Shrinkage for Extreme Partial Least Squares

Julyan Arbel, Stéphane Girard, Hadrien Lorenzo

Shrinkage for Extreme Partial Least Squares

J. ARBEL, S. GIRARD,
Univ. Grenoble Alpes, Inria, CNRS,
Grenoble INP, LJK 38000 Grenoble, France.
julyan.arbel@inria.fr, stephane.girard@inria.fr
& H. LORENZO
Aix Marseille Univ, CNRS,
I2M, Marseille, France.
hadrien.lorenzo@inria.fr

October 20, 2023

Abstract

This research focuses on dimension-reduction techniques for modeling conditional extreme values. Specifically, we investigate the idea that extreme values of a response variable can be explained by nonlinear functions derived from linear projections of an input random vector. In this context, the estimation of projection directions is examined, as approached by the Extreme Partial Least Squares (EPLS) method–an adaptation of the original Partial Least Squares (PLS) method tailored to the extreme-value framework. Further, a novel interpretation of EPLS directions as maximum likelihood estimators is introduced, utilizing the von Mises-Fisher distribution applied to hyperballs. The dimension reduction process is enhanced through the Bayesian paradigm, enabling the incorporation of prior information into the projection direction estimation. The maximum a posteriori estimator is derived in two specific cases, elucidating it as a regularization or shrinkage of the EPLS estimator. We also establish its asymptotic behavior as the sample size approaches infinity. A simulation data study is conducted in order to assess the practical utility of our proposed method. This clearly demonstrates its effectiveness even in moderate data problems within high-dimensional settings. Furthermore, we provide an illustrative example of the method's applicability using French farm income data, highlighting its efficacy in real-world scenarios.

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

MSC 2020 subject classification: 62G32, 62H25, 62H12, 62E20.

1 Introduction

In modern statistical regression situations, one has to deal with problems where the dimension p of the covariates X is large, and where the size n of the dataset is insufficient to provide reliable estimations. Using standard (parametric or nonparametric) regression techniques in such situations may yield overfitting and therefore unstable estimations. This curse of dimensionality [Gee11] may be mitigated by identifying a low-dimensional subspace of the covariates X that maintains a strong link between the projected covariates and the response variable Y. As an example, Partial Least Squares (PLS) regression [Wol75] aims at estimating linear combinations of X coordinates having a high covariance with Y. Even though PLS has been initially developed within the chemometrics field [MN92], it has also received considerable attention in the statistical literature, see for instance [NT00]. Sliced Inverse Regression (SIR) [Li91] is an alternative method to estimate a so-called central dimension reduction subspace based on an inverse regression model, *i.e.* when X is written as a function of Y. Several extensions have been developed for PLS and SIR, see [CHS13, LCT07] and [CFG17, CGS14, Por16] among others or [GLS22] for a review. While the above-mentioned methods adopt the frequentist point of view, there also exist a number of works in the literature based on Bayesian approaches. In [RBL11], the authors model the response variable Y in terms of the predictors X using a mixture model whose parameters are estimated with a Markov chain Monte Carlo (MCMC) procedure. The converse point of view is adopted in [MLM10]: X is modeled as a function of Y thanks to an inverse mixture model, the estimation also requiring an MCMC method. A similar approach is proposed in [CLL21] using a Bayesian inverse regression through Gaussian processes and MCMC procedures.

The curse of dimensionality is exacerbated when modeling conditional extremes since tail events are rare by nature. Nonparametric estimators of extreme conditional features [DGG13, DSUC23, GSUC21] are thus impacted both by the scarcity of extremes and the high dimensional setting. Recently, some works have introduced dimension-reduction tools dedicated to conditional extremes. One can mention [APSZ21, Gar18] who propose extreme analogs of the central dimension reduction subspace. In [WLX22], a semi-parametric approach is introduced for the estimation of extreme conditional quantiles based on a tail single-index model. The dimension reduction direction is estimated by fitting a misspecified linear quantile regression model. Extreme-PLS (EPLS) [BEG23] is a dimension reduction method relying on PLS principles for estimating the linear combinations of X that best explain the extreme values of Y.

In this work, we develop shrinkage versions of the EPLS method for high-dimensional settings. More specifically, the EPLS estimator is interpreted as a maximum likelihood estimator associated with a von Mises-Fisher likelihood (Section 2). The latter distribution, which naturally arises for modeling directional data distributed on the unit sphere [MJ09], is here adapted to hyperballs. Two prior distributions are introduced on the dimension reduction direction in Section 3: A conjugate one based on the von Mises-Fisher distribution and a second one using the Laplace distribution (both defined on the unit sphere) to enforce sparsity. It is shown that the maximum a posteriori (MAP) estimator is available in

closed form, its computation does not require MCMC methods and can be interpreted as a shrinkage version of the initial EPLS estimator. Convergence results are also established when the sample size tends to infinity. The behavior of the two proposed estimators is illustrated on simulated data in Section 4, while an application on French farm income data is described in Section 5 to assess the influence of various parameters on field-grown carrot production. The functions to compute Shrinkage Extreme Partial Least-Squares (SEPaLS) estimators are available in the R package `SEPaLS`[1] [LGA23], while the R code replicating the figures can be found online[2]. A discussion is provided in Section 6 and proofs are postponed to the Appendix.

2 Extreme Partial Least Squares without shrinkage

Throughout, $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product on $\mathbb{R}^p$, $\| \cdot \|_2$ is the corresponding quadratic norm and $S^{p-1} = \{x \in \mathbb{R}^p, \|x\|_2 = 1\}$ is the associated unit sphere. Moreover, for any set $\{z_1, \ldots, z_n\}$, $z_{1:n}$ denotes the vector $(z_1^\top, \ldots, z_n^\top)^\top$. Plus, two sequences of random variables (A_n) and (B_n) (where (B_n) is almost surely non-zero) are equivalent in probability if $A_n/B_n \xrightarrow{\mathbb{P}} 1$ which is denoted by $A_n \overset{\mathbb{P}}{\sim} B_n$. Also, we write $A_n = o_{\mathbb{P}}(B_n)$ if $A_n/B_n \xrightarrow{\mathbb{P}} 0$.

We first recall in Subsection 2.1 the derivation of the EPLS estimator from a statistical regression model and, in Subsection 2.2, the extreme-value assumptions necessary to establish its asymptotic properties. Subsection 2.3 is dedicated to the presentation of the von Mises-Fisher distribution on the sphere and to its adaptation to hyperballs. Based on these, we then reinterpret the EPLS direction as a maximum likelihood estimator and derive its asymptotic properties in Subsection 2.4.

2.1 EPLS model

The following single-index inverse regression model is introduced in [BEG23]:

($\mathbf{A_0}$) $X = g(Y)\beta + \varepsilon$, where X and ε are p-dimensional random vectors, Y is a real random variable, $g : \mathbb{R} \to \mathbb{R}$ is an unknown link function, $\beta \in S^{p-1}$ is the unknown direction of interest.

Model ($\mathbf{A_0}$) is referred to as an inverse regression model since the covariates X are written as functions of the response variable Y. Let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be an n sample with same distribution as (X, Y). The EPLS estimator of the unit direction β is obtained by maximizing with respect to $u \in S^{p-1}$ the empirical covariance between $\langle u, X \rangle$ and Y conditionally on large values of Y. More specifically, the conditional covariance maximization problem is equivalent to

$$\hat{\beta}(y_n) = \underset{\|u\|_2=1}{\operatorname{argmax}} \langle u, \hat{v}(y_n) \rangle = \frac{\hat{v}(y_n)}{\|\hat{v}(y_n)\|_2}, \tag{1}$$

[1] `https://github.com/hlorenzo/SEPaLS/`

[2] `https://github.com/hlorenzo/SEPaLS_simus/`

where, for any threshold $y_n \in \mathbb{R}$, $\hat{v}(y_n)$ is defined by

$$\hat{v}(y_n) = \sum_{i=1}^{n} X_i \Phi_i(y_n, Y_{1:n}), \tag{2}$$

with, for all $i \in \{1, \ldots, n\}$,

$$\Phi_i(y_n, Y_{1:n}) = \frac{1}{n} \left(\hat{\bar{F}}(y_n) Y_i - \hat{m}_Y(y_n) \right) \mathbf{1}\{Y_i \geq y_n\},$$

the following first-order empirical moment

$$\hat{m}_Y(y_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i \mathbf{1}\{Y_i \geq y_n\},$$

and $\hat{\bar{F}}$ the empirical survival function of Y. See [BEG23] for details. The asymptotic properties of the EPLS estimator can be established under some assumptions on the distribution tails, described hereafter.

2.2 Extreme-value framework

Three assumptions on the link function g and the distribution tail of Y and ε are considered. They rely on the notion of regularly-varying ($\mathcal{RV}$) functions. Recall that $\varphi \in \mathcal{RV}_\theta$ ($\theta \in \mathbb{R}$) if and only if φ is positive and

$$\lim_{y \to \infty} \frac{\varphi(ty)}{\varphi(y)} = t^\theta,$$

for all $t > 0$. We refer to [BGT89] for a detailed account of regular variations.

$(\mathbf{A_1})$ The density function f of Y belongs to $\mathcal{RV}_{-1/\gamma-1}$, with $0 < \gamma < 1$;

$(\mathbf{A_2})$ $g \in \mathcal{RV}_c$ with $c > 0$ and $2\gamma(c+1) < 1$;

$(\mathbf{A_3})$ There exists $q > 1/(\gamma c)$ such that $\mathbb{E}(\|\varepsilon\|_2^q) < \infty$.

Assumption $(\mathbf{A_1})$ implies that $\bar{F} \in \mathcal{RV}_{-1/\gamma}$ which in turn is equivalent to assuming that the distribution of Y is in the Fréchet maximum domain of attraction with positive tail-index γ, see [BGT89, Theorem 1.5.8] and [dHF07, Theorem 1.2.1]. This domain of attraction consists of heavy-tailed distributions, such as Pareto, Burr and Student distributions, see [BGST04] for further examples. The larger γ is, the heavier the tail. The restriction to $\gamma < 1$ ensures that the first-order moment $\mathbb{E}(|Y|\mathbf{1}\{Y \geq y\})$ exists for all $y \in \mathbb{R}$. Assumption $(\mathbf{A_2})$ ensures that the link function g ultimately behaves like a power function. Finally, $(\mathbf{A_3})$ can be interpreted as an assumption on the tail of $\|\varepsilon\|_2$. It is satisfied, for instance, by distributions with exponential-like tails such as Gaussian, Gamma or Weibull distributions.

2.3 Two von Mises-Fisher distributions

The von Mises-Fisher distribution vMF/S(μ, κ) on the unit sphere S^{p-1}, $p \geq 2$, is defined by its probability density function [WW56]:

$$f_{\mathrm{vMF/S}}(x|\mu, \kappa) = c_p(\kappa) \exp\left(\kappa \langle \mu, x \rangle\right) \mathbf{1}\{\|x\|_2 = 1\},$$

where $\mu \in S^{p-1}$ is a location parameter and $\kappa \geq 0$ is a concentration parameter. The normalizing constant is given by:

$$c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)} \text{ if } \kappa > 0 \text{ and } c_p(0) = \frac{\Gamma(p/2)}{(2\pi)^{p/2}} \text{ otherwise,} \tag{3}$$

where $I_q(\cdot)$ is the modified Bessel function of the first kind and order $q \geq 0$ defined on $\mathbb{R}_+$ by

$$\kappa \mapsto I_q(\kappa) = \sum_{\ell=0}^{\infty} \frac{1}{\Gamma(q + \ell + 1)\ell!} \left(\frac{\kappa}{2}\right)^{2\ell+q}, \tag{4}$$

see [AS72, Chapter 9], with $\Gamma(\cdot)$ the Gamma function. The von Mises-Fisher distribution on the unit sphere is widely used in the analysis of directional data and can be considered as a spherical analog of the multivariate Gaussian distribution [Mar75]. Let us also recall that, for all $\mu \in S^{p-1}$, vMF/S$(\mu, 0)$ is the uniform distribution on the unit sphere (and thus, $c_p(0)$ coincides with the inverse of the sphere surface) and that μ is the mode of the vMF/S(μ, κ) distribution for all $\kappa > 0$. We propose the following adaptation of this distribution on balls:

Definition 1. *The von Mises-Fisher distribution* vMF/B(μ, r, κ) *on the p-dimensional ball, $p \geq 2$, of radius $r > 0$ is defined by its probability density function:*

$$f_{\mathrm{vMF/B}}(x|\mu, r, \kappa) = \frac{2\pi c_{p+2}(\kappa)}{r^p} \exp\left(\frac{\kappa \langle \mu, x \rangle}{r}\right) \mathbf{1}\{\|x\|_2 \leq r\},$$

where $\mu \in S^{p-1}$ is a location parameter and $\kappa \geq 0$ is a concentration parameter.

We refer to Lemma 1 in the Appendix for a proof that $f_{\mathrm{vMF/B}}(\cdot|\mu, r, \kappa)$ integrates to one. The next paragraph shows that the vMF/B distribution plays a central role in the interpretation of the EPLS estimator as a maximum likelihood estimator.

