Tail Inverse Regression: dimension reduction for prediction of extremes

Anass Aghbalou, François Portier, Anne Sabourin, Chen Zhou

Tail Inverse Regression: dimension reduction for prediction of extremes

Anass Aghbalou[1] **François Portier**[2] **Anne Sabourin**[3] **Chen Zhou**[4]

[1] *LTCI, Télécom Paris, Institut polytechnique de Paris, France., e-mail:* anass.aghbalou@telecom-paris.fr

[2] *Ensai, CREST - UMR 9194, Rennes, France, e-mail:* francois.portier@gmail.com

[3] *Université Paris Cité, CNRS, MAP5, F-75006 Paris, France, e-mail:* anne.sabourin@u-paris.fr

[4] *Erasmus University Rotterdam, Rotterdam, Netherlands; Tinbergen Institute, Rotterdam, Netherlands, e-mail:* zhou@ese.eur.nl

Abstract: We consider the problem of supervised dimension reduction with a particular focus on extreme values of the target $Y \in \mathbb{R}$ to be explained by a covariate vector $X \in \mathbb{R}^p$. The general purpose is to define and estimate a projection on a lower dimensional subspace of the covariate space which is sufficient for predicting exceedances of the target above high thresholds. We propose an original definition of Tail Conditional Independence which matches this purpose. Inspired by Sliced Inverse Regression (SIR) methods, we develop a novel framework (TIREX, Tail Inverse Regression for EXtreme response) in order to estimate an extreme sufficient dimension reduction (SDR) space of potentially smaller dimension than that of a classical SDR space.We prove the weak convergence of tail empirical processes involved in the estimation procedure and we illustrate the relevance of the proposed approach on simulated and real world data.

MSC2020 subject classifications: Primary 62G32, 62H25; secondary 62G08, 62G30.
Keywords and phrases: Dimension reduction, Empirical processes, Extreme events, Inverse regression, Supervised learning.

1. Introduction

Dimension reduction is a crucial matter in supervised learning problems where the goal is to predict a *dependent variable* $Y \in \mathbb{R}$ or summaries of it, when the dimension p of the *covariate vector* $X \in \mathbb{R}^p$ is large. In this paper we consider dimension reduction for prediction of tail events, by which we mean events of the kind $\{Y > y\}$, for arbitrarily large values of y. This stylized statistical problem relates to a wide range of practical applications such as supervised anomaly detection, system monitoring with a large number of sensors, prediction of extreme weather conditions or financial risk management. For instance, in financial risk management, a typical concern is to identify risk factors, which will be further used to explain extreme events such as financial market crashes, see *e.g.* Fama and French (1993, 2015). Risk factors are often lower dimensional functionals based on a large number of stock returns. Identifying such risk factors that can predict financial market crashes is therefore an example of dimension reduction for the problem of predicting tail events.

Our focus on extreme values connects our work with the field of Extreme Value Theory (EVT) which has been successfully applied to model tail events with potentially catastrophic impact. Statistical inference in this framework is performed using the most extreme realizations of the random variable under consideration. We refer the interested reader to the monographs Beirlant et al. (2006); De Haan and Ferreira (2007); Resnick (2013, 2007). Notice that the curse of dimensionality is particularly troublesome in extreme value analysis where only a small fraction of the data, reflected by the low probability $\mathbb{P}(Y > y)$, is used for inference. Before proceeding further we remark that the method proposed in this study,

although motivated by and formulated in an EVT framework, does not rely on the minimal assumptions typically required in EVT such as a power law decay. It is in fact a local method related to any small range of Y and as such, it could be easily adapted to tackle the problem of dimension reduction for prediction of Y within low probability regions of other shapes. However in view of the importance of applications towards risk management, we concentrate on this specific tail region.

Dimension reduction in EVT. The subject of dimension reduction for extremes has inspired numerous recent works. The vast majority of them are devoted to the unsupervised setting, *i.e.* analyzing the extremes of a high dimensional random vector. Such studies can be divided into the following categories: clustering methods (Chautru (2015); Chiapino et al. (2020); Janßen and Wan (2020)), support identification, (Goix, Sabourin and Clémençon (2016); Goix, Sabourin and Clémençon (2017); Chiapino and Sabourin (2016); Chiapino, Sabourin and Segers (2019); Simpson, Wadsworth and Tawn (2020); Meyer and Wintenberger (2021)), Principal Component Analysis of the angular component of extremes (Cooley and Thibaud (2019); Jiang, Cooley and Wehner (2020); Drees and Sabourin (2021)), and graphical models for extremes (Hitz and Evans (2016); Engelke and Hitz (2020); Asenova, Mazo and Segers (2021)); see also Engelke and Ivanovs (2021) and the references therein.

By contrast, our approach takes place in the supervised setting. Our main informal assumption is that *a low dimensional orthogonal projection PX is sufficient for predicting extreme values of Y*. In other words the extreme values of Y can be entirely explained by a limited number of linear combinations of the components of X. In this setting, the only existing works are, to our best knowledge, Gardes (2018) and Bousebata, Enjolras and Girard (2023). In Gardes (2018), the informal assumption emphasized above is made precise by a specific notion of *tail conditional independence*, reported in Equation (3.2) below. Dimension reduction is considered under this condition. Gardes (2018) demonstrates the usefulness of such a reduction for statistical estimation of large conditional quantiles. Even though we follow in the footsteps of Gardes (2018) in terms of informal goal, our framework differs significantly from Gardes (2018)'s on several key aspects. First, the specific definition of tail conditional independence that we propose (See Definition 2 in Section 3) is not equivalent to Gardes (2018)'s condition (3.2). We carry out an in-depth comparison of both conditions and we show that neither one of them implies the other, in Section B from the supplementary material. Second, our assumption is motivated by a downstream task (predicting the occurrence of a tail event) which is different from, although related to the one motivating Gardes (2018) (estimation of extreme conditional quantiles). Third, the statistical guarantees brought by Gardes (2018) are obtained under the assumption that the dimension reduction space is already known. In the cited reference an estimation method is indeed proposed for the dimension reduction space, however its statistical properties are only analyzed *via* simulations. Instead, we bring statistical guarantees regarding the estimation of a sufficient projection subspace itself. We discuss qualitatively the positive impact it may have for prediction of tail events in Remark 1. Lastly, the computational cost of TIREX depends only polynomially on the ambient dimension p, which is not the case with the current estimation method in Gardes (2018), as discussed in Section 6.

Another study related to our work is the recently published paper Bousebata, Enjolras and Girard (2023), where the authors adopt a partial least square strategy to uncover the relation between linear combinations of covariates and the extreme values of the target. Their model assumptions differ from ours substantially: the inverse regression model assumed in Bousebata, Enjolras and Girard (2023) implies a single-index relationship between extreme values of the response and the covariates. In addition, the model requires regular variation of

the dependent variable Y and of the link function. Lastly, the model relies on finite variance of Y. In contrast, our approach is somewhat 'free' from most restrictions on the distribution of (X, Y) except from the well-known linearity condition and constant variance condition, typically needed for SIR. Such conditions concern only the distribution of the covariates. Since we do not impose regular variation, we can handle not only thin-tailed but also extremely heavy-tailed dependent variables with no finite variance or even mean.

Sufficient Dimension Reduction and inverse methods. The underlying assumption of a sufficient linear projection subspace has been formalized under the notion of Sufficient Dimension Reduction (SDR) space (Cook (2009)). Many classical approaches to supervised dimension reduction rely on a linear regression model between X and Y. This is the case *e.g.* for Principal component regression (Hotelling (1957)), Partial least squares (Wold (1966)), Canonical correlation analysis (Thompson (1984)) or penalized methods with sparsity inducing regularization such as the Lasso (Jenatton, Audibert and Bach (2011)). Differently, *SDR* builds upon a *linear dimension reduction* assumption: only a small number of *linear* combinations of covariates is useful for predicting the dependent variable. In other words, there exists a linear subspace E (a SDR) of a moderate dimension $d \leq p$, such that

$$\mathbb{P}\left(Y \leq t \mid X\right) = \mathbb{P}(Y \leq t | PX), \qquad \forall t \in \mathbb{R}, \quad \text{almost surely}, \tag{1.1}$$

where P is the orthogonal projector on E, *i.e.* Y depends on X only through $PX \in \mathbb{R}^d$. This framework relies heavily on the notion of conditional independence Dawid (1979); Constantinou and Dawid (2017): Condition (1.1) characterizes the fact that Y is conditionally independent from X given PX. One major advantage of this approach is that it strikes a balance between interpretability of the dimension reduction based on linear operations and flexibility of the generative model – no assumption is made regarding the dependence structure between PX and Y.

Under the assumption that there exists a non trivial subspace E such that (1.1) holds, a natural idea is to estimate such a subspace first, and then use only the variable PX to predict Y, thus reducing the dimensionality of the regression problem. The estimation problem based on SDR can also be viewed as a specific case of semi-parametric M-estimation (Delecroix, Hristache and Patilea (2006)). Alternatively, one may consider derivative based methods, relying on the fact that the gradient of the regression curve belongs to E (Härdle and Stoker (1989); Hristache et al. (2001); Xia et al. (2007); Dalalyan, Juditsky and Spokoiny (2008)). Recently, the framework of Reproducing Kernel Hilbert Spaces (RKHS) has been employed to estimate SDR spaces by means of covariance operators (Fukumizu, Bach and Jordan (2004); Fukumizu et al. (2009)).

The family of methods to which our work relates most is the inverse regression paradigm initiated by Li (1991), including the Sliced Inverse Regression (SIR) strategy and its second order variant Sliced Average Variance Estimate (SAVE) (Cook and Weisberg (1991)). The main idea underlying these methods is that under appropriate assumptions the inverse regression curve $\mathbb{E}\left[X|Y\right]$ and its second moment variant – the columns of the conditional covariance matrix $\mathbb{V}\mathrm{ar}\left[X|Y\right]$ – almost surely belong to the minimal SDR. Cumulative slicing estimation (CUME), proposed in Zhu, Zhu and Feng (2010) and further analyzed in Portier (2016), aims at recovering the largest possible subspace of the minimal SDR. It is achieved by estimating the conditional expectation and variance of X, conditioning on 'slices' of the target Y, in the form of $\mathbb{1}\{Y < y\}$, and then aggregating such conditional expectations and variances by integration with respect to y.

A well-known restriction of the SIR strategy is that it relies on a so-called *linearity condition* (LC) regarding the covariates, namely equation (2.1) in the next section, see Hall and

Li (1993) for a justification. The required condition is satisfied in particular if the covariates form an elliptical random vector or are independent (Cook (2009); Eaton (1986)). There are various extensions of SIR permitting to overcome this restriction. Using RKHS, it has been proposed to transform the data in a way that LC is approximately satisfied (Wu (2008); Yeh, Huang and Lee (2008)). Another possibility allowing to depart from elliptical covariates is to apply the SIR methodology and its higher order variants to score functions of the explanatory variables (Babichev et al. (2018)). Finally, the high dimensional case $p > n$ calls for regularization methods which permit in addition to perform feature selection (Li and Yin (2008)). All these extensions are out of the scope of the present paper, in which we restrict ourselves to the original SIR and SAVE methods, thus leaving room for several improvement in further works. For estimation purposes we consider a variant of CUME.

Contributions and outline. Our contributions are twofold. First, we develop in Section 3 a modified version of Gardes (2018)'s probabilistic setting regarding tail conditional independence. In particular we explain in Remark 1 the relevance of our definition for the purpose of predicting tail events and its connections to the statistical learning framework of imbalanced classification. We discuss thoroughly the distinctions between the two alternative definitions for tail conditional independence in Section B where we also provide examples of models satisfying one or the other. Second, we show in Section 4 that our definition permits to extend inverse regression principles and methods to this extreme values setting (theorems 1, 2). We derive an asymptotic analysis for our proposed estimation strategy TIREX stemming from inverse regression, using specific tools from the theory of empirical processes (Section 5). We illustrate the finite sample performance of TIREX with simulated and real world data sets in Section 6, in particular we demonstrate empirically the usefulness of TIREX for tail events prediction. The code developed for TIREX is available online [1] and some technical proofs and additional comments are deferred to the supplementary material.

We start-off in Section 2 by recalling the necessary background regarding conditional independence of random variables, SDR spaces, and inverse regression.

2. Background: dimension reduction space and Sliced Inverse Regression

Conditional independence of random variables Y and V given W is defined *e.g.* in Constantinou and Dawid (2017) as follows: the conditional distribution of Y given (V, W) is the same as the conditional distribution of Y given W, almost surely. Several characterizations are recalled below, the equivalence of which are proved in Constantinou and Dawid (2017), Proposition 2.3.

Definition 1 (conditional independence). Let Y, V, W be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in arbitrary measure spaces. The variables Y and V are called conditionally independent given W, a property denoted by $Y \perp\!\!\!\perp V \mid W$, if the equivalent conditions below are satisfied.

(CI-1) For all $A_Y \in \sigma(Y)$, $\mathbb{P}(Y \in A \mid V, W) = \mathbb{P}(Y \in A \mid W)$, almost surely.

(CI-2) For all real-valued functions f and g, measurable and bounded,

$$\mathbb{E}[f(Y)g(V) \mid W] = \mathbb{E}[f(Y) \mid W]\,\mathbb{E}[g(V) \mid W], \quad \text{a.s.}$$

(CI-3) For any real-valued function g, measurable and bounded,

$$\mathbb{E}[g(V) \mid Y, W] = \mathbb{E}[g(V) \mid W], \quad \text{a.s.}$$

[1] https://github.com/anassag/TIREX

Notice that the existence of regular versions of conditional probability distributions is not required in Definition 1. However in this paper, Y is real valued, thus the existence of such a regular version for the conditional distribution of Y given (V, W) is granted. As a consequence we may write, without additional precautions, expressions of the kind '$\mathbb{P}(Y \in A \mid V = v, W = w)$'. The latter quantity is defined as the value of the conditional probability kernel at point $((v, w), A)$.

In the context of supervised dimension reduction, we consider $V = X$ and search for a projection $W = PX$ of X on a lower dimensional subspace E satisfying the above conditions. We assume for simplicity that the covariance matrix $\Sigma = \mathbb{C}\mathrm{ov}[X]$ is invertible and for ease of presentation we introduce a standardized covariate vector $Z = \Sigma^{-1/2}(X - m)$ where $m = \mathbb{E}[X]$. We consider in the remaining of this paper the problem of regressing Y on Z, which amounts to assuming that both m and Σ are known, so that the vector Z is observed. Thus $\mathbb{V}\mathrm{ar}[Z] = I_p$ and $\mathbb{E}[Z] = 0$. A SDR space (Cook (2009); Cook and Ni (2005)) is a subspace E of $\mathbb{R}^p$ such that $Y \perp\!\!\!\perp Z \mid PZ$ where P is the orthogonal projection on E, which is equivalent to condition (1.1) in the introduction. Our results easily extend to general covariates X (see *e.g.* Cook and Weisberg (1991)) at the price of an additional notational burden, see Section E in the supplementary material. Notice already that in terms of non-standardized covariates X, a subspace $\tilde{E}$ of $\mathbb{R}^p$ with associated orthogonal projector $\tilde{P}$ is a SDR space for the pair (X, Y) if and only if $\tilde{E} = \Sigma^{-1/2}E$ where E is a SDR space for Z.

A central space is a SDR subspace E_c for the pair (Z, Y) of minimal dimension. In our context of finite dimensional covariates a central space always exists since the ambient space $\mathbb{R}^p$ itself is a SDR space. Uniqueness is not guaranteed in general but holds true under mild assumptions ensuring that an intersection of SDR spaces is a SDR space (see *e.g.* Portier and Delyon (2013), Theorem 1). In such a case one may refer without ambiguity to *the* central space.

First and second order inverse methods, respectively named SIR (Li (1991)) and SAVE (Cook and Weisberg (1991)) are two of many methods to estimate SDR spaces. Both rely on the fact that under appropriate assumptions detailed below, first and second moments of the covariate vector, conditioning upon the target, belong to a SDR space. In the sequel, E is a SDR space, and P denotes the orthogonal projection on E. Then $Q = I_p - P$ is the orthogonal projection on $E^{\perp}$, the orthogonal complement of E. The required conditions are the Linearity Condition (LC):

$$\mathbb{E}[Z \mid PZ] = PZ \quad \text{a.s.} \tag{2.1}$$

and the additional Constant Conditional Variance (CCV),

$$\mathbb{V}\mathrm{ar}[Z|PZ] \text{ is constant} \quad \text{a.s.} \tag{2.2}$$

Under both LC and CCV, we have that $\mathbb{E}[\mathbb{V}\mathrm{ar}[Z|PZ]] = \mathbb{E}\left[ZZ^T\right] - \mathbb{E}\left[PZ(PZ)^T\right] = I_p - P$ and therefore the constant matrix in (2.2) is necessarily the projection $Q = I_p - P$ on the orthogonal complement of E.

Notice that LC and CCV depend on an unknown SDR space. Assuming that LC holds for all orthogonal projectors is in fact equivalent to assuming that the covariate vector Z is spherically symmetric, *i.e.* $Z = \rho U$ where $\rho \perp\!\!\!\perp U$, ρ is a non negative random variable and U is uniformly distributed over the unit sphere of $\mathbb{R}^p$, as proved in Eaton (1986). Among spherical variables with finite second moment, CCV is equivalent to being Gaussian ((Bryc, 2012, Theorem 4.1.4)).

The following proposition in Li (1991) encapsulates the main idea of SIR. We give below the (classical) proof for the sake of completeness.

Proposition 1 (SIR principle)**.** *If E is an SDR space for which LC (2.1) is satisfied, then $Q(\mathbb{E}[Z|Y]) = 0$ a.s., that is, $\mathbb{E}[Z|Y] \in E$ a.s.*

Proof. By the tower rule from conditional expectation,

$$\begin{aligned}\mathbb{E}[Z \mid Y] &= \mathbb{E}\left[\mathbb{E}\big(Z \,|\, Y, PZ\big) \,\middle|\, Y\right] = \mathbb{E}\left[\mathbb{E}(Z \,|\, PA) \mid Y)\right] \\ &= \mathbb{E}[PZ \mid Y] = P\,\mathbb{E}[Z \mid Y]\end{aligned}$$

where the second equality comes from conditional independence and the third one follows from the linearity condition (2.1). Thus $Q\mathbb{E}[Z|Y] = 0$. □

The SIR method advocated first by Li (1991) consists in estimating first conditional expectations $C_h = \mathbb{E}[Z \mid Y \in I(h)]$, $h = 1, \ldots, H$, where $I(h), h = 1, \ldots, H$ are called slices and form a partition of the sample range of Y (or the support of the density function if Y is continuous). From Proposition 1, those estimates lie in the vicinity of the SDR space. Next, performing a Principal Component Analysis (PCA) on the C_h's, one obtains a good approximation of E. More precisely, the SIR estimate of E is given by the largest eigenvectors associated to the SIR matrix,

$$M_{\text{SIR}} = \sum_{h=1}^{H} p_h^{-1} C_h C_h^T,$$

where $p_h = \mathbb{P}(Y \in I(h))$; see Li (1991). Various estimation procedures of SDR spaces are proposed in Cook and Ni (2005); Zhu, Zhu and Feng (2010). In the latter reference, the matrix

$$M_{\text{CUME}} = \mathbb{E}\big[m(Y)m(Y)^T\big], \tag{2.3}$$

with $m(y) = \mathbb{E}[Z\mathbb{1}\{Y \le y\}]$, is introduced as an alternative to the SIR matrix. One advantage of this approach is that the slicing parameter h is no longer needed. In addition the estimate of the matrix M_{CUME} benefits from the aggregating effect of the expectation sign which is typically associated with variance reduction.

A pitfall of SIR is that it is not guaranteed that the C_h's span the entire space E, so that SIR may be inconsistent. This may happen in particular when the regression function $\mathbb{E}[Y|Z]$ admits some symmetry properties (Li, 1991, Remark 4.5), a phenomenon referred to as the SIR pathology. In this case, Li (1991) and Cook and Weisberg (1991) recommend to use higher order moments such as the conditional variance of Z given Y to obtain a second order matrix with wider range. This second order method requires that CCV (2.2) is satisfied in addition to LC, in which case the following result holds. Here and throughout, $\text{span}(M)$ stands for the column space of matrix M.

Proposition 2 (SAVE principle)**.** *If E is an SDR space for which LC (2.1) and CCV (2.2) are satisfied, then*

$$Q\Big(\mathbb{E}\big[ZZ^\top \,|\, Y\big] - I_p\Big) = 0 \qquad a.s.,$$

in other words $\text{span}\left(\mathbb{E}[ZZ^\top \,|\, Y] - I_p\right) \subset E \quad a.s.$

Proof. We reformulate here the arguments of Cook and Weisberg (1991) in our notational framework for convenience. An immediate consequence of assumptions (2.1) and (2.2) is that $\mathbb{E}\left[ZZ^T|PZ\right] = Q + PZZ^TP$. From a conditioning argument and the conditional independence assumption, $\mathbb{E}\left[ZZ^T|Y\right] = Q + P\mathbb{E}\left[ZZ^T|Y\right]P$. Rearranging gives $\mathbb{E}\left[ZZ^T|Y\right] - I_p = P(\mathbb{E}\left[ZZ^T|Y\right] - I_p)P$, thus $Q(\mathbb{E}\left[ZZ^T|Y\right] - I_p) = 0$.

□

Notice that propositions 1 and 2 together imply that $Q(\mathbb{V}\mathrm{ar}\,[Z \mid Y] - I_p) = 0$. Finally for estimation purpose the extension of the CUME method to the second order framework is termed CUVE (cumulative variance estimation) by Zhu, Zhu and Feng (2010). In the case of standardized covariates, it consists in estimating the matrix $M_{\mathrm{CUVE}} = \mathbb{E}\big[W(Y)W(Y)^\top\big]$, where $W(y) = \mathbb{V}\mathrm{ar}\,[Z\mathbb{1}\{Y \le y\}] - F_Y(y)I_p$ is a second order moments matrix which column space is included in $\tilde{E}$. The latter fact is obtained by a slight modification of the argument leading to the SAVE principle.

