Bernoulli **23**(4B), 2017, 3178–3196
DOI: 10.3150/16-BEJ841

Influence functions for penalized M-estimators

MARCO AVELLA-MEDINA

GSEM and Research Center for Statistics, University of Geneva, Boulevard du Pont d'Arve 40, CH-1211 Geneva, Switzerland. E-mail: marco.avella@unige.ch

We study the local robustness properties of general nondifferentiable penalized M-estimators via the influence function. More precisely, we propose a framework that allows us to define rigorously the influence function as the limiting influence function of a sequence of approximating estimators. We show that it can be used to characterize the robustness properties of a wide range of sparse estimators and we derive its form for general penalized M-estimators including lasso and adaptive lasso type estimators. We prove that our influence function is equivalent to a derivative in the sense of distribution theory.

Keywords: distribution theory; implicit function theorem; lasso; regularization; robust statistics

1. Introduction

Sparse models have become very popular in recent years. Since the introduction of penalized methods to study them in the linear model (Breiman [5], Tibshirani [28]), many extensions and algorithms have been proposed. Fan and Lv [11] and Tibshirani [29] provide good reviews. Asymptotic properties of lasso-type estimators have been studied in the fixed dimensional parameter case (Knight and Fu [19], Fan and Li [10], Zou [36]), as well as in the high dimensional set up where the number of parameters is allowed to grow at an even faster rate than the sample size (Bühlmann and van de Geer [6]).

Given the increasing importance that sparsity inducing penalties play in modern statistics, the need for a clear understanding of the robustness properties of these type of procedures is evident. Robust statistics develops a theoretical framework that allows us to take into account that the models used for fitting the data are only idealized approximations of reality. It provides methods that still give reliable results when slight deviations from the stochastic assumptions on the model occur. Book-length expositions can be found in Huber [17] and second edition by Huber and Ronchetti [18], Hampel *et al.* [13] and Maronna *et al.* [24].

Some authors have suggested sparse estimators that limit the impact of contamination in the data (e.g., Sardy *et al.* [26], Wang *et al.* [31], Li *et al.* [20], Lozano and Meinshausen [22] and Fan *et al.* [9] among many others). These procedures rely on the intuition that a loss function that defines robust estimators in the well understood unpenalized fixed dimensional M-estimation set up, should also define robust estimators when it is penalized by a deterministic function. In the linear model for instance, Fan *et al.* [9] show that under very mild conditions on the error term their estimator satisfies the oracle properties.

One of the main lines of research in the robustness literature was opened by Hampel [12] who considered local robustness, that is, the impact of moderate distributional deviations from ideal

1350-7265 © 2017 ISI/BS

models on a statistical procedure. In this setting, the quantities of interest are viewed as functionals of the underlying distribution. Typically their linear approximation is studied to assess the behavior of estimators in a neighborhood of the model. In this approach the influence function plays a crucial role in describing the local stability of the functional analyzed. It allows for an easy assessment of the relative influence of individual observations on the value of an estimate. If it is unbounded, a single outlier could cause trouble. If a statistical functional $T(F)$ is sufficiently regular, a von Mises expansion (von Mises [30]) yields

$$T(G) \approx T(F) + \int \mathrm{IF}(z; F, T)\,\mathrm{d}(G - F)(z), \tag{1.1}$$

where $\mathrm{IF}(z; F, T)$ denotes the influence function of the functional T at the distribution F. Considering the approximation (1.1) over an ε neighborhood of the model $\mathcal{F}_\varepsilon = \{G | G = (1 - \varepsilon)F + \varepsilon H, H \text{ arbitrary}\}$, we see that the influence function can be used to linearize the asymptotic bias in a neighborhood of the ideal model. Therefore, a bounded influence function implies a bounded approximate bias.

The goal of this article is to give a formal definition of the influence function for a wide class of penalized M-estimators, that covers most of the existing proposals. This requires developing a new framework. Indeed, the typical tools used to derive the influence function of M-estimators suffer from a major problem when considering penalized M-estimators: they cannot handle non-differentiable penalty functions which are necessary for achieving sparsity (Fan and Li [10]). Note that in a recent paper Wang *et al.* [33] give the form of the influence function for their penalized M-estimator. However, it is derived in a limited set up without developing an appropriate rigorous framework. Further details can be found in Section 5.

Our work provides a number of contributions to the existing literature. First, we introduce an influence function defined through a sequence of approximating functionals and show that it is uniquely defined for penalized M-estimators and two-stage penalized M-estimators. The former class covers lasso and group lasso type estimators. The latter includes adaptive lasso type estimators. We compute the influence function of all these important examples. Second, we show that the two main features of the influence functions for M-estimators can also be valid for the influence function of penalized M-estimators, that is, (a) it allows to assess the relative influence of individual observations towards the value of an estimate; (b) it allows an immediate and simple, informal assessment of the asymptotic properties of an estimate (Huber and Ronchetti [18], pages 14–15). Third, we show that our limiting influence function can be viewed as a distributional derivative in the sense of Schwartz [27]. This opens the door for further research exploiting the tools of distribution theory, which to the best of our knowledge, has essentially not been used in the statistical literature previously. Finally, a key step in our theoretical argument is the innovative use of Berge's maximum theorem. This is a powerful tool that could have more applications in statistics.

The rest of the article is organized as follows. Section 2 introduces the general framework and provides the main results regarding our influence function as well as some examples. Section 3 extends the results to two step penalized M-estimators. Section 4 provides some numerical illustrations of the local robustness of penalized estimators and its relation to our influence function. Section 5 establishes the connection between our influence function and distributional derivatives. Finally, Section 6 concludes by discussing some potential research directions. A selection of the proofs of our main results is given at the end of the article. Additional proofs and auxiliary results are provided in the supplementary material Avella-Medina [1].

2. Penalized M-estimators and influence functions

2.1. Background

Consider a collection of n observations $\{z_1, \dots, z_n\}$ drawn from a common distribution F over the space $\mathcal{Z}$, a parameter space Θ which is an open subset of $\mathbb{R}^d$ and a loss function $L : \mathcal{Z} \times \mathbb{R}^d \mapsto \mathbb{R}$. Let $\hat{F} = \frac{1}{n}\sum_{i=1}^n \Delta_{z_i}$ be the empirical distribution of F corresponding to the observed sample, where Δ_z is the distribution probability that assigns mass 1 at the point z and 0 elsewhere. Then the value $E_{\hat{F}}[L(Z,\theta)] = \frac{1}{n}\sum_{i=1}^n L(z_i,\theta)$ serves as a measure of fit between the a parameter vector $\theta \in \mathbb{R}^p$ and the observed data. This empirical loss function can be seen as an estimator of the unknown population risk function $E_F[L(Z,\theta)]$. We study the functionals resulting from the minimization of the regularized risk

$$\Lambda_\lambda(\theta; F) = E_F\big[L(Z,\theta)\big] + p(\theta;\lambda) \tag{2.1}$$

with respect to θ, where $p(\cdot;\lambda)$ is a continuous penalty function with regularization parameter λ. Important illustrations of such functionals are considered in Examples 1, 2 and 3 of this paper.

