Asymmetric influence measure for high dimensional regression

Amadou Barry\textsuperscript{a,b}, Nikhil Bhagwat\textsuperscript{c}, Bratislav Misic\textsuperscript{c}, Jean-Baptiste Poline\textsuperscript{c,d,e}, and Celia M. T. Greenwood\textsuperscript{a,b,e,f}

\textsuperscript{a}Departments of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Qu\textsuperscript{e}bec, Canada; \textsuperscript{b}Lady Davis Institute, Jewish General Hospital, Montreal, Qu\textsuperscript{e}bec, Canada; \textsuperscript{c}Faculty of Medicine, Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McConnell Brain Imaging Centre, McGill University, Montreal, Qu\textsuperscript{e}bec, Canada; \textsuperscript{d}Henry H. Wheeler Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California, Berkeley, California, USA; \textsuperscript{e}Ludmer Centre for Neuroinformatics & Mental Health, McGill University, Montreal, Qu\textsuperscript{e}bec, Canada; \textsuperscript{f}Departments of Oncology and Human Genetics, McGill University, Montreal, Qu\textsuperscript{e}bec, Canada

**ABSTRACT**

Identification of influential observations is crucial in data analysis, particularly with high dimensional datasets, where the number of predictors is higher than the sample size. These rich datasets with extensive detail are increasingly exploited and analyzed in multiple fields of science, e.g., genomics, neuroscience, finance, etc. Unfortunately, classical diagnostic statistical tools are not tailored for identifying influential observations in high dimensional setup. In this paper, we use the concept of expectiles to develop an influence measure in high dimensional regression. The influence measure is based on the asymmetric marginal correlation, and its derived asymptotic distribution is used to define a threshold based on statistical principles. Our comprehensive simulation results display the favorable qualities of this influence measure under various scenarios. The usefulness of the proposed measure is illustrated through the analysis of a neuroimaging dataset. An R package implementing the procedure is publicly available on GitHub (https://github.com/AmBarry/hidetify).

**ARTICLE HISTORY**

Received 30 December 2019
Accepted 19 October 2020

**KEYWORDS**

Asymmetric correlation; expectile; high dimensional regression; influence measure; outlier; quantile

1. Introduction

Statistical inference and prediction rely on the calculated values of various statistics or estimators. Frequently, these calculated values are biased by the contribution of one or more influential observations in the dataset, that often stand out from the bulk of the observations in some way. A sample with such observations is often called a contaminated dataset if there is reason to believe that these observations do not accurately represent the signals of interest. Under these conditions, statistical analysis and interpretation based on these estimators will be most likely erroneous. Thus, it is crucial to be able to detect such influential observations and to limit their adverse effect.

**CONTACT** Amadou Barry\textsuperscript{a} amadou.barry@mcgill.ca Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Purvis Hall, 1020 Pine Ave. West, Montreal, Qu\textsuperscript{e}bec H3A 1A2, Canada.

Supplemental data for this article is available online at https://doi.org/10.1080/03610926.2020.1841793.

© 2020 Taylor & Francis Group, LLC
For classical linear regression, there is an extensive literature on identifying influential observations and reducing their impact. These efforts led to the development of several influence measures, most of which are available in standard statistical software. These measures are generally based on the estimated coefficients and residuals of the regression model; for example, Cook’s distance measure (Cook and Weisberg 1982; Cook 1977) uses the estimated coefficients and the leave-one-out approach to identify influential observations. The Cook’s distance of an observation is defined as the difference between the full sample coefficient estimate and the coefficient estimate on the sample excluding that observation, and observations with large values of Cook’s distance are identified as influential. Cook suggested identifying these influential observations by determining a threshold based on the Fisher distribution. For more details, Chatterjee and Hadi (1986) presented an excellent literature review, including a broad class of influence measures and a description of the relationship between them and Cook’s measure.

Bias and inference problems associated with influential observations also occur in high-dimensional datasets. Indeed, high dimensionality increases the likelihood of an observation or multiple observations to be influential and amplifies their potential impact on downstream analysis. Unfortunately, traditional diagnostic measures, such as Cook’s distance, are not directly applicable in high dimension due to singularities. The gram matrix, derived from the design matrix (whose number of columns is often larger than the number of rows), is not invertible and the parameters must be estimated by regularization methods which are computationally demanding and often produce very unstable estimators. In addition, it becomes more difficult or impossible to use visualization techniques to identify influential observations when the data is very large.

Recently, the emergence of high-dimensional data in a number of fields such as genetics, medical imaging, astronomy and finance has spawned a new scientific literature on high-dimensional influence measures. In this regard, there is She and Owen’s (2010) paper, which suggest a novel penalty-based measure applicable when the number of predictors is larger than the number of observations. She and Owen (2010) introduced, into the classical regression model, an indicator variable for each observation, and a new parameter—the one mean shift parameter—in addition to the parameter of interest. This new parameter is subject to a penalty that simultaneously identifies the influential observations while finding a sparse solution.

Subsequently, using the leave-one-out approach, Zhao et al. (2013) proposed a new influence measure (high dimensional influence measure) that captures each individual’s contribution to the marginal correlation between the response variable and each predictor in the model. This new influence measure is based on the concept of “sure independence screening (SIS)” which is a variable selection method based on the marginal correlation between predictors and the response variable (Fan and Lv 2008). The high dimensional influence measure (HIM) defines the contribution of an individual observation as the difference between the full sample marginal correlation and the marginal correlation derived from the sample excluding that observation. Another important contribution of Zhao et al. (2013) paper is the derivation of the asymptotic distribution of the HIM measure, defining the statistical threshold beyond which the observations are
Zhao et al. (2013) showed the usefulness of the HIM measure for subsequent analyzes such as coefficient estimation, variable selection and screening. A few years later, Wang and Li (2017) also used the concept of correlation to develop an influence measure in the context of high-dimensional regression. Their influence measure is a combination of the leave-one-out idea and the correlation based on the characteristic function introduced by Székely, Rizzo, and Bakirov (2007). Since the asymptotic distribution of their influence measure is unknown, they used an empirical bootstrap distribution to define a statistical cutoff threshold. Unfortunately, this solution is computationally demanding, particularly in high dimensions.

Recently, Jeong and Kim (2018) investigated the impact of influential observations on estimation and variable selection for the LASSO. They derived an analytic expression of this influence on the LASSO estimator and concluded that the impact of influential observations is more severe in the high dimensional framework compared to the classical framework.

Influence measures have also been developed for high dimensional multivariate models. These influence measures adapt the Mahalanobis distance, either by penalizing the covariance matrix (Fritsch et al. 2012), or by replacing the covariance matrix by its diagonal (Ro et al. 2015) or by associating it with other statistical methods (Filzmoser, Maronna, and Werner 2008). However, in this manuscript our interest is in identifying influential observations in high dimensional univariate regression models, and therefore we do not provide further detail on these methods based on the covariance matrix.

With the exception of Zhao et al. (2013) influence measure, the proposed influence measures in high-dimensional regression do not offer any significant statistical decision rule for the identification of influential observations. It should be noted that the statistical distribution of the influence measure which gives the cutoff threshold is crucial in the determination of the influential observations. In the absence of this distribution, the decision rule is usually determined by an heuristic choice or by other subjective approaches, such as graphical visualization of the influence measure values.

In this paper, we use expectiles to introduce a new class of high dimensional influence measures: the asymmetric high dimensional influence measure (asymHIM). The expectiles, as do quantiles, describe the distribution function of a random variable. An expectile is expressed as a weighted average whose associated weight is a random variable; the expectiles, therefore, generalize the mean, which is an expectile of level $\tau = 0.5$, in the same way that quantiles generalize the median, i.e., a quantile of level 0.5. Expectile and expectile regression (ER) were introduced simultaneously by Newey and Powell (1987). Expectile and ER are attracting more and more interest, as evidenced by recent publications (Schnabel and Eilers 2009; De Rossi and Harvey 2009; Sobotka and Kneib 2012; Sobotka et al. 2013; Waltrup et al. 2015). Today, ER is being applied to different type of models: Bayesian (Majumdar and Paul 2016; Waldmann, Sobotka, and Kneib 2017; Xing and Qian 2017), non-parametric (Righi, Yang, and Ceretta 2014; Yang and Zou 2015), non-linear (Kim and Lee 2016), neural network (Xu et al. 2016; Jiang et al. 2017), support vector machine (Farooq and Steinwart 2017), and smoothing and spline with penalty (Schulze Waltrup and Kauermann 2017).
Our asymHIM influence measure, which is a generalization of Zhao et al. (2013) influence measure, is based on asymmetric correlations. In this regard, we introduce the asymmetric covariance that measures the variability of the data around the expectile level $s$:

$$s = 0.5$$

The asymmetric covariance becomes the classical covariance and the asymHIM measure corresponds to the HIM measure. Our asymHIM measure captures the heterogeneity of the impact of the observations on the variability of the data. Through this property, it increases the probability of identifying influential observations and therefore provides greater power than Zhao et al. (2013) influence measure (HIM).

To illustrate, consider a sample with 100 observations, where the response variable of the first 10 observations is contaminated (according to the design of Model I of Section 3). Figure 1 shows the asymHIM influence measure values of the first 20 observations for three values of $\tau \in \{0.25, 0.5, 0.75\}$. The influentialness of the 10 contaminated observations is better captured by the asymmetric influence measure (asymHIM) defined at level $\tau = 0.25$ or $\tau = 0.75$ than with the influence measure (HIM) defined at level $\tau = 0.5$.

