IRTG 1792 Discussion Paper 2018-060

RESIDUAL'S INFLUENCE INDEX (RINFIN), BAD LEVERAGE AND UNMASKING IN HIGH DIMENSIONAL L2-REGRESSION

Yannis G. Yatracos*

* Cyprus University of Technology, Cyprus

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

<u>R</u>ESIDUAL'S <u>INF</u>LUENCE <u>IN</u>DEX (*RINFIN*), BAD LEVERAGE AND UNMASKING IN HIGH DIMENSIONAL L_2-REGRESSION

Yannis G. Yatracos

Faculty of Communication and Media Studies

Cyprus University of Technology

October 29, 2018

Summary

In linear regression of Y on $\mathbf{X}(\in R^p)$ with parameters $\beta(\in R^{p+1})$, statistical inference is unreliable when observations are obtained from gross-error model, $F_{\epsilon,G} = (1-\epsilon)F + \epsilon G$, instead of the assumed probability $F; G$ is gross-error probability, $0 < \epsilon < 1$. When G is unit mass at $(\mathbf{x}, y)$, *Residual's Influence Index*, *RINFIN*$(\mathbf{x}, y; \epsilon, \beta)$, measures the difference in small $\mathbf{x}$-perturbations of L_2-residual, $r(\mathbf{x}, y)$, for model F and for $F_{\epsilon,G}$ via r's $\mathbf{x}$-partial derivatives. Asymptotic properties are presented for sample *RINFIN* that is successful in extracting *indications* for influential and bad leverage cases in microarray data and simulated, high dimensional data. Its performance improves as p increases and can also be used in multiple response linear regression. *RINFIN*'s advantage is that, whereas in influence functions of L_2-regression coefficients each $\mathbf{x}$-coordinate and $r(\mathbf{x}, y)$ appear in a sum as product with moderate size when $(\mathbf{x}, y)$ is bad leverage case and masking makes $r(\mathbf{x}, y)$ nearly vanish, *RINFIN*'s $\mathbf{x}$-partial derivatives convert the product in sum allowing for unmasking.

Some key words: Big Data, Data Science, Influence Function, Leverage, Masking, Residual's Influence Index (*RINFIN*)

AMS 2010 subject classifications: 62-07, 62-09, 62J05, 62F35, 62G35

1 Introduction

Tukey (1962, p.60) wrote: "Procedures of *diagnosis*, and procedures to extract *indications* rather than extract conclusions, will have to play a large part in the future of data analyses and graphical techniques offer great possibilities in both areas." This philosophy is widely adopted nowadays in Data Science and motivates this work.

Cleaning high dimensional data is a crucial step before the statistical analysis. In linear regression of Y on $\mathbf{X}$ and parameters β, it is often erroneously assumed that the data follows probability F instead of the gross-error model $F_{\epsilon,G} = (1-\epsilon)F + \epsilon G$ (Huber, 1964); G is gross-error probability, $0 < \epsilon < 1, Y \in R, \mathbf{X} \in R^p,\ \beta \in R^{p+1}$. One or more cases from G may influence the analysis and their identification and removal will improve statistical inference for the F-population. When $\mathbf{x}$ is far away from the bulk of F's factor space, $(\mathbf{x}, y)$ is called *leverage case* (Rousseeuw and Leroy, 1987). A "*bad*" leverage case from G forces the regression hyperplane determined by F (the F-regression) and *the associated F-residuals* to change drastically when $\mathbf{x}$ becomes more remote. The goal of this work is to provide *a simple and easy to implement procedure* extracting indications for influential/bad leverage cases (from G) in least squares (L_2) regression.

The empirical influence function of a non-robustified estimator suffers from the *masking* effect. For example, in simple, linear L_2-regression with sample $(x_1, y_1), \ldots, (x_n, y_n)$, the influence function of the slope at (x, y) is $C \cdot r \cdot (x - \bar{x}_n)$; r is the residual of $(x, y), C$ is independent of x, y. If x is bad leverage *and* is masked due to few neighboring values in the sample, the difference $(x - \bar{x}_n)$ will have large absolute value whereas r may be near zero due to masking and the absolute value of the influence, $|r \cdot (x - \bar{x}_n)|$, may be moderate. To the contrary, the x-derivative of the influence function measures *local* influence and separates the factors of the influence function obtaining instead the sum of $C \cdot \hat{\beta}(x - \bar{x}_n)$ and $C \cdot r$ which has large absolute value even when x is masked and r is near 0; $\hat{\beta}$ is the L_2-estimate of the slope. The influence index introduced herein inherits this advantage in multiple, linear L_2-regression being the sum of influence functions and their derivatives. This holds also for L_2-regression with diagonal matrix, W, of weights independent of $\mathbf{x}, r$.

Changes in regression residuals for small $\mathbf{x}$-perturbations under models F and $F_{\epsilon,\mathbf{x},y}$

where the derivative of the influence function appears naturally, are used to detect leverage cases; $F_{\epsilon,\mathbf{x},y}$ is gross-error model with G unit mass at $(\mathbf{x}, y)$. L_2-*Residual's Influence Index*, $RINFIN(\mathbf{x}, y; \epsilon, \beta)$, is the sum of squared differences for the $\mathbf{x}$-partial derivatives of the F-residual and the $F_{\epsilon,\mathbf{x},y}$-residual at $(\mathbf{x}, y)$. For gross-error model with a *group* of remote $\mathbf{x}$-*neighboring* cases $(\mathbf{x}, y)$ drawn with probability ϵ and group average $(\bar{\mathbf{x}}, \bar{y})$, $RINFIN(\bar{\mathbf{x}}, \bar{y}; \epsilon, \beta)$ measures the group's influence avoiding masking of influential cases from the group's members and with ϵ its factor. Asymptotic properties of $RINFIN(\mathbf{x}, y; \epsilon, \hat{\beta}_n)$ are presented; n is the sample size, $\hat{\beta}_n$ is β's L_2-estimate.

Our goal is to look for indications of leverage cases from G in $F_{\epsilon,G}$. Every case $(\mathbf{x}, y)$ in the sample is used to calculate its sample *RINFIN*-value. Since the percentage of G-observations in $F_{\epsilon,\mathbf{x},y}$ is expected to be 10% or less, potential bad leverage cases in the sample are those $(\mathbf{x}, y)$ with the 10% larger sample $RINFIN(\mathbf{x}, y; 1/n, \hat{\beta}_n)$ values and especially those with the same ordering when the squared differences in the *RINFIN* sum are replaced by absolute values obtaining *RINFINABS* values. $RINFIN(\mathbf{x}, y; 1/n, \hat{\beta}_n)$ is successful with the microarray data used in Zhao *et al.* (2016, 2013) for which $n = 120$ and $p = 1500$. In simulations with gross-error normal mixtures F, G and fixed sample size n, the misclassification proportion of G-cases using $RINFIN(\mathbf{x}, y; 1/n, \hat{\beta}_n)$ decreases to zero as p increases, $p < n$. The *blessing of high dimensionality* is due to the "separation" of the mixtures'components measured, *e.g.*, by their Hellinger's distance, as p increases (Yatracos 2017, 2013, Section 8, Proposition 8.1). When n is smaller than p, sample *RINFIN* is calculated sequentially, for the y-response on subvectors of $\mathbf{x}$-covariates with dimension $q < n$. For each case, the total of its $\frac{p}{q}$ sample *RINFIN* values is its total residual influence index. *RINFIN* can also be used with multiple response linear regression, adding for $(\mathbf{x}, y)$ the sample *RINFIN*-values for each response.

With the recent flood of Big Data, there is need in regression problems for new influence measures in outlier detection. She and Owen (2011) have as goals outlier identification and robust coefficient estimation, both achieved using a nonconvex sparsity criterion. Zhao *et al.* (2013) propose a high dimensional influence measure (*HIM*) based on marginal correlations between the response and the individual covariates and the leave-one-out observation idea (Weisberg, 1985). Zhao *et al.* (2016) propose a novel procedure, for *multiple* influential point detection (*MIP*).

In Genton and Ruiz-Gazen (2010), an *observation* is influential "whenever a change in its value leads to a radical change in the estimate" and the *hair-plot* is used for *visual identification.* Local and global influence measures are proposed using partial derivative of the *estimate.* For a *particular* regression model, Flores (2015) introduced *leverage constants* to determine bad leverage cases.

The Influence Function has been used in outlier detection by Campbell (1978) and Boente *et al.* (2002). Rousseeuw and van Zomeren (1990) used standardized Least Trimmed Squares residuals against robust distance to classify observations in regression. Hubert, Rousseeuw and Van Aelst (2008) present a survey of High Breakdown Robust methods to detect outlying observations. The influence of observations in estimates' values has been also studied by several authors, among others by Cook (1977), Welsch and Kuh (1977), Belsley *et al.* (1980), Cook and Weisberg (1980), Ruppert and Carroll (1980), Huber (1981), Velleman and Welsch (1981), Welsch (1982), Hawkins *et al.* (1984), Carroll and Ruppert (1985), Hampel (1985), Hampel *et. al.* (1986), Ronchetti (1987), Hadi and Simonoff (1993) and Genton and Hall (2016).

