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

Handling cellwise outliers in static fixed effects panel data models

Sophia van Megen (507991)

Supervisor:	M. Zhelonkin
Second assessor:	W. Wang
Date final version:	24th March 2024

Abstract

In this study, we tackle the challenge of outliers in static fixed effects panel data by developing a both cellwise and casewise robust methodology. Our approach integrates the adapted Gervini-Yohai outlier filter for its outlier detection capabilities, followed by imputation of filtered outliers using coordinate-wise medians. To address both cellwise and casewise outliers effectively, we then incorporate the robust casewise WMS method, noted for its asymptotic efficiency and regression properties.

Our evaluation through a Monte Carlo simulation study contrasts our method with the classical Within Estimator, focusing on the bias and standard deviation impacts on regression parameters. We explore the model's resilience to local contamination through the empirical influence function (EIF). This investigation is crucial as outliers, including vertical, horizontal, or leverage points, can significantly distort statistical analysis outcomes in panel data, a context where atypical observations or errors often happen. Classical OLS-based methods such as the Within Groups estimator when faced with contaminated data, can yield inaccurate parameter estimates, underscoring the need for our cellwise robust approach.

Subsequently, GY filtering is applied to the numerical variables in the fixed effects model by Grossman, Pierskalla and Boswell Dean (2017) and sensitivity analysis is performed.

Our research fills a gap in the application of cellwise robust estimation techniques to panel data models in econometrics, addressing the critical issue of managing cellwise outliers to prevent incorrect interpretations and ensure reliable statistical analysis.

Contents

Chapter 1

Introduction

Outliers can significantly influence results, specifically in the statistical analysis of fixed effects panel data models. Although the effectiveness of robust estimators in linear regression models is well recognised, the extension of these robust methods to panel data models in econometrics is limited. We argue that handling outliers is especially crucial in the panel data context since large panels are likely to contain atypical observations or gross errors - including typing, recording or computation errors (Bramati & Croux, 2007). Next to that, the panel data may include vertical, horizontal or leverage outliers (Bakar & Midi, 2015). Also, outliers can be concentrated in blocks in a way that the fraction of outliers per cross-section constitutes at least half of the observations over time periods (Aquaro & Čížek, 2013). Because of this, classical OLS-based methods can be seriously affected due to contaminated data. Consequently, commonly used estimators, such as the within-group least squares estimators used in fixed effects panel data models, often result in incorrect parameter estimates (Beyaztas & Bandyopadhyay, 2022). Additionally, visually examining panel data - next to being a subjective approach - is not as straightforward as it is with cross-sectional data, especially when dealing with multiple regressors, which can lead to erroneous interpretations if not handled properly.

The robust statistics methodology, as developed in books Hampel (1968) and Huber and Ronchetti (2009) finds a robust fit that works well for the majority of the data. Casewise contamination, which means that each observation is either an outlier or not contaminated at all, is a crucial assumption underlying this approach (Raymaekers & Rousseeuw, 2023). This concept is also referred to as rowwise contamination because the data points are usually in rows in a data matrix while columns represent the variables.

Various interpretations of casewise contamination are available, but the most prevalent one is the Tukey-Huber contamination model (THCM), where a small portion of the cases might be contaminated and follows an arbitrary distribution. Under the THCM, one believes that a case either comes from the central model, or from a model that is unrelated to the control model. Consequently, methods developed under this model often either fully trust or downweight all aspects of a case simultaneously. Yet, there are instances where certain elements of a case are suspect, while others remain valid. As Raymaekers and Rousseeuw (2023) write, fully downweighting such cases might result in losing valuable insights from the uncontaminated parts. This led to the consideration of an alternative approach, focusing on cellwise outliers, where individual elements within a case are evaluated for contamination rather than the entire

case. Alqallaf, van Aelst, Yohai and Zamar (2009) first formulated the cellwise outlier paradigm. Their contamination model, the Fully Independent Contamination Model (FICM), assumes we observe data vector $\mathbf{d}$, given by

$$\mathbf{d} = (\mathbf{I} - \mathbf{B}_\epsilon)\mathbf{d_0} + \mathbf{B}_\epsilon\widetilde{\mathbf{d}}, \tag{1.1}$$

where $\mathbf{B}_\epsilon$ is a diagonal matrix with diagonal values of either zero or one. $\mathbf{d_0} \sim F_0$ represents the uncontaminated original data vector, while $\widetilde{\mathbf{d}} \sim \widetilde{F}$ represents the data vector from the contaminated distribution. Therefore, the model assumes that the data was initially generated by the clean distribution F_0, but then contaminated by substituting some cells with arbitrary values. In a particular row, it may be that either no, a few, or all cells are contaminated. This contamination model allows for some elements of the observed random vector $\mathbf{d}$ to be contaminated, while others remain unaffected. As Raymaekers and Rousseeuw (2023) write, the concept of cellwise outliers represents a paradigm shift from the traditional view of rowwise outliers. Alqallaf et al. (2009) specify that in multivariate statistics, cells are linked to the coordinate system, and orthogonal or other linear transformations can alter these cells. Therefore, the effect of outliers is propagated in the case of cellwise outliers. Take for example a standard multivariate normal distribution, with 4 dimensions and observation $(10, 0, 0, 0)$. An orthogonal transformation can move this point to either $(\sqrt{50}, \sqrt{50}, 0, 0)$ or $(5, 5, 5, 5)$. In this case, any orthogonally invariant casewise detection method yields the same result in the three situations. In the cellwise paradigm however, $(10, 0, 0, 0)$ contains one contaminated cell, $(\sqrt{50}, \sqrt{50}, 0, 0)$ contains two, and $(5, 5, 5, 5)$ has four (Raymaekers & Rousseeuw, 2019).

If we assume a situation with K regressors, where cells are independently contaminated with probability ϵ, then the probability that a row is contaminated in at least one of its cells equals

$$P[\text{row is contaminated}] = 1 - (1 - \epsilon)^K. \tag{1.2}$$

This probability grows rapidly as dimension K increases (Raymaekers & Rousseeuw, 2023).

Traditional robust estimation methods primarily address casewise outliers, where entire observations are anomalous. Most of these methods for robust linear regression rely on the classical Tukey-Huber contamination model (THCM), where a small portion of the cases might be contaminated. The field of robust estimators for panel data models with fixed effects is novel, yet some researchers recently introduced robust and positive breakdown point alternatives to Within Groups LS estimators. Bramati and Croux (2007) extended well known robust regression estimators, namely the least trimmed squares (LTS) estimator by Rousseeuw (1984) and a combination of M- and S-estimates: MS estimates by Maronna and Yohai (2000). Aquaro and Čížek (2013) introduced another robust estimation method for linear panel data with fixed effects, using first-difference and pairwise-difference data transformations. This method applies Gervini and Yohai (2002)'s efficient weighted LS estimator and Čížek (2013)'s reweighted LTS estimator. Following this, Víšek (2015) developed a robust algorithm that employs a smooth decrease of the influence of outliers through weighting of the order statistics of the squared residuals. Additionally, Muhammad, Shamshuddeen and Baoku (2021) propose a Weighted Least Squares (WLS) method, building upon the MM-centering approach initially introduced by Bakar and Midi (2015). This method demonstrates significant resilience against both leverage points and

vertical outliers. More recently, Beyaztas and Bandyopadhyay (2022)'s approach extends the M-estimation methods with various loss functions, incorporating iterative procedures to obtain the tuning parameters.

These methods, however, often inadequately handle cellwise outliers since detection of outliers is based on detecting outlying rows. Because of this, the presence of cellwise outliers can significantly distort the results of traditional casewise robust estimators. This led to a surge in research of cellwise robust methods that can handle such outliers effectly. Most of the statistical research on cellwise outliers focuses on the FICM by (Alqallaf et al., 2009). This model is very useful when the number of considered regressors is small (Saraceno & Agostinelli, 2021). Other papers including Raymaekers and Rousseeuw (2019) allow the cellwise outlying values to depend on some systematic effect or structure, showing a pattern in their deviations. In this case, some underlying factor or process, rather than random aberrations, influences these structured or adversarial outliers. In this paper, we restrict ourselves to the FICM setting for simplicity.

Several papers have studied cellwise outliers in regression methods, although recent research in cellwise robust estimators for static fixed effects panel data methods is lacking. Agostinelli and Yohai (2016) proposed a fully cellwise robust approach for linear mixed models. Their method uses composite τ-estimators that aggregate low-dimension likelihoods, often pairs, to approximate the full likelihood. These robust procedures are resilient against outliers in both the THCM and the FICM, demonstrating high breakdown points and offering a significant advancement in handling high-dimensional data scenarios (Agostinelli & Yohai, 2016).

There is also extensive research into iterative approaches to robust cellwise methods in linear regression. These typically involve a three-step process: detecting outliers, imputing outliers, and then applying a casewise robust method. Estimators that include the first two stages, for example Farcomeni (2014), perform well under FICM, but are not sufficiently robust under THCM (Leung, Zhang & Zamar, 2016).

To detect cellwise outliers, the most direct approach involves examining each variable individually. A more sophisticated method for identifying marginal outliers involves employing a univariate filter, as introduced in Gervini and Yohai (2002). The Gervini-Yohai (GY) filter measures the disparity between the empirical and a reference distribution, subsequently computing an adaptive cutoff value (Agostinelli, Leung, Yohai & Zamar, 2015b). While detecting marginally outlying observations is quick, only extreme cellwise outliers will be reliably detected, as correlations between variables are neglected (Raymaekers & Rousseeuw, 2023). This led to the introduction of methods that consider more than just the marginal contributions. Leung, Yohai and Zamar (2017) extend the univariate GY filter to a bivariate filter in the context of estimating location and scatter. More recently, Saraceno and Agostinelli (2021) further expand the filter to multiple dimensions. They show that the univariate-bivariate filter by Leung et al. (2017) is a special case, if an appropriate statistical depth function is used.

A different method is the Detect Deviating Cells (DDC) algorithm by Rousseeuw and Van den Bossche (2018) for numerical predictors and response variables. DDC uses robust simple linear regressions of each regressor on other regressors in the data matrix (Raymaekers & Rousseeuw, 2023). A residual of each cell is calculated by the difference between the original data and its prediction, and used to flag cellwise outliers. DDC can deal with high dimensions and has

no restriction on the number of clean rows (Rousseeuw & Van den Bossche, 2018). Walach, Filzmoser, Kouřil, Friedeckỳ and Adam (2020) apply a similar technique to metabolomics data, focusing on aggregating deviations in pairwise log-ratios to identify outliers.