3. Tail conditional independence, Extreme SDR space

3.1. Definition for Tail Conditional Independence

The focus on the largest values of the target variable Y suggests to weaken the classical definition of conditional independence, so that the equivalent conditions (CI-1)-(CI-3) hold only for Y exceeding a high threshold tending to its right endpoint. Namely, in a similar (but different) manner as in Gardes (2018) we define tail conditional independence as a variant of condition (CI-1) from Definition 1. In the sequel the right endpoint (*i.e.* the supremum) of the support of the random variable Y is denoted by y^+. The limits as $y \to y^+$ as understood as the limits as $y \to y^+, y < y^+$. We assume that $\mathbb{P}(Y > y) \to 0$ as $y \to y^+$, in particular we exclude the case of point masses at y^+.

Definition 2 (Tail Conditional Independence (TCI)). Let Y, V, W be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that Y is real valued, Borel measurable, while V and W take their values in arbitrary measure spaces. We say that Y is *tail conditionally independent from V given W* and write $Y {}_\infty\!\perp\!\!\!\perp V|W$, if

$$\frac{\mathbb{E}\left|\,\mathbb{P}(Y > y \mid V, W) - \mathbb{P}(Y > y \mid W)\,\right|}{\mathbb{P}(Y > y)} \xrightarrow[y \to y^+]{} 0. \tag{3.1}$$

Contrary to conditional independence, tail conditional independence is not symmetric: $Y {}_\infty\!\perp\!\!\!\perp V \mid W$ does not imply that $V {}_\infty\!\perp\!\!\!\perp Y \mid W$.

In Gardes (2018)'s work, tail conditional independence is defined in a somewhat more technical manner, see Definition 1 from the cited reference. However a necessary condition (see Equation (2) in that paper) is the almost sure convergence of the $\sigma(V, W)$-measurable ratio,

$$\frac{\mathbb{P}(Y > y \mid V, W) - \mathbb{P}(Y > y \mid W)}{\mathbb{P}(Y > y \mid W)} \xrightarrow[y \to y^+]{} 0, \quad \text{a.s.} \tag{3.2}$$

In the sequel we refer to our notion of tail conditional independence defined in (3.1) as TCI, while we write TCI-G to refer to L. Gardes' condition (3.2). Both definitions are motivated by similar but different downstream tasks, namely prediction of extreme values for TCI in connection to the AM risk criterion (see Remark 1 below), versus estimation of large conditional quantiles (see Section 3.1 in Gardes (2018)).

In Subsection 3.2 below we work out a generic example where TCI holds and on this occasion, we discuss briefly the differences between TCI and TCI-G. In order not to interrupt the flow of ideas a more thorough comparison between the two definitions is relegated to the supplementary material (Section B).

In practice TCI allows for an extension of the SIR framework to handle extreme values (Section 4). Whether it is possible to obtain a similar extension with TCI-G is an open question. We conjecture a negative answer because our Tail Inverse Regression principles theorems 1, 2 rely on a specific consequence of TCI, namely Property (iii) from Proposition 3

below. In spirit our definition for TCI and the subsequent Tail inverse regression framework developed in Section 4 below is compatible with the main notions underlying graphical models for extremes (Engelke and Hitz (2020)) and One component regular variation (Hitz and Evans (2016)). These connections are further detailed in Remarks 5 and 6 from Section 4.

Meanwhile the next remark sheds light on the relevance of the proposed definition of TCI for statistical learning applications.

Remark 1 (TCI and Imbalanced Classification). Predicting exceedances over arbitrarily high thresholds y may be viewed as a family of binary classification problems indexed by y. Indeed for fixed y, consider the binary target $T = \mathbb{1}\{Y > y\}$ with marginal class probability $\pi = \pi_y = \mathbb{P}(Y > y)$. The goal is thus to predict T, by means of the covariate vector $X = (V, W)$ where $V \in \mathbb{R}^{p-d}, W \in \mathbb{R}^d$. As $y \to y^+$, $\pi_y \to 0$. This is a typical instance of *class imbalance*, a well documented potential issue in binary classification which has been the subject of several works in the statistical learning literature, see *e.g.* the recent papers Menon et al. (2013) or Xu et al. (2020) and the references therein. A classifier is a binary function h defined on $\mathbb{R}^p$. Given a family of candidate classifiers $h \in \mathcal{H}$ the goal is to select a 'good' candidate based on a training set and an appropriate notion of a theoretical risk and its empirical counterpart. When π is so close to zero that the probability of a classification error $\mathbb{P}(h(X) \neq T, T = 1)$ is negligible compared with $\mathbb{P}(h(X) \neq T, T = 0)$, the traditional $0-1$ risk $R(h) = \mathbb{P}(h(X) \neq T)$ is driven by the latter term and tends to favor the trivial classifier $h \equiv 0$. One standard approach aiming at granting more importance to the minority class when required by the application context (*e.g.* if the event $\{T = 1\}$, although rare, has an overwhelming impact) is to consider the *Arithmetic Mean Risk* (AM risk in short), see *e.g.* Menon et al. (2013),

$$\mathcal{R}_{\mathrm{AM}}(h) = \frac{1}{2}\Big[\mathbb{P}(h(X) = 1 \mid T = 0) + \mathbb{P}(h(X) = 0 \mid T = 1)\Big]. \tag{3.3}$$

Generalizations to arbitrary weight vectors are considered in Xu et al. (2020). In a dimension reduction context consider the classes

$$\begin{aligned}\mathcal{H} &= \{h : \mathbb{R}^p \to \{0,1\},\ \text{measurable w.r.t. } \mathcal{B}(\mathbb{R}^p)\}\ ,\\ \mathcal{H}_W &= \{h \in \mathcal{H} :\ \forall (v,w) \in \mathbb{R}^{p-d} \times \mathbb{R}^d,\ h(v,w) = \tilde{h}(w),\ \tilde{h} \text{ is measurable w.r.t. } \mathcal{B}(\mathbb{R}^d)\}.\end{aligned}$$

Let us refer to the classification problem attached respectively to $\mathcal{H}$ and $\mathcal{H}_W$ as the *full problem* and the *reduced problem*. The Bayes classifier for each problem are respectively minimizers of the AM risk over the full family $\mathcal{H}$ and the reduced one $\mathcal{H}_W$,

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{R}_{\mathrm{AM}}(h)\ ; \qquad h^*_W \in \arg\min_{h \in \mathcal{H}_W} \mathcal{R}_{\mathrm{AM}}(h).$$

The main ingredient of the subsequent analysis are the regression functions $\eta(x) = \mathbb{P}(T = 1 \mid X = x)$ and $\eta_W(w) = \mathbb{P}(T = 1 \mid W = w)$. A modification of standard arguments (see the supplementary material, Section A) yields explicit expressions for the Bayes classifiers $h^*(x) = \mathbb{1}\{\eta(x) > \pi\}$, $h^*_W(x) = \mathbb{1}\{\eta_W(w) > \pi\}$. In addition the Bayes risks are

$$\begin{aligned}\mathcal{R}_{\mathrm{AM}}(h^*) &= \mathbb{E}\left[\min\left(\frac{\eta(X)}{\pi}, \frac{1-\eta(X)}{1-\pi}\right)\right]\ ;\\ \mathcal{R}_{\mathrm{AM}}(h^*_W) &= \mathbb{E}\left[\min\left(\frac{\eta_W(W)}{\pi}, \frac{1-\eta_W(W)}{1-\pi}\right)\right].\end{aligned} \tag{3.4}$$

Because $\mathcal{H}_W \subset \mathcal{H}$ we must have $\mathcal{R}_{\mathrm{AM}}(h^*_W) \geq \mathcal{R}_{\mathrm{AM}}(h^*)$. The difference between the two may be seen as a bias term: the price to pay for dimension reduction. Indeed for any random

choices $\widehat{h} \in \mathcal{H}$, $\widehat{h}_W \in \mathcal{H}_W$, which are typically the outputs of a statistical learning algorithm applied respectively to the full covariate space and the reduced one, the excess risk for the reduced problem decomposes as

$$\mathcal{R}_{\mathrm{AM}}(\widehat{h}_W) - \mathcal{R}_{\mathrm{AM}}(h^*) = \underbrace{\mathcal{R}_{\mathrm{AM}}(\widehat{h}_W) - \mathcal{R}_{\mathrm{AM}}(h^*_W)}_{A} + \underbrace{\mathcal{R}_{\mathrm{AM}}(h^*_W) - \mathcal{R}_{\mathrm{AM}}(h^*)}_{B} .$$

The first term (A) in the right-hand side is the excess risk stemming from the particular choice of the learning algorithm, which typically increases with the dimension of the input W. In particular when $p - d$ is large, the excess risk term A will be typically less than its counterpart in the full problem $\mathcal{R}_{\mathrm{AM}}(\widehat{h}) - \mathcal{R}_{\mathrm{AM}}(h^*)$. The second term (B) is the bias term above mentioned. The bias-variance compromise is in favour of dimensionality reduction *via* projection on the second variable W whenever $A + B \leq \mathcal{R}_{\mathrm{AM}}(\widehat{h}) - \mathcal{R}_{\mathrm{AM}}(h^*)$.

We now derive an upper bound on the bias term B which is closely connected to our definition of TCI. Notice that for any finite set $\mathcal{X}$ and any pair of real functions (f, g) it holds that $|\min_{x\in\mathcal{X}} f(x) - \min_{x\in\mathcal{X}} g(x)| \leq \max_{x\in\mathcal{X}} |f(x) - g(x)|$. This, combined with (3.4) above and Jensen inequality, implies that

$$\begin{aligned} B = \mathcal{R}_{\mathrm{AM}}(h^*_W) - \mathcal{R}_{\mathrm{AM}}(h^*) &\leq \mathbb{E}\left| \min\left(\frac{\eta_W(W)}{\pi}, \frac{1-\eta_W(W)}{1-\pi}\right) - \min\left(\frac{\eta(X)}{\pi}, \frac{1-\eta(X)}{1-\pi}\right)\right| \\ &\leq \mathbb{E}\Big\{ \max\left(\frac{\eta(X) - \eta_W(W)}{\pi}, \frac{(1-\eta(X)) - (1-\eta_W(W))}{1-\pi}\right)\Big\} \\ &= \mathbb{E}\left|\frac{\eta(X) - \eta_W(W)}{\pi}\right|, \end{aligned} \tag{3.5}$$

where the latter identity holds whenever $\pi \leq 1/2$. Now, with $T = \mathbb{1}\{Y > y\}$,

$$\mathbb{E}\left|\frac{\eta(X) - \eta_W(W)}{\pi}\right| = \frac{\mathbb{E}\left|\mathbb{P}\left(Y > y \mid V, W\right) - \mathbb{P}\left(Y > y \mid W\right)\right|}{\mathbb{P}\left(Y > y\right)}.$$

One recognizes the TCI criterion in the latter expression. Thus TCI means that the bias term B vanishes as $y \to y^+$, so that *projection on W is relevant for the problem of predicting the rare event* $\{Y > y\}$, for large values of y. The cut-off value y above which $\mathcal{R}_{\mathrm{AM}}(\hat{h}_W) \leq \mathcal{R}_{\mathrm{AM}}(\hat{h})$, that is $A + B \leq \mathcal{R}_{\mathrm{AM}}(\hat{h}) - \mathcal{R}_{\mathrm{AM}}(h^*)$ (in expectation or with high probability), depends on two main factors: (i) the rate of convergence of B_y to zero and (ii) the sensitivity of the learning algorithm to the curse of dimensionality for a given sample size. Indeed both excess risks $\mathcal{R}_{\mathrm{AM}}(\hat{h}) - \mathcal{R}_{\mathrm{AM}}(h^*)$ and $\mathcal{R}_{\mathrm{AM}}(\hat{h}_W) - \mathcal{R}_{\mathrm{AM}}(h^*_W)$ typically converge to zero (in expectation or in probability) with the sample size, at a different rate which depends on the respective dimensions p, d. Precise quantification of this cut-off point for specific learning algorithms and finite sample sizes is outside the scope of the present paper and left for future research.

3.2. Examples and discussion

In this section we provide a generic example based on a mixture model where the TCI condition (3.1) is satisfied under mild assumptions. We discuss an alternative additive model in Remark 2. We consider several particular instances of the generic mixture model and on this occasion we discuss the similarities and differences between TCI and the TCI-G condition (3.2) proposed in Gardes (2018). Some technical proofs are deferred to Section B in

the supplementary material, as well as additional comments, examples and counter-examples allowing for a better understanding of the differences between the two definitions.

Our leading example is constructed as follows: Let the target Y be distributed according to a mixture

$$Y = BY_1 + (1-B)Y_2\,,$$

where B is a Bernoulli variable with parameter $\theta \in (0,1)$, and Y_1, Y_2 are real variables defined through their conditional survival functions

$$S_1(y,V) = \mathbb{P}\left(Y_1 > y \mid V\right)\,, \qquad S_2(y,W) = \mathbb{P}\left(Y_2 > y \mid W\right).$$

Here, the covariate variables V, W are respectively valued in $\mathbb{R}^{p-d}$ and $\mathbb{R}^d$ with marginal distributions that we denote by P_V and P_W. The full covariate vector is $X = (V,W) \in \mathbb{R}^p$. We assume that the variables (B,V,W) are independent. Notice that independence between V and W ensures that the Linearity Condition and Constant Conditional Variance condition are automatically satisfied. In this context, straightforward calculations (as detailed in the supplementary material, Section B) show that

$$\frac{|\mathbb{P}\left(Y > y \mid V,W\right) - \mathbb{P}\left(Y > y \mid W\right)|}{\mathbb{P}\left(Y > y\right)} = \frac{\theta(S_1(y,V) - S_1(y))}{\theta S_1(y) + (1-\theta)S_2(y)}.$$

The TCI condition is that the expectancy of the above ratio vanishes as $y \to y^+$ and it is not difficult to imagine several models for (Y_1, V) and (Y_2, W) for which it is the case, as exemplified below.

Remark 2 (Variant: additive model). The mixture model described here is by no means the only option to construct examples of variables(Y,V,W) satisfying the TCI assumption. Another natural example is an additive model $Y = Y_1 + Y_2$, where Y_1 and Y_2 are respectively driven by V and W, while Y_1 has lighter tails than Y_2. The mathematical derivations are somewhat more intricate because convolutions are involved instead of sums of distribution functions. However special cases can be worked out. In the supplementary material we consider $Y_1 = V \in \mathbb{R}$, $Y_2 = W\zeta \in \mathbb{R}$ where ζ is heavy-tailed and V, W have a compact support which is bounded away from 0 and we show that TCI holds. More general statements might be obtained using results regarding sums of regularly varying random variables (Jessen and Mikosch (2006)). We leave this question to further works.

As an example in the generic mixture model described above, consider the case where Y_1 and Y_2 are themselves defined as multiplicative mixtures

$$Y_1 = \sum_{i=1}^{p-d} M_i^{(1)} V_i \epsilon_i\,, \qquad Y_2 = \sum_{j=1}^{d} M_j^{(2)} W_j \zeta_j\,, \tag{3.6}$$

where $M^1 = (M_1^1, \ldots, M_{p-d}^1)$ is a multinomial vector with weight parameter $\pi^1 = (\pi_1^1, \ldots, \pi_{p-d}^1)$, that is $\sum_{i=1}^{p-d} M_i^1 = 1$ and $\mathbb{P}(M_i^1 = 1) = \pi_i^1$; M_2 is as well a multinomial variable with parameter $\pi^2 = (\pi_1^2, \ldots, \pi_d^2)$; and the variables ϵ_i, $i \le p-d$ and ζ_j, $j \le d$ are multiplicative noises, with different tail behaviour. Assume for simplicity that all ϵ_j's (*resp.* ζ_j's) share the same survival function S_ϵ (*resp.* S_ζ) and that for all $s,t > 0$,

$$\lim_{y\to\infty} S_\epsilon(s^{-1}y)/S_\zeta(t^{-1}y) = 0. \tag{3.7}$$

Condition (3.7) is satisfied *e.g.* with Pareto noises, $S_\epsilon(y) = y^{-\alpha_1}, S_\zeta(y) = y^{-\alpha_2}$ with $\alpha_1 > \alpha_2 > 0$, or with Exponential versus Pareto noises, $S_\epsilon(y) = e^{-\alpha_1 y}, S_\zeta(y) = y^{-\alpha_2}$, $\alpha_1, \alpha_2 > 0$.

The random vectors $M^1, M^2, \epsilon, \zeta, V, W$ are independent. Finally the covariate vectors V and W are made of independent components V_i, W_j, with nonnegative, bounded support included in an interval $[a, b]$ with $0 \leq a < b < \infty$.

In this generic example, Y_1 has a lighter tail than Y_2, so that it is the main risk factor regarding large values of Y, and it is intuitively desirable for a formal definition of tail conditional independence to be such that Y is tail conditionally independent from V given W here.

We now consider two special cases regarding the marginal distributions of the covariates V_j, W_j. recall that $[a, b]$ contains the support of each V_i and each W_j.

(i) As a first go assume that $a > 0$. Then both TCI and TCI-G hold. The proof is deferred to the supplementary material, Section B.4.

(ii) Assume now that $a = 0$, more specifically that each variable V_j, W_j follows a binary Bernoulli distribution with parameter $\tau \in (0, 1)$ (the choice of a common τ merely simplifies the notations). In Section B.5 from the supplementary material we show that TCI-G does not hold, while TCI does.

Notice that the difference between the two cases concerns only the marginal distribution of the covariate, namely whether $\mathbb{P}(W_j = 0) > 0$ is key. This seemingly minor variation results in fact in potential failure of TCI-G, while TCI remains true. The main conclusions of our comparison between the two definitions (TCI and TCI-G) in the supplementary material, Section B, may be summarized as follows.

1. Neither condition implies the other in general, except for discrete covariates where TCI-G implies TCI.
2. TCI-G criterion concerns the additional information brought by V regarding the probability of the event $Y > y$, *after* conditioning on W. The criterion is satisfied if the additional information is negligible, for *all* possible values $W = w$, even those values such that the conditional distribution of Y given $W = w$ is shorter tailed than the marginal distribution of Y. Indeed TCI-G is primarily designed for quantile regression, and the focus is not on the tail of Y's distribution, but instead on the tails of the conditional distributions of Y given W. This is the informal reason why TCI-G is not satisfied in the example above, Case (ii).
3. In constrast TCI is designed for prediction of extreme values of Y. It is an integrated version of TCI-G with respect to the variable (V, W), with a weight function granting more importance to w's such that the ratio $\mathbb{P}(Y > y \mid W = w)/\mathbb{P}(Y > y)$ is large as $y \to y^+$. In words, TCI is comparatively more sensitive to values w such that the conditional probability given $W = w$ of an exceedance $Y > y$ is large.

3.3. *Technical consequences of* TCI*, parallel with traditional conditional independence*

Definition 2 implies equivalent weak formulations of the traditional conditions (CI-1,CI-2,CI-3) reviewed in the background section.

Proposition 3. *If $Y_\infty \perp\!\!\!\perp V \mid W$ in the sense of Definition 2, then the following equivalent conditions* (i), (ii), (iii) *hold.*

(i) For any real-valued functions g and h, measurable and bounded, we have

$$\frac{\mathbb{E}\Big[g(V)h(W)\Big(\mathbb{P}(Y > y \mid V, W) - \mathbb{P}(Y > y \mid W)\Big)\Big]}{\mathbb{P}(Y > y)} \xrightarrow[y \to y^+]{} 0.$$

(ii) For any real-valued functions g and h, measurable and bounded, we have

$$\frac{\mathbb{E}\Big[\, h(W)\Big(\,\mathbb{E}\left[\mathbb{1}\{Y>y\}g(V)\mid W\right]-\mathbb{E}\left[\mathbb{1}\{Y>y\}\mid W\right]\mathbb{E}\left[g(V)\mid W\right]\,\Big)\,\Big]}{\mathbb{P}\left(Y>y\right)} \xrightarrow[y\to y^+]{} 0.$$

(iii) For any real-valued functions g and h, measurable and bounded, we have

$$\frac{\mathbb{E}\Big[\, h(W)\mathbb{1}\{Y>y\}\,\Big(\mathbb{E}\left[g(V)\mid Y,W\right]-\mathbb{E}\left[g(V)\mid W\right]\Big)\,\Big]}{\mathbb{P}\left(Y>y\right)} \xrightarrow[y\to y^+]{} 0.$$

Remark 3 (Relevance of Proposition 3 for our purpose). From a technical perspective, Property (iii) in Proposition 3 is key to obtain the tail analogues of the SIR and SAVE principles (Theorems 1, 2 in Section 4). This is not surprising insofar as the traditional condition (CI-3) for conditional independence in Definition 1 is central to prove the SIR/SAVE principles.

Whether the converse implication from Proposition 3 holds true in general, *i.e.* whether Conditions (i), (ii), (iii) imply TCI remains an open question which is not directly relevant for our purposes and thus left for future works.

