Single-index Extreme-PLS regression

Meryem Bousebata[1,2], Geoffroy Enjolras[2] & Stéphane Girard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.*
[2] *Univ. Grenoble Alpes, CERAG, 38000 Grenoble, France.*
meryem.bousebata@inria.fr, geoffroy.enjolras@grenoble-iae.fr, stephane.girard@inria.fr

Résumé. L'objectif de cette communication est de proposer une nouvelle approche, appelée Single-index Extreme-PLS, pour la réduction de dimension en régression qui soit adaptée aux queues de distributions. Nous nous intéressons à la combinaison linéaire des prédicteurs qui explique au mieux les valeurs extrêmes de la variable réponse dans un contexte de régression inverse non linéaire. La normalité asymptotique de l'estimateur Single-index Extreme-PLS est établie sous des hypothèses modérées. Les performances de la méthode sont évaluées par simulations numériques. Une analyse statistique de données de revenu agricole français, considérant des rendements céréaliers extrêmes, est fournie à titre d'illustration.

Mots-clés. Valeurs extrêmes, Réduction de dimension, Régression inverse non linéaire, Partial Least Squares.

Abstract. The goal of this communication is to propose a new approach, called Single-index Extreme-PLS, for dimension reduction in regression and adapted to distribution tails. The objective is to find a linear combination of predictors that best explain the extreme values of the response variable in a non-linear inverse regression model. The asymptotic normality of the Single-index Extreme-PLS estimator is established under mild assumptions. The performance of the method is assessed on simulated data. A statistical analysis of French farm income data, considering extreme cereal yields, is provided as an illustration.

Keywords. Extreme value, Dimension reduction, Non-linear inverse regression, Partial Least Squares.

1 Introduction

Context. Regression analysis is widely used to study the relationship between a response variable Y and an explanatory p-dimensional vector X starting from an-sample. When p grows, a dimension reduction becomes necessary to show only the most relevant directions of high-dimensional data. There exist a number of statistical models for dimension reduction in regression problems. One of the most popular is Partial Least Squares (PLS) regression, introduced by (Wold, 1975), that combines the characteristics of Principal Component Analysis (PCA) and multiple regression. Its purpose is to find linear combinations of the X coordinates highly correlated with Y. Sliced Inverse Regression

(SIR) is an alternative method for dimension reduction in regression which explores the simplicity of the inverse regression view of X against Y (Li, 1991). It aims at replacing X by its projection onto a subspace of smaller dimension without loss of information. At the same time, there is a growing interest for the modelling of conditional extremes, *i.e.* extremes depending on a covariate. One can mention for instance the estimation of conditional extreme quantiles or more generally, the tail of conditional distributions (Gardes & Girard, 2010). In this communication, we aim to deal with these two lines of works (dimension reduction in regression and conditional extremes) by looking for a linear combination $\beta^t X$ of the covariates that best explains the extreme values of Y. More precisely, we propose a single-index approach to find a direction $\hat{\beta}$ maximizing the covariance between $\beta^t X$ and Y given Y exceeds a high threshold y. This adaptation of the PLS estimator to the extreme-value framework, referred to as Single-index extreme-PLS (SIEPLS), is achieved in the context of a non-linear inverse regression model. In practice, $\hat{\beta}$ allows to quantify the effect of the covariates on the extreme values of Y in a simple and interpretable way. Plotting Y against the projection $\hat{\beta}^t X$ also provides a visual interpretation of conditional extremes. Moreover, working on the pair $(\hat{\beta}^t X, Y)$ should yield improved results for most estimators dealing with conditional extreme values thanks to the dimension reduction achieved thanks to the projection step. From the theoretical point of view, the asymptotic normality of $\hat{\beta}$ is established without linearity or independence requirements.

An inverse model. Let us first consider the following single-index non linear inverse regression model:

(M) $X = g(Y)\beta + \varepsilon$, where X is a p-dimensional random vector, Y is a real random variable, $g : \mathbb{R} \to \mathbb{R}$ is the link function and ε is p-dimensional random vector of error. The parameter $\beta \in \mathbb{R}^p$ is an unknown unit vector, ε may depend on Y and g is an unknown function.

