Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics and Management Science

Cellwise Robust Instrumental Variable Estimators

Desiderius de Gouw (449799)

Supervisor:	Mikhail Zhelonkin
Second assessor:	Jens Klooster
Date final version:	30th April 2024

Abstract

This paper examines the bias and variance of the endogenous coefficient from robust estimators in an Instrumental Variable model under cellwise contamination. Instead of robustifying the normal equations of the instrumental variable model, this paper robustifies the solutions to these equations following the approach in Freue et al. (2013). The resulting estimator uses cellwise robust scatter matrices as building blocks. The scatter matrices used are the Two-Step General S-estimator (TSGS), Cellwise MCD estimator and the Stahel-Donoho estimator with cellwise weights, along with their casewise counterparts. The cellwise robust estimators outperform the casewise robust estimators in terms of bias and variance, but not in terms of efficiency. When outliers are marginal the Cellwise MCD performs best, while the TSGS estimator performs best when outliers are extreme. For small dimensions, the TSGS provides reliable results when 83% of observations are expected to contain an extreme cellwise outliers. The TSGS estimator yields reliable results when applied to a real-life dataset to assess the relationship between inflation and trade openness. The matrix with flagged cells or weights is an insightful, convenient byproduct of the cellwise estimators.

Contents

List of Acronyms

2SGS 2-Step General S-estimator

2SLS 2-Stage Least Squares

AR(3) Third-order Autoregressive model

BF Bivariate Filter

CRIV Cellwise Robust Instrumental Variable

DDC *DetectingDeviatingCells* algorithm

FDCM Fully Dependent Contamination Model

FICM Fully Independent Contamination Model

GSE Generalized S-Estimator

IF Influence Function

IV Instrumental Variable

MAD Median Absolute Deviation

MMAD Modified Median Absolute Deviation

MCD Minimum Covariance Determinant

MD Mahalanobis Distance

MedSE Median Squared Error

MLE Maximum Likelihood Estimator

MSE Mean Squared Error

NA Not Available

OLS Ordinary Least Squares

PCA Principal Component Analysis

PDS Positive Definite Symmetric

RIV Robust Instrumental Variable

SD Stahel-Donoho

SDC Stahel-Donoho Components

SDM Stahel-Donoho Maximizing

THCM Tukey-Huber Contamination Model

TSGS Two-Step Generalized S-estimator

UBF Univariate-Bivarite Filter

UF Univariate Filter

1 Introduction

In recent years attention in robust statistics shifted from casewise outliers to cellwise outliers (Raymaekers and Rousseeuw, 2024). The difference is that for casewise contamination the whole observation is treated as an outlier, while for cellwise contamination only some of the components of an observation are treated as outliers. Alqallaf et al. (2009) were the first to introduce this formally and since then many alternatives to detect cellwise outliers were proposed (Gervini and Yohai, 2002; Leung et al., 2017; Saraceno and Agostinelli, 2021; Rousseeuw and Van Den Bossche, 2018). Additionally, new estimators of location and scale specifically catered to cellwise contamination were investigated (Danilov et al., 2012; Raymaekers and Rousseeuw, 2023; Van Aelst et al., 2011). Cellwise robust alternative estimators for linear regression soon followed (Öllerer et al., 2016; Bottmer et al., 2022; Leung et al., 2017; Filzmoser et al., 2020). For time series, Raymaekers and Rousseeuw (2024) show that the method in Leung et al. (2016) performs well for an AR(3) model and lastly, factor analysis under cellwise contamination has been studied by Hubert et al. (2019) amongst others.

One of the classical econometric methods that has not been studied under cellwise contamination is the Instrumental Variable (IV) regression. IV regression solves the issue of endogenous independent variables in an Ordinary Least Squares (OLS) regression, where the endogeneity leads to inconsistent estimates and renders conventional diagnostic tests invalid (Heij et al., 2004). Instrumental Variables are also a common tool used to determine the causal or treatment effects in economics and political science, since randomized experiments are often infeasible (Angrist and Krueger, 2001). In such a case, instrumental variables can provide a source of exogenous variation in an endogenous variable. Outliers in endogenous and other variables often negatively influence the estimation results. Young (2022) and Lal et al. (2023) show that results of IV regressions in economics and political science literature, respectively, can be misleading due to outliers in the data. These misleading results can be costly when they are used, for example, to determine a national policy and hence Instrumental Variable estimators benefit from robust methods.

An attempt of robustifying the IV estimator under casewise contamination is given in Freue et al. (2013), who propose the Robust Instrumental Variable (RIV) estimator that uses the highly robust S-estimator as its building block. Their idea is based on the regression estimator from Croux et al. (2003). Instead of robustifying the first order conditions of the OLS estimator, they robustify the solution to these equations. Freue et al. (2013) show that their estimator is consistent under weak distributional assumptions and asymptotically normal under mild regularity conditions. Additionally, the estimator has a high breakdown point and bounded

influence function. Lastly, Freue et al. (2013) present an iterative algorithm called L_1-RIV that allows exogenous dummy variables to be included in the model while simultaneously using only continuous variables as the inputs for the robust scatter estimator. Highly robust and efficient methods such as the Minimum Covariance Determinant (MCD) from Rousseeuw (1985) often fail in the presence of dummy variables, however, the L_1-RIV allows them to be used alongside dummy variables. These favourable characteristics of the RIV estimator are the motivation to explore the performance of the method under cellwise contamination.

This paper investigates estimation in an Instrumental Variable model under cellwise contamination. Specifically, it investigates the performance of the natural robustification proposed by Freue et al. (2013) under cellwise contamination. It aims to answer the following research question: **Can robustifying the solution equations of the IV estimator provide robust estimates under cellwise contamination?** As the natural robustification is based on robust covariance estimation, this paper compares the performance of different casewise robust covariance estimators and explores cellwise robust alternatives. Since the L_1-RIV algorithm from Freue et al. (2013) does not limit potential estimators to be able to handle dummy variables, I chose estimators that use conceptually distinct approaches and have a cellwise counterpart. The estimators that fulfil these conditions are the S-estimator from Rousseeuw and Yohai (1984), the Minimum Covariance Determinant (MCD) from Rousseeuw (1985) and the Stahel-Donoho (SD) estimator from Stahel (1981) and Donoho (1982). Their cellwise robust alternatives are the Two-Step Generalized S-estimator (TSGS) from Leung et al. (2017), the Cellwise MCD from Raymaekers and Rousseeuw (2023) and the SD estimator with cellwise weights from Van Aelst et al. (2011). The focus of the paper lies on the accuracy of estimation of the endogenous coefficient, since that is often the variable of interest in an IV setting.

The performances of these estimators are investigated for three model specifications. The first model contains one endogenous, control and instrumental variable, the second model contains one endogenous and instrumental variable and five control variables and the third model contains one endogenous variable, three instruments and three control variables. These three scenarios respectively represent a practical scenario with a small dataset, a scenario with a larger dataset to assess the curse of dimensionality for cellwise robust estimators and the scenario where highly correlated lagged values are used as instruments. The data sets will be contaminated with marginal outliers, i.e. outliers that in the tail of the target distribution ($\pm$ 3 standard deviations), and extreme outliers ($\pm$ 10 standard deviations). The performance of the estimators will be measured in terms of bias, variance and Mean Squared Error (MSE) of the coefficients for the endogenous variable and I briefly examine the efficiencies of the estimators.

My findings indicate that the model specifications are less relevant than the contamination magnitude. For marginal outliers, the Cellwise MCD performs best, but breaks down when the contamination rate is higher than 10%. All other estimators are already biased or inaccurate when contamination levels are higher than 5%. When the outliers are extreme and hence easy to flag as outliers, the TSGS performs best and only becomes inaccurate and biased if the contamination rate is higher than 25%. Note that this corresponds to a case where 83% of observations contain an outlying cell. The Cellwise MCD also performs well when the outliers are extreme. The SD estimator with cellwise weights performs poorly across all model specifications and types of outliers. The additional cellwise robustness comes at the cost of efficiency, although the loss in efficiency is not large. The TSGS performs well in an applied setting and the matrices with weights from the filtering step in the TSGS and the subset selection of the Cellwise MCD estimator prove to be insightful byproducts of the proposed estimator.

The main contribution to literature is applying cellwise robust covariance matrices in an IV setting. The cellwise robust covariance matrices have already been applied to regression (Leung et al., 2016), PCA (Hubert et al., 2019), discriminant analysis (Aerts and Wilms, 2017), hence the, yet unexplored, application to other econometric models is a logical addition to existing literature. This paper could also be interpreted as an assessment of the performance of the methods proposed in Freue et al. (2013) under a new contamination model.

The paper is structured as follows: Section 2 elaborates on cellwise contamination and the need for cellwise robust alternatives and Section 2.1 discusses current literature on scatter estimators for cellwise contamination. Section 3 formalizes the Instrumental Variable model, Section 3.2 investigates the effect of outliers in the IV model and Section 3.3 describes the robustification of Freue et al. (2013). Section 4 proposes the Cellwise Robust Instrumental Variable (CRIV) estimator and elaborates on the robust covariance matrices used as building blocks. Section 5 describes the simulation setup and analyzes the results of the cellwise estimators compared to their casewise counterparts in different scenarios. Section 6 applies the new estimator to a real-life dataset. Section 7 gives a summary and suggestions for future research.

2 Cellwise Contamination

This section first discusses the source of cellwise contamination, formalizes the definitions of cellwise contamination and argues why casewise robust estimators might not yield reliable results under cellwise contamination. Concrete examples are given to support the last statement. Then Section 2.1 briefly examines the current literature concerning cellwise robust location and scatter matrices and the main ideas of the available estimators.

Suppose a researcher is collecting data on households and part A of the data is collected from a statistics bureau and part B is collected through a survey. Both parts are measured separately and may contain outliers in variables where the other part does not. The result is that observations will contain outliers in either part A or B, in both or in none, such that downweighting the whole observation will also downweight the uncontaminated data cells.

As is common in robust statistics literature, the basis of analysis is the Tukey-Huber Contamination Model (THCM) $\boldsymbol{F}_\epsilon = (1-\epsilon)\boldsymbol{F} + \epsilon\boldsymbol{G}$, where $\boldsymbol{F}$ denotes the target distribution and $\boldsymbol{G}$ denotes an unknown distribution that is not of interest to the researcher (Tukey, 1962; Huber, 1964). Hence the observed random variable $\boldsymbol{F}_\epsilon$ is a mixture of the target distribution and another distribution. Here ϵ determines the extent to which the mixture leans towards the target distribution and it can be interpreted as the probability that the observed variable is contaminated. Note that it would typically make sense to analyse a data set if $(1-\epsilon) \geq 0.5$ as otherwise $\boldsymbol{F}$ is not the dominant model anymore. If $\boldsymbol{F}_\epsilon$ is a multivariate random variable of dimension k then the model becomes $\boldsymbol{F}_\epsilon = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{F} + \boldsymbol{B}\boldsymbol{G}$, where $\boldsymbol{B}$ is a diagonal matrix with elements $b_1, b_2, .., b_k$. In line with the literature, $\boldsymbol{F}$, $\boldsymbol{G}$ and $\boldsymbol{B}$ are assumed to be independent in this paper (Alqallaf et al., 2009).

Within the THCM, casewise contamination can be defined by setting $b_1 = b_2 = \cdots = b_k \in \{0, 1\}$ and $P(\boldsymbol{B} = \boldsymbol{I}) = \epsilon$. That is, every component in $\boldsymbol{F}_\epsilon$ either comes from the unknown distribution $\boldsymbol{G}$ with probability ϵ or comes from $\boldsymbol{F}$ with probability $1 - \epsilon$. This is also called the Fully Dependent Contamination Model (FDCM) and it has the convenient property that the number of contaminated cases remains equal after affine equivariant transformations, that is why methods designed to be robust under the FDCM often share that property (Alqallaf et al., 2009).

Cellwise contamination, also known as Fully Independent Contamination Model (FICM), is different from the FDCM in the sense that each diagonal element b_j is 1 or 0 independently from each other. The probability that the diagonal entries are 0 or 1 is assumed to be the same for all elements, such that the only assumption is that $P(b_j = 1) = \epsilon$ for all $j = 1, 2, \ldots, k$. An important difference between the two models is that the probability that an observation is contaminated is much larger in the FICM compared to the FDCM. For the former it holds that $P[\text{case is contaminated}] = 1 - (1-\epsilon)^k$, which increases rapidly in k and ϵ, while for the latter this probability remains ϵ independent of the dimension.

Table 1 shows the probabilities that a case is contaminated under FICM for different values of ϵ and k. For example, in a dataset with 15 variables where only 5% of the cells per variable are contaminated, one should expect 54% of the cases to contain contaminated cells. Downweighting

	$k=2$	$k=4$	$k=8$	$k=10$	$k=15$	$k=20$
$\epsilon=0.05$	0.10	0.19	0.34	0.40	0.54	0.64
$\epsilon=0.1$	0.19	0.34	0.57	0.65	0.79	0.88
$\epsilon=0.15$	0.28	0.48	0.73	0.80	0.91	0.96
$\epsilon=0.2$	0.36	0.59	0.83	0.89	0.96	0.99
$\epsilon=0.25$	0.44	0.68	0.90	0.94	0.99	1.00
$\epsilon=0.3$	0.51	0.76	0.94	0.97	1.00	1.00

Table 1: The probability that a case is contaminated for different cellwise contamination rates ϵ and number of parameters k

cases containing a contaminated cell, as casewise robust estimator generally do, results in the loss of information from many uncontaminated cells within those cases. This illustrates the need for alternative estimators when the data is suspected to be contaminated according to the FICM.

Another illustration of the need for cellwise robust estimators can be found in Figure 1. In the left panel three out of fifteen observations are contaminated by outliers, while in the right panel the cells within a variable have been contaminated with probability of 3/15 such that the contamination rate is the same. In the left panel only three observations are contaminated, while in the right panel only two observations are totally clean. It is not a good idea to downweight all contaminated observations as the estimate will then depend strongly on the few uncontaminated observations.

Another reason we need estimators specifically constructed for cellwise contamination is that estimators that are affine equivariant will not necessarily be robust anymore. This is because linear combinations of the data might change the fraction of contaminated observations. Formally, as defined by Alqallaf et al. (2009): let $\tilde{\boldsymbol{F}} = \boldsymbol{A}\boldsymbol{F}_\epsilon + \boldsymbol{b}$ then $\tilde{\boldsymbol{F}} = \boldsymbol{A}(\boldsymbol{I}-\boldsymbol{B})\boldsymbol{F} + \boldsymbol{A}\boldsymbol{B}\boldsymbol{G} + \boldsymbol{b}$ does not necessarily follow the same distribution as $\boldsymbol{F}$. Only when $\boldsymbol{A}$ is diagonal, and hence $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$ will the distribution of $\tilde{\boldsymbol{F}}$ be the same as the distribution of $\boldsymbol{F}$. This implies that the estimator is only scale equivariant and not affine equivariant.

2.1 Cellwise robust scatter estimators

Various estimators of scatter have been proposed under cellwise contamination but in general three different approaches are used for estimation. The first approach flags outlying cells and then treats the flagged cells as missing, the second approach finds a subset of cells that implies the lowest determinant of the resulting covariance matrix and the third approach uses cellwise weights based on the outlyingness of the cell. In other words, the value of outlying cells are changed, outlying cells are either included or excluded or the extent to which the cells are included depends on how outlying they are.

The first approach is a two-step approach that uses a filter to flag outliers and then sets the values of these cells to Not Available (NA). Some examples of filters used in the literature are the univariate filter from Gervini and Yohai (2002), the bivariate filter of Leung et al. (2017) and a multivariate filter from Saraceno and Agostinelli (2021). Rousseeuw and Van Den Bossche (2018) detect deviating cells using correlations to predict values and then flag the cells with large residuals. These methods can also be combined to flag outliers as is done in Leung et al. (2017), who then set the outliers to NA and apply the Generalized S-Estimator (GSE) of Danilov et al. (2012) to the resulting data set with missing values to obtain a robust covariance matrix.

The second approach is proposed in Raymaekers and Rousseeuw (2023), who adjust the MCD estimator of Rousseeuw (1985) to be robust under cellwise contamination. The method is called Cellwise MCD and the general idea is the same as its casewise counterpart. Their algorithm loops through all columns to compute the zero-one weights of each cell while penalising the number of zero weights. Results prove that the algorithm can flag cellwise outliers well and performs best in simulations when compared to Gaussian and Spearman rank-based estimators from Öllerer and Croux (2015), the Gnanadesikan-Kettenring estimator from Tarr et al. (2016), the 2-Step General S-estimator (2SGS) from Agostinelli et al. (2015) and the Detection-Imputation algorithm from Raymaekers and Rousseeuw (2021). The Cellwise MCD estimator yielded estimates close to the true covariance matrix when the outliers were marginal

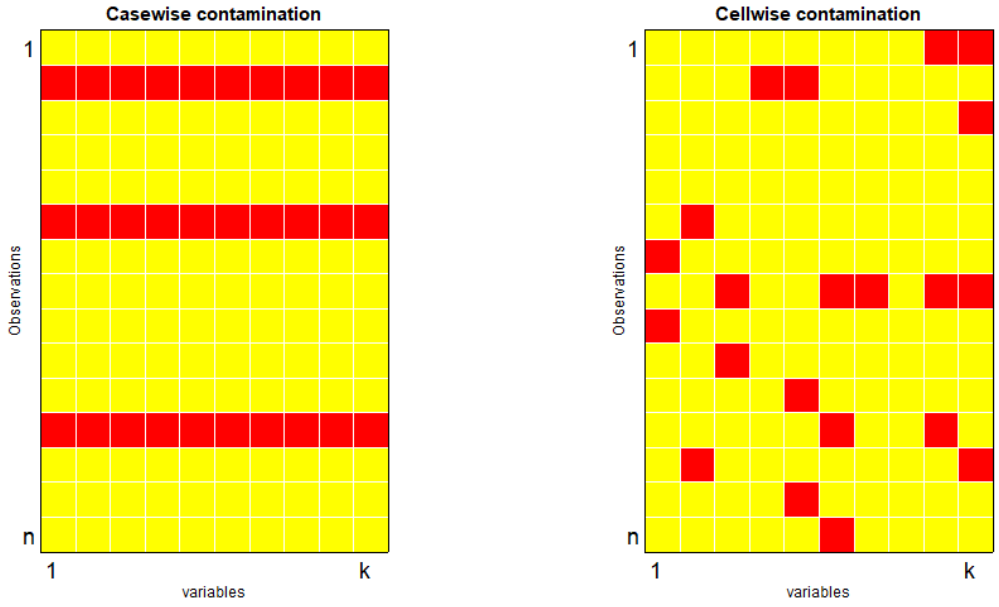


Figure 1: The left panel shows the outlying cells under casewise contamination which are by definition all in the same row, the right panel shows outlying cells under cellwise contamination. Note that while the contamination rate is the same, only two rows are unaffected in the cellwise contaminated scenario.

and the efficiency of the averaged 90% compared to the Maximum Likelihood estimator.

Lastly, Van Aelst et al. (2011) propose a Stahel-Donoho estimator (SD-Estimator) with cellwise weights as an extension to the casewise Stahel-Donoho location and scatter estimator of Stahel (1981) and Donoho (1982). Instead of weighting observations as in the original SD-estimator, they weigh each component of each observation separately. The cellwise weights for component j are a weighted average of the original SD weight for the whole observation and the degree of outlyingness of the observation in the direction of variable j. The SD-estimator performs well for contamination under the FICM model as the weights assigned to cells that are not outlying are higher and hence the resulting estimator becomes more efficient.

3 The Instrumental Variable model its robustification

This section describes the idea behind the IV model, formalizes the model, provides a list of necessary conditions that need to hold and derives the 2-Stage Least Squares (2SLS) estimator. Section 3.2 examines the behaviour of the non-robust 2SLS estimator when there are outliers in the data and illustrate the need for robust estimators. Section 3.3 explains the natural robustification of Freue et al. (2013) and Section 3.3.1 describes the L_1-RIV algorithm that allows for exogenous dummy variables.

3.1 The 2-Stage Least Squares Estimator

Instrumental variables can be used to solve a system of equations or to overcome measurement errors in the data, but more recently it has become pivotal to disclose causal relationships in the presence of endogenous regressors (Angrist and Krueger, 2001). One possible source of endogeneity is the omission of relevant variables, such that a regressor becomes correlated with the error term. If there are endogenous variables in the model, OLS will yield inconsistent estimates. To eliminate the endogeneity issue, instruments are used to predict the exogenous variation in endogenous variables such that these predictions can be used as exogenous regressors.

Suppose that a researcher wants to estimate a regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$ and some control variables $\boldsymbol{X}_2$, but he finds that $\boldsymbol{X}_1$ is endogenous, i.e. $E[\boldsymbol{X}_1'\boldsymbol{\varepsilon}_2] \neq 0$. Consider the instrument $\boldsymbol{Z}$ that is correlated with $\boldsymbol{X}_1$ but uncorrelated with $\boldsymbol{\varepsilon}_2$. In the first stage, he can regress $\boldsymbol{X}_1$ on the instrument $\boldsymbol{Z}$ and the control variables $\boldsymbol{X}_2$ and obtain the predictions $\hat{\boldsymbol{X}}_1$. These predictions are the part of $\boldsymbol{X}_1$ that is uncorrelated with $\boldsymbol{\varepsilon}_2$. I assume that $\boldsymbol{\varepsilon}_1$ is uncorrelated with $\boldsymbol{\varepsilon}_2$ and both follow a normal distribution with mean zero and variance σ_1 and σ_2 respectively. Formally, the model described above are summarized in equations (1)

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{Z}'\boldsymbol{\gamma} + \boldsymbol{X}_2'\boldsymbol{\delta} + \boldsymbol{\varepsilon}_1, \\ \boldsymbol{y} &= \boldsymbol{X}_1'\boldsymbol{\beta}_1 + \boldsymbol{X}_2'\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2. \end{aligned} \tag{1}$$

Here $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ is a $n \times (p_1 + p_2)$ matrix where the i-th row is given by $(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, $\boldsymbol{Z}$ is a $n \times q$ matrix where the i-th row is given by $(\boldsymbol{z}_i - \bar{\boldsymbol{z}})$ and $\boldsymbol{y}' = (y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$. Let p_1, p_2 and q be the number of endogenous, control and instrumental variables respectively.

The instrument $\boldsymbol{Z}$ needs to fulfil three conditions if it is to be considered a valid instrument. First, the instrument needs to be uncorrelated with the error term, that is $E[\boldsymbol{Z}'\boldsymbol{\varepsilon}_2] = 0$. The instrument also needs to be sufficiently correlated with the endogenous regressor, i.e. $E[\boldsymbol{X}'\boldsymbol{Z}] = \boldsymbol{Q}_{\boldsymbol{XZ}}$, where $\boldsymbol{Q}_{\boldsymbol{XZ}}$ is full rank. The third condition is that the instrument is stable in the sense that $E[\boldsymbol{Z}'\boldsymbol{Z}] = \boldsymbol{Q}_{\boldsymbol{ZZ}}$ where $\boldsymbol{Q}_{\boldsymbol{ZZ}}$ is full rank.

To show why the coefficient of the endogenous variable $\boldsymbol{\beta}_1$ will be inconsistent if OLS is applied to the second equation, let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. Then by definition of an endogenous variable $E[\boldsymbol{X}'\boldsymbol{\varepsilon}_2] \neq 0$. Then the expected value of the estimator is not equal to its true value, which is shown in the bottom equation of (2)

$$\begin{aligned} E\left[\hat{\boldsymbol{\beta}}_{OLS}\right] &= E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\right], \\ &= E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2\right)\right], \\ &= \boldsymbol{\beta} + E\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}_2\right] \neq \boldsymbol{\beta}. \end{aligned} \tag{2}$$

The last equation holds because of the endogeneity of $\boldsymbol{X}_1$ and the assumption of stable regressors, i.e. $E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] = \boldsymbol{Q}_{\boldsymbol{XX}} \neq 0$, where $\boldsymbol{Q}_{\boldsymbol{XX}}$ is full rank.

By projecting $\boldsymbol{Z}$ onto $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$, one is left with predictions of $\boldsymbol{X}$ that are uncorrelated with $\boldsymbol{\varepsilon}_2$, because $E\left[\hat{\boldsymbol{X}}'\boldsymbol{\varepsilon}_2\right] = E\left[\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}\boldsymbol{Z}'\right)^{-1}\boldsymbol{Z}'\boldsymbol{\varepsilon}_2\right] = 0$ due to the exogeneity of the instrument. In the next step, the exogenous predictions can be used as regressors in the second equation of (1). This two-step procedure is known as the 2SLS estimator and yields $\boldsymbol{\beta}_{2SLS} = \left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}$, which is consistent as shown in equation (3)

$$\begin{aligned} E\left[\hat{\boldsymbol{\beta}}_{2SLS}\right] &= E\left[\left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}\right], \\ &= E\left[\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2\right)\right], \\ &= \boldsymbol{\beta} + \left(\boldsymbol{Q}_{\boldsymbol{ZX}}'\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{Q}_{\boldsymbol{ZX}}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{ZX}}'\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}E\left[\boldsymbol{Z}'\boldsymbol{\varepsilon}_2\right] = \boldsymbol{\beta}. \end{aligned} \tag{3}$$

Thus the consistent IV estimator is given by

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} = \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}. \tag{4}$$

3.2 The effect of outliers in the IV model

This subsection investigates what happens to 2SLS coefficients when there are outliers present in the data. The goal is to show the need for robust estimators in that case. I assume the model follows (1) and contains one exogenous variable, one endogenous variable and one instrument, i.e. $p_1 = p_2 = q = 1$.