Proofs follow in the Appendix where $\mathcal{E}$-matrix is introduced to obtain in simple form the influence functions of regression coefficients when the **X**-covariates are uncorrelated.

2 The Tools-The Derivative of the Influence Function

Hampel (1971) introduced the influence function, $IF(\mathbf{x};T,F)$, of a functional T with real values,

$$IF(\mathbf{x};T,F)=\lim_{\epsilon\to 0}\frac{T[(1-\epsilon)F+\epsilon\Delta_{\mathbf{x}}]-T(F)}{\epsilon}, \tag{1}$$

when this limit exists; $\mathbf{x}(\in R^p)$, F is a probability, $\Delta_{\mathbf{x}}$ is probability with all its mass at $\mathbf{x}$, $0<\epsilon<1$.

$IF(\mathbf{x};T,F)$ determines the "bias" in the value of T at F due to an ϵ-perturbation of F

with $\Delta_{\mathbf{x}}$:

$$T[(1-\epsilon)F+\epsilon\Delta_{\mathbf{x}}]-T(F)=\epsilon IF(\mathbf{x};T,F)+o(\epsilon)\approx\epsilon IF(\mathbf{x};T,F), \tag{2}$$

" $\approx''$ is used since

$$\lim_{\epsilon\to 0}\frac{T[(1-\epsilon)F+\epsilon\Delta_{\mathbf{x}}]-T(F)}{\epsilon IF(\mathbf{x};T,F)}=1.$$

Definition 2.1 *(Hampel, 1971) The Breakdown Point is the upper bound on ϵ for which linear approximation (2) can be used.*

Discussing further concepts related to the influence function, Hampel (1974, p. 389) introduced *local-shift-sensitivity*,

$$\lambda^* = \sup_{\mathbf{x}\neq\mathbf{y}}\frac{|IF(\mathbf{x};T,F)-IF(\mathbf{y};T,F)|}{||\mathbf{x}-\mathbf{y}||}, \tag{3}$$

as "a measure for the *worst* (approximate) effect of wiggling the observations"; $||\cdot||$ is a Euclidean distance in R^p.

Unlike the extensive use of Breakdown Point, local-shift-sensitivity was never fully exploited. One reason is that, in reality, it is a "global" measure as supremum over all $\mathbf{x},\mathbf{y}$. Thus, λ^* cannot be used to study T's bias for $\mathbf{x}$'s small perturbation in the ϵ-mixture, from $\mathbf{x}$ to $\mathbf{x}+\mathbf{h}$, $||\mathbf{h}||$ small,

$$T[(1-\epsilon)F+\epsilon\Delta_{\mathbf{x}+\mathbf{h}}]-T[(1-\epsilon)F+\epsilon\Delta_{\mathbf{x}}]. \tag{4}$$

When F is defined on the real line, (4) is evaluated at neighboring points $x,\ x+h,\ x\in R,\ h\in R, |h|$ small.

Lemma 2.1

$$\lim_{h\to 0}\lim_{\epsilon\to 0}\frac{T[(1-\epsilon)F+\epsilon\Delta_{x+h}]-T[(1-\epsilon)F+\epsilon\Delta_{x}]}{\epsilon h}=\frac{dIF(x;T,F)}{dx}=IF'(x;T,F), \tag{5}$$

when the limit exists.

Remark 2.1 *Under mild conditions, e.g. for any function g for which the derivative g' exists at x and for the functional*

$$T(F)=\int g(y)dF(y),$$

the limits in ϵ and h can be interchanged in (5) without affecting the limit.

$IF'(x;T,F)$ is used to approximate (4) for small $\epsilon,\ |h|$:

$$T[(1-\epsilon)F+\epsilon\Delta_{x+h}]-T[(1-\epsilon)F+\epsilon\Delta_x]\approx\epsilon h IF'(x;T,F); \tag{6}$$

(6) is *the Tool* used to approximate L_2 residuals of gross-error models and determine *RINFIN*. When (6) is used, the Influence Function's derivative is always evaluated at F.

Examples of IF' follow.

Example 2.1 Let F be a probability on the real line, $T(F)$ is the mean of F, its influence function is

$$IF(x;T,F)=x-T(F)$$

with derivative a constant.

Example 2.2 Consider a simple linear regression model, $Y=\beta_0+\beta_1X+e$, with error e having mean zero and finite second moment, F is the joint distribution of (X,Y).

The influence functions for the L_2 -parameters $\beta_0(F),\ \beta_1(F)$, obtained at F are

$$IF(x,y;\beta_0(F),F)=[y-\beta_0(F)-\beta_1(F)x]\frac{EX^2-xEX}{Var(X)}=r(x,y;F)\frac{EX^2-xEX}{Var(X)}, \tag{7}$$

$$IF(x,y;\beta_1(F),F)=[y-\beta_0(F)-\beta_1(F)x]\frac{x-EX}{Var(X)}=r(x,y;F)\frac{x-EX}{Var(X)}; \tag{8}$$

EU and $Var(U)$ denote, respectively, U's mean and variance. The x-derivatives of (7), (8) are

$$IF'_{x,0}=\frac{\partial IF(x,y;\beta_0(F),F)}{\partial x}=-\beta_1(F)\frac{EX^2-xEX}{Var(X)}-r(x,y;F)\frac{EX}{Var(X)}, \tag{9}$$

$$IF'_{x,1}=\frac{\partial IF(x,y;\beta_1(F),F)}{\partial x}=-\beta_1(F)\frac{x-EX}{Var(X)}-r(x,y;F)\frac{1}{Var(X)}. \tag{10}$$

Observe in (7), (8) the *multiplicative* effects of r with $(x-EX)$ and EX^2-xEX and their conversions to *additive* effects in (9), (10).

Remark 2.2 *The y-derivatives of L_2-influence functions (7), (8) are, respectively, $(EX^2-xEX)/Var(X)$ and $(x-EX)/Var(X)$. Thus, y-derivatives of influence functions do not provide information for $r(x,y;F)$ and their sample versions are maximized at the extreme x-values in the sample.*

3 Residuals, Influence, Leverage Cases and *RINFIN*

3.1 Least Squares Regression and Influence Functions

Let $(\mathbf{X}, Y)$ follow probability model F,

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e; \tag{11}$$

$\mathbf{X} = (X_1, ...X_p)^T$ is the covariates' vector, Y is the response, $\beta = (\beta_0, .., \beta_p)^T = (\beta_0(F), .., \beta_p(F))^T$.

The Model Assumptions:

($\mathcal{A}1$) The error, e, has mean zero and finite second moment.

($\mathcal{A}2$) Case $(\mathbf{x}, y)$ is mixed with cases from model F with probability ϵ (model $F_{\epsilon,\mathbf{x},y}$).

The L_2-regression coefficients β are obtained minimizing Ee^2; E denotes expected value. *RINFIN* has a simple form when an additional assumption is used:

($\mathcal{A}3$) $X_1, \ldots, X_p$ are uncorrelated random variables.

Notation

The j-th regression coefficient obtained by L_2-minimization at model $F_{\epsilon,\mathbf{u},v}$ is denoted by $\beta_j(F_{\epsilon,\mathbf{u},v}),\ j = 0, 1, ..., p$, and their vector by $\beta(F_{\epsilon,\mathbf{u},v})$.

Denote the L_2- residuals for model $F_{\epsilon,\mathbf{u},v}$ at $(\mathbf{x}, y)$ by

$$r(\mathbf{x}, y; F_{\epsilon,\mathbf{u},v}) = y - \beta_0(F_{\epsilon,\mathbf{u},v}) - \sum_{j=1}^{p} \beta_j(F_{\epsilon,\mathbf{u},v}) x_j; \tag{12}$$

r is also used to denote $r(\mathbf{x}, y; F)$.

For $|h|$ small, let

$$\mathbf{x}_{i,h} = \mathbf{x} + (0, \ldots, h, \ldots, 0), \tag{13}$$

such that $(\mathbf{x}_{i,h}, y), (\mathbf{x}, y + h)$ are small perturbations of $(\mathbf{x}, y)$ making it more extreme.

The influence function of β_j is evaluated at $(\mathbf{x}, y)$ for F, thus use

$$IF_j = IF(\mathbf{x}, y; \beta_j, F), \qquad IF'_{v,j} = \frac{\partial IF(\mathbf{x}, y; \beta_j, F)}{\partial v}, \qquad v = y, x_i,\ i = 1, \ldots, p, \tag{14}$$

i.e., in words, $IF'_{v,j}$ is the derivative of IF_j with respect to v, $j = 0, 1, \ldots, p$.

