Sliced Inverse Regression with Interaction (Siri) Detection

for non-Gaussian BN learning

Jun S. Liu
Department of Statistics
Harvard University

Joint work with Bo Jiang

General: Regression and Classification

	Covariates	Responses
Ind 1	$x_{11}, x_{12}, \ldots, x_{1p}$	Y_1
Ind 2	$x_{21}, x_{22}, \ldots, x_{2p}$	Y_2
$\vdots$		$\vdots$
Ind N	$x_{N1}, x_{N2}, \ldots, x_{NP}$	Y_N

THE LAB WHERE
THEY STUDY DRUG INTERACTION

Variable Selection with Interaction

Let $Y \in R$ be a univerate response variable and $\boldsymbol{X} \in R^p$ be a vector of p continuous predictor variables

$Y = X_1 \times X_2 + \epsilon,\ \epsilon \sim N(0, \sigma^2),\ \boldsymbol{X} \sim \text{MVN}(0, I_p)$

Suppose $p = 1000$. How to find X_1 and X_2?

One step forward selection :~ 500,000 interaction terms

Variable Selection with Interaction

Let $Y \in R$ be a univerate response variable and $\boldsymbol{X} \in R^p$ be a vector of p continuous predictor variables

$Y = X_1 \times X_2 + \epsilon,\ \epsilon \sim N(0, \sigma^2),\ \boldsymbol{X} \sim \text{MVN}(0, I_p)$

Suppose $p = 1000$. How to find X_1 and X_2 ?

One step forward selection :~ 500,000 interaction terms

Is there any *marginal* relationship between Y and X_1 ?

[Y|X] ? [X|Y] ? Who is behind the *bar*?

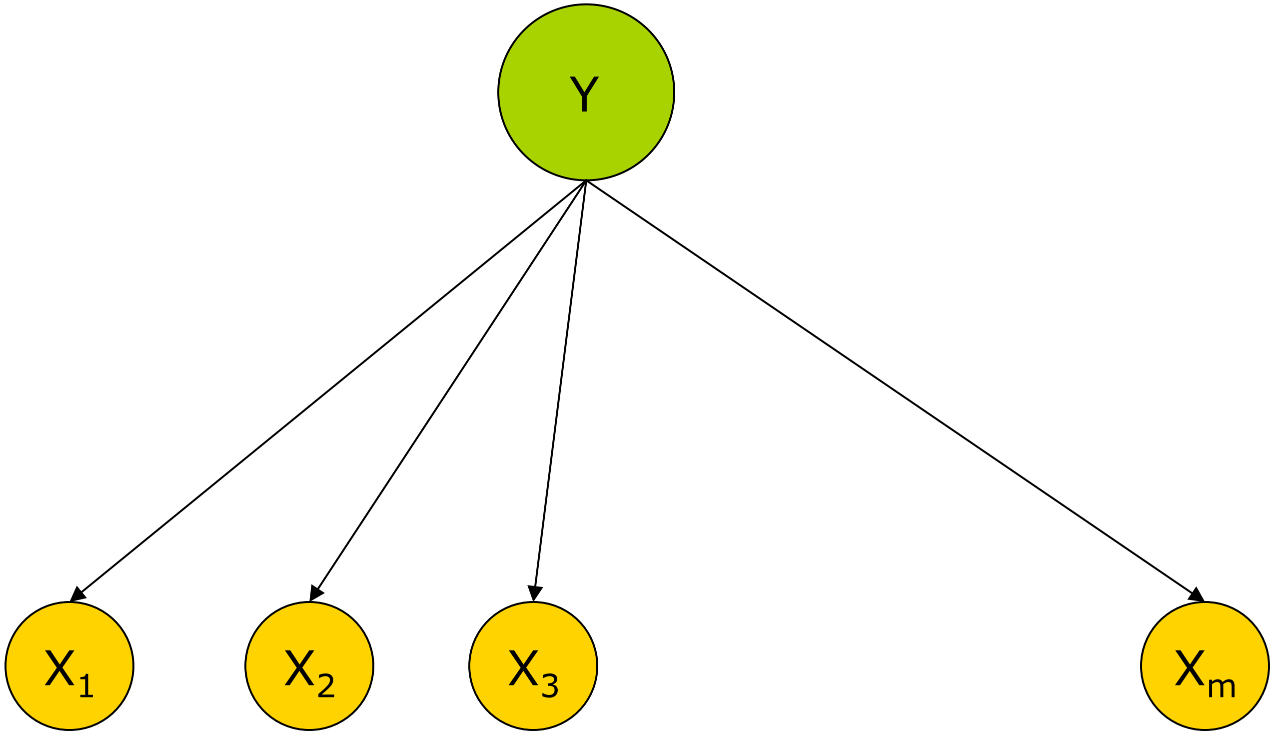
General: Regression and Classification

	Covariates	Responses
Ind 1	$x_{11}, x_{12}, \ldots, x_{1p}$	Y_1
Ind 2	$x_{21}, x_{22}, \ldots, x_{2p}$	Y_2
$\vdots$		$\vdots$
Ind N	$x_{N1}, x_{N2}, \ldots, x_{NP}$	Y_N

$$P(Y \mid \mathbf{X}) = P(\mathbf{X} \mid Y)P(Y) / P(\mathbf{X})$$

How to model this?

Naïve Bayes model

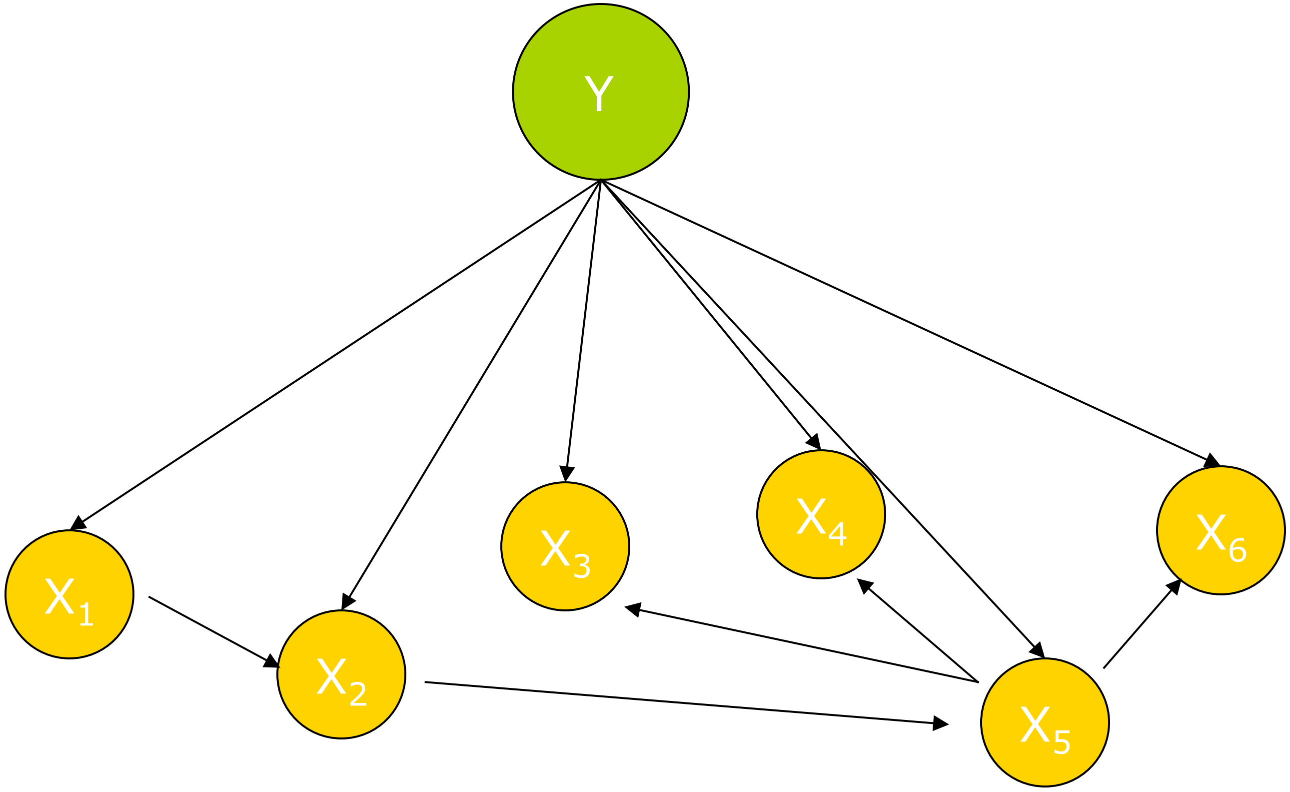


$$P(Y|X_1,\cdots,X_m)=\frac{P(Y)\prod_{j=1}^{m}P(X_j|Y)}{P(X_1,\cdots,X_m)}.$$

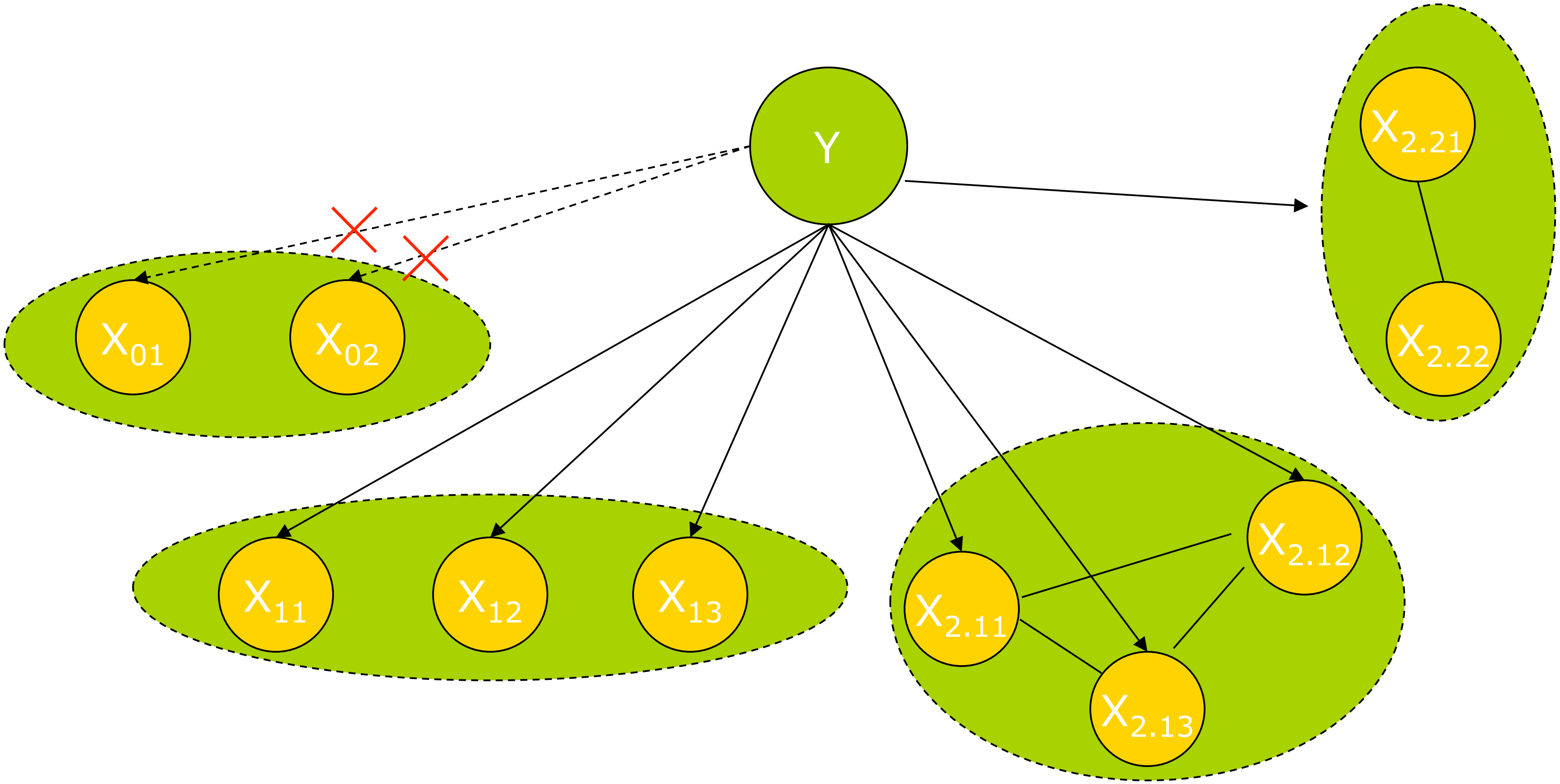
(Augmented) Naïve Bayes Model

- BEAM: Bayesian Epistasis Association Mapping (Zhang and Liu 2007): discrete univariate response and discrete predictors
- (Augmented) Naïve Bayes Classifier with Variable Selection and Interaction Detection (Yuan Yuan et al.): discrete univariate response and continuous (but discretized) predictors
- Bayesian Partition Model for eQTL study (Zhang et al. 2010): continuous multivariate responses and discrete predictors
- Sliced Inverse Regression with Interaction Detection (SIRI): continuous univariate response and continuous predictors

Tree-Augmented Naïve Bayes



Augmented Naïve Bayes



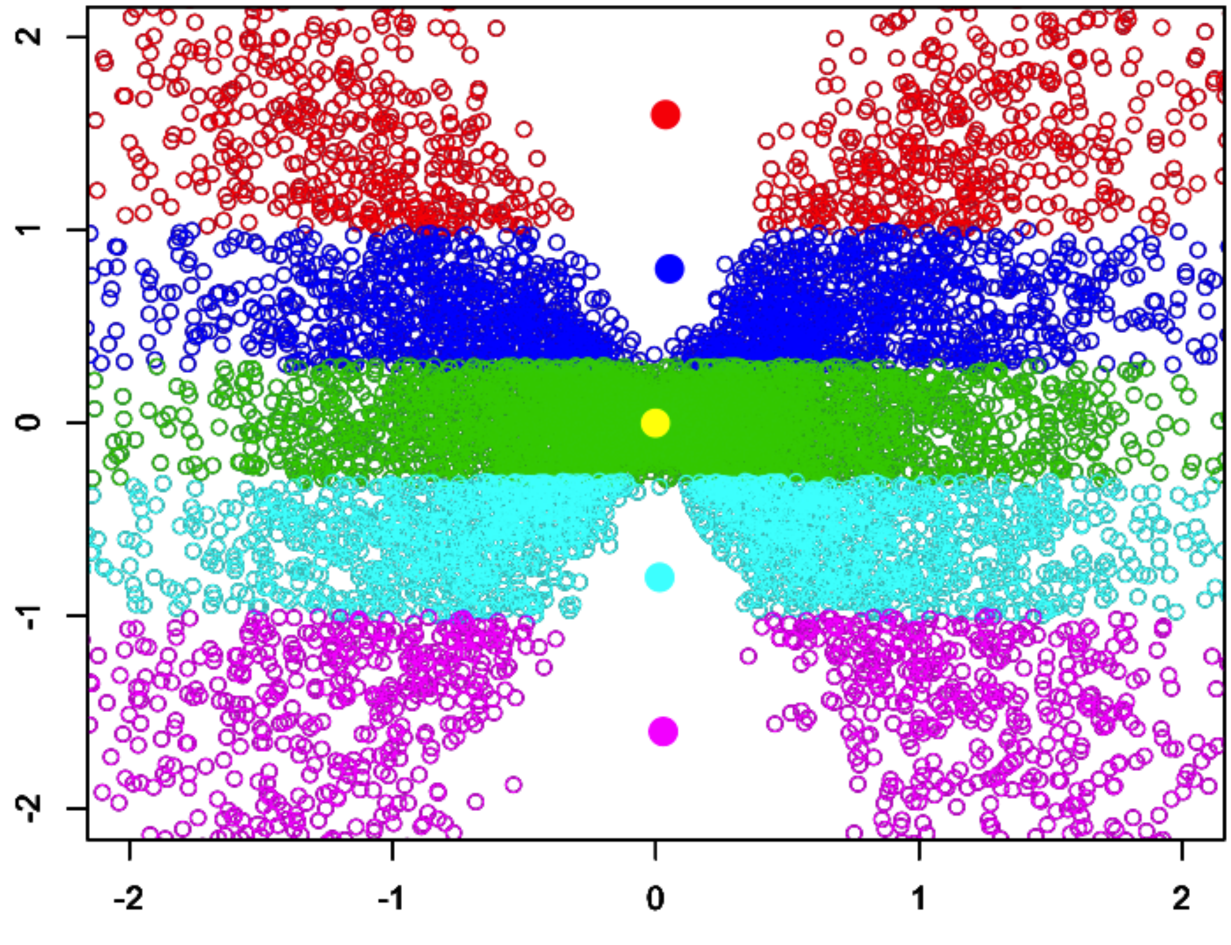
How about continuous covariates?

- We may discretize Y, and discretize each X
- Or discretize Y, assuming joint Gaussian distributions on X?
- Sound familiar?

An observation:

$$Y = X_1 \times X_2 + \epsilon,\ \epsilon \sim N(0,\sigma^2),\ \boldsymbol{X} \sim \mathrm{MVN}(0, I_p)$$

y



Sliced Inverse Regression (SIR, Li 1991)

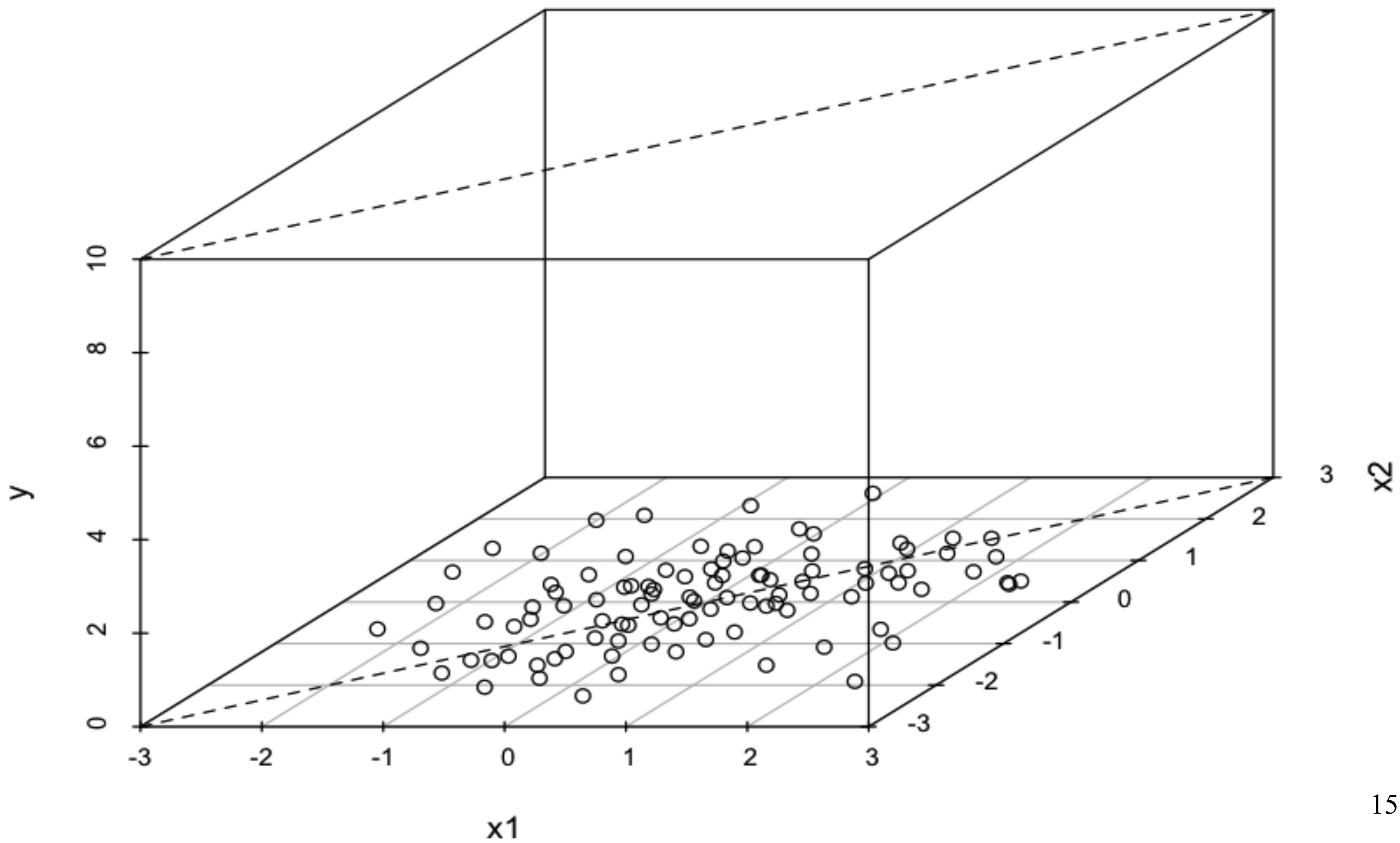
SIR is a tool for dimension reduction in multivariate statistics