The simulation setup is as follows: I simulate $r = 1.000$ samples of $n = 100$ observations from a normal distribution with mean zero and variance one. The covariance matrix is such that the exogenous variable and the instrument are not correlated with the error term, while the endogenous variable has a correlation of 0.9 with the instrument, i.e. the instrument is a strong instrument, and 0.2 with the error term. The dependent variable is computed as $y_i = 1 + 2x_{i,End} + 2x_{i,Exo} + \varepsilon_i$. Lastly, in each sample one variable is contaminated by replacing $\epsilon\%$ of the data with $\gamma \in \{3, 5, 7, 10\}$ and then 2SLS is applied to obtain the coefficients. If the instrument was rendered invalid after contamination, a new set was simulated and contaminated.

Similar to the approach used in Freue et al. (2013), Monte Carlo Median Squared Error (MedSE) of the coefficient estimates with respect to the true values are computed to assess the bias of the estimates. Formally, MedSE $= \text{median}(||\hat{\beta}^1 - \beta)||, ||\hat{\beta}^2 - \beta||, \ldots, ||\hat{\beta}^r - \beta)||)$, where $||.||$ denotes the Euclidean norm. The median is used because the estimator is expected to break down and the mean has a breakdown point of one observation, such that the mean will not be

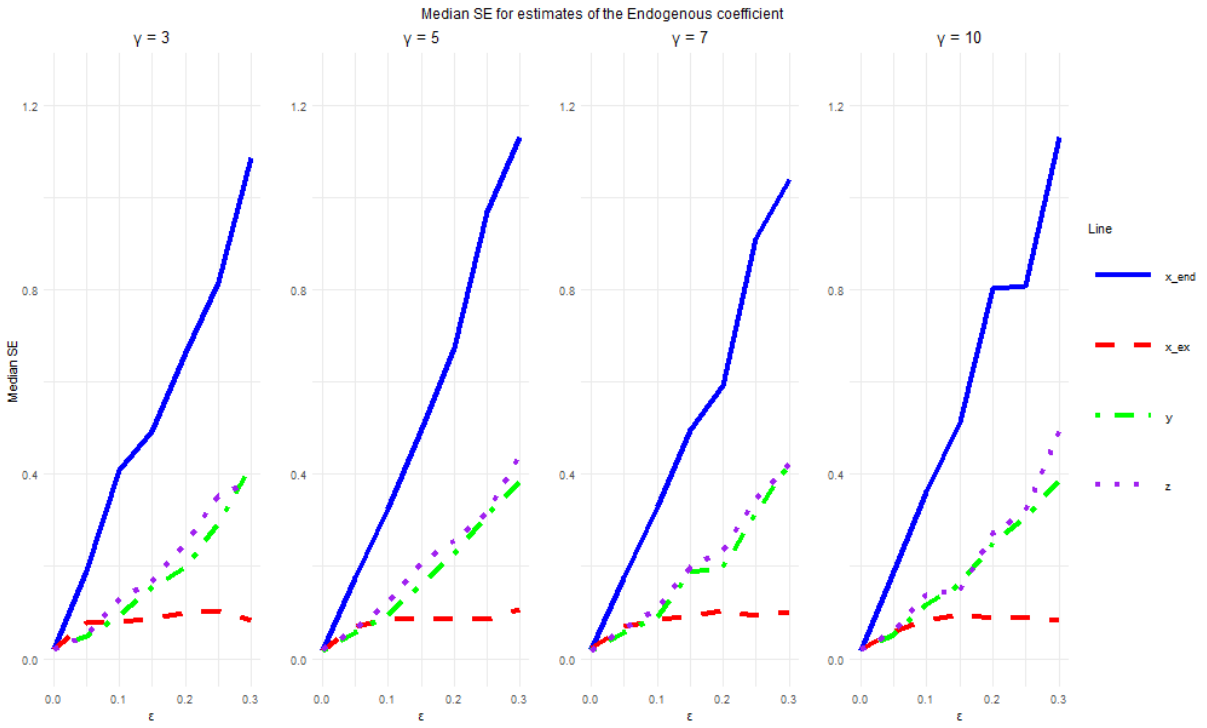


Figure 2: Median squared bias of the endogenous coefficient in an IV model for different contamination levels γ and for contamination in all variables separately

able to show the meaningful patterns that the median will presumably show.

Figure 2 depicts the results of contaminating the variables separately and running 2SLS on the contaminated data sets. Most importantly, we can see that the contamination in the endogenous variable causes the MedSE to increase the most, while contamination in the exogenous variable leads to a relatively low median Squared Error that does not increase much with the contamination rate. Contamination in the instruments and the dependent variable also increases the MedSE, albeit less than contamination in the endogenous variable. These results indicate that effect of contamination varies depending on which variables are contaminated. Figures 9 and 10 in the Appendix show the same figures for the intercept and exogenous variable respectively.

3.3 Ordinary IV estimator and its natural robustification

Because the model in equations (1) uses centered variables, the IV estimator solution equation can be rewritten in terms of the covariance matrix. The covariance matrix $\boldsymbol{\Sigma}$ of the variables $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y})$ is split into

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Xy}} \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Zy}} \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{yX}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{yZ}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{yy}} \end{bmatrix}. \tag{5}$$

Then the estimator can be written as in equation (6)

$$\hat{\boldsymbol{\beta}}_{OIV} = \left[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}}\right]^{-1}\left[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XZ}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Zy}}\right]. \tag{6}$$

Rewriting the estimators in this manner reveals a natural way to robustify the estimator, namely by replacing the sample estimate $\hat{\boldsymbol{\Sigma}}$ with robust alternatives. Denote the robust estimators of location and scatter by $\boldsymbol{S}$ for a sample of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y})$ and split them in the same way as done in equation (5). Then the Robust Instrumental Variable (RIV) estimator is given in equation (7)

$$\hat{\boldsymbol{\beta}}_{RIV} = \left[\boldsymbol{S}_{\boldsymbol{XZ}}\boldsymbol{S}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{S}_{\boldsymbol{ZX}}\right]^{-1}\left[\boldsymbol{S}_{\boldsymbol{XZ}}\boldsymbol{S}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{S}_{\boldsymbol{Zy}}\right]. \tag{7}$$

Freue et al. (2013) use the robust S-estimator to compute the location and scatter matrix $\hat{\Sigma}$ and show that the estimates is consistent as long as $\boldsymbol{S}_{\boldsymbol{Z\varepsilon}} = 0$, even if the whole S-estimator is not consistent. Additionally, they derive the Influence Function (IF) of the RIV estimator and the asymptotic variance, which can be uses to approximate the standard errors. For the details I refer to the paper of Freue et al. (2013), the important practical note is that the method to

compute the asymptotic variances has been implemented in their `R` package `riv` and is readily available.

It is important to note that although the estimator assumes centered variables, in practice this does not mean that variables should first be centered. If a dataset is contaminated, estimating the true location of the data set can be complicated. However, covariance matrices are generally location invariant, implying that shifting non-centered data will not change the covariance estimates and hence will not change the estimates of the coefficients, except for the intercept.

3.3.1 Dummy variables in the Robust IV

This section briefly outlines the L_1-RIV algorithm from Freue et al. (2013) to allow for exogenous dummy variables in the RIV estimator. First the problem of including dummy variables in the regular RIV estimators are explained and then the algorithm is described.

Suppose the model in equation (1) contains a dummy variable and the bottom equation becomes

$$\boldsymbol{y} = \boldsymbol{X_1}'\beta_1 + \boldsymbol{X_2}'\beta_2 + \boldsymbol{C}'\beta_3 + \boldsymbol{\varepsilon}_2, \tag{8}$$

where $\boldsymbol{C}$ is a $n \times d$ matrix with dummy variables, for convenience I assume here that $d = 1$. Stromberg (1993) show that it is impossible to compute the S-estimator exactly and therefore need to be approximated. Proposed estimators often use subsampling to speed up the minimization of the objective function (Salibian-Barrera and Yohai, 2006). However, Maronna and Yohai (2000) state that the subsampling algorithms may fail due to collinearity of the dummy variables. This happens for example when all dummy values are either zero or one in the subsample.

To solve this problem, Maronna and Yohai (2000) propose an algorithm that iteratively computes the coefficients of the continuous variable with an S-estimator and the coefficients of the dummy variables with a monotone M-estimator such as the L_1 estimator. The idea described above was also adopted in Freue et al. (2013). At each iteration, the partial residuals from the regression of $\boldsymbol{y}$ on the dummy variables are used as dependent variables in the RIV procedure to compute the coefficients of the continuous variables. Then the partial residuals of the regression of $\boldsymbol{y}$ on the continuous variables are used to compute the coefficients of the dummy variables. This process is iterated until convergence.

The iterative procedure is formally defined by equation (9)

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{1,2}^{(k)} &= RIV\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y} - \boldsymbol{C}'\hat{\boldsymbol{\beta}}_3^{(k-1)}\right), \\ \hat{\boldsymbol{\beta}}_3^{(k)} &= L_1\left(\boldsymbol{C}, \boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1^{(k-1)} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2^{(k-1)}\right),\end{aligned} \tag{9}$$

where k denotes the $k-th$ iteration, $RIV(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y})$ denotes the RIV procedure as described above and $L_1(\boldsymbol{X}, \boldsymbol{y})$ is an M-regression that minimizes the absolute value of the residuals. To initialize the algorithm, the effect of the dummy variables is removed from all other continuous variables and then these "clean" variables are used to estimate the coefficients $\hat{\boldsymbol{\beta}}_{1,2}^{(0)}$ of the continuous variables with an S-estimator. The resulting coefficients are then used to compute the residuals, which are then regressed on the dummy variables to obtain $\hat{\boldsymbol{\beta}}_3^{(0)}$. For more detailed description of the iterative algorithm I refer to Maronna and Yohai (2000).

4 Cellwise Robust IV estimator

The Cellwise Robust Instrumental Variable (CRIV) estimators proposed in this thesis will follow the same approach as Freue et al. (2013). Instead of robustifying the model equations, I robustify the model solutions and estimate the coefficients using robust covariance matrices as in equation (7). This method allows to compare the performance of the RIV estimator based on different covariance matrices. The benchmark covariance estimator is the S-estimator as used in Freue et al. (2013). Additionally, the MCD estimator of Rousseeuw (1985) and the SD estimator of Stahel (1981) and Donoho (1982) are included in the analysis to compare the cellwise estimators to their casewise counterparts. The contending cellwise estimators are the Two-Step Generalized S-estimator (TSGS) from Leung et al. (2017), the Cellwise MCD from Raymaekers and Rousseeuw (2023) and the Stahel-Donoho estimator with cellwise weights from Van Aelst et al. (2011). The methods are explained in the next sections along with their properties. In each section, assume that $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is a data set with n observations and each $\boldsymbol{x}_i$ is a vector of dimension p.

4.1 S-Estimator

The S-estimator was originally proposed in a regression setting by Rousseeuw and Yohai (1984), but is easily translated to compute a covariance matrix. Following the definition in Danilov et al. (2012), define $\boldsymbol{\mu} \in \Re^p$ and $\boldsymbol{\Sigma} \in \text{PDS}\ (p)$, the set of positive definite symmetric matrices and define $S_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the solution for s to equation (10)

$$\frac{1}{N}\sum_{i=1}^{n} \rho\left(\frac{d\left(\boldsymbol{x_i}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)}{c_p s}\right) = \delta, \tag{10}$$

where $d(\boldsymbol{x_i}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$ is the Mahalanobis Distance (MD), c_p is chosen such that $E(\rho\left(\frac{||\boldsymbol{X}||}{c_p}\right)) = \delta$ with $\delta \in [0,1]$ and $\boldsymbol{X}$ in this case has a multivariate normal density. The definition of the S-estimator for location and scale $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is then given in equation (11)

$$\begin{aligned} \left(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n\right) &= \arg\min_{\hat{\boldsymbol{\mu}}, |\hat{\boldsymbol{\Sigma}}|=1} \boldsymbol{\theta}_n (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \hat{s}_n &= S_n \left(\hat{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\Sigma}}_n\right), \\ \hat{\boldsymbol{\Sigma}}_n &= \hat{s}_n \tilde{\boldsymbol{\Sigma}}_n. \end{aligned} \tag{11}$$

Rousseeuw and Yohai (1984) state that the ρ function needs to satisfy the following conditions: $\rho(0) = 0$, ρ is symmetric, ρ is continuously differentiable and there exists $c > 0$ such that ρ is monotonically increasing on $[0, c]$ and constant for values larger than c. The function ρ that used by Freue et al. (2013) and in this paper is the Tukey bisquare loss function tuned to have a breakdown point of 50%. Formally, the Tukey bisquare loss function is given in equation (12) and can be tuned to achieve a breakdown point of 50% by setting $c \simeq 1.547$

$$\begin{aligned} \rho &= \frac{x^2}{2} - \frac{x^4}{sc^2} + \frac{x^6}{6c^4}, && \text{for } |x| \leq c, \text{ and} \\ \rho &= \frac{c^2}{6}, && \text{for } |x| \geq c. \end{aligned} \tag{12}$$

Davies (1987) showed that S-estimators are strongly consistent for the semiparametric elliptical model. The RIV estimator from Freue et al. (2013) inherits the consistency of the S-estimator above.

4.2 Two-Step Generalized S-estimator

The predecessor of the TSGS was introduced by Agostinelli et al. (2015), it is called the 2SGS and it is based on the following idea: first the procedure applies the adaptive univariate filter from Gervini and Yohai (2002), also called the Gervini-Yohai filter, to detect cellwise outliers and sets these to NA. In step two, the Generalized S-estimator of Danilov et al. (2012) is applied to the incomplete data set, i.e. the set with NA values. Leung et al. (2017) extend the first step of this procedure by combining the Gervini-Yohai filter with a bivariate filter and then intersect the resulting set of outliers with a set from another outlying detection method called the *DetectingDeviatingCells* algorithm (DDC) from Rousseeuw and Van Den Bossche (2018). The cells that are deemed outliers by both methods are set to NA and then the GSE computes the scatter matrix. Leung et al. (2017) call this the *UBF-DDC GSE* estimator, however, here I will refer to it as the Two-Step Generalized S-estimator (TSGS). Note that the difference between the 2SGS and TSGS is that the latter combines multiple filters and a detection method, while the

former only uses a univariate filter. The TSGS is available in the `R` package `GSE`. This subsection will briefly touch upon each filter and the Generalized S-estimator, for detailed treatment I recommend aforementioned literature.

4.2.1 The Univariate Filter (UF)

The idea for the adaptive Univariate Filter (UF) in Gervini and Yohai (2002) is the following: standardize the values using a robust location and scatter estimator and compare the standardized values to a reference distribution. Formally, let $z_i = (x_i - T)/S$ be standardized values using the median as location estimator T and the Median Absolute Deviation (MAD) as scatter estimator S following the choices in Agostinelli et al. (2015) and Leung et al. (2017). Both papers also took the Normal distribution as reference distribution, i.e. $F = \Phi$, which is done here too. The proportion of values that will be flagged by the filter is given by equation (13)

$$d_n = \sup_{t \geq \eta} \{F^+(t) - F_n^+(t)\}^+. \tag{13}$$

Here F^+ is the reference distribution of $|Z|$, $F_n^+ = \frac{1}{n}\sum_{i=1}^n I(|\boldsymbol{z}_i| \leq t)$ is the empirical distribution function of the absolute values of the standardized variable, $\{g\}^+$ refers to the positive part in g and $\eta = (F^+)^{-1}(\alpha)$ is the positive quantile of the distribution function of F. In order to detect only large outliers, $\alpha = 0.95$ is used. Once d_n is known, the filter will flag the $\lfloor d_n n \rfloor$ largest absolute standardized values. The filter is shown to be consistent as long as the tail of the reference distribution is equal or heavier than the tail of the true distribution.

4.2.2 The Bivariate Filter (BF)

The Bivariate Filter (BF) compares the pairwise Mahalanobis Distances, defined as $MD_i = (\boldsymbol{z}_i - \boldsymbol{T})'\boldsymbol{C}^{-1}(\boldsymbol{z}_i - \boldsymbol{T})$, to a reference distribution instead of the standardized values. The location estimator $\boldsymbol{T}$ is the coordinate wise median and the scatter estimator $\boldsymbol{C}$ is the bivariate Gnanadesikan-Kettenring estimator with MAD scale defined as $C_{0n,jk} = \frac{1}{4}(MAD(\{\boldsymbol{z}_{ij} - \boldsymbol{z}_{ik}\})^2 - MAD(\{\boldsymbol{z}_{ij} - \boldsymbol{z}_{ik}\})^2)$ for $k = 1, \ldots, p$, where $MAD(\{\boldsymbol{z}_i\})$ denotes the MAD of $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$. The empirical distribution of the pairwise Mahalanobis Distance is then given by $G_n(t) = \frac{1}{n}\sum_{i=1}^n I(MD_i \leq t)$ and is compared to the reference function $G(t)$ which is taken to be the chi-squared distribution with two degrees of freedom, i.e. $G = \chi_2^2$. Then the proportion of observations that will be flagged as a bivariate outlier is defined by equation (14)

$$d_n = \sup_{t \geq \eta} \{G(t) - G_n(t)\}^+. \tag{14}$$

Now $\eta = G^{-1}(\alpha)$ and $\alpha = 0.85$ because the goal of the bivariate filter is to flag the moderate outliers, since the univariate filter will flag the large outliers. Again $\lfloor d_n n \rfloor$ observations with the largest Mahalanobis Distances will be flagged as outliers and set to NA. The filter is shown to be consistent in Leung et al. (2017).

4.2.3 The Univeriate-Bivariate Filter (UBF)

Leung et al. (2017) combine the univariate and bivariate filter to obtain the Univariate-Bivarite Filter (UBF) in the following way: first the univariate filter is applied to each variable and then the bivariate filter is applied to all the cells that have not been flagged by the univariate filter. If a pair is flagged by the bivariate filter, then it is not yet clear which cell needs to be flagged as outlier. To find the cells that should be flagged, consider the set $J = \{(i, j, k) : MD_i^{(jk)}$ is flagged as bivariate outlier$\}$ with all flagged pairs from the bivariate filter. For each cell (i, j) the number of flagged pairs in which it is involved is counted. The cells that have a large count are probably outliers. The count $m_{ij} = \#\{k : (i, j, k) \in J\}$ for cell (i, j) in a clean observation follows a binomial distribution, $m \sim Bin(\sum_{k \neq j} U_{ik}, \delta)$, with δ the fraction of cellwise outliers undetected by the Gervini-Yohai filter. Leung et al. (2017) flag an outlier if $m_{ij} > c_{ij}$, i.e. the count of flagged pairs exceeds c_{ij}, which is the 0.99-quantile of the aforementioned binomial distribution with the conservative choice of $\delta = 0.1$. This choice of delta yielded good results in the simulation study and in practice, hence it is used in this paper as well. This combination of these filters UBF is also consistent.

4.2.4 The DetectDeviatingCells algorithm

As the *DetectDeviatingCells* algorithm is only a small part of this thesis and the procedure itself is too comprehensive to deal with in its entirety here, I only outline the core steps of the method. A step-by-step description can be found in Rousseeuw and Van Den Bossche (2018). The first step is standardizing each column of the data separately. Following this, univariate outlier detection is applied to identify extreme values using a predetermined cutoff. Correlations between variables are computed, and connected variables are identified based on a specified correlation threshold. Predicted values for each cell are then calculated using a combination rule applied to the connected variables. To counteract any shrinkage caused by the prediction process, adjustments are made to the predicted values. Cells are flagged as outliers based on standardized residuals, and an imputed matrix is generated, replacing flagged cells with predicted values. Rowwise outliers are identified using a criterion derived from the distribution of standardized residuals. Finally, the standardization process is reversed to obtain the final

imputed matrix, along with lists of flagged outliers for further analysis.

4.2.5 Generalized S-estimator

The Generalized S-Estimator (GSE) from Danilov et al. (2012) was originally constructed to compute a robust scatter matrix for an incomplete data set and a target distribution from an elliptical family. As the output of step one is a data matrix with NA values, the GSE allows the estimation of our robust covariance matrix after outlying cells have been flagged.

To define missing or NA values in $\boldsymbol{X}$, we construct an auxiliary matrix $\boldsymbol{U}$ that has entries of zero or one, where zero indicates a missing or NA value. The dimension of the observed part of observation i is then given by $p_i = p(\boldsymbol{u}_i) = \sum_{j=1}^{p} \boldsymbol{u}_j$. For p-dimensional vectors $\boldsymbol{u}$ and $\boldsymbol{m}$ and a $p \times p$ matrix $\boldsymbol{\Sigma}$, the subvectors obtained from the one entries in $\boldsymbol{u}$ are denoted by $\boldsymbol{m}^{(\boldsymbol{u})}$ and $\boldsymbol{\Sigma}^{(\boldsymbol{u})}$. The covariance matrix is normalized such that $|\boldsymbol{\Sigma}^*| = 1$ where $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}/|\boldsymbol{\Sigma}|^{1/p}$. The partial square Mahalanobis is then defined as $d(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\Sigma}) = (\boldsymbol{x}^{(\boldsymbol{u})} - \boldsymbol{m}^{(\boldsymbol{u})})^T (\boldsymbol{\Sigma}^{(\boldsymbol{u})})^{-1} (\boldsymbol{x}^{(\boldsymbol{u})} - \boldsymbol{m}^{(\boldsymbol{u})})$.

Generalizing the S-estimator to allow for missing values yields a definition similar to equations (10) and (11). Let $\hat{\boldsymbol{\Omega}}_n$ be a $p \times p$ positive definite initial estimator for $\boldsymbol{\Sigma}_0$. Also, given the location $\boldsymbol{m}$ and scatter matrix $\boldsymbol{\Sigma}$, we have that $S_n^*(\boldsymbol{m}, \boldsymbol{\Sigma})$ is the solution to equation (15)

$$\sum_{i=1}^{n} c_{p(\boldsymbol{u})} \rho \left(\frac{d(\boldsymbol{x}_i^{(\boldsymbol{u_i})}, \boldsymbol{m}^{(\boldsymbol{u_i})}, \boldsymbol{\Sigma}^{*(\boldsymbol{u_i})})}{s |\hat{\boldsymbol{\Omega}}_n^{(\boldsymbol{u_i})}|^{1/p(\boldsymbol{u_i})} c_{p(\boldsymbol{u_i})}} \right) = \frac{1}{2} \sum_{i=1}^{n} c_{p(\boldsymbol{u})}, \tag{15}$$

where c_p is chosen as in section 4.1 and $p(\boldsymbol{u}) = \sum_{j=1}^{p} \boldsymbol{u}_j$. The multivariate location and scatter estimators are defined in equation (16)

$$(\hat{\boldsymbol{m}}_n, \tilde{\boldsymbol{\Sigma}}_n) = \arg\min_{\boldsymbol{m}, \boldsymbol{\Sigma}} S^*(\boldsymbol{m}, \boldsymbol{\Sigma}). \tag{16}$$

Lastly, the GSE estimator of scatter is defined in the upper line of equation (17), where $\hat{s}_n$ satisfies the lower equation (17)

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_n &= \hat{s}_n \tilde{\boldsymbol{\Sigma}}_n, \\ \sum_{i=1}^{n} c_{p(\boldsymbol{u})} \rho \left(\frac{d(\boldsymbol{x}_i^{(\boldsymbol{u_i})}, \hat{\boldsymbol{m}}^{(\boldsymbol{u_i})}, \tilde{\boldsymbol{\Sigma}}^{*(\boldsymbol{u_i})})}{c_{p(\boldsymbol{u_i})} \hat{s}_n} \right) &= \frac{1}{2} \sum_{i=1}^{n} c_{p(\boldsymbol{u})}. \end{aligned} \tag{17}$$

Danilov et al. (2012) show that the estimator is consistent and they also use the Tukey bisquare rho function tuned to a breakdown point of 50%.

4.3 Minimum Covariance Determinant estimator

The conceptual idea of the casewise Minimum Covariance Determinant (MCD) estimator from Rousseeuw (1985) is to find a subset of $H \leq N$ of observations such that the covariance matrix of the h observations has the lowest determinant of all possible subsets. The resulting set H^* with the lowest determinant is calculated iteratively using concentration steps that decrease the determinant at each iteration. The mean and covariance matrix of H^* are the MCD estimates of location and scatter. To increase the efficiency of the estimator, the resulting estimators of location and scatter are used to compute weights based on statistical distances. Then the reweighted mean and covariance of all observations are computed and yield an estimator that uses many more observations and hence is more efficient. The MCD used in this paper is the reweighted MCD.

