Biometrika (2018), **105**, 1, *pp.* 31–44
Printed in Great Britain
doi: 10.1093/biomet/asx070
Advance Access publication 25 December 2017

Robust and consistent variable selection in high-dimensional generalized linear models

By MARCO AVELLA-MEDINA

Sloan School of Management, Massachusetts Institute of Technology, 30 Memorial Drive, Cambridge, Massachusetts 02142, U.S.A.

mavella@mit.edu

and ELVEZIO RONCHETTI

Research Center for Statistics, University of Geneva, Boulevard du Pont d'Arve 40, 1205 Geneva, Switzerland

elvezio.ronchetti@unige.ch

Summary

Generalized linear models are popular for modelling a large variety of data. We consider variable selection through penalized methods by focusing on resistance issues in the presence of outlying data and other deviations from assumptions. We highlight the weaknesses of widely-used penalized M-estimators, propose a robust penalized quasilikelihood estimator, and show that it enjoys oracle properties in high dimensions and is stable in a neighbourhood of the model. We illustrate its finite-sample performance on simulated and real data.

Some key words: Contamination neighbourhood; Generalized linear model; Infinitesimal robustness; Lasso; Oracle estimator; Robust quasilikelihood.

1. Introduction

Penalized methods have proved to be a good alternative to traditional methods for variable selection, particularly in high-dimensional problems. By allowing simultaneous estimation and variable selection, they overcome the high computational cost of variable selection in cases where the number of covariates is large. Since their introduction for the linear model (Frank & Friedman, 1993; Breiman, 1995; Tibshirani, 1996) many extensions have been proposed (Efron et al., 2004; Zou & Hastie, 2005; Yuan & Lin, 2006; Tibshirani, 2011). When the number of parameters is fixed, the asymptotic properties of penalized estimators have been studied by Knight & Fu (2000), Fan & Li (2001) and Zou (2006). A large literature deals with the high-dimensional case, where the number of parameters is allowed to grow as the sample size increases (Bühlmann & van de Geer, 2011). These results provide strong arguments in favour of such procedures.

The above approaches to variable selection rely on assumptions that may not be satisfied for real data, and they are typically badly affected by the presence of a few outlying observations. Robust statistical methods (Huber, 1981; Hampel et al., 1986; Maronna et al., 2006; Huber & Ronchetti, 2009) give reliable results when slight deviations from the stochastic assumptions on the model occur. Several authors have suggested sparse estimators that limit the impact of

contamination in the data (Sardy et al., 2001; Wang et al., 2007; Li et al., 2011; Fan et al., 2014; Lozano et al., 2016). These procedures rely on the intuition that a loss function which defines robust estimators in the 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. (2014) showed that under very mild conditions on the error term, their estimator satisfies the oracle properties. Wang et al. (2013) and Alfons et al. (2013) studied the finite-sample breakdown of their proposals for the linear model. The derivation of the influence function has been explored in Wang et al. (2013) and Öllerer et al. (2015) and by Avella-Medina (2017) in a more general framework.

In this paper, we first formalize the robustness properties of general penalized M-estimators. When the dimension of the parameter is fixed, we characterize robust penalized M-estimators that have a bounded asymptotic bias in a neighbourhood of the assumed model and establish their oracle properties in shrinking neighbourhoods. Then we propose a class of robust penalized estimators for high-dimensional generalized linear models, in cases where the dimension of the parameter exceeds the sample size. We show that our estimators are consistent for variable selection, are asymptotically normally distributed, and can behave as well as a robust oracle estimator. We characterize the robustness of penalized M-estimators by showing that they have bounded bias in a contamination neighbourhood of the model. This is a characterization in the spirit of the infinitesimal robustness of Hampel et al. (1986), but unlike in the usual approach we do not establish our results based on the form of the influence function.

2. PENALIZED M-ESTIMATORS

2·1. *The oracle estimator and its robustness properties*

Consider a loss function $\rho_n : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{Z}$ denotes some sample space, and let the observations $Z^{(n)} = \{z_1, \ldots, z_n\}$ be drawn from a common distribution F over $\mathcal{Z}$. The quantity $\rho_n(\beta; Z^{(n)}) = n^{-1} \sum_{i=1}^n \rho(\beta; z_i)$ measures the fit between a parameter vector $\beta \in \mathbb{R}^d$ and the data, and is an estimator of the unknown population risk function $E_F\{\rho_n(\beta; Z^{(n)})\}$. We study the estimators resulting from the minimization of the regularized risk

$$\rho_n(\beta; Z^{(n)}) + \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|) \tag{1}$$

with respect to β, where $p_\lambda(\cdot)$ is a continuous penalty function with regularization parameter λ. We suppose that the true underlying parameter vector is sparse, and without loss of generality we write $\beta_0 = (\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}}$, where $\beta_1 \in \mathbb{R}^k$ and $\beta_2 = 0 \in \mathbb{R}^{d-k}$ with $k < n$ and $k < d$. In this sparse setting, the oracle estimator has played an important role in the theoretical analysis of many penalized estimators. It is the ideal unpenalized estimator we would use if we knew the support $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$ of the true parameter β_0. This estimator can also be used for a simple robustness assessment of more complicated penalized estimators. Indeed, the oracle estimator, whose robustness properties can easily be assessed with standard tools, serves as a benchmark for a best possible procedure that is unfortunately unattainable. Let us discuss this in more detail. Consider a set-up where the cardinality of $\mathcal{A}$ is a fixed number k. An M-estimator $T(F_n)$ of β (Huber, 1964; Huber & Ronchetti, 2009), where F_n denotes the empirical distribution function, is defined as a solution to $\sum_{i=1}^n \Psi\{z_i, T(F_n)\} = 0$. This class of estimators is a generalization of the class of maximum likelihood estimators defined by differentiable likelihoods. If the statistical

functional $T(F)$ is sufficiently regular, a von Mises expansion (von Mises, 1947) yields

$$T(G) \approx T(F) + \int \text{IF}(z; F, T)\, \mathrm{d}(G - F)(z), \tag{2}$$

where $\text{IF}(z; F, T)$ denotes the influence function of the functional T at the distribution F (Hampel, 1974). Considering the approximation (2) over an ϵ-neighbourhood of the model $\mathcal{F}_\epsilon = \{F_\epsilon = (1-\epsilon)F + \epsilon G,\ G \text{ an arbitrary distribution}\}$, we see that the influence function can be used to linearize the asymptotic bias in a neighbourhood of the ideal model. Therefore, a bounded influence function implies a bounded approximate bias, and in that sense an M-estimator characterized by a bounded score equation implies infinitesimal robustness (Hampel et al., 1986). Throughout the paper we call such M-estimators and their penalized counterparts robust. This definition is also justified by the derivation of the influence function for penalized M-estimators in Avella-Medina (2017). Since likelihood-based estimators generally do not have a bounded score function, they are not robust in this sense. It follows that we could expect a penalized M-estimator based on a loss function with a bounded derivative to behave better in a neighbourhood of the model than the classical oracle estimator; that is, a robust penalized M-estimator could beat the oracle estimator under contamination. Clearly, an appropriate benchmark estimator under contamination will only be given by a robust estimator that remains stable in $\mathcal{F}_\epsilon$.

