Electronic Journal of Statistics
Vol. 16 (2022) 3996–4030
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2038

Robust sieve M-estimation with an application to dimensionality reduction

Julien Bodelet

University of Zürich
Andreasstrasse 15, 8050 Zürich, Switzerland
e-mail: julien.bodelet@jacobscenter.uzh.ch

Davide La Vecchia

University of Geneva,
Blv. Pont d'Arve 40, 1211 Geneva, Switzerland
e-mail: davide.lavecchia@unige.ch

Abstract: We propose a sieve M-estimation procedure which combines the flexibility of semiparametric inference with the stability and reliability of infinitesimal robustness. We derive the asymptotic theory of the proposed estimators, studying their convergence rate. In the context of functional magnetic resonance imaging (fMRI) data analysis, we illustrate how to apply our procedure to conduct inference on a semiparametric dynamic factor model. Monte Carlo simulations and real data analysis exemplify the stability of our estimators, providing a comparison with the extant, non robust and routinely applied sieve M-estimators.

MSC2020 subject classifications: Primary 62G05, 62F35; secondary 62H25.
Keywords and phrases: Dynamic factor model, functional magnetic resonance imaging, Huber loss function, outliers, semiparametric modeling.

Received May 2021.

Contents

1. Introduction

Sieve estimation represents a powerful technique to conduct semi- and nonparametric inference on complex, functional parameters. The method is intuitive and flexible: the estimators are based on the optimization of an empirical criterion over a sequence of approximating parameter spaces (the so-called sieve spaces), which are less complex than (and dense in) the original space; see e.g. Grenander [14] for a book-length introduction.

One attractive feature of sieve estimation is that the solution to the optimization problem is often obtained using M-estimation theory. Then, the theory of empirical processes yields the asymptotic properties of the resulting estimators. For book-length presentation we refer to Van Der Vaart and Wellner [31], van de Geer [29], and Kosorok [19].

Besides its use in theoretical statistics, sieve estimation is applied also in other fields. Just to mention some examples, sieve method is common in econometrics (for a survey see e.g. Chen [3] and Chen et al. [5]), in finance (see e.g. Fengler et al. [10]), in the analysis of functional magnetic resonance imaging (see e.g. Park et al. [23] and van Bömmel et al. [28]), in environmetrics (see e.g. Muller and Phillips [22]), in epidemiology for cohort studies (see e.g. Zhou et al. [33]), and in biostatistics for survival analysis (see e.g. Cao et al. [2]).

Two important issues arise from the application of sieve estimation in these research areas: (i) the modeling may require the specification of a non-convex sieve space; (ii) anomalous records (occasionally, we use the word outliers) have a non negligible impact on the estimates. With this regard, we remark that, while it is clear that a suitable and fine tuned (e.g., via judicious selection of the sieve space and of its dimension) sieve estimation can yield accurate inference in the absence of anomalous records, there are no theoretical results on how one can reduce and control the impact of outliers.

We aim at filling these gaps, explaining how to tackle (i) and (ii) by means of robust sieve M-estimation on non-convex sieve spaces. The developed theory

sheds light on the resistance of sieve estimators to outliers and on their rate of convergence.

The paper has the following structure. In Section 2 we motivate our research by the analysis of fMRI data. We also give an overview of our method. The theory is developed in Section 3, Section 4 and Section 5, where we characterize the robustness principle in the context of sieve M-estimation and we derive the rate of convergence of the proposed estimators. Our result complements the asymptotics already available for non-robust sieve M-estimators on non-convex sieve space (see van de Geer [29]) and for general (robust and non-robust) sieve M-estimators on convex sieve space (see van de Geer [30] and Van Der Vaart and Wellner [31]). In Section 6, we illustrate how to apply our method, using both simulated and real fMRI data. In Section 7, we mention some further, possible, theoretical developments and additional applications. All proofs, computational aspects and additional numerical results are available in Appendix.

2. Motivation and preview of our results

2.1. Statistical analysis of fMRI data

We consider data of type $\{y_{t,j}\}$, with $t = 1, ..., T$ and $j = 1, ..., J$, for $T, J \in \mathbb{N}$. The data are collected over time (t) and are observed at changing locations (j). For instance, this kind of data arise in the statistical analysis of fMRI records; see e.g. Lazar [21] for a book-length introduction. The fMRI is a noninvasive method of recording brain signals via a scanner, which measures changes, due to the oxygenation of the hemoglobin, in the magnetic field over several regions (the so-called voxel, namely volumetric pixels) of the brain. Images are typically acquired in axial slices, which are perpendicular to the longitudinal axis of the body. The outcome of this data acquisition procedure is the so-called Blood Oxygen Level Dependent (henceforth, BOLD) signal.

Two main statistical issues characterize the analysis of fMRI data. First, a huge number of time series is stored in large datasets of 4Dimensional (3D in the space and 1D in time) records and becomes the object of the statistical analysis. Traditional multivariate time series models for such datasets would require as many parameters as the number of observations, and as a consequence they are helpless. Thus, novel techniques which deal with the high-dimensionality of the data are needed. Among them, factor models allow for low-dimensional representation of the high-dimensional data; see Park et al. [23] and Hallin and Lippi [15]. Second, the recording of fMRI data is subject to an unavoidable contamination, which is entailed, e.g., by the scanner failures or by the fact that the experimental subjects breathe and move suddenly during the data acquisition process. This contamination may generate anomalous records. One approach commonly applied to deal with these records is the pre-processing: the procedure involves identifying and removing the outliers from the recorded signal. The resulting "cleaned" data are the input for further statistical analysis (e.g. for dimension reduction techniques). We refer to Lazar [21], Chapter 2,

for more details. The drawback of pre-processing is that the "cleaned" data depends on the (subjective) choices made for the identification and the removal of outliers.

Recently, the literature has focused on the problem arising from the analysis of high-dimensional data in the presence of outliers; see e.g. Fan et al. [8], Avella-Medina and Ronchetti [1] and related papers. These inferential procedures rely on regularization techniques (a robust loss function is combined with a penalty) and allow simultaneous estimation and variable selection when the number of observable covariates is large. When the covariates are latent, dimension reduction via factor models (see, among others, Forni et al. [12], Park et al. [23], Hallin and Lippi [15]) represents a powerful toolkit for the statistical analysis of high-dimensional fMRI data. However, the sensitivity of factor models to the presence of anomalous records is only partially explored. Many (if not most) extant procedures define a robust principal component analysis (e.g. using robust estimates of the covariance matrix) to guard against outliers; see, among others, Croux and Haesbroeck [6], Salibián-Barrera et al. [25], Fan and Kim [7], Fan et al. [9] and reference therein. In this paper, we propose a different approach to define a robust dimensionality reduction for fMRI data: we combine the basic principles of infinitesimal robustness of M-estimation (Hampel et al. [16]) with the flexibility of sieve methods on non-convex sieve spaces. Our construction has the same spirit as the approach in Peña and Yohai [24], where generalized dynamic principal components are defined for (possibly robust) dimensionality reduction. However, differently from Peña and Yohai [24], we derive the asymptotic theory, proving consistency of our estimates.

As a potential benefit of our methodology, we mention that the developed inferential procedure may enable clinicians to detect anomalous physiological patterns with increased confidence, avoiding the risks related to data pre-processing and obtaining stable (in the sense of outliers resistant) estimates of the latent factors which determine the brain activity.

2.2. Modeling and estimation

We assume that the BOLD signal $\{y_{t,j}\}$ depends on observable and non random covariates, say $\{\boldsymbol{\xi}_j\}$ with $\boldsymbol{\xi}_j \in [0,1]^d$, and on a latent L-dimensional process $\boldsymbol{\mathcal{Z}}_t = (\mathcal{Z}_{t,1}, ..., \mathcal{Z}_{t,L})^\top$:

$$y_{t,j} = m_0(\boldsymbol{\xi}_j) + \sum_{l=1}^{L} \mathcal{Z}_{t,l} m_l(\boldsymbol{\xi}_j) + \epsilon_{t,j}, \quad t = 1, ..., T \text{ and } \; L \ll J, \tag{2.1}$$

where $m_0, m_1, ..., m_L$ are unknown real-valued functions of the covariates and defined on a subset of $\mathbb{R}^d$. We assume that $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_J$ are fixed (they represent the voxel position) and the errors $\epsilon_{1,1}, ..., \epsilon_{T,J}$ are independent with mean zero, constant variance and they are symmetrically distributed.

The model in (2.1) encodes the temporal dimension of the $\{y_{t,j}\}$ in the factors $(\boldsymbol{\mathcal{Z}}_t)$ and the spatial dimension in the factor loadings $(\boldsymbol{m}(\boldsymbol{\xi}_j))$. Since the number

(L) of factors is much smaller than the number (J) of time series, the model in (2.1) achieves dimension reduction and it is a special case of the generalized dynamic factor model considered by Forni et al. [12]. We refer to Park et al. [23] for a related discussion.

As far as the estimation of the $(L+1)$-tuple of functions $\boldsymbol{m}=(m_0, m_1, ..., m_L)^\top$ is concerned, the extant inferential procedure is based on an orthonormal basis of functions $\phi_1, \phi_2, ..., \phi_K : [0,1]^d \to \mathbb{R}$, where K is chosen as function of J and T; e.g., a popular choice is a tensor B-spline basis with K knots. Then, the $(L+1)$-tuple $\boldsymbol{m}$ is approximated by $\boldsymbol{\mathcal{A}}\boldsymbol{\phi}$ where $\boldsymbol{\mathcal{A}}$ is a $(L+1)\times K$ matrix and $\boldsymbol{\phi}=(\phi_1, ..., \phi_K)^\top$. This characterizes the proposed inferential method as a sieve method; more details are available in Section 6.

The estimators $\hat{\boldsymbol{\mathcal{Z}}}_t$ and $\hat{\boldsymbol{\mathcal{A}}}$ are defined as the solutions to the following problem:

$$\underset{\boldsymbol{\mathcal{A}}, \boldsymbol{z}_1, ..., \boldsymbol{z}_T}{\arg\min} \sum_{j=1}^{J}\sum_{t=1}^{T} \gamma\{y_{t,j} - (1, \boldsymbol{z}_t^\top)\boldsymbol{\mathcal{A}}\boldsymbol{\phi}(\boldsymbol{\xi}_j)\}, \tag{2.2}$$

where $\boldsymbol{z}_t$ are L-dimensional vectors and γ is a loss function. The squared loss function is routinely applied in (2.2): its first order conditions are such that the resulting $\hat{\boldsymbol{\mathcal{A}}}$ and $\hat{\boldsymbol{\mathcal{Z}}}_t$ are the ordinary least squares (OLS) estimates; for implementation detail we refer to Park et al. [23]. Now, it is well known that in the setting of independently and identically distributed observations, the OLS estimates are sensitive to outliers: even a small number of anomalous records can induce a large bias; see e.g. Hampel et al. [16]. In the next numerical examples we illustrate that the same issue is observable for the OLS estimates in model (2.1) for fMRI data. Anticipating some of the results of the next sections, we illustrate how to cope with outliers by means of sieve M-estimates which are a robustified version of the OLS sieve estimates. These estimates are obtained using the Huber loss function; see equations (3.3) and (3.4) below.

2.3. Glancing at the stability of our method

We perform a sensitivity analysis to study the behaviour of the widely-applied OLS and of our robust sieve M-estimators for the model (2.1). The aim of this section is twofold. First, we illustrate the instability of the routinely applied sieve estimators. Second, we glance at the robustness of the robust inferential procedure developed in this paper. Additional and more detailed numerical studies are available in Section 6.

We set $d=2$ and $L=1$, so we have the model $y_{t,j} = m_0(\boldsymbol{\xi}_j) + \mathcal{Z}_{t,j}^{\eta,\nu} m_1(\boldsymbol{\xi}_j) + \epsilon_{t,j}$, where $\mathcal{Z}_{t,j}^{\eta,\nu} = (1-\eta_{t,j})\mathcal{Z}_t + \eta_{t,j}\nu\tilde{\mathcal{Z}}_t$, with $\mathcal{Z}_t = \beta\mathcal{Z}_{t-1} + \sigma u_t$ and $\tilde{\mathcal{Z}}_t = 4/\{1+\exp[-0.1(t-T/2)]\}$. The error terms are Gaussian, namely $\epsilon_{t,j} \sim \mathcal{N}(0,1)$, $u_t \sim \mathcal{N}(0,1)$, and u_t and $\epsilon_{t,j}$ are independent, for any t and j. Moreover, we set $\beta=-0.5$ and $\sigma=1$. The $\{\eta_{t,j}\}$ in $\{\mathcal{Z}_{t,j}^{\eta,\nu}\}$ is a double array of Bernoulli random variables, with parameter η, which take value 1 if the (tj)-th BOLD signal is contaminated and zero otherwise. Thus, the signal $\{y_{t,j}\}$ is randomly contaminated by anomalous values in the latent factor process. $\tilde{\mathcal{Z}}_t$ is a sigmoid

over the time. The motivation behind this setting is to mimic a potential change of the condition of the fMRI experiment. For example the patient changed his position.

In our simulation design, we set two values of η, namely $\eta = 0$ and $\eta = 0.05$. For $\eta = 0$ we have a sample which does not contain contaminated values: we label this kind of sample as "clean sample". In contrast, for $\eta = 0.05$, the simulated BOLD signal contains some outliers and we label it "contaminated sample". The positive scalar ν is equal either to 5 or to 10 and it represents the intensity of the contamination, while $\{\tilde{\mathcal{Z}}_t\}$ are the contaminating values. To be realistic, for the functions m_0 and m_1, we chose two horizontal images of the brain functions, with the aim of mimicking fMRI data (see Section 6 for details). Finally, we set $T = 100$, $J = 64 \times 64$ and we use 2-dimensional B-splines basis functions with $K = 18 \times 18$.

We conduct inference on the latent process $\{\mathcal{Z}_t\}$, on the function $\boldsymbol{m} = (m_0, m_1)^\top$ and on the Euclidean parameter (β, σ). We study the sensitivity to outliers comparing the behaviour of the OLS and of the robust sieve estimates applied on the clean and the contaminated sample. The outcomes are available in Figures 1 and 2. For the sake of visualization, we focus on one slice of the brain.

From Figure 1, we see that the instability of OLS estimated factor increases with ν. In the presence of anomalous records, the OLS estimated factor becomes more persistent (a spurious trend appears) and less volatile. In contrast, the robust estimation method yields a stable inference: the estimates of the latent factor in the clean and in the contaminated sample are virtually indistinguishable.

In Figure 2 we compare the OLS and robust estimates of the function m_0 (similar results are available for m_1), when $\nu = 5$. For the sake of visualisation, in each plot, we display only the higher (above the third quartile) values of $\hat{m}_0$. The plots illustrate that already a moderate contamination can heavily bias the OLS estimates, inducing an artificial activated area (characterized by high estimates of m_0) in the presence of outliers. In contrast, the robust estimate of m_0 (bottom panels) remains stable, providing sensible information on the true activated areas, even in the presence of contamination.

To elaborate further, for each estimation method we compute the relative prediction error

$$\mathrm{PE} = \frac{\sum_{t=1}^{T} \|\hat{m}_0 + \sum_{l=1}^{L} \hat{\mathcal{Z}}_{t,l}\hat{m}_l - m_0 - \sum_{l=1}^{L} \mathcal{Z}_{t,l} m_l\|_J^2}{\sum_{t=1}^{T} \|m_0 + \sum_{l=1}^{L} \mathcal{Z}_{t,l} m_l\|_J^2} \tag{2.3}$$

where we write $\|f\|_J^2 = \sum_{j=1}^{J} f(\boldsymbol{\xi}_j)^2/J$, $\hat{\mathcal{Z}}_t$ is the estimated latent factor at time t and $\hat{\boldsymbol{m}} = (\hat{m}_0, \hat{m}_1)^\top$. For the OLS method, the PE increases exponentially fast, taking values $0.101, 0.145, 0.232$, for $\nu = 0, 5, 10$, respectively. So, we argue that OLS is fitting essentially the ouliers in the sense that it adjusts the estimates to fit the anomalous records. In contrast, for the same values of ν, our robust procedure implies a PE which remains essentially flat, taking values $0.103, 0.131, 0.165$.

