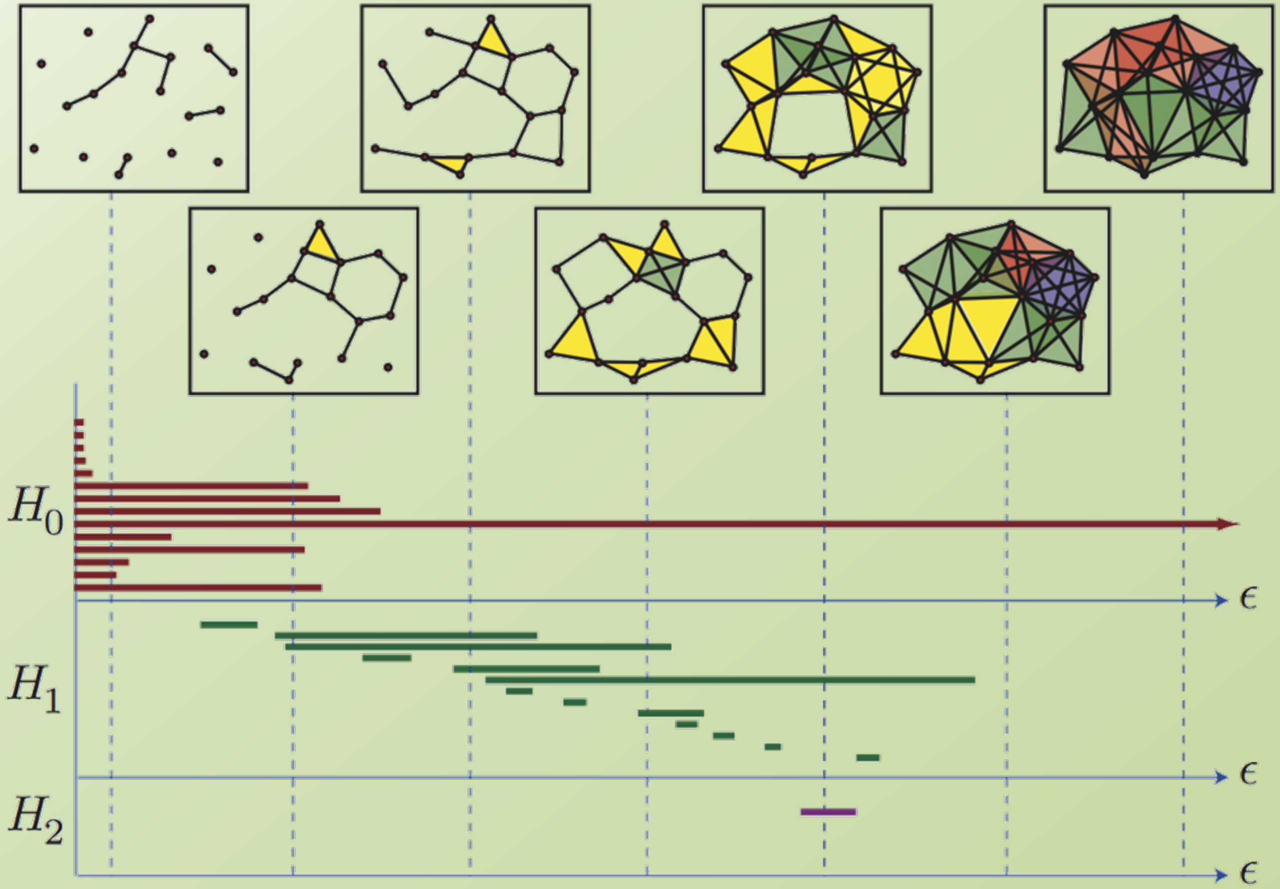
- Student SIR: How to set the dimension d of the central subspace. We have proposed the use of BIC which is not stable when the sample size is too small

- Knockoff SIR: Extend the analysis to the case $n < 2p$.

- Extend our results to the multivariate case when Y is a vector

- Open question: Is there a statistical tool that can exclude the existence of a link between a predictor space X and a response variable Y?

Computational Topology and dimension reduction



Are there any topological problems to dimension reduction?
Use persistence diagrams as an a priori for dimension reduction