Let $Y \in R$ be a univerate response variable and $\boldsymbol{X} \in R^p$ be a vector of p continuous predictor variables

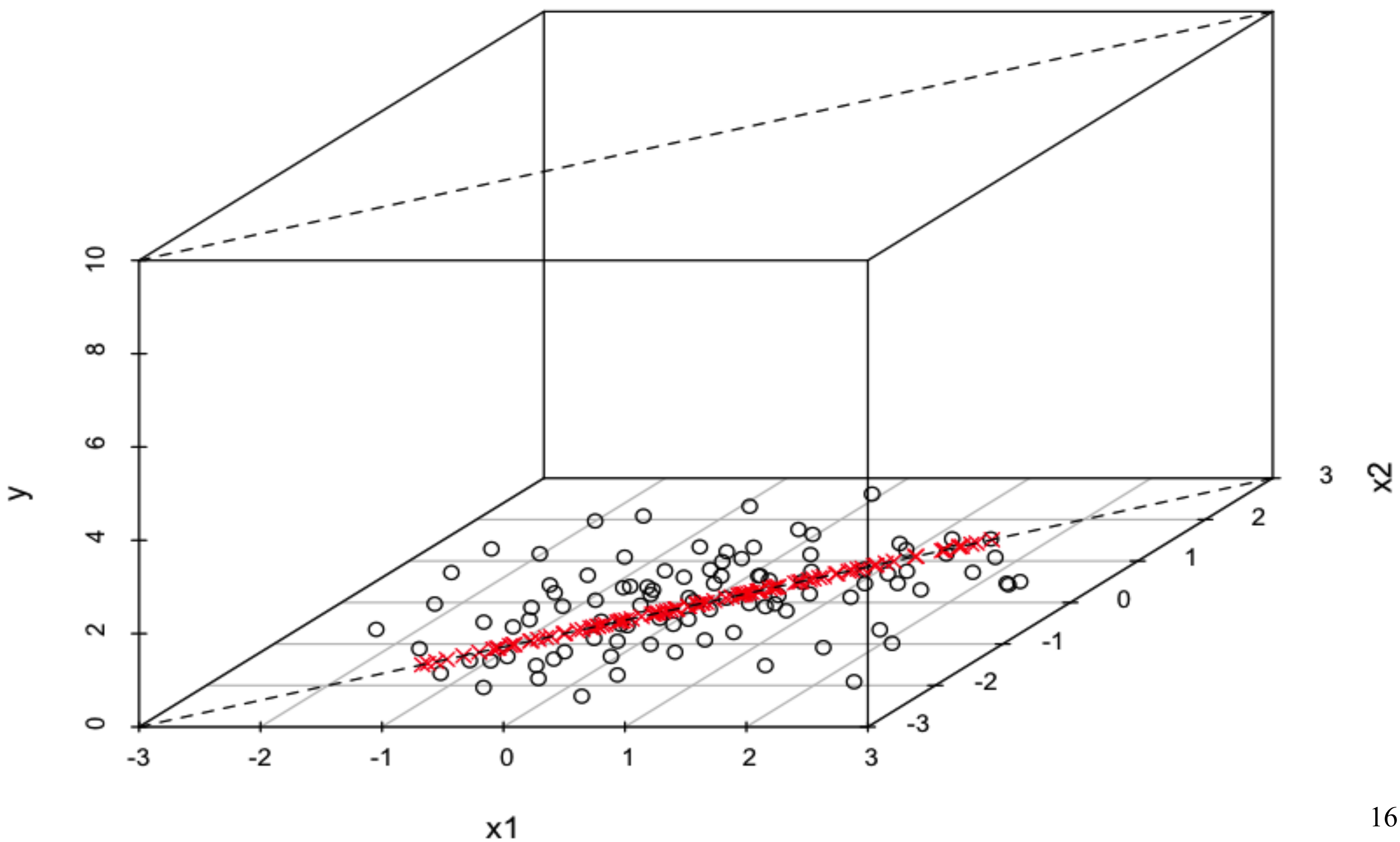
$$Y = f(\beta_1^T \boldsymbol{X}, \ldots, \beta_K^T \boldsymbol{X}, \epsilon)$$

f is an unknown function and ϵ is the error with finite variance

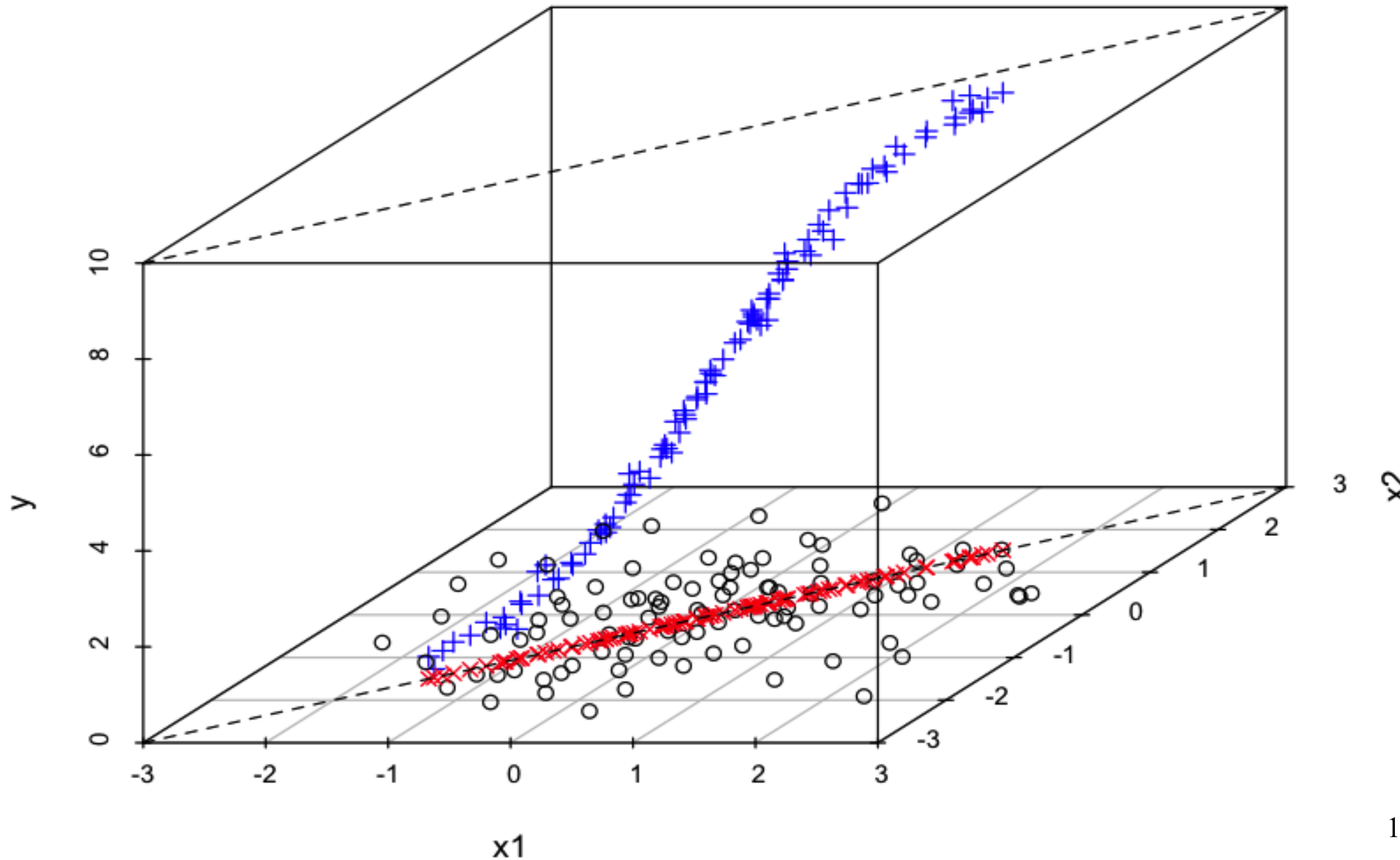
How to identify unknown projection vectors $\beta_1, \ldots, \beta_K$?