2·2. *Some fixed-parameter asymptotics*

We now provide an asymptotic analysis of estimators obtained as the solution to problem (1) when the parameter dimension is fixed and $n \to \infty$. In particular, we study the asymptotic behaviour of robust penalized M-estimators at the ϵ-contamination model $\mathcal{F}_\epsilon$ under the following conditions:

Condition 1. $E_F\{\Psi_i(\beta_0)\} = 0$, where $\Psi_i(\beta) = \Psi(\beta; z_i)$ is bounded and continuous in a neighbourhood $\mathcal{O}$ of β_0 uniformly in z_i;

Condition 2. for all β in $\mathcal{O}$, $\Psi_i(\beta)$ has two continuous derivatives and

$$M(\beta) = E_F\left\{\frac{\partial \Psi_i(\beta)}{\partial \beta^{\mathrm{T}}}\right\} < \infty, \quad Q(\beta) = E_F\left\{\Psi_i(\beta)\Psi_i^{\mathrm{T}}(\beta)\right\} < \infty,$$

where $M(\beta)$ is positive definite;

Condition 3. for all $\beta \in \mathcal{O}$, $\|\partial \Psi_i(\beta)/\partial \beta\|_\infty < \infty$;

Condition 4. $P(t; \lambda) = \lambda^{-1} p_\lambda(t)$ is increasing and concave in $t \in [0, \infty)$ and has a continuous derivative such that $P'(0+) = P'(0+; \lambda) > 0$ is independent of λ. In addition, $P'(t; \lambda)$ is decreasing in $\lambda \in (0, \infty)$.

Conditions 1 and 2 are analogous to conditions A_0–A_3 of Clarke (1983) and are standard in robust statistics. Condition 1 implies that the functional defined by the minimizers of $\rho(z_i; \beta)$ is Fisher-consistent and has a bounded influence function. Condition 2 imposes regularity conditions on the terms appearing in the expression for the asymptotic variance of the M-estimators. Condition 3 requires a bounded second derivative of ρ and implies second-order infinitesimal robustness as defined in La Vecchia et al. (2012); see that paper for a discussion of the benefits of higher-order robustness. Condition 4 is similar to Condition 1 of Fan & Lv (2011) and defines

a class of penalty function conditions that includes the lasso penalty and the smoothly clipped absolute deviation penalty of Fan & Li (2001). The latter takes the form

$$p_\lambda^{\mathrm{SCAD}}(t) = \begin{cases} \lambda|t|, & |t| \leqslant \lambda, \\ -(t^2 - 2a\lambda|t| + \lambda^2 + 1)/\{2(a-1)\}, & \lambda < |t| \leqslant a\lambda, \\ (a+1)\lambda^2/2, & |t| > a\lambda, \end{cases}$$

where $a > 2$ is fixed. Following Lv & Fan (2009) and Zhang (2010), we define the local concavity of the penalty P at $v = (v_1, \ldots, v_q)^{\mathrm{T}} \in \mathbb{R}^q$ with $\|v\|_0 = q$ as

$$\kappa(P; \lambda, v) = \lim_{\epsilon \to 0+} \max_{1 \leqslant j \leqslant q} \sup_{t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} \frac{P'(t_2; \lambda) - P'(t_1; \lambda)}{t_2 - t_1}.$$

Condition 4 implies that $\kappa(P; \lambda, v) \geqslant 0$, and by the mean value theorem $\kappa(P; \lambda, v) = \max_{1 \leqslant j \leqslant q} = P''(|v_j|; \lambda)$ if the second derivative of P is continuous. For $p_\lambda^{\mathrm{SCAD}}$, clearly $\kappa(P; \lambda, v) = 0$ unless some component of $|v|$ takes values in $[\lambda, a\lambda]$.

Theorem 1 below shows that for a given tuning parameter, and under the above regularity conditions, robust penalized M-estimators have a bounded asymptotic bias in $\mathcal{F}_\epsilon$. Theorem 2 states results on oracle properties for general penalized M-estimators in a shrinking ϵ-neighbourhood. We use the notation $\Psi_i(\beta) = \partial\rho(z_i; \beta)/\partial\beta$, $M(\beta) = \partial E_F\{\Psi_i(\beta)\}/\partial\beta$ and $Q(\beta) = E_F\{\Psi_i(\beta)\Psi_i^{\mathrm{T}}(\beta)\}$. We let M_{11} and Q_{11} denote submatrices of $M(\beta_0)$ and $Q(\beta_0)$ appearing in the asymptotic variance of the oracle estimator.

THEOREM 1. *Under Conditions 1–4, for any $\lambda > 0$, the asymptotic bias of the penalized M-estimator is of order $O(\epsilon)$ in $\mathcal{F}_\epsilon$.*

THEOREM 2. *Assume Conditions 1–4, and let $\epsilon = o(\lambda_n)$ with $\lambda_n n^{1/2} \to 0$, $\lambda_n n \to \infty$ and $\kappa(P; \lambda_n, \beta_1) \to 0$. Further, let* v *be a k-dimensional vector with $\|\mathrm{v}\|_2 = 1$. Then if the data are generated under the contamination model $\mathcal{F}_\epsilon$, there is a minimizer of* (1) *satisfying the following oracle properties as $n \to \infty$:*

(a) *sparsity, i.e.,* $\mathrm{pr}(\hat{\beta}_2 = 0) \to 1$;
(b) *asymptotic normality, i.e.,* $n^{1/2}\mathrm{v}^{\mathrm{T}} M_{11} Q_{11}^{-1/2}(\hat{\beta}_1 - \beta_1) \to N(0, 1)$ *in distribution.*

Theorem 2 is in the spirit of the oracle properties of Fan & Li (2001). As pointed out by a referee, when using such estimators one should also be aware of their limitations. Indeed, many authors have shown that valid post-selection inference typically relies on uniform signal strength conditions on all nonzero regression coefficients, often called the beta-min condition in the literature. In the fixed-parameter asymptotic scenario considered above, it requires that the nonzero coefficients be asymptotically larger than $O(n^{-1/2})$. As shown by Leeb & Pötscher (2005, 2008, 2009), in the presence of weaker signals, the distribution of estimators satisfying the oracle properties can be highly nonnormal regardless of the sample size. Although the beta-min assumption can be too stringent in some applications (Martinez et al., 2011; Bühlmann & Mandozzi, 2014), we limit our investigation to establishing oracle properties for our estimators. In the high-dimensional setting, there have been recent proposals for uniform post-selection inference that do not require minimal signal conditions on the coefficients. Representative work in this direction includes Belloni et al. (2012), Javanmard & Montanari (2014), Zhang & Zhang (2014), Belloni et al. (2015) and Lee et al. (2016). Developing new results in this direction is left for future research.