In this paper, we developed a computationally efficient influence measure, and derived its asymptotic distribution to determine a statistical cutoff threshold. We evaluated the corresponding $p$-value by applying a multiple testing procedure. We adopted the Bonferroni test to control the family-wise error at the nominal level $\alpha = 5\%$. We provided a publicly available statistical package to simplify the implementation of the influence measure. With its computational efficiency, the asymHIM measure is very easy to implement and will be a very useful instrument in the toolkit of data scientists using high dimensional regression. The asymmetric points (expectiles) provide our influence measure a good ability to identify influential observations with greater power.
In the next section (Section 2), we introduce the expectile statistic, the asymmetric covariance, and the asymmetric correlation. We present the new asymmetric influence measure as well as its asymptotic distribution which will then allow choosing a cutoff threshold for influential observations based on sound statistical principles. In Section 3, we undertake an extensive simulation to assess the performance of our influence measure with respect to the HIM measure. In Section 4, we apply our asymmetric high dimensional influence measure to the Autism Brain Imaging Data Exchange (ABIDE) neuroimaging dataset, Di Martino et al. (2014), and assess its impact on the downstream analysis to predict brain maturity level. Finally, in Section 5, we present the conclusions as well as future avenues of research. We have built a R package for the influential measure which is publicly available on our GitHub (https://github.com/AmBarry/hidetify).

2. Models and methods

Notation

Vectors are written in lower bold letters, \( \mathbf{x} \in \mathbb{R}^{p} \) and matrices are represented in capital bold letters, \( \mathbf{X} \in \mathbb{R}^{n \times p} \). Estimated quantities are represented with a hat \( \hat{X} \), the inverse matrix is noted \( X^{-1} \) and \( X^T \) is the transposed matrix. \( I_{p \times p} \) or \( I_{p} \) is the identity matrix and is noted as \( I \) when the dimension is implicitly known. The symbol \( 1(t < 0) \) is the indicator function equal to 1 if \( t < 0 \) and 0 otherwise and the vectors \( 1 \) and \( 0 \) are constant vectors filled respectively with 1 and 0. The norm \( \| \cdot \| \) corresponds to the \( l_2 \) norm, and \( |S| \) is the number of elements in the set \( S \). We denote by \( S_{\text{inf}} \) the set of influential observations, and \( n_{\text{inf}} = |S_{\text{inf}}| \) its cardinality. The set \( \hat{S}_{\text{inf}} \) is its estimator, and \( S_{\text{inf}}^{C} \) is its complement. We define \( \text{Supp}(\beta) = \{ j | \beta_j \neq 0, \text{ for } j = 1, \ldots, p \} \) as the set of non-zero coefficients (active set) and \( \text{Supp}^{C}(\beta) = \{ 1, \ldots, p \} \setminus \text{Supp}(\beta) \) its complement.

Let \( \mathbf{y} = (y_1, \ldots, y_n)^T \) be the \( n \times 1 \) response vector and \( \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \) the \( n \times p \) associated design matrix with \( \mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T, \ j \in \{1, \ldots, p\} \). The observation formed by the pair \( (y_i, \mathbf{x}_i^T) \) of each subject \( 1 \leq i \leq n \) is assumed to be generated from the following regression model:

\[
  y = \beta_0 + \mathbf{x}^T \beta_1 + \epsilon, \tag{1}
\]

where the random error \( \epsilon \sim \mathcal{N}(0, 1) \) and \( \beta = (\beta_0, \beta_1)^T \) is the parameter vector. Notice that observations generated by model (1) are error free and not contaminated.

2.1. High dimensional influence measure (HIM)

Under the classical setup \( (n > p) \), Cook defined the influence measure (Cook’s distance) using the ordinary least square estimator (OLS), \( \hat{\beta}_{\text{OLS}} = (X^T X)^{-1} X^T y \). In the high dimension regression setting, the classical Cook’s distance or any other OLS based influence measure is no longer applicable, because the gram matrix is not invertible and the OLS estimator is unstable. Additionally, most of the high dimension regression estimators rely on a regularization parameter and are computationally expensive. Instead, Zhao et al. (2013) used marginal correlation between the response and the predictors to
define an influence measure in high dimension regression (HIM). Fan and Lv (2008) showed that marginal correlation is a componentwise regression estimator, which is a specific case of ridge regression with a large regularization parameter. The marginal correlation is easy to compute, and such computational advantage is critical for high dimensional data analysis.

Specifically, define $q_j = \frac{E \left[ x_{ij} \right]}{\sigma_{x_{ij}}} \frac{E \left[ y_i \right]}{\sigma_{y_i}}$ as the marginal correlation between the $j$th predictor and the response, for $j = 1, \ldots, p$, and $\mu_{x_{ij}} = E[x_{ij}], \mu_{y_i} = E[y_i], \sigma_{x_{ij}}^2 = \text{Var}[x_{ij}]$, and $\sigma_{y_i}^2 = \text{Var}[y_i]$. Denote by $\hat{q}_j = \frac{n^{-1} \sum_{i=1}^n (x_{ij} - \hat{\mu}_{x_{ij}})(y_i - \hat{\mu}_{y_i})}{(\hat{\sigma}_{x_{ij}} \hat{\sigma}_{y_i})}$ the sample estimate of $q_j$, and $\hat{\mu}_{x_{ij}}, \hat{\mu}_{y_i}, \hat{\sigma}_{x_{ij}},$ and $\hat{\sigma}_{y_i}$ the sample estimates of $\mu_{x_{ij}}, \mu_{y_i}, \sigma_{x_{ij}},$ and $\sigma_{y_i}$, respectively. Employing the leave-one-out approach, Zhao et al. (2013) quantified the influence of the $k$th observation in high dimension regression by the following measure:

$$D_k = \frac{1}{p} \sum_{j=1}^p (\hat{q}_j - \hat{q}_j^{(k)})^2, \quad 1 \leq k \leq n,$$

where $\hat{q}_j^{(k)}$ is the sample marginal correlation with the $k$th observation removed. The $k$th observation is identified as an influential observation if its measure $D_k$ is large, where large is defined by a statistical threshold. When there are no influential observations in the sample, Zhao et al. (2013) showed that $n^2 D_k \sim \chi^2(1)$, where $\chi^2(1)$ is a chi-square distribution with one degree of freedom.

In the following section we introduce a new measure we call asymmetric high dimensional influence measure (asymHIM) using the theory of expectiles. AsymHIM measure generalizes the HIM measure and increases its sensitivity.

### 2.2. Asymmetric high dimensional influence measure (asymHIM)

We start by introducing the expectile function and presenting some of its properties.

The expectile is an asymmetric weighted average which characterizes the cumulative distribution function of a random variable in the same way a quantile does. The expectile of a random variable $Y$ is defined as the solution $\mu_\tau(Y) = \frac{E \left[ \psi_\tau(Y - \mu_\tau) \right]}{E \left[ \psi_\tau(Y - \mu_\tau) \right]}$ which minimizes the following risk function

$$\mathbb{E} \{ R_\tau(Y - \theta) \}$$

over $\theta \in \mathbb{R}$ for a fixed value of $\tau \in (0, 1)$. The function $R_\tau(\cdot)$, of the form

$$R_\tau(t) = |t - 1(t \leq 0)| \cdot t^2,$$

is an asymmetric square loss function that assigns weights $\tau$ and $1 - \tau$ to positive and negative deviations, respectively.

By equating the first derivative of (2) to zero, the expectile can also be defined the solution of

$$\mu_\tau = \mathbb{E} \left[ \frac{\psi_\tau(Y - \mu_\tau)}{\mathbb{E} \left[ \psi_\tau(Y - \mu_\tau) \right]} \right].$$
where $\psi_\tau(t) = |t - 1(t \leq 0)|$ is the piecewise linear check function. Notice that, when $\tau = 0.5$ then $\psi_{0.5}(Y - \mu_{0.5}) = 0.5$ and $\mu_{0.5} = \mu = \mathbb{E}[Y]$ is the expectation of the random variable $Y$. Given a random sample, $\{(y_i)\}_{i=1}^n$, the $r$th empirical expectile

$$\hat{\mu}_r = \frac{1}{n} \sum_{i=1}^n \frac{\psi_\tau(y_i - \hat{\mu}_r)}{\sum_{i=1}^n \psi_\tau(y_i - \hat{\mu}_r)} y_i$$

is the solution which minimizes the empirical loss function

$$\frac{1}{n} \sum_{i=1}^n \mathcal{R}_\tau(y_i - \theta). \quad (3)$$

The sample expectile is computed iteratively using the iterative reweighted least square (IRLS) algorithm. Since the asymmetric square loss function is convex and continuously differentiable the IRLS algorithm converges very fast. A package (expectreg package) to estimate expectiles and other expectile models is implemented in R Core Team (2018) by Sobotka et al. (2014), Sobotka and Kneib (2012), Schnabel and Eilers (2009).

We therefore propose to use the expectile function to define an asymmetric covariance and an asymmetric correlation function. An asymmetric covariance measures the covariance between variables after centering them at the expectiles of their distribution. Using this approach, the asymmetric covariance and the asymmetric correlation between the response variable and the predictor $j$ are, for a fixed $\tau \in (0,1)$, defined as:

$$\text{Cov}_\tau(x_{ij}, y_i) := \text{Cov}_{ij} = \mathbb{E}[(x_{ij} - \mu_\tau(x_{ij}))(y_i - \mu_\tau(y_i))]$$

and

$$\rho_\tau(x_{ij}, y_i) := \rho_{ij} = \frac{\text{Cov}_{ij}(x_{ij}, y_i)}{\sigma_\tau(x_{ij})\sigma_\tau(y_i)},$$

where $\sigma_\tau^2(y_i) = \mathbb{E}[(y_i - \mu_\tau(y_i))(y_i - \mu_\tau(y_i))]$ is the asymmetric variance and $\sigma_\tau^2(x_{ij})$ is defined similarly. Notice that, when $\tau = 0.5$ then $\mu_{0.5} = \mu, \text{Cov}_{0.5j} = \mathbb{E}[(x_{ij} - \mu(x_{ij}))(y_i - \mu(y_i))]$, and $\rho_{0.5j} = \frac{\text{Cov}_{0.5j}}{\sigma_{0.5}(x_{ij})\sigma_{0.5}(y_i)}$ are the mean, the covariance and the classical correlation, respectively.