Proof of Proposition 3. We first show the equivalence (ii)⇔(iii) by proving that the left-hand sides of the two conditions are identical. Indeed if g and h are bounded and measurable, then

$$\begin{aligned}\mathbb{E}\Big[h(W)\mathbb{1}\{Y>y\}\mathbb{E}\left[g(V)\mid Y,W\right]\Big] &= \mathbb{E}\Big[h(W)\mathbb{1}\{Y>y\}g(V)\Big]\\ &= \mathbb{E}\Big[h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\}g(V)\mid W\right]\Big],\end{aligned}$$

while

$$\mathbb{E}\Big[h(W)\mathbb{1}\{Y>y\}(\mathbb{E}\left[g(V)\mid W\right])\Big] = \mathbb{E}\Big[h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\}\mid W\right]\mathbb{E}\left[g(V)\mid W\right]\Big].$$

To show that (ii)⇒(i), note that

$$\begin{aligned}\mathbb{E}\Big[g(V)h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\}\mid V,W\right]\Big] &= \mathbb{E}\Big[g(V)h(W)\mathbb{1}\{Y>y\}\Big]\\ &= \mathbb{E}\Big[h(W)\mathbb{E}\left[g(V)\mathbb{1}\{Y>y\}\mid W\right]\Big]\\ &= \mathbb{E}\Big[h(W)\mathbb{E}\left[g(V)\mid W\right]\mathbb{E}\left[\mathbb{1}\{Y>y\}\mid W\right]\Big]+r_1(y)\\ &= \mathbb{E}\Big[g(V)h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\}\mid W\right]\Big]+r_1(y),\end{aligned}$$

where $\lim_{y\to y^+} r_1(y)/\mathbb{P}\left(Y>y\right)=0$ by Condition (ii).

The argument for the converse implication (ii)⇐(i) is similar:

$$\begin{aligned}\mathbb{E}\left[h(W)\mathbb{1}\{Y>y\}g(V)\right] &= \mathbb{E}\Big[h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\}\mid V,W\right]g(V)\Big]\\ &= \mathbb{E}\Big[h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\}\mid W\right]g(V)\Big]+r_2(y),\end{aligned}$$

where $\lim_{y\to y^+} r_2(y)/\mathbb{P}\left(Y>y\right)=0$ under condition (i).

Finally we show that Property (i) from Proposition 3 is satisfied under the TCI assumption from Definition 2. Let g,h be bounded, measurable functions defined on $\mathcal{V},\mathcal{W}$

respectively and let $\|g\|_\infty$ and $\|h\|_\infty$ denote their supremum norm. By Jensen's inequality,

$$\begin{aligned}&\mathbb{P}\left(Y > y\right)^{-1} \left|\mathbb{E}\Big[g(V)h(W)\Big(\mathbb{P}\left(Y > y \mid V, W\right) - \mathbb{P}\left(Y > y \mid W\right)\Big)\Big]\right| \\ &\leq \|g\|_\infty \|h\|_\infty \mathbb{P}\left(Y > y\right)^{-1} \mathbb{E}\left|\mathbb{P}\left(Y > y \mid V, W\right) - \mathbb{P}\left(Y > y \mid W\right)\right|,\end{aligned}$$

where the right hand side tends to zero under Condition (3.1) from Definition 2.

□

3.4. Extreme dimension reduction spaces

In the context of statistical regression, we now define extreme sufficient dimension reduction subspaces in a similar fashion to the usual SDR spaces.

Definition 3 (Extreme SDR space and extreme central space).

- An extreme SDR space for the pair (Z, Y) is a subspace E_e of $\mathbb{R}^p$ such that $Y_\infty \perp\!\!\!\perp Z \mid P_e Z$, where P_e is the orthogonal projection on E_e. In other words E_e is called an extreme SDR space whenever

$$\mathbb{E}\left|\frac{\mathbb{P}(Y > y|Z) - \mathbb{P}(Y > y|P_e Z)}{\mathbb{P}\left(Y > y\right)}\right| \xrightarrow[y \to y^+]{} 0. \tag{3.8}$$

- An extreme central space $E_{e,c}$ for the pair (Z, Y) is an extreme SDR space of minimum dimension.

Investigating sufficient conditions ensuring uniqueness of an extreme central space is left for future studies. Instead, in the present paper we shall consider *an* extreme SDR space E_e and we shall show that under appropriate assumptions, inverse extreme regression objects, namely limits of conditional expectations $\mathbb{E}\left[Z \mid Y > y\right]$ (Theorem 1) and second order variants (Theorem 2) belong to E_e. In particular they belong to any extreme *central* space.

Remark 4 (Relationship between the central space and its extreme counterpart). Because Equation (3.8) holds true for any $y \in \mathbb{R}$ when E_e is chosen as a (non extreme) SDR space for the pair (Z, Y), any SDR space for (Z, Y) is an extreme SDR space. Thus, upon uniqueness of the central space E_c and the extreme central space $E_{e,c}$, it holds that $E_{e,c} \subset E_c$. Examples of other dimension reduction subspaces more specific than E_c but not related to the extreme value of Y include the central mean subspace (Cook and Li, 2002) and the central quantile subspace Christou (2020).

4. Tail Inverse Regression

In the sequel, we consider an extreme SDR space $E_e \subset \mathbb{R}^p$ for the pair (Z, Y) in the sense of Definition 3. That is, we assume that $Y_\infty \perp\!\!\!\perp Z \mid P_e Z$ as in Definition 2, where P_e is the orthogonal projection on E_e. Also we define $Q_e = I_p - P_e$. In order to adapt the SIR strategy to this tail conditional independence framework, we show the following result which is a 'tail version' of the SIR principle (Proposition 1). In the remainder of this paper let $\|\cdot\|$ denote any norm on a finite dimensional vector space.

Theorem 1 (TIREX1 principle)**.** *Assume the following conditions regarding the pair* (Z, Y) *and the extreme SDR space* E_e*.*

1. *(Uniform integrability):*
   The random variables $g_{1,A}(Z) = \|Z\|\mathbb{1}\{\|Z\| > A\}$, $g_{2,A}(Z) = \mathbb{E}\left[\|Z\|\mathbb{1}\{\|Z\| > A\} \mid P_e Z\right]$ *indexed by* $A \in \mathbb{R}$ *satisfy*

$$\lim_{A\to\infty} \limsup_{y\to y^+} \mathbb{E}\left[g_{k,A}(Z) \mid Y > y\right] = 0, \qquad k = 1, 2\,; \tag{4.1}$$

2. *(LC) The standardized vector* Z *satisfies the linearity condition* (2.1) *relative to* P_e*;*
3. *(Convergence of conditional expectations) For some* $\ell \in \mathbb{R}^p$*,*

$$\mathbb{E}\left[Z \mid Y > y\right] \xrightarrow[y\to y^+]{} \ell. \tag{4.2}$$

Then $\ell \in E_e$*.*

Proof. We need to show that $Q_e\ell = 0$. By continuity of the projection operator Q_e it is enough to show that $Q_e\mathbb{E}\left[Z \mid Y > y\right] \to 0$ as $y \to y^+$. On the other hand the linearity condition (LC) (2.1) ensures that $Q_e\mathbb{E}\left[Z \mid P_e Z\right] = Q_e P_e Z = 0$ almost surely. Thus letting $p_y = \mathbb{P}(Y > y)$ one may write

$$\begin{aligned} Q_e\mathbb{E}\left[Z \mid Y > y\right] &= p_y^{-1} Q_e\mathbb{E}\left[Z\mathbb{1}\{Y > y\}\right] \\ &= p_y^{-1}\mathbb{E}\Big[\big(Q_e\mathbb{E}\left[Z|P_eZ, Y\right] - Q_e\mathbb{E}\left[Z \mid P_eZ\right]\big)\mathbb{1}\{Y > y\}\Big], \end{aligned}$$

because the second term of the difference inside the expectation of the second line is zero.

Let $A > 0$ and consider separately the case when $Z \le A$ and $Z > A$, so that

$$\begin{aligned} &Q_e\mathbb{E}\left[Z\mathbb{1}\{Y > y\}\right] \\ &= Q_e\mathbb{E}\Big[\big(\mathbb{E}\left[Z\mathbb{1}\{\|Z\| \le A\} \mid P_eZ, Y\right] - \mathbb{E}\left[Z\mathbb{1}\{\|Z\| \le A\} \mid P_eZ\right]\big)\mathbb{1}\{Y > y\}\Big] \\ &\quad + Q_e\mathbb{E}\Big[\big(\mathbb{E}\left[Z\mathbb{1}\{\|Z\| > A\} \mid P_eZ, Y\right] - \mathbb{E}\left[Z\mathbb{1}\{\|Z\| > A\} \mid P_eZ\right]\big)\mathbb{1}\{Y > y\}\Big]. \end{aligned}$$

For the first term of the above display, using Condition (ii) of Proposition 3 with $h = 1$ and $g(Z) = Z\mathbb{1}\{\|Z\| < A\}$, we obtain that

$$p_y^{-1}\, Q_e\mathbb{E}\Big[\big(\mathbb{E}\left[Z\mathbb{1}\{\|Z\| \le A\} \mid P_eZ, Y\right] - \mathbb{E}\left[Z\mathbb{1}\{\|Z\| \le A\} \mid P_eZ\right]\big)\mathbb{1}\{Y > y\}\Big] \xrightarrow[y\to y^+]{} 0. \tag{4.3}$$

For the second term corresponding to $Z > A$, we use that $\|Q_e z\| \le \|z\|$, the Jensen inequality and the triangular inequality, which yields

$$\begin{aligned} &\Big\|Q_e\mathbb{E}\Big[\big(\mathbb{E}\left[Z\mathbb{1}\{\|Z\| > A\} \mid P_eZ, Y\right] - \mathbb{E}\left[Z\mathbb{1}\{\|Z\| > A\} \mid P_eZ\right]\big)\mathbb{1}\{Y > y\}\Big]\Big\| \\ &\le \mathbb{E}\big(\{\mathbb{E}\left[\|Z\|\mathbb{1}\{\|Z\| > A\} \mid P_eZ, Y\right] + \mathbb{E}\left[\|Z\|\mathbb{1}\{\|Z\| > A\} \mid P_eZ\right]\}\mathbb{1}\{Y > y\}\big) \\ &= \mathbb{E}\left[g_{1,A}(Z)\mathbb{1}\{Y > y\}\right] + \mathbb{E}\left[g_{2,A}(Z)\mathbb{1}\{Y > y\}\right] \end{aligned}$$

By (4.3) and the previous decomposition, we have shown that

$$\limsup_{y\to y^+} \|Q_e\mathbb{E}\left[Z|Y > y\right]\| \le \limsup_{y\to y^+} \mathbb{E}\left[g_{1,A}(Z) \mid Y > y\right] + \limsup_{y\to y^+} \mathbb{E}\left[g_{2,A}(Z)|Y > y\right].$$

By further letting $A \to \infty$, by Assumption (4.1), the right-hand side is arbitrarily small. This shows that $\lim_{y\to y^+} Q_e\mathbb{E}\left[Z \mid Y > y\right] = 0$ and the proof is complete.

□

Remark 5 (special case: Tail conditional distribution). A particular framework justifying the existence of the limit ℓ (Condition (E.5) in the statement of Theorem 1) is the following. Assume that the covariate Z admits a *tail conditional distribution* given Y, in the sense that the distribution of Z conditional to $Y > y$ converges as $y \to y^+$. In other words assume that there is a probability distribution μ, that we may call the *tail conditional distribution* of Z given Y, such that for all bounded, continuous function g defined on $\mathbb{R}^p$,

$$\mathbb{E}\left[g(Z) \mid Y > y\right] \xrightarrow[y \to y^+]{} \mu(g) := \int_{\mathbb{R}^p} g \, \mathrm{d}\mu.$$

By virtue of proposition 2.20 in Van der Vaart (1998), if the uniform integrability condition (4.1) is satisfied regarding the functions $g_{1,A}$ and if Z admits a tail conditional distribution μ relative to Y, then it holds that

$$\mathbb{E}\left[Z \mid Y > y\right] \xrightarrow[y \to y^+]{} m_\mu := \int z d\mu(z),$$

so that condition (E.5) automatically holds with $\ell = m_\mu$.

Remark 6 (relationships with graphical models for extremes). The above notion of tail conditional distribution reveals a connection between the present work and graphical modeling approaches in EVT. Namely, assuming a tail conditional distribution of Z given Y, and requiring in addition that the random variable Y is regularly varying, is equivalent to assuming *one-component regular variation* of the pair (Y, Z), a concept first introduced by Hitz and Evans (2016). See in particular their Theorem 1.4, where the pair (X, Y) plays the role of the pair (Y, Z) in the present work.

The notion of conditional independence at extreme levels also plays an important role in Engelke and Hitz (2020). However our work departs significantly from the latter, in so far as the general context in the cited reference is that of unsupervised learning. All considered variables play a symmetric role –there is no target variable nor covariate –, and they rely on an assumption of joint multivariate regular variation of the considered random vector which is by no means necessary in our context.

Remark 7 (Special case: extreme central space). Upon uniqueness of the extreme central space $E_{e,c}$, under the assumptions of Theorem 1 we obtain that $\ell \in E_{e,c}$.

Remark 8 (Sufficient condition for uniform integrability). Using the fact that for any $\varepsilon > 0$, $\mathbb{1}\{\|Z\| > A\} \leq \|Z\|^\varepsilon / A^\varepsilon$, a sufficient condition for the uniform integrability condition (4.1) is that

$$\limsup_{y \to y^+} \frac{\mathbb{E}\left[\|Z\|^{1+\varepsilon} \, \mathbb{1}\{Y > y\}\right]}{\mathbb{P}\left(Y > y\right)} < \infty,$$

for some $\varepsilon > 0$.

A natural strategy in view of Theorem 1 is to consider empirical counterparts of the conditional expectations $\mathbb{E}\left[Z \mid Y > y\right]$ for large values of y so as to estimate the limit value ℓ, which belongs to any extreme SDR space. Asymptotic statistical guarantees for this approach are derived in Section 5. However an obvious limitation of Theorem 1 is that it recovers a single direction within an extreme SDR space, namely the line $\{t\ell, t \in \mathbb{R}\}$ in the case where $\ell \neq 0$. If a unique extreme central space exists and if this subspace is one dimensional, then indeed the generated line and the extreme central space coincide. To consider situations where the minimum dimension of an extreme SDR space is greater than one, we develop an extreme analogue of the SAVE framework by considering conditional second order moments. The main result justifying this approach is encapsulated in Theorem 2 below.

Theorem 2 (TIREX2 principle)**.** *Assume* (Z, Y) *and the extreme SDR space* E_e *satisfy the assumptions of Theorem 1 and that in addition,*

1. *(second order uniform integrability):*

$$\lim_{A\to\infty} \limsup_{y\to y^+} \mathbb{E}\left[h_{k,A}(Z) \mid Y > y\right] = 0, \qquad k = 1, 2\,, \tag{4.4}$$

   where $h_{1,A}(Z) = \|Z\|^2 \mathbb{1}\{\|Z\| > A\}$ *and* $h_{2,A}(Z) = \mathbb{E}\left[\|Z\|^2 \mathbb{1}\{\|Z\| > A\} \,\middle|\, P_e Z\right]$ *for* $A \in \mathbb{R}$*;*
2. *(CCV) The standardized vector* Z *satisfies the constant variance condition (2.2) relative to* P_e*;*
3. *(Convergence of conditional expectations) For some* $S \in \mathbb{R}^{p\times p}$*,*

$$\mathbb{E}\big[ZZ^\top \mid Y > y\big] \xrightarrow[y\to y^+]{} S + \ell\ell^\top. \tag{4.5}$$

Then $\operatorname{span}(S - I_p) \subset E_e$, i.e. $Q_e(S - I_p) = 0$.

Notice that the existence of $\ell = \lim \mathbb{E}\left[Z \mid Y > y\right]$ is part of the assumptions of Theorem 1 and that in the latter framework, $Q_e \ell\ell^\top = 0$. Thus condition (E.7) is equivalent to requiring that $\mathbb{V}\mathrm{ar}\left[Z \mid Y > y\right]$ converges to some limit variance S as $y \to y^+$ and the conclusion can be rephrased as $Q_e(\,\mathbb{E}[ZZ^\top \mid Y > y\,] - I_p\,) \to 0$ as $y \to y^+$, or equivalently $Q_e(\mathbb{V}\mathrm{ar}\left[Z|Y > y\right] - I_p) \to 0$. The technique of the proof is similar to that for Theorem 1. The key is to observe that the Constant Conditional Variance assumption allows to introduce a difference $(\,\mathbb{E}[ZZ^\top \mid P_e Z, Y] - \mathbb{E}[ZZ^\top \mid P_e Z]\,)\mathbb{1}\{Y > y\}$ which is asymptotically negligible because of the TCI assumption. The details are gathered in the supplement material, Section C.

5. Estimation

This section is devoted to the statistical implementation of our main results from Section 4. Theorems 1 and 2 show that the quantities ℓ and S in the limits of the two statements are key to estimate the extreme SDR space, because $\ell \in E_e$ and $\operatorname{span}(S - I_p) \subset E_e$. A natural first idea would be to use as an estimate an empirical version of the quantities $\mathbb{E}[Z \mid Y > y\,]$ or $\mathbb{E}[ZZ^\top \mid Y > y\,]$ for a high threshold y growing with the sample size n. A typical choice of such a threshold is the quantile of Y at a probability level $1 - k/n$, where $k = k(n)$ is an intermediate sequence such that $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$. Here we propose a refinement of this strategy integrating out the latter quantities over varying quantiles at probability levels $1 - uk/n$ for $u \in (0, 1)$. Such a refinement follows the proven approaches based on the CUME and CUVE matrices described in the background section 2. For this purpose we introduce and prove the asymptotic normality of the empirical processes associated with the first and second order method, that are respectively the specialisation of the SIR/CUME and the SAVE/CUVE processes to extreme regions of the target Y.

Even though the first order method is potentially less fruitful than the second order one since the limit ℓ in Theorem 1 is a single vector, it helps build the intuition about the statistical theory for both the first order and second order methods. In addition, the first order method turns out to be more stable in some of our experiments.

Some notations are introduced in Section 5.1. We provide asymptotic theory for the first and second order empirical processes in Section 5.2. Section 5.3 summarizes the methods we suggest for estimating E_e.

5.1. Framework and notations

For any right-continuous cumulative distribution function H (be it empirical or not), we shall denote by H^- the left-continuous inverse of H, $H^-(u) = \inf\{x \in \mathbb{R} : H(x) \geq u\}$. Recall that with these conventions, for $u \in [0,1]$ and $x \in \mathbb{R}$, we have

$$H(x) \geq u \iff x \geq H^-(u). \tag{5.1}$$

For any i.i.d. sample $(T_i)_{i\leq n}$ associated with a real random variable T with cumulative distribution H, we use the standard definition of the empirical distribution function,

$$\hat{H}(x) = n^{-1}\sum_{i=1}^{n} \mathbb{1}\{T_i \leq x\}. \tag{5.2}$$

For notational and mathematical convenience we shall work with the negative target $\tilde{Y} = -Y$ and denote the *c.d.f.* of $\tilde{Y}$ as F , which we assume to be continuous in the remainder of this paper. For simplicity we write k instead of $k(n)$ for the intermediate sequence defined at the beginning of this section, as is customary in extreme value statistics. Consider the first order and second order inverse regression functions $C_n(u), B_n(u)$,

$$C_n(u) = \frac{n}{k}\mathbb{E}\left[Z\mathbb{1}\{\tilde{Y} < F^-(uk/n)\}\right] , \tag{5.3}$$

$$B_n(u) = \frac{n}{k}\mathbb{E}\left[(ZZ^\top - I_p)\mathbb{1}\{\tilde{Y} < F^-(uk/n)\}\right]. \tag{5.4}$$

The empirical versions of (5.3) and (5.4) based on an independent sample (Z_i, Y_i) identically distributed as the pair (Z, Y) are

$$\hat{C}_n(u) = \frac{1}{k}\sum_{i=1}^{n} Z_i\mathbb{1}\{\tilde{Y}_i \leq \hat{F}^-(uk/n)\}\,, \tag{5.5}$$

$$\hat{B}_n(u) = \frac{1}{k}\sum_{i=1}^{n}(Z_iZ_i^\top - I_p)\mathbb{1}\{\tilde{Y}_i \leq \hat{F}^-(uk/n)\}. \tag{5.6}$$

Extensions to the more realistic situation where the pair (X, Y) is observed with the mean and covariance of X unknown are gathered in Section E from the supplementary material.

5.2. Main result

The remainder of this section aims at establishing the weak convergence of the (tail) empirical processes associated with TIREX, respectively $\sqrt{k}(\hat{C}_n(u) - C_n(u))$ and $\sqrt{k}(\hat{B}_n(u) - B_n(u))$ in the space of bounded functions $\ell^\infty([0,1])$. This is achieved in Corollary 1.

A key point of our analysis, which follows from the continuity of F, is that the functions $C_n(u), B_n(u)$ and their estimates $\hat{C}_n(u), \hat{B}_n(u)$ are invariant under the transformation $U = F(\tilde{Y})$. More precisely, with the latter notation, we have the following identities

$$C_n(u) = \frac{n}{k}\mathbb{E}\big[\, Z\mathbb{1}\{U < uk/n\}\,\big]\,, B_n(u) = \frac{n}{k}\mathbb{E}\big[\,(ZZ^\top - I_p)\mathbb{1}\{U < uk/n\}\,\big],$$

and the processes $\hat{C}_n(u), \hat{B}_n(u)$ remain the same when constructed from the initial sample $(X_i, \tilde{Y}_i)$ or when constructed from the uniformized sample (X_i, U_i). Indeed for $u \in [0,1]$, it holds that

$$\mathbb{1}\{\tilde{Y}_i \leq \hat{F}^-(uk/n)\} = \mathbb{1}\{U_i \leq \hat{F}_U^-(uk/n)\}, \qquad \text{a.s.},$$

where $\hat{F}_U$ is the empirical distribution function associated with the uniform sample $U_1, \ldots, U_n$, see Fact D.1 in the supplementary material for a short proof. These facts are a known feature of rank based estimators; see for instance Fermanian, Radulovic and Wegkamp (2004) in the copula estimation context and Portier (2016) in the standard SIR context.