Let $T(F) = \theta_{F,\lambda} = \theta^* = \operatorname{argmin}_\theta \Lambda_\lambda(\theta; F)$. In robust statistics, we are interested in constructing and studying smooth and bounded functionals T, because they yield stable minimizers within small neighborhoods of F. A bounded derivative $\nabla T(F)$ of $T(F)$ implies that the functional $T(F)$ cannot vary arbitrarily in small neighborhoods of F. One general approach to robustness is the one based on the influence function (Hampel *et al.* [13]). Given a distribution space $\mathcal{F}$, a parameter space Θ and a functional $T : \mathcal{F} \mapsto \Theta$, the influence function of T at a point $z \in \mathcal{Z}$ for a distribution F is defined as

$$\mathrm{IF}(z; F, T) = \lim_{\varepsilon\to 0+} \frac{T(F_\varepsilon) - T(F)}{\varepsilon},$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_z$. It has the heuristic interpretation of describing the effect of an infinitesimal contamination at the point z on the estimate, standardized by the mass of contamination. The well studied M-functionals are defined implicitly as a root of

$$\int \Psi\big(z, T(F)\big)\,\mathrm{d}F = 0.$$

They are a special case of the minimizers of (2.1) when the penalty function is set to be zero and the risk function is differentiable with respect to the parameter. The standard argument for showing the existence and deriving the form influence function of M-functionals, is to use an appropriate implicit function theorem. In our set up this would require that Λ_λ has two derivatives respect to θ. However, it is well known that a penalty function has to be singular at the origin in order to achieve sparsity (Fan and Li [10] and Negahban *et al.* [25]). Therefore new tools that can deal with nondifferentiable penalty functions are required to derive influence functions of many modern penalized estimators. We propose to define the influence function as the limit of the influence functions of a sequence of differentiable penalized M-estimators that converge to the penalized M-estimator of interest.

2.2. Limiting influence function

We will require the following set of assumptions for the derivation of our theoretical results:

(A1) $E_F[L(Z,\theta)]$ is continuous in Θ and has two bounded derivatives with respect to θ denoted by $E_F[\psi(Z,\theta)]$ and $E_F[\dot{\psi}(Z,\theta)]$ respectively, with $E_F[\dot{\psi}(Z,\theta)]$ nonsingular.

(A2) There is a unique point $\theta_0 \in \Theta$ such that $E_F[L(Z,\theta_0)] < \inf_{\theta\in\Theta\setminus U} E_F[L(Z,\theta)]$ for every open set U containing θ_0.

(A3) There is a unique θ^* in an open ball $\mathcal{O}$ containing θ_0. The functions $\psi(z,\theta)$ and $E_F[\dot{\psi}(Z,\theta)]$ are continuous at θ^* for all $z \in \mathbb{R}^d$.

Assumptions (A1) and (A2) are regularity conditions on the population risk. Namely, they require some smoothness and a well separated maximum. Assumption (A3) imposes regularity conditions on the score function ψ and a unique regularized risk minimum on a neighborhood of the global risk minimum. Obviously when both the population risk and the penalty function are convex we can take $\mathcal{O} = \Theta$. Nonconvex loss functions and penalty functions restrain the size of the ball $\mathcal{O}$.

As discussed in the previous subsection, when the penalty function is sufficiently smooth, a simple application of the implicit function theorem establishes the existence and the form of the influence function of the minimizer of (2.1). This result is stated in the following lemma.

Lemma 1. *Assume* (A1)–(A3). *Let* $p : \Theta \to \mathbb{R}$ *be twice differentiable and* $S := E_F[\dot{\psi}(Z,\theta^*)] + \nabla^2 p(\theta^*;\lambda)$ *be invertible. Then the influence function of* $T(F)$ *exists for all* $z \in \mathbb{R}^d$ *and we have*

$$\mathrm{IF}(z;F,T) = -S^{-1}\big(\psi\big(z,\theta^*\big) + \nabla p\big(\theta^*;\lambda\big)\big).$$

Note that S is a fixed matrix that depends on the distribution F as well as on the loss and penalty functions. From Lemma 1, we conclude that just as for M-estimators, a bounded derivative for the loss function is required in order to obtain bounded influence estimators in the penalized setting. When the penalty function in (2.1) is not differentiable, the conditions of Lemma 1 do not hold. We therefore propose to study the limiting form of the influence function of penalized M-estimators obtained using smooth penalty functions p_m such that $\lim_{m\to\infty} p_m = p$. Such penalized M-estimators, denoted by $T(F;p_m)$, are defined as the minimizers of

$$\Lambda_\lambda(\theta;F,p_m) = E_F\big[L(Z,\theta)\big] + p_m(\theta;\lambda). \tag{2.2}$$

We let $\mathrm{IF}_{p_m}(z;T,F)$ be the influence function of $T(F;p_m)$ and define the influence function of $T(F)$ as

$$\mathrm{IF}(z;F,T) := \lim_{m\to\infty} \mathrm{IF}_{p_m}(z;F,T). \tag{2.3}$$

A natural question that arises from this definition is whether the limit depends on the limiting sequence $\{p_m\}$ chosen. In order to answer this question, we first show that the limiting functional $T(F)$ is unique. The use of Berge's maximum theorem (Berge [4]) is a key step in our proof. It is an innovative tool in the statistical literature. The lemma is crucial for the uniqueness argument of the limiting influence function, stated in Proposition 1. While completing this article, the author noticed that Machado [23] had also used Berge's maximum theorem for the derivation of qualitative robustness for model selection criteria based on M-estimators.

Lemma 2. *Let $C^{\infty}(\Theta) = \{f : \Theta \to \mathbb{R} |\ f$ continuous and infinitely differentiable$\}$ and consider a sequence $\{p_m\}_{m\geq 1} \in C^{\infty}(\Theta)$ converging to p in the Sobolev space $W^{2,2}(\Theta)$ when $m \to \infty$. Suppose that for the problem* (2.1)*, each of the approximating problems* (2.2) *resulting from $\{p_m\}_{m\geq 1}$ satisfy* (A1)–(A3). *Then we have* $\lim_m T(F; p_m) = T(F)$.

Proposition 1. *Under the conditions of Lemma* 2*, the limiting influence function defined in* (2.3) *does not depend on the choice of p_m.*

Remark 2.1. The uniqueness of (2.3) holds for any local minimum as long as they are well separated. Indeed (A3) can be relaxed by simply considering a unique minimizer contained in an open ball $\mathcal{O}$ that does not necessarily contain θ_0. We can see from the proofs of Lemma 2 and Proposition 1 that this is enough for them to hold.

Example 1 (Lasso type penalties). Without loss of generality, we will assume that the tuning parameter λ is such that the resulting estimators are sparse. More specifically, we consider $\theta_{F,\lambda} = \theta^* = (\theta_1^{*T}, \theta_2^{*T})^T$ with $\theta_1^* \in \mathbb{R}^s$, $s < d$ and $\theta_2^* = 0$. The following proposition gives the form of the influence function of estimators that arise when considering a general class of penalty functions. It covers as special cases convex penalties such as the lasso (Tibshirani [28]) and nonconvex penalties such as the scad (Fan and Li [10]).

Proposition 2. *Denote by $\theta^* = T(F)$ the penalized M-functional obtained as the minimizer of* (2.1) *with penalty functions of the form $p_\lambda(\theta) = \sum_{j=1}^{p} p_{\lambda,j}(|\theta_j|)$, where $p_{\lambda,j}(\cdot)$ are differentiable functions. Then under* (A1)–(A3) *the influence function* (2.3) *of $T(F)$ has the form*

$$\mathrm{IF}(z; F, T) = -S^{-1}\big(\psi(z, \theta^*) + \phi_\lambda(\theta^*)\big),$$

where $S^{-1} = \mathrm{blockdiag}\{(M_{11} + P_\lambda)^{-1}, 0\}$, $M_{11} = E_F[\dot{\psi}_{11}(Z, T(F))]$, *$P_\lambda$ is a diagonal matrix with diagonal elements $p''_{\lambda,j}(|\theta_j^*|)$ for $j = 1, \ldots, s$, and $\phi_\lambda(\theta^*)$ is a d dimensional vector with components* $p'_{\lambda,j}(|\theta_j^*|)\,\mathrm{sgn}(\theta_j^*)$ *for $j = 1, \ldots, s$ and* 0 *elsewhere.*

Example 2 (Group lasso type penalties). We now give the form of the influence function of penalized M-estimators achieving sparsity for grouped variables via group lasso type penalties (e.g., Yuan and Lin [34], Wang *et al.* [32], Huang *et al.* [14]). We suppose, as in the previous example, that the regularization parameter λ is such that the resulting regularized θ^* is sparse.