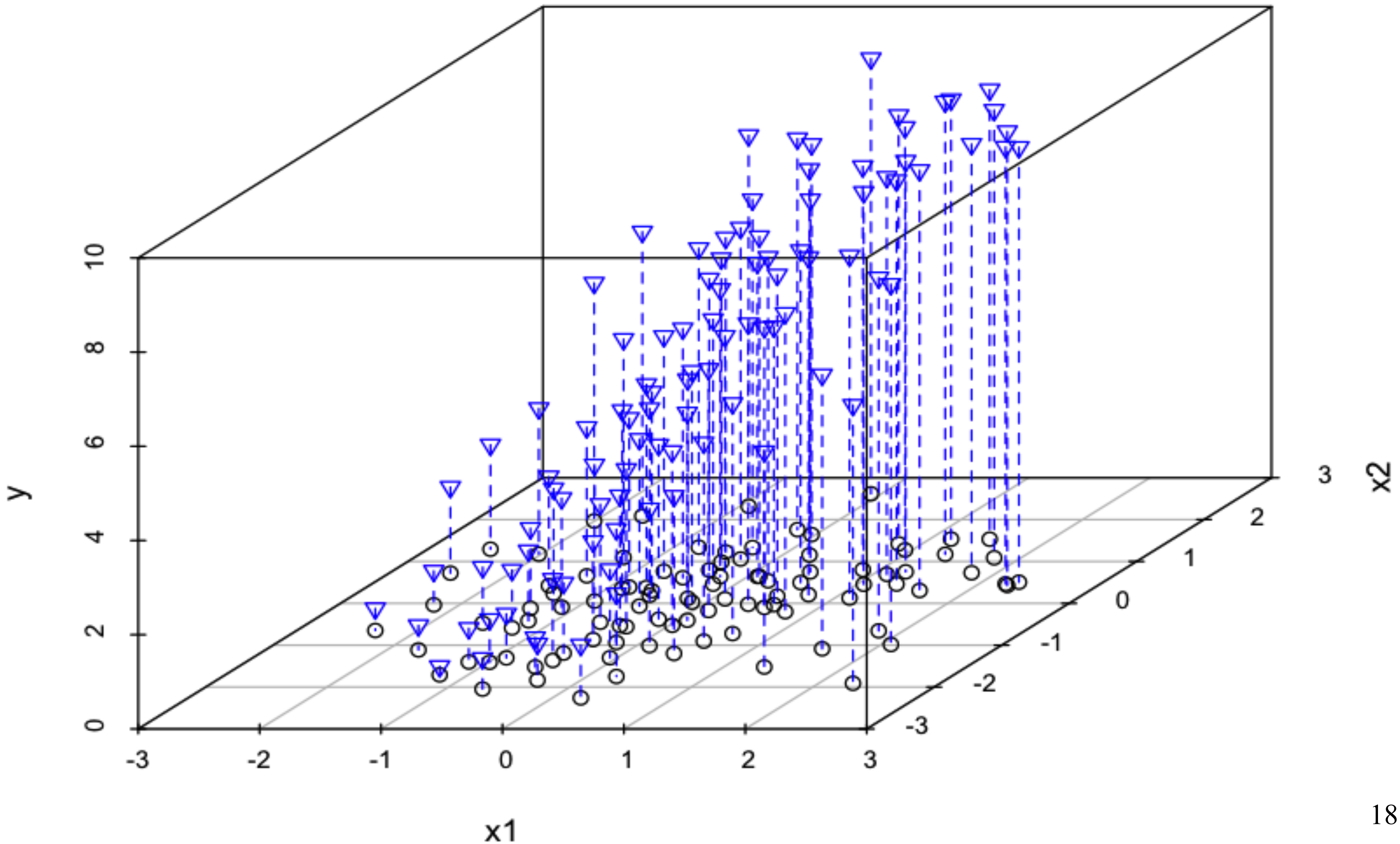


y
0
2
4
6
8
10
x1
-3
-2
-1
0
1
2
3
x2
-3
-2
-1
0
1
2
3

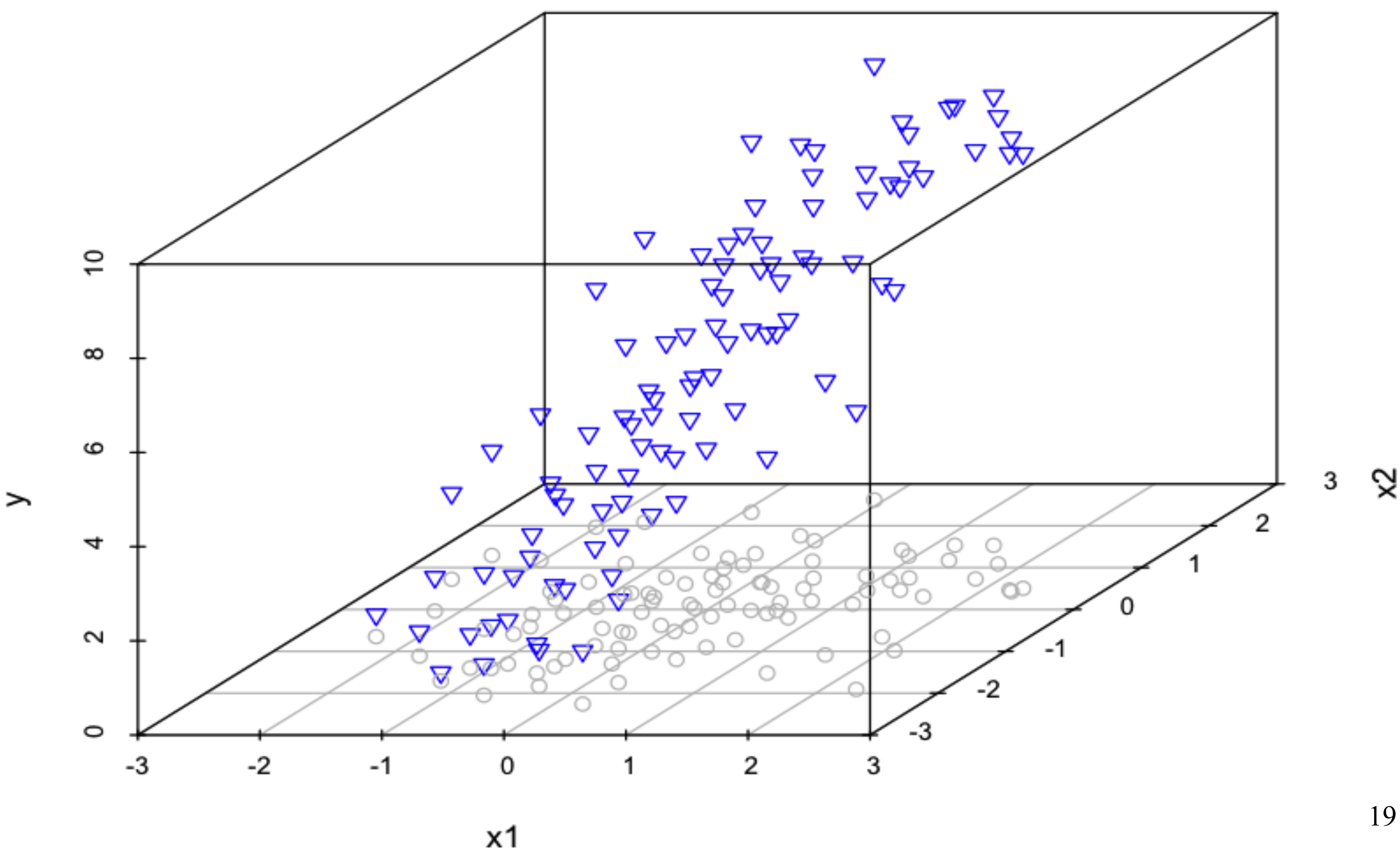


y
0
2
4
6
8
10
-3
-2
-1
0
1
2
3
x1
-3
-2
-1
0
1
2
3
x2

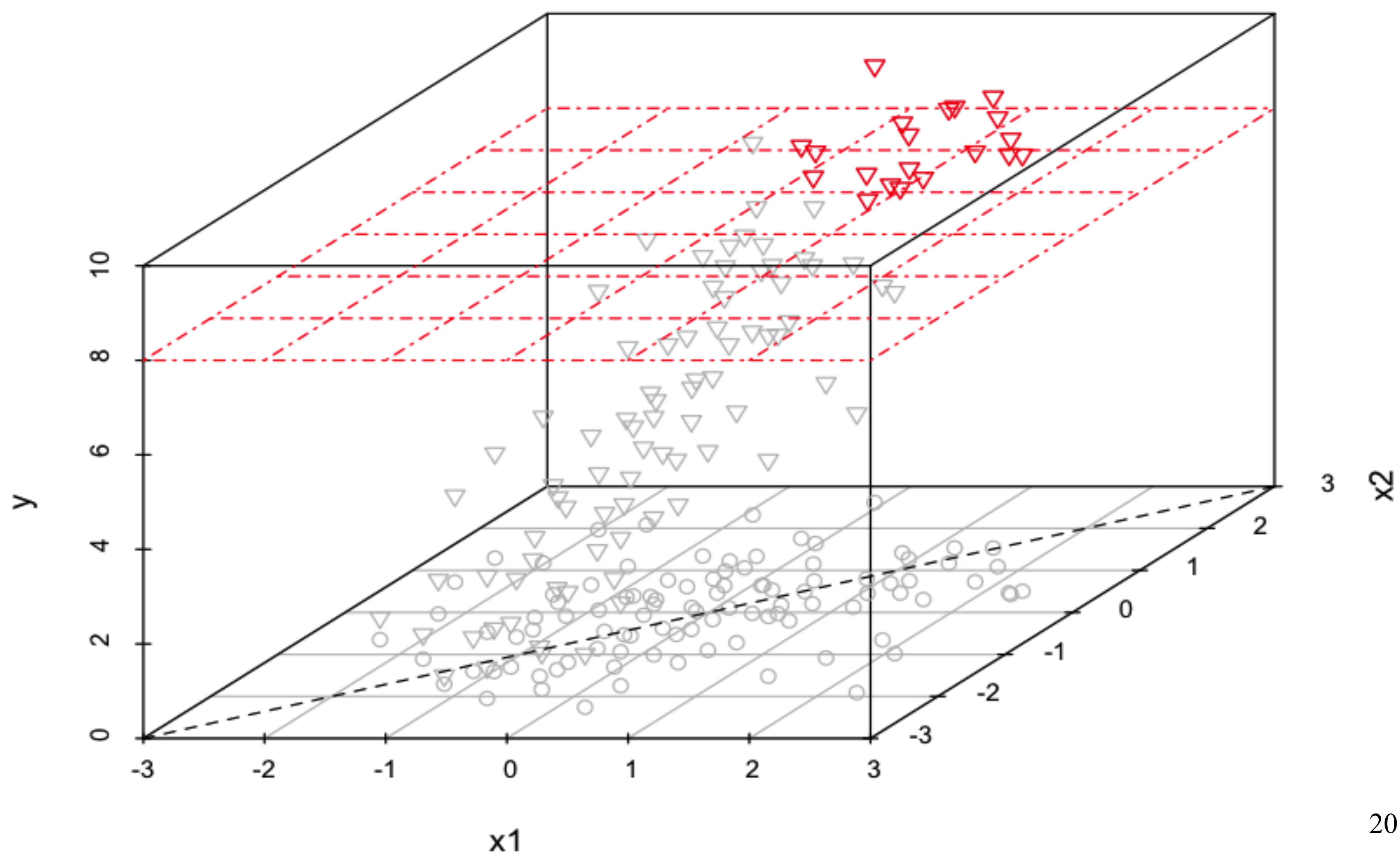

y
0
2
4
6
8
10
-3
-2
-1
0
1
2
3
x1
-3
-2
-1
0
1
2
3
x2



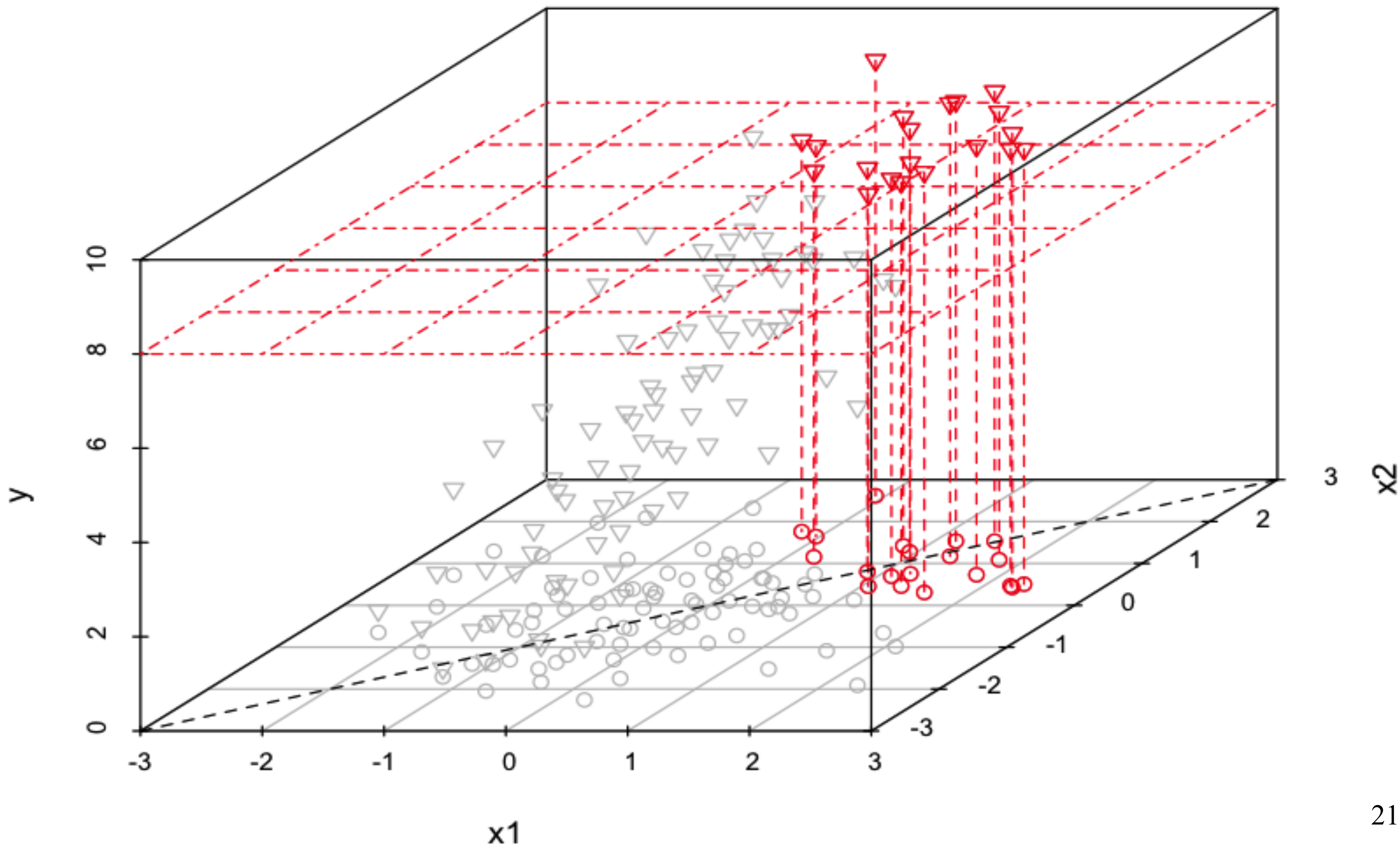
10
8
6
4
2
0
y
-3
-2
-1
0
1
2
3
x1
3
2
1
0
-1
-2
-3
x2



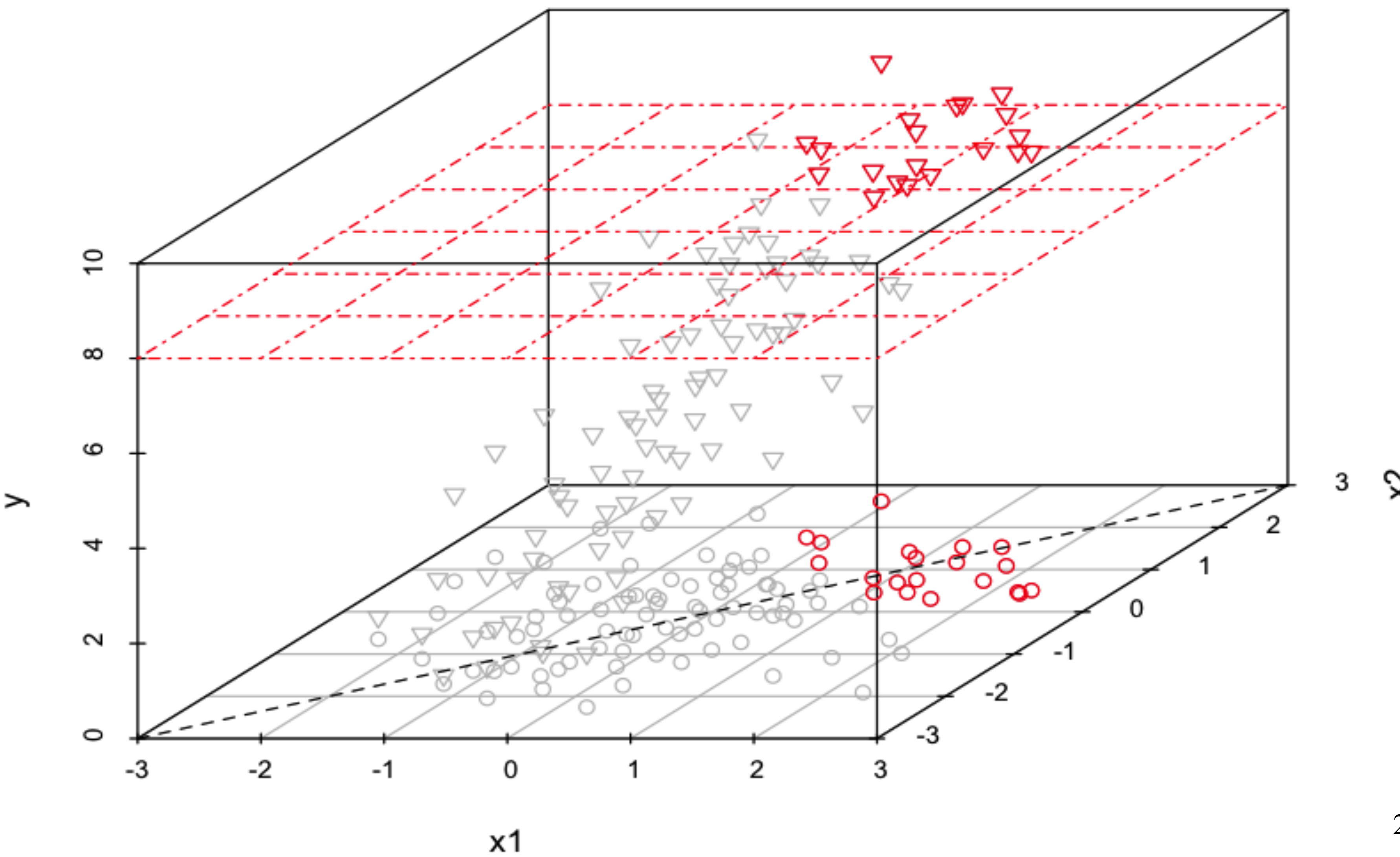
y
0
2
4
6
8
10
x1
-3
-2
-1
0
1
2
3
x2
-3
-2
-1
0
1
2
3

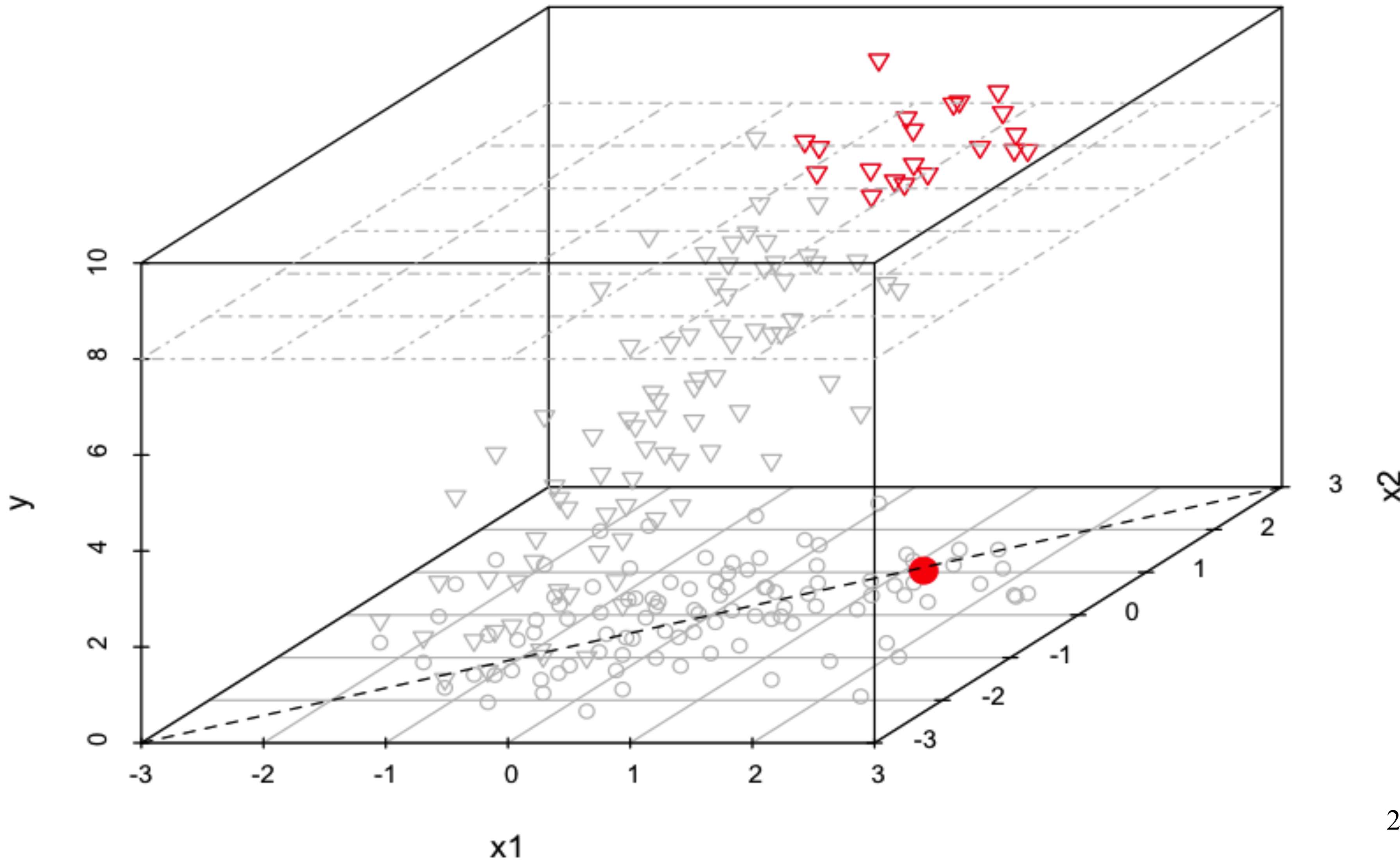


10
8
6
4
2
0
y
-3
-2
-1
0
1
2
3
x1
3
2
1
0
-1
-2
-3
x2

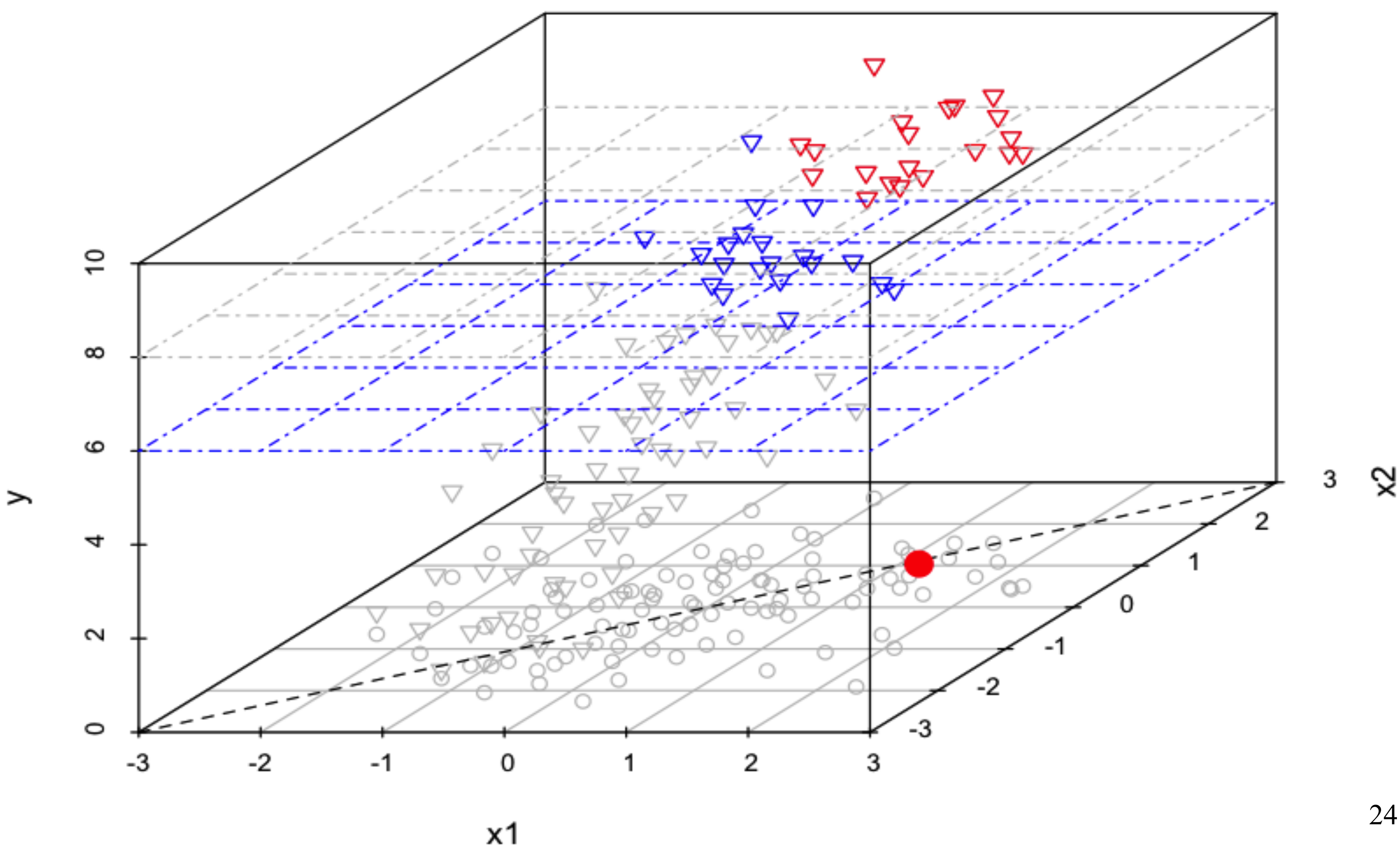


y
0
2
4
6
8
10
x1
-3
-2
-1
0
1
2
3
x2
-3
-2
-1
0
1
2
3

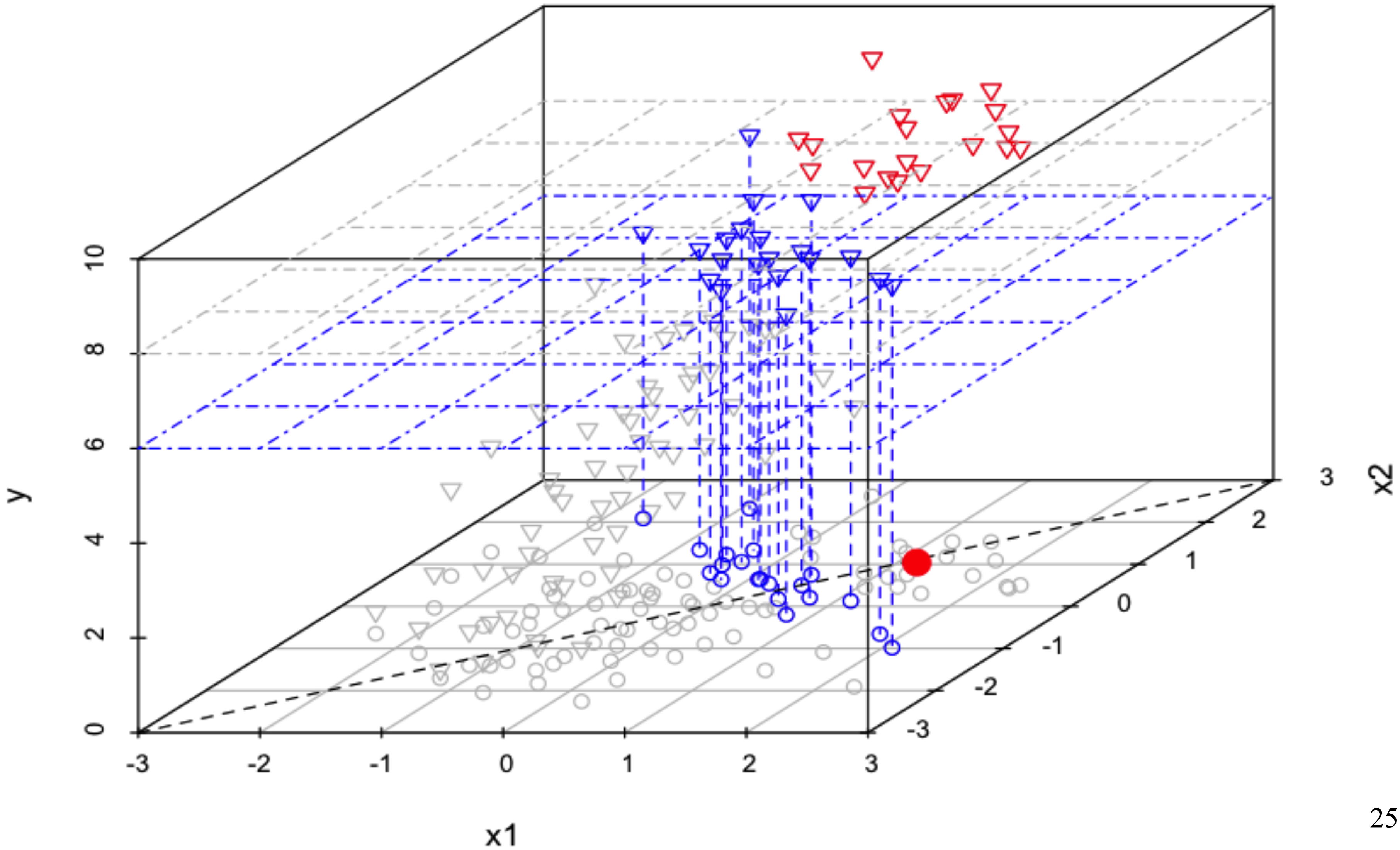

y
0
2
4
6
8
10
x1
-3
-2
-1
0
1
2
3
x2
-3
-2
-1
0
1
2
3



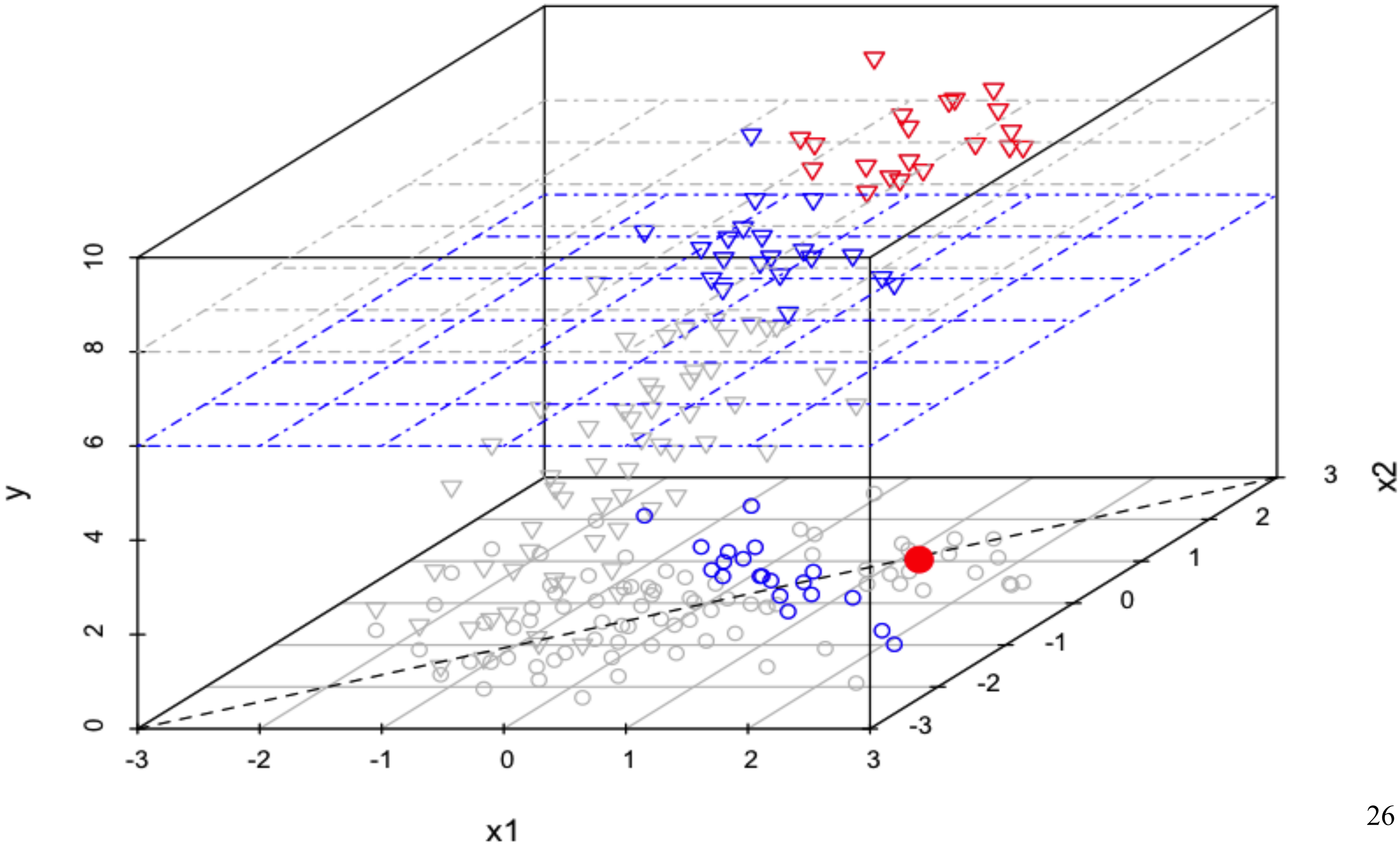
y
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
x1
x2
3
2
1
0
-1
-2
-3