Similar inverse regression models were used to establish the theoretical properties of SIR (Bernard-Michel, Gardes & Girard, 2008). Under model (**M**), we aim at estimating β by maximizing the covariance between $\beta^t X$ and Y conditionally on large values of Y. Indeed, roughly speaking, when Y is large, provided the distribution tail of ε is negligible, one has $X \simeq g(Y)\beta$ leading to the approximate single-index model $Y \simeq g^{-1}(\beta^t X)$. Note that the considered model does not require a linear conditional mean or a conditional independence assumption. The paper is organized as follows. In Section 2, the SIEPLS approach is introduced in the framework of a single-index model and heavy-tailed distributions. Some preliminary properties are stated in order to justify the above heuristics from a theoretical point of view. The associated estimator is exhibited in Section 3 and its asymptotic distribution is established under mild assumptions. In Section 4, the performances of the method are investigated through a simulation study and is applied to assess the influence of various parameters on cereal yields collected on French farms.

2 SIEPLS approach

Let us denote by $w(y)$ the unit vector maximizing the covariance between w^tX and Y given that Y exceeds a large threshold y:

$$w(y) = \arg\max_{\|w\|=1} \operatorname{cov}(w^tX, Y|Y \geq y). \tag{1}$$

This optimization problem benefits from a closed-form solution given in the next proposition and obtained by solving the constrained optimization problem using Lagrange multipliers. For all $y \in \mathbb{R}$, let us denote by $\bar{F}(y) = \mathbb{P}(Y \geq y)$ the survival function of Y and the tail-moments, whenever they exist, $m_Y(y) = \mathbb{E}(Y\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}$, $m_X(y) = \mathbb{E}(X\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}^p$, $m_{XY}(y) = \mathbb{E}(XY\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}^p$.

Proposition 1. *Suppose that* $\mathbb{E}\|X\| < \infty$, $\mathbb{E}|Y| < \infty$ *and* $\mathbb{E}\|XY\| < \infty$. *Then, the solution of the optimization problem (1) is:*

$$w(y) = v(y)/\|v(y)\| \text{ where } v(y) = \bar{F}(y)m_{XY}(y) - m_X(y)m_Y(y). \tag{2}$$

Let us note that the solution (2) is invariant with respect to the scaling and location of X. Besides, when ε is centered and independent of Y, we recover the classical PLS framework and it is easily shown that $w(y) = \pm\beta$ for all $y \in \mathbb{R}$. In the following, no assumption is made on the (in)dependence between Y and ε, but additional assumptions on the link function g and the distribution tail of Y are considered:

($\mathbf{A}_1$) Y is a random variable with density function f regularly varying at infinity with index $-\frac{1}{\gamma} - 1$, $\gamma \in (0, 1)$ *i.e.* for all $t > 0$,

$$\lim_{y\to\infty} \frac{f(ty)}{f(y)} = t^{-\frac{1}{\gamma}-1}.$$

This property is denoted for short by $f \in RV_{-1/\gamma-1}$.

($\mathbf{A}_2$) $g \in RV_c$ with $c > 0$.

($\mathbf{A}_3$) There exists $q > 1/(\gamma c)$ such that $\mathbb{E}(\|\varepsilon\|^q) < \infty$.

Let us note that ($\mathbf{A}_1$) implies that $\bar{F} \in RV_{-1/\gamma}$ which is equivalent to assuming that the distribution of Y is in the Fréchet maximum domain of attraction, with extreme-value index $\gamma > 0$, see de Haan & Ferreira (2007). In other words, ($\mathbf{A}_1$) entails that Y has a right heavy-tail. The restriction to $\gamma < 1$ ensures that $\mathbb{E}|Y|$ exists. Assumption ($\mathbf{A}_2$) means that the link function asymptotically behaves like a power function. Finally, ($\mathbf{A}_3$) is a technical assumption which is satisfied for instance by Gaussian distributions.

In order to assess the convergence of $w(y)$ to β as $y \to \infty$, the squared cosine of the angle between the above unit vectors is defined as: $\cos^2(w(y), \beta) = (w(y)^t\beta)^2$. A value close to 0 implies a weak proximity ($w(y)$ is almost orthogonal to β) while a value close to 1 means a high colinearity.

