Shrinkage for Extreme Partial Least Squares

Julyan Arbel, Stéphane Girard, Hadrien Lorenzo

Shrinkage for Extreme Partial Least-Squares

Julyan Arbel[(1)], Stéphane Girard[(1,⋆)] & Hadrien Lorenzo[(2)]

[(1)] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.
[(2)] Aix Marseille Univ, CNRS, I2M, Marseille, France.
⋆ Corresponding author, stephane.girard@inria.fr

Abstract

This work focuses on dimension-reduction techniques for modelling 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

Partial Least Squares (PLS). 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 (Geenens, 2011) 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 (Wold, 1975) 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 (Martens and Næs, 1992), it has also received considerable attention in the statistical literature, see for instance Naik and Tsai (2000). Sliced Inverse Regression (SIR, Li, 1991) 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 Cook et al. (2013), Li et al. (2007) and Chiancone et al. (2017), Coudret et al. (2014), Portier (2016) among others or Girard et al. (2022) 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 Reich et al. (2011), 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 Mao et al. (2010): X is modelled as a function of Y thanks to an inverse mixture model, the estimation also requiring an MCMC method. A similar approach is proposed in Cai et al. (2021) using a Bayesian inverse regression through Gaussian processes and MCMC procedures.

Extreme Partial Least Squares (EPLS). The curse of dimensionality is exacerbated when modelling conditional extremes since tail events are rare by nature. Nonparametric estimators of extreme conditional features (Daouia et al., 2013, 2023, Girard et al., 2021) 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 Aghbalou et al. (2024), Gardes (2018) who propose extreme analogues of the central dimension reduction subspace. In Xu et al. (2022), 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 Partial

PLS	$\hat{\beta}$	maximizes covariance between $\langle \beta, X \rangle$ and Y
EPLS	$\hat{\beta}_{\text{ml}}(y)$	a PLS estimator for values of Y larger than y
SEPaLS	$\hat{\beta}^{\text{c}}_{\text{map}}(y)$	an EPLS estimator with conjugate prior
	$\hat{\beta}^{\text{s}}_{\text{map}}(y)$	an EPLS estimator with sparse prior

Figure 1: Different Partial Least Squares approaches discussed here with their adaptations to the extreme and shrinkage frameworks.

Least Squares (EPLS, Bousebata et al., 2023) is a dimension reduction method relying on PLS principles for estimating the linear combinations of X that best explain the extreme values of Y. See also Girard and Pakzad (2024) for an adaptation of EPLS to functional covariates.

Shrinkage EPLS, contributions, and outline. In this work, we develop two shrinkage versions of the EPLS method for high-dimensional settings under the common acronym SEPaLS. The starting point consists of recognizing the EPLS estimator as a maximum likelihood estimator associated with a von Mises–Fisher likelihood (Section 2). The latter distribution, which naturally arises for modelling directional data distributed on the unit sphere (Mardia and Jupp, 1999), 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. Proposition 4 and Proposition 6 show 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. See Figure 1 for a summary of the different PLS adaptations. Convergence results are also established when the sample size tends to infinity, in Proposition 2, Proposition 5, and Proposition 7. 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 estimators are available in the R package `SEPaLS`[1] (Lorenzo et al., 2023), while the R code replicating the figures can be found online[2]. A discussion is provided in Section 6 and proofs are postponed to Appendix A.

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

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

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 Bousebata et al. (2023):

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

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, see Bernard-Michel et al. (2009), Cook (2007) for similar inverse models in the SIR framework. Under model ($\mathbf{A_0}$), if the distribution tail of ε is negligible compared to the one of $g(Y)$, then $X \simeq g(Y)\beta$ for large values of Y, leading to the approximate single-index forward model $Y \simeq g^{-1}(\langle \beta, X \rangle)$. Finally, let us stress that no independence assumption is made on (X, Y, ε). Let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be an n sample with same distribution as (X, Y).

Definition 1 (EPLS estimator of the unit direction β, Bousebata et al., 2023)**.** *The EPLS estimator $\hat{\beta}(y_n)$ of the unit direction β is obtained by maximizing with respect to $\beta \in S^{p-1}$ the empirical covariance between $\langle \beta, X \rangle$ and Y conditionally on values of Y larger than y_n:*

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

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.

Note that the EPLS estimator focuses on large values of Y. It could be easily adapted to the lower distribution tail by considering $-Y$ and thus replacing the indicator functions $\mathbf{1}\{Y_i \geq y_n\}$ by $\mathbf{1}\{Y_i \leq -y_n\}$. The asymptotic properties of the EPLS estimator can be established under some assumptions on the upper 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 functions. Recall that φ is regularly-varying with index $\theta \in \mathbb{R}$ 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 Bingham et al. (1987) for a detailed account of regular variations.

($\mathbf{A_1}$) The density function f of Y is regularly-varying of index $-1/\gamma_Y - 1$, with $0 < \gamma_Y < 1$.

($\mathbf{A_2}$) The link function g is regularly-varying of index $c > 0$ and $2\gamma_Y(c+1) < 1$.

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

Assumption ($\mathbf{A_1}$) implies that the survival function $\bar{F}$ is regularly-varying with index $-1/\gamma_Y$, 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_Y, see Bingham et al. (1987, Theorem 1.5.8) and Haan and Ferreira (2007, Theorem 1.2.1). This domain of attraction

consists of heavy-tailed distributions, such as Pareto, Burr and Student distributions, see Beirlant et al. (2004) for further examples. The larger γ_Y is, the heavier the tail. The restriction to $\gamma_Y < 1$ ensures that the first-order moment $\mathbb{E}(Y\mathbf{1}\{Y \geq y\})$ exists for all $y \geq 0$. Assumption $(\mathbf{A_2})$ ensures that the link function g ultimately behaves like a power function. Combined with $(\mathbf{A_1})$, it implies that $g(Y)$ is heavy-tailed with tail-index $\gamma_{g(Y)} := c\gamma_Y$. 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. More specifically, $\mathbb{E}(\|\varepsilon\|^q) < \infty$ implies that the tail-index associated with $\|\varepsilon\|$ is such that $\gamma_{\|\varepsilon\|} < 1/q$. Condition $(\mathbf{A_3})$ thus imposes that $\gamma_{g(Y)} > \gamma_{\|\varepsilon\|}$, meaning that $g(Y)$ has an heavier right tail than $\|\varepsilon\|$. Under model $(\mathbf{A_0})$, the tail behaviors of $|\beta^t X|$ and $\|X\|$ are thus driven by $g(Y)$, *i.e.*, $\gamma_{\|X\|} = \gamma_{g(Y)}$, which is the desired property. Finally, condition $2\gamma_Y(c+1) < 1$ implies the existence of $\mathrm{var}(XY\mathbf{1}\{Y \geq y\})$ for all $y \geq 0$.