Proposition 3. *Denote by $\theta^* = T(F)$ the penalized M-functional obtained as the minimizer of* (2.1) *with group penalty functions of the form $p_\lambda(\theta) = \sum_{g=1}^{G} p_{\lambda,g}(\|\theta_{(g)}\|_2)$, where $p_{\lambda,g}(\cdot)$ are* *differentiable functions, $\theta = (\theta_{(1)}, \dots, \theta_{(G)})$ and each $\theta_{(g)}$ is a subvector of θ corresponding to the gth group of variables. Then under* (A1)–(A3) *the influence function* (2.3) *of $T(F)$ has the form*

$$\mathrm{IF}(z; F, T) = -S^{-1}\big(\psi\big(z, \theta^*\big) + \phi_\lambda\big(\theta^*\big)\big),$$

where $S^{-1} = \text{blockdiag}\{(M_{11} + P_\lambda)^{-1}, 0\}$, $M_{11} = E_F[\dot{\psi}_{11}(Z, T(F))]$ and P_λ is a block diagonal matrix with blocks $p''_{\lambda,g}(\|\theta_{(g)}\|_2)(\theta_{(g)}\theta_{(g)}^T - \|\theta_{(g)}\|_2 I_{|g|})/\|\theta_{(g)}\|_2^3$ where

$$\phi_\lambda\big(\theta^*\big) = \begin{cases} p'_{\lambda,g}\big(\|\theta^*_{(g)}\|\big)\theta^*_j/\|\theta^*_{(g)}\|_2, & \textit{if}\, \theta^*_{(g)} \neq 0, \\ 0, & \textit{if}\, \theta^*_{(g)} = 0. \end{cases}$$

$I_{|g|}$ is a diagonal matrix of size $|g|$, i.e. the cardinality of group g, and $p'_{\lambda,g}(t)$ and $p''_{\lambda,g}(t)$ are the first two derivatives of $p_{\lambda,g}(t)$ for $t > 0$.

It is clear from Propositions 2 and 3 that a bounded ψ function is necessary for a Penalized M-estimator to have bounded limiting influence function. The fact that the influence function has some zero components is rather surprising. Further discussion on this feature can be found in the last two sections.

3. Two-stage penalized M-estimators

We can extend the results of the previous section to a class of two-stage penalized M-estimators. Important examples of the adaptive lasso estimators are discussed below. The following set up can be viewed as a direct extension of the framework provided by Zhelonkin *et al.* [35]. Let F be the distribution function of $Z = (Z^{(1)}, Z^{(2)})$ and let $\theta = (\theta_1, \theta_2)$ be a vector defining the arguments of the first and second stages, with $\theta_1 \in \Theta_1 \subset \mathbb{R}^{d_1}$, $\theta_2 \in \Theta_2 \subset \mathbb{R}^{d_2}$. We consider penalized M-estimators $(\theta_1^*, \theta_2^*) = (S(F), T(F))$ defined by

$$\theta_1^* = \operatorname*{argmin}_{\theta_1}\big\{E_F\big[L^{(1)}\big(Z^{(1)}, \theta_1\big)\big] + p^{(1)}(\theta_1; \gamma)\big\}, \tag{3.1}$$

$$\theta_2^* = \operatorname*{argmin}_{\theta_2}\big\{E_F\big[L^{(2)}\big(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1^*\big)\big] + p^{(2)}\big(\theta_2, \theta_1^*; \lambda\big)\big\}, \tag{3.2}$$

where $L^{(i)}$ and $p^{(i)}$ denote respectively, the loss and penalty functions in the ith stage. For the theoretical argument, we adapt assumptions (A1)–(A3) to this set up in the following straightforward way:

(A1′) The loss function $L^{(1)}$ and $L^{(2)}$ are such that:

(i) $E_F[L^{(1)}(Z^{(1)}, \theta_1)]$ is continuous in Θ_1 and has two derivatives with respect to θ_1 denoted by $E_F[\psi^{(1)}(Z^{(1)}, \theta_1)]$ and $E_F[\dot{\psi}^{(1)}(Z^{(1)}, \theta_1)]$ respectively, with $E_F[\dot{\psi}^{(1)}(Z^{(1)}, \theta_1)]$ nonsingular.

(ii) $E_F[L^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)]$ is continuous in $\Theta_1 \times \Theta_2$ and has two derivatives with respect to θ_2 denoted respectively, by $E_F[\psi^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)]$ and

$E_F[\dot\psi^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)]$, where $E_F[\dot\psi^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)]$ is nonsingular. $E_F[\psi^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)]$ has a derivative with respect to θ_1 denoted by $E_F[\psi_1^{(2)}(Z^{(2)}, \theta_2, Z^{(1)}, \theta_1)]$.

(A2′) There are unique points $\theta_{01} \in \Theta_1$ and $\theta_{02} \in \Theta_2$ such that:

(i) $E_F[L^{(1)}(Z^{(1)}, \theta_{01})] < \inf_{\theta_1 \in \Theta_1 \setminus U_1} E_F[L^{(1)}(Z^{(1)}, \theta_1)]$ for every open set U_1 containing θ_{01}.

(ii) $E_F[L^{(2)}(Z^{(2)}, \theta_{02}, Z^{(1)}, \theta_{01})] < \inf_{\theta_2 \in \Theta_2 \setminus U_2} E_F[L^{(2)}(Z^{(2)}, \theta_{02} Z^{(1)}, \theta_{01})]$ for every open set U_2 containing θ_{02}.

(A3′) There are open subsets of Θ_1 and Θ_2 denoted by $\mathcal{O}_1$ and $\mathcal{O}_2$, and unique solutions $\theta_1^* \in \mathcal{O}_1$ and $\theta_2^* \in \mathcal{O}_2$ such that:

(i) $\theta_{01} \in \mathcal{O}_1$, and the functions $\psi^{(1)}(z^{(1)}, \theta_1)$ and $E_F[\dot\psi^{(1)}(Z^{(1)}, \theta_1)]$ are continuous at θ_1^* for all $z^{(1)} \in \mathbb{R}^{d_1}$.

(ii) $\theta_{02} \in \mathcal{O}_2$, and the functions $\psi^{(2)}(z^{(2)}, \theta_2, z^{(1)}, \theta_1)$ and $E_F[\dot\psi^{(2)}(Z^{(2)}, \theta_2; Z^{(1)}, \theta_1)]$ are continuous at θ^* for all z.

We first provide the influence function of (3.2) for sufficiently smooth penalty functions.

Lemma 3. *Denote by $\theta^* = (S(F), T(F))$ the estimators defined by (3.1)–(3.2), and assume (A1′)–(A3′). Let $p^{(i)} : \Theta_i \to \mathbb{R}$ be twice differentiable with respect to θ_i and $\nabla_{\theta_2} p^{(2)}$ be differentiable with respect to θ_1. Let also $S := E_F[\dot\psi^{(2)}(Z^{(2)}, \theta_2^*, Z^{(1)}, \theta_1^*)] + \nabla_{\theta_2,\theta_2} p^{(2)}(\theta_2^*, \theta_1^*; \lambda)$ be invertible. Then the influence function of $T(F)$ exists for all $z \in \mathbb{R}^d$ and we have*:

$$\begin{aligned}\mathrm{IF}(z; F, T) = {}& -S^{-1}\big(\psi^{(2)}\big(Z^{(2)}, \theta_2^*, Z^{(1)}, \theta_1^*\big) + \nabla_{\theta_2} p^{(2)}\big(\theta_2^*, \theta_1^*; \lambda\big) \\ & + \big(E_F\big[\psi_1^{(2)}\big(Z^{(2)}, \theta_2^*, Z^{(1)}, \theta_1^*\big)\big] + \nabla_{\theta_2,\theta_1} p^{(2)}\big(\theta_2^*, \theta_1^*; \lambda\big)\big)\,\mathrm{IF}\big(z^{(1)}; F, S\big)\big),\end{aligned}$$

where $\mathrm{IF}(z^{(1)}; S, F)$ has the form given in Lemma 1.