The corresponding sample estimate of the asymmetric covariance and correlation function, for a fixed $\tau \in (0,1)$, can be defined as:

$$\hat{\text{Cov}}_\tau(x_{ij}, y_i) := \hat{\text{Cov}}_{ij} = n^{-1} \sum_{i=1}^n (x_{ij} - \hat{\mu}_\tau(x_{ij}))(y_i - \hat{\mu}_\tau(y_i))$$

and

$$\hat{\rho}_\tau(x_{ij}, y_i) := \hat{\rho}_{ij} = \frac{\hat{\text{Cov}}_{ij}}{\hat{\sigma}_\tau(x_{ij})\hat{\sigma}_\tau(y_i)},$$

where $\sigma_\tau^2(x_{ij}) = n^{-1} \sum_{i=1}^n (x_{ij} - \hat{\mu}_\tau(x_{ij}))^2$ is the empirical asymmetric variance and $\hat{\sigma}_\tau^2(y_i)$ is defined similarly.

Therefore, using the leave-one-out principle, we introduce an asymmetric high dimensional influence measure for a subject $k \in \{1, \ldots, n\}$ and for a single $\tau$ as:

$$D_{tk} := \frac{1}{p} \sum_{j=1}^p (\hat{\rho}_{ij} - \hat{\rho}_{tk})^2, \quad (4)$$
where $\hat{\rho}_{ij}^{(k)}$ is the marginal asymmetric correlation with the $k$th observation removed and is defined as:

$$\hat{\rho}_{ij}^{(k)} = \frac{\sum_{l=1, l \neq k}^{n} (x_{ij} - \hat{\mu}_{ij}^{(k)}(x_{ij})) (y_{ij} - \hat{\mu}_{ij}^{(k)}(y_{ij}))}{(n - 1) \hat{\sigma}_{ij}^{(k)}(x_{ij}) \hat{\sigma}_{ij}^{(k)}(y_{ij})}, \quad j = 1, \ldots, p, \ k = 1, \ldots, n.$$  

The estimates $\hat{\mu}_{ij}^{(k)}(x_{ij}), \hat{\mu}_{ij}^{(k)}(y_{ij}), \hat{\sigma}_{ij}^{(k)}(x_{ij}), \text{ and } \hat{\sigma}_{ij}^{(k)}(y_{ij})$ are the sample estimates with the $k$th observation removed.

The asymHIM measure retains the same computational advantages as the HIM measure of Zhao et al. (2013). As noted by Zhao et al. (2013), the impact of the influence measure is not limited to the marginal correlation alone. On the contrary, any substantial change in the marginal correlation will have implications on downstream analyses such as variable selection and parameter estimation. In Section 3, we show the effect of identifying influential observations with asymHIM measure on parameter estimation and variable selection.

Several different strategies can be proposed for using the asymmetric influence measure, depending on the value of $\tau$. For example, we could choose a sequence of $\tau$ covering the left, right and center of the distribution and select the best measure based on criteria such as true positive rate $\text{TPR}_{\text{inf}}$ and false positive $\text{FPR}_{\text{inf}}$ (which are defined in Section 3). We could also use the max and the min of that sequence as an influence measure. We tested all these scenarios and the simulation results, which are not shown here, were the same for any single $\tau$. However, when summarizing across values of $\tau$, we found that the min influence measure had small true positive rate ($\text{TPR}_{\text{inf}}$), whereas the max influence measure demonstrated inflated false positive rate ($\text{FPR}_{\text{inf}}$). Among all the scenarios, the sum based asymmetric influence measure led to a better tradeoff in terms of $\text{TPR}_{\text{inf}}$ rate and $\text{FPR}_{\text{inf}}$ rate. Therefore, in the following, we propose the sum based asymmetric influence measure as an instrument for identifying influential observations in high dimensional regression.

Therefore, we define our asymmetric high dimensional influence measure (asymHIM) for a subject $k$ and for a sequence of asymmetric points $(\tau_1, \ldots, \tau_q)$ as:

$$\text{asymD}_k = \frac{1}{p} \sum_{l=1}^{q} \sum_{j=1}^{p} (\hat{\rho}_{ij} - \hat{\rho}_{ij}^{(k)})^2 = \sum_{l=1}^{q} D_{\tau_l,k}, \quad k = 1, \ldots, n,$$

where $\hat{\rho}_{ij}$ and $\hat{\rho}_{ij}^{(k)}$ are the sample asymmetric correlation estimates for the full sample and with the $k$th observation removed and for a fixed $\tau_l, l = 1, \ldots, q$. This new defined measure will help capture possible asymmetry in the data across several different values of $\tau$ across the distributional range $(0, 1)$. The asymHIM influence measure did not perform as well when the sequence of $\tau$s included the extremes of the distribution (e.g., expectiles such as 0.95, 0.09, 0.1, 0.05); these expectiles have a higher probability of being influential observations themselves or being affected by them.

In our explorations, we noted that the asymHIM influence measure based on 5 values, i.e., $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ or 3 values of $\tau \in \{0.25, 0.5, 0.75\}$ gave better true positive and false positive rates, when compared to larger series of values for $\tau$. Therefore, in Section 3, we present the results of the asymHIM measure based on 5
(denoted asymHIM5) and 3 (asymHIM3) values of \( \tau \). The term asymHIM refers to either of the two variants. We also present in the Supplementary material file the impact of using a larger number of expectiles, especially when some of the expectiles lie at the extreme end of the distribution.

### 2.3. Asymptotic properties of asymHIM

An important step in developing an influence measure is the derivation of its asymptotic distribution. Toward that goal, we establish in this section the asymptotic distribution of the asymHIM measure under the following conditions.

**H1.** For any fixed \( \tau \in (0, 1) \) and \( 1 \leq j \leq p, \rho_{ij} \) is constant and does not change as \( p \) increases.

**H2.** The asymmetric covariance matrix \( \Sigma_\tau = \text{Cov}_\tau(x) = \mathbb{E}[(x - \mu_\tau(x))(x - \mu_\tau(x))^\top] \), with the eigen decomposition \( \Sigma_\tau = \sum_{j=1}^{p} \lambda_j u_j u_j^\top \), is assumed to verify \( \lambda_j = O(p^r) \) for some \( 0 \leq r < 2 \) and for any fixed \( \tau \in (0, 1) \).

**H3.** The predictor \( x_i^\top \) follows a multivariate normal distribution and the random noise \( e_i \) follows a normal distribution.

The stated conditions are similar to those stated by Zhao et al. (2013) Condition **H1** assumes a fixed correlation \( \rho_{ij} \) between the response and the predictor \( j \) for any fixed \( \tau \) and independently of \( p \). Condition **H2** permits high values of the eigenvalues of the covariance matrix \( \Sigma_\tau \), but at a rate controlled by the dimensionality \( p \). A sufficient condition would be that \( \max_{1 \leq j \leq p} \lambda_j \) be bounded. The normality assumption, Condition **H3**, is used, among others, to ensure independence across columns of the matrix \( X \). For reading convenience we introduce the following notation: \( l_{ij}^\tau = l_{ij}^\tau(x_{ij}), l_\tau = l_\tau(y_i), \sigma_{ij} = \sigma_{\tau}(x_{ij}) \) and \( \sigma_\tau = \sigma_\tau(y_i) \).

**Theorem 1.** Assume conditions **H1–H3** and that the following parameters, \( \mu_{ij}, \mu_\tau, \sigma_{ij} \) and \( \sigma_\tau \) are known. When there are no influential points and \( \min(n, p) \to \infty \), then

\[
n^2\text{asymD}_k = n^2 \sum_{j=1}^{q} D_{qj} \sim \chi_2^2(q),
\]

where \( \chi_2^2(q) \) is the chi-square distribution with \( q \) degrees of freedom, and \( q \) is the number of asymmetric points. Notice that the number of asymmetric points is adjusted by the number of degrees of freedom of the chi-square distribution.

**Theorem 1** is stated under the assumption that the parameters \( \mu_{ij}, \mu_\tau, \sigma_{ij} \) and \( \sigma_\tau \) are known. However, in reality they are unknown and are replaced by their estimator. We now replace them by their \( \sqrt{n} \) consistent estimators and show that the asymptotic result continues to hold. We choose their corresponding sample moment estimates: \( \hat{\mu}_{ij}, \hat{\mu}_\tau, \hat{\sigma}_{ij} \) and \( \hat{\sigma}_\tau \). Notice that we used robust estimators to compute the asymmetric influence measure in practice. We estimated \( \mu_{ij} \) and \( \mu_\tau \) by the corresponding empirical quantile of level \( \tau \) and \( \sigma_{ij} \) and \( \sigma_\tau \) by the median absolute deviation (MAD) estimator. To derive the next result we introduce the following quantities:
\[ (Q_{ft}, R_{ft}) = \left( (\hat{\mu}_{ft} - \mu_{ft})/\sigma_{ft}, \sigma_{ft}/\hat{\sigma}_{ft} \right) \text{ and } (Q_{r}, R_{r}) = \left( (\hat{\mu}_{r} - \mu_{r})/\sigma_{r}, \sigma_{r}/\hat{\sigma}_{r} \right). \]

In addition, let:

\[
S_{Q_{ft}} = \limsup_{n \to \infty} \mathbb{E} \left[ n^{1/2} Q_{1_{ft}} \right]^8, \quad S_{R_{ft}} = \limsup_{n \to \infty} \mathbb{E} \left[ n^{1/2} (R_{1_{ft}} - 1) \right]^8,
\]
\[
S_{Q_{r}} = \limsup_{n \to \infty} \mathbb{E} \left[ n^{1/2} Q_{r} \right]^8, \quad S_{R_{r}} = \limsup_{n \to \infty} \mathbb{E} \left[ n^{1/2} (R_{y_{r}} - 1) \right]^8.
\]

We make the following additional assumption.