The formal definition of the MCD estimator can be expressed as follows: let $\boldsymbol{X}$ be the data matrix, X be the set analogous to the matrix and $H \subseteq X$. Then the sample mean and covariance matrix of subset H are given by equations (18) and (19) respectively. The final estimators for location and scatter $\boldsymbol{T}_{MCD}$ and $\boldsymbol{S}_{MCD}$ are the mean and covariance matrix of the solution H_{MCD} to equation (20). The right side of equations (18) and (19) display the estimators as a weighted mean and covariance matrix of all observations where the weights w_i are one if observation $i \in H$ and zero otherwise

$$\boldsymbol{T}_H = \frac{1}{h}\sum_{i \in H} \boldsymbol{x}_i \qquad = \frac{1}{h}\sum_{i=1}^{n} w_i(\boldsymbol{x}_i), \tag{18}$$

$$\boldsymbol{S}_H = \frac{1}{h}\sum_{i \in \boldsymbol{H}} (\boldsymbol{x}_i - \boldsymbol{T}_H)(\boldsymbol{x}_i - \boldsymbol{T}_H) \qquad = \frac{1}{h}\sum_{i=1}^{n} w_i((\boldsymbol{x}_i - \boldsymbol{T}_H)(\boldsymbol{x}_i - \boldsymbol{T}_H)), \tag{19}$$

$$H_{MCD} = \arg\min_{H:|H|=h} det(\boldsymbol{S}_H). \tag{20}$$

The algorithm used to compute the MCD estimator is called the `Fast-MCD` algorithm and is proposed in Rousseeuw and Driessen (1999). [1] The algorithm starts with an initial (random) subset $H_1 \subset X$ and uses these observations to compute the mean and covariance matrix. Then based on these estimates the algorithm computes the distances of all observations and orders these distances from small to large. The h observations with the smallest distances are used as the new subset H_2. The resulting determinant of the covariance matrix of H_2 is at least small

[1] Faster algorithms have been proposed by Hubert et al. (2012) and De Ketelaere et al. (2020) in which the initial subsets are not random, but computed with known estimators that can handle different types of contamination. Although these methods are faster, I stuck with the `Fast-MCD` algorithm as it did not cause computational issues and the implementations of the faster algorithms were not readily available in `R`.

as the previous determinant. Formally, $\det(H_m) \geq \det(H_{m+1})$ where equality holds if and only if $\boldsymbol{T}_m = \boldsymbol{T}_{m+1}$ and $\boldsymbol{S}_m = \boldsymbol{S}_{m+1}$. The proof can be found in the Appendix of Rousseeuw and Driessen (1999).

For the regular MCD with data following a normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, Butler et al. (1993) have shown that $\boldsymbol{T}_{MCD}$ is Fisher consistent with the mean $\boldsymbol{\mu}$. The estimator of scale $\boldsymbol{S}_{MCD}$ is not Fisher consistent with $\boldsymbol{\Sigma}$, but it needs a consistency correction c_α (Butler et al., 1993). Additionally, Pison et al. (2002) show that the MCD also needs a small sample correction c_{np} with $c_{np} \to 1$ if $n \to \infty$. The maximum breakdown point of the MCD estimator is attained when subset size $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ and the resulting breakdown point for both location and scatter is $\frac{1}{n}\left\lfloor \frac{n-p+1}{2} \right\rfloor$. The Influence Function of the MCD is shown to be bounded in Croux and Haesbroeck (1999).

4.4 Cellwise MCD estimator

The objective function in the Cellwise MCD in Raymaekers and Rousseeuw (2023) differs from the casewise MCD because the former is computed with cellwise weights instead of casewise weights. The weights are set to zero during the minimization of the objective function if the positive contribution of that data cell to the objective is too large. To make sure that the algorithm does not flag too many cells as outliers a penalty term is added that penalizes the number of flagged cells. The resulting estimators of location and scatter are computed using data cells that have not been flagged as outliers.

As we have seen in equations (18) and (19), the regular MCD estimators of location and scatter can be expressed as a weighted mean and covariance matrix, where all observations in subset H_{MCD} have weight one and all other observations in $X \setminus H_{MCD}$ have weight zero. Then minimizing the negative log likelihood is equivalent to minimizing the determinant of the weighted covariance matrix. When each cell obtains its own weight w_{ij} and a term that penalizes the number of cells with a zero weight is added, the objective function that we minimize is given in equation (21)

$$\sum_{i=1}^{n}(\ln|\boldsymbol{\Sigma}^{(\boldsymbol{w}_i)}| + d^{(\boldsymbol{w}_i)}\ln 2\pi + MD^2(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})) + \sum_{j=1}^{d} q_j||1_d - \boldsymbol{W}_{\cdot j}||_0, \\ \text{such that } \lambda_d(\Sigma) \geq a \text{ and } ||\boldsymbol{W}_{\cdot j}||_0 \geq h \text{ for all } j = 1, \ldots, n. \tag{21}$$

Hhere $\boldsymbol{w}_i$ is the vector with weights for observation i, $d^{(w_i)}$ is the dimension of the weighted observation (i.e. the number of cells that have weight one), q_j is the hyperparameter defining

how strongly each flagged outlier is penalized, the operator $||\boldsymbol{A}||_0$ counts the number of non-zero elements in matrix $\boldsymbol{A}$ and $\text{MD}(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\left(\boldsymbol{x}_i^{(\boldsymbol{w}_i)} - \boldsymbol{\mu}_i^{(\boldsymbol{w}_i)}\right)' \left(\Sigma^{(\boldsymbol{w}_i)}\right)^{-1} \left(\boldsymbol{x}_i^{(w_i)} - \boldsymbol{\mu}_i^{(\boldsymbol{w}_i)}\right)}$ is the partial Mahalanobis Distance as defined in Danilov et al. (2012). The first constraint ensures that the resulting matrix $\hat{\boldsymbol{\Sigma}}$ is non-singular and the second constraint ensures at most $n-h$ cells are flagged as outliers per variable. Note that setting $q_j = 0$ and $w_{ij} = w_i$ for all $j = 1, \ldots, k$ yields the casewise MCD of Rousseeuw (1985).

The choice for the penalty term q_j flows naturally from the set up of the objective function. To determine whether the weight of a cell is zero or one, the algorithm computes the difference in objective function when the weights of the cell are set to one and zero and the option yielding the lowest value is chosen. After rewriting (21) to $\sum_{i=1}^{n} \tilde{L}(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{q})$ with

$$\tilde{L}(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{q}) = \ln |\boldsymbol{\Sigma}^{(\boldsymbol{w}_i)}| + d^{(\boldsymbol{w}_i)} \ln 2\pi + MD^2(\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sum_{j=1}^{d} q_j |1 - w_{ij}|, \tag{22}$$

the resulting difference in the objective function when setting $w_{ij} = 1$ or $w_{ij} = 0$ is

$$\Delta_{ij} = \ln C_{ij} + \ln 2\pi + \frac{(x_{ij} - \hat{x}_{ij})^2}{C_{ij}} - q_j. \tag{23}$$

Here $\hat{x}_{ij}$ is the expected value of x_{ij} conditional on the cells that have a weight of one and C_{ij} is the estimate of the conditional variance given all other observations. Since the default weight is set to one to ensure a higher efficiency, we only set the weight to zero if the cell is an outlier. That is the case if the standardized residual $\frac{(x_{ij} - \hat{x}_{ij})^2}{C_{ij}}$ is beyond a certain threshold. Here the residual follows a chi-square distribution with one degree of freedom, such that the natural choice of q_j becomes as in equation (24), where the conditional variance C_{ij} is approximated by the initial estimator $\hat{\boldsymbol{\Sigma}}_0$ such that C_j is the diagonal element of variable j

$$q_j = \chi^2_{1,0.99} + \ln 2\pi + \ln C_j. \tag{24}$$

The properties of the Cellwise MCD are repeated here from Raymaekers and Rousseeuw (2023). The location estimate of the Cellwise MCD is Fisher consistent, but the scale estimate is not since the penalty term will always cause some observations to be flagged as outliers. The upper bounds of the breakdown points of estimators in casewise contaminated samples also hold under cellwise contamination as the former could be considered a special case of the latter. The cellwise implosion breakdown point is 1 because the first constraint in equation (21) ensures that it does not implode. The location and cellwise explosion breakdown point are the same at $\frac{n-h+1}{n}$. The influence function of the Cellwise MCD is yet to be investigated.

4.5 Stahel-Donoho estimator

The Stahel-Donoho (SD) estimator was originally proposed by Stahel (1981) and Donoho (1982) independently and assigns weights to observations based on their outlyingness. Here I follow the description of the SD estimator as given in Van Aelst et al. (2011). If μ and σ are shift and scale equivariant, univariate estimators of location and scatter, then the outlyingness for any $y \in \Re^p$ is defined as in equation (25)

$$r(\boldsymbol{y}, \boldsymbol{X}) = \sup_{\boldsymbol{a} \in S^p} \frac{|\boldsymbol{a}'\boldsymbol{y} - \mu(\boldsymbol{a}'\boldsymbol{X})|}{\sigma(\boldsymbol{a}'\boldsymbol{X})}, \tag{25}$$

where the set S^p denotes the set of linear combinations of dimension p normalized to 1. The SD estimator of location and scale is defined as

$$\begin{aligned} \boldsymbol{T}_{SD} &= \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}, \\ \boldsymbol{S}_{SD} &= \frac{\sum_{i=1}^n w_i (x_i - \boldsymbol{T}_{SD})(x_i - \boldsymbol{T}_{SD})}{\sum_{i=1}^n \sqrt{w_{ij}}}, \end{aligned} \tag{26}$$

where the weights depend on the outlyingness of the observation. The original SD estimator computes the weights for each observation based on a Huber-type weight function as advocated by Maronna and Yohai (1995). The weight function is given in equation (27) with the outlyingness as input

$$w(r) = I_{r \leq c} + (\frac{c}{r})^2 I_{r > c}, \tag{27}$$

where c is a threshold beyond which the outlyingness is Huberized and is taken to be $c = min(\sqrt{\chi^2_p}(0.5), 4)$. Stahel (1981) proved that as long as the estimators μ and σ in (25) have an asymptotic breakdown point of 0.5, then the resulting estimator would inherit this breakdown point. Therefore the estimators for the location and scatter used are often the median and a Modified Median Absolute Deviation (MMAD) defined in equation (28), with the correction factor $\beta = \Phi^{-1}(\frac{1}{2}((n + p - 1)/2n + 1)$ and $\Phi^{-1}(.)$ the inverse CDF of the standard normal distribution

$$MMAD(\boldsymbol{a}'\boldsymbol{X}) = \frac{|\boldsymbol{a}'\boldsymbol{X} - Median(\boldsymbol{a}'\boldsymbol{X})|_{\lceil (n+p-1)/2 \rceil : n} + |\boldsymbol{a}'\boldsymbol{X} - Median(\boldsymbol{a}'\boldsymbol{X})|_{(\lfloor (n+p-1)/2 \rfloor + 1) : n}}{2\beta}. \tag{28}$$

4.6 Stahel-Donoho estimator with cellwise weights

The weights for the SD estimator with cellwise weights are computed with the same weight function $w(r)$ as the classical SD estimator, but the cellwise outlyingness r_{ij} becomes a weighted average of original the outlyingness of the observation r_i and the outlyingness of observation i within variable j denoted as c_{ij}. Equation (29) shows the adapted outlyingness and equation (30) gives the definition of c_{ij}.

$$r_{ij} = \alpha_{ij} r_i + (1 - \alpha_{ij}) c_{ij} \tag{29}$$

$$c_{ij} = \frac{|x_{ij} - Median(\boldsymbol{X}_j)|}{MAD^*(\boldsymbol{X}_j)} \tag{30}$$

Since the outlyingness in the direction of the variable j (c_{ij}) is a subset of the directions considered in (25), we have that $c_{ij} \leq r_i$ such that $r_{ij} \leq r_i$ and hence the cellwise weights are larger or equal to the casewise weights. Van Aelst et al. (2011) propose two relevant options to determine the parameters α_{ij}:

1. $\alpha_{ij} = (max_{k=1}^{p} c_{ik})^{-1} c_{ij}$, which implies that a_{ij} is large if outlyingness in variable j is high relative to the outlyingness of other components. The cell receives approximately the weight of the original estimator. On the other hand, if the component is not an outlier and hence c_{ij} is small, then α_{ij} will be close to 1 and the weight of the cell is increased. The resulting estimator from this option is called the Stahel-Donoho Components (SDC).

2. $\alpha_{ij} = (max_{k=1}^{p} |u_{ik}|)^{-1} |u_{ij}|$, where $|u_{ij}|$ is the direction that maximizes r_i. This implies that components that are responsible for the regular outlyingness obtain a lower weight. This is called the Stahel-Donoho Maximizing (SDM).

Once the outlyingness and weights have been computed, one can compute the resulting estimator for location and scatter from equations (31) and (32)

$$\boldsymbol{T}_{SD^*,j} = \frac{\sum_{i=1}^{n} w_{ij} x_{ij}}{\sum_{i=1}^{n} w_{ij}}, \tag{31}$$

$$\boldsymbol{S}_{SD^*,jk} = \frac{\sum_{i=1}^{n} \sqrt{w_{ij}} \sqrt{w_{ik}} (x_{ij} - \boldsymbol{T}_{SD^*,j})(x_{ij} - \boldsymbol{T}_{SD^*,j})}{\sum_{i=1}^{n} \sqrt{w_{ij}} \sqrt{w_{ik}}}. \tag{32}$$

Note that setting all weights $w_{ij} = w_i = w(r_i)$ for all $j = 1, \ldots, p$ results in the classic SD estimator. The breakdown point is still 0.5 due to the choice of the median and MAD for the cellwise outlyingness.

Casewise Estimators:	S-estimator	MCD	Stahel-Donoho
Cellwise Estimators:	TSGS	Cellwise MCD	SD with cellwise weights

Table 2: Overview of the robust covariance estimators that are used as building blocks for the CRIV estimator in equation (33)

4.7 Overview of the estimators

Before continuing to the simulation experiment I present an overview of the estimators used in the following sections. Table 2 shows which casewise and cellwise estimators are used in the simulation exercise and Table 7 in the Appendix shows the detailed specifications of tuning parameters and used packages. Each method is used to estimate a cellwise robust scatter matrix $\hat{\boldsymbol{C}}$, which is then used to compute the Cellwise Robust Instrumental Variable (CRIV) estimator as in equation (33)

$$\hat{\boldsymbol{\beta}}_{CRIV} = \left[\hat{\boldsymbol{C}}_{\boldsymbol{XZ}}\hat{\boldsymbol{C}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{C}}_{\boldsymbol{ZX}}\right]^{-1}\left[\hat{\boldsymbol{C}}_{\boldsymbol{XZ}}\hat{\boldsymbol{C}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{C}}_{\boldsymbol{Zy}}\right]. \tag{33}$$

5 Simulation

The simulation setup here is similar to the simulation setup in Section 3.2. The goal of the simulation exercise is to compare the performance of cellwise robust scatter estimators to casewise robust scatter estimators. For each scenario, $r = 1.000$ samples of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\varepsilon})$ with size $n = 250$ are generated from a multivariate normal distribution with mean zero. The covariance matrices used to simulate the data sets can be found in the Appendix A.2. After the variables have been simulated, the dependent variable is computed according to $\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{\beta} + \varepsilon$. Lastly, the dataset is contaminated according to the scheme described below. The focus is laid on the estimation of the coefficient of the endogenous variable, the true value for the coefficient is two in all scenarios, i.e. $\beta_{End} = 2$.

I look at three different scenarios: in the first scenario there is only one endogenous, one instrumental and one control variable as a baseline. The dimensions are kept small and this is meant to give an impression of the estimators' performances when the model is simple. Often a model is regressed with and without control variables, scenario 1 is supposed to represent the regression without the control variables. An example of this is given in Table 6 of Alesina and Zhuravskaya (2011).

In scenario 2 the model contains five control variables to see what happens to the estimators when the dimension increases. The main reason this model specifications with a higher number of control variables is investigated is that the curse of dimension affects the datasets heavily under

	Cont. magnitude k	Cont. rate ϵ	Endogenous	Instrumental	Control
Scenario 1	$k = 3$ and $k = 10$	$\{0, 0.05, \ldots, 0.3\}$	1	1	1
Scenario 2	$k = 3$ and $k = 10$	$\{0, 0.05, \ldots, 0.3\}$	1	1	5
Scenario 3	$k = 3$ and $k = 10$	$\{0, 0.05, \ldots, 0.3\}$	1	3	3

Table 3: Overview of the simulation scenarios that are adopted in this thesis. Simulation is done with both magnitudes seperately for all contamination rates. The number of variables for each type of variable are given in the last three columns.

cellwise contamination. As visible in Table 1, if 10% of the cells in a variable are contaminated, then 57% of the cases is expected to contain an outlier. Cellwise robust estimators should be able to deal with this high number of contaminated cases, which is what is investigated in this scenario.

Scenario 3 considers a dataset of the same size as scenario 2, but now there are three instruments that are highly correlated with the endogenous variable. This is supposed to mimic the situation where lagged variables are used as instruments for endogenous variables, which is becoming increasingly popular in political science and economics (Bellemare et al., 2017). If the variable is persistent then the lagged values are highly correlated with the endogenous variable and are thus strong instruments, given the instruments fulfil the other conditions. In the IV model without contamination, strong instruments yield more accurate estimates since $\boldsymbol{\Sigma_{XZ}}$ is larger. This statement is what I intend to analyze in scenario 3 under cellwise contamination. One practical implication could be that if estimates become more accurate with additional instruments, it can be beneficial to invest into finding for more instruments. Table 3 provides an overview for the different simulation scenarios investigated in this section.

The data is asymmetrically contaminated by replacing cells $x_{ij}^{cont} = k$ with $k = 3$ and $k = 10$ to examine the performance of the estimators when the outliers are marginal and extreme respectively. It can be that an observation is from another distribution, but it is close enough to tail of the target distribution that it might not be flagged as an outlier. These outliers in the distributional neighbourhood can influence outcomes, hence they are examined here. Each variable is contaminated separately with a probability of ϵ, where ϵ will vary from 0 to 0.3 with 0.05 increments. Specifically, for each sample a matrix with the same size as the full data set $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y})$ is constructed by individually simulating the columns from a random binomial distribution with probability ϵ. The entries that are one will be replaced by k. In all repetitions, the same contaminated data set is used to compute each scatter matrix estimate. Other forms of contamination such as symmetric contamination are also interesting, but I choose to focus on the implications of different contamination magnitudes.

The performance measures considered here are the bias, variance and MSE with respect to

the coefficient for the endogenous variable. The bias is computed as the average distance the estimate is from the true value, i.e. $bias = \frac{1}{r}\sum_{j=1}^{r} \hat{\boldsymbol{\beta}}_{CRIV,End} - \boldsymbol{\beta}_{End}$. The variance measures the average squared distance from the true value and can be interpreted as the accuracy of the estimator. Formally, it is computed as $variance = \frac{1}{r}\sum_{j=1}^{r}(\hat{\boldsymbol{\beta}}_{CRIV,End} - \boldsymbol{\beta}_{End})^2$. The Mean Squared Error (MSE) is computed from the bias and variance and acts as a summary of both measurements defined by $MSE = bias^2 + variance$. Additionally, Section 5.4 will investigate the efficiency of the proposed estimators.

5.1 Scenario 1: one endogenous, one instrumental and one control variable

The simulation results for scenario 1 are given in Figures 3 and 4. In general, all estimators perform reasonably well when the contamination rate is 10% or lower, since their bias and variance are roughly close to its variance under no contamination. This holds for the case of marginal outliers ($k = 3$) and extreme outliers ($k = 10$). The cellwise estimators perform much better in the latter case when the contamination rate increases. In both figures the downward trend when the contamination rate increases is caused by the fact that outliers distort the covariance between variables towards zero, hence the estimates go to zero. The SD estimators suffer from this already at low contamination rates. Van Aelst et al. (2011) mention in their paper that the SDM and SDC methods have difficulties in identifying all outlying components in structural outliers. The consequense is that some outlying cells will receive a large weight, which will increase the bias of the estimate. This is likely what happens here too. It is important to note that the SDC estimates were practically similar to the normal SD estimates, hence we report only on SDM estimates.

Results for the case where the outliers are marginal are shown in Figure 3. The estimators cannot distinguish properly between outlying values and extreme values of the target distribution. There is no estimator that remains unbiased when the contamination rate is 10% or higher. The variances of the estimates for almost all estimators increase sharply with the contamination rate, except for the SD estimator, however, its estimates are heaviliy biased towards zero. The TSGS and Cellwise MCD do not necessarily perform better than their casewise counterparts based on the MSEs, while the SDM estimator performs worse than its casewise counterpart in all regards.

When outliers are extreme the performance of the cellwise estimators increases strongly with the TSGS yielding the best results. The biases are generally lower compared to the situation where outlying cells are marginal outliers. Casewise estimators break down when the contamination rate increases but the variance of TSGS estimates remains stable with the contamination

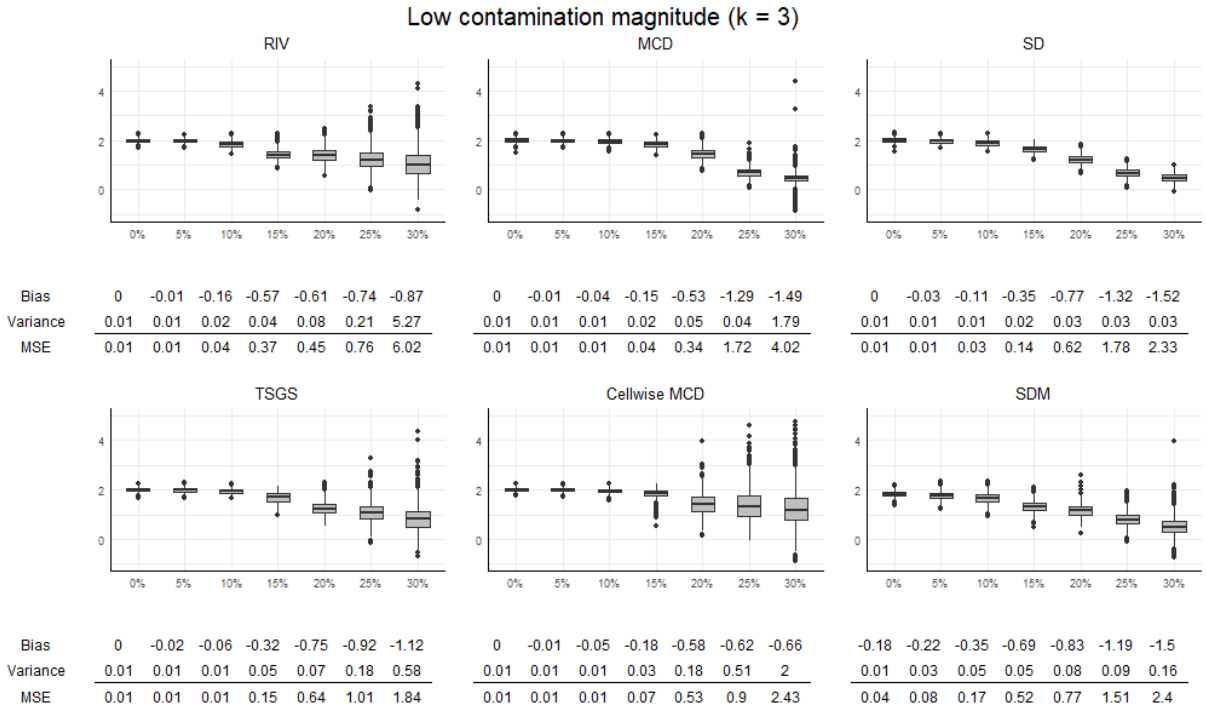


Figure 3: Boxplots with coefficients for the endogenous variable in scenario 1 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true $\beta = 2$ and the covariance estimators used are found above each graph.

rate of up to 25% without becoming severely biased. The Cellwise MCD estimates performs well until 20% contamination but become inaccurate when the rate increases beyond that. The classic RIV estimates have a larger bias and variance compared to cellwise methods, which is evidence that the CRIV can be preferable in certain situations.

5.2 Scenario 2: one endogenous, one instrumental and five control variables

The results for scenario two are depicted in Figures 5 and 6. In this scenario the number of control variables has increased. This implies that the probability that an observation contains an outlier is increased from 0.34 to 0.57 when the contamination rate is 10% due to the increase in dimension. As all casewise estimators are tuned to have a breakdown point of 50%, they are not expected to perform well at a contamination rate than 10%.

The results for the moderate outliers are given in Figure 5. Both MCD estimators remain relatively unbiased with a low variance at a contamination rate of 10%, while all other estimators become biased. This may be because the estimates of the variance of the subset of included cells becomes more accurate when the dimension is larger. Overall the estimators cannot distinguish marginal outliers from uncontaminated cells and no method performs well at a contamination rate of 15% and higher.

