Tilburg University

Extreme Value Statistics in Semi-Supervised Models

Ahmed, Hanan; Einmahl, John; Zhou, Chen

Publication date:
2021

Document Version
Early version, also known as pre-print

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Ahmed, H., Einmahl, J., & Zhou, C. (2021). *Extreme Value Statistics in Semi-Supervised Models*. (CentER Discussion Paper; Vol. 2021-007). CentER, Center for Economic Research.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

No. 2021-007

**EXTREME VALUE STATISTICS IN
SEMI-SUPERVISED MODELS**

By

Hanan Ahmed, John H.J. Einmahl, Chen Zhou

2 March 2021

ISSN 0924-7815
ISSN 2213-9532

EXTREME VALUE STATISTICS IN SEMI-SUPERVISED MODELS

BY HANAN AHMED, JOHN H.J. EINMAHL* AND CHEN ZHOU

*Dept. of Econometrics and OR and CentER
Tilburg University
PO Box 90153, 5000 LE Tilburg
The Netherlands
h.e.g.ahmed@tilburguniversity.edu; j.h.j.einmahl@tilburguniversity.edu*

*Econometric Institute
Erasmus University
P.O. Box 1738, 3000 DR Rotterdam
The Netherlands
zhou@ese.eur.nl*

Abstract We consider extreme value analysis in a semi-supervised setting, where we observe, next to the n data on the target variable, $n + m$ data on one or more covariates. This is called the semi-supervised model with n labeled and m unlabeled data. By exploiting the tail dependence between the target variable and the covariates, we derive an estimator for the extreme value index of the target variable in this setting and establish its asymptotic behavior. Our estimator substantially improves the univariate estimator, based on only the n target variable data, in terms of asymptotic variance whereas the asymptotic bias remains unchanged. We present a simulation study in which the asymptotic results are confirmed and also an extreme quantile estimator is derived and its improved performance is shown. Finally the estimation method is applied to rainfall data in France.

1. Introduction. The semi-supervised model, initially introduced in machine learning, deals with unbalanced datasets, when the labeled data are harder (more expensive or more time consuming) to obtain than the unlabeled data. Consider a dataset with one variable of interest, sometimes referred to as the target variable or outcome variable, and one or more covariates. The difficulty for collecting labeled data stems from collecting the target variable, whereas unlabeled data containing only the covariates, i.e. with the target variable missing, can be easily collected. Semi-supervised learning focuses on uncovering the (non-linear) relation between the target variable and the covariates. Estimations and predictions based on such relations and using the additional unlabeled data often show substantially improved performance. For example, for classification analysis see Vapnik (2013) and Zhu and Goldberg (2009); for regression analysis see Wasserman and Lafferty (2008), Azriel et al. (2016) and Chakraborty and Cai (2018).

Semi-supervised inference aims at estimating parameters or quantities regarding the target variable under the semi-supervised model. Zhang et al. (2019) investigates the general semi-supervised framework and shows how to use the unlabeled data to improve the estimation of the mean of the target variable. In this paper we consider semi-supervised inference in extreme value statistics.

*John Einmahl holds the Arie Kapteyn Chair 2019–2022 and gratefully acknowledges the corresponding research support.

JEL Codes: C13, C14

MSC 2010 subject classifications: Primary 62G05, 62G20, 62G30, 62G32; secondary 60F05, 60F17, 60G70

Keywords and phrases: Asymptotic normality, extreme value index, semi-supervised inference, tail dependence, variance reduction

Extreme value statistics deals with estimation of parameters or quantities related to the tail of a distribution, only making semi-parametric assumptions on this tail. Consequently, most of extreme value methods start with a relatively large number of observations n , but select only $k \ll n$ extreme observations from the full sample for statistical inference. Two techniques are often used in selecting the extreme observations: the peaks-over-threshold (POT) approach which selects the highest k observations, and the block maxima (BM) approach which splits the full sample into k blocks and selects the maxima of each block. Since only k observations are used in estimation, typically consistent estimators have a speed of convergence of $1/\sqrt{k}$. In practice, to obtain accurate estimators for tail parameters/quantities, one needs a sample with a relatively large sample size n to guarantee a sufficient number of extreme observations. In contrast, the semi-supervised model is greatly suitable for statistics of extremes in case data on the target variable are hard to obtain.

Recently, Ahmed and Einmahl (2019) considers improving the estimation of the extreme value index when the distributions of the target variable and the covariates are both heavy-tailed. The extreme value index γ describes the tail heaviness of a probability distribution. If $\gamma > 0$ the distribution is heavy tailed and has an infinite right endpoint, if $\gamma = 0$ the distribution is light tailed and may have an infinite or finite endpoint, and if $\gamma < 0$ the endpoint is finite, see, e.g., Beirlant et al. (2004) or de Haan and Ferreira (2006) for a thorough treatment of extreme value theory and the corresponding statistical inference.

The main goal of this paper is to derive in this semi-supervised setting the improved pseudo-maximum likelihood estimator (MLE) for a general extreme value index $\gamma \in \mathbb{R}$ and to establish its asymptotic behavior. For ease of explanation of our estimator, let us assume that there is only one covariate. We estimate γ for the variable of interest initially (that is ignoring the covariate) using the pseudo-MLE $\hat{\gamma}$, see Smith (1987) and Drees et al. (2004). Then, we choose a number g and for the covariate we transform the labeled data empirically (using all the labeled and unlabeled data) such that they obtain an artificial extreme value index g . Using the transformed covariates of the labeled data, we estimate the *known* g by the pseudo-MLE $\hat{g}$, say, and use the difference $\hat{g} - g$ to adapt and substantially improve the initial estimator $\hat{\gamma}$ for the extreme value index γ of the variable of interest. For this adaptation the tail dependence between the target variable and the covariate is crucial. Precise estimation of γ is important for describing the tail heaviness, but it is even more important when estimating extreme quantiles or very small tail probabilities. We further demonstrate the improved performance of an extreme quantile estimator under the semi-supervised model by a simulation study.

Compared to Ahmed and Einmahl (2019), this study has at least three improvements. Firstly, we provide a general result in the context of the relevant tail quantile process (see Lemma 6.5). Based on the tail quantile process result, one may improve most estimators based on the POT approach in extreme value statistics under the semi-supervised model. Secondly, we impose no assumptions on the tail of the covariates. When analyzing the tail of the target variable, it is crucial to assume regularity in its tail such as the max-domain of attraction condition in extreme value analysis. However, requiring such conditions for the covariates can be restrictive in applications. Thirdly and most important, our main result is valid for a broader class of distributions for the target variable: we deal with a general extreme value index $\gamma \in \mathbb{R}$ whereas Ahmed and Einmahl (2019) only handles the case $\gamma > 0$. Extending the range of the extreme value index is particularly important for applications where the sign of γ is not known beforehand. For example, when analyzing extreme weather, various studies find that the extreme value index is around zero for different meteorologic variables: for hourly surge level on the English east coast (Coles and Tawn (1991)), for hourly maximum wind speed in Sheffield, UK (Coles and Walshaw (1994)), for wave height and still water level on the Dutch coast (de Haan and de Ronde (1998)) and for daily rainfall in North Holland, The Netherlands (Buishand et al. (2008)).

This paper is organized as follows. In Section 2, for clearness of the exposition, we first introduce our adapted estimator for the extreme value index in the semi-supervised model with one covariate and we establish its asymptotic normality. In Section 3 we consider the general multivariate semi-supervised setting and present and establish asymptotic normality of the adapted estimator. Section 4 is devoted to a simulation study for the setting with one covariate. The improved performance, in terms of variance, of the adapted estimator compared with the initial estimator is shown. An application to rainfall in France can be found in Section 5 and the detailed proofs are deferred to Section 6.

2. Main results: one covariate. Let F be a bivariate distribution function with marginals F_1 and F_2 . We assume that F_1 is in the max-domain of attraction of an extreme-value distribution G_γ , where γ is the extreme value index, our parameter of interest. Let the pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from F , and let $(Y_{n+1}, \dots, Y_{n+m})$ be a random sample from F_2 , independent from the n pairs. This is the semi-supervised model. Assume that the tail copula R of (X_1, Y_1) exists:

$$(1) \quad R(x, y) = \lim_{t \downarrow 0} \frac{1}{t} \mathbb{P}(1 - F_1(X_1) \leq tx, 1 - F_2(Y_1) \leq ty), (x, y) \in [0, \infty]^2 \setminus \{(\infty, \infty)\}.$$

Denote the order statistics of $X_i, i = 1, \dots, n$, with $X_{1:n} \leq \dots \leq X_{n:n}$, and similarly for the $Y_i, i = 1, \dots, n$. We estimate $\gamma > -\frac{1}{2}$ with the often used pseudo-MLE $\hat{\gamma}$ based on $X_{n-k:n}, \dots, X_{n:n}$, for $k \in \{1, \dots, n-1\}$; see Section 3.4 in de Haan and Ferreira (2006).