Proposition 2. *Assume* (**M**), ($\mathbf{A}_1$), ($\mathbf{A}_2$) *and* ($\mathbf{A}_3$) *hold with* $\gamma(c+1) < 1$. *Then,*

$$\cos^2(w(y), \beta) = 1 - O\left\{ \left(\frac{1}{g(y)\bar{F}^{1/q}(y)} \right)^2 \right\} \to 0,$$

as $y \to \infty$.

In view of assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_2$), the function $y \mapsto g(y)\bar{F}^{1/q}(y)$ is regularly varying with index $c - 1/(q\gamma) > 0$ from ($\mathbf{A}_3$). Unsurprisingly, the above convergence rates are large when c is large (*i.e.* the link function is rapidly increasing), q is large (*i.e.* the noise ε is small) or/and γ is large (*i.e.* the tail of Y is heavy). The estimation of $w(y)$ from data distributed from model (**M**) is addressed in the following section.

3 SIEPLS: Population version

Let (X_i, Y_i), $1 \leq i \leq n$ be independent and identically distributed random variables from model (**M**) and let $y_n \to \infty$ as the sample size n tends to infinity. The solution (2) is estimated by its empirical counterpart introducing

$$\hat{v}(y_n) = \hat{\bar{F}}(y_n)\hat{m}_{XY}(y_n) - \hat{m}_X(y_n)\hat{m}_Y(y_n),$$

with $\hat{\bar{F}}$ the empirical survival function and

$$\hat{m}_{XY}(y_n) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \mathbb{1}_{\{Y_i \geq y_n\}}, \hat{m}_Y(y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\{Y_i \geq y_n\}}, \hat{m}_X(y_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}_{\{Y_i \geq y_n\}}.$$

Our main result is the following:

Theorem 1. *Assume* (**M**), ($\mathbf{A}_1$), ($\mathbf{A}_2$) *and* ($\mathbf{A}_3$) *hold with* $2\gamma(c+1) < 1$. *Let* $y_n \to \infty$ *such that* $n\bar{F}(y_n) \to \infty$ *and* $n\bar{F}(y_n)^{1-2/q}/g^2(y_n) \to 0$ *as* $n \to \infty$. *Then,*

$$\sqrt{n\bar{F}(y_n)} \left(\frac{\hat{v}(y_n)}{\|v(y_n)\|} - \beta \right) \xrightarrow{d} \xi\beta,$$

with $\xi \sim \mathcal{N}(0, \lambda(c,\gamma))$ *and where* $\lambda(c,\gamma)$ *is a constant.*

Assumption $n\bar{F}(y_n) \to \infty$ ensures that the variance of the estimator tends to zero while condition $n\bar{F}(y_n)^{1-2/q}/g^2(y_n) \to 0$ entails that the bias (bounded above by $1/(g(y_n)\bar{F}^{1/q}(y_n))$, see Proposition 2) is asymptotically small compared to the standard deviation $1/\sqrt{n\bar{F}(y_n)}$. Finally, Theorem 1 shows that the estimated direction $\hat{v}(y_n)$ is asymptotically aligned with the true direction β.

4 Numerical results

4.1 Simulated data

We consider a sample of size $n = 1000$ and dimension $p \in \{3, 30\}$ from model (M) with a link function $g(t) = t^c$, $t > 0$, $c \in \{1/4, 1/2, 1, 3/2, 2\}$. The results (available in Bousebata, Enjolras & Girard (2021)) are not reported here for lack of space reasons, they will provided during the presentation.