**H4.** For all \( 1 \leq j \leq p, (Q_{ft}, R_{ft}) \) are the same symmetric function of \( \{x_{ij} = (x_{ij} - \mu_{ij})/\sigma_{ij}, \text{ for } 1 \leq t \leq n\} \) and \( (Q_{r}, R_{r}) \) are also the same symmetric function of \( \{\hat{x}_{ij} = (\hat{y}_{ij} - \mu_{ij})/\sigma_{ij}, \text{ for } 1 \leq t \leq n\} \). We assume that \( S_{Q_{r}}, S_{R_{r}}, S_{Q_{ft}}, \text{ and } S_{R_{ft}} \) are finite.

What we mean by the “same” is that, for all \( 1 \leq j \leq p, ((\hat{\mu}_{ij} - \mu_{ij})/\sigma_{ij}, \sigma_{ij}/\hat{\sigma}_{ij}) = f(\hat{x}_{ij}, ..., \hat{x}_{n_{ij}}), \) where \( f(x_1, ..., x_p) = (f_1(x_1, ..., x_p), f_2(x_1, ..., x_p)) \), and \( f_1 \) and \( f_2 \) are symmetric functions. Condition **H4** is used by Zhao et al. (2013) and the definition is to emphasize the symmetric properties which will be used in the proof.

**Proposition 2.** For a fixed \( \tau \in (0, 1) \) and \( j = 1, ..., p \), assume that \( \hat{\mu}_{ij}, \mu_{ij}, \sigma_{ij}, \text{ and } \hat{\sigma}_{ij} \) are \( \sqrt{n} \) consistent and satisfy assumption **H4**. Substituting \( \mu_{ij}, \mu_{ij}, \sigma_{ij} \) and \( \sigma_{ij} \) with their corresponding estimates in \( D_{ikb} \), **Theorem 1** continues to hold under the same conditions.

The proofs of **Theorem 1** and **Proposition 2** are in the Supplementary material file.

In their paper, Zhao et al. (2013) showed that the influence measure (when \( \tau = 0.5 \)) is a function of two main terms that play similar roles to those of the residuals and the diagonal elements of the hat matrix in the expression of the classical Cook’s distance. They also showed that the influence measure identifies the influential observation with probability one as \( n \) and \( p \to \infty \). All these statistical properties remain for the asymmetric influence measure.

### 3. Simulation

#### 3.1. Design

In this section, we evaluated the performance of the asymHIM measure through extensive simulations. The simulated samples are generated from model equation (1) and the contaminated data are generated by the following model

\[ \hat{y}_i = \hat{x}_i^\top \hat{\beta} + \hat{\epsilon}_i, \quad 1 \leq i \leq \tilde{n}, \]

where \( \tilde{n} = 0.1n \) is the number of subjects with contaminated data. Following Zhao et al. (2013) contamination scheme, we contaminated the data according to two different schemes: Model I and Model II.

In Model I, only the response variable of the \( \tilde{n} \) subjects is altered and the predictors are left unchanged, i.e., \( \hat{x}_i = x_i \). The alteration is generated from the following equation:

\[ \hat{y}_i = x_i^\top \hat{\beta} + \epsilon_i, \] where \( \hat{\beta} = \beta + \kappa \gamma \). The parameter \( \gamma \) allows the contamination of the response variable by creating a false relationship between the response variable and the variables without any prediction power. The parameter \( \kappa \in (0, 0.4, 0.8, 1.2, 1.6) \) controls the degree of contamination, with larger values resulting in points that display more deviation. When \( \kappa = 0 \), there is no contamination and the dataset is clean. That is, there
are no apriori-known influential points in the dataset. The values of $\beta$ and $\gamma$ are set below.

In Model II, only the predictors $\tilde{x}_i$ of the $n$ subjects are altered and the response variable is unaltered, i.e., $\tilde{y}_i = y_i$. The predictors contamination is set up under four different scenarios according to the degree of contamination, and the contaminated region ($S$). We contaminated either the first $\tilde{p}_1, \tilde{p}_2$ predictors, $S_1 = \{1, ..., \tilde{p}_1\}$ and $S_2 = \{1, ..., \tilde{p}_2\}$, or the last $\tilde{p}_1, \tilde{p}_2$ predictors, $S_3 = \{p - \tilde{p}_1 + 1, ..., p\}$ and $S_4 = \{p - \tilde{p}_2 + 1, ..., p\}$, using two degree of contamination $\tilde{p}_1 = 0.1p$ and $\tilde{p}_2 = 0.3p$. The contamination is formalized by the following equation: $\tilde{x}_i = x_i^T + 30\kappa z_i^T$ where $z_i = (z_{i1}, ..., z_{ip})^T$, $z_{ij} = \tilde{v}_{ij}1(j \in S)$ and $\tilde{v}_{ij} \sim \mathcal{N}(0, 1)$. A summary of the contamination schemes is presented in Table 1.

To account for the interaction between the contaminated region $S \in \{S_1, ..., S_4\}$, and the predictive region, we set the number of non-zero coefficients for $\beta$ to $\tilde{p}_c = 0.01p$ and distribute the values in three different ways: $\beta_1 = (\beta_1, ..., \beta_{\tilde{p}_c}, 0, ..., 0)^T$, $\beta_2 = (0, ..., 0, 0, \beta_{p-\tilde{p}_c+1}, ..., \beta_p)^T$, and $\beta_3 = (\beta_1, 0, ..., 0, \beta_{101}, 0, ..., 0, \beta_{201}, 0, ..., 0, \beta_{p-99}, 0, ..., 0)^T$.

For example, when $\beta = \beta_1$ and the contaminated regions are $S_1$ or $S_2$ then the important predictors are contaminated, and when the contaminated regions are $S_3$ or $S_4$ then the contamination takes place only among variables with no predictive power. The values of the parameter $\gamma$ are defined in accordance with the values of the parameter $\beta$

- $\gamma_1 = (0, ..., 0, 1, ..., 1)^T$
- $\gamma_2 = (1, ..., 1, ..., 1, 0, ..., 0)^T$
- $\gamma_3 = (0, 1, ..., 1, 0, 1, ..., 1, 0, 1, ..., 1)^T$

We set the number of subjects $n \in (100, 500)$ and the number of predictors $p \in (1000, 5000)$. The simulation is carried out with 200 replications for each samples. We estimated the asymmetric correlation at 5 $\tau \in (0.3, 0.4, 0.5, 0.6, 0.7)$, and 3 different points $\tau \in (0.25, 0.5, 0.75)$.

We evaluated the performance of the asymHIM measure according to two criteria. In the first criterion, we measured its power to retrieve all influential observations with the true positive rate ($TPR_{inf}$). For the second criterion, we computed the false positive rate $FPR_{inf}$ to assess its weakness to falsely report as influential observations that are not. Formally, we have:

\[
TPR_{inf} = \frac{|S_{inf} \cap \hat{S}_{inf}|}{n_{inf}} \quad \text{and} \quad FPR_{inf} = \frac{|S_{inf} \cap \hat{S}_{inf}|}{n - n_{inf}},
\]

where $S_{inf}$ is the set of influential observations, $n_{inf}$ its size, $\hat{S}_{inf}$ its estimator generated by a influence measure. We generated the influential set $\hat{S}_{inf}$ estimator by controlling the family-wise error at the nominal level $\alpha = 5\%$. Among the multiple testing procedures we chose the Bonferroni test.
In the second criterion, we evaluated the performance of the asymHIM influence measure in relation to its impact on coefficient estimation and variable selection. To do this, we fit LASSO models to the raw data and to the clean data obtained after applying the influence measure. Then we evaluated the accuracy of the coefficient estimates and their support (the set of non-zero coefficients estimator). The accuracy of the coefficient estimates is evaluated by the following error function:

\[
\text{ERR} = \sqrt{\| \hat{\beta} - \beta \|_2^2}
\]

where \( \beta \) and \( \hat{\beta} \) are respectively the parameter, and its estimator and, \( \| \cdot \|_2 \) is the \( l_2 \) norm.

We evaluated the influence measure in terms of variable selection by comparing the support of the parameter (the set of non-zero parameters) with that of the parameter estimator. For this we have defined two criteria: the true positive rate (TPR) and the false positive rate (FPR). The TPR is the percentage of the true non-zero coefficients selected by the model and the FPR is the percentage of the false non-zero coefficients selected by the model. Let \( \text{Supp}(\beta) = \{j | \beta_j \neq 0, \text{ for } j = 1, ..., p \} \) and \( \text{Supp}(\hat{\beta}) = \{1, ..., p \} \setminus \text{Supp}(\beta) \) then the TPR and the FPR are defined as:

\[
\text{TPR} = \frac{| \text{Supp}(\beta) \cap \text{Supp}(\hat{\beta}) |}{| \text{Supp}(\beta) |} \quad \text{and} \quad \text{FPR} = \frac{| \text{Supp}(\beta) \cap \text{Supp}(\hat{\beta}) |}{| \text{Supp}(\hat{\beta}) |}.
\]

Simulations were conducted using high performance computing clusters provided by Calcul Quebec and Compute Canada. All computations are performed with the R (v3.5.0) statistical programming language R Core Team (2018). The implemented R package \texttt{hidetify} that comes with this manuscript is publicly available on Github at https://github.com/AmBarry/hidetify. Recall that asymHIM5 is the asymHIM measure based on 5 values of \( \tau \in (0.3, 0.4, 0.5, 0.6, 0.7) \) and asymHIM3 is the asymHIM measure based on 3 values of \( \tau \in (0.25, 0.5, 0.75) \). In the following, asymHIM refers to asymHIM5 and asymHIM3 at the same time. Note that the simulation results for the parameter values \( \beta_2 \) and \( \beta_3 \) are in the Supplementary material file.

### 3.2. Results

#### 3.2.1. TPR\textsubscript{inf} and FPR\textsubscript{inf}.

Figure 2 reports the results on the TPR\textsubscript{inf} and FPR\textsubscript{inf} of the three influence measures: asymHIM5, asymHIM3 and HIM, in Model I. The two columns show results for different sample sizes (100 and 500) and the two rows are different numbers of predictors (1000 and 5000). We have on the x-axis, the degree of contamination of the datasets and on the y-axis the values of the TPR\textsubscript{inf} and FPR\textsubscript{inf} of the influence measures. The TPR\textsubscript{inf} and FPR\textsubscript{inf} results of Model II are reported in Figure 3, where the first column (A) shows results for \( n = 100 \) and the second (B) shows results for \( n = 500 \). An effective influence measure will have large TPR\textsubscript{inf} \sim 1 and small FPR\textsubscript{inf} \sim 0.