Unsurprisingly, bounded-influence estimators are obtained only by taking loss functions with bounded derivatives in both stages. The expression obtained is very similar to the one derived in the unpenalized set up by Zhelonkin *et al.* [35]. We are now ready to state the uniqueness of the limiting two-stage estimator (3.2) and its influence function.

Lemma 4. *For $i = 1, 2$, let $\{p_{m_i}^{(i)}\}$ be a sequence in $C^\infty(\Theta_i)$ converging to $p^{(i)}$ in the Sobolev space $W^{2,2}(\Theta_i)$ when $m_i \to \infty$ and assume $\{\nabla_{\theta_2} p_m^{(2)}\}$ is differentiable with respect to θ_1. Consider the sequence of problems of the (2.1) implied by (3.1)–(3.2) and $\{p_{m_i}^{(i)}\}$ for $i = 1, 2$. Assume that each of them satisfy (A1′)–(A3′). Then we have $\lim_m T(F; p_m^{(2)}) = T(F)$.*

Proposition 4. *Under the conditions of Lemma 4, the limiting influence function (2.3) of (3.2) does not depend on the choice of $p_m = (p_m^{(1)}, p_m^{(2)})$.*

Example 3 (Adaptive lasso type penalties). The adaptive lasso of Zou [36] is a popular two stage procedure that improves on the results of the lasso by ensuring the oracle properties of Fan and Li [10] under milder conditions. The tools developed above allow us to derive the influence function of adaptive lasso type estimators.

Proposition 5. *Let θ be the d dimensional parameter of interest and let $\theta^{(0)} = S(F)$ be an initial estimate of θ, with $s^{(0)}$ non zero components, defined by* (3.1). *For $j = 1, \dots, d$ and some nonnegative function w, define the weights $w_j = w(|\theta_j^{(0)}|)$. Denote by $\theta^* = T(F)$ the penalized M-estimators obtained as the minimizer of* (3.2) *with a penalty function of the form $p(\theta, \theta^{(0)}; \lambda) = \lambda \sum_{j=1}^{p} w_j |\theta_j|$ and loss function $L(Z, \theta, \theta^{(0)})$. Then under* (A1′)–(A3′) *the influence function* (2.3) *of $T(F)$ has the form*

$$\mathrm{IF}(z; F, T) = -S^{-1}\big(\psi\big(z, \theta^*\big) + \phi_\lambda\big(\theta^*, \theta^{(0)}\big) + \varphi_\lambda\big(\theta^*, \theta^{(0)}\big)\,\mathrm{IF}(z; F, S)\big),$$

where $S^{-1} = \mathrm{blockdiag}\{M_{11}^{-1}, 0\}$, $M_{11} = E_F[\dot{\psi}^{(1)}(Z, T(F))]$, $\phi_\lambda(\theta^*, \theta^{(0)})$ is a d dimensional vector with components $\lambda w'(|\theta_j^{(0)}|)\,\mathrm{sgn}(\theta_j^{(0)})\,\mathrm{sgn}(\theta_j^*)$ for $j = 1, \dots, d$ where w' denotes the derivative of w, and $\varphi_\lambda(\theta^*, \theta^{(0)}) = \mathrm{blockdiag}\{H_\gamma, 0\}$ is a matrix with $H_\gamma = \{\nabla_{\theta_k^{(0)}} \phi_{\lambda,j}(\theta^*, \theta^{(0)})\}_{j,k=1}^{s^{(0)}}$.

The form of $\mathrm{IF}(z; F, T)$ depends on the choice of the initial estimator. A bounded influence estimator $T(F)$ can only be obtained by taking a bounded influence initial estimator $S(F)$ and choosing a loss function defining a bounded ψ function in the second stage. Among the non-sparse initial estimates proposed in the literature, Zou [36] proposed to use maximum likelihood estimates for the fixed parameter case where d does not vary with n. In the high dimensional set up where $d > n$, Huang *et al.* [15] proposed to use an initial zero consistent estimate, e.g. marginal least squares. For those cases the influence function of $S(F)$ is simply proportional to $\psi^{(1)}$ as they are well-known M-estimators. Lasso estimators have also been proposed as initial estimates. See, for instance, Fan *et al.* [9] and Avella-Medina and Ronchetti [2], in the context of high dimensional penalized likelihood and robust quasilikelihood estimation, respectively. Note that for $w_j = 1/|\theta_j^{(0)}|$ the usual convention is that for $\theta_j^{(0)} = 0$ we define $w_j = \infty$ and $\infty \cdot 0 = 0$. Hence for this choice of w_j, a coefficient set to zero in the first step will never appear in $T(F)$.

4. Numerical illustrations

4.1. Orthogonal design

In the special case of the linear model with orthogonal design, the asymptotic bias of the lasso, group lasso and adaptive lasso estimators under a contamination neighborhood $(1 - \varepsilon)F + \varepsilon G$ can be computed explicitly. In particular, it can be seen that if $h = \int X_j Y \,\mathrm{d}(G - F)$ is bounded and ε is sufficiently small, the bias of jth component of the lasso estimator $\hat{\theta}_j$ is very well approximated by the limiting influence function. A similar conclusion can be reached for the bias of the group and adaptive lasso estimators.

We illustrate this point in an orthogonal design linear model simulation study. Specifically, we simulated a sparse linear model $y_i = x_i^T \beta + u_i$ for $i = 1, \dots, 100$, where the x_i's are i.i.d. standard normal variables, $\beta = (1, 1, 1, 0, 0, 0, 0, 0, 0)^T$, $u_i \sim (1 - b) \cdot N(0, 1) + b \cdot N(0, 10)$ and $b \sim \mathrm{Bernoulli}(\varepsilon)$. Figure 1 illustrates how the influence function approximates the asymptotic bias of the lasso, group lasso and adaptive lasso estimators of β_1 for $\varepsilon \in (0, 0.2)$ and tuning