After detection of outlying cells, methods often replace the flagged cells by NA's, a process called 'snipping' by Farcomeni (2014). Next, these cells are imputed, for example by estimates of expected values conditional on the observed part of the data, as done in Štefelová, Alfons, Palarea-Albaladejo, Filzmoser and Hron (2021). Their method estimates such conditional expected values by linear regression models, fitted using the casewise robust MM-estimator. Agostinelli et al. (2015b) propose the General Sequential Element (GSE) method that resamples filtered data to compute an initial estimate. Subsequently, the GSE method iteratively performs robust imputation and estimation steps till convergence. In response to Agostinelli et al. (2015b), Agostinelli, Leung, Yohai and Zamar (2015a) propose an alternative that uses a faster initial estimator. Their method imputes filtered cells using coordinate-wide medians to obtain an initial estimator.

In this study, we develop a cellwise robust approach for static fixed effects panel data, prioritising computational efficiency and robustness. We incorporate the adapted Gervini-Yohai outlier filter (Leung et al., 2016), due to its superior detection performance. Subsequently, we impute filtered outliers using coordinate-wise medians for computational feasibility, as suggested by Agostinelli et al. (2015a). The final phase employs the robust casewise WMS method by Bramati and Croux (2007), chosen for its asymptotic efficiency and, regression and affine equivariance properties. The WMS method is similiar to REWLS and RLTS estimators by Aquaro and Čížek (2013) and has similar breakdown points (Beyaztas & Bandyopadhyay, 2022). Additionally, other methods, such as the method by Agostinelli and Yohai (2016) are more computationally intensive. Our objective is to efficiently address both cellwise and casewise outliers, while minimising the computational burden.

We assess our method by conducting a Monte Carlo simulation study, comparing it with the classical Within Estimator used in panel data with fixed effects. This comparison focuses on analysing the bias and standard deviation of the regression parameters. To make inference, we need the asymptotic distribution of the estimator. Therefore, we estimate the bias and standard error estimates over simulations. We assess the models' resilience against local contamination through the empirical influence function (EIF) of the beta estimates. Moreover, we plot the t-test statistic and the EIF of the t-test statistic for the beta estimates with local contamination for both the cellwise robust model and the classical WG estimator. The EIF functions demonstrate that hypothesis testing becomes unreliable in the presence of cellwise outliers.

Lastly, we implement the GY filter on the fixed effects model as discussed in Grossman et al. (2017) to evaluate the significance of addressing cellwise outliers in practical applications. We also conduct a sensitivity analysis. This sensitivity study reveals that cellwise outliers can significantly affect parameter estimates, underscoring the need of their consideration.

The remainder of this paper is structured as follows: Section 2 briefly outlines crucial robustness properties to be assessed for robust methods for static fixed effects panel data. Section 3 specifies the methodology of the proposed estimator, the contamination model, and performance measures used to evaluate the estimators. Subsequently, Section 4 describes the way the

simulations are performed and the calculation of the empirical influence functions. It also reports the obtained results and discusses their interpretation. Section 5 studies the GY filter using empirical data. Finally, Section 6 summarises the results and discusses limitations and suggestions for future work.

Chapter 2

Robustness measures

We define several robustness measures to assess the models' resilience against outliers. In this research, we focus on understanding the sensitivity of the estimator to small perturbations in the data, known as local robustness. Local robustness refers to the impact of adding or modifying a small number of observations in the dataset and observing how these minor changes influence the estimator's behavior. The Influence Function (IF) and the Empirical Influence function (EIF), its empirical counterpart, provide insights into the estimator's stability in the presence of local perturbations.

On the other hand, the breakdown point offers insight into global robustness, which represents the maximum proportion of outlier data that the estimator can handle, before yielding arbitrarily or infinitely large biases. Global robustness assesses the estimator's robustness in scenarios where a significant portion of the data is contaminated. In papers that study casewise robust methods, the breakdown point is defined as the number of cases or rows that are contaminated.

We are particularly interested in the performance of models in the presence of cellwise outliers. Until recently, most research has focused on the Tukey-Huber contamination model (THCM), where a small portion of the data rows is contaminated. In this research, the focus shifts to the Fully Independent Contamination Model (FICM).

2.1 Empirical influence functions

The empirical influence function, $EIF(x, T_n, X)$, evaluates the influence of an observation x on the estimator T_n in a sample X with n observations (Hampel, 1974). The empirical influence function (EIF), or sensitivity curve, is the finite-sample version of the influence function (IF). We require a functional form T of the estimator such that $T(F_n) = T_n$, where F_n is the empirical distribution of the sample, $X = x_1, \ldots, x_n$. The IF of an estimator T is defined as

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T\left\{(1-\epsilon)F + \epsilon\Delta_x\right\} - T(F)}{\epsilon}, \tag{2.1}$$

where F is the cumulative distribution function, and Δ_x is the Dirac delta function, i.e. the point mass function at x. The EIF is the sample version of (2.1), where F is replaced by F_n and ϵ by $1/n$.

Consider a sample, $X = x_1, x_2, \ldots, x_n$, and an estimator T_n based on this sample. The EIF at an observation $x*$ for the estimator T_n is given by

$$EIF(x*; T_n, X) = n \cdot \{T_n(x_1, \ldots, x_{i-1}, x*, x_{i+1}, \ldots, x_n) - T_n(x_1, \ldots, x_n)\}. \tag{2.2}$$

A possible outlying value has a large influence on the estimator when $EIF(x, T_n, X)$ is high, which means the estimator is not robust.

2.2 Breakdown point

Let us examine a dataset, $n \times d$ matrix $\mathbf{X}$, where n represents the number of data points and d the number of dimensions. Our goal is to estimate the unknown location vector by the estimator $\hat{\mu}$. We define the finite-sample cellwise breakdown value of $\hat{\mu}$ at $\mathbf{X}$ as the smallest fraction of contaminated cells per column that pushes the estimator to infinity. Specify $\mathbf{X}_m$ as any contaminated sample with no more than m outlying cells in each column. Hence, the finite-sample cellwise breakdown value of the location estimator $\hat{\mu}$ at $\mathbf{X}$ is defined by (Raymaekers & Rousseeuw, 2023) as

$$\epsilon_n^*(\hat{\boldsymbol{\mu}}, \mathbf{X}) = \min \left\{ \frac{m}{n} : \sup_{\mathbf{X}_m} \|\hat{\boldsymbol{\mu}}(\mathbf{X}_m) - \hat{\boldsymbol{\mu}}(\mathbf{X})\| = \infty \right\}, \tag{2.3}$$

where $\| \cdot \|_2$ denotes the L_2 norm, or Euclidean norm, of the vector. The equation is a finite-sample adaptation of the asymptotic cellwise breakdown value, that was originally proposed by (Alqallaf et al., 2009).

The casewise breakdown value upper bound for an estimator is, clearly, also an upper bound for the cellwise breakdown value of that estimator (Raymaekers & Rousseeuw, 2023). Because of this, (2.3) can be $\left\lfloor \frac{(n+1)}{2} \right\rfloor / n$ maximum for a translation equivariant estimator, by Theorem 2.1 of Lopuhaa and Rousseeuw (1991). We leave further investigation of the breakdown point for cellwise robust estimators - specifically, in the linear panel data model with fixed effects - to further research.

2.3 Tukey-Huber Contamination Model (THCM)

Before defining the methods used to mitigate cellwise outliers, let us first define the contamination methods considered in this research.

Most robust linear regression methods are based on the classical Tukey-Huber contamination model (THCM) (Leung et al., 2016). The THCM has a significant influence in the general approach of most robust statistical procedures: to identify outlying cases. For panel data models with fixed effects, this is generally the only contamination model applied. THCM assumes that a small fraction of cases is contaminated, reflecting the presence of outliers. This casewise contamination model describes the contamination process as a mixture of two distributions: one being the standard model and the other for outliers. Robust statistical analysis aims to conduct inference on the dominant part of the mixture, filtering out outliers generated by the outlier distribution (Alqallaf et al., 2009). The method by Huber (1964) and Tukey (1962) is specified

as follows:

$$\mathbf{d} = (1 - \epsilon)\mathbf{d_0} + \epsilon\widetilde{\mathbf{d}}. \tag{2.4}$$

Here, $\mathbf{d}$ represents a random vector. Under THCM, $\mathbf{d} \sim F$ is modeled as a mixture of two distributions, where F_0 is the distribution of the uncorrupted data $\mathbf{d_0}$, $\widetilde{F}$ is an arbitrary outlier distribution for $\widetilde{\mathbf{d}}$ and ϵ $(0 \leq \epsilon \leq 1)$ is the proportion of contamination.

2.4 Fully Independent Contamination Model (FICM)

Contamination, however, can occur differently in the sense that individual components in $\mathbf{X}$ are independently contaminated. In high-dimensional datasets often used in applications, where variables are collected seperately, each is susceptible to contamination independently (Leung et al., 2016). This cellwise contamination can significantly impact the effectiveness of traditional (casewise) robust methods. Alqallaf et al. (2009) introduce a new model to address this issue, known as the Fully Independent Contamination Model (FICM). Alqallaf et al. (2009) formulate the FICM as follows

$$\mathbf{d} = (\mathbf{I} - \mathbf{B}_\epsilon)\mathbf{d_0} + \mathbf{B}_\epsilon\widetilde{\mathbf{d}}, \tag{2.5}$$

where $\mathbf{B}_\epsilon = \text{diag}(b_1, b_2, \ldots, b_k)$ is a diagonal matrix and each b_j is an i.i.d. $Bin(1, \epsilon)$ random variables. Stated differently, each value in $\mathbf{d}$ has a probability of ϵ of being independently contaminated. That makes the probability that at least one element of $\mathbf{d}$ is contaminated equal to

$$\hat{\epsilon} = 1 - (1 - \epsilon)^k. \tag{2.6}$$

In a data matrix with regressors and response variables, this equals the probability that at least one element in a row is contaminated. Therefore even a small ϵ results in a large $\hat{\epsilon}$ in the case of multiple dimensions (Raymaekers & Rousseeuw, 2023). Alqallaf et al. (2009) show that the breakdown point of all traditional 0.5 breakdown point affine equivariant location estimators is $1 - 0.5^{1/k}$ as $k \to \infty$ (Leung et al., 2016). This explains the need for another, cellwise robust method to tackle both cellwise and casewise outliers.

Chapter 3

Methodology

3.1 Static panel data model with fixed effects

The static fixed effects panel data model is used to analyse data that varies across individuals and entities over time. This model is particularly useful for examining the impact of variables that change over time while controlling for unobserved heterogeneity that is constant over time but varies between individuals. The essence of the fixed effects model is to allow for individual-specific intercepts, capturing the effect of omitted variables that do not change over time for each individual.