Define for $i = 1, \dots, n$,

$$(2) \quad \tilde{Y}_i = \begin{cases} \frac{\left(1 - \left(F_{n+m}(Y_i) - \frac{1}{2(n+m)}\right)\right)^{-g} - 1}{g}, & g \neq 0, \\ -\log \left(1 - \left(F_{n+m}(Y_i) - \frac{1}{2(n+m)}\right)\right), & g = 0, \end{cases}$$

where F_{n+m} is the empirical distribution function based on $Y_l, l = 1, \dots, n+m$, and $g > -\frac{1}{2}$ is a number we may choose that mimics an extreme value index. Let the order statistics of $\tilde{Y}_i, i = 1, \dots, n$, be denoted by $\tilde{Y}_{1:n} \leq \dots \leq \tilde{Y}_{n:n}$, and let $\hat{g}$ be the pseudo-MLE of g based on $\tilde{Y}_{n-k:n}, \dots, \tilde{Y}_{n:n}$, using the same k as before. Of course, since we choose and hence know g , there is no direct need to estimate it. We will show below, however, that the dependence of the difference $\hat{g} - g$ and $\hat{\gamma}$, helps to improve the estimator of γ in the semi-supervised setting.

For the asymptotic theory, we assume that $m = m(n)$ and

$$(3) \quad k \rightarrow \infty, \frac{k}{n} \rightarrow 0, \sqrt{\frac{n}{n+m}} \rightarrow \nu \in (0, 1), \quad \text{as } n \rightarrow \infty.$$

We begin with establishing the joint asymptotic normality of $\hat{\gamma}$ and $\hat{g}$, a crucial result for deriving and showing asymptotic normality of our semi-supervised estimator (SSE) of γ . For that purpose we need the usual second order condition on the marginal distribution F_1 . Let $U_1 = F_1^{-1}(1 - 1/\cdot)$ be the tail quantile corresponding to F_1 . We assume that there exist a positive scale function a , a positive or negative function A , with $\lim_{t \rightarrow \infty} A(t) = 0$, and $\rho \leq 0$, such that for $x > 0$,

$$(4) \quad \lim_{t \rightarrow \infty} \frac{\frac{U_1(tx) - U_1(t)}{a(t)} - \frac{x^{\gamma-1}}{\gamma}}{A(t)} = \Psi(x), \quad \gamma \in \mathbb{R},$$

where

$$\Psi(x) = \begin{cases} \frac{x^{\gamma+\rho}-1}{\gamma+\rho}, & \rho < 0, \\ \frac{1}{\gamma} x^\gamma \log x, & \gamma \neq \rho = 0, \\ \frac{1}{2} \log^2 x, & \gamma = \rho = 0, \end{cases}$$

see de Haan and Ferreira (2006), p. 46.

PROPOSITION 2.1. Assume $\gamma > -\frac{1}{2}$ and choose $g > -\frac{1}{2}$. Assume that F_2 is continuous, (1), (3) and (4) hold, and $\sqrt{k}A(\frac{n}{k}) \rightarrow \lambda \in \mathbb{R}$, as $n \rightarrow \infty$, then

$$\left(\sqrt{k}(\hat{\gamma} - \gamma), \sqrt{k}(\hat{g} - g)\right) \xrightarrow{d} N\left(\left[\frac{\lambda(1+\gamma)}{(1-\rho)(1+\gamma-\rho)}, 0\right], \Sigma\right)$$

where

$$\Sigma = \begin{bmatrix} (1+\gamma)^2 & (1-\nu^2)(1+\gamma)(1+g)R_g \\ (1-\nu^2)(1+\gamma)(1+g)R_g & (1-\nu^2)(1+g)^2 \end{bmatrix},$$

$$\text{with } R_g = R(1, 1) + \frac{g-\gamma}{\gamma+g+1} \left((2\gamma+1) \int_0^1 \frac{R(s, 1)}{s^{1-\gamma}} ds - (2g+1) \int_0^1 \frac{R(1, t)}{t^{1-g}} dt \right).$$

Based on Proposition 2.1, we derive the SSE of γ . For this derivation only, take $\lambda = 0$. Then the approximate bivariate normal distribution of $(\hat{\gamma}, \hat{g} - g)$, has mean $[\gamma, 0]$ and estimated covariance matrix

$$\frac{1}{k} \hat{\Sigma} = \frac{1}{k} \begin{bmatrix} (1+\hat{\gamma})^2 & (1 - \frac{n}{n+m})(1+\hat{\gamma})(1+g)\hat{R}_g \\ (1 - \frac{n}{n+m})(1+\hat{\gamma})(1+g)\hat{R}_g & (1 - \frac{n}{n+m})(1+g)^2 \end{bmatrix},$$

where $\hat{R}_g$ is the estimator of R_g , obtained by replacing γ with $\hat{\gamma}$ and the tail copula R with its natural estimator

$$(5) \quad \hat{R}(x, y) = \frac{1}{k} \sum_{i=1}^n 1_{[X_i \geq X_{n-[kx]+1:n}, Y_i \geq Y_{n-[ky]+1:n}]}, \quad x, y \geq 0,$$

see, e.g., Drees and Huang (1998). Maximizing the thus obtained approximate likelihood function of the single “data point” $(\hat{\gamma}, \hat{g} - g)$ with respect to the unknown γ we obtain as SSE for γ :

$$(6) \quad \hat{\gamma}_g = \hat{\gamma} - \frac{1+\hat{\gamma}}{1+g} \hat{R}_g(\hat{g} - g).$$

Now we present the main result of this section, the asymptotic normality of the SSE.

THEOREM 2.1. Under the conditions of Proposition 2.1, as $n \rightarrow \infty$,

$$(7) \quad \sqrt{k}(\hat{\gamma}_g - \gamma) \xrightarrow{d} N\left(\frac{\lambda(1+\gamma)}{(1-\rho)(1+\gamma-\rho)}, (1+\gamma)^2 [1 - (1-\nu^2)R_g^2]\right).$$

REMARK 2.1. Note that the asymptotic bias of the SSE $\hat{\gamma}_g$ is the same as that of the pseudo-MLE $\hat{\gamma}$ (in Proposition 2.1). Therefore, when comparing both estimators we can and will focus on the (relative) reduction of the asymptotic variance which is equal to $(1-\nu^2)R_g^2$. We will see in Section 4 that this variance reduction is substantial in “standard” semi-supervised settings.

3. Main results: multiple covariates. In this section we consider the more general situation with $d-1$ covariates where $d > 2$. Consider a d -variate distribution F , with marginals $F_1, \dots, F_d$. We assume again that (only) F_1 is in the max-domain of attraction of an extreme-value distribution G_γ . Let F_- be the distribution function of the last $d-1$ components of a random vector with distribution function F . Let $(X_1, Y_{1,2}, \dots, Y_{1,d}), \dots, (X_n, Y_{n,2}, \dots, Y_{n,d})$ be a random sample of size n from F and let $(Y_{n+1,2}, \dots, Y_{n+1,d}), \dots, (Y_{n+m,2}, \dots, Y_{n+m,d})$ be a random sample of size m from F_- , independent of the d -variate random sample of size n . This is the multivariate semi-supervised setting.

Then, for fixed $j = 2, \dots, d$, we use all data for the covariates $\{Y_{i,j}\}_{i=1}^{n+m}$ to obtain $\{\tilde{Y}_{i,j}\}_{i=1}^n$ as in (2), where we may choose a number $g > -\frac{1}{2}$, that mimics an extreme value index, as before. For $k \in \{1, \dots, n-1\}$, let, similarly as in the previous section, $\hat{\gamma}$ and $\hat{g}_j$, $j = 2, \dots, d$, be the pseudo-MLEs of γ and $(d-1)$ times of g , respectively. Assume the existence of the tail copula R_{ij} of the i^{th} and the j^{th} component,

$$(8) \quad R_{ij}(x, y) = \lim_{t \downarrow 0} \frac{1}{t} \mathbb{P}(1 - F_i(Y_{1,i}) \leq tx, 1 - F_j(Y_{1,j}) \leq ty),$$

where $(x, y) \in [0, \infty]^2 \setminus \{(\infty, \infty)\}$, $1 \leq i, j \leq d$. Here $Y_{1,1}$ is understood as X_1 . Again, we first consider the joint asymptotic normality of $\hat{\gamma}$, and $\hat{g}_j, j = 2, \dots, d$.