We now state our main result which is formulated in terms of a generic random pair (V, Y), an $i.i.d.$ sample thereof $(V_i, Y_i), i \leq n$, and a measurable function $h : \mathbb{R}^r \to \mathbb{R}^q$, where Y is the response variable as above, the covariate V is a random vector of size $r \in \mathbb{N}^*$ and h is such that the random vector $h(V)$ has finite second moments. Define

$$D_n(u) = \frac{n}{k}\mathbb{E}\left[h(V)\mathbb{1}\{\tilde{Y} < F^-(uk/n)\}\right],$$
$$\hat{D}_n(u) = \frac{1}{k}\sum_{i=1}^{n} h(V_i)\mathbb{1}\{\tilde{Y}_i \leq \hat{F}^-(uk/n)\}.$$

Weak convergence of the TIREX1 and TIREX2 processes (Corollary 1) is obtained upon setting $V = Z$ and respectively $h_C(z) = z$ and $h_B(z) = \mathrm{vec}(zz^\top - I_p)$, where for any matrix $M \in \mathbb{R}^{r\times s}$, $\mathrm{vec}(M)$ denotes the vector of size $r \times s$ obtain by concatenating the columns of M.

Theorem 3 (Tail empirical process for a generic pair (V, Y))**.** *Suppose that the distribution function F of $\tilde{Y} = -Y$ is continuous and that, letting $U = F(\tilde{Y})$, it holds that*

1. *for any $j \in \{1, \ldots, q\}$, the functions $u \mapsto \mathbb{E}\left[h(V)_j\mathbb{1}\{U \leq u\}\right]$ and $u \mapsto \mathbb{E}\left[h(V)_j^2\mathbb{1}\{U \leq u\}\right]$ are differentiable on $(0, 1)$ with a continuous derivative at 0,*
2. *for all $M \geq 0$, $S(M) := \lim_{\delta\to 0}\mathbb{E}\left[h(V)h(V)^\top\mathbb{1}\{\|V\| \geq M\} \,\middle|\, U \leq \delta\right]$ exists and is such that $\lim_{M\to\infty} S(M) = 0$,*
3. *as $\delta \to 0$, $\mathbb{E}\left[h(V) \mid U \leq \delta\right]$ converges to a limit $\nu \in \mathbb{R}^q$.*

Then we have as $n \to \infty, k \to \infty, k/n \to 0$,

$$\left\{\sqrt{k}(\widehat{D}_n(u) - D_n(u))\right\}_{u\in[0,1]} \rightsquigarrow \{W(u)\}_{u\in[0,1]},$$

where W is a Gaussian process with mean zero and covariance function

$$(s, t) \mapsto s \wedge t\left(\Xi - \nu\nu^\top\right), \tag{5.7}$$

with ν as in the 3^{textrd} Condition of the statement and

$$\Xi = S(0) = \lim_{\delta\to 0}\mathbb{E}\left[h(V)h(V)^\top \,\middle|\, U \leq \delta\right] \in \mathbb{R}^{q\times q}. \tag{5.8}$$

Corollary 1 (Weak convergence of the TIREX1 and TIREX2 processes)**.** *By choosing the pair $(V, Y) = (Z, Y)$ and assuming that the function $h_C(z) = z$ (*resp. *$h_B(z) = \mathrm{vec}(zz^\top - I_p)$) satisfies the assumptions of Theorem 3, the TIREX1 process $\sqrt{k}(\hat{C}_n(u) - C_n(u))$ (*resp. *the TIREX2 process $\sqrt{k}(\hat{B}_n(u) - B_n(u))$ converges weakly in $\ell^\infty(0, 1)$ to a tight Gaussian process W_C (*resp. *W_B) with covariance function given by (5.7) with $V = Z$ and $h = h_C$ (*resp. *$h = h_B$)*

Proof of Theorem 3. Consider the pseudo-empirical version of $D_n(u)$,

$$\widetilde{D}_n(u) = k^{-1}\sum_{i=1}^{n} h(V_i)\mathbb{1}\{U_i \leq uk/n\} = k^{-1}\sum_{i=1}^{n} h(V_i)\mathbb{1}\{\tilde{Y}_i \leq F^-(uk/n)\}. \tag{5.9}$$

Notice that $\widetilde{D}_n$ is not observed but serves as an intermediate quantity through the following key identity:

$$\hat{D}_n(u) = \widetilde{D}_n\Big(\frac{n}{k}\hat{F}_U^-(uk/n)\Big),$$

where $\hat{F}_U$ is the empirical *c.d.f.* associated with the sample $(U_i, i \leq n)$. Introducing the process

$$\widetilde{\Gamma}(u) = \sqrt{k}\left(\widetilde{D}_n(u) - D_n(u)\right), \quad u \in [0,1], \tag{5.10}$$

we have the following decomposition

$$\sqrt{k}(\hat{D}_n(u) - D_n(u)) = \widetilde{\Gamma}\Big(\frac{n}{k}\,\hat{F}^-\big(uk/n\big)\Big) + \sqrt{k}\Big(D_n\Big(\frac{n}{k}\,\hat{F}_U^-\big(uk/n\big)\Big) - D_n(u)\Big). \tag{5.11}$$

In the remainder of the proof, we show that the first term can be replaced by $\widetilde{\Gamma}(u)$, while the second term can be replaced by $-\nu\hat{\gamma}_1(u)$ where $\hat{\gamma}_1$ is the tail empirical process for uniform random variables,

$$\hat{\gamma}_1(u) = \sqrt{k}\left(\frac{n}{k}\hat{F}_U(uk/n) - u\right). \tag{5.12}$$

Finally we show that the process $(\hat{\gamma}_1(u), \widetilde{\Gamma}(u))_{u\in[0,1]}$ converges jointly to a Gaussian process.

Intermediate results, uniform tail processes

The main tools that we use in our proof of Theorem 3 concern the weak convergence of the tail empirical (quantile) process associated with a uniform response variables. Many approaches have been considered to handle the behavior of such processes, see Csorgo et al. (1986) for general empirical processes and Einmahl and Mason (1988) for the tail version. For the sake of completeness we provide in the supplementary material (Section D.3) a different, direct proof of Lemma 1 below, relying on 'classes of function changing with n' (Van Der Vaart and Wellner (2013))

Lemma 1. *Under the assumptions of Theorem 3, the process $\widetilde{\Gamma}$ defined in (5.10) converges weakly in $\ell^\infty(0,1)$ to a tight Gaussian process $\widetilde{W}$ with covariance function*

$$(u_1, u_2) \mapsto (u_1 \wedge u_2)\Xi\,,$$

where Ξ is defined in (5.8)

An immediate consequence of Lemma 1, obtained upon setting $V = Z$ and $h(V) = 1$, is the weak convergence of the tail empirical process for uniform random variables introduced in 5.12.

Corollary 2. *As $n \to \infty$, $k \to \infty$, and $k/n \to 0$, the uniform tail empirical process (5.12) weakly converges to a standard Brownian motion W_1.*

Combining Corollary 2 and an appropriate variant of Vervaat's lemma (see Section D.2 from the supplementary material) we obtain in Section D.4 from the same supplement, the following result.

Lemma 2. *As $n \to \infty$, $k \to \infty$,*

$$\sup_{u\in(0,1]}\left|\sqrt{k}\left(\frac{n}{k}\hat{F}_U^-(uk/n) - u\right) + \hat{\gamma}_1(u)\right| = o_{\mathbb{P}}(1).$$

Separate and joint convergence in Decomposition (5.11)

We now show the following three relations: as $n \to \infty$,

$$\sup_{u\in(0,1]} \left|\widetilde{\Gamma}\Big(\frac{n}{k}\,\hat{F}_U^-\big(uk/n\big)\Big) - \widetilde{\Gamma}(u)\right| = o_{\mathbb{P}}(1), \tag{5.13}$$

$$\sup_{u\in(0,1]} \left|\sqrt{k}\left(D_n\Big(\frac{n}{k}\,\hat{F}_U^-\big(uk/n\big)\Big) - D_n(u)\right) + \nu\,\hat{\gamma}_1(u)\right| = o_{\mathbb{P}}(1), \tag{5.14}$$

$$\begin{pmatrix}\hat{\gamma}_1(u)\\ \widetilde{\Gamma}(u)\end{pmatrix} \rightsquigarrow W'(u), \tag{5.15}$$

where W' is a centered Gaussian process on $(0,1]$ with covariance function $(s,t)\mapsto s\wedge t\,\Xi'$. Here $\Xi' = S'(0)$ is the limit second moment matrix from Lemma 1 applied to $h'(V) = (1, h(V))$. More specifically, with this choice of h', we have

$$\Xi' = \begin{pmatrix}1 & \nu^\top\\ \nu & \Xi\end{pmatrix} \in \mathbb{R}^{(q+1)\times(q+1)}$$

where $\Xi = \lim_{\delta\to 0}\mathbb{E}\left[h(V)h(V)^\top \,|\, U\le\delta\right]$.

We first prove (5.13). From Lemma 1, the process $\widetilde{\Gamma}$ is tight, whence asymptotically equi-continuous, meaning that

$$\lim_{\delta\downarrow 0}\limsup_n \mathbb{P}\Big(\sup_{|s-t|\le\delta}|\widetilde{\Gamma}(s)-\widetilde{\Gamma}(t)| > \epsilon\Big) = 0.$$

Also, from Lemma 2 and Corollary 2, $\sup_{u\in(0,1]}|(n/k)\hat{F}_U^-(uk/n) - u| = o_{\mathbb{P}}(1)$. Combining the two yields (5.13).

To prove (5.14), we apply the mean value theorem to get that

$$\begin{aligned}
&\sqrt{k}\left\{D_n\Big(\frac{n}{k}\hat{F}_U^-\big(uk/n\big)\Big) - D_n(u)\right\}\\
&= \frac{n}{\sqrt{k}}\left\{\mathbb{E}\Big[h(V)\mathbb{1}\{U\le u_n\}\Big]_{u_n=\hat{F}_U^-(uk/n)} - \mathbb{E}\Big[h(V)\mathbb{1}\{U\le uk/n\}\Big]\right\}\\
&= \frac{n}{\sqrt{k}}\tilde{g}(\tilde{U}_{u,n})\Big\{\hat{F}_U^-\big(uk/n\big) - uk/n\Big\}\\
&= \sqrt{k}\,\tilde{g}(\tilde{U}_{u,n})\Big\{\frac{n}{k}\,\hat{F}_U^-\big(uk/n\big) - u\Big\},
\end{aligned}$$

where $\tilde{g}(x)$ is the derivative of $x\mapsto\mathbb{E}\left[h(V)\mathbb{1}\{U\le x\}\right]$ at point x and $\tilde{U}_{u,n}$ lies on the line segment between $\hat{F}_U^-(uk/n)$ and uk/n. Lemma 2 and Corollary 2 imply that $\tilde{U}_{u,n}\to 0$ in probability uniformly over $u\in[0,1]$, thus by continuity of $\tilde{g}$ at 0, $g(\tilde{U}_{u,k}) = \tilde{g}(0) + o_{\mathbb{P}}(1)$. We can further calculate $\tilde{g}(0)$ based on assumption 3 in Theorem 3 as follows,

$$\tilde{g}(0) = \lim_{u\to 0}\mathbb{E}\left[h(V)\mathbb{1}\{U\le u\}\right]/u = \lim_{u\to 0}\mathbb{E}\left[h(V)\,|\,U\le u\right] = \nu.$$

Therefore, the relation (5.14) is proved by applying Lemma 2, and the Slutsky's lemma. Finally, (5.15) follows from applying Lemma 1 to the function $h'(V) = (1, h(V))$.

Conclusion

By combining the decomposition in (5.11) with the relations (5.13)-(5.15), we obtain that, as $n\to\infty$,

$$\left\{\sqrt{k}\left(\widehat{D}_n(u) - D_n(u)\right)\right\}_{u\in[0,1]} \rightsquigarrow W := (-\nu, I_q)\,W'$$

which is a Gaussian process with covariance function

$$\begin{aligned}\Sigma(s,t) &= s \wedge t \, (-\nu, I_q) \begin{pmatrix} 1 & \nu^\top \\ \nu & \Xi \end{pmatrix} \begin{pmatrix} -\nu^\top \\ I_q \end{pmatrix} \\ &= s \wedge t \left(\Xi - \nu\nu^\top\right).\end{aligned}$$

□

5.3. Proposed estimation method

This section summarizes the main steps of the first and second order methods that we propose based on the processes $\hat{C}_n$ and $\hat{B}_n$. We first introduce TIREX1 and TIREX2 matrices in parallel with the matrix M_{CUME} defined in (2.3) in our framework, following the integral based methods proposed by Zhu, Zhu and Feng (2010), see also Portier (2016). In line with the CUME (2.3) matrix, we define

$$\begin{aligned}M_{\text{TIREX1}} &= \int_0^1 C_n(u) C_n(u)^\top \, \mathrm{d}u \,, \\ M_{\text{TIREX2}} &= \int_0^1 B_n(u) B_n(u)^\top \, \mathrm{d}u \,,\end{aligned} \tag{5.16}$$

where C_n and B_n are defined in (5.3) and (5.4) respectively. We omit the dependency of the matrices on n, k for convenience. An easy but important observation which underlies our strategy for estimating an extreme SDR space is the following lemma.

Lemma 3 (Consistency of the TIREX matrices)**.**
(i) *Under the assumptions of Theorem 1,*

$$M_{TIREX1} \longrightarrow \frac{1}{3}\ell\ell^\top \quad \text{as } n \to \infty.$$

(ii) *Under the assumptions of Theorem 2,*

$$M_{TIREX2} \longrightarrow \frac{1}{3}(S - I_p + \ell\ell^\top)^2 \quad \text{as } n \to \infty.$$

Proof. Under the assumptions of the first statement, for fixed u, $C_n(u)C_n(u)^\top \to u^2\ell\ell^\top$ as $n \to \infty$. The result follows by dominated convergence on $(0,1)$, which applies by virtue of Condition (4.1). Indeed this uniform integrability assumption ensures that for some constant $A > 0$, for n large enough, for all $u \in (0,1)$,

$$\begin{aligned}\|C_n(u)\| &= \left\| u\mathbb{E}\left[Z \,|\, \tilde{Y} < F^-(uk/n) \right] \right\| \\ &\leq u(A + \mathbb{E}\left[\|Z\| \mathbb{1}\{\|Z\| > A\} \,|\, \tilde{Y} < F^-(uk/n) \right] \leq u(A+1).\end{aligned}$$

The argument for the second statement is similar, up to a call to Condition (4.4) instead of (4.1). □

As a consequence of Lemma 3, both column spaces of M_{TIREX1} and M_{TIREX2} are asymptotically included in E_e. The column space of M_{TIREX1} has dimension one while that of M_{TIREX2} can be of any dimension not higher than that of E_e. We propose the following estimation procedures based respectively on the processes $\hat{C}_n$ and $\hat{B}_n$.

TIREX1

1. Choose $k \ll n$ and $1 \leq d \leq p$.
2. Compute the estimated TIREX1 matrix, $\widehat{M}_{\text{TIREX1}} = \int_0^1 \hat{C}_n(u)\hat{C}_n^\top(u)\,\mathrm{d}u$ using the identity given in (6.1).
3. Perform an eigen decomposition of $\widehat{M}_{\text{TIREX1}}$ and keep the first d eigenvectors $(e_i, i \leq d)$.
4. ouptut: $\hat{E}_e = \text{span}(\{e_i, i \leq d\})$.

Choosing $d > 1$ is not immediately justified because the limit of M_{TIREX1} is a rank one matrix $\ell\ell^\top/3$ as indicated in Lemma 3. However, empirical evidence suggests that allowing $d > 1$ can be useful to recover more components among the extreme central subspace basis. This is why we include this option in the algorithm.

TIREX2

1. Choose $k \ll n$ and $1 \leq d \leq p$.
2. Compute the estimated TIREX2 matrix, $\widehat{M}_{\text{TIREX2}} = \int_0^1 \hat{B}_n(u)\hat{B}_n^\top(u)\,\mathrm{d}u$ using the identity given in (6.2).
3. Perform an eigen decomposition of $\widehat{M}_{\text{TIREX2}}$ and keep the first d eigenvectors $(e_i, i \leq d)$ associated with the highest eigen values.
4. ouptut: $\hat{E}_e = \text{span}(\{e_i, i \leq d\})$

We make the following remarks regarding the relationships between our main theoretical result Corollary 1 and the proposed estimation methods TIREX1 and TIREX2.

Remark 9 (Asymptotic normality of the TIREX matrices). The asymptotic normality of the random matrices $\sqrt{k}(\widehat{M}_{\text{TIREX1}} - M_{\text{TIREX1}})$ and $\sqrt{k}(\widehat{M}_{\text{TIREX2}} - M_{\text{TIREX2}})$ could be obtained as a further consequence of Corollary 1 with straightforward calculations. This can be achieved by using the Delta-method as in the proof of Portier (2016), Proposition 5. For the sake of conciseness we leave the detailed proof to interested readers.

Remark 10 (Bias term). Notice that the TIREX matrices M_{TIREX} are deterministic but subasymptotic quantities which depend on the the choice of the ratio k/n. The ultimate goal in view of Lemma 3 would be to obtain the limit distribution of $\sqrt{k}(\widehat{M}_{\text{TIREX1}} - \frac{1}{3}\ell\ell^\top)$ and $\sqrt{k}(\widehat{M}_{\text{TIREX2}} - \frac{1}{3}(S - I_p + \ell\ell^\top))$. An obvious way to do so is to assume that the bias terms $\sqrt{k}(M_{\text{TIREX1}} - \frac{1}{3}\ell\ell^\top)$ and $\sqrt{k}(M_{\text{TIREX2}} - \frac{1}{3}(S - I_p + \ell\ell^\top))$ converge to zero in probability, and use Slutsky's lemma.

Remark 11 (Principal Component Analysis of the TIREX matrices). The output of the TIREX methods is the eigen spaces of the estimated TIREX matrices. An important final step is to show that such eigen spaces converges to the space spanned by the limits $1/3\ell\ell^\top$ and $1/3(S - I_p + \ell\ell^\top)$. A possible starting point would be to use results from perturbation theory, see *e.g.* (Zwald and Blanchard, 2005, Theorem 3) where the Frobenius norm of the error is controlled by the inverse of a spectral gap.Since this problem is left aside even in the traditional inverse regression literature we leave this question to further research while demonstrating the performance of the TIREX algorithms by numerical experiments.

Remark 12 (Choices of d, k). The choice of the intermediate sequence k of extreme order statistics is a standard issue in extreme value statistics. In our experiments (Section 6) we propose to choose k by cross-validation. Theoretical investigation regarding this strategy is beyond the scope of this paper. Similarly, the choice of d in the PCA decomposition of the

matrix $\widehat{M}_{\text{TIREX2}}$ is a recurrent question in the PCA literature, which is also left to further research. In practice a natural and widely used strategy is an elbow method applied to the plot of the estimated eigen values. In the supervised learning context, we recommend to choose d by cross-validation. More generally (outside the supervised learning context), testing for the rank of the underlying matrix is a convenient method to infer the value of d. Such an approach has been succcessfully employed in the SDR literature (Portier and Delyon, 2014) where the test statistics are usually based on the eigenvalues amplitude. Finally recall that the limit of the matrix M_{TIREX1} has rank one, so that the default choice of $d = 1$ in the first order method is legitimate. Investigating theoretical guarantees for choosing the value of d in the TIREX context is beyond the scope of this paper and left for future work.

6. Experiments

This section focuses on the practical usefulness of TIREX for finite sample sizes based on simulated and real data. We first give some details about the implementation of TIREX (Section 6.1) and discuss its computational complexity. We discuss the improvement brought by TIREX over the estimation method proposed by Gardes (2018). Second, with synthetic datasets of various dimensions, we explore the estimation performance of TIREX1 and TIREX2 for various values of k, as measured by a distance between the estimated and true extreme SDR spaces (Section 6.2). On this occasion we compare the estimation performance of TIREX with that of its closest alternatives, namely Gardes (2018)'s method, CUME and CUVE. Finally in Section 6.3 we compare TIREX with several existing dimension reduction tools for predicting tail events on several real data sets of relatively high dimension.

6.1. TIREX implementation

In a preliminary step common to all our experiments, the covariates are empirically standardized and we set $\hat{Z}_i = \hat{\Sigma}^{-1/2}(X_i - \hat{m})$ with $\hat{m} = n^{-1}\sum_{i=1}^n X_i$ and $\hat{\Sigma} = n^{-1}\sum_{i=1}^n (X_i - \hat{m})(X_i - \hat{m})^T$. Working with empirically standardized covariates to estimate an extreme SDR space E_e is equivalent to working with raw covariates to estimate $\tilde{E}_e = \Sigma^{-1/2}E_e$ up to remainder terms of order $O_{\mathbb{P}}(1/\sqrt{n})$, see Section E.3 in the supplementary material. By abuse of notation we use the same symbols in the present section to denote both the empirical processes constructed with the $\hat{Z}_i$'s and the Z_i's.

We start off by deriving an explicit, computationally efficient formula for the matrices $\widehat{M}_{\text{TIREX1}}, \widehat{M}_{\text{TIREX2}}$. Let $(\hat{Z}_{(1)}, Y_{(1)}), (\hat{Z}_{(2)}, Y_{(2)}), \ldots, (\hat{Z}_{(n)}, Y_{(n)})$ be such that $Y_{(1)} \geq \cdots \geq Y_{(n)}$. From the definition of $\hat{C}_n$, we have $\hat{C}_n(u) = \frac{1}{k}\sum_{i=1}^{\lceil ku \rceil} \hat{Z}_{(i)}$. This implies that $\hat{C}_n$ is piece-wise constant, more precisely for $j \in \{1, \ldots, n\}$, whenever $u \in ((j-1)/k, j/k]$, we have $k\hat{C}_n(u) = \sum_{i=1}^j \hat{Z}_{(i)} := \hat{S}_j$. Since $\widehat{M}_{\text{TIREX1}} = \sum_{j=1}^k \int_{(j-1)k}^{j/k} \hat{C}_n(u)\hat{C}_n(u)^\top du$, it follows that

$$\widehat{M}_{\text{TIREX1}} = \frac{1}{k^3}\sum_{j=1}^k \hat{S}_j \hat{S}_j^\top . \tag{6.1}$$

Evaluating the latter display requires $O(n\log(n))$ operations for sorting the Y's values; kd operations to compute the $\hat{S}_j$, $j = 1, \ldots, k$ (because $\hat{S}_j$ can be deduced from $\hat{S}_{j-1}$ with one operation); and $O(kd^2)$ operations to compute the matrix $\widehat{M}_{\text{TIREX1}}$. The overall cost is then of order $n\log(n) + kd^2$. Similar arguments regarding the second order matrix $\widehat{M}_{\text{TIREX2}}$ lead

to the expression

$$\widehat{M}_{\text{TIREX2}} = \frac{1}{k^3}\sum_{j=1}^{k} \hat{T}_j \hat{T}_j^T, \tag{6.2}$$

with $T_j = \sum_{i=1}^{j}(\hat{Z}_{(i)}\hat{Z}_{(i)}^T - I_p)$.