3. Robust generalized linear models

3·1. *A robust M-estimator*

Generalized linear models (McCullagh & Nelder, 1989) include standard linear models and can be used to model both discrete and continuous responses belonging to the exponential family. The response variables $Y_1, \ldots, Y_n$ are drawn independently from the densities $f(y_i; \theta_i) = \exp[\{y_i\theta_i - b(\theta_i)\}/\phi + c(y_i, \phi)]$, where $b(\cdot)$ and $c(\cdot)$ are specific functions and ϕ is a nuisance parameter. Thus $E(Y_i) = \mu_i = b'(\theta_i)$, $\mathrm{var}(Y_i) = v(\mu_i) = \phi b''(\theta_i)$ and $g(\mu_i) = \eta_i = x_i^{\mathrm{T}}\beta_0$, where $\beta_0 \in \mathbb{R}^d$ is the vector of parameters, $x_i \in \mathbb{R}^d$ is the set of explanatory variables and $g(\cdot)$ is the link function.

We construct our penalized M-estimators by penalizing the class of loss functions proposed by Cantoni & Ronchetti (2001), which can be viewed as a natural robustification of the quasilikelihood estimators of Wedderburn (1974). The robust quasilikelihood is

$$\rho_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} Q_M(y_i, x_i^{\mathrm{T}}\beta), \tag{3}$$

where the functions $Q_M(y_i, x_i^{\mathrm{T}}\beta)$ can be written as

$$Q_M(y_i, x_i^{\mathrm{T}}\beta) = \int_{\tilde{s}}^{\mu_i} v(y_i, t)w(x_i)\,\mathrm{d}t - \frac{1}{n}\sum_{j=1}^{n}\int_{\tilde{t}}^{\mu_j} E\big\{v(y_i, t)\big\}w(x_j)\,\mathrm{d}t$$

with $v(y_i, t) = \psi\{(y_i - t)/\sqrt{v(t)}\}/\sqrt{v(t)}$, $\tilde{s}$ such that $\psi\{(y_i - \tilde{s})/\sqrt{v(\tilde{s})}\} = 0$ and $\tilde{t}$ such that $E[\psi\{(y_i - \tilde{s})/\sqrt{v(\tilde{s})}\}] = 0$. The function $\psi(\cdot)$ is bounded and protects against large outliers in the responses, and $w(\cdot)$ downweights leverage points. The estimator $\hat{\beta}$ of β_0 derived from the minimization of this loss function is the solution to the estimating equation

$$\Psi^{(n)}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Psi\big(y_i, x_i^{\mathrm{T}}\beta\big) = \frac{1}{n}\sum_{i=1}^{n}\left\{\psi(r_i)\frac{1}{\sqrt{v(\mu_i)}}\,w(x_i)\frac{\partial\mu_i}{\partial\beta} - a(\beta)\right\} = 0, \tag{4}$$

where $r_i = (y_i - \mu_i)/\sqrt{v(\mu_i)}$ and $a(\beta) = n^{-1}\sum_{i=1}^{n} E\{\psi(r_i)/\sqrt{v(\mu_i)}\}w(x_i)\partial\mu_i/\partial\beta$ ensures Fisher consistency and can be computed as in Cantoni & Ronchetti (2001). Although in our analysis we consider only the robust quasilikelihood loss, the results and discussion in the rest of this section are more general. Indeed, as detailed in the Supplementary Material, they also apply to the class of bounded deviance losses introduced in Bianco & Yohai (1996) and the class of robust Bregman divergences of Zhang et al. (2014). We relegate to the Supplementary Material a review of existing robust estimation procedures in fixed dimensions.

3·2. *Asymptotic analysis in high dimensions*

We consider estimators which minimize (1) where $\rho_n(\beta)$ is as in (3). Theorems 1 and 2 of § 2 apply in this case. Here we want to go beyond these results and present an asymptotic analysis for when the number of parameters d diverges to infinity and can exceed the sample size n. For the choice of the penalty function $p_{\lambda_n}(\cdot)$ we propose using the ℓ_1-norm in a first stage. We then use the resulting lasso estimates for the construction of adaptive lasso estimators. This choice of initial estimators has been shown to yield variable selection consistency in the linear model in Fan et al. (2014). Given the initial lasso estimates $\tilde{\beta}$, we define the weights of the adaptive lasso

estimator as

$$\hat{w}_j = \begin{cases} 1/|\tilde{\beta}_j|, & |\tilde{\beta}_j| > 0 \\ \infty, & |\tilde{\beta}_j| = 0 \end{cases} \quad (j = 1, \ldots, d). \tag{5}$$

Hence variables that are shrunk to zero by the initial estimator are not included in the adaptive lasso minimization. The robust adaptive lasso minimizes the penalized robust quasilikelihood

$$\rho_n(\beta) + \lambda_n \sum_{j=1}^{d} \hat{w}_j |\beta_j|, \tag{6}$$

where we define $\hat{w}_j|\beta_j| = 0$ whenever $\hat{w}_j = \infty$ and $\beta_j = 0$. Let $\{(x_i, y_i)\}_{i=1}^n$ denote independent pairs, each having the same distribution. We assume the following conditions:

Condition 5. there is a sufficiently large open set $\mathcal{O}$ containing the true parameter β_0 such that for all $\beta \in \mathcal{O}$, the matrices $M(\beta)$ and $Q(\beta)$ satisfy

$$0 < c_1 < \lambda_{\min}\{M(\beta)\} \leqslant \lambda_{\max}\{M(\beta)\} < c_2 < \infty, \quad Q(\beta) < \infty,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues;

Condition 6. $|\psi'(r)|$ and $|\psi'(r)r|$ are bounded. Furthermore, for all $\beta \in \mathcal{O}$,

$$\frac{\partial Q_M(y,\mu)}{\partial \eta} \leqslant A_1(x), \quad \frac{\partial^2 Q_M(y,\mu)}{\partial \eta^2} \leqslant A_2(x)$$

where $\mu = g^{-1}(x^{\mathrm{T}}\beta)$ and $A_1(x), A_2(x) < \infty$. Moreover, for all $1 \leqslant j,k \leqslant d$ we have $\big|A_2(x)|x_j x_k|\big| < \infty$;