Consider a dataset with $i = 1, \ldots, N$ individuals observed over $t = 1, \ldots, T$ time periods. The static fixed effects model can be represented by the following equation:

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \epsilon_{it}, \tag{3.1}$$

where y_{it} is the dependent variable for individual i at time t, α_i is the unobserved individual-specific effect, $\mathbf{x}_{it}$ is a $K \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $K \times 1$ vector of coefficients to be estimated, and ϵ_{it} is the idiosyncratic error term. The α_i captures the unobserved heterogeneity across individuals that could bias the estimates of $\boldsymbol{\beta}$ if not properly accounted for.

In later sections, we describe the traditional estimation of the fixed effects model and propose a cellwise robust alternative.

3.2 Detection: Adapted Gervini-Yohai filter

The first step in handling cellwise outliers is detecting them. Gervini and Yohai (2002) introduce a univariate filter to identify marginal outliers. The Gervini-Yohai (GY) filter measures the disparity between the empirical distribution and a reference distribution, subsequently computing an adaptive cutoff value.

While detecting marginally outlying observations is quick, only extreme cellwise outliers will be reliably detected, as correlations between variables are neglected (Raymaekers & Rousseeuw, 2023). Therefore, considering more than marginal contributions is essential. We consider a univariate plus bivariate filter in this research, as proposed by Leung et al. (2017) in the context of estimating location and scatter. We use univariate and bivariate filter because of our emphasis

on cellwise outliers, compared to cellwise outliers. Compared to a multivariate filter, this also significantly reduces the computational complexity.

pointed out that filtering the variables based solely on their value may be too limiting as no correlation with other variables is taken into account. A not-so-large contaminated cell that passes the univariate filter could be flagged when viewed together with other correlated components, especially for highly correlated data. To overcome this deficiency, we introduce a consistent bivariate filter and use

3.2.1 Univariate filter

The univariate filter aims to detect large cellwise outliers by looking at marginals. Gervini and Yohai (2002) propose the use of an adaptive cutoff, instead of a fixed cutoff value. Leung et al. (2017) assume a consistent filter for a distribution F_0 as one that asymptotically will not mistakenly classify a cell as an outlier if the data is derived from F_0. For a clean dataset, this means that the proportion of flagged outliers approaches zero as the sample size tends to infinity: $\lim_{n\to\infty} f_n = 0 \quad \text{a.s.}\,[F_0]$. Here, f_n is used to denote the proportion of flagged outliers.

To illustrate the adapted GY filter, we consider $x_1, \ldots, x_n$, a random (univariate) sample of observations. We introduce the location estimator T_{0n} and estimator of scatter S_{0n}, for which we adopt the median and median absolute deviation (MAD) in this paper, respectively as in Leung et al. (2017, 2016); Agostinelli et al. (2015b). The standardised version of x_i is $z_i = \frac{x_i - T_{0n}}{S_{0n}}$ and F the chosen reference distribution for z_i. Although F_0 is unknown in practice, a standard normal distribution is typically assumed. In our case, we know the distribution F_0 is standard normal since we simulate the data. We define adapted cutoff values by Gervini and Yohai (2002) as follows

$$F_{+n}(t) = \frac{1}{n}\sum_{i=1}^{n} I(|z_i| \leq t), \tag{3.2}$$

where F_{+n} is the empirical distribution function. We define the fraction of outliers by

$$\begin{aligned} d_n &= \sup_{t\geq\eta} \{F_+(t) - F_{+n}(t)\}_+ \\ &= \max_{i>i_0} \left\{ F_+(|z|_{(i)}) - \frac{i-1}{n} \right\}_+ . \end{aligned} \tag{3.3}$$

Here $\{F\}_+$ denotes the positive component of the distribution of $|z|$ when $z \sim F$. The term $|Z|_{(i)}$ refers to the order statistics of $|z_i|$. We define i_0 as the maximum index for which $|z|_{(i)}$ is less than threshold η. η is a quantile of F_+, $\eta = (F_+)^{-1}(\alpha)$. We set $\alpha = 0.95$ to detect large outliers in the univariate filter, as is the default in Leung et al. (2017, 2016); Agostinelli et al. (2015b). We then flag $\lfloor nd_n \rfloor$ - the greatest integer not exceeding nd_n - cells with the highest standardised values as cellwise outliers, subsequently replacing these with NA's. We then determine the adaptive cutoff value for each z_i as

$$t_n = \min \{t : F_{+n}(t) \geq 1 - d_n\}, \tag{3.4}$$

which corresponds to $t_n = Z_{(i_n)}$, with $i_n = n - \lfloor nd_n \rfloor$. In practice, this means we flag x_i when $|z_i| \geq t_n$.

3.2.2 Consistent bivariate filter

Consider a random sample of bivariate observations $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where each $\mathbf{x}_i = (x_{i1}, x_{i2})'$. We introduce a pair of initial location and scatter estimators,

$$\mathbf{T}_{0n} = \begin{pmatrix} T_{0n,1} \\ T_{0n,2} \end{pmatrix} \text{ and } \mathbf{C}_{0n} = \begin{pmatrix} C_{0n,11} & C_{0n,12} \\ C_{0n,21} & C_{0n,22} \end{pmatrix}.$$

Following the univariate approach, we use the coordinate-wise median for $\mathbf{T_{0n}}$. For $\mathbf{C_{0n}}$, we use the bivariate estimator with MAD scale by Gnanadesikan and Kettenring (1972). Specifically, we define the initial scatter estimators as

$$C_{0n,jk} = \frac{1}{4}\left\{\mathrm{MAD}(x_{ij} + x_{ik})^2 - \mathrm{MAD}(x_{ij} - x_{ik})^2\right\}, \tag{3.5}$$

where $\mathrm{MAD}(\{y_i\})$ is the MAD of $y_1, \ldots, y_n$. We note that $C_{0n,jj} = \mathrm{MAD}(\{\mathbf{x}_j\})^2$, consistent with our choice of coordinate-wise dispersion estimators.

Then, we define the pairwise (squared) Mahalanobis distances as $D_i = (\mathbf{x}_i - \mathbf{T}_{0n})'\mathbf{C}_{0n}^{-1}(\mathbf{x}_i - \mathbf{T}_{0n})$. The empirical distribution for these distances is

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(D_i \leq t). \tag{3.6}$$

Outlying points x_i are then identified by comparing $G_n(t)$ with a reference distribution G. In this study, we use the chi-squared distribution with two degrees of freedom, $G = \chi^2_2$. The fraction of flagged bivariate outliers is noted as

$$d_n = \sup_{t \geq \eta}\{G(t) - G_n(t)\}_+, \tag{3.7}$$

where $\eta = G^{-1}(\alpha)$. For bivariate filtering aimed at moderate outliers, we set $\alpha = 0.85$, though other values of α may be considered. Subsequently, we flag $\lfloor nd_n \rfloor$ observations with the greatest pairwise Mahalanobis distances as bivariate outliers. We define the adaptive cutoff value for the distances for the bivariate filter in the same way as in (3.4).

3.2.3 Univariate and bivariate filter

We begin the detection of large cellwise outliers with the univariate filter. The bivariate filter then aims to detect moderate cellwise outliers by incorporating information about the correlation structure of the data (Leung et al., 2016). When viewed marginally, a moderately contaminated cell passes the filter, but can be flagged when other variables are taken into account, especially for highly correlated data (Rousseeuw & Van den Bossche, 2018). We describe the following adapted approach from Leung et al. (2017) adapted for static fixed effects panel data. We assume a $NT \times K$ matrix, which is displayed in R as follows

$$\mathbb{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,K} \\ x_{1,1}^{(2)} & x_{1,2}^{(2)} & \cdots & x_{1,K}^{(2)} \\ x_{2,1}^{(2)} & x_{2,2}^{(2)} & \cdots & x_{2,K}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1}^{(2)} & x_{N,2}^{(2)} & \cdots & x_{N,K}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,1}^{(T)} & x_{1,2}^{(T)} & \cdots & x_{1,K}^{(T)} \\ x_{2,1}^{(T)} & x_{2,2}^{(T)} & \cdots & x_{2,K}^{(T)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1}^{(T)} & x_{N,2}^{(T)} & \cdots & x_{N,K}^{(T)} \end{pmatrix},$$

where each $x_{i,j}^{(t)}$ can be independently contaminated. We must note that for simplicity, we assume pooled data for the filter, meaning we do not distinguish between time periods and cross-sectional units. To illustrate, if we would take into account averages over time periods, for example, to take into account deviations between cross-sections, we would have $k \times N$ values of T_0 to estimate, whereas, in the current situation, we have K.

We apply the univariate filter separately to each variable in $\mathbb{X}$, consisting of N individuals observed over T time periods. We use the initial location and dispersion estimators $\mathbf{T}_{0,NT} = (T_{0,NT,1}, \ldots, T_{0,NT,k})$ and $\mathbf{S}_{0,NT} = (S_{0,NT,1}, \ldots, S_{0,NT,k})$. We introduce an auxiliary binary matrix $\mathbb{U}$, with zeros indicating filtered entries in $\mathbb{X}$ across both individuals and time periods. We then proceed to evaluate all pairs of variables for each individual at each time period to identify bivariate outliers, which helps filter cells with moderate contamination.

For each individual at each time period, denoted as (i, t), and for a selected pair of variables (x_{itj}, x_{itp}), we set $\mathbf{x}_{it}^{(jp)} = (x_{itj}, x_{itp})$, with $1 \leq j < p \leq k$. We then compute an initial pairwise scatter matrix estimator $\mathbf{C}_{0,NT}^{(jp)}$ for this pair. We compute $\mathbf{d}_{it}^{(jp)} = (\mathbf{x}_{it}^{(jp)} - \mathbf{T}_{0,NT}^{(jp)})^t (\mathbf{C}_{0,NT}^{(jp)})^{-1} (\mathbf{x}_{it}^{(jp)} - \mathbf{T}_{0,NT}^{(jp)})$, the pairwise Mahalanobis distances, and apply bivariate filtering to these pairwise distances, excluding those with flagged components by the univariate filter: $\{\mathbf{d}_{it}^{(jp)} : u_{itj} = 1, u_{itp} = 1\}$. We repeat this procedure for all variable pairs, where $1 \leq j < p \leq k$, for each individual at each time period.

We let

$$J = \{(i, t, j, p) : \mathbf{d}_{it}^{(jp)} \text{ is flagged as a bivariate outlier}\},$$

be a set of triplets, identifying pairs of cells detected by the bivariate filter across individuals and time periods ($i = 1, \ldots, N$, $t = 1, \ldots, T$). To determine the cells (i, t, j) that should be detected as cellwise outliers, we first count the number of flagged pairs involving cell (i, t, j) for each individual, at each time period, which corresponds to row $i \times t$ in $\mathbb{X}$:

$$m_{it,j} = \#\{p : (i, t, j, p) \in J\}. \tag{3.8}$$

Cells with large $m_{it,j}$ are most probably univariate outliers. If we assume that $x_{it,j}$ is not cellwise contaminated, then $m_{it,j}$ is approximately binominally distributed, $\text{Bin}(\sum_{k \neq j} u_{itk}, \delta)$, under the FICM. Here, δ is the overall fraction undetected by the univariate filter. The observation $x_{it,j}$ is flagged if

$$m_{it,j} > c_{it,j},$$

where $c_{it,j}$ is the 0.99-quantile of the respective binomial distribution. We use a conservative $\delta = 0.10$ as is used in Leung et al. (2016, 2017). These values prove to have good results in simulation and real data (Leung et al., 2017).