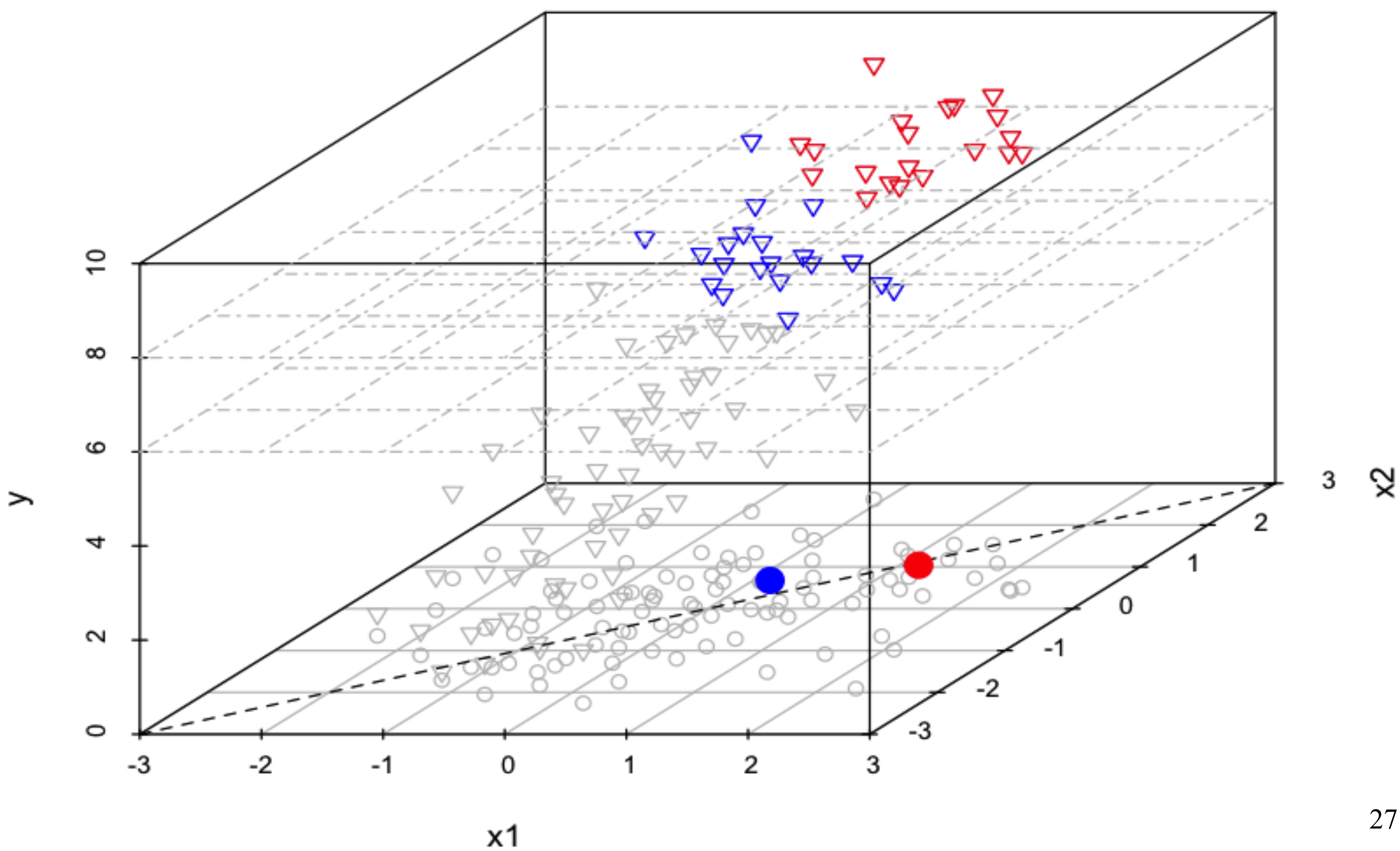
y
x1
x2
0
2
4
6
8
10
-3
-2
-1
0
1
2
3
-3
-2
-1
0
1
2
3



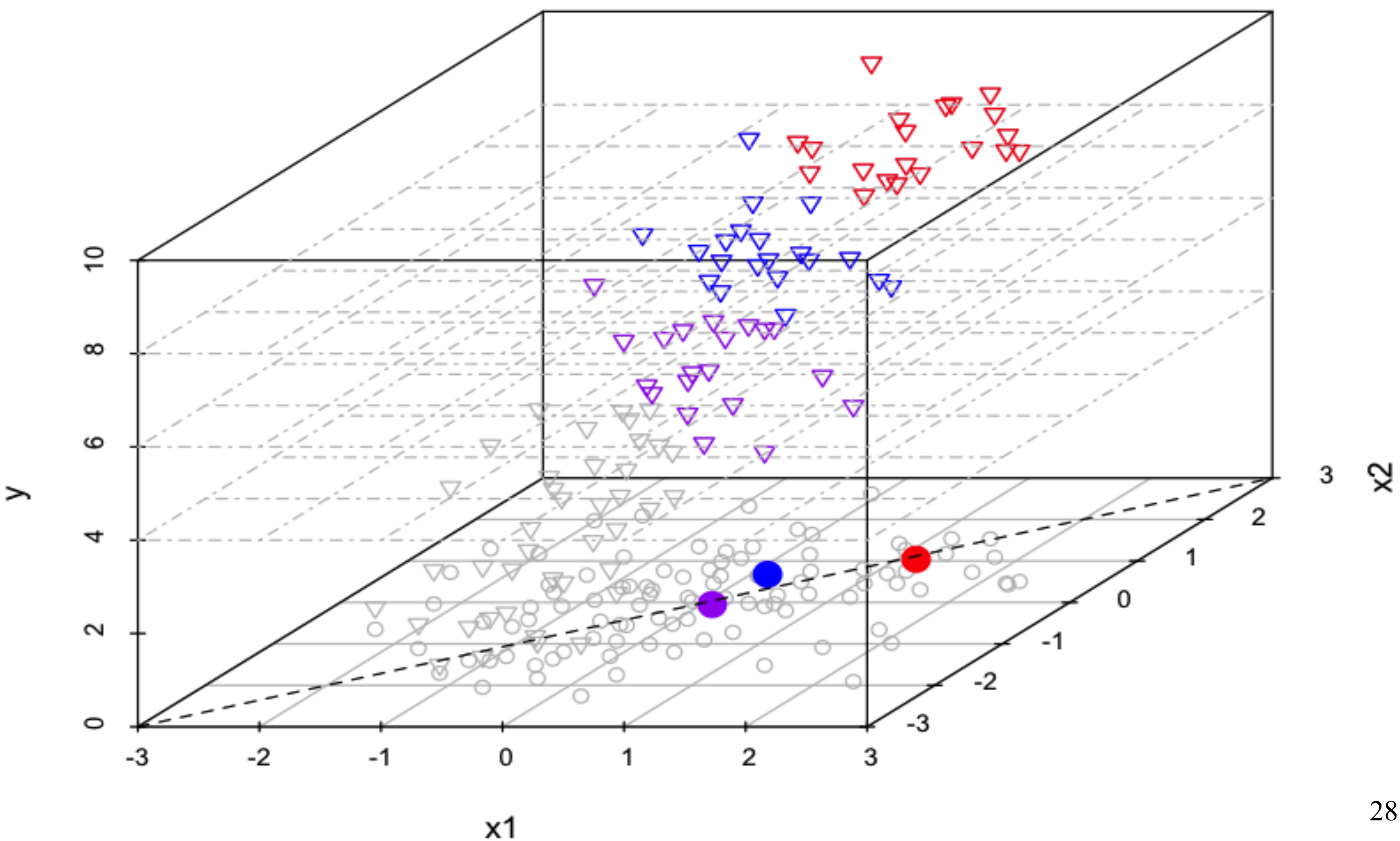
y
0
2
4
6
8
10
x1
-3
-2
-1
0
1
2
3
x2
-3
-2
-1
0
1
2
3



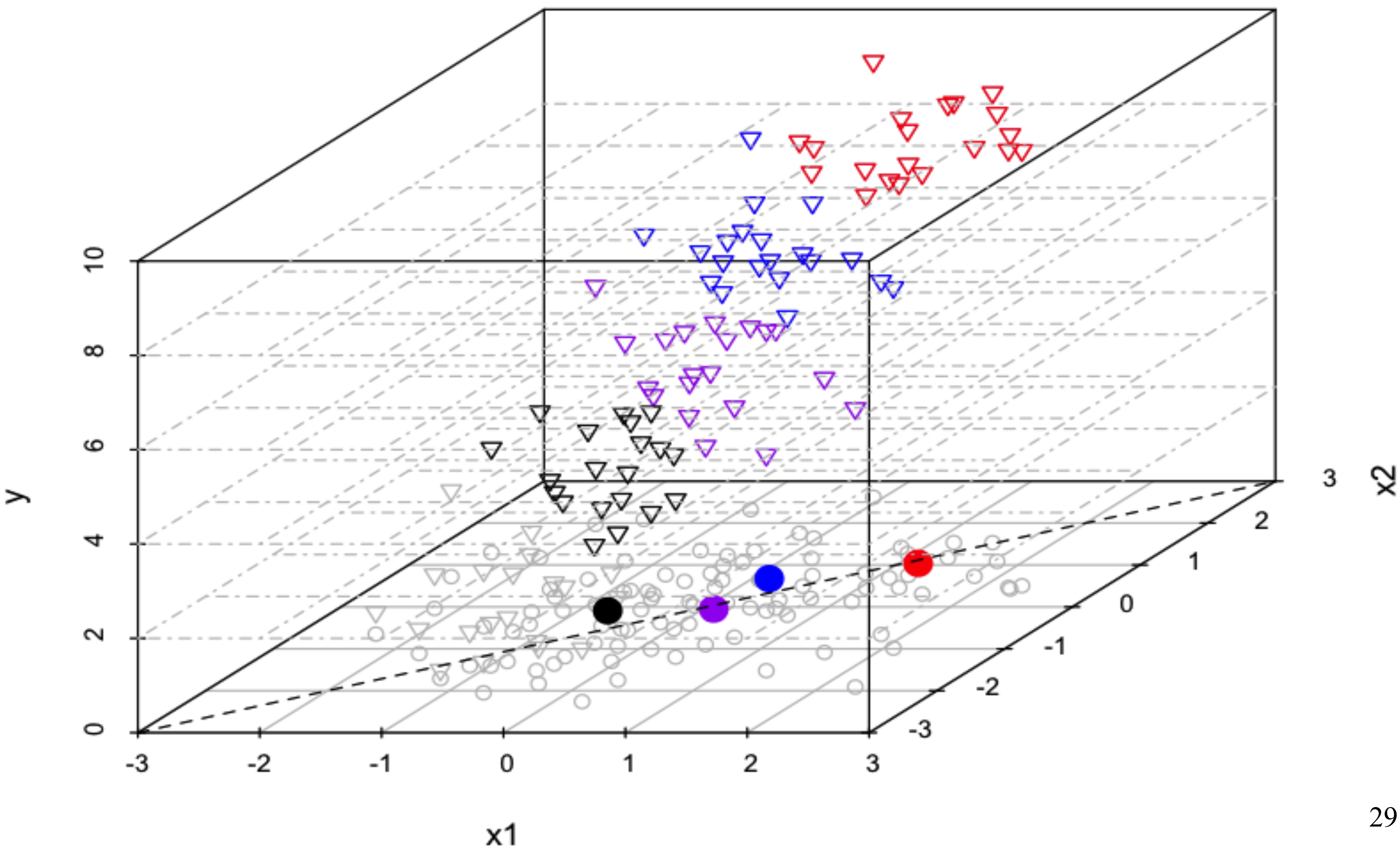
y
10
8
6
4
2
0
x1
-3
-2
-1
0
1
2
3
x2
3
2
1
0
-1
-2
-3



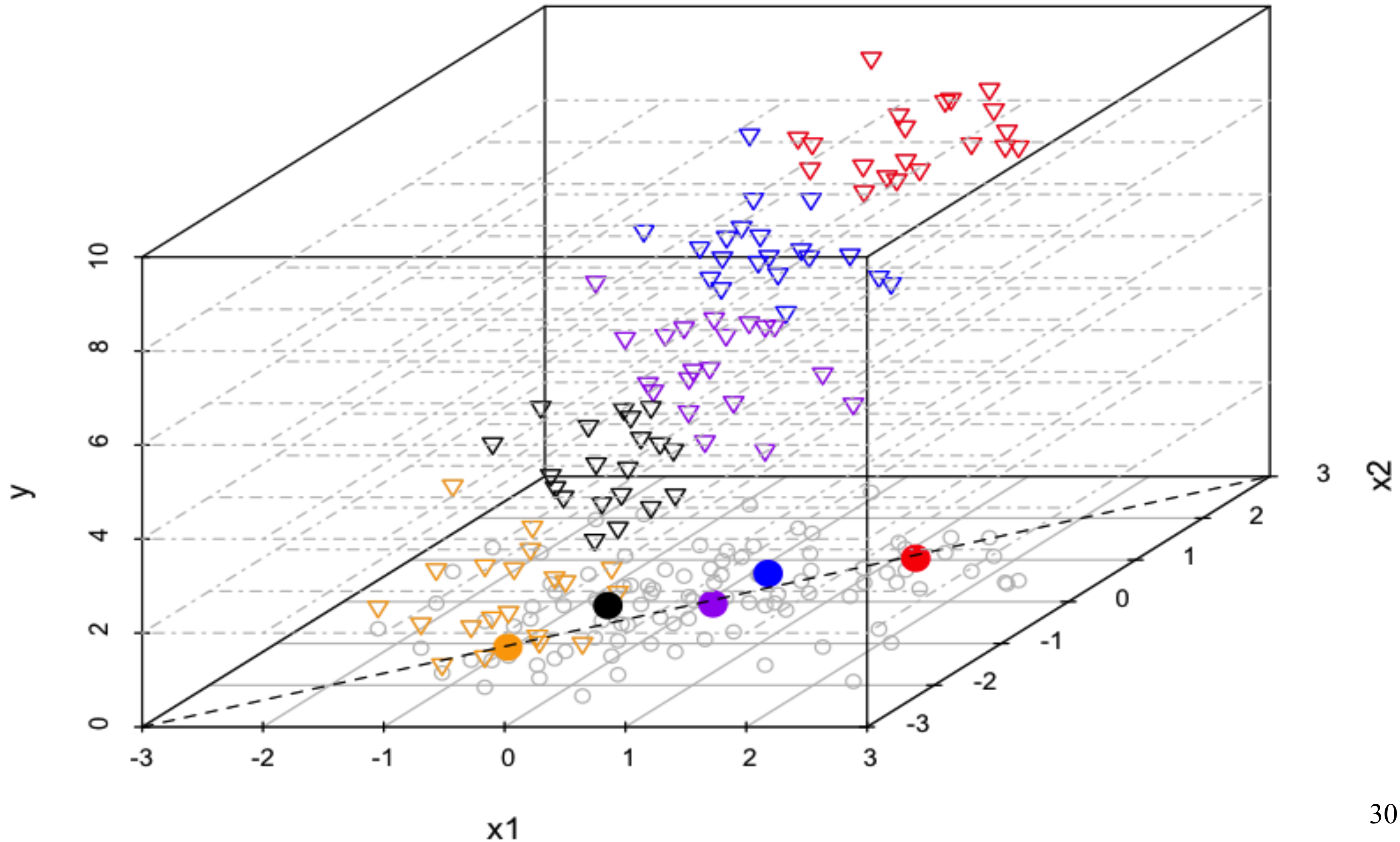
10
8
6
4
2
0
y
-3
-2
-1
0
1
2
3
x1
3
2
1
0
-1
-2
-3
x2



y
0
2
4
6
8
10
x1
-3
-2
-1
0
1
2
3
x2
-3
-2
-1
0
1
2
3



y
0
2
4
6
8
10
x1
-3
-2
-1
0
1
2
3
x2
-3
-2
-1
0
1
2
3



10
8
6
4
2
0
y
-3
-2
-1
0
1
2
3
x1
3
2
1
0
-1
-2
-3
x2

SIR Algorithm

Let Σ_{xx} be the covariance matrix of $\boldsymbol{X}$. Standarize $\boldsymbol{X}$ to :

$$\boldsymbol{Z} = \Sigma_{xx}^{-1/2}\{\boldsymbol{X} - E\boldsymbol{X}\}$$

Divde the range of y_i into S nonoverlapping slices $H_{s \in \{1,\ldots,S\}}$

n_s is the number of observations within each slice

Compute the mean of z_i over all slices $\bar{z}_s = n_s^{-1}\sum_{i \in H_s} z_i$, and calculate the estimate for $Cov\{E(\boldsymbol{X}|Y)\}$:

$$\hat{M} = n^{-1}\sum_{s=1}^{S} n_s \bar{z}_s \bar{z}_s^T$$

Identify largest K eigenvalues of $\hat{M}$, $\hat{\lambda}_k$ and corresponding eigenvectors $\hat{\eta}_k$. Then,

$$\hat{\beta}_k = \hat{\Sigma}_{xx}^{-1/2}\hat{\eta}_k \quad (k = 1, \ldots, K)$$

SIR with Variable Selection

Only a subset of predictors are relevant: $\beta_1, \ldots, \beta_K$ are sparse

Backward subset selection (Cook 2004, Li et al. 2005)

Shrinkage estimates of $\beta_1, \ldots, \beta_K$ using L_1- or L_2-penalty :

Regularized SIR (RSIR, Zhong et al. 2005)

Sparse SIR (SSIR, Li 2007)

Correlation Pursuit (Zhong et al. 2012) : A forward selection and backward elimination procedure motivated by F-test in stepwise regression

$$F_{1,n-d-1} = (n-d-1)\frac{(\hat{R}^2_{d+1} - \hat{R}^2_d)}{1 - \hat{R}^2_{d+1}}$$

Correlation Pursuit (COP)

Let A be the current set of selected predictors and $\hat{\lambda}_k^A$ the kth largest eigenvalue estimated by SIR based on predictors in A

For jth predictor ($j \notin A$), X_j, define statistic

$$COP_k^{A+j} = n\frac{(\hat{\lambda}_k^{A+j} - \hat{\lambda}_k^A)}{1 - \hat{\lambda}_k^{A+j}}$$