2.4 Maximum likelihood estimation

We first prove that the EPLS estimator, initially introduced by maximizing some empirical covariance, can also be interpreted as a maximum likelihood (ML) estimator. It is thus denoted by $\hat{\beta}_{\mathrm{ml}}(y_n)$ in the sequel.

Proposition 1. *The EPLS estimator* (1) *is the ML estimator of* β *in the following model:*

(i) $X_1, \dots, X_n$ *are independent and, for all* $i \in \{1, \dots, n\}$, X_i *given* $(Y_{1:n}, \varepsilon_i)$ *is* vMF/B(β, r_i, κ_i) *distributed, with location parameter* β*, radius* $r_i = |g(Y_i)| + \|\varepsilon_i\|_2$ *and concentration parameter* $\kappa_i = \theta_n r_i \Phi_i(y_n, Y_{1:n})$*, where* $\theta_n > 0$ *is an arbitrary parameter.*

(ii) $(Y_{1:n}, \varepsilon_{1:n})$ *is distributed according to some arbitrary density* $p(\cdot, \cdot)$ *on* $\mathbb{R}^n \times \mathbb{R}^{pn}$ *that does not depend on* β*.*

This formalism opens the door to the construction of shrinkage estimators for β based on the Bayesian paradigm in Section 3. Before that, the next Proposition provides a consistency result on the ML estimator (1).

Proposition 2. *Assume* $(\mathbf{A_0})$, $(\mathbf{A_1})$, $(\mathbf{A_2})$ *and* $(\mathbf{A_3})$ *hold. Let* $y_n \to \infty$ *such that* $n\bar{F}(y_n) \to \infty$ *and* $n\bar{F}(y_n)^{1-2/q}/g^2(y_n) \to 0$ *as* $n \to \infty$*. Then,*

$$\sqrt{n\bar{F}(y_n)}\left(\hat{\beta}_{\mathrm{ml}}(y_n) - \beta\right) \xrightarrow{\mathbb{P}} 0.$$

We refer to [BEG23] for a discussion of the assumptions on the (y_n) sequence. Let us simply recall that $n\bar{F}(y_n)$ represents the effective number of observations used in the ML estimator. It is thus natural that the associated rate of convergence is of order $\sqrt{n\bar{F}(y_n)}$.

3 Shrinkage for Extreme Partial Least Squares

A prior distribution $\pi(\cdot)$ is introduced on β and the shrinkage effect on the maximum a posteriori (MAP) estimator is investigated. The posterior distribution is established in Subsection 3.1 and MAPs are derived for two particular cases of priors in Subsection 3.2 and Subsection 3.3.

3.1 Posterior distribution

Combining Bayes' rule with Proposition 1 makes it possible to derive the posterior distribution of β. See Appendix for a detailed proof.

Proposition 3. *Let* $\theta_n > 0$ *and* $\pi(\cdot)$ *a prior distribution on* $\beta \in S^{p-1}$*. Then, under the model (i), (ii) of Proposition 1, the posterior distribution of* β *is given by*

$$p(\beta | X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) \propto \pi(\beta) \exp\left(K_n \langle \beta, \hat{\beta}_{\mathrm{ml}}(y_n) \rangle\right),$$

where we set $K_n := \theta_n \|\hat{v}(y_n)\|_2$*.*

The mode of the above posterior distribution is referred to as the SEPaLS estimator in the sequel. Its existence is ensured as soon as $\pi(\cdot)$ is continuous on S^{p-1}, since a continuous function on a compact domain attains its maximum value within that domain. We focus on the computation of the SEPaLS estimator for two particular choices of $\pi(\cdot)$ described in the next two subsections.

3.2 Conjugate prior

We first assume a vMF/S prior distribution for the direction $\beta \in S^{p-1}$, with location parameter $\mu_0 \in S^{p-1}$ and concentration parameter $\kappa_0 \geq 0$. The unit vector μ_0 can be interpreted as a prior on β while κ_0 is the confidence level on this prior. A graphical representation in dimension $p = 3$ of the density isocontours associated with this distribution is provided on the top of Figure 1 for $\mu_0 = (1, 0, 0)^\top$ and $\kappa_0 \in \{0, 1, 10, 100\}$. On the leftmost panel, the density is uniform on the unit sphere, and it becomes more peaked around $(1, 0, 0)^\top$ as κ_0 increases. Proposition 3 entails that the posterior distribution is written for any $\beta \in S^{p-1}$ as:

$$p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) \propto \exp\left(\langle \beta, K_n \hat{\beta}_{\mathrm{ml}}(y_n) + \kappa_0 \mu_0 \rangle\right),$$

which is still a vMF/S distribution. As expected, since the von Mises-Fisher distribution belongs to the exponential family, considering the associated conjugate prior for β yields a posterior distribution of the same type [NAGP05, TML14]. The following Corollary is easily derived.

Corollary 1. *Let $\theta_n > 0$, $K_n := \theta_n \|\hat{v}(y_n)\|_2$ and set $\pi := \mathrm{vMF/S}(\mu_0, \kappa_0)$, with $\mu_0 \in S^{p-1}$ and $\kappa_0 \geq 0$, as prior distribution on β. Then, under the model (i), (ii) of Proposition 1, the posterior distribution of β is given by*

$$\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n} \sim \mathrm{vMF/S}(\mu_n, \kappa_n),$$

with location parameter μ_n equal to the MAP estimator,

$$\mu_n = \hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n) = \frac{K_n \hat{\beta}_{\mathrm{ml}}(y_n) + \kappa_0 \mu_0}{\|K_n \hat{\beta}_{\mathrm{ml}}(y_n) + \kappa_0 \mu_0\|_2},$$

and concentration parameter $\kappa_n = \|K_n \hat{\beta}_{\mathrm{ml}}(y_n) + \kappa_0 \mu_0\|_2$.

In this conjugate framework, the computation of the MAP estimator is straightforward since the mode of the vMF/S distribution coincides with the location parameter: $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n)$ is a linear combination of the prior direction μ_0 with the EPLS estimator $\hat{\beta}_{\mathrm{ml}}(y_n)$. Letting $\kappa_0 \to \infty$ yields $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n) \to \mu_0$, the EPLS estimator is shrunk towards the prior direction. In contrast, setting $\kappa_0 = 0$ amounts to assuming a uniform prior distribution for the direction β and we thus recover the EPLS framework. This behavior is illustrated on the bottom panel of Figure 1 with $\hat{\beta}_{\mathrm{ml}} \propto (3/2, -1, 1/2)^\top$ and $K_n = 1$.

We show in the next Proposition that a similar situation arises when $K_n \overset{\mathbb{P}}{\sim} c\sqrt{n\bar{F}(y_n)} \to \infty$ (where $c > 0$) and the rate of convergence of $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n)$ to β is provided.

Proposition 4. *Under the assumptions of Proposition 2, let $c > 0$ and*

$$\theta_n \overset{\mathbb{P}}{\sim} c\sqrt{n\bar{F}(y_n)}/\|\hat{v}(y_n)\|_2,$$

as $n \to \infty$, then,

$$\sqrt{n\bar{F}(y_n)}\left(\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n) - \beta\right) \xrightarrow{\mathbb{P}} (\kappa_0/c)\, P^{\perp}_{\beta}(\mu_0),$$

where $P^{\perp}_{\beta}(\mu_0)$ denotes the projection of μ_0 on the hyperplane orthogonal to β.

Comparing Proposition 2 and Proposition 4, it appears that the MAP estimator converges to β at a slightly slower rate than the MLE. The two convergence rates however coincide when $P_\beta^\perp(\mu_0) = 0$ *i.e.* when $\mu_0 = \beta$, meaning that the prior distribution is centered on the true (unknown) direction.

3.3 Sparse prior

The EPLS method can be adapted to take into account the information that only a few covariates in X are useful to explain the extreme values of the response variable Y. To this end, consider a Laplace(λ) distribution on the unit sphere:

$$\pi(\beta|\lambda) = \frac{1}{b_p(\lambda)} \exp(-\lambda\|\beta\|_1)\mathbf{1}\{\|\beta\|_2 = 1\}, \text{ with } b_p(\lambda) = \int_{\|x\|_2=1} \exp(-\lambda\|x\|_1)\mathrm{d}x \tag{5}$$

as a prior for $\beta \in S^{p-1}$, where $\lambda \geq 0$ is a concentration parameter. We refer to [Tib96] for the introduction of the Laplace prior in the regression context and to [CK10, VvGBL13] for sparse versions of PLS in a non-extreme context. A graphical representation of the density isocontours of the Laplace distribution in dimension $p = 3$ is provided on the top of Figure 2 for $\lambda \in \{0, 0.2, 0.4, 0.6\}$. On the leftmost panel, the density is nearly uniform on the unit sphere, and it becomes more peaked around the three vertices $(1, 0, 0)^\top$, $(0, 1, 0)^\top$ and $(0, 0, 1)^\top$ as λ increases.

As a consequence of Proposition 3, the posterior distribution can be written as

$$p(\beta|X_{1:n}, Y_{1:n}, \varepsilon_{1:n}) \propto \exp\left(K_n\langle\beta, \hat{\beta}_{\mathrm{ml}}(y_n)\rangle - \lambda\|\beta\|_1\right), \tag{6}$$

for any $\beta \in S^{p-1}$. Although this is not a classical distribution on the unit sphere, the MAP can be computed in closed form:

Corollary 2. *Let $\theta_n > 0$, $K_n := \theta_n\|\hat{v}(y_n)\|_2$ and set $\pi(\cdot|\lambda)$ as the Laplace prior distribution* (5) *on β. Then, under the model (i), (ii) of Proposition 1, the MAP estimator of β is:*

$$\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y_n) = \tilde{\beta}(y_n)/\|\tilde{\beta}(y_n)\|_2, \text{ with } \tilde{\beta}_j(y_n) = S_\lambda(K_n\hat{\beta}_{\mathrm{ml},j}(y_n)),\ j \in \{1, \ldots, p\},$$

and where $S_\lambda(\cdot)$ is the shrinkage operator defined as $S_\lambda(x) = \mathrm{sign}(x)\,(|x| - \lambda)\,\mathbf{1}\{|x| > \lambda\}$, $x \in \mathbb{R}$.

The MAP is obtained by shrinking the coordinates of $\hat{\beta}_{\mathrm{ml}}(y_n)$ associated with the EPLS estimator towards zero. See Figure 3 for an illustration of the shrinkage operator and [CK10, Theorem 3] for a similar result in a non-extreme framework. Note that, when the concentration parameter is set to $\lambda = 0$, we recover the EPLS method. The behavior of the $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ estimator is illustrated on the bottom panel of Figure 2 with $\hat{\beta}_{\mathrm{ml}} \propto (3/2, -1, 1/2)^\top$ and $K_n = 1$. When λ is small, both estimates $\hat{\beta}_{\mathrm{ml}}$ and $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ are superimposed. When λ increases, $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ gets closer and closer to the vertex $(1, 0, 0)^\top$.

Similarly to the conjugate case, when $K_n \overset{\mathbb{P}}{\sim} c\sqrt{n\bar{F}(y_n)} \to \infty$ (where $c > 0$), the rate of convergence of $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y_n)$ to β can be established.

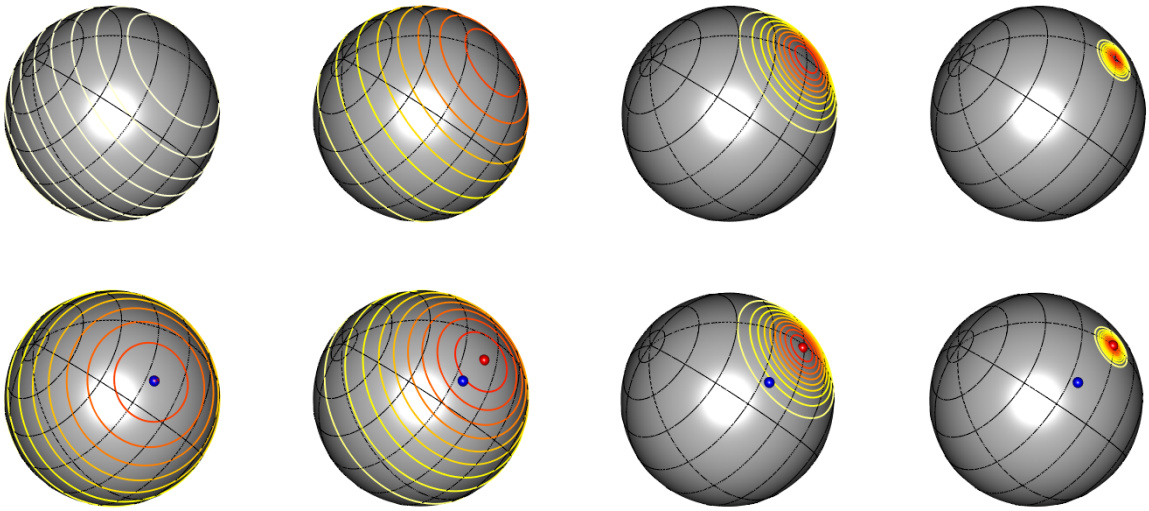

Figure 1: Isocontour plots of the vMF/S(μ_0, κ_0) prior density in dimension $p = 3$ (top) and of the resulting posterior density (bottom) for $\kappa_0 \in \{0, 1, 10, 100\}$ (from left to right). The prior direction is set to $\mu_0 = (1, 0, 0)^\top$. The estimators $\hat{\beta}_{\text{ml}}$ and $\hat{\beta}^{\text{c}}_{\text{map}}$ are depicted by blue and red points respectively.