Condition 7. for all $\beta \in \mathcal{O}$,

$$\big\|X_2^{\mathrm{T}}\Upsilon'(X\beta)X_1\{X_1^{\mathrm{T}}\Upsilon'(X\beta)X_1\}^{-1}\big\|_\infty \leqslant \min\left\{c_3\frac{P'(0+)}{P'(s_n;\lambda_n)}, O(n^{\alpha_1})\right\},$$

where $s_n = 2^{-1}\min\{|\beta_{0j}| : \beta_{0j} \neq 0\}$, $\Upsilon(X\beta) = \partial\{\sum_{i=1}^n Q_M(y,\mu_i)\}/\partial\eta$, $c_3 \in (0,1)$ and $\alpha_1 \in (0, 1/2)$. The expression $\Upsilon'(\cdot)$ is to be understood as the derivative with respect to η_i of its corresponding diagonal element.

Conditions 5 and 6 impose regularity on Ψ and are analogous to Conditions 1 and 2. Obviously these conditions are violated by most likelihood loss functions. However, for robust estimates, $\Psi(\cdot)$ is bounded and $w(x)$ typically goes to zero as $\|x\|^{-1}$. Our condition requires a slightly stronger version, namely that $w(x)$ goes to zero as $\min_{1\leqslant j\leqslant d}(x_j^{-2})$. Condition 7 is similar to equation (16) in Condition 2 of Fan & Lv (2011). When the lasso penalty is used, the upper bound in Condition 7 becomes the strong irrepresentability condition of Zhao & Yu (2006). This

condition is significantly weaker when a folded concave penalty function is used, because the upper bound in Condition 7 can grow to infinity at a rate of $O(n^{\alpha_1})$. This occurs for instance when the smoothly clipped absolute deviation penalty of Fan & Li (2001) is used, as long as the minimum signal condition $s_n \gg \lambda_n$ holds.

The following theorem shows that given an initial consistent estimate, the robust adaptive lasso enjoys oracle properties when $k \ll n$, $\log d = O(n^{\alpha})$ for some $\alpha \in (0, 1/2)$, and $s_n \gg \lambda_n$.

THEOREM 3. *Assume Conditions 4–6 and let $\tilde{\beta}$ be a consistent initial estimator with rate $r_n = \{(k \log d)/n\}^{1/2}$ in ℓ_2-norm defining weights of the form (5). Suppose that $\log d = O(n^{\alpha})$ for $\alpha \in (0, 1/2)$. Further, let the nonsparsity dimensionality be of order $k = o(n^{1/3})$ and assume that the minimum signal is such that $s_n \gg \lambda_n$ with $\lambda_n (nk)^{1/2} \to 0$ and $\lambda_n r_n \gg \max\{(k/n)^{1/2}, n^{-\alpha}(\log n)^{1/2}\}$. Finally, let v be a k-dimensional vector with $\|\mathrm{v}\|_2 = 1$. Then there exists a minimizer $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^{\mathrm{T}}$ of (6) such that as $n \to \infty$, the following properties hold:*

(a) *sparsity, i.e.,* $\mathrm{pr}(\hat{\beta}_2 = 0) \to 1$;
(b) *asymptotic normality, i.e.,* $n^{1/2}\, \mathrm{v}^{\mathrm{T}} M_{11} Q_{11}^{-1/2} (\hat{\beta}_1 - \beta_1) \to N(0, 1)$ *in distribution.*

As in Zou (2006), the existence of an initial consistent estimator is key to obtaining variable selection consistency in Theorem 3. Trying to apply our proof strategy to the lasso penalty leads to the requirements $\lambda_n \gg (k/n)^{1/2}$ and $\lambda_n (nk)^{1/2} \to 0$, which cannot be met simultaneously. In fact, it has been shown that in general the lasso cannot be consistent for variable selection (Meinshausen & Bühlmann, 2006; Yuan & Lin, 2006; Zhao & Yu, 2006; Zou, 2006).

Theorem 4 below shows that the robust lasso is indeed consistent. Combined with Theorem 3, it implies that using the robust lasso as initial estimator for its adaptive counterpart yields an estimator that satisfies the oracle properties.

THEOREM 4. *Denote by $\hat{\beta}$ the robust lasso estimator obtained by solving (6) with $\tilde{w}_j = 1$ $(j = 1, \ldots, d)$ and $\lambda_n = O\{(n^{-1} \log d)^{1/2}\}$. Further, let $k = o(n^{1/2})$ and $\log d = o(n^{1/2})$. Then, under Conditions 4–6 we have that*

$$\|\hat{\beta} - \beta\|_2 = O\left\{ \left(\frac{k \log d}{n} \right)^{1/2} \right\}$$

with probability at least $1 - 4\exp(-\gamma \kappa_n)$, where γ is some positive constant and $\kappa_n = \min(n/k^2, \log d)$.

Since the robust quasilikelihood function is not convex, the proof of Theorem 4 relies on results given in Loh & Wainwright (2015) for penalized estimators with nonconvexity. The result also holds if we replace the lasso penalty by a decomposable penalty as defined in Loh & Wainwright (2015), and it does not require minimal signal strength conditions as in Theorem 3.

We show next that we can also construct robust estimators satisfying the oracle properties by combining the robust quasilikelihood loss with nonconvex penalty functions such as those proposed by Fan & Li (2001) and Zhang (2010).

THEOREM 5. *Assume Conditions 4–7 and let $\log d = O(n^{\alpha})$ for $\alpha \in (0, 1/2)$. Further, let $k = o(n^{1/3})$ and assume that the minimum signal is such that $s_n \gg \lambda_n$ with $\lambda_n (nk)^{1/2} \to 0$ and $\lambda_n \gg \max\{(k/n)^{1/2}, n^{-\alpha}(\log n)^{1/2}\}$, and that $\lambda_n \kappa_0 = o(1)$ where $\kappa_0 = \max_{\delta \in N_0} \kappa(P; \lambda_n, \delta)$ and $N_0 = \{\delta \in R^k : \|\delta - \beta_1\|_\infty \leqslant s_n\}$. Let v be a k-dimensional vector with $\|\mathrm{v}\|_2 = 1$. Then there*

exists a minimizer $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^{\mathrm{T}}$ *of* (6) *such that as* $n \to \infty$, *the following properties hold:*

(a) *sparsity, i.e.,* $\mathrm{pr}(\hat{\beta}_2 = 0) \to 1$;
(b) *asymptotic normality, i.e.,* $n^{1/2}\, \mathrm{v}^{\mathrm{T}} M_{11} Q_{11}^{-1/2} (\hat{\beta}_1 - \beta_1) \to N(0,1)$ *in distribution.*

Theorems 3–5 state that our estimators have the same properties as the penalized likelihood estimators for generalized linear models (Fan & Lv, 2011) when the data are generated by the assumed model. In particular, the oracle properties of Theorems 3 and 5 are established under the same uniform signal strength conditions as in Fan & Lv (2011). Thus the discussion at the end of § 2·2 also applies to these distributional results.