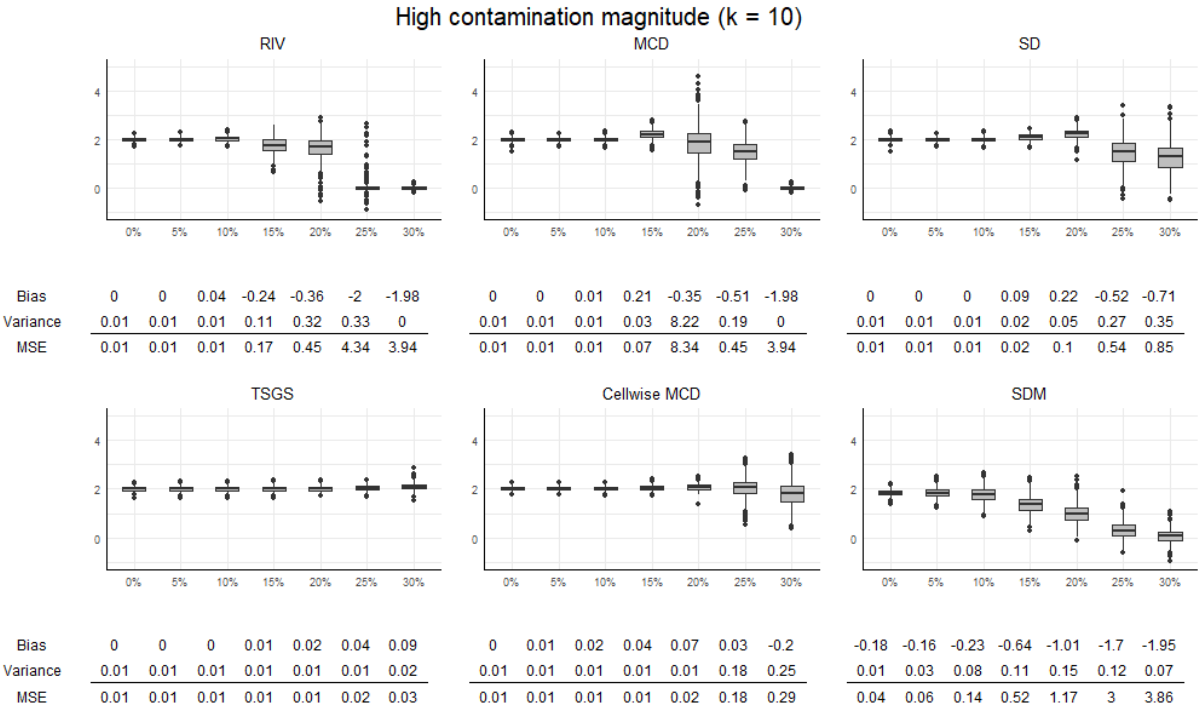


Figure 4: Boxplots with coefficients for the endogenous variable in scenario 1 for contamination rates from 0% to 30%. The contamination magnitude is high and set at $k = 10$.The true $\beta = 2$ and the covariance estimators used are found above each graph.

Figure 6 shows the results for contamination with extreme outliers. The RIV becomes already slightly biased at 10% while the casewise MCD and SD perform relatively well compared to their performance with marginal outliers. The main improvement can be seen in the performances of the TSGS and Cellwise MCD, they remain unbiased until contamination hits 25% and the variance remains equal to the case when there is no contamination. The bias of the TSGS is lower than the bias of the Cellwise MCD such that the former performs best in this case. Nonetheless, these results imply that both methods provide reliable estimates even though 83% of the observations is expected to be contaminated and shows that when cellwise contamination is suspected, cellwise robust estimators should be used.

Compared to scenario 1, the variances of the estimates are larger in scenario 2. This indicates that the curse of dimensionality outweighs the increased efficiency from more variables. For example, when $k = 10$ the TSGS estimators' variance remains stable when the contamination rate is higher than 20% in scenario 1, while the estimates are less accurate at the same levels of contamination in scenario 2. For the MCD based estimators this difference in accuracy is only apparent when the outliers are marginal, i.e. the cellwise MCD suffers less from the curse of dimensionality when outliers are sure to be flagged. Based on the results in scenarios 1 and 2, the Cellwise MCD performs better when the outliers are only marginal, while the TSGS outperforms all estimators when outliers are extreme. Both results are in line with the findings

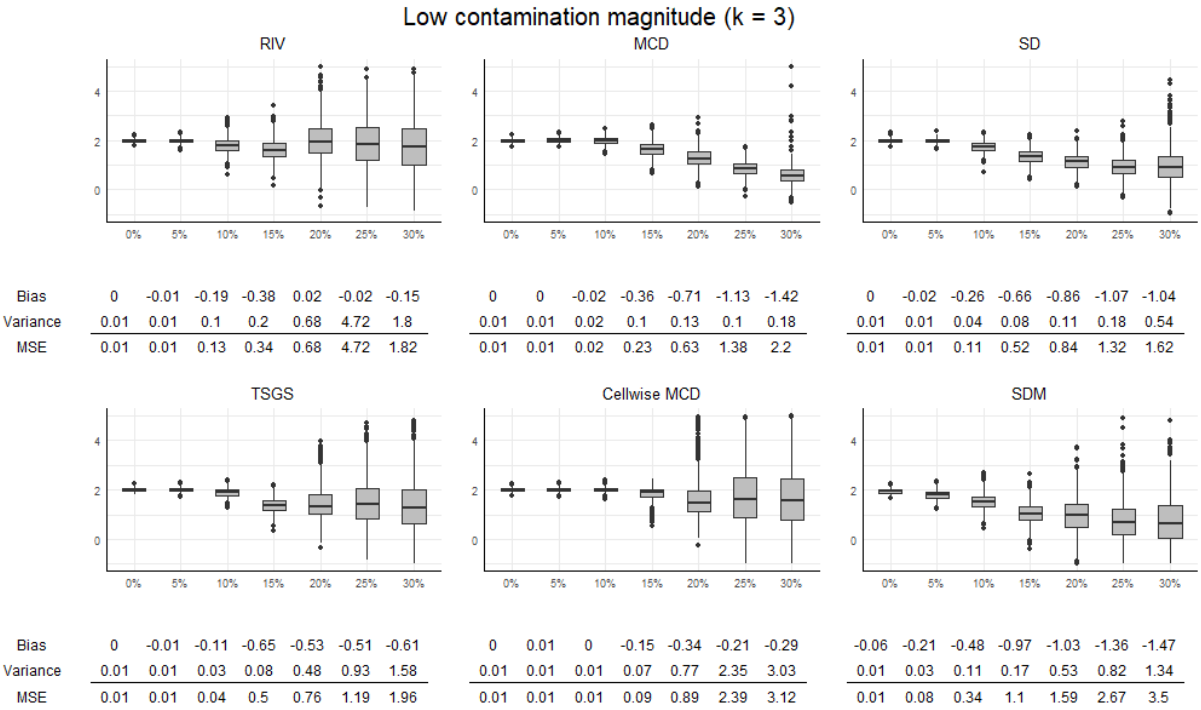


Figure 5: Boxplots with coefficients for the endogenous variable in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true $\beta = 2$ and the covariance estimators used are found above each graph.

in Raymaekers and Rousseeuw (2023).

5.3 Scenario 3: one endogenous, three instrumental and three control variables

The results for the last scenario are shown in Figures 7 and 8. The casewise methods still are not performing well when contamination hits levels above 5%. The difference with scenario 2 is that there are now three variables that are highly correlated with the endogenous variable, hence estimates are expected to be more accurate than in scenario 2. The results for all estimators are in line with aforementioned statement, since in scenario 3 the variance is lower for both the marginal and extreme outlier case than in scenario 2. Besides that the results are similar to scenario 2, for marginal outliers the Cellwise MCD performs best and for extreme outliers the TSGS estimator performs best. However, for marginal outliers the estimators become inaccurate when the contamination rate is higher than 10%.

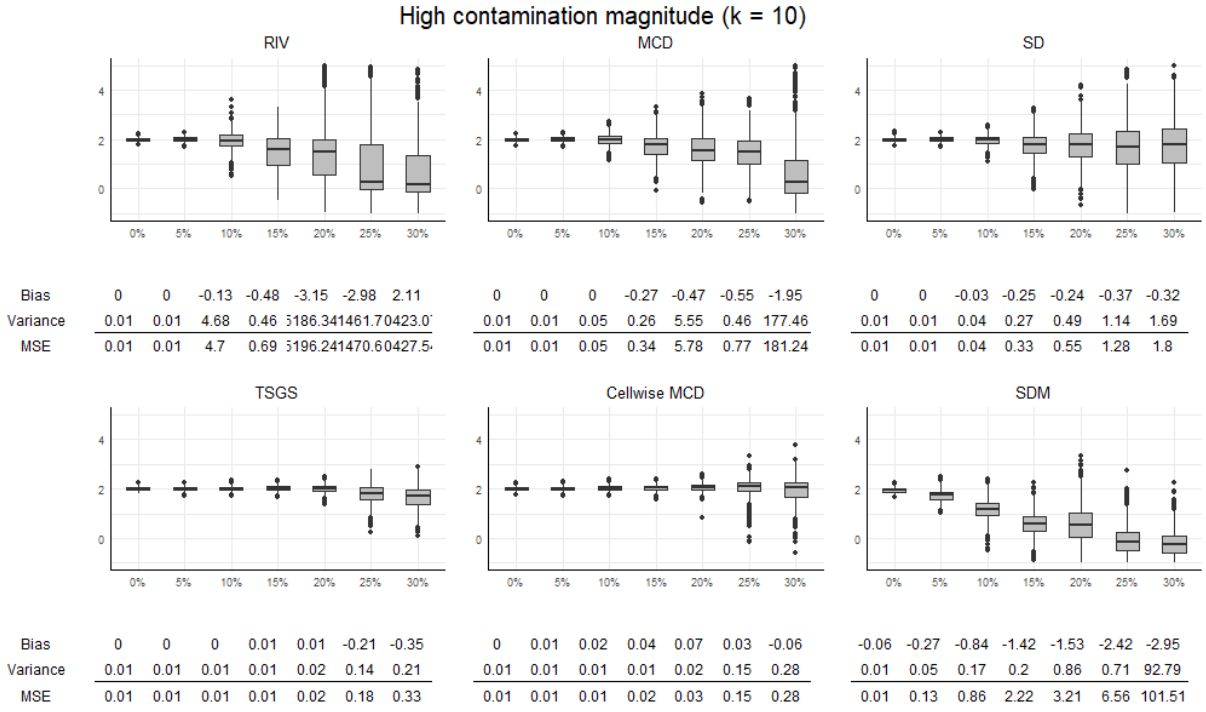


Figure 6: Boxplots with coefficients for the endogenous variable in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is high and set at $k = 10$.The true $\beta = 2$ and the covariance estimators used are found above each graph.

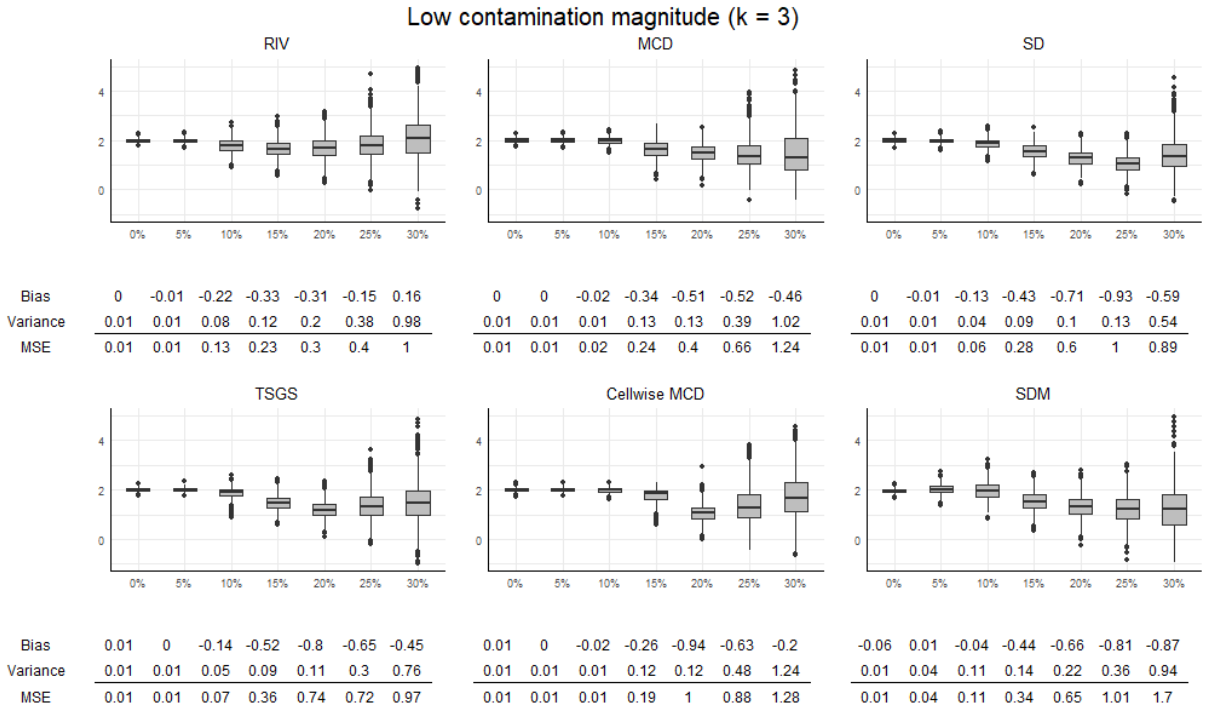


Figure 7: Boxplots with coefficients for the endogenous variable in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true $\beta = 2$ and the covariance estimators used are found above each graph.

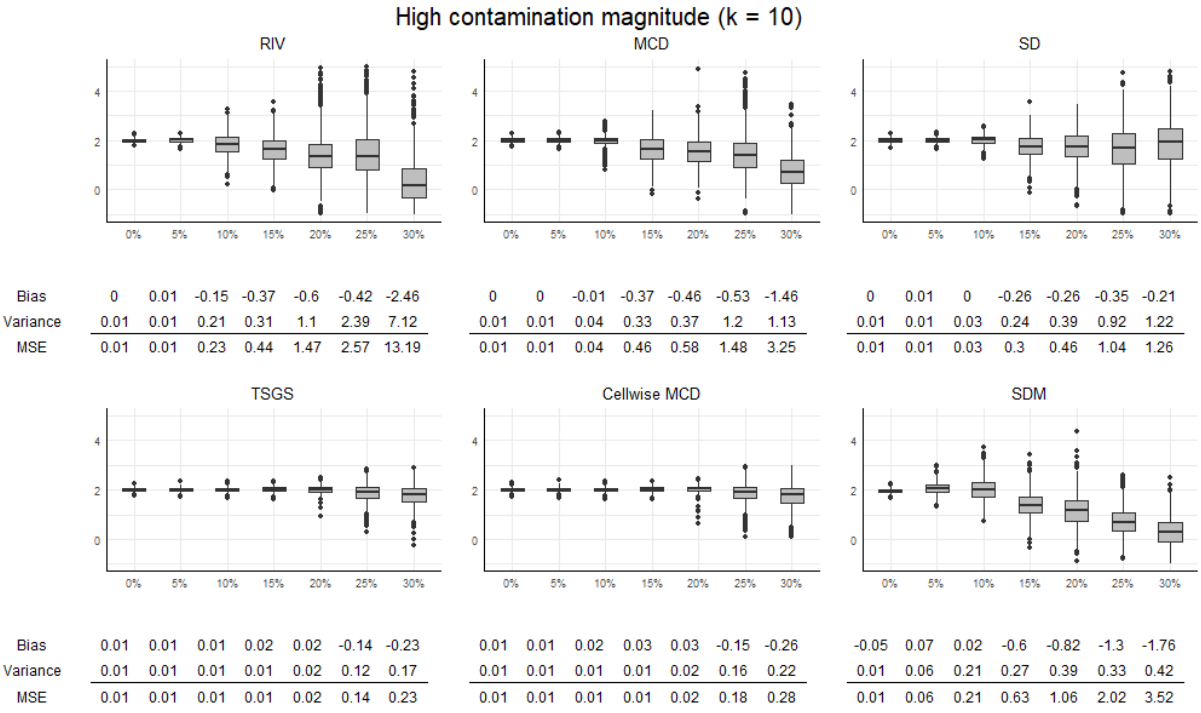


Figure 8: Boxplots with coefficients for the endogenous variable in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is high and set at $k = 10$.The true $\beta = 2$ and the covariance estimators used are found above each graph.

5.4 A note on efficiency

Another interesting property is the efficiency of the estimator. As many robust estimators downweight or exclude observations or in our case data cells, this is paid for with efficiency: less information is used hence the estimate is less efficient. To investigate the efficiency of the estimators, the variance of the estimates is compared to the most efficient estimator when the dataset is clean. Since the datasets in all three scenarios come from a normal distribution, the most efficient estimator is the Maximum Likelihood Estimator (MLE). The MLE is the regular covariance matrix and the covariance between two variables are computed as in equation (34)

$$Cov(X_i, X_j) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_{X_i})(X_j - \mu_{X_j}), \tag{34}$$

where μ_X is the regular mean of X. The results are given in Table 4. The top row in each scenario depicts the efficiency of each estimator relative to the MLE, that is, the variance of the MLE estimator is divided by the variance of the robust estimators. The second row of each scenario shows the variances. The efficiency increases as the dimension increase from scenario 1 to scenario 2 and 3. For scenario 1 the Cellwise MCD is the most efficient at 57.69% and for the larger covariance matrices in scenarios 2 and 3 the S-estimator of the RIV yields the highest efficiency around 88.32%. However, the Cellwise MCD is almost as efficient at 86.62%. The

	MLE	RIV	MCD	SD	TSGS	CellMCD	SDM	SDC
Scenario 1:	100.00%	53.49%	32.89%	38.03%	50.79%	**57.69%**	35.11%	52.43%
	(0.0137)	(0.0257)	(0.0418)	(0.0361)	(0.0270)	**(0.0238)**	(0.0391)	(0.0262)
Scenario 2:	100.00%	**88.36%**	37.30%	46.37%	71.26%	75.86%	45.65%	59.87%
	(0.0134)	**(0.0151)**	(0.0359)	(0.0289)	(0.0188)	(0.0176)	(0.0293)	(0.0223)
Scenario 3:	100.00%	**88.32%**	49.32%	57.34%	79.17%	86.62%	53.65%	54.75%
	(0.0134)	**(0.0152)**	(0.0272)	(0.0234)	(0.0170)	(0.0155)	(0.0250)	(0.0245)

Table 4: The efficiencies of the estimators in the different scenarios relative to the MLE. The variances are given in brackets below the efficiencies.

efficiency of the TSGS estimator is in the top three for scenarios 2 and 3 and is relatively close to the top three in scenario 1 such that the efficiency loss is not large for the TSGS either.

6 Practical Application

This section applies the proposed estimator to a real world dataset. The goal of this section is twofold: first to compare the results of the CRIV estimator to the RIV estimator in an applied setting and second to illustrate how the matrices with flagged cells can enhance the interpretation of casewise weights.

The application is from Romer (1993) and it investigates the link between trade openness and inflation using cross-sectional data. The idea of the paper stems from the following reasoning: in an open economy, a positive monetary shock will cause the exchange rate of a country to depreciate and this will reduce the incentive to increase output. If a country were to coordinate a positive monetary shock with another country such that both currencies will not depreciate, there will be an incentive to increase output in both countries. By coordinating their policies the two countries de facto function as one larger economy and are hence less open. This suggests a inverse relationship between trade openness and the inefficiently high inflation caused by unexpected monetary shocks, for example monetary expansions without precommitment.

The data used to estimate the relationship are the inflation, as measured by the average annual change in log GDP denominator since 1973, and openness, as measured by the average share of import in GDP since 1973. The dataset runs until 1992. Since there are some countries that are outlying, Romer (1993) chose to regress log inflation on openness to reduce the influence of outliers. The instrument that is used is the logarithm of the land area in square miles. Graphical representations of the dataset are given in Section A.5 in the Appendix.

The estimation results are given in Table 5. The OIV estimator yields a negative coefficient for openness which is statistically significant at the 5% level. The RIV also estimates a negative

	OIV	RIV	MCD	SD	TSGS	CellMCD	SDM
Intercept	2.9898	2.7793	2.8533	2.7481	2.8584	2.5141	-0.0094
	(0.1610)	(0.2160)	-	-	-	-	-
Openness	-1.3158	-1.1171	-1.3388	-1.0682	-1.4673	-0.6068	8.0344
	(0.3992)	(0.6741)	-	-	-	-	-

Table 5: Estimation results for regressing the logarithm of inflation on trade openness using the logarithm of land area as an instrument. Standard errors are given in parenthesis where available.

coefficient, but it is not significant anymore as the p-value has increased to 0.0995. The TSGS estimator shows that the coefficient is lower than its casewise alternative. Although the IF of the TSGS estimator has not been derived yet, we could take the standard errors from the RIV estimator as proxy for the standard errors of the TSGS estimator. Although the influence function of the TSGS estimator has to be conditioned on the filtered cells, the main difference between the TSGS estimator and S-estimator is in the input, not in the procedure of computation such that it is reasonable to assume the influence functions are similar and hence the asymptotic variance. Using the standard errors from the S-estimator, the TSGS estimate becomes significant at the 5% level. This is a reaffirmation of a negative relation between openness and inflation, something that is later confirmed in other research: Alfaro (2005) investigate the same period with different techniques and find that in the long run there is a negative relationship and Bowdler and Malik (2017) find a statistically significant negative effect in a more recent sample.

Another benefit of using cellwise robust methods is that the methods naturally provide weights for each data cell, while the RIV estimator only provides weights for the full observation. For the TSGS estimator, one could also run the UBF without estimating the variance, however, it is convenient that the matrices with flagged cells are a byproduct of the CRIV estimation. The same goes for the Cellwise MCD matrix of flagged cells. Nonetheless, the major advantage of these matrices is that they imply which component of the observation is outlying and which is not, whereas the RIV weights only show which observation is outlying.

To illustrate this, Table 6 shows the weights of the RIV estimator and flagged cells from the TSGS and CellwiseMCD estimator. The table only depicts the possible outliers, i.e. observations with the lowest RIV weights. The outliers were chosen by graphical examination, but the point here is to show the additional insights obtained from the CRIV methods, not to determine whether the observations are actually outliers. For example, Argentina and Bahrein both receive a zero weight from the RIV estimator but both cellwise estimators indicate that they are outlying for different reasons. Bahrein is a small country that depends heavily on imports, while Argentina has seen soaring inflation rates throughout the sample period. Although this makes sense theoretically, the outcomes of these flagging procedures provide statistical grounds.

Another insight that would be harder to obtain from the RIV weights only, is that most countries in Latin America have been flagged due to their extremely high inflation in the sample period. This showcases the added value of the CRIV methods.

		TSGS:			Cellwise MCD:		
country	RIV Weight	Openness	Log Land	Inflation	Openness	Log Land	Inflation
Argentina	0	0	0	1	0	0	1
Bahrein	0	1	1	0	1	1	0
Barbados	0	0	1	0	1	1	0
Bolivia	0	0	0	1	0	0	1
Brazil	0	0	0	1	0	0	1
Hong Kong	0	1	1	0	1	1	0
Israel	0	0	0	1	0	0	1
Jordan	0	1	0	0	1	0	0
Lesotho	0	1	0	0	1	0	0
Malta	0	1	1	0	1	1	0
Mauritania	0	0	0	0	1	0	0
Singapore	0	1	1	0	1	1	0
Swaziland	0	1	0	0	1	0	0
Guyana	0	0	0	0	1	0	0
Zaire	0.0001	0	0	1	0	0	1
Botswana	0.0002	0	0	0	1	0	0
Luxembourg	0.0006	1	1	0	1	0	0
Chile	0.0014	0	0	1	0	0	1
Mauritius	0.0019	0	1	0	0	0	0
Peru	0.0029	0	0	1	0	0	1
Japan	0.0031	0	0	0	0	0	0
United States	0.0031	0	0	0	0	0	0

Table 6: Table with an overview of RIV weights and TSGS and Cellwise MCD flagged cells for outlying countries. 1 means flagged as outlier, 0 means regular data cell.

7 Conclusion

This paper has investigated whether robustifying the solution equations to the Instrumental Variable model under cellwise contamination yields robust estimates. In the spirit of Freue et al. (2013), the IV estimator was rewritten as a function of the covariance matrix of the endogenous, exogenous, instrumental and dependent variable. Replacing the covariance matrix with robust estimators of scatter, the IV estimator is robustified in a natural way. The robust estimators used in this paper are three different kinds of estimators that have a popular casewise alternative. The first is the Two-Step Generalized S-Estimator which is a cellwise alternative to the S-estimator, the second is the Cellwise MCD which is based on the popular MCD estimator and lastly the SD estimator with cellwise weights is used as the cellwise counterpart to the Stahel-Donoho estimator.

The performance of the estimators was assessed in an extensive simulation study. The simulation study involved three different model specifications, two different types of outliers and seven contamination rates. A model with only one endogenous, instrumental and control variable was the starting point, then it was extended to have more control variables and finally some control variables were substituted for instruments. All three models were simulated and contaminated with marginal outliers close to tail of the target distribution and extreme outliers. The contamination rate ranged from 0% to 30%. Performance was measured with bias, variance and summarized in the Mean Squared Error and the efficiency of the estimators was examined.