2.3 Two von Mises–Fisher distributions

The von Mises–Fisher distribution $\mathrm{vMF_S}(\mu, \kappa)$ on the unit sphere S^{p-1}, $p \geq 2$, is defined by its probability density function (Watson and Williams, 1956):

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

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 Abramowitz and Stegun (1965, 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 analogue of the multivariate Gaussian distribution (Mardia, 1975). Let us also recall that, for all $\mu \in S^{p-1}$, $\mathrm{vMF_S}(\mu, 0)$ 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 $\mathrm{vMF_S}(\mu, \kappa)$ distribution for all $\kappa > 0$. We propose the following adaptation of this distribution on balls:

Definition 2 (von Mises–Fisher distribution on the ball)**.** *The von Mises–Fisher distribution* $\mathrm{vMF_B}(\mu, \mathrm{r}, \kappa)$ *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 Appendix A for a proof that $f_{\mathrm{vMF_B}}(\cdot|\mu, r, \kappa)$ integrates to one. The next paragraph shows that the $\mathrm{vMF_B}$ 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 estimator. It is thus denoted by $\hat{\beta}_{\mathrm{ml}}(y_n)$ in the sequel.

Proposition 1 (EPLS estimator as a maximum likelihood estimator)**.** *The EPLS estimator of Definition 1 is the maximum likelihood estimator of* β*, denoted by* $\hat{\beta}_{\mathrm{ml}}(y_n)$*, in the following model:*

(i) $X_1, \ldots, X_n$ *are independent and, for all* $i \in \{1, \ldots, n\}$*,* X_i *given* $(Y_{1:n}, \varepsilon_i)$ *is* $\mathrm{vMF_B}(\beta, \mathrm{r_i}, \kappa_\mathrm{i})$ *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*.*

The next proposition provides a consistency result on the EPLS maximum likelihood estimator (Definition 1 and Proposition 1).

Proposition 2 (EPLS consistency)**.** *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 Bousebata et al. (2023) for a discussion of the assumptions on the (y_n) sequence. Let us simply note that the associated rate of convergence is faster than $\sqrt{n\bar{F}(y_n)}$. Even though the exact rate is not available there, this result will reveal sufficient for deriving the exact rates of convergence associated with the shrunk estimators, see Proposition 5 and Proposition 7 hereafter.

3 Shrinkage for Extreme Partial Least Squares

The result of item (i) in Proposition 1 opens the door to the construction of shrinkage estimators for β based on the Bayesian paradigm, referred to as Shrinkage for Extreme Partial Least Squares (SEPaLS) estimators. A prior distribution $\pi(\cdot)$ is introduced on the direction parameter β and the shrinkage effect of the maximum a posteriori (MAP) estimator is investigated. The posterior distribution is established in Subsection 3.1 and MAP estimators are derived for two particular cases of priors, a conjugate one based on the von Mises–Fisher distribution on the sphere in Subsection 3.2, and a sparse one based on the Laplace distribution in Subsection 3.3. In both cases, the implementation of the method requires selecting both the shrinkage parameter associated with the prior as well as the threshold y_n. A data-driven method is described in Section 5 on the real data application.

3.1 Posterior distribution

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

Proposition 3 (SEPaLS posterior distribution)**.** *Let $\theta_n > 0$ and $\pi(\cdot)$ a prior distribution on the direction parameter $\beta \in S^{p-1}$. 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 $\mathrm{vMF_S}$ prior

We first assume a $\mathrm{vMF_S}$ 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 2a for $\mu_0 = (1, 0, 0)^\top$ and $\kappa_0 \in \{0, 1, 10\}$. 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 $\mathrm{vMF_S}$ 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 (Nunez-Antonio and Gutiérrez-Pena, 2005, Taghia et al., 2014). The following proposition is easily derived.

Proposition 4 (MAP with conjugate prior)**.** *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 $\mathrm{vMF_S}$ 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 2a 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.

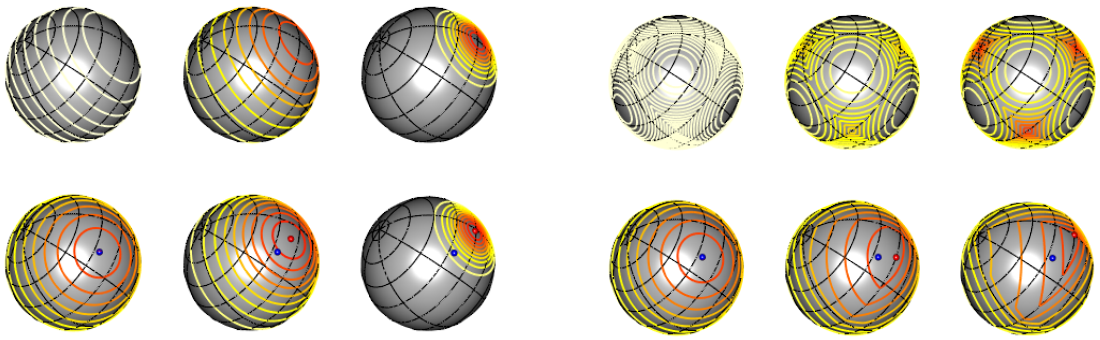
Proposition 5 (MAP consistency under conjugate prior)**.** *Under the assumptions of Proposition 2, let $c > 0$ and*

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

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 $P^{\perp}_{\beta}(\mu_0) := \mu_0 - \langle \mu_0, \beta \rangle \beta$ denotes the projection of μ_0 on the hyperplane orthogonal to β.



$\kappa_0 \approx 0$ $\kappa_0 = 1$ $\kappa_0 = 10$ $\lambda \approx 0$ $\lambda = 0.3$ $\lambda = 0.6$

(a) Conjugate $\mathrm{vMF}_{\mathrm{S}}(\mu_0, \kappa_0)$ prior. (b) Sparse Laplace(λ) prior.

Figure 2: Isocontour plots of (a) the von Mises–Fisher $\mathrm{vMF}_{\mathrm{S}}(\mu_0, \kappa_0)$ and (b) the Laplace(λ) prior densities (top) and of the resulting posterior density (bottom) in dimension $p = 3$. The estimators $\hat{\beta}_{\mathrm{ml}}$ and $\hat{\beta}_{\mathrm{map}}$ are depicted by blue and red points respectively.

It appears that $\hat{\beta}^{\mathrm{c}}_{\mathrm{map}}(y_n)$ converges to β at the $\sqrt{n\bar{F}(y_n)}$ rate which is the classical convergence rate of most of extreme-value estimators since $n\bar{F}(y_n)$ is the effective number of tail observations involved in the estimator. The MAP estimator can however reach a faster convergence rate when $P^{\perp}_{\beta}(\mu_0) = 0$ *i.e.* when $\mu_0 = \beta$, meaning that the prior distribution is centred on the true (unknown) direction.

3.3 Sparse Laplace 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:

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

as a prior for $\beta \in S^{p-1}$, where $\lambda \geq 0$ is a concentration parameter. We refer to Tibshirani (1996) for the introduction of the Laplace prior in the regression context and to Chun and Keleş (2010), Vidaurre et al. (2013) 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 2b for $\lambda \in \{0, 0.3, 0.6\}$. 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 posterior distribution does not correspond to a classical distribution on the unit sphere, the MAP can be computed in closed form:

Proposition 6 (MAP with sparse prior)**.** *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:*

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

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 Theorem 3 of Chun and Keleş (2010) for a similar result in a non-extreme framework. The zero coordinates in $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y_n)$ correspond to covariates in X that have no impact on the extreme values of Y. 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 2b 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.