Since the oracle estimator defined by our loss function is robust in a neighbourhood of the model, so should be our penalized estimator. It can be seen from the proofs that the oracle properties will hold under F_ϵ if $E_{F_\epsilon}\{\psi_i(\beta_0)\} = 0$. This implies consistency of the robust estimator over a broader class of distributions than those covered by the usual likelihood-based counterparts, because the boundedness of ψ can ensure consistency for heavy-tailed distributions. In particular, we require milder conditions on the distribution of the responses than in Fan & Lv (2011), where the existence of a moment generating function is necessary. Our conditions are also milder than the fourth moment assumption of van de Geer & Müller (2012).

To illustrate this, consider a Poisson regression where a fraction ϵ of responses come from a heavy-tailed overdispersed Poisson mixture. Specifically, let the contamination distribution G be such that for $U_i \sim G$, U_i takes values in the natural numbers for all $i = 1, \ldots, n$ and we have $E(U_i) = \mu_i$ and $\mathrm{pr}(|U_i| > q) = c_\kappa q^{-\kappa}$, where c_κ is a constant depending only on $\kappa \in (0, 3/4)$. In this case the distribution of the responses does not have a moment generating function and has infinite variance. Still, in this scenario our estimator satisfies the oracle properties because $E_{F_\epsilon}\{\psi_i(\beta_0)\} = 0$.

3·3. *Weak oracle properties in a contamination neighbourhood*

The previous asymptotic analysis generalizes to a shrinking contamination neighbourhood where $\epsilon \to 0$ when $n \to \infty$ as in Theorem 2, if we let $\epsilon = o(\lambda_n)$ in Theorem 5. Here we give a result where the contamination neighbourhood does not shrink, but instead produces a small nonasymptotic bias on the estimated nonzero coefficients. If this bias is not too large and the minimum signal is large enough, we could expect to obtain correct support recovery and bounded bias. In this sense we could expect a robust estimator that behaves as well as a robust oracle.

THEOREM 6. *Assume Conditions 4–7 and let* $\log d = O(n^{1-2\alpha})$ *for* $\alpha \in (0, 1/2)$, *with* $k = o(n)$ *nonzero parameters and a minimum signal such that* $s_n \geqslant n^{-\zeta} \log n$. *Let* λ_n *satisfy* $p'_{\lambda_n}(s_n) = o(n^{-\zeta} \log n)$ *and* $\lambda_n \gg n^{-\alpha} \log^2 n$. *In addition, suppose that* $\lambda_n \kappa_0 = o(\tau_0)$, *where* $\kappa_0 = \max_{\delta \in N_0} \kappa(P; \lambda_n, \delta)$ *and* $\tau_0 = \min_{\delta \in N_0} \lambda_{\min}[n^{-1} X_1^{\mathrm{T}}\{\Upsilon''(X\beta)\}X_1]$. *Then, for n large enough, there is a contamination neighbourhood where the robust penalized quasilikelihood estimator* $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^{\mathrm{T}}$ *satisfies, with probability at least* $1 - 2\{kn^{-1} + (d-k)\exp(-n^{1-2\alpha}\log n)\}$, *the following weak oracle properties:*

(a) *sparsity, i.e.,* $\hat{\beta}_2 = 0$;
(b) *in the* ℓ_∞*-norm,* $\|\hat{\beta}_1 - \beta_1\|_\infty = O(n^{-\zeta} \log n + \epsilon)$.

Theorem 6 can be viewed as an extension to a robust penalized M-estimator in a contamination neighbourhood of the weak oracle properties introduced in Theorem 2 of Fan & Lv (2011). Moreover, it is in the spirit of the infinitesimal robustness approach to robust statistics, whereby the

impact of moderate distributional deviations from ideal models is assessed by approximating and bounding the resulting bias; see Hampel et al. (1986). It can be seen from the proof of Theorem 6 that as long as $E_{F_\epsilon}\{\psi_i(\beta_0)\} = 0$, we obtain the stronger result $\|\hat{\beta}_1 - \beta_1\|_\infty = O(n^{-\zeta}\log n)$.

3·4. *Fisher scoring coordinate descent*

An appropriate algorithm is required for the computation of (6) with $\hat{w}_j = 1$ $(j = 1, \ldots, d)$. We propose a coordinate descent-type algorithm based on successive expected quadratic approximation of the quasilikelihood about the current estimates. Specifically, for a given value of the tuning parameter, we successively minimize via coordinate descent the penalized weighted least squares loss

$$\|W(z - X\beta)\|_2^2 + \lambda\|\beta\|_1, \tag{7}$$

where $W = \mathrm{diag}(W_1, \ldots, W_n)$ is a weight matrix and $z = (z_1, \ldots, z_n)^{\mathrm{T}}$ is a vector of pseudo-data with components

$$W_i^2 = E\{\psi(r_i)r_i\}v(\mu_i)^{-1}w(x_i)\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2, \quad z_i = \eta_i + \frac{\psi(r_i) - E\{\psi(r_i)\}}{E\{\psi(r_i)r_i\}}v(\mu_i)^{1/2}\frac{\partial\eta_i}{\partial\mu_i}.$$

These are the robust counterparts of the usual expressions appearing in the iterative reweighted least squares algorithm; cf. Appendix E.3 in Heritier et al. (2009). Our coordinate descent algorithm is therefore a sequence of three nested loops: (i) outer loop, decrease λ; (ii) middle loop, update W and z in (7) using the current parameters $\hat{\beta}_\lambda$; and (iii) inner loop, run the coordinate descent algorithm on the weighted least squares problem (7).