The final step is to perform an eigen-decomposition of the estimated matrix $\widehat{M}_{\text{TIREX1}}$ (*resp.* $\widehat{M}_{\text{TIREX2}}$). Given the alleged dimension d of E_e, the vector space generated by the d eigen vectors associated to the d largest eigenvalues of the matrix (with multiplicities, assuming uniqueness of the corresponding eigen space for simplicity) constitutes the TIREX estimate $\widehat{E}_e$. The non standard SDR space can be estimated by multiplying the obtained directions by $\hat{\Sigma}^{-1/2}$.

Computational complexity.

Evaluating (6.1) requires $O(n\log(n))$ operations for sorting the Y's values; kp operations to compute the $\hat{S}_j$, $j = 1, \ldots, k$ (because $\hat{S}_j$ can be deduced from $\hat{S}_{j-1}$ with one operation); and $O(kp^2)$ operations to compute the matrix $\widehat{M}_{\text{TIREX1}}$. The overall cost is then of order $n\log(n) + kp^2$. Similarly the overall cost for $\widehat{M}_{\text{TIREX2}}$ is of order $n\log(n) + kp^4$. Finally the eigen-decompostion based on SVD requires $O(p^3)$ operations.

In contrast the estimation procedure proposed in Gardes (2018) relies on an optimization strategy over a $p-d$-dimensional grid where d is the reduced dimension, and has an important computational cost when $d > 1$ according to the author (see Sections 3.2 and 4.1 of the cited reference). The existing implementation of Gardes (2018)'s method is restricted to $d = 1$ and the experiments conducted in that paper are limited to $p = 4$. Whether it is possible to bypass the curse of dimensionality in Gardes (2018)'s framework remains an open question. For these reasons we limit our comparison with Gardes (2018)'s method in our experiments to low dimensional examples, Models A,C, introduced below.

6.2. Performance for tail SDR estimation, synthetic data

We consider three particular instances of the mixture model presented in Section 3.2. The heavy tailed noise variables $\zeta_j, j \leq d$ follow identical Pareto distributions, $\mathbb{P}(\zeta_j > t) = t^{-\alpha_2}$ with $\alpha_2 = 10$. The short-tailed noise variables $\epsilon_j, j \leq p-d$ are exponentially distributed, $\mathbb{P}(\epsilon_j > t) = e^{-\alpha_1 t}, t > 0$, with rate parameter $\alpha_1 = 10$. The variables $(\zeta_j, j \leq d; \epsilon_j, j \leq p-d)$ are independent.

Model A. We consider Case (i) from the generic example (continuous covariates) with $\theta = 0.5$, $a = 1, b = 10$. For simplicity we take all covariate variables uniformly distributed over the interval $[a, b] = [1, 10]$. Recall that in this context, both TCI and TCI-G hold. Then according to both definitions the d-dimensional subspace of $\mathbb{R}^p$ generated by the canonical basis vectors $(e_{p-d+1}, \ldots, e_p)$ is an extreme SDR space. We set $p = 2$, $d = 1$.

Model B. Here we set $p = 30, d = 5$, all other setup remains unchanged comparing with Model A.

Model C. We use the distribution described in Case (ii) from Section 3.2, where the covariates are Bernoulli variables. In this context, TCI holds but TCI-G does not. We set the Bernoulli parameter to $\tau = 0.5$. To maintain the comparability between TIREX and Gardes (2018) we set $p = 2, d = 1$.

Experimental setting.

The sample size is set to $n = 10^4$ for Models A and C, and to $n = 10^5$ for Model B. The TIREX matrices following (6.1) and (6.2) are computed for 150 different values of k within the range $[\![n/100, n]\!]$. The orthogonal projection on the subspace generated by their first d eigen vectors constitutes our estimates $\hat{P}_e$. In other words we consider for simplicity that d is known by the user, as discussed in Remark 12. The quality of the estimator is measured by the squared Frobenius norm of the error, $\|\hat{P}_e - P_e\|_F^2$. We evaluate the squared bias $\|P_e - \mathbb{E}[\hat{P}_e]\|_F^2$, the variance $\mathbb{E}[\|\hat{P}_e - \mathbb{E}[\hat{P}_e]\|_F^2]$, and the MSE $\mathbb{E}[\|\hat{P}_e - P_e\|_F^2]$ using TIREX, based on $N = 200$ repetitions. Thus the maximum relative error of the MSE estimate, *i.e.* the maximum standard deviation of the estimate divided by the estimate itself, over all models and all values of k, is 0.11, which is sufficiently small for a qualitative interpretation of the results.

In addition we compare the relative performances of TIREX1 and Gardes (2018)'s method for Models A and C. We leave TIREX2 outside the comparison because our results (Figure 1) show that TIREX1 is a better option in this setting. To alleviate the computational cost we perform only $N = 100$ repetitions and we estimate the projectors for two values of k, namely $k = n^{2/3} \approx 464$ as recommended in Gardes (2018) and $k = 2000$ which is close to the value minimizing the MSE with TIREX1 for both models, considering our results below.

Results. Figure 1 displays the squared bias, variance and MSE for TIREX1 and TIREX2 as a function of k. The curves illustrate the typical bias-variance trade-off in Extreme Value Analysis regarding the choice of k, and confirm the findings of Corollary 1. Small values of k are associated with large variance, while large values result in a large bias. Notice that choosing $k = n$ with TIREX1 (*resp.* TIREX2) amounts to applying the standard SIR method CUME (*resp.* CUVE). Our results show that the MSE in this case is typically much larger (due to the bias) than with moderate k's, namely with $k \approx 2000$ for $n = 10^4$ and $k \approx 15000$ for $n = 10^5$.

In some cases, comparatively larger variances occur for $k \approx n/2$. We interpret this as an unstable transitional regime between two extremal behaviors: On the one hand, for small values of k, only the very largest values of Y are selected. These are mostly generated by the second component Y_2 of the mixture model, the heavy-tailed one. On the other hand when k is large, both components Y_1, Y_2 are equally involved in the computation of M_{TIREX}.

The variance attached to the second order method TIREX2 tends to be larger than that of the first order method TIREX1. However, when the dimension of the extreme SDR space is greater than one (Model B), TIREX1 fails to recover more than one direction, and TIREX2 is preferable. This fact illustrates the conclusion of Theorem 1, see also Lemma 3, where a single vector (or a rank-one matrix) is identified in the limit. TIREX2 does not suffer from this flaw since the associated limit in Lemma 3 is a matrix offering potentially more than one direction in the SDR space. As a conclusion, one should definitely prefer TIREX1 over TIREX2 when the extreme values of Y are known to be explained by a single linear combination of $Z_1, \ldots, Z_p$. Otherwise it is necessary to resort to TIREX2 to discover additional directions, even though the estimates may have a higher variance.

Table 1 displays the results of the comparison with Gardes (2018)'s method in terms of MSE and execution time. In Model A where Gardes (2018)'s assumptions are satisfied, Gardes (2018)'s method performs better than TIREX for the two values of k considered. However its execution time, even in this low dimensional setting is several orders of magnitude higher than that of TIREX. In Model C, as suggested by the theory, Gardes (2018)'s method fails to recover the tail SDR space (in the sense of TCI, not TCI-G). By contrast, TIREX can recover the tail SDR space within very short execution time.

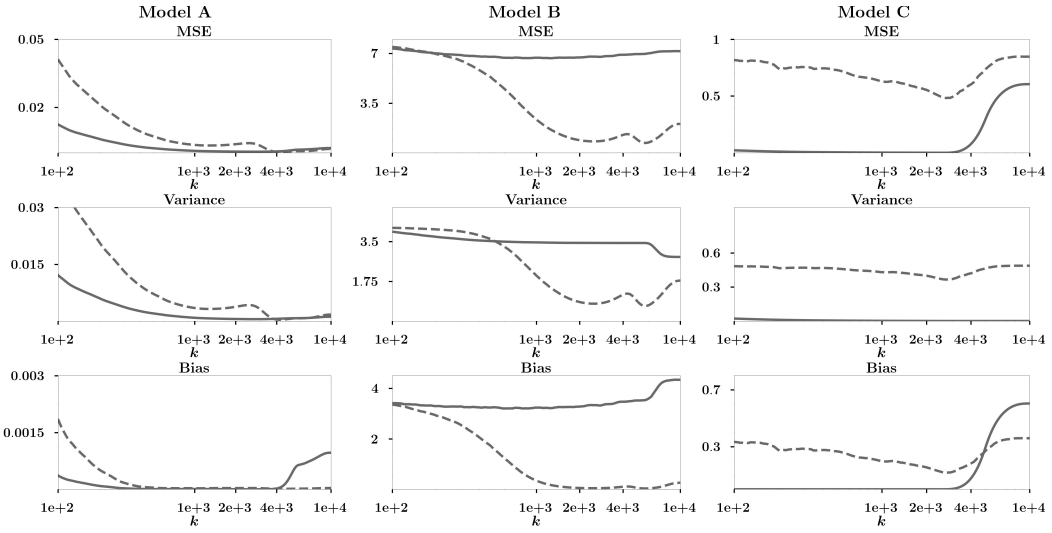

FIG 1. *Performance in terms of Frobenius norm of the error, as a function of k, with TIREX1 (solid line) and TIREX2 (dotted line), in Models A,B,C. Mean squared error, bias and variance computed over* 100 *repetitions.*

	Model A, TIREX1	Model A, Gardes (2018)	Model C, TIREX1	Model C, Gardes (2018)
$k = 464$	2.10^{-3} (2 *s*)	4.10^{-4} (6 *h*)	4.10^{-3} (2.3 *s*)	1 (4.3 *h*)
$k = 2000$	5.10^{-4} (1.7 *s*)	5.10^{-5} (6.5 *h*)	9.10^{-4} (3.2 *s*)	0.8 (8.5 *h*)

TABLE 1

MSE for TIREX and Gardes (2018)'s method in Models A and C, 100 *replications. Execution times on a standard laptop are in brackets, with h and s indicating hour and second respectively.*

6.3. Predicting tail events with TIREX on real datasets

We now investigate the relevance of TIREX as a dimension reduction tool for predicting unusually large values of Y. As explained in Remark 1, this may be viewed as a classification task: predict an exceedance $\{Y > y\}$ with the help of p covariates $X \in \mathbb{R}^p$. Reducing the dimension allows to escape the curse of the dimensionality using the projected covariates, however it generally induces a bias which may influence the (weighted) risk of an error. The most important observation in Remark 1 is that, if $Y_\infty \perp\!\!\!\perp X \mid P_e X$, the bias term vanishes in the limit $y \to y^+$. Since TIREX aims precisely at estimating P_e such that $Y_\infty \perp\!\!\!\perp X \mid P_e X$, a reasonable hope is that it would generally perform better than other dimension reduction algorithms targeting different reduction subspaces $P \neq P_e$ that would not enjoy this property.

Experimental setting.

We follow a two-steps procedure: first, run a dimension reduction algorithm (TIREX or another existing method) and project the covariates X_i on the estimated SDR space; second apply a classification algorithm to predict the event $Y_i > y$ with the help of the projected covariates. For all dimension reduction methods entering the comparison, the dimension of the reduced subspace is set to $d = 2$.

Throughout our experiments the second step is fixed: We use a nearest neighbors algorithm with a number of neighbors set to 5. In the end the performance of the competing

dimension reduction methods is measured in terms of the AM risk (3.3) and the AUC (Area under the ROC Curve) of the nearest neighbors classifier trained on the reduced covariates. The number of observations k in TIREX is selected based on 5-fold cross-validation with the AUC criterion.

Competitors.

TIREX is compared with several alternative methods using the full dataset for estimation, not only the subset associated with the largest values of the target. Namely we consider in a supervised setting the standard SDR estimates obtained with the CUME and CUVE methods introduced in Section 2. In an unsupervised setting we consider routinely used methods available in the Python Scikit-learn package Pedregosa et al. (2011), namely Principal Component Analysis (PCA), Singular Value Decomposition (SVD) which is a non-centered version of PCA, Locally Linear Embedding (LLE), and Isomap (IMP). The latter two methods are non-linear generalizations of of PCA (Roweis and Saul (2000), Tenenbaum, Silva and Langford (2000), see also Chojnacki and Brooks (2009); Bengio et al. (2003)) which are widely applied in many contexts such as data visualization Elgammal and Lee (2004); Tenenbaum, Silva and Langford (2000), or classification Vlachos et al. (2002), among others. Considering the dimensions $p \in [\![18, 103]\!]$ of the datasets described below, Gardes (2018)'s method for dimension reduction could not be included in the comparison for the algorithmic complexity reasons described above.

Data sets.

Eight datasets are used. Three of them come from the UCI repository[2]: *Residential* (372 apartment sale prices, with 103 covariates); *crime* (1994 per capita violent crimes with 122 socio-economic covariates); *Parkinsons* (5875 voice recordings along with 25 attributes). Three other datasets come from the Delve repository[3]: *Bank* (8192 rejection rates of different banks, with 32 features each); *CompAct* (8192 CPU's times with 27 covariates); *PUMA32* (8192 angular accelerations of a robot arm, with 32 attributes). Finally, two other data are obtained from the LIACC repository[4]: *Ailerons* (13750 control action on the ailerons of an aircraft with 40 attributes) and *Elevator* (16559 control action on the elevators of an aircraft with 18 attributes).

Results.

For all datasets, y is chosen equal to the 0.98-quantile of the target $(Y_i)_{i=1,\dots,n}$ except for *Residential* where the 0.90-quantile has been used to counterbalance the small sample size. The results in terms of AM risk and AUC are summarized in Tables 2 and 3 respectively. In the vast majority of cases, TIREX1 or TIREX2 performs better than the other methods. On these examples, TIREX1 is often superior to TIREX2, which indicates that the added flexibility introduced by the second order moments does not compensate for the increased variance.

[2] https://archive.ics.uci.edu
[3] http://www.cs.toronto.edu/~delve/data/datasets.html
[4] https://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html

	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.

Supplementary Material

The supplementary material placed below the bibliography contains proofs, additional examples and discussions regarding existing notions of Tail Conditional Independence, and extensions to non-standardized covariates.

References

ASENOVA, S., MAZO, G. and SEGERS, J. (2021). Inference on extremal dependence in the domain of attraction of a structured Hüsler–Reiss distribution motivated by a Markov tree with latent variables. *Extremes* **24** 461–500.

BABICHEV, D., BACH, F. et al. (2018). Slice inverse regression with score functions. *Electron. J. Stat.* **12** 1507–1543.

BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J. and TEUGELS, J. L. (2006). *Statistics of extremes: theory and applications*. John Wiley & Sons.

BENGIO, Y., PAIEMENT, J.-F., VINCENT, P., DELALLEAU, O., ROUX, N. and OUIMET, M. (2003). Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. In *NeurIPS proceedings* **16**. MIT Press.

BOUSEBATA, M., ENJOLRAS, G. and GIRARD, S. (2023). Extreme partial least-squares. *J. Multivar. Anal.* **194** 105101.

BRYC, W. (2012). *The normal distribution: characterizations with applications* **100**. Springer Science & Business Media.

CHAUTRU, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electron. J. Stat.* **9** 383–418.

Chiapino, M. and Sabourin, A. (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *NFMCP workshop proceedings* 132–147. Springer.

Chiapino, M., Sabourin, A. and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes* **22** 193–222.

Chiapino, M., Clémençon, S., Feuillard, V. and Sabourin, A. (2020). A multivariate extreme value theory approach to anomaly clustering and visualization. *Comput. Stat.* **35** 607–628.

Chojnacki, W. and Brooks, M. J. (2009). A note on the locally linear embedding algorithm. *Intern. J. Pattern Recognit. Artif. Intell.* **23** 1739–1752.

Christou, E. (2020). Central quantile subspace. *Stat. Comput.* **30** 677–695.

Constantinou, P. and Dawid, A. P. (2017). Extended conditional independence and applications in causal inference. *Ann. Stat.* **45** 2618–2653.

Cook, R. D. (2009). *Regression graphics: Ideas for studying regressions through graphics* **482**. John Wiley & Sons.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Stat.* **30** 455–474.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Am. Stat. Assoc.* **100** 410–428.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *J. Am. Stat. Assoc.* **86** 328–332.

Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika* **106** 587–604.

Csorgo, M., Csorgo, S., Horváth, L. and Mason, D. M. (1986). Weighted empirical and quantile processes. *Ann. Probab.* **14** 31–85.

Dalalyan, A. S., Juditsky, A. and Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. *J. Mach. Learn. Res.* **9** 1647–1678.

Dawid, A. P. (1979). Conditional independence in statistical theory. *J. R. Stat. Soc. Series B Stat. Methodol.* **41** 1–15.

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

Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Stat. Plan.* **136** 730–769.

Drees, H. and Sabourin, A. (2021). Principal component analysis for multivariate extremes. *Electron. J. Stat.* **15** 908–943.

Eaton, M. L. (1986). A characterization of spherical distributions. *J. Multivar. Anal.* **20** 272–276.

Einmahl, J. and Mason, D. (1988). Strong limit theorems for weighted quantile processes. *Ann. Probab.* **16** 1623–1643.

Elgammal, A. and Lee, C.-S. (2004). Inferring 3D Body Pose from Silhouettes Using Activity Manifold Learning. In *CVPR'04 proceedings* 681–688. IEEE Computer Society, USA.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *J. R. Stat. Soc. Series B Stat. Methodol.* **82** 871–932.

Engelke, S. and Ivanovs, J. (2021). Sparse Structures for Multivariate Extremes. *Annu. Rev. Stat. Appl.* **8** 241–270.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33** 3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *J. Financ. Econ.*

116 1–22.

Fermanian, J.-D., Radulovic, D. and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli* **10** 847–860.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **5** 73–99.

Fukumizu, K., Bach, F. R., Jordan, M. I. et al. (2009). Kernel dimension reduction in regression. *Ann. Stat.* **37** 1871–1905.

Gardes, L. (2018). Tail dimension reduction for extreme quantile estimation. *Extremes* **21** 57–95.

Goix, N., Sabourin, A. and Clémençon, S. (2016). Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. In *AISTATS proceedings* **51** 75–83. PMLR.

Goix, N., Sabourin, A. and Clémençon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *J. Multivar. Anal.* **161** 12–31.

Hall, P. and Li, K.-C. (1993). On almost Linearity of Low Dimensional Projections from High Dimensional Data. *Ann. Stat.* **21** 867–889.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Am. Stat. Assoc.* **84** 986–995.

Hitz, A. and Evans, R. (2016). One-component regular variation and graphical modeling of extremes. *J. Appl. Probab.* **53** 733–746.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *Br. J. Stat. Psychol.* **10** 69–79.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Stat.* **29** 1537–1566.

Janssen, A. and Wan, P. (2020). k-means clustering of extremes. *Electron. J. Stat.* **14** 1211–1233.

Jenatton, R., Audibert, J.-Y. and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.* **12** 2777–2824.

Jessen, H. A. and Mikosch, T. (2006). Regularly varying functions. *Publications de L'Institut Mathematique* **80** 171–192.

Jiang, Y., Cooley, D. and Wehner, M. F. (2020). Principal Component Analysis for Extremes and Application to US Precipitation. *J. Clim.* **33** 6441–6451.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **86** 316–327.

Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131.

Menon, A., Narasimhan, H., Agarwal, S. and Chawla, S. (2013). On the statistical consistency of algorithms for binary classification under class imbalance. In *ICML proceedings* 603–611. PMLR.

Meyer, N. and Wintenberger, O. (2021). Sparse regular variation. *Adv. Appl. Probab.* **53** 1115–1148.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830.

Portier, F. (2016). An Empirical Process View of Inverse Regression. *Scand. J. Stat. Theory Appl.* **43** 827–844.

Portier, F. and Delyon, B. (2013). Optimal transformation: A new approach for covering

the central subspace. *J. Multivar. Anal.* **115** 84–107.

Portier, F. and Delyon, B. (2014). Bootstrap testing of the rank of a matrix via least-squared constrained estimation. *J. Am. Stat. Assoc.* **109** 160–172.

Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.

Resnick, S. I. (2013). *Extreme values, regular variation and point processes*. Springer.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

Segers, J. (2015). Hybrid copula estimators. *J. Stat. Plan.* **160** 23–34.

Simpson, E. S., Wadsworth, J. L. and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika* **107** 513–532.

Tenenbaum, J. B., Silva, V. d. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science* **290** 2319–2323.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation. Quantitative Applications in the Social Sciences* **47**. Sage.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge university press.

Van Der Vaart, A. W. and Wellner, J. A. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G. and Koudas, N. (2002). Non-linear dimensionality reduction techniques for classification and visualization. In *ACM SIGKDD proceedings* 645–651.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis; proceedings* (P. Krishnaiah, ed.) 391–420. Academic press.

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *J. Comput. Graph. Stat.* **17** 590–610.

Xia, Y. et al. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Stat.* **35** 2654–2690.

Xu, Z., Dan, C., Khim, J. and Ravikumar, P. (2020). Class-Weighted Classification: Trade-offs and Robust Approaches. In *ICML proceedings* 10544–10554. PMLR.