Influence functions of L_2 regression coefficients at F are solutions of the equations' system:

$$IF_0 + IF_1 EX_1 + ... + IF_p EX_p = r(\mathbf{x}, y; F), \tag{15}$$

$$IF_0 EX_i + ... + IF_j EX_i X_j + ... + IF_p EX_i X_p = x_i r(\mathbf{x}, y; F), \ i = 1, \ldots, p. \tag{16}$$

Equations (15) and (16) are obtained by interchanging in the normal equations,

$$\frac{\partial E_H (Y - \beta_0 - \beta_1 X_1 - \ldots - \beta_p X_p)^2}{\partial \beta_i} = 0, \ i = 0, 1, \ldots, p, \tag{17}$$

the expected value with the partial derivatives for $i = 0, 1, \ldots, p$. The obtained equations are evaluated at the models $H = F$ and $H = (1-\epsilon)F + \epsilon \Delta_{(\mathbf{x},y)}$, the equations for the i-th partial derivative for both models are subtracted, are divided by ϵ and when $\epsilon \to 0$ the Influence Functions appear in the left side of equations (15) and (16) and in the right side are the remaining terms.

The influence functions in (15) and (16) are now provided *in closed form* when, in addition, ($\mathcal{A}3$) holds. With an additional assumption on the error, e, influence functions of L_1-regression coefficients have also been obtained (Yatracos, 2018, Proposition 3.2).

Proposition 3.1 *For regression model (11) with assumptions ($\mathcal{A}1$)-($\mathcal{A}3$) and notation (14), the influence functions of L_2-regression coefficients at $(\mathbf{x}, y)$ for model F are:*

$$IF_0 = r[1 - p + \sum_{j=1}^{p} \frac{EX_j^2 - x_j EX_j}{\sigma_j^2}], \qquad IF_j = r\frac{x_j - EX_j}{\sigma_j^2}, \qquad j = 1, \ldots, p; \tag{18}$$

$r = r(\mathbf{x}, y; F)$, σ_j^2 *is the variance of* X_j, $j = 1, \ldots, p$.

3.2 Perturbations of L_2-Residuals for models F and $F_{\epsilon,\mathbf{x},y}$

The goal is to compare small $(\mathbf{x}, y)$-residual changes in L_2 regressions for $F_{\epsilon,\mathbf{x},y}$ and F :
i) when $(\mathbf{x}_{i,h}, y)$ replaces $(\mathbf{x}, y)$ in the ϵ-mixture, *i.e.*, under $F_{\epsilon,\mathbf{x},y}$ and $F_{\epsilon,\mathbf{x}_{i,h},y}$:
$r(\mathbf{x}_{i,h}, y; F_{\epsilon,\mathbf{x}_{i,h},y}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y})$ (in (20)) and
ii) when $(\mathbf{x}, y+h)$ replaces $(\mathbf{x}, y)$ in the ϵ-mixture, *i.e.*, under $F_{\epsilon,\mathbf{x},y}$ and $F_{\epsilon,\mathbf{x},y+h}$:
$r(\mathbf{x}, y+h; F_{\epsilon,\mathbf{x},y+h}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y})$ (in (24)).

A Lemma used repeatedly to calculate residuals'differences *i)*, *ii)* follows.

Lemma 3.1 *For regression model (11) with assumptions ($\mathcal{A}1$), ($\mathcal{A}2$) and $\epsilon, |h|$ both small it holds for $0 \leq j \leq p:$*

$$\beta_j(F_{\epsilon,\mathbf{x},y}) \approx \beta_j(F) + \epsilon IF_j, \qquad \beta_j(F_{\epsilon,\mathbf{x}_{i,h},y}) \approx \beta_j(F_{\epsilon,\mathbf{x},y}) + \epsilon h \frac{\partial IF(\mathbf{x}, y; \beta_j, F)}{\partial x_i}. \tag{19}$$

Proposition 3.2 *For regression model (11) with ($\mathcal{A}1$), ($\mathcal{A}2$), $\mathbf{x}_{i,h}$ the perturbation of $\mathbf{x}$ (see 13) and for ϵ and $|h|$ both small:*

a) the difference of $(\mathbf{x}, y)$-residuals at $F_{\epsilon,\mathbf{x}_{i,h},y}$ and $F_{\epsilon,\mathbf{x},y}$ is:

$$r(\mathbf{x}_{i,h}, y; F_{\epsilon,\mathbf{x}_{i,h},y}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y}) + \beta_i h \approx -\epsilon h[IF_i + \frac{\partial IF_0}{\partial x_i} + \sum_{j=1}^{p} x_j \frac{\partial IF_j}{\partial x_i}], \ i = 1, \ldots, p, \tag{20}$$

b) the difference of $(\mathbf{x}, y)$-residuals at $F_{\epsilon,\mathbf{x},y+h}$ and $F_{\epsilon,\mathbf{x},y}$ is:

$$r(\mathbf{x}, y + h; F_{\epsilon,\mathbf{x},y+h}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y}) - h \approx -\epsilon h[\frac{\partial IF_0}{\partial y} + \sum_{j=1}^{p} x_j \frac{\partial IF_j}{\partial y}]. \tag{21}$$

Remark 3.1 *The right side in (20) involves influence functions and their derivatives. An index using it to detect bad leverage is less affected by masking than diagnostics based solely on Influence Functions, as explained in the Introduction.*

The right sides of (20) and (21) are obtained below for uncorrelated covariates.

Corollary 3.1 *Under the assumptions of Proposition 3.2 and ($\mathcal{A}3$), with $r = r(\mathbf{x}, y; F):$*

a_1)

$$r(\mathbf{x}_{i,h}, y; F_{\epsilon,\mathbf{x}_{i,h},y}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y}) + \beta_i h \approx -\epsilon h\{2\frac{r(x_i - EX_i)}{\sigma_i^2} - \beta_i[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}]\}. \tag{22}$$

a_2) If, in addition, $|x_i|$ is large,

$$r(\mathbf{x}_{i,h}, y; F_{\epsilon,\mathbf{x}_{i,h},y}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y}) \approx \epsilon h \cdot 3\beta_i \frac{(x_i - EX_i)^2}{\sigma_i^2}, \tag{23}$$

b)

$$r(\mathbf{x}, y + h; F_{\epsilon,\mathbf{x},y+h}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y}) - h \approx -\epsilon h[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}]. \tag{24}$$

3.3 x-Influence and Residual's Influence Index *RINFIN*$(\mathbf{x}, y; \epsilon, \beta)$

Influence is determined using the distance of residuals' partial derivatives at $(\mathbf{x}, y)$ for model F and gross-error model $F_{\epsilon,\mathbf{x},y}$. The larger the distance is, the larger the influence of $(\mathbf{x}, y)$ is.

$\mathbf{x}$-*Influence on* L_2-*Residuals*

For $(\mathbf{x}_{i,h}, y)$ and $(\mathbf{x}, y)$ both under model F,

$$\frac{r(\mathbf{x}_{i,h}, y; F) - r(\mathbf{x}, y; F)}{h} + \beta_i = 0, \qquad i = 1, \ldots, p. \tag{25}$$

For gross-error models $F_{\epsilon,\mathbf{x},y}$, $F_{\epsilon,\mathbf{x}_{i,h},y}$, the difference in partial derivatives of residuals is obtained from (20) for small ϵ,

$$\lim_{h \to 0} \frac{r(\mathbf{x}_{i,h}, y; F_{\epsilon,\mathbf{x}_{i,h},y}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y})}{h} + \beta_i \approx -\epsilon[IF_i + \frac{\partial IF_0}{\partial x_i} + \sum_{j=1}^{p} x_j \frac{\partial IF_j}{\partial x_i}], \quad i = 1, \ldots, p. \tag{26}$$

From (25) and (26), the right side of (26) measures influence of $\mathbf{x}$'s i-th coordinate in the residual's derivative and provides the motivation for defining influence.

Definition 3.1 *For gross-error model* $F_{\epsilon,\mathbf{x},y}$,
a) the influence of $\mathbf{x}$*'s i-th coordinate in the* L_2*-residual is*

$$INF(i) = \epsilon \cdot |IF_i + \frac{\partial IF_0}{\partial x_i} + \sum_{j=1}^{p} x_j \frac{\partial IF_j}{\partial x_i}|, \; i = 1, \ldots, p. \tag{27}$$

b) The L_2*-Residual Influence Index (RINFIN) is*

$$RINFIN(\mathbf{x}, y; \epsilon, \beta) = \epsilon \cdot \sum_{i=1}^{p} (IF_i + \frac{\partial IF_0}{\partial x_i} + \sum_{j=1}^{p} x_j \frac{\partial IF_j}{\partial x_i})^2 \tag{28}$$

Remark 3.2 *When in (28) the squares are replaced by absolute values, RINFINABS*$(\mathbf{x}, y; \epsilon, \beta)$ *is obtained. It can be used to confirm RINFIN's ordering as described in Section 4.2.*

Assuming in addition $(\mathcal{A}3)$, (28) becomes (using (45) in the Appendix):

$$RINFIN(\mathbf{x}, y; \epsilon, \beta) = \epsilon \cdot \sum_{i=1}^{p} \{2\frac{r(\mathbf{x}, y; F)(x_i - EX_i)}{\sigma_i^2} - \beta_i[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}]\}^2. \tag{29}$$

Remote $\mathbf{x}$'s have large *RINFIN*$(\mathbf{x}, y; \epsilon, \beta)$.