This algorithm differs from that of Friedman et al. (2010) in the quadratic approximation step, where we compute expected weights. This step ensures that W has only positive components in Poisson and binomial regression when using Huberized residuals, which guarantees the convergence of the inner loop. For classical penalized loglikelihood regression, the algorithms coincide. The initial value of the tuning parameter in our algorithm is $\lambda_0 = n^{-1}\|WX^{\mathrm{T}}z\|_\infty$, where W and z are computed with $\beta = 0$. We then run our coordinate descent algorithm and solve our lasso problem along an entire path of tuning parameters. The middle loop uses current parameters as warm starts. In the inner loop, after a complete cycle through all the variables, we iterate only on the current active set, i.e., the set of nonzero coefficients. If another complete cycle does not change this set, the inner loop has converged; otherwise the process is repeated. The use of warm starts and active set cycling speeds up computations, as discussed in (Friedman et al., 2010, p. 7).

3·5. *Tuning parameter selection*

We choose the tuning parameter λ_n based on a robust extended Bayesian information criterion. Specifically, we select the parameter λ_n that minimizes

$$\mathrm{EBIC}(\lambda_n) = \rho_n(\hat{\beta}_{\lambda_n}) + \left(\frac{\log n}{n} + \gamma\frac{\log d}{n}\right)|\mathrm{supp}\,\hat{\beta}_{\lambda_n}|, \tag{8}$$

where $|\mathrm{supp}\,\hat{\beta}_{\lambda_n}|$ denotes the cardinality of the support of $\hat{\beta}_{\lambda_n}$ and $0 \leqslant \gamma \leqslant 1$ is a constant. We use $\gamma = 0{\cdot}5$. We write $\hat{\beta}_{\lambda_n}$ to stress the dependence of the minimizer of (6) on the tuning parameter. In an unpenalized set-up, a Schwartz information criterion was considered by Machado (1993), who provided theoretical justification for it by proving model selection consistency and robustness. In the penalized sparse regression literature, Lambert-Lacroix & Zwald (2011) and

Li et al. (2011) used this criterion to select the tuning parameter. In high dimensions Chen & Chen (2008, 2012) and Fan & Tang (2013) showed the benefits of minimizing (8) in a penalized likelihood framework.

4. NUMERICAL EXAMPLES

4·1. *Large outliers in the responses*

We explore the behaviour of classical and robust versions of both the lasso and the adaptive lasso in a simulated generalized linear model. For the robust estimators we use the quasilikelihood defined by (3) and (4) with $\psi(r) = \psi_c(r) = \min\{c, \max(-c, r)\}$, the Huber function, and $w(x_i) = 1$. We consider the Poisson regression model with canonical link $g(\mu_i) = \log \mu_i = x_i^{\mathrm{T}}\beta_0$, where $\beta_0^{\mathrm{T}} = (1{\cdot}8, 1, 0, 0, 1{\cdot}5, 0, \ldots, 0)$ and the covariates x_{ij} were generated from the standard uniform distribution with correlation $\mathrm{cor}(x_{ij}, x_{ik}) = \rho^{|j-k|}$ and $\rho = 0{\cdot}5$ $(j, k = 1, \ldots, d)$. The responses Y_i were generated according to $\mathcal{P}(\mu_i)$, the Poisson distribution with mean μ_i, and a perturbed distribution of the form $(1 - b)\mathcal{P}(\mu_i) + b\mathcal{P}(\nu\mu_i)$, where b is a Bernoulli random variable $\mathcal{B}(1, \epsilon)$. The latter represents a situation where the distribution of the data lies in a small neighbourhood of the assumed model $\mathcal{P}(\mu_i)$. We set $\nu = 5, 10$ and $\epsilon = 0, 0{\cdot}05$. The sample sizes and dimensions were taken to be $n = 50, 100, 200$ and $d = 100, 400, 1600$, respectively, and the number of replications was set to 500.

Table 1 shows the performances of the classical and robust lasso, adaptive lasso and oracle estimators, with tuning parameters selected by minimizing (8). The robust estimators are very competitive when there is no contamination and remain stable under contamination. In particular, the robust adaptive lasso performs almost as well as the robust oracle in terms of L_2-loss and recovers the true support when $n = 100$ or 200 even under contamination. In this example the robust adaptive lasso outperforms the classical oracle estimator under contamination. Further results can be found in the Supplementary Material.

4·2. *Earthquake data example*

We analyse a time series from the Northern California Earthquake Data Center, available from http://quake.geo.berkeley.edu and previously studied in Fryzlewicz & Nason (2004). The data consist of $n = 1024$ consecutive observations of the number of earthquakes of magnitude 3 or greater which occurred each week, with the last week under consideration being 29 November to 5 December 2000. The time series is plotted in Fig. 1.

Our aim is to estimate the intensity function of a Poisson process. More precisely, denoting by Y_i the number of earthquakes that occurred in the ith week and letting $X_i = i/n$, we assume that the pairs (Y_i, X_i) follow the model $Y_i \mid X_i \sim \mathcal{P}\{f_0(X_i)\}$. Our goal is to estimate the unknown intensity function f_0, assumed to be positive. We suppose that $\log f_0$ can be well approximated by a linear combination of known wavelet functions, and so the estimation of f_0 boils down to the estimation of d coefficients. We consider four different estimators. The first applies the Haar–Fisz transform followed by thresholding as in Fryzlewicz & Nason (2004). We used the function denoise.poisson of the R package haarfisz (Fryzlewicz, 2010; R Development Core Team, 2018). The second method is the adaptive Poisson lasso estimator of Ivanoff et al. (2016), with the data-driven weights implemented in their code available at http://pbil.univ-lyon1.fr/members/fpicard/software.html. Finally, we consider our robust lasso and robust adaptive lasso estimators for Poisson regression with tuning parameter selected by minimizing (8). We scaled the lasso penalty with $2^{s/2}\lambda$, where s is the scale of the wavelet coefficients, as proposed by Sardy et al. (2004). For all the estimators we used the Daubechies wavelet basis where $d = n$.

Table 1. *Summary of the performance of the penalized estimators computed under different high-dimensional Poisson regression scenarios; the median of each measure over* 500 *simulations is given, with its median absolute deviation in parentheses*