Yeh, Y.-R., Huang, S.-Y. and Lee, Y.-J. (2008). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans. Knowl. Data. Eng.* **21** 1590–1603.

Zhu, L.-P., Zhu, L.-X. and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Am. Stat. Assoc.* **105** 1455–1466.

Zwald, L. and Blanchard, G. (2005). On the convergence of eigenspaces in kernel principal component analysis. *NeurIPS proceedings* **18**.

Tail Inverse Regression: dimension reduction for prediction of extremes. Supplementary material.

This supplement contains proofs, additional examples and discussions regarding existing notions of Tail Conditional Independence, and extensions to non-standardized covariates. Section and equation numbers in the supplement start with a letter, to distinguish them from those in the paper.

Appendix A: Proofs for Remark 1

In this section, for the sake of completeness, we prove two facts regarding classification with the AM risk in the *full problem* defined in Remark 1 from the main paper. First the classifier

$$h^*(x) = \mathbb{1}\{\eta(x) > \pi\} \tag{A.1}$$

is a minimizer of the AM risk ; Second, the associated Bayes risk is given by

$$\mathcal{R}_{\mathrm{AM}}(h^*) = \mathbb{E}\left[\min\Big(\frac{\eta(X)}{\pi}, \frac{1-\eta(X)}{1-\pi}\Big)\right]. \tag{A.2}$$

We introduce the AM loss function

$$\ell_{\mathrm{AM}}(\hat{t}, t) = \frac{1}{1-\pi}\mathbb{1}\{\hat{t}=1, t=0\} + \frac{1}{\pi}\mathbb{1}\{\hat{t}=0, t=1\}$$

so that for any classifier, $\mathcal{R}_{\mathrm{AM}}(h) = \mathbb{E}\left[\ell_{\mathrm{AM}}(h(X), T)\right]$. Consider now the conditional AM risk

$$\widetilde{\mathcal{R}_{\mathrm{AM}}}(h, x) = \mathbb{E}\left[\ell_{\mathrm{AM}}(h(X), T) \mid X = x\right],$$

thus $\mathcal{R}_{\mathrm{AM}}(h) = \mathbb{E}\left[\widetilde{\mathcal{R}_{\mathrm{AM}}}(h, X)\right]$. We also have

$$\begin{aligned}\widetilde{\mathcal{R}_{\mathrm{AM}}}(h, x) &= \frac{1}{1-\pi}\mathbb{1}\{h(x)=1\}(1-\eta(x)) + \frac{1}{\pi}\mathbb{1}\{h(x)=0\}\eta(x) \\ &= \frac{1-\eta(x)}{1-\pi} + \mathbb{1}\{h(x)=0\}\Big[\frac{\eta(x)}{\pi} - \frac{1-\eta(x)}{1-\pi}\Big].\end{aligned} \tag{A.3}$$

Also, the classifier in (A.1) may be written equivalently as $h^*(x) = \mathbb{1}\Big\{\frac{\eta(x)}{\pi} > \frac{1-\eta(x)}{1-\pi}\Big\}$. Thus for any classifier h, we may write the difference in conditional risks as

$$\begin{aligned}\widetilde{\mathcal{R}_{\mathrm{AM}}}(h, x) - \widetilde{\mathcal{R}_{\mathrm{AM}}}(h^*, x) &= \frac{\eta-\pi}{\pi(1-\pi)}\left[\mathbb{1}\{h(x)=0\} - \mathbb{1}\{h^*(x)=0\}\right] \\ &= \left|\frac{\eta-\pi}{\pi(1-\pi)}\right| \mathbb{1}\{h(x) \neq h^*(x)\}\end{aligned}$$

The latter display is nonnegative, which shows that h^* defined in (A.1) indeed minimizes the AM risk. Turning to our second claim, notice that we may write, using (A.3),

$$\begin{aligned}\widetilde{\mathcal{R}_{\mathrm{AM}}}(h^*, x) &= \begin{cases}\eta(x)/\pi & \text{if } \eta(x)/\pi > (1-\eta(x))/(1-\pi) \\ (1-\eta(x))/(1-\pi) & \text{otherwise}\end{cases} \\ &= \min\Big(\frac{\eta(x)}{\pi}, \frac{1-\eta(X)}{1-\pi}\Big).\end{aligned}$$

This proves (A.2).

Appendix B: Proofs for Section 3.2 and additional comments

In this section we provide the full proofs regarding our examples and counter-examples from Section 3.2 regarding the generic mixture model. On this occasion we conduct a thorough comparison between the two definitions of tail conditional independence TCI and TCI-G, see Equations 3.1 and 3.2 in the main paper. For convenience write $S(y) = \mathbb{P}(Y > y); S(y, W) = \mathbb{P}(Y > y|W); S(y, W, V) = \mathbb{P}(Y > y|W, V)$. The relevant quantities are respectively the ratios

$$R(y, V, W) = \frac{S(y, V, W) - S(y, W)}{S(y)} \ , \text{ and } \tilde{R}(y, V, W) = \frac{S(y, V, W) - S(y, W)}{S(y, W)}. \tag{B.1}$$

The TCI condition is that $\mathbb{E}|R(y, V, W)| \to 0$ as $y \to y^+$, whereas TCI-G means that $\tilde{R}(y, V, W) \to 0$ as $y \to y^+$, almost surely. Notice already that our criterion (3.1) is an integrated version of (3.2), with a weight function

$$\rho(y, W) = S(y, W)/S(y), \tag{B.2}$$

such that $\rho(y, W) \geq 0$ and $\mathbb{E}\left[\rho(y, W)\right] = 1$ for all y. Namely, TCI means that

$$\mathbb{E}\left|\tilde{R}(y, V, W)\rho(y, W)\right| \xrightarrow[y\to y^+]{} 0 \tag{B.3}$$

B.1. Additional notations regarding the generic mixture model from Section 3.2

We introduce in the context of Section 3.2 from the main paper the additional notations

$$S_1(y) = \mathbb{P}(Y_1 > y) = \int S_1(y, v)\, \mathrm{d}P_V(v) \ , \quad S_2(y) = \mathbb{P}\left(Y_2 > y\right) = \int S_2(y, w)\, \mathrm{d}P_W(w).$$

With these notations, using the independence assumption regarding the pair (V, W) we may write

$$S(y, v, w) = \theta S_1(y, v) + (1 - \theta)S_2(y, w) \ ; \quad S(y, w) = \theta S_1(y) + (1 - \theta)S_2(y, w) \ ;$$
$$S(y) = \theta S_1(y) + (1 - \theta)S_2(y).$$

Thus, the ratios R, $\tilde{R}$ defined at the beginning of this section and involved in TCI and TCI-G write respectively

$$R(y, v, w) = \frac{\theta(S_1(y, v) - S_1(y))}{\theta S_1(y) + (1 - \theta)S_2(y)} \, , \qquad \tilde{R}(y, v, w) = \frac{\theta(S_1(y, v) - S_1(y))}{\theta S_1(y) + (1 - \theta)S_2(y, w)} \, . \tag{B.4}$$

Notice already that

$$|R(y, v, w)| \leq \frac{\theta}{1 - \theta} \frac{S_1(y, v) + S_1(y)}{S_2(y)} \, , \tag{B.5}$$

$$|\tilde{R}(y, v, w)| \leq \frac{\theta}{1 - \theta} \left(\frac{S_1(y, v)}{S_2(y, w)} + \int \frac{S_1(y, v')}{S_2(y, w)}\, \mathrm{d}P_V(v') \right) . \tag{B.6}$$

Finally, specializing to the case where Y_1 and Y_2 follow the mixture model described in the same section of the main paper, the conditional survival functions for Y_1, Y_2 are, for $y > b$,

$$S_1(y,v) = \sum_{i=1}^{p-d} \mathbb{1}\{v_i > 0\}\pi_i^1 S_\epsilon(y/v_i)\ , \qquad S_2(y,w) = \sum_{j=1}^{d} \mathbb{1}\{w_j > 0\}\pi_j^2 S_\zeta(y/w_j) \tag{B.7}$$

We now discuss the main differences between the two definitions. Natural questions to ask are *(i)* whether one definition is more appropriate than the other depending on the context ; *(ii)* whether one condition is stronger than the other, possibly under additional assumptions.

As for Question *(i)*, in spirit, as reflected by the equivalent condition (B.3), TCI is comparatively more sensitive to values $W = w$ such that the conditional probability of an exceedance $Y > y$ is large, which is an appealing feature for identifying tail risk factors as described in the introduction. On the other hand, one advantage of TCI-G's scaling is that the ratio $\tilde{R}$ introduced at the beginning of this section is a *relative* deviation, which is arguably easily interpretable. However TCI-G's criterion takes into account *all* possible values w, even those such that the conditional distribution of Y given $W = w$ is shorter tailed than the marginal distribution of Y. The focus in TCI-G is not exactly on the tail of Y's distribution, but rather on the tails of the conditional distributions of Y given W.

Before turning to Question *(ii)*, we discuss the differences between the two conditions in terms of mode of convergence.

B.2. Convergence almost-surely or in expectation in TCI-G *or* TCI

Almost sure convergence $\tilde{R}(y,V,W) \to 0$ as $y \to y^+$ implies $\mathbb{E}|\tilde{R}(y,V,W)| \to 0$. Indeed by conditioning on W, we have $\mathbb{E}\,[\tilde{R}(y,V,W)] = 0$ so that, denoting by z_+ (resp. z_-) the negative (resp. positive) part of a real z, it holds that $\mathbb{E}\,[\tilde{R}(y,V,W)_+] = \mathbb{E}\,[\tilde{R}(y,V,W)_-]$. As a consequence

$$\mathbb{E}\,|\tilde{R}(y,V,W)| = 2\mathbb{E}\,[\tilde{R}(y,V,W)_-].$$

However for all y, v, w, $\tilde{R}(y,v,w) \geq -1$ so that $0 \leq \tilde{R}(y,V,W)_- \leq 1$. By dominated convergence, if (3.2) holds, then also $\mathbb{E}\,[\tilde{R}(y,V,W)_-] \to 0$ and the above display implies that $\mathbb{E}\,|\tilde{R}(y,V,W)| \to 0$ as well. This argument is not valid regarding the tail behaviour of $R(y,V,W)$ because it is not true that $R(y,V,W) \geq -1$ almost surely.

We are now ready to examine Question *(ii)*, that is, whether one condition (TCI or TCI-G) implies the other, in general or under simplifying assumptions.

B.3. Special case: discrete covariates with finite support

In order to build up the intuition, consider the special case where the covariates have a finite support. This is a sensible assumption for real life applications where observations are discretized.

We thus consider here finitely supported covariates $V \in \{v_1, \ldots, v_m\}$, $W \in \{w_1, \ldots, w_n\}$. Denote $p(v_i) = \mathbb{P}(V = v_i), p(w_j) = \mathbb{P}(W = w_j), p(v_i, w_j) = \mathbb{P}(V = v_i, W = w_j)$. Assume for simplicity that for all $(i,j) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$, we have $p(v_i, w_j) > 0$.

First, in this case, almost sure convergence and convergence in expectation are equivalent for both ratios R and $\tilde{R}$ introduced at the beginning of this section. In other words

$$\mathbb{E}|R(y,V,W)| \xrightarrow[y\to y^+]{} 0 \iff |R(y,V,W)| \xrightarrow[y\to y^+]{} 0, \text{ almost surely ;} \tag{B.8}$$

$$\mathbb{E}|\tilde{R}(y,V,W)| \xrightarrow[y\to y^+]{} 0 \iff |\tilde{R}(y,V,W)| \xrightarrow[y\to y^+]{} 0, \text{ almost surely .} \tag{B.9}$$

Indeed

$$\mathbb{E}|R(y,V,W)| = \sum_{i=1}^{m}\sum_{j=1}^{n} p(v_i,w_j)\Big|\frac{S(y,v_i,w_j)-S(y,w_j)}{S(y)}\Big|.$$

The latter display converges to 0 as $y \to y^+$ if and only if each terms in the finite summation does, that is, if and only if $\forall (i,j), R(y,v_i,w_j) \to 0$ as $y \to y^+$. This proves (B.8), and the argument for (B.9) is similar.

Second, TCI-G implies TCI, meaning that our definition is weaker than Gardes (2018)'s in this discrete setting. To see this, in view of the equivalence between L^1 and almost sure convergences, it is enough to show that the ratio $R(y,v_i,w_j)/\tilde{R}(y,v_i,w_j)$ is uniformly upper bounded when y, i and j vary. However for all (y,i,j),

$$\frac{R(y,v_i,w_j)}{\tilde{R}(y,v_i,w_j)} = \rho(y,w_j) = S(y,w_j)/S(y) = \frac{S(y,w_j)}{\sum_{k=1}^{n} p(w_k)S(y,w_k)} \le \frac{1}{p(w_j)} \le 1/\min_{k\le n} p(w_k) < \infty.$$

As a consequence, if $\tilde{R}(y,V,W) \to 0$ almost surely, then also $R(y,V,W) \to 0$ almost surely as $y \to y^+$ and the result follows.

B.4. Example in the mixture model where both TCI *and* TCI-G *hold*

We consider the setting of Section 3.2 from the main paper, and in particular the case where the lower bound of the support of each W_j is positive, $a > 0$.

We verify that the upper bounds (B.5) and (B.6) uniformly converge to 0. First, using (B.7), we have

$$\frac{S_1(y)+S_1(y,v)}{S_2(y)} \le 2\frac{\sup_{v\in[a,b]^{p-d}} S_1(y,v)}{\inf_{w\in[a,b]^d} S_2(y,w)} \le 2\frac{S_\epsilon(y/b)}{S_\zeta(y/a)},$$

where the right-hand-side converges to 0 as $y \to \infty$ under Condition (3.7). Thus the upper bound in (B.5) uniformly converges to 0 and TCI holds by dominated convergence.

Turning to $\tilde{R}$, we also have

$$\sup_{(v,w)\in[a,b]^p} \frac{S_1(y,v)}{S_2(y,w)} \le \frac{\sup_{v\in[a,b]^{p-d}} S_1(y,v)}{\inf_{w\in[a,b]^d} S_2(y,w)} \le \frac{S_\epsilon(y/b)}{S_\zeta(y/a)} \xrightarrow[y\to\infty]{} 0.$$

Thus, by dominated convergence the right-hand-side of (B.6) converges to 0 as $y \to \infty$ so that TCI-G holds as well.

In the general case the situation is much more complex and it turns out that neither condition implies the other, as revealed by the counter-examples constructed in the next two subsections.

B.5. Counter-example in the mixture model where TCI holds but TCI-G does not

In contrast to the latter subsection, we now consider the case where the support of the W_j's includes 0, so that $a = 0$. Namely we take each variable V_j, W_j following a binary Bernoulli distribution with parameter $\tau \in (0,1)$Thus $\mathbb{P}(W = (0,\dots,0)) = (1-\tau)^d > 0$. Notice already that the right-hand side of (B.6) is not bounded because $S_2(y, w = (0,\dots,0)) = 0$ for $y > 0$. Also, from (B.4),

$$\tilde{R}(y, v, w = (0,\dots,0)) = \frac{S_1(y,v) - S_1(y)}{S_1(y)}.$$

In this specific example Y_1 and Y_2 have point masses at 0 and we have for $y > 0$, $S_1(y) = \sum_j \pi_j^1 \tau S_\epsilon(y) = \tau S_\epsilon(y)$ while for $v_1 = (1,\dots,1)$, $S_1(y, v_1) = \sum_j \pi_j^1 S_\epsilon(y) = S_\epsilon(y)$. Thus in the above display, $\tilde{R}(y, v_1, 0) = (1-\tau)/\tau$ for all $y > 1$ and TCI-G does not hold.

Finally we show that TCI holds by examining the right-hand side of (B.5). The argument above shows that

$$\frac{S_1(y,v) + S_1(y)}{S_2(y)} \leq \frac{(1+\tau)S_\epsilon(y)}{\tau S_\zeta(y)}\,.$$

This proves uniform convergence to 0 in (B.5) under Condition (3.7) and concludes the argument.

B.6. Counter-example where TCI-G holds but TCI does not

In this example we depart from the mixture model forming the basis of the two latter examples. The idea behind is to build the survival functions in such a way that $\limsup_{y\to\infty} \rho(y, W) = \infty$ (see (B.2) for the definition of ρ), with probability one, while TCI-G holds.

In addition to the notations introduced at the beginning of this section, we introduce the ratio

$$q(y, v, w) = S(y, v, w)/S(y, w).$$

Thus $\tilde{R}(y, v, w) = q(y, v, w) - 1$ and $R(y, v, w) = (q(y, v, w) - 1)\rho(y, w)$. We denote respectively by $P_W, P_{V,W}$ the marginal distribution of W and the joint distribution of (V, W). Here we define V, W as independent uniform variables, $P_W = P_V = \mathcal{U}_{[-1/2,1/2]}$ and $P_{V,W} = P_V \otimes P_W$. We shall build (S, ρ, q) such that $\mathrm{h}|q(y, V, W) - 1| \to 0$ as $y \to \infty$, almost surely, (so that TCI-G holds) while $\limsup \mathbb{E}\left[|q(y, V, W) - 1|\rho(y, W)\right] > 0$ as $y \to \infty$ (so that TCI does not hold).

The functions $S(y), q(y, v, w), \rho(y, w)$ define a joint distribution of (Y, V, W) with no mass at the right end point of Y if conditions (B.10) (B.11) and (B.12) below hold.

$$S \text{ is non-increasing}\,, \quad \lim_{y\to y^+} S(y) = 0, \quad S(y) \geq 0; \tag{B.10}$$

$$P_W\text{-almost surely, the function } y \mapsto \rho(y, W)S(y) \text{ is non-increasing, and}$$
$$\lim_{y\to y^+} \rho(y, W)S(y) = 0, \quad \rho(y, W) \geq 0, \quad \mathbb{E}\left[\rho(y, W)\right] = 1\,, \forall y; \tag{B.11}$$

$$P_{V,W}\text{-almost surely, the function } y \mapsto q(y, V, W)\rho(y, W)S(y) \text{ is non-increasing, and}$$
$$\lim_{y\to y^+} q(y, V, W)\rho(y, W)S(y) = 0, \quad q(y, V, W) \geq 0, \quad \mathbb{E}\left[q(y, V, W) \mid W\right] = 1\,, \forall y. \tag{B.12}$$

B.6.1. Construction of $S(y), \rho(y,w)$

We let $S(y) = e^{-y}$, $y \geq 0$, and we construct ρ such that $\mathbb{P}\left(\limsup_{y\to\infty} \rho(y,w) = \infty\right) = 1$ while (B.11) is satisfied. To this end define for $n \geq 2$, and $0 \leq j \leq n$,

$$L_n = \sum_{k<n, k\geq 2} k^2 \; ; \qquad L_{n,j} = L_n + jn. \tag{B.13}$$

Thus $L_2 = 0, L_3 = 4$, $L_n \leq n^3$ for $n \geq 2$ and $L_{n,n} = L_{n+1}$. Also $\mathbb{R}_+ = \sqcup_{n\geq 2} \sqcup_{0\leq j<n} [L_{n,j}, L_{n,j+1})$. Also for $y \geq 0$, we denote by $(n(y), j(y))$ the unique pair of integers such that $y \in [L_{n,j}, L_{n,j+1})$.

For $n \geq 2, 0 \leq j < n$, we define $\rho(y,w)$ for $y \in [L_{n,j}, L_{n,j+1})$ and $w \in [-1/2, 1/2]$ as follows: let $I_{n,j} = [1/2 - (j+1)/n, 1/2 - j/n]$, then

$$\rho(y,w) = 1 + w + \frac{n}{4\pi} \sin\Big(\pi(y - L_{n,j})/n\Big) \left[\mathbb{1}\{w \in I_{n,j}\} - \mathbb{1}\{w \notin I_{n,j}\}/(n-1)\right]. \tag{B.14}$$

Notice that for all $w \in [-1/2, 1/2]$, the function $y \mapsto \rho(y,w)$ is continuous. Also for all $w \in [-1/2, 1/2]$ we have $\limsup_y \rho(y,w) = +\infty$. Indeed for any fixed $n \geq 2$, let j such that $w \in I_{n,j}$. Then letting

$$y_n = L_{n,j} + n/2,$$

we have $\rho(y_n, w) = w + n/(4\pi) \geq n/(4\pi) - 1/2$. The sequence y_n converges to ∞ and is such that $\rho(y_n, w) \to \infty$ as $n \to \infty$, which proves the claim.

We now verify that the conditions gathered in (B.11) hold.

1. First for all $y \geq 0$,

$$\mathbb{E}\left[\rho(y, W)\right] = 1 + \mathbb{E}\left[W\right] + \frac{n}{4\pi} \sin\Big(\pi(y - L_{n,j})/n\Big) \left[1/n - (n-1)/(n(n-1))\right] = 1.$$

2. We show that for all y, $\rho(y, W) \geq 1/3$ almost surely. By construction, $\rho(y, W) \geq 1/2 - \frac{n(y)}{4(n(y)-1)\pi}$. Since $m/(m-1) \leq 2$ for $m \geq 2$, we obtain

$$\rho(y, W) \geq 1/2 - \frac{2}{4\pi} \geq 1/2 - 1/6 = 1/3.$$

3. We now show that $y \mapsto S(y)\rho(y,w)$ is non increasing for all $w \in [-1/2, 1/2]$. Since both S and ρ are continuous functions of y, with derivatives from the right which we denote respectively $S'(y)$ and $\rho'(y,w)$, we need to show that $\rho'(y,w) < -\rho(y,w)S'(y)/S(y)$. Here $S'(y)/S(y) = -1$, and from the above point we obtain $-\rho(y,w)S'(y)/S(y) \geq 1/3$. To conclude we show that

$$\forall y \geq 0, w \in [-1/2, 1/2], \rho'(y,w) \leq 1/4.$$

   Let $y > 0$ and $(n, j) = (n(y), j(y))$ as above. On the one hand if $w \in I_{n(y),j(y)}$ we have $0 \leq \rho'(y, W) \leq 1/4$. On the other hand if $w \notin I_{n(y),j(y)}$, we have $\rho'(y,w) < 0$. In both cases $\rho'(y,w) \leq 1/4 \leq 1/3 \leq -\rho(y,w)S'(y)/S(y)$, which conludes the argument.
4. Finally we verify that $\lim_{y\to\infty} \rho(y, W)S(y) = 0$, almost surely. To see this, notice that for all $y > 0, w \in [-1/2, 1/2], |\rho(y,w)| \leq 3/2 + \frac{n(y)}{4\pi}$. Now since $L_n \geq n^2$, $\{n : L_n \leq y\} \subset \{n : n^2 \leq y\}$, so that $n(y) = \sup\{n : L_n \leq y\} \leq \sup\{n : n^2 \leq y\} \leq \sqrt{y}$. Thus $|\rho(y,w)|e^{-y} \leq (3/2 + \sqrt{y}/(4\pi))e^{-y} \to 0$ as $y \to \infty$.