Correlation Pursuit (COP)

Let A be the current set of selected predictors and $\hat{\lambda}_k^A$ the k th largest eigenvalue estimated by SIR based on predictors in A

For j th predictor $(j \notin A)$, X_j, define statistic

$$COP_k^{A+j} = n \frac{(\hat{\lambda}_k^{A+j} - \hat{\lambda}_k^A)}{1 - \hat{\lambda}_k^{A+j}}$$

If $j \notin A$, $COP_k^{A+j} (k = 1, \ldots, K)$ are *asymptotically* i.i.d. $\chi^2(1)$, and $COP_{1:K}^{A+j} = \sum_{k=1}^{K} COP_k^{A+j}$ is *asymptotically* $\chi^2(K)$

Correlation Pursuit (COP)

Let A be the current set of selected predictors and $\hat{\lambda}_k^A$ the k th largest eigenvalue estimated by SIR based on predictors in A

For j th predictor $(j \notin A)$, X_j, define statistic

$$COP_k^{A+j} = n \frac{(\hat{\lambda}_k^{A+j} - \hat{\lambda}_k^A)}{1 - \hat{\lambda}_k^{A+j}}$$

If $j \notin A$, $COP_k^{A+j} (k = 1, \ldots, K)$ are *asymptotically* i.i.d. $\chi^2(1)$,

and $COP_{1:K}^{A+j} = \sum_{k=1}^{K} COP_k^{A+j}$ is *asymptotically* $\chi^2(K)$

The stepwise procedure is *consistent* if $p = O(n^r), r < 1/2$

Correlation Pursuit (COP)

Let A be the current set of selected predictors and $\hat{\lambda}_k^A$ the kth largest eigenvalue estimated by SIR based on predictors in A

For jth predictor $(j \notin A)$, X_j, define statistic

$$COP_k^{A+j} = n \frac{(\hat{\lambda}_k^{A+j} - \hat{\lambda}_k^A)}{1 - \hat{\lambda}_k^{A+j}}$$

If $j \notin A$, $COP_k^{A+j} (k = 1, \ldots, K)$ are *asymptotically* i.i.d. $\chi^2(1)$,

and $COP_{1:K}^{A+j} = \sum_{k=1}^{K} COP_k^{A+j}$ is *asymptotically* $\chi^2(K)$

The stepwise procedure is *consistent* if $p = O(n^r), r < 1/2$

Dimension K and threshold in forward selection (backward elimnation) are chosen by cross-validation

SIR via MLE

Let A be the set of relevant predictors and $C = A^C$, $d = |A|$

$$\boldsymbol{X}_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

SIR via MLE

Let A be the set of relevant predictors and $C = \neg A$, $d = |A|$

$$\boldsymbol{X}_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

$$\boldsymbol{X}_C | \boldsymbol{X}_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

SIR via MLE

Let A be the set of relevant predictors and $C = A^C$, $d = |A|$

$$\boldsymbol{X}_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

$$\boldsymbol{X}_C | \boldsymbol{X}_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

$\mu_s = \alpha + \Gamma \gamma_s$, where $\gamma_s \in R^K$ and Γ is a $d \times K$ orthogonal matrix

SIR via MLE

Let A be the set of relevant predictors and $C = A^C$, $d = |A|$

$$\boldsymbol{X}_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

$$\boldsymbol{X}_C | \boldsymbol{X}_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

$\mu_s = \alpha + \boldsymbol{V}^{\boldsymbol{K}}$, belongs to a K-dimensional affine space $(K < d)$

SIR via MLE

Let A be the set of relevant predictors and $C = A^C$, $d = |A|$

$$\boldsymbol{X}_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

$$\boldsymbol{X}_C | \boldsymbol{X}_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

$\mu_s = \alpha + \boldsymbol{V}^K$, belongs to a K-dimensional affine space $(K < d)$

MLE of the span of subspace $\boldsymbol{V}^K$ coincides with SIR directions (Cook 2007, Szretter and Yohai 2009)

SIR via MLE

Let A be the set of relevant predictors and $C = \neg A$, $d = |A|$

$$\boldsymbol{X}_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

$$\boldsymbol{X}_C | \boldsymbol{X}_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

$\mu_s = \alpha + \boldsymbol{V}^K$, belongs to a K-dimensional affine space $(K < d)$

MLE of the span of subspace $\boldsymbol{V}^K$ coincides with SIR directions (Cook 2007, Szretter and Yohai 2009)

Given current A and predctor $X_{j \notin A}$, we want to test

$H_0: X_j$ is irrelevant, *vs.* $H_1: X_j$ is relevant

SIR via MLE

Let A be the set of relevant predictors and $C = A^C$, $d = |A|$

$$\boldsymbol{X}_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

$$\boldsymbol{X}_C | \boldsymbol{X}_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

$\mu_s = \alpha + \boldsymbol{V}^K$, belongs to a K-dimensional affine space $(K < d)$

MLE of the span of subspace $\boldsymbol{V}^K$ coincides with SIR directions (Cook 2007, Szretter and Yohai 2009)

Given current A and predctor $X_{j \notin A}$, we want to test

$H_0: X_j$ is irrelevant, *vs.* $H_1: X_j$ is relevant

$$\frac{P_{M_1}(\boldsymbol{X}|Y)}{P_{M_0}(\boldsymbol{X}|Y)} = \frac{P_{M_1}(X_j|\boldsymbol{X}_A, Y)}{P_{M_0}(X_j|\boldsymbol{X}_A, Y)}$$

SIR via MLE

Let A be the set of relevant predictors and $C = A^C$, $d = |A|$

$$\boldsymbol{X}_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

$$\boldsymbol{X}_C | \boldsymbol{X}_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

$\mu_s = \alpha + \boldsymbol{V}^K$, belongs to a K-dimensional affine space $(K < d)$

MLE of the span of subspace $\boldsymbol{V}^K$ coincides with SIR directions (Cook 2007, Szretter and Yohai 2009)

Given current A and predctor $X_{j \notin A}$, we want to test

$H_0: X_j$ is irrelevant, *vs.* $H_1: X_j$ is relevant

$$LR_j = \frac{P_{\hat{M}_1}(X_j | \boldsymbol{X}_A, Y)}{P_{\hat{M}_0}(X_j | \boldsymbol{X}_A, Y)}$$

LR Test vs. COP

Given current A, the likelihood ratio (LR) test statistic of $H_0: X_j$ is irrelevant, *vs.* $H_1: X_j$ is relevant

$$2\mathrm{LR}_j = -n\left(\sum_{k=1}^{K} log(1-\hat{\lambda}_k^{A+j}) - \sum_{k=1}^{K} log(1-\hat{\lambda}_k^{A})\right)$$

$$= n\sum_{k=1}^{K} log\left(1+\frac{\hat{\lambda}_k^{A+j}-\hat{\lambda}_k^{A}}{1-\hat{\lambda}_k^{A+j}}\right)$$

Under $H_0: X_j$ is irrelevant

$$COP_k^{A+j} = n\frac{(\hat{\lambda}_k^{A+j}-\hat{\lambda}_k^{A})}{1-\hat{\lambda}_k^{A+j}} \to_p \chi^2(1), \quad \frac{(\hat{\lambda}_k^{A+j}-\hat{\lambda}_k^{A})}{1-\hat{\lambda}_k^{A+j}} \to_p 0$$

LR Test vs. COP

Given current A, the likelihood ratio (LR) test statistic of $H_0: X_j$ is irrelevant, *vs.* $H_1: X_j$ is relevant

$$2\mathrm{LR}_j = -n\left(\sum_{k=1}^{K} log(1-\hat{\lambda}_k^{A+j}) - \sum_{k=1}^{K} log(1-\hat{\lambda}_k^{A})\right)$$

$$= n\sum_{k=1}^{K} log\left(1+\frac{\hat{\lambda}_k^{A+j}-\hat{\lambda}_k^{A}}{1-\hat{\lambda}_k^{A+j}}\right)$$

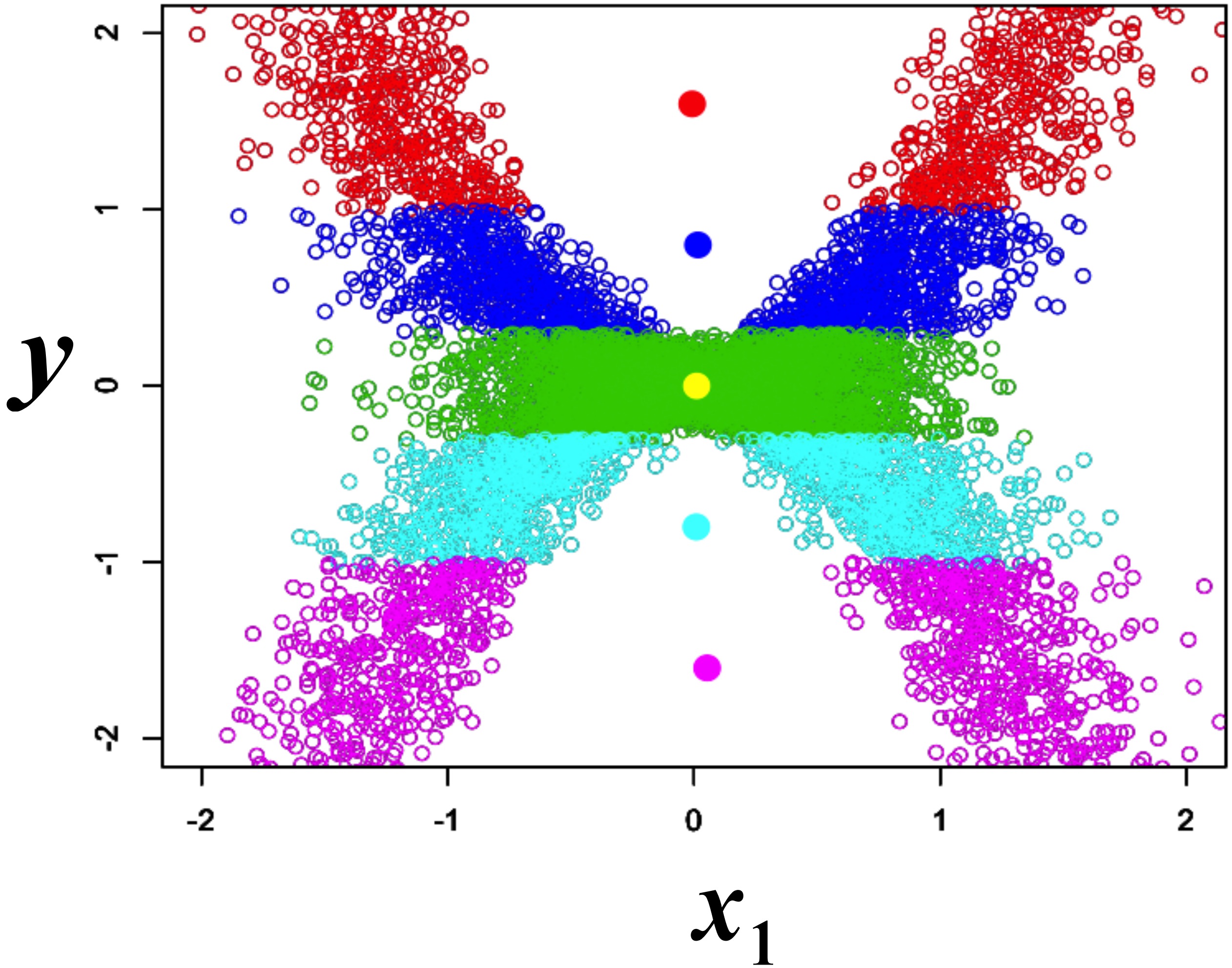
Under $H_0: X_j$ is irrelevant

$$COP_k^{A+j} = n\frac{(\hat{\lambda}_k^{A+j}-\hat{\lambda}_k^{A})}{1-\hat{\lambda}_k^{A+j}} \rightarrow_p \chi^2(1), \quad \frac{(\hat{\lambda}_k^{A+j}-\hat{\lambda}_k^{A})}{1-\hat{\lambda}_k^{A+j}} \rightarrow_p 0$$

$$\boxed{2\mathrm{LR}_j \rightarrow_p COP_{1:K}^{A+j} = \sum_{k=1}^{K} COP_k^{A+j} \rightarrow_p \chi^2(K)}$$

Beyond the First-order

- $E(X_1|Y)=0$



An Augmented Model

- For $h = 1, \ldots, H$,

$$\begin{aligned} \mathbf{X}_{\mathcal{A}} | Y \in S_h &\sim \mathrm{N}\left(\mu_h, \Sigma_h\right), \\ \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}}, Y \in S_h &\sim \mathrm{N}\left(\alpha + \beta' \mathbf{X}_{\mathcal{A}}, \Sigma_0\right). \end{aligned}$$

- If $\mathrm{Cov}\left(\mathbf{X}\right)$ is not degenerate, the *minimum* set $\mathcal{A}$ satisfying the above model is unique.
- Likelihood ratio test for $H_0 : \mathcal{A} = \mathcal{C}$ v.s. $H_1^* : \mathcal{A} = \mathcal{C} \cup \{j\}$,

$$L^*_{j|\mathcal{C}} = \frac{P_{\mathcal{M}_1^*}\left(\mathbf{x}|y\right)}{P_{\mathcal{M}_0}\left(\mathbf{x}|y\right)} = \frac{P_{\mathcal{M}_1^*}\left(x_j|\mathbf{x}_{\mathcal{C}}, y\right)}{P_{\mathcal{M}_0}\left(x_j|\mathbf{x}_{\mathcal{C}}, y\right)}$$

Likelihood Ratio Test

- Assume $\left[\widehat{\sigma}_j^{(h)}\right]^2$ is the estimated variance by regressing X_j on $X_{\mathcal{C}}$ in slice S_h and $\widehat{\sigma}_j^2$ is the overall estimated variance. Then,

$$\widehat{D}^*_{j|\mathcal{C}} = \frac{2}{n}\log\left(L^*_{j|\mathcal{C}}\right) = \log\widehat{\sigma}_j^2 - \sum_{h=1}^{H}\frac{n_h}{n}\log\left[\widehat{\sigma}_j^{(h)}\right]^2$$

- When $\mathcal{A}\subset\mathcal{C}$ and $|\mathcal{C}| = d$, $n\widehat{D}^*_{j|\mathcal{C}} \xrightarrow{d} \chi^2_{(H-1)(d+2)}$

$$\boxed{\widehat{D}^*_{j|\mathcal{C}} \sim \log\left(1+\frac{Q_0}{\sum_{h=1}^{H}Q_h}\right) - \sum_{h=1}^{H}\frac{n_h}{n}\log\left(\frac{Q_h/n_h}{\sum_{h=1}^{H}Q_h/n}\right)}$$

where $Q_0 \sim \chi^2_{(H-1)(d+1)}$ and $Q_h \sim \chi^2_{n_h-d-1}$ $(h = 1,\ldots,H)$ are mutually independent according to *Cochran's theorem*.

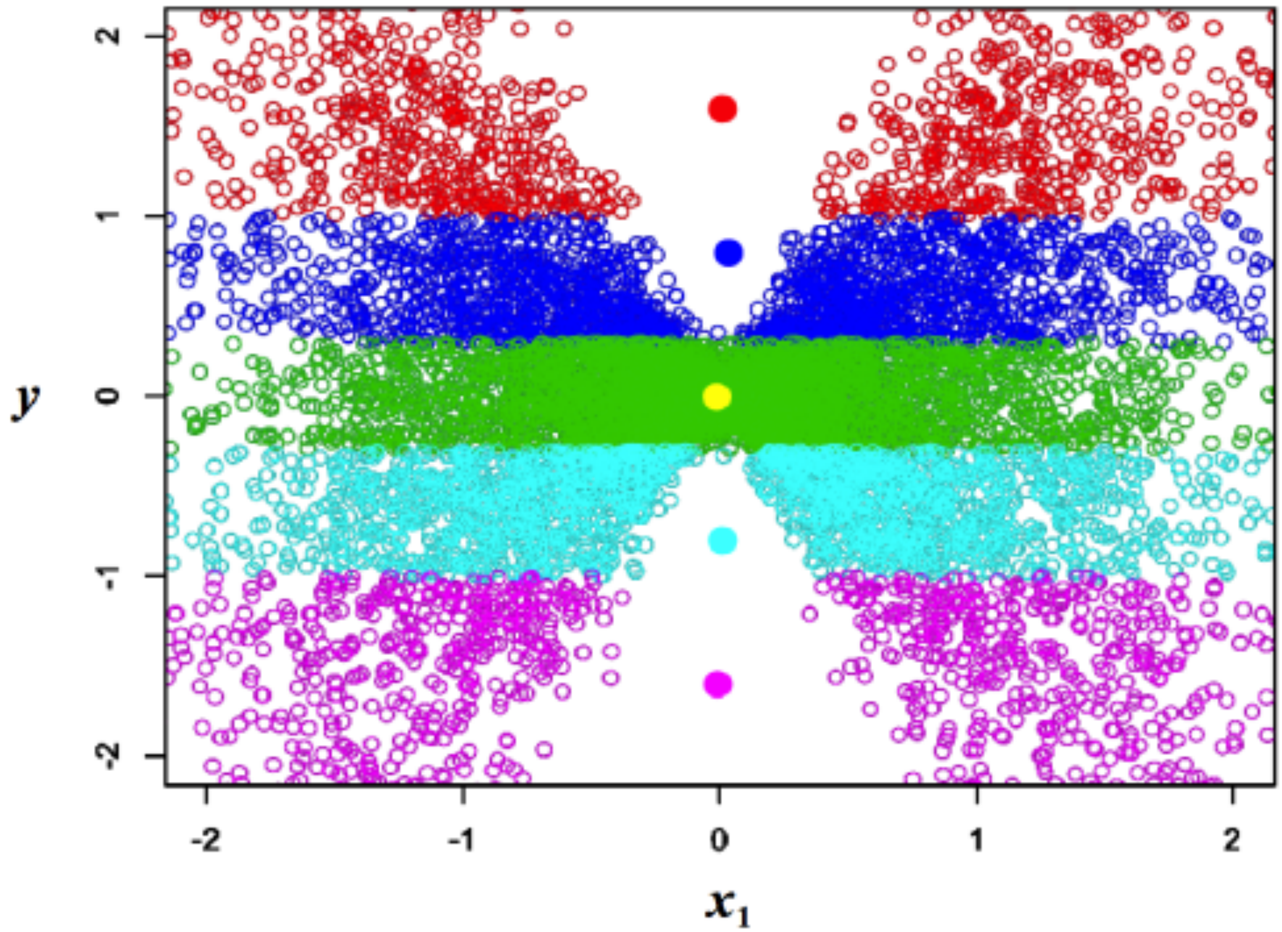
- As $n \to \infty$, we have

$$
\begin{aligned}
& \widehat{D}^*_{j|\mathcal{C}} \xrightarrow{\text{a.s.}} D^*_{j|\mathcal{C}} \\
= \; & \log\left(1 + \frac{\mathrm{Var}(M_j) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Var}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)'}{\mathbb{E}\left(V_j\right)}\right) \\
+ \; & \log \mathbb{E}\left(V_j\right) - \mathbb{E}\log\left(V_j\right)
\end{aligned}
$$

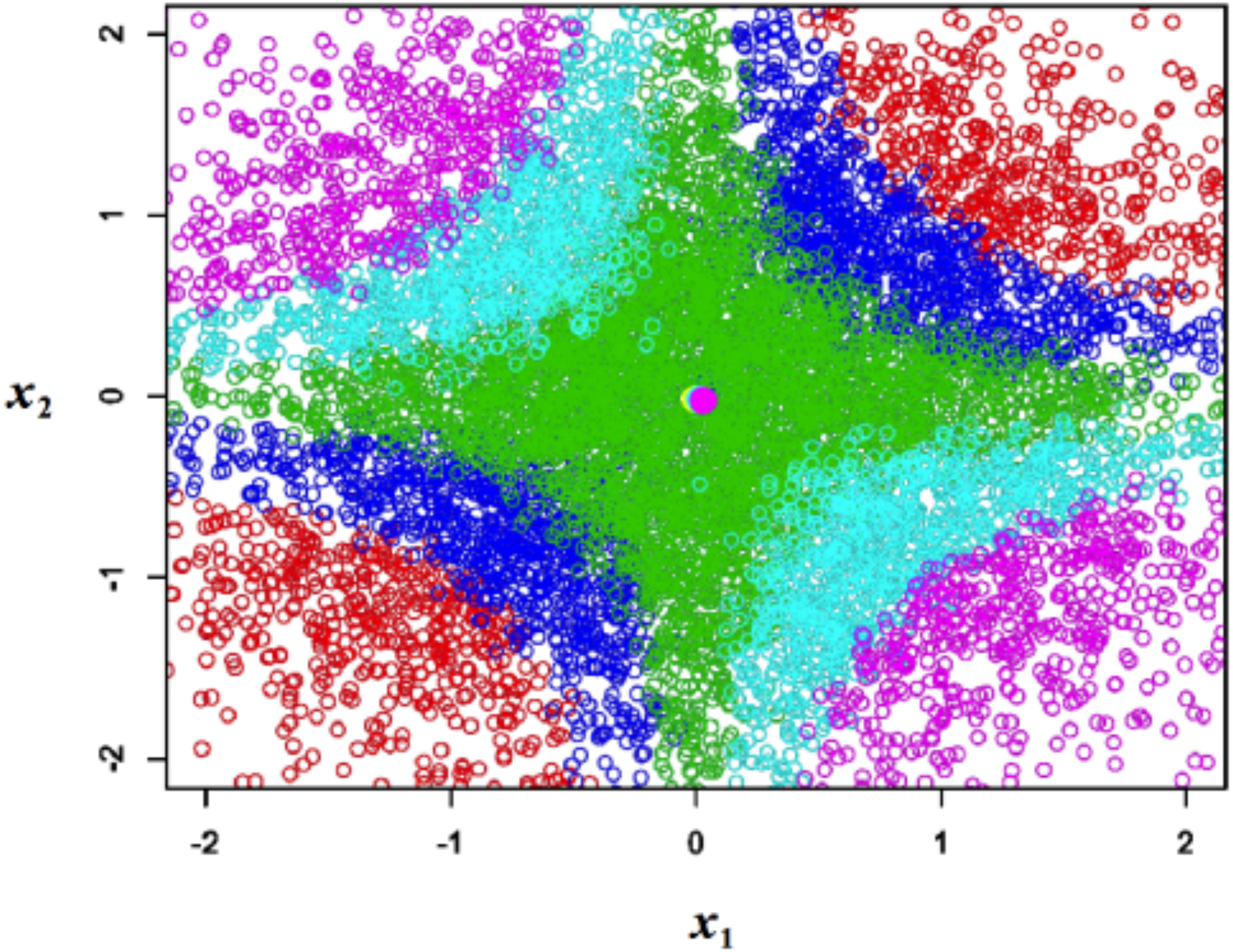
  where $M_j = \mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, \mathbb{S}(Y)\right)$ and $V_j = \mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}, \mathbb{S}(Y)\right)$.
- $\widehat{D}^*_{j|\mathcal{C}} \xrightarrow{\text{a.s.}} 0$ if and only if $\mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, \mathbb{S}(Y)\right) = \mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}\right)$ and $\mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}, \mathbb{S}(Y)\right) = \mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}\right)$
- A stepwise procedure based on $\widehat{D}^*_{j|\mathcal{C}}$ is consistent when $p = O(n^\gamma)$ with $\gamma < 1/2$.