3.3 Imputation: coordinate-wise medians

After detecting cellwise outliers using the adapted Gervini-Yohai approach, we impute the filtered cells. We introduce a similar approach to that of Agostinelli et al. (2015a), imputing the filtered cells using coordinate-wise medians. We calculate the median across columns, disregarding the NA's. Then, this median replaces the NA's in that particular column. Such an imputation approach is based on a pooled specification, assuming that each variable's distribution can be independently considered, simplifying the imputation process for panel data with missing values. We apply the casewise robust method by Bramati and Croux (2007) to the imputed data.

3.4 Cellwise robust estimation

We first specify the general formulation of the linear panel data model with fixed effects, individuals ($i = 1, \ldots, N$) and time points ($t = 1, \ldots, T$). We present the relationship as $y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it}$, where y_{it} denotes the dependent variable. The $K \times 1$ vector $\mathbf{x}_{it}$ displays the explanatory variables. The model includes a $K \times 1$ vector of regression coefficients $\boldsymbol{\beta}$. The fixed individual-specific effects α_i are time-invariant and unobserved. The error terms ε_{it} are assumed to be uncorrelated through time and cross-sections.

In matrix form, often referred to as the stacked form representation, the linear static fixed effects panel data model is expressed as

$$\mathbf{y} = \mathbf{e}_T \otimes \boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.9}$$

where $\mathbf{y}$ is an $NT \times 1$ vector containing all observations, and $\mathbf{X}$ is an $NT \times K$ matrix of explanatory variables. Here, $\boldsymbol{\alpha}$ is an $N \times 1$ vector representing individual effect coefficients, $\mathbf{e}_T$ is a $T \times 1$ vector consisting of ones, and $\otimes$ denotes the Kronecker product.

3.4.1 General Within Groups estimator

Our analysis involves contrasting robust estimators with the traditional Within Groups (WG) estimators. In this approach, the dataset undergoes an initial transformation where each time

series is centered. This centering looks as follows:

$$\tilde{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}, \tag{3.10}$$

and similarly for the explanatory variables:

$$\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}. \tag{3.11}$$

Post-centering, we obtain

$$\tilde{y}_{it} = \boldsymbol{\beta}'\tilde{\mathbf{x}}_{it} + \text{error}_{it}, \tag{3.12}$$

where the fixed effects α_i are removed through the centering process. Applying Ordinary Least Squares (OLS) estimator of $\tilde{y}_{it}$ on $\tilde{\mathbf{x}}_{it}$ yields the Within Groups estimator $\hat{\beta}_{WG}$. While the primary focus is on estimating β, the fixed effects parameters are also derivable. For a comprehensive discussion on the classical Within Groups estimator, Baltagi and Baltagi (2008) provide an extensive review.

3.4.2 Equivariance properties

Before introducing robust alternatives to the within estimator for panel data, we discuss equivariance properties that a regression estimator in panel data should satisfy (Rousseeuw & Leroy, 2005). Let us define R as an estimator for the regression coefficient β in (3.9). An estimator R in panel regression is scale equivariant if the following condition holds:

$$R(\mathbf{x}_{it}, cy_{it}) = cR(\mathbf{x}_{it}, y_{it}) \tag{3.13}$$

for any scalar c. Additionally, the estimator is regression equivariant if

$$R(\mathbf{x}_{it}, y_{it} + \mathbf{x}_{it}\gamma) = R(\mathbf{x}_{it}, y_{it}) + \gamma, \tag{3.14}$$

where γ is a $K \times 1$ constant vector. Moreover, R is affine equivariant if it satisfies

$$R(\mathbf{A}\mathbf{x}_{it}, y_{it}) = \mathbf{A}^{-1}R(\mathbf{x}_{it}, y_{it}), \tag{3.15}$$

for all individuals $i = 1, \ldots, N$ and time points $t = 1, \ldots, T$, and for any $K \times K$ nonsingular matrix $\mathbf{A}$ (Bramati & Croux, 2007). The classical Within Groups estimator fulfills these three equivariance criteria.

3.4.3 The Within Groups MS (WMS) Estimator

Outlying observations disrupt the classical Within Groups estimator. Bramati and Croux (2007) built the Within Groups MS (WMS) estimator to deal with casewise outliers. After detection and imputation of cellwise outliers, the WMS estimator is used to estimate the regression parameters and deal with casewise outliers. To obtain a robust alternative to the Within Groups estimator, we robustly estimate the center of each time-series, using the median, for both the dependent

and independent variables, and subtract it from each observation in the block. Afterwards, we use a robust regression method on the centered data to obtain robust coefficient estimates (Bramati & Croux, 2007).

Bramati and Croux (2007)'s WMS estimator is a particular application of the MS regression estimator by Maronna and Yohai (2000). The approach robustly estimates both continuous and categorical regressors and is affine and regression equivariant. Moreover, the method actively alternates between M-estimators for categorical variables and S-estimators for continuous variables. M-estimators are quick to calculate, though they may lack robustness against leverage points. However, since categorical and particularly dummy variables rarely present large leverage points, we can efficiently apply M-estimators for regression on these variables. The M-estimator diminishes the impact of categorical variables compared to continuous variables. The S-estimator is robust, also with regards to leverage points. Therefore, the S-estimator is applied to continuous variables. (Bramati & Croux, 2007) implemented the WMS estimator for numerical variables only, therefore we only implement the method for numerical variables.

Let us assume that the fixed effects α_i are known. We obtain an S-estimate of regression by minimising the M-estimator of scale derived from the regression residuals $r_{it}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} - \alpha_i$ (Rousseeuw & Leroy, 2005). We determine the M-estimator of scale S by solving the equation for s:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_S\left\{\frac{r_{it}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{s}\right\} = b, \tag{3.16}$$

Here, $\rho_S(\cdot)$ is an even, symmetric, and continuously differentiable loss function with $\rho_S(0) = 0$ (Bramati & Croux, 2007). The value of b is set to be equal to $\mathbb{E}[\rho_S(\varepsilon)]$, where ε denotes the standard normal distribution, to ensure consistent estimates for the regression scale parameter. We then define the S-estimator of regression as the argument that minimises S, that is

$$\hat{\boldsymbol{\beta}}_S(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{\beta}} S\left\{r_1(\boldsymbol{\alpha}, \boldsymbol{\beta}), \ldots, r_{NT}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right\}. \tag{3.17}$$

S-estimators are more efficient than the LTS estimator and have a higher breakdown point in regression analysis (Bramati & Croux, 2007). For the loss function, we employ the Tukey Biweight function ($\rho(q)$), defined as

$$\rho(q) = \begin{cases} \frac{c^2}{6}\left[1 - \left\{1 - \left(\frac{q}{c}\right)^2\right\}^3\right], & \text{for } |q| \leq c, \\ \frac{c^2}{6}, & \text{for } |q| > c, \end{cases}$$

where q represents the residual in a regression model, and c is a tuning constant determining the threshold for outlier identification (Rousseeuw & Leroy, 2005). c is chosen such that the regression estimator has an overall breakdown point of 25%, as in Bramati and Croux (2007). The S-estimator can be computationally heavy when the dimension of $\boldsymbol{\beta}$ is large. We do not optimise $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ simultaneously, because $\boldsymbol{\alpha}$ can contain many elements, making it time intensive to compute the S-estimator. Alternatively, consider the scenario where $\boldsymbol{\beta}$ is known.

The estimator $\hat{\boldsymbol{\alpha}}$, defined as $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)$, is derived as an M-estimator of regression

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho_M \left\{ r_{it}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right\}, \tag{3.18}$$

which implies

$$\hat{\boldsymbol{\alpha}}_i(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\alpha}_i} \sum_{t=1}^{T} \rho_M \left(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \boldsymbol{\alpha}_i \right), \text{ for } i = 1, \ldots, N. \tag{3.19}$$

Given the presence of many fixed effects, quick estimation is crucial. We suggest using $\rho_M(\cdot) = |\cdot|$, the absolute value loss function. This provides a clear formula for the above expression:

$$\hat{\boldsymbol{\alpha}}_i(\boldsymbol{\beta}) = \text{median}_t \left(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} \right), \text{ for } i = 1, \ldots, N. \tag{3.20}$$

The WMS estimator for a linear panel data model, denoted as $\hat{\boldsymbol{\beta}}_{\text{WMS}}$, is then defined as

$$\hat{\boldsymbol{\beta}}_{\text{WMS}} = \arg\min_{\boldsymbol{\beta}} S \left[r_1 \left\{ \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta} \right\}, \ldots, r_{NT} \left\{ \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta} \right\} \right], \tag{3.21}$$

where

$$r_{it} \left\{ \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta} \right\} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \text{median}_t(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}). \tag{3.22}$$

These equivariance properties remain intact by jointly minimising $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

3.5 Quantile-Quantile plots

Ideally, we would like to prove the asymptotic properties of an estimator analytically. However, the complexity of the cellwise robust estimator compels us to investigate the asymptotic distribution numerically, by comparing it directly with the normal distribution. We plot Quantile-Quantile (QQ) plots to illustrate how closely the asymptotic distribution of the GY-WMS BC estimator aligns with the normal distribution, without outliers.

For the robust estimator, the QQ plot of the estimator should adhere to a straight line, indicating conformity with the normal distribution. Deviations from this line, particularly in the plot's tails, show a deviation from the normal distribution.

3.5.1 Asymptotic distribution

We evaluate the asymptotic distribution of the GY-WMS BC estimator numerically. In the Quantile-Quantile (QQ) plots depicted in Figure 3.1 and Figure 3.2, we plot the GY-WMS BC estimator for the scenario with uncorrelated and correlated regressors, respectively.