In general the performance of the estimators depended most heavily on whether the outliers were marginal or extreme. This makes sense as extreme outliers are easier to flag with cellwise robust methods. All casewise robust methods were outperformed by their counterparts in all cases, except for the Stahel-Donoho estimator with cellwise weights. This indicates that it is worthwhile switch to CRIV when cellwise contamination is suspected to be present in a dataset. As expected, the general pattern is that the estimates are decreasing towards zero when the contamination rate increases.

In case of marginal outliers, the estimators could not provide reliable results when the contamination rate was higher than 10%. However, the performance of the casewise and cellwise MCD estimators was equal to their performance under no contamination until that threshold. For most scenarios and contamination rates, the TSGS yielded a lower MSE than the regular S-estimator. Performance also decreased with dimensions as the MSE for all estimators was lower for the small dataset (scenario 1) as compared to larger datasets (scenarios 2 and 3).

In the presence of extreme outliers, the results were more favourable. Whereas the variance of the estimates from the casewise estimators exploded at 10% contamination or more, the cellwise robust estimators yielded good results up until a contamination rate of 20% or higher sometimes. This indicates that these methods can provide reliable estimates even if 83% of the observations are expected to contain an outlying cell. The TSGS had the best performance in all scenarios, with even similar results at a 25% contamination rate to no contamination in the first scenario. The Cellwise MCD also provided reliable results until a contamination rate of 20% and is hence a robust estimator when outliers are extreme.

The application of the CRIV estimator to estimate the relation between trade openness and inflation shows that the TSGS estimator yields reliable estimates and could potentially shift estimates to a higher significance level, assuming the influence function is similar to the S-estimator. Additional insight is gained from the cellwise weights, which show that some countries are outliers because of their size or because their inflation extremely high.

Instead of robustifying the model solutions as done here, there are different approaches to obtaining robust IV estimates. There is an extensive strand of literature that has been concerned with these different approaches under casewise contamination, it would be interesting to see the comparison of these methods to the robustification approach used here. To name a few, Ronchetti and Trojani (2001) robustify the orthogonality conditions in a Generalized Method of Moments model, Maronna and Yohai (1997) provide the τ-estimator to estimate coefficients in a simultaneous equation model and Wagenvoort and Waldmann (2002) consider the two stage estimator and robustify both stages. For example, the latter could be done using the Cellwise robust M regression from Filzmoser et al. (2020). The possibility to cater these and other approaches to cellwise contamination is left for future research.

Another suggestion for future research is to investigate different specifications and tuning of the cellwise robust estimators. The TSGS can also be applied with the Rocke-type weight function in the second step instead of the Tukey bisquare function or the bivariate filter could be extended to the multivariate filter from Saraceno and Agostinelli (2021). Additionally, the Cellwise MCD can be tuned with a higher or lower penalty term for assigning zero weights to cells and it may be that the size of the subset h can be optimized for certain scenarios. In that same spirit, the number of subsamples of the SDM estimator could also be optimized, although a larger amount of subsamples also implies longer computation times. Lastly, the initial location and scatter estimators used for all estimators was the default estimator. Choosing specific initial estimators might lead to increases in estimator performance.

Last remark on potential future research revolves around the properties of the resulting estimators. The TSGS, Cellwise MCD and SDM are consistent estimators, however, their (asymptotic) distributions are not known. The theoretical influence functions of the aforementioned estimators also have not been researched either and deriving equations for the estimates of the standard errors would assist greatly in inference.

References

Aerts, S. and Wilms, I. (2017). Cellwise robust regularized discriminant analysis. *Statistical Analysis and Data Mining*, 10(6):436–447.

Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, 24:441–461.

Alesina, A. and Zhuravskaya, E. (2011). Segregation and the quality of government in a cross section of countries. *The American Economic Review*, 101(5):1872–1911.

Alfaro, L. (2005). Inflation, openness, and exchange-rate regimes: The quest for short-term commitment. *Journal of Development Economics*, 77(1):229–249.

Alqallaf, F., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85.

Bellemare, M. F., Masaki, T., and Pepinsky, T. B. (2017). Lagged explanatory variables and the estimation of causal effect. *The Journal of Politics*, 79(3):949–963.

Bottmer, L., Croux, C., and Wilms, I. (2022). Sparse regression for large data sets with outliers. *European Journal of Operational Research*, 297(2):782–794.

Bowdler, C. and Malik, A. (2017). Openness and inflation volatility: Panel data evidence. *The North American Journal of Economics and Finance*, 41:57–69.

Butler, R. W., Davies, P. L., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 21(3):1385–1400.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2):161–190.

Croux, C., Van Aelst, S., and Dehon, C. (2003). Bounded influence regression using high breakdown scatter matrices. *Annals of the Institute of Statistical Mathematics*, 55(2):265–285.

Danilov, M., Yohai, V. J., and Zamar, R. H. (2012). Robust estimation of multivariate location and scatter in the presence of missing data. *Journal of the American Statistical Association*, 107(499):1178–1186.

Davies, P. L. (1987). Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15(3):1269–1292.

De Ketelaere, B., Hubert, M., Raymaekers, J., Rousseeuw, P. J., and Vranckx, I. (2020). Real-time outlier detection for large datasets by RT-DetMCD. *Chemometrics and Intelligent Laboratory Systems*, 199:103957.

Donoho, D. L. (1982). *Breakdown Properties of Multivariate Location Estimators.* PhD thesis, Harvard University.

Filzmoser, P., Höppner, S., Ortner, I., Serneels, S., and Verdonck, T. (2020). Cellwise robust M regression. *Computational Statistics & Data Analysis*, 147:106944.

Freue, G. V. C., Ortiz-Molina, H., and Zamar, R. H. (2013). A natural robustification of the ordinary instrumental variables estimator. *Biometrics*, 69(3):641–650.

Gervini, D. and Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics*, 30(2):583–616.

Heij, C., de Boer, P., Franses, P. H., Kloek, T., and van Dijk, H. K. (2004). *Econometric Methods with Applications in Business and Economics.* Oxford University Press.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Hubert, M., Rousseeuw, P. J., and Van den Bossche, W. (2019). MacroPCA: an all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, 61:459–473.

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3):618–637.

Lal, A., Lockhart, M., Xu, Y., and Zu, Z. (2023). How much should we trust instrumental variable estimates in political science? Practical advice based on over 60 replicated studies. *Political Analysis. (forthcoming).*

Leung, A., Yohai, V. J., and Zamar, R. H. (2017). Multivariate location and scatter matrix estimation under cellwise and casewise contamination. *Computational Statistics & Data Analysis*, 111:59–76.

Leung, A., Zhang, H., and Zamar, R. H. (2016). Robust regression estimation and inference in the presence of cellwise and casewise contamination. *Computational Statistics & Data Analysis*, 99:1–11.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341.

Maronna, R. A. and Yohai, V. J. (1997). Robust estimation in simultaneous equations models. *Journal of Statistical Planning and Inference*, 57(2):233–244.

Maronna, R. A. and Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, 89(1):197–214.

Öllerer, V., Alfons, A., and Croux, C. (2016). The shooting S-estimator for robust regression. *Computational Statistics*, 31:829–844.

Öllerer, V. and Croux, C. (2015). Robust high-dimensional precision matrix estimation. In Nordhausen, K. and Taskinen, S., editors, *Modern Nonparametric, Robust and Multivariate Methods*, pages 325–350. Springer International Publishing, Cham.

Pison, G., Van Aelst, S., and Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, 55:111–123.

Raymaekers, J. and Rousseeuw, P. J. (2021). Handling cellwise outliers by sparse regression and robust covariance. *Journal of Data Science, Statistics, and Visualisation*, 1(3).

Raymaekers, J. and Rousseeuw, P. J. (2023). The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association. Advance online publication.*

Raymaekers, J. and Rousseeuw, P. J. (2024). Challenges of cellwise outliers. *Econometrics and Statistics. Advance online publication.*

Romer, D. (1993). Openness and inflation: Theory and evidence. *The Quarterly Journal of Economics*, 108(4):869–903.

Ronchetti, E. and Trojani, F. (2001). Robust inference with GMM estimators. *Journal of Econometrics*, 101(1):37–69.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications Vol. B*, pages 283–297.

Rousseeuw, P. J. and Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.

Rousseeuw, P. J. and Van Den Bossche, W. (2018). Detecting deviating data cells. *Technometrics*, 60(2):135–145.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In Franke, J., Härdle, W., and Martin, D., editors, *Robust and Nonlinear Time Series Analysis*, volume 26, pages 256–272, New York, NY. Springer.

Salibian-Barrera, M. and Yohai, V. J. (2006). A fast algorithm for s-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2):414–427.

Saraceno, G. and Agostinelli, C. (2021). Robust multivariate estimation based on statistical depth filters. *TEST*, 30.

Stahel, W. A. (1981). Breakdown of covariance estimators. *Research Report 31, Fachgruppe fur Statistik*.

Stromberg, A. J. (1993). Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM Journal on Scientific Computing*, 14(6):1289–1299.

Tarr, G., Müller, S., and Weber, N. (2016). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis*, 93:404–420.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.

Van Aelst, S., Vandervieren, E., and Willems, G. (2011). Stahel-Donoho estimators with cellwise weights. *Journal of Statistical Computation and Simulation*, 81:1–27.

Wagenvoort, R. and Waldmann, R. (2002). On B-robust instrumental variable estimation of the linear model with panel data. *Journal of Econometrics*, 106:297–324.

Young, A. (2022). Consistency without inference: Instrumental variables in practical application. *European Economic Review*, 147:104–112.

A Appendix

A.1 IV under contamination

Figures 9 and 10 show the Median Squared Error for the intercept and exogenous variables respectively. The estimates of the intercept are mostly affected by contamination in the endogenous variable and the dependent variable. For the coefficient of the exogenous variable contamination in the exogenous variable yields the largest increase in median squared error.

A.2 Simulation details

This subsection dives into the details of simulating the data set used in the simulation exercise. First the endogenous, exogenous and instrumental variables are simulated along with the error terms. Then these are used to compute the dependent variable y. The first scenario considers one endogenous variable, one exogenous variable and one instrument. The second scenario considers multiple exogenous variables while keeping the number of endogenous and instrumental variables at one. The third and last scenario considers three instruments and three control variables. The first element of the β vector is the intercept, the second element is the coefficient for the endogenous variable and the others are for the control variables.

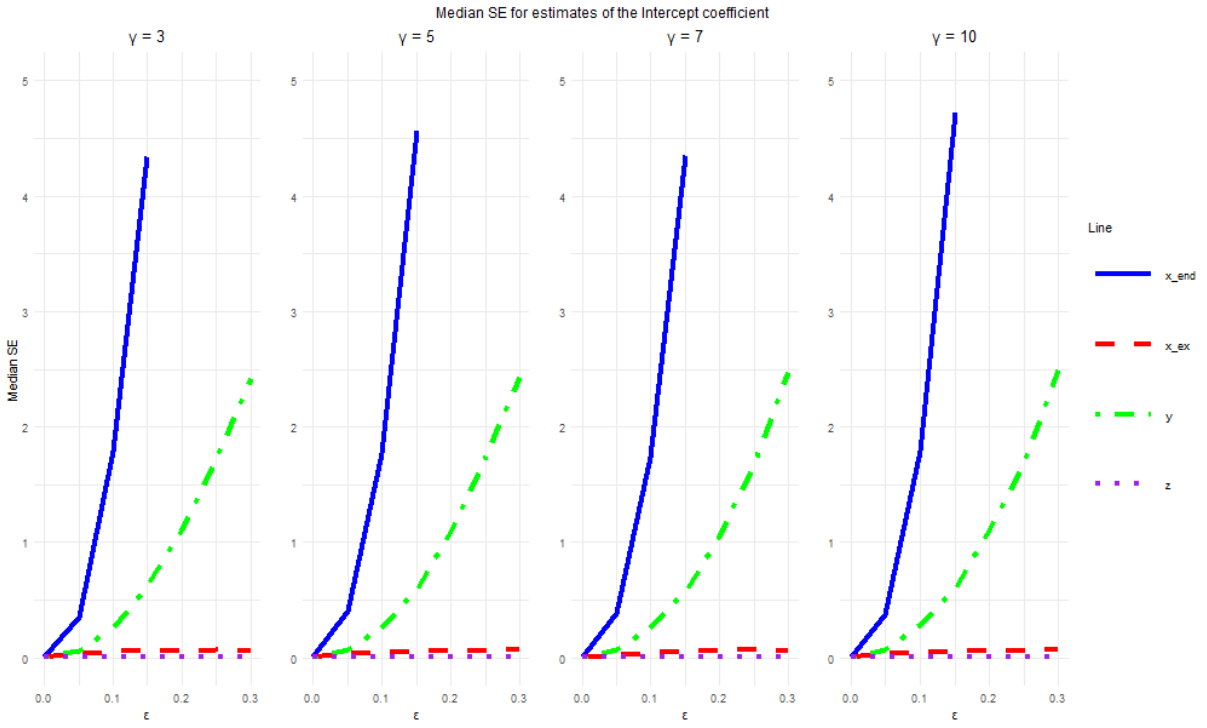


Figure 9: Median squared bias of the intercept in an IV model for different contamination levels γ and for contamination in all variables separately

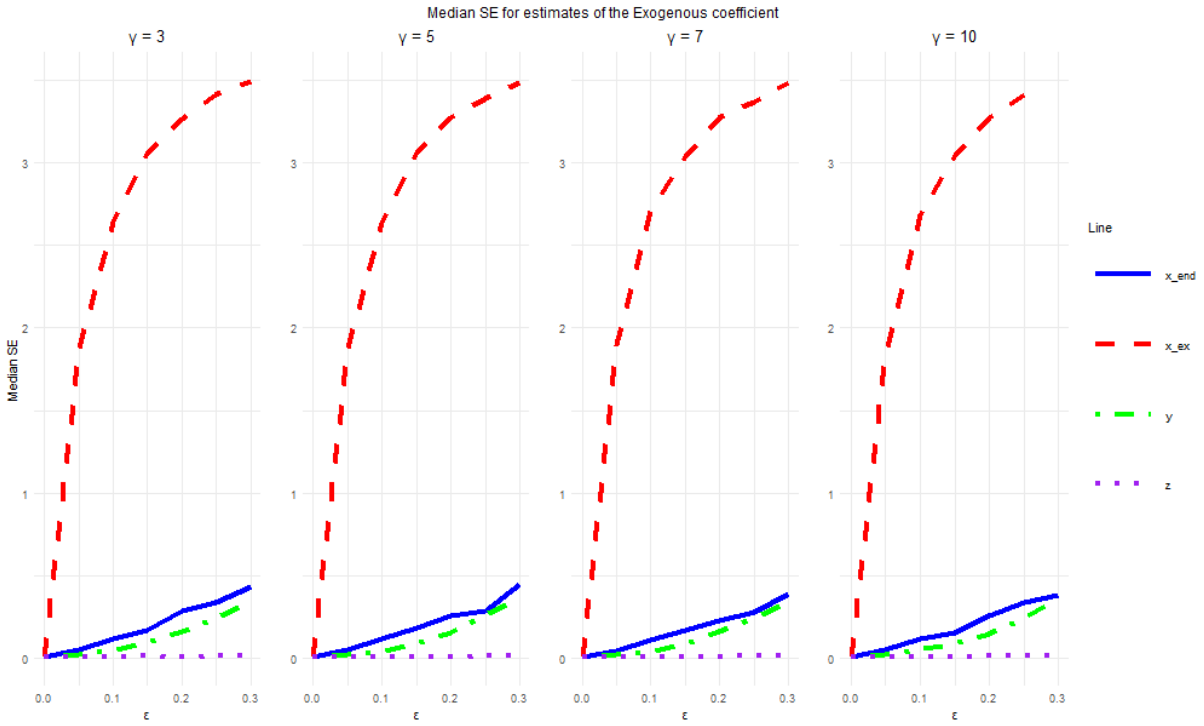


Figure 10: Median squared bias of the exogenous coefficient in an IV model for different contamination levels γ and for contamination in all variables separately

A.2.1 Scenario 1

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \beta = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0 & 0.9 & 0.2 \\ 0 & 1 & 0 & 0 \\ 0.9 & 0 & 1 & 0 \\ 0.2 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \boldsymbol{Z} \\ \boldsymbol{\varepsilon} \end{matrix} \tag{35}$$

A.2.2 Scenario 2

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \beta = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 4 \\ 0.5 \\ 0.5 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0.25 & 0.2 & 0.15 & 0.1 & 0 & 0.9 & 0.2 \\ 0.25 & 1 & 0.5 & 0.4 & 0.3 & 0 & 0 & 0 \\ 0.2 & 0.5 & 1 & 0.2 & 0.1 & 0 & 0 & 0 \\ 0.15 & 0.4 & 0.2 & 1 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0.3 & 0.1 & 0.1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0.9 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_{2,1} \\ \boldsymbol{X}_{2,2} \\ \boldsymbol{X}_{2,3} \\ \boldsymbol{X}_{2,4} \\ \boldsymbol{X}_{2,5} \\ \boldsymbol{Z} \\ \boldsymbol{\varepsilon} \end{matrix} \tag{36}$$

A.2.3 Scenario 3

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \beta = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 0.5 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} 1 & 0.2 & 0.1 & 0 & 0.9 & 0.8 & 0.7 & 0.2 \\ 0.2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.9 & 0 & 0 & 0 & 1 & 0.9 & 0.8 & 0 \\ 0.8 & 0 & 0 & 0 & 0.9 & 1 & 0.7 & 0 \\ 0.7 & 0 & 0 & 0 & 0.8 & 0.7 & 1 & 0 \\ 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_{2,1} \\ \boldsymbol{X}_{2,2} \\ \boldsymbol{X}_{2,3} \\ \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \boldsymbol{Z}_3 \\ \boldsymbol{\varepsilon} \end{matrix} \tag{37}$$

A.3 Computation details of robust scatter estimators

Table 7 contains the estimators and their specific specifications in terms of tuning and hyper-parameters.

A.4 Boxplots for intercept and coefficients of the control variables

The graphs in this sections are the same graphs exhibited in the main text, but then for the other coefficients and the intercept provided here for completeness.

Name	Specifications	`R` Package	Source
S-estimator	Tukey Bisquare, $c = 1.547$	`RIV`	Freue et al.
MCD	Objective function (20) with $h = 0.6n$	`rrcov`	Rousseeuw
SD	$c = 1.96$	`rrcov`	Stahel; Donoho
TSGS	Tukey Bisquare loss function	`GSE`	Danilov et al.
CellMCD	Objective function (21) with $h = 0.6$	`cellMCD`	Raymaekers and Rousseeuw
SDC	$\alpha_{ij} = (max_{k=1}^{p} c_{ik})^{-1} c_{ij}$	*	Van Aelst et al.
SDM	$\alpha_{ij} = (max_{k=1}^{p} \lvert u_{ik} \rvert)^{-1} \lvert u_{ij} \rvert$	*	Van Aelst et al.

Table 7: Overview with all estimators used in the thesis, * code was obtained directly from Stephan van Aelst

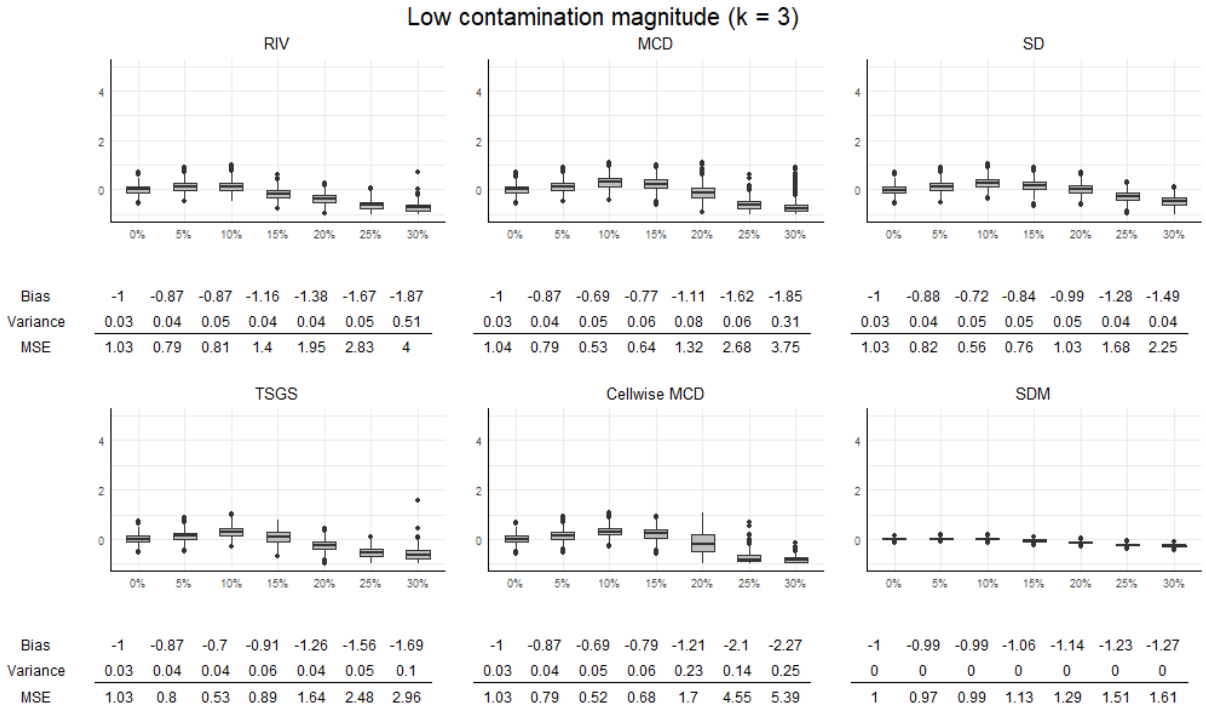


Figure 11: Boxplots with intercept in scenario 1 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

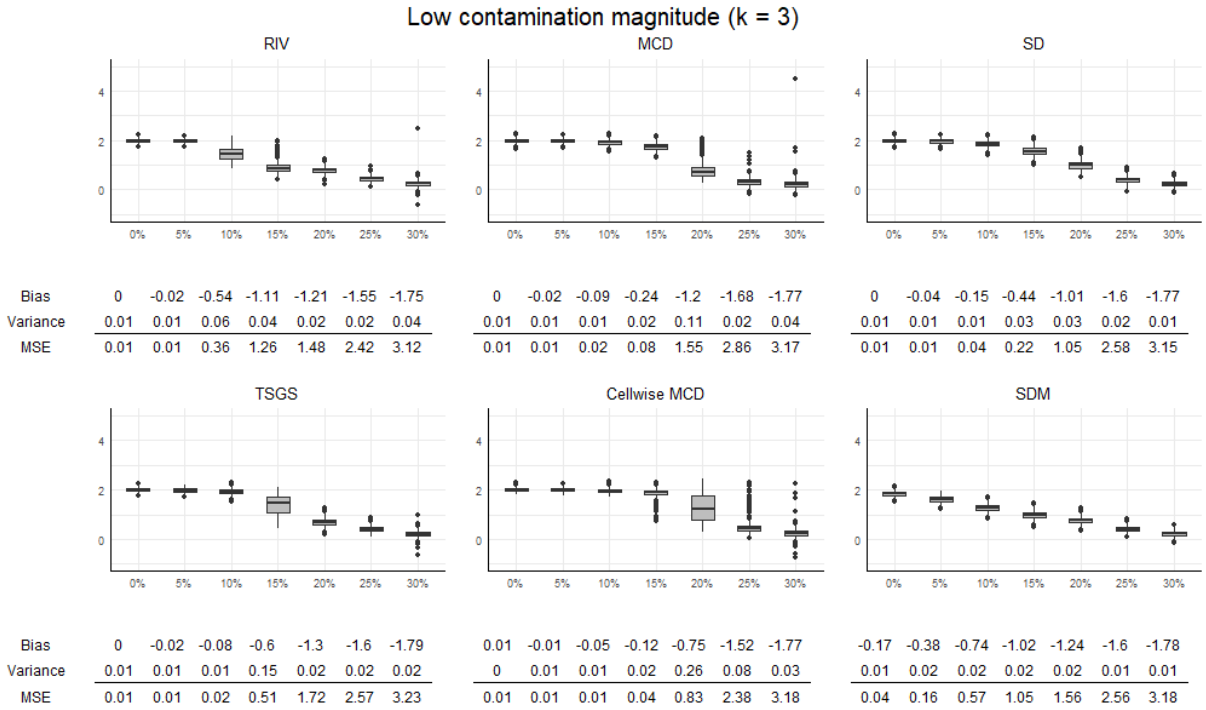


Figure 12: Boxplots with coefficients for the control variable ($\boldsymbol{X}_2$) in scenario 1 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