Proposition 3.3 *Under ($\mathcal{A}1$)-($\mathcal{A}3$), with G unit mass at $(\mathbf{x}, y)$, $\epsilon = 1/n$,*

$$\lim_{|x_i| \to \infty} RINFIN(\mathbf{x}, y; \epsilon, \beta) = \infty. \tag{30}$$

Remark 3.3 *For the sample version of Proposition 3.3, with G in reality discrete probability on bad leverage cases and with masking occuring, with $\epsilon = 1/n$ because of the way sample* RINFIN *is calculated, the lower bound used in the Proof will be roughly 1/9 of that without masking. This can be used to provide more indications about masking and bad leverage.*

y-Influence on L_2-Residuals

For $(\mathbf{x}, y+h)$ and $(\mathbf{x}, y)$ both under model F,

$$\frac{r(\mathbf{x}, y+h; F) - r(\mathbf{x}, y; F)}{h} = 1, \qquad i = 1, \ldots, p. \tag{31}$$

Proposition 3.4 *For models F, $F_{\epsilon,\mathbf{x},y}$, $F_{\epsilon,\mathbf{x},y+h}$, ϵ small and L_2 regression under $(\mathcal{A}1) - (\mathcal{A}3)$:*

$$\lim_{h \to 0} \frac{r(\mathbf{x}, y+h; F_{\epsilon,\mathbf{x},y+h}) - r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y})}{h} - 1 \approx -\epsilon[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}]. \tag{32}$$

Remark 3.4 *From (32), the y-influence index is*

$$\sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}; \tag{33}$$

it is maximized for cases in the extremes of the $\mathbf{x}$-coordinates. Thus, RINFIN is restricted to the influence of factor space cases.

3.4 Large Sample Properties of $RINFIN(\mathbf{x}, y; \epsilon, \hat{\beta}_n)$

The equations' system (15) and (16) can be written in matrix notation

$$\tilde{\mathcal{E}} \cdot \mathbf{IF} = \mathbf{q}(\mathbf{x}, y; \beta); \tag{34}$$

$\tilde{\mathcal{E}}$ is the symmetric matrix of EX_i, EX_iX_j and 1, $1 \le i,j \le p$, **IF** is the vector of β-influence functions and

$$\mathbf{q} = (r(\mathbf{x}, y; F), x_1 r(\mathbf{x}, y; F), \ldots, x_p r(\mathbf{x}, y; F))^T.$$

Consistency of $RINFIN(\mathbf{x}, y; \epsilon, \hat{\beta}_n)$ and its asymptotic distribution follow from the properties of the least squares estimates $\hat{\beta}_n$ of β. For the next proposition the notation is changed: $\mathbf{X}(\in R^{p+1})$ will have as first coordinate 1, $\tilde{\mathcal{E}}$ is $E\mathbf{X}\mathbf{X}^T$ however $\mathbf{x}(\in R^p)$ will still denote a factor space vector.

Proposition 3.5 *Let $(\mathbf{X}, Y), (\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)$ be independent, identically distributed random vectors with form $\mathbf{X}^T = (1, X_1, \ldots, X_p) \in R^{p+1}$, $Y \in R$,*

$$Y = \mathbf{X}^T \beta + \epsilon. \tag{35}$$

Let $\hat{\beta}_n$ be the least squares estimate of β.

a) Assume that i) Rank $\tilde{\mathcal{E}}$ = Rank $E\mathbf{X}\mathbf{X}^T = p+1$, ii) $E\mathbf{X}\epsilon = \mathbf{0}$, iii) $E\epsilon^2 < \infty$. Then, for every $(\mathbf{x}, y) \in R^{p+1}$, $RINFIN(\mathbf{x}, y; \epsilon, \hat{\beta}_n)$ is consistent estimate for $RINFIN(\mathbf{x}, y; \epsilon, \beta)$, $\epsilon > 0$.

b) Assume in addition to i) and ii) in a):
iv) $E\epsilon^4 < \infty$ and $E||\mathbf{X}||_2^4 < \infty$; $||\mathbf{u}||_2$ denotes the Euclidean L_2 norm of vector $\mathbf{u}$.
v) For at least one β-coordinate, e.g. the i-th:

$$g_i = \frac{\partial RINFIN(\mathbf{x}, y; \epsilon, \beta)}{\partial \beta_i} \neq 0. \tag{36}$$

Then, $RINFIN(\mathbf{x}, y; \epsilon, \hat{\beta}_n)$ is asymptotically normal:

$$\sqrt{n}[RINFIN(\mathbf{x}, y; \epsilon, \hat{\beta}_n) - RINFIN(\mathbf{x}, y; \epsilon, \beta)] \xrightarrow{\mathcal{D}} N(0, \mathbf{g}^T V \mathbf{g}); \tag{37}$$

$V = \tilde{\mathcal{E}}^{-1} E(\mathbf{X}_i \mathbf{X}_i^T \epsilon_i^2) \tilde{\mathcal{E}}^{-1}$ is the Covariance matrix of the asymptotic normal distribution of $\hat{\beta}_n$ and $\mathbf{g}$ has coordinates g_i in (36), $i = 0, 1, \ldots, p$.

Remark 3.5 *RINFIN's advantage, i.e. making additive the effects of $\mathbf{x}$ and r, remains for L_2-regression with diagonal weight matrix, W, independent of $\mathbf{x}, r$; Proposition 3.5 still* *holds with known $V(W)$ in (37). When W depends on $\mathbf{x}, r$, the decomposition of the influence function in Dollinger and Staudte (1991, Theorem 3, Equation (2)) indicates that RINFIN's advantage may not hold, depending on the form of the weights.*

4 ***RINFIN*** in Action

4.1 RINFIN and Simulations, $p < n$

Data $(\mathbf{X}, Y)$ from F follows linear model (11) with $\beta = (1.5, .5, 0, 1, 0, 0, 1.5, 0, 0, 0, 1, 0, \ldots, 0)$; when $p < 11$, β's first p coordinates are used. $\mathbf{X}$ is obtained from p-dimensional normal distribution, $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma}$'s entries $\Sigma_{i,j} = .5^{|j-i|}, 1 \leq i, \ j \leq p$, as in Alfons *et al.* (2013, p.11). For gross-error model, $F_{\epsilon,G}$, the proportion ϵ is 10%. For each contaminated $\mathbf{X}$ (from G) the first $\gamma \cdot p$ coordinates are *independent*, normal with mean μ and variance 1, $0 < \gamma \leq 1$. Various values for γ, p and μ are used and p is smaller than the sample size n. The regression errors are independent, standard normal random variables.

The simulations follow the spirit in Khan *et al.* (2007). Each of the $N = 100$ simulated samples has size $n = 100$. Cases $1 - 10$ are contaminated and compared with those having the 10 larger sample *RINFIN*-values for calculating the misclassification proportion.

COMPLETE CONTAMINATION ($\gamma = 1$)				
p	$\mu = .5$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
10	0.857	0.624	0.320	0.117
30	0.802	0.394	0.079	0.003
50	0.775	0.254	0.016	0.000
70	0.728	0.162	0.000	0.000
90	0.740	0.208	0.009	0.000

Table 1: Average misclassification proportion with RINFIN's orderings

In Table 1, the misclassification proportion decreases as p increases except for an anomaly when $p = 90$ due to its proximity to $n = 100$. By increasing n to 150 cases this anomaly disappears, *e.g.*, for $\mu = 1$ the misclassification proportion is 0.105.

PARTIALLY CONTAMINATED DATA IN THE FIRST $\gamma \cdot p$ X-COORDINATES						
p	$\mu = 1, \gamma = .2$	$\mu = 1, \gamma = .4$	$\mu = 1, \gamma = .6$	$\mu = 1.5, \gamma = .2$	$\mu = 1.5, \gamma = .4$	$\mu = 1.5, \gamma = .6$
10	0.859	0.834	0.747	0.811	0.695	0.550
30	0.822	0.753	0.599	0.719	0.516	0.296
50	0.804	0.676	0.506	0.663	0.364	0.164
70	0.787	0.612	0.416	0.598	0.250	0.089
90	0.784	0.605	0.435	0.611	0.294	0.116

Table 2: Average misclassification proportion with RINFIN's ordering

In Table 2, for *fixed* contamination proportion in the first $\gamma \cdot p$ **x**-coordinates, $\gamma(< 1)$, the *RINFIN* misclassification proportion decreases as p increases. The anomaly is still observed when $p = 90$. The blessing of high dimensionality is observed in both Tables 1, 2.