We conclude that both for the scenario with correlated and uncorrelated regressors, the estimators' distribution closely adheres to the normal distribution. For the estimator of β_2 with uncorrelated regressors, we see a slight downward tilt at the start, suggesting that the lower end of the estimator's distribution has fewer extreme low values than the normal distribution. For the scenario with correlated regressors, we see a bit more deviation from the normal distribution when looking at some of the tails of the estimators. Specifically, we see that the upper tails of

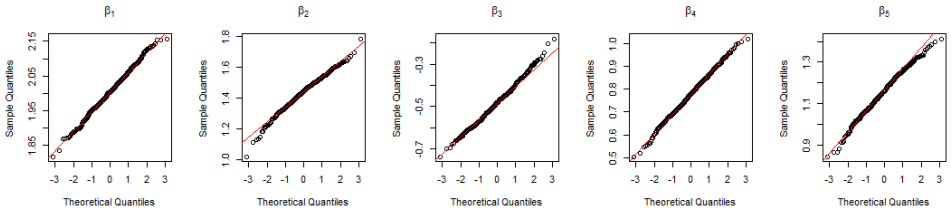

Figure 3.1: QQ plot of the GY-WMS BC estimator compared to a normal distribution, for uncorrelated regressors

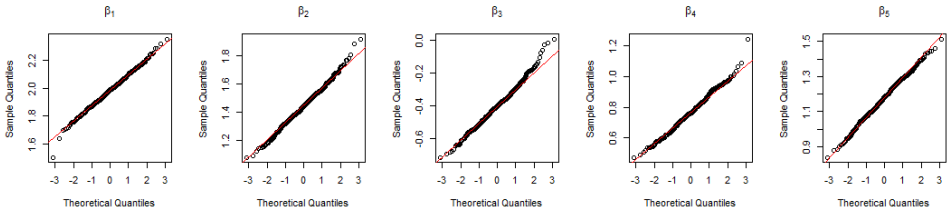

Figure 3.2: QQ plot of the GY-WMS BC estimator compared to a normal distribution, for correlated regressors

$\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$ tilt up slightly, suggesting that the upper end of the distribution has more extreme values than the normal distribution predicts.

3.6 Performance measures

The β parameters in the linear static fixed effects model are of main interest, because these parameters describe the influence of the explanatory variables on the outcome variable. We can use various performance measures to assess an estimator T_n. In this paper, we focus on the bias and the variance.

3.6.1 Bias

Bias represents the difference between an estimator's expected value $\hat{\theta}$ and the true parameter value θ. For an estimator T_n aimed at estimating a parameter θ, bias is expressed as $\text{Bias}(T_n) = E[T_n] - \theta$, where $E[T_n]$ denotes the expected value of T_n. Bias reflects a systematic, as opposed to random, error in estimation. An ideal estimator is unbiased, where $\text{Bias}(T_n) = 0$. Additionally, an estimator is considered consistent if it converges in probability to the true parameter value as the sample size tends to infinity.

3.6.2 Variance

Variance measures the dispersion within a dataset, representing the average squared deviation from the mean and thus indicating how the estimate changes over different simulations. We denote the variance as $\text{Var}(T_n) = E[\{T_n - E(T_n)\}^2]$, or $\text{Var}(T_n) = E(T_n^2) - \{E(T_n)\}^2$. Alternatively, we can describe it as the covariance of the estimate T_n A lower variance indicates that the estimates are close to each other across simulations, while a high variance suggests a wider

spread. If there is no systematic bias, the point estimates tend to be nearer to the true parameter value as the sample size increases. Therefore, the variance of the estimate decreases when the sample size increases.

In the Monte Carlo simulation, we calculate the sample version of the standard deviation, the standard errors, by calculating the standard deviation of the betas across simulations. For the empirical study, we use a bootstrapping technique where we bootstrap new data and use the β estimates of those bootstrapped samples to compute the standard deviation. Specifically, we apply a type of block bootstrapping, where we resample entire entities (countries) of the panel data to account for temporal correlation between the variables.

Chapter 4

Simulation study

4.1 Linear static fixed effects model

We initiate the simulation by setting a reproducible seed, to ensure consistency in our results. Our panel dataset includes $N = 100$ individuals, each observed over $T = 5$ time periods. To show the effect on a multivariate problem, $X_1, X_2, X_3, X_4,$ and $X_5 \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ for every individual at each time point. We generate dependent variable y_{it} as

$$y_{it} = \beta_0 + \alpha_i + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \beta_3 x_{it,3} + \beta_4 x_{it,4} + \beta_5 x_{it,5} + \epsilon_{it}, \tag{4.1}$$

where $\alpha_i \sim \mathcal{N}(0, 1)$ denotes the individual-specific effect and $\epsilon_{it} \sim \mathcal{N}(0, 1)$ represents the idiosyncratic error term. We set the coefficients $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1.5$, $\beta_3 = -0.5$, $\beta_4 = 0.8$, and $\beta_5 = 1.2$. Subsequently, we draw observations from the model described above.

Across 500 simulations, we estimate the fixed effects model using the Within Groups estimator provided by the `plm` package. We also estimate the beta parameters using the Within Groups estimator in the code provided by Bramati and Croux (2007), to provide a fair comparison between the classical and the casewise robust method. The provided code was written in Matlab and is translated into R for this paper.

4.2 Correlated regressors

Next to independent and identically distributed regressors, we also study a scenario with correlated regressors, to reflect common occurences in empirical datasets more accurately. We use a Toeplitz matrix to systematically structure the covariance among regressors. While often associated with temporal or spatial autocorrelation, here, the Toeplitz matrix structures the covariance among regressors based on their relative positions to each other, not temporal lags. Therefore, the correlation between X_1 and X_2 is the same as between X_3 and X_4, following a systematic pattern of correlation that decreases with the increase in distance between variable indices.

The Toeplitz covariance matrix is constructed as follows:

$$\mathbf{\Sigma_X} = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{K-1} \\ \rho & 1 & \rho & \dots & \rho^{K-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{K-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{K-1} & \rho^{K-2} & \rho^{K-3} & \dots & 1 \end{pmatrix}, \tag{4.2}$$

where K represents the number of regressors, and ρ denotes the base correlation between adjacent regressors, set at 0.5 for our simulation study. This structure ensures that regressors are correlated due to inherent characteristics or interactions among the variables.

4.3 Cellwise outliers

Only 1% of cells in the dataset are contaminated, without restrictions on the number of contaminated rows. We insert outliers according to the FICM by Alqallaf et al. (2009); replacing selected cells' values with samples drawn from an alternative normal distribution, with $\mu = 10$ or $\mu = -10$, depending on the outlier sign. The standard deviation σ equals 1. This contamination is equivalent to adding a magnitude of (negative) 10 to the uncontaminated cells to keep interpretation simple.

4.4 Estimation

4.4.1 WMS estimator

We take a subsample I of size K from the set $\{(x_{it} - \text{median}_t x_{it}, y_{it} - \text{median}_t y_{it}) | 1 \leq i \leq N, 1 \leq t \leq T\}$. For this subsample, a parameter $\hat{\boldsymbol{\beta}}_I$ exactly fits the observations within I. The fit of $\hat{\boldsymbol{\beta}}_I$ over the entire dataset is evaluated using

$$s_I = S\left[r_1\left\{\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}_I), \hat{\boldsymbol{\beta}}_I\right\}, \dots, r_{NT}\left\{\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}_I), \hat{\boldsymbol{\beta}}_I\right\}\right]. \tag{4.3}$$

We generate $N_{\text{samp}} = 500$ of such subsamples I, and choose the subsample with the minimal s_I. The chosen N_{samp} ensures, with a certain probability, the presence of at least one casewise outlier-free subsample based on the outlier proportion ε in the data, according to Bramati and Croux (2007). We have to take into account, however, that Bramati and Croux's procedure initially only takes casewise contamination into account. In the GY filtering process, we aim to tackle all cellwise outliers, but most likely, some cases will still be contaminated.

Assuming the absence of cellwise outliers, the likelihood of obtaining at least one casewise outlier-free subsample approaches $1 - \left\{1 - (1-\varepsilon)^K\right\}^{N_{\text{samp}}} > 0$ as the number of observations increases (Bramati & Croux, 2007).

The optimal fit $\hat{\boldsymbol{\beta}}_0$, found through the minimal s_I, is supposedly near the global solution $\hat{\boldsymbol{\beta}}_{\text{WMS}}$. Starting with $\hat{\boldsymbol{\beta}}_0$, an iterative algorithm refines the estimate towards minimising (3.21). At iteration $(k+1)$, we define

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} S\left[r_1\left\{\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}^{(k)}), \boldsymbol{\beta}\right\}, \ldots, r_{NT}\left\{\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}^{(k)}), \boldsymbol{\beta}\right\}\right]. \tag{4.4}$$

The corresponding first-order condition is given by

$$\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}\mathbf{x}_{it}\left\{y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} - \hat{\boldsymbol{\alpha}}_i(\hat{\boldsymbol{\beta}}^{(k)})\right\} = 0. \tag{4.5}$$

Here, $w_{it} = W\left[r_{it}\left\{\boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}^{(k)}), \boldsymbol{\beta}\right\}\right]$ are weights derived from the weighting function $W(r) = \frac{\rho\{S(r)\}}{r}$. Due to the unknown $\boldsymbol{\beta}$ to compute the weights, w_{it} is approximated using $W\left[r_{it}\left\{\boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}^{(k)}), \hat{\boldsymbol{\beta}}^{(k)}\right\}\right]$, making (4.5) linear and solvable for $\hat{\beta}^{(k+1)}$. Subsequently, we compute $\hat{\alpha}(\hat{\beta}^{(k+1)})$ as in 3.20 and the weights, after which the iteration continues. As suggested by Maronna and Yohai (2000), the process iterates a fixed number of times, selecting the $\hat{\boldsymbol{\beta}}^{(k)}$ that minimises the objective function in (3.21). In our case, we take $M = 20$ as in Bramati and Croux (2007).

4.5 Empirical influence function

We want to assess the influence of a single outlier in $x_{it,j}$ or y_{it} on the β parameters of the Within Groups estimator to show why a (cellwise) robust method is needed to handle outliers. Then, we compare this EIF to the effect of a single (cellwise) outlier on the WMS estimator with the Gervini-Yohai filter and coordinate-wise median imputation. We evaluate the empirical influence function (EIF) on the β parameters of the before-mentioned estimators. This is done for cellwise outliers in $\mathbf{X}$ and vertical outliers, respectively.

4.5.1 Simulation of cellwise contamination in X

We introduce a single contaminated cell in the matrix of independent variables $\mathbf{X}$ during each simulation run. This approach enables us to methodically examine the empirical influence function (EIF) by observing the variation in estimated β coefficients in response to outliers of different magnitudes. We consider the same samples as used for the estimation procedure in Section 4.1.

For the precise evaluation of the EIF concerning β_j, we evaluate 500 simulation runs. Within these iterations, contamination is placed in $x^*_{it,j}$, a random cell within column j of $\mathbf{X}$. Column j directly relates to the β^*_j we aim to examine. We can therefore analyse how an outlier in a given column impacts its corresponding estimator. The contaminated values introduced in $x^*_{it,j}$ span from -6 to 6, covering a spectrum of potential outlier magnitudes.