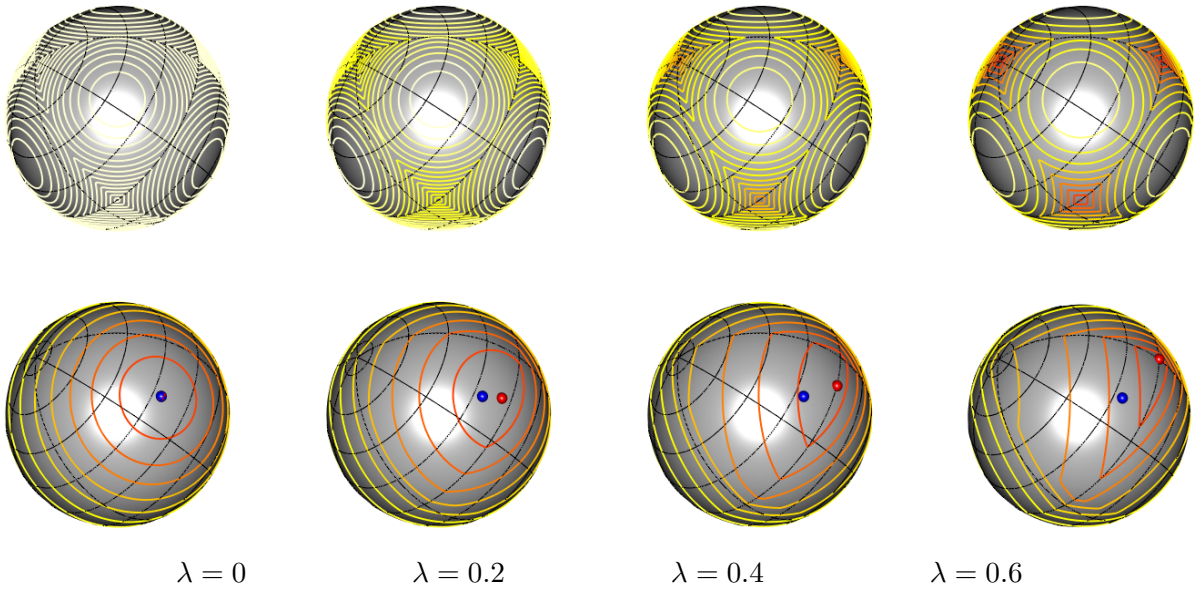

Figure 2: Isocontour plots of the Laplace(λ) prior density in dimension $p = 3$ (top) and of the resulting posterior density (bottom) for $\lambda \in \{0, 0.2, 0.4, 0.6\}$ (from left to right). The estimators $\hat{\beta}_{\text{ml}}$ and $\hat{\beta}^{\text{s}}_{\text{map}}$ are depicted by blue and red points respectively.

Proposition 5. *Under the assumptions of Proposition 2, let $c > 0$ and*

$$\theta_n \overset{\mathbb{P}}{\sim} c\sqrt{n\bar{F}(y_n)}/\|\hat{v}(y_n)\|_2,$$

as $n \to \infty$, then, for all $j \in \{1, \ldots, p\}$ such that $\beta_j \neq 0$,

$$\sqrt{n\bar{F}(y_n)}\left(\hat{\beta}^{\mathrm{s}}_{\mathrm{map},j}(y_n) - \beta_j\right) \xrightarrow{\mathbb{P}} (\lambda/c)\left(\|\beta\|_1\beta_j - \mathrm{sign}(\beta_j)\right).$$

Otherwise, if $\beta_j = 0$, then $\hat{\beta}^{\mathrm{s}}_{\mathrm{map},j}(y_n) = 0$ with probability tending to 1.

It appears that the null coordinates of β are recovered with large probability thanks to the Laplace prior. Similarly to the conjugate case, the MAP estimator usually converges to β at a slower rate than the MLE. Both convergence rates are the same when the non-zero coordinates of β all coincide: $\beta_j = \mathrm{sign}(\beta_j)/\|\beta\|_1$ for all $j \in \{1, \ldots, p\}$ such that $\beta_j \neq 0$.

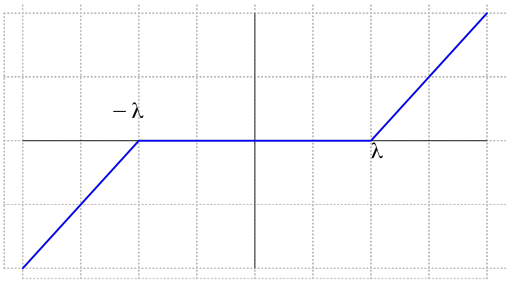

Figure 3: Plot of the shrinkage operator $S_\lambda(\cdot)$ for any $\lambda > 0$.

4 Illustration on simulated data

4.1 Experimental design

The behavior of the SEPaLS estimators $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}$ and $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ is illustrated on the regression model $(\mathbf{A_0})$ with power link function: $t > 0 \mapsto g(t) = t^c$, $c \in \{1/4, 1/2, 1\}$. The output variable Y is distributed from a Pareto distribution with survival function $\bar{F}(y) = (y/2)^{-1/\gamma}$, $y \geq 2$ and with tail-index $\gamma = 1/5$. Each margin $\varepsilon^{(j)}$, $j \in \{1, \ldots, p\}$ of the error ε is simulated as the absolute value of a $\mathcal{N}(0, \sigma^2)$ random variable and depending on Y using the Clayton copula, an Archimedean copula [Nel07, Section 4], defined for all $(u, v) \in [0, 1]^2$ by

$$C_\theta(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta},$$

where $\theta \geq 0$ is a parameter tuning the dependence between the margins. Equivalently, the joint cumulative distribution function of ε is given for all $x \in \mathbb{R}_+^p$ by the one-factor model [KJ13]:

$$F_\varepsilon(x) = \int_0^1 \prod_{j=1}^{p} \frac{\partial C_\theta}{\partial v}(2\Psi(x_j/\sigma) - 1, v)\mathrm{d}v,$$

where Ψ denotes the cumulative distribution function of the standard Gaussian distribution. Note that $C_0(u, v) = uv$ represents the independence copula while, as $\theta \to \infty$, $C_\theta(u, v) \to \min(u, v)$ which represents the co-monotonicity copula. The dependence between the margins is assessed using Kendall's tau $\tau(\theta) = \theta/(\theta+2) \in [0, 1)$ and is thus limited to positive values. We shall also consider the associated rotated copula defined by $\tilde{C}_\theta(u, v) = v - C_\theta(1 - u, v)$ whose Kendall's tau is negative and given by $\tilde{\tau}(\theta) = -\tau(\theta) \in (-1, 0]$, for all $\theta \geq 0$. Here, $\theta \in \{1/2, 8\}$ leads to four possible values of the Kendall's tau: $\{-0.8, -0.2, 0.2, 0.8\}$.

The standard deviation σ is selected such that the Signal to Noise Ratio (SNR), defined as SNR:$=$ $g(\bar{F}^{-1}(1/n))/\sigma$, is equal to 10. Note that $g(\bar{F}^{-1}(1/n))$ represents the approximate maximum value of g on a n-sample from the distribution with associated survival function $\bar{F}$.

The sample size is fixed to $n = 500$ and two dimensions are considered: $p \in \{30, 300\}$. The true direction is $\beta = (1, 1, 0, \ldots, 0)^\top/\sqrt{2}$ for both dimensions.

The location parameter μ_0 of the prior vMF/S distribution (conjugate case) is set either to β, which corresponds to a perfect prior, or to $\tilde{\beta} := (1, \ldots, 1, 0, \ldots, 0)^\top/\sqrt{p/2}$, which is far from the true one, see Subsection 3.2 for details. Four values of the concentration parameter are investigated: $\kappa_0 \in \{0, 10^{-4}, 3.10^{-3}, 10^{-2}\}$. In the case of the Laplace prior (sparse case), we let $\lambda \in \{0, 10^{-4}, 5.10^{-4}, 10^{-3}\}$. In both situations, we set $\theta_n := 1$ since this parameter does not play any role in practice.

4.2 Performance assessment

Let us define a "Proximity Criterion", PC in the following, between the theoretical vector β and its MAP estimator computed on $N = 1000$ replications, as follows:

$$\mathrm{PC}(y) = \frac{1}{N}\sum_{r=1}^{N} \left\langle \hat{\beta}_{\mathrm{map}}^{(r)}(y), \beta \right\rangle^2, \tag{7}$$

where $\hat{\beta}_{\mathrm{map}}^{(r)}$ denotes the MAP estimate on the r^{th} replication under either the conjugate or the sparse prior. Clearly PC $\in [0, 1]$ and the closer PC is to 1, the larger the proximity is. In practice, PC$(Y_{n-k+1,n})$ is computed as a function of the number of exceedances $k \in \{1, \ldots, 100\}$, where $Y_{n-k+1,n}$ denotes the $(n - k + 1)^{\mathrm{th}}$ largest observation from the sample $\{Y_1, \ldots, Y_n\}$.

4.3 Results

Conjugate prior. The proximity criterion $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}$ and β is drawn as a function of $k \in \{1, \ldots, 100\}$ on Figures 4–6 considering 96 configurations in dimension $d = 30$: $\kappa_0 \in \{0, 10^{-4}, 3.10^{-3}, 10^{-2}\}$, $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$, $c \in \{1, 1/2, 1/4\}$ and $\mu_0 \in \{\beta, \tilde{\beta}\}$, see Subsection 4.1 for details. Unsurprisingly, when $\mu_0 = \beta$ *i.e.* when the prior direction points towards the true one, the regularization improves the results of the original EPLS estimator (obtained when $\kappa_0 = 0$). Moreover, it reduces the sensitivity with respect to the number of exceedances k, the dependence degree τ, and the exponent c of the link function. In all situations, one can obtain $\mathrm{PC} \simeq 1$ with $\kappa_0 = 10^{-2}$. In contrast, when $\mu_0 = \tilde{\beta}$, the prior direction is ill-adapted since $\langle \tilde{\beta}, \beta \rangle^2 = 4/p \simeq 0.13$ and too large values of κ_0 deteriorate the EPLS estimator. As expected, the choice of μ_0 is of primary importance in the conjugate prior.

Sparse prior. Similarly, the proximity criterion $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ and β is drawn as a function of $k \in \{1, \ldots, 100\}$ on Figures 7–9 in 96 configurations: $\lambda \in \{0, 10^{-4}, 5.10^{-4}, 10^{-3}\}$, $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$, $c \in \{1, 1/2, 1/4\}$ and $d \in \{30, 300\}$. Here, the regularization always improves the results of the original EPLS estimator (obtained when $\lambda = 0$) since the true direction β is rather sparse, it only has two non-zero coordinates. Enforcing sparsity allows to obtain $\mathrm{PC} \simeq 0.8$ (resp. $\mathrm{PC} \simeq 0.6$) in dimension $p = 30$ (resp. $d = 300$) with exponents $c \geq 1/2$. The case of small exponents ($c = 1/4$) appears to be more complicated, the maximum value of PC depending on the dimension p and on the dependence degree τ.

5 Application to real data

The SEPaLS method is illustrated on data extracted from the Farm Accountancy Data Network (FADN)[3]. This dataset targets French farms described by numerous qualitative and quantitative variables over the period 2000–2015. Here, we focus on the $n = 598$ farms producing field-grown carrots. The response variable Y is the production of carrots (in quintals) and the covariate X is made of $p = 259$ continuous variables including meteorological and economic measurements. Our goal is to investigate, among the 259 collected factors, which ones may influence the upper tail of Y, *i.e.* are linked to large productions of carrots.

Three visual checks are first carried out in Figure 10 to verify whether the heavy-tail hypothesis on Y is realistic. The histogram of the $\{Y_1, \ldots, Y_n\}$ on the top left panel is skewed to the right and has a heavy right tail. Besides, the Hill estimator [Hil75]

$$\hat{\gamma}(k) = \frac{1}{k} \sum_{i=1}^{k} \log(Y_{n-i+1,n} / Y_{n-k,n})$$

[3] Available at: `https://agreste.agriculture.gouv.fr/agreste-web/servicon/I.2/listeTypeServicon/` (in French).

of the tail-index γ is drawn on the top right panel as a function of $k \in \{1, \dots, 500\}$. The resulting graph is stable on the range $k \in \{160, \dots, 280\}$ and points towards $\gamma \simeq 0.72$. Finally, selecting $k = 199$ (this choice is discussed below), the associated quantile-quantile plot of the log-excesses $\log(Y_{n-i+1,n}/Y_{n-k,n})$ against the quantiles $\log(k/i)$ of the unit exponential distribution, $i \in \{1, \dots, k\}$ exhibits a linear trend (bottom panel) which is further empirical evidence that the heavy-tail assumption is appropriate, see [BGST04, pp.109–110].