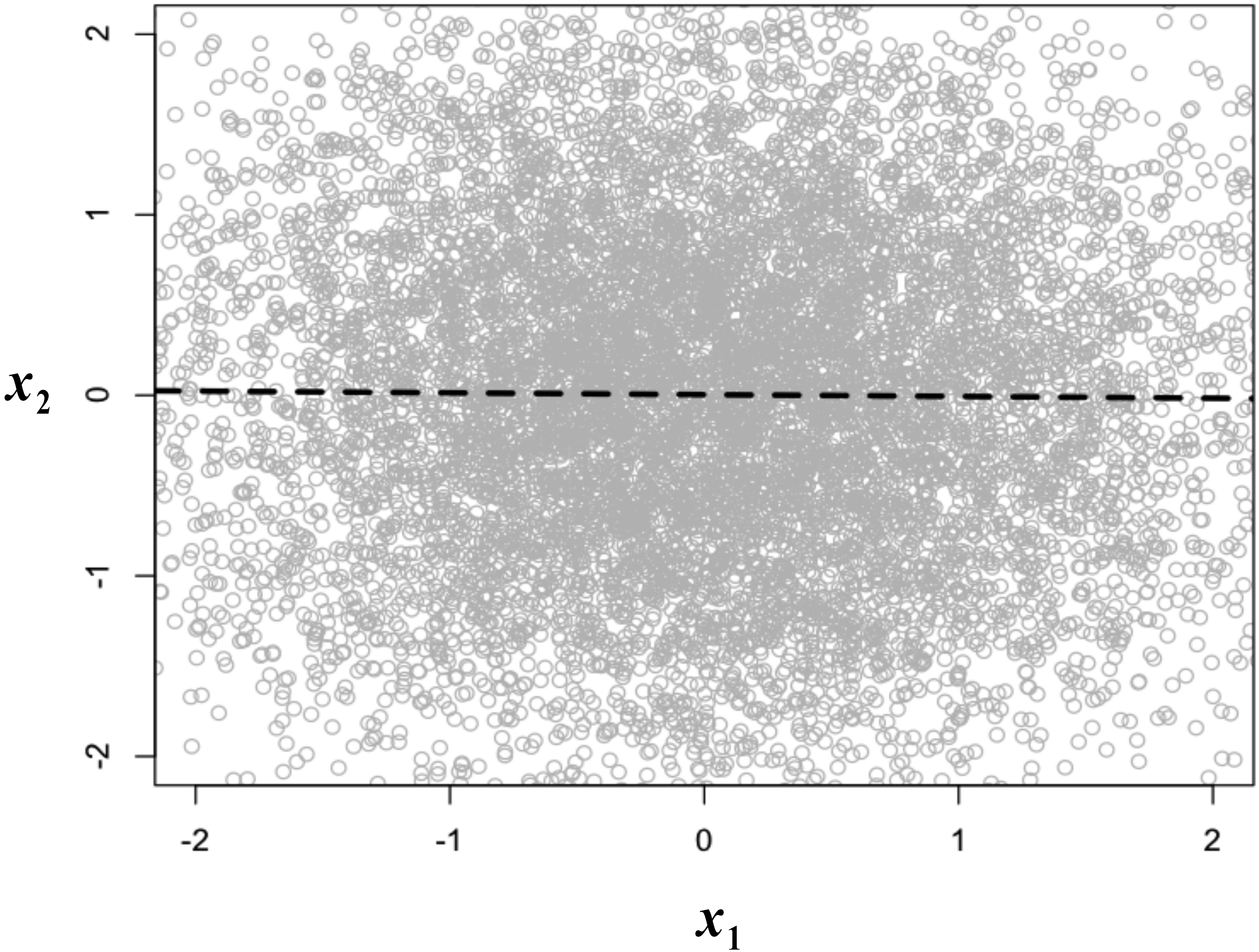
Example revisit

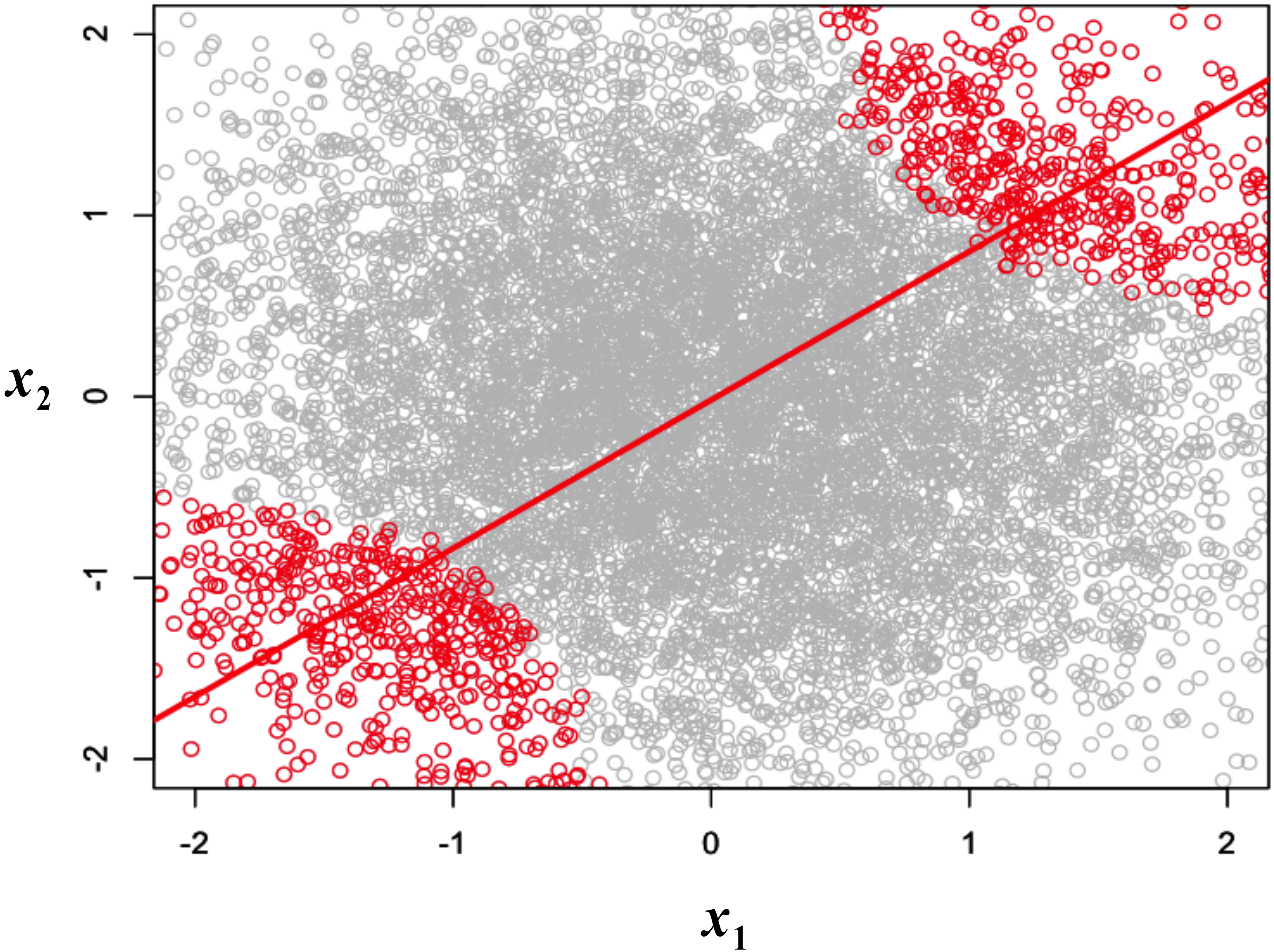
- $\mathrm{Var}(X_1|Y)$ depends on Y.



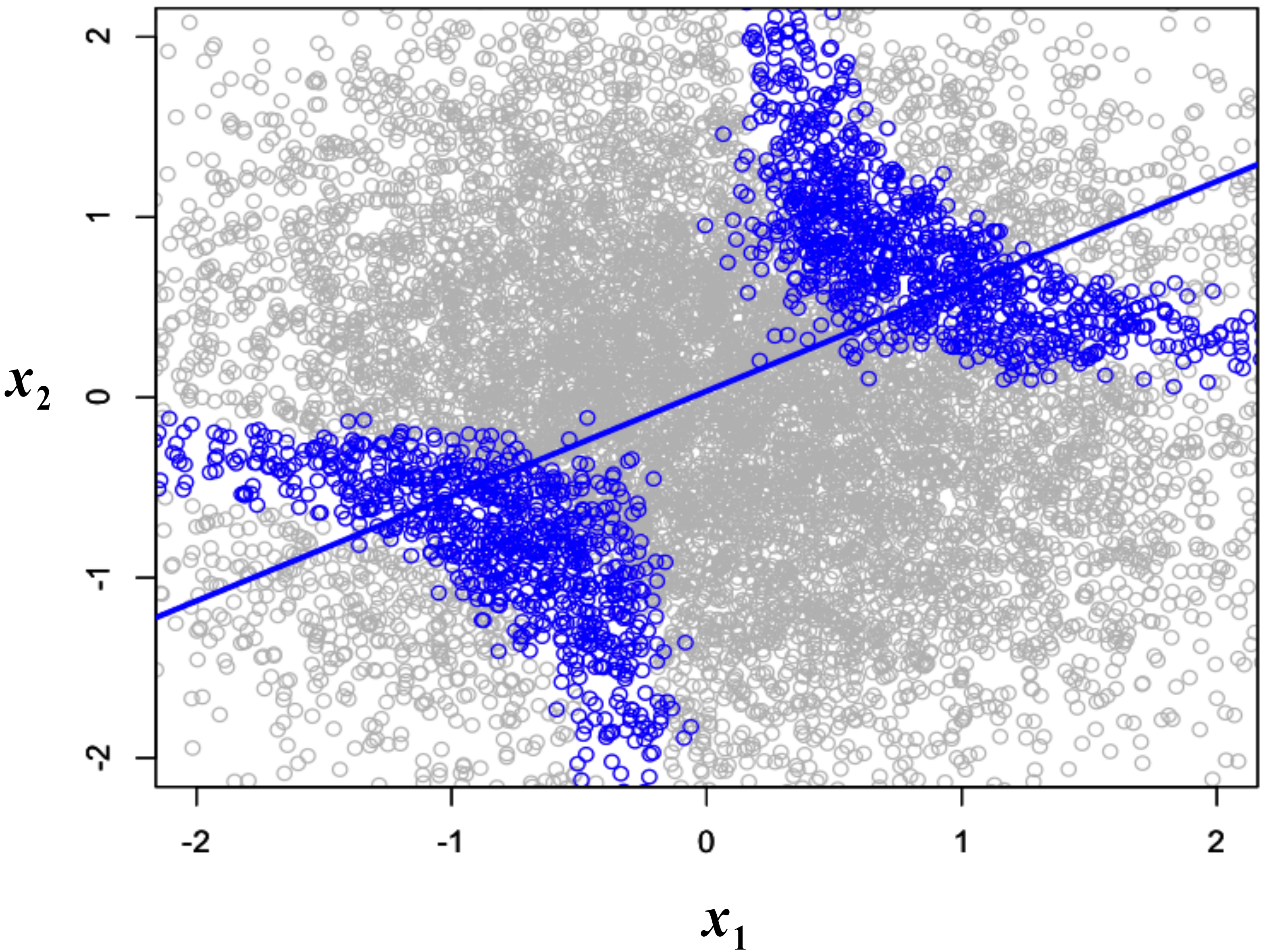
- $\mathbb{E}(X_2|X_1, Y)$ depends on Y.