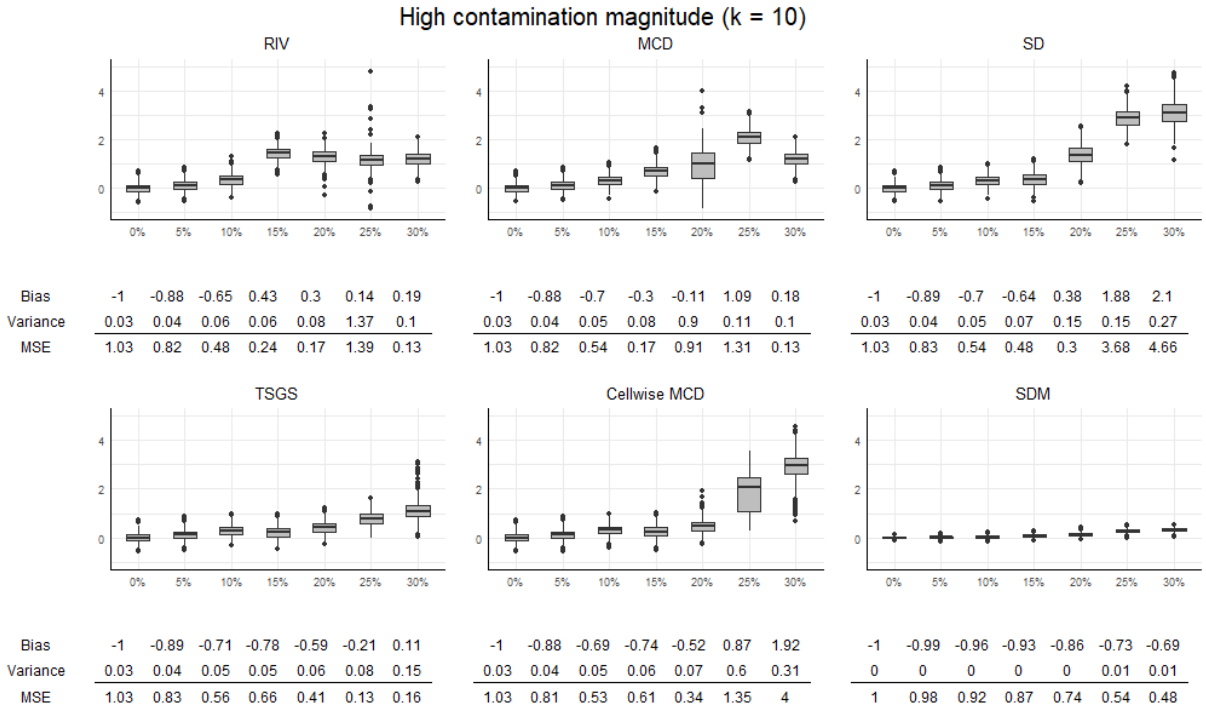


Figure 13: Boxplots with intercept in scenario 1 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

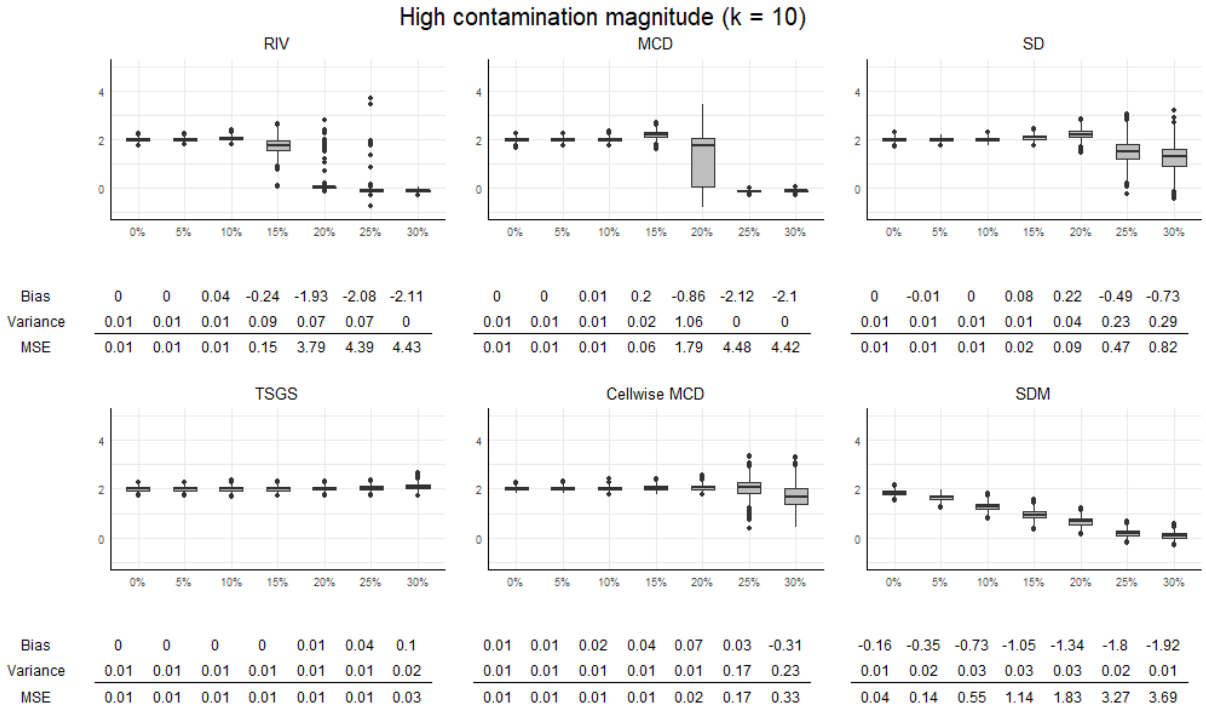


Figure 14: Boxplots with coefficients for the control variable ($\boldsymbol{X}_2$) in scenario 1 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

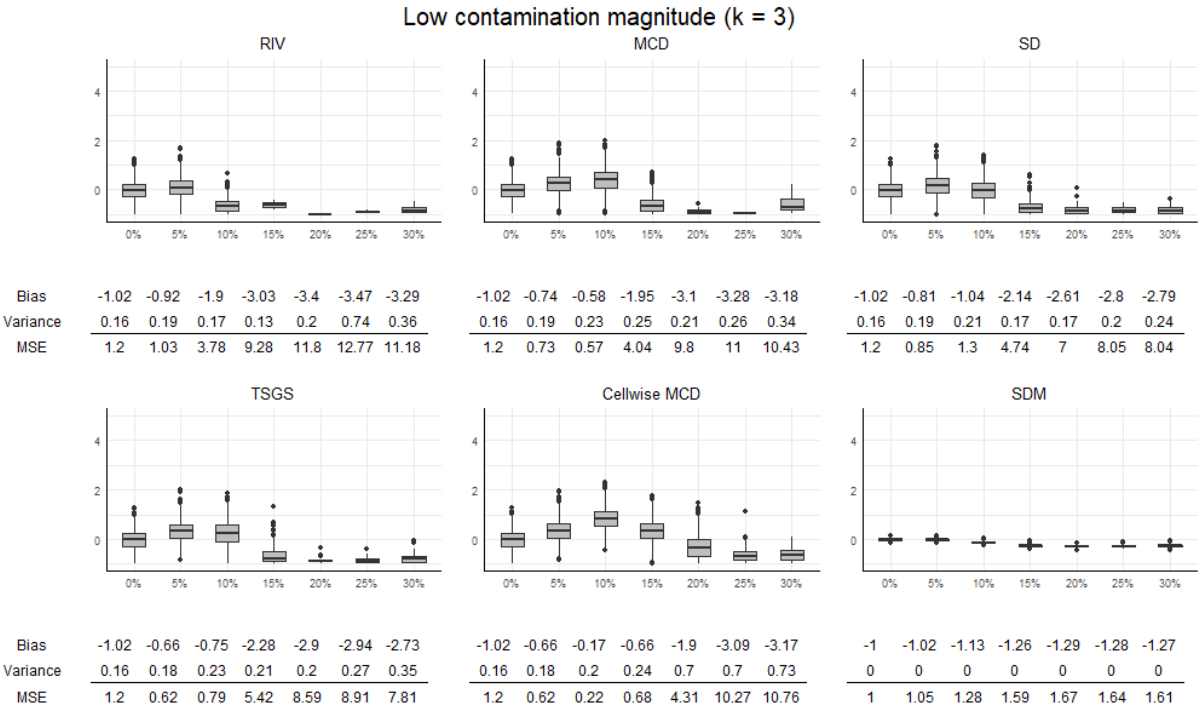


Figure 15: Boxplots with intercept in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

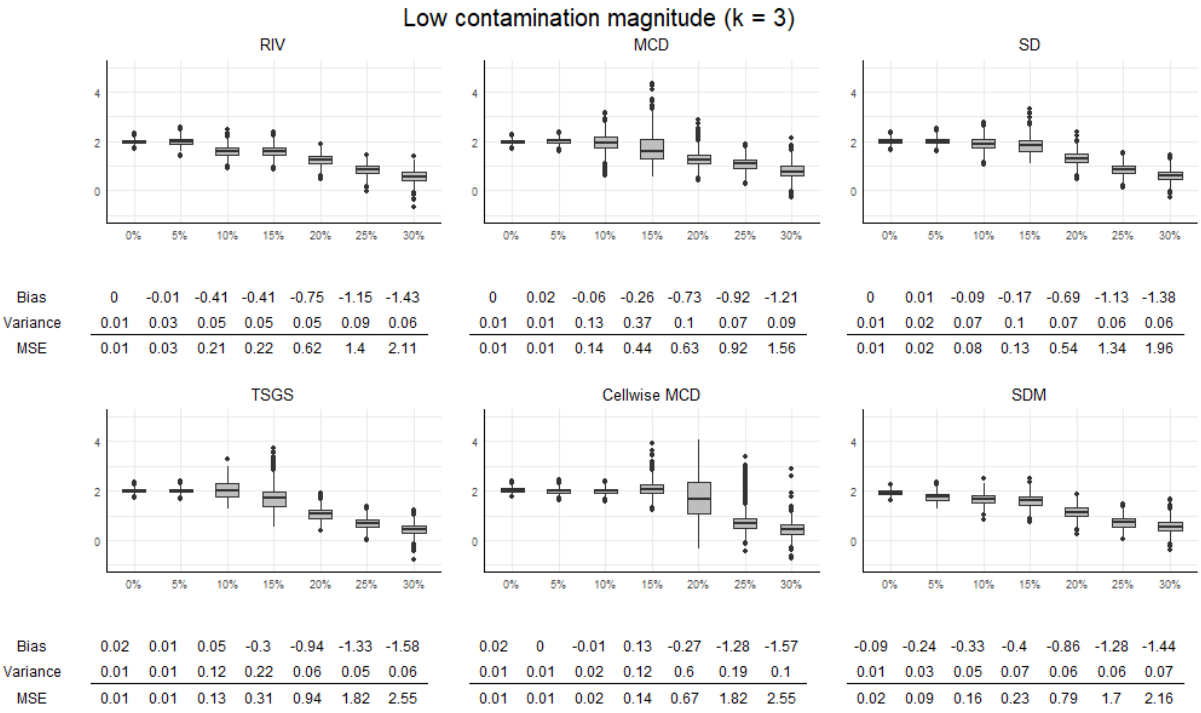


Figure 16: Boxplots with coefficients for control variable 1 ($\boldsymbol{X}_{2,1}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

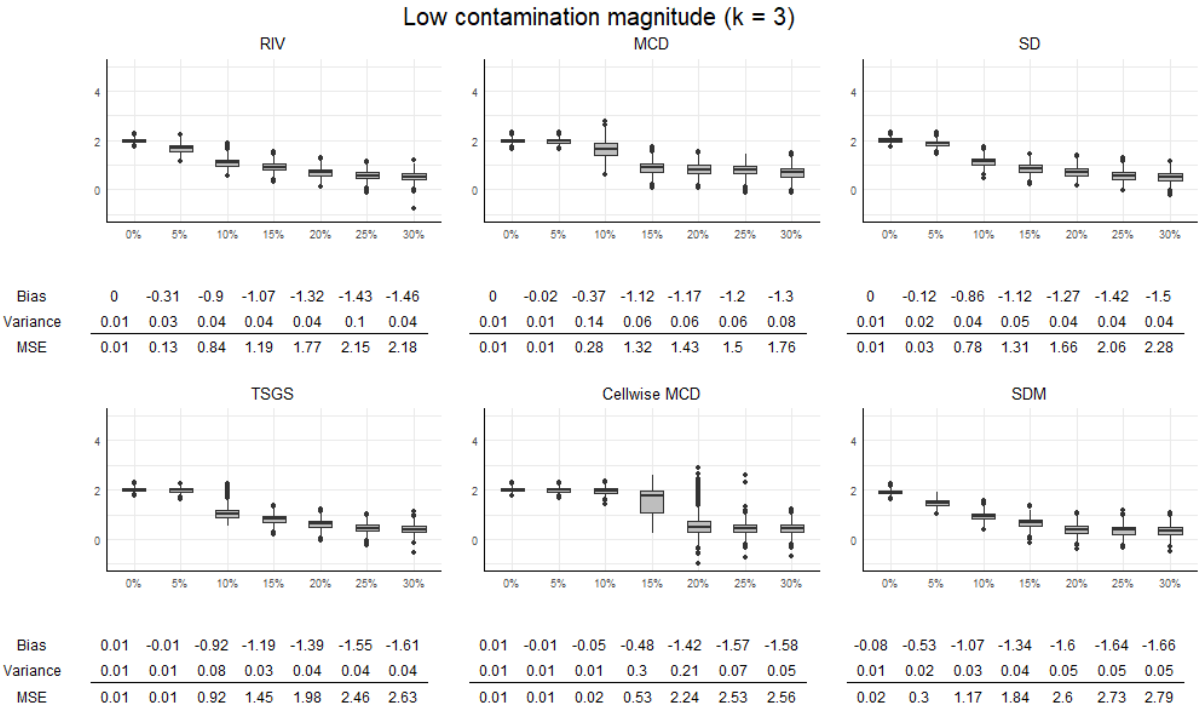


Figure 17: Boxplots with coefficients for control variable 2 ($\boldsymbol{X}_{2,2}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

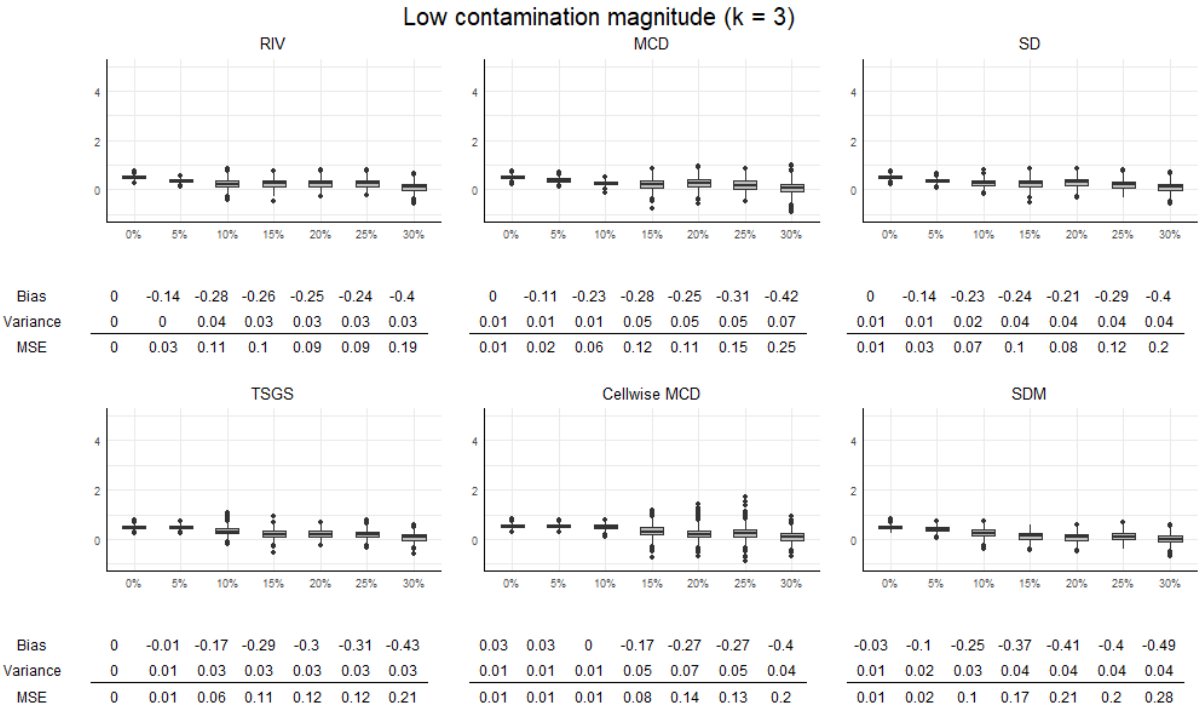


Figure 18: Boxplots with coefficients for control variable 3 ($\boldsymbol{X}_{2,3}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

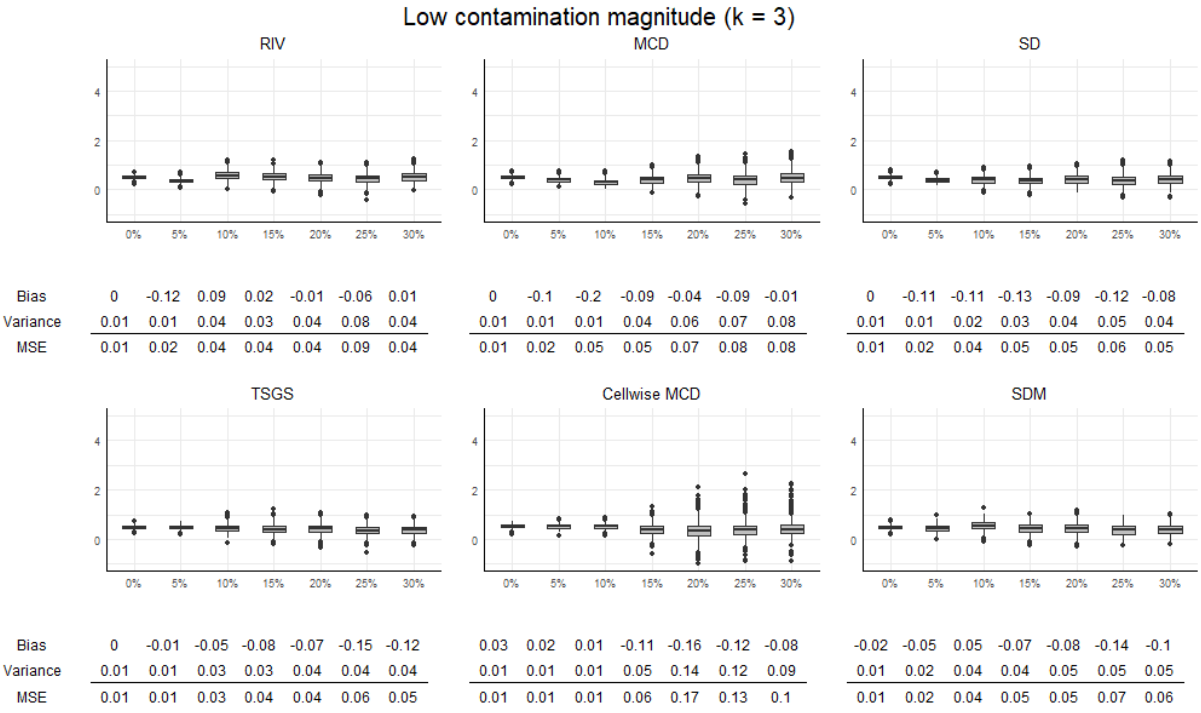


Figure 19: Boxplots with coefficients for control variable 4 ($\boldsymbol{X}_{2,4}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

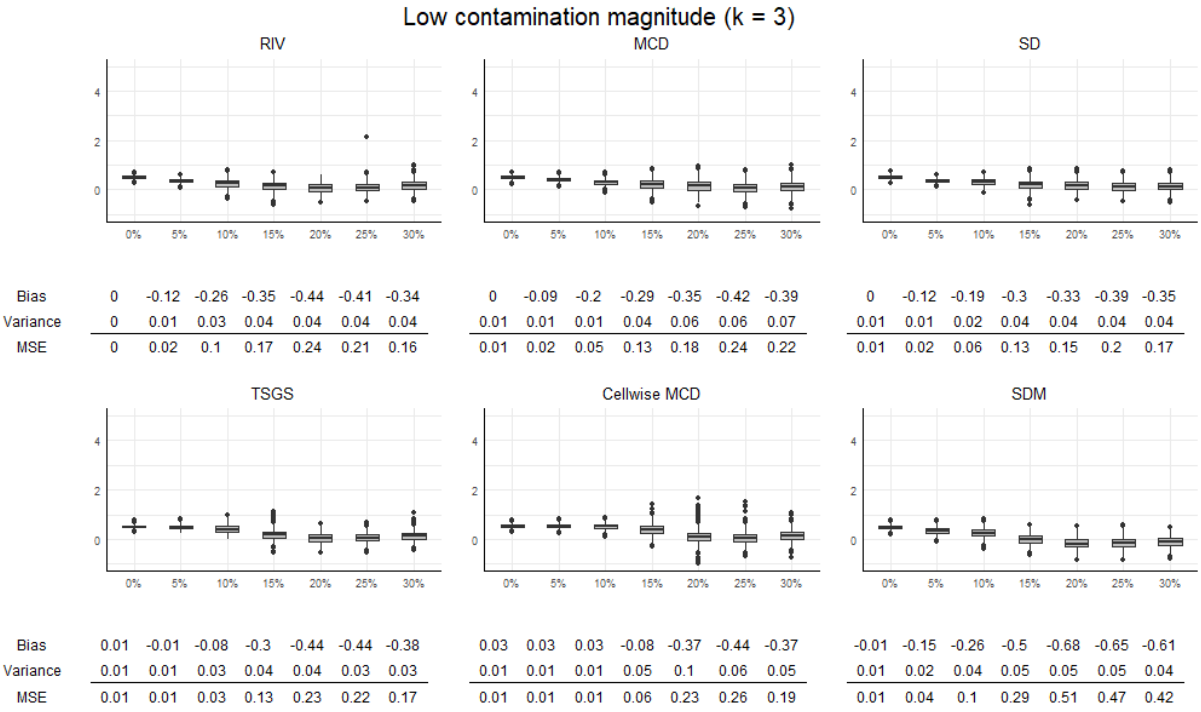


Figure 20: Boxplots with coefficients for control variable 5 ($\boldsymbol{X}_{2,5}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

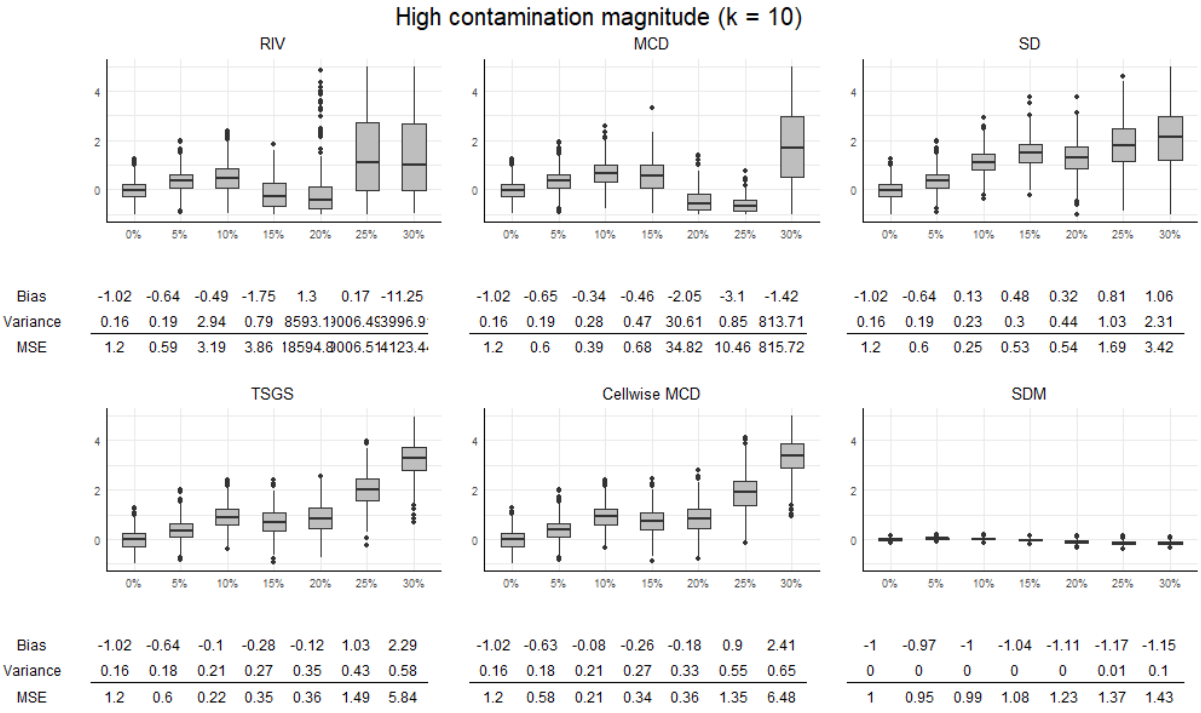


Figure 21: Boxplots with intercept in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