Method	L_2-loss	Size	#FN	#FP	L_2-loss	Size	#FN	#FP	L_2-loss	Size	#FN	#FP
$n = 50$, $p = 100$		$\epsilon = 0$, $\nu = 0$				$\epsilon = 0{\cdot}05$, $\nu = 5$				$\epsilon = 0{\cdot}05$, $\nu = 10$		
Lasso	0·66	5	0	2	1·42	16	0	13	2·11	19	0	16
	(0·20)	(1·48)	(0)	(1·49)	(0·37)	(4·45)	(0)	(4·45)	(0·45)	(1·48)	(0)	(2·22)
Robust lasso	0·61	7	0	4	0·84	10	0	7	0·89	12	0	9
	(0·19)	(2·97)	(0)	(2·97)	(0·38)	(5·93)	(0)	(5·93)	(0·53)	(8·90)	(0)	(8·90)
Adaptive lasso	0·31	3	0	0	1·55	8	0	5	2·46	11	1	9
	(0·17)	(0)	(0)	(0)	(0·66)	(4·45)	(0)	(4·45)	(0·62)	(2·97)	(1·48)	(2·97)
Robust adaptive lasso	0·36	3	0	0	0·55	3	0	1	0·61	4	0	1
	(0·19)	(0)	(0)	(0)	(0·40)	(0)	(0)	(1·48)	(0·59)	(1·48)	(0)	(1·48)
Oracle	0·30	3	0	0	0·68	3	0	0	1·16	3	0	0
	(0·13)				(0·43)				(0·69)			
Robust oracle	0·27	3	0	0	0·31	3	0	0	0·29	3	0	0
	(0·13)				(0·14)				(0·13)			
$n = 100$, $p = 400$												
Lasso	0·53	4	0	1	1·19	27	0	24	1·87	30	0	27
	(0·13)	(1·48)	(0)	(1·48)	(0·24)	(4·45)	(0)	(4·45)	(0·30)	(1·48)	(0)	(1·48)
Robust lasso	0·50	6	0	3	0·59	8	0	5	0·59	8	0	5
	(0·12)	(2·97)	(0)	(2·97)	(0·16)	(4·45)	(0)	(4·45)	(0·16)	(4·45)	(0)	(4·45)
Adaptive lasso	0·19	3	0	0	1·28	12	0	9	2·06	18	0	15
	(0·22)	(0)	(0)	(0)	(0·31)	(4·45)	(0)	(4·45)	(0·36)	(4·45)	(0)	(4·45)
Robust adaptive lasso	0·22	3	0	0	0·30	3	0	0	0·32	3	0	0
	(0·12)	(0)	(0)	(0)	(0·18)	(0)	(0)	(0)	(0·18)	(0)	(0)	(0)
Oracle	0·18	3	0	0	0·52	3	0	0	0·90	3	0	0
	(0·08)				(0·26)				(0·45)			
Robust oracle	0·19	3	0	0	0·21	3	0	0	0·21	3	0	0
	(0·09)				(0·10)				(0·09)			
$n = 200$, $p = 1600$												
Lasso	0·41	4	0	1	0·99	40	0	37	1·63	41	0	38
	(0·08)	(1·28)	(0)	(1·48)	(0·14)	(2·97)	(0)	(2·97)	(0·20)	(1·48)	(0)	(1·48)
Robust lasso	0·39	4	0	1	0·45	5	0	2	0·46	5	0	2
	(0·08)	(1·48)	(0)	(1·48)	(0·08)	(1·48)	(0)	(1·48)	(0·09)	(2·97)	(0)	(2·97)
Adaptive lasso	0·13	3	0	0	1·13	18	0	15	1·80	27	0	24
	(0·06)	(0)	(0)	(0)	(0·18)	(4·45)	(0)	(4·45)	(0·24)	(4·45)	(0)	(4·45)
Robust adaptive lasso	0·17	3	0	0	0·23	0	0	0	0·23	3	0	0
	(0·10)	(0)	(0)	(0)	(0·16)	(0)	(0)	(0)	(0·13)	(0)	(0)	(0)
Oracle	0·13	3	0	0	0·37	3	0	0	0·71	3	0	0
	(0·06)				(0·18)				(0·34)			
Robust oracle	0·13	3	0	0	0·14	3	0	0	0·15	3	0	0
	(0·06)				(0·06)				(0·07)			

L_2-loss, is $\|\hat{\beta} - \beta_0\|_2$; Size, the number of selected variables in the final model; #FN, the number of missed true variables; #FP, the number of false variables included in the model.

Figure 2 shows the fits of the intensity function for the four methods considered. We display only two time windows for ease of presentation. The first is for weeks 201 to 400 as illustrated in Fryzlewicz & Nason (2004), and the second covers weeks 401 to 600. Overall, the fit of the adaptive Poisson lasso is similar to that of our two robust procedures, though some differences can be observed around weeks 220 and 445. In the first case, only the robust procedures seem to detect

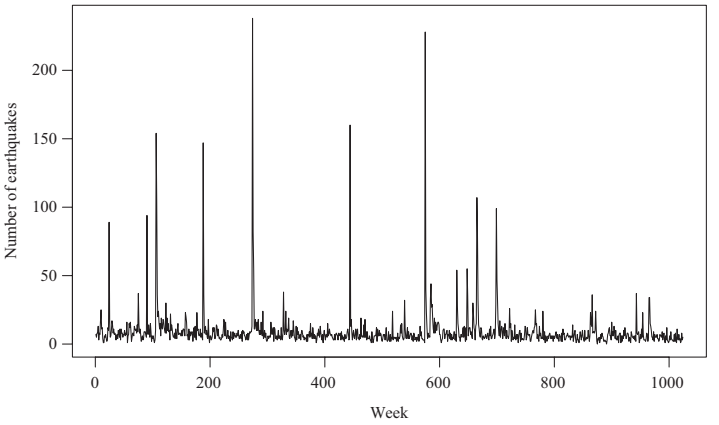

Fig. 1. Number of earthquakes of magnitude 3 or greater which occurred in Northern California in 1024 consecutive weeks.

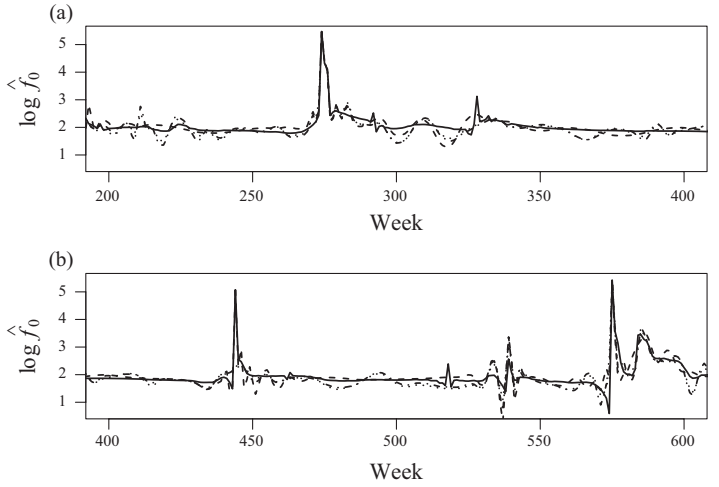

Fig. 2. Intensity function estimates for (a) weeks 201–400 and (b) weeks 401–600: adaptive Poisson lasso (dashed); Haar–Fisz transform followed by soft thresholding (solid); robust lasso (dotted); and robust adaptive lasso (dash-dot).

some small fluctuations in the intensity function; in the second case, only the robust procedures keep a flat fit while both nonrobust procedures yield a large spike. This suggests that our robust method plays a complementary role to the classical approach, by providing useful information in a data analysis. Close inspection of the data for these periods can provide information on possible anomalous phenomena.