In the following, we focus on the sparse estimator $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ since the use of $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}$ would require an initial guess for β_0 which is not obvious in this application context. The next two conditional tail correlation measures are introduced to interpret the results obtained with $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$:

$$\rho(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle, Y | Y \geq y) = \frac{\mathrm{cov}(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle, Y | Y \geq y)}{\sigma(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle | Y \geq y)\sigma(Y | Y \geq y)}, \qquad \text{(Figure 11, top panel)}, \tag{8}$$

$$\rho(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle, X^{(j)} | Y \geq y) = \frac{\mathrm{cov}(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle, X^{(j)} | Y \geq y)}{\sigma(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle | Y \geq y)\sigma(X^{(j)} | Y \geq y)}, \quad \text{(Figure 11, bottom panel)}, \tag{9}$$

with $j \in \{1, \dots, p\}$. The role of the tail correlation measure (8) is to assess the correlation in the tail between the response variable Y and the summary $\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle$ of the predictors built by the SEPaLS method. It is computed at the threshold $y = Y_{n-k+1,n}$ and plotted on Figure 11 as a function of the number of exceedances k for several levels of shrinkage λ. Note that, when k is small, the correlation vanishes for a wide range of λ values since, in this case, the prior weight is too large compared to the likelihood one. The global maximum is located at $k = 199$ which corresponds to a stable region of the Hill estimator according to Figure 10. The maximum correlation ($\rho \simeq 0.79$) is reached at $\lambda = 353$.

The role of the tail correlation measure (9) is to assess the correlation in the tail between the summary $\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle$ of the predictors built by the SEPaLS method and the initial ones $X^{(j)}$, $j \in \{1, \dots, p\}$. It is computed at the threshold $y = Y_{n-k+1,n}$ and plotted on the bottom left panel of Figure 11 as a function of the number of exceedances k for $\lambda = 353$. All correlation curves feature a nice stability with respect to k, especially in the neighborhood of $k = 199$.

In the sequel, we thus select $k = 199$ and $\lambda = 353$. With these choices, only 5 coordinates of $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ out of 259 are estimated to non-zero values, see the bottom right panel of Figure 11 for an illustration and Table 1 for a description of the selected variables. Meteorological variables are discarded since large productions of carrots do not seem to depend on weather conditions. Remarking on Figure 10 that the summary variable $\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle$ is positively correlated with the high values of Y, one can conclude that, unsurprisingly, large productions are associated with large cultivated areas (`SUD4CARO`), large amounts of work both in terms of time (`UTASA`, `UTATO`) and remuneration charges (`FPERS`), and large investments in supplies (`CHRFO`).

Table 1: Real data example. Description of the 5 selected variables (out of 259) associated with 598 farms producing field-grown carrots in France from 2000 to 2015. The last column displays the corresponding non-zero coordinates of $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$. $^{(\star)}$ UTA: amount of work associated with one full-time working person during one year.

Selected variables	**Description**	**Units**	$\hat{\beta}^{\mathrm{s}}_{\mathrm{map},j}$
SUD4CARO	Area cultivated with field-grown carrots	hectares	0.978
UTASA	Salaried work	UTA$^{(\star)}$	0.158
UTATO	Salaried and not salaried work	UTA$^{(\star)}$	0.124
CHRFO	Actual cost of stored supplies	euros	0.038
FPERS	Remuneration charges	euros	0.026

6 Discussion

We proposed a Bayesian interpretation of the EPLS model to introduce prior information on the direction of dimension reduction for extreme values. Two examples of shrinkage priors are provided: a conjugate von Mises-Fisher prior allowing to consider an initial guess on the direction, and a Laplace prior enforcing sparsity on the estimated direction. Finite sample experiments demonstrate that the proposed method is effective in high dimension ($d = 300$ on simulated data and $d \simeq 260$ on real data) with moderate sample sizes ($n = 500$ on simulated data and $n \simeq 600$ on real data).

Here, we limited ourselves to the estimation of one single direction, but the SEPaLS method can easily be adapted to the estimation of multiple directions using the iterative procedure described in [BEG23, Section 4]. We also focused on prior distributions yielding explicit shrinkage estimators. It would be of interest to investigate the use of other priors: either uninformative priors such as Jeffreys' one [Jef46] or other shrinkage priors [VEOM19] can be considered. The computation of the posterior mode estimate would rely on an MCMC procedure.

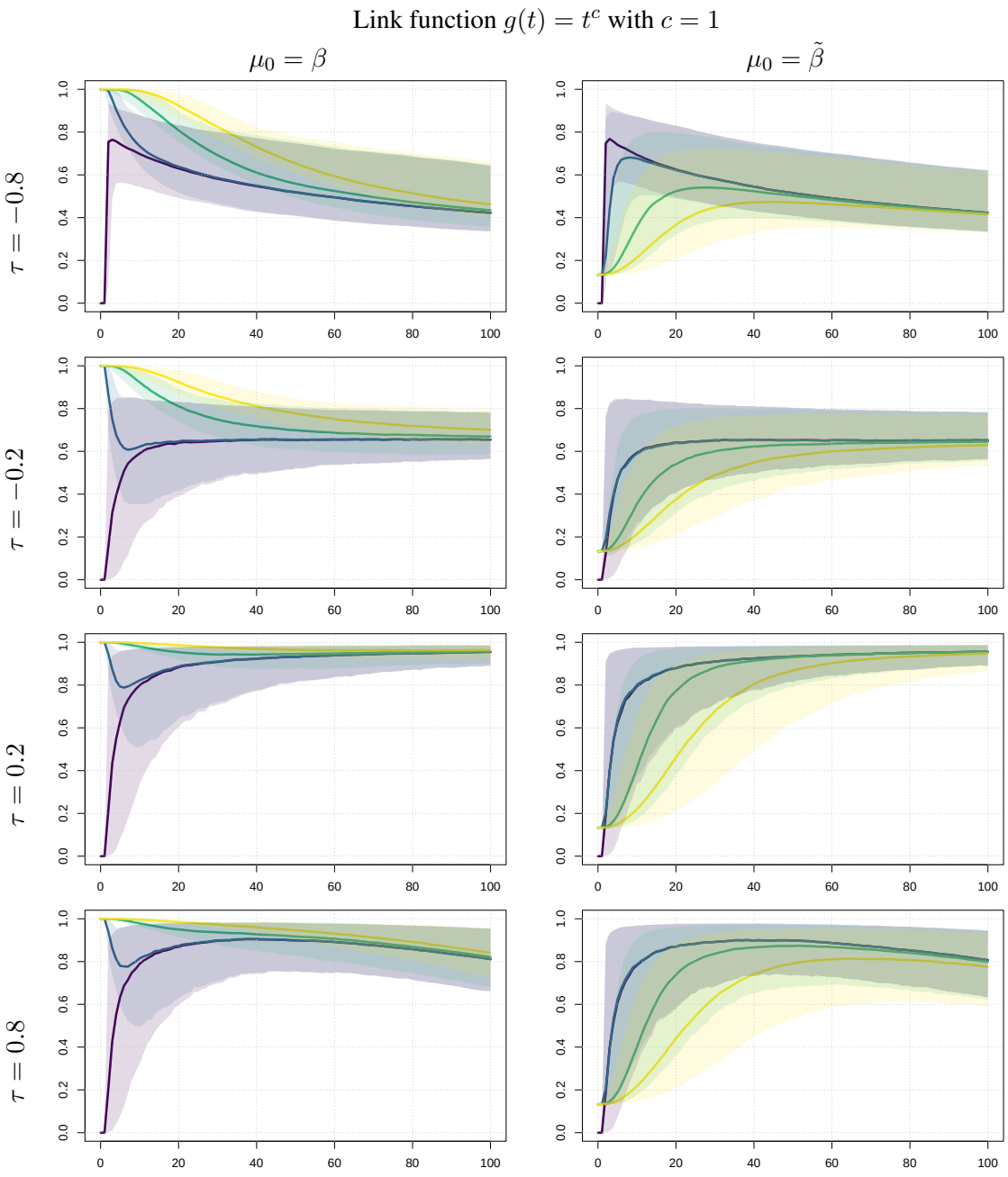

Figure 4: Finite sample behavior of the SEPaLS estimator computed with the conjugate prior on simulated data in dimension $d = 30$ from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1$. Vertically: $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}$ and β for a prior direction $\mu_0 = \beta$ (left) or $\mu_0 = \tilde{\beta}$ (right) as a function of the number $k \in \{1, \ldots, 100\}$ of exceedances (horizontally). The concentration parameter is $\kappa_0 \in \{0, 10^{-4}, 3.10^{-3}, 10^{-2}\}$, respectively in violet, blue, green and yellow. Colored areas correspond to 90% confidence intervals, lines to medians.

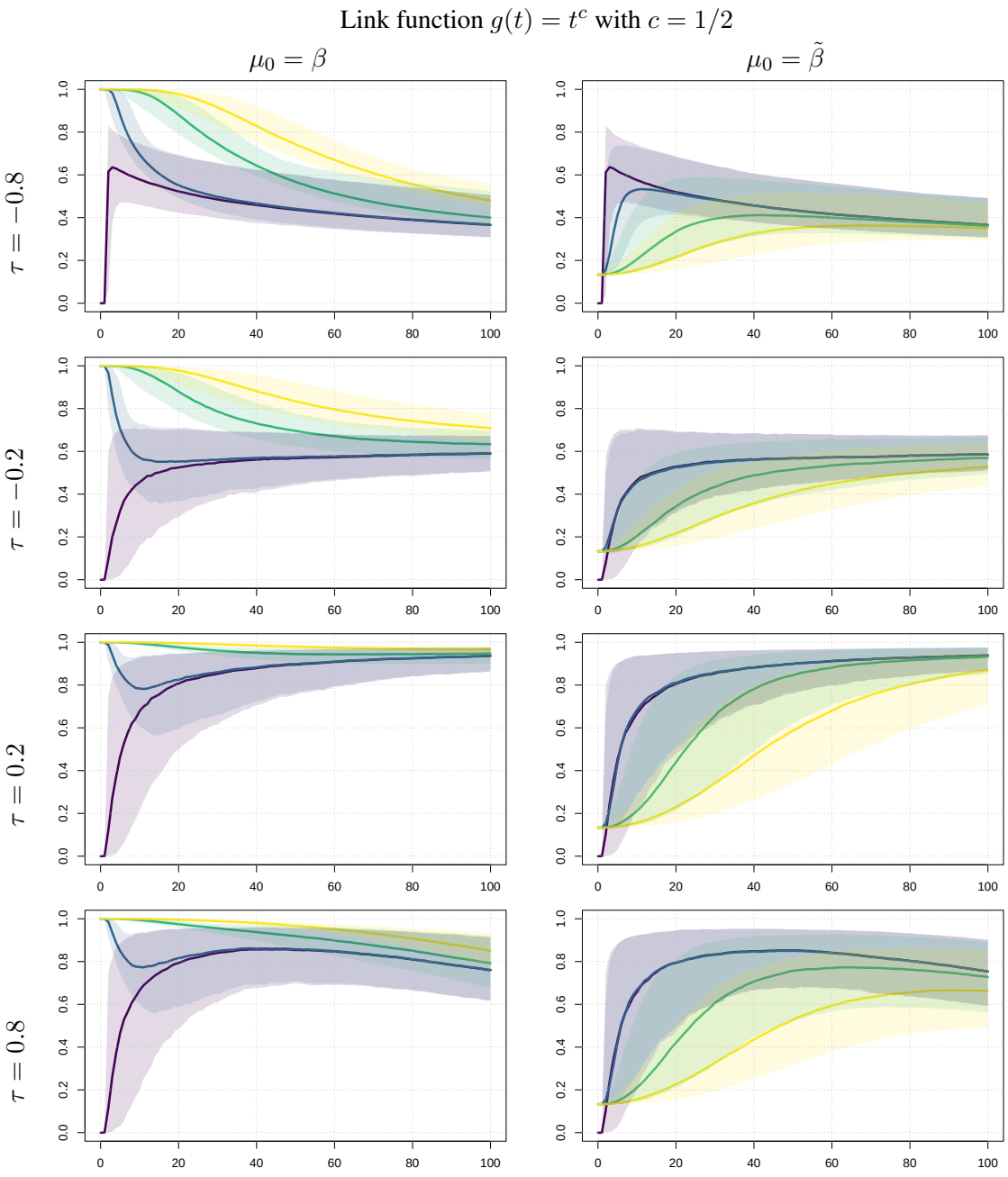

Figure 5: Finite sample behavior of the SEPaLS estimator computed with the conjugate prior on simulated data in dimension $d = 30$ from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1/2$. Vertically: $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}$ and β for a prior direction $\mu_0 = \beta$ (left) or $\mu_0 = \tilde{\beta}$ (right) as a function of the number $k \in \{1, \ldots, 100\}$ of exceedances (horizontally). The concentration parameter is $\kappa_0 \in \{0, 10^{-4}, 3.10^{-3}, 10^{-2}\}$, respectively in violet, blue, green and yellow. Colored areas correspond to 90% confidence intervals, lines to medians.