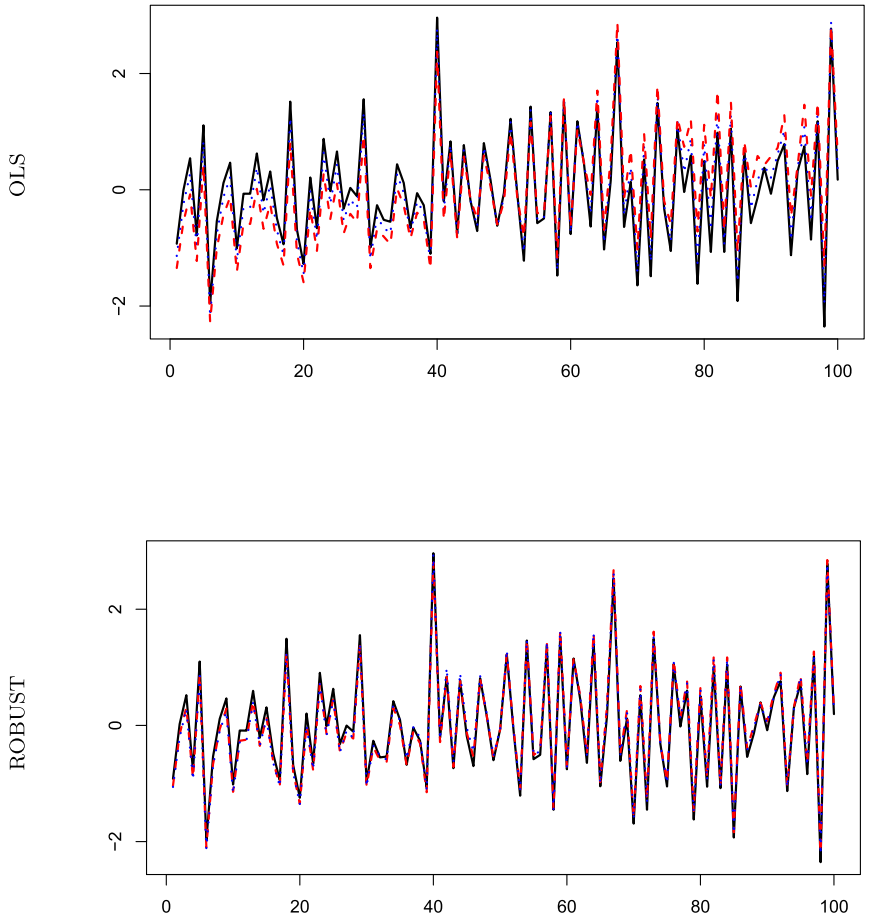

FIG 1. *Estimates of the factors $\mathcal{Z}_t$ under no contamination (continuous line), $\nu = 5$ (dashed line) and $\nu = 10$ (dotted-dashed).*

Once m_0, m_1 and the latent factors are estimated, it is common practice to fit a time series model on them; see e.g. Park et al. [23] for an application in the fMRI context. In this spirit, we investigate if the stability of $\{\hat{\mathcal{Z}}_t\}$ implies also stable maximum likelihood estimates of the parameters of the AR(1) model for the latent factor. For $\eta = 0$, the OLS yields $\hat{\beta} = -0.513$, however, in the presence of contamination ($\eta = 5\%$) the estimates display a large bias: $\hat{\beta} = -0.438$ for $\nu = 5$, and $\hat{\beta} = -0.187$, for $\nu = 10$. Differently, filtering the latent factors via our robust sieve M-estimator yields stable estimates of both β and σ. Indeed, for $\eta = 0$, the robust procedure yields $\hat{\beta} = -0.511$, and, in the presence of contamination the estimates are virtually unchanged: $\hat{\beta} = -0.479$, for $\nu = 5$, and $\hat{\beta} = -0.477$, for $\nu = 10$. We observe a similar pattern for the estimator of σ.

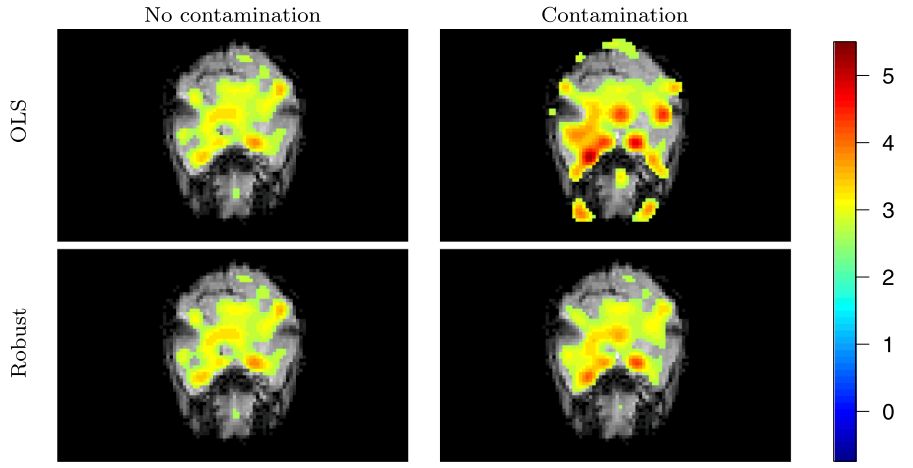

FIG 2. *Plot of the higher values of the estimates of m_0, under no contamination and with contamination ($\nu = 10$), with a percentage $\eta = 5\%$ of contaminated data.*

2.4. The way ahead

The results discussed in Section 2.3 illustrate the benefits that the use of Huber loss function can bring in the statistical analysis of fMRI data and, more generally, in the setting of sieve M-estimation. However, the motivating example raises some questions.

First, the properties of the Huber loss function are well-known when the parameter space is a finite-dimensional Euclidean space; see Hampel et al. [16]. It is not clear if (and how) the Huber loss function yields estimates which are resistant to outliers also in the context of sieve estimation, where the parameter space is infinite dimensional. Second, in model (2.1), the sieve space $\mathcal{G}_S = \{g : \{1, ..., T\} \times [0,1]^d \to \mathbb{R} : \; g(t, \boldsymbol{\xi}) = (1, \boldsymbol{z}_t^\top)\boldsymbol{\mathcal{A}}\boldsymbol{\phi}(\boldsymbol{\xi}), \boldsymbol{z}_t \in \mathbb{R}^L, \boldsymbol{\mathcal{A}} \in \mathbb{R}^{(L+1)\times K}\}$ is non-convex: for any two functions $g_1, g_2 \in \mathcal{G}_S$, the function $g_1 + g_2$ is not necessarily in $\mathcal{G}_S$. To our knowledge, the extant asymptotic results addressing the consistency of sieve M-estimators are either for convex spaces (see Shen and Wong [26] and van de Geer [30]) or for non-convex spaces with non robust estimating function (see van de Geer [29]). So, for the sake of completeness, we should prove that our robust sieve M-estimators on $\mathcal{G}_S$ are consistent. Third, we do not know if (and in which sense) the use of the Huber loss function should be preferred to the use of any other loss function. For instance, one can think of applying the ℓ_1 loss function and obtain a median-type sieve M-estimator, or a loss function defining a redescending M-estimator; see Hampel et al. [16] for several example on loss functions yielding robust M-estimators. In the next sections, we answer all these questions.

3. Setting

3.1. Notation and basic notions

We consider observations that consist in a response variable $y_i \in \mathbb{R}$ as well as fixed covariates $\boldsymbol{x}_i$ which belong to some vector space $\mathcal{X}$. We moreover assume that the pair $(y_i, \boldsymbol{x}_i), i = 1, ..., n$ are drawn from a common distribution such that:

$$y_i = g_0(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, ..., n \tag{3.1}$$

where ϵ_i are independently symmetrically distributed errors with expectation 0 and finite variance. We assume that the covariates are non random and we denote by $Q_n = \sum_{i=1}^n \delta_{\boldsymbol{x}_i}/n$ their empirical marginal distribution, where $\delta_{\boldsymbol{x}}$ denotes the distribution that assigns mass 1 at the point $\boldsymbol{x}$ and 0 elsewhere. We also assume that $g_0 \in \mathcal{G}$, with $\mathcal{G} = \mathcal{B} \times \mathcal{H}$ being the Cartesian product of a finite dimensional space $\mathcal{B}$ and an infinite dimensional space $\mathcal{H}$. In the case where there is no Euclidean parameter, the model is fully non-parametric and $\mathcal{B} = \emptyset$. We define a collection of spaces $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_S$ which approximate $\mathcal{G}$. The sequence is called a sieve if it is dense in the original parameter space: for any $g \in \mathcal{G}$ there exists a projection of g, say $\pi_S g \in \mathcal{G}_S$, such that, for a suitable pseudo-distance d, we have $d(g, \pi_S g) \to 0$, as $S \to \infty$. We refer to Grenander [14, Ch. 8].

Often sieve spaces, are indexed by a growing Euclidian space $\Theta_S \subseteq \mathbb{R}^S$.Then $\mathcal{G}_S = \mathcal{G}(\Theta_S) = \{g_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta_S\}$ where $S \ll n$ and it is allowed to increase with n. For instance, if $\mathcal{G}$ is a set of continuous functions, a standard choice is the set of the linear combinations of orthonormal basis of functions $\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_S)^T$ (e.g., tensor B-spline basis) and the sieve space is $\mathcal{G}_S = \{g_{\boldsymbol{\theta}} = \boldsymbol{\theta}^\top \boldsymbol{\phi} : \boldsymbol{\theta} \in \Theta_S\}$.

To estimate g, we consider the class of sieve estimators, such that the sieve M-estimator $\hat{\boldsymbol{\theta}} \in \Theta_S$ is the:

$$\arg\min_{\boldsymbol{\theta} \in \Theta_S} \sum_{i=1}^n \gamma(y_i - g_{\boldsymbol{\theta}}(\boldsymbol{x}_i)), \tag{3.2}$$

for some loss function γ. We can equivalently set the following:

Definition 1 (Sieve M-estimator)**.** *Given a differentiable loss function $\gamma : \mathbb{R} \to \mathbb{R}$, a sieve M-estimator $\hat{\boldsymbol{\theta}}$ is the solution to the system of estimating equations:*

$$\sum_{i=1}^n \boldsymbol{\Psi}(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \mathbf{0}, \tag{3.3}$$

where the estimating function $\boldsymbol{\Psi} : \mathbb{R} \times \mathcal{X} \times \Theta_S \to \mathbb{R}^S$. Denoting $\psi(u) = \partial_u \gamma(u)$ the derivative of γ and $\nabla_{\boldsymbol{\theta}} g_{\boldsymbol{\theta}}(x)$ the gradient of $g_{\boldsymbol{\theta}}(x) : \Theta_S \to \mathbb{R}$ with respect to $\boldsymbol{\theta}$ it yields that

$$\boldsymbol{\Psi}(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \psi(y_i - g_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) \nabla_{\boldsymbol{\theta}} g_{\boldsymbol{\theta}}(\boldsymbol{x}_i).$$

Many kinds of loss functions can be considered. Typically, the square function is applied, so $\psi(u) = u$, which defines the OLS sieve M-estimator. To achieve resistance to outliers, in Section 4.2, we propose (and justify theoretically) to use

$$\psi^{(c)}(u) = \max(\min(u, c), -c). \tag{3.4}$$

in (3.3). Although the proposed estimating function is standard in parametric robust regression estimation (see Hampel et al. [16]), its use for sieve M-estimation on (possibly non-convex) sieve spaces requires additional theoretical investigation; see Section 5.

Finally, we notice that we can cast model (2.1) in the shape as in (3.1) where $\boldsymbol{\theta}^\top = (\boldsymbol{z}_1^\top, ..., \boldsymbol{z}_T^\top, \boldsymbol{\alpha}^\top)$, with $\boldsymbol{\alpha}$ being the stack form of $\boldsymbol{\mathcal{A}}$. Note that (2.2) defines a sieve estimator as in equation (3.2), where we replace the unique sum (in n) with a double sum (one in the time dimension and the other in the cross section).

3.2. Sketch of the theoretical construction

In the usual sieve estimation, one makes the assumption that the actual distribution of the data $(y, \boldsymbol{x})$ coincides with a reference distribution P^0, such that $E_{P^0}(y|\boldsymbol{x}) = g_0(\boldsymbol{x})$, for $g_0 \in \mathcal{G}$ and $\mathrm{Var}_{P^0}(\epsilon|\boldsymbol{x}) \leq C$. We relax this assumption and derive a class of sieve estimators which remain stable even when the reference distribution captures only the behavior of the majority of the data, while some outliers have a different (unspecified) distribution.

To formalize these ideas, let $\mathcal{M}$ denote the space of all probability measures G for $(y, \boldsymbol{x})$, such that the marginal for the covariates coincides with Q_n. For any $P^0 \in \mathcal{M}$, we define $\mathcal{V}_\eta(P^0)$ as the set of all distributions P such that $d_k(P, P^0) < \eta$, where d_k is the Kolmogorov distance. We label as P^{cont} the actual distribution of $(y, \boldsymbol{x})$ and we assume that it does not coincide with P^0, rather it belongs to $\mathcal{V}_{\eta_0}(P^0)$, for $\eta_0 \in [0, 1)$. Then, for each sieve space $\mathcal{G}_S$ with S fixed, we let $\hat{g}_n$ (or $\hat{g}_n^{\mathrm{cont}}$) denote the sieve estimation of g_0 under P^0 (or P^{cont}). Within this setting, our aim is the definition of sieve M-estimators which are stable when $P^{\mathrm{cont}} \in \mathcal{V}_{\eta_0}(P^0)$. This prevents large changes of the estimators in the presence of local departures from P^0.

To proceed further, let us call $\mathcal{L}_2(Q_n)$-norm the norm $\|\cdot\|_n$ defined as $\|g\|_n^2 = \sum_{i=1}^n g(\boldsymbol{x}_i)^2/n$ for any function $g : \mathcal{X} \to \mathbb{R}$. The triangle inequality yields

$$\|\hat{g}_n^{\mathrm{cont}} - g_0\|_n \leq \|\hat{g}_n^{\mathrm{cont}} - \hat{g}_n\|_n + \|\hat{g}_n - g_0\|_n. \tag{3.5}$$

The term $\|\hat{g}_n^{\mathrm{cont}} - \hat{g}_n\|_n$ in (3.5) expresses the changes in the estimates of g due to the deviation of P^{cont} from the reference distribution P^0: it is based on a fixed sieve space, with a fixed S. The term $\|\hat{g}_n - g_0\|_n$ is related to the behaviour of the sieve space as S diverges and it is obtained under the reference model P^0. In Section 4, we show how to bound $\|\hat{g}_n^{\mathrm{cont}} - \hat{g}_n\|_n$ by the theory of robust M-estimators. Then, in Section 5, we bound $\|\hat{g}_n - g_0\|_n$ developing the asymptotics for robust sieve M-estimators on non-convex spaces.