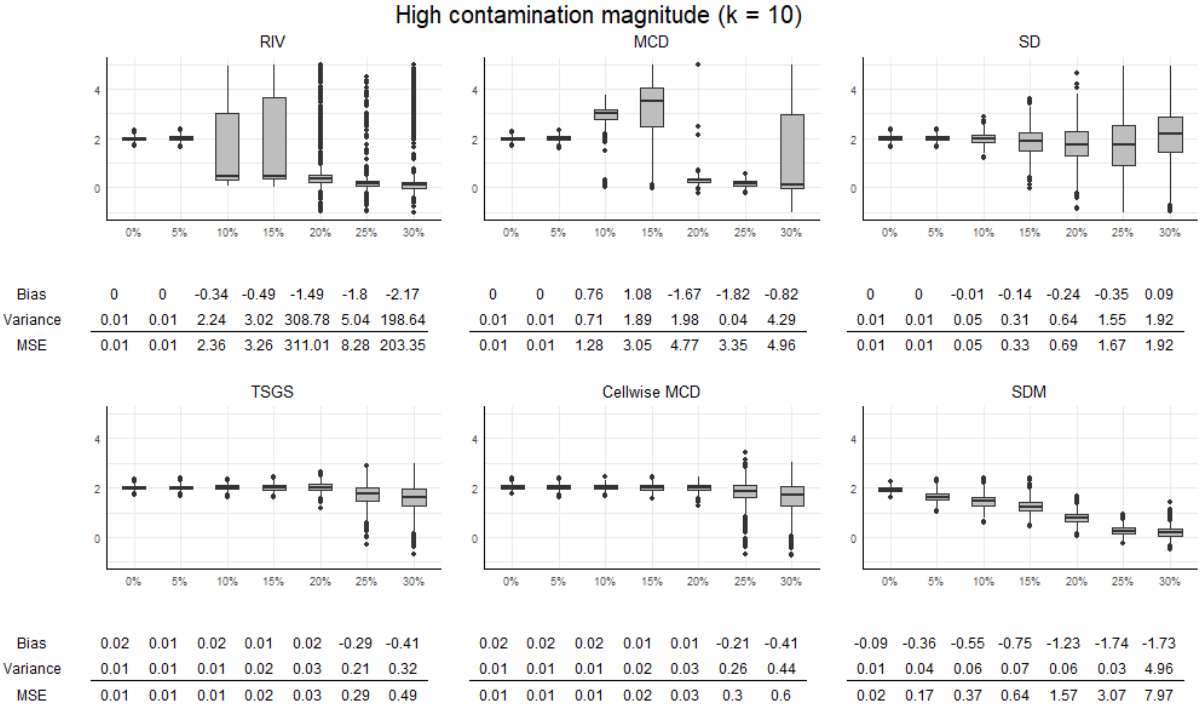


Figure 22: Boxplots with coefficients for control variable 1 ($\boldsymbol{X}_{2,1}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

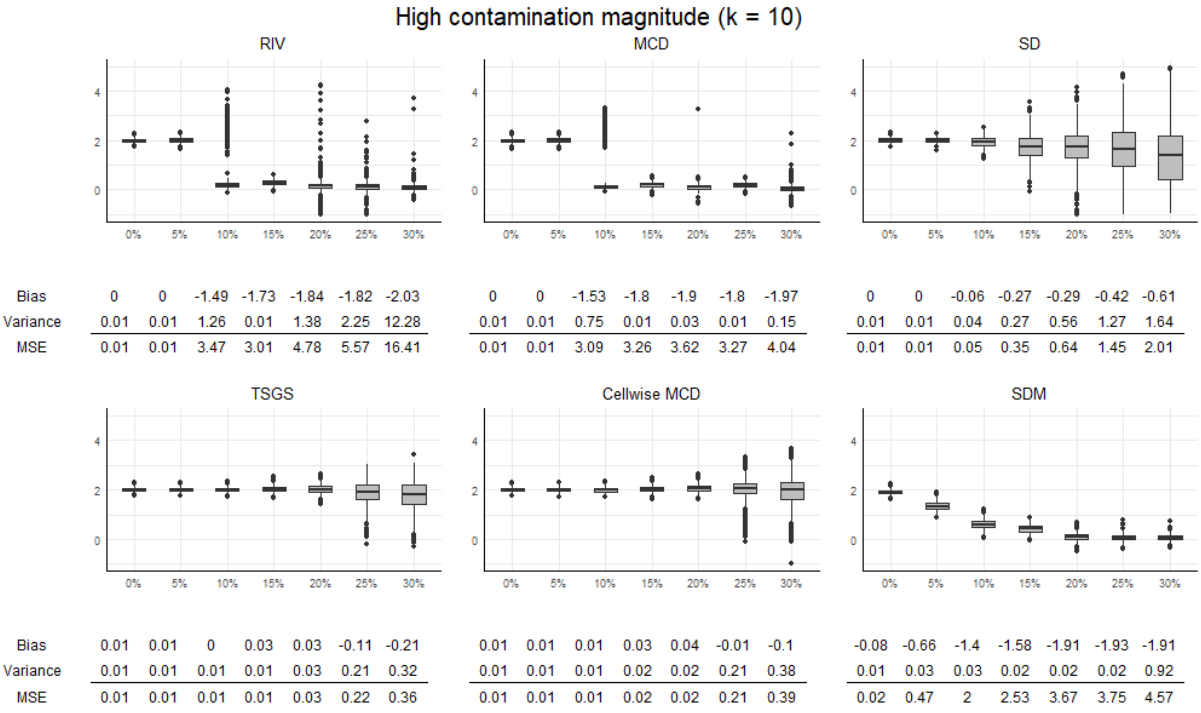


Figure 23: Boxplots with coefficients for control variable 2 ($\boldsymbol{X}_{2,2}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

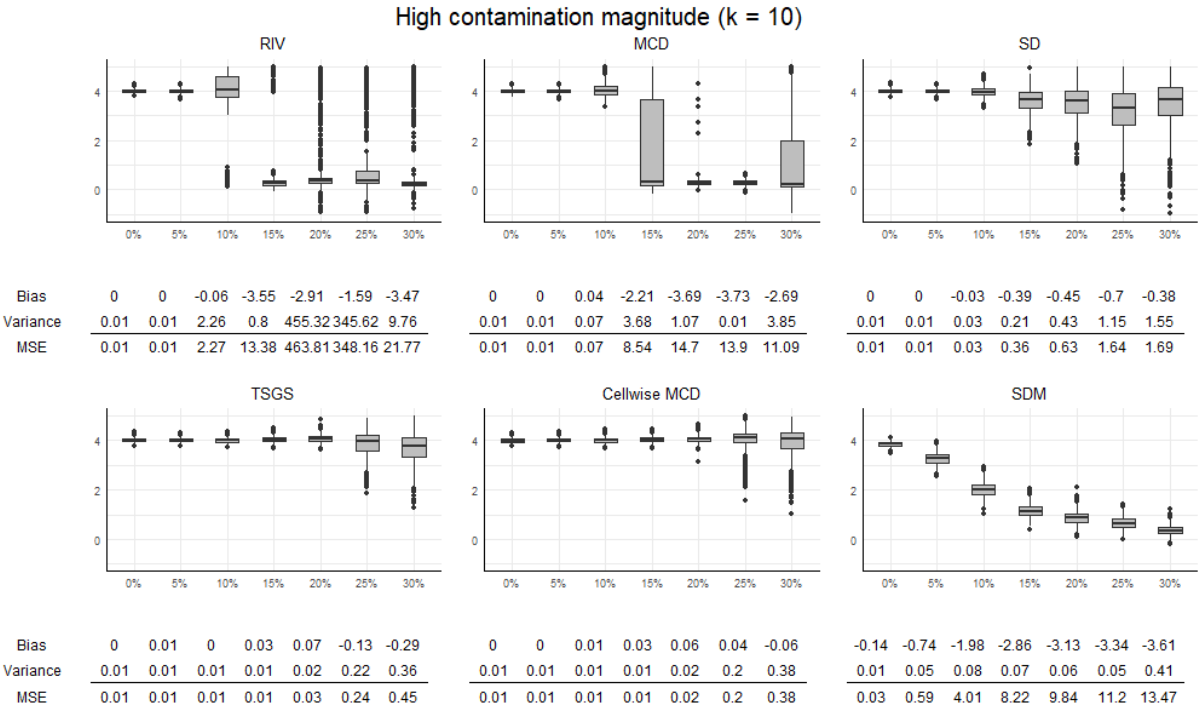


Figure 24: Boxplots with coefficients for control variable 3 ($\boldsymbol{X}_{2,3}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

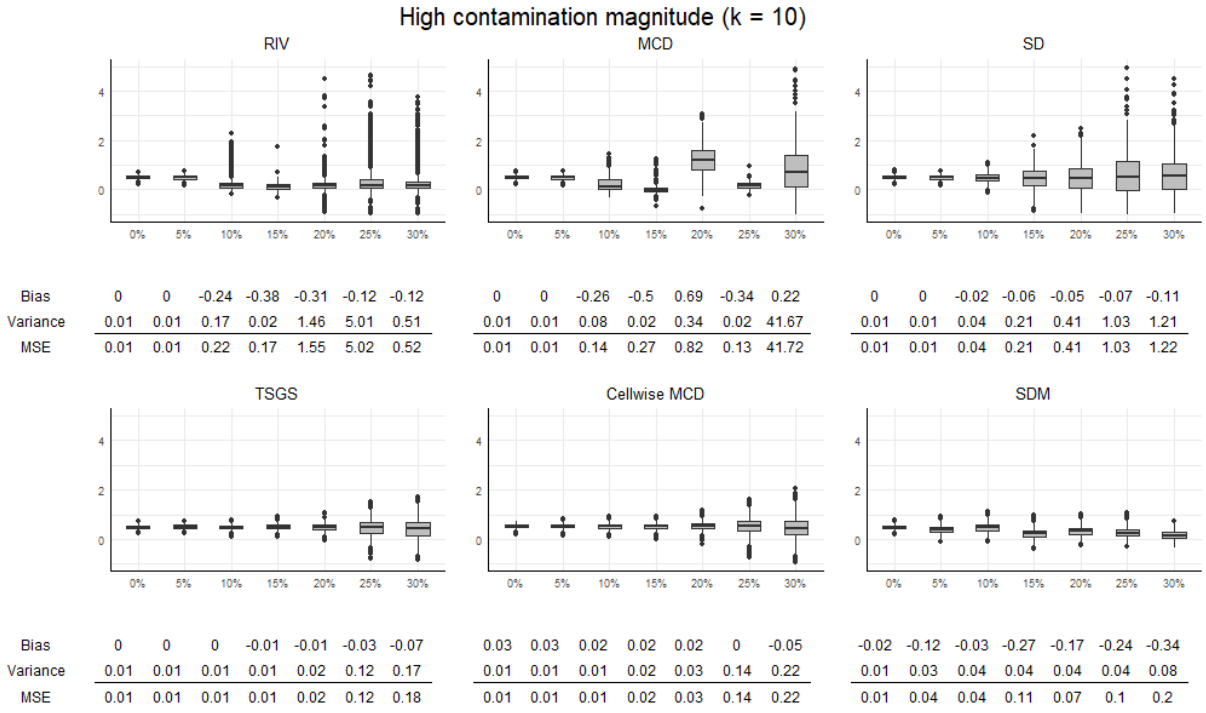


Figure 25: Boxplots with coefficients for control variable 4 ($\boldsymbol{X}_{2,4}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

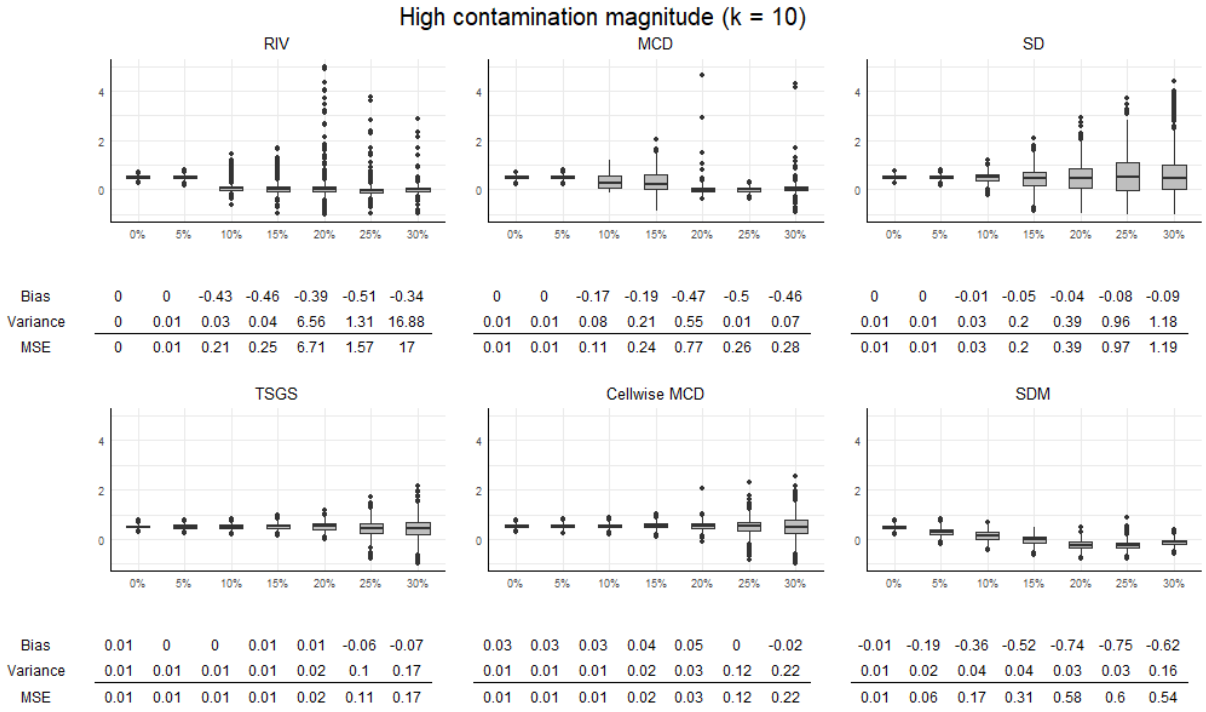


Figure 26: Boxplots with coefficients for control variable 5 ($\boldsymbol{X}_{2,5}$) in scenario 2 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

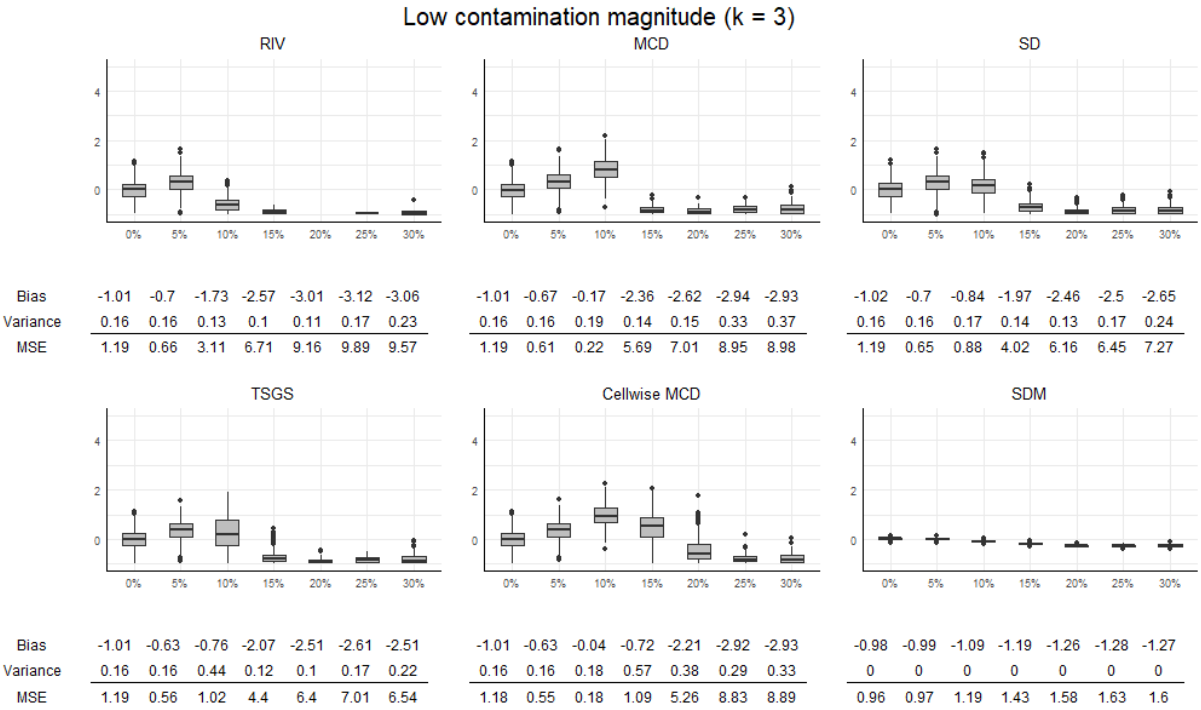


Figure 27: Boxplots with intercept in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

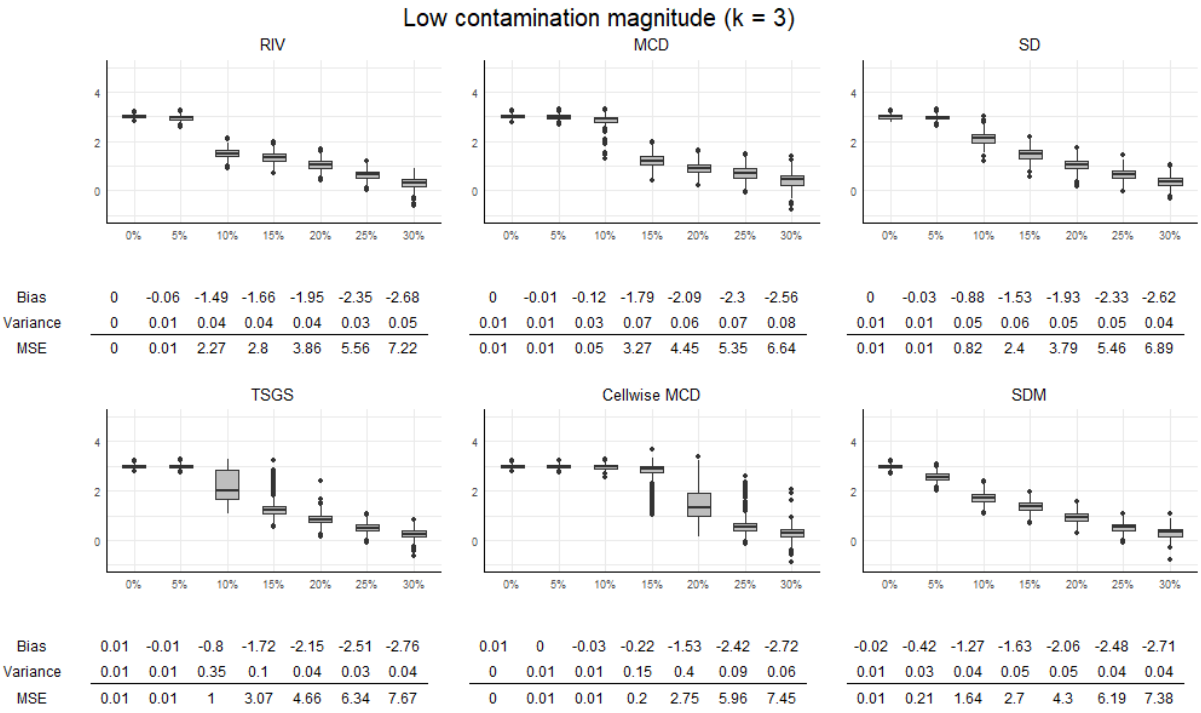


Figure 28: Boxplots with coefficients for control variable 1 ($\boldsymbol{X}_{2,1}$) in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

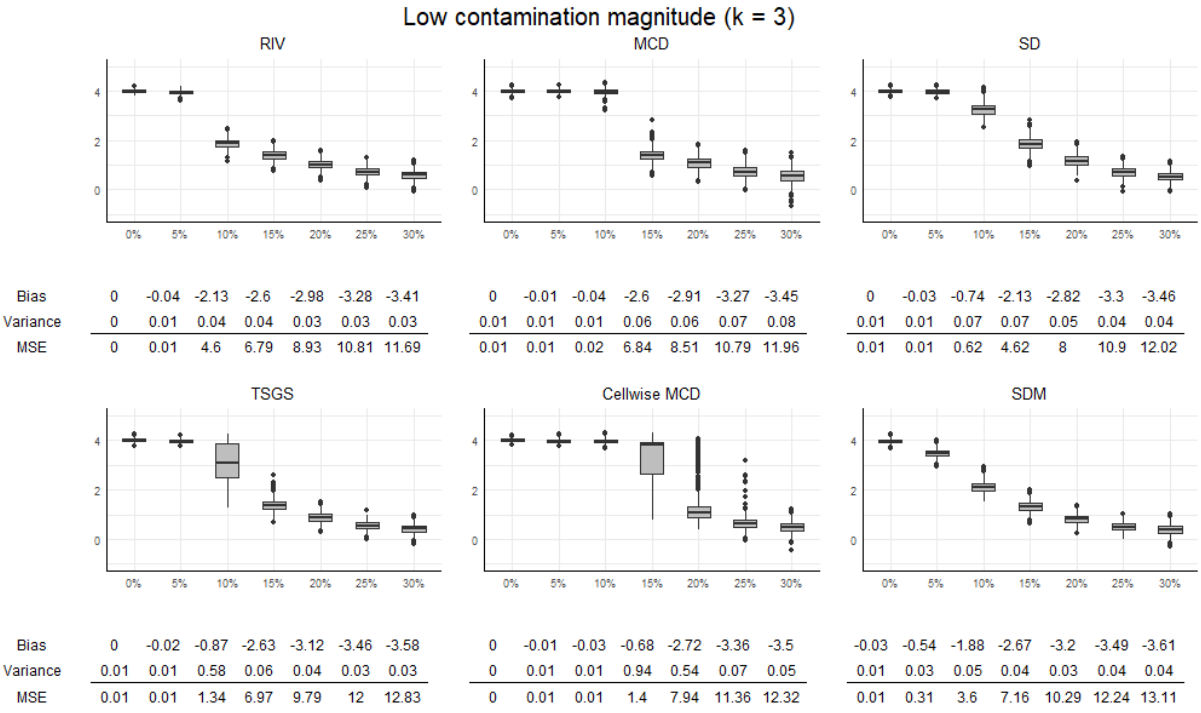


Figure 29: Boxplots with coefficients for control variable 2 ($\boldsymbol{X}_{2,2}$) in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

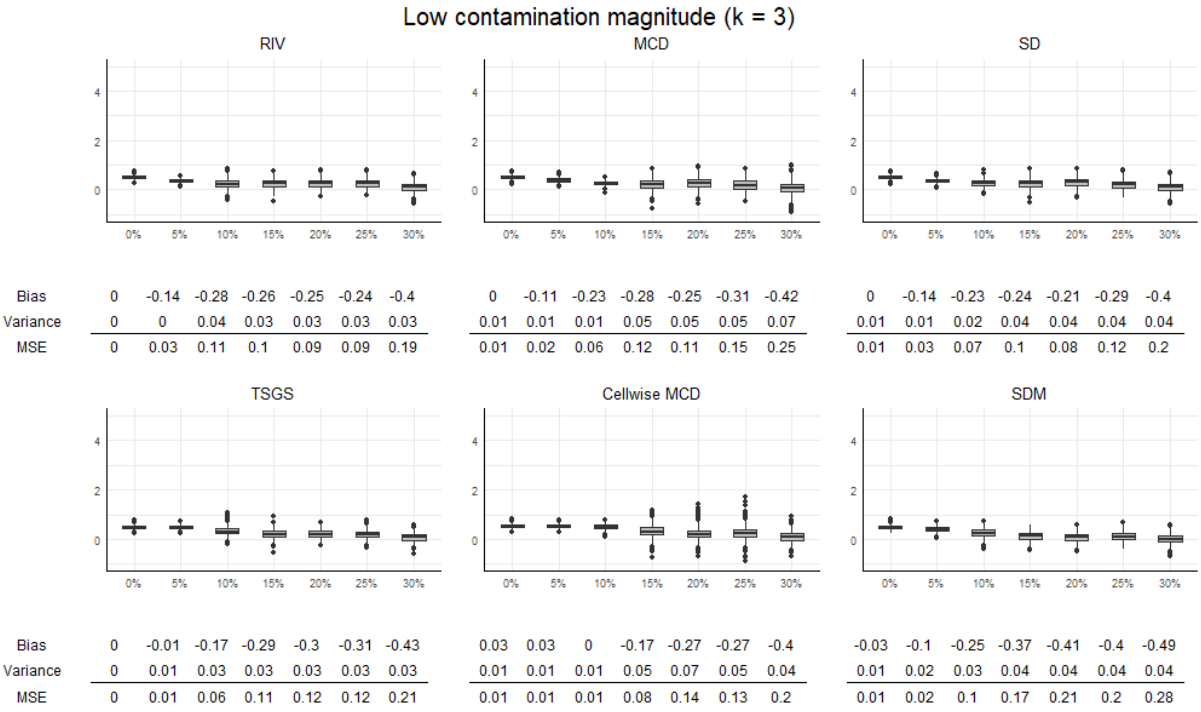


Figure 30: Boxplots with coefficients for control variable 3 ($\boldsymbol{X}_{2,3}$) in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 3$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