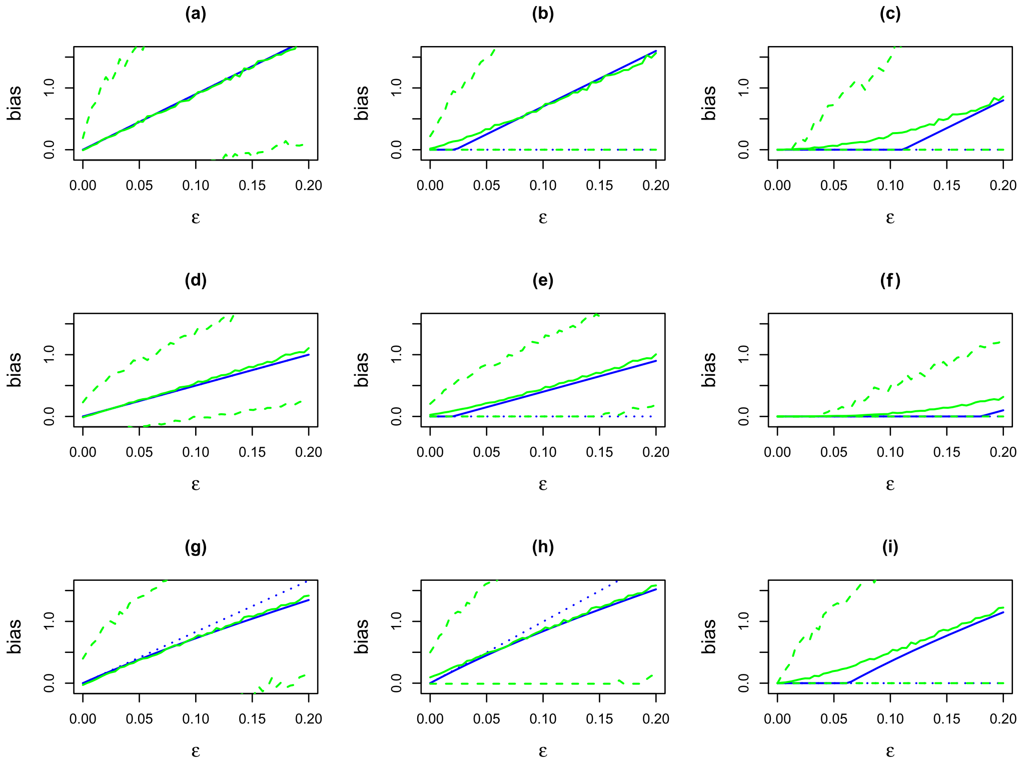

Figure 1. The solid dark curves are the bias of the respective minimizers of (2.1), the dotted lines are the bias approximations of the influence function, the solid light curves are the empirical mean biases and the dashed lines represent the quantiles 0.025 and 0.975 respectively. Plots (a)–(c) show the behavior of the bias of the lasso for the tuning parameters $\lambda = \{0.2, 1.2, 2\}$ and contamination parameter $\varepsilon \in (0, 0.2)$. Plots (d)–(f) and (g)–(i) show the same results for the group lasso and the adaptive lasso, respectively.

parameter $\lambda = \{0.2, 1.2, 2\}$. The green lines report the mean and the quantiles 0.025 and 0.975 of the empirical biases obtained over 1000 replications. For the group lasso, we considered the groups $(\beta_1, \beta_2, \beta_3)$, $(\beta_4, \beta_5, \beta_6)$ and $(\beta_7, \beta_8, \beta_9)$. The least squares estimator was used to construct the weights of the adaptive lasso.

We see that the linearized asymptotic bias obtained from the influence function is quite accurate for small values of ε, especially for small and large values of λ. Given that the influence function gives an insensitive approximation to the asymptotic bias for the coefficients estimated as 0, the accuracy of this approximation will depend on how much contamination the estimator can absorb before the estimated null coefficients become non null. If a penalized estimator is insensitive to contamination in terms of the model selected, then the linearized asymptotic bias will be exact. This question seems to be related to the definition of qualitative model selection robustness proposed by Machado [23] and is beyond the scope of this work.

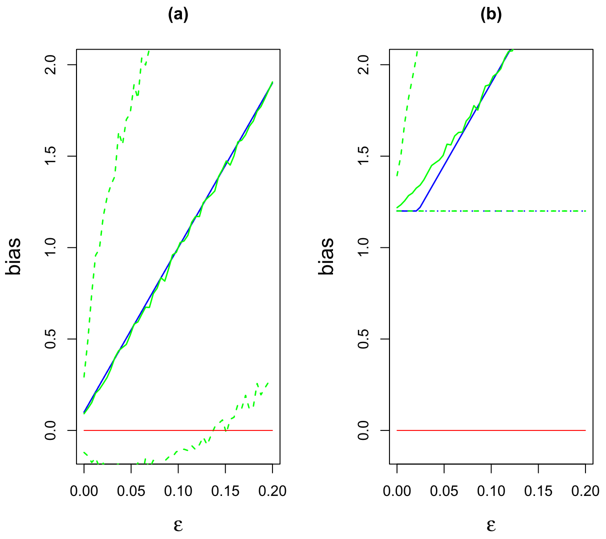

Figure 2. The solid horizontal lines indicate the zero bias benchmark with respect to the true value of the parameter β_1. The solid dark lines indicate the bias of the minimizers of (2.1), the dotted lines the bias approximation of the influence function, the solid light curvesthe empirical mean biases and the light dashed curves represent the quantiles 0.025 and 0.975, respectively.

Note that in theory the $\sqrt{(k \log p)/n}$ rates of convergence of penalized estimators are obtained for tuning parameters of order $O(\sqrt{(\log p)/n})$ (Loh and Wainwright [21]). Therefore, in practice we expect to choose small values of λ.

Figure 2 further stresses this point. It provides another picture of plots (a) and (b) showed in Figure 1 by taking into account the distance of the lasso functional and the true value of the parameter β_1. It is clear from (b) that $\lambda = 1.2$ is already too large a tuning parameter.

4.2. High dimensional Poisson regression

We consider now a more sophisticated example where an analytic expression of the asymptotic bias cannot be computed. We show instead how the estimators behave in terms of $L2$-loss, that is, $\|\hat{\beta} - \beta\|_2$ when the contamination increases. The results of Sections 2 and 3 imply that penalized M-estimators defined by a loss function having a bounded derivative will have a bounded influence function. Therefore such estimators are expected to have a bounded bias in a contamination neighborhood of the model. Typical likelihood based procedures on the other hand will in general have an unbounded score function and will therefore be very sensitive to small contaminations. We will show that this is exactly what happens when we consider a high dimensional regression problem. Specifically, we simulated a Poisson regression model with canonical link $g(\mu_i) = \log \mu_i = 1.8x_{i1} + x_{i2} + 1.5x_{i5}$, for $i = 1, \dots, 100$. The covariates x_{ij} were generated from standard uniforms with correlation $\text{cor}(x_{ij}, x_{ik}) = \rho^{|j-k|}$ and $\rho = 0.5$ for

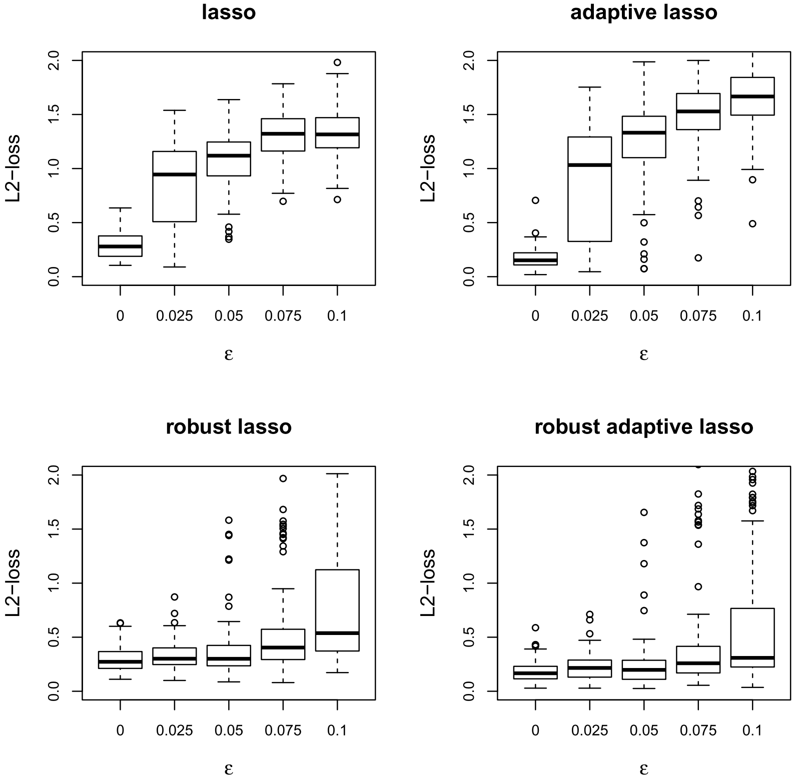

Figure 3. The boxplots show the performance measured by the $L2$-loss of classical and robust counterparts of the lasso and adaptive lasso as the contamination increases. In all the simulations, we set $\nu = 5$.