The power (TPR\textsubscript{inf}) of the influence measures increases with the degree of contamination, while the FPR\textsubscript{inf} remains relatively stable when the data is contaminated, particularly for asymHIM3 and HIM measures. Our influence measures (asymHIM5 and
asymHIM3) have higher power than the HIM measure ($\tau = 0.5$) and the difference is greater in Model II where the power of the HIM measurement can be below 30%, (Figure 3). Our influence measures have larger $FPR_{\text{inf}}$ than the HIM measure. However, the $FPR_{\text{inf}}$ of the asymHIM3 influence measure is below the nominal level 5% and is below 2% when the sample size increases ($n = 500$), Figure 3. Finally, in Model II, the results are similar according to the contaminated regions ($S_1 - S_4$):

Overall, our influence measures perform better than the HIM measurement ($\tau = 0.5$) particularly in Model II. The HIM influence measure has smaller $FPR_{\text{inf}}$ but in terms of $TPR_{\text{inf}}$ and $FPR_{\text{inf}}$ ratio, the asymHIM3 influence measure, whose power is comparable to that of asymHIM5 and its $FPR_{\text{inf}}$ to that of HIM, is more efficient.

Results for parameter settings $\beta = \beta_2$ and $\beta = \beta_3$ lead to the same conclusions and are in the Supplementary material file.

### 3.2.2. Variable selection and parameter estimation

Table 2 and Table 3 report results of criteria ERR, TPR and FPR which serve to evaluate the impact of the correlation-based influence measures on the downstream statistical analyzes such as parameter estimation and variable selection.

In Model I, we observe that the TPR and FPR of all the methods are, relatively, at their optimal value when the dataset is clean ($\kappa = 0$). The results show that the bias (ERR) of the LASSO increases significantly with respect to the degree of contamination ($\kappa$), while the bias of the LASSO + asymHIM and LASSO + HIM decreases. Overall, the bias of the LASSO + asymHIM, especially that of the LASSO + asymHIM3, is smaller than that of the LASSO + HIM. The LASSO + asymHIM has a better performance in
Figure 3. $TPR_{inf}$ and $FPR_{inf}$ of the different influence measures (asymHIM5, AsymHIM3, HIM), in Model II according to the sample size (100, 500), the number of predictors (1000, 5000), the parameter value $\beta_1$, the contaminated region ($S_1 - S_4$), and the degree of contamination ($\kappa$). First column (A) show results for $n = 100$ and second column (B) for $n = 500$. 
Table 2. Impact is measured by bias (ERR), true positive rate (TPR) and false positive rate (FPR) according to the sample size \( n \in (100, 500) \), the number of predictors \( p \in (1000, 5000) \), the parameter value \( \beta_1 \), and the degree of contamination \( \kappa \in (0, 0.4, 0.8, 1, 1.6) \).

| Criterion | Method  | \( p = 1000 \) | \( n = 100 \) | \( n = 500 \) | \( p = 5000 \) | \( n = 100 \) | \( n = 500 \) |
|-----------|---------|----------------|----------------|----------------|----------------|----------------|----------------|
|           |         | \( \kappa = 0 \) | 0.4 | 0.8 | 1.2 | 1.6 | 0 | 0.4 | 0.8 | 1.2 | 1.6 |
| ERR       | LASSO   | 0.64           | 2.92 | 3.7 | 4.09 | 4.37 | 5.68 | 7.52 | 8.11 | 8.57 | 9.11 |
|           | asymHIM5| 0.8            | 1.2 | 1.08 | 1.01 | 0.96 | 6.6 | 6.85 | 6.73 | 6.74 | 6.73 |
|           | asymHIM3| 0.74           | 1.18 | 1.03 | 0.97 | 0.92 | 6.31 | 6.76 | 6.65 | 6.63 | 6.62 |
|           | HIM     | 0.64           | 1.5 | 1.29 | 1.15 | 1.07 | 5.74 | 6.88 | 6.76 | 6.59 | 6.53 |
| TPR       | LASSO   | 0.92           | 0.26 | 0.06 | 0.02 | 0.01 | 0.37 | 0.03 | 0.01 | 0 | 0 |
|           | asymHIM5| 0.9           | 0.8 | 0.85 | 0.85 | 0.87 | 0.2 | 0.15 | 0.17 | 0.17 | 0.17 |
|           | asymHIM3| 0.91           | 0.79 | 0.86 | 0.86 | 0.87 | 0.26 | 0.18 | 0.2 | 0.21 | 0.21 |
|           | HIM     | 0.91           | 0.71 | 0.74 | 0.8 | 0.83 | 0.36 | 0.16 | 0.19 | 0.22 | 0.24 |
| FPR       | LASSO   | 0.02           | 0.02 | 0.01 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 |
|           | asymHIM5| 0.02          | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
|           | asymHIM3| 0.02          | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
|           | HIM     | 0.02           | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
terms of TPR, while the FPR of the LASSO is relatively smaller than that of the LASSO + asymHIM and LASSO + HIM which have same FPR values.

In general, the LASSO + asymHIM estimate has lower bias (ERR) and its support (TPR) includes most often the parameter support. The results for $b_2$ and $b_3$ in Model I are similar and are in the Supplementary material file.

The results of Model II, Table 3, depend on the overlap between the support of the parameter and the contaminated region, $S_1 - S_4$. When the support of the parameter $\beta$ overlap with the contaminated region then the LASSO + asymHIM method outperforms the LASSO + HIM and the LASSO methods in terms of ERR, TPR and FPR. Otherwise, the three methods perform similarly in terms of ERR and TPR and LASSO + asymHIM fall behind the LASSO + HIM and the LASSO methods in terms of FPR. For example, when the support of the parameter $\beta$ is that of $b_1$ and the contaminated regions are $S_1$ and $S_2$ then LASSO + asymHIM outperforms the other methods and performs as well as the other methods when the contaminated regions are $S_3$ and $S_4$.

When the support of the parameter $\beta$ is that of $b_2$ and the contaminated regions are $S_3$ and $S_4$ then the LASSO + asymHIM outperforms the other methods and performs similarly otherwise. Finally, when the support of the parameter $\beta$ is that of $b_3$ then the support overlap with all the contaminated regions $S_1 - S_4$. In this case, the LASSO + asymHIM performs better than the other methods in terms of ERR and TPR and its FPR is slightly higher than the other methods. We also observe in Model II the same phenomenon observed in Model I, when $n = 500$ and $p = 1000$. The results of $b_2$ and $b_3$ are in the Supplementary material file.

We observe that the LASSO method slightly outperforms other methods when there is no crossover between the predictive variables with non-zero coefficients and the contaminated region $S \in \{S_1, ..., S_4\}$. We also notice a poor performance of all methods (high ERR, small TPR), when $p = 1000$, and $n = 500$ (Table 2 and Table 3).

### 3.2.3. Computation time

We used the microbenchmark package (Mersmann 2018), with 100 replications, to evaluate the computation time of the influence measures from a sample of size $n = 500$ with $p = 5000$ predictors. It takes on average 5 seconds to the HIM measure to identify influential observations from a sample of size $n = 500$ with $p = 5000$ predictors, 7 seconds for the asymHIM3 measure, and only 11 seconds for the asymHIM5 measure (Figure 4).

### 3.2.4. Masking and swamping effects

We also simulated a third model, where both the response and the predictors of the first 10 observations are contaminated. The response is contaminated according to the scheme of Model I and the predictors are contaminated according to the scheme of Model II. We applied our asymHIM measure and Zhao et al. (2013) influence measure (HIM) to detect the influential observations, and neither method performed well. Results, which are not shown here, displayed high false positive rates ($\text{FPR}_{\text{inf}} > 50\%$) for both methods. That is, both methods falsely reported as influential many good
Table 3. Impact is measured by bias (ERR), true positive rate (TPR), and false positive rate (FPR) according to the sample size $n \in (100, 500)$, the number of predictors $p \in (1000, 5000)$, the parameter value $\beta_1$, the contaminated region $\left(S_1 - S_4\right)$ and the degree of contamination, $\kappa \in (0, 0.4, 0.8, 1.2, 1.6)$.