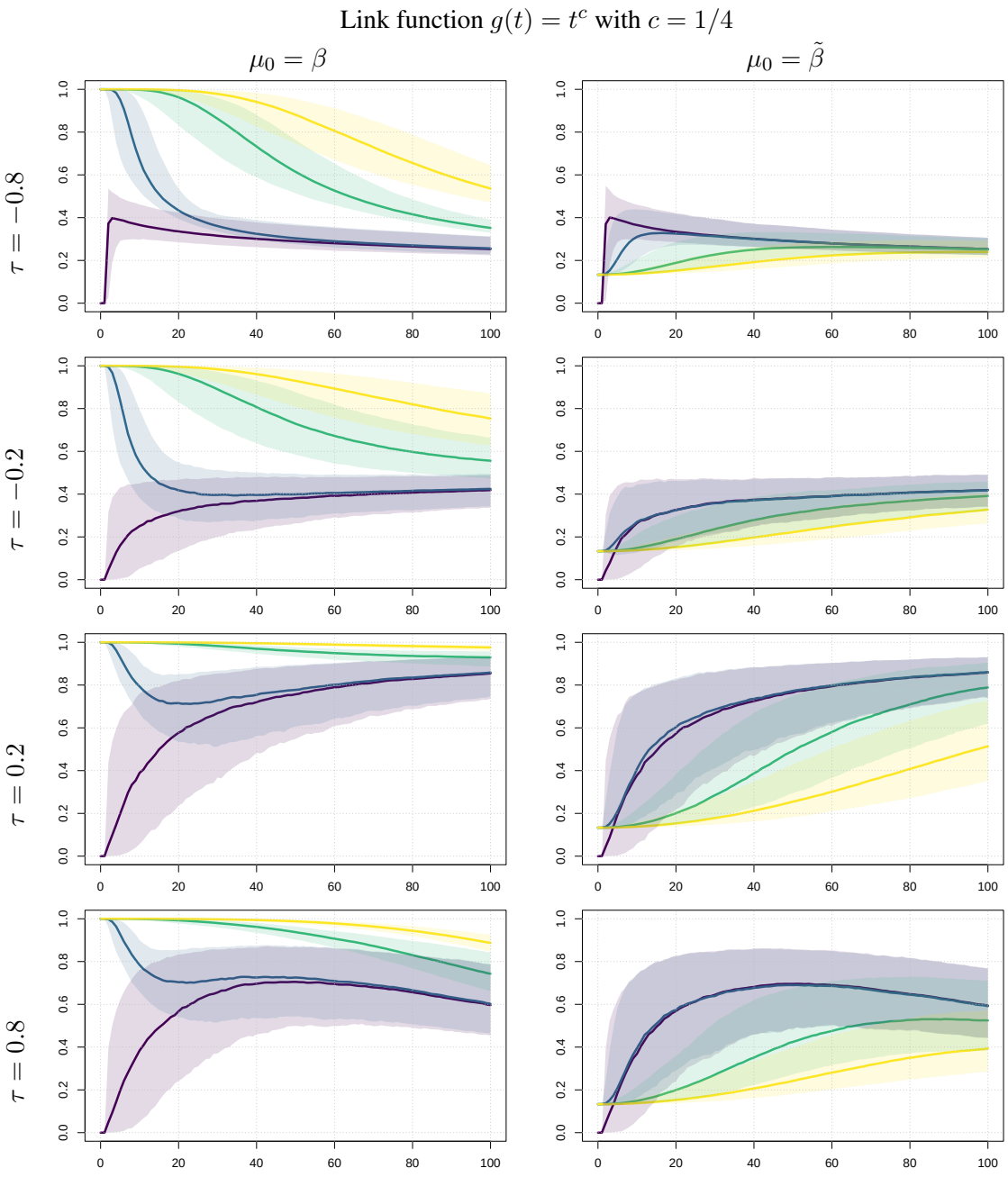

Figure 6: Finite sample behavior of the SEPaLS estimator computed with the conjugate prior on simulated data in dimension $d = 30$ from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1/4$. Vertically: $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}$ and β for a prior direction $\mu_0 = \beta$ (left) or $\mu_0 = \tilde{\beta}$ (right) as a function of the number $k \in \{1, \ldots, 100\}$ of exceedances (horizontally). The concentration parameter is $\kappa_0 \in \{0, 10^{-4}, 3.10^{-3}, 10^{-2}\}$, respectively in violet, blue, green and yellow. Colored areas correspond to 90% confidence intervals, lines to medians.

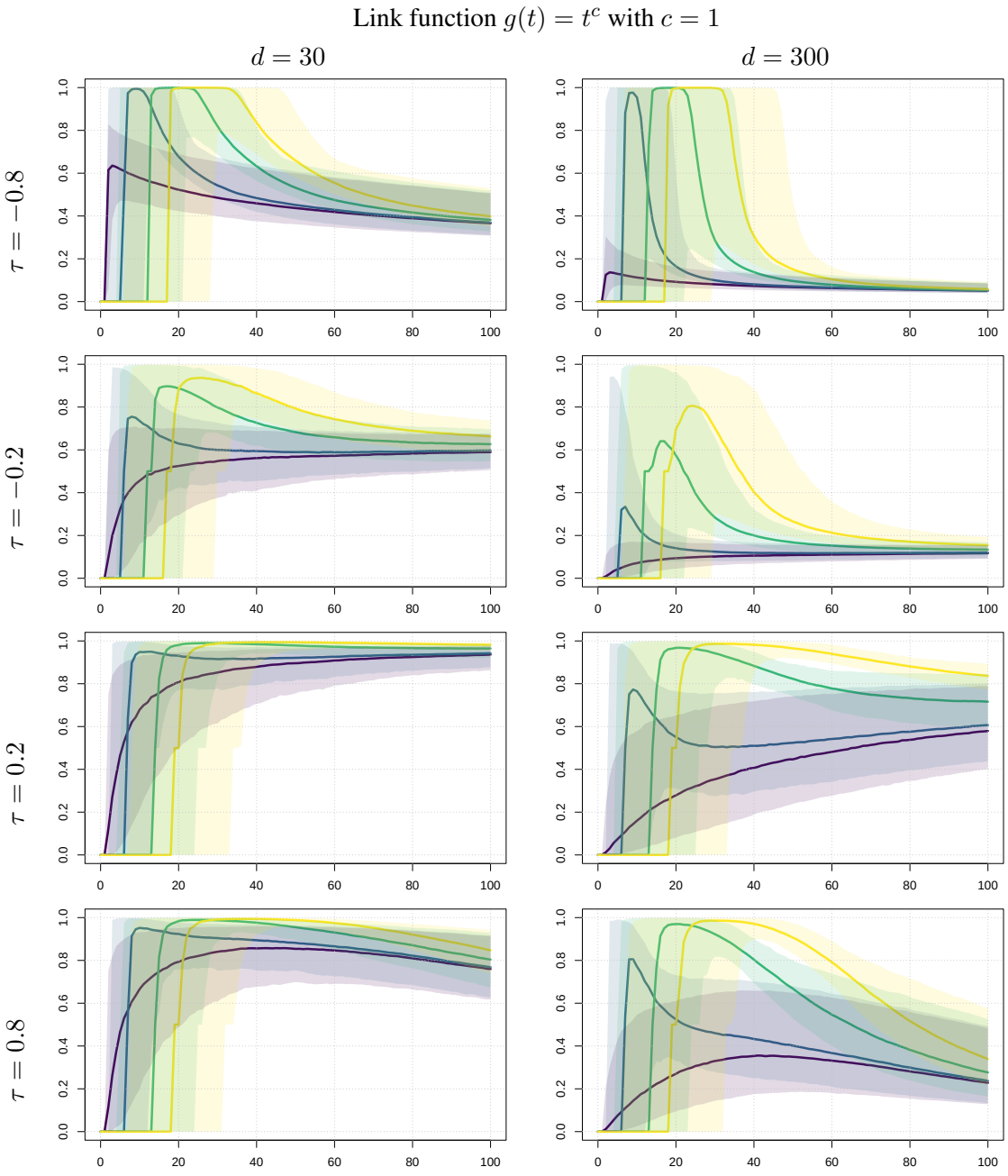

Figure 7: Finite sample behavior of the SEPaLS estimator computed with the sparse prior on simulated data in dimension from a Pareto distribution ($\gamma = 1/5,\ \ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1$. Vertically: $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ and β for $d = 30$ (left) and $d = 300$ (right) as a function of the number $k \in \{1, \ldots, 100\}$ of exceedances (horizontally). The concentration parameter is $\lambda \in \{0, 10^{-4}, 5.10^{-4}, 10^{-3}\}$, respectively in violet, blue, green and yellow. Colored areas correspond to 90% confidence intervals, lines to medians.

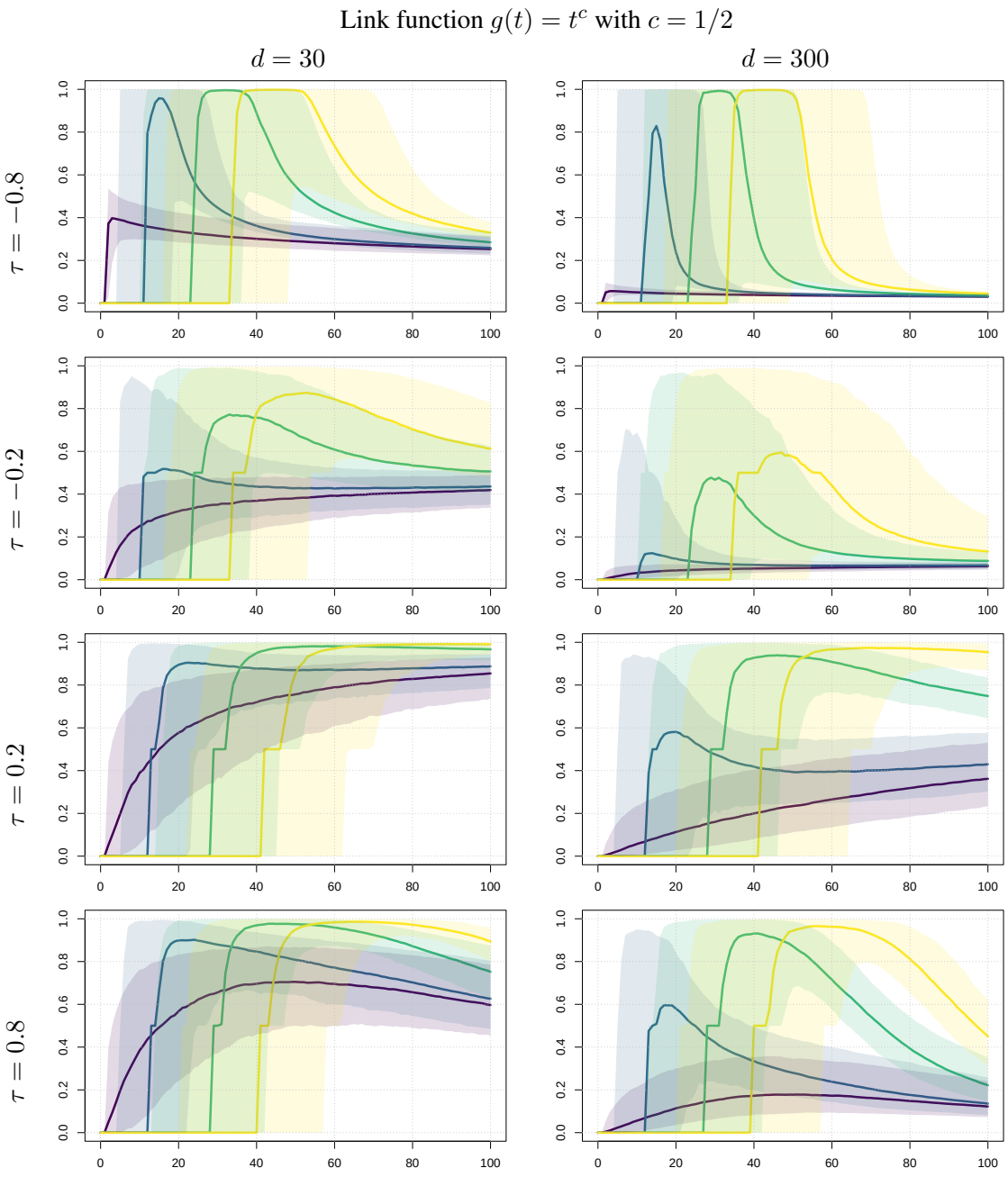

Figure 8: Finite sample behavior of the SEPaLS estimator computed with the sparse prior on simulated data in dimension from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1/2$. Vertically: $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ and β for $d = 30$ (left) and $d = 300$ (right) as a function of the number $k \in \{1, \ldots, 100\}$ of exceedances (horizontally). The concentration parameter is $\lambda \in \{0, 10^{-4}, 5.10^{-4}, 10^{-3}\}$, respectively in violet, blue, green and yellow. Colored areas correspond to 90% confidence intervals, lines to medians.