Proposition 7 (MAP consistency under sparse prior)**.** *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 $j \in \{1, \dots, p\}$ 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 converges to β at the usual $\sqrt{n\bar{F}(y_n)}$ rate. The convergence rate is higher when the non-zero coordinates of β all coincide: $\beta_j = \mathrm{sign}(\beta_j)/\|\beta\|_1$ for all $j \in \{1, \dots, p\}$ such that $\beta_j \neq 0$.

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: $t > 0 \mapsto g(t) = t^c$, $c \in \{1, 1/2, 1/4\}$. The output variable Y is distributed from a Pareto distribution with survival function $\bar{F}(y) = (y/2)^{-1/\gamma_Y}$, $y \geq 2$ and with tail-index $\gamma_Y = 1/5$. 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 (Nelsen, 2007, 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 (Krupskii and Joe, 2013):

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

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, defined as $g(\bar{F}^{-1}(1/n))/\sigma$, 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 $\mathrm{vMF_S}$ 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. 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 is irrelevant to the inference.

4.2 Performance assessment

Let us define a similarity measure R between the theoretical vector β and its MAP estimator computed on $N = 1\,000$ replications as follows:

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

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

4.3 Results

Conjugate prior. The similarity measure $R(Y_{n-k+1,n})$ between $\hat{\beta}_{\text{map}}^{\text{c}}$ and β is represented as a function of $k \in \{1, \ldots, 100\}$ on Figure 3 for the choice of parameter $c = 1$. See Figure 6 and Figure 7 in Appendix B for the cases $c \in \{1/2, 1/4\}$. Each of these figures considers 32 configurations in dimension $p = 30$: $\kappa_0 \in \{0, 10^{-4}, 3.10^{-3}, 10^{-2}\}$, $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$, 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 shrinkage 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 $R \simeq 1$ 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 similarity measure $R(Y_{n-k+1,n})$ between $\hat{\beta}_{\text{map}}^{\text{s}}$ and β is represented as a function of $k \in \{1, \ldots, 100\}$ on Figures 8–10 in Appendix B for the cases $c \in \{1, 1/2, 1/4\}$. Each of these figures considers 32 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 $p \in \{30, 300\}$. Here, the shrinkage 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 $R \simeq 0.8$ (resp. $R \simeq 0.6$) in dimension $p = 30$ (resp. $p = 300$) with exponents $c \geq 1/2$. The case of small exponents ($c = 1/4$) appears to be more complicated, the maximum value of R 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. A similar study could be achieved on the small productions of carrots by focusing on the upper tail of $1/Y$.

Three visual checks are first carried out in Figure 4 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 (Hill, 1975)

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

of the tail-index γ_Y is drawn on the top right panel as a function of $k \in \{1, \ldots, 500\}$. The resulting graph is stable on the range $k \in \{160, \ldots, 280\}$ and points towards $\gamma_Y \simeq 0.72$. 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, $i \in \{1, \ldots, k\}$, exhibits a linear trend (bottom panel) which is further empirical evidence that the heavy-tail assumption is appropriate, see Beirlant et al. (2004, 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 $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$:

$$\rho(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle, Y | Y \geq y) = \frac{\mathrm{cov}(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle, Y | Y \geq y)}{\sigma(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle | Y \geq y)\sigma(Y | Y \geq y)}, \text{ (see Figure 5a)}, \tag{8}$$

$$\rho(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle, X^{(j)} | Y \geq y) = \frac{\mathrm{cov}(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle, X^{(j)} | Y \geq y)}{\sigma(\langle X, \hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y)\rangle | Y \geq y)\sigma(X^{(j)} | Y \geq y)}, \text{ (see Figure 5b)}, \tag{9}$$

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

with $j \in \{1, \ldots, p\}$. 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 5a 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 4. 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 $X^{(j)}$, $j \in \{1, \ldots, p\}$. It is computed at the threshold $y = Y_{n-k+1,n}$ and plotted on Figure 5b as a function of the number of exceedances k for $\lambda = 353$. All correlation curves feature nice stability with respect to k, especially in the neighbourhood 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 Figure 5c 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 4 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`).

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 ($p = 300$ on simulated data and $p \simeq 260$ on real data) with moderate sample sizes ($n = 500$ on simulated data and $n \simeq 600$ on real data). In this study, we limited ourselves to the estimation of a single direction. However, the SEPaLS method could be adapted to estimate multiple directions using the iterative procedure described in Bousebata et al. (2023, Section 4). We also focused on prior distributions that yield

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

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.

closed-form shrinkage estimators. It would be of interest to investigate the use of other priors, such as uninformative priors like Jeffreys' prior (Harold, 1946) or other shrinkage priors (van Erp et al., 2019). In such cases, the estimators would not be in closed form, and their computation would rely on MCMC procedures. This would make the process more computationally intensive than the estimation procedure considered here, especially because one must constrain the MCMC algorithm to sample on the sphere in dimension p.

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). J. Arbel acknowledges the support of ANR-21-JSTM-0001 grant. 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.