$j, k = 1, \dots, 250$. The response variables Y_i were generated according to a perturbed Poisson distribution of the form $(1 - b) \cdot \mathcal{P}(\mu_i) + b \cdot \mathcal{P}(\nu\mu_i)$ where $b \sim$ Bernoulli(ε). We set $\nu = 5, 10$ and $\varepsilon = \{0, 0.025, 0.05, 0.075, 0.1\}$. For each combination of ε and ν we generated 100 data sets. The tuning parameters were chosen by 5-fold cross-validation. We consider classical and robust counterparts of the lasso and the adaptive lasso based on the robust quasilikelihood of Cantoni and Ronchetti [7] as in Avella-Medina and Ronchetti [2]. The estimators where computed with the coordinate descent algorithm described in the latter paper.

Figures 3 and 4 clearly show that the robust penalized estimators are stable under moderate contamination whereas their classical counterparts are not. When $\nu = 10$, even very small contaminations completely ruin the performance of the classical procedures. It is interesting to note that this particular robust estimator seems that have its breakdown point somewhere around $\varepsilon = 0.075$ as its performance starts deteriorating at that point. This example illustrates the fact that the influence function only provides information about the local robustness properties.

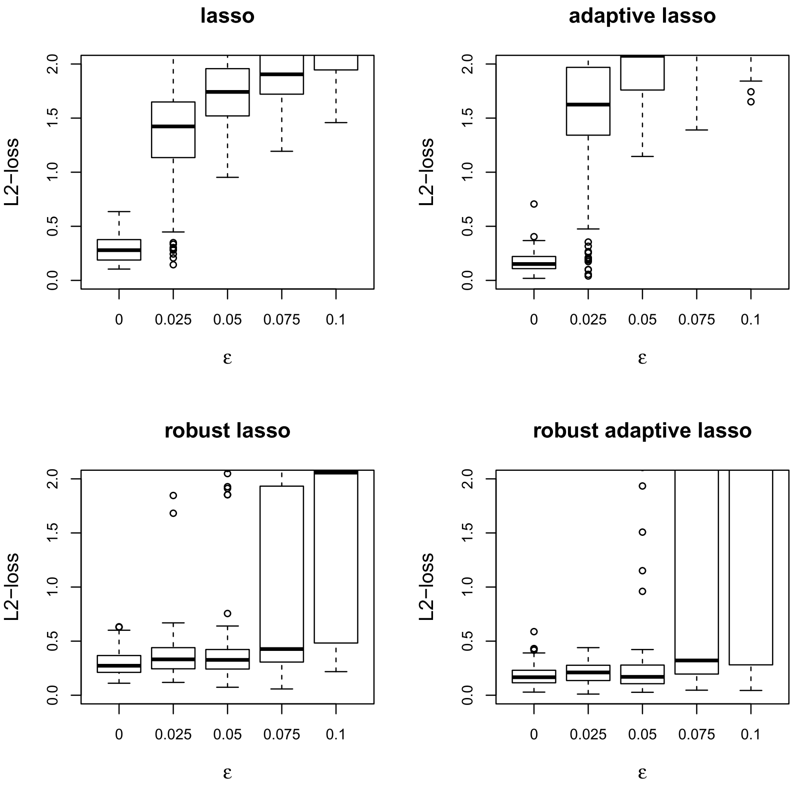

Figure 4. The boxplots show the performance measured by the $L2$-loss of classical and robust counterparts of the lasso and adaptive lasso as the contamination increases. In all the simulations we set $\nu = 10$.

5. Connections to distribution theory

Wang *et al.* [33] studied the robustness properties of their proposed robust penalized estimator by calculating its finite sample breakdown point and influence function. It can be seen that their derivation of the influence function (Theorem 3) could easily be extended to more general loss functions. In their proof they implicitly use distributions in the sense of Schwartz [27], since they require the first two derivatives of the absolute value function. They use $\operatorname{sgn}(x)$ as first derivative of $|x|$ and the Dirac delta function $\delta(x)$ as its second derivative. These derivatives are justified by the theory of distributions. However, working explicitly with the Dirac delta function understood informally as

$$\delta(x) = \begin{cases} +\infty, & \text{if } x = 0, \\ 0, & \text{otherwise}, \end{cases}$$

and inverting a matrix containing such an expression, is not fully satisfactory from a formal mathematical standpoint. Interestingly, the expression obtained in Theorem 3 in Wang *et al.* [33]

is the same as the one we give in Proposition 3. This suggests that a more careful treatment of the problem with a rigorous use of differentiation in the sense of distribution theory will yield the same influence function.

At first glance, the theory of distributions seems to provide a natural and rigorous way of tackling nondifferentiable penalties. However, the theory suffers from at least two major drawbacks for the purposes of deriving influence functions with direct computations. The product of two distributions cannot be consistently defined in general. This makes the manipulation of distributions delicate. Furthermore, to the best of our knowledge there is no implicit function theorem for distributions. This closes the door to the derivation of the influence function as the derivative of an implicit function as in Lemma 1. We relegate to the Appendix an example where we explicitly compute the distributional derivative of the lasso and scad functionals. Extending this approach to more general problems does not look obvious. We can however show that the influence functions derived in Section 2 using the limiting influence function can be viewed as distributional derivatives. Before giving this result in Proposition 6, we need an intermediate result that is interesting on its own and concerns the continuity of $T(F_\varepsilon)$ with respect to ε. Its proof uses Berge's maximum theorem and is similar to the proof of Lemma 2.

Lemma 5. *Under* (A1)–(A3), *the penalized M-estimator $T(F_\varepsilon)$ resulting from the minimization of* (2.1) *is continuous with respect to $\varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon})$, $\bar{\varepsilon} > 0$.*

Proposition 6. *Under the assumptions of Proposition* 1, *the influence function* (2.3) *of the minimizer $T(F)$ of* (2.1) *is the distributional derivative of $T(F_\varepsilon)$ with respect to ε evaluated at* 0.

6. Discussion

We introduced the idea of calculating the influence function of penalized M-estimators with the help of a sequence of approximating functionals. In Sections 2 and 3, we justified the validity of such an approach and derived the limiting influence functions of general penalized M-estimators. In particular, these computations show that the local robustness properties are a direct result of the form and boundedness of the derivative of the loss function. This reflects the intuition that the sources of local instability for M-estimators and their penalized counterparts should be the same. In Section 4, we illustrated via numerical examples the bias problem arising in a contamination neighborhood and its relation to the influence function. Finally, in Section 5, we showed that the influence function can be viewed as a derivative in the sense of distribution theory.