4.2 Real data

Our approach is applied to data extracted from the Farm Accountancy Data Network (FADN), an annual database of commercial-sized farm holdings. This dataset of $n = 949$ observations contains significant accounting and financial information about French professional farm incomes in 2014. Our goal is to investigate the impact of various factors on farm yields (expressed in quintals per hectare). The response variable Y is the inverse of the wheat yield (in quintals/hectare), as we are interested in the analysis of low yields, and the covariate X includes 12 continuous variables: selling prices (euro/quintal), pesticides, fertilizers, crop insurance purchased, insurance claims, farm subsidies, seeds and seedlings costs, works and services purchase for crops, other insurance premiums, farm income taxes, farmer's personal social security cost (euro/hectare) and temperature average (degree Celsius). A number of visual checks of whether the heavy-tailed assumption makes sense for these data have been implemented (Hill plot and quantile-quantile plot). The estimator SIEPLS $\hat{v}(y_n)$ is computed for each $y_n = Y_{n-k+1,n}$. For the sake of interpretation, we define the conditional correlation between the projected covariate $\hat{v}(y_n)^t X$ and each coordinate $X^{(j)}$ of the covariate as:

$$\rho(X^t\hat{v}(y_n), X^{(j)}|Y \geq y_n) = \frac{\mathrm{cov}(X^t\hat{v}(y_n), X^{(j)}|Y \geq y_n)}{\sigma(X^t\hat{v}(y_n)|Y \geq y_n)\sigma(X^{(j)}|Y \geq y_n)}.$$

Results are depicted on Figure 1 for the 12 considered covariates. Note that the 150 largest inverse wheat yields are mainly consequences of operational costs (fertilisers, pesticides, seeds and seedlings), structural costs (claims, purchase of an insurance policy, farm subsidies, social security cost) and supplementary costs (works and services purchase). This result may be explained by the fact that, in 2014, yields were strongly impacted by production costs, despite mild winter temperatures. Finally, two estimators (linear and non-linear) of the conditional mean $\mathbb{E}(\hat{v}(y_n)^t X|Y)$ have been computed. A positive trend appears for large values of Y in accordance to the inverse regression model (**M**).

Acknowledgements. This work is supported by the French National Research Agency (ANR) in the framework of the Investissements d'Avenir Program (ANR-15-IDEX-02).

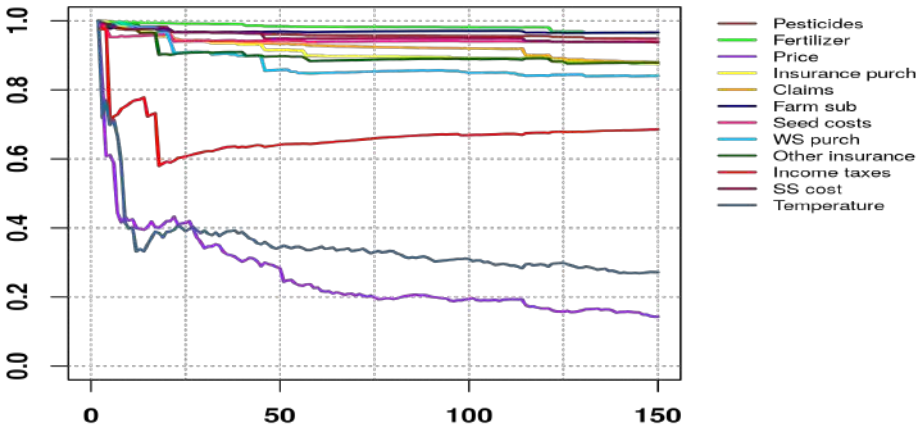

Figure 1: Graph of the estimated conditional correlation function $y \mapsto \rho(X^t\hat{v}(y), X^{(j)}|Y \geq y)$ for $j = 1, \ldots, 12$ (horizontally: number of exceedances k, vertically: conditional correlation estimated by its empirical counterpart using the threshold $y = Y_{n-k+1,n}$).

References

Bernard-Michel, C., Gardes, L., & Girard, S. (2008). A note on sliced inverse regression with regularizations. *Biometrics*, 64(3), 982–984.

Bingham, N. H., Goldie, C. M., & Teugels, J. L. (1989). Regular variation (Vol. 27). *Cambridge university press*.

Bousebata, M., Enjolras, G., & Girard, S. (2021). Extreme Partial Least-Squares regression, *Submitted*, `https://hal.inria.fr/hal-03165399`

de Haan, L., & Ferreira, A. (2007). Extreme value theory: An introduction. *Springer Science & Business Media*.

Gardes, L., & Girard, S. (2010). Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2), 177–204.

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

Nelsen, R. B. (2007). An introduction to copulas. *Springer Science & Business Media*.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, 12(S1), 117–142.