Although our theoretical results do not cover the functional regression set-up considered in this example, we show in the Supplementary Material that our methods can be useful in this situation as well. In particular, we demonstrate by simulation that our estimators can be competitive with the Haar–Fisz and Poisson lasso estimators at the model but are more reliable under contamination.

ACKNOWLEDGEMENT

We thank the editor, the associate editor, two referees, E. Cantoni, S. Sardy, and R. E. Welsch for comments that have led to significant improvements of the manuscript. Most of this research was conducted while the first author was a PhD student at the Research Centre for Statistics at

the University of Geneva. The first author was partially supported by the Swiss National Science Foundation and the second author by the EU Framework Programme Horizon 2020.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of all the theorems, additional numerical results, and a discussion on alternative approaches to robust estimation for generalized linear models.

References

Alfons, A., Croux, C. & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Statist.* **7**, 226–48.

Avella-Medina, M. (2017). Influence functions for penalized M-estimators. *Bernoulli* **23**, 3778–96.

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–429.

Belloni, A., Chernozhukov, V. & Kato, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102**, 77–94.

Bianco, A. M. & Yohai, V. J. (1996). Robust estimation in the logistic regression model: Missing links. In *Robust Statistics, Data Analysis and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*, ed. H. Rieder. New York: Springer, pp. 17–34.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–84.

Bühlmann, P. & Mandozzi, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comp. Statist.* **29**, 407–30.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.

Cantoni, E. & Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Am. Statist. Assoc.* **96**, 1022–30.

Chen, J. & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–71.

Chen, J. & Chen, Z. (2012). Extended BIC for small-*n*-large-*p* sparse GLM. *Statist. Sinica* **22**, 555–74.

Clarke, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.* **11**, 1196–205.

Efron, B., Hastie, T. J. Johnstone, I. M. & Tibshirani, R. J. (2004). Least angle regression. *Ann. Statist.* **32**, 407–99.

Fan, J. Fan, Y. & Barut, E. (2014). Adaptive robust variable selection. *Ann. Statist.* **57**, 324–51.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Info. Theory* **57**, 5467–84.

Fan, Y. & Tang, C. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc.* B **75**, 531–52.

Frank, L. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–35.

Friedman, J. H. Hastie, T. J. & Tibshirani, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1–22.

Fryzlewicz, P. (2010). *haarfisz: a Haar-Fisz algorithm for Poisson intensity estimation*. R package version 4.5.

Fryzlewicz, P. & Nason, G. P. (2004). A Haar–Fisz algorithm for Poisson intensity estimation. *J. Comp. Graph. Statist.* **13**, 621–38.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69**, 383–93.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Heritier, S., Cantoni, E., Copt, S. & Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. Chichester: Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Huber, P. J. & Ronchetti, E. M. (2009). *Robust Statistics*. New York: Wiley, 2nd ed.

Ivanoff, S., Picard, F. & Rivoirard, V. (2016). Adaptive Lasso and group-Lasso for functional Poisson regression. *J. Mach. Learn. Res.* **17**, 1–46.

Javanmard, A. & Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–909.

KNIGHT, K. & FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–78.
LA VECCHIA, D., RONCHETTI, E. & TROJANI, F. (2012). Higher-order infinitesimal robustness. *J. Am. Statist. Assoc.* **107**, 1546–57.
LAMBERT-LACROIX, S. & ZWALD, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Statist.* **5**, 1015–53.
LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 907–27.
LEEB, H. & PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Economet. Theory* **21**, 21–59.
LEEB, H. & PÖTSCHER, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Economet.* **142**, 201–11.
LEEB, H. & PÖTSCHER, B. M. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Mult. Anal.* **100**, 2065–82.
LI, G., PENG, H. & ZHU, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statist. Sinica* **21**, 391–419.
LOH, P.-L. & WAINWRIGHT, M. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16**, 559–616.
LOZANO, A., MEINSHAUSEN, N. & YANG, E. (2016). Minimum distance lasso for robust high-dimensional regression. *Electron. J. Statist.* **10**, 1296–340.
LV, J. & FAN, J. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–528.
MACHADO, J. (1993). Robust model selection and M-estimation. *Economet. Theory* **9**, 478–93.
MARONNA, R., MARTIN, R. D. & YOHAI, V. J. (2006). *Robust Statistics: Theory and Methods*. New York: Wiley.
MARTINEZ, J. G., CARROLL, R. J., MÜLLER, S., SAMPSON, J. N. & CHATTERJEE, N. (2011). Empirical performance of cross-validation with oracle methods in a genomics context. *Am. Statistician* **65**, 223–8.
MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall/CRC, 2nd ed.
MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.
ÖELLERER, V., CROUX, C. & ALFONS, A. (2015). The influence function of penalized regression estimators. *Statistics* **49**, 741–65.
R DEVELOPMENT CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
SARDY, S., ANTONIADIS, A. & TSENG, P. (2004). Automatic smoothing with wavelets for a wide class of distributions. *J. Comp. Graph. Statist.* **13**, 399–421.
SARDY, S., TSENG, P. & BRUCE, A. (2001). Robust wavelet denoising. *IEEE Trans. Sig. Proces.* **49**, 1146–52.
TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
TIBSHIRANI, R. J. (2011). Regression shrinkage and selection via the lasso: A retrospective. *J. R. Statist. Soc.* B **73**, 273–82.
VAN DE GEER, S. & MÜLLER, P. (2012). Quasi-likelihood and/or robust estimation in high dimensions. *Statist. Sci.* **27**, 469–80.
VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18**, 309–48.
WANG, H., LI, G. & JIANG, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-lasso. *J. Bus. Econ. Statist.* **25**, 347–55.
WANG, X., JIANG, Y., HUANG, M. & ZHANG, H. (2013). Robust variable selection with exponential squared loss. *J. Am. Statist. Assoc.* **108**, 632–43.
WEDDERBURN, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–47.
YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.
ZHANG, C., GUO, X., CHENG, C. & ZHANG, Z. (2014). Robust-BD estimation and inference for varying dimensional general linear models. *Statist. Sinica* **24**, 515–32.
ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.
ZHANG, C. H. & ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B **76**, 217–42.
ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.
ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.
ZOU, H. & HASTIE, T. J. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **76**, 301–20.

[*Received on* 12 *June* 2015. *Editorial decision on* 24 *September* 2017]