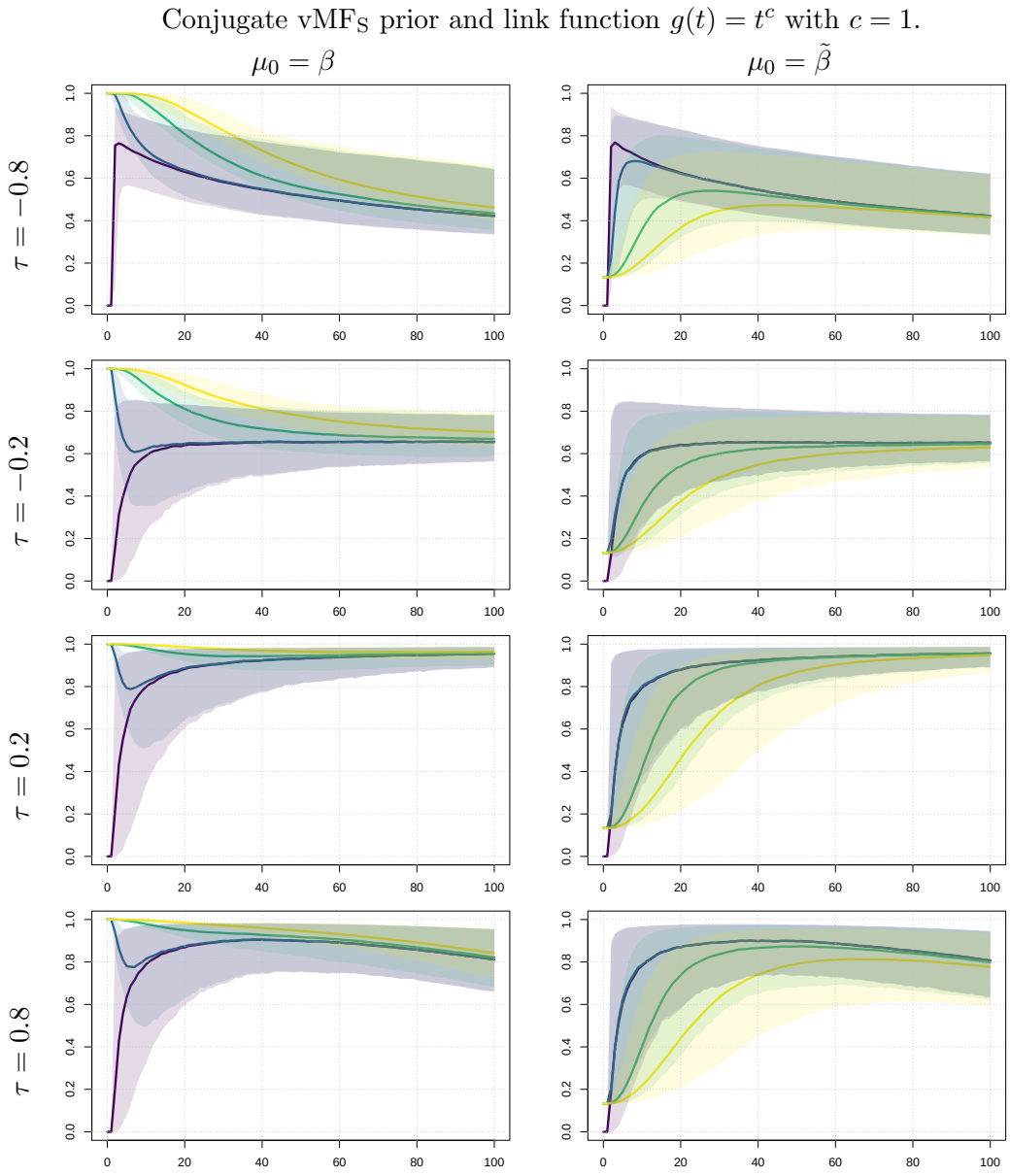


Figure 3: Finite sample behavior of the SEPaLS estimator computed with the conjugate prior on simulated data in dimension $p = 30$ from a Pareto distribution ($\gamma_Y = 1/5,\ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1$. Vertically: $R(Y_{n-k+1,n})$ 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. Coloured areas correspond to 90% confidence intervals.

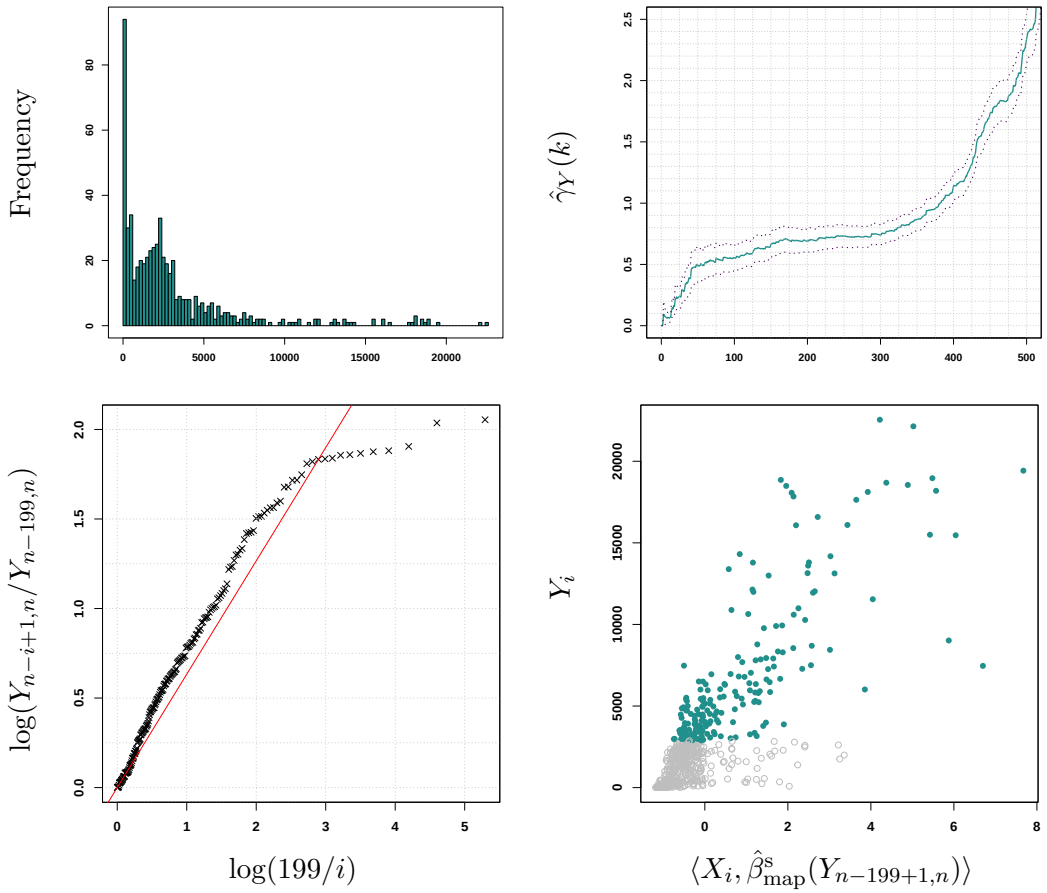
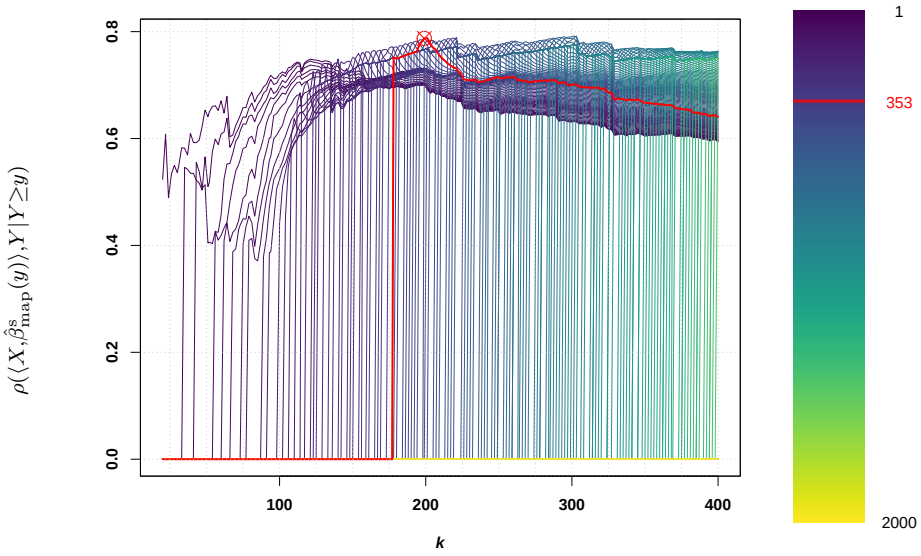
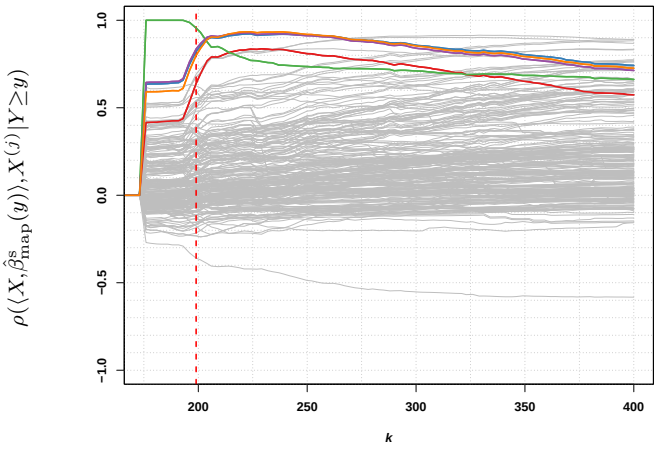