Let us conclude this paper with a short discussion of appealing future directions for research. We believe that the limiting influence function introduced in this paper could be useful for asymptotic distributional considerations. In the case of Fisher consistent M-estimators, substituting G by the empirical distribution $\hat{F}$ in (1.1), we have

$$\sqrt{n}\big(T(\hat{F}) - T(F)\big) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(z_i; F, T)$$

because $\int \mathrm{IF}(z; F, T)\, \mathrm{d}F(z) = 0$. Then by the central limit theorem we obtain that $\sqrt{n}(T(\hat{F}) - T(F))$ is asymptotically normally distributed with mean 0 and variance $V(F, T) = \int \mathrm{IF}(z;$

$F, T)\,\mathrm{IF}^T(z; F, T)\,\mathrm{d}F(z)$. A rigorous general argument can be found in Huber [16] and Huber and Ronchetti [18]. For M-estimators, the conditions for Fréchet differentiability of Clarke [8] guarantee the validity of the von Mises expansion and imply good robustness properties as discussed in Bednarski [3]. Using these formulas and the influence functions derived in Sections 2 and 3, we see that for penalized M-estimators, the approximation (1.1) leads to the expression

$$\sqrt{n}\big(\hat{\theta} - \theta^*\big) \rightarrow_d \mathcal{N}\big(0, V(F, T)\big),$$

where $V(F, T)$ is block diagonal with 0 entries for the elements corresponding to θ_2. All the analysis in Sections 2 and 3 was developed for a fixed λ. If we assume that there is a true underlying set of parameters θ_0, this implies in general that $\theta^* \neq \theta_0$. Note however that if we work instead with a λ_n tending to zero at an appropriate rate, the heuristic asymptotic distribution would match the oracle properties (Fan and Li [10]) for the class of lasso type estimators satisfying such properties. Examples of this type of estimators were derived for instance in Fan and Lv [11] and Avella-Medina and Ronchetti [2]. The von Mises expansion could be therefore used to assess informally the asymptotic properties of an important class of penalized M-estimators. A more careful study of this phenomenon is required for a better understanding of the conditions under which it holds and is left for future research. Still, as pointed out in Hampel *et al.* [13], page 85: "…it is usually easier to verify the asymptotic normality in another way instead of trying to assess the necessary regularity conditions to make this approach rigorous".

7. Proofs

Proof of Lemma 2

Let $b(p) = \inf_{\theta \in \mathcal{O}} \Lambda_\lambda(\theta; F, p)$ and $f(\theta, p) = b(p) - \Lambda_\lambda(\theta; F, p)$. The mapping $\Gamma : W^{2,2} \mapsto \Theta$, $\Gamma p = \{\theta | \theta \in \Theta, f(\theta, p) \leq 0\}$ is closed by construction (Berge [4], Example, page 111). Therefore, Γp is compact for any p, which implies that Γ is continuous (Berge [4], Example, page 109). From Proposition 7 in the supplementary document [1], $M(p) = \max\{-\Lambda_\lambda(\theta; F, p) | \theta \in \Gamma p\}$ is continuous in $W^{2,2}$ and the mapping $\Phi p = \{\theta | \theta \in \Gamma p, -\Lambda_\lambda(\theta; F, p) = M(p)\}$ is upper hemicontinuous from $W^{2,2}$ to $\mathcal{O}$. Since $\theta_m = \phi p_m = \{\theta | \theta \in \Phi p_m, -\Lambda_\lambda(\theta; F, p) = \sup M(p_m)\}$ is single valued and upper hemicontinuous, it is continuous. Therefore, we have $\lim_m \theta_m = \theta^* = \phi p = \lim_m \phi p_m$.

Proof of Proposition 1

For ease of notation, we will write $\psi_p = \psi(Z; T(F; p_m))$, $S_p = E_F[\dot{\psi}(Z; T(F; p_m))] + \nabla^2 p_m(T(F; p_m))$, $D = F - \Delta_z$ and $\mathrm{IF}(p_m) = \mathrm{IF}_{p_m}(z; F, T)$. Further let $\{p_m\}_{m \geq 1}$ and $\{p'_m\}_{m \geq 1}$ be to sequences in $C^\infty(\Theta)$ converging to p in $W^{2,2}(\Theta)$. Then Lemma 1 and (A3) guarantee that

$$\begin{aligned}\mathrm{IF}\big(p'_m\big) - \mathrm{IF}(p_m) &= S_p^{-1} E_D[\psi_p] - S_{p'}^{-1} E_D[\psi_{p'}] \\ &= S_{p'}^{-1} E_D[\psi_p - \psi_{p'}] + S_p^{-1}(S_p - S_{p'}) S_p^{-1} E_D[\psi_p] + o\big(\|S_p - S_{p'}\|\big).\end{aligned}$$

By Lemma 2, we have $\lim_m T(F; p'_m) = \lim_m T(F; p_m) = T(F)$. Therefore, $E_D[\psi_p - \psi_{p'}] \to 0$ and $\|S_p - S_{p'}\| \to 0$. Hence, (A3) and $\lim p_m = \lim p'_m = p$ give $\lim_m[\mathrm{IF}(p_m) - \mathrm{IF}(p'_m)] = 0$.

Proof of Proposition 2

From Proposition 1, it suffices to show that the limiting influence function of a smooth approximation of the problem has the desired form. One possible infinitely differentiable approximation for the absolute value is

$$s_m(t) = \frac{2}{m}\log\bigl(e^{tm}+1\bigr) - t \underset{m\to\infty}{\longrightarrow} |t|.$$

Its first two derivatives have the form

$$s'_m(t) := \frac{2e^{tm}}{e^{tm}+1} - 1 \underset{m\to\infty}{\longrightarrow} \operatorname{sgn}(t) = \begin{cases} -1, & \text{if } t<0,\\ +1, & \text{if } t>0,\\ 0, & \text{otherwise} \end{cases} \tag{7.1}$$

and

$$s''_m(t) = \frac{2me^{tm}}{(e^{tm}+1)^2} \underset{m\to\infty}{\longrightarrow} \begin{cases} 0, & \text{if } t\neq 0,\\ +\infty, & \text{otherwise.} \end{cases} \tag{7.2}$$

Defining $p_m(\theta) = \sum_{j=1}^p p_{\lambda,j}(s_m(\theta_j))$, from Lemma 1 and the notation of Proposition 1 we have

$$\mathrm{IF}_{p_m}(z; T, F) = -S_{p_m}^{-1}\bigl(\psi\bigl(z; T(F; p_m)\bigr) + \nabla p_m\bigl(T(F; p_m)\bigr)\bigr).$$

Remember that for a partitioned matrix A, if all the necessary inverses exist, the elements of A^{-1} are

$$\left.\begin{array}{ll} A^{11} = \bigl(A_{11} - A_{12}A_{22}^{-1}A_{21}\bigr)^{-1}, & A^{22} = \bigl(A_{22} - A_{21}A_{11}^{-1}A_{12}\bigr)^{-1},\\ A^{12} = -A^{11}A_{12}A_{22}^{-1}, & A^{21} = -A_{22}^{-1}A_{21}A^{11} \end{array}\right\}. \tag{7.3}$$

Since (7.2) implies $\nabla^2 p_m(T(F))_{jj} \to 0$ for $j = 1, \dots, s$ and $\nabla^2 p_m(T(F))_{jj} \to \infty$ for $j = s+1, \dots, d$, (7.3) yields

$$\begin{aligned} S_{p_m}^{-1} &= \begin{bmatrix} E_F\bigl[\dot\psi_{11}\bigl(Z, T(F)\bigr)\bigr] + \nabla^2 p_m^1\bigl(T(F)\bigr) & E_F\bigl[\dot\psi_{12}\bigl(Z, T(F)\bigr)\bigr] \\ E_F\bigl[\dot\psi_{21}\bigl(Z, T(F)\bigr)\bigr] & E_F\bigl[\dot\psi_{22}\bigl(Z, T(F)\bigr)\bigr] + \nabla^2 p_m^2\bigl(T(F)\bigr) \end{bmatrix}^{-1} \\ &\underset{m\to\infty}{\longrightarrow} \begin{bmatrix} (M_{11} + P_\lambda)^{-1} & 0 \\ 0 & 0 \end{bmatrix} = S^{-1}. \end{aligned} \tag{7.4}$$

Hence from (7.1), (7.2) and (7.4), it is easily seen that $\lim_m \mathrm{IF}_{p_m}(z; F, T) = \mathrm{IF}(z; F, T)$ has the claimed form.