4. Robust inference

4.1. The sieve reference distribution

For fixed S, we see a sieve M-estimator $\hat{\boldsymbol{\theta}}$ as a M-functional of some distribution P:

$$\hat{\boldsymbol{\theta}}(P) : \int \boldsymbol{\Psi}(y, \boldsymbol{x}, \hat{\boldsymbol{\theta}}(P))dP = \mathbf{0}. \tag{4.1}$$

We define the random measure $P_n^0 = n^{-1}\sum_{i=1}^n \delta_{(y_i,\boldsymbol{x}_i)}$ where $(y_i, \boldsymbol{x}_i) \sim P^0$, as well as the contaminated random measure $P_n^{\text{cont}} = n^{-1}\sum_{i=1}^n \delta_{(y_i,\boldsymbol{x}_i)}$ where $(y_i, \boldsymbol{x}_i) \sim P^{\text{cont}}$. Then by Glivenko-Cantelli lemma, $P_n^0 \Rightarrow P^0$ and $P_n^{\text{cont}} \Rightarrow P^{\text{cont}}$ weakly with $d_k(P_n^0, P^0) = O_P(n^{-1/2})$ and $d_k(P_n^{\text{cont}}, P^{\text{cont}}) = O_P(n^{-1/2})$ (Dvoretzky-Kiefer-Wolfowitz inequality).

The sieve M-estimator defined in (3.3) is an M-functional (see, e.g., Hampel et al. [16] and van de Geer [29] among others) of the empirical distribution function P_n: $\hat{\boldsymbol{\theta}}(P_n)$ is the solution to $\int \boldsymbol{\Psi}(y, \boldsymbol{x}, \boldsymbol{\theta})dP_n = \mathbf{0}$.

We assume that the following Lipschitz condition holds

$$\|\hat{g}_n^{\text{cont}} - \hat{g}_n\|_n \leq C_g \|\hat{\boldsymbol{\theta}}(P_n^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P_n^0)\| \tag{4.2}$$

for some constant $C_g \in \mathbb{R}^+$. Then we are going to prove that a bound on the Euclidean norm $\|\hat{\boldsymbol{\theta}}(P_n^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P_n^0)\|$ implies that $\|\hat{g}_n^{\text{cont}} - \hat{g}_n\|_n$ remains bounded. Thanks to this property, the outliers cannot induce large changes in the sieve M-estimates of g_0.

To control $\|\hat{\boldsymbol{\theta}}(P_n^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P_n^0)\| = \|\hat{\boldsymbol{\theta}}(P^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P^0)\| + O_P(n^{-1/2})$ in (4.2), we introduce the sieve reference distribution $P^{\boldsymbol{\theta}^*}$. This represents a contrived device which we apply, in tandem with the triangle inequality, to write

$$\|\hat{\boldsymbol{\theta}}(P^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P^0)\| \leq \|\hat{\boldsymbol{\theta}}(P^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P^{\boldsymbol{\theta}^*})\| + \|\hat{\boldsymbol{\theta}}(P^{\boldsymbol{\theta}^*}) - \hat{\boldsymbol{\theta}}(P^0)\|. \tag{4.3}$$

The latter expression is the stepping stone of our definition of robustness. Indeed, we are going to show that both terms on the right side of (4.3) can be approximated by the influence function of the sieve M-estimator. Thus, a bounded influence function will characterize the robustness of the sieve estimators over a neighborhood of the sieve reference distribution.

We now introduce the sieve reference distribution. To begin with, we define a collection of semi-parametric probability models

$$\mathcal{P}_S = \{P^{\boldsymbol{\theta}} \in \mathcal{M} : P^{\boldsymbol{\theta}}(y|\boldsymbol{x}) = P^0(y - g_{\boldsymbol{\theta}}(\boldsymbol{x}) + g_0(\boldsymbol{x})|\boldsymbol{x}), \boldsymbol{\theta} \in \Theta_S\}.$$

Any $P^{\boldsymbol{\theta}} \in \mathcal{P}_S$ is such that $E_{P^{\boldsymbol{\theta}}}(y|\boldsymbol{x}) = g_{\boldsymbol{\theta}}(\boldsymbol{x})$ and, $\text{Var}_{P^{\boldsymbol{\theta}}}(y|\boldsymbol{x}) \leq C$. Also we remark that $P^{\boldsymbol{\theta}}$ is symmetric and continuous, moreover $\hat{\boldsymbol{\theta}}$ is Fisher consistent over $\mathcal{P}_S$: $\hat{\boldsymbol{\theta}}(P^{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ for $P^{\boldsymbol{\theta}} \in \mathcal{P}_S$ and $d_k(P^{\boldsymbol{\theta}}, P^0) \leq \|g_{\boldsymbol{\theta}} - g_0\|_n$. Indeed, one can see that $d_k(P^{\boldsymbol{\theta}}(y|\boldsymbol{x}), P^0(y|\boldsymbol{x})) \leq |g_{\boldsymbol{\theta}}(\boldsymbol{x}) - g_0(\boldsymbol{x})|$, so $d_k(P^{\boldsymbol{\theta}}, P^0) \leq \int |g_{\boldsymbol{\theta}}(\boldsymbol{x}) - g_0(\boldsymbol{x})|dQ_n \leq \|g_{\boldsymbol{\theta}} - g_0\|_n$.

Definition 2 (Sieve reference distribution)**.** *We call sieve reference distribution the distribution $P^{\boldsymbol{\theta}^*}$ belonging to $\mathcal{P}_S$ where $\boldsymbol{\theta}^* = \arg\inf_{\boldsymbol{\theta}\in\Theta_S} \|g_{\boldsymbol{\theta}} - g_0\|_n$.*

We notice that, by Definition 2, $P^{\boldsymbol{\theta}^*}$ minimizes the Kolmogorov distance among all elements in $\mathcal{P}_S$. This yields that $d_k(P^{\boldsymbol{\theta}^*}, P^0) \leq \Delta_n$ where we define $\Delta_n = \|g_{\theta^*} - g_0\|_n$ and thus $P^0 \in \mathcal{V}_{\Delta_n}(P^{\boldsymbol{\theta}^*})$. Since $P^{\text{cont}} \in \mathcal{V}_{\eta_0}(P^0)$ we have $P^{\text{cont}} \in \mathcal{V}_{\eta_0+\Delta_n}(P^{\boldsymbol{\theta}^*})$. Therefore, we have that P^{cont} and P^0 are both in a neighborhood of $P^{\boldsymbol{\theta}^*}$. Then we control the behavior of the functional $\hat{\boldsymbol{\theta}}$ by bounding its changes over the neighborhood. Following Hampel et al. [16], in the next section we show how to achieve this goal by bounding the influence function of $\hat{\boldsymbol{\theta}}$ at $P^{\boldsymbol{\theta}^*}$.

4.2. Influence function for sieve M-estimators

The sieve reference distribution is an helpful device in the characterization of the robustness of sieve M-estimators. To see this, let $D[0,1]$ be the space of empirical distribution functions on $[0,1]$, that is the space of right continuous real valued functions on $[0,1]$ which have left hand limits. We define $\epsilon_i(\theta) = y_i - g_\theta(x_i)$, for $\theta \in \Theta_S$. Then, for any $P^\theta \in \mathcal{P}_S$, we have $E_{P^\theta}[\epsilon_i(\theta)] = 0$, $\text{Var}_{P^\theta}[\epsilon_i(\theta)] \leq C < \infty$ and $\{\epsilon_i(\theta)\}$ are i.i.d. random variables (r.v.). Let us define $P^{\boldsymbol{\theta}^*}[\epsilon_1(\boldsymbol{\theta}^*)], ..., P^{\boldsymbol{\theta}^*}[\epsilon_n(\boldsymbol{\theta}^*)]$ as i.i.d. r.v. with distribution function (d.f.) uniform (U) on the unit interval $[0,1]$. U_n is the empirical d.f. corresponding to $P^{\boldsymbol{\theta}^*}[\epsilon_1(\boldsymbol{\theta}^*)], ..., P^{\boldsymbol{\theta}^*}[\epsilon_n(\boldsymbol{\theta}^*)]$, it follows that $P_n = U_n \circ P^{\boldsymbol{\theta}^*}$. We define the induced functional $\boldsymbol{\tau} : D[0,1] \to \mathbb{R}^S$ as $\boldsymbol{\tau}(U_n) = \hat{\boldsymbol{\theta}}(U_n \circ P^{\boldsymbol{\theta}^*}) = \hat{\boldsymbol{\theta}}(P_n)$.

We label by $\mathcal{U}$ the class of d.f. on $[0,1]$. Then, for any $F \in \mathcal{U}$, we have $\boldsymbol{\tau}(F) = \hat{\boldsymbol{\theta}}(F \circ P^{\boldsymbol{\theta}^*})$, provided that the latter is well-defined. The statistical functional $\hat{\boldsymbol{\theta}}$ induces a functional $\boldsymbol{\tau}$ on the space of d.f.'s with mass concentrated on $[0,1]$. For this reason, we focus on d.f.'s concentrated on [0,1] and view them as elements of $D[0,1]$. We refer to Fernholz [11] for a book-length presentation.

To study the behavior of the M-functional $\hat{\boldsymbol{\theta}}$ in the presence of local departures from the reference model $P^{\boldsymbol{\theta}^*}$, we consider the first-order von Mises expansion of the induced functional $\boldsymbol{\tau}$. Thus, for $U \in \mathcal{U}$, we set $\boldsymbol{\tau}(U) = \hat{\boldsymbol{\theta}}(U \circ P^{\boldsymbol{\theta}^*}) = \hat{\boldsymbol{\theta}}(P^{\boldsymbol{\theta}^*})$ and using the first order von Mises expansion we have:

$$\boldsymbol{\tau}(U + tH) - \boldsymbol{\tau}(U) = \boldsymbol{\tau}'_U(H) + \mathbf{Rem}(tH), \tag{4.4}$$

where $H \in \mathcal{U}$ and $t \in \mathbb{R}$. Let $\mathcal{S}$ be a collection of subsets of $\mathcal{U}$. We recall, see Fernholz [11, page 16], that $\boldsymbol{\tau}$ is $\mathcal{S}$-differentiable if, for $t \to 0$, there exists $\boldsymbol{\tau}'_U(H)$ such that $t^{-1}\ \mathbf{Rem}(tH) \to 0$ uniformly in $H \in \mathcal{H}$, for all $\mathcal{H} \in \mathcal{S}$. The linear function $\boldsymbol{\tau}'_U$ is called the $\mathcal{S}$-derivative of $\boldsymbol{\tau}$ at U. So, we state the following

Definition 3 (Sieve Influence Function)**.** *Let $P \in \mathcal{M}$, $\hat{\boldsymbol{\theta}}$ be a sieve M-estimator as in Definition 1 and let $\boldsymbol{\tau}$ its induced functional. We call sieve influence function (IF^s) the $\mathcal{S}$-derivative (provided that there exists) $\boldsymbol{\tau}'_U$ evaluated at $(P - P^{\boldsymbol{\theta}^*}) \circ (P^{\boldsymbol{\theta}^*})^{-1} \in \mathcal{U}$, that is:*

$$\boldsymbol{IF}^s(P; P^{\boldsymbol{\theta}^*}, \hat{\boldsymbol{\theta}}) = \boldsymbol{\tau}'_U[(P - P^{\boldsymbol{\theta}^*}) \circ (P^{\boldsymbol{\theta}^*})^{-1}]. \tag{4.5}$$

The expression in (4.5) follows from the usual approach for statistical functionals on $D[0,1]$, see e.g. Fernholz [11, page 39], keeping in mind that $P^{\boldsymbol{\theta}^*}$ is an

element of $\mathcal{P}_S$, indexed by $\boldsymbol{\theta}^* \in \Theta_S$, where the Euclidean parameter is related to the sieve space $\mathcal{G}_S$.

Eq. (4.4) implies that the $\boldsymbol{IF}^s$ provides an approximation to the changes in $\boldsymbol{\tau}$ due to perturbation of $P^{\boldsymbol{\theta}^*}$:

$$\begin{aligned}\boldsymbol{\tau}(U + t[(P - P^{\boldsymbol{\theta}^*}) \circ (P^{\boldsymbol{\theta}^*})^{-1}]) - \boldsymbol{\tau}(U) &\simeq \boldsymbol{\tau}'_U[(P - P^{\boldsymbol{\theta}^*}) \circ (P^{\boldsymbol{\theta}^*})^{-1}] \\ &\simeq \boldsymbol{IF}^s(P; P^{\boldsymbol{\theta}^*}, \hat{\boldsymbol{\theta}}) \simeq \hat{\boldsymbol{\theta}}(P) - \hat{\boldsymbol{\theta}}(P^{\boldsymbol{\theta}^*}),\end{aligned} \tag{4.6}$$

where $\simeq$ has to be interpreted up to a remainder as in (4.4) (see Hampel et al. [16] for a book-length discussion).

Given a sample of size n, a standard calculation (see Fernholz [11], page 40), yields $\boldsymbol{IF}^s(P_n, P^{\boldsymbol{\theta}^*}, \hat{\boldsymbol{\theta}}) = n^{-1} \sum_{i=1}^n \boldsymbol{IF}^s(\delta_{(y_i, \boldsymbol{x}_i)}, P^{\boldsymbol{\theta}^*}, \hat{\boldsymbol{\theta}})$, which motivates the following

Definition 4 (Robustness Principle)**.** *A sieve M-estimator $\hat{\boldsymbol{\theta}}$ defined as in (3.3) is robust if it induces a functional on $D[0,1]$ which has a bounded $\boldsymbol{IF}^s$, namely if:*

$$\sup_{y \in \mathbb{R}} \|\boldsymbol{IF}^s(\delta_{(y,\boldsymbol{x})}, P^{\boldsymbol{\theta}^*}, \hat{\boldsymbol{\theta}})\| \leq C^*, \tag{4.7}$$

for all $\boldsymbol{x} \in \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. Since we have (see, e.g., Hampel et al. [16])

$$\boldsymbol{IF}^s(\delta_{(y,\boldsymbol{x})}; P^{\boldsymbol{\theta}^*}, \hat{\boldsymbol{\theta}}) = \left\{ - \int \nabla_{\boldsymbol{\theta}} \boldsymbol{\Psi}(y, \boldsymbol{x}, \boldsymbol{\theta}^*) dP^{\boldsymbol{\theta}^*} \right\}^{-1} \boldsymbol{\Psi}(y, \boldsymbol{x}, \boldsymbol{\theta}^*), \tag{4.8}$$

a robust sieve M-estimator is obtained using a bounded, in the Euclidean norm, $\boldsymbol{\Psi}(y, \boldsymbol{x}, \boldsymbol{\theta})$.

Robustness of the sieve estimator $\hat{\boldsymbol{\theta}}$ implies its Hadamard differentiability ($\mathcal{S}$ is thus the set of all compact subsets of $\mathcal{U}$), yielding that for any $P \in \mathcal{V}_\eta(P^{\boldsymbol{\theta}^*})$ we have that $\|\hat{\boldsymbol{\theta}}(P) - \hat{\boldsymbol{\theta}}(P^{\boldsymbol{\theta}^*})\| \leq C^* \eta + o(\eta) = O(\eta)$, where $C^* > 0$. Since $P^0 \in \mathcal{V}_{\Delta_n}(P^{\boldsymbol{\theta}^*})$ and $P^{\text{cont}} \in \mathcal{V}_{\eta_0 + \Delta_n}(P^{\boldsymbol{\theta}^*})$, we get $\|\hat{\boldsymbol{\theta}}(P^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P^0)\| \leq \|\hat{\boldsymbol{\theta}}(P^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P^{\boldsymbol{\theta}^*})\| + \|\hat{\boldsymbol{\theta}}(P^{\boldsymbol{\theta}^*}) - \hat{\boldsymbol{\theta}}(P^0)\| = O(\eta_0 + \Delta_n)$, thus for $n^{-1/2} < \Delta_n$ we obtain

$$\|\hat{\boldsymbol{\theta}}(P_n^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P_n^0)\| \leq O_P(\eta_0 + \Delta_n). \tag{4.9}$$

In the next section, we explain how this result is useful to derive the rate of convergence of $g_{\hat{\boldsymbol{\theta}}(P_n^{\text{cont}})}$ to g_0, as a function of the radius of contamination η_0.