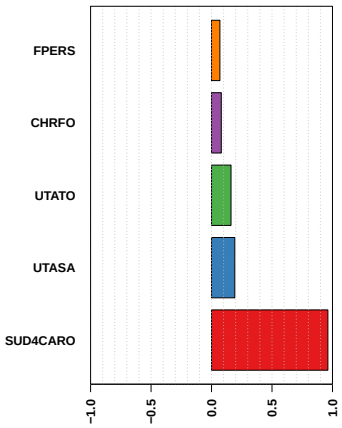
Figure 4: Real data example. Top left: Histogram of $\{Y_1, \ldots, Y_n\}$. Top right: Hill plot $k \in \{1, \ldots, 500\} \mapsto \hat{\gamma}_Y(k)$ and associated confidence intervals (dotted lines). Bottom left: Quantile-quantile plot (horizontally: $\log(k/i)$, vertically: $\log(Y_{n-i+1,n}/Y_{n-k,n})$, 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 $k = 199$ depicted in green. Points below the threshold ($Y_i \leq Y_{n-k+1,n}$) are depicted in gray.



(a) Correlation $\rho(\langle X, \hat{\beta}^{s}_{map}(y)\rangle, Y|Y \geq y)$.



(b) Correlation $\rho(\langle X, \hat{\beta}^{s}_{map}(y)\rangle, X^{(j)}|Y \geq y)$.



(c) Non-zero coordinates of $\hat{\beta}^{s}_{map}$.

Figure 5: Real data example. (a) 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. (b) 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 $j \in \{1, \ldots, 259\}$. (c) Non-zero coordinates of $\hat{\beta}^{s}_{map}(Y_{n-k+1,n})$ for the selected pair $(k, \lambda) = (199, 353)$. The colour code is the same for both left and right panels.

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Dover Publications.

Aghbalou, A., Portier, F., Sabourin, A., and Zhou, C. (2024). Tail inverse regression: Dimension reduction for prediction of extremes. *Bernoulli*, 30(1):503–533.

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, Chichester, England, UK.

Bernard-Michel, C., Gardes, L., and Girard, S. (2009). Gaussian Regularized Sliced Inverse Regression. *Statistics and Computing*, 19(1):85–98.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular Variation*. Encyclopedia of Mathematics and its Applications. Cambridge University Press.

Bousebata, M., Enjolras, G., and Girard, S. (2023). Extreme Partial Least-Squares. *Journal of Multivariate Analysis*, 194:105101.

Cai, X., Lin, G., and Li, J. (2021). Bayesian inverse regression for supervised dimension reduction with small datasets. *Journal of Statistical Computation and Simulation*, 91(14):2817–2832.

Chiancone, A., Forbes, F., and Girard, S. (2017). Student Sliced Inverse Regression. *Computational Statistics & Data Analysis*, 113:441–456.

Chun, H. and Keleş, S. (2010). Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. *Journal of the Royal Statistical Society: Series B*, 72(1):3–25.

Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, 22(1):1–26.

Cook, R. D., Helland, I. S., and Su, Z. (2013). Envelopes and Partial Least Squares Regression. *Journal of the Royal Statistical Society: Series B*, 75(5):851–877.

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

Daouia, A., Gardes, L., and Girard, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli*, 19(5B):2557–2589.

Daouia, A., Stupfler, G., and Usseglio-Carleve, A. (2023). Inference for extremal regression with dependent heavy-tailed data. *The Annals of Statistics*, 51(5):2040–2066.

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

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

Girard, S., Lorenzo, H., and Saracco, J. (2022). Advanced topics in Sliced Inverse Regression. *Journal of Multivariate Analysis*, 188:104852.

Girard, S. and Pakzad, C. (2024). Functional Extreme Partial Least-Squares. hal-04488561.

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

Haan, L. and Ferreira, A. (2007). *Extreme Value Theory*. Springer, New York, NY, USA.

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

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, 3(5):1163–1174.

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

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414):316–327.

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

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

Mao, K., Liang, F., and Mukherjee, S. (2010). Supervised Dimension Reduction Using Bayesian Mixture Modeling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 501–508.

Mardia, K. V. (1975). Distribution Theory for the Von Mises-Fisher Distribution and Its Application. In *A Modern Course on Statistical Distributions in Scientific Work*, pages 113–130. Springer Netherlands, Dordrecht.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. John Wiley & Sons, Chichester, England, UK.

Martens, H. and Næs, T. (1992). *Multivariate Calibration*. Wiley, Hoboken, NJ, USA.

Naik, P. and Tsai, C.-L. (2000). Partial Least Squares Estimator for Single-Index Models. *Journal of the Royal Statistical Society: Series B*, 62(4):763–771.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer, New York, NY, USA.

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

Portier, F. (2016). An Empirical Process View of Inverse Regression. *Scandinavian Journal of Statistics*, 43(3):827–844.

Reich, B. J., Bondell, H. D., and Li, L. (2011). Sufficient Dimension Reduction via Bayesian Mixture Modeling. *Biometrics*, 67(3):886–895.

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

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50.

Vidaurre, D., van Gerven, M. A. J., Bielza, C., Larrañaga, P., and Heskes, T. (2013). Bayesian Sparse Partial Least Squares. *Neural Computation*, 25(12):3318–3339.

Watson, G. S. and Williams, E. J. (1956). On the Construction of Significance Tests on the Circle and the Sphere. *Biometrika*, 43(3):344–352.

Wold, H. (1975). Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. *Journal of Applied Probability*, 12:117–142.

Xu, W., Wang, H. J., and Li, D. (2022). Extreme Quantile Estimation Based on the Tail Single-index Model. *Statistica Sinica*, 32(2):893–914.

A 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 centred 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 $\mathrm{vMF_B}$ 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

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

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 Bousebata et al. (2023, 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 5. Let $\sigma_n = 1/\sqrt{n\bar{F}(y_n)}$. Combining Proposition 4 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 $\sigma_n \to 0$ and $1/K_n \overset{\mathbb{P}}{\sim} \sigma_n/c \to 0$ as $n \to \infty$, a first order Taylor expansion yields:

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

since $\|\beta\|_2 = 1$, 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 Proposition 6. 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

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

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

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

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 $\lambda - K_n|\hat{\beta}_{\mathrm{ml},j}(y)|$ remains and the problem can be rewritten as

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

One can recognise a problem of minimization of projection on the vector of negative terms $\big(\lambda - K_n|\hat{\beta}_{\mathrm{ml},j}(y)|\big)_{j\in J_-}$ which is solved for positive terms $(b_j)_{j\in J_-}$ defined by

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

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

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

The result is thus proved.