PROPOSITION 3.1. *Assume $\gamma > -\frac{1}{2}$ and choose $g > -\frac{1}{2}$. Assume that $F_j, j = 2, \dots, d$, is continuous, (3), (4), and (8) hold, and as $n \rightarrow \infty$, $\sqrt{k}A(\frac{n}{k}) \rightarrow \lambda \in \mathbb{R}$, then*

$$\left(\sqrt{k}(\hat{\gamma} - \gamma), \sqrt{k}(\hat{g}_2 - g), \dots, \sqrt{k}(\hat{g}_d - g) \right) \xrightarrow{d} N \left(\left[\frac{\lambda(\gamma + 1)}{(1 - \rho)(1 + \gamma - \rho)}, 0, \dots, 0 \right], \Sigma_d \right),$$

with $\Sigma_d = \Gamma \Gamma^T \circ H$ (" $\circ$ " is the Hadamard or entrywise product), where

$$\Gamma = \begin{bmatrix} 1 + \gamma \\ 1 + g \\ \cdot \\ \cdot \\ \cdot \\ 1 + g \end{bmatrix}, H = \begin{bmatrix} 1 & h_{12} & \dots & h_{1d} \\ h_{12} & 1 - \nu^2 & \dots & h_{2d} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ h_{1d} & h_{2d} & \dots & 1 - \nu^2 \end{bmatrix},$$

$$h_{1i} = (1 - \nu^2) \left[R_{1i}(1, 1) + \frac{g - \gamma}{\gamma + g + 1} \left[(2\gamma + 1) \int_0^1 \frac{R_{1i}(s, 1)}{s^{1-\gamma}} ds - (2g + 1) \int_0^1 \frac{R_{1i}(1, t)}{t^{1-g}} dt \right] \right], \text{ and}$$

$$h_{ij} = (1 - \nu^2) R_{ij}(1, 1), \quad i = 2, \dots, d, j = i + 1, \dots, d.$$

Very similar to the bivariate case, let $\lambda = 0$ and derive the SSE of γ by using the approximate multivariate normal distribution of $(\hat{\gamma}, \hat{g}_2 - g, \dots, \hat{g}_d - g)$, with mean $[\gamma, 0, \dots, 0]$, and variance $\frac{1}{k} \hat{\Sigma}_d = \frac{1}{k} \hat{\Gamma} \hat{\Gamma}^T \circ \hat{H}$, where for the estimation of R_{ij} , $\hat{R}_{ij}$ is defined like in (5). By maximizing the approximate likelihood function of $(\hat{\gamma}, (\hat{g}_2 - g), \dots, (\hat{g}_d - g))$ with respect to γ , we obtain the SSE in this multivariate setting:

$$(9) \quad \hat{\gamma}_g = \hat{\gamma} + \frac{1 + \hat{\gamma}}{1 + g} \sum_{j=2}^d \frac{\hat{H}_{1j}^{-1}}{\hat{H}_{11}^{-1}} (\hat{g}_j - g),$$

where $\hat{H}_{ij}^{-1}$ is the entry in the i^{th} row and j^{th} column of the inverse of the matrix $\hat{H}$. The following theorem shows the asymptotic behavior of the improved estimator $\hat{\gamma}_g$.

THEOREM 3.1. *Assume that H is invertible. Then under the conditions of Proposition 3.1, as $n \rightarrow \infty$,*

$$(10) \quad \sqrt{k}(\hat{\gamma}_g - \gamma) \xrightarrow{d} N \left(\left(\frac{\lambda(1 + \gamma)}{(1 - \rho)(1 + \gamma - \rho)}, \sigma^2 \right), \right),$$

where

$$\sigma^2 = (1 + \gamma)^2 \left(1 + \frac{1}{(H_{11}^{-1})^2} \left[2 \sum_{i=1}^d \sum_{j=i+1}^d H_{1i}^{-1} H_{1j}^{-1} h_{ij} + (1 - \nu^2) \sum_{j=2}^d (H_{1j}^{-1})^2 \right] \right).$$

4. Simulation study. In this section we perform for the one-covariate setting a simulation study. First we investigate the finite sample performance of our novel SSE of γ and then we compare in detail the variances of the SSE with those of the pseudo-MLE based on $X_{n-k:n}, \dots, X_{n:n}$ only. In addition, we estimate an extreme quantile by substituting the SSE $\hat{\gamma}_g$ and a similar SSE $\hat{\sigma}_g$ of the scale a , in the generic formula of the extreme quantile estimator. Again, we compare the variances of this SSE and the classical estimator based on the pseudo-MLEs of γ and a .

We begin with simulating data from the bivariate Cauchy distribution restricted to the first quadrant. This Cauchy density is proportional to

$$(1 + xS^{-1}x^T)^{-3/2}$$

where S is a 2×2 scale matrix with 1 on the diagonal and s off-diagonal. For s we take two values: 0 and 0.8. These data are denoted by $(\tilde{X}_i, Y_i)$. To obtain our data (X_i, Y_i) , where the X_i have extreme value index γ , we transform the $\tilde{X}_i$, as follows

$$X_i = \begin{cases} \frac{(1-F_s(\tilde{X}_i))^{-\gamma}-1}{\gamma}, & \gamma \neq 0, \\ -\log(1-F_s(\tilde{X}_i)), & \gamma = 0, \end{cases}$$

where F_s is the distribution function of $\tilde{X}_i$. Simulations are performed for values of γ that are negative, positive or 0.

First, we generate 500 samples of sizes $n = 500, m = 1000$, for $s = 0.8$, and estimate γ using the SSE and the pseudo-MLE for $k = 1, \dots, 499$. We depict the root mean squared error (RMSE) based on these 500 samples as a function of k . We consider $\gamma = -0.25, 0$, and, 0.25, and take $g = 0$. The RMSE of the SSE (indicated by AMLE in Figure 1) is indeed substantially lower than that of the pseudo-MLE for the different values of γ .

FIGURE 1. RMSE using the pseudo-MLE and the SSE-MLE. From left to right: $\gamma = -0.25, 0, 0.25$.

Next, we focus on the (relative) variance reduction of the SSE in comparison to the pseudo-MLE. We use the following values of n and m (and k):

- $n = 1000, m = 500$ (less unlabeled than labeled data) and $k = 250$,
- $n = 1000, m = 1000$ (equal number of unlabeled and labeled data) and $k = 250$,
- $n = 500, m = 1000$ (more unlabeled than labeled data) and $k = 125$.

Table 1 shows the empirical percentages of variance reduction for different values of γ and g . The results are based on 10,000 replications. We observe that the variance reduction ranges from 10% to more than 30%, hence indeed the SSE has a substantially smaller variance than

the pseudo-MLE. By comparing the three panels, we observe that the variance reduction increases substantially with the ratio of the number of unlabeled data m and the number of labeled data n , which is in line with the asymptotic theory. Observe that the actual choice of g does not have a large influence as long as it is somewhat close to γ , a choice that is in practice often feasible.

TABLE 1
Variance reduction for different extreme value indices

(i) $(n, m) = (1000, 500)$									
g	γ				g	γ			
	-0.3	-0.2	-0.1	0		0	0.1	0.2	0.3
-0.25	13.8%	14.0%	13.4%	12.0%	0	15.4%	15.6%	15.4%	14.8%
-0.125	12.6%	14.2%	14.8%	14.1%	0.125	15.3%	15.9%	16.1%	15.8%
0	10.4%	13.0%	14.6%	15.4%	0.25	14.7%	15.5%	16.1%	16.4%
(ii) $(n, m) = (1000, 1000)$									
	-0.3	-0.2	-0.1	0		0	0.1	0.2	0.3
-0.25	20.2%	21.1%	20.5%	18.8%	0	23.6%	23.5%	23.1%	22.2%
-0.125	18.9%	21.5%	22.4%	22.0%	0.125	23.4%	24.1%	24.3%	25.0%
0	16.3%	19.9%	22.3%	23.6%	0.25	22.4%	23.7%	24.6%	24.6%
(iii) $(n, m) = (500, 1000)$									
	-0.3	-0.2	-0.1	0		0	0.1	0.2	0.3
-0.25	22.4%	25.8%	25.7%	24.5%	0	30.8%	30.2%	29.9%	29.4%
-0.125	21.6%	26.2%	28.0%	28.0%	0.125	31.0%	30.4%	31.0%	30.8%
0	18.6%	24.6%	28.1%	30.8%	0.25	30.0%	30.1%	31.4%	31.5%

Next we investigate in more detail the sensitivity of the variance reduction to the choice of g using a wider range of values of g , including cases where γ and g have opposite sign. In these simulations, we take $s = 0$ for the bivariate Cauchy distribution. The results, based on 10,000 replications, for the aforementioned values of n, m and k are presented in Figure 2. Generally, there is always variance reduction, but for $|g - \gamma|$ relatively large the reduction is lower than when g is closer to γ . Observe that for the present range of γ (-0.3 to 0.3) the choice $g = 0$ yields an almost maximal variance reduction.

Finally we study in more detail the effect of the size of m , the number of unlabeled data, on the variance reduction; again we take $s = 0$. We consider the case where $n = 500$ and let m vary; we choose $g = 0$. The results are based on 500 replications. Table 2 shows that the variance reduction approximately doubles when m ranges from 500 to 10,000.