5. Asymptotics

To our knowledge, asymptotic results for sieve M-estimators are available for convex sieve spaces and for general loss functions. For non-convex spaces, the well-known asymptotic results are available for the least squares (OLS) sieve M-estimator (see van de Geer [29], Park et al. [23] and reference therein). We could not find theoretical results which guarantee the consistency of sieve M-estimators for generic loss functions (like e.g. the Huber loss function proposed in

this paper) and for non-convex sieve spaces. Thus, for the sake of completeness, we prove the consistency of sieve M-estimators for a large class of convex loss functions, when $\mathcal{G}$ and $\mathcal{G}_S$ are allowed to be non-convex.

We build on van de Geer [29] and, for simplicity of notation, we remove the $\boldsymbol{\theta}$ index. To achieve consistency, the dimension of the sieve space S has to be an increasing function of n. To emphasize the dependence on n and simplify the notation, we write $S_n = S$ and denote by $\mathcal{G}_n$ (rather than by $\mathcal{G}_{S_n}$) the sieve approximation to $\mathcal{G}$. In what follows, all the expected values $E(\cdot)$ are taken with respect to P^0. Moreover, we recall that $\hat{g}_n = g_{\hat{\boldsymbol{\theta}}(P_n^0)}$ and we use the following notation,

$$\int \gamma_g dP_n^0 = n^{-1} \sum_{i=1}^n \gamma\{y_i - g(\boldsymbol{x}_i)\}, \quad \int \gamma_g dP^0 = n^{-1} \sum_{i=1}^n E[\gamma\{y_i - g(\boldsymbol{x}_i)\}]. \tag{5.1}$$

Therefore we can write the solutions g_0 and $\hat{g}_n$ as,

$$g_0 = \arg\min_{g \in \mathcal{G}} \int \gamma_g dP^0, \quad \hat{g}_n = \arg\min_{g \in \mathcal{G}_n} \int \gamma_g dP_n^0. \tag{5.2}$$

Recall that $\|\cdot\|_n$ denotes the $\mathcal{L}_2(Q_n)$ norm $\|g\|_n = \{\sum_{i=1}^n g(\boldsymbol{x}_i)^2/n\}^{1/2}$ for some $g \in \mathcal{G}$. The solution in the sieve space is given by:

$$g_n^* = \arg\inf_{g \in \mathcal{G}_n} \|g - g_0\|_n, \tag{5.3}$$

noting that $g_n^* = g_{\boldsymbol{\theta}^*}$. Thus, g_n^* is the projection of g_0 on the sieve space. We define the sieve error (or approximation error) by

$$\Delta_n := \|g_n^* - g_0\|_n \tag{5.4}$$

and $\|\hat{g}_n - g_n^*\|_n$ is the estimation error.

Following van de Geer [29], we consider the space $\mathcal{F}_n$ and $\hat{f}_n \in \mathcal{F}_n$:

$$\mathcal{F}_n = \left\{ \frac{g - g_n^*}{1 + b_n \|g - g_n^*\|_n} : g \in \mathcal{G}_n \right\}, \quad \hat{f}_n = \frac{\hat{g}_n - g_n^*}{1 + b_n \|\hat{g}_n - g_n^*\|_n}. \tag{5.5}$$

where $\{b_n\}$ is a strictly positive sequence. This yields $\|f\|_n \leq b_n^{-1}$ for all $f \in \mathcal{F}_n$. Our theoretical developments rely on the following assumptions.

Condition 1. *For some $\kappa_1 > 0$ and for all $|u| \leq \kappa_1$, $E[\gamma(\epsilon_i + u) - \gamma(\epsilon_i)] \geq \kappa_1 u^2$.*

Condition 2. *For some constant C_3 and κ_2 and for all $|u| \leq C_3$, $E[\gamma(\epsilon_i + u) - \gamma(\epsilon_i)] \leq \kappa_2 u^2$.*

Condition 3. *$E[\gamma(\epsilon_i - u)]$ has a unique minimum at $u = 0$.*

Condition 4. *There exists a constant C_2 such that $\sup_{f \in \mathcal{F}_n} |f|_\infty \leq C_2 < \infty$ where $|\cdot|_\infty$ denotes the supremum norm.*

Condition 5. $|\gamma\{\epsilon_i - f(\boldsymbol{x}_i)\} - \gamma\{\epsilon_i - \tilde{f}(\boldsymbol{x}_i)\}| \leq |V_i||f(\boldsymbol{x}_i) - \tilde{f}(\boldsymbol{x}_i)|, \forall f, \tilde{f} \in \mathcal{F}_n$ *where* V_i *are uniformly sub-Gaussian, that is*

$$\max_{i=1,\dots,n} C_1^2 E\left(e^{|V_i|^2/\kappa_3} - 1\right) \leq \sigma_0^2.$$

Conditions 1 and 2 ensure that the loss function behaves locally quadratically around 0 and are always satisfied by the Huber loss function, defined by:

$$\gamma(u) = \begin{cases} u^2, \text{ if } |u| \leq c \\ 2c|u| - c^2, \text{ if } |u| > c \end{cases}$$

Condition 2 is not met by the estimating function of the median (see Appendix B). Condition 3 is always satisfied by the Huber loss function whereas for the least absolute deviation loss we have to assume that there exists a unique median at 0. In Condition 5, if γ is Lipschitz continuous (as it is the case of the Huber or the absolute value loss function), the V_i's can be constants and there is no need for an assumption on the tail weight of the error distribution.

To prove consistency we need two fundamental ingredients: a basic inequality and an entropy bound; we refer to van de Geer [29] for a book-length explanation. The following lemma states the basic inequality.

Lemma 1 (Basic inequality)**.** *Assume Conditions 1-4 hold. Then for some constant* $\lambda > 0$ *depending on* C_2, C_3 *and* κ_1,

$$\|\hat{f}_n\|_n^2 \leq -\lambda^{-1} \int (\gamma_{g_n^* + \hat{f}_n} - \gamma_{g_n^*}) d(P_n^0 - P^0) + 2\Delta_n b_n^{-1} + \left(\kappa_2 \lambda^{-1} - 1\right) \Delta_n, \quad (5.6)$$

where g_0 *and* $\hat{g}_n$ *are defined as in* (5.2), g_n^* *in* (5.3), $\hat{f}_n$ *in* (5.5) *and* Δ_n *as in* (5.4).

The first term (the integral) in (5.6) is standard in the theory of M-estimation, whereas the additional terms $2\Delta_n b_n^{-1} + \left(\kappa_2 \lambda^{-1} - 1\right) \Delta_n$ represent the bias due to the sieve. This term cannot be ignored: the sieve space is not assumed to be convex. The next theorem shows that under suitable conditions on $\{b_n\}$, consistency can be achieved with the same rate as the OLS sieve M-estimator.

Let $\mathcal{F}_n(\delta) := \{f \in \mathcal{F}_n : \|f\|_n \leq \delta\}$ and $H(\delta, \mathcal{F}_n, Q_n)$ denotes the δ-entropy of $\mathcal{F}_n$ with respect to the $\mathcal{L}_2(Q_n)$-metric. Then, we state the following

Theorem 1. *Suppose that Conditions 1-5 hold. For an appropriate constant* c_1, *assume that there exists a function* $\Omega(\delta) \geq \max[\int_{\delta^2/c_1}^{\delta} H^{1/2}\{u, \mathcal{F}_n(\delta), Q_n\} du, \delta]$ *such that* $\Omega(\delta)/\delta^2$ *is a non-decreasing function of* δ, *where* $0 < \delta < c_1$. *Then for a constant* c_2 *depending on* λ, C_1, C_2 *and* σ_0, *and for a sequence* $\{\delta_n\}$ *such that* $\sqrt{n}\delta_n^2 \geq c_2 \Omega(\delta_n)$, *by taking* $b_n^{-1} = \max(2\delta_n, \|g_n^* - g_0\|_n)$ *we obtain the following rate of convergence:*

$$\|\hat{g}_n - g_0\|_n = O_P(\delta_n + \Delta_n). \quad (5.7)$$

In the literature on sieve estimation, Δ_n decreases as $S_n \to \infty$ and δ_n typically is a function of both n and S_n. The exact specifications of Δ_n and δ_n

depend on the space $\mathcal{G}$ and on its sieve approximation $\mathcal{G}_n$. We refer to van de Geer [29], Chapter 10 (p. 185), for a book-length discussion.

The theorem applies to sieve M-estimator obtained using the Huber loss function and it implies that the resulting robust sieve M-estimator has the same rates of convergence as for the OLS sieve M-estimator (see van de Geer [29]). We emphasize that, if the Huber loss function is applied, the results in Theorem 1 is valid under weaker assumption: thanks to the boundedness of the estimating function, there is no need for assuming a sub-Gaussian distribution for the errors (see Condition 5).

In Section 4 we have seen that using the Huber functions leads to a bound on $\|\hat{\boldsymbol{\theta}}(P^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P^0)\|$. Since $\|\hat{\boldsymbol{\theta}}(P_n^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P_n^0) = \|\hat{\boldsymbol{\theta}}(P^{\text{cont}}) - \hat{\boldsymbol{\theta}}(P^0)\| + O_P(n^{-1/2})$ and $n^{-1/2} \leq \delta_n$ for sufficiently large n, we conclude that $\|\hat{g}_n^{\text{cont}} - g_0\|_n = O_p(\eta_0 + \delta_n + \Delta_n)$.

6. Application to dimensionality reduction

6.1. Dynamic semiparametric factor model

Let us come back to the semiparametric model (2.1) and let us see how Theorem 1 applies to the model.

The objects of our inference are the latent process $(\mathcal{Z}_{t,1}, ..., \mathcal{Z}_{t,L})$ and the $(L+1)$-tuple $(m_0, ..., m_L)$ of unknown real-valued functions. To estimate them, Park et al. [23] apply a sieve M-estimator, as defined using the least squares loss function. Formula (4.8) implies that the sieve influence function of the resulting M-estimator is unbounded, thus the estimator is not robust in the sense of Definition 4. To achieve robustness of the estimated (space spanned by the) factors and of the estimates of $\boldsymbol{m}$, we propose to replace the OLS estimator by its robust counterpart. The latter is obtained replacing the least squares loss function with the Huber loss function. We refer to Appendix B for the derivation of its sieve influence function and for a discussion about the problem of uniqueness of the estimated latent factors—which are subject to the usual sign indeterminacy.

The following corollary makes use of Theorem 1 to show consistency of the robust sieve M-estimator. To state our next result, we need the same assumptions in Park et al. [23]. We state them:

Assumption A1. *The variables* $\varepsilon_{1,1}, \ldots, \varepsilon_{T,J}$, *and* $\boldsymbol{\mathcal{Z}}_1, \ldots, \boldsymbol{\mathcal{Z}}_T$ *are independent. The process* $\{\boldsymbol{\mathcal{Z}}_t\}$ *is allowed to be nonrandom.*

Assumption A2. *We assume that* $E\varepsilon_{t,j} = 0$ *for* $1 \leq t \leq T, 1 \leq j \leq J$, *and for* $c > 0$ *small enough* $\sup_{1\leq t\leq T, 1\leq j\leq J} E \exp\left(c\varepsilon_{t,j}^2\right) < \infty$.

Assumption A3. *The functions* ϕ_k *may depend on the increasing indices* T *and* J, *but are normed so that* $\int_{[0,1]^d} \phi_k^2(\xi) d\xi = 1$ *for* $k = 1, \ldots, K$. *Furthermore, it holds that* $\sup_{\xi\in[0,1]^d} \|\phi(\xi)\| = O\left(K^{1/2}\right)$.

Assumption A4. *The vector of functions* $\boldsymbol{m}$ *can be approximated by* ϕ_k*, i.e.,*

$$\delta_K := \sup_{\boldsymbol{\xi}\in[0,1]^d} \inf_{\boldsymbol{\mathcal{A}}\in\mathbb{R}^{(L+1)\times K}} \|\boldsymbol{m}(\boldsymbol{\xi}) - \boldsymbol{\mathcal{A}}\boldsymbol{\phi}(\boldsymbol{\xi})\| \to 0$$

as $K \to \infty$*. We denote by* $\boldsymbol{\mathcal{A}}^*$ *the* $\boldsymbol{\mathcal{A}}$ *that fulfills* $\sup_{\boldsymbol{\xi}\in[0,1]^d} \|\boldsymbol{m}(\boldsymbol{\xi}) - \boldsymbol{\mathcal{A}}\boldsymbol{\phi}(\boldsymbol{\xi})\| \leq 2\delta_K$*.*

Assumption A5. *There exist constants* $0 < C_L < C_U < \infty$ *such that all eigenvalues of the matrix* $T^{-1}\sum_{t=1}^T \mathcal{Z}_t\mathcal{Z}_t^\top$ *lie in the interval* $[C_L, C_U]$ *with probability tending to one.*

Assumption A6. *The minimization in (2.2) runs over all values of* $(\boldsymbol{z}_t, \boldsymbol{\mathcal{A}})$ *with*

$$\sup_{\xi\in[0,1]^d} \max_{1\leq t\leq T} \|(1, \boldsymbol{z}_t^\top)\boldsymbol{\mathcal{A}}\boldsymbol{\phi}(\xi)\| \leq M_T$$

where the constant M_T *fulfils (with probability tending to one)* $\max_{1\leq t\leq T}\|\boldsymbol{\mathcal{Z}}_t\| \leq M_T/C_m$*, for a constant* C_m *such that* $\sup_{\xi\in[0,1]^d}\|\boldsymbol{m}(\xi)\| < C_m$*.*

Assumption A7. *It holds that* $\kappa^2 = (K+T)M_T^2\log(JTM_T)/(JT) \to 0$*. The dimension* L *is fixed.*

Corollary 1. *Suppose that model (2.1) holds and that* $(\hat{\boldsymbol{\mathcal{Z}}}_t, \hat{\boldsymbol{\mathcal{A}}})$ *is defined by the minimization problem (2.2), where the loss function* $\gamma(\cdot)$ *satisfies Condition 1-4 above. Under Assumptions (A1)-(A8),*

$$T^{-1}\sum_{t=1}^T \|(1, \hat{\boldsymbol{\mathcal{Z}}}_t^T)\hat{\boldsymbol{\mathcal{A}}} - (1, \boldsymbol{\mathcal{Z}}_t^T)\boldsymbol{\mathcal{A}}^*\|^2 = O_p(\kappa^2 + \delta_K^2), \tag{6.1}$$

where

$$\delta_K \equiv \sup_{\boldsymbol{\xi}\in[0,1]^d} \inf_{\boldsymbol{\mathcal{A}}\in\mathbb{R}^{(L+1)\times K}} \|m(\boldsymbol{\xi}) - \boldsymbol{\mathcal{A}}\boldsymbol{\phi}(\boldsymbol{\xi})\|.$$

Some comments. (1) The independence assumption in A1 and in the zero expected value condition in A2 can be weakened assuming that $\epsilon_{j,t}$ is a martingale difference with subgaussian (or subexponential) tails, conditionally to the past values of $\boldsymbol{\mathcal{Z}}_s$, for $1 \leq s \leq t$. The process $\{\boldsymbol{\mathcal{Z}}_t\}$ is allowed to be nonrandom and A5 imposes a condition on the boundedness of the eigenvalues of the matrix $T^{-1}\sum_{t=1}^T \boldsymbol{\mathcal{Z}}_t^\top\boldsymbol{\mathcal{Z}}_t$. Assumption A6 is merely technical: it simplifies the proof of the Corollary 1; we do not exclude that one may prove the same result weakening this assumption. Finally A7 is fairly standard in high-dimensional statistics: it requires that the number of parameters grows slower than the number of observations. We refer to Park et al. [23] p. 293 for an additional discussion on these assumptions.