4.5.2 Simulation of vertical outliers

To examine the effect of vertical outliers on an estimator $\hat{\beta}_j$, we now contaminate y^*_{it} in each simulation run by integer values from -6 to 6. The dataset is simulated as described in Section 4 and repeated for 500 simulations.

4.5.3 Calculation of the EIF for $\hat{\beta}_j$

We evaluate the impact of these outliers on the Within Groups (WG) estimator using the `plm` package in R. Subsequently, we compare the EIF for the WG estimator to the EIF of the robust method. The cellwise robust WMS estimator, in combination with the GY filter and coordinate-wise mean imputation, aims to mitigate the influence of cellwise outliers. Therefore, we expect the EIF of the latter not to break down.

For each contamination level, we compute and store the estimated β_j coefficients from the cellwise robust method and the WG method. This process is repeated across all simulations, such that we capture the variability of β estimates in response to cellwise contamination.

To estimate the EIF, we also calculate the non-contaminated $\hat{\beta}_j$ across simulations. The EIF_{β_j} is then computed by $N \times (\hat{\beta}_j^* - \hat{\beta}_j)$, as derived from (2.2). Subsequently, we plot boxplots for this EIF_{β_j} across contamination values $x_{it,j}^*$ ranging from integers -6 to 6.

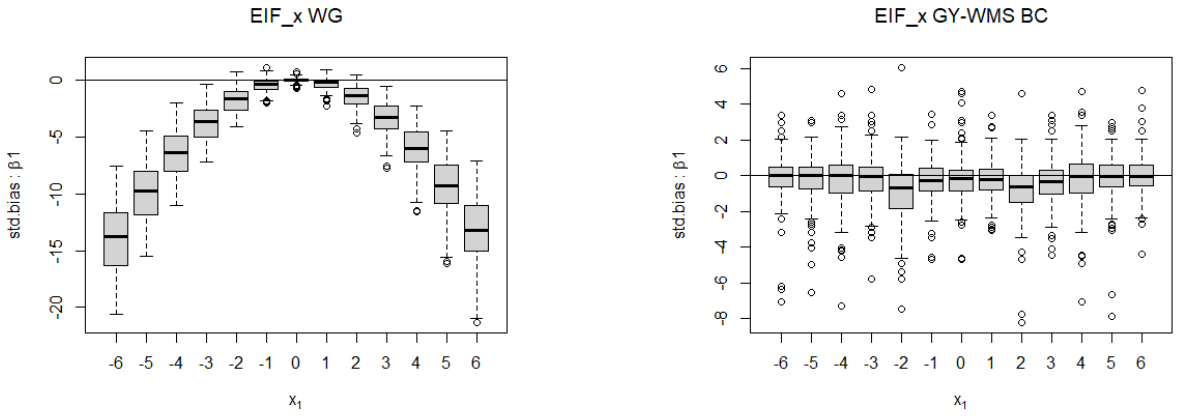

Figure 4.1: EIF plots for β_1 with outliers in $x_{it,1}^*$, for $M = 100$

Figure 4.1 illustrates the EIF plots for β_1 for the WG estimator and the GY-WMS estimator, respectively, in the case of a single contaminated $x_{it,1}^*$. These plots represent the empirical influence functions for data with uncorrelated regressors. Therefore, we assume the plots for β_1 to be representative of behavior of the other beta estimators. On the y-axis, we observe the standardised bias of β_1, while the variation of the estimates is depicted by the boxplots.

By variation or spread, we mean the interquartile range (IQR) reflected in the length of the box and whiskers' length, which extends to the largest (absolute) data point within 1.5 times the IQR. As anticipated, the WG estimator breaks down completely as $x_{it,1}^*$ increases in absolute value. This breakdown is evident from the increased absolute standardised bias and the widened spread.

Regarding the GY-WMS BC estimator, we observe relatively consistent standardised bias across outlier values. However, a notable increase in bias and standard deviation occurs when $x_{it,1}^*$ equals -2 or 2. This increase can be attributed to the fact that at this level, $x_{it,1}^*$ is not yet considered an outlier by the model and is factored into the beta estimate calculation. Once the outlier's absolute value exceeds 3, the observation is disregarded in the beta estimate calculation. Consequently, the bias and spread of the estimator decrease again, approaching the scenario where the median of $x_{it,1}^*$ approaches 0. Outliers in X_i are more likely to introduce

bias in the estimation of β_i, because they directly influence the slope of the regression line, potentially distorting the estimated relationship between X_i and y_{it}. The variance is increased by adding leverage.

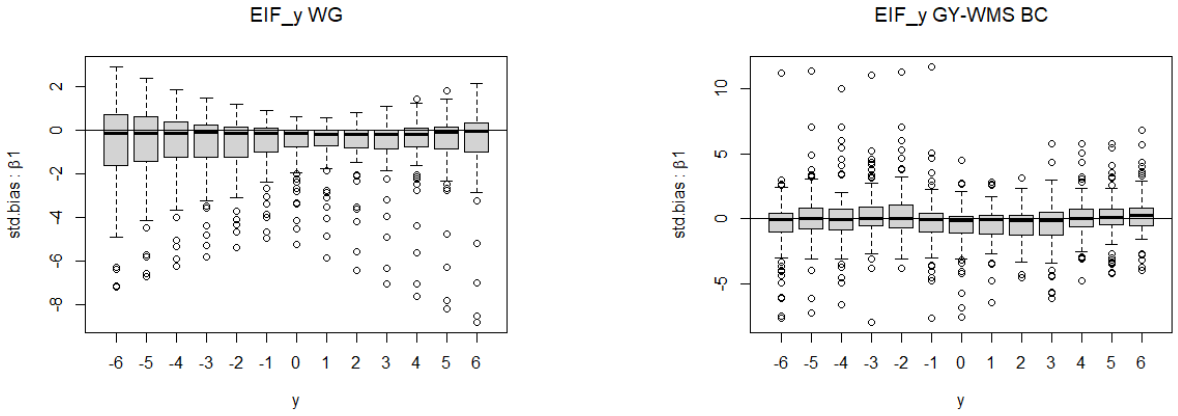

Figure 4.2: EIF plots for β_1 with outliers in y_{it}^*, for M = 100

Figure 4.2 displays the EIF for β_1 of the WG estimator and the GY-WMS BC estimator, respectively. In this scenario, we analyse a so-called vertical outlier, a single contaminated y_{it}^*. We expect that the contamination of a single response variable will have a less pronounced impact on the estimation of β_1 compared to the contamination of $x_{ij,1}^*$, due to the direct relationship between β_1 and X_1. For the WG estimator, we can see that the spread of the standardised bias becomes larger as the absolute value of the vertical outlier increases. We also see that, generally, there seem to be more outliers in the EIF values that are smaller than the true beta parameter. Interestingly, the median of the EIF function for the WG estimator stays close to 0, even for outliers up to (negative) 6. This can be explained by the fact that an outlier in y_{it}^* does not directly influence β_1 as an outlier in x_{it}^* does. Additionally, the WG estimator focuses on within-unit changes, where the bias impact of a one-off outlier is mitigated. However, since the WG estimator uses within-entity changes to estimate β, any significant deviation in these changes, such as those caused by an outlier, will increase the spread of the estimated coefficients.

The pattern for the GY-WMS BC estimator is difficult to identify. In general, we can deduct that, similar to the WG estimator, the β estimate seems to be relatively unbiased. Furthermore, the spread of the EIF values is relatively small all over the spectrum of y_{it}^*. This finding suggests that the estimator is indeed robust against leverage outliers.

4.5.4 Calculation of the EIF for the level of the t-test for $\hat{\beta}_j$

Next, we calculate the EIF to assess the sensitivity of the level of the t-test to small changes in the data. The t-test's level equals the probability of rejecting the null hypothesis when it is true. If the EIF indicates that the t-test level is highly sensitive to slight changes in the data, it may suggest that the test results are not reliable in the case of outliers. We compute the EIF for the t-test level for both the WG estimator and the GY-WMS BC estimator. Our aim is to mitigate the effect of cellwise outliers using the cellwise robust estimator, therefore we expect only the WG estimator to break down. As above, we calculate and store the estimated $\hat{\beta}_j$ coefficients of

the two estimators. This process is repeated for $M = 100$ simulations.

The formula of the level of the t-test for $\hat{\beta}_j$ is

$$t_{\beta_j} = \frac{\hat{\beta}_j - \beta_{j0}}{SE(\hat{\beta}_j)}, \tag{4.6}$$

where β_{j0} represents the beta parameter under the null hypothesis and $SE(\hat{\beta}_j)$ denotes the standard error of the beta estimate. However, standard deviation estimates for the cellwise robust method may not accurately reflect the true variability and uncertainty of the beta coefficients. To address this, we employ a bootstrapping approach by resampling $\hat{\beta^b}_j$ 25 times and computing the standard error of $\hat{\beta^1}_j, \ldots, \hat{\beta^B}_j$. This provides a more robust alternative for estimating standard errors.

One must note that detection followed by imputation of flagged cells may result in reduced variability in the dataset. This occurs because the imputed values are coordinate-wise medians of the columns, which inherently have less variation than the true values. Consequently, standard errors derived from the imputed data may be underestimated, potentially leading to a higher absolute value of the t-test. Moreover, outliers can impact the estimator in non-normal ways, violating the normality assumption for the distribution of the estimator.

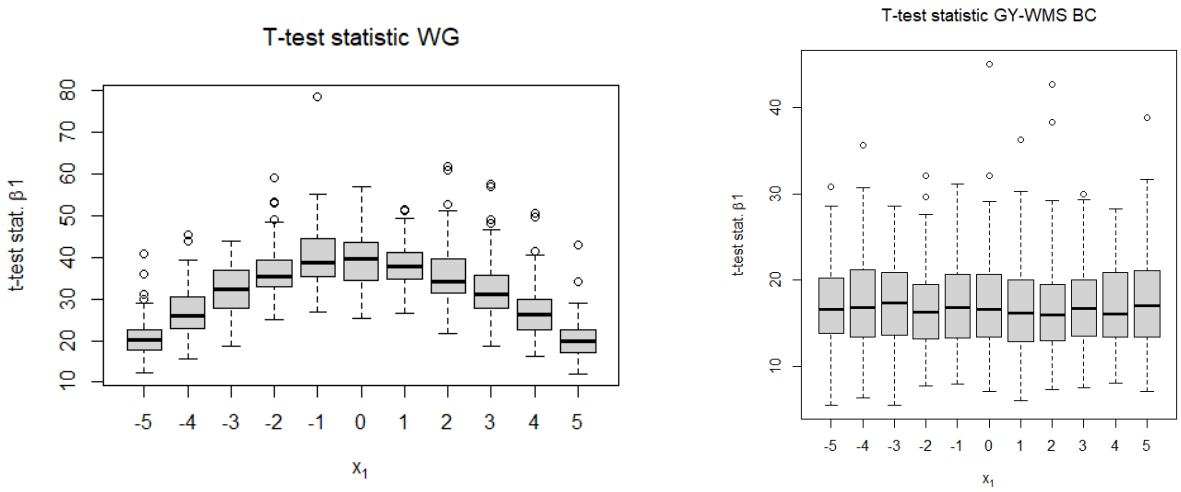

Figure 4.3: Plots of the t-test statistic for WG and GY-WMS BC estimators, respectively, for β_1 with outliers in $x^*_{it,1}$, for $M = 100$ in the case of uncorrelated regressors.