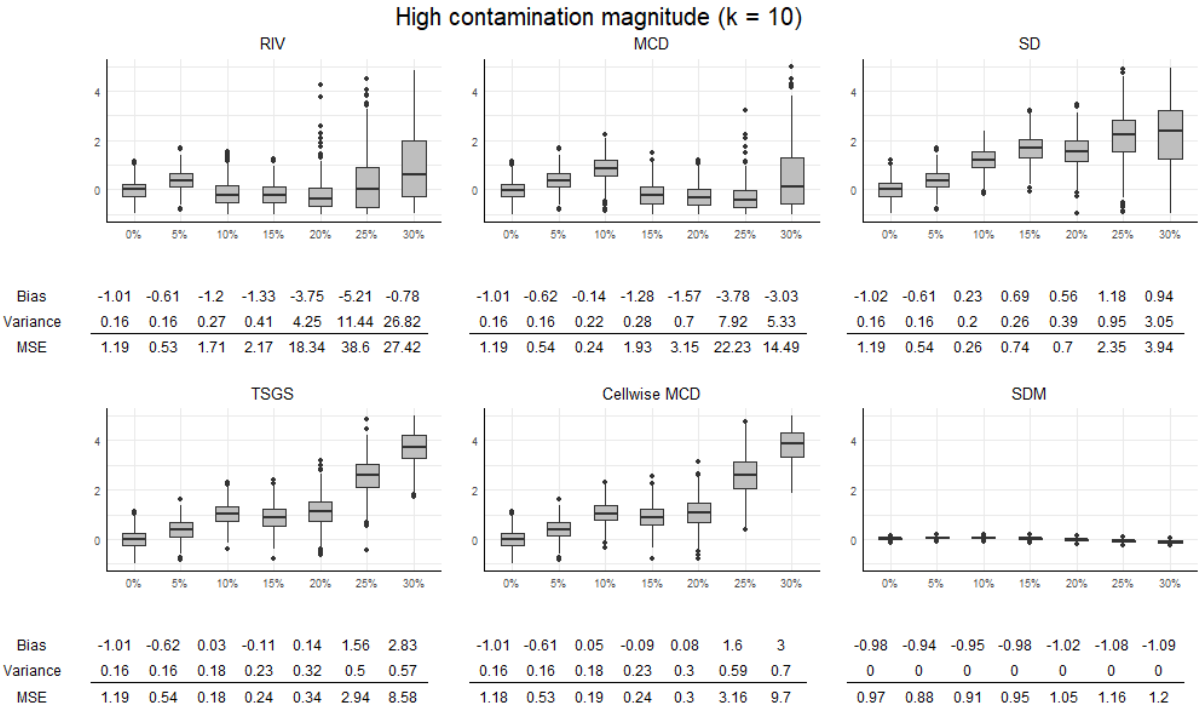


Figure 31: Boxplots with intercept in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

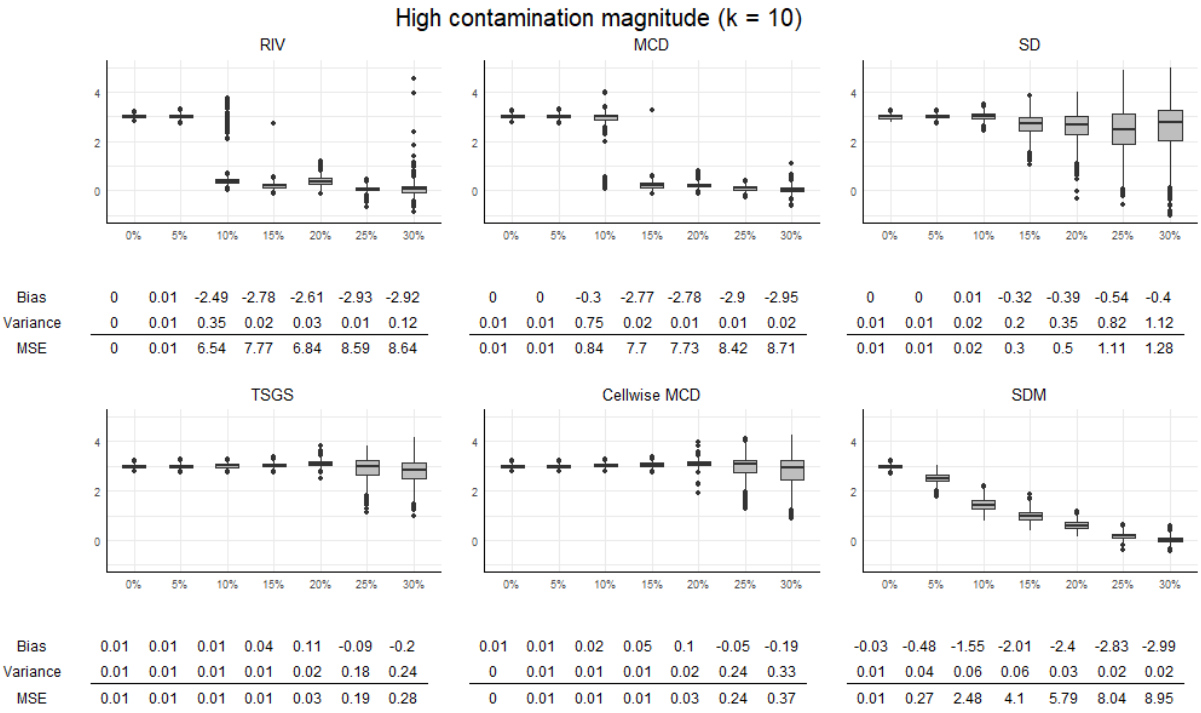


Figure 32: Boxplots with coefficients for control variable 1 ($\boldsymbol{X}_{2,1}$) in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

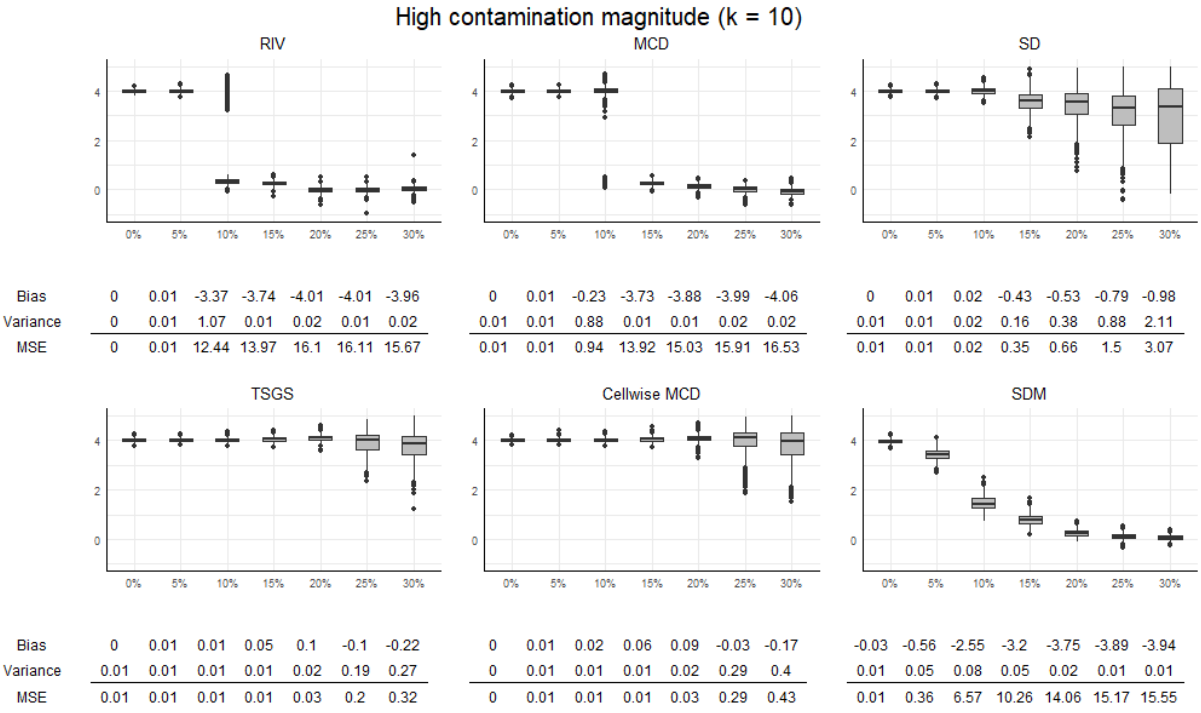


Figure 33: Boxplots with coefficients for control variable 2 ($\boldsymbol{X}_{2,2}$) in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

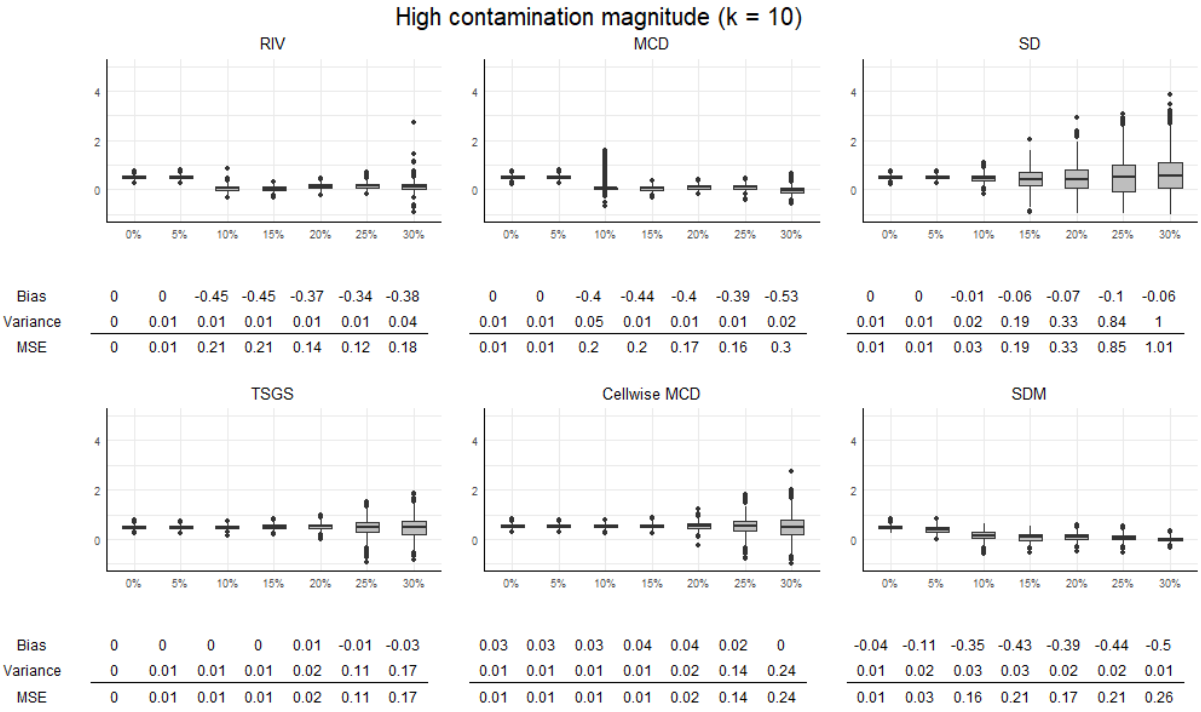


Figure 34: Boxplots with coefficients for control variable 3 ($\boldsymbol{X}_{2,3}$) in scenario 3 for contamination rates from 0% to 30%. The contamination magnitude is low and set at $k = 10$.The true beta can be found in subsection A.2 and the covariance estimators used are found above each graph.

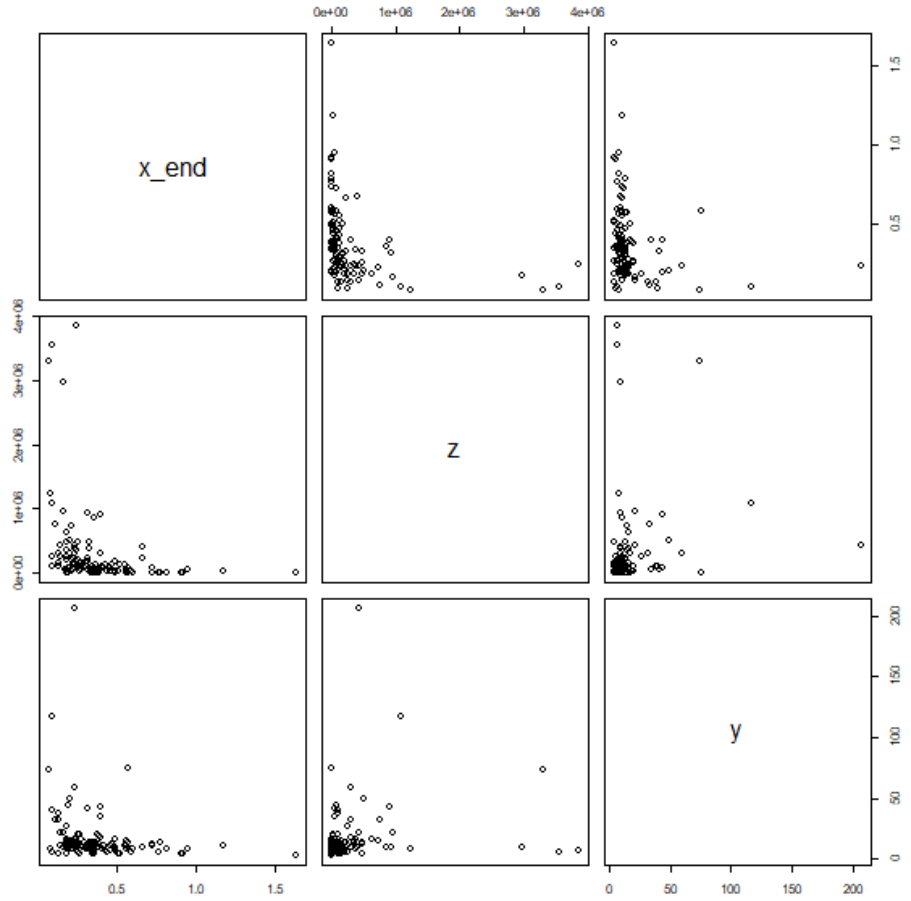


Figure 35: Matrix of scatterplots for the raw variables, i.e. the variables before the logarithmic transformation

A.5 Practical Application

This section contains more information on the dataset that is used in Section 6. The data used to estimate the relationship are the inflation, as measured by the average annual change in log GDP denominator since 1973, and openness, as measured by the average share of import in GDP since 1973. The instrument that is used is the logarithm of the land area in square miles. Since there are some countries that are outlying, Romer (1993) chose to regress log inflation on Openness to reduce the influence of outliers.

Figures 35 and 36 show scatterplots of matrices for the raw and logarithmic variables respectively. Figure 37 shows three histograms of the endogenous variable, the instrument and the dependent variable after the transformation. All three variables contain some outliers and are skewed towards the direction of these outliers. Finally, Figure **??** shows cellmaps with outlying components for the first and second half of the sample.

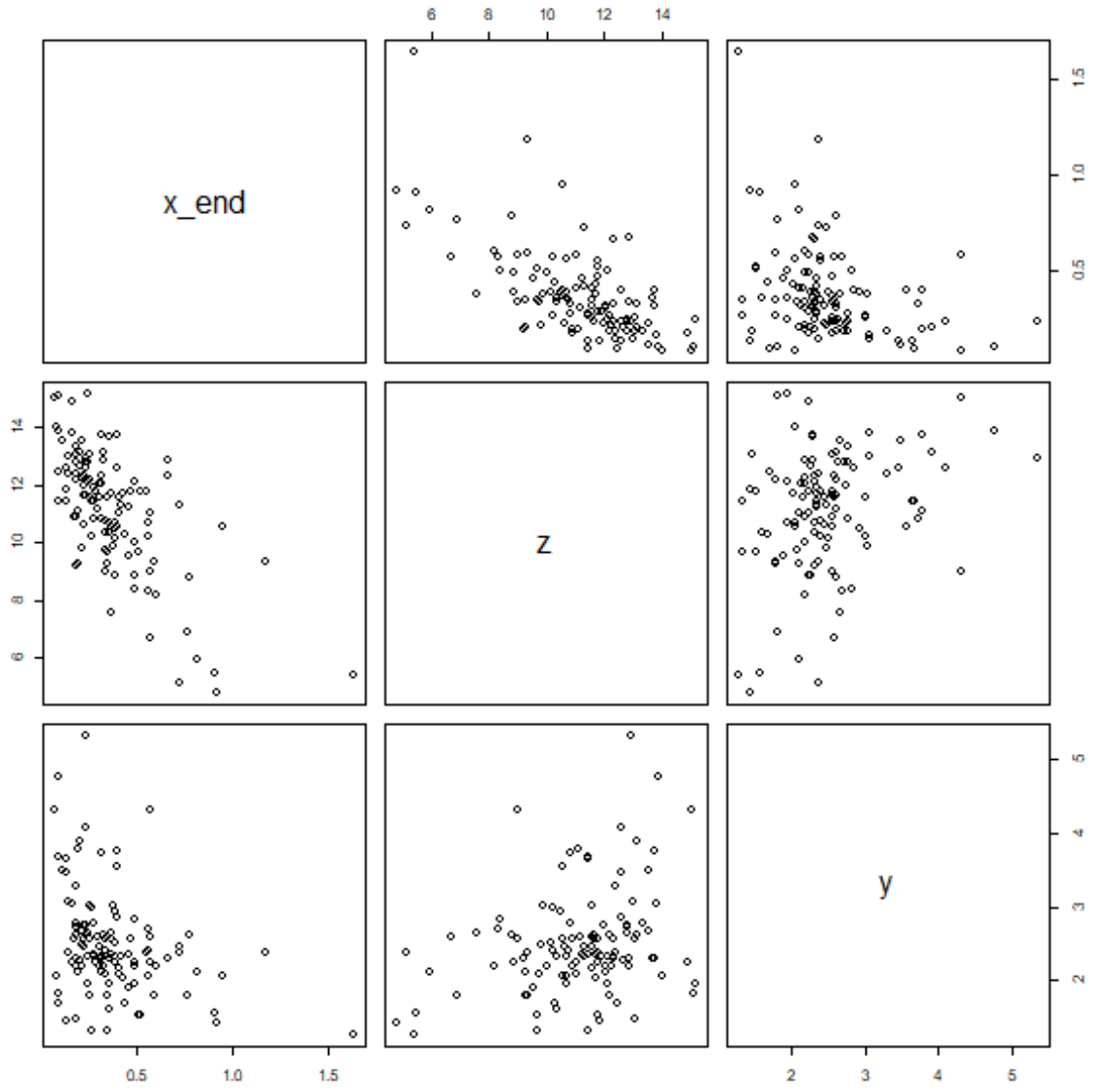


Figure 36: Matrix of scatterplots for the transformed variables, i.e. the variables after the logarithmic transformation

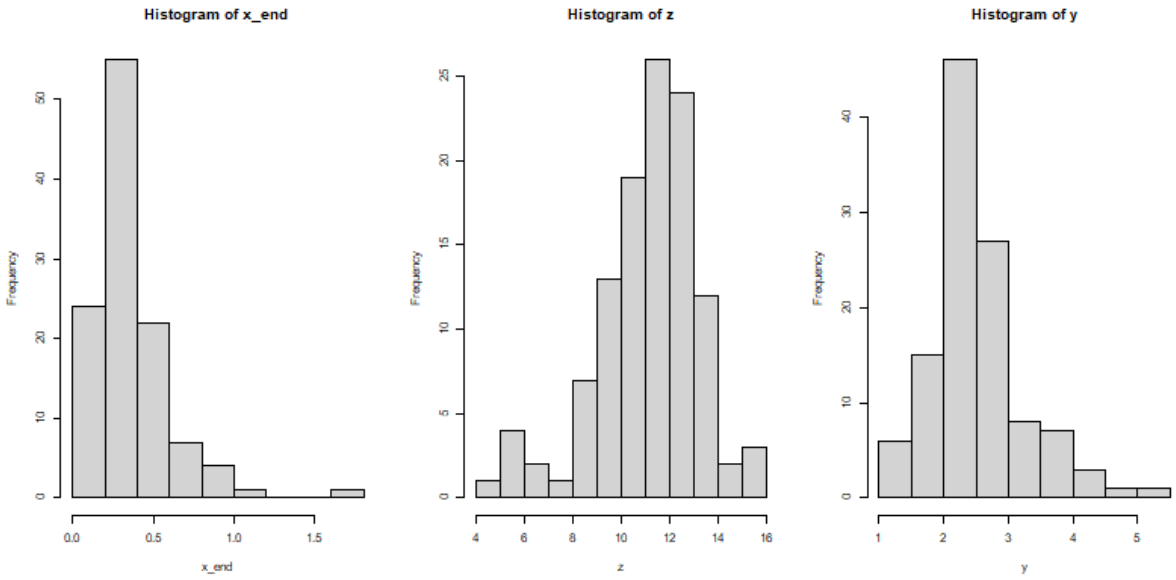
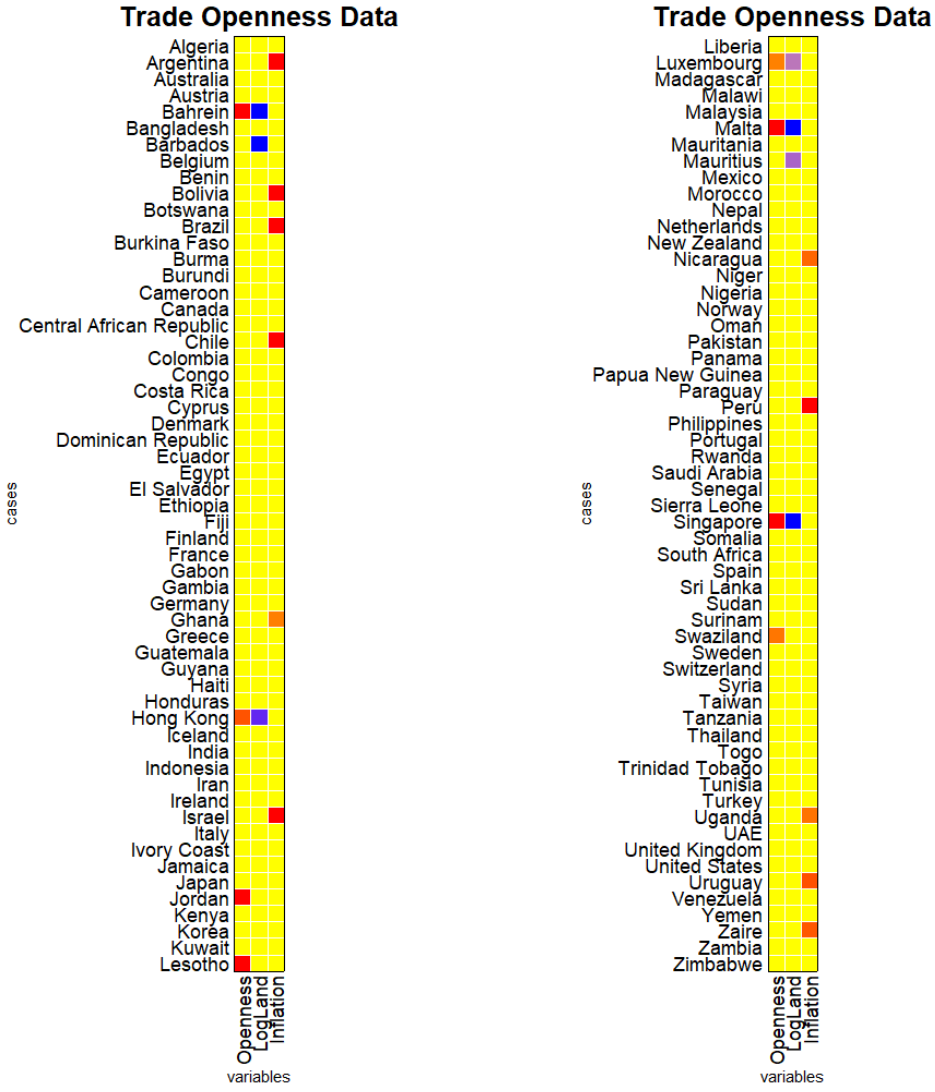


Figure 37: Histogram of the three variables after the logarithmic transformation of Inflation (y) and Land Area (z).



(a) First half of the sample.

(b) Second half of the sample.

Figure 38: Cellmap with outlying components. Red colour indicates an outlier on the right side of the distribution (large values), while a blue colour indicates an outlier on the left side of the distribution (small values).