TABLE 2
Variance reduction for different numbers of unlabeled data m

m	γ						
	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
500	6.0%	8.3%	9.6%	11.4%	11.7%	11.9%	11.8%
1000	8.7%	12.3%	14.4%	15.6%	12.9%	13.0%	15.5%
5000	11.2%	16.4%	19.5%	22.5%	21.8%	22.0%	21.7%
10000	11.9%	17.5%	20.8%	24.6%	23.0%	23.2%	22.9%

The last part of this section is devoted to extending the semi-supervised estimation approach to scale estimation and in particular to extreme quantile estimation. Here we confine ourselves to the choice $g = 0$. The scale $a = a(n/k)$ can be estimated with $\hat{\sigma}$, based on

FIGURE 2. *Variance reduction for various combinations of γ and g* (i) $(n, m) = (1000, 500)$, $k = 250$ (ii) $(n, m) = (1000, 1000)$, $k = 250$ (iii) $(n, m) = (500, 1000)$, $k = 125$

$X_{n-k:n}, \dots, X_{n:n}$; see Section 3.4 in de Haan and Ferreira (2006) for definitions and results. Somewhat similar to the derivation of the SSE $\hat{\gamma}_g$, the SSE of the scale can be derived as:

$$\hat{\sigma}_0 = \hat{\sigma} \left(1 - \hat{S}_0(\tilde{\sigma} - 1) \right),$$

where $\tilde{\sigma}$ is the pseudo-MLE of the scale parameter (which is equal to 1) based on $\tilde{Y}_{n-k:n}, \dots, \tilde{Y}_{n:n}$, and

$$\begin{aligned} \hat{S}_0 = \frac{1}{2} & \left((3\hat{\gamma}_0 - 1) \int_0^1 \frac{\hat{R}(1, t)}{t} dt + \hat{\gamma}_0 \int_0^1 \frac{\hat{R}(1, t)}{t} \log t dt \right. \\ & \left. - (2\hat{\gamma}_0 + 1)^2 \int_0^1 \frac{\hat{R}(s, 1)}{s^{1-\hat{\gamma}_0}} ds + 2(\hat{\gamma}_0 + 2) \hat{R}(1, 1) \right). \end{aligned}$$

For very small $p \in (0, 1)$, an extreme quantile is defined to be $x_p = F_1^{-1}(1 - p)$. It is usually estimated by

$$(11) \quad \hat{x}_p = X_{n-k:n} + \hat{\sigma} \frac{\left(\frac{k}{np}\right)^{\hat{\gamma}} - 1}{\hat{\gamma}},$$

see Section 4.3 in de Haan and Ferreira (2006). Using our SSEs of γ and a , we define the SSE of x_p as

$$(12) \quad \hat{x}_{p,0} = X_{n-k:n} + \hat{\sigma}_0 \frac{\left(\frac{k}{np}\right)^{\hat{\gamma}_0} - 1}{\hat{\gamma}_0}.$$

For n and m as in Table 1, we again simulate 10,000 times from the bivariate Cauchy distribution with $s = 0$. For γ ranging from -0.3 to 0.3 , we aim to estimate the extreme quantile x_p for $p = \frac{1}{n}$. Table 3 shows the reduction in the variance of $\hat{\gamma}_0, \hat{\sigma}_0$, and $\hat{x}_{p,0}$, respectively, when compared with the pseudo-MLE estimators based on only $X_{n-k:n}, \dots, X_{n:n}$. Remarkably, the variance reduction when estimating the extreme quantile x_p ranges from 15% to more than 30%. It is much higher than the reductions when estimating γ or a and in almost all cases it even exceeds the sum of both reductions.

TABLE 3
Variance reduction

	γ	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
$\hat{\gamma}_0$ vs. $\hat{\gamma}$	$n > m$	6.6%	7.6%	9.3%	8.9%	9.9%	9.8%	9.5%
	$n = m$	9.4%	12.0%	13.6%	13.1%	14.5%	13.1%	13.9%
	$n < m$	8.7%	12.9%	14.5%	16.4%	16.5%	16.8%	16.6%
$\hat{\sigma}_0$ vs. $\hat{\sigma}$	$n > m$	4.5%	4.7%	5.4%	4.8%	5.5%	5.4%	5.3%
	$n = m$	6.1%	7.0%	7.5%	6.7%	7.8%	7.7%	7.6%
	$n < m$	6.1%	8.1%	8.4%	8.1%	8.9%	8.9%	8.7%
$\hat{x}_{p,0}$ vs. $\hat{x}_p$	$n > m$	15.7%	15.8%	17.5%	16.0%	16.3%	15.6%	14.9%
	$n = m$	22.6%	25.3%	25.2%	23.5%	23.4%	22.3%	21.2%
	$n < m$	27.6%	30.6%	32.5%	31.2%	31.6%	31.0%	30.5%

5. Application. In this section, we demonstrate an application using the SSE for analyzing forecasted precipitation data.

The national French weather service, Météo France, produces daily forecasted precipitation (in mm) at very high resolution ($0.1^\circ \times 0.1^\circ$) covering the mainland of France, between 2012 and 2017. To improve the forecasting model, meteorologists want to check if the forecasted precipitation shares the same distribution as the observed precipitation at the same

To further analyze the variance reduction factor, we plot the histogram of the estimated variance reduction factors across all 123 locations in Figure 3. The variance reduction using the SSE compared to the pseudo-MLE ranges from 9% to 27%, and is on average 16.54%. This confirms the improved performance of the SSE for the forecasted precipitation data.

FIGURE 3. *Histogram of the variance reduction factor across 123 stations*

6. Proofs. We first present proofs for the one-covariate (bivariate) case and then extend the proofs to the multivariate case. The asymptotic normality for $\sqrt{k}(\hat{\gamma} - \gamma)$, the first component of the pair in Proposition 2.1, is established in Drees et al. (2004). However, we cannot directly use that proof, since we have to keep track of the joint behavior of $\hat{\gamma}$ and $\hat{g}$. Nevertheless, we mimic that proof for both $\hat{\gamma}$ and $\hat{g}$, with observing that $\hat{g}$ is based on dependent observations. In this respect the proof has to be adapted substantially. We begin with various lemmas which are needed for the main proofs.

Let C be a copula corresponding to the distribution function of $(-X_1, -Y_1)$. Let $(V_{1,1}, V_{1,2}), \dots, (V_{n,1}, V_{n,2})$ be a random sample of size n from C and $V_{n+1,2}, \dots, V_{n+m,2}$ be a random sample of size m from the uniform- $(0, 1)$ distribution, independent of the random sample from C . Clearly all the $V_{i,j}, i = 1, \dots, n, j = 1, 2$, have also a uniform- $(0, 1)$ distribution. Write $X_i = F_1^{-1}(1 - V_{i,1}), i = 1, \dots, n$, and $Y_l = F_2^{-1}(1 - V_{l,2}), l = 1, \dots, n + m$. Then $(X_i, Y_i), i = 1, \dots, n$, and $Y_{n+1}, \dots, Y_{n+m}$ have the distributions as specified in the beginning of Section 2.

Consider the following uniform empirical distribution functions:

$$\Gamma_{n,j}(s) = \frac{1}{n} \sum_{i=1}^n 1_{[0,s]}(V_{i,j}), \quad 0 \leq s \leq 1, j = 1, 2,$$

$$\Gamma_{n+m}(t) = \frac{1}{n+m} \sum_{l=1}^{n+m} 1_{[0,t]}(V_{l,2}), \quad 0 \leq t \leq 1.$$

The corresponding uniform tail empirical processes are

$$w_{n,j}(s) = \sqrt{k} \left[\frac{n}{k} \Gamma_{n,j} \left(\frac{k}{n} s \right) - s \right], \quad 0 \leq s \leq 1, j = 1, 2,$$

$$w_{n+m}(t) = \sqrt{\frac{(n+m)k}{n}} \left[\frac{n}{k} \Gamma_{n+m} \left(\frac{k}{n} t \right) - t \right], \quad 0 \leq t \leq 1.$$

Define the Gaussian vector of processes (W_1, W_2, W_3) , where W_j , $j = 1, 2, 3$, is a standard Wiener process on $[0, T]$, $T > 0$, with covariances:

$$\text{Cov}(W_1(s), W_2(t)) = R(s, t), \quad 0 \leq s, t \leq T,$$

$$\text{Cov}(W_1(s), W_3(t)) = \nu R(s, t), \quad 0 \leq s, t \leq T,$$

$$\text{Cov}(W_2(s), W_3(t)) = \nu(s \wedge t), \quad 0 \leq s, t \leq T.$$

Let I denote the identity function. Then we have on $(D[0, T])^3$, for all $0 \leq \delta < \frac{1}{2}$, as $n \rightarrow \infty$,

$$(13) \quad \left(\frac{w_{n,1}}{I^\delta}, \frac{w_{n,2}}{I^\delta}, \frac{w_{n+m}}{I^\delta} \right) \xrightarrow{d} \left(\frac{W_1}{I^\delta}, \frac{W_2}{I^\delta}, \frac{W_3}{I^\delta} \right).$$

The proof of (13) is given in Ahmed and Einmahl (2019); note that in there $T = 1$, but the proof for arbitrary $T > 0$ follows similarly. Now using a Skorohod construction we obtain from (13) that

$$(14) \quad \sup_{0 < s \leq T} \frac{|w_{n,j}(s) - W_j(s)|}{s^\delta} \xrightarrow{a.s.} 0, j = 1, 2, \quad \text{and} \quad \sup_{0 < s \leq T} \frac{|w_{n+m}(s) - W_3(s)|}{s^\delta} \xrightarrow{a.s.} 0.$$

The processes in (14) are different from those in (13) but we keep the same notation, since the new vector $(w_{n,1}, w_{n,2}, w_{n+m})$ has the same distribution as the old vector and also the new vector (W_1, W_2, W_3) has the same distribution as the old vector. In the sequel the X_i and Y_i are transformations as above of the uniform-(0,1) random variables on which the $w_{n,j}$ are based. We continue with the processes satisfying (14).