4.2 RINFIN and Real, High Dimensional Data, $p > n$

RINFIN is used for the microarray data in Zhao *et al.* (2016), obtained from Chiang et al.(2006) and previously analyzed by Zhao *et al.* (2013): 120 twelve-week-old male offspring were selected for tissue harvesting from the eyes; the data was kindly communicated to us by Leng (2017). The microarray contains over 30,000 different probe sets. Probe gene *TR32* is used as the response and the covariates are 1500 genes mostly correlated with it.

Since $n = 120 < p = 1500$, *RINFIN* is calculated for the response *TR32* and 100 **x**-covariates selected sequentially, in blocks, with coordinates $100(j-1)+1, \ldots, 100j,\ 1 \leq j \leq 15$. For each of the 120 cases, the total of its fifteen *RINFIN* values is its index, providing ordering of all the cases. In Table 3, cases with the higher 16 *RINFIN*-values are provided, more than 10% of the cases in order to get an idea of the differences in the values.

Indications for leverage cases from G in the gross-error model are given for cases 80, 95, 32, 120 and 59, after which the spacings are significantly reduced. In Table 4, the highest 16 *RINFINABS*-values are provided. Cases 80, 95, 32, 120 and 59 have still the same order as in Table 3, but the order of the remaining cases changes.

MICROARRAY DATA								
CASE	80	95	32	120	59	64	85	112
TOTAL RINFIN	824,471	146,639	40,295	24,749	14,802	12,849	12,582	11,683
CASE	38	40	24	117	27	28	84	90
TOTAL RINFIN	11,680	10,973	10,476	8,478	7,516	6,214	5,689	5,536

Table 3: Cases with the higher *RINFIN* values

MICROARRAY DATA								
CASE	80	95	32	120	59	85	38	112
TOTAL RINFIN	1744.5	797.4	488.1	379.4	319.3	285.6	282.1	273.6
CASE	64	24	40	27	117	6	84	28
TOTAL RINFIN	261.8	259.4	254.4	228.7	226	193.5	191.9	191.4

Table 4: Most inluential cases with *RINFINABS*

Cases 80, 95, 32, 120 and 59, are also supported by diagnostics *HIM* and *MIP*. According to Leng (2017), diagnostic *HIM* (Zhao *et al.*, 2013) finds 15 influential points with indices:

$$80, 95, 120, 32, 75, 70, 107, 28, 59, 38, 67, 27, 17, 51, 98;$$

diagnostic *MIP* (Zhao *et al.*, 2016) finds 7 influential points with indices:

$$80, 95, 120, 32, 75, 28, 59.$$

5 Appendix- Proofs and $\mathcal{E}$-matrix

Proof of Lemma 2.1: Equality (5) is obtained by adding and subtracting $T(F)$ in the numerator of its left side and by taking first the limit with respect to ϵ. $\square$

To proceed with the proof of Proposition 3.1 the general form of a symmetric, $(n+1)$ by $(n+1)$ matrix $\mathcal{E}_n$ is introduced. $\mathcal{E}_n$'s entries are motivated by the expected values in the equations' system (15), (16) when the n covariates are uncorrelated. $\mathcal{E}_n$'s main diagonal and its method of construction make it different from existing categories of matrices. $\mathcal{E}_n$'s cofactors are obtained and used to determine in *closed form* the Influence Functions of

L_2-regression coefficients. A similar result for least absolute deviation (L_1) regression coefficients also holds (Yatracos, 2018).

$\mathcal{E}$-MATRIX AND ITS COFACTORS

Under assumption ($\mathcal{A}3$), the coefficients in the system of equations (15), (16) form $\mathcal{E}_n$-matrix; n is the covariates' dimension. As an illustration, for real numbers a, b, c, A, B, C,

$$\mathcal{E}_3 = \begin{pmatrix} 1 & a & b & c \\ a & A & ab & ac \\ b & ba & B & bc \\ c & ca & cb & C \end{pmatrix}.$$

For $\mathcal{E}_3$, the corresponding linear regression model with uncorrelated covariates $X_1,\ X_2,\ X_3$ provides $a = EX_1,\ b = EX_2,\ c = EX_3$ and $A = EX_1^2,\ B = EX_2^2,\ C = EX_3^2$.

Definition 5.1 *$\mathcal{E}_n$-matrix with real entries has form:*

$$\mathcal{E}_n = \begin{pmatrix} 1 & a_1 & a_2 \dots & a_n \\ a_1 & A_1 & a_1 a_2 \dots & a_1 a_n \\ a_2 & a_2 a_1 & A_2 \dots & a_2 a_n \\ \dots & & & \\ a_n & a_n a_1 & a_n a_2 \dots & A_n \end{pmatrix}. \tag{38}$$

Notation: $\mathcal{E}_{n,-k}$ denotes the matrix obtained from $\mathcal{E}_n$ by deleting its k-th column and k-th row, $2 \le k \le n+1$.

Property of $\mathcal{E}_n$-matrix: Deleting the k-th row and the k-th column of $\mathcal{E}_n$-matrix, the obtained matrix $\mathcal{E}_{n,-k}$ is $\mathcal{E}_{n-1}$ matrix formed by $\{1, a_1, \dots, a_n\} - \{a_{k-1}\},\ 2 \le k \le n+1$.

The cofactors of $\mathcal{E}_n$-matrix are needed to solve the system of equations (15), (16).

Proposition 5.1 *a) The determinant of $\mathcal{E}_n$-matrix (38) is*

$$|\mathcal{E}_n| = \Pi_{m=1}^n (A_m - a_m^2). \tag{39}$$

b) Let $C_{i+1,j+1}$ be the cofactor of element $(i+1, j+1)$ in $\mathcal{E}_n$, then:

$$C_{i+1,j+1} = 0,\ \text{if } i > 0, j > 0, i \ne j, \qquad C_{1,j+1} = -a_j \Pi_{k \ne j}(A_k - a_k^2).$$

$$C_{i+1,1} = -a_i \Pi_{j\neq i}(A_j - a_j^2), \ \textit{if } i > 0, \qquad C_{1,1} = |\mathcal{E}_n| + \sum_{k=1}^{n} a_k^2 |\mathcal{E}_{n,-k}|.$$

Proof for Proposition 5.1: *a)* Induction is used.

For $n = 1$, the determinant is $A_1 - a_1^2$.

For $n = 2$, the determinant is

$$(A_1A_2 - a_1^2a_2^2) - a_1 \cdot (a_1A_2 - a_1a_2^2) + a_2 \cdot (a_1^2a_2 - A_1a_2) = A_1A_2 - a_1^2A_2 + a_1^2a_2^2 - A_1a_2^2$$

$$= A_2(A_1 - a_1^2) - a_2^2(A_1 - a_1^2) = (A_1 - a_1^2)(A_2 - a_2^2).$$

Assume that (39) holds also for $\mathcal{E}_n$. It is enough to show (39) holds for

$$\mathcal{E}_{n+1} = \begin{pmatrix} 1 & a_1 & a_2 \ldots & a_n & a_{n+1} \\ a_1 & A_1 & a_1a_2 \ldots & a_1a_n & a_1a_{n+1} \\ a_2 & a_2a_1 & A_2 \ldots & a_2a_n & a_2a_{n+1} \\ \ldots & & & & \\ a_n & a_na_1 & a_na_2 \ldots & A_n & a_na_{n+1} \\ a_{n+1} & a_{n+1}a_1 & a_{n+1}a_2 \ldots & a_{n+1}a_n & A_{n+1} \end{pmatrix}.$$

$|\mathcal{E}_{n+1}|$ is obtained using line $(n+1)$ and its cofactors $C_{n+1,1}, \ldots, C_{n+1,n+1}$:

$$|\mathcal{E}_{n+1}| = a_{n+1}C_{n+1,1} + a_{n+1}a_1C_{n+1,2} + \ldots + a_{n+1}a_nC_{n+1,n} + A_{n+1}C_{n+1,n+1}. \tag{40}$$

Observe that for $2 \leq j \leq n$, cofactor $C_{n+1,j}$ is obtained from a matrix where the last column is a multiple of its first column by a_{n+1}, thus,

$$C_{n+1,j} = 0, \ j = 2, \ldots, n. \tag{41}$$

For the matrix in cofactor $C_{n+1,1}$, observe that in its last column a_{n+1} is common factor and if taken out of the determinant the remaining column is the vector generating $\mathcal{E}_n$, i.e. $\{1, a_1, \ldots, a_n\}$. With $n-1$ successive interchanges to the left, this column becomes first and $\mathcal{E}_n$ appears. Thus,

$$C_{n+1,1} = (-1)^{n+2}(-1)^{n-1} \cdot a_{n+1}|\mathcal{E}_n| = -a_{n+1}|\mathcal{E}_n|. \tag{42}$$

In cofactor $C_{n+1,n+1}$, the determinant is that of $\mathcal{E}_n$,

$$C_{n+1,n+1} = (-1)^{2(n+1)}|\mathcal{E}_n| = |\mathcal{E}_n|. \tag{43}$$

From (40)-(43) it follows that

$$|\mathcal{E}_{n+1}| = -a_{n+1}^2|\mathcal{E}_n| + A_{n+1}|\mathcal{E}_n| = \Pi_{m=1}^{n+1}(A_m - a_m^2).$$

b) We now work with $\mathcal{E}_n$. For $i > 0, j > 0, i \neq j$, after deleting row $(j+1)$ the remaining of column $(j+1)$ in the cofactor is a multiple of column 1, thus $|C_{i+1,j+1}|$ vanishes.

For $C_{1,j+1}$, using column $j+1$ to calculate $\mathcal{E}_n$, it holds:

$$a_jC_{1,j+1}+A_jC_{j+1,j+1} = |\mathcal{E}_n| \to a_jC_{1,j+1} = -a_j^2\Pi_{k\neq j}(A_k - a_k^2) \to C_{1,j+1} = -a_j\Pi_{k\neq j}(A_k - a_k^2).$$

For $C_{i+1,1}, i > 0$, after deletion of row $(i+1)$ in $\mathcal{E}_n$ the remaining of column $(i+1)$ in the cofactor's matrix is multiple of a_i and the basic vector creating $\mathcal{E}_{n,-i}$. Column 1 of $\mathcal{E}_n$ is also deleted and for column $(i+1)$ in the cofactor's matrix to become first column $(i-1)$ exchanges of columns are needed. Thus,

$$C_{i+1,1} = (-1)^{i+2} \cdot a_i \cdot (-1)^{i-1}\Pi_{k\neq i}(A_k - a_k^2) = -a_i \cdot \Pi_{k\neq i}(A_k - a_k^2).$$

For $C_{1,1}$ we express $|\mathcal{E}_n|$ as sum of cofactors along the first row of $\mathcal{E}_n$,

$$C_{1,1} + a_1C_{1,2} + \ldots + a_nC_{1,n} = |\mathcal{E}_n|$$

$$\to C_{1,1} = \Pi_{k=1}^n(A_k - a_k^2) + a_1^2\Pi_{k\neq 1}(A_k - a_k^2) + \ldots + a_n^2\Pi_{k\neq n}(A_k - a_k^2). \qquad \square$$

Proof of Proposition 3.1: For system of equations (15), (16) and matrix $\mathcal{E}_p$ with $a_j = EX_j,\ A_j = EX_j^2,\ j = 1,\ldots,p$, from Proposition 5.1 with $r = r(\mathbf{x}, y; F)$,

$$IF_j = \frac{C_{1,j+1}r + C_{j+1,j+1}rx_j}{|\mathcal{E}_p|} = r\frac{-EX_j\Pi_{k\neq j}\sigma_k^2 + x_j\Pi_{k\neq j}\sigma_k^2}{\Pi_{k=1}^p\sigma_k^2} = r\frac{x_j - EX_j}{\sigma_j^2},\ j = 1,\ldots,p.$$

$$IF_0 = \frac{C_{1,1}r + \sum_{j=1}^p C_{1,j+1}rx_j}{|\mathcal{E}_p|} = r\frac{\Pi_{k=1}^p\sigma_j^2 + \sum_{j=1}^p(EX_j)^2\Pi_{k\neq j}\sigma_k^2 - \sum_{j=1}^p x_jEX_j\Pi_{k\neq j}\sigma_k^2}{\Pi_{k=1}^p\sigma_k^2}$$

$$= r[1 + \sum_{j=1}^p \frac{EX_j^2 - \sigma_j^2 - x_jEX_j}{\sigma_j^2} = r[1 - p + \sum_{j=1}^p \frac{EX_j^2 - x_jEX_j}{\sigma_j^2}]. \qquad \square$$

Lemma 5.1 *For the influence functions (18) with $r = r(\mathbf{x}, y; F)$ it holds:*

a)

$$IF_0 + \sum_{j=1}^{p} x_j IF_j = r[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}], \tag{44}$$

b)

$$IF_i + IF'_{x_i,0} + \sum_{j=1}^{p} x_j IF'_{x_i,j} = 2\frac{r \cdot (x_i - EX_i)}{\sigma_i^2} - \beta_i[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}] \tag{45}$$

$$\approx -3\beta_i \frac{(x_i - EX_i)^2}{\sigma_i^2}, \text{ if } |x_i - EX_i| \text{ is very large}, \tag{46}$$

c)

$$IF'_{y,0} + \sum_{j=1}^{p} x_j IF'_{y,j} = 1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}. \tag{47}$$

Proof of Lemma 5.1: *a)* From (18),

$$IF_0 + \sum_{j=1}^{p} x_j IF_j = r[1 - p + \sum_{j=1}^{p} \frac{EX_j^2 - x_j EX_j}{\sigma_j^2}] + \sum_{j=1}^{p} x_j \frac{r(x_j - EX_j)}{\sigma_j^2}$$

$$= r[1 - p + \sum_{j=1}^{p} \frac{EX_j^2 - 2x_j EX_j + x_j^2}{\sigma_j^2}] = r[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}].$$

b) Proof is provided for $i = 1$. Since

$$IF_0 = r[1 - p + \sum_{j=1}^{p} \frac{EX_j^2 - x_j EX_j}{\sigma_j^2}], \qquad IF_j = r\frac{x_j - EX_j}{\sigma_j^2}, \ j = 1, \dots, p,$$

$$IF'_{x_1,0} = -\beta_1[1 - p + \sum_{j=1}^{p} \frac{EX_j^2 - x_j EX_j}{\sigma_j^2}] - r\frac{EX_1}{\sigma_1^2}$$

$$IF'_{x_1,1} = -\beta_1 \frac{x_1 - EX_1}{\sigma_1^2} + \frac{r}{\sigma_1^2} \rightarrow x_1 IF'_{x_1,1} = -\beta_1 \frac{x_1^2 - x_1 EX_1}{\sigma_1^2} + r\frac{x_1}{\sigma_1^2}$$

$$IF'_{x_1,j} = -\beta_1 \frac{x_j - EX_j}{\sigma_j^2} \rightarrow x_j IF'_{x_1,j} = -\beta_1 \frac{x_j^2 - x_j EX_j}{\sigma_j^2}, \qquad j \neq 1.$$

Thus,

$$x_1 IF'_{x_1,1} + x_2 IF'_{x_1,2} + \ldots + x_p IF'_{x_1,p} = r\frac{x_1}{\sigma_1^2} - \beta_1 \sum_{j=1}^{p} \frac{x_j^2 - x_j EX_j}{\sigma_j^2}$$

$$\rightarrow IF_1 + IF'_{x_1,0} + x_1 IF'_{x_1,1} + x_2 IF'_{x_1,2} + \ldots + x_p IF'_{x_1,p}$$

$$= 2\frac{r(x_1 - EX_1)}{\sigma_1^2} - \beta_1[1 - p + \sum_{j=1}^{p} \frac{x_j^2 - 2x_j EX_j + EX_j^2}{\sigma_j^2}]$$

$$= 2\frac{r(x_1 - EX_1)}{\sigma_1^2} - \beta_1[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}].$$

Since

$$r(x_1 - EX_1) = y(x_1 - EX_1) - \beta_1 x_1(x_1 - EX_1) - (x_1 - EX_1)\sum_{j=2}^{p} \beta_j x_j$$

$$= y(x_1 - EX_1) - \beta_1(x_1 - EX_1)^2 - \beta_1(x_1 - EX_1)EX_1 - (x_1 - EX_1)\sum_{j=2}^{p} \beta_j x_j,$$

if $|x_1 - EX_1|$ is very large dominating all the other terms, then

$$IF_1 + IF'_{x_1,0} + x_1 IF'_{x_1,1} + x_2 IF'_{x_1,2} + \ldots + x_p IF'_{x_1,p} \approx -3\beta_1 \frac{(x_1 - EX_1)^2}{\sigma_1^2}.$$

c) From (18),

$$IF'_{y,0} = 1 - p + \sum_{j=1}^{p} \frac{EX_j^2 - x_j EX_j}{\sigma_j^2}, \qquad IF'_{y,j} = \frac{x_j - EX_j}{\sigma_j^2},\ j = 1, \ldots, p.$$

Thus,

$$IF'_{y,0} + \sum_{j=1}^{p} x_j IF'_{y,j} = 1 - p + \sum_{j=1}^{p} \frac{EX_j^2 - x_j EX_j + x_j^2 - x_j EX_j}{\sigma_j^2} = 1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}. \quad \Box$$

Proof of Lemma 3.1: Use approximations (2), (6). $\Box$

Proof of Proposition 3.2: *a)* Is provided for $i = 1$ using repeatedly Lemma 3.1:

$$r(\mathbf{x}_{1,h}, y; F_{\epsilon,\mathbf{x}_{1,h},y}) = y - \beta_0(F_{\epsilon,\mathbf{x}_{1,h},y}) - \beta_1(F_{\epsilon,\mathbf{x}_{1,h},y})(x_1 + h) - \ldots - \beta_p(F_{\epsilon,\mathbf{x}_{1,h},y})x_p$$

$$\approx y - \{\beta_0(F_{\epsilon,\mathbf{x},y}) + \epsilon h IF'_{x_1,0}\} - \{\beta_1(F_{\epsilon,\mathbf{x},y}) + \epsilon h IF'_{x_1,1}\}(x_1 + h) - \ldots - \{\beta_p(F_{\epsilon,\mathbf{x},y}) + \epsilon h IF'_{x_1,p}\}x_p$$

$$= r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y}) - \beta_1(F_{\epsilon,\mathbf{x},y})h - \epsilon h[IF'_{x_1,0} + x_1 IF'_{x_1,1} + x_2 IF'_{x_1,2} + \ldots + x_p IF'_{x_1,p}] - \epsilon h^2 IF'_{x_1,1}$$

$$= r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y}) - \beta_1 h - \epsilon h[IF_1 + IF'_{x_1,0} + x_1 IF'_{x_1,1} + x_2 IF'_{x_1,2} + \ldots + x_p IF'_{x_1,p}] - \epsilon h^2 IF'_{x_1,1}.$$

b) Lemma 3.1 is also used.

$$r(\mathbf{x}, y + h; F_{\epsilon,\mathbf{x},y+h}) = y + h - \beta_0(F_{\epsilon,\mathbf{x},y+h}) - \beta_1(F_{\epsilon,\mathbf{x},y+h})x_1 - \ldots - \beta_0(F_{\epsilon,\mathbf{x},y+h})x_p$$

$$\approx y + h - \{\beta_0(F_{\epsilon,\mathbf{x},y}) + \epsilon h IF'_{y,0}\} - \{\beta_1(F_{\epsilon,\mathbf{x},y}) + \epsilon h IF'_{y,1}\}x_1 - \ldots - \{\beta_p(F_{\epsilon,\mathbf{x},y}) + \epsilon h IF'_{y,p}\}x_p$$

$$= r(\mathbf{x}, y; F_{\epsilon,\mathbf{x},y}) + h - \epsilon h[IF'_{y,0} + \sum_{j=1}^{p} x_j IF'_{y,j}]. \qquad \Box$$

Proof of Corollary 3.1: a_1) The right side of (22) follows from (45).

a_2) If $|x_i|$ is large and $|h|$ is small, $\beta_i h$ and $\epsilon h^2 IF'_{x_i,i}$ are of smaller order than the remaining terms and (46) implies (23).

b) The right side of 24) follows from (47). $\square$

Proof of Proposition 3.3:

$$\lim_{|x_i|\to\infty} RINFIN(\mathbf{x}, y; \epsilon, L_2) \geq \epsilon \cdot \lim_{|x_i|\to\infty} \{ 2\frac{r(\mathbf{x}, y)(x_i - EX_i)}{\sigma_i^2} - \beta_i[1 + \sum_{j=1}^{p} \frac{(x_j - EX_j)^2}{\sigma_j^2}] \}^2$$

$$\approx \lim_{|x_i|\to\infty} 3^2 \beta_i^2 \frac{(x_i - EX_i)^4}{\sigma_i^4} = \infty;$$

the last approximation follows from (46). $\square$.

Proof of Proposition 3.4: Follows from (24) dividing both its sides by h and taking the limit with h converging to zero. $\square$

Lemma 5.2 *For regression model (11) under ($\mathcal{A}1$), ($\mathcal{A}2$), $(\mathbf{x}, y) \in R^{p+1}$,*

$$INF[i] = \epsilon \cdot \{2r[e^*_{i0} + \sum_{k=1}^{p} e^*_{ik} x_k] - \beta_i(e^*_{00} + 2\sum_{j=1}^{p} x_j e^*_{j0} + \mathbf{x}^T \mathcal{E}^* \mathbf{x})\}, \ i = 1, \dots, p. \tag{48}$$

Proof of Lemma 5.2: From (34),

$$\mathbf{IF} = \mathcal{E}^* \cdot \mathbf{q}, \quad \mathcal{E}^* = (e^*_{ij}) = \tilde{\mathcal{E}}^{-1}, \ 0 \leq i, j \leq p. \tag{49}$$

The Influence Function of β_j has form

$$IF_j = \sum_{k=0}^{p} e^*_{jk} q_k(\mathbf{x}, y; \beta) = r e^*_{j0} + r \sum_{k=1}^{p} e^*_{jk} x_k, \ j = 0, 1, \dots, p. \tag{50}$$

For $j = 0, 1, \dots, p, \ i = 1, \dots, p$

$$\frac{\partial IF_j}{\partial x_i} = e^*_{j0} \frac{\partial r}{\partial x_i} + \sum_{k=1}^{p} e^*_{jk} \frac{\partial \ (x_k \cdot r)}{\partial x_i} = -\beta_i(e^*_{j0} + \sum_{k=1}^{p} e^*_{jk} x_k) + r e^*_{ji}$$

$$\sum_{j=1}^{p} x_j \frac{\partial IF_j}{\partial x_i} = -\beta_i \sum_{j=1}^{p} x_j e^*_{j0} - \beta_i \sum_{j=1}^{p} x_j \sum_{k=1}^{p} e^*_{jk} x_k + r \sum_{j=1}^{p} x_j e^*_{ji}$$

$$INF(i) = \epsilon \cdot [IF_i + \frac{\partial IF_0}{\partial x_i} + \sum_{j=1}^{p} x_j \frac{\partial IF_j}{\partial x_i}]$$

$$= \epsilon \cdot [re^*_{i0} + r\sum_{k=1}^{p} e^*_{ik}x_k - \beta_i(e^*_{00} + \sum_{k=1}^{p} e^*_{0k}x_k) + re^*_{0i} - \beta_i \sum_{j=1}^{p} x_j e^*_{j0} - \beta_i \sum_{j=1}^{p} x_j \sum_{k=1}^{p} e^*_{jk}x_k + r\sum_{j=1}^{p} x_j e^*_{ji}]$$

$$= \epsilon \cdot \{2r[e^*_{i0} + \sum_{k=1}^{p} e^*_{ik}x_k] - \beta_i(e^*_{00} + 2\sum_{j=1}^{p} x_j e^*_{j0} + \sum_{j=1}^{p}\sum_{k=1}^{p} x_j e^*_{jk}x_k)\}. \qquad \square$$

Proof of Proposition 3.5: *a)* Conditions *i)-iii)* imply that the least squares estimate $\hat{\beta}_n$ is consistent estimate of β. From (28) and (48), $RINFIN(\mathbf{x}, y; \epsilon, \beta)$ is continuous function of β and therefore $RINFIN(\mathbf{x}, y; \epsilon, \hat{\beta}_n)$ is consistent estimate of $RINFIN(\mathbf{x}, y; \epsilon, \beta)$. *b)* Conditions *i), ii), iv)* imply that $\hat{\beta}_n$ has asymptotically multivariate normal distribution with covariance matrix $\tilde{\mathcal{E}}^{-1}E(\mathbf{X}_i\mathbf{X}_i^T\epsilon_i^2)\tilde{\mathcal{E}}^{-1}$. From (28) and (48), $RINFIN(\mathbf{x}, y; \epsilon, \beta)$ has continuous first partial derivatives at β which are not all zero from *v)*. Thus, $RINFIN(\mathbf{x}, y; \epsilon, \beta)$ has non-zero diferential at β. The result follows from Serfling (1980, Corollary in section 3.3, p. 124). $\square$

References

[1] Alfons, A., Croux, C. and Gelper, S. (2013) Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.* **7**, 226-248.

[2] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity.* Wiley, New York.

[3] Boente, G., Pires, A. M. and Rodrigues, I. M. (2002) Influence functions and outlier detection under the common principal components model: A robust approach *Biometrika* **89**, 861-875.

[4] Campbell, N. A. (1978) The influence function as an aid in outlier detection in discriminant analysis. *Appl. Statist.* **27**, 251-258.

[5] Carroll, R. J. and Ruppert, D. (1985) Transformations in Regression: A robust Analysis. *Technometrics* **27**, 1-12.

[6] Chiang, A. P., Beck J. S., Yen, H. J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D.Y., Braun, T. A., Kim, K. Y., Huang, J. Elbedour, K., Carmi,

R., Slusarski, D. C., Casavant, T. L., Stone, E. M., and Sheffield, V. C. (2006) Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet-biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 6287-6292.

[7] Cook, R. D. (1977) Detection of influential observations in linear regression. *J. Amer. Statist. Assoc.*, **74**, 169-174.

[8] Cook, R. D. and Weisberg, S. (1980) Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics* **22**, 495-508.

5North-Holland, Amsterdam.

[9] Dollinger, M. B. and Staudte, R. G. (1991) Influence Functions of Iteratively Reweighted Least Squares Estimators. *J. Amer. Statist. Assoc.* **86**, 709-716.

[10] Flores, S. (2015) Sharp non-aymptotic performance bounds for l_1 and Huber robust regression estimators. *Test*, **24**, 796-812.

[11] Genton, M. G. and Hall, P. (2016) A tilting approach to ranking influence *JRSS B*, **78**, 77-97.

[12] Genton, M. G. and Ruiz-Gazen, A. (2010) Visualizing influential observations in dependent data. *J. Comp. and Graph. Stat.* **19**, 808-825.

[13] Hadi, A. S. and Simonoff, J. S.(1993) Procedures for the Identification of Multiple Outliers in Linear Models. *J. Amer. Statist. Assoc.* **88**, 1264-1272.

[14] Hampel, F. R.(1985) The breakdown point of the mean combined with some rejection rules. *Technometrics* **27**, 95-107.

[15] Hampel, F. R.(1974) The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383-394.

[16] Hampel, F. R.(1971) A general qualitative definition of robustness *Ann. Math. Stat.* **42**, 1887-1896.

[17] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A.(1986) *Robust Statistics: The Approach Based on Influence Functions.* Wiley, N.Y.

[18] Hawkins, D. M., Bradu, D. and Kass, G. V.(1984) Location of several outliers in multiple regression data using elemental sets. *Technometrics* **26**, 197-208.

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

[20] Huber, P. J.(1964) Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **35**, 73-101.

[21] Hubert, M., Rousseeuw, P. J. and Van Aelst, S. (2008) High-Breakdown Robust Multivariate Methods. *Stat. Science* **23**, 92-119.

[22] Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *J. Amer. Statist. Assoc.* **102**, 1289-1299.

[23] Leng, C. (2017) Private communication.

[24] Ronchetti, E.(1987) Bounded Influenc Inference in Regression: A Review. *Statistical Data Analysis Based on the L_1-norm and Related Methods.*, p. . Editor Dodge, Y., North-Holland.

[25] Rousseeuw, P. J. and van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.*, **85**, 633-639.

[26] Rousseeuw, P. J. and Leroy, A. M.(1987) *Robust Regression & Outlier Detection.* Wiley, New York.

[27] Ruppert, D. and Carroll, R. J. (1980) Trimmed Least Squares Estimation in the Linear Model. *J. Amer. Statist. Assoc.* **75**, 828-838.

[28] Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics* Wiley, New York.

[29] She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *J. Amer. Statist. Assoc* **106**, 626-639.

[30] Tukey, J.W.(1962) The Future of Data Analysis. *Ann. Math. Stat.* **33**, 1-68.

[31] Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. *American Statistician* **35**, 234-242.

[32] Weisberg, S. S. (1985) *Applied Linear Regression* (2nd ed.) Wiley, New York.

[33] Welsch, R. E. (1982). Influence functions and regression diagnostics. In *Modern Data Analysis.* Academic Press, New York.

[34] Welsch, R. E. and Kuh, E. (1977). Linear regression diagnostics. Technical report 923-77. Sloan School of Management, Massachusetts Institute of Technology.

[35] Yatracos, Y. G. (2018) The Derivative of Influence Function, Location Breakdown Point, Group Influence and Regression Residuals' Plots. Under revision, https://arxiv.org/pdf/1607.04384v3.pdf

[36] Yatracos, Y. G. (2017) Discussion of "Random-projection ensemble classification" by T. I. Cannings and R.J. Samworth. *J. Roy. Statist. Soc. B* **79**, 1026-1027.

[37] Yatracos, Y. G. (2013) Detecting clusters in the data from variance decompositions of its projections. *Journal of Classification* **30, 1,** 30-55.

[38] Zhao, J., Leng, C., Li. L., and Wang, H. (2013). High-dimensional influence measure. *Ann. Stat.* **41**, 2639-2667.

[39] Zhao, J., Liu, C., Niu, L and Leng, C. (2016) Multiple influential point detection in high dimensional spaces. https://arxiv.org/abs/1609.03320

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.

018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.

019 "Lasso, knockoff and Gaussian covariates: a comparison" by Laurie Davies, May 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

020 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin and Yanli Zhu, May 2018.
021 "LASSO-Driven Inference in Time and Space" by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018.
022 " Learning from Errors: The case of monetary and fiscal policy regimes" by Andreas Tryphonides, June 2018.
023 "Textual Sentiment, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018.
024 "Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance" by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018.
025 "Construction of Non-asymptotic Confidence Sets in 2 -Wasserstein Space" by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018.
026 "Large ball probabilities, Gaussian comparison and anti-concentration" by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018.
027 "Bayesian inference for spectral projectors of covariance matrix" by Igor Silin, Vladimir Spokoiny, June 2018.
028 "Toolbox: Gaussian comparison on Eucledian balls" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.
029 "Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification" by Nikita Puchkin, Vladimir Spokoiny, June 2018.
030 "Gaussian Process Forecast with multidimensional distributional entries" by Francois Bachoc, Alexandra Suvorikova, Jean-Michel Loubes, Vladimir Spokoiny, June 2018.
031 "Instrumental variables regression" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.
032 "Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective" by Li Guo, Yubo Tao and Wolfgang Karl Härdle, July 2018.
033 "Optimal contracts under competition when uncertainty from adverse selection and moral hazard are present" by Natalie Packham, August 2018.
034 "A factor-model approach for correlation scenarios and correlation stress-testing" by Natalie Packham and Fabian Woebbeking, August 2018.
035 "Correlation Under Stress In Normal Variance Mixture Models" by Michael Kalkbrener and Natalie Packham, August 2018.
036 "Model risk of contingent claims" by Nils Detering and Natalie Packham, August 2018.
037 "Default probabilities and default correlations under stress" by Natalie Packham, Michael Kalkbrener and Ludger Overbeck, August 2018.
038 "Tail-Risk Protection Trading Strategies" by Natalie Packham, Jochen Papenbrock, Peter Schwendner and Fabian Woebbeking, August 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

039 "Penalized Adaptive Forecasting with Large Information Sets and Structural Changes" by Lenka Zbonakova, Xinjue Li and Wolfgang Karl Härdle, August 2018.
040 "Complete Convergence and Complete Moment Convergence for Maximal Weighted Sums of Extended Negatively Dependent Random Variables" by Ji Gao YAN, August 2018.
041 "On complete convergence in Marcinkiewicz-Zygmund type SLLN for random variables" by Anna Kuczmaszewska and Ji Gao YAN, August 2018.
042 "On Complete Convergence in Marcinkiewicz-Zygmund Type SLLN for END Random Variables and its Applications" by Ji Gao YAN, August 2018.
043 "Textual Sentiment and Sector specific reaction" by Elisabeth Bommes, Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, September 2018.
044 "Understanding Cryptocurrencies" by Wolfgang Karl Härdle, Campbell R. Harvey, Raphael C. G. Reule, September 2018.
045 "Predicative Ability of Similarity-based Futures Trading Strategies" by Hsin-Yu Chiu, Mi-Hsiu Chiang, Wei-Yu Kuo, September 2018.
046 "Forecasting the Term Structure of Option Implied Volatility: The Power of an Adaptive Method" by Ying Chen, Qian Han, Linlin Niu, September 2018.
047 "Inferences for a Partially Varying Coefficient Model With Endogenous Regressors" by Zongwu Cai, Ying Fang, Ming Lin, Jia Su, October 2018.
048 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin, Yanli Zhu, October 2018.
049 "Strict Stationarity Testing and GLAD Estimation of Double Autoregressive Models" by Shaojun Guo, Dong Li, Muyi Li, October 2018.
050 "Variable selection and direction estimation for single-index models via DC-TGDR method" by Wei Zhong, Xi Liu, Shuangge Ma, October 2018.
051 "Property Investment and Rental Rate under Housing Price Uncertainty: A Real Options Approach" by Honglin Wang, Fan Yu, Yinggang Zhou, October 2018.
052 "Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective" by Qingliang Fan, Wei Zhong, October 2018.
053 "The impact of temperature on gaming productivity: evidence from online games" by Xiaojia Bao, Qingliang Fan, October 2018.
054 "Topic Modeling for Analyzing Open-Ended Survey Responses" by Andra-Selina Pietsch, Stefan Lessmann, October 2018.
055 "Estimation of the discontinuous leverage effect: Evidence from the NASDAQ order book" by Markus Bibinger, Christopher Neely, Lars Winkelmann, October 2018.
056 "Cryptocurrencies, Metcalfe's law and LPPL models" by Daniel Traian Pele, Miruna Mazurencu-Marinescu-Pele, October 2018.
057 "Trending Mixture Copula Models with Copula Selection" by Bingduo Yang, Zongwu Cai, Christian M. Hafner, Guannan Liu, October 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

058 "Investing with cryptocurrencies – evaluating the potential of portfolio allocation strategies" by Alla Petukhina, Simon Trimborn, Wolfgang Karl Härdle, Hermann Elendner, October 2018.

059 "Towards the interpretation of time-varying regularization parameters in streaming penalized regression models" by Lenka Zbonakova, Ricardo Pio Monti, Wolfgang Karl Härdle, October 2018.

060 "Residual's Influence Index (Rinfin), Bad Leverage And Unmasking In High Dimensional L2-Regression" by Yannis G. Yatracos, October 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.