Figure 4.3 depicts the test statistics for the levels of the t-test for the WG estimator and the GY-WMS BC estimator, respectively. For the WG estimator, we see a similar parabolic shape as we saw for the EIF of the β_1 estimates themselves. We note that the null hypothesis of $\beta_1 = 0$ is rejected for all outliers as the t-test statistic exceeds 1.96. The value of the t-test statistic for the WG estimator, however, is halved as soon as we reach the outlier value of -5 or 5. We conclude that only one outlying cell already impacts the reliability of the t-test significantly for the WG estimator. For the GY-WMS BC estimator, however, the t-test statistic is significantly lower overall, with a fairly constant median of 16 across outlier values. The standard errors of the beta estimates are almost twice as large for the robust estimator, which explains why the

t-test statistic is approximately half the size for GY-WMS BC.

For the t-test, $EIF_{t_{\beta_j}}$ is computed by $N \times (\hat{t}^*_{\beta_j} - \hat{t}_{\beta_j})$, as derived from (2.2). We display the boxplots for the $EIF_{t_{\beta_j}}$ in Figure 4.4 for the outliers in $x^*_{it,j}$. To compute the value for $\hat{t}_{\beta_j}$ for every simulation ($M = 100$), we again compute standard errors through bootstrapping with $B = 25$. Because of the computation time, we reduce the range of values of the outliers from -5 to 5.

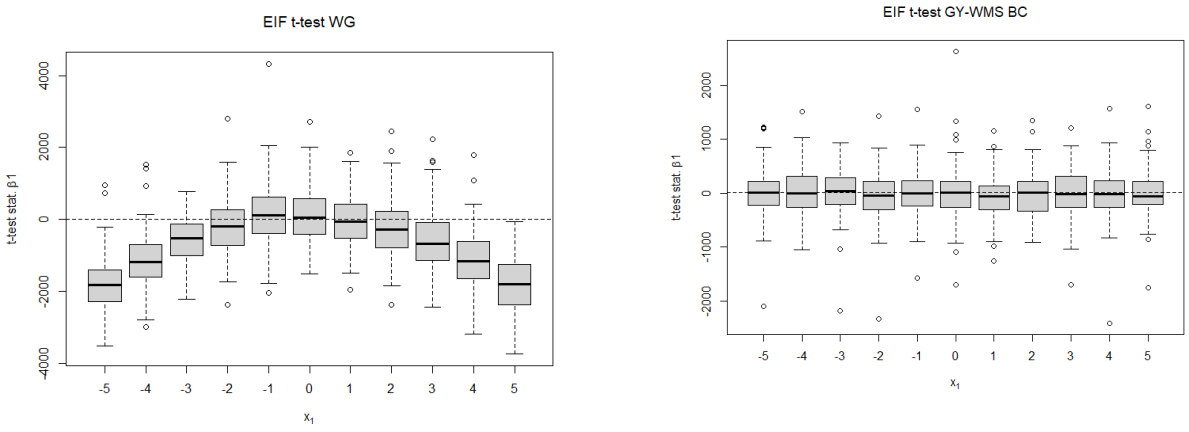

Figure 4.4: EIF plots for t_{β_1} with outliers in $x^*_{it,1}$, for $M = 100$

We observe that the plots for the EIF of the t-test look similar to those of the t-test statistic in shape. We note that the EIF for the WG estimator is strongly biased at the tails, as $x^*_{it,j}$ exceeds the absolute value of 2. For the GY-WMS BC estimator, the EIF remains relatively unbiased. For both estimators, however, we note that the spread of the EIF is extremely large. For the WG estimators, values range from -4000 and 4000, while for the GY-WMS BC estimator they are about half this size: -2000 to 2000. We explain this through the larger average variance of the WMS estimator. Therefore, we conclude that cellwise outliers significantly affect the reliability of the t-test statistic.

One drawback of this analysis is the small bootstrapping and simulation sample, which results in a larger variation of the empirical influence function values.

4.6 Simulation study estimates

We describe the outcomes of the simulation studies using two plots for uncorrelated and correlated regressors. Table 4.1 and Table 4.2 illustrate the beta estimates for the Monte Carlo simulation of the static panel data model with fixed effects, for uncorrelated and correlated regressors, respectively. Note that the tables include estimates for the WG estimator as in the commonly used `plm` package in R ("WG") and for the WG estimator as in the code by Bramati and Croux (2007) ("WG BC"). The aim here is to demonstrate that the efficiency of the code can also affect the simulation results. Next, we incorporate the casewise robust estimator by Bramati and Croux (2007) ("GY-WMS BC"), to explain that a casewise robust method is not sufficient to tackle cellwise outliers. Lastly, we include the proposed cellwise robust method, the GY-WMS BC estimator ("GY-WMS BC"). We distinguish between an uncontaminated scenario

and a scenario where 1% of the cells are contaminated. The standard deviations in the table are shown in parentheses.

As anticipated, in the ideal, uncontaminated scenario with independent regressors, the WG estimator excels, in both bias and variance. In our examined uncontaminated case with correlated regressors, this classical estimator also outperforms the alternatives. However, this advantage shifts when faced with cellwise outliers. The GY-WMS BC estimator stands out as the only one resilient to even 1% of cellwise outliers. While we do see an increased variance compared to the WG estimator in the uncontaminated scenario, it remains robust in the presence of 1% cellwise contamination. By the fact that the WMS BC estimator also breaks down in case of cellwise outliers, we conclude that to effectively tackle cellwise outliers, a casewise robust estimator is not sufficient.

4.6.1 Uncontaminated scenario

	Clean data				*1% cellwise contamination*			
	WG	WG BC	WMS BC	GY-WMS BC	WG	WG BC	WMS BC	GY-WMS BC
$\beta_1 = 2$	1.998	2.002	1.996	1.941	0.801	0.803	1.981	1.938
	(0.047)	(0.067)	(0.072)	(0.104)	(0.183)	(0.191)	(0.092)	(0.105)
$\beta_2 = 1.5$	1.500	1.497	1.491	1.445	0.610	0.604	1.392	1.444
	(0.051)	(0.075)	(0.079)	(0.098)	(0.150)	(0.159)	(0.290)	(0.100)
$\beta_3 = -0.5$	-0.502	-0.500	-0.497	-0.482	-0.196	-0.197	-0.198	-0.481
	(0.052)	(0.071)	(0.070)	(0.082)	(0.086)	(0.090)	(0.074)	(0.084)
$\beta_4 = 0.8$	0.795	0.793	0.789	0.763	0.323	0.319	0.352	0.762
	(0.053)	(0.073)	(0.078)	(0.091)	(0.105)	(0.108)	(0.131)	(0.090)
$\beta_5 = 1.2$	1.198	1.201	1.195	1.161	0.489	0.484	0.867	1.160
	(0.051)	(0.072)	(0.075)	(0.095)	(0.131)	(0.134)	(0.353)	(0.097)

Table 4.1: Beta estimates (means and standard deviations) for the Monte Carlo simulation, comparing clean data with 1% cellwise contamination for $N = 100$, $T = 5$, $M = 500$.

For the uncontaminated scenario, we can see that the beta estimates are all relatively unbiased. As expected, WG and WG BC perform best with regards to bias. This can be understood by considering that in this scenario, we simulate data based on the assumptions of the WG estimator. Minor variations can be attributed to our sample size restriction of 500 observations. WMS BC follows closely, with a deviation of less than 1% from the true beta. Finally, the bias of the GY-WMS BC remains minimal in the uncontaminated scenario, deviating by 5% or less from the true betas. WG performs best in terms of standard deviation for the uncontaminated scenario, with standard deviations ranging from 0.047 for $\hat{\beta}_1$ to 0.053 for $\hat{\beta}_4$. Notably, the standard deviations of WG BC are up to 50% higher than those of WG, which can be attributed to modifications in `plm` for code efficiency. WMS BC shows a slightly higher standard deviation than WG BC, a result of the trade-off between efficiency and robustness. Moreover, GY-WMS BC shows the largest standard deviation, nearly doubling that of WG for certain parameters. There is a clear robustness trade-off, where in the uncontaminated scenario, the WG estimators outperform. For the contaminated scenario, we observe that WG and WG BC break down entirely with only 1% cellwise contamination, producing estimates far from the true betas. WMS BC already shows a significant improvement over the classical method, yet the estimates still vary considerably from the true parameters. The only estimator that remains

relatively unbiased, with a maximum deviation of 5% from the true parameter, is the GY-WMS BC. The estimates produced by the GY-WMS estimator for the contaminated scenario closely resemble those from the uncontaminated scenario. This can be attributed to the fact that the outliers are generated from a normal distribution with $\mu = 10$ or $\mu = -10$, values substantial enough to be detected by the GY filter.

Regarding the standard deviation of the classical estimators, we again find it significantly higher compared to the uncontaminated scenario. Specifically, it doubles for $\hat{\beta}_2$ and even escalates from 0.047 to 0.183 for the WG estimator of $\hat{\beta}_1$. Interestingly, the standard deviations for WMS BC estimator skyrocket for $\hat{\beta}_2$ and $\hat{\beta}_5$, yet remain relatively low for $\hat{\beta}_1$ and $\hat{\beta}_3$ with coefficients of 0.092 and 0.074, respectively.