(2) The rate of convergence in Corollary 1 is the same as the one in Theorem 2 of Park et al. [23] for OLS sieve M-estimator. Therefore, we conclude that the robust sieve M-estimator obtained using the Huber loss function has the

same rate of convergence as the widely and routinely applied sieve OLS M-estimator. However, differently from the latter, the former guarantees stability of the estimates in the presence of outliers.

(3) As pointed out by an anonymous referee, one may want to measure the accuracy of our consistent sieve estimator via its limiting distribution, e.g. with the aim of defining confidence intervals. To that purpose, one needs to derive the limiting distribution (typically an asymptotic normality result is desired) of the sieve M-estimators in the setting of dynamic semiparametric factor model (2.1). Unfortunately, to the best of our knowledge, there does not yet exist such a general theory. Asymptotic normality is available for OLS sieve M-estimators (see e.g. Huang [17] for pointwise asymptotic normality of the spline series OLS estimator) and for real-valued (smooth) mappings of sieve M-estimators (see Chen [3] for a general overview of the problem and Chen et al. [4] for some results related to the asymptotic normality of plug-in sieve M-estimators).

(4) To prove Corollary 1, $\{\boldsymbol{\mathcal{Z}}_t\}$ can be either non random or a stochastic process. Our method produces robust estimates of the unobserved factors. In practice, one often conducts additional inference on the estimated values by time series modeling. This is the approach that we illustrate in the motivating example of section 2.3, where the univariate latent factor is modeled by an autoregressive process. More generally, for L-dimensional factors ($L > 1$) as those analyzed in the simulation exercise of section 6.2 (Model M2), one may consider a Vector Autoregressive (VAR) process for the (mean-adjusted) process. In these cases, one needs to characterize the behaviour of the stochastic process $\{\boldsymbol{\mathcal{Z}}_t\}$. The main problem with these time series approaches is if the inference based on the estimated factors coincides (in some statistical sense, see below) with the one based on the unobserved factors. With this question in mind, as in Park et al. [23], we introduce the following assumption:

Assumption A8 (Assumption A8)**.** *$\{\boldsymbol{\mathcal{Z}}_t\}$ is a strictly stationary sequence with $E(\boldsymbol{\mathcal{Z}}_t) = 0$, $E(\|\boldsymbol{\mathcal{Z}}_t\|^\gamma) < \infty$ for some $\gamma > 2$. It is strongly mixing with $\sum_{i=1}^{\infty} \alpha(i)^{(\gamma-2)/\gamma} < \infty$. The matrix $E\boldsymbol{\mathcal{Z}}_t\boldsymbol{\mathcal{Z}}_t^{\mathrm{T}}$ has full rank. The process $\{\boldsymbol{\mathcal{Z}}_t\}$ is independent of $(\epsilon_{1,1}, \ldots, \epsilon_{T,J})$.*

Under additional assumptions on function $\boldsymbol{m}$ and on the growth of K, J, T, the Theorem 3 in Park et al. [23] implies that a VAR process can be successfully applied to model the estimated L-dimensional factors. Specifically, Theorem 3 implies that the difference between the VAR parameter estimated using $\hat{\boldsymbol{\mathcal{Z}}}_t$ and the (unfeasible) VAR parameter estimated using $\boldsymbol{\mathcal{Z}}_t$ is asymptotically negligible. We refer to Park et al. paper, p.294 and p.295 for further details.

6.2. Simulation exercises

To gain further insights on the behavior of the classical (namely, OLS) and robust estimators, we consider an extensive simulation study, with different sample sizes and different model complexities. Precisely, we generate two classes of semiparametric models and use the PE (see equation 2.3) to measure the

overall performance of the different estimation methods. The first model, M1, is a single factor model ($L = 1$) with a univariate covariate. The presence of only one factor in M1 allows us to explore the behavior of the OLS and robust estimator using various performance measures (see, PE but also $d_z, d_m, d_\beta, d_\rho$ defined below). The second model, M2, is a potentially multivariate latent factor ($L = 1, 2, 3$) with bivariate covariates. For M2, we study the PE using various sample sizes and number of factors.

For each model, we use $N = 500$ Monte Carlo runs. For computations we make use of B-splines basis functions. In each run, we consider a clean sample (as generated using the underlying factor model under the reference assumptions, see details below) and a contaminated sample, where outliers are obtained replacing randomly a percentage η of the original datum $(y_{t,j})$ by $y^{\text{cont}} = \sigma^2_{\text{out}}\tilde{t}$, with $\tilde{t}$ being the realization of a r.v having a student t-distribution with five degrees-of-freedom. For the robust estimator, we consider the Huber loss function with tuning constant c. Here below, we itemize the main aspects of the underlying data generating process for each model.

Model M1. The clean samples are generated by:

$$y_{t,j} = m_0(\xi_j) + \mathcal{Z}_t m_1(\xi_j) + \epsilon_{t,j} \tag{6.2}$$

with an ARMA structure on the latent: $\mathcal{Z}_t = \beta\mathcal{Z}_{t-1} - \rho u_{t-1} + u_t$ with $u_t \sim \mathcal{N}(0,1)$. The error $\epsilon_{t,j} \sim \mathcal{N}(0,1)$ is independent of u_t, while ξ defines a grid over $[0,1]$ and it is such that $\xi_j = j/(J+1)$. We set $m_0(\xi) = \cos(2\pi\xi)$, $m_1(\xi) = \sin(2\pi\xi)/[2 - \sin(2\pi\xi)]$, $\beta = 0.9$ and $\rho = 0.3$.

In Table 1, we show the outcomes of the simulation study, presenting the PE as well as additional stability measures related to the estimates of the model parameters. Specifically, we set $d_z = \sum_{t=1}^{T}(\hat{\mathcal{Z}}_t - \mathcal{Z}_t)^2 / \sum_{t=1}^{T}\mathcal{Z}_t^2$; $d_m = \|\hat{\boldsymbol{m}} - \boldsymbol{m}\|_J^2 / \|\boldsymbol{m}\|_J^2$; $d_\beta = |\hat{\beta} - \beta|^2/|\beta|^2$; $d_\rho = |\hat{\rho} - \rho|^2/|\rho|^2$. In line with the theoretical developments, we see that the classical and robust estimator behaves similarly when there is no contamination. However, under contamination, the performance of the OLS estimator deteriorates, while the digits of our robust estimators remain essentially stable. For example, PE is equal to 0.21 for both estimators under no contamination. As the contamination rate increases, PE is driven to 0.65 for the OLS estimator, while it remains bounded by 0.25 for the robust estimator. Similar results hold for the other measures. These digits confirm, in a larger Monte Carlo exercise, the results obtained in the sensitivity analysis of Section 2.3.

To illustrate graphically the stability of the robust estimator, in Figure 3, we display the functional boxplots (see Sun and Genton [27]) of the estimates of the function m_1. The plots show that, in the absence of contamination, the OLS and the robust method yield similar estimates, having similar variability and similar median curve (the curve with maximal functional depth, representing a typical estimate). However, in the presence of contamination, the OLS estimates suffer from a large variability, as can be remarked looking at the distance between the first and third functional quartile (the spread is larger than in the clean

TABLE 1. *Summary of the performance of the estimators computed for model M1 under different scenarios; the median of each measure over 500 simulations is given, with its median absolute deviation in parentheses; we chose* $T = 100$, $J = 120$, $K = 18$, $\sigma_{out} = 8$ *and* $c = 3$.

			OLS					Robust		
η	0%	1%	2.5%	5%	7.5%	0%	1%	2.5%	5%	7.5%
PE	0.21	0.27	0.37	0.51	0.65	0.21	0.21	0.22	0.23	0.25
	(0.02)	(0.12)	(0.21)	(0.28)	(0.33)	(0.03)	(0.05)	(0.06)	(0.09)	(0.13)
d_z	0.41	0.46	0.54	0.65	0.76	0.41	0.41	0.42	0.43	0.45
	(0.15)	(0.16)	(0.21)	(0.24)	(0.27)	(0.15)	(0.15)	(0.16)	(0.17)	(0.18)
d_m	0.34	0.36	0.4	0.48	0.56	0.34	0.34	0.34	0.34	0.35
	(0.21)	(0.21)	(0.23)	(0.26)	(0.29)	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)
d_β	0.07	0.08	0.1	0.15	0.21	0.08	0.07	0.08	0.08	0.08
	(0.08)	(0.11)	(0.20)	(0.32)	(0.39)	(0.08)	(0.08)	(0.12)	(0.13)	(0.13)
d_ρ	0.34	0.39	0.5	0.71	0.92	0.34	0.34	0.35	0.36	0.37
	(0.29)	(0.32)	(0.48)	(0.66)	(0.73)	(0.29)	(0.30)	(0.33)	(0.35)	(0.36)

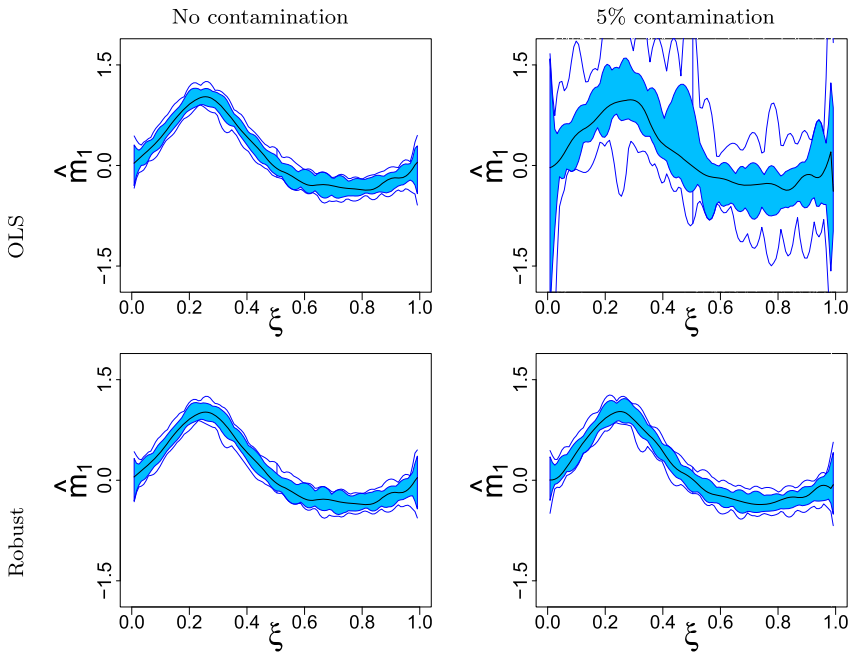

FIG 3. *Functional boxplots of the estimates of* m_1. *Results are obtained from* 500 *replications of model M1.*

sample and the quartile curves become wiggle). Moreover, looking at the median functional curve, with and without contamination, we remark that the OLS estimation method entails large changes in the typical estimate of m_1— e.g. look at the behaviour of the median functional curve of the OLS estimator, in the interval $\xi \in [0.6, 0.8]$, with and without contamination. In contrast, the estimates based on our robust method remains essentially unchanged in the presence of contamination. Figure 4 confirms this pattern, displaying the stability of the robust estimates via the boxplot of the PE. Finally, Figure 5 shows the boxplots for $\hat{\beta}$, where the parameter is estimated using the maximum likelihood method on $\{\hat{Z}_t\}$ and the estimates of latent factors are obtained using either the OLS or our robust method. In agreement with (and in complement to) the results in Section 2.3, the plots illustrate that the robust sieve M-estimation guarantees the numerical stability of the $\hat{\beta}$ in the presence of contamination.

Model M2. The clean samples are generated by a L-dimensional factor models with VAR(1) structure (with $L = 1, 2, 3$):

$$y_{t,j} = m_0(\boldsymbol{\xi}_j) + \sum_{l=1}^{L} \mathcal{Z}_{t,l} m_l(\boldsymbol{\xi}_j) + \epsilon_{t,j} \tag{6.3}$$

where $\epsilon_{t,j} \sim N(0, 10)$, $\boldsymbol{\xi}_j$ are bivariate uniforms over $[0, 1]^2$. $\boldsymbol{\mathcal{Z}}_t = \boldsymbol{\beta}_L \boldsymbol{\mathcal{Z}}_{t-1} + \boldsymbol{u}_t$ is

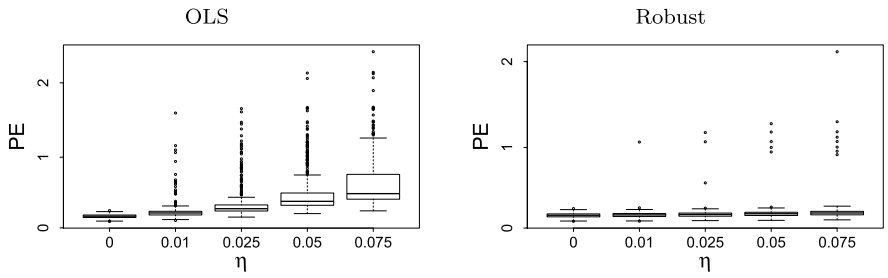

FIG 4. *Boxplots of prediction errors PE for* $\eta = 0, 1, 2.5, 5$ *and* 7.5% *contamination. Results are obtained from* 500 *replications of model M1.*

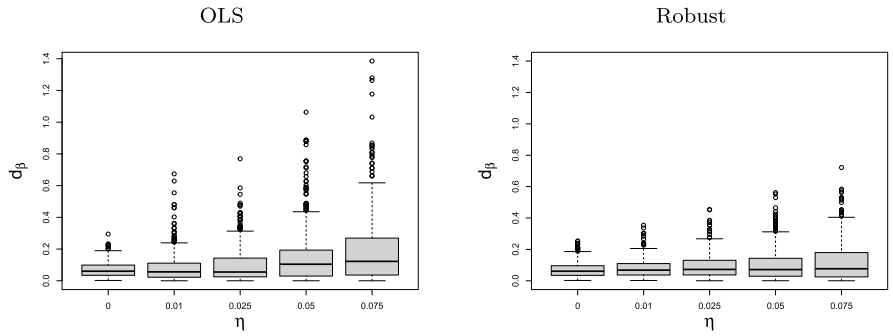

FIG 5. *Boxplots of* $d_\beta = |\hat{\beta} - \beta|^2/|\beta|^2$ *for* $\eta = 0, 1, 2.5, 5$ *and* 7.5% *contamination. Results are obtained from* 500 *replications of model M1.*

a L-dimensional Vector Autoregressive (VAR) process with $\boldsymbol{u}_t$ generated from a $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_L)$. We consider the following 2-dimensional functions: $m_0(\xi_1, \xi_2) = \cos(2\pi\xi_2)$, $m_1(\xi_1, \xi_2) = 1.34\sin(2\pi\xi_1)$, $m_2(\xi_1, \xi_2) = 8.32((\xi_1 - 0.5)^2 + (\xi_2 - 0.5)^2 - 1.36)$, $m_3(\xi_1, \xi_2) = 8.17((\xi_1 - 0.5)^3 - (\xi_2 - 0.5))$. $\boldsymbol{\beta}_L$ are $L \times L$ matrices. We choose $\beta_1 = 0.95$, $\boldsymbol{\beta}_2$ is the 2×2-matrix with rows from the top are equal to $(0.95, -0.2)$, $(0, 0.8)$, and $\boldsymbol{\beta}_3$ is the 3×3-matrix with rows $(0.95, -0.2, 0)$, $(0, 0.8, 0.1)$ and $(0.1, 0, 0.6)$. We generate data for $L = 1, 2, 3$ for three different settings with different T, J, K and we report the prediction error PE in Table 2. We flag that, while setting 1 is low dimensional ($T < J$), setting 2 is a high-dimensional setting, with a large number of time series $J > T$ in. In Setting 3 we push even further the time series dimension T and total number of time series J, obtaining a big data framework with $T \times J = 1,250,000$ observations.