B.6.2. Construction of $q(y, v, w)$

Recall $n(y)$ from the beginning of the above paragraph. Define

$$q(y,v,w) = 1 + v\left[\mathbb{1}\Big\{w > \frac{1}{2} - \frac{1}{n(y)}\Big\} + \mathbb{1}\Big\{w \le \frac{1}{2} - \frac{1}{n(y)}\Big\} \exp\Big(-\frac{y + \left\lceil \frac{1}{1/2-w} \right\rceil}{4}\Big)\right] \tag{B.15}$$

We now verify that the function $y \mapsto q(y, v, w)S(y, w) = S(y, v, w)$ is non increasing. Notice already that all the other constraints gathered in (B.12) are satisfied. Since for fixed (v, w), both $y \mapsto q(y, v, w)$ and $y \mapsto S(y, w)$ are continuous, it is enough to verify that the derivative from the right of $y \mapsto q(y, v, w)S(y, w)$ is negative or null, that is (since $q \ge 1/2$ is positive), we need to ensure that

$$\frac{q'(y,v,w)}{q(y,v,w)} \le -\frac{S'(y,w)}{S(y,w)}. \tag{B.16}$$

With our definition of $S(y, w)$ from Subsection B.6.1,

$$-S'(y,w)/S(y,w) = (\rho(y,w) - \rho'(y,w))/\rho(y,w) = 1 - \rho'(y,w)/\rho(y,w) \ge 1 - \frac{1/4}{1/3} = 1/4.$$

If we denote $y(w) = y - \left\lceil \frac{1}{1/2-w} \right\rceil$, we have

$$q'(y,v,w)/q(y,v,w) = -\frac{1}{4}\mathbb{1}\Big\{w \le \frac{1}{2} - \frac{1}{n(y)}\Big\} v \exp(-y(w)/4)/(1 + v\exp(-y(w)/4)).$$

The above display is always less than $1/4$ so that (B.16) holds for all $y > 0$ and $v, w \in [-1/2, 1/2]$ and (B.12) is satisfied. This fact combined with the argument in Subsection B.6.1 implies that the functions (S, ρ, q) define a proper joint distribution for (Y, V, W).

B.6.3. Conclusion

We have constructed a joint distribution for (Y, V, W) in Sections B.6.1, B.6.2, such that $P_{V,W}$-almost surely, $q(y, V, W) \to 1$ as $y \to \infty$, as can be seen immediately from the definition of q in (B.15). Thus (Y, V, W) satisfy TCI-G. However, for all $n \ge 0$, let $y_n = L_n + n/2$ (see Subsection B.6.1), so that by construction $n(y_n) = n$. Notice that $I_{n,0} = [1/2 - 1/n, 1/2]$ and for $w \in I_{n,0}$, we have $\rho(y_n, w) = 1 + w + n/(4\pi) \ge n/16$ and $q(y_n, v, w) = 1 + v$. Thus

$$\begin{aligned}
\mathbb{E}\left[|R(y_n, V, W)|\right] &= \mathbb{E}\left[|q(y_n, V, W) - 1|\rho(y_n, W)\right] \\
&\ge \mathbb{E}\left[|q(y_n, V, W) - 1|\rho(y_n, W)\mathbb{1}\{W > 1/2 - 1/n\}|\right] \\
&= \mathbb{P}\left(W > 1/2 - 1/n\right)\mathbb{E}\left[|q(y_n, V, W) - 1|\rho(y_n, W) \mid W \in I_{n,0}\right] \\
&\ge \frac{1}{n}\mathbb{E}\left[|V|n/16\right] \ge \mathbb{E}\left[|V|\right]/16 = 1/64.
\end{aligned}$$

We have shown that $\limsup_{y\to\infty} \mathbb{E}\left[|R(y, V, W)|\right] > 0$, so that TCI does not hold, which concludes the counter-example.

B.7. Additive Mixture Model (Remark 2 in the main paper)

We end this section devoted to examples with a full derivation of the additive mixture example mentioned in Remark 2 from the main paper. We consider here an additive mixture $Y = Y_1 + Y_2$. The first (light-tailed) component is $Y_1 = V \in [a, b]$ $(-\infty < a < b < \infty)$; and the second (heavy-tailed) one is $Y_2 = W\xi$ where $W \in [c, d]$ with $0 < c < d < \infty$, and ξ has a continuous survival function $S_\xi(y) := 1 - F_\xi(y)$ satisfying $q(y) = y^\alpha S_\xi(y) \to C$ as $y \to \infty$, for some $\alpha, C > 0$. In addition, we assume that V and W are independent.

We show that TCI holds, that is $Y_\infty \perp\!\!\!\perp V \mid W$. Introducing the function

$$g(v, w, y) = w^{-\alpha} y^\alpha S_\xi[(y - v)/w],$$

we have that

$$\begin{aligned} &\sup_{[v,w]\in[a,b]\times[c,d]} \big|g(v, w, y) - C\big| \\ &= \sup_{v,w\in[a,b]\times[c,d]} \left|\Big(1 - \frac{v}{y}\Big)^{-\alpha}\Big(\frac{y - v}{w}\Big)^{\alpha} S_\xi\Big[\frac{y - v}{w}\Big] - c\right| \xrightarrow[y\to\infty]{} 0. \end{aligned} \tag{B.17}$$

The last limit relation follows from $\frac{y-v}{w} \to \infty$ and $1 - v/y \to 1$, uniformly for $v, w \in [a, b] \times [c, d]$ as $y \to \infty$. We have that

$$\begin{aligned} \mathbb{P}(Y > y|V, W) &= S_\xi((y - V)/W) = W^\alpha y^{-\alpha} g(V, W, y) \\ \mathbb{P}(Y > y|W) &= W^\alpha y^{-\alpha} \int_a^b g(v, W, y) f_1(v)\, \mathrm{d}v, \\ \mathbb{P}(Y > y) &= y^{-\alpha} \int_c^d \int_a^b w^\alpha g(v, w, y) f_1(v) f_2(w)\, \mathrm{d}v\, \mathrm{d}w. \end{aligned}$$

Thus

$$\frac{\mathbb{P}\left(Y > y|X\right) - \mathbb{P}\left(Y > y|W\right)}{\mathbb{P}\left(Y > y\right)} = \frac{W^\alpha \left\{g(v, W, y) - \int_a^b g(v, W, y) f_1(v)\, \mathrm{d}v\right\}}{\int_c^d \int_a^b w^\alpha g(v, w, y) f_1(v) f_2(w)\, \mathrm{d}v\, \mathrm{d}w}$$

By (B.17) and dominated convergence, $\int_c^d \int_a^b g(v, w, y) f_1(v) f_2(w)\, \mathrm{d}v\, \mathrm{d}w \to c\mathbb{E}(W^\alpha)$ as $y \to \infty$. Regarding the numerator, Cauchy's inequality implies that

$$\begin{aligned} &\mathbb{E}\Big|W^\alpha \left\{g(v, W, y) - \int_a^b g(v, W, y) f_1(v)\, \mathrm{d}v\right\}\Big| \\ \leq &\sqrt{\mathbb{E}W^{2\alpha}} \sqrt{\mathbb{E}\left\{g(v, W, y) - \int_a^b g(v, W, y) f_1(v)\, \mathrm{d}v\right\}^2}. \end{aligned}$$

The right-hand side tends to zero by noting that $\mathbb{E}W^{2\alpha} < \infty$ and applying the dominated convergence theorem twice to the second term. The proof is complete.

Appendix C: Proof of Theorem 2

We need to show that

$$Q_e \mathbb{E}\left[ZZ^\top - I \;\middle|\; Y > y\right] \xrightarrow[y\to y^+]{} 0. \tag{C.1}$$

Notice first that from (LC) (2.1) and (CCV) (2.2) it holds that $Q_e(\mathbb{V}\mathrm{ar}\left[Z \mid P_e Z\right] - I_p) = -Q_e P_e = 0$. Thus also

$$Q_e \mathbb{E}\left[ZZ^\top - I_p \mid P_e Z\right] = Q_e(\mathbb{V}\mathrm{ar}\left[Z \mid P_e Z\right] - I_p) + Q_e E[Z \mid P_e Z] E[Z \mid P_e Z]^\top = Q_e P_e = 0.$$

As a consequence

$$\begin{aligned}
&Q_e \mathbb{E}\left[(ZZ^\top - I_p)\mathbb{1}\{Y > y\}\right] = Q_e \mathbb{E}\left[\mathbb{E}\Big((ZZ^\top - I_p)\mathbb{1}\{Y > y\} \mid P_e Z, Y\Big)\right] \\
&= Q_e \mathbb{E}\Big[\Big(\mathbb{E}\left[ZZ^\top - I_p \mid P_e Z, Y\right] - \mathbb{E}\left[ZZ^\top - I_p \mid P_e Z\right]\Big)\mathbb{1}\{Y > y\}\Big] \\
&= Q_e \mathbb{E}\Big[\Big(\mathbb{E}\left[ZZ^\top \mid P_e Z, Y\right] - \mathbb{E}\left[ZZ^\top \mid P_e Z\right]\Big)\mathbb{1}\{Y > y\}\Big]
\end{aligned}$$

Thus in order to show (C.1) it is sufficient to show that for all pair $(i,j) \in \{1,\dots,p\}^2$, writing $p_y = \mathbb{P}(Y > y)$,

$$p_y^{-1} \mathbb{E}\Big[\Big(\mathbb{E}\left[Z_i Z_j \mid P_e Z, Y\right] - \mathbb{E}\left[Z_i Z_j \mid P_e Z\right]\Big)\mathbb{1}\{Y > y\}\Big] \xrightarrow[y \to y^+]{} 0 \tag{C.2}$$

Fixing $i, j \le p$ and following the same path as in Theorem 1 we decompose the left-hand side of (C.2) for any $A > 0$ as a sum $C_1(A,y) + C_2(A,y)$ where

$$\begin{aligned}
C_1(A,y) = p_y^{-1} \mathbb{E}\Big[\Big(&\mathbb{E}\left[Z_i Z_j \mathbb{1}\{\|Z\| \le A\} \mid P_e Z, Y\right] - \dots \\
&\dots \mathbb{E}\left[Z_i Z_j \mathbb{1}\{\|Z\| \le A\} \mid P_e Z\right]\Big)\mathbb{1}\{Y > y\}\Big], \\
C_2(A,y) = p_y^{-1} \mathbb{E}\Big[\Big(&\mathbb{E}\left[Z_i Z_j \mathbb{1}\{\|Z\| > A\} \mid P_e Z, Y\right] - \dots \\
&\dots \mathbb{E}\left[Z_i Z_j \mathbb{1}\{\|Z\| > A\} \mid P_e Z\right]\Big)\mathbb{1}\{Y > y\}\Big].
\end{aligned}$$

Point (iii) of Proposition 3 with $h = 1$ and $g(Z) = Z_i Z_j \mathbb{1}\{\|Z\| \le A\}$ ensures that $C_1(A,y) \to 0$ as $y \to y^+$ for any fixed A. On the other hand, using that $|Z_i Z_j| \le \frac{1}{2}(|Z_i|^2 + |Z_j|^2) \le \frac{1}{2}\|Z\|_2^2 \le c\|Z\|^2$ for some constant c we may bound $|C_2(A,y)|$ as follows,

$$\begin{aligned}
|C_2(A,y)| &\le p_y^{-1} c \mathbb{E}\Big[\mathbb{E}\left[\|Z\|^2 \mathbb{1}\{\|Z\| > A\} \mid P_e Z, Y\right] \mathbb{1}\{Y > y\}\Big] + \dots \\
&\dots p_y^{-1} c \mathbb{E}\Big[\mathbb{E}\left[\|Z\|^2 \mathbb{1}\{\|Z\| > A\} \mid P_e Z\right] \mathbb{1}\{Y > y\}\Big] \\
&= p_y^{-1} c \Big(\mathbb{E}\left[\|Z\|^2 \mathbb{1}\{\|Z\| > A\} \mathbb{1}\{Y > y\}\right] + \mathbb{E}\Big(\mathbb{E}\left[\|Z\|^2 \mathbb{1}\{\|Z\| > A\} \mid P_e Z\right] \mathbb{1}\{Y > y\}\Big)\Big) \\
&= c\mathbb{E}\left[h_{1,A}(Z) \mid Y > y\right] + \mathbb{E}\left[h_{2,A}(Z) \mid Y > y\right].
\end{aligned}$$

Hence, in view of condition (4.4) for any $\epsilon > 0$ there exists some $A > 0$ such that

$$\limsup_{y \to y^+} |C_2(A,y)| \le \epsilon,$$

whence $\limsup_{y \to y^+} |C_2(A,y)| + |C_1(A,y)| \le \epsilon$, which shows (C.2) and completes the proof.

Appendix D: Proofs and auxiliary results for Section 5

D.1. Inverse of empirical c.d.f.*'s and order statistics*

The following general fact is used on several occasions in our proofs:

Fact D.1. For $u \in (0,1], \hat{H}^-(u) = T_{(\lceil nu \rceil)}$ and for $z \in [0, n-1]$:

$$(\hat{H}(T_i) < (z+1)/n) \Leftrightarrow (T_i \leq \hat{H}^-(z/n)).$$

Proof. The first statement follows from the definition of $\hat{H}^-$. Thus, using (5.1),

$$\begin{aligned}\hat{H}(T_i) < (z+1)/n) &\iff T_i < \hat{H}^-((z+1)/n) = T_{(\lceil z+1 \rceil)} = T_{(\lceil z \rceil + 1)} \\ &\iff T_i \leq T_{(\lceil z \rceil)} = \hat{H}^-(z/n).\end{aligned}$$

□

D.2. Vervaat's Lemma

We quote Lemma 4.3 in Segers (2015), which is a variant of "Vervaat's lemma", i.e., the functional delta method for the mapping sending a monotone function to its inverse.

Lemma D.1. *Let $G : \mathbb{R} \to [0,1]$ be a continuous distribution function. Let $0 < r_n \to \infty$ and let $\hat{G}_n$ be a sequence of random distribution functions such that, in $\ell^\infty(\mathbb{R})$, we have $r_n(\hat{G}_n - G) \rightsquigarrow \beta \circ G$, as $n \to \infty$, where β is a random element of $\ell^\infty([0,1])$ with continuous trajectories. Then $\beta(0) = \beta(1) = 0$ almost surely and as $n \to \infty$,*

$$\sup_{u \in [0,1]} r_n |(G\{\hat{G}_n^-(u)\} - u) + (\hat{G}_n\{G^-(u)\} - u)| = o_{\mathbb{P}}(1).$$

D.3. Proof of Lemma 1

Because we only need to show that for any $u \in \mathbb{R}^p$, $v^T \widetilde{\Gamma} \rightsquigarrow v^T \widetilde{W}$, to prove that $\widetilde{\Gamma}$ is asymptotically tight we may consider the case where $q = 1$, i.e., $h(V) \in \mathbb{R}$. Denoting by ψ the derivative of $x \mapsto \mathbb{E}\left[h(V)^2 \mathbb{1}\{U \leq x\}\right]$, there exist by assumption positive constants (c_0, δ_0) such that for all $\delta \leq \delta_0$, $\psi(\delta) \leq c_0$. Similarly, because we assume the existence of $\Xi = S(0)$ (Assumption 2. in Theorem 3), there exist positive constants (c_1, δ_1) such that for all $\delta \leq \delta_1$, $\mathbb{E}[h(V)^2 \,|\, U < \delta_1] \leq c_1$. We assume in the following argument that $k/n \leq \delta_0 \wedge \delta_1$.

We apply Theorem 2.11.23 in Van Der Vaart and Wellner (2013) (Classes of functions changing with n) with

$$\begin{aligned}f_{n,u}(V,U) &= \sqrt{\frac{n}{k}} h(V) \mathbb{1}\{U \leq uk/n\}, \\ \mathcal{F}_n &= \{f_{n,u} \,:\, u \in [0,1]\}, \\ F_n(V,U) &= \sqrt{\frac{n}{k}} |h(V)| \mathbb{1}\{U \leq k/n\}.\end{aligned}$$

We start by verifying equation 2.11.21 in Van Der Vaart and Wellner (2013). First, we have

$$\mathbb{E}\left[F_n(V,U)^2\right] \leq c_1.$$

Second, for any $\eta > 0$ and $M > 0$, it holds that (for n, k large enough)

$$\begin{aligned}&\mathbb{E}\left[F_n(V,U)^2 \mathbb{1}\{F_n(V,U) > \eta\sqrt{n}\}\right] \\ &= \left(\frac{n}{k}\right) \mathbb{E}\left[|h(V)|^2 \mathbb{1}\{U \leq k/n\} \mathbb{1}\{|h(V)| \mathbb{1}\{U \leq k/n\} > \eta\sqrt{k}\}\right] \\ &\leq \left(\frac{n}{k}\right) \mathbb{E}\left[|h(V)|^2 \mathbb{1}\{U \leq k/n\} \mathbb{1}\{|h(V)| > \eta\sqrt{k}\}\right] \\ &\leq \left(\frac{n}{k}\right) \mathbb{E}\left[|h(V)|^2 \mathbb{1}\{U \leq k/n\} \mathbb{1}\{|h(V)| > M\}\right].\end{aligned}$$

Hence

$$\limsup_{n\to\infty} \mathbb{E}\left[F_n(V,U)^2 \mathbb{1}\{F_n(U) > \eta\sqrt{n}\}\right] \leq S(M).$$

But M is arbitrary so the latter display is arbitrarily small. Third, by the mean value theorem, whenever $u \leq t$, $\exists \tilde{t} \in (u,t)$ such that

$$\begin{aligned}\mathbb{E}\left[(f_{n,u}(V,U) - f_{n,t}(V,U))^2\right] &= \binom{n}{k} \mathbb{E}\left[h(V)^2 \mathbb{1}\{uk/n \leq U \leq tk/n\}\right] \\ &= \psi(\tilde{t}k/n)(t-u) \\ &\leq c_0(t-u).\end{aligned}$$

This implies that

$$\sup_{|u-t|\leq \delta_n} \mathbb{E}\left[(f_{n,u}(V,U) - f_{n,t}(V,U))^2\right] \to 0, \text{ as } \delta_n \to 0.$$

It remains to check the entropy condition for the class $\mathcal{F}_n$. Let $0 < \epsilon < 1$, and denote by $u_i = i\epsilon$, $i = 0, \ldots, N$ and $u_{N+1} = 1$ with $N = \lfloor 1/\epsilon \rfloor$. Denote respectively by $f^+_{n,u}$ and $f^-_{n,u}$ the positive and negative parts of $f_{n,u}$ and by $\mathcal{F}^+_n, \mathcal{F}^-_n$ the associated classes. The functions (f^+_{n,u_i}) (resp. (f^-_{n,u_i})) forms an (ϵ, L_2)-bracketing of $\mathcal{F}^+_n$ (resp. $\mathcal{F}^-_n$), i.e., for any $u \in [0,1]$, there exists i such that

$$f^+_{n,u_i} \leq f^+_{n,u} \leq f^+_{n,u_{i+1}},$$

and

$$\mathbb{E}\left[(f^+_{n,u_{i+1}}(V,U) - f^+_{n,u_i}(V,U))^2\right] \leq c_0\epsilon.$$

Similar inequalities remain valid for $\mathcal{F}^-_n$. Hence considering the functions $f_{n,i} = f^+_{n,u_i} - f^-_{n,u_{i+1}}$, we have that for $u \in [u_i, u_{i+1}]$, $i = 0, \ldots, N$,

$$f_{n,u}(\,\cdot\,) = f^+_{n,u}(\,\cdot\,) - f^-_{n,u}(\,\cdot\,) \in [f_{n,i}(\,\cdot\,), f_{n,i+1}(\,\cdot\,)]\,,$$

thus there exists $C > 0$ such that

$$\mathcal{N}_{[\,]}(\epsilon \|F_n\|_{L_2(P)}, \mathcal{F}_n, L_2(P)) \leq C/\epsilon^2.$$

The entropy condition is satisfied as for all $\delta_n \to 0$,

$$\int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[\,]}(\epsilon \|F_n\|_{L_2(P)}, \mathcal{F}_n, L_2(P))}\, d\epsilon \to 0. \tag{D.1}$$

Consequently, the process $\widetilde{\Gamma}$ is tight. Finally the covariance functions at $s \leq t$ are given by

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left[\widetilde{\Gamma}_h(s), \widetilde{\Gamma}_h(t)\right] &= \mathbb{E}\left[n/k h(V) h(V)^\top \mathbb{1}\{U \leq sk/n\}\right] - \cdots \\ &\qquad n/k \mathbb{E}\left[h(V)\mathbb{1}\{U \leq sk/n\}\right] \mathbb{E}\left[h(V)\mathbb{1}\{U \leq tk/n\}\right] \\ &= s\mathbb{E}\left[h(V)h(V)^\top \,\middle|\, U \leq sk/n\right] - \cdots \\ &\qquad k/n\; st \mathbb{E}\left[h(V) \mid U \leq sk/n\right] \mathbb{E}\left[h(V) \mid U \leq tk/n\right]\end{aligned}$$

The first term in the right-hand side converges to $s\,\Xi = (s \wedge t)\,\Xi$ while the second term goes to zero from assumption 3. in Theorem 3's statement. This concludes the proof.