For convenience we introduce the following notation. Let f_n, h_n be positive functions on $[l_n, u_n]$. Then we write, as $n \rightarrow \infty$,

$$f_n \stackrel{\mathbb{P}}{\asymp} h_n|_{l_n}^{u_n},$$

if both $f_n(s) = O_{\mathbb{P}}(h_n(s))$ and $h_n(s) = O_{\mathbb{P}}(f_n(s))$ hold uniformly for $s \in [l_n, u_n]$. This notation is useful for the following lemma, which can be found in Shorack and Wellner (2009), p. 419.

LEMMA 6.1. *Let $\Gamma_{n,j}^{-1}, j = 1, 2$, be the empirical quantile functions corresponding to $\Gamma_{n,j}, j = 1, 2$, respectively. Then, as $n \rightarrow \infty$,*

$$\Gamma_{n,j} \stackrel{\mathbb{P}}{\asymp} I|_{\Gamma_{n,j}^{-1}(1/(2n))}^1 \quad \text{and} \quad \Gamma_{n,j}^{-1} \stackrel{\mathbb{P}}{\asymp} I|_{1/(2n)}^1, \quad j = 1, 2.$$

The following lemma states the weighted convergence of the tail quantile processes corresponding to $\Gamma_{n,j}^{-1}, j = 1, 2$, to the processes $-W_1$ and $-W_2$ in (14).

LEMMA 6.2. *Let $\Gamma_{n,j}^{-1}$, be the empirical quantile functions corresponding to $w_{n,j}$, $j = 1, 2$, in (14) and let W_j be as in (14), $j = 1, 2$. Then for any $\delta < \frac{1}{2}$, as $n \rightarrow \infty$,*

$$(15) \quad \sup_{\frac{1}{2k} \leq s \leq 1} \frac{|\sqrt{k}(\frac{n}{k}\Gamma_{n,j}^{-1}(\frac{k}{n}s) - s) + W_j(s)|}{s^\delta} \xrightarrow{\mathbb{P}} 0.$$

PROOF. Write $v_{n,j}(s) = \sqrt{k}(\frac{n}{k}\Gamma_{n,j}^{-1}(\frac{k}{n}s) - s)$, $j = 1, 2$. Theorem 2.3 in Einmahl (1992) yields, as $n \rightarrow \infty$,

$$(16) \quad \sup_{\frac{1}{2k} \leq s \leq 1} \frac{|v_{n,j}(s) - W_{n,j}(s)|}{s^\delta} \xrightarrow{\mathbb{P}} 0,$$

where $W_{n,j}$ is an appropriate sequence of standard Wiener processes. Let W be a standard Wiener process and let $\varepsilon > 0$. It is well-known that there exist an $\eta > 0$, such that

$$(17) \quad \mathbb{P} \left(\sup_{0 < s \leq \eta} \frac{|W(s)|}{s^\delta} \geq \frac{\varepsilon}{2} \right) \leq \frac{\varepsilon}{2}.$$

Combining (16) and (17) yields, for large n ,

$$(18) \quad \mathbb{P} \left(\sup_{\frac{1}{2k} \leq s \leq \eta} \frac{|v_{n,j}(s)|}{s^\delta} \geq \varepsilon \right) \leq \varepsilon.$$

Combining (17), (18), (14), and Lemma 1 in Vervaat (1972), yields (15). $\square$

The next lemma is very similar to Lemma 3.1 in Drees et al. (2004), but the lemma therein cannot be used here because we need specifically the approximation with the present W_1 in order to obtain the joint behavior of $\hat{\gamma}$ and $\hat{g}$.

LEMMA 6.3. *Let $\varepsilon > 0$. Assume that (3) and (4) hold and $\sqrt{k}A(\frac{n}{k}) = O(1)$, as $n \rightarrow \infty$. Then for suitably chosen functions a and A in (4), as $n \rightarrow \infty$,*

$$\sup_{\frac{1}{2k} \leq s \leq 1} s^{\gamma+1/2+\varepsilon} \left| \sqrt{k} \left(\frac{X_{n-[ks]:n} - U_1(\frac{n}{k})}{a(\frac{n}{k})} - \frac{s^{-\gamma} - 1}{\gamma} \right) - s^{-\gamma-1}W_1(s) - \sqrt{k}A\left(\frac{n}{k}\right)\Psi(s^{-1}) \right| \xrightarrow{\mathbb{P}} 0.$$

PROOF. From (4) we obtain inequality (2.3.17) in de Haan and Ferreira (2006): for any $\theta, \delta > 0$ to be specified later, there exists $t_0 = t_0(\theta, \delta)$ such that for all $t, tx \geq t_0$,

$$\left| \frac{\frac{U_1(tx) - U_1(t)}{a(t)} - \frac{x^\gamma - 1}{\gamma}}{A(t)} - \Psi(x) \right| \leq \theta x^{\gamma+\rho} \max(x^\delta, x^{-\delta}).$$

We replace tx by $1/\Gamma_{n,1}^{-1}(\frac{k}{n}s)$ and t by $\frac{n}{k}$. Then we have, writing $\check{s} = \frac{n}{k}\Gamma_{n,1}^{-1}(\frac{k}{n}s)$, with probability tending to 1, as $n \rightarrow \infty$,

$$(19) \quad \left| \frac{X_{n-[ks]:n} - U_1(\frac{n}{k})}{a(\frac{n}{k})} - \frac{\check{s}^{-\gamma} - 1}{\gamma} - A\left(\frac{n}{k}\right)\Psi(\check{s}^{-1}) \right| \leq \left| A\left(\frac{n}{k}\right) \right| \theta \check{s}^{-\gamma-\rho} \cdot \max(\check{s}^{-\delta}, \check{s}^\delta).$$

Define $f(s) = \frac{s^{-\gamma-1}}{\gamma}$. Then by a Taylor expansion for some $\check{\Theta}_n(s)$ between $\check{s}$ and s we have

$$f(\check{s}) - f(s) = f'(s)(\check{s} - s) + \frac{f''(\check{\Theta}_n(s))}{2}(\check{s} - s)^2.$$

Lemma 6.1 implies $\check{\Theta}_n \asymp I|_{\frac{1}{2k}}^{\mathbb{P}}$ and thus $f''(\check{\Theta}_n) \asymp I^{-\gamma-2}|_{\frac{1}{2k}}^{\mathbb{P}}$. Next, by Lemma 6.2 and the fact that for all $\delta_1 < \frac{1}{2}$, $\sup_{0 \leq s \leq 1} |W_1(s)|/s^{\delta_1} = O_{\mathbb{P}}(1)$, we have that, as $n \rightarrow \infty$,

$$\sup_{\frac{1}{2k} \leq s \leq 1} (\check{s} - s)^2 / s^{2\delta_1} = O_{\mathbb{P}}\left(\frac{1}{k}\right).$$

This and again Lemma 6.2 with $\delta = \delta_1$ yield, as $n \rightarrow \infty$, uniformly for all $\frac{1}{2k} \leq s \leq 1$,

$$f(\check{s}) - f(s) = -s^{-\gamma-1} \frac{1}{\sqrt{k}} \left(-W_1(s) + s^{\delta_1} O_{\mathbb{P}}(1) \right) + s^{-\gamma-2+2\delta_1} O_{\mathbb{P}}\left(\frac{1}{k}\right).$$

Choose δ_1 such that $\frac{1-\varepsilon}{2} < \delta_1 < \frac{1}{2}$. Then $\delta_1 > \frac{1}{2} - \varepsilon$ and $2\delta_1 + \varepsilon > 1$. Hence, as $n \rightarrow \infty$,

$$\sup_{\frac{1}{2k} \leq s \leq 1} s^{-\frac{3}{2}+\varepsilon+2\delta_1} \leq \max\left(1, (2k)^{\frac{3}{2}-\varepsilon-2\delta_1}\right) = o(\sqrt{k}).$$

Therefore, as $n \rightarrow \infty$, uniformly for all $\frac{1}{2k} \leq s \leq 1$,