□

Proof of Proposition 7. Let us recall the notation introduced in the proof of Proposition 5: $\sigma_n = 1/\sqrt{n\bar{F}(y_n)}$. Combining Proposition 6 and Proposition 2, it follows that $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}(y_n) = \tilde{\beta}(y_n)/\|\tilde{\beta}(y_n)\|_2$ 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,

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

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. □

B Appendix: Additional figures

We provide below additional figures corresponding to the illustration on simulated data presented in Section 4. They correspond to the use of the conjugate prior with parameter $c \in \{1/2, 1/4\}$ (while the case $c = 1$ can be found in the main text), and the sparse prior with parameter $c \in \{1, 1/2, 1/4\}$.

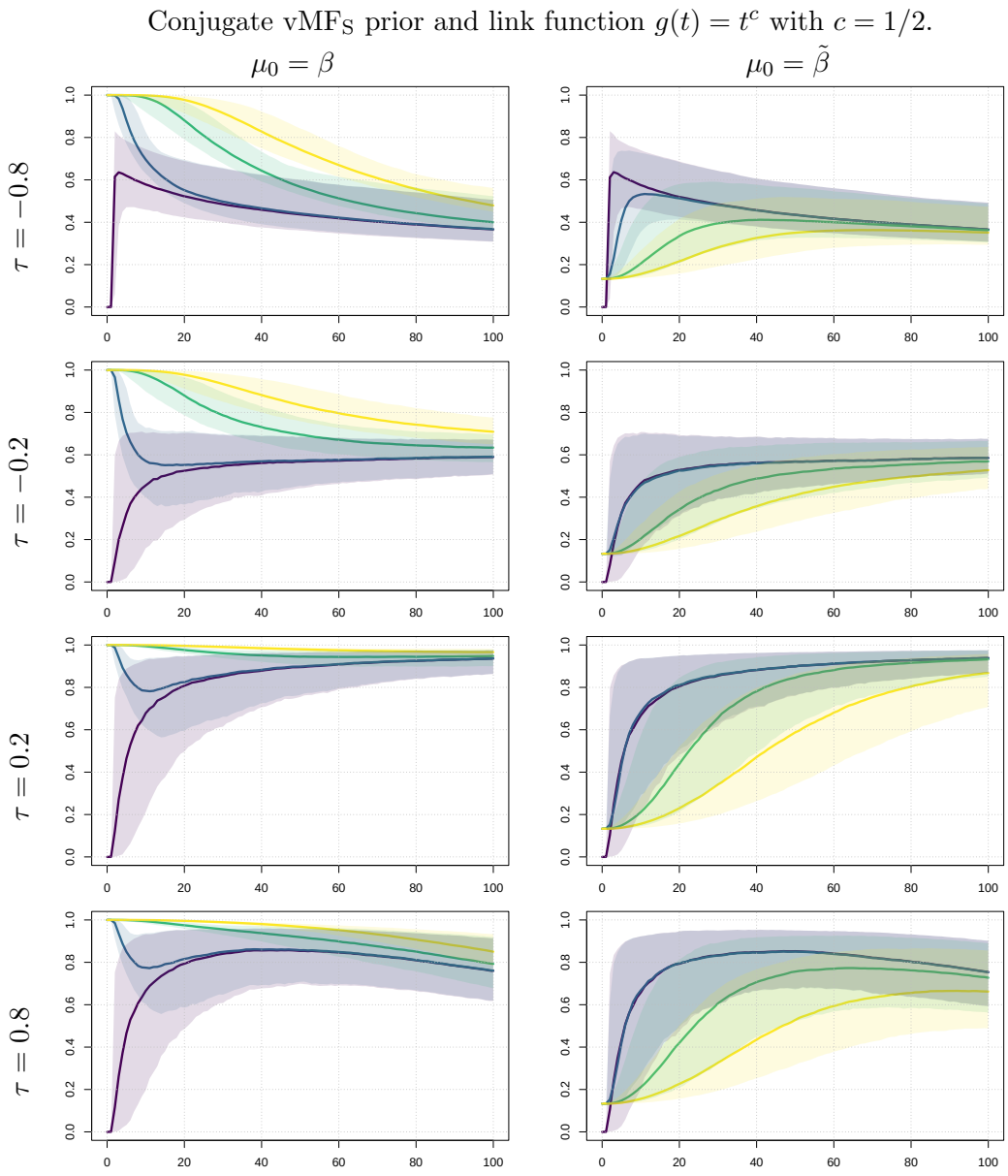


Figure 6: Finite sample behavior of the SEPaLS estimator computed with the conjugate prior on simulated data in dimension $p = 30$ from a Pareto distribution ($\gamma_Y = 1/5,\ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1/2$. Vertically: $R(Y_{n-k+1,n})$ 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. Coloured areas correspond to 90% confidence intervals.