For all settings, we see that the OLS and the robust estimators perform similarly without contamination. However, we remark that our robust approach outperforms the OLS estimator in the presence of contamination. For instance, in setting 1, the estimators yield comparable PE (0.814 and 0.853 for the OLS and the robust estimator, respectively), when $\eta = 0$. Nevertheless, when $\eta = 7.5\%$, the PE of the OLS estimator increases to 1.51, while the robust estimator yields

TABLE 2. *Summary of the performance of the estimators under model M2 under different scenarios; the median of the relative prediction error PE for each measure over 500 simulations is given, with its median absolute deviation in parentheses; we chose* $\sigma_{out} = 40$ *and* $c = 10$.

	OLS					Robust				
η	0%	1%	2.5%	5%	7.5%	0%	1%	2.5%	5%	7.5%
Setting 1	$T = 200,\ J = 100,\ K = 5 \times 5$									
L=1	0.814	0.958	1.114	1.324	1.51	0.853	0.865	0.89	0.931	0.984
	(0.132)	(0.164)	(0.206)	(0.242)	(0.285)	(0.138)	(0.144)	(0.156)	(0.169)	(0.195)
L=2	0.072	0.087	0.107	0.137	0.166	0.075	0.076	0.079	0.083	0.088
	(0.012)	(0.015)	(0.018)	(0.020)	(0.023)	(0.013)	(0.014)	(0.014)	(0.016)	(0.017)
L=3	0.067	0.081	0.101	0.133	0.163	0.071	0.072	0.074	0.078	0.083
	(0.012)	(0.015)	(0.018)	(0.020)	(0.023)	(0.012)	(0.012)	(0.014)	(0.015)	(0.017)
Setting 2	$T = 200,\ J = 625,\ K = 8 \times 8$									
L=1	0.345	0.388	0.451	0.548	0.611	0.361	0.362	0.363	0.376	0.385
	(0.085)	(0.098)	(0.117)	(0.152)	(0.159)	(0.088)	(0.088)	(0.089)	(0.089)	(0.093)
L=2	0.031	0.038	0.05	0.074	0.097	0.033	0.033	0.034	0.036	0.04
	(0.008)	(0.009)	(0.008)	(0.010)	(0.010)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)
L=3	0.028	0.034	0.046	0.069	0.094	0.03	0.03	0.031	0.033	0.037
	(0.007)	(0.008)	(0.008)	(0.008)	(0.009)	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)
Setting 3	$T = 500,\ J = 2500,\ K = 15 \times 15$									
L=1	0.183	0.212	0.243	0.285	0.328	0.191	0.196	0.197	0.199	0.204
	(0.024)	(0.026)	(0.032)	(0.035)	(0.040)	(0.026)	(0.025)	(0.026)	(0.024)	(0.026)
L=2	0.016	0.022	0.034	0.057	0.081	0.017	0.018	0.019	0.021	0.025
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
L=3	0.015	0.020	0.032	0.056	0.080	0.015	0.016	0.017	0.020	0.024
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

a PE which remains fairly stable, being about 0.984. Similar considerations hold for the other settings, even the high-dimensional ones.

6.3. Real data exercise

We illustrate the applicability of our method via an exercise on dimensionality reduction for a real dataset of fMRI records. We consider an open source data set coming from a study on either adult or young patients, in which subjects are exposed to a suddenly appearing peripheral target. They are asked to inhibit the strong urge to saccade toward the target and instead to look toward the mirror location. The data are collected at the Brain Imaging Research Center, at the University of Pittsburgh and is publicly available at https://www.openfmri.org/; see in Geier et al. [13].

It is well known that outliers are more likely to affect data collected on young subjects. An immaturity of the brain and a weaker ability to concentrate typically produce some issues like head motions and mind wandering, causing outliers and anomalous records. As a result, we think that our robust procedure can be helpful for the analysis of the fMRI data of young patients. Thus, we select a 12 year-old male (the seventh subject in the Geier et al. [13] dataset, which contains 38 patients).

We apply model (2.1), where, as in the Monte Carlo experiments, we take the voxel index (i_1, i_2, i_3) as covariate and we set $L = 2$. We focus on 15 slices of the brain, so $J = 64 \times 64 \times 15$ and $T = 180$. To implement our robust sieve M-estimator, we need to select the tuning constant c. To this end, we propose to adapt the data-driven approach of La Vecchia et al. [20] (developed in the univariate time series setting) to our high-dimensional context. Essentially, the method relies on the empirical stability of the estimates: it ensures that the selected tuning constant is such that the resulting estimating function coincides with the classical estimating function, in absence of contamination, whereas, in the presence of contamination, it yields the bounded estimating function closest to the non-robust (unbounded) one. The algorithm and the additional numerical aspects related to its application to the current dataset are available in the supplementary materials. Using the selected c, we compute the robust estimates. For the sake of comparison, we also compute the OLS estimates of the latent factors and of the functions m_0, m_1 and m_2. Our code for the estimation of robust semiparametric factor models is available at https://github.com/JulienSB/RobDSFM.

In Figure 6, we display the robust estimates of the functions and the absolute difference between the robust and OLS estimates. Specifically, for $l = 0, 1, 2$, first we compute $\hat{\Delta}_{m_l} = |\hat{m}_l^{OLS}(\boldsymbol{\xi}) - \hat{m}_l^{ROB}(\boldsymbol{\xi})|$, for $\boldsymbol{\xi} \in [0, 1]^3$, then we plot the brain areas where $\hat{\Delta}_{m_l} \geq t_{m_l}$, with t_{m_l} being the third quartile of $\hat{\Delta}_{m_l}$. The right panels illustrate that the OLS estimates seem to emphasize the brain activity in some regions. Looking at m_0, the deviations of the OLS estimates from the robust ones are close to the borders of the head. It is well-known (see [32]) that these deviations are typically due to head motion. For m_1 and m_2,

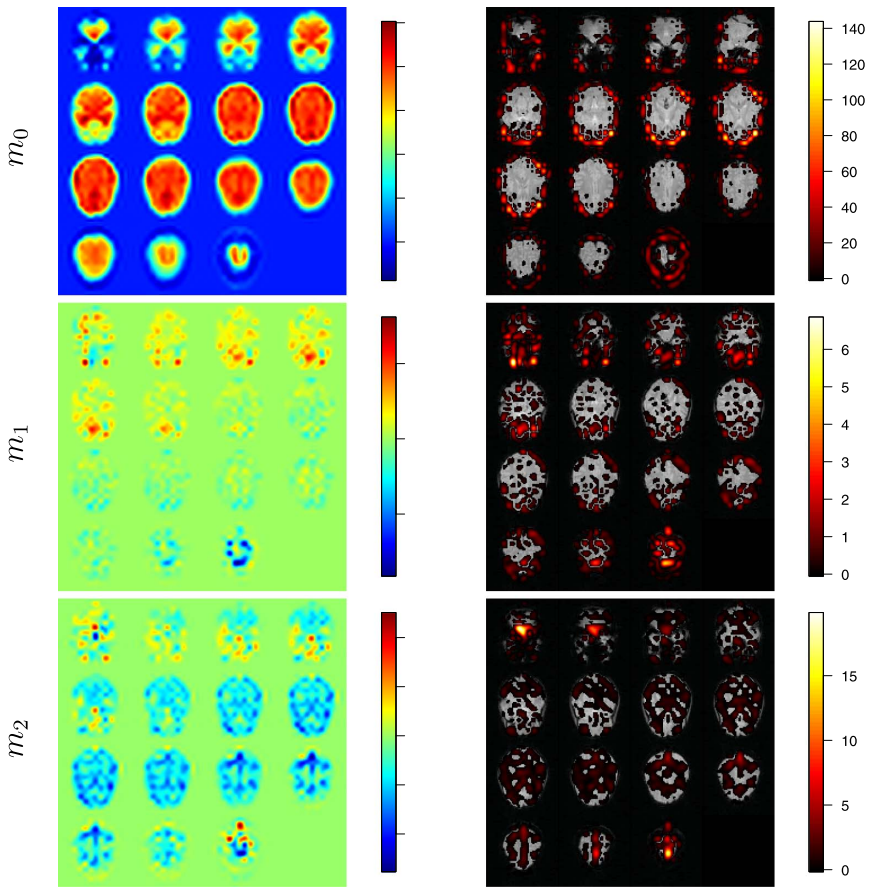

FIG 6. *Left panels: Robust estimates of* m_0, m_1 *and* m_2. *Right panels:* $\hat{\Delta}_{m_l} \geq t_{m_l}$ *for* $l = 1, 2, 3$.

the deviations are due to movements of the eyes and/or to some body movements. Similar considerations hold for the estimates of the latent factors: see Figure 7, where we plot the robust estimates of the factors and their difference with OLS estimates. To investigate the robustness and the stability of our estimates, we perform a sensitivity analysis. Mimicking the logic applied in the numerical examples of Section 6.2, we add outliers to the data by moving the most influential points (as detected by the Huber weights, unreported). Then, we compare our robust estimates to the estimates obtained using the non-robust OLS method. To perform the sensitivity analysis, we randomly replace a percentage η of the original data $\{y_{t,j}\}$ by a random variable having a t-student distribution, with 5 degrees-of-freedom. For the sake of comparison, we compute the predicted value $\hat{y}^{\eta}_{t,j}$ for different values of $\eta \in \{0\%, 2.5\%, 5\%, 7.5\%, 10\%\}$ and, for each method, we calculate the Relative Mean Square Error, defined as

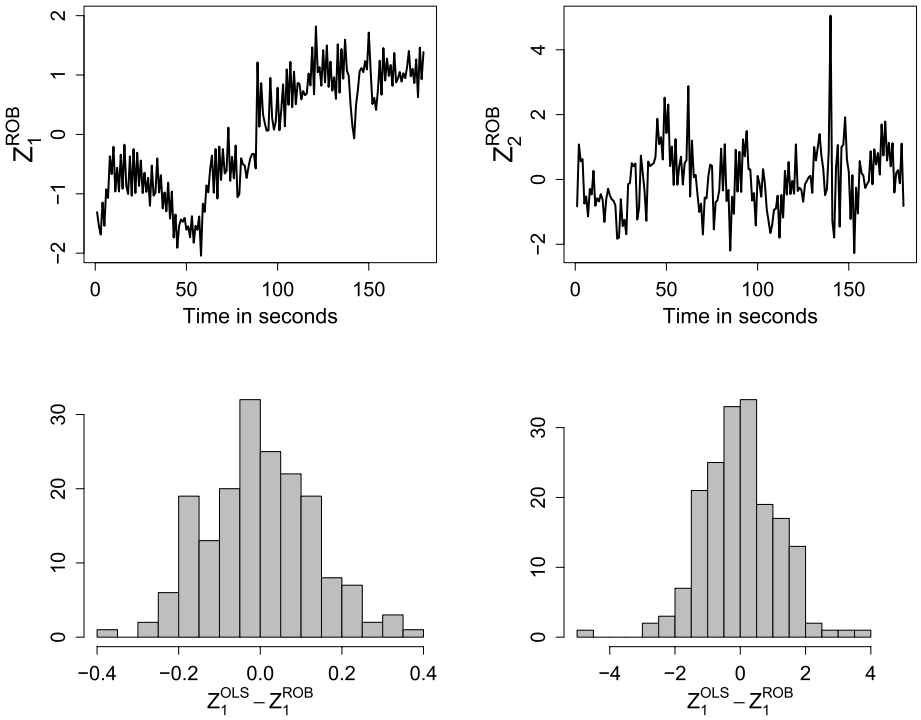

FIG 7. *Robust estimates of the $L = 2$ factors (upper panels) and histograms of the difference between least squares and robust estimates (lower panels).*

$\mathrm{RMSE}(\eta) = \sum_{t=1}^{T}\sum_{j=1}^{J}(\hat{y}_{t,j}^{\eta} - \hat{y}_{t,j}^{0})^2 / \sum_{t=1}^{T}\sum_{j=1}^{J}(\hat{y}_{t,j}^{0})^2$, where $\hat{y}_{t,j}^{\eta}$ denotes the predicted value of $y_{t,j}$ using the η-contaminated data. In line with the theoretical developments, the robust estimates yield a stable and bounded RMSE, whilst the OLS estimator entails a RMSE which diverges (exponentially) fast as η grows—the plot is available in the supplementary material.

7. Conclusion and outlook

We develop a robust sieve M-estimation procedure, when the sieve space is non-convex, and we illustrate the use of our method for the statistical analysis of large fMRI dataset. Further developments can be considered in terms of applications and theory. We mention some possible research directions.

On the applications side, we foresee that our inference procedure can be applied also in the context of longitudinal studies (consisting of fMRI data about many patients), as in van Bömmel et al. [28]. Other potential applications are in finance, for the modelling of the implied volatility (see Fengler et al. [10] and Park et al. [23]), and in economics, for forecasting inflation (see Chen et al. [5]). On the computational side, in this paper, we propose the use of a unique

bounding constant (c) in the Huber function ψ_c. However, in many applications (like e.g in the analysis of fMRI data of several patients) one might consider different values of c, for different patients and/or for different areas of the brain. To this end, new algorithms need to be developed for an efficient numerical implementation of these robust inference procedures.

On the theoretical side, an open problem is related to the derivation of a limiting distribution of the proposed robust sieve M-estimators in the context of model (2.1). Another aspect that deserves further investigation is related to the possibility of modeling the spatial dependence enconded in the factor loadings. Indeed, similarly to the approach adopted for the latent factors, one may conjecture that under suitable assumptions (like Assumption A8), the spatial dependence can be modeled by operating on the estimated factor loadings. However, this additional modeling step needs a careful theoretical justification that, at present, we do not have.