(20)

$$\begin{aligned} f(\check{s}) &= f(s) + \frac{1}{\sqrt{k}} s^{-\gamma-1} \left(W_1(s) + s^{\delta_1} O_{\mathbb{P}}(1) + s^{-1+2\delta_1} O_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right) \right) \\ &= f(s) + \frac{1}{\sqrt{k}} s^{-\gamma-1} \left(W_1(s) + s^{1/2-\varepsilon} \left(s^{\delta_1-1/2+\varepsilon} O_{\mathbb{P}}(1) + s^{-3/2+\varepsilon+2\delta_1} O_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right) \right) \right) \\ &= \frac{s^{-\gamma}-1}{\gamma} + \frac{1}{\sqrt{k}} s^{-\gamma-1} \left(W_1(s) + s^{1/2-\varepsilon} O_{\mathbb{P}}(1) \right). \end{aligned}$$

From the mean value theorem, for some $\Theta_n(s)$ between $\check{s}$ and s

$$\Psi(\check{s}^{-1}) = \Psi(s^{-1}) - \Psi'(1/\Theta_n(s))(\Theta_n(s))^{-2}(\check{s} - s).$$

As above, $\Theta_n \asymp I|_{\frac{1}{2k}}^{\mathbb{P}}$, which implies that as $n \rightarrow \infty$, uniformly for $\frac{1}{2k} \leq s \leq 1$,

$$|\Psi'(1/\Theta_n(s))(\Theta_n(s))^{-2}| = s^{-\gamma-\rho-1}(1 + |\log s|)O_{\mathbb{P}}(1).$$

Hence, using Lemma 6.2 with $\delta = \delta_1$ (as above), we have uniformly for $\frac{1}{2k} \leq s \leq 1$,

$$A\left(\frac{n}{k}\right) (\Psi(\check{s}^{-1}) - \Psi(s^{-1})) = \frac{1}{\sqrt{k}} A\left(\frac{n}{k}\right) s^{-\gamma-\rho-1+\delta_1} (1 + |\log s|) O_{\mathbb{P}}(1).$$

With δ_1 chosen as above, we have that as $n \rightarrow \infty$, uniformly for $\frac{1}{2k} \leq s \leq 1$,

$$(21) \quad A\left(\frac{n}{k}\right) \Psi(\check{s}^{-1}) = A\left(\frac{n}{k}\right) \Psi(s^{-1}) + \frac{1}{\sqrt{k}} s^{-\gamma-\varepsilon-\frac{1}{2}} O_{\mathbb{P}}(1).$$

Next consider the right-hand side of (19), where we take $\delta < 1/2$. Using Lemma 6.1, it can be bounded, uniformly for $\frac{1}{2k} \leq s \leq 1$, by

$$\begin{aligned} &\theta \left| A\left(\frac{n}{k}\right) \right| s^{-\gamma-\rho-\delta} O_{\mathbb{P}}(1) = \theta \sqrt{k} \left| A\left(\frac{n}{k}\right) \right| \frac{1}{\sqrt{k}} s^{-\gamma-\delta} O_{\mathbb{P}}(1) \\ (22) \quad &= \theta \frac{1}{\sqrt{k}} s^{-\gamma-\varepsilon-1/2} s^{\varepsilon+1/2-\delta} O_{\mathbb{P}}(1) = \theta \frac{1}{\sqrt{k}} s^{-\gamma-\varepsilon-1/2} O_{\mathbb{P}}(1). \end{aligned}$$

Now, plugging (20), (21), and (22) into inequality (19) and noting that $\theta > 0$ can be chosen arbitrarily small, we obtain the statement in the lemma. $\square$

Define

$$Z_n(s) = \sqrt{k} \left(\frac{X_{n-[ks]:n} - X_{n-k:n}}{a\left(\frac{n}{k}\right)} - \frac{s^{-\gamma} - 1}{\gamma} \right).$$

Then for functions a and A as in Lemma 6.3, for any $\varepsilon > 0$, uniformly for $\frac{1}{2k} \leq s \leq 1$,

$$Z_n(s) = s^{-\gamma-1} W_1(s) - W_1(1) + \sqrt{k} A\left(\frac{n}{k}\right) \Psi(s^{-1}) + o_{\mathbb{P}}(1) s^{-\gamma-1/2-\varepsilon}.$$

Hence for $\gamma > -\frac{1}{2}$,

$$(23) \quad \sup_{\frac{1}{2k} \leq s \leq 1} s^{\gamma+1/2+\varepsilon} |Z_n(s)| = O_{\mathbb{P}}(1).$$

PROPOSITION 6.1. *Under the conditions of Lemma 6.3, for $\gamma > -\frac{1}{2}$ and $\gamma \neq 0$, as $n \rightarrow \infty$.*

$$\sqrt{k}(\hat{\gamma} - \gamma) - \frac{(\gamma + 1)^2}{\gamma} \int_0^1 (s^\gamma - (2\gamma + 1)s^{2\gamma}) Z_n(s) ds = o_{\mathbb{P}}(1),$$

and, for $\gamma = 0$,

$$\sqrt{k}\hat{\gamma} + \int_0^1 (2 + \log s) Z_n(s) ds = o_{\mathbb{P}}(1).$$

PROOF. Using Lemma 6.3 and (23) above in conjunction with Lemma 3.2 in Drees et al. (2004), the result is obtained following the same steps as in the proof of Proposition 3.1 in Drees et al. (2004). $\square$

To study the asymptotic behavior of $\hat{g}$ we need the following result. Define

$$\tilde{w}_n(s) = \frac{n}{\sqrt{k}} \left(\Gamma_{n+m} \left(\Gamma_{n,2}^{-1} \left(\frac{k}{n} s \right) \right) - \frac{k}{n} s \right) \quad \text{and} \quad \tilde{W}(s) = \nu W_3(s) - W_2(s).$$

LEMMA 6.4. *Assume that F_2 is continuous and k satisfies (3), then for any $0 \leq \delta < \frac{1}{2}$, as $n \rightarrow \infty$,*

$$\sup_{\frac{1}{2k} \leq s \leq 1} \frac{|\tilde{w}_n(s) - \tilde{W}(s)|}{s^\delta} \xrightarrow{\mathbb{P}} 0.$$

PROOF. We have

$$\tilde{w}_n(s) = \sqrt{\frac{n}{n+m}} w_{n+m} \left(\frac{n}{k} \Gamma_{n,2}^{-1} \left(\frac{k}{n} s \right) \right) + \frac{n}{\sqrt{k}} \left(\Gamma_{n,2}^{-1} \left(\frac{k}{n} s \right) - \frac{k}{n} s \right).$$

Define $\hat{s} = \frac{n}{k} \Gamma_{n,2}^{-1} \left(\frac{k}{n} s \right)$. From Lemma 6.2 with $j = 2$, (3) and (17), we see that it suffices to show that, as $n \rightarrow \infty$,

$$(24) \quad \sup_{\frac{1}{2k} \leq s \leq 1} \frac{|w_{n+m}(\hat{s}) - W_3(s)|}{s^\delta} \xrightarrow{\mathbb{P}} 0.$$

Let $s_0 \in (0, 1)$. We first handle the region $s \geq s_0$. Obviously we have $1/s^\delta \leq 1/s_0^\delta$. By Lemma 6.2, as $n \rightarrow \infty$,

$$(25) \quad \sup_{\frac{1}{2k} \leq s \leq 1} |\hat{s} - s| \xrightarrow{\mathbb{P}} 0.$$

Using this, (14), and the uniform continuity of W_3 we obtain, as $n \rightarrow \infty$,

$$\sup_{s_0 \leq s \leq 1} \frac{|w_{n+m}(\hat{s}) - W_3(s)|}{s^\delta} \xrightarrow{\mathbb{P}} 0.$$

It remains to show that for $\varepsilon > 0$ there exists $s_0 \in (0, 1)$ such that for large n

$$\mathbb{P} \left(\sup_{\frac{1}{2k} \leq s \leq s_0} \frac{|w_{n+m}(\hat{s}) - W_3(s)|}{s^\delta} \geq 3\varepsilon \right) \leq 3\varepsilon.$$

Using again (17), for this it suffices to show that

$$\mathbb{P} \left(\sup_{\frac{1}{2k} \leq s \leq s_0} \frac{|w_{n+m}(\hat{s})|}{s^\delta} \geq 2\varepsilon \right) \leq 2\varepsilon.$$

Using Lemma 6.1, the proof is complete if we show that for all $\varepsilon > 0, \kappa > 0$ there exists $s_0 \in (0, 1)$ such that for large n

$$\mathbb{P} \left(\sup_{\frac{1}{2k} \leq s \leq s_0} \frac{|w_{n+m}(\hat{s})|}{\hat{s}^\delta} \geq 2\kappa \right) \leq \varepsilon.$$

We have

$$\mathbb{P} \left(\sup_{\frac{1}{2k} \leq s \leq s_0} \frac{|w_{n+m}(\hat{s})|}{\hat{s}^\delta} \geq 2\kappa \right) \leq \mathbb{P} \left(\sup_{0 < t \leq 2s_0} \left| \frac{w_{n+m}(t)}{t^\delta} \right| \geq \kappa \right) + \mathbb{P}(\hat{s} > 2s_0).$$

From (14) and (17), we have that for small enough $s_0 \in (0, 1)$ the first term on the right is bounded by $\varepsilon/2$ for large n , and using (25) we obtain that the second term on the right also does not exceed $\varepsilon/2$ for large n . $\square$

In the following we prove a result for the tail quantile process based on $\{\tilde{Y}_i\}_{i=1}^n$ instead of $\{X_i\}_{i=1}^n$. The proof of the next lemma uses Lemma 6.4, which is very similar to but easier than that of Lemma 6.3, and hence will be omitted.