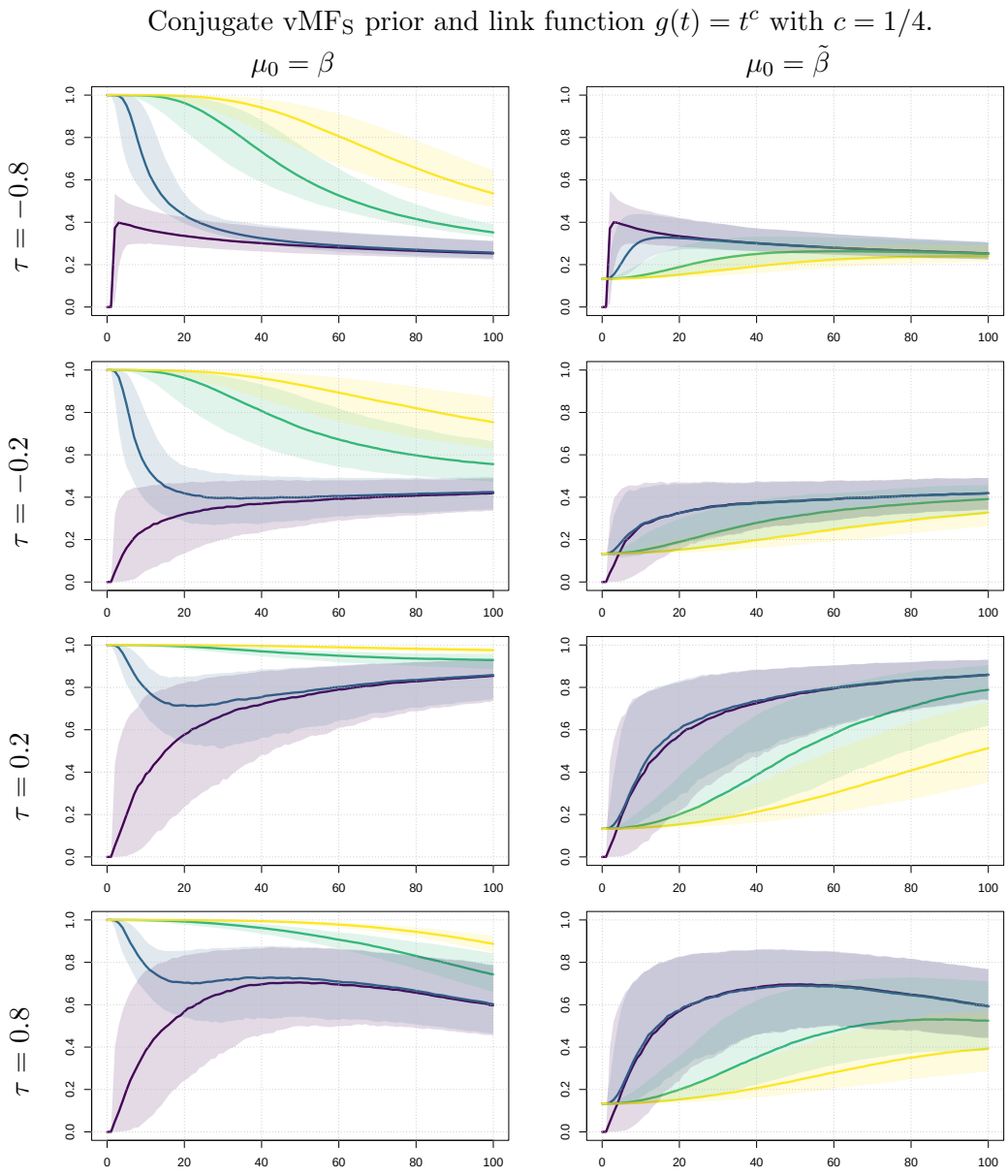


Figure 7: Finite sample behavior of the SEPaLS estimator computed with the conjugate prior on simulated data in dimension $p = 30$ from a Pareto distribution ($\gamma_Y = 1/5,\ a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1/4$. Vertically: $R(Y_{n-k+1,n})$ 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. Coloured areas correspond to 90% confidence intervals.

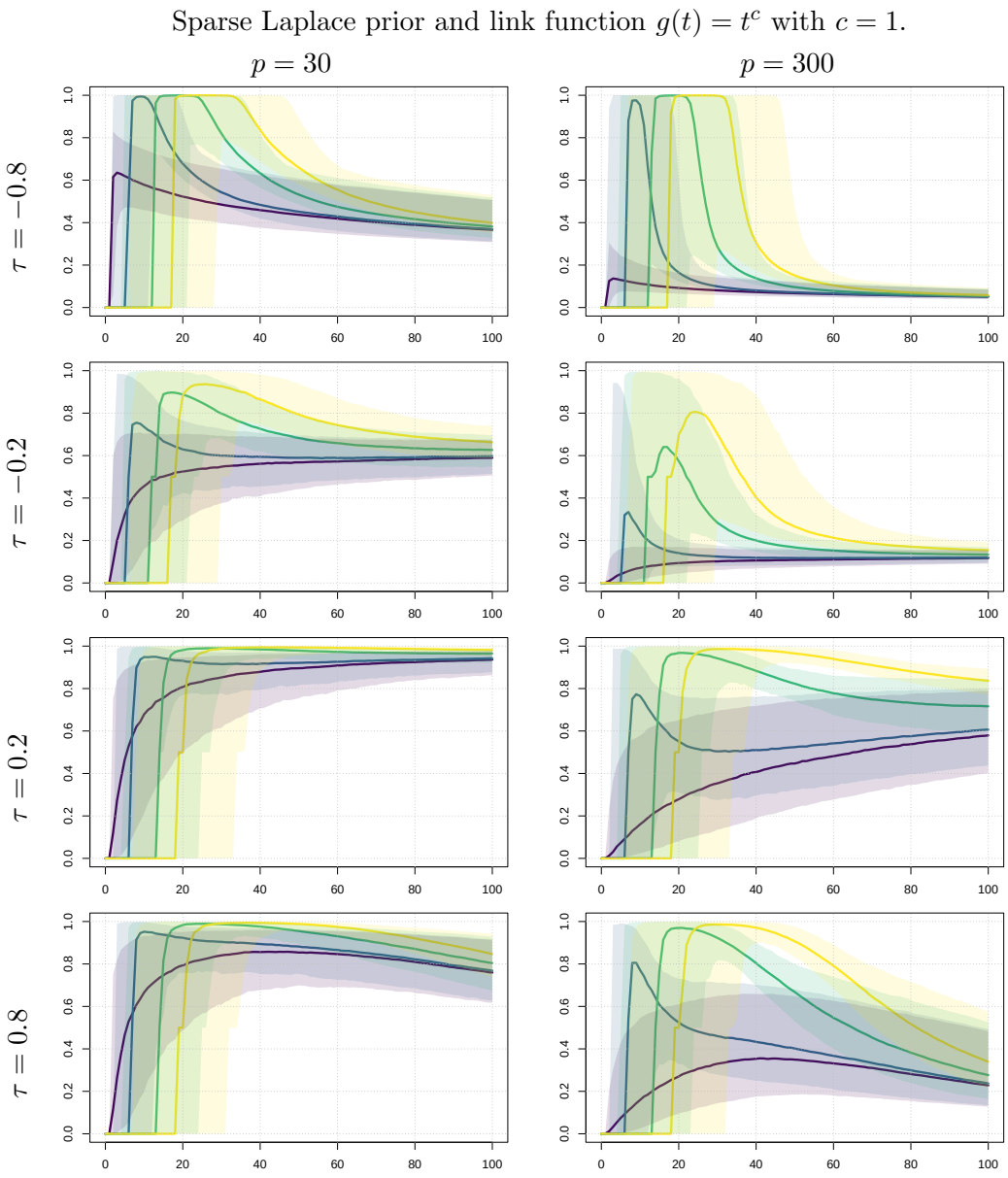


Figure 8: Finite sample behavior of the SEPaLS estimator computed with the sparse prior on simulated data in dimension $p = 30$ (left) and $p = 300$ (right) from a Pareto distribution ($\gamma_Y = 1/5$, $a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1$. Vertically: $R(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ and β for 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. Coloured areas correspond to 90% confidence intervals.