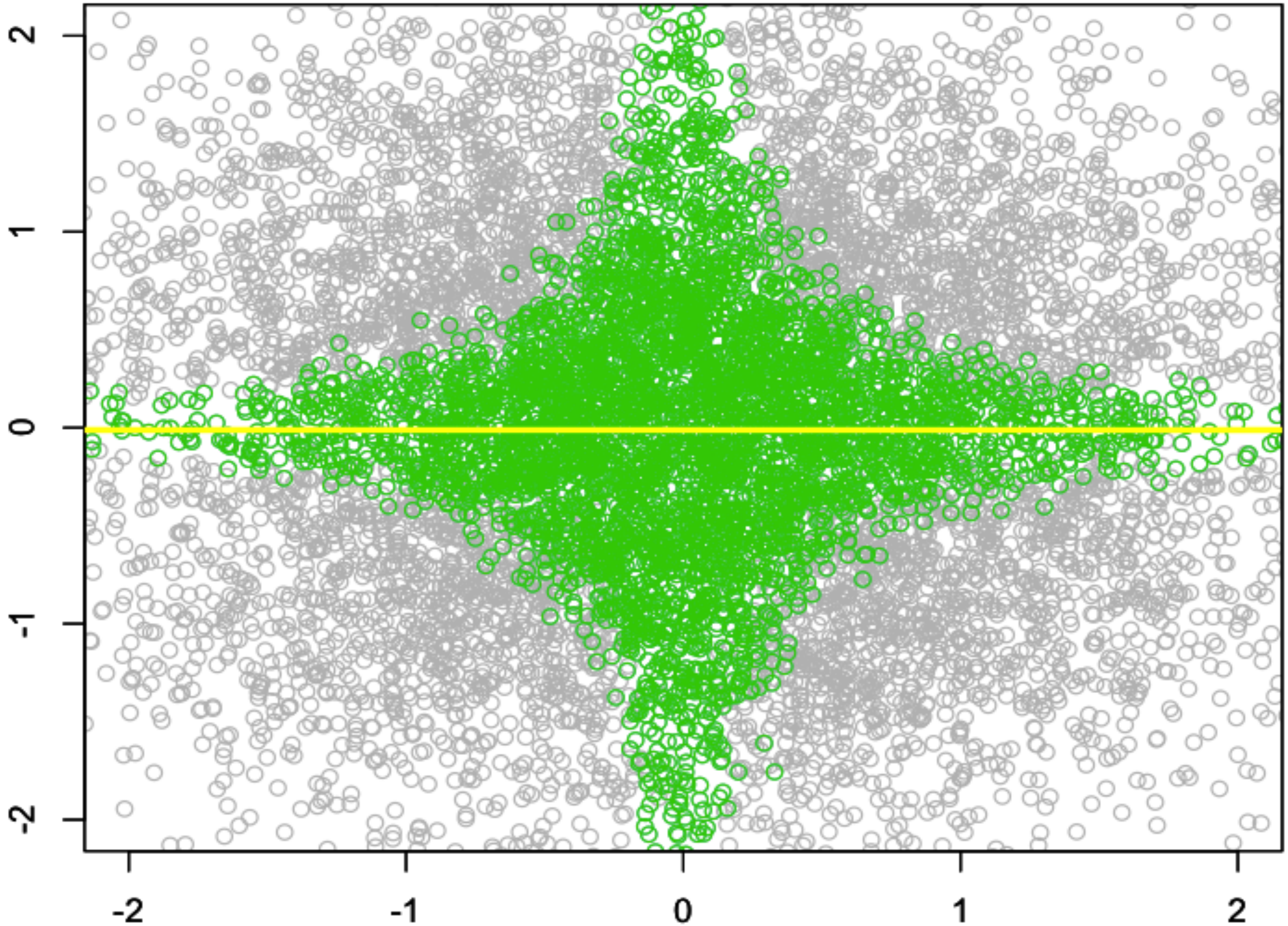



2
1
0
-1
-2
-2
-1
0
1
2
x_2
x_1



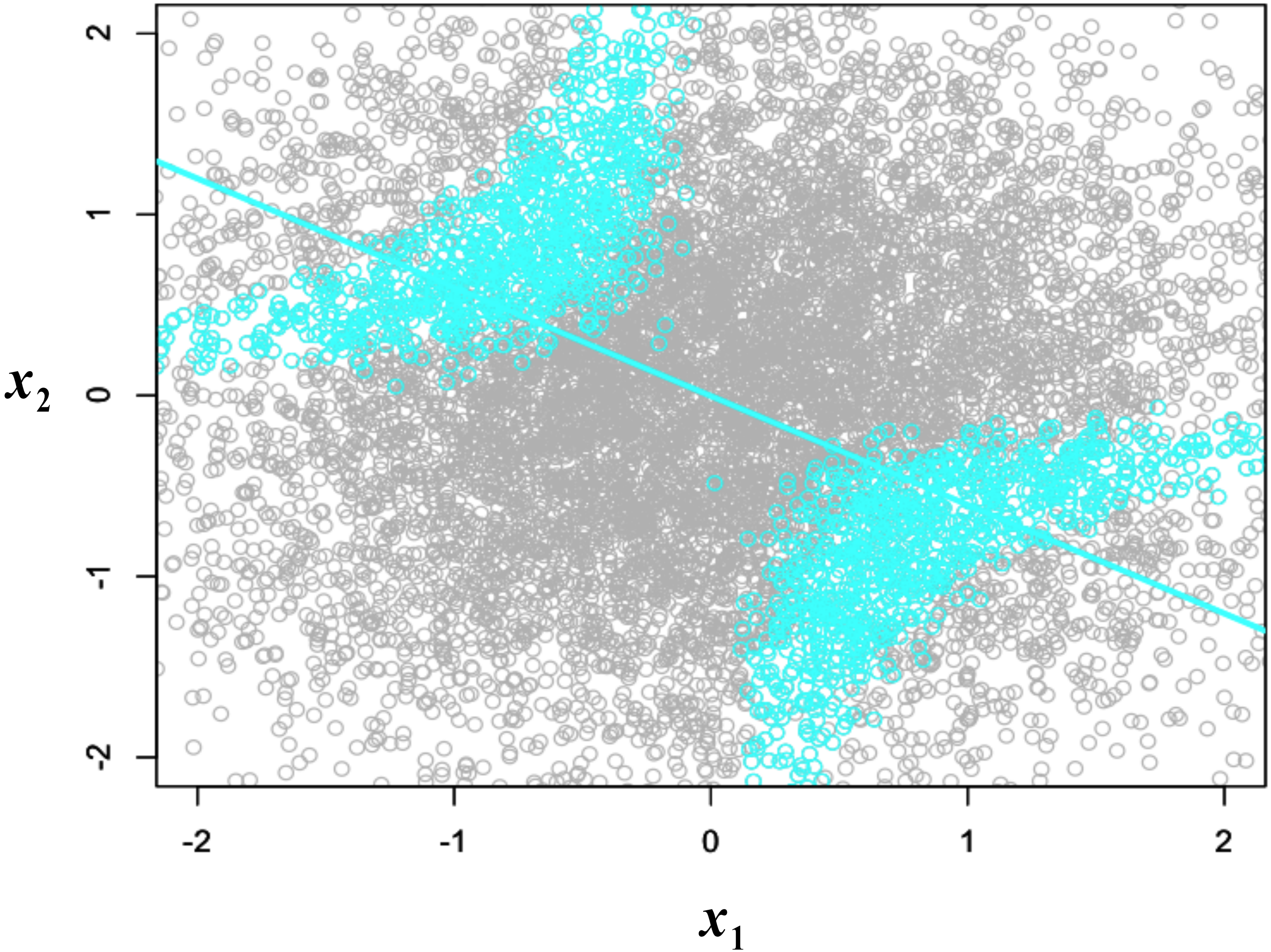
2
1
0
-1
-2
x_2
-2
-1
0
1
2
x_1


x_2
2
1
0
-1
-2
-2
-1
0
1
2
x_1

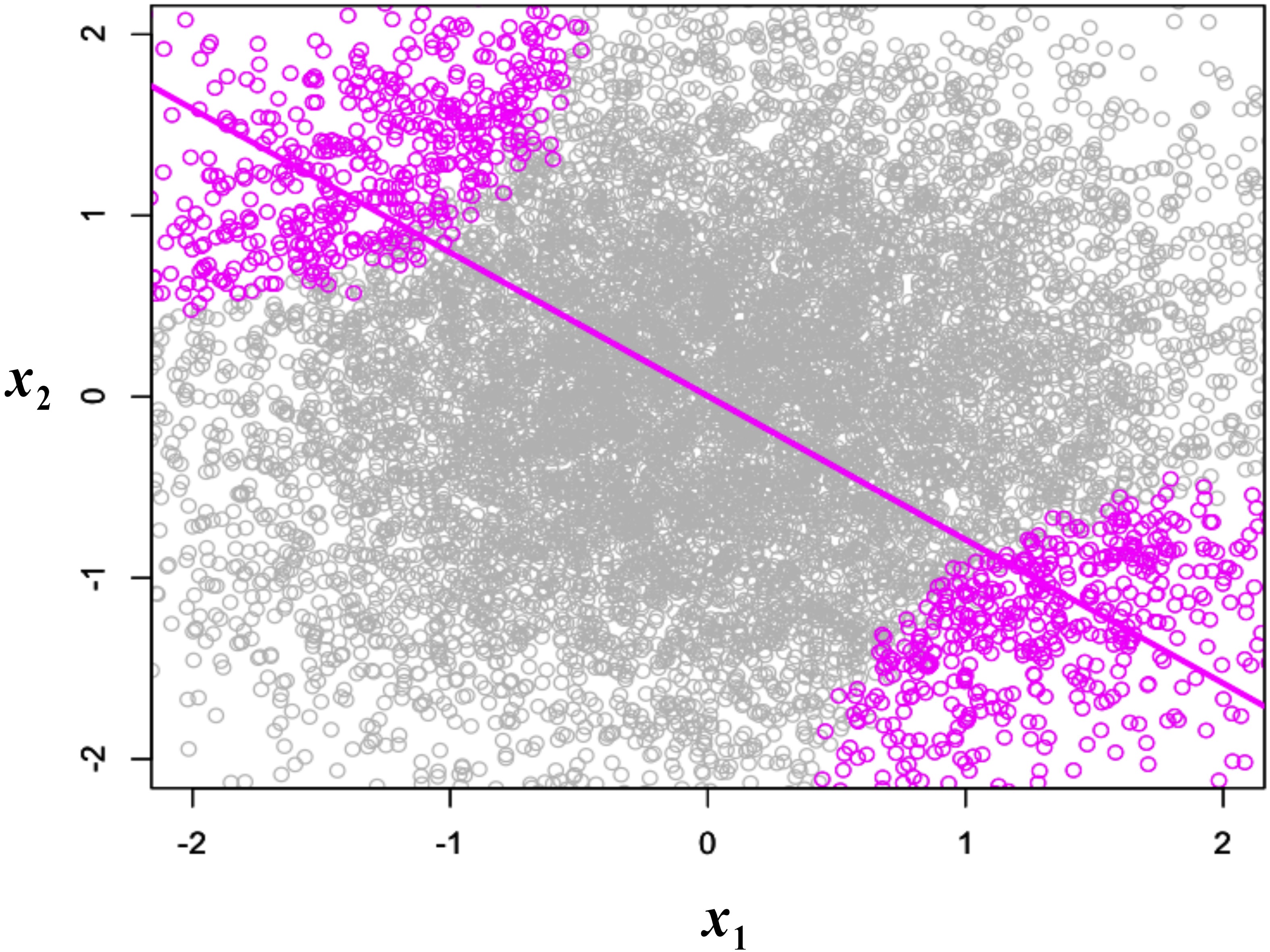


x_2

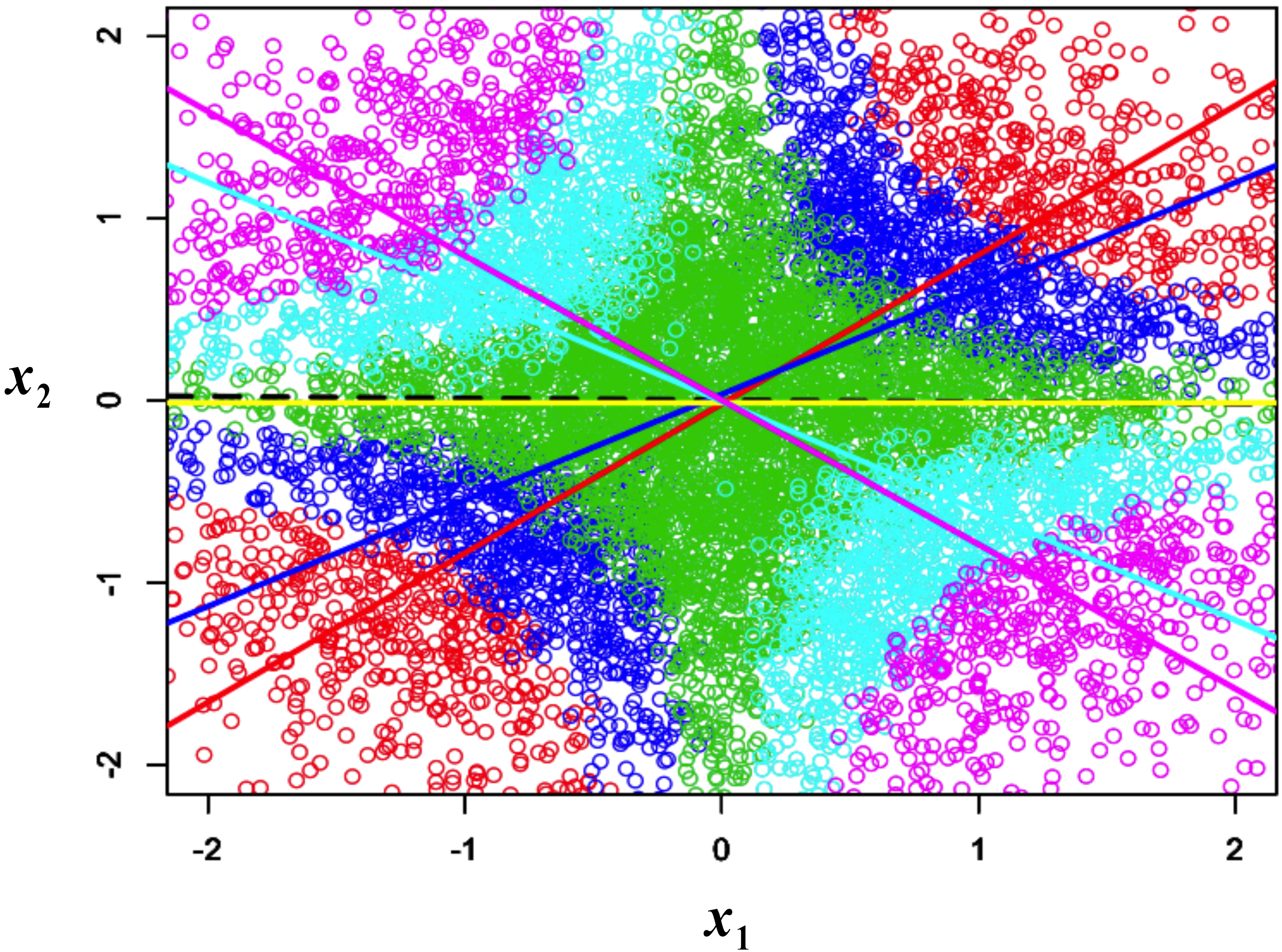
x_1



x_2
2
1
0
-1
-2
-2
-1
0
1
2
x_1



2
1
0
-1
-2
x_2
-2
-1
0
1
2
x_1


2
1
0
-1
-2
x_2
-2
-1
0
1
2
x_1

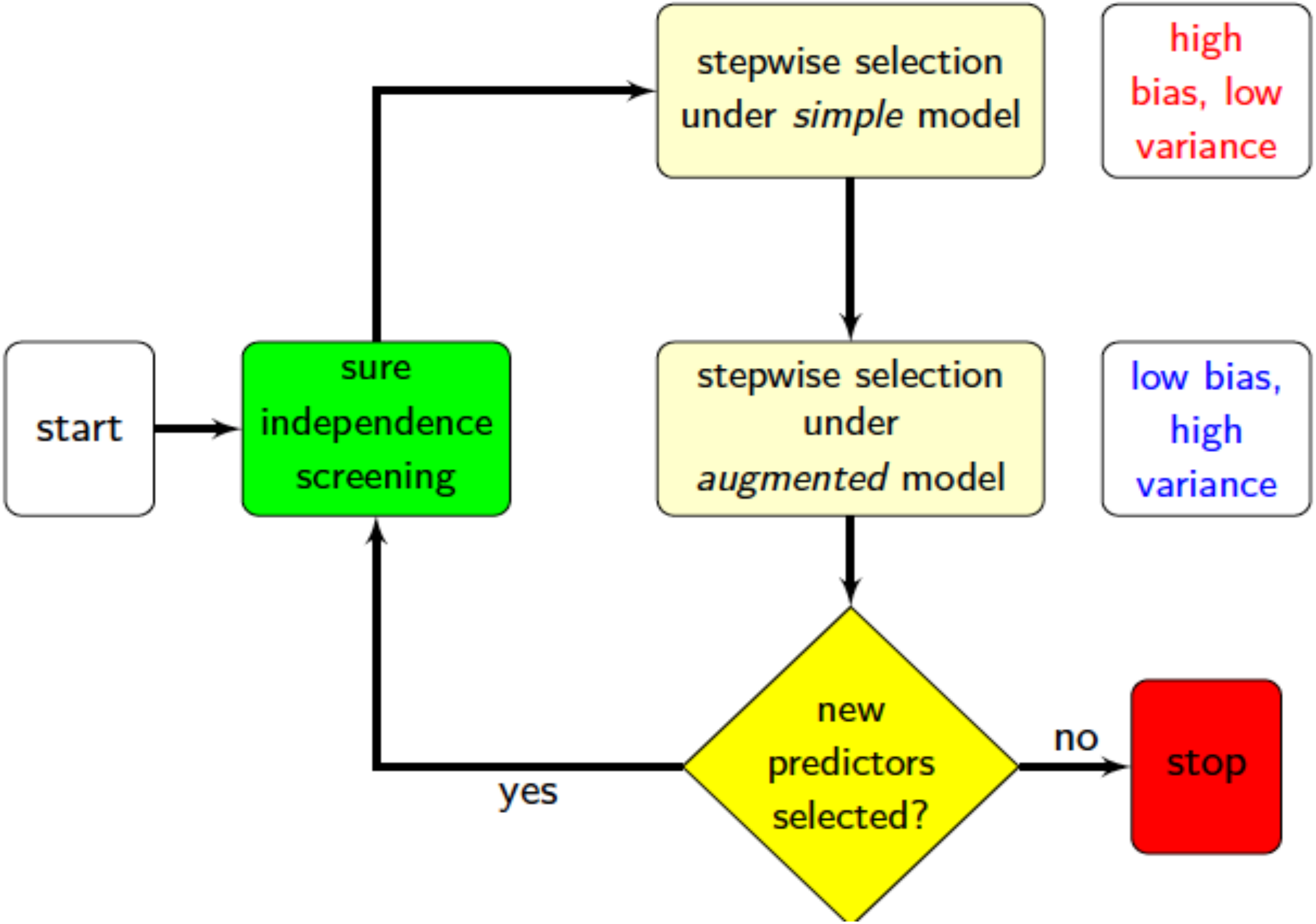
Sure independence screening (SIS) when p>>n

- Independence screening: first independently selects variables based on their marginal relationships with the response and then applies refined methods in the second step.
- Rank predictors according to $\widehat{D}^*_{j|\mathcal{C}}$ with $\mathcal{C} = \emptyset$:

$$\begin{aligned}\widehat{D}^*_{j|\mathcal{C}=\emptyset} \quad &\xrightarrow{\text{a.s.}} \quad \log\left(1 + \frac{\mathrm{Var}\left(\mathbb{E}(X_j|\mathbb{S}(Y))\right)}{\mathbb{E}\left(\mathrm{Var}(X_j|\mathbb{S}(Y))\right)}\right) \\ &+ \quad \log \mathbb{E}\left(\mathrm{Var}(X_j|\mathbb{S}(Y))\right) - \mathbb{E}\log\left(\mathrm{Var}(X_j|\mathbb{S}(Y))\right)\end{aligned}$$

- The first $n-1$ variables have a high probability to include the true predictors $\mathcal{A}$ (almost surely under moderate conditions) even when $\log(p) = O(n^\gamma)$ with $\gamma < 1$.

Siri: An interweaving strategy



Theoretical Properties

- Under moderate conditions, a stepwise procedure with forward selection and backward elimination is consistent when $p = O(n^\gamma)$ with $\gamma < 1/2$.
- By choosing the threshold appropriately, the addition step will not stop selecting variables until all the true predictors have been included.
- Once all the true predictors have been included, all the redundant variables will be removed from the selected variables.

Conditions for consistency

- For $j \in \mathcal{A}$, we have

$$X_j | \mathbf{X}_{\mathcal{A}-\{j\}}, Y \in S_h \sim \mathrm{N}\left(\alpha_j^{(h)} + \beta' \mathbf{X}_{\mathcal{A}-\{j\}}, \sigma_j^2\right).$$

- Condition 1 (detectability): Let $\alpha_j(Y) = \sum_h^H \alpha_j^{(h)} \mathbb{I}(Y \in S_h)$. There exist $\xi > 0$ and $\kappa > 0$ such that

$$\mathrm{Var}\left(\alpha_j(Y)\right) \geq \xi n^{-\kappa} \text{ for } j \in \mathcal{A}.$$

- Condition 2 (dependency): The eigenvalues of $\mathrm{Var}\left(\mathbf{X}\right)$ and $\mathrm{Var}\left(\mathbf{X} | Y \in S_h\right)$ $(h = 1, \ldots, H)$ have positive lower and upper bounds.
- Condition 3 (dimensionality): $\lim_{n \to \infty}(p) = \infty$ and $p = o(n^\rho)$ with $\rho > 0$ and $2\rho + 2\kappa < 1$.

Consistency of Stepwise Procedure

- Under Condition 1-3, as $n \to \infty$, there exist constants $c > 0$ and $\kappa \geq 0$ such that

$$\Pr\left(\min_{\mathcal{C}:\mathcal{C}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{C}^c} \widehat{D}_{j|\mathcal{C}} \geq cn^{-\kappa}\right) \to 1, \text{ and}$$

$$\Pr\left(\max_{\mathcal{C}:\mathcal{C}^c \cap \mathcal{A} = \emptyset} \max_{j \in \mathcal{C}^c} \widehat{D}_{j|\mathcal{C}} < Cn^{-\kappa}\right) \to 1 \text{ for any } C > 0.$$

- If we choose $t_a = cn^{-\kappa}$ and $t_d = (c/2)n^{-\kappa}$, then the addition step will not stop selecting variables until all the true predictors have been included.

Implementation Issues

- Under Condition 1-3, as $n \to \infty$, there exist constants $c > 0$ and $\kappa \geq 0$ such that

$$\Pr\left(\min_{\mathcal{C}:\mathcal{C}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{C}^c} \widehat{D}_{j|\mathcal{C}} \geq cn^{-\kappa}\right) \to 1, \text{ and}$$

$$\Pr\left(\max_{\mathcal{C}:\mathcal{C}^c \cap \mathcal{A} = \emptyset} \max_{j \in \mathcal{C}^c} \widehat{D}_{j|\mathcal{C}} < Cn^{-\kappa}\right) \to 1 \text{ for any } C > 0.$$

- If we choose $t_a = cn^{-\kappa}$ and $t_d = (c/2)n^{-\kappa}$, then the addition step will not stop selecting variables until all the true predictors have been included.
- Once all the true predictors have been included, all the redundant variables will be removed from the selected variables.

