Tail Inverse Regression: dimension reduction for prediction of extremes

Anass Aghbalou, François Portier, Anne Sabourin, Chen Zhou

7 décembre 2023

Introduction

- $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, p grand
- Objectif : classification des $\{Y > y\}$ extrêmes
- Exhiber un espace pertinent de dimension $d << p$
- Contribution majeur : TIREX1, TIREX2
- Algorithmes concurrents
- Algorithme rapide, et utilisant peu d'hypothèses

Sommaire

- Sufficient Dimension Reduction (SDR). Espace E, avec P la projection sur cet espace :

$$\mathbb{P}(Y \leq t|X) = \mathbb{P}(Y \leq t|PX)$$

- But : trouver un espace SDR minimal
- Existence garantie, unicité ?

- Standardisation des variables : $Z = \Sigma^{-1}(X - m)$, $\Sigma = \mathbb{C}ov(X)$, $m = \mathbb{E}(X)$
- Condition de linéarite (LC) : p.s. $\mathbb{E}(Z|PZ) = PZ$
- Covariance conditionelle constante (CCV) : $\mathbb{V}(Z|PZ)$ constante p.s.

Proposition : Si E est un SDR vérifiant (LC), alors $\mathbb{E}(Z|Y) \in E$ presque sûrement.

$$M_{SIR} = \sum_{h=1}^{H} \mathbb{P}(Y \in I(h)) C_h C_h^T$$

- $C_h = \mathbb{E}(Z|Y \in I(h))$, $(I(h))_{h \in [\![1,H]\!]}$ partition
- Réduction de dimension
- SIR pathology

Proposition : Si E est un SDR qui vérifie (LC) et (CCV), alors $\mathbb{E}(ZZ^T|Y) - I_p \subseteq E$ presque sûrement.

$$M_{CUVE} = \mathbb{E}(W(Y)W(Y)^T)$$

- avec $W(y) = \mathbb{V}(Z\mathbb{1}(Y \leq y)) - F_Y(y)I_p$

- supremum de Y : y^+
- Hypothèse : $\mathbb{P}(Y > y) \xrightarrow[y \to y^+]{} 0$
- Adapatation des SDR : SDR extrême :

$$\frac{\mathbb{E}|\mathbb{P}(Y > y|Z) - \mathbb{P}(Y > y|P_e Z)|}{\mathbb{P}(Y > y)} \xrightarrow[y \to y^+]{} 0$$

Algorithm TIREX 1 $O(nlog(n) + kp^2)$

Require: Choisir $d \leq p$ et $k << n$ $(k \approx n^{2/3})$

1: $\hat{Z}$ = standardisation empirique de X

2: Trier $\hat{Z}$ en fonction de leur réponse Y : $\left(\hat{Z}_{(1)}, Y_{(1)}\right), \left(\hat{Z}_{(2)}, Y_{(2)}\right), \ldots, \left(\hat{Z}_{(n)}, Y_{(n)}\right)$ tels que $Y_{(1)} \geq \cdots \geq Y_{(n)}$.

3: Calculer

$$\widehat{M}_{\mathrm{TIREX1}} = \frac{1}{k^3} \sum_{j=1}^{k} \hat{S}_j \hat{S}_j^\top \; avec \; \hat{S}_j = \sum_{i=1}^{j} \hat{Z}_{(i)}$$

4: Trouver les vecteurs propres associés aux plus grandes valeurs propres $e_i, 1 \leq i \leq d$ de $\hat{M}_{TIREX1}$ avec une PCA

5: Renvoyer $Vect(e_i, i \leq d)$

Algorithm TIREX 2 $O(nlog(n) + kp^4)$

Require: Choisir $d \leq p$ et $k << n$ ($k \approx n^{2/3}$)

1: $\hat{Z}$ = standardisation empirique de X
2: Trier $\hat{Z}$ en fonction de leur réponse Y : $\left(\hat{Z}_{(1)}, Y_{(1)}\right), \left(\hat{Z}_{(2)}, Y_{(2)}\right), \ldots, \left(\hat{Z}_{(n)}, Y_{(n)}\right)$ tels que $Y_{(1)} \geq \cdots \geq Y_{(n)}$.
3: Calculer

$$\widehat{M}_{\mathrm{TIREX2}} = \frac{1}{k^3} \sum_{j=1}^{k} \hat{T}_j \hat{T}_j^\top \; avec \; \hat{T}_j = \sum_{i=1}^{j} \left(\hat{Z}_{(i)} \hat{Z}_{(i)}^T - I_p\right)$$

4: Trouver les vecteurs propres associés aux plus grandes valeurs propres $e_i, 1 \leq i \leq d$ de $\hat{M}_{TIREX1}$ avec une PCA
5: Renvoyer $Vect(e_i, i \leq d)$

Protocole :

- Utiliser une méthode de réduction de dimension
- Entrainer un classifieur sur cet espace de dimension réduit (dans le papier, KKN a été choisi)

Métriques d'évaluation :

- AUC : aire sous la courbe ROC. (classifieur naïf constamment égal à 0 : $AUC = q_\alpha$)
- AMR : $\frac{1}{2}[\mathbb{P}(h(X) = 1 | T = 0) + \mathbb{P}(h(X) = 0 | T = 1)]$ (classifieur naïf constamment égal à 0 : $AMR = 0.5$)

A. Aghbalou, F. Portier, A. Sabourin, C. Zhou/Tail Inverse Regression for Extremes

	TIREX1	TIREX2	CUME	CUVE	PCA	SVD	LLE	IMP
Bank	0.434	**0.378**	0.42	0.392	0.418	0.474	0.486	0.432
Crime	**0.412**	0.5	0.471	0.47	0.502	0.469	0.47	0.5
CompAct	**0.208**	0.279	0.287	0.313	0.242	0.243	0.271	0.253
Residential	**0.158**	0.353	0.421	0.447	0.479	0.479	0.49	0.49
Parkinsons	**0.252**	0.346	0.268	0.346	0.469	0.469	0.455	0.47
Puma32	0.492	0.501	0.5	0.5	0.5	0.5	0.501	**0.49**
Elevators	**0.446**	**0.446**	0.471	0.463	0.5	0.5	0.5	0.5
Ailerons	**0.307**	0.329	0.314	0.33	0.498	0.499	0.498	0.501

TABLE 2
AM risk of the nearest neighbors classifier with reduced covariates obtained with different dimension reduction methods.