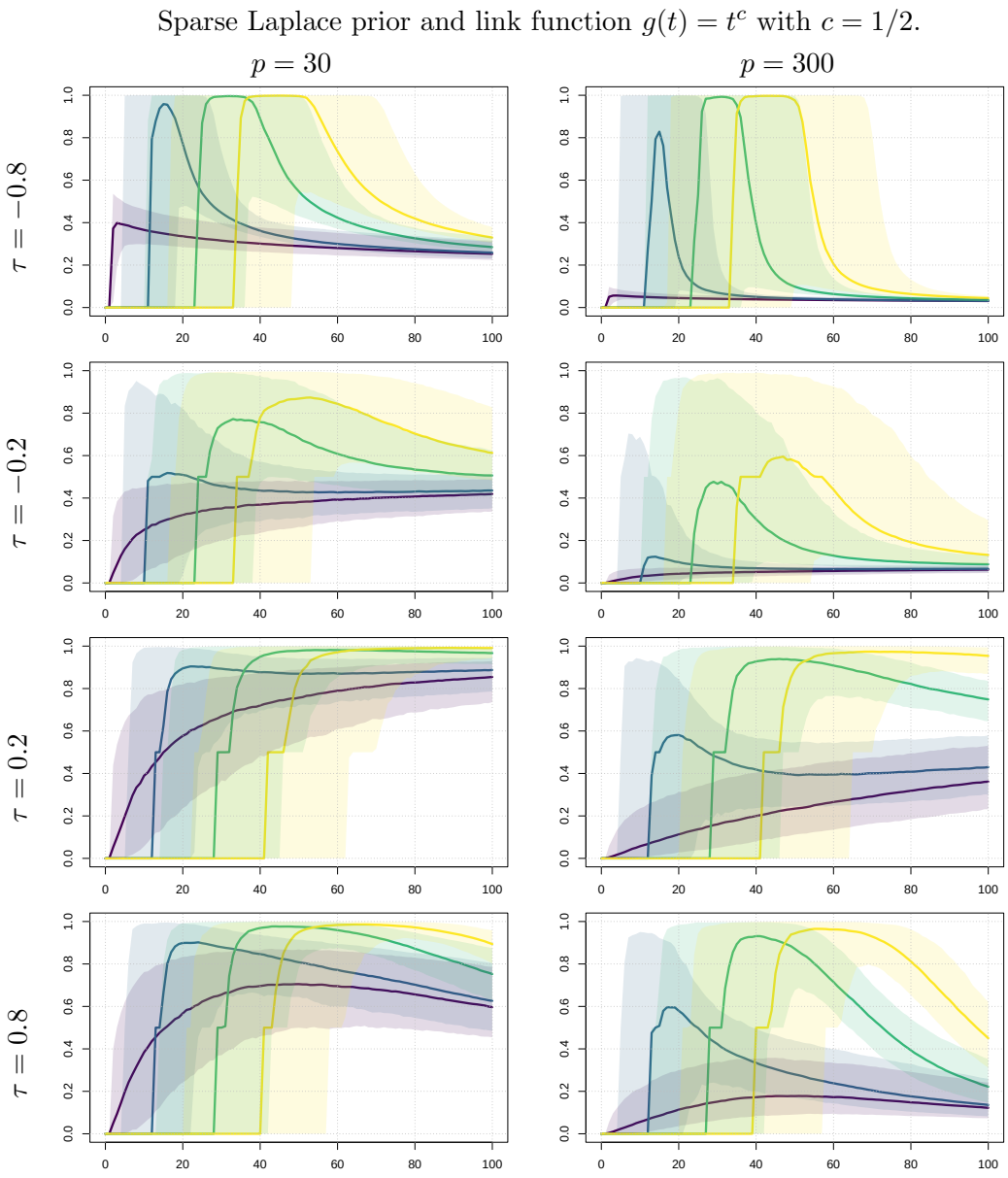


Figure 9: Finite sample behavior of the SEPaLS estimator computed with the sparse prior on simulated data in dimension $p = 30$ (left) and $p = 300$ (right) from a Pareto distribution ($\gamma_Y = 1/5$, $a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1/2$. Vertically: $R(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ and β for 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. Coloured areas correspond to 90% confidence intervals.

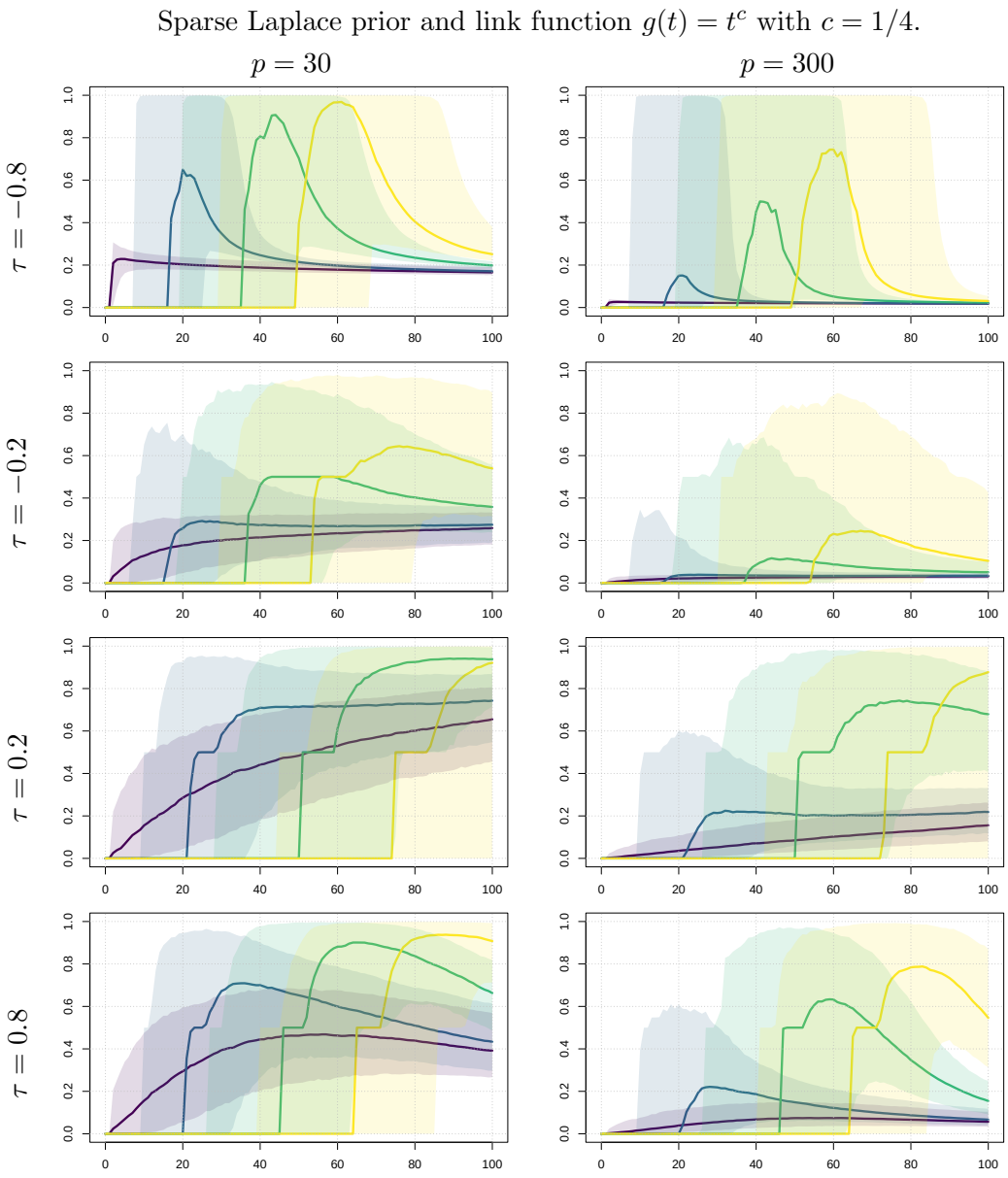


Figure 10: Finite sample behavior of the SEPaLS estimator computed with the sparse prior on simulated data in dimension $p = 30$ (left) and $p = 300$ (right) from a Pareto distribution ($\gamma_Y = 1/5$, $a = 2$) and a (rotated) Clayton copula with Kendall's tau $\tau \in \{-0.8, -0.2, 0.2, 0.8\}$ (from top to bottom). The power of the link function $g(t) = t^c$ is fixed to $c = 1/4$. Vertically: $R(Y_{n-k+1,n})$ between $\hat{\beta}^{\mathrm{s}}_{\mathrm{map}}$ and β for 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. Coloured areas correspond to 90% confidence intervals.