| Subset | Criterion | Method | $\kappa = 0$ | 0.4 | 0.8 | 1.2 | 1.6 | 0 | 0.4 | 0.8 | 1.2 | 1.6 | 0 | 0.4 | 0.8 | 1.2 | 1.6 | 0 | 0.4 | 0.8 | 1.2 | 1.6 |
|--------|-----------|--------|-------------|-----|-----|-----|-----|---|-----|-----|-----|-----|---|-----|-----|-----|-----|---|-----|-----|-----|-----|
| ERR    | LASSO     | 0.64   | 3.17        | 3.19 | 3.19 | 3.2 | 5.68 | 7.23 | 7.25 | 7.25 | 7.25 | 0.24 | 3.01 | 3.13 | 3.13 | 3.13 | 0.64 | 7.08 | 7.08 | 7.09 | 7.08 |
|        | asymHIM5  | 0.8    | 2.86        | 0.88 | 0.88 | 0.89 | 6.6 | 7.18 | 6.79 | 6.79 | 6.79 | 0.29 | 2.8 | 0.6 | 0.3 | 0.3 | 0.74 | 7.01 | 0.81 | 0.8 | 0.8 |
|        | asymHIM3  | 0.74   | 2.92        | 0.8  | 0.8  | 0.8  | 6.31 | 7.18 | 6.61 | 6.61 | 6.63 | 0.27 | 2.8 | 0.62 | 0.28 | 0.28 | 0.69 | 7.04 | 0.78 | 0.75 | 0.75 |
|        | HIM       | 0.64   | 3.12        | 3.08 | 3.29 | 2.92 | 5.74 | 7.25 | 7.19 | 7.15 | 7.05 | 0.24 | 2.93 | 3.11 | 3.12 | 3.13 | 0.64 | 7.09 | 7.08 | 7.08 | 7.08 |
| TPR    | LASSO     | 0.92   | 0.04        | 0.01 | 0.01 | 0.01 | 0.37 | 0    | 0    | 0    | 0    | 0.98 | 0.46 | 0.05 | 0.02 | 0.01 | 0.98 | 0.03 | 0    | 0    | 0 |
|        | asymHIM5  | 0.9    | 0.14        | 0.88 | 0.89 | 0.89 | 0.2  | 0.01 | 0.15 | 0.15 | 0.14 | 0.98 | 0.52 | 0.9 | 0.98 | 0.98 | 0.98 | 0.98 | 0.05 | 0.98 | 0.98 | 0.98 |
|        | asymHIM3  | 0.91   | 0.13        | 0.89 | 0.89 | 0.89 | 0.26 | 0.01 | 0.19 | 0.19 | 0.19 | 0.99 | 0.55 | 0.89 | 0.98 | 0.98 | 0.98 | 0.98 | 0.05 | 0.97 | 0.98 | 0.98 |
|        | HIM       | 0.91   | 0.05        | 0.06 | 0.07 | 0.1  | 0.36 | 0    | 0    | 0    | 0.02 | 0.98 | 0.57 | 0.14 | 0.04 | 0.02 | 0.98 | 0.02 | 0    | 0    | 0 |
|        |           |        |             |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| ERR    | LASSO     | 0.64   | 3.17        | 3.19 | 3.2  | 3.18 | 5.68 | 7.22 | 7.25 | 7.23 | 7.25 | 0.24 | 2.99 | 3.13 | 3.13 | 3.13 | 0.64 | 7.08 | 7.08 | 7.08 | 7.09 |
|        | asymHIM5  | 0.8    | 1.02        | 0.87 | 0.89 | 0.95 | 6.6 | 6.78 | 6.75 | 6.81 | 6.83 | 0.29 | 1.31 | 0.3 | 0.3 | 0.3 | 0.74 | 2.96 | 0.79 | 0.8 | 0.8 |
|        | asymHIM3  | 0.74   | 0.93        | 0.8  | 0.8  | 0.83 | 6.31 | 6.64 | 6.62 | 6.62 | 6.65 | 0.27 | 1.4 | 0.28 | 0.28 | 0.28 | 0.69 | 3.17 | 0.74 | 0.74 | 0.74 |
|        | HIM       | 0.64   | 2.82        | 2.86 | 2.62 | 2.4  | 5.74 | 7.17 | 7.06 | 6.99 | 6.93 | 0.24 | 2.76 | 3.04 | 3.05 | 3.03 | 0.64 | 7.04 | 7.07 | 7.04 | 7.01 |
| TPR    | LASSO     | 0.92   | 0.04        | 0.01 | 0.01 | 0.01 | 0.37 | 0    | 0    | 0    | 0    | 0.98 | 0.45 | 0.06 | 0.02 | 0.02 | 0.98 | 0.04 | 0    | 0    | 0 |
|        | asymHIM5  | 0.9    | 0.83        | 0.89 | 0.88 | 0.88 | 0.2  | 0.14 | 0.16 | 0.14 | 0.14 | 0.98 | 0.76 | 0.98 | 0.98 | 0.98 | 0.98 | 0.74 | 0.98 | 0.98 | 0.98 |
|        | asymHIM3  | 0.91   | 0.84        | 0.89 | 0.89 | 0.89 | 0.26 | 0.01 | 0.2  | 0.2  | 0.18 | 0.99 | 0.75 | 0.98 | 0.98 | 0.98 | 0.98 | 0.72 | 0.98 | 0.98 | 0.98 |
|        | HIM       | 0.91   | 0.16        | 0.13 | 0.22 | 0.3  | 0.36 | 0.01 | 0.03 | 0.06 | 0.08 | 0.98 | 0.56 | 0.19 | 0.08 | 0.08 | 0.98 | 0.09 | 0.01 | 0.01 | 0.01 |
|        |           |        |             |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| ERR    | LASSO     | 0.64   | 0.63        | 0.63 | 0.63 | 0.63 | 5.68 | 5.62 | 5.62 | 5.61 | 5.6 | 0.24 | 0.24 | 0.24 | 0.24 | 0.24 | 0.64 | 0.63 | 0.63 | 0.63 | 0.63 |
|        | asymHIM5  | 0.8    | 0.86        | 0.87 | 0.88 | 0.89 | 6.6 | 6.7 | 6.76 | 6.79 | 6.81 | 0.29 | 0.3 | 0.3 | 0.3 | 0.3 | 0.74 | 0.78 | 0.8 | 0.8 | 0.8 |
|        | asymHIM3  | 0.74   | 0.78        | 0.8  | 0.8  | 0.8  | 6.31 | 6.51 | 6.6 | 6.6 | 6.62 | 0.27 | 0.28 | 0.28 | 0.28 | 0.28 | 0.69 | 0.72 | 0.75 | 0.75 | 0.75 |
|        | HIM       | 0.64   | 0.68        | 0.68 | 0.68 | 0.69 | 5.74 | 5.95 | 6.05 | 6.07 | 6.13 | 0.24 | 0.26 | 0.26 | 0.26 | 0.26 | 0.64 | 0.66 | 0.67 | 0.68 | 0.68 |

(continued)
Table 3. Continued.

| Subset | Criterion | Method | $\kappa = 0$ | 0.4 | 0.8 | 1.2 | 1.6 | 0.4 | 0.8 | 1.2 | 1.6 | 0.4 | 0.8 | 1.2 | 1.6 | 0.4 | 0.8 | 1.2 | 1.6 |
|--------|-----------|--------|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| TPR    | LASSO     | 0.92   | 0.92         | 0.92| 0.92| 0.37| 0.39| 0.39| 0.39| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98|
|        | asymHIM5  | 0.9    | 0.89         | 0.89| 0.89| 0.2 | 0.16| 0.15| 0.15| 0.14| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98|
|        | asymHIM3  | 0.91   | 0.9          | 0.89| 0.89| 0.26| 0.22| 0.19| 0.19| 0.19| 0.99| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98|
|        | HIM       | 0.91   | 0.91         | 0.91| 0.9| 0.36| 0.33| 0.31| 0.3| 0.29| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98|
| S3     | FPR       | 0.02   | 0.02         | 0.02| 0.02| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02|
|        | LASSO     | 0.02   | 0.02         | 0.02| 0.02| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02|
|        | asymHIM5  | 0.02   | 0.02         | 0.02| 0.02| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02|
|        | asymHIM3  | 0.02   | 0.02         | 0.02| 0.02| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02|
|        | HIM       | 0.02   | 0.02         | 0.02| 0.02| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02| 0.02|
| S4     | ERR       | 0.64   | 0.62         | 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62| 0.62|
|        | LASSO     | 0.8    | 0.85         | 0.86| 0.9 | 0.95| 6.6 | 6.73| 6.77| 6.79| 6.83| 0.29| 0.29| 0.3 | 0.3 | 0.3 | 0.3 | 0.74| 0.77| 0.79| 0.8 |
|        | asymHIM5  | 0.74   | 0.78         | 0.79| 0.81| 0.83| 6.31| 6.57| 6.62| 6.67| 6.72| 0.27| 0.27| 0.28| 0.28| 0.28 | 0.28 | 0.69| 0.73| 0.75| 0.74|
|        | asymHIM3  | 0.64   | 0.66         | 0.67| 0.68| 0.69| 5.74| 5.82| 5.93| 6.01| 6.07| 0.24| 0.25| 0.25| 0.25| 0.25 | 0.25 | 0.64| 0.66| 0.67| 0.67|
|        | HIM       | 0.91   | 0.9          | 0.9 | 0.9 | 0.9 | 0.36| 0.35| 0.32| 0.3 | 0.28| 0.98| 0.98| 0.98| 0.98| 0.98 | 0.98 | 0.98| 0.98| 0.98|
| TPR    | LASSO     | 0.92   | 0.92         | 0.92| 0.92| 0.37| 0.42| 0.43| 0.43| 0.42| 0.98| 0.98| 0.98| 0.98| 0.98| 0.98 | 0.98| 0.98| 0.98|
|        | asymHIM5  | 0.9    | 0.89         | 0.89| 0.89| 0.88| 0.2 | 0.16| 0.15| 0.15| 0.13| 0.98| 0.98| 0.98| 0.98| 0.98 | 0.98 | 0.98| 0.98| 0.98|
|        | asymHIM3  | 0.91   | 0.9          | 0.89| 0.89| 0.89| 0.26| 0.21| 0.2 | 0.19| 0.18| 0.99| 0.98| 0.98| 0.98| 0.98 | 0.98 | 0.98| 0.98| 0.98|
|        | HIM       | 0.91   | 0.91         | 0.9 | 0.9 | 0.9 | 0.36| 0.35| 0.32| 0.3 | 0.28| 0.98| 0.98| 0.98| 0.98| 0.98 | 0.98 | 0.98| 0.98| 0.98|
| FPR    | LASSO     | 0.02   | 0.02         | 0.02| 0.02| 0.02| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02 | 0.02 | 0.02| 0.02| 0.02|
|        | asymHIM5  | 0.02   | 0.02         | 0.02| 0.02| 0.02| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02 | 0.02 | 0.02| 0.02| 0.02|
|        | asymHIM3  | 0.02   | 0.02         | 0.02| 0.02| 0.02| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02 | 0.02 | 0.02| 0.02| 0.02|
|        | HIM       | 0.02   | 0.01         | 0.01| 0.01| 0.01| 0.01| 0.01| 0.01| 0.01| 0.01| 0.02| 0.02| 0.02| 0.02| 0.02 | 0.02 | 0.02| 0.02| 0.02|
observations. Notice that Model III was also simulated by Zhao et al. (2013), but they did not report the $FPR_{inf}$ rate of their HIM measure.