	TIREX1	TIREX2	CUME	CUVE	PCA	SVD	LLE	IMP
Bank	**0.771**	0.696	0.698	0.684	0.736	0.689	0.608	0.65
Crime	0.666	0.67	0.616	0.686	0.678	**0.773**	0.672	0.661
CompAct	0.893	0.887	**0.899**	0.871	0.876	0.874	0.868	0.885
Residential	**0.902**	0.827	0.674	0.745	0.667	0.659	0.666	0.694
Parkinsons	**0.901**	0.818	0.852	0.82	0.742	0.753	0.743	0.748
Puma32	**0.711**	0.578	0.616	0.515	0.587	0.577	0.537	0.547
Elevators	0.686	**0.694**	0.615	0.672	0.528	0.537	0.514	0.514
Ailerons	**0.853**	0.834	0.828	0.832	0.502	0.515	0.514	0.525

TABLE 3
AUC of the nearest neighbors classifier with reduced covariates obtained with different dimension reduction methods.

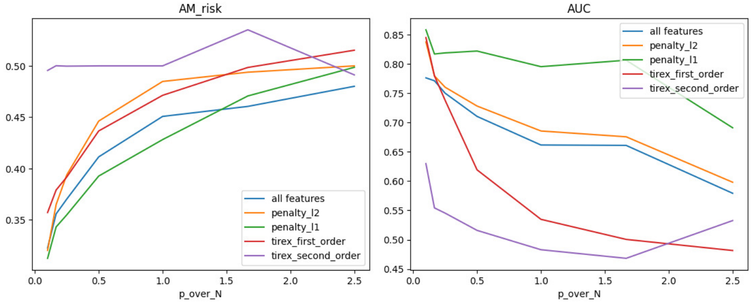
Arithmetic Mean Risk (AMR) et AUC sur des datasets réels avec KKN au bout de la réduction de dimension

Description :

- Features : $X_1, X_2, ..., X_p \sim \mathcal{N}(0, 1)$
- $Y = X_1 + X_2 + \varepsilon$ où $\varepsilon \sim \mathcal{N}(0, 1)$
- Satisfait LC et CCV.
- Labels : $\mathbb{1}(Y > q_\alpha)$

Expériences :

- Comparer les performances de TIREX avec une Régression Logistique non pénalisée au bout (entrainé sur cet espace de features réduit)
- Faire varier le nombre de points de données N
- Faire varier α dans q_α
- Visualiser le SDR extrême estimé

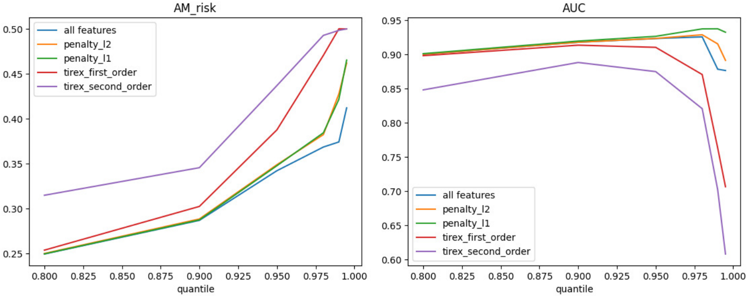


AMR et AUC en fonction de $\frac{p}{N}$ avec $p = 500$, $d = 2$, $\alpha = 0.9$, $200 \leq N \leq 5000$

- SDR extrême de dimension 1 qui est $Vect(1, 1, 0, 0, ..., 0)$
- On peut également voir la Régression Logistique comme un estimateur de SDR de dimension 1

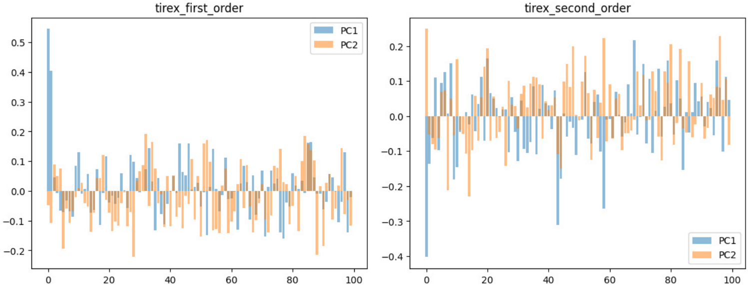
Expérience synthétique

Résultats



AMR et AUC en fonction de q_α avec $N = 5000$, $p = 100$, $d = 2$

- Globalement le même classement qu'auparavant



SDR extrêmes estimés par TIREX1 et TIREX2 avec $N = 10000$, $p = 100$, $d = 2$, $\alpha = 0.9$

- TIREX1 a bien "appris" car le 1^{er} vecteur propre de son SDR est bien proche de $(1, 1, 0, 0, ..., 0)$
- TIREX2 semble avoir un peu appris mais l'estimation est très peu précise