Appendix A: Proofs of Theorems

Proof of Lemma 1. Because γ is convex, for $\alpha \in [0,1]$, it holds that:

$$\gamma\{y_i - g(\boldsymbol{x}_i)\} - \gamma\{y_i - g_n^*(\boldsymbol{x}_i)\} \geq \frac{1}{\alpha}\left(\gamma[(1-\alpha)\{y_i - g_n^*(\boldsymbol{x}_i)\} + \alpha\{y_i - g(\boldsymbol{x}_i)\}]\right.$$
$$\left. -\gamma\{y_i - g_n^*(\boldsymbol{x}_i)\}\right).$$

So by taking $\alpha = (1 + b_n\|g - g_n^*\|_n)^{-1}$ we obtain

$$\gamma_g(\boldsymbol{x}_i) - \gamma_{g_n^*}(\boldsymbol{x}_i) \geq (1 + b_n\|g - g_n^*\|_n)\left(\gamma_{g_n^*+f}(\boldsymbol{x}_i) - \gamma_{g_n^*}(\boldsymbol{x}_i)\right),$$

where $f = \frac{g - g_n^*}{(1+b_n\|g-g_n^*\|_n)} \in \mathcal{F}_n$ and thus

$$\int (\gamma_{g_n^*+\hat{f}_n} - \gamma_{g_n^*}) dP_n^0 \leq \int (\gamma_{\hat{g}_n} - \gamma_{g_n^*}) dP_n^0 \leq 0.$$

For n big enough, there is a constant C_3, such that $\|g_n^* - g_0\|_\infty < C_3$. Assuming $\kappa_1/(C_2 + C_3) \leq 1$, it yields that for $f \in \mathcal{F}_n$:

$$\begin{aligned}
&\int (\gamma_{g_n^*+f} - \gamma_{g_n^*}) dP^0 \\
&= n^{-1}\sum_{i=1}^n E\left(\gamma\left[\epsilon_i - \{f(\boldsymbol{x}_i) + g_n^*(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)\}\right]\right. \\
&\left. -\gamma[\epsilon_i - \{g_n^*(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)\}]\right) \\
&= n^{-1}\sum_{i=1}^n E\left(\gamma\left[\epsilon_i - \{f(\boldsymbol{x}_i) + g_n^*(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)\}\right] - \gamma(\epsilon_i)\right) \\
&- E\left(\gamma[\epsilon_i - \{g_n^*(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)\}] - \gamma(\epsilon_i)\right) \\
&\geq n^{-1}\sum_{i=1}^n E\left(\gamma\left[\epsilon_i - (C_2 + C_3)^{-1}\kappa_1\{f(\boldsymbol{x}_i) + g_n^*(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)\}\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&-\kappa_2\{g_n^*(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)\}^2 \\
&\geq n^{-1}\sum_{i=1}^{n}(C_2+C_3)^{-2}\kappa_1^3\{f(\boldsymbol{x}_i)+g_n^*(\boldsymbol{x}_i)-g_0(\boldsymbol{x}_i)\}^2 - \kappa_2\{g_n^*(\boldsymbol{x}_i)-g_0(\boldsymbol{x}_i)\}^2,
\end{aligned}
$$

where the third inequality follows from Condition 2 and Condition 3 and the last inequality from Condition 1, 4. We denote Δ_n the approximation error $\|g_n^* - g_0\|_n$.Then by the reverse triangle inequality and setting $\lambda = (C_2+C_3)^{-2}\kappa_1^3$:

$$
\begin{aligned}
&\int(\gamma_{g_n^*+f} - \gamma_{g_n^*})dP^0 \geq \lambda\|f + (g_n^* - g_0)\|_n^2 - \kappa_2\Delta_n^2 \\
&\geq \lambda(\|f\|_n - \Delta_n)^2 - \kappa_2\|g_n^* - g_0\|_n^2 \geq \lambda(\|f\|_n^2 + \Delta_n^2 - 2\|f\|_n\Delta_n) - \kappa_2\Delta_n^2 \\
&\geq \lambda(\|f\|_n^2 + \Delta_n^2 - 2\Delta_n b_n^{-1}) - \kappa_2\Delta_n^2 = \lambda\left\{\|f\|_n^2 + \left(1-\lambda^{-1}\kappa_2\right)\Delta_n^2 - 2\Delta_n b_n^{-1}\right\}
\end{aligned}
$$

since $\|f\|_n \leq b_n^{-1}$. We obtain finally:

$$
\lambda\left\{\|f\|_n^2 + \left(1-\lambda^{-1}\kappa_2\right)\Delta_n^2 - 2\Delta_n b_n^{-1}\right\} \leq \int(\gamma_{g_n^*+f} - \gamma_{g_n^*})dP^0
$$

The lemma follows by substracting $\int(\gamma_{\hat{g}_n} - \gamma_{g_n^*})dP_n^0 \leq 0$ on the left hand side of the the previous inequality. □

Proof of Theorem 1. First, note that an inequality for $\hat{f}_n$ will be binding for $\|\hat{g}_n - g_n^*\|_n$ only if b_n^{-1} is big enough. Indeed, $\|\hat{f}_n\| \leq \delta$ implies only $\|\hat{g}_n - g_n^*\|_n \leq \delta + \delta b_n\|\hat{g}_n - g_n^*\|_n$. Thus we should require that $b_n^{-1} \geq 2\delta$ to obtain $\|\hat{g}_n - g_n^*\|_n \leq \delta + 2^{-1}\|\hat{g}_n - g_n^*\|_n$ which implies $\|\hat{g}_n - g_n^*\|_n \leq 2\delta$. Let us call $c = \min\left(\frac{\kappa_2}{\lambda} - 1, 0\right)$ and in a slight abuse of notation, we write $\|V\|_n^2 = \frac{1}{n}\sum_{i=1}^n V_i^2$. Making use of Lemma 1, on the set where

$$
c\|g_n^* - g_0\|_n + 2\|g_n^* - g_0\|_n b_n^{-1} \leq -\lambda^{-1}\int(\gamma_{g_n^*+\hat{f}_n} - \gamma_{g_n^*})d(P_n^0 - P^0),
$$

we use Lemma 8.5 in van de Geer [29] to obtain:

$$
\begin{aligned}
& P\left(\|\hat{g}_n - g_n^*\|_n > 2\delta_n \cap \|V\|_n^2 \leq \varsigma^2\right) \leq P\left(\|\hat{f}_n\|_n^2 > \delta_n^2 \cap \|V\|_n^2 \leq \varsigma^2\right) \\
\leq\ & P\left\{\sup_{f\in\mathcal{F}_n} -2\lambda^{-1}\int(\gamma_{g_n^*+f} - \gamma_{g_n^*})d(P_n^0 - P^0) \geq \delta_n^2 \cap \|V_i\|_n^2 \leq \varsigma^2\right\} \\
\leq\ & c_2\exp\left(-\frac{n\delta_n^4}{c_2^2 4\delta_n^2}\right) \leq c_3\exp\left(-\frac{n\delta_n^2}{c_3^2}\right),
\end{aligned}
$$

where c_3 is a constant depending on c_2. Then we obtain:

$$
\begin{aligned}
P\left(\|\hat{g}_n - g_n^*\|_n > 2\delta_n\right) &\leq P\left(\|\hat{g}_n - g_n^*\|_n > 2\delta_n \cap \|V\|_n^2 \leq \varsigma^2\right) + P\left(\|V\|_n^2 > \varsigma\right) \\
&\leq c_3\exp\left(-\frac{n\delta_n^2}{c_3^2}\right) + P\left(\|V\|_n^2 > \varsigma\right)
\end{aligned}
$$

Since the V_i are sub-Gaussian, $P\left(\|V\|_n > \varsigma\right)$ is small for $\varsigma > \varsigma_0$. Indeed Berstein's inequality with $\varsigma = 2\varsigma_0$ implies:

$$P\left(\|V\|_n^2 > 2\varsigma_0^2\right) \leq \exp\left(-\frac{n\varsigma_0^2}{12K_1}\right).$$

On the set where

$$c\|g_n^* - g_0\|_n + 2\|g_n^* - g_0\|_n b_n^{-1} \geq -\lambda^{-1}\int(\gamma_{g_n^*+\hat{f}_n} - \gamma_{g_n^*})d(P_n^0 - P^0),$$

we have

$$\|\hat{f}_n\|_n \leq 4\|g_n^* - g_0\|_n b_n^{-1} + 2c\|g_n^* - g_0\|_n^2.$$

If $\|g_n^* - g_0\|_n \leq b_n^{-1}$, then $\|\hat{f}_n\|_n^2 \leq (4+2c)b_n^{-2}$. This inequality will be always binding for $\|\hat{g}_n - g_n^*\|_n$ for n sufficiently large, since b_n^{-1} is strictly decreasing. On the other side $\|g_n^* - g_0\|_n > b_n^{-1}$ implies that $\|\hat{f}_n\|_n^2 \leq (4+2c)\|g_n^* - g_0\|_n^2$. In order to get a binding inequality for $\|\hat{g}_n - g_n^*\|_n$, we should impose $b_n^{-1} \geq 2(4+2c)\|g_n^* - g_0\|_n^2$, which is satisfied for n sufficiently large if $\|g_n^* - g_0\|_n = O(b_n^{-1})$. We choose $b_n^{-1} = \max\{2\delta_n, \|g_n^* - g_0\|_n\}$ to satisfy the former conditions so that:

$$\|\hat{g}_n - g_n^*\|_n = O_P\left(\delta_n + \|g_n^* - g_0\|_n + b_n^{-1}\right) = O_P\left(\delta_n + \|g_n^* - g_0\|_n\right).$$

Since $\|\hat{g}_n - g_0\|_n \leq 2\|\hat{g}_n - g_n^*\|_n + 2\|g_n^* - g_0\|_n$, the theorem follows. □

Proof of Corollary 1. The proof follows along the same lines as in the proof of Theorem 2 in Park et al. [23] and by applying Theorem 1. □

Appendix B: Additional technical material

B.1. Checking Conditions 1 and 2 for the Huber loss function

We prove that Conditions 1 and 2 are satisfied for the Huber loss function defined in (3.4).

Proof. Let's condsider $\kappa_1, K_3 \leq b$

$$\begin{aligned}
& E\{\gamma(\epsilon_i + u) - \gamma(\epsilon_i)\} = \int_{-\infty}^{\infty} \gamma(\epsilon_i + u) - \gamma(\epsilon_i) dP \\
= & \int_b^{\infty} 2b(\epsilon_i + u) - b^2 - (2b\epsilon_i - b^2) dP + \int_{-\infty}^{-b} -2b(\epsilon_i + u) - b^2 + (2b\epsilon_i + b^2) dP \\
& + \int_{-b}^{b} \epsilon_i^2 + u^2 + 2u\epsilon_i - \epsilon_i^2 dP \\
= & \int_b^{\infty} 2bu\, dP + \int_{-\infty}^{-b} -2bu\, dP + \int_{-b}^{b} u^2 + 2u\epsilon_i dP \\
= & \; 2bu\{P(\epsilon_i \geq b) - P(\epsilon_i \leq -b)\} + u^2 P(-b \leq \epsilon_i \leq b) + \int_{-b}^{b} 2u\epsilon_i dP
\end{aligned}$$

$$= \quad u^2\alpha P(-b \le \epsilon_i \le b) + 2bu\{P(\epsilon_i \ge b) - P(\epsilon_i \le -b)\} +$$
$$+ \quad u^2(1-\alpha)P(-b \le \epsilon_i \le b) + \int_{-b}^{b} 2u\epsilon_i dP.$$

For Condition 1 we know that $\int_{-\infty}^{\infty} \gamma(\epsilon_i + u) - \gamma(\epsilon_i)dP > 0$. As we can choose α as small as we want, there exists $\alpha > 0$ such that $2bu\{P(\epsilon_i \ge b) - P(\epsilon_i \le -b)\} + u^2(1-\alpha)P(-b \le \epsilon_i \le b) + \int_{-b}^{b} 2u\epsilon_i dP > 0$, and then we choose $\kappa_1 = \alpha P(-b \le W \le b)$ to get $E\{\gamma(\epsilon_i + u) - \gamma(\epsilon_i)\} \ge \kappa_1 u^2$. For Condition 2, we can choose α possibly greater than 1, and $\kappa_2 = \alpha P(-b \le \epsilon_i \le b)$ to obtain $E[\gamma(\epsilon_i + u) - \gamma(\epsilon_i)] \le \kappa_2 u^2$. □

B.2. Identification, estimation and implementation of the semiparametric factor model

B.2.1. Identification and estimation

For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, we write for brevity $\boldsymbol{A}_{i,.} = (\boldsymbol{A}_{i,1}, ..., \boldsymbol{A}_{i,m})^\top$, $\boldsymbol{A}_{.,j} = (\boldsymbol{A}_{1,j}, ..., \boldsymbol{A}_{n,j})^\top$ and we define a vector operator by $\mathrm{vec}(\boldsymbol{A}) = (\boldsymbol{A}_{1,.}^\top, ..., \boldsymbol{A}_{n,.}^\top) \in \mathbb{R}^{nm}$. We can rewrite equation (3.1) in a more compact form:

$$y_{t,j} = (1, \boldsymbol{\mathcal{Z}}_{t,.}^\top)\boldsymbol{m}(\boldsymbol{\xi}_{t,j}) + \epsilon_{t,j} \tag{B.1}$$

for $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{T\times L}$. For identification, we follow Fengler et al. [10] and we assume that the functions m_l are orthonormal in $\mathcal{L}^2([0,1]^d)$ that is $\int m_l^2(\boldsymbol{\xi})d\boldsymbol{\xi} = 1$ for any $l = 0, 1, ..., L$ and $\int m_l(\boldsymbol{\xi})m_{l'}(\boldsymbol{\xi})d\boldsymbol{\xi} = 0$ for $l \ne l'$. We also assume that the factors are ordered according to their variances as in principal component analysis (PCA): $\boldsymbol{\mathcal{Z}}_{.,1}^\top\boldsymbol{\mathcal{Z}}_{.,1} > \boldsymbol{\mathcal{Z}}_{.,2}^\top\boldsymbol{\mathcal{Z}}_{.,2} > ... > \boldsymbol{\mathcal{Z}}_{.,L}^\top\boldsymbol{\mathcal{Z}}_{.L}$. The estimator $\hat{\boldsymbol{\theta}} = (\mathrm{vec}(\hat{\boldsymbol{\mathcal{Z}}}), \mathrm{vec}(\hat{\boldsymbol{\mathcal{A}}}))^\top$ satisfies then:

$$\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) = \sum_{j=1}^{J}\sum_{t=1}^{T} \gamma\{y_{t,j} - (1, \boldsymbol{z}_{t,.}^\top)\boldsymbol{\mathcal{A}}\boldsymbol{\phi}(\boldsymbol{\xi}_{t,j})\} = \min_{\boldsymbol{\theta}\in\Theta_S} !, \tag{B.2}$$

for some loss function γ. From now on, we restrict the parameter space Θ_S to be the set of all $\boldsymbol{\theta} = (\mathrm{vec}(\boldsymbol{z}), \mathrm{vec}(\boldsymbol{\mathcal{A}}))^\top$ such that $\boldsymbol{z}_{.,1}^\top\boldsymbol{z}_{.,1} > ... > \boldsymbol{z}_{.,L}^\top\boldsymbol{z}_{.,L}$ and such that $\boldsymbol{\mathcal{A}}$ is an orthonormal matrix of size $(L+1)\times K$.

To emphasize the dependence on $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{z}$, we sometimes write $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{z}, \boldsymbol{\mathcal{A}})$. In fact, the solution to (3.2) is not unique and as in PCA one can show that it is determined only up to a multiplicative constant. Specifically, if $\boldsymbol{\theta}(\hat{\boldsymbol{\mathcal{Z}}}, \hat{\boldsymbol{\mathcal{A}}})$ is a solution, $(\hat{\boldsymbol{\mathcal{Z}}}\boldsymbol{B}, \boldsymbol{B}'\hat{\boldsymbol{\mathcal{A}}})$ is also a solution, for any diagonal matrix $\boldsymbol{B} \in \mathbb{R}^{L\times L}$ with diagonal elements 1 or -1 (i.e. $\boldsymbol{B} \in \{\boldsymbol{B} = \mathrm{diag}(b_l), b_l^2 = 1, l = 1, ..., L\}$) and with

$$\boldsymbol{B}' = \begin{pmatrix} 1 & 0 \\ 0 & \boldsymbol{B}^{-1} \end{pmatrix}. \tag{B.3}$$

Then, we proceed as in functional PCA (see for e.g. [18]), we introduce the unobservable matrix of random signs $\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}\in\mathcal{B}} \sum_{t=1}^{T} \|\hat{\boldsymbol{\mathcal{Z}}}_{t,.}\boldsymbol{B} - \boldsymbol{\mathcal{Z}}_{t,.}\|$ and $\hat{\boldsymbol{\theta}}$ will denote in fact $\boldsymbol{\theta}(\hat{\boldsymbol{\mathcal{Z}}}\hat{\boldsymbol{B}}, \hat{\boldsymbol{B}}'\hat{\boldsymbol{\mathcal{A}}})$, which is then unique.