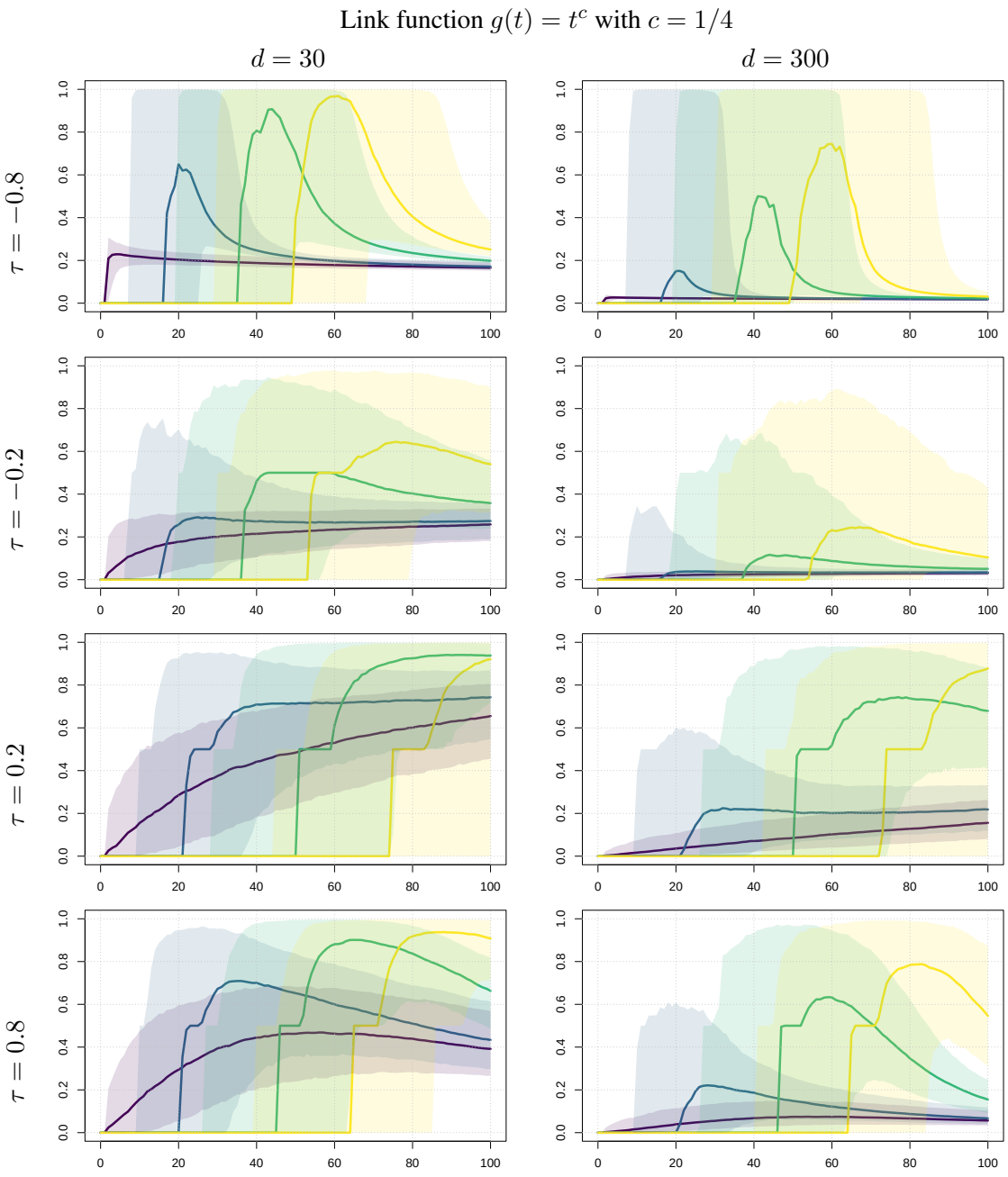

Figure 9: Finite sample behavior of the SEPaLS estimator computed with the sparse prior on simulated data in dimension from a Pareto distribution ($\gamma = 1/5,\ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1/4$. Vertically: $\mathrm{PC}(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ and β for $d = 30$ (left) and $d = 300$ (right) as a function of the number $k \in \{1, \ldots, 100\}$ of exceedances (horizontally). The concentration parameter is $\lambda \in \{0, 10^{-4}, 5.10^{-4}, 10^{-3}\}$, respectively in violet, blue, green and yellow. Colored areas correspond to 90% confidence intervals, lines to medians.

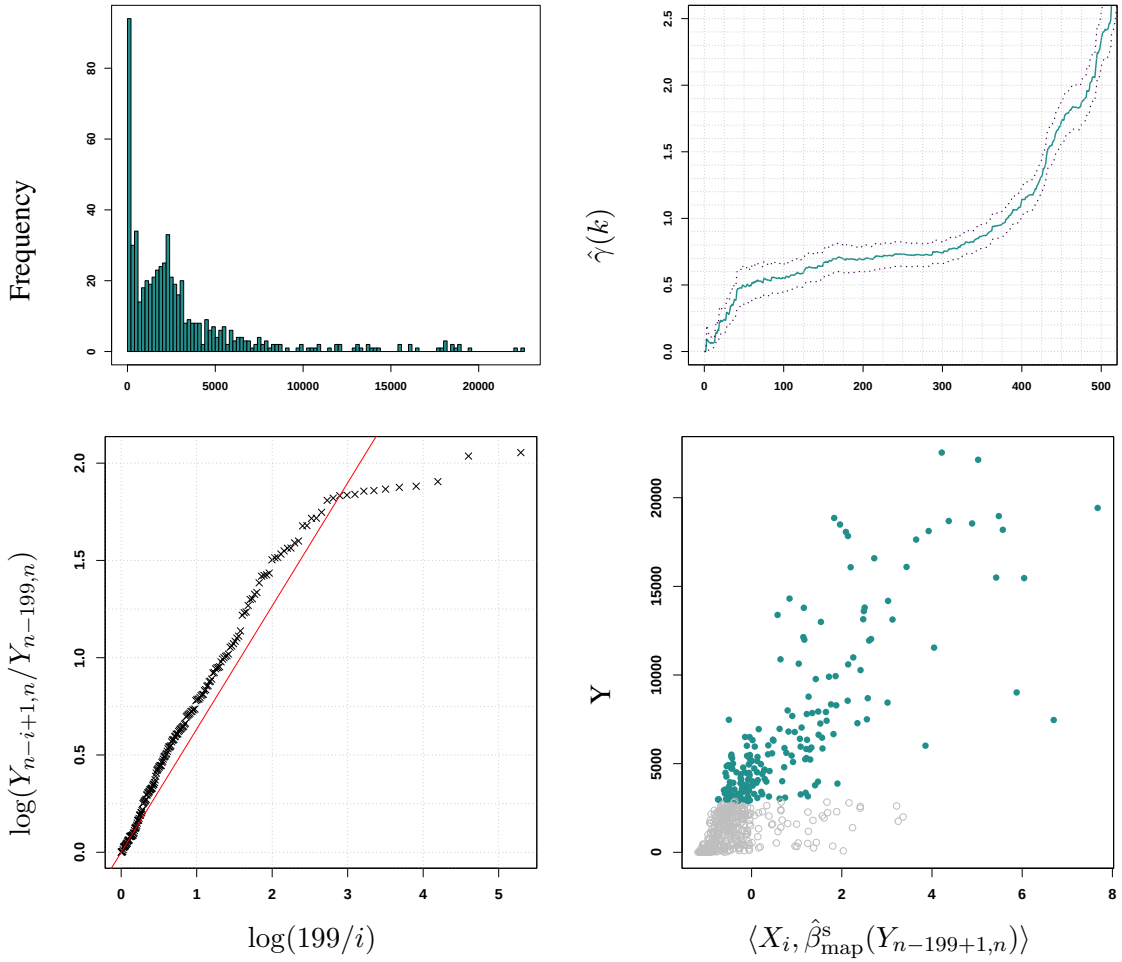

Figure 10: Real data example. Top left: Histogram of $\{Y_1, \ldots, Y_n\}$. Top right: Hill plot $k \in \{1, \ldots, 500\} \mapsto \hat{\gamma}(k)$ and associated confidence intervals (dotted lines). Bottom left: Quantile-quantile plot (horizontally: $\log(k/i)$, vertically: $\log(Y_{n-i+1,n}/Y_{n-k,n})$, for $i \in \{1, \ldots, k\}$) drawn with $k = 199$, the regression line is superimposed in red. Bottom right: Scatter-plot $(\langle X_i, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(Y_{n-k+1,n})\rangle, Y_i)$, $i \in \{1, \ldots, n\}$ with $k = 199$. Points below the threshold ($Y_i \leq Y_{n-k+1,n}$) are depicted in gray.

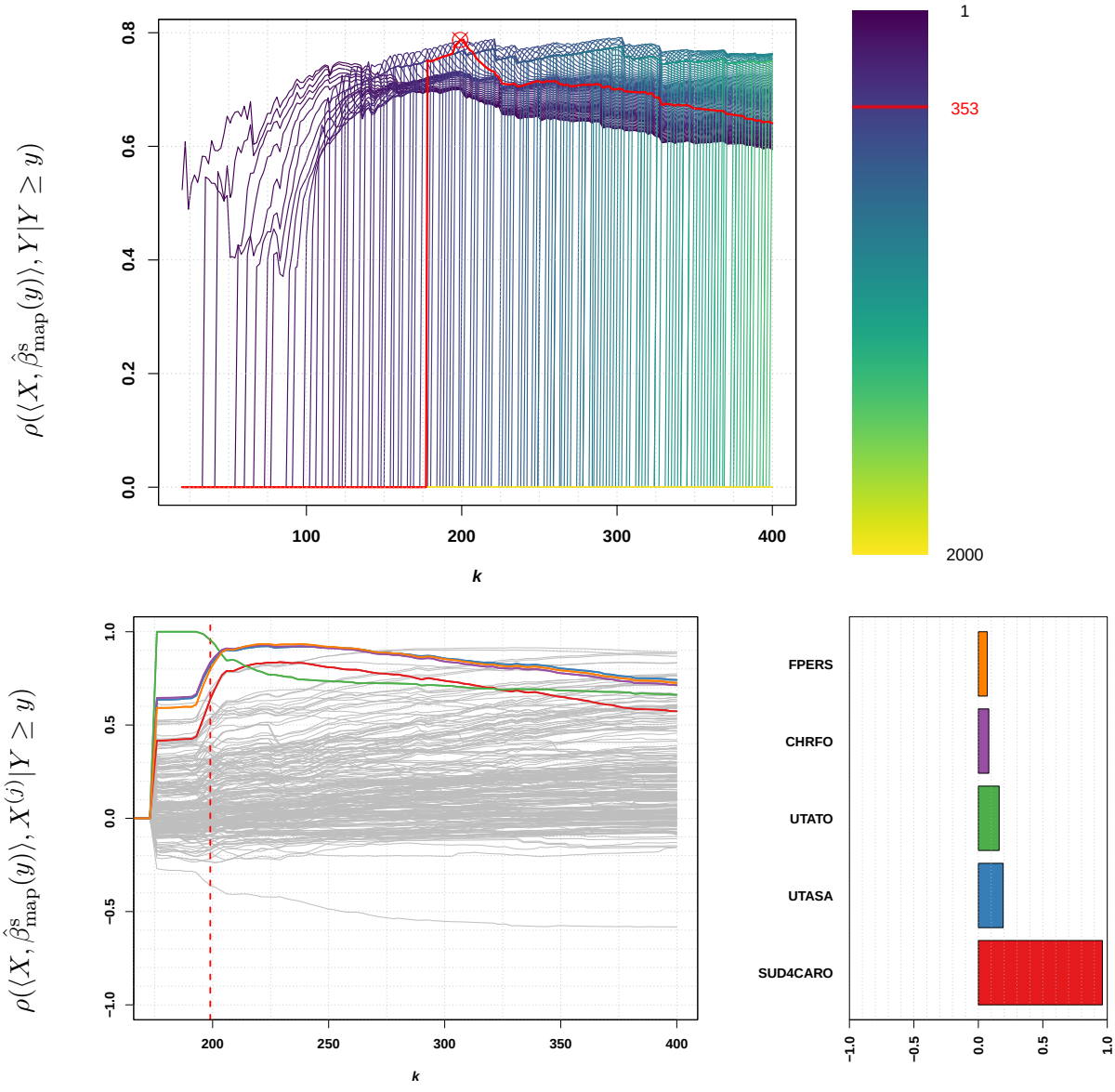

Figure 11: Real data example. Top: Correlation $\rho(\langle X, \hat{\beta}^{s}_{map}(y)\rangle, Y|Y \geq y)$ defined in Equation (8) computed at $y = Y_{n-k+1,n}$ as a function of $k \in \{20, \ldots, 400\}$ for 200 evenly distributed values of λ in $\{1, \ldots, 2000\}$. The selected pair $(k, \lambda) = (199, 353)$ is depicted in red. Bottom left: Correlation $\rho(\langle X, \hat{\beta}^{s}_{map}(y)\rangle, X^{(j)}|Y \geq y)$ defined in Equation (9) computed at $y = Y_{n-k+1,n}$ as a function of $k \in \{175, \ldots, 400\}$ for $\lambda = 353$ and $j \in \{1, \ldots, 259\}$. Bottom right: Non-zero coordinates of $\hat{\beta}^{s}_{map}(Y_{n-k+1,n})$ for the selected pair $(k, \lambda) = (199, 353)$. The color code is the same for both left and right panels.