Simulation I (linear)

$$Y = \boldsymbol{X}_p \beta + \epsilon,\ \epsilon \sim N(0,1),\ Cov(X_i, X_j) = 0.5^{|i-j|}$$

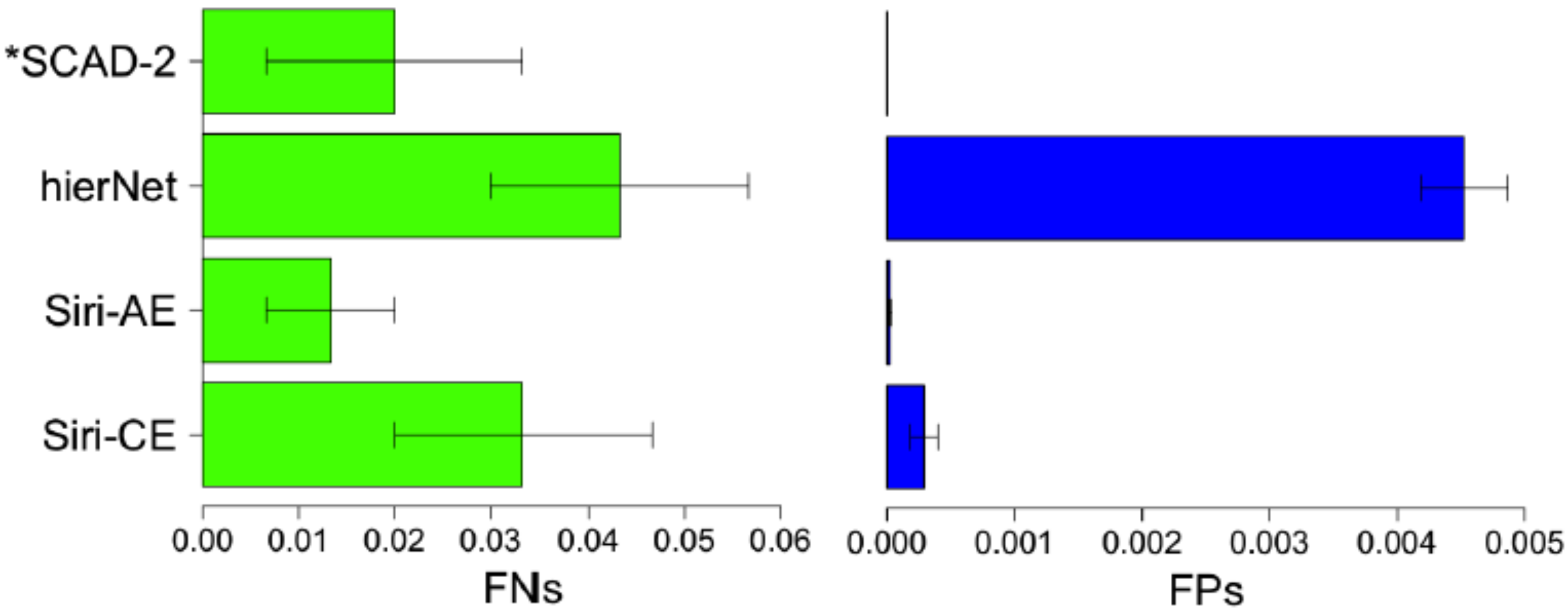
$$n = 200,\ p = 1000,\ \beta = (3,1.5,1,1,2,1,0.9,1,1,1,0,\ldots,0)^T$$

Method	FP(0, 990)	FN(0, 10)
SIRI-C [CV minimizing classification error]	**1.86 (0.222)**	**1.66 (0.117)**
SIRI-M [CV minimizing mean square error]	**0.76 (0.120)**	**1.75 (0.114)**
COP	**1.62 (0.165)**	**1.67 (0.118)**
SIS-SCAD	**0.10 (0.030)**	**0.64 (0.069)**
LASSO	**5.40 (0.188)**	**0.00 (0.000)**

Simulation II: hierarchical interactions

- $\mathbf{X} \sim \mathrm{MVN}\left(\mathbf{0}, \mathbb{I}_p\right)$ with $n = 200$, $p = 1000$, and

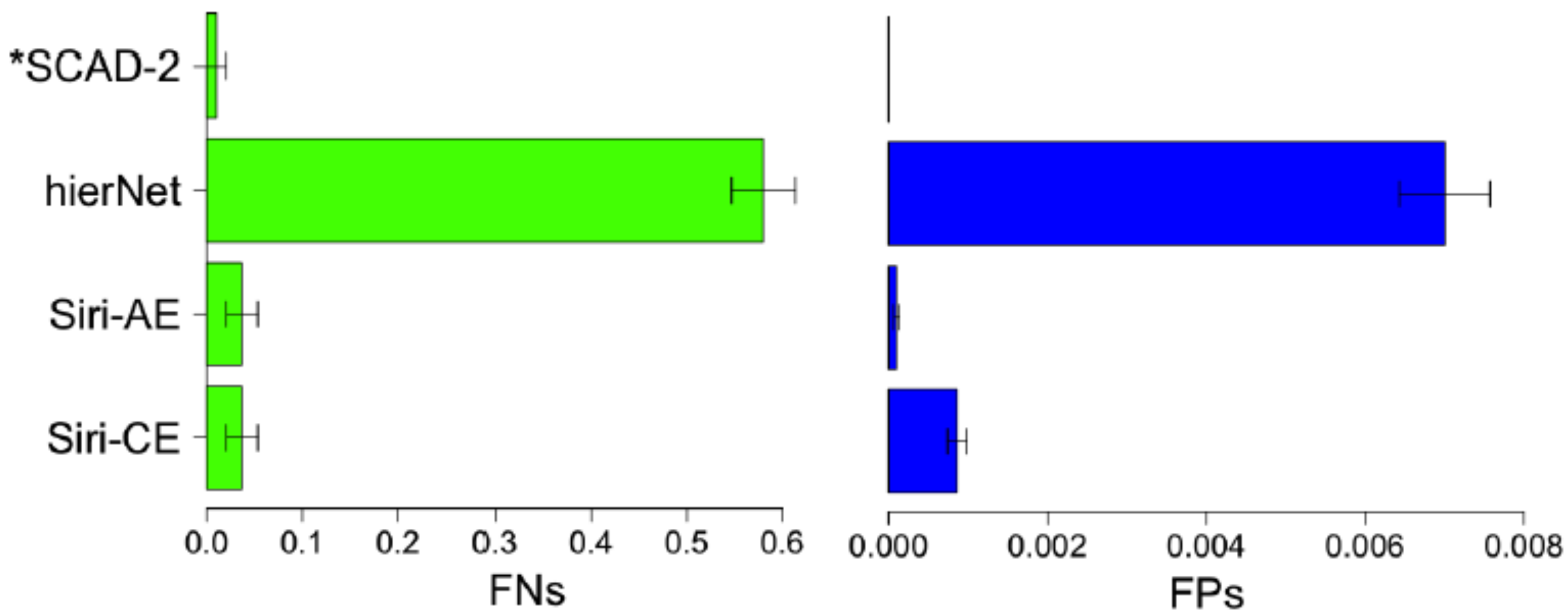
$$Y = X_1 + X_1 X_2 + X_1 X_3 + 0.2\epsilon$$



Simulation III: non-hierarchical

- $\mathbf{X} \sim \mathrm{MVN}(\mathbf{0}, \mathbb{I}_p)$ with $n = 200$, $p = 1000$, and

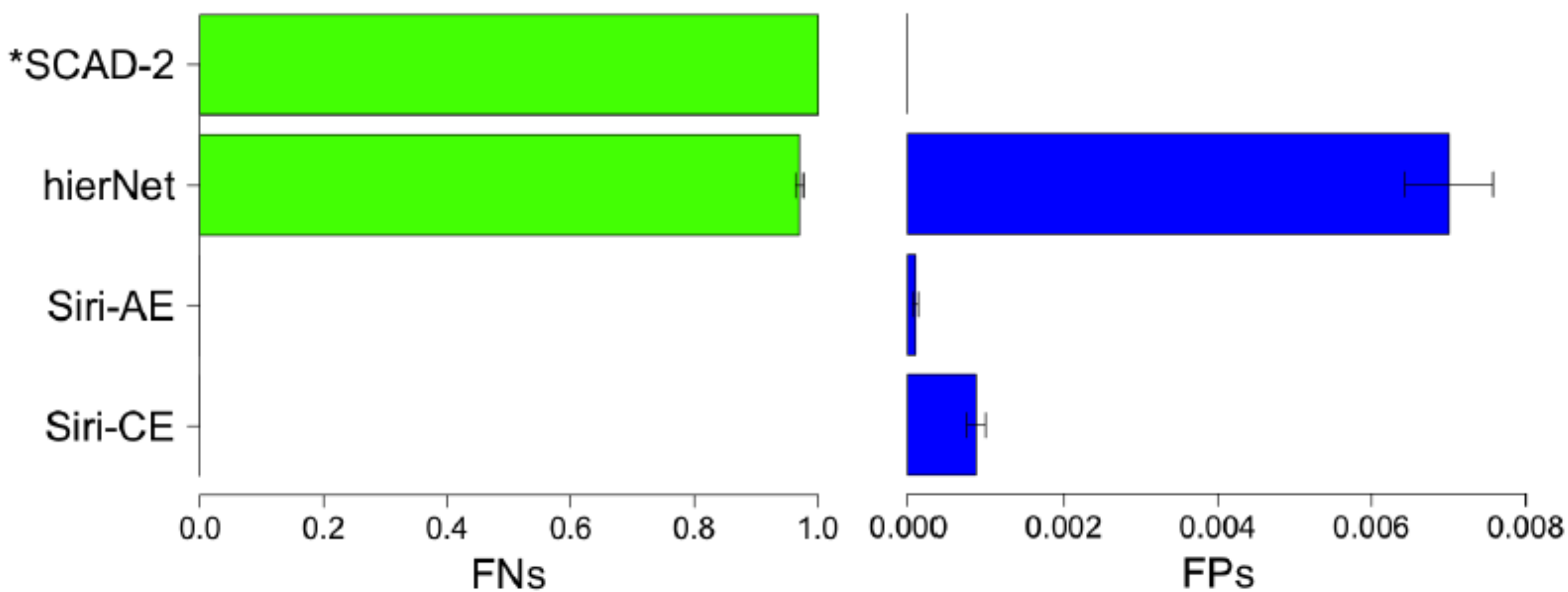
$$Y = X_1X_2 + X_1X_3 + 0.2\epsilon$$



Simulation IV: Non-multiplicative

- $\mathbf{X} \sim \mathrm{MVN}(\mathbf{0}, \mathbb{I}_p)$ with $n = 200$, $p = 1000$, and

$$Y = \frac{X_1}{X_2 + X_3} + 0.2\epsilon$$



Simulation V (heteroscedastic, single index)

$$Y = \frac{0.2\epsilon}{1.5 + \sum_{j=1}^{8} X_j}, \quad \boldsymbol{X}_p \sim \text{indepdent normal}$$

$$n = 1000, \quad p = 1000$$

Method	FP(0, 992)	FN(0, 8)
SIRI-C	2.00 (0.163)	0.42 (0.138)
SIRI-M	0.43 (0.079)	4.60 (0.274)
COP	1.26 (0.128)	3.32 (0.192)
SIS-SCAD	3.23 (0.356)	8.00 (0.000)
LASSO	0.64 (0.255)	8.00 (0.000)

Simulation VI (hub with linear effect)

$Y = X_1 + X_1 \times (X_2 + X_3) + 0.2\epsilon$, $\boldsymbol{X}_p \sim$ indepdent normal

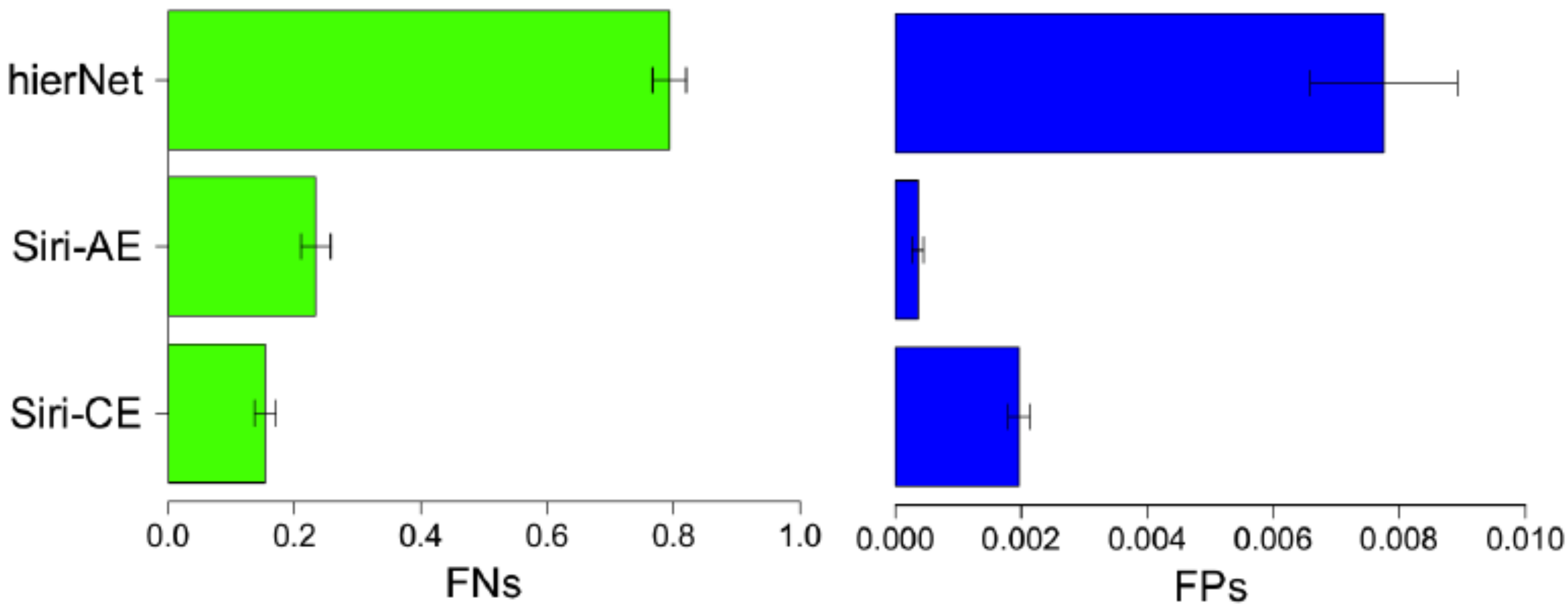
$n = 200$, $p = 1000$

Method	FP(0, 997)	FN(0, 3)
SIRI-C	0.39 (0.115)	0.12 (0.046)
SIRI-M	0.03 (0.017)	0.04 (0.020)
SIS-SCAD-2	0.00 (0.000)	0.45 (0.068)

Simulation VII (three-way interaction)

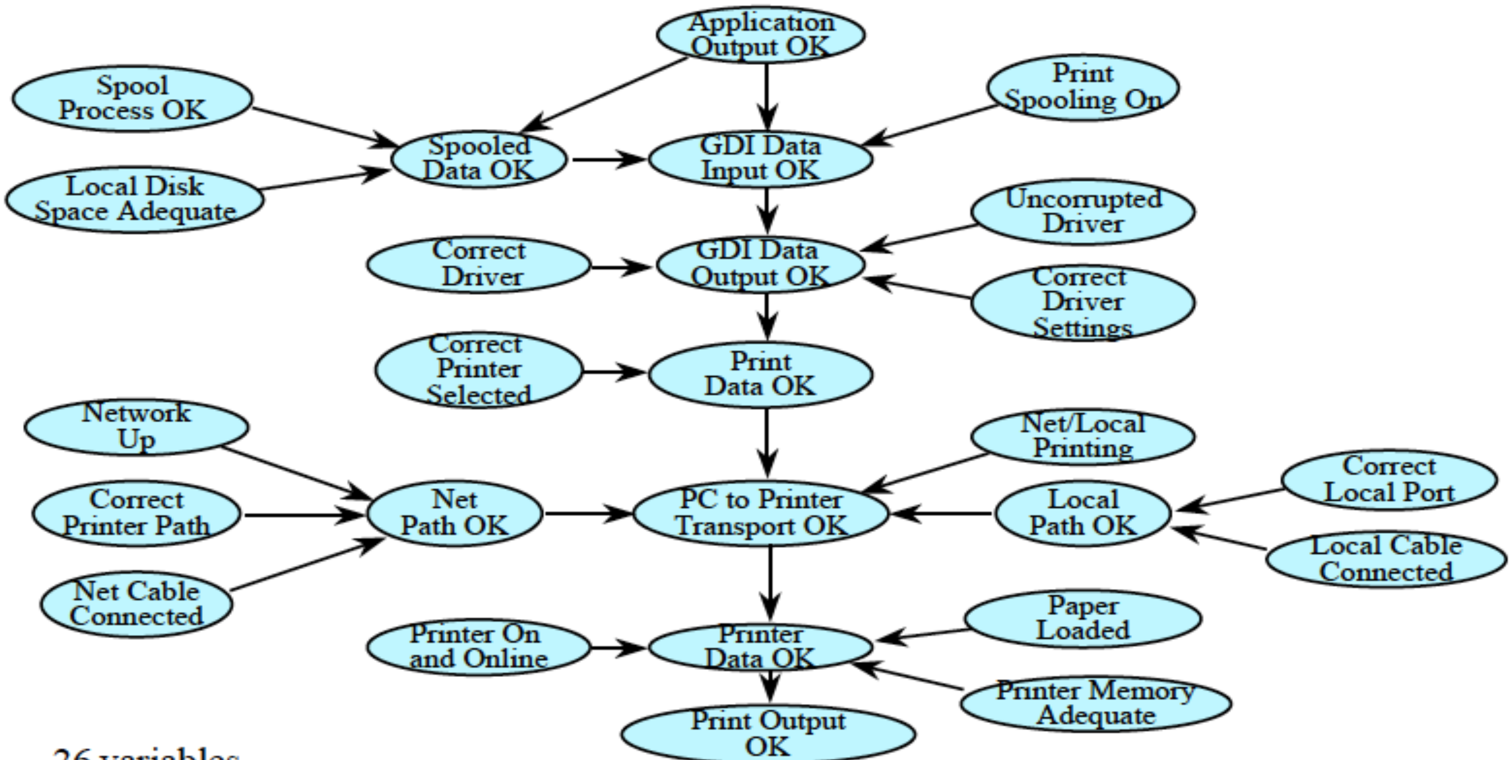
$$Y = X_1 \times X_2 \times X_3 + 0.2\epsilon\,,\ \boldsymbol{X}_p \sim \text{indepdent normal}$$

$$n = 500\,,\ p = 1000$$



Bayesian Networks

Example: Printer Troubleshooting



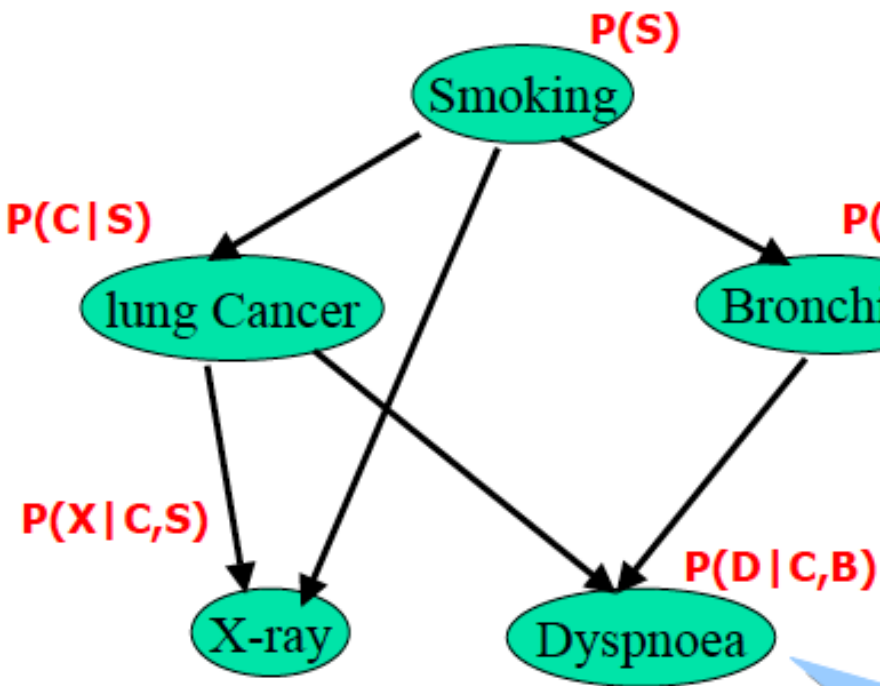
26 variables

Instead of 2^{26} parameters we get

$99 = 17x1 + 1x2^1 + 2x2^2 + 3x2^3 + 3x2^4$

[Heckerman, 95]

Bayesian Network: $\mathbf{BN = (G, \Theta)}$



G - directed acyclic graph (DAG)
nodes – random variables
edges – direct dependencies

Θ - set of parameters in all conditional probability distributions (CPDs)

CPD:

C	B	D=0	D=1
0	0	0.1	0.9
0	1	0.7	0.3
1	0	0.8	0.2
1	1	0.9	0.1

CPD of node X: P(X|parents(X))

Compact representation of joint distribution in a **product form** (chain rule):

$$P(S, C, B, X, D) = P(S)\, P(C|S)\, P(B|S)\, P(X|C,S)\, P(D|C,B)$$

$$\mathbf{1+2+2+4+4=13 \text{ parameters instead of } 2^5=32}$$

Learning BN structures

- Global approach (Score/likelihood based):
  - Posterior inference:

$$P(G \mid Data) \propto \int P(Data \mid \theta_G, G)\, p(\theta_G \mid G)\, p(G)\, d\theta_G$$

  - Or score-based criterion
    - AIC = $-2\log P(Data \mid \hat{\theta}_G, G) + 2p_G$
    - BIC = $-2\log P(Data \mid \hat{\theta}_G, G) + \log(n)\, p_G$

Learning structures

- Local approaches: using conditional independence statements as constraints.
  - Represented by “Inductive causality” (IC) algorithm due to Pearl (2000).

1. First, the skeleton of the network (undirected graph underlying the network structure) is learned by recursively testing the conditional independence between nodes.
2. Set all direction of the arcs that are part of a v-structure, which is a triplet of nodes incident on a converging connection $X_j \rightarrow X_i \leftarrow X_k$
3. Set the directions of the order arcs as needed to satisfy the acyclicity constraint.

Finding Markov blanket for each note using Growth-Shrink (GS) algorithm

- It is like a stepwise regression. For each note X_i, we treat it as the response variable and
  - (a) gradually add variables that are predictive of X_i;
  - (b) Backward removing those "redundant" X_j's obtained from the growth phase.

Discussion

- Cross-validation to select the dimension and thresholds
- Back to full Bayesian model with dynamic slicing
  - We want to have flexibility in choosing slicing boundaries
  - Connection with Mutual-Information Criterion (MIC)
  - Many interesting possibilities
- Robustness to the distribution of predictors

Acknowledgment

Bo Jiang – who did all the work

Dr. Tingting Zhang, Dr. Wenxuan Zhong

Joseph K. Blitzstein