LEMMA 6.5. *Let $\varepsilon > 0$. Assume that F_2 is continuous and that (3) holds, then, as $n \rightarrow \infty$,*

$$\sup_{\frac{1}{2k} \leq s \leq 1} s^{g+1/2+\varepsilon} \left| \sqrt{k} \left(\frac{\left(\tilde{Y}_{n-[ks]:n} - \frac{(\frac{n}{k})^g - 1}{g} \right)}{(\frac{n}{k})^g} - \frac{s^{-g} - 1}{g} \right) + s^{-g-1} \tilde{W}(s) \right| \xrightarrow{\mathbb{P}} 0.$$

Define

$$H_n(s) := \sqrt{k} \left(\frac{\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}}{(\frac{n}{k})^g} - \frac{s^{-g} - 1}{g} \right).$$

Then for any $\varepsilon > 0$, uniformly for $s \in [\frac{1}{2k}, 1]$,

$$H_n(s) = \tilde{W}(1) - s^{-g-1} \tilde{W}(s) + o_{\mathbb{P}}(1) s^{-g-1/2-\varepsilon}.$$

Hence for $g > -\frac{1}{2}$,

$$(26) \quad \sup_{\frac{1}{2k} \leq s \leq 1} s^{g+1/2+\varepsilon} |H_n(s)| = O_{\mathbb{P}}(1).$$

Next we show a version of Lemma 3.2 in Drees et al. (2004) based on $\{\tilde{Y}_i\}_{i=1}^n$.

LEMMA 6.6. Assume that F_2 is continuous and k satisfies (3). Let g_n be a sequence of random variables such that

$$(27) \quad g_n = g + O_{\mathbb{P}}(k^{-1/2}).$$

Then, if $-1/2 < g < 0$ or $g > 0$, as $n \rightarrow \infty$,

$$(28) \quad \mathbb{P} \left(1 + g_n \frac{\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}}{\left(\frac{n}{k}\right)^g} \geq C_n s^{-g}, \text{ for all } s \in \left[\frac{1}{2k}, 1\right] \right) \rightarrow 1,$$

for some random variables $C_n > 0$ such that $1/C_n = O_{\mathbb{P}}(1)$.

If $g = 0$, as $n \rightarrow \infty$,

$$(29) \quad \mathbb{P} \left(1 + g_n \left(\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n} \right) \geq \frac{1}{2}, \text{ for all } s \in \left[\frac{1}{2k}, 1\right] \right) \rightarrow 1,$$

and

$$(30) \quad \sup_{s \in [0,1]} \tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n} = O_{\mathbb{P}}(\log k).$$

PROOF. Consider first $-1/2 < g < 0$ or $g > 0$. Applying Lemma 6.1 to Γ_{n+m} and $\Gamma_{n,2}^{-1}$ yields, as $n \rightarrow \infty$,

$$\Gamma_{n+m} \left(\Gamma_{n,2}^{-1} \left(\frac{k}{n} I \right) \right) \underset{\mathbb{P}}{\asymp} \frac{k}{n} I \Big|_{\frac{1}{2k}}.$$

Define $G_n(s) = \Gamma_{n+m}(\Gamma_{n,2}^{-1}(\frac{k}{n}s)) + \frac{1}{2(n+m)}$, $s \in (0, 1]$. Hence, as $n \rightarrow \infty$,

$$(31) \quad G_n \underset{\mathbb{P}}{\asymp} \frac{k}{n} I \Big|_{\frac{1}{2k}}.$$

Observe that for $g \neq 0$

$$\begin{aligned} s^g \left(1 + g_n \frac{\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}}{\left(\frac{n}{k}\right)^g} \right) &= s^g \left(1 + \frac{g_n}{g} \frac{[(G_n(s))^{-g} - (G_n(1))^{-g}]}{\left(\frac{n}{k}\right)^g} \right) \\ &= \frac{g_n}{g} \left(\frac{G_n(s)}{(ks)/n} \right)^{-g} + s^g \left[\left(1 - \left(\frac{G_n(1)}{k/n} \right)^{-g} \right) - \left(\frac{g_n}{g} - 1 \right) \left(\frac{G_n(1)}{k/n} \right)^{-g} \right] \\ &=: T_1(s) + s^g [T_2 - T_3]. \end{aligned}$$

From (31) and $g_n/g \xrightarrow{\mathbb{P}} 1$, we have that $1/\inf_{s \in [1/(2k), 1]} T_1(s) = O_{\mathbb{P}}(1)$, as $n \rightarrow \infty$. Lemma 6.4 for $s = 1$ yields that $T_2 = O_{\mathbb{P}}(1/\sqrt{k})$ and hence, since $g > -1/2$, $\sup_{s \in [1/(2k), 1]} s^g \cdot T_2 \xrightarrow{\mathbb{P}} 0$. By the assumption on g_n and again (31) we obtain similarly $\sup_{s \in [1/(2k), 1]} s^g \cdot T_3 \xrightarrow{\mathbb{P}} 0$. This yields (28).

In case $g = 0$, for $1/(2k) \leq s \leq 1$,

$$(32) \quad \tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n} = -\log G_n(s) + \log G_n(1) \leq 2 \log A_n - \log s,$$

with

$$A_n = \max \left(\sup_{s \in [\frac{1}{2k}, 1]} \frac{G_n(s)}{\frac{k}{n}s}, \sup_{s \in [\frac{1}{2k}, 1]} \frac{\frac{k}{n}s}{G_n(s)} \right).$$

If $g_n \geq 0$, then $1 + g_n \left(\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n} \right) \geq 1$. If $g_n < 0$, then for $1/(2k) \leq s \leq 1$,

$$1 + g_n \left(\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n} \right) \geq 1 + g_n(2 \log A_n + \log 2 + \log k).$$

Since, as $n \rightarrow \infty$, $A_n = O_{\mathbb{P}}(1)$ and $g_n = O_{\mathbb{P}}(k^{-1/2})$, we obtain (29). Finally, the sup in (30) is attained at $s = 1/(2k)$. Hence, (32) yields (30). $\square$

From Lemma 6.5, (26), and Lemma 6.6 we obtain, following the proof of Proposition 3.1 in Drees et al. (2004), the following analogue of Proposition 6.1.

PROPOSITION 6.2. *Assume that F_2 is continuous and k satisfies (3). For $g > -\frac{1}{2}$ and $g \neq 0$, as $n \rightarrow \infty$,*

$$\sqrt{k}(\hat{g} - g) - \frac{(g+1)^2}{g} \int_0^1 (s^g - (2g+1)s^{2g})H_n(s)ds = o_{\mathbb{P}}(1).$$

and, for $g = 0$,

$$\sqrt{k}\hat{g} + \int_0^1 (2 + \log s)H_n(s)ds = o_{\mathbb{P}}(1).$$

PROOF OF PROPOSITION 2.1. Combining (14), Propositions 6.1 and 6.2 we obtain, as $n \rightarrow \infty$,

$$\left(\sqrt{k}(\hat{\gamma} - \gamma), \sqrt{k}(\hat{g} - g) \right) \xrightarrow{d} \left(\Omega, \tilde{\Omega} \right),$$

where

$$\Omega = \frac{(\gamma+1)^2}{\gamma} \int_0^1 (s^\gamma - (2\gamma+1)s^{2\gamma})(s^{-\gamma-1}W_1(s) - W_1(1))ds + \frac{\lambda(\gamma+1)}{(1-\rho)(1+\gamma-\rho)}$$

and

$$\tilde{\Omega} = \frac{(g+1)^2}{g} \int_0^1 (t^g - (2g+1)t^{2g})(\tilde{W}(1) - t^{-g-1}\tilde{W}(t))dt.$$