7 Appendix: Proofs

This first Lemma establishes that $f_{\mathrm{vMF/B}}(\cdot|\mu, r, \kappa)$ is a proper density function integrating to one.

Lemma 1. *Let $p \geq 2$. For all $\mu \in S^{p-1}$, $r > 0$ and $\kappa \geq 0$,*

$$\int_{\|x\|_2 \leq r} \frac{1}{r^p} \exp\left(\frac{\kappa\langle \mu, x\rangle}{r}\right) \mathrm{d}x = \frac{1}{2\pi c_{p+2}(\kappa)},$$

where $c_{p+2}(\kappa)$ is defined in (3)*.*

Proof (of Lemma 1). The change of variable $x \mapsto y = x/r$ leads to

$$\int_{\|x\|_2 \leq r} \frac{1}{r^p} \exp\left(\frac{\kappa\langle \mu, x\rangle}{r}\right) \mathrm{d}x = \int_{\|y\|_2 \leq 1} \exp(\kappa\langle \mu, y\rangle)\mathrm{d}y,$$

and switching to polar coordinates yields

$$\begin{aligned}
\int_{\|y\|_2 \leq 1} \exp(\kappa\langle \mu, y\rangle)\mathrm{d}y &= \int_0^1 \rho^{p-1} \int_{S^{p-1}} \exp(\rho\kappa\langle \mu, u\rangle)\mathrm{d}u\mathrm{d}\rho, \\
&= \int_0^1 \frac{\rho^{p-1}}{c_p(\rho\kappa)}\mathrm{d}\rho \\
&= \frac{(2\pi)^{p/2}}{\kappa^{p/2-1}} \int_0^1 \rho^{p/2} I_{p/2-1}(\rho\kappa)\mathrm{d}\rho \\
&= \frac{(2\pi)^{p/2}}{\kappa^p} \int_0^\kappa t^{p/2} I_{p/2-1}(t)\mathrm{d}t.
\end{aligned}$$

From the definition of the modified Bessel function (4) as a power series with infinite radius of convergence, one has:

$$\begin{aligned}
\int_0^\kappa t^{p/2} I_{p/2-1}(t)\mathrm{d}t &= \sum_{\ell=0}^\infty \left(\frac{1}{2^{2\ell+p/2-1}\Gamma(p/2+\ell)\ell!} \int_0^\kappa t^{2\ell+p-1}\mathrm{d}t\right) \\
&= \sum_{l=0}^\infty \frac{\kappa^{2\ell+p}}{2^{2\ell+p/2-1}\Gamma(p/2+\ell)\ell!\,(2\ell+p)}.
\end{aligned}$$

Taking account of $(p/2+\ell)\Gamma(p/2+\ell) = \Gamma(p/2+\ell+1)$, it follows

$$\int_0^\kappa t^{p/2} I_{p/2-1}(t)\mathrm{d}t = \kappa^{p/2} \sum_{\ell=0}^\infty \frac{1}{\Gamma(p/2+\ell+1)\ell!} \left(\frac{\kappa}{2}\right)^{2\ell+p/2} = \kappa^{p/2} I_{p/2}(\kappa),$$

leading to

$$\int_{\|x\|_2 \leq r} \frac{1}{r^p} \exp\left(\frac{\kappa\langle \mu, x\rangle}{r}\right) \mathrm{d}x = \frac{(2\pi)^{p/2}}{\kappa^{p/2}} I_{p/2}(\kappa) = \frac{1}{2\pi c_{p+2}(\kappa)},$$

which concludes the proof.

Proof (of Proposition 1). For any $\theta_n > 0$, in view of (2), the optimization problem (1) can be rewritten as:

$$\hat{\beta}(y_n) = \underset{\|\beta\|_2=1}{\operatorname{argmax}} \exp\left(\theta_n \langle \beta, \hat{v}(y_n)\rangle\right) = \underset{\|\beta\|_2=1}{\operatorname{argmax}} \prod_{i=1}^{n} \exp\left(\theta_n \langle \beta, X_i\rangle \Phi_i(y_n, Y_{1:n})\right). \tag{10}$$

Under model $(\mathbf{A_0})$, the triangle inequality yields $\|X_i\|_2 \leq |g(Y_i)| + \|\varepsilon_i\|_2$, and thus, conditionally on (Y_i, ε_i), X_i belongs to the ball centered at 0 with radius $r_i := |g(Y_i)| + \|\varepsilon_i\|_2$. The optimization problem (10) can be rewritten in terms of densities associated with the vMF/B distribution as

$$\hat{\beta}(y_n) = \underset{\|\beta\|_2=1}{\operatorname{argmax}} \prod_{i=1}^{n} f_{\mathrm{vMF/B}}\left(X_i|\beta, r_i = |g(Y_i)| + \|\varepsilon_i\|_2, \kappa_i = \theta_n r_i \Phi_i(y_n, Y_{1:n})\right).$$

It appears that $\hat{\beta}$ can be interpreted as the estimator maximizing the likelihood conditionally on $(Y_{1:n}, \varepsilon_{1:n})$. Since the density $p(\cdot,\cdot)$ of $(Y_{1:n}, \varepsilon_{1:n})$ does not depend on β, one also has

$$\hat{\beta}(y_n) = \underset{\|\beta\|_2=1}{\operatorname{argmax}} \left(\prod_{i=1}^{n} f_{\mathrm{vMF/B}}\left(X_i|\beta, r_i = |g(Y_i)| + \|\varepsilon_i\|_2, \kappa_i = \theta_n r_i \Phi_i(y_n, Y_{1:n})\right) \right) p(Y_{1:n}, \varepsilon_{1:n}),$$

and thus $\hat{\beta}(y_n)$ can also be viewed as the unconditional maximum likelihood estimator of β.

The next Lemma will reveal useful in the proof of Proposition 2 below.

Lemma 2. *Let (σ_n) and (c_n) be positive real sequences with $\sigma_n \to 0$ as $n \to \infty$. Let A be a random vector in $\mathbb{R}^p$, $b \in S^{p-1}$ a non-random vector, and (B_n) a sequence of random vectors in $\mathbb{R}^p$ such that*

$$\sigma_n^{-1}\left(\frac{B_n}{c_n} - b\right) \xrightarrow{d} A.$$

Then,

$$\sigma_n^{-1}\left(\frac{B_n}{\|B_n\|_2} - b\right) \xrightarrow{\mathbb{P}} P_b^{\perp}(A),$$

where $P_b^{\perp}(A) := A - \langle b, A\rangle b$ denotes the projection of A on the hyperplane orthogonal to b.

Proof (of Lemma 2). Let $\epsilon_n := \sigma_n^{-1}\left(\frac{B_n}{c_n} - b\right) - A$. From the assumption of convergence in distribution, we have that ϵ_n converges in distribution to a Dirac mass at 0. Clearly,

$$\|B_n\|_2^2 = c_n^2 \|b + \sigma_n(A + \epsilon_n)\|_2^2 = c_n^2\left(1 + 2\sigma_n\langle b, A + \epsilon_n\rangle + \mathcal{O}_{\mathbb{P}}(\sigma_n^2)\right),$$

and inverting the latter equality yields

$$c_n = \|B_n\|_2\left(1 - \sigma_n\langle b, A + \epsilon_n\rangle + \mathcal{O}_{\mathbb{P}}(\sigma_n^2)\right).$$

Replacing in the expression of $B_n = c_n(b + \sigma_n(A + \epsilon_n))$, we obtain

$$\begin{aligned} B_n &= \|B_n\|_2\left(1 - \sigma_n\langle b, A + \epsilon_n\rangle + \mathcal{O}_{\mathbb{P}}(\sigma_n^2)\right)\left(b + \sigma_n(A + \epsilon_n)\right) \\ &= \|B_n\|_2\left(b + \sigma_n(A + \epsilon_n - b\langle b, A + \epsilon_n\rangle) + \mathcal{O}_{\mathbb{P}}(\sigma_n^2)\right), \end{aligned}$$

and therefore

$$\sigma_n^{-1}\left(\frac{B_n}{\|B_n\|_2}-b\right)=A+\epsilon_n-b\langle b,A+\epsilon_n\rangle+\mathcal{O}_{\mathbb{P}}(\sigma_n^2)\xrightarrow{\mathbb{P}}A-b\langle b,A\rangle=P_b^{\perp}(A),$$

which is the desired result.

Proof (of Proposition 2). From [BEG23, Theorem 1], one has

$$\sqrt{n\bar{F}(y_n)}\left(\frac{\hat{v}(y_n)}{\|v(y_n)\|_2}-\beta\right)\xrightarrow{d}\xi\beta,$$

with ξ a centered Gaussian random variable and where

$$v(y_n):=\bar{F}(y_n)\mathbb{E}(XY\mathbf{1}_{\{Y\geq y_n\}})-\mathbb{E}(X\mathbf{1}_{\{Y\geq y_n\}})\mathbb{E}(Y\mathbf{1}_{\{Y\geq y_n\}}).$$

The result follows from Lemma 2 applied with $\sigma_n=1/\sqrt{n\bar{F}(y_n)}$, $B_n=\hat{v}(y_n)$, $c_n=\|v(y_n)\|_2$, $b=\beta$, $A=\xi\beta$ and therefore $P_b^{\perp}(A)=0$.

Proof (of Proposition 3). In view of Bayes' rule, the posterior distribution of β is given by

$$p(\beta|X_{1:n},Y_{1:n},\varepsilon_{1:n})\propto\pi(\beta)p(Y_{1:n},\varepsilon_{1:n})\prod_{i=1}^{n}f_{\mathrm{vMF/B}}\left(X_i|\beta,r_i=|g(Y_i)|+\|\varepsilon_i\|_2,\kappa_i=\theta_n r_i\Phi_i(y_n,Y_{1:n})\right).$$

Since $p(Y_{1:n},\varepsilon_{1:n})$ does not depend on β, the posterior distribution can be simplified as

$$\begin{aligned}p(\beta|X_{1:n},Y_{1:n},\varepsilon_{1:n})&\propto\pi(\beta)\prod_{i=1}^{n}f_{\mathrm{vMF/B}}\left(X_i|\beta,r_i=|g(Y_i)|+\|\varepsilon_i\|_2,\kappa_i=\theta_n r_i\Phi_i(y_n,Y_{1:n})\right)\\&\propto\pi(\beta)\prod_{i=1}^{n}\exp\left(\theta_n\langle\beta,X_i\rangle\Phi_i(y_n,Y_{1:n})\right)\\&=\pi(\beta)\exp\left(\theta_n\|\hat{v}(y_n)\|_2\langle\beta,\hat{\beta}_{\mathrm{ml}}(y_n)\rangle\right),\end{aligned}$$

and the result is proved.

Proof (of Proposition 4). Let $\sigma_n=1/\sqrt{n\bar{F}(y_n)}$. Combining Corollary 1 and Proposition 2, it follows

$$\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n)=\frac{\beta+\sigma_n\varepsilon_n+(\kappa_0/K_n)\mu_0}{\|\beta+\sigma_n\varepsilon_n+(\kappa_0/K_n)\mu_0\|_2},$$

where $\varepsilon_n:=\sigma_n^{-1}(\hat{\beta}_{\mathrm{ml}}(y_n)-\beta)\xrightarrow{\mathbb{P}}0$. Taking account of $\sigma_n\to0$ and $1/K_n\overset{\mathbb{P}}{\sim}\sigma_n/c\to0$ as $n\to\infty$, a first order Taylor expansion yields:

$$\|\beta+\sigma_n\varepsilon_n+(\kappa_0/K_n)\mu_0\|_2^2=1+2(\kappa_0/K_n)\langle\mu_0,\beta\rangle+o_{\mathbb{P}}(\sigma_n)+o_{\mathbb{P}}(1/K_n),$$

and therefore

$$1/\|\beta+\sigma_n\varepsilon_n+(\kappa_0/K_n)\mu_0\|_2=1-(\kappa_0/K_n)\langle\mu_0,\beta\rangle+o_{\mathbb{P}}(\sigma_n)+o_{\mathbb{P}}(1/K_n).$$

Replacing, we get

$$\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n) = \beta + (\kappa_0/K_n)(\mu_0 - \langle \mu_0, \beta \rangle \beta) + o_{\mathbb{P}}(\sigma_n) + o_{\mathbb{P}}(1/K_n),$$

or equivalently,

$$\sigma_n^{-1}(\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n) - \beta) = \kappa_0/(\sigma_n K_n)(\mu_0 - \langle \mu_0, \beta \rangle \beta) + o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1/(\sigma_n K_n)),$$

and the result is proved under the assumption that $\sigma_n K_n \xrightarrow{\mathbb{P}} c > 0$ as $n \to \infty$.