D.4. Proof of Lemma 2

We apply Lemma D.1 (Vervaat) to the distribution functions

$$\hat{G}_n(u) = \begin{cases} 0 & \text{for } u < 0 \\ \hat{F}_U(uk/n)/\hat{F}_U(k/n) & \text{for } 0 \le u \le 1 \\ 1 & \text{for } 1 < u \end{cases}, \qquad G(u) = \begin{cases} 0 & \text{for } u < 0 \\ u & \text{for } 0 \le u \le 1 \\ 1 & \text{for } 1 < u \end{cases}.$$

The quantile functions of $\hat{G}_n$ and G are respectively, for any $u \in [0,1]$,

$$\hat{G}_n^-(u) = \frac{\hat{F}_U^-(u\hat{F}_U(k/n))}{k/n}, \qquad G^-(u) = u,$$

Now we prove that the conditions of Lemma D.1 are satisfied with $r_n = \sqrt{k}$ and β a Brownian bridge with covariance function $u_1 \wedge u_2 - u_1 u_2$. Define

$$a_n = \frac{k/n}{\hat{F}_U(k/n)}$$

and write

$$\begin{aligned} \sqrt{k}(\hat{G}_n(u) - u) &= \left(\frac{\sqrt{k}}{\hat{F}_U(k/n)}\right)\left(\hat{F}_U(uk/n) - u\hat{F}_U(k/n)\right) \\ &= a_n\sqrt{k}\left(\frac{n}{k}\hat{F}_U(uk/n) - u\frac{n}{k}\hat{F}_U(k/n)\right) \\ &= a_n\sqrt{k}\left((\frac{n}{k}\hat{F}_U(uk/n) - u) - u(\frac{n}{k}\hat{F}_U(k/n) - 1)\right) \\ &= a_n\left(\hat{\gamma}_1(u) - u\hat{\gamma}_1(1)\right) \\ &= a_n\hat{\gamma}_2(u), \end{aligned}$$

where $\hat{\gamma}_1$ is defined in (5.12) and

$$\hat{\gamma}_2(u) = \hat{\gamma}_1(u) - u\hat{\gamma}_1(1). \tag{D.2}$$

Now use that $a_n \to 1$ in probability and that $\sup_{u\in[0,1]} |\hat{\gamma}_2(u)| = O_{\mathbb{P}}(1)$ (both are consequences of Corollary 2) to conclude (invoking Slutsky's lemma) that

$$\begin{aligned} \sqrt{k}(\hat{G}_n(u) - u) &= \hat{\gamma}_2(u) + (a_n - 1)\hat{\gamma}_2(u) \\ &= \hat{\gamma}_2(u) + o_{\mathbb{P}}(1), \end{aligned} \tag{D.3}$$

where the stochastic convergence $o_{\mathbb{P}}(1)$ is uniform in $u \in [0,1]$. In particular that $\sqrt{k}(\hat{G}_n(u) - u)$ weakly converges to a Brownian bridge with covariance function $u_1 \wedge u_2 - u_1 u_2$. The conclusion of Lemma D.1 is that

$$\sup_{u\in(0,1]} \left|\hat{\gamma}_3(u) + \sqrt{k}(\hat{G}_n(u) - u)\right| = o_{\mathbb{P}}(1),$$

with

$$\hat{\gamma}_3(u) = \sqrt{k}(\hat{G}_n^-(u) - G^-(u)) = \sqrt{k}\left(\frac{n}{k}\hat{F}_U^-(u\hat{F}_U(k/n)) - u\right). \tag{D.4}$$

Consequently, using (D.3),

$$\sup_{u\in(0,1]} |\hat{\gamma}_3(u) + \hat{\gamma}_2(u)| = o_{\mathbb{P}}(1). \tag{D.5}$$

Remark that

$$\sqrt{k}\left((n/k)\hat{F}_U^-(uk/n) - u\right) = \hat{\gamma}_3(ua_n) + u\sqrt{k}(a_n - 1). \tag{D.6}$$

and that, as $\hat{\gamma}_1(1) = \sqrt{k}((n/k)\hat{F}_U(k/n) - 1)$,

$$\sqrt{k}(a_n - 1) = -a_n\hat{\gamma}_1(1) \tag{D.7}$$

Using the triangle inequality and (D.6), we get

$$\begin{aligned}
&\left|\sqrt{k}\left((n/k)\hat{F}_U^-(uk/n) - u\right) + \hat{\gamma}_1(u)\right| \\
&= \left|\hat{\gamma}_3(ua_n) + u\sqrt{k}(a_n - 1) + \hat{\gamma}_1(u)\right| \\
&\leq |\hat{\gamma}_3(ua_n) + \hat{\gamma}_2(ua_n)| + \left|u\sqrt{k}(a_n - 1) + \hat{\gamma}_1(u) - \hat{\gamma}_2(ua_n)\right| \\
&= |\hat{\gamma}_3(ua_n) + \hat{\gamma}_2(ua_n)| + |\hat{\gamma}_1(u) - \hat{\gamma}_1(ua_n)|,
\end{aligned}$$

where the last line is deduced from (D.7) and $\hat{\gamma}_2(u) = \hat{\gamma}_1(u) - u\hat{\gamma}_1(1)$. Whenever $u \in [0, 1/2]$, we have, with probability going to 1, that $ua_n \in [0, 1]$. Moreover, because $a_n \to 0$ in probability, there exists $\delta_n \to 0$ such that the event $|u - ua_n| \leq |a_n| \leq \delta_n$ has probability going to 1. On these events, it holds

$$\begin{aligned}
\sup_{u\in(0,1/2]} |\hat{\gamma}_3(ua_n) + \hat{\gamma}_2(ua_n)| &\leq \sup_{u\in(0,1]} |\hat{\gamma}_3(u) + \hat{\gamma}_2(u)| = o_{\mathbb{P}}(1) \\
\sup_{u\in(0,1/2]} |\hat{\gamma}_1(u) - \hat{\gamma}_1(ua_n)| &= \sup_{u\in(0,1], v\in(0,1], |u-v|\leq\delta_n} |\hat{\gamma}_1(u) - \hat{\gamma}_1(v)| = o_{\mathbb{P}}(1).
\end{aligned}$$

We have used (D.5) and the asymptotic equicontinuity of $\hat{\gamma}_1$. Consequently we have shown that, whenever $n \to \infty$, $k \to \infty$, we have

$$\sup_{u\in(0,1/2]} \left|\sqrt{k}\left((n/k)\hat{F}_U^-(uk/n) - u\right) + \hat{\gamma}_1(u)\right| = o_{\mathbb{P}}(1).$$

To obtain the stated result, apply this with $2k$ in place of k.

Appendix E: Extension to non-standardized covariates

In this section we extend our inverse regression framework to the case of non-standardized covariates X. Section E.1 recalls standard results for that matter. In Section E.2 the extensions of the TIREX1 and TIREX2 principles are presented. The proofs of these results are omitted since they follow from classical arguments from non-standardized covariates combined with our proofs with standardized covariates from Section 3. In Section E.3 we show that estimating the mean vector and covariance matrix for standardization does not change the asymptotic behavior of the latter tail processes.

E.1. SIR and SAVE principles with non-standardized covariates

We first recall some necessary background from the theory of inverse regression with non-standardized covariates, as exposed *e.g.* in Cook and Weisberg (1991).

E.1.1. SDR spaces

Recall from Section 2 that in terms of non-standardized covariates $X = m + \Sigma^{1/2}Z$, a subspace $\tilde{E}$ of $\mathbb{R}^p$ is a SDR space for the pair (X, Y) if and only if $\tilde{E} = \Sigma^{-1/2}E$ where E is a SDR space for the pair (Z, Y). We denote in the sequel by $\tilde{P}$ the orthogonal projector onto such a SDR space $\tilde{E}$ and we define $\tilde{Q} = I_p - \tilde{P}$.

E.1.2. Linearity and constant variance conditions

Conditions LC (2.1) and CCV (2.2) regarding the standardized variable Z are respectively equivalent to

$$\mathbb{E}\left[X|\tilde{P}X\right] = b + B\tilde{P}X \tag{E.1}$$

for some $b \in \mathbb{R}^p$ and $B \in \mathbb{R}^{p\times p}$, and

$$\mathbb{V}\mathrm{ar}\left[X|\tilde{P}X\right] \text{ is constant} \quad \text{a.s.} \tag{E.2}$$

E.1.3. SIR principle and CUME matrix

The extension of the SIR principle (Proposition 1) in terms of non-standardized covariates, is that under condition (E.1), it holds that

$$\Sigma^{-1}(\mathbb{E}\left[X|Y\right] - m) \in \tilde{E}. \tag{E.3}$$

As a consequence the CUME matrix defined in (2.3) must be replaced with the matrix $\tilde{M}_{\mathrm{CUME}} = \mathbb{E}\left[\tilde{m}(Y)\tilde{m}(Y)^T\right]$, with

$$\tilde{m}(y) = \mathbb{E}\left[(X - m)\mathbb{1}\{Y \le y\}\right],$$

in which case it holds that

$$\mathrm{span}(\tilde{M}_{\mathrm{CUME}}) \subset \Sigma\tilde{E} = \Sigma^{1/2}E.$$

E.1.4. SAVE principle

The parallel statement of Proposition 2 is that under conditions (E.1) and (E.2), we have

$$\mathrm{span}(\Sigma^{-1}(\mathbb{V}\mathrm{ar}\left[X \mid Y\right] - \Sigma)) \subset \tilde{E} \quad a.s., \tag{E.4}$$

or equivalently $\mathrm{span}(\Sigma^{-1}(\mathbb{E}\left[(X - m)(X - m)^\top \mid Y\right] - \Sigma)) \subset \tilde{E}$.

E.2. TIREX principles with non-standardized covariates

It follows from Definition 3 that E_e is an extreme SDR space for the pair (Z, Y) if and only if $\tilde{E}_e = \Sigma^{-1/2}E_e$ is an extreme SDR space for the pair (X, Y), in the sense that, denoting by $\tilde{P}_e$ the orthogonal projection on $\tilde{E}_e$, $Y_\infty \perp\!\!\!\perp X \mid \tilde{P}_e X$.

We now state the analogue statement to Theorem 1 in terms of the non-standardized covariate X.

Proposition E.1 (non-standardized TIREX1 principle)**.** *The assumptions of Theorem 1 are equivalent to*

1. $\lim_{A\to\infty} \limsup_{y\to y^+} \mathbb{E}\left[\tilde{g}_{k,A}(X) \mid Y > y\right] = 0$, $k = 1, 2$ *where* $\tilde{g}_{1,A}(X) = \|X\|\mathbb{1}\{\|X\| > A\}$ *and* $\tilde{g}_{2,A}(X) = \mathbb{E}\left[\|X\|\mathbb{1}\{\|X\| > A\} \,\middle|\, \tilde{P}_e X\right]$, *where* $\tilde{P}_e$ *is the orthogonal projector on* $\tilde{E}_e = \Sigma^{-1/2} E_e$.
2. *The covariate vector satisfies the non-standardized linearity condition* (E.1)
3. *For some* $\tilde{\ell} \in \mathbb{R}^p$, *with* $m = \mathbb{E}[X]$

$$\mathbb{E}[X \mid Y > y] - m \xrightarrow[y\to y^+]{} \tilde{\ell}. \tag{E.5}$$

In such a case $\tilde{\ell} = \Sigma^{1/2}\ell$ *where* ℓ *is the limit defined in Theorem 1 and the conclusion is that*

$$\Sigma^{-1}\tilde{\ell} \in \tilde{E}_e.$$

Proposition E.2 (non-standardized TIREX2 principle)**.** *Assume that* (X, Y) *and the extreme SDR space satisfy the assumptions of Proposition E.1 (non-standardized TIREX1 principle) and that in addition,*

1. *(second order uniform integrability):*

$$\lim_{A\to\infty} \limsup_{y\to y^+} \mathbb{E}\left[\tilde{h}_{k,A}(X) \,\middle|\, Y > y\right] = 0, \qquad k = 1, 2\,, \tag{E.6}$$

   where $\tilde{h}_{1,A}(X) = \|X\|^2\mathbb{1}\{\|X\| > A\}$ *and* $\tilde{h}_{2,A}(X) = \mathbb{E}\left[\|X\|^2\mathbb{1}\{\|X\| > A\} \,\middle|\, \tilde{P}_e X\right]$,
2. *(CCV) The covariate vector* X *satisfies the non-standardized constant variance condition* (E.2) *relative to* $\tilde{P}_e$,
3. *(Convergence of conditional expectations) For some* $\tilde{S} \in \mathbb{R}^{p\times p}$,

$$\mathbb{E}\left[XX^\top \,\middle|\, Y > y\right] \xrightarrow[y\to y^+]{} \tilde{S} + \tilde{\ell}\tilde{\ell}^\top, \tag{E.7}$$

   where $\tilde{\ell}$ *is the limit appearing in Proposition E.1.*

Then

$$\operatorname{span}(\Sigma^{-1}(\tilde{S} - \Sigma)) \subset \tilde{E}_e,$$

i.e. $\tilde{Q}_e\Sigma^{-1}(\tilde{S} - \Sigma) = 0$.

E.3. Estimation with non-standardized covariates

Consider the non-standardized versions of the matrices $M_{\text{TIREX1}}, M_{\text{TIREX2}}$ from Section 5 defined as follows:

$$\begin{aligned} \tilde{M}_{\text{TIREX1}} &= \int_0^1 C_n^m(u) C_n^m(u)^\top \,\mathrm{d}u\,, \text{ with} \\ C_n^m(u) &= \frac{n}{k}\mathbb{E}\left[(X - m)\mathbb{1}\{\tilde{Y} < F^-(uk/n)\}\right]\,, \end{aligned} \tag{E.8}$$

and

$$\begin{aligned} \tilde{M}_{\text{TIREX2}} &= \int_0^1 B_n^{m,\Sigma}(u) B_n^{m,\Sigma}(u)^\top \,\mathrm{d}u\,, \text{ with} \\ B_n^{m,\Sigma}(u) &= \frac{n}{k}\mathbb{E}\left[\left((X - m)(X - m)^\top - \Sigma\right)\mathbb{1}\{\tilde{Y} < F^-(uk/n)\}\right]. \end{aligned} \tag{E.9}$$

In view of Propositions E.1 and E.2, under the same assumptions therein, span($\tilde{M}_{\text{TIREX1}}$) and span($\tilde{M}_{\text{TIREX2}}$) become close to $\Sigma\tilde{E}_e$ as $n \to \infty$, in the sense that

$$\lim_{n\to\infty} \tilde{Q}_e \Sigma^{-1} \tilde{M}_{\text{TIREX1}} = \lim_{n\to\infty} \tilde{Q}_e \Sigma^{-1} \tilde{M}_{\text{TIREX2}} = 0,$$

where $\tilde{Q}_e$ is the orthogonal projector on $\tilde{E}_e^{\perp}$.

Notice that we can write $C_n^m, B_n^{m,\Sigma}$ in terms of C_n, B_n as follows:

$$\begin{aligned} C_n^m(u) &= \Sigma^{1/2} C_n(u) \\ B_n^{m,\Sigma}(u) &= \Sigma^{1/2} B_n(u) \Sigma^{1/2} \end{aligned} \tag{E.10}$$

Despite the apparent simplicity of (E.10), in the estimation step with unknown covariate's mean and covariance, one must replace m and Σ in definitions (E.8) and (E.9) with some estimates, *e.g.* the empirical ones which we denote by $\hat{m}, \hat{\Sigma}$. Namely we consider the processes

$$\begin{aligned} \hat{C}_n^{\hat{m}}(u) &= \frac{1}{k} \sum_{i=1}^{n} (X_i - \hat{m}) \mathbb{1}\{\tilde{Y}_i \le \hat{F}^{-}(uk/n)\} \,, \\ \hat{B}_n^{\hat{m},\hat{\Sigma}}(u) &= \frac{1}{k} \sum_{i=1}^{n} \Big((X_i - \hat{m})(X_i - \hat{m})^{\top} - \hat{\Sigma} \Big) \mathbb{1}\{\tilde{Y}_i \le \hat{F}^{-}(uk/n)\}] \end{aligned} \tag{E.11}$$

and define the non-standardized TIREX1 and TIREX2 tail empirical processes respectively as

$$\sqrt{k}\Big(\hat{C}_n^{\hat{m}} - C_n^m\Big) \text{ and } \sqrt{k}\Big(\hat{B}_n^{\hat{m},\hat{\Sigma}} - B_n^{m,\Sigma}\Big). \tag{E.12}$$

We assume that the conditions for the central limit theorem regarding the estimators $\hat{m}$ and $\hat{\Sigma}$ are met. For instance, we assume that X admits fourth order moments, an assumption which is needed anyway for the weak convergence of the TIREX2 process, see Corollary 1. Thus we work under the assumption that

$$\hat{m} = m + O_{\mathbb{P}}(1/\sqrt{n}) \,; \quad \hat{\Sigma} = \Sigma + O_{\mathbb{P}}(1/\sqrt{n}). \tag{E.13}$$

Proposition E.3 (Weak convergence of non-standardized TIREX processes)**.** *Under Assumption* (E.13)*,*

1. *The standardized TIREX1 process $\sqrt{k}(\hat{C}_n - C_n)$ converges weakly in $\ell^\infty([0,1])$ to a tight Gaussian process W_1 if and only if its non-standardized version defined in* (E.12) *converges weakly, in the same space, to the Gaussian process $\Sigma^{1/2} W_1$.*
2. *If weak convergence of the TIREX1 process holds true, then the standardized TIREX2 process $\sqrt{k}(\hat{B}_n - B_n)$ converges weakly in $\ell^\infty([0,1])$ to a tight Gaussian process W_2 if and only if its non-standardized version defined in* (E.12) *converges weakly, in the same space, to the Gaussian process $\Sigma^{1/2} W_2 \Sigma^{1/2}$.*

Proof of Proposition E.3.

1. Substituting $X - m$ with $\Sigma^{1/2} Z$ we obtain

$$\begin{aligned} \hat{C}^{\hat{m}}(u) &= \frac{1}{k} \sum_{i=1}^{n} \Sigma^{1/2} (Z_i + m - \hat{m}) \mathbb{1}\{\tilde{Y}_i \le \hat{F}^{-}(uk/n)\} \\ &= \Sigma^{1/2} \Big\{ \hat{C}_n(u) + \Delta_n(u)(m - \hat{m}) \Big\} \end{aligned} \tag{E.14}$$

where $\hat{C}_n$ is defined in (5.5) in terms of Z and

$$\Delta_n(u) := \frac{n}{k}\hat{F}\big(\hat{F}^{-}(uk/n)\big) \leq \frac{n}{k}\hat{F}\big(\hat{F}^{-}(k/n)\big) = \frac{n}{k}\hat{F}(\tilde{Y}_{(k)}) = 1. \tag{E.15}$$

Combining the latter upper bound, (E.14) and (E.10) we obtain

$$\sqrt{k}\left(\hat{C}_n^{\hat{m}} - C_n^{m}\right) = \Sigma^{1/2}\sqrt{k}(\hat{C}_n(u) - C_n(u)) + R_n(u), \tag{E.16}$$

where $\sup_{u\in[0,1]} R_n(u) = O_{\mathbb{P}}(\sqrt{k/n}) = o_{\mathbb{P}}(1)$ and the main term $\sqrt{k}(\hat{C}_n(u) - C_n(u))$ is the standardized TIREX1 process. The first assertion of the statement follows from the Slutsky's lemma.

2. The argument for the second order method is similar though the computation is more involved. We have

$$\begin{aligned}\hat{B}_n^{\hat{m},\hat{\Sigma}}(u) = \Sigma^{1/2}\Bigg\{\frac{1}{k}\sum_{i\leq n}\Big(\big(Z_i + \Sigma^{-1/2}(m-\hat{m})\big)\big(Z_i + \Sigma^{-1/2}(m-\hat{m})\big)^{\top} - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\Big) \times \cdots \\ \cdots \mathbb{1}\{\tilde{Y}_i \leq \hat{F}^{-}(uk/n)\}\Bigg\}\Sigma^{1/2} \\ = \Sigma^{1/2}\left\{\hat{B}_n(u) + A_{1,n}\Delta_n(u) + A_{2,n}(u)\right\}\Sigma^{1/2}\end{aligned}$$

with $\Delta_n(u) \leq 1$ as in (E.15) and

$$\begin{aligned}A_{1,n} &= \big(I_p - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\big) + \Sigma^{-1/2}(m-\hat{m})(m-\hat{m})^{\top}\Sigma^{-1/2}\,, \\ A_{2,n} &= \Sigma^{-1/2}(m-\hat{m})\hat{C}_n^{\top}(u) + \hat{C}_n(u)(m-\hat{m})^{\top}\Sigma^{-1/2}.\end{aligned}$$

Under the assumption that the TIREX1 empirical process converges weakly we have that $\sup_u \hat{C}_n(u) = O_{\mathbb{P}}(1)$, and using (E.13) and (E.10) we obtain

$$\sqrt{k}\Big(\hat{B}_n^{\hat{m},\hat{\Sigma}}(u) - B_n^{m,\Sigma}(u)\Big) = \Sigma^{1/2}\sqrt{k}\Big(\hat{B}_n(u) - B_n(u)\Big)\Sigma^{1/2} + R'_n(u)$$

with $\sup_u R'_n(u) = O_{\mathbb{P}}(\sqrt{k/n}) = o_{\mathbb{P}}(1)$. The second assertion follows.

□