Since the Wiener processes involved have mean zero, we obtain immediately the mean of the limiting pair. Also the individual variances of the limiting pair follow readily, see Drees et al. (2004). It remains to determine the covariance. Note that $Cov(W_1(s), -\tilde{W}(t)) = (1 - \nu^2)R(s, t)$. We have that

$$\begin{aligned} Cov(\Omega, \tilde{\Omega}) &= (1 - \nu^2) \frac{(\gamma+1)^2(g+1)^2}{\gamma g} \\ &\quad \cdot \int_0^1 \int_0^1 (s^\gamma - (2\gamma+1)s^{2\gamma})(t^g - (2g+1)t^{2g}) \left(\frac{R(s, t)}{s^{\gamma+1}t^{g+1}} - \frac{R(s, 1)}{s^{\gamma+1}} - \frac{R(1, t)}{t^{g+1}} + R(1, 1) \right) ds dt \\ &= (1 - \nu^2) \frac{(\gamma+1)^2(g+1)^2}{\gamma g} \int_0^1 \int_0^1 \left(\frac{1}{st} - \frac{(2g+1)}{st^{1-g}} - \frac{(2\gamma+1)}{s^{1-\gamma}t} + \frac{(2\gamma+1)(2g+1)}{s^{1-\gamma}t^{1-g}} \right) R(s, t) \\ &\quad - \left(\frac{t^g}{s} - \frac{(2g+1)t^{2g}}{s} - \frac{(2\gamma+1)t^g}{s^{1-\gamma}} + \frac{(2\gamma+1)(2g+1)t^{2g}}{s^{1-\gamma}} \right) R(s, 1) \\ &\quad - \left(\frac{s^\gamma}{t} - \frac{(2g+1)s^\gamma}{t^{1-g}} - \frac{(2\gamma+1)s^{2\gamma}}{t} + \frac{(2\gamma+1)(2g+1)s^{2\gamma}}{t^{1-g}} \right) R(1, t) \\ &\quad + (s^\gamma t^g - (2g+1)s^\gamma t^{2g} - (2\gamma+1)s^{2\gamma} t^g + (2\gamma+1)(2g+1)s^{2\gamma} t^{2g}) R(1, 1) ds dt. \end{aligned}$$

Using a change of variables and the first order homogeneity of R , we obtain:

$$\begin{aligned} \int_0^1 \int_0^1 \frac{R(s,t)}{st} ds dt &= \int_0^1 \int_0^t \frac{R(s,t)}{st} ds dt + \int_0^1 \int_0^s \frac{R(s,t)}{st} dt ds \\ &= \int_0^1 \int_0^1 \frac{R(ts,t)}{st} ds dt + \int_0^1 \int_0^1 \frac{R(s,ts)}{st} dt ds = \int_0^1 \frac{R(s,1)}{s} ds + \int_0^1 \frac{R(1,t)}{t} dt, \\ \int_0^1 \int_0^1 \frac{R(s,t)}{s^{1-\gamma}t} ds dt &= \int_0^1 \int_0^t \frac{R(s,t)}{s^{1-\gamma}t} ds dt + \int_0^1 \int_0^s \frac{R(s,t)}{s^{1-\gamma}t} dt ds \\ &= \int_0^1 \int_0^1 \frac{R(ts,t)}{(st)^{1-\gamma}} ds dt + \int_0^1 \int_0^1 \frac{R(s,ts)}{s^{1-\gamma}t} dt ds = \frac{1}{1+\gamma} \left[\int_0^1 \frac{R(s,1)}{s^{1-\gamma}} ds + \int_0^1 \frac{R(1,t)}{t} dt \right], \end{aligned}$$

and similarly

$$\int_0^1 \int_0^1 \frac{R(s,t)}{st^{1-g}} ds dt = \frac{1}{1+g} \left[\int_0^1 \frac{R(s,1)}{s} ds + \int_0^1 \frac{R(1,t)}{t^{1-g}} dt \right].$$

Also,

$$\begin{aligned} \int_0^1 \int_0^1 \frac{R(s,t)}{s^{1-\gamma}t^{1-g}} ds dt &= \int_0^1 \int_0^t \frac{R(s,t)}{s^{1-\gamma}t^{1-g}} ds dt + \int_0^1 \int_0^s \frac{R(s,t)}{s^{1-\gamma}t^{1-g}} dt ds \\ &= \int_0^1 \int_0^1 \frac{R(ts,t)}{s^{1-\gamma}t^{1-\gamma-g}} ds dt + \int_0^1 \int_0^1 \frac{R(s,ts)}{s^{1-\gamma-g}t^{1-g}} dt ds \\ &= \frac{1}{\gamma+g+1} \left[\int_0^1 \frac{R(s,1)}{s^{1-\gamma}} ds + \int_0^1 \frac{R(1,t)}{t^{1-g}} dt \right]. \end{aligned}$$

Substituting the expressions for these four integrals involving $R(s,t)$ in the formula for $Cov(\Omega, \hat{\Omega})$ above, we obtain that this covariance is equal to $(1-\nu^2)(\gamma+1)(g+1)R_g$. $\square$

PROOF OF THEOREM 2.1. From the uniform consistency of $\hat{R}$ on $[0, 1]^2$, it can be shown that $\hat{R}_g \xrightarrow{\mathbb{P}} R_g$. Using the latter convergence in combination with Proposition 2.1 we obtain that, as $n \rightarrow \infty$,

$$\sqrt{k}(\hat{\gamma}_g - \gamma) = \sqrt{k}(\hat{\gamma} - \gamma) - \frac{1+\gamma}{1+g} R_g \sqrt{k}(\hat{g} - g) + o_{\mathbb{P}}(1).$$

Now Proposition 2.1 in conjunction with the continuous mapping theorem yields (7). $\square$

The proof of Proposition 3.1 can be given along the same lines as that of Proposition 2.1 and will be omitted. Note that the lemmas and propositions needed for the proof of Proposition 2.1 are of univariate nature and that hence immediately very similar lemmas can be stated (and proved) in the more-covariates case. Once these results are given, only a straightforward covariance calculation remains; cf. Ahmed and Einmahl (2019) for the joint weak convergence of all the tail empirical processes involved.

PROOF OF THEOREM 3.1. From the uniform consistency of the tail copula estimators we obtain $\hat{H}_{1j}^{-1} \xrightarrow{\mathbb{P}} H_{1j}^{-1}, j = 1, \dots, d$, which in combination with Proposition 3.1 yields that, as $n \rightarrow \infty$,

$$\sqrt{k}(\hat{\gamma}_g - \gamma) = \sqrt{k}(\hat{\gamma} - \gamma) + \frac{1+\gamma}{1+g} \sum_{j=2}^d \frac{H_{1j}^{-1}}{H_{11}^{-1}} \sqrt{k}(\hat{g}_j - g) + o_{\mathbb{P}}(1).$$

Now Proposition 3.1 and the continuous mapping theorem yield (10). $\square$

REFERENCES

- AHMED, H. and EINMAHL, J. H. J. (2019). Improved estimation of the extreme value index using related variables. *Extremes* **22** 553–569.
- AZRIEL, D., BROWN, L. D., SKLAR, M., BERK, R., BUJA, A. and ZHAO, L. (2016). Semi-supervised linear regression. *arXiv preprint arXiv:1612.02391*.
- BEIRLANT, J., GOEGBEUR, Y., SEGERS, J. and TEUGELS, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley.
- BUISHAND, T., DE HAAN, L. and ZHOU, C. (2008). On spatial extremes: with application to a rainfall problem. *The Annals of Applied Statistics* **2** 624–642.
- CHAKRABORTTY, A. and CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics* **46** 1541–1572.
- COLES, S. G. and TAWN, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Series B (Methodological)* **53** 377–392.
- COLES, S. G. and WALSHAW, D. (1994). Directional modelling of extreme wind speeds. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **43** 139–157.
- DE HAAN, L. and DE RONDE, J. (1998). Sea and wind: multivariate extremes at work. *Extremes* **1** 7.
- DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: an Introduction*. Springer.
- DREES, H., FERREIRA, A. and DE HAAN, L. (2004). On maximum likelihood estimation of the extreme value index. *The Annals of Applied Probability* 1179–1201.
- DREES, H. and HUANG, X. (1998). Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis* **64** 25–46.
- EINMAHL, J. H. J. (1992). The a.s. behavior of the weighted empirical process and the LIL for the weighted tail empirical process. *The Annals of Probability* **20** 681–695.
- OESTING, M. and NAVEAU, P. (2020). Spatial Modeling of Heavy Precipitation by Coupling Weather Station Recordings and Ensemble Forecasts with Max-Stable Processes. *arXiv preprint arXiv:2003.05854*.
- SHORACK, G. R. and WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics*. SIAM.
- SMITH, R. L. (1987). Estimating tails of probability distributions. *The Annals of Statistics* **15** 1174–1207.
- VAPNIK, V. (2013). *The Nature of Statistical Learning Theory*. Springer.
- VERVAAT, W. (1972). Functional central limit theorems for processes with positive drift and their inverses. *Probability Theory and Related Fields* **23** 245–253.
- WASSERMAN, L. and LAFFERTY, J. D. (2008). Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems* 801–808.
- ZHANG, A., BROWN, L. D. and CAI, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics* **47** 2538–2566.
- ZHU, X. and GOLDBERG, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* **3** 1–130.