4.6.2 Correlated data estimates

	Clean data				*1% cellwise contamination*			
	WG	WG BC	WMS BC	GY-WMS BC	WG	WG BC	WMS BC	GY-WMS BC
$\beta_1 = 2$	2.004	2.001	1.996	1.984	0.905	0.907	2.099	1.984
	(0.058)	(0.083)	(0.087)	(0.114)	(0.220)	(0.228)	(0.267)	(0.109)
$\beta_2 = 1.5$	1.494	1.503	1.496	1.445	0.746	0.745	1.146	1.446
	(0.068)	(0.099)	(0.103)	(0.130)	(0.200)	(0.205)	(0.466)	(0.126)
$\beta_3 = -0.5$	-0.493	-0.499	-0.488	-0.401	0.198	0.198	-0.038	-0.391
	(0.067)	(0.092)	(0.097)	(0.113)	(0.105)	(0.111)	(0.094)	(0.113)
$\beta_4 = 0.8$	0.794	0.795	0.791	0.768	0.439	0.438	0.324	0.767
	(0.064)	(0.092)	(0.099)	(0.110)	(0.144)	(0.144)	(0.261)	(0.108)
$\beta_5 = 1.2$	1.203	1.204	1.199	1.179	0.563	0.567	1.183	1.181
	(0.060)	(0.083)	(0.087)	(0.110)	(0.165)	(0.173)	(0.403)	(0.110)

Table 4.2: Beta estimates (means and standard deviations) for the Monte Carlo simulation, comparing clean data with 1% cellwise contamination for $N = 100$, $T = 5$, $M = 500$.

Next, we evaluate the scenario where regressors are correlated using the Toeplitz matrix. Table 4.2 presents the results in a similar way as above.

Remarkably, WG BC performs slightly better compared to WG in terms of bias, though the difference between the two is less than 1%. Similar to uncorrelated scenario, WMS BC performs slightly worse on bias, followed by GY-WMS BC. GY-WMS BC deviates less than 4% from the true parameters for all parameters except $\hat{\beta}_4$, where the estimate is -0.401 in contrast to the true β_4 of -0.5. This discrepancy suggests that the method may incorrectly classify certain variables as outliers when they should instead be identified as correlated values.

We observe a marginally higher standard deviation for estimators in the correlated scenario compared to the uncorrelated scenario. Similar to the uncorrelated scenario, the WG BC estimator's standard deviation increases up to 50% compared to WG, likely for the same reasons as previously noted. Furthermore, the GY-WMS BC estimator's standard deviation doubles that of WG, which can be attributed partially to less efficient coding and the robustness-efficiency trade-off.

As in the scenario with uncorrelated regressors, WG and WG BC break down completely with just 1% of cellwise contamination. For all betas, their estimates differ significantly from the true betas. In terms of bias, WMS BC significantly outperforms the classical methods, yet it still falls short when compared to the GY-WMS BC method, which surpasses all others by

producing estimates very close to or even better than those in the uncontaminated scenario, except for β_3, which is estimated at -0.391 compared to the true value of -0.5. This can again be attributed to the fact that some cells are incorrectly recognised as outliers and imputed, or that some outliers are missed by either the GY filtering or the WMS estimator.

We again note that the classical estimators' standard deviations are significantly higher compared to the uncontaminated scenario and the contaminated scenario with uncorrelated regressors. We observe in particular that the standard deviation of WMS BC explodes, taking on values of 0.466 and 0.403 for the standard deviation of $\hat{\beta}_2$ and $\hat{\beta}_5$, respectively. This can be explained by the fact that the estimator is not used to dealing with either cellwise outliers or correlated variables. Finally, GY-WMS BC performs best in this scenario, with standard deviations ranging from 0.108 to 0.113, because of its ability to deal with both marginally large cellwise outliers and correlated cellwise outliers, thanks to the GY filter.

Chapter 5

Empirical study

To explore the effects of regional government fragmentation on the quality of social services such as health and education, Grossman et al. (2017) investigate a dataset manually recording the number of top-tier regional government entities across Sub-Saharan Africa, from 1960 (or the year of independence) to 2012. The primary variable is 'N. Regional Gov pc', which represents the annual count of top-tier regional governments in a country per 1 million inhabitants. This approach isolates the impact of government fragmentation on service delivery quality from other institutional and economic influences. The analysis reveals a large number of regional governments from as small as two in São Tomé to 112 in Uganda, with an average of 12.6. When adjusted for population size, the number of regional governments per 1 million people in Sub-Saharan Africa ranges from 0.06 in Nigeria to 361.7 in Seychelles, averaging at 9.5, and showing a pronounced right skew with 90% of country years falling under 10. (Grossman et al., 2017).

In the paper, the authors aim to mitigate endogeneity issues through two-way fixed effects models and the use of instrumental variables. For this empirical study, we use their fixed effects model B because it most closely aligns with our paper and we do not explicitly address endogeneity problems in this research.

Fixed effects in this study account for various aspects of a country's governmental structure, including administrative layers, overall state capacity, historical factors such as colonialism, and legal origins (Grossman et al., 2017). On the other hand, year fixed effects are necessary to address long-term improvements in public services and to capture global policy influences. Control variables are included to help address some time-varying factors such as politically motivated changes in government structure.

To evaluate the hypothesis regarding a positive (yet nonlinear) relation between government fragmentation and service quality, Grossman et al. (2017) employ, among others, the following model:

$$y_{it} = \alpha_i + \gamma_t + \beta_1 x_{it-1} + \delta \log(\text{govpc}_{it-5}) + \epsilon_{it}. \tag{5.1}$$

Here, the outcome variable y_{it} is the service provision index as a function of country and year fixed effects α_i and γ_t, respectively, along with a set of control variables x_{it}. To capture the non-linear effect of the main target variable, the log of the measure of regional government units per capita dgovpc_{it} is incorporated. Because changes in the number of regional governments may

not yield immediate effects, the main independent variable is lagged by five years. Throughout their analysis, Grossman et al. (2017) adjust for clustering of standard errors at the country level to accommodate for arbitrary serial correlation and heteroskedasticity.

5.1 Replication

In our research, we include FE model B as in Table 1 in Grossman et al. (2017). The model includes both categorical and numerical variables.

One limitation to our research is that we can not apply the WMS estimator for categorical variables, as it was not yet implemented by Bramati and Croux (2007). The GY filter can also only filter numerical variables. Therefore, we split the dataset into two subsets, one consisting of categorical variables and one with the numerical variables. In this research, we disregard potential outliers in the categorical variables. Subsequently, we apply the GY filter to this dataset and estimate the parameters using the classical WG estimator. We compare the dataset used by the authors to the dataset filtered by the GY filter

To compute the standard errors, the authors use clustered standard errors. In our paper, we bootstrap the standard errors. We do so by using block bootstrapping, bootstrapping complete entities from the dataset. We do this, because of this study's explicit focus on over-time variation within cases. One downside of this approach is that we do not account for the correlation of variables between entities.

5.1.1 Block bootstrapping entities

To bootstrap standard errors, we resample complete entities (countries) with replacement from the original dataset. We perform 500 bootstrap samples in Python and calculate the beta estimates for each one in R, both with and without the GY filter with default setting. We perform the same method for the sensitivity analysis described in the next section. The samples are bootstrapped in Python and the estimation of the fixed effects model is performed in R. Subsequently, the standard errors are calculated using the beta estimates computed from the bootstrap samples.

5.1.2 Sensitivity Analysis

To illustrate the potential effect of cellwise outliers. We add 0.5% of cellwise outliers to the numerical variables of the dataset. The approach is the same as in the Monte Carlo simulation described earlier. We only contaminate the numerical variables to distinguish the effect of applying the GY filter, which can only be applied for numerical variables, after contamination. Subsequently, we apply the GY filter to illustrate what effect cellwise outliers have if they are accounted for, versus the situation for which they are not.

5.2 Results

We compare the results for model B in Table 1 of Grossman et al. (2017). Specifically, we contrast the estimates of the fixed effects within estimator ("GPB FE"), as replicated from

(Grossman et al., 2017), to estimates of the within estimator after the GY filter ("GY FE"). The results are depicted in Table 5.1. Please note that our comparison includes a column for the original estimates with clustered standard errors, as replicated from Grossman et al. (2017), and a column with bootstrapped standard errors, respectively. We focus on the bootstrapped standard errors for an equal comparison between the scenario with and without GY filter.

We conclude that filtering out cellwise outliers with GY filter leads to similar estimates as the FE model as used in GPB, with higher standard errors. We therefore conclude that there are little cellwise outliers in the dataset. Therefore, we propose to conduct a sensitivity analysis with 0.5% cellwise outliers. The difference between parameter estimates in the contaminated scenario with and without GY filter is significant. Whereas unaccounted-for cellwise contamination results into an insignificant target variable, log(N local gov PC), results with the GY filter are fairly similar to the GPB FE estimates. As regards to the standard deviation, we see that the filter increases the variance for some parameters while diminishing it for others. This suggests that cellwise outliers are to be accounted for, even with a small contamination percentage. One must, however, take into account that contaminating only 0.5% of the variables, can lead to very different estimates and standard deviations depending on which cells are contaminated.

	Grossman et al. (2017) data			*Sensitivity analysis*	
	GPB FE (clus. se)	GPB FE (bt. se)	GY FE (bt. se)	GPB FE (bt. se)	GY FE (bt. se)
log(Population)	2.290^{***}	2.290^{***}	2.037^{**}	1.659^{**}	2.033^{**}
	(0.531)	(0.728)	(0.632)	(0.717)	(0.836)
Urbanization	-0.024^{**}	-0.024^{**}	$-0.026^{\cdot}$	-0.018^{*}	$-0.024^{\cdot}$
	(0.008)	(0.011)	(0.016)	(0.009)	(0.017)
log(GDP pc)	$0.271^{\cdot}$	$0.271^{\cdot}$	0.337^{*}	0.069^{***}	$0.313^{\cdot}$
	(0.148)	(0.181)	(0.164)	(0.108)	(0.198)
Intrastate conflict	0.011	0.011	0.042	0.027	0.044
	(0.053)	(0.060)	(0.051)	(0.063)	(0.059)
State elections	0.224	0.224	0.232	$0.303^{\cdot}$	0.235
	(0.142)	(0.191)	(0.160)	(0.211)	(0.200)
Polity 2	0.016	0.016	0.016	0.020	0.017
	(0.014)	(0.015)	(0.014)	(0.015)	(0.015)
log(Oil value pc)	-0.066^{*}	-0.066^{*}	$-0.061^{\cdot}$	-0.043^{*}	$-0.061^{\cdot}$
	(0.031)	(0.041)	(0.035)	(0.031)	(0.052)
Foreign aid pc	0.000	0.000	0.003^{*}	0.000	0.003^{*}
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
log(N local gov pc) 5-year lag	0.388^{**}	0.388^{**}	0.380^{*}	0.000	0.383^{*}
	(0.147)	(0.197)	(0.153)	(0.107)	(0.220)

Note: $^{\cdot}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table 5.1: Results for the regular dataset and the sensitivity analysis of the beta coefficients for the FE model with and without GY filter

Chapter 6

Conclusion

In our research, we develop a both cellwise and casewise robust methodology for static panel data with fixed effects. We take the adapted Gervini-Yohai outlier filter (Leung et al., 2016) for its recognised performance in outlier detection. Subsequently, we impute filtered outliers using coordinate-wise medians for computational feasibility, as suggested by Agostinