Farnè, Matteo; Vouldis, Angelos T.

Working Paper

A methodology for automatised outlier detection in high-dimensional datasets: An application to euro area banks' supervisory data

ECB Working Paper, No. 2171

Provided in Cooperation with:
European Central Bank (ECB)

Suggested Citation: Farnè, Matteo; Vouldis, Angelos T. (2018) : A methodology for automatised outlier detection in high-dimensional datasets: An application to euro area banks' supervisory data, ECB Working Paper, No. 2171, ISBN 978-92-899-3276-9, European Central Bank (ECB), Frankfurt a. M.,
https://doi.org/10.2866/357467

This Version is available at:
https://hdl.handle.net/10419/183353

EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Matteo Farnè, Angelos T. Vouldis

A methodology for automatised outlier detection in high-dimensional datasets: an application to euro area banks' supervisory data

No 2171 / July 2018

Abstract

Outlier detection in high-dimensional datasets poses new challenges that have not been investigated in the literature. In this paper, we present an integrated methodology for the identification of outliers which is suitable for datasets with higher number of variables than observations. Our method aims to utilise the entire relevant information present in a dataset to detect outliers in an automatized way, a feature that renders the method suitable for application in large dimensional datasets. Our proposed five-step procedure for regression outlier detection entails a robust selection stage of the most explicative variables, the estimation of a robust regression model based on the selected variables, and a criterion to identify outliers based on robust measures of the residuals' dispersion. The proposed procedure deals also with data redundancy and missing observations which may inhibit the statistical processing of the data due to the ill-conditioning of the covariance matrix. The method is validated in a simulation study and an application to actual supervisory data on banks' total assets.

Keywords: Outlier detection; Robust regression; Variable selection; High dimension; Missing data; Banking data

JEL classification codes: C18, C81, G21

Non-technical summary

The analysis of high-dimensional data sets is very much affected by the presence of outliers which may distort the conclusions reached when standard econometric methods are applied. On the other hand, large datasets are becoming more and more common in banking and finance calling for the adaptation of existing methods or the adoption of new approaches. In this context, outlier detection in high-dimensional datasets poses new challenges, while offering new possibilities for enhanced outlier detection methods which have not been investigated in the literature.

In this paper, we present an integrated methodology for the identification of outliers which is suitable for 'fat' datasets i.e. high-dimensional data sets with attributes exceeding the number of observations. A typical example of such dataset is represented by the supervisory statistics collected in the context of the ECB Banking Supervision, whereby extremely granular information for banks' activities, risks and profitability is provided for the systemically significant and less significant institutions.

Our method aims to utilise the whole relevant information present in such a dataset to detect outliers in an automatized way, a feature that renders the method suitable for application in high-dimensional data sets. For example, when one tries to identify outliers in the variable "notional amount of derivatives as a percentage of assets", the variable "asset size" may provide critical complementary information for the identification of outliers of the former variable. In fact, it may happen that the value of one variable can be considered an outlier conditional on another one, but not in the univariate sense. This fact keeps its validity if the conditioning variables are more than one. This type of outlier detection, which is based on deviations from the regression hyperplane representing the bulk of the data, can produce valuable insights as a by-product and inform further analytical work.

Our proposed five-step procedure for regression outlier detection entails a robust selection stage of the most explicative variables, the estimation of a robust model based on the selected variables, and a criterion to identify outliers based on robust measures of the residuals' dispersion. In addition, the proposed procedure deals also with data redundancy and missing issues which may inhibit the statistical processing of the data due to the ill-conditioning of the covariance matrix. The method is validated in a simulation study and it is also applied to actual supervisory data on banks' total assets.

1. Introduction

Data quality is a prerequisite for any reliable, quantitative type of analysis and the large data sets which are becoming more common present specific challenges to the task of monitoring and ensuring data quality. One critical aspect of data quality monitoring is outlier detection i.e. the identification of values which either are obviously mistaken or seem to be unjustified from a business side perspective. Classical statistical methods, which underpin analytical tools supporting analysis of large data sets, are sensitive to the presence of outliers and may be led, consequently, to present a distorted picture of the reality due to the presence of outlier values leading to erroneous conclusions.

On the one hand, a challenge for outlier detection in large datasets is that the development of purely automatized and efficient processes is required (see for example Maciá-Pérez et al. 2015). While large datasets present a number of challenges regarding data processing and analysis e.g. related to the difficulty of examining the data, it offers also the possibility to utilise the richness of the dataset for outlier detection. While the uncertainty of a large dataset may be larger compared to a smaller one (except e.g. in the case of surveys, where larger data allow more precise answers to be obtained from the data), the size of the dataset compels the development of methods which should make the optimal usage of the existing data.

A strand of research has focused on outlier identification on *datums* that comprise vectors of numerical (or also categorical) attributes i.e. outliers when all dimensions of each *datum* are considered (hereafter called 'multi-dimensional outliers'). Indicative references from this vast literature include Otey et al. (2006), Koufakou and Georgiopoulos (2010), and Kutsuma and Yamamoto (2017).[1] However, the identification of multi-dimensional outliers presupposes a "clean" dataset with respect to the individual values in the sense that large deviations of particular observations of specific variables from their expected values represent the true behaviour of the respective variables rather than outliers. In the opposite case, the detection of multi-dimensional outliers will also be affected by distortions caused by outliers in individual values. Furthermore, in a

[1] See also the special issue of *Data Mining and Knowledge Discovery* (Volume 20, Issue 2, March 2010) which is entirely devoted to the detection of multi-dimensional outliers.

large-dimensional dataset, some of these outliers may not be detected by multi-dimensional methods as they may be "hidden" within the granular dataset and not affect significantly the multi-dimensional distance metrics used by the respective outlier detection method. Therefore, it is critical to address the issue of outliers in single variables.

The contribution of this paper is to propose a multi-step "integrated procedure" for single variable outlier detection which draws on the full information present in a multi-dimensional dataset combining robust methods to generate insights about the nature of the detected outliers and the structure of the dataset. The proposed method includes both a robust variable selection step and a robust regression step. We formulate and implement an integrated approach in different variants which differ with respect to the combination of the methods employed in these two key steps. Furthermore, we perform a controlled assessment of its performance under different data features in order to infer the optimal calibration of the method. In addition, we apply the method to a large dataset of supervisory banking data from the largest European banks, collected by the European Central Bank (ECB).

Furthermore, the paper formulates a generic methodology for detecting outliers in "fat" datasets, while formalising the procedure to deal with practical problems such as data redundancy. The procedure to deal with data redundancy uses a formal criterion to exclude variables, based on a measure of the statistical 'importance' of each variable.

The simplest approach to detect outliers for one single variable uses distribution-based techniques. However, such univariate approach which utilises only the sample values for a single variable may not be sufficient in this specific context. The main reason is that outliers could remain unnoticed because information contained in the values of some other related variables is not utilised. For example, in an application to banking data, as will be presented below, when one tries to identify outliers in the variable "notional amount of derivatives as a percentage of assets", the variable "asset size" may provide critical complementary information for the identification of outliers of the former variable. If for example, a small bank's balance sheet contains a relatively large percentage of derivatives, this value may be considered to represent an outlier because in general it is mainly large and complex banks that use derivatives extensively. Therefore, in such case the derivative amount

value can be considered an outlier conditional on the asset size of the bank, even though neither the derivative amount nor the asset size could be considered outliers in the univariate sense. In other words, a variable conditional on a set of related variables can represent an outlier even though neither of the variables individually represents an outlier. This type of outlier detection in the regression sense can produce valuable insights as a by-product and inform further analytical work.

Therefore, there is clearly ground for regression-based outlier detection in high-dimensional datasets in this sense, taking into account the properties of the empirical distribution for each variable, as well as statistical information derived from an appropriately defined set of related variables. However, when aiming to detect outliers in this manner, a number of new issues arise. First, the identification of the explanatory variables should be automatized robustly given the large number of variables and the impossibility of selecting the regressors based purely on economic criteria and expert judgment. Second, this identification of relevant regressors necessitates the 'cleaning' of the initial data set from highly correlated subsets of variables, a procedure which should also be automated. This approach follows the research strand which incorporates machine learning techniques into economic analysis, given the limitations of classical model-based econometric procedures to tackle large datasets (Varian 2014).

The existing literature has provided algorithms for regression outlier detection. However, these do not seem to be particularly suited for high-dimensional datasets with attributes exceeding the number of observations ("fat" data), especially if there are missing observations, collinear variables, or observations with outstanding magnitude and variability.

The paper is structured as follows. Section 2 reviews the literature on outlier detection. Section 3 presents our proposed methodology for regression outlier detection in large dimensions. Section 4 conducts a simulation study against an appropriate set of benchmarks to assess the performance of the proposed method in a controlled environment. Section 4 provides an application to a selection of ECB supervisory data, presenting and discussing outlier detection results. Finally, Section 6 concludes.

2. Related literature

Outlier detection is a critical step in the statistical analysis of large data sets. Hawkins (1980) provided the definition of outlier, intended as '*an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism*'. Barnett and Lewis (1994) define outlier(s) as '*an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*'. These general definitions may be applied and assume different meaning in various contexts. Many statistical branches include outlier detection as a relevant topic, as widely described in Huber (2004).

2.1 Distribution-based methods

Most of standard non-robust statistical methods are based on distributional assumptions which are strongly affected by the presence of outliers. For example, classical multivariate linear and non-linear regressions, clustering, principal components, and factor analysis all are based on standard means, covariance and correlation matrices, which are not robust against outliers. A point that is often missed by researchers applying classical statistical or econometric methods is that the presence of outliers affects the estimation of the model and cannot be conducted ex-post after a model has been estimated; on the contrary, the 'distorted' model may fit the data fairly well, but the insights it offers about the underlying data have been irrevocably determined by the presence of outliers. Consequently, the outlier detection stage should precede the statistical analysis of the data, or, alternatively, a robust version of above mentioned methods should be used.

Robust methods have been first developed to limit the effect of the presence of outliers for location parameters estimation. The Huber estimator (or M-estimator, see Huber (1964)) was the seminal idea, which opened the path to robust likelihood methods. Another seminal contribution is the Hampel identifier which proposes a quantile-based outlier identification rule (Hampel 1968). The use of distribution-based outlier detection methods is the most direct approach to univariate outlier detection. Essentially, one needs to define a central value, which can be the mean (as in the case of 3-sigma rule)

or the median (as in Hampel identifier). The box plot rule, flagging as outliers all the values being more than 1.5 times the interquartile range, is often used.

Distribution-based methods can also be applied to the multivariate context when adjusted appropriately. The analogous of 3-sigma rule in the multivariate context is the Mahalanobis distance (Mahalanobis 1936) defined as $D = \sqrt{(\boldsymbol{x} - \overline{\boldsymbol{x}})' S^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}})}$, where S is the covariance matrix, $\boldsymbol{x}$ is a data vector and $\overline{\boldsymbol{x}}$ is the mean vector. In order to robustify this distance, which may be conditioned by outliers, the MCD (Minimum Covariance Determinant) estimator is computed (Rousseeuw and van Driessen 1999), where only the $100(1-\alpha)\%$, $\alpha \in [0,0.5]$, observations giving the smallest possible determinant of S are kept. In this way masking or swamping problems can be effectively addressed. MCD estimates can be computed by Fast-MCD, proposed in the same paper, or by an alternative procedure called Bacon EEM algorithm (Beguin and Hulliger, 2008).

Hubert et al. (2008) proposes robust estimations of the mean vector and covariance matrix together with χ^2 approximation to obtain cut-off values. In Cerioli (2010) the size properties of the same methodology are improved by using finite-sample distributional results for the definition of distances and the cut-off values. Robust multivariate mean and covariance estimation has been recently approached in an alternative way by Maronna and Zamar (2012) and Pena and Prieto (2012). Maronna and Yohai (2016) provides a general overview on the topic.

2.2 Regression-based methods

Besides distribution-based methods, regression-model based methods can be used in a multivariate context. Rousseeuw and Leroy (1987) provide a detailed description of outlier detection methods in relationship to regression methods, highlighting two different approachs. First, there is the outlier diagnostics approach which works with residuals produced by a standard OLS regression or a standard OLS regression where some points have been omitted. Studentised residuals computed from the *leave-one-out* approach (Weisberg 1985) or the application of the Cook distance and DFFITS can be effective to detect influential objects, i.e. the observations which are affecting significantly the

regression plane. In addition, the properties of the 'hat' matrix H which transforms the observed values to the predicted ones ($\hat{y} = Hy$) can be investigated for outlier identification. For example, if a diagonal element of H is 'large', meaning that $\frac{\partial \hat{y}_i}{\partial y_i}$ is large (as defined by some criterion), then i is an influential point (e.g. Henderson and Velleman 1981). Unfortunately, these methods are effective when the number of outliers and observations is small (e.g. one or two outliers), however they become computationally infeasible when the number of outliers is high. Therefore, the scope for their application in a 'fat' large data set is very limited.

The second approach is to apply robust regression methods which are less sensitive to the presence of outliers compared to the OLS. Several approaches have been proposed and their effectiveness can be assessed using the concept of the 'breakdown point' i.e. the fraction of distorted points which can have an arbitrarily large effect on the regression operator (see Hubert et al. 2008).

Three very well-known robust regression estimation methods are: i) *Least Trimmed Squares* (LTS), which estimates regression coefficients minimizing the trimmed squares of residuals (Rousseeuw and Leroy 1987, 2006); ii) the *MM-estimation,* which is a 3-stage procedure, based on M-estimation, ensuring at a given breakdown point much higher efficiency than LTS (Yohai, 1987); and iii) the *GS-estimation,* which is based on the optimization of a generalized function of the residual scale, guaranteeing minimum max-bias (Croux et al., 1994). In addition, Salibian-Barrera and Zamar (2002) propose a bootstrap approach to robust regression. For a general overview on robust regression we refer to Hubert et al. (2008) and Rousseeuw and Hubert (2011).

A rather different approach, described in Au et al. (2008), proposes a methodology for irregularity detection in time series data by using a decomposition of time series, clustering of time series and LASSO regression method for components selection. This approach has some similarities with this paper, given that it can be used for large cross sections, however, it is assumed that there is a high degree of homogeneity among the individual variables (time series), so that they can be conveniently classified into clusters, which is not the case for supervisory data, because the degree of heterogeneity among variables is extremely high.

2.3 Methods for high-dimensional data

The emergence of high-dimensional datasets creates the need for the development of new data analysis techniques. We know that 'Big data' are large datasets possessing the so called 5 Vs: volume, velocity, variety, variability, veracity. They can also be viewed as data sets for which standard statistical techniques are not effective and whose size poses logistical challenges for existing database software tools. As Varian (2014) notes, the size of the data requires automated techniques for the identification of subsets of variables which are statistically linked. In addition, high-dimensional data sets may exhibit types of statistical relationships, such as highly correlated variables, which would affect negatively the application of standard statistical techniques, for example by rendering correlation matrices ill-conditioned. In such datasets, even in case they are not so high-dimensional to be defined as 'Big', the procedures for eliminating statistically redundant information should be automatized. For example, the decision to exclude one of two highly correlated variables should be based on a specific computational criterion because it is not possible to decide by inspecting the variables in question. Furthermore, 'fat' datasets require some kind of variable selection when working in a multivariate context, since the number of potential regressors may be larger than the number of observations. In this direction, and considering these practical issues, there is a need to develop single-variable outlier detection methodologies in a multivariate context simultaneously robust to the presence of missing data, high collinearity, innovation outliers and a dimension larger than the sample dimension.

The most traditional method for variable selection is LASSO (Tibshirani, 1996): even if LASSO gives the sparsest possible solution and has a certain protection from noise (see Xu et al., 2009, Yang and Xu, 2013), it is not robust against corrupted measurements. For this reason, some variants of the LASSO like the extended LASSO (Nasrabadi et al., 2011) and the fused LASSO (Kim et al., 2015) have been proposed. In addition, there are alternative methods exploiting Huber penalty (Owen, 2007) or minimum distance estimators (Lozano et al., 2016).

A recent approach by (Yi and Huang, 2016) integrates Huber/quantile losses and the elastic net penalty for robust regression estimation, thus encompassing both LASSO and ridge penalties. This procedure is based on a semi-smooth Newton coordinate descent algorithm which provides regression coefficients estimates, but no robust estimate of the residual scale. This happens because the objective is not optimized via a reweighted least squares method. Consequently, this approach (from now labelled "HQ") will be used as the robust variable selection step to identify the most explicative covariates, so that the usual robust regression methods like LTS, GS and MM may be subsequently used to perform outlier detection.

A competitive family of regression outlier detection methods in high dimensions is based on the optimization of a function which selects covariates and identifies outliers robustly at the same time. These methods include the robust version of the Least Angle Regression (LARS) by Efron et al. (2004), which was developed in (Khan et al., 2007), and SPARSE-LTS (Alfons et al., 2016), a method which optimizes an objective composed by a trimmed sum of squares and an $l_1$penalty. SPARSE-LTS will be the main competitor of our integrated approach throughout the paper.

For the sake of completeness, we mention that there are several multivariate outlier detection methods for high-dimensional data which do not employ multivariate regression. Among others we mention Entropy Fast Detection (Liu et al., 2013), Local Search Algorithm (He et al, 2006), a rank-order approach for high-dimensional databases (Texeira et al. 2008), a cluster-based similarity method (Christy et al. 2015), and Cell-DROS method (Van Hieu and Meesad 2016). In the same context, Todorov et al. (2011) and Templ et al. (2017) are two multivariate outlier identification methods for datasets affected by missing values and structural zeros respectively. Despite that, our aim is to identify the outliers in one variable, taking advantage of the knowledge of other related variables. For this reason, we will not take into account this family of methods.

In this paper, our reference model setting has the distinctive feature that neither y nor x are directly perturbed, while the regression coefficients β are. This case may be very common for banking data, where the usual relationships among variables may be amplified or annihilated for specific banks. In

addition, we allow for the presence of missing data, high collinearity, innovation outliers (i.e. outliers with arbitrarily amplified innovation error) and a dimension p (i.e. the number of quantitative variables or attributes) larger than the sample dimension n (i.e. the number of observations). Therefore, it is of interest to study the behaviour of SPARSE-LTS and our integrated procedure under these conditions. Our integrated approach takes successfully into account all these issues at the same time being an "integrated" methodology for regression outlier detection.

3. An integrated outlier detection methodology

3.1. Outlier detection in supervisory data

As already explained, the literature has not paid sufficient attention to the development of integrated, automatized methodologies which can deal with large datasets and perform multivariate outlier detection in them. 'Integrated' here refers to the incorporation of all required steps for multivariate outlier detection, including the preparation steps of the data set, such as standardisation, dealing with completeness issues and 'cleaning' the data from redundant information. This paper aims to fill this gap, focusing on 'fat' large datasets, and employing banking supervisory data to test the proposed method. Our data set does not strictly possess the size which would lead to its characterisation as 'Big data', however the techniques we develop are aimed for application in a 'Big data' context; given also the increase in the supervisory data set that we use, which will growth significantly with time, as banks submit an expanded data set at a quarterly frequency.

Zhang et al (2010) provide a survey of outlier detection techniques focusing on specific type of data, namely data from wireless sensor networks. They point out that, in this specific field, the approaches adopted for outlier detection neglect in various ways the properties which characterise the underlying data; for example, the range of dependencies considered does not correspond to the actual dependencies characterising the sensor data. In general, the development of integrated methodologies should be customised for the type of data in which it will be applied. The dimensions, dependence structures, range of values, and the statistical distributions properties all influence the optimal selection of methods.

Outliers in supervisory data can be succinctly defined as those data that significantly deviate from the normal pattern of scaled data. This definition is based on the fact that supervisory data (e.g. the various types of loans) when scaled by the banks' size (total assets) are comparable across banks of different sizes and represent the composition of the activities and corresponding risks that the bank is undertaking. Differences in the composition of activities among subsets of banks may exist e.g. because different banks follow different business models (Farnè and Vouldis, 2016). Furthermore, there may be scale effects in the composition of activities, e.g. large banks with more complex strategies may, for example, make more extended use of derivative instruments. Despite the existence of such distinctive features which can characterise subsets of banks, and may lead to various reported variables following distinct statistical distributions within such subsets, we expect that there is also an aggregate statistical distribution for almost all reported items when they are standardised. The bulk of these distributions (usually strongly skewed, as pointed out by Hubert and van der Veeken, 2008) is expected to be located within a subset of [0,1], given the normalisation with respect to assets[2] and the fact that most items are lower than the bank's size – although a few variables, such as the notional amounts of derivatives, may exceed unity i.e. their size may be higher than the bank's total assets.

3.2. Problem definition

Our problem can be formulated in an abstract way as follows. Let X be a $n \times p$ matrix containing a set of p variables with n observations. In our case, each observation index refers to a bank-reference date combination, while the variables cover a wide range of aspects about the banks' activities, performance or risk-taking.

Each observation vector is denoted as

$$x_i = \left(x_{i1,\cdots,}x_{ip}\right), i = 1,\cdots,n.$$

Each observation can be thought of as referring to either a specific bank, in the case that our sample contains observations for n banks, or a bank-date pair, in the case of panel data, when more than one

[2] The total asset value represents an upper bound for almost all the items present in banks' balance sheets, irrespectively of whether they represent assets, liabilities, off balance sheet items or flow components such as profit. Some of these constraints are driven by accounting identities and some others by the business reality.

reference date is present in the dataset. In the more general case, we could also have an unbalanced panel of banks, i.e. a changing cross section of banks, if we have data from different reference periods.

Each variable vector is denoted as

$$x^j = \left(x_{1j,\cdots,}x_{nj}\right), j = 1, \cdots, p.$$

The present paper focuses on outliers for specific variables (see Farnè and Vouldis (2016) for the analysis of outlier data vectors). Consequently, our aim is to identify outlier data points meaning specific data points x_{ij} that come from another population compared to the other elements of the vector x^j.

The outlier detection process for a variable x^j takes as input a certain information set Ω^j. The simplest approach is to use only the sample of variable x^j in order to identify the outliers within that sample i.e. $\Omega^j = x^j$. An alternative, which is explored here, is to extend the information set and include also other "related" variables. The idea then would be to utilise statistical relationship(s) between a relevant set of variables in order to define the outliers. For example, one could identify the loan-to-assets ratio as being an outlier for a bank based solely on the sample distribution of loan-to-assets for the whole set of banks. However, one could utilise additional information, e.g. trading assets-to-assets in order to identify outliers, given that we would expect a negative relationship between these two variables, and, therefore, a bank which is characterized by relatively high values for both of these variables is an outlier. It is clear in this example, that an outlier detection algorithm which utilises this additional information would identify values of dubious quality more reliably.

A salient feature of our approach is that it is suited for high-dimensional large data, such as those collected under the ECB Banking Supervision institutional set-up, i.e. data where $p \gg n$. In other words, the granularity of the data set is very large compared to the number of entities which are reporting the supervisory data.

The manipulation of this information provides a challenge to develop robust methodologies for outlier detection. Specifically, due to the large size of the data, the selection of relevant variables which could be used to inform the outlier detection procedure for each specific variable is not straightforward and

needs to be automatized. Consequently, we aim to identify outlier one-dimensional data points in this large data set taking into account the relations which exist among the variables in this data set, while automatizing all the steps in this procedure.

3.3. A five-step procedure

Our proposed methodology consists of five (5) steps, entailing robust covariates selection, estimation of a robust model based on the selected variables and a criterion to identify outliers based on robust measures of the residuals' distribution. Statistically related information is therefore identified and then utilised to spot outliers for each variable. This procedure can be applied consecutively to all the variables of a data set, therefore enabling the identification of outliers for all variables.

Specifically, our proposed multivariate outlier detection methodology consists of the following steps, which will be explained below in detail:

Step 1: Standardization of variables or normalization (optional).

Step 2: Selection of "closely related" variables ("determinants") using a procedure based on the semi-smooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression of Yi and Huang (2016) (HQ procedure). Such method estimates the regression coefficients by minimizing a Huber loss plus a classical l_1 penalty. In this way, we can identify the most relevant L determinants in a robust way.

Step 3: Using the selected set of variables we run a robust regression. In this step, three possible alternatives could be utilised:

- *least trimmed squares* (LTS) estimation, which identifies the $100 \times (1 - \alpha)\%$ most concentrated observations and estimates the model on those;
- MM-estimator (Yohai, 1987), the most efficient estimator given a level of breakdown, which minimizes a Huber function of the residuals and uses a robust initialization of the coefficients β and the residual scale σ;

- GS-estimator (Croux et al., 1994), which estimates the residual scale minimizing a Tukey biweight function of scaled residuals.

More details about these methods are provided below.

Step 4: Calculate in a robust way the dispersion of residuals.

Step 5: Identify outliers based on the calculated dispersion.

These steps are described in detail below.

Step 1: Standardisation (optional)

The aim of standardisation is to abstract from differences in the first two moments of the distributions followed by the variables in the dataset. Standardisation is performed by subtracting a robust estimator of the mean of each variable and then dividing the result by a robust measure of variance. This step is optional. The main criterion for its application is whether we would like to retain or abstract from the background ('confounding') characteristics of the Pvariables. Each of the described methods has its natural standardization pre-processing step, which is used by default. In principle, our outlier detection method could work with both standardized and non-standardized input data.

Step 2: Selection of determinants[3] using results of LASSO estimations

This step aims to select, in an automatized way, the statistical determinants of the variable in question. Specifically, for each variable x^j, we would like to search within the set of remaining variables $V^j = \{x^k : k = 1, \cdots, p; k \neq j\}$ and select a subset of its "closely related" variables ("determinants") which can be used to predict the values of x^j. Given the "fat" nature of the dataset, the aim here is to select a relatively small number of 'determinants'. This is accomplished by estimating the following linear model with the set V^jat the righthand side

$$x^j = \alpha + V^j\beta + \varepsilon$$

[3] The term "determinants" here does not imply causality in an economic sense but refers to the set of variables which are statistically related to and can be used to estimate a statistically "plausible" value for the dependent variable.

using the semi-smooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression (Yi and Huang, 2016):

$$\beta = \{\beta_k\}_{k=1,k\neq j}^{min} \sum_{i=1}^{n}\Big(\rho_\tau(r_i) + \lambda \sum_{k=1,k\neq j}^{p} |\beta_k|\Big),$$

where $r_i = x_{ij} - \sum_{k=1,k\neq j}^{p} \beta_k x_{ij}$ and $\rho_\tau(t) = \frac{1}{2}\{|t| + (2\tau - 1)t^2\}$ is the Huber weight function. This minimization procedure performs robust variable selection and shrinkage at the same time. The nonnegative regularization parameter λ determines the weight given to exclude variables which do not possess "explanatory" power over x^j i.e. as λ increases, the number of nonzero β_j components decreases. Therefore, the HQ procedure allows to identify a parsimonious subset of variables which are statistically relevant as predictors of x^j from the large data set V^j.

The model is estimated for a range of λ values, and for each variable we select the estimated models with maximum 3 determinants. Let us denote by ${D_1}^j$, ${D_2}^j$, and ${D_3}^j$ the sets of determinant variables for x^j, with one, two and three members, respectively.[4]

Step 3: Calculation of statistical relationship between the examined variable and its predictors using robust regression

After having identified subsets of variables which are linked statistically with x^j, we estimate the corresponding statistical models in a way which is robust to the presence of outliers. There is a number of robust regression methods, which identify statistical relationships between variables without being affected by the presence of outliers to the degree that this is the case for the standard OLS regression. These techniques differ from the standard OLS with respect to the function that is being minimised, which is not the sum of squared residuals but an alternative that is less sensitive to extreme values. Specifically, the LTS estimator (Rousseeuw 1984) is given by

$$\min_{\beta} \sum_{i=1}^{h} (r^2)_{i:n}$$

[4] See also Davidson and Tayi (2009) for an alternative approach.

where $(r^2)_{i:n}$ denotes the squared lower residual from the set which remains after the lower $i-1$ ones have been removed:$(r^2)_{1:n} \leq (r^2)_{2:n} \leq \cdots \leq (r^2)_{n:n}$. Consequently, the above expression contains the h lower residuals. The parameter h is critical for the outcome of the estimation; relatively high values of h could be 'permissive' allowing outliers to influence the results, while relatively low values of h could be excessively 'strict'. The first case entails the risk of 'masking' leading to Type II errors of the algorithm i.e. non-identification of existing outliers while the latter case could lead to 'swamping' i.e. Type I errors i.e. mislabeling of non-outlying data as outliers.

The computational complexity is also a significant factor when the method is intended to be applied to a large data set. The FAST-LTS algorithm proposed in Rousseeuw and van Driessen (1999) provides an efficient implementation of the LTS regression.[5]

Alternative robust regression methods tested in place of LTS are the MM and GS estimators. The MM type is based on the following optimization problem:

$$\underset{\beta = \{\beta_j\}}{min} \sum_{i=1}^{n} \rho_\tau({r_i}^2),$$

where $\rho_\tau(t)$ is a Tukey bi-weight function with an efficiency level of 95%, initialized by M-estimates of coefficients and residual scale at a breakdown point of 50%. The minimization problem is solved via reweighted least squares. MM-estimator is proved to have much higher efficiency than LTS given a specific breakdown point. For further details, we refer to Yohai (1987).

The GS type, instead, optimizes the residual scale based on the constraint that the sum of a bi-weight function of rescaled residuals equals a particular value. In symbols: $\min \sigma(\beta)$, given that the following equation holds:

$$\binom{n}{2}^{-1} \sum_{i<j} \rho\left(\frac{r_i - r_j}{\sigma(\beta)}\right) = \left(\binom{n}{2} - \binom{h_p}{2} + 1\right) / \binom{n}{2}$$

[5] The approach proposed in Zioutas et al. 2009 avoids the need to pre-define *h* but at an increased computational cost.

where $h_p = \left(\frac{n+p+1}{2}\right)$ and $\rho(t) = \min\left(\frac{3t^2}{c^2} - \frac{3t^4}{c^4} + \frac{t^6}{c^6}, 1\right)$ is the Tukey biweight error function with $c = 0.9958$. For further details, we refer to Croux et al. (1994).

As explained in Section 2, we compare the results of our proposed method to those of SPARSE-LTS (Alfons et al., 2013), which is based on the following minimization objective:

$$Q(H, \beta) = \sum_{i \in H} {r_i}^2 + \alpha n \sum_j |\beta_j|,$$

where H is the subsample of length αn, $\alpha \in [0.5,1]$, producing the minimum for Q respect to β. The optimal β is the sparse-LTS solution. The minimum is computed via a Sparse Fast-LTS algorithm, which computes at each step the LASSO estimate on the $100 \times (1 - \alpha)\%$ most concentrated observations. We remark that SPARSE-LTS does not need any variable selection step like our Step 2, because variable selection and robust model estimation are performed contemporaneously minimizing $Q(H, \beta)$.

Step 4: Robust estimation of residual dispersion

A robust calculation of the residuals dispersion should exclude the impact of outliers i.e. it should be conducted by first assigning binary weights to the observations, assigning zero weights to the observations which are deemed to be outliers from an initial dispersion estimate.

Specifically, the dispersion estimate is given by the standard consistent dispersion formula adjusted with binary weights which exclude extreme values:

$$\hat{\sigma} = \sqrt{\frac{(\sum_{i=1}^{n} w_i {\hat{r}_i}^2)}{\sum_{i=1}^{n} w_i - p^j}}$$

where the weights w_i will be defined below and p^j is the number of explanatory variables in the regression equation considered (i.e. 1, 2 or 3).

The weights w_i are defined by utilising an initial estimate of the dispersion:

$$\hat{\sigma}_0 = \sqrt{median_i({\hat{r}_i}^2)}/\phi^{-1}(0.75)$$

The above expression assumes that the residuals follow a normal distribution $N(0, \sigma^2)$ (where σ is the real value of the dispersion). Consequently, following Ruppert (2010, p. 118) this expression provides an estimation of the population dispersion. Rousseeuw and Leroy (2003, p. 202) propose in addition to multiply the second term with the finite-sample correction factor $[1+\frac{5}{n-p}]$.

Having calculated this initial dispersion, the weights can be defined at a pre-specified level e.g. 1%, considering our normality assumption for the residuals: $w_i = 1$ if $\left|\frac{\hat{r}_i}{\hat{\sigma}_0}\right| \leq 2.5$, and 0 otherwise.

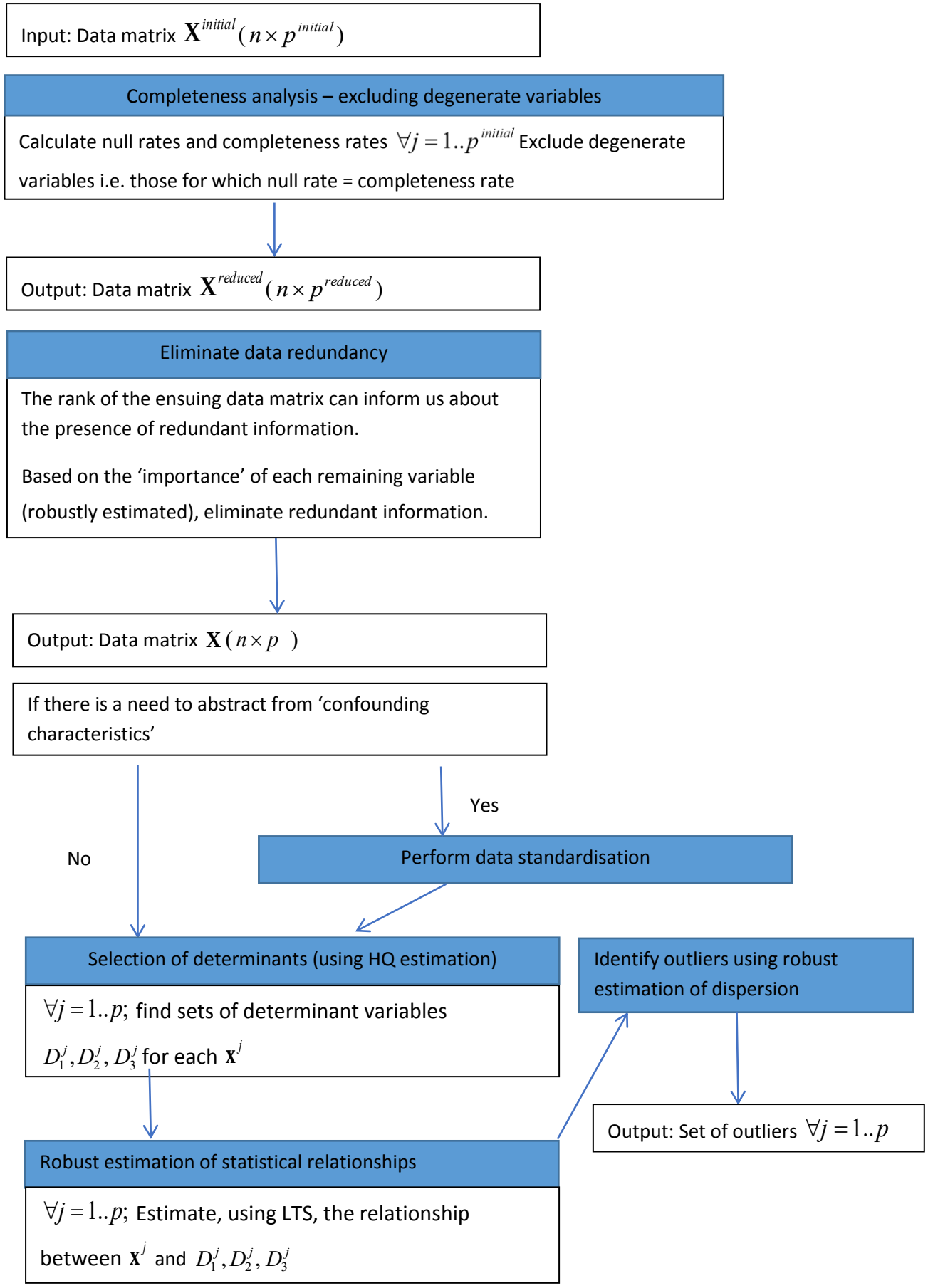
Alternative considered methods have different estimation procedures for residual dispersion. Concerning MM estimator, the i −th observation is flagged as a regression outlier if the estimated robustness weight $\hat{w}_i = 0$. Robustness weights vary from 0 to 1, and are exploited to estimate the residual scale. Concerning the GS type, robust Mahalanobis distances of the residuals are computed, observations exceeding $\sqrt{\chi^2_{L,1-\alpha/2}}$, $\alpha = 0.05$, are flagged as regression outliers, and the residual scale is robustly estimated accordingly. SPARSE-LTS, instead, identifies regression outliers by a reweighting step, which then provides final estimates of coefficients and residual scale, in a specular way respect to LTS.

Step 5: Identification of outliers based on the estimated dispersion

With the dispersion estimation $\hat{\sigma}$ in hand, the outliers can be specified at e.g. 1% level by the condition $\left|\frac{\hat{r}_i}{\hat{\sigma}}\right| \leq 2.5$. The underlying assumption in this step is that the set of residuals follow a normal distribution with $\hat{\sigma}$ as the standard deviation.

Chart 1 below illustrates graphically the different steps of the proposed algorithm, including the data preparation phase which is presented in detail in Appendix A.

Chart 1: Schematic representation of the proposed methodology



4. Simulation study

In this section we conduct a simulation study to test the validity of our method. Specifically, the three variants of our integrated approach, namely the versions utilising LTS, MM, and GS in the third step respectively (will be called HQ+LTS, HQ+MM, HQ+GS henceforth), are compared to each other and also to two benchmark methods. The first benchmark method is a robust 3-sigma rule where location and scale parameters are estimated in a robust way using logistic psi-functions.[6] The second benchmark is the SPARSE-LTS method.

Our generated $n \times p$ data matrix X contains p variables each with n observations, we assume that $p > n$ (i.e. "fat" data). The first column of X contains our response variable y of which we would like to identify outliers, utilising also the additional information contained in X.

It is assumed that some of the remaining variables in X provide supplementary information which can enhance the detection of outliers in y due to the existence of relevant statistical relationships. Therefore, a number of L determinants is randomly selected across the remaining $p - 1$ variables. We call L the set of "determinant" indices and X_L refers to the matrix which contains the relevant columns from X. A number of $\alpha \times n$ outliers is randomly generated across the n observations. We call O the generated set of outlier indices i.e. a set of indices in the set $1, \dots, n$ of observation indices. We call D the generated set of determinants indices and D' the generated set of non-determinants indices in the set $2, \dots, p$ of variable indices.

The simulation proceeds as follows. First, we generate the $p - 1$ covariates according to a multivariate normal distribution $MVN(0, \Sigma)$ with the null vector as mean and Σ, a matrix with unitary diagonal and all the off-diagonal elements equal to ρ, as covariance matrix. For the response variable y, i.e. the first column of X, we set a linear model without intercept where the covariates are only the L chosen determinants. In symbols:

$$y_i = X_{i,D}\beta + \varepsilon_i, i \notin O,$$

[6] Using the MATLAB functions *mloclogist* and *mscalelogist*.

where $X_{i,D}, i \notin O$, is the $1 \times L$ vector of the chosen determinants for observation i, β is a generated L-dimensional vector, $\beta_q \sim U(\beta_{min}, \beta_{max}), \forall q = 1, \dots, L$, and $\varepsilon_i \sim N(0,1)$, $\forall i = 1, \dots, n$.

Regression outliers in the response variable y are generated by multiplying the coefficients in β by an amplifying term m randomly drawn from the set (m_{min}, m_{max}):

$y_i = mX_{i,D}\beta + \varepsilon_i, i \in O$.In order to test for the presence of multicollinearity in the data, we allow for the possibility of the covariates to be correlated with a correlation coefficient ρ_X which is set to belong to $\{0.3, 0.7\}$. In order to test for the presence of sparsity in the data, we establish that a prescribed percentage γ_X, within the range $\{0.3, 0.7\}$, of the entries in X_D, (the matrix containing as rows all observations and as columns only the non-determinants) may be randomly set to 0. In addition, we establish that a prescribed percentage γ_{X_L}, within the range $\{0.3, 0.7\}$, of the rows in X_D (the matrix containing as columns only the determinants) may be randomly set to 0.

In the case where innovation outliers are allowed, the above equations become

$$y_i = X_{i,D}\beta + \varepsilon_i, i \notin O, \qquad \varepsilon_i \sim N(0,1)$$

and

$$y_i = mX_{i,D}\beta + \varepsilon_i, i \in O, \qquad \varepsilon_i \sim N(0, m).$$

Therefore, in the last equation, both regression and innovation outliers are present. We also distinguish the case in which the coefficients β are allowed also to be negative. In that case, the sign is decided by throwing a dice ξ before they are generated. Concretely, if ξ=1, then $\beta \sim U(5,15)$ while if ξ=0 then $\beta \sim U(-15, -5)$. We also set $m_{min,} = 1,\ m_{max} = 19$.

We focus on the following simulation settings:

- case 1: $p = 200, n = 100, L = 3, \alpha = 0.1, \rho_X = 0, \gamma_{X_L} = 0, \gamma_X = 0$, Presence of innovation outliers=NO, Mixed sign of coefficients=NO.
- case 2: $p = 200, n = 100, L = 3, \alpha = 0.1, \rho_X = 0, \gamma_{X_L} = 0, \gamma_X = 0,$ Presence of innovation outliers=YES, Mixed sign of coefficients=NO.
- case 3: $p = 100, n = 50, L = 3, \alpha = 0.1, \rho_X = 0, \gamma_{X_L} = 0, \gamma_X = 0.7,$ Presence of innovation outliers=YES, Mixed sign of coefficients=YES.

- case 4: $p = 100, n = 50, L = 3, \alpha = 0.1, \rho_X = 0, \gamma_{X_L} = 0.7, \gamma_X = 0.7,$ Presence of innovation outliers=YES, Mixed sign of coefficients=YES.
- case 5: $p = 200, n = 100, L = 3, \alpha = 0.1, \rho_X = 0.3, \gamma_{X_L} = 0, \gamma_X = 0.7,$ Presence of innovation outliers=YES, Mixed sign of coefficients=NO.
- case 6: $p = 200, n = 100, L = 3, \alpha = 0.1, \rho_X = 0.3, \gamma_{X_L} = 0, \gamma_X = 0.7,$ Presence of innovation outliers=YES, Mixed sign of coefficients=YES.

For each setting, a number of $N = 100$ replicates have been generated. For each simulation set-up, we compute the following performance measures:

- masking rate, which is the rate of non-identified outliers over the true ones (type I errors);
- swamping rate, which is the rate of erroneously identified outliers over the recovered ones (type II errors);
- aggregate error rate, which is the mean between masking and swamping rates.

All measures range from 0 to 1.

Table 1: masking, swamping and aggregate rates for cases 1-6 and each performed method ($m = 19$). Numbers in bold show the lower values for each column.

	Case 1			Case 2			Case 3		
	Masking	Swamping	Aggregate	Masking	Swamping	Aggregate	Masking	Swamping	Aggregate
HQ+LTS	**0.01**	0.10	0.05	**0.14**	0.10	0.12	**0.17**	0.13	0.15
HQ+GS	0.05	0.02	0.04	0.30	**0.03**	0.16	0.26	**0.03**	**0.15**
HQ+MM	0.01	0.06	0.04	0.16	0.08	**0.12**	0.18	0.13	0.16
SPARSE-LTS	0.01	0.02	**0.02**	0.29	0.06	0.18	0.28	0.10	0.19
Logistic	0.15	**0.01**	0.08	0.99	0.93	0.96	0.96	0.85	0.91

	Case 4			Case 5			Case 6		
	Masking	Swamping	Aggregate	Masking	Swamping	Aggregate	Masking	Swamping	Aggregate
HQ+LTS	**0.02**	0.10	0.06	**0.10**	0.10	0.10	**0.12**	0.11	0.11
HQ+GS				0.23	**0.01**	0.12	0.26	**0.03**	0.15
HQ+MM	**0.02**	0.11	0.06	0.11	0.06	**0.08**	0.12	0.08	**0.10**
SPARSE-LTS	**0.02**	**0.02**	**0.02**	0.19	0.03	0.11	0.23	0.05	0.14
Logistic	0.03	0.61	0.32	0.98	0.79	0.88	0.93	0.57	0.75

We report in Table 1, masking, swamping and aggregate rates for each of the six cases and for each method. As it can be observed, the HQ+LTS method is the best performer with respect to the masking rate for all cases. Furthermore, HQ+GS performs best with respect to the swamping rate for all cases but Case 4 among regression methods. At the same time, HQ+MM is often prevailing with respect to the aggregate rate, even if relevant exceptions are Cases 1 and 4, when it is outperformed by

SPARSE-LTS. The observed pattern is that for complex cases our integrated approach outperforms the benhchmark methods, while the benchmark method SPARSE-LTS is the best performer when only regression outliers are present and when the number of zero observations is very high.

We present the results in more detail below, showing how the two rates, relevant for assessing the performance of the outlier detection methods, change while the disturbance parameter $\boldsymbol{m}$ increases. The logistic method performs clearly worse than all other methods both with respect to swamping and masking rates, therefore we do not include these results below.

In Figure 1 we report the masking rate for case 1. All variants of our method perform very well with HQ+GS performing slightly worse. The swamping rate for case 1 is reported in Figure 2. We see that among regression methods SPARSE-LTS and HQ+GS exhibit the best results regarding the swamping rate, which is in both cases lower than 5%. On the contrary, HQ+LTS and HQ+MM are characterised by larger rates. Therefore, if there are no innovation outliers, SPARSE-LTS has a distinct advantage. We remark that variable selection works perfectly for both methods (HQ and SPARSE-LTS), while the logistic method converges slowly to a masking rate of 15% and a swamping rate smaller than 2% as the disturbance coefficient m increases. Therefore, the benchmark method SPARSE-LTS proves to be the superior one in this simple case, showing an aggregate error rate smaller than 2%.

Figure 1: Masking rate for case 1 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.

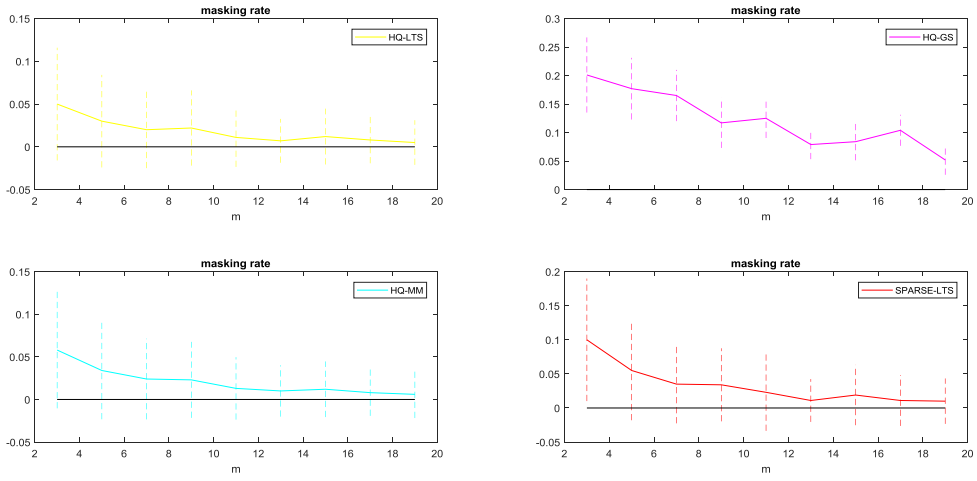
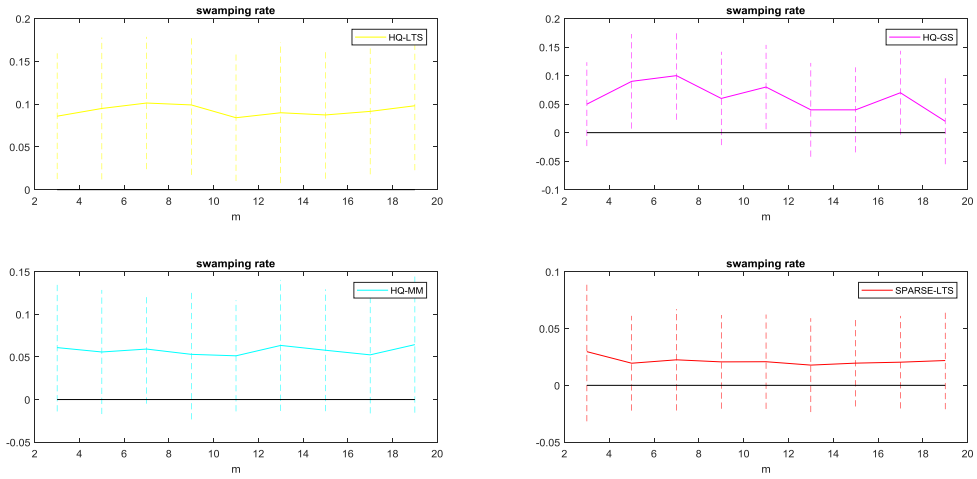


Figure 2: Swamping rate for case 1 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.



Case 2 introduces innovation outliers and the results as presented in Figures 3 and 4 show the corresponding masking and swamping rates. Convergence is slower than before and now SPARSE-LTS is not the best-performing method anymore, neither with respect to the masking nor and the swamping rate. Instead HQ+LTS performs best with respect to the masking rate criterion, with a 15% masking rate, followed by HQ+MM. From the swamping rate side, HQ+GS exhibit the better performance, while HQ+LTS shows the worst at 10%. Covariate recovery is almost perfect, while the logistic method goes completely wrong, with a masking rate close to 100%. Therefore, in this case both the HQ+MM and the HQ+LTS variants of our proposed method seem optimal if the aggregate measure is considered.

Figure 3: Masking rate for case 2 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.

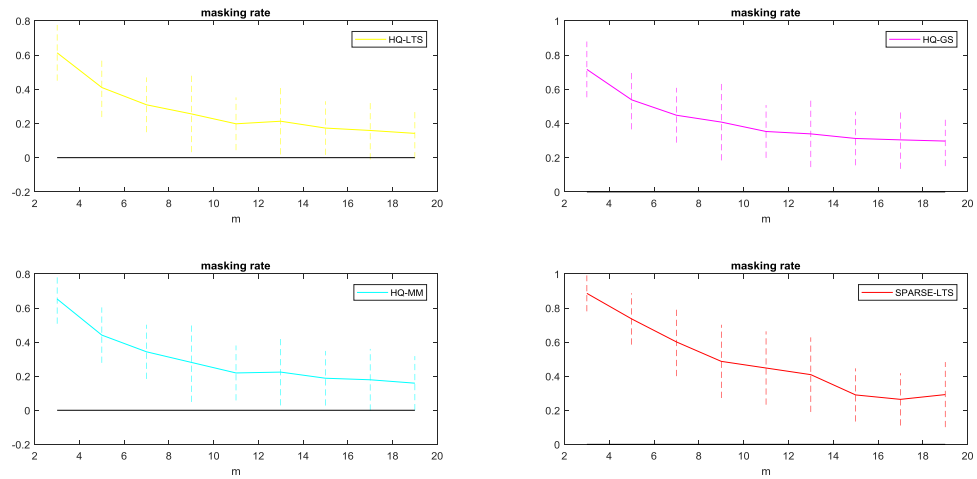
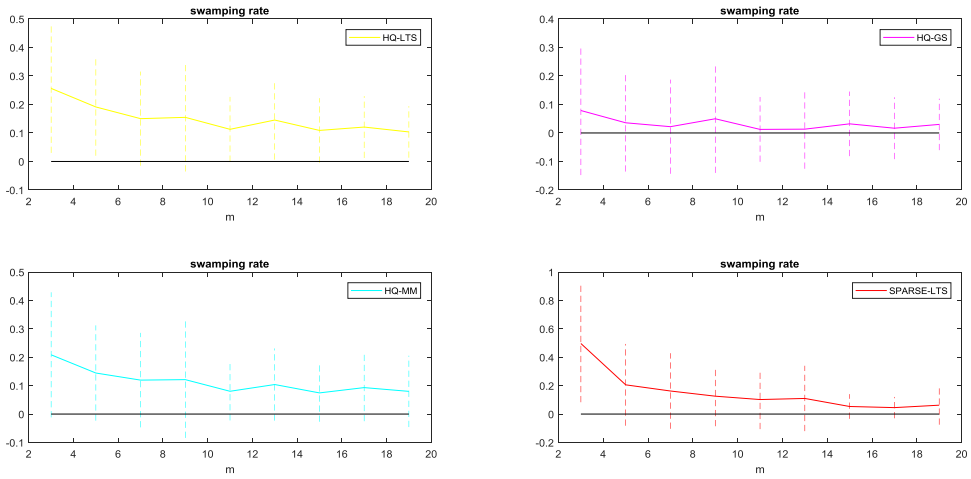
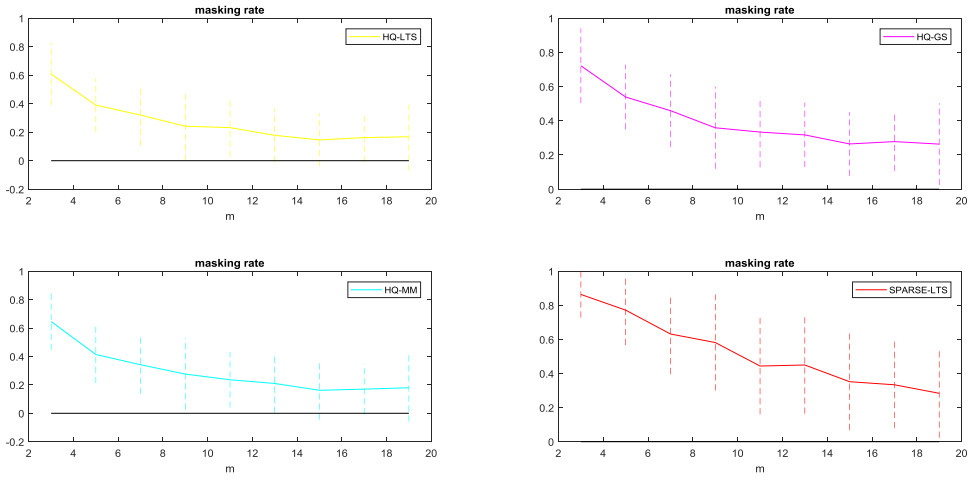


Figure 4: Swamping rate for case 2 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.



When sparsity is introduced (case 3), Figures 5 and 6 show that both HQ+LTS and HQ+MM feature a masking rate below 20% and a swamping rate slightly above 10%. HQ+GS exhibits a more uneven performance, with a higher masking rate and a very low swamping rate. SPARSE-LTS shows the worst performance among regression methods, with a masking rate around 30% and a swamping rate around 10%. The other benchmark method fails completely, exhibiting masking rates above 90%.

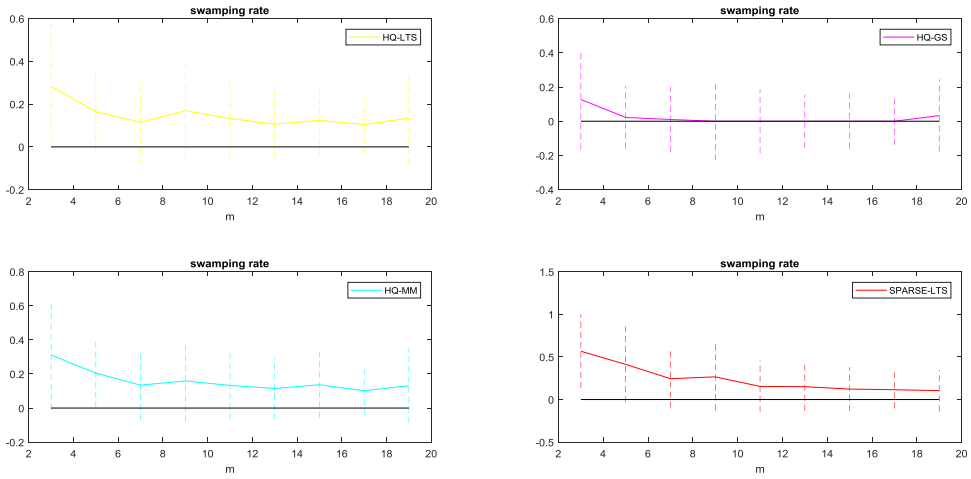
Figure 5: Masking rate for case 3 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.



The average number of masked or swamped determinants is never above 0.1 both for HQ and SPARSE-LTS. Overall, all three variants of our proposed approach are close with respect to the aggregate performance and the optimal choice is determined whether the user is more 'risk averse' (i.e. aims to minimise the masking rate) in which case HQ+LTS or HQ+MM would be the preferred

choices or more focused to avoid over-identification of outliers in which case HQ+GS would be the preferred choice.

Figure 6: Swamping rate for case 3 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.



Case 4 is designed to test how the presence of missing rows in the matrix of determinants affects outlier detection. In that case, the GS method becomes completely infeasible, because the linear system behind the procedure has not a unique solution. The reason is that the condition number of the sample covariance matrix is close to infinity. On the contrary, the other regression methods are robust to such perturbation, still showing satisfactory results. Specifically, Figures 7 and 8 show that the masking rate is for all methods close to 0.

Figure 7: Masking rate for case 4 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.

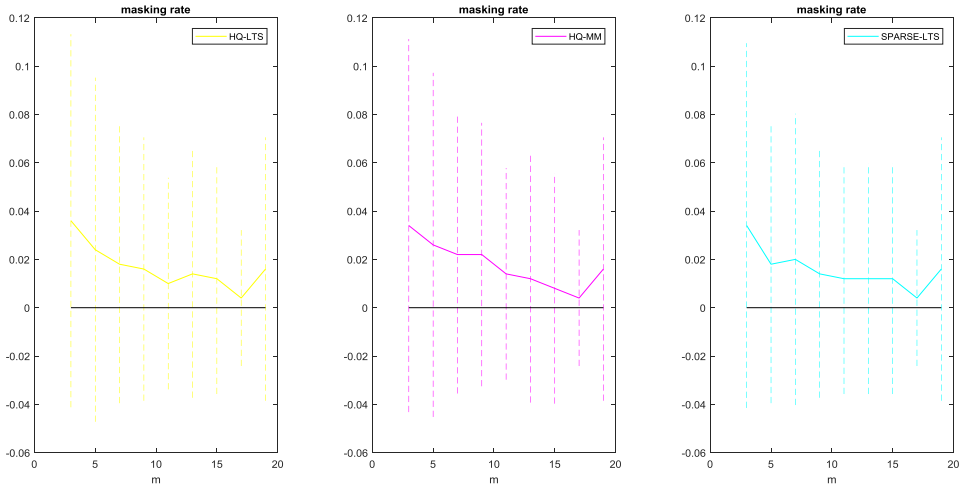
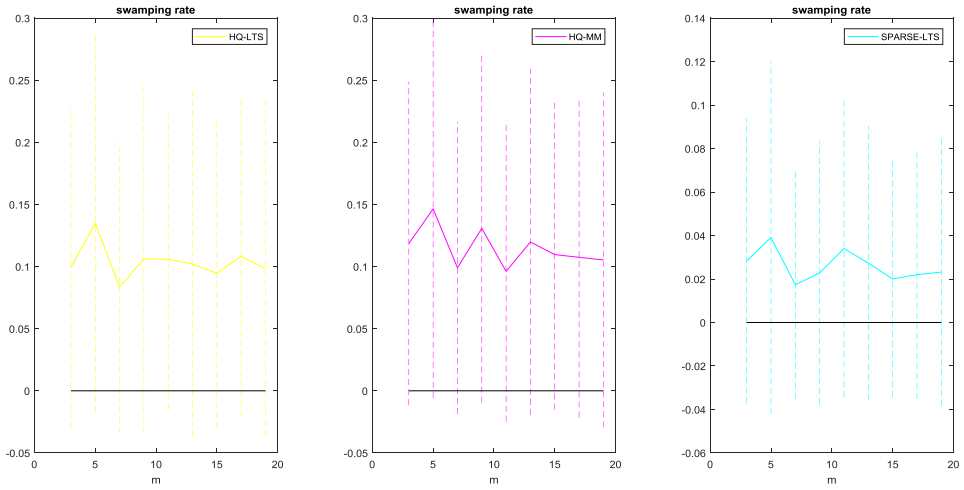


Figure 8: Swamping rate for case 4 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.



The swamping rate is significantly better for SPARSE-LTS, standing at 2% against the 10% of HQ+LTS and the 11% of HQ+MM. The same holds for the aggregate rate, which is equal to 2% for SPARSE-LTS and 6% for HQ+LTS and HQ+MM. Variable selection is perfect for SPARSE-LTS, while HQ shows an average number of masked determinants lower than 0.1. The logistic method converges to a masking rate of 2% and a swamping rate of 60%, thus being completely ineffective. If we combine a zeros rate of 0.7 and a correlation level of 0.3 among the covariates (case 5), HQ+LTS and HQ+MM also converge to the lowest masking rates, in both cases around 10%, while HQ+GS is again higher at 23% (Figure 9). However HQ+GS exhibits the lower degree of swamping at 1%, while for HQ+MM swamping is around 6% and for HQ+LTS around 10% (Figure 10). SPARSE-LTS shows a swamping rate around 3% and a masking rate around 20%. On aggregate, HQ+MM shows the best overall error rate at around 8%, followed by HQ+LTS, SPARSE-LTS and HQ+GS by a small margin. Variable selection is still almost perfect for both methods (HQ and SPARSE-LTS).

Figure 9: Masking rate for case 5 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.

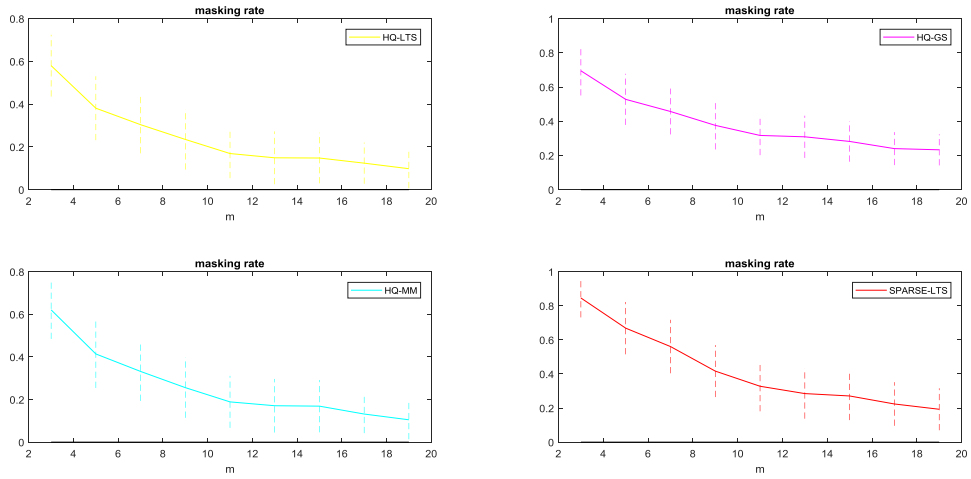
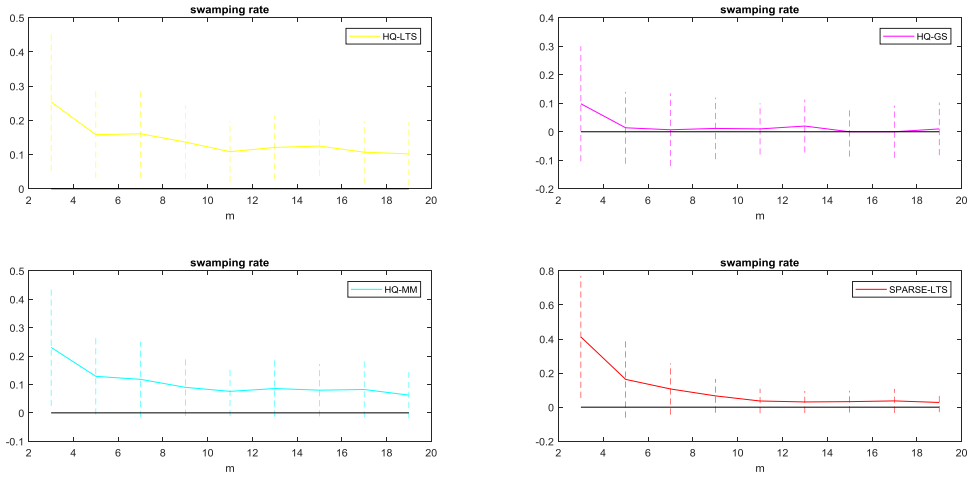


Figure 10: Swamping rate for case 5 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.



A similar pattern is shown if coefficients are allowed to have different signs (case 6), as exemplified in Figures 11 and 12. Therefore, for cases 5 and 6 it turns out that HQ+MM is the optimal option if one aims to minimise both the swamping and the masking rate with an equal weight, while HQ+MM outperforms the other methods from the masking rate side and HQ+GS from the swamping rate side.

Figure 11: Masking rate for case 5 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.

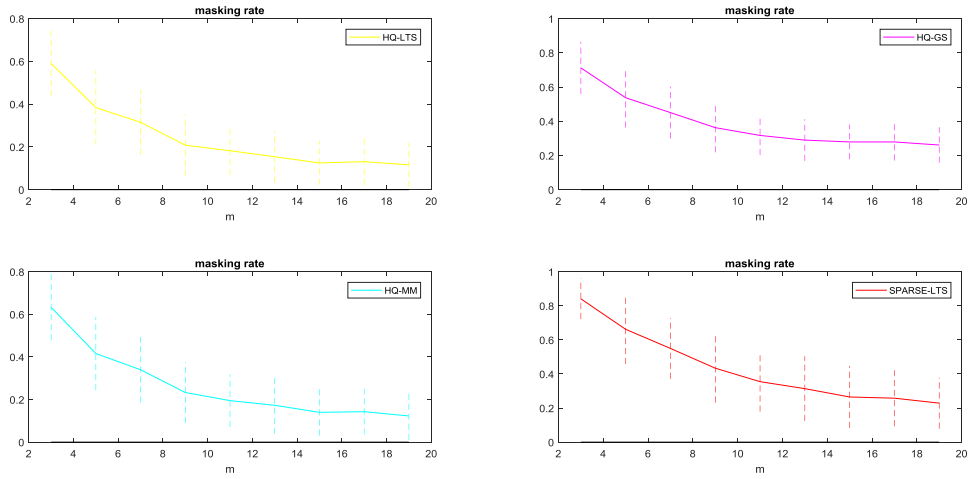
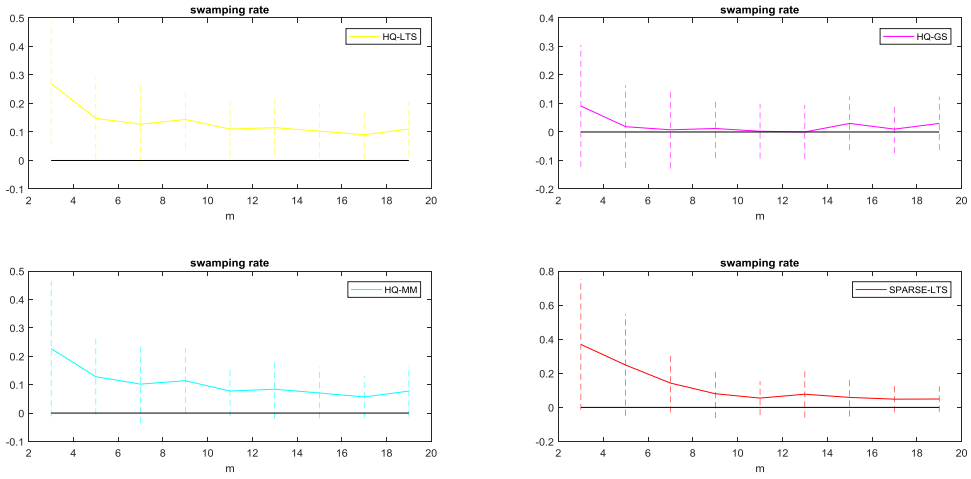


Figure 12: Swamping rate for case 6 (constant m in the x-axis). Dashed lines show the one standard deviation intervals.



To sum up, we can conclude that the performance of SPARSE-LTS gets worse as the degree of sparsity and collinearity increases. As a consequence, for real data with a lot of zero entries or collinear variables, HQ+LTS and HQ+MM represent the best approaches, because they ensure a low masking rate. In addition, in case of missing rows in the matrix of determinants, GS regression becomes infeasible, while HQ+LTS shows also a lower swamping rate than HQ+MM. Furthermore, it is observed that the performance of SPARSE-LTS improves considerably as sparsity, collinearity and missing rows are present into the matrix of determinants. This occurs because in that case the impact of innovation outliers is minimized.

Note that if the fraction of outliers α increases, masking and swamping effects are amplified accordingly without affecting the relative performance of competing methods. If both p and n are increased, the performance is quite similar, and the detected patterns remain unaltered, given that the ratio $\frac{p}{n}$ remains the same. We remark that the computational cost of SPARSE-LTS increases considerably, because it requires $O(n^2)$ iterations due to the cost of sorting squared residuals (Rousseeuw and van Driessen, 1999). On the contrary, the HQ procedure only requires $O(pn)$ iterations (Yi and Huang, 2016), thus being much faster as both p and n are in the order of 10^3- What is more, as p increases it is not easy to select a model with a prescribed number of determinants via SPARSE-LTS, because the trimmed sum of squares is very sensitive to the value of λ. On the contrary, the HQ method is less affected by this drawback, due to the use of the Huber loss function.

5. An application to supervisory banking data

The methodology is applied using real data, specifically to the supervisory data submitted by the European banks to the Single Supervisory Mechanism (SSM) operating within the European Union. The broader context is the institutional changes at the European level, brought about by the financial crisis leading to the harmonised supervision of large banks in Europe and the centralised collection of their data by the European Central Bank (ECB) and specifically its supervisory function.

More broadly, during the last years, economic policy institutions, like international organisations, finance ministries or central banks face the challenge to incorporate the newly emerging large data sets into their mode of operation.[7] The challenges pertain both to the conceptual and the technical dimensions. On the conceptual aspect, the sheer amount of information contained in large data sets poses methodological questions such as the selection between theoretically grounded or a-theoretical, data-driven approaches. A report from the Bank of International Settlements notes that "*conventional*

[7] For an overview, focusing on central banks, see Hammer, Kostroch, Quiros et al. (2017) and Nymand-Andersen (2016). The workshops on big data organised by the European Central Bank in 2014 (https://www.ecb.europa.eu/pub/conferences/html/20140407_workshop_on_using_big_data.en.html) , the Bank of England (see Bholat 2015 for a synopsis), and the Riksbank in 2015 (http://www.riksbank.se/en/Press-and-published/Notices/2015/The-Riksbank-organises-a-workshop-on-big-data/) testify for this increasing interest in the central banking community.

structured data sources" appear to be … more effectively mobilised than "new" big data sources" highlighting the methodological exploration which takes place with respect to analysing such datasets (Bank of International Settlements 2015). Moreover, on the technical front, the standard econometric techniques for analysing data need to be extended, or new ones should be adopted, borrowed from other data-rich fields, dictated again by the size of such data sets. See Varian (2014) for an overview of techniques suited for big datasets.

At the European level, the assumption of the supervision of euro area banks by the ECB, starting from 2014, has as one of its implications, the collection of large sets of supervisory data from a number of countries within one institution, a feature which is unique at a world scale. The supervisory data, reported by the banks to the ECB, cover, with a high degree of granularity, the different activities undertaken by the banks. A number of breakdowns are present in the data e.g. with respect to counterparties, products, geographical areas, accounting portfolios, risk classes and types of risk. This exceptional granularity is the distinctive feature of this particular dataset and is also present in big datasets collected in other finance-related fields, for example in banking sector's credit registry datasets (where detailed profiles for each borrower are recorded) or behavioural data utilised by insurers (where various dimensions of behaviour are aggregated, aiming to provide insights on expected insurance claims) or finally data from over-the-counter derivatives (with many-attribute non-standardised contracts).

Monitoring the quality of these supervisory data, from a statistical perspective (as opposed to a compliance or regulatory perspective) is a task with critical implications for the quality of supervision. Improved data quality enhances the reliability of analyses and supervisory decisions which are informed using these data. This is a challenging task due to the sheer amount of these data and the substantial heterogeneity contained therein. [8]

Our sample consists of a cross section of 365 SSM banks. A subset of these institutions, specifically 124 of them, have been labelled as 'Significant Institutions' in the context of the SSM i.e. they

[8] Studies investigating the biases caused by outliers on the measurement of financial risk include Chatterjee and Jacques (1994) and Grane and Veiga (2014). Seaver and Triantis (1995), Johnson and McGinnis (2008) and Bellini (2012) investigate the impact of outlier points on technical efficiency measurement.

possess systemic importance with respect to the European or their domestic banking sector i.e. based on the consolidated group size or on their size at the individual SSM country level. The reference date of the data is 2014 Q4.[9]

We analyse a selected set of FINREP templates which contains all basic breakdowns of the banks' balance sheet, therefore providing very granular information on banks' activities. 'FINREP' stands for 'Financial Reporting' and includes data on the prudential scope of consolidation (as opposed to the accounting scope of consolidation) using IFRS accounting concepts. FINREP templates were initially developed as a part of a guideline from the Committee of European Banking Supervisors (CEBS) in 2005. During the following years, the European Banking Authority (EBA) which superseded CEBS continued to coordinate the development of the templates. In the context of the SSM, FINREP data have a much more elevated status since they form part of the mandatory reporting framework, given that there is a relevant legal requirement incorporated into European Law with the Capital Requirement Regulation (CRR), No 575/2013, Article 99 on 26 June 2013.

The templates can be found at the EBA website.[10] Henceforth, we will refer to each data point as F *x.y* [r*z*,c*w*] where *x*, *y*, *z*, and *w* refer to the sheet numbering (the first two), row numbering and column numbering, respectively e.g. F 1.1 [r380,c10] refers to the data point 'Total assets'.

The selected templates contain information on the breakdown of the balance sheet across at least four dimensions, namely accounting portfolios, instruments (e.g. debt, equity, derivatives), products (e.g. types of loan products) and counterparties (e.g. governments, central banks, other financial institutions, non-financial corporations). This granularity is present for both the asset and the liability sides. The main accounting categories for assets which are included are as follows: i) Loans and receivables & held-to-maturity assets (Loans and HtM); ii) Assets held for trading (HfT); iii) Assets at Fair value through profit or loss (FVPL); iv) Available-for-sale assets (AfS). These categories differ

[9] The reference date is the first one in which FINREP was submitted, therefore a change in completion rates and other characteristics of the submitted data may be present in subsequent submissions.
[10] The URL address is: https://www.eba.europa.eu/documents/10180/359626/Annex+III+-+FINREP+templates+IFRS.xlsx/049e48a4-e7c2-44c6-89b1-4086447bced9

with respect to their valuation and to the extent that they affect the profits of the bank. Both liabilities valued at amortised cost and FV liabilities are included.

Specifically, our input set consists of the following templates (in parenthesis their main contents): F 1.1 (overview of assets), F 4.1 (HfT assets), F 4.2 (FVPL assets), F 4.3 (AfS assets), F 4.4 (loans and HtM assets), F 5.0 (loans and advances by product e.g. on demand, credit card, leases loans etc), F 8.1 (overview of liabilities), F 9.1 (off-balance sheet items e.g. loan commitments and guarantees), F 10.0 (derivatives – trading), F 11.1 (derivatives – hedge accounting).

The practical application of the multivariate outlier detection has to take into account two additional issues: the data availability issue (missing data and differential completeness rates among variables) and the existence of a number of (mainly linear) identities among the variables of the data sets. Both these features are strongly present in our data set. These practical issues are generic enough to be of wider interest for other applications, therefore their treatment is elaborated in the Appendix of our paper.

5.1 Results and discussion

In this section, we present the results of our approach with respect to detecting outliers on the banks' size (measured as the value of total assets). The five-step procedure that was formulated in Section 2.3 is applied to the data set of 433 variables which remain after the exclusion of degenerate and highly correlated variables (see the Appendix for the details). The resulting data matrix has 77% of missing data, while the mean absolute correlation coefficient is 0.051.

The histogram below shows the distribution of the log of banks' size, expressed in millions. It transpires that the median value of banks' size is 11.3bn (the mean is 10.2bn) and the log of size follows a distribution which is close to normal (as can be inferred from the p-value of the Jarque-Bera test, which equals 0.43 and allows the non-rejection of the null hypothesis of a normal distribution). The fact that the size follows a log normal distribution is intuitive, given the existence of a large

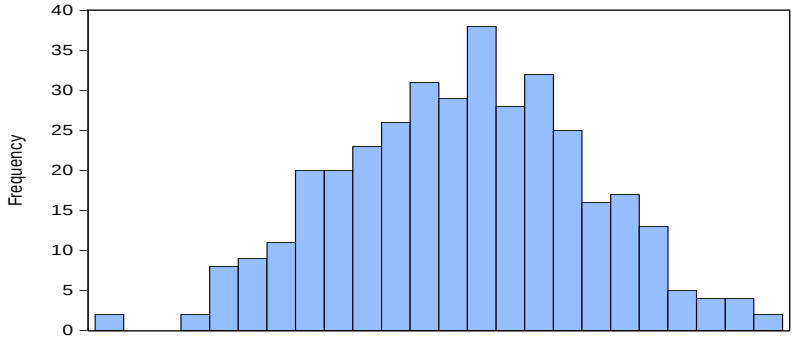
number of small- and medium-sized banks, high variances (also reflecting the size dispersion of the countries in the sample e.g. with respect to their GDP) and the non-negativity of the size variable.

Figure 13 allows us to spot immediately two outliers, at the left tail of the distribution, which imply that the banks report total assets as equal to 1.8mn and 2.0mn respectively (two implausibly low figures). Therefore, this first level check was able to identify two outliers, caused probably by the use of a wrong unit e.g. values expressed in million euros, instead of euros as required. We now proceed with second level checks for outliers based on the results of our proposed multivariate outlier detection approach.

We perform HQ+LTS on our data matrix, following the assessment of the relative merits of the different variants of our proposed method, as was presented in Section 3. We set $L = 3$. The matrix of recovered determinants presents a mean absolute correlation equal to 0.67, a degree of sparsity of 35% and a percentage of missing rows of 20%.

The procedure finds (see Table 2) that the optimal determinants of the (normalized by the maximum value) size variable is the (normalized by each bank's assets) amortised cost of debt securities issued, the carrying amount reported for hedging derivatives at the asset side, and the notional amount of total hedging derivatives.[11] This is an intuitive result since these variables contain basic information about the use of debt securities and derivatives and these instruments is more widely used by large banks.

Figure 13: Histogram of the size of banks in logs (there is no x-axis values to ensure data anonymization).



Source: Supervisory data, authors' calculations

[11] i.e. data points F 10.0 [r10,c10], F 10.0 [r320, c30] and F 11.0 [r500,c30].

Table 2: Estimated equation linking banks' size with the three top covariates.

Dependent variable: **log-size**	Estimated coefficients
Constant	18.4521 ***
	(156.418)
Hedge Derivatives	104.9235 ***
(Carrying amount)	(5.522)
Assets	
Debt securities	6.1278 ***
(Amortised cost)	(5.548)
Hedge derivatives	3.7063 ***
(Notional amount)	(5.391)
Total hedging	

R^2=0.4644

Note: The table reports the Least Trimmed squares estimation results. The t-statistic is reported in parenthesis. The asterisks *,** and *** indicate statistical significance at 10%, 5% and 1% respectively.

Using the fitted values of the estimated equation we identify the outliers, following the Step 5 of Section 2.3. The two outliers which were identified by simply plotting the histogram of values are still identified i.e. outliers with a clear innovation outlier component (using the terminology of Section 3). In addition, based on the estimated relationship between size and its covariates, we also identify 14 additional outliers.

Figure 14 provides a comparison between the histograms of log-size in the non-outlier and the outlier set. It can thus be observed that detected outliers are large-sized banks: among those 14, only 3 have total assets below the general mean over all the 365 banks.

Figure 14: Histograms of log-size in the non-outlier (left) and the outlier set (right).

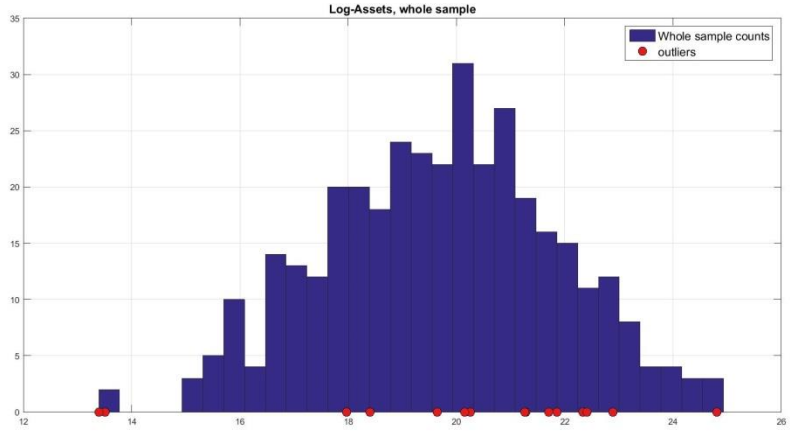


Figure 15: Histogram of the estimated standardized residuals in the outlier set

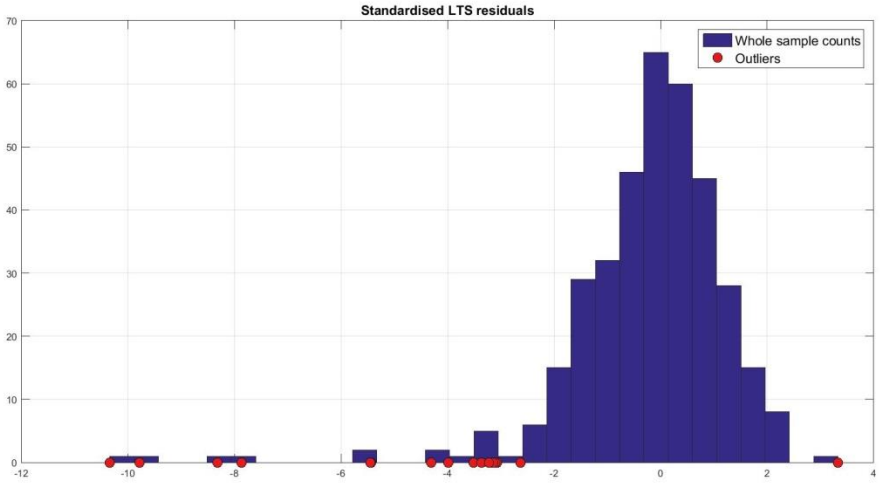


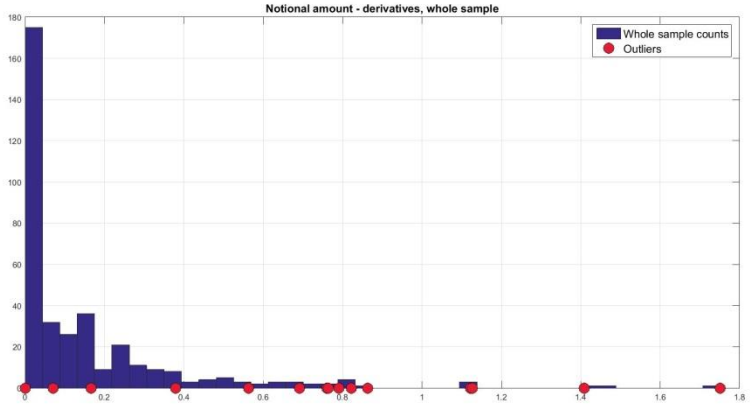
Figure 15 provides the histogram of the estimated standardized residuals in the outlier set, providing information on the size of each bank relative to what would be expected by the size determinants. The only bank with a positive residual presents a very large value of total assets compared to the relatively small values of the three covariates, or in other words, does not utilise the instruments represented in the set of covariates to the extent that would be expected by its size. On the contrary, eleven of the remaining banks show a value for notional amounts of hedge derivatives larger than 0.5, and seven of them show a value larger than 0.5 for debt securities at amortised cost. Six banks possess a relatively high amount of debt securities and notional amounts of hedge derivatives. Four banks even show a notional amount of hedge derivatives larger than 1. In general, the 13 banks with a negative residual contain more hedging derivatives than expected by their total assets.

When the HQ+MM method is used, 19 outliers are identified, instead of 16 as with the HQ+LTS method. This difference in the number of identified outliers is consistent with our simulation study of Section 3 and the generally higher swamping rates characteristic of HQ+MM compared to the HQ+LTS when sparsity increases (as in case 4 of Table 1). Importantly, the outliers are in almost all cases (except from one) common among the two methods.

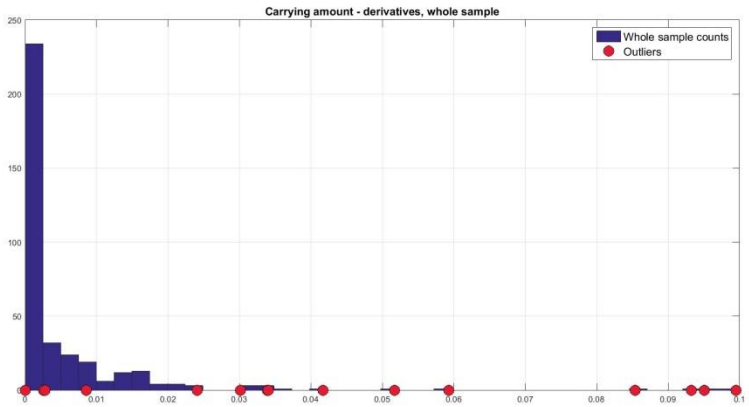
The insights offered by our proposed methodology are illustrated further by comparing the histograms of recovered covariates for non-outlier banks and for the set of outliers comparatively (see Figure 16). This comparison is used to delineate the nature of the identified outliers. It is clear that the outlier set contains extreme values for all the determinants, showing completely different distributions compared to the whole sample. The values are much higher with respect to those expected from the sample distributions, leading to the negative residuals presented in Figure 15. The outlier discrimination capability of the proposed procedure, with respect to uncovering unusual relationships between log-size and the identified set of determinants, is well illustrated via these graphs.

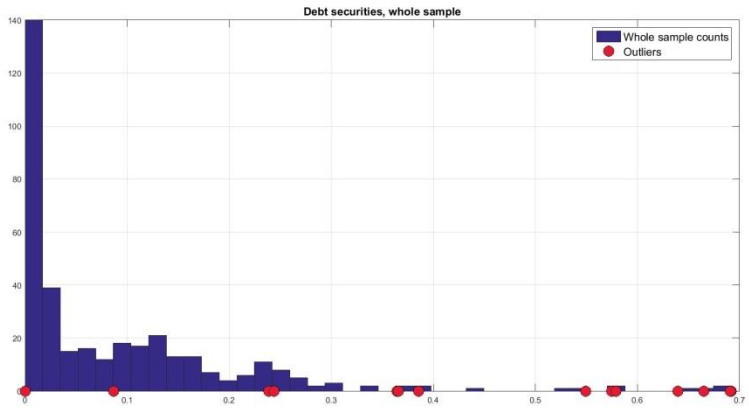
Overall, it is important to stress that the proposed outlier detection technique generates useful insights about the nature of the identified outliers: the analysis is able to offer a richer explanation why the specific value represents an outlier, offering value added to the data quality analysis.[12]

Figure 16: Notional and carrying amounts of hedge derivatives, debt securities at amortised cost: histograms comparing the distributions of non-outlier and outlier banks.



[12] One could investigate further whether the reported values are indeed justifiable, maybe due to idiosyncratic features of their strategy, or they represent simply wrong data. A statistical check for data quality can be formulated starting from our procedure, aimed at testing data reliability.





6. Conclusion

In this paper, we address the problem of outlier detection in a high-dimensional context by presenting an integrated, automatized procedure involving a preparatory phase and a multi-step outlier detection algorithm. The procedure is suitable for application to high-dimensional datasets, i.e. datasets with a very large number of variables compared to the number of observations for each variable. In addition, the proposed procedure can be easily adapted for the purpose of pattern recognition in high-dimensional datasets.

The proposed integrated outlier detection algorithm is a five-step procedure, which incorporates, data standardisation, data selection (via the heuristics of Yi and Huang (2016)), estimation of a robust regression model based on the selected variables (via the Least Trimmed Regression), robust calculation of dispersion and finally a criterion to identify outliers. The methodological advance lies in the effective combination of these techniques which is shown through a simulation study to be

particularly useful for collinear and sparse “fat” data, performing better than some robust alternatives with respect to minimising both swamping and masking rates.

The above procedure is then applied to a supervisory dataset, consisting of balance sheet data, which are submitted by the Euro Area banks in the context of the ECB Banking Supervision. Algebraic data redundancy measures, such as the rank of the correlation matrix, point to the need to address the dense correlation structure present in the raw data set. In addition, it is found that the high number of accounting identities leads to a concentration of pairwise correlation values close to 1. As a result, the data redundancy procedure led to the removal of a significant amount of variables from the initial dataset. Application of the determinants’ identification regression to the resulting dataset finds that variables related to the use of debt securities and derivatives represent the most closely related subset of variables for explaining the size variable. Consequently, the outlier criterion identified outlier size values by detecting the banks showing unexpected relationships between size and derivatives-based variables. This test case shows that this type of outlier detection analysis is a promising way to both enhancing the quality monitoring and identifying patterns in high-dimensional data sets.

An important aspect to be considered is the interdependence of the various steps. The interaction of the different steps is important in similar decision making and signal processing algorithms. Further work exploring the possibility to develop a one-step method for regression outlier detection in presence of missing data, collinearity and outlying variances would be of great interest.

Appendix A: Data preparation

Data completeness and data redundancy

Each of the ten (10) FINREP templates of our input data set is essentially a large matrix with its rows corresponding to different banks and the columns to different variables. The first step of our statistical analysis is the aggregation of the individual FINREP templates, containing all the banks in rows and all the variables in consecutive columns. A matching step of the banks in different templates to align the rows (banks) in all templates is performed before applying the outlier detection techniques.

This section will describe the data completeness analysis per template, and the approach followed. The results of this step possess an interest of their own because the differential data completeness per template and per data point reflect the significance of the respective templates and data points and provide insights for a business analysis; therefore, it is a procedure which provides value added for data sets with a similar structure. Subsequently, the data redundancy reduction step has been performed in the aggregated matrix.

We now present the completeness analysis and the procedure we follow to exclude redundant information from our data set. Each template k is a matrix containing $p(k)$ columns (variables) and $n(k)$ rows (banks). The dimensions of each template are presented in first two rows of Table A-1. As it is apparent the different 'blocks' (i.e. templates) of our data set seem to be heterogeneous with respect to their dimensions. In addition, heterogeneity is present with respect to the submission rate of each template by the different banks; some templates have been submitted by fewer banks, reflecting that their contents are not as relevant for the whole set of banks under examination. Specifically, the number of banks across templates ranges from 207 to 365. As expected, the templates F 1. 1, F 4.4, and F 8.1 which present the basic items for each bank (asset composition, loans and liabilities) are reported from all banks. On the other hand, the template F 4.2 containing assets assessed using "fair value via profit and loss" (FVPL) is reported from only 207 banks. The number of numeric variables for a single template ranges from 31 to 215. The overall number of variables is 1039.

The issue of data completeness is critical, especially for data sets with a large cross-sectional dimension in which heterogeneity is expected regarding the availability of information for the different entities. In our case, we take into account the convention underpinning the submission of supervisory data in the SSM context, namely that banks are required to report all required data (e.g. in contrast to household surveys where the different cross-sectional units are not usually required to respond to all questions) and interpret all missing data as zeros i.e. that the bank does not have any activity in the respective category.

Table A-1: Completeness statistics per template

Template	(1) F 1.1	(2) F 4.1	(3) F 4.2	(4) F 4.3	(5) F 4.4	(6) F 5.0	(7) F 8.1	(8) F 9.1	(9) F 10.0	(10) F 11.1
Number of banks	365	254	207	347	365	362	365	357	329	272
Number of variables	58	31	33	76	168	60	215	24	177	197
Mean completion rates	50%	38%	39%	42%	35%	59%	21%	67%	29%	26%
Null rate	6%	14%	18%	11%	11%	8%	8%	7%	7%	14%
Zero rate	45%	50%	54%	49%	60%	22%	74%	16%	58%	76%
Number of degenerate points	20	1	1	6	33	0	89	0	54	64
Percentage of degenerate points	34%	3%	3%	7%	19%	0%	41%	0%	30%	32%
Percentage of banks with no missing values	1%	20%	48%	8%	6%	9%	1%	10%	1%	12%
Final number of variables, after excluding degenerate points	38	30	32	70	135	60	126	24	123	133
Rank of the correlation matrix	29	27	26	46	76	54	90	21	97	80

Source: Supervisory data, authors' calculations

The data completeness analysis involves the computation of descriptive statistics both with respect to variables and banks. Specifically, we compute for each variable the completeness rate, defined as the percentage of non-missing values respect to the number of banks, and the percentage of non-zeros both respect to the number of non-missing values and to the overall number of banks. Moreover, we compute for each template the proportion of banks having complete records (i.e. no missing values).

We perform checks on missing data, both row-wise and column-wise. Firstly, for each variable we compute the number of missing and non-missing values and the completion rate is derived as the ratio between the non-missing values and $n(k)$ (see third row of Table A-1).

It is noted that the higher degree of completeness is found in template 8, a template containing the asset side balance sheet items, followed by the loans and receivables items (template 5) and the asset breakdown overview (template 1). Templates 1 and 5 contain the most fundamental items of the asset side in banks' balance sheet, therefore it is expected that the majority of banks will report most of these items. In addition, template 8 is the smallest one as regards its size (only 24 variables) and the off-balance sheet contains items common to many banks, hence its high completeness rate. On the other hand, the lowest degree of completeness is found in the liability breakdown (template 7). The overall mean completeness rate across all data-points is 33.34%. The minimum is 2.43%, while the maximum completeness rate (100%) is reached only by seven (7) data points out of 1039 (see Table A-2). As it can be observed, this set of points contains some fundamental asset items which one would expect exist for every bank.

Table A-2: List of points with a 100% completeness rate

Data point in ITS template	Content
F 1.1 [r10,c10]	Cash, cash balances at central banks and other demand deposits
F 1.1 [r270,c10]	Tangible assets, like property, plant and equipment
F 1.1 [r360,c10]	Other assets
F 1.1 [r380,c10]	Total assets
F 4.2 [r190,c10]	Financial assets designated at fair value through profit and loss
F 4.3 [r190,c30]	Available for sale financial assets
F 5.0 [r80,c30]	Loans and advances to credit institutions

Besides completeness rates which measure the percentage of missing elements we also quantify the percentage of zero elements within the subset of non-missing data, which we call the null rate. Specifically, we compute for each variable contained in the input templates the number of zero values,

from which we can derive the rate of zero elements as the ratio between zeros and the number of non-missing elements. Consequently, we derive the null rate as the ratio between the number of zero and the total number of data points. We report the average null rate and zero rate across templates in Table A-1.

The null rate is considerably smaller than the zeros rate, and this difference is related to the degree of completeness rate. The mean null rate across templates is 10% (minimum=0% and maximum=31%). The templates on "Held for trading" (HfT) assets (F 4.1), FVPL assets (F 4.2) and derivatives for trading (F 10.0) exhibit the largest percentage of zeros across the set of templates. The mean zeros rate is 61% and gets its maximum values in templates containing liabilities (F 08.1) and derivatives in hedge accounting (F 11.1). Overall the template on derivatives for hedge accounting is the one which is the most sparsely populated, reflecting the lack of use of complex financial instruments from many small banks.

If the number of zero and non-missing values coincide, we flag that data point as degenerate one i.e. this data point always has the value of zero when it is filled, consequently it is an activity which is not undertaken by any bank. So, the degenerate variables are the only ones having the null rate coinciding with the completeness rate, i.e. having a zeros rate equal to 1.

The number and percentage of degenerate variables across templates are presented in Table A-1 (sixth and seventh rows). The percentage is highest for the general templates of the asset and liability side (templates 1 and 7) meaning that there are some specific items in these templates which are not used by any of the reporting banks. We can see that templates with the product breakdown of loans (F 5.0) and off-balance sheet items (F 9.1) have no degenerate variables. Both of these templates are relatively small, refer to subsets of banks' core business and they do not contain extremely specialised items. It is also noted that the mean percentage of degenerate points is equal to 17% and there is a quite polarised situation: templates F 4.1, F4.2, F 4.3, F 5.0, F 9.1 are much below the average with respect to the percentage of degenerate points, while the templates F 1.1, F 8.1, F 10.0, F 11.1 are well above the average. The last two templates contain derivative items with a very detailed breakdown,

and therefore it is not surprising that some of these specialised derivative instruments are not found in any of the reporting banks. The total number of degenerate variables for all templates is 268.

The following variables have a completeness rate equal to unity and no zero elements: F 1.1 [r010,c010] (Cash balances), F 1.1 [r270,c10] (Tangible assets) F 1. 1 [r360,c10] (Other assets) and F 1.1 [r380,c10] (Total assets). Given the fundamental nature and the wide use of these items this comes as no surprise.

A specific analysis concerning missing values can be performed also row-wise, i.e. on the banks. The percentage of banks being complete (i.e. having no missing values) is computed for each template and the percentage of complete records for each template is reported in Table A-1. In general, it is observed that the row-wise completeness is not directly related to the column-wise completeness. The reason is that these completeness rates are driven by various factors. While the column-wise completeness is mainly driven by the degree of use of the activities contained in each template, the row-wise completeness is driven by the degree of diversification of each bank with respect to the different activities. The percentages of banks with no missing variables within each template is generally lower than 10% reflecting that most banks do not possess a degree of diversification which corresponds to the granularity of the templates.

Given that for all data points in our input set, the rule is that banks do not report them if the respective item is not present in their balance sheets, we set all missing values equal to 0.[13] At the same time, we exclude all the degenerate variables. The number of data-points which can be used in the analysis, after this step, becomes thus equal to 1039-268=771.

The second-to-last row of Table A-1 presents the number of variables remaining in each template after the degenerate points are excluded. This number can be interpreted as the "volume" of

[13] Of course it may be the case that a bank has not reported a data point by mistake, but we ignore this effect, given that it is also in the interest of the bank to report complete data. Regarding the handling of missing values and their substitution with zeros, our approach differs from that usually applied in statistical datasets. This approach is justified of course given the institutional context of these data within the SSM framework.

information contained in each template (without any inference about their economic significance) since it abstracts from non-actually-existing information corresponding to degenerate data points.

Data selection

The aggregated dataset that remains after degenerate variables are excluded is not yet ready to be used for the multivariate outlier detection approach. The reason is that many variables are involved in (accounting) identities, therefore presenting a correlation which is very close to unity. We calculate the rank of the modules after the exclusion of degenerate variables (see last row of Table A-1) and this testifies to the existence of a number of redundant variables (as the rank is always lower than the number of the remaining variables, in some cases to a large degree). Figure 1 presents the histogram of correlation values among the variables in this set. It can be observed that there is a large concentration of values in ranges close to 1 and to -0.8 (more extreme values are present at the positive upper bound, reflecting the existence of accounting identities).

Therefore, an automatic procedure for dealing with this dense structure of high correlated variables by excluding those which provide redundant information has to be applied. The procedure aims to keep the more "fundamental" variables of the data set, i.e. those variables which are related to a relatively large number of other variables. For example, if there is a choice between keeping a more general category e.g. "loans and advances to households" and one of its subcategories e.g. "loans and advances to households – real estate collateralized loans", we would prefer to keep the more general one, on the condition that it is strongly related with more of the remaining variables. However, this procedure has to be automatized given the large number of variables.

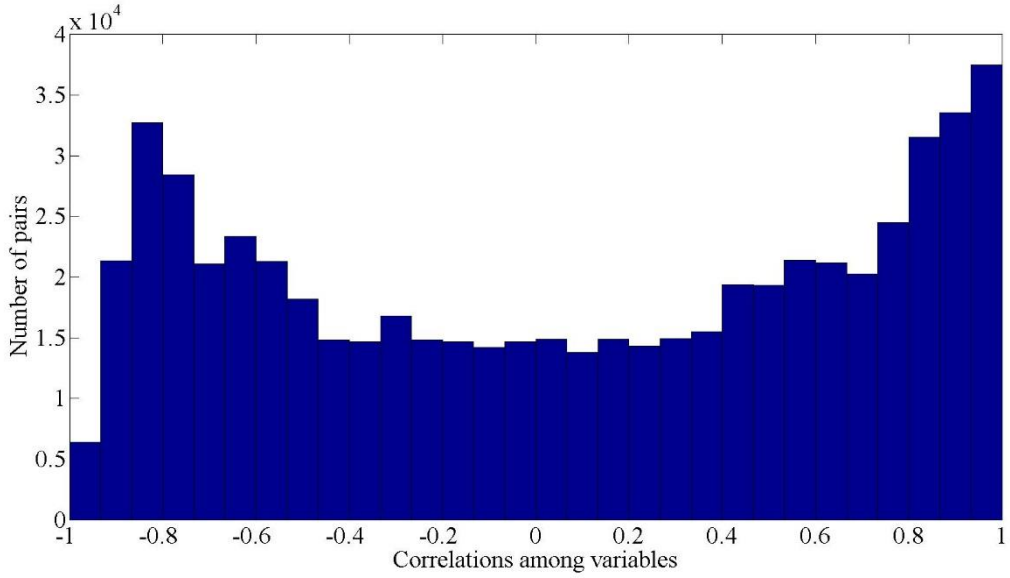
In order to automatize this selection process, we define a measure of the "importance" of each variable within the data set. We re-condition the sample correlation matrix applying the shrinkage estimate of Schäfer and Strimmer, 2005, which is able to drastically reduce the numerical instability of the raw correlation matrix. We define the "importance" $I(j)$ of each variable j as the linear combination of the correlation absolute values with the other variables of a data set:

$I(j) = \sum_{\substack{k=1,\cdots,P, \\ k \neq j}} Corr(j,k).$

Consequently, we order the variables in a non-increasing order based on their $I(j)$. We define a level of correlation C^*, which triggers the need to select one of two correlated variables (e.g. 0.7) and then apply the following selection algorithm:

```
1 begin
2 for k = j + 1 to N
3       if Corr(j,k) > C* then
4             if I(j) < I(k) then
5                   delete variable j from the dataset
6             else
7                   delete variable k from the dataset
8             endif
9       endif
10 end;
11 end;
```

A remarkable number of variables are eliminated, and the final set consists of 433 variables. This large number of variables removed is expected, given the correlation values distribution that is presented in Figure A-1.



Source: Supervisory data, authors' calculations

Figure A-1: Histogram of correlation values (after degenerate points are excluded)

The above procedure is able to identify a number of identities that are present in the data. These identities consist in relationships of more than two data points, for example the breakdown of "Intangible Assets" into "Goodwill" and "Other Intangible Assets". One important advantage of the proposed approach is that the identities among variables located in different templates, a feature which is very important when validation rules for big data sets, such as those contained in FINREP and COREP, have to be formulated.

References

Alfons, A., Croux, C., and Gelper, S. 2013 Sparse least trimmed squares regression for analyzing high-dimensional large data sets. The Annals of Applied Statistics, 7(1), 226–248.

Au, S., Duan, R., Hesar, S., Jiang, W. 2008. A framework of irregularity enlightenment for data pre-processing in data mining, Annals of Operations Research, vol. 174, 47-66.

Barnett, V. and Lewis, T. 1994. Outliers in Statistical Data, New York: John Wiley Sons, 1994.

Bank for International Settlements, 2015. Central banks' use of interest in "big data". Irving Fisher Committee on Central Bank Statistics (IFC) Report, Bank for International Settlements.

Bellini, T. 2012. Forward search outlier detection in data envelopment analysis, European Journal of Operational Research, 216, 2012, pp. 200-207.

Bholat, D. 2015. Big data and central banks, Bank of England, Quarterly Bulletin 2015Q1

Béguin, C., and Hulliger, B. 2008. The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. Survey Methodology, 34(1), 91.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., Sander, J. 2000. LOF: Identifying Density-based Local Outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104.

Cerioli A., Multivariate Outlier Detection With High-Breakdown Estimators, Theory and Methods, Journal of the American Statistical Association, 105, 2010, pp. 147-156.

Çetin, M. 2009. Robust model selection criteria for robust Liu estimator, European Journal of Operational Research, 199, 2009, pp. 21-24.

Chatterjee, S., Jacques, W., 1994. An outlier-resistant approach to risk estimation. Financial Analysts Journal 50 (5), 69-75.

Christy, A., Gandhi, G. M., Vaithyasubramanian, S. 2015. Cluster Based Outlier Detection Algorithm for Healthcare Data. Procedia Computer Science, 50, 209-215.

Croux, C., Rousseeuw, P. J., & Hössjer, O. 1994. Generalized S-estimators. Journal of the American Statistical Association, 89(428), 1271-1281.

Davidson, I., Tayi, G. 2009. Data preparation using data quality matrices for classification mining, European Journal of Operational Research, 197, 2009, pp. 764-772.

Duan, L., Xu, L., Liu, Y., Lee, J. 2009. Cluster-based outlier detection, Annals of Operations Research, vol. 168, 151-168.

Farnè, M. and Vouldis, A. 2016. "Business models of the banks in the euro area", ECB Working Paper Series n. 2070, 26th May 2017.

García-Escudero L. A., Gordaliza A., Matrán C., Mayo-Iscar A. 2010. A review of robust clustering methods, Advances in Data Analysis and Classification, 4, pp. 89-109.

Grane, A., Veiga, H. 2014. Outliers, GARCH-type models and risk measures: A comparison of several approaches, Journal of Empirical Finance, 26, 26-40.

Hammer, C., Kostroch, D. C., Quiros, G. et al. (2017). Big data; potential, challenges and statistical implications. Technical report, International Monetary Fund.

Hampel F. R. Contributions to the theory of robust estimation. PhD thesis, University of California, Berkeley, 1968.

Hawkins, D. 1980. Identification of Outliers. Chapman and Hall.

Henderson, H. V. and Velleman, P. 1981. Building multiple regression models interactively. Biometrics, vol. 37, 391-411.

He, Z., Deng, S., Xu, X., & Huang, J. Z. 2006. A fast-greedy algorithm for outlier mining. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 567-576. Springer Berlin Heidelberg.

C. Hennig 2015, What are the true clusters? Pattern Recognition Letters, 64, pp. 53–62.

P. J. Huber 1964, Robust Estimation of a Location Parameter, Annals of Mathematical Statistics, 35, 1, pp. 73-101.

P. J. Huber 2004, Robust Statistics, John Wiley & Sons, 2nd edition.

M. Hubert, P. J. Rousseeuw, P. J., and Van Aelst, S. 2008, "High-Breakdown Robust Multivariate Methods", Statistical Science, 23, 92-119.

Hubert, Mia, and Stephan Van der Veeken 2008. "Outlier detection for skewed data." Journal of Chemometrics 22.3-4, pp. 235-246.

A. L. Johnson, L. F. McGinnis, 2008. Outlier detection in two-stage semiparametric DEA models, European Journal of Operational Research, 187, pp. 629-635.

Kim, Hyon-Jung, Esa Ollila, and Visa Koivunen, 2015. New robust LASSO method based on ranks. Signal Processing Conference (EUSIPCO), 23rd European. IEEE, 2015.

Khan, J.A., Van Aelst, S. and Zamar, R.H. 2007. Robust linear model selection based on least angle regression. Journal of the American Statistical Association, 102(480), 1289–1299.

Koufakou, A., Georgiopoulos, M., 2010. A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Mining and Knowledge Discovery, 20, 259-289.

Kutsuma, T., Yamamoto, A. 2017. Outlier detection using binary decision diagrams. Data Mining and Knowledge Discovery, 31, 548-572.

Lin S., Brown D., 2006. An outlier-based data association method for linking criminal incidents. Decision Support Systems, 41, pp. 604-615.

Liu, B., Fan, W., & Xiao, T, 2013. A Fast Outlier Detection Method for Big Data. In Asian Simulation Conference (pp. 379-384). Springer Berlin Heidelberg.

Lozano, A C., Nicolai M, Yang, E., 2016. "Minimum Distance Lasso for robust high-dimensional regression." Electronic Journal of Statistics 10. 1., 1296-1340.

Mahalanobis P., 1936. On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India, pp. 49–55.

Maciá-Pérez F., Berna-Martinez J., Fernández Oliva A., Ortega, M. Abreu, 2015. Algorithm for the detection of outliers based on the theory of rough sets. Decision Support Systems, 75, pp. 63-75.

Maronna, R., Yohai V. J, 1976. Robust estimation of multivariate location and scatter. Wiley StatsRef: Statistics Reference Online, pp 1-12.

Maronna, Ricardo A., and Ruben H. Zamar, 2002. "Robust estimates of location and dispersion for high-dimensional datasets." Technometrics 44.4, 307-317.

Nasrabadi, N M., Trac D. T., Nguyen N, 2011. Robust lasso with missing and grossly corrupted observations. Advances in Neural Information Processing Systems.

Nymand-Andersen P., 2016. Big data: The hunt for timely insights and decision uncertainty. Irving Fisher Committee on Central Bank Statistics (IFC) Working Papers, Bank for International Settlements, 2016.

Otey, M., Ghoting, A., Parthasarathy, S., 2006. Fast distributed outlier detection in mixed-attribute data sets. Data Mining and Knowledge Discovery 12, 203-228.

Owen, Art B., 2007. "A robust hybrid of lasso and ridge regression." Contemporary Mathematics 443, 59-72.

Peña, Daniel, and Francisco J. Prieto, 2001. "Multivariate outlier detection and robust covariance matrix estimation." Technometrics 43.3, 286-310.

Ramaswamy S., Rastogi R., Shim. K., 2000. Efficient algorithms for mining outliers from large data sets. Proc. ACMSIGMOD Int. Conf. on Management of data.

Rousseeuw P. J., 1984. Least median of squares regression. Journal of the American Statistical Association, 79, pp. 871-880.

Rousseeuw, P. J., and Hubert M., 2011. "Robust statistics for outlier detection." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, pp. 73-79.

Rousseeuw P.J., Leroy A. M., 2017. Robust Regression and Outlier Detection, John Wiley.

Rousseeuw, P.J., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41, 3, pp. 212-223.

Rousseeuw, P.J., van Driessen, K., 2006. Computing LTS regression for large data sets. Data Mining and Knowledge Discovery, 12, pp. 29-45.

D. Ruppert, 2010. Statistics and Data Analysis for Financial Engineering, Springer.

Salibian-Barrera, M., and Zamar, R.H., 2002. "Bootstrapping robust estimates of regression." Annals of Statistics, 556-582.

Schäfer, J., and K. Strimmer., 2005. A shrinkage approach to large-scale covariance estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol., 4-32.

Seaver B.L., Triantis, K. P., 1995. The impact of outliers and leverage points for technical efficiency measurement using high breakdown procedures. Management Science, 41, 6, pp. 937-956.

She, A. O., Outlier detection using nonconvex penalized regression, Journal of the American Statistical Association, 196, 494, 2011, pp. 626-639.

Teixeira, C., Orair, G. H., Meira Jr, W., Parthasarathy, S., 2008. An efficient algorithm for outlier detection in high dimensional real databases. Tech. Rep.

Templ, M., K. Hron, P. Filzmoser, 2017. Exploratory tools for outlier detection in compositional data with structural zeros. Journal of Applied Statistics 44.4, 734-752.

R. Tibshirani, Regression shrinkage and selection via Lasso, 1996. Journal of the Royal Statistical Society B, 58, pp. 267-288.

Todorov, Valentin, Matthias Templ, Peter Filzmoser, 2011. Detection of multivariate outliers in business survey data with incomplete information. Advances in Data Analysis and Classification 5, 137-56.

Varian H., 2014. Big Data: New Tricks for Econometrics, Journal of Economic Perspectives. 28, 2, pp. 3-27.

Wang H., Li G., Jiang, G., 2007. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. J. Bus. Econ. Stat., vol. 25, pp. 347–355.

Weisberg S., 1985. Applied Linear Regression (2nd ed.), New York: Wiley.

Wu N., Zhang, J., 2006. Factor-analysis based anomaly detection and clustering. Decision Support Systems, 42, pp. 375-389.

Van Hieu, D., Meesad, P., 2016. Cell-DROS: A Fast Outlier Detection Method for Big Datasets. International Journal of Advances in Soft Computing & Its Applications, 8(3).

Huan X., Caramanis C., Mannor S., 2009. Robust regression and lasso. Advances in Neural Information Processing Systems.

Yang, W, Huan X, 2013. A Unified Robust Regression Model for Lasso-like Algorithms. ICML (3).

Yi, C. and Huang, J., 2016. Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression, Journal of Computational and Graphical Statistics.

Yohai V.J., 1987. High Breakdown-Point and High Efficiency Robust Estimates for Regression, The Annals of Statistics, Vol. 15, No. 2, pp. 642-656.

Zhang, Y., Meratnia N., Havinga P., 2010. Outlier detection techniques for wireless sensor networks: A survey. IEEE Communications Surveys and Tutorials, 12, 2, pp. 1-11.

G. Zioutas, L. Pitsoulis, A. Avramidis, 2009. Quadratic mixed integer programming and support vectors for deleting outliers in robust regression. Annals of Operations Research, pp. 339-353.

Acknowledgements

This paper should not be reported as representing the views of the European Central Bank (ECB). We would like to thank an anonymous reviewer for constructive comments on a previous version of the paper. The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Matteo Farnè
University of Bologna, Bologna, Italy; email matteo.farne2@unibo.it

Angelos T. Vouldis
European Central Bank, Frankfurt am Main, Germany; email angelos.vouldis@ecb.europa.eu

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

PDF ISBN 978-92-899-3276-9 ISSN 1725-2806 doi:10.2866/357467, QB-AR-18-051-EN-N