Under this third model, the poor performance of both methods can be explained by the strong contamination of the data (response and predictors) and by the violation of the underlying asymptotic assumption. The contamination in Model III occurs simultaneously in the response and the predictors, and their effects are multiplicative since the influence measure is based on the marginal correlation of the response and the predictors. Additionally, the influence measure of each subject $k \in \{1, \ldots, n\}$ is affected by the contribution of the influential observations, even though the theory assumes that there are no influential observations in the dataset.

Hence, when there are influential observations in the dataset, and their contributions are very large—which is the case when the effect is multiplicative - then influence measures derived from the good observations are swamped. To illustrate the problem, consider a fixed $\tau$, and a single predictor ($p = 1$), then the influence measure of the $k$-th observation is:

$$n^2 D_{ck} = n^2 \left[ n^{-1} \sum_{i=1}^{n} x_iy_i - (n-1)^{-1} \sum_{i=1, i\neq k}^{n} x_iy_i \right]^2$$

$$= \left[ (n-1)^{-1} \sum_{i=1}^{n} x_iy_i - x_ky_k \right]^2$$

$$= \left[ (n-1)^{-1} \sum_{i \in S_{inf}} x_iy_i + (n-1)^{-1} \sum_{i \in S_{inf}} x_iy_i - x_ky_k \right]^2.$$ 

Then, the $k$-th observation is identified as a good observation when the contribution of the influential observations are bounded, that is:

$$(n-1)^{-1} \sum_{i \in S_{inf}} x_iy_i \leq \left( \frac{x_k^2 - \frac{(1)}{n} \left( 1 - x \right)^{1/2}}{n} \right) + x_ky_k - (n-1)^{-1} \sum_{i \in S_{inf}} x_iy_i,$$
where $\chi^2_{1-\alpha}(1)$ is the $100(1-\alpha)\%$ quantile of a $\chi^2(1)$ distribution. However, the influential observations, by definition, have high correlations and this upper bound is difficult to verify.

In summary, with the above equation, we can see how good observations can be swamped by the presence of at least one single influential observation, and this swamping effect is exacerbated when the contamination effect is multiplicative, i.e., response and predictor are contaminated simultaneously. A similar problem arises in classical statistics (Bendre 1989; Nurunnabi, Hadi, and Imon 2014; Roberts, Martin, and Zheng 2015), and many solutions exist in this framework. Some of those solutions have been extended to a high dimension regression setting (Zhao et al. 2019; Wang et al. 2018; Wang et al. 2019), and their adaptation to the asymHIM measure may help mitigate the swamping effect in general.

4. Real data application

4.1. Materials and methods

Structural and functional neuroimaging datasets are high dimensional, and vertex-wise measurements are strongly correlated, and are often contaminated due to acquisition, and preprocessing artifacts (Fritsch et al. 2012). Indeed, the brain is a spatially embedded system, so there is a huge amount of spatial autocorrelation. This means that an influential observation will likely be extreme on many features. We applied our asymmetric high dimensional influence measure (asymHIM3) to the Autism Brain Imaging Data Exchange (ABIDE) neuroimaging dataset, and conducted a downstream analysis to predict brain maturity based on cortical thickness. Brain maturity prediction is a well studied subject in neuroscience to establish a baseline for normal brain development against which neurodevelopmental disorders can be assessed (Khundrakpam, Tohka, and Evans 2015).

4.2. Participants

The ABIDE dataset comprises 573 control and 539 autism spectrum disorder (ASD) individuals from 16 international sites (Di Martino et al. 2014). The neuroimaging data of these individuals were obtained from DataLad repository (http://fcon_1000.projects.nitrc.org/indi/abide/). Only a subset of individuals is used in this work due to failures resulting from the image processing pipeline. The demographic description of these individuals is provided in Table 4. In this paper, we used 542 controls from the ABIDE study to illustrate the asymHIM influence measure.

4.3. MR image processing and cortical thickness measurements

The MR images (T1-weighted scans) were processed using the FreeSurfer 6.0 pipeline (Fischl 2012) deployed on CBrain - a high-performance computing facility (Sherif et al. 2014). FreeSurfer delineates the cortical surface from a given MR scan and quantifies thickness measurements on this surface for each brain hemisphere. The cortical thickness measurements for each MR image are computed using FreeSurfer 6.0 pipeline
The pipeline consists of 1) affine registration to the MNI305 space (Collins et al. 1994); 2) bias field correction; 3) removal of skull, cerebellum, and brainstem regions from the MR image; 3) estimation of white matter surface based on MR image intensity gradients between the white and gray matter; and 4) estimation of pial surface based on intensity gradients between the gray matter and cerebrospinal fluid (CSF). The distance between the white and pial surfaces provides the thickness estimate at a given location of cortex. For detailed descriptions refer to (Fischl 2012; Dale, Fischl, and Sereno 1999). The individual cortical surfaces are then projected onto a common space (i.e., fsaverage) characterized by 163,842 vertices per hemisphere to establish inter-individual correspondence.

4.4. Data cleaning and age prediction

We applied Zhao et al. (2013) influence measure (HIM), and our influence measure (asymHIM) to identify, and remove influential observations from the ABIDE dataset, using a 5% Bonferroni test correction. Our influential measure identified 48 subjects as influential, while the HIM measure identified 21 subjects, a subset of our 48 subjects (see Figure 5). Then, we fit an elastic net model, with the R package glmnet (Friedman, Hastie, and Tibshirani 2010), to the three resulting datasets: Full, HIM-cleaned, and asymHIM-cleaned, to predict age based on cortical thickness measurements. After cleaning, the Full dataset contained 542 subjects, the HIM-cleaned dataset had 521 subjects, and the asymHIM-cleaned dataset contained 494 subjects. In each of the 3 datasets, the age response is predicted with 299,569 cortical vertex features, 149,778 cortical vertices from the right hemisphere, and 149,791 cortical vertices from the left hemisphere. Following Khundrakpam, Tohka, and Evans (2015), we estimated the elastic net model using the hyper-parameter \( \theta = 0.5 \) balancing the \( L_1 \) and \( L_2 \) penalization. To simultaneously select the best regularization parameter and evaluate the prediction accuracy, we applied a 10-fold nested cross-validation (CV) loops (Ambroise and McLachlan 2002). In the inner 10-fold CV loop, we selected the best regularization parameter, and then age prediction accuracy was evaluated in the outer 10-fold CV loop. To measure goodness of fit, we used the mean absolute error (MAE) between the chronological and estimated age in the inner loop, and the correlation coefficient between the chronological and estimated age in the outer loop. Finally, since every 10-fold CV is random, we repeated the analysis 100 times and report the distribution of results as violin plots.

4.5. Results

We summarize the application results in term of sparsity, goodness of fit and prediction accuracy (Figure 6). Despite the small differences, the cleaned dataset based models achieve better sparsity with 355 and 344 cortical vertices selected on average when using the asymHIM and the HIM measures, respectively. Without cleaning the dataset, the

| Group  | Subjects (N) | Age (std) | Sex (M/F) |
|--------|--------------|-----------|-----------|
| Controls | 542          | 16.88 (7.7)| 444 / 98  |
| ASD    | 508          | 17.24 (8.48)| 446 / 62  |
The model selects on average 384 cortical vertices. The cleaned dataset based models also show better goodness of fit criteria with on average 3.37 and 3.49 years of MAE with the asymHIM and the HIM measure, respectively, while the full dataset based model has 3.54 years of MAE on average (Figure 6).

In terms of estimation of the chronological age, the full dataset based model has higher correlation than the cleaned dataset based models, with 0.78, 0.75 and 0.75 average correlation, respectively, for the full dataset based model and for cleaned dataset based models. However, the full dataset based model has also higher residual variance (12.65 on average) than the cleaned dataset based models (10.65 and 11.62, respectively for the asymHIM and HIM measure). The lower estimated correlation displayed by the cleaned dataset based models might be related to the removed influential observations. Indeed, the two influence measures (asymHIM and HIM) are correlation based, and observations with high marginal correlation have higher probability to be flagged as influential and removed from the dataset. Similarly, the high variance displayed by the full dataset based model can be related to the presence of the influential observations.

5. Conclusion

In this paper, we introduced the asymmetric influence measure (asymHIM) to identify influential observations in high dimensional regression. We derived its asymptotic distribution, which yields a statistical decision rule for identifying influential observations. The availability of an asymptotic distribution avoids needing to use subjective methods such as visualization, or computationally demanding methods such as the bootstrap, to
define appropriate thresholds for influential observations. The existence of a statistical threshold facilitates systematic application of the influence measure during quality control and preprocessing step, which will contribute to the improvement of reproducibility.

Figure 6. Elasticnet results for the ABIDE dataset. Violin plots show number of selected cortical vertices (top left), mean absolute error (MAE, top right), correlation between chronical and predicted age (bottom left), and residual variance between chronical and predicted age (bottom right) across 100 repeated cross-validations.
The simulations showed that the asymHIM measure outperforms the HIM measure of Zhao et al. (2013) in term of $\frac{TPR_{\inf}}{FPR_{\inf}}$ ratio. We explored several sequences of asymmetric points with different ranges. We noticed that the asymHIM measure did not perform as well in the presence of extreme expectiles which have a higher probability of being influential. The exploration of several sequences of expectiles demonstrated that the asymHIM measure with three points inside the interquartile range had the best performance, and use of asymmetric points (expectiles) outside the interquartile range gave results that were overly prone to influential observations.

We used both asymHIM and HIM measures to detect influential observations present among the controls in the ABIDE dataset and then to predict brain maturity. The prediction model derived from the clean datasets displayed more sparsity, better mean absolute error between chronological and predicted age, and smaller residual variance, than the prediction model derived from the full dataset. We noticed that the correlation decreased for the cleaned datasets relative to the full dataset based model; such patterns are often observed even in simple models where elimination of an influential point may lead to an attenuated slope estimate. This in itself is not a bad thing - as we can argue this prevents us from “overestimating” the association between cortical thickness and age.