Proof of Proposition 3

The difficulty of deriving the influence function for group lasso type penalties comes from the fact that $\sqrt{|t|}$ is not differentiable. Following the proof of Proposition 2, we will approximate $\sqrt{t}$ by $\sqrt{s_m(t)},\ t \geq 0$. This yields

$$\big(\sqrt{s_m(t)}\big)' = \frac{e^{tm}-1}{(e^{tm}+1)s_m^2(t)} \underset{m\to\infty}{\longrightarrow} \begin{cases} t^{-1/2}, & \text{if } t > 0, \\ \infty, & \text{for } t \downarrow 0 \end{cases} \tag{7.5}$$

and

$$\big(\sqrt{s_m(t)}\big)'' = \frac{2me^{tm}}{(e^{tm}+1)^2 s_m^{1/2}(t)} - \frac{(e^{tm}-1)^2}{2(e^{tm}+1)^2 s_m^{3/2}(t)} \underset{m\to\infty}{\longrightarrow} \begin{cases} -t^{-3/2}, & \text{if } t > 0, \\ -\infty, & \text{for } t \downarrow 0. \end{cases}$$

Note that after some simplifications

$$\Big(\big(\sqrt{s_m(t)}\big)''\Big)^{-1}\big(\sqrt{s_m(t)}\big)' = \frac{s_m(t)(e^{tm}+1)(e^{tm}-1)}{4me^{tm}s_m(t) - -(e^{tm}-1)^2} \underset{m\to\infty}{\longrightarrow} -t, \qquad t \geq 0. \tag{7.6}$$

Therefore for the penalty $p_m(\theta) = \sum_{g=1}^{G} p_{\lambda,g}\big(\sqrt{s_m(\|\theta_{(g)}\|_2^2)}\big)$ we have $\nabla^2 p_{m,g}(T(F)_{(g)})_{jj} \to \infty$ for $j = 1, \ldots, |g|$ if $T(F)_{(g)} = 0$ and

$$\big(E_F\big[\dot{\psi}_{22}\big(Z, T(F)\big)\big] + \nabla^2 p_m^2\big(T(F)\big)\big)^{-1} \nabla p_m^2\big(T(F)\big) \underset{m\to\infty}{\longrightarrow} 0. \tag{7.7}$$

The same argument used in (7.4) gives

$$S_{p_m}^{-1} \underset{m\to\infty}{\longrightarrow} \begin{bmatrix} (M_{11}+P_\lambda)^{-1} & 0 \\ 0 & 0 \end{bmatrix} = S^{-1}. \tag{7.8}$$

Finally, from Lemma 1, we have

$$\begin{aligned} \mathrm{IF}_{p_m}(z; T, F) = {} & \big(E_F\big[\dot{\psi}\big(Z; T(F; p_m)\big)\big] + \nabla^2 p_m\big(T(F; p_m)\big)\big)^{-1} \\ & \times \big(\psi\big(z; T(F; p_m)\big) + \nabla p_{\lambda,m}\big(T(F; p_m)\big)\big). \end{aligned} \tag{7.9}$$

Thus, the claimed results follows using (7.5)–(7.8) when taking the limit of (7.9) as $m \to \infty$.

Proof of Proposition 5

From Lemma 4 and Proposition 4, we know that we only need to take a sequence of approximating influence functions. First note that $\nabla_{\theta^{(0)}} E_F[\psi(Z, T(F))] = 0$ and $\nabla_{\theta_j,\theta_j} w_j|\theta_j| = 0$ for $\theta_j \neq 0$. Then an argument similar to the one given in Proposition 2 completes the proof.

Proof of Proposition 6

By Lemma 5, $T(F_\varepsilon)$ is a continuous function of ε in a neighborhood of 0. Therefore, $T(F_\varepsilon)$ is locally integrable. Lemma 7 in the supplementary document [1] tells us that $T(F_\varepsilon)$ has a distributional derivative given by the limiting form of the derivatives of $T(F_\varepsilon; p_m)$.

Acknowledgments

The author would like to thank the Editor and two anonymous referees for their valuable comments and Elvezio Ronchetti for numerous discussions and suggestions, including a key contribution to an early development of the idea of limiting influence function.

Supplementary Material

Supplement to "Influence functions for penalized M-estimators" (DOI: 10.3150/16-BEJ841SUPP; .pdf). The supplementary file is organized as follows. Appendices A and B contain some definitions and results from Berge [4] and from distribution theory. Appendix C provides the proofs of Lemmas 1, 3, 4 and 5 as well as Proposition 4. Appendix D gives a direct computation of the influence function via distributional derivatives for the lasso and scad functionals in the orthogonal linear model. Finally, Appendix E is a discussion of the empirical influence function in high dimensions.

References

[1] Avella-Medina, M. (2016). Supplement to "Influence functions for penalized M-estimators." DOI:10.3150/16-BEJ841SUPP.

[2] Avella-Medina, M. and Ronchetti, E. (2015). Robust and consistent variable selection in high dimensional generalized linear models. Unpublished manuscript.

[3] Bednarski, T. (1993). Fréchet differentiability of statistical functionals and implications to robust statistics. In *New Directions in Statistical Data Analysis and Robustness* (*Ascona*, 1992). *Monte Verità* 25–34. Basel: Birkhäuser. MR1280271

[4] Berge, C. (1997). *Topological Spaces*: *Including a Treatment of Multi-Valued Functions*, *Vector Spaces and Convexity*. Mineola, NY: Dover. MR1464690

[5] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720

[6] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Heidelberg: Springer. MR2807761

[7] Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96** 1022–1030. MR1947250

[8] Clarke, B.R. (1986). Nonsmooth analysis and Fréchet differentiability of M-functionals. *Probab. Theory Related Fields* **73** 197–209. MR0855222

[9] Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *Ann. Statist.* **42** 324–351. MR3189488

[10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[11] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368

[12] Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657

[13] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. New York: Wiley. MR0829458

[14] Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147

[15] Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618. MR2469326

[16] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley*, *Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Berkeley, CA: Univ. California Press. MR0216620

[17] Huber, P.J. (1981). *Robust Statistics*. New York: Wiley. MR0606374

[18] Huber, P.J. and Ronchetti, E.M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR2488795

[19] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[20] Li, G., Peng, H. and Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statist. Sinica* **21** 391–419. MR2796868

[21] Loh, P.-L. and Wainwright, M.J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

[22] Lozano, A. and Meinshausen, N. (2013). Minimum distance estimation for robust high-dimensional regression. Preprint. Available at arXiv:1307.3227.

[23] Machado, J.A.F. (1993). Robust model selection and M-estimation. *Econometric Theory* **9** 478–493. MR1241985

[24] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics*: *Theory and Methods*. *Wiley Series in Probability and Statistics*. Chichester: Wiley. MR2238141

[25] Negahban, S.N., Ravikumar, P., Wainwright, M.J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

[26] Sardy, S., Tseng, P. and Bruce, B. (2001). Robust wavelet denoising. *IEEE Trans. Signal Process.* **49** 1146–1152.

[27] Schwartz, L. (1959). *Théorie des Distributions*, *Volume* 2. *Publications de l'Institut de Mathématique de L'Université de Strasbourg*, *No. IX-X. Nouvelle É dition*, *Entiérement Corrigée*, *Refondue et Augmentée*. Paris: Hermann. MR0209834

[28] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

[29] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 273–282. MR2815776

[30] von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Stat.* **18** 309–348. MR0022330

[31] Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econom. Statist.* **25** 347–355. MR2380753

[32] Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23** 1486–1494.

[33] Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013). Robust variable selection with exponential squared loss. *J. Amer. Statist. Assoc.* **108** 632–643. MR3174647

[34] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

[35] Zhelonkin, M., Genton, M.G. and Ronchetti, E. (2012). On the robustness of two-stage estimators. *Statist. Probab. Lett.* **82** 726–732. MR2899513

[36] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Received January 2015 and revised March 2016