Denoting $\boldsymbol{\alpha} = \mathrm{vec}(\boldsymbol{\mathcal{A}})^T$, it can be shown by some algebraic computations that

$$\boldsymbol{IF}^s(P_n; P^{\boldsymbol{\theta}}, \vartheta) = -E_{P^{\boldsymbol{\theta}}} \begin{pmatrix} \nabla_{zz}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) & \nabla_{\alpha}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z})^\top \\ \nabla_{z\alpha}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) & \nabla_{\alpha\alpha}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) \end{pmatrix}^{-1} \begin{pmatrix} \nabla_{z}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) \\ \nabla_{\alpha}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}), \end{pmatrix} \quad \text{(B.4)}$$

where, denoting by $\boldsymbol{\phi}_{t,j}$ the $K \times 1$ vector $\boldsymbol{\phi}(\boldsymbol{\xi}_{t,j})$ and by $\boldsymbol{A}$ the $L \times K$ matrix obtained by deleting the first row of $\boldsymbol{\mathcal{A}}$,

$$\nabla_{\alpha}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) = \sum_{j=1}^{J}\sum_{t=1}^{T} \psi(\epsilon_{t,j}) \otimes (1, \boldsymbol{z}_{t,\cdot}^\top)^\top,$$

$$\nabla_{\alpha\alpha}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) = \sum_{j=1}^{J}\sum_{t=1}^{T} \dot{\psi}(\epsilon_{t,j})\boldsymbol{\phi}(\boldsymbol{\xi}_{t,j})\boldsymbol{\phi}_{t,j}^\top \otimes (1, \boldsymbol{z}_{t,\cdot}^\top)^\top (1, \boldsymbol{z}_{t,\cdot}^\top),$$

$\nabla_z\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) = \{-\sum_{j=1}^{J} \psi(\epsilon_{1,j})\boldsymbol{A}\boldsymbol{\phi}_{1,j}, ..., -\sum_{j=1}^{J} \psi(\epsilon_{T,j})\boldsymbol{A}\boldsymbol{\phi}_{T,j}\}$. Let $\tilde{\boldsymbol{I}}_L$ be a $(L+1)\times L$ matrix such that $\tilde{\boldsymbol{I}}_L^\top = (0, \boldsymbol{I}_L)$ and $\boldsymbol{I}_L$ denote the identity matrix of dimension L. Then we obtain: $\nabla_{z\alpha}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}) = [\sum_{j=1}^{J} \dot{\psi}(\epsilon_{1,j})(\boldsymbol{\phi}_{1,j} \otimes (1, \boldsymbol{z}_{1,\cdot}^\top)^\top)\boldsymbol{\phi}_{1,j}^\top\boldsymbol{A}^\top - \psi(\epsilon_{1,j})\boldsymbol{\phi}_{1,j} \otimes \tilde{I}_L, ..., \sum_{j=1}^{J} \dot{\psi}(\epsilon_{T,j})(\boldsymbol{\phi}_{T,j} \otimes (1, \boldsymbol{z}_{T,\cdot}^\top)^\top)\boldsymbol{\phi}_{T,j}^\top\boldsymbol{A}^\top - \psi(\epsilon_{T,j})\boldsymbol{\phi}_{T,j} \otimes \tilde{I}_L]$. Finally, $\nabla_{zz}\Gamma(\boldsymbol{\alpha}, \boldsymbol{z})$ is a $TL \times TL$ matrix that consists of T diagonal blocks $\sum_{j=1}^{J} \psi(\epsilon_{t,j})\boldsymbol{A}\boldsymbol{\phi}_{t,j}\boldsymbol{\phi}_{t,j}^\top\boldsymbol{A}^\top$ for $t = 1, ..., T$.

B.2.2. Implementation of the robust estimation procedure

For the implementation of our estimation procedure, one can follow the approach in Park et al. [23]. We implement the method mentioned on page 286 of Park et al. [23] paper, which builds on the algorithm proposed by Fengler et al. [10]. Specifically, to find a solution $\hat{\boldsymbol{\theta}}$ to (B.2), we adopt an iterative algorithm:

(i) choose an initial value for $\boldsymbol{z}^{(0)}$,
(ii) minimize $\Gamma(\boldsymbol{\alpha}, \boldsymbol{z}^{(0)})$ with respect to $\boldsymbol{\alpha}$, and call the minimizer $\boldsymbol{\alpha}^{(1)}$,
(iii) minimize $\Gamma(\boldsymbol{\alpha}^{(1)}, \boldsymbol{z})$ with respect to $\boldsymbol{z}$.

Iterate (ii) and (iii) until convergence. To solve each of these steps, one can use a Newton-type algorithm and use the partial derivatives of Γ computed above in (B.4). A relevant implementation aspect of the aforementioned algorithm is related to the selection of the tuning constant c. We provide a criterion (and an algorithm) to deal with that.

Tuning constant selection. Let $\hat{\boldsymbol{\theta}}$ be the estimator corresponding to the non-robust loss function γ and $\hat{\boldsymbol{\theta}}_c$ the estimator corresponding to γ_c the robustified loss function with tuning constant c. Choose $c_1 = \max\{|\psi(y_i - g_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}_i))|, i = 1, ..., n\}$ where ψ is the non-robust score function. Define a spaced grid $c_1 > c_2 > ... > c_m > 0$ and compute the correspondent estimates, $\hat{\boldsymbol{\theta}}_{c_i}, i = 1, ..., m$. Compute the absolute variations $AV_{c_i} = \|g_{\hat{\boldsymbol{\theta}}_{c_i}}(\boldsymbol{x}_i) - g_{\hat{\boldsymbol{\theta}}_{c_{i+1}}}(\boldsymbol{x}_i)\|_n$. Then we select

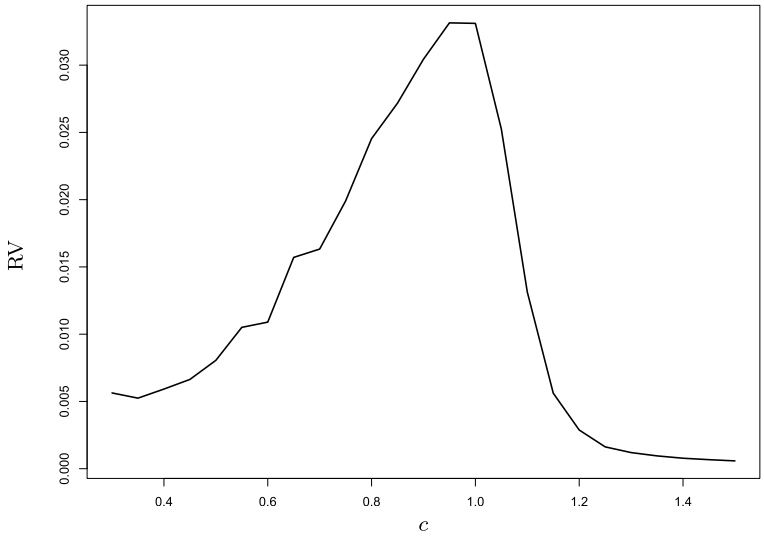

FIG 8. *Relative variation (RV) computed on the real data.*

the optimal value $c^* = \{\max c_i : RV_{c_k} < \iota, \text{ for all } c_k \leq c_i\}$, where $\iota > 0$ is some threshold value. Choosing the first point c_1 would be equivalent as using the non-robust method.

By design, this method leads to choose the tuning constant c the closest to c_1 (i.e. to the non-robust case) through a grid of points $c_1 > c_2 > ... > c_n$ given that the variation of the point estimates is smaller than an acceptable value ι. Intuitively, we find the closest tuning constant to the standard case while we require stability of our estimate.

In Figure 8 we display the outcomes of the application of the described algorithm to the real data example in Section 6.3. The relative variations (RV) are computed for the data over an equidistant spaced grid of tuning constants $c_1 = 1.5, c_2 = 1.4, ..., c_{27} = 0.2$. The plot suggest to choose $c = 0.5$: smaller values of c do not yield large changes in the RV, which remains pretty stable for $c < 0.5$.

B.3. RMSE for real-data analysis

To perform the sensitivity analysis, we randomly replace a percentage η of the original data $\{y_{t,j}\}$ by a random variable having a t-student distribution, with 5 degrees-of-freedom. For the sake of comparison, we compute the predicted value $\hat{y}^{\eta}_{t,j}$ for different values of $\eta \in \{0\%, 2.5\%, 5\%, 7.5\%, 10\%\}$ and, for each method, we calculate the Relative Mean Square Error, defined as

$$\text{RMSE}(\eta) = \sum_{t=1}^{T}\sum_{j=1}^{J}(\hat{y}^{\eta}_{t,j} - \hat{y}^{0}_{t,j})^2 / \sum_{t=1}^{T}\sum_{j=1}^{J}(\hat{y}^{0}_{t,j})^2,$$

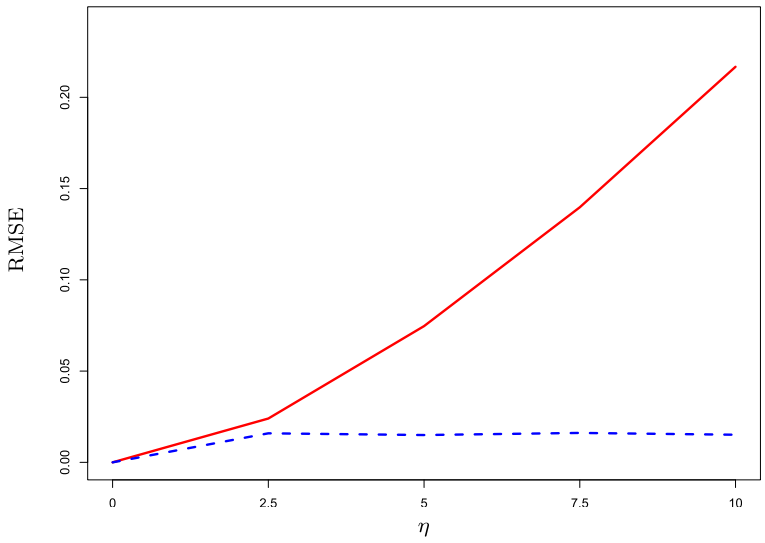

FIG 9. *RMSE(η) for robust (dashed line) and OLS (continuous line) estimation method, for different values of η (on the x-axis).*

where $\hat{y}^{\eta}_{t,j}$ denotes the predicted value of $y_{t,j}$ using the η-contaminated data. In Figure 9, we display the results for the robust and least squares procedures. In line with the theoretical developments, the robust estimates yield a stable and bounded RMSE, whilst the OLS estimator entails a RMSE which diverges (exponentially) fast as η grows. The plot confirms that even a small amount of data perturbation may induce large changes in the OLS estimates, whilst our robust procedure yields stable inference.

References

[1] Avella-Medina, M. and Ronchetti, E. (2017). Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika*, 105(1):31–44. MR3768863

[2] Cao, Y., Huang, J., Liu, Y., and Zhao, X. (2016). Sieve estimation of cox models with latent structures. *Biometrics*, 72(4):1086–1097. MR3591593

[3] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

[4] Chen, X., Liao, Z., and Sun, Y. (2014). Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics*, 178:639–658. MR3132457

[5] Chen, X., Racine, J., and Swanson, N. R. (2001). Semiparametric ARX neural-network models with an application to forecasting inflation. *IEEE Transactions on Neural Networks*, 12(4):674–683.

[6] Croux, C. and Haesbroeck, G. (2000). Principal component analysis based

on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618. MR1789812

[7] Fan, J. and Kim, D. (2017). Robust high-dimensional volatility matrix estimation for high-frequency factor model. *Journal of the American Statistical Association*, (just-accepted). MR3862356

[8] Fan, J., Li, Q., and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265. MR3597972

[9] Fan, J., Wang, K., Zhong, Y., and Zhu, Z. (2018). Robust high dimensional factor models with applications to statistical machine learning. *arXiv preprint arXiv:1808.03889*. MR4255196

[10] Fengler, M. R., Härdle, W. K., and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics*, 5(2):189–218. MR2183565

[11] Fernholz, L. T. (1983). *Von Mises calculus for statistical functionals*. Springer Science & Business Media. MR0713611

[12] Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4):540–554.

[13] Geier, C. F., Terwilliger, R., Teslovich, T., Velanova, K., and Luna, B. (2010). Immaturities in reward processing and its influence on inhibitory control in adolescence. *Cerebral Cortex*, 20(7):1613–1629.

[14] Grenander, U. (1981). *Abstract inference*. Wiley New York. MR0599175

[15] Hallin, M. and Lippi, M. (2013). Factor models in high-dimensional time series—a time-domain approach. *Stochastic processes and their applications*, 123(7):2678–2695. MR3054541

[16] Hampel, F., Ronchetti, Elvezio Rousseeuw, P., and Stahel, W. (1986). *Robust statistics: the approach based on Influence Functions*. Wiley. MR0829458

[17] Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635. MR2012827

[18] Kokoszka, P. and Reimherr, M. (2013). Asymptotic normality of the principal components of functional time series. *Stochastic Processes and their Applications*, 123(5):1546–1562. MR3027890

[19] Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer, New York. MR2724368

[20] La Vecchia, D., Camponovo, L., and Ferrari, D. (2015). Robust heart rate variability analysis by generalized entropy minimization. *Computational Statistics & Data Analysis*, 82:137–151. MR3282171

[21] Lazar, N. (2008). *The statistical analysis of functional MRI data*. Springer Science & Business Media.

[22] Muller, N. Z. and Phillips, P. C. (2008). Sinusoidal modeling applied to spatially variant tropospheric ozone air pollution. *Environmetrics*, 19(6):567–581. MR2528541

[23] Park, B. U., Mammen, E., Härdle, W., and Borak, S. (2009). Time series

modelling with semiparametric factor dynamics. *Journal of the American Statistical Association*, 104(485):284–298. MR2504378

[24] Peña, D. and Yohai, V. J. (2016). Generalized dynamic principal components. *Journal of the American Statistical Association*, 111(515):1121–1131. MR3561936

[25] Salibián-Barrera, M., Van Aelst, S., and Willems, G. (2006). Principal components analysis based on multivariate mm estimators with fast and robust bootstrap. *Journal of the American Statistical Association*, 101(475):1198–1211. MR2328307

[26] Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615. MR1292531

[27] Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334. MR2847798

[28] van Bömmel, A., Song, S., Majer, P., Mohr, P. N., Heekeren, H. R., and Härdle, W. K. (2014). Risk patterns and correlated brain activities. Multidimensional statistical analysis of fMRI data in economic decision making study. *Psychometrika*, 79(3):489–514. MR3257499

[29] van de Geer, S. (2000). *Empirical processes in M-estimation*. Cambridge university Press.

[30] van de Geer, S. (2002). M-estimation using penalties or sieves. *Journal of Statistical Planning and Inference*, 108(1-2):55–69. MR1947391

[31] Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer. MR1385671

[32] Van Dijk, K. R., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage*, 59(1):431–438.

[33] Zhou, Q., Zhou, H., and Cai, J. (2017). Case-cohort studies with interval-censored failure time data. *Biometrika*, 104(1):17–29. MR3626480