In simulations, we noted poor performance of both the asymHIM and HIM measures when the response and the predictors are simultaneously contaminated. This poor performance can be explained by the presence of influential observations which violate the asymptotic assumption, and the fact that the simultaneous contamination amplifies the swamping effect multiplicatively. We are currently exploring different alternatives such as random group deletions to mitigate the dual phenomenon of the swamping and masking effects. We have promising preliminary results which will be presented in the near future.

Finally, it will be interesting to extend the proposed work in this article to other high dimensional complex models, such as generalized linear models or models containing complex dependence structures to evaluate the benefits associated with identification of influential points in those situations.

**Acknowledgements**

I would like to thank Bowei Xiao, Tianyuan Lu, Ting Zhang, Yixiao Zeng, Kaiqiong Zhao, Kevin McGregor, Antonio Ciampi, Lai Jiang, and Kathleen Klein from Celia Greenwood’s Lab. at the Lady Davis Institute for their comments, discussions and for reviewing the first version of the manuscript. I would also like to thank Dr. Brent Richards and Dr. Claudia Kleinman at the Lady Davis Institute, and the members of their Lab. for their comments and discussions on the preliminary results of the manuscript at the journal club.

**Funding**

This work was funded by the Healthy Brains for Healthy Lives Discovery grant to Drs. Bratislav Misic, Jean-Baptiste Poline, Celia M. T. Greenwood & Alan C. Evans.

**ORCID**

Amadou Barry [http://orcid.org/0000-0003-3396-1724](http://orcid.org/0000-0003-3396-1724)
References

Ambroise, C., and G. J. McLachlan. 2002. Selection bias in gene extraction on the basis of micro-array gene-expression data. Proceedings of the National Academy of Sciences of the United States of America 99 (10):6562–66. doi:10.1073/pnas.102102699.

Bendre, S. M. 1989. Masking and swamping effects on tests for multiple outliers in normal sample. Communications in Statistics - Theory and Methods 18 (2):697–710. doi:10.1080/03610928908829928.

Chatterjee, S., and A. S. Hadi. 1986. Influential observations, high leverage points, and outliers in linear regression. Statistical Science 1 (3):379–93. doi:10.1214/ss/1177013622.

Collins, D. L., P. Neelin, T. M. Peters, and A. C. Evans. 1994. Automatic 3d intersubject registration of MR volumetric data in standardized talairach space. Journal of Computer Assisted Tomography 18 (2):192–205.

Cook, R. D. 1977. Detection of influential observation in linear regression. Technometrics 19 (1):15–18.

Cook, R. D., and S. Weisberg. 1982. Residuals and Influence in Regression. New York: Chapman & Hall. ISBN 978-0-412-24280-9.

Dale, A. M., B. Fischl, and M. I. Sereno. 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage 9 (2):179–94. doi:10.1006/nimg.1998.0395.

De Rossi, G., and A. Harvey. 2009. Quantiles, expectiles and splines. Journal of Econometrics 152 (2):179–85. doi:10.1016/j.jeconom.2009.01.001.

Di Martino, A., C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al. 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry 19 (6):659–67. doi:10.1038/mp.2013.78.

Fan, J., and J. Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (5):849–911. doi:10.1111/j.1467-9868.2008.00674.x.

Fischl, B. 2012. FreeSurfer. NeuroImage 62 (2):774–81. doi:10.1016/j.neuroimage.2012.01.021.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1):1–22. http://www.jstatsoft.org/v33/i01/.

Fritsch, V., G. Varoquaux, B. Thyreau, J.-B. Poline, and B. Thirion. 2012. Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Medical Image Analysis 16 (7):1359–70. doi:10.1016/j.media.2012.05.002.

Jeong, J., and C. Kim. 2018. Effect of outliers on the variable selection by the regularized regression. Communications for Statistical Applications and Methods 25 (2):373–43. doi:10.29220/CSAM.2018.25.2.235.

Jiang, C., M. Jiang, Q. Xu, and X. Huang. 2017. Expectile regression neural network model with applications. Neurocomputing 247:73–86. doi:10.1016/j.neucom.2017.03.040.

Khundrakpam, B. S., J. Tohka, and A. C. Evans, Brain Development Cooperative Group. 2015. Prediction of brain maturity based on cortical thickness at different spatial resolutions. NeuroImage 111:350–59. doi:10.1016/j.neuroimage.2015.02.046.

Kim, M., and S. Lee. 2016. Nonlinear expectile regression with application to value-at-risk and expected shortfall estimation. Computational Statistics and Data Analysis 94:1–19. doi:10.1016/j.csda.2015.07.011.

Majumdar, A., and D. Paul. 2016. Zero expectile processes and bayesian spatial regression. Journal of Computational and Graphical Statistics 25 (3):727–47. doi:10.1080/10618600.2015.1062014.
Mersmann, O. 2018. *Microbenchmark: accurate timing functions*. R package version 1.4-6. https://CRAN.R-project.org/package=microbenchmark.

Newey, W. K., and J. L. Powell. 1987. Asymmetric least squares estimation and testing. *Econometrica* 55 (4):819–47.

Nurunnabi, A. A. M., A. S. Hadi, and A. H. M. R. Imon. 2014. Procedures for the identification of multiple influential observations in linear regression. *Journal of Applied Statistics* 41 (6):1315–31. doi:10.1080/02664763.2013.868418.

R Core Team. 2018. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Righi, M. B., Y. Yang, and P. S. Ceretta. 2014. Nonparametric expectile regression for conditional autoregressive expected shortfall estimation. *Contemporary Studies in Economic and Financial Analysis* 96:83–95. doi:10.1108/1569-37592014000096003.

Ro, K., C. Zou, Z. Wang, and G. Yin. 2015. Outlier detection for high-dimensional data. *Biometrika* 102 (3):589–99. doi:10.1093/biomet/asv021.

Roberts, S., M. A. Martin, and L. Zheng. 2015. An adaptive, automatic multiple-case deletion technique for detecting influence in regression. *Technometrics* 57 (3):408–17. doi:10.1080/00401706.2014.946152.

Schnabel, S. K., and P. H. C. Eilers. 2009. Optimal expectile smoothing. *Computational Statistics and Data Analysis* 53 (12):4168–77. doi:10.1016/j.csda.2009.05.002.

Schulze Waltrup, L., and G. Kauermann. 2017. Smooth expectiles for panel data using penalized splines. *Statistics and Computing* 27 (1):271–82. doi:10.1007/s11222-015-9621-2.

She, Y., and A. B. Owen. 2010. Outlier detection using nonconvex penalized regression. http://arxiv.org/abs/1006.2592.

Sherif, T., P. Rioux, M.-E. Rousseau, N. Kassis, N. Beck, R. Adalat, S. Das, T. Glatard, and A. C. Evans. 2014. CBRAIN: A web-based, distributed computing platform for collaborative neuroimaging research. *Frontiers in Neuroinformatics* 8, 54. doi:10.3389/fninf.2014.00054.

Sobotka, F., and T. Kneib. 2012. Geoadditive expectile regression. *Computational Statistics and Data Analysis* 56 (4):755–67. doi:10.1016/j.csda.2010.11.015.

Sobotka, F., G. Kauermann, L. Schulze Waltrup, and T. Kneib. 2013. On confidence intervals for semiparametric expectile regression. *Statistics and Computing* 23 (2):135–48. March doi:10.1007/s11222-011-9297-1.

Sobotka, F., S. Schnabel, L. Schulze Waltrup, P. Eilers, T. Kneib, and G. Kauermann. 2014. *Expectreg: Expectile and Quantile Regression*. R package version 0.39.

Szekely, G. J., M. L. Rizzo, and N. K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35 (6):2769–94. doi:10.1214/009053607000000505.

Waldmann, E., F. Sobotka, and T. Kneib. 2017. Bayesian regularisation in geoadditive expectile regression. *Statistics and Computing* 27 (6):1539–53. doi:10.1007/s11222-016-9703-9.

Waltrup, L. S., F. Sobotka, T. Kneib, and G. Kauermann. 2015. Expectile and quantile regression—David and Goliath? *Statistical Modelling: An International Journal* 15 (5):433–56. doi:10.1177/1471082X14561155.

Wang, T., and Z. Li. 2017. Outlier detection in high-dimensional regression model. *Communications in Statistics - Theory and Methods* 46 (14):6947–58. doi:10.1080/03610926.2016.1140783.

Wang, T., Q. Li, B. Chen, and Z. Li. 2018. Multiple outliers detection in sparse high-dimensional regression. *Journal of Statistical Computation and Simulation* 88 (1):89–107. doi:10.1080/00949655.2017.1379521.

Wang, T., Q. Li, Q. Zang, and Z. Li. 2019. A multiple-case deletion approach for detecting influential points in high-dimensional regression. *Communications in Statistics - Simulation and Computation* 48 (7):2065–82. doi:10.1080/03610918.2018.1433840.

Xing, J.-J., and X.-Y. Qian. 2017. Bayesian expectile regression with asymmetric normal distribution. *Communications in Statistics - Theory and Methods* 46 (9):4545–55. doi:10.1080/03610926.2015.1088030.
Xu, Q., X. Liu, C. Jiang, and K. Yu. 2016. Nonparametric conditional autoregressive expectile model via neural network with applications to estimating financial risk. *Applied Stochastic Models in Business and Industry* 32 (6):882–908. doi:10.1002/asmb.2212.

Yang, Y., and H. Zou. 2015. Nonparametric multiple expectile regression via er-boost. *Journal of Statistical Computation and Simulation* 85 (7):1442–58. doi:10.1080/00949655.2013.876024.

Zhao, J., C. Leng, L. Li, and H. Wang. 2013. High-dimensional influence measure. *The Annals of Statistics* 41 (5):2639–67. doi:10.1214/13-AOS1165.

Zhao, J., C., Liu, L. Niu, and C. Leng. 2019. Multiple influential point detection in high dimensional regression spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81 (2):385–408. doi:10.1111/rssb.12311.