Proof (of Corollary 2). In view of (6), the MAP estimator is given by:

$$\begin{aligned}
\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y_n) &= \operatorname*{argmin}_{\|\beta\|_2^2=1} \lambda \|\beta\|_1 - K_n \langle \beta, \hat{\beta}_{\mathrm{ml}}(y_n) \rangle \\
&= \operatorname*{argmin}_{\|\beta\|_2^2=1} \sum_{j=1}^{p} \Big(\lambda |\beta_j| - K_n \beta_j \hat{\beta}_{\mathrm{ml},j}(y_n) \Big) \\
&= \operatorname*{argmin}_{\|\beta\|_2^2=1} \sum_{j=1}^{p} |\beta_j| \Big(\lambda - K_n \operatorname{sign}(\beta_j) \hat{\beta}_{\mathrm{ml},j}(y_n) \Big).
\end{aligned}$$

Introducing $b_j = |\beta_j|$ and $s_j = \operatorname{sign}(\beta_j)$ so that $\beta_j = s_j b_j$, the above optimization problem can be rewritten as

$$\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y_n) = \operatorname*{argmin}_{b,s} \sum_{j=1}^{p} b_j (\lambda - K_n s_j \hat{\beta}_{\mathrm{ml},j}(y_n)) \;\text{ s.t. } \|b\|_2^2 = 1,\ b_j \geq 0,\ |s_j| = 1,\ j \in \{1, \ldots, p\}.$$

Clearly, the solution w.r.t. s is given by $s_j = \operatorname{sign}(\hat{\beta}_{\mathrm{ml},j}(y_n))$ for all $j \in \{1, \ldots, p\}$ and therefore

$$\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y) = \operatorname*{argmin}_{b \in \mathbb{R}^p} C(b), \;\; \text{s.t. } \|b\|_2^2 = 1,\ b_j \geq 0,\ j \in \{1, \ldots, p\}$$

where

$$C(b) = \sum_{j=1}^{p} b_j (\lambda - K_n |\hat{\beta}_{\mathrm{ml},j}(y_n)|).$$

Let us introduce the two sets of indices

$$J_+ = \Big\{ j \in \{1, \ldots, p\} \,;\, \lambda - K_n |\hat{\beta}_{\mathrm{ml},j}(y)| \geq 0 \Big\} \text{ and } J_- = \Big\{ j \in \{1, \ldots, p\} \,;\, \lambda - K_n |\hat{\beta}_{\mathrm{ml},j}(y)| < 0 \Big\},$$

such that $C(b) = C_+(b) - C_-(b)$ where

$$C_+(b) = \sum_{j \in J_+} b_j \big(\lambda - K_n |\hat{\beta}_{\mathrm{ml},j}(y)| \big) \quad \text{and} \quad C_-(b) = \sum_{j \in J_-} b_j \big| \lambda - K_n |\hat{\beta}_{\mathrm{ml},j}(y)| \big|.$$

The minimum of the non-negative term $C_+(b)$ is reached for $b_j = 0, \forall j \in J_+$. The negative term $C_-(b)$ corresponding to negative values of $\lambda - K_n|\hat{\beta}_{\text{ml},j}(y)|$ remains and the problem can be rewriten as

$$\hat{\beta}^{\text{s}}_{\text{map}}(y) = \underset{b\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{j\in J_-} b_j \left(\lambda - K_n|\hat{\beta}_{\text{ml},j}(y)|\right) \quad \text{s.t.} \quad \|b\|_2^2 = 1 \quad \text{and} \quad \begin{cases} b_j \geq 0, & j \in \{1,\dots,p\}, \\ b_j = 0, & j \in J_+. \end{cases}$$

One can recognize a problem of minimization of projection on the vector of negative terms $\left(\lambda - K_n|\hat{\beta}_{\text{ml},j}(y)|\right)_{j\in J_-}$ which is well-known to be solved for positive terms $(b_j)_{j\in J_-}$ defined by

$$\forall j \in J_-,\ b_j = \left(K_n|\hat{\beta}_{\text{ml},j}(y)| - \lambda\right)/\sqrt{\gamma} \quad \text{where} \quad \gamma = \sum_{j\in J_-} \left(K_n|\hat{\beta}_{\text{ml},j}(y)| - \lambda\right)^2.$$

One can notice that $\gamma = \|S_\lambda(K_n|\hat{\beta}_{\text{ml}}(y)|)\|_2^2$, and therefore

$$\hat{\beta}^{\text{s}}_{\text{map}}(y) = S_\lambda(K_n|\hat{\beta}_{\text{ml}}(y)|)/\|S_\lambda(K_n|\hat{\beta}_{\text{ml}}(y)|)\|_2.$$

The result is thus proved.

Proof (of Proposition 5). Let us recall the notation introduced in the proof of Proposition 4: $\sigma_n = 1/\sqrt{n\bar{F}(y_n)}$. Combining Corollary 2 and Proposition 2, it follows that $\hat{\beta}^{\text{s}}_{\text{map}}(y_n) = \tilde{\beta}(y_n)/\|\tilde{\beta}(y_n)\|_2$ with, for all $j \in \{1,\dots,p\}$:

$$\tilde{\beta}_j(y_n) = S_\lambda\left(K_n(\beta_j + \sigma_n\varepsilon_{j,n})\right),$$

where $\varepsilon_n \xrightarrow{\mathbb{P}} 0$. Two cases arise:

- If $\beta_j = 0$ then, clearly, $\tilde{\beta}_j(y_n) = 0$ with probability tending to one, since $K_n\sigma_n \xrightarrow{\mathbb{P}} c$ and $\varepsilon_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.
- If $\beta_j \neq 0$, then $K_n \xrightarrow{\mathbb{P}} \infty$ and $K_n\sigma_n \xrightarrow{\mathbb{P}} c$ entail $|K_n(\beta_j + \sigma_n\varepsilon_{j,n})| \xrightarrow{\mathbb{P}} \infty$ as $n \to \infty$ and, therefore, with probability tending to one,

$$\tilde{\beta}_j(y_n) = \operatorname{sign}(\beta_j)\left(K_n(|\beta_j| \pm \sigma_n\varepsilon_{j,n}) - \lambda\right) = \beta_j K_n\left(1 - \frac{\lambda}{|\beta_j|K_n}(1 + o_{\mathbb{P}}(1))\right). \tag{11}$$

As a consequence, one has, with probability tending to one,

$$\begin{aligned}\|\tilde{\beta}(y_n)\|_2^2 &= K_n^2 \sum_{\beta_j\neq 0} \beta_j^2 \left(1 - \frac{\lambda}{|\beta_j|K_n}(1 + o_{\mathbb{P}}(1))\right)^2 \\ &= K_n^2\left\{1 + \sum_{\beta_j\neq 0} \beta_j^2 \left(\frac{\lambda^2}{\beta_j^2 K_n^2}(1 + o_{\mathbb{P}}(1)) - \frac{2\lambda}{|\beta_j|K_n}(1 + o_{\mathbb{P}}(1))\right)\right\},\end{aligned}$$

since $\|\beta\|_2 = 1$. It follows that

$$\|\tilde{\beta}(y_n)\|_2^2 = K_n^2\left(1 - \frac{2\lambda\|\beta\|_1}{K_n}(1 + o_{\mathbb{P}}(1))\right),$$

with probability tending to one, leading to

$$\frac{1}{\|\tilde{\beta}(y_n)\|_2} = \frac{1}{K_n}\left(1 + \frac{\lambda\|\beta\|_1}{K_n}(1 + o_{\mathbb{P}}(1))\right).$$

Combining with (11), one has, for all $j \in \{1, \ldots, p\}$ such that $\beta_j \neq 0$,

$$\frac{\tilde{\beta}_j(y_n)}{\|\tilde{\beta}(y_n)\|_2} = \beta_j\left(1 + \frac{\lambda}{K_n}\left(\|\beta\|_1 - \frac{1}{|\beta_j|}\right)(1 + o_{\mathbb{P}}(1))\right),$$

or equivalently,

$$\sigma_n^{-1}\left(\frac{\tilde{\beta}_j(y_n)}{\|\tilde{\beta}(y_n)\|_2} - \beta_j\right) = \frac{\lambda}{K_n\sigma_n}\left(\|\beta\|_1 - \frac{1}{|\beta_j|}\right)\beta_j\,(1 + o_{\mathbb{P}}(1)),$$

and $K_n\sigma_n \xrightarrow{\mathbb{P}} c$ proves the result.

Acknowledgements

This work is partially supported by the French National Research Agency (ANR) in the framework of the Investissements d'Avenir Program (ANR-15-IDEX-02). S. Girard acknowledges the support of the Chair "Stress Test, Risk Management and Financial Steering", led by the French Ecole Polytechnique and its Foundation and sponsored by BNP Paribas.

References

[APSZ21] A. Aghbalou, F. Portier, A. Sabourin, and C. Zhou. Tail inverse regression for dimension reduction with extreme response. arxiv. preprint, 2021.

[AS72] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1972.

[BEG23] M. Bousebata, G. Enjolras, and S. Girard. Extreme partial least-squares. *Journal of Multivariate Analysis*, 194(10510):1, 2023.

[BGST04] J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of Extremes: Theory and Applications*. Wiley, 2004.

[BGT89] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*. university press, Cambridge, 1989.

[CFG17] A. Chiancone, F. Forbes, and S. Girard. Student sliced inverse regression. *Computational Statistics & Data Analysis*, 113:441–456, 2017.

[CGS14] R. Coudret, S. Girard, and J. Saracco. A new sliced inverse regression method for multivariate response. *Computational Statistics & Data Analysis*, 77:285–299, 2014.

[CHS13] R. D. Cook, I. S. Helland, and Z. Su. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B*, 75(5):851–877, 2013.

[CK10] H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B*, 72(1):3–25, 2010.

[CLL21] X. Cai, G. Lin, and J. Li. Bayesian inverse regression for supervised dimension reduction with small datasets. *Journal of Statistical Computation and Simulation*, 91:1–16, 2021.

[DGG13] A. Daouia, L. Gardes, and S. Girard. On kernel smoothing for extremal quantile regression. *Bernoulli*, 19:2557–2589, 2013.

[dHF07] L. de Haan and A. Ferreira. *Extreme value theory: An introduction*. Springer, Science and Business Media, 2007.

[DSUC23] A. Daouia, G. Stupfler, and A. Usseglio-Carleve. *Inference for extremal regression with dependent heavy-tailed data*. The Annals of Statistics, to appear, 2023.

[Gar18] L. Gardes. Tail dimension reduction for extreme quantile estimation. *Extremes*, 21(1):57–95, 2018.

[Gee11] G. Geenens. Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys*, 5:30–43, 2011.

[GLS22] S. Girard, H. Lorenzo, and J. Saracco. Advanced topics in sliced inverse regression. *Journal of Multivariate Analysis*, 188(10485):2, 2022.

[GSUC21] S. Girard, G. Stupfler, and A. Usseglio-Carleve. Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models. *The Annals of Statistics*, 49(6):3358–3382, 2021.

[Hil75] B. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[Jef46] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London: Series A*, 186(1007):453–461, 1946.

[KJ13] P. Krupskii and H. Joe. Factor copula models for multivariate data. *Journal of Multivariate Analysis*, 120:85–101, 2013.

[LCT07] L. Li, R. D. Cook, and C. L. Tsai. Partial inverse regression. *Biometrika*, 94(3):615–625, 2007.

[LGA23] H. Lorenzo, S. Girard, and J. Arbel. *SEPaLS: Shrinkage for Extreme Partial Least-Squares in R*. R package, 2023.

[Li91] K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[Mar75] K. V. Mardia. Distribution theory for the von mises-fisher distribution and its application. In Kotz G. P. and Ord and S., editors, *Patil*. A Modern Course on Statistical Distributions in Scientific Work. NATO Advanced Study Institutes Series, vol 17. Springer, Dordrecht, 1975.

[MJ09] K. V. Mardia and P. E. Jupp. *Directional statistics*. John Wiley & Sons, New-York, 2009.

[MLM10] K. Mao, F. Liang, and S. Mukherjee. Supervised dimension reduction using Bayesian mixture modeling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 501–508, 9, 2010. PMLR.

[MN92] H. Martens and T. Naes. *Multivariate calibration*. John Wiley & Sons, New-York, 1992.

[NAGP05] G. Nunez-Antonio and E. Gutiérrez-Pena. A Bayesian analysis of directional data using the von mises-fisher distribution. *Communications in Statistics-Simulation and Computation*, 34(4):989–999, 2005.

[Nel07] R. B. Nelsen. *An introduction to copulas*. Springer, Science & Business Media, 2007.

[NT00] P. Naik and C. L. Tsai. Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B*, 62(4):763–771, 2000.

[Por16] F. Portier. An empirical process view of inverse regression. *Scandinavian Journal of Statistics*, 43(3):827–844, 2016.

[RBL11] B. J. Reich, H. D. Bondell, and L. Li. Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics*, 67(3):886–895, 2011.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[TML14] J. Taghia, Z. Ma, and A. Leijon. Bayesian estimation of the von-mises fisher mixture model with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1701–1715, 2014.

[VEOM19] S. Van Erp, D. L. Oberski, and J. Mulder. Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50, 2019.

[VvGBL13] D. Vidaurre, M. A. J. van Gerven, C. Bielza, and T. Larra naga, P. & Heskes. Bayesian sparse partial least squares. *Neural Computation*, 25(12):3318–3339, 2013.

[WLX22] H. J. Wang, D. Li, and W. Xu. Extreme quantile estimation based on the tail single-index model. *Statistica Sinica*, 32:1–22, 2022.

[Wol75] H. Wold. Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12:117–142, 1975.

[WW56] G. S. Watson and E. J. Williams. On the construction of significance tests on the circle and the sphere. *Biometrika*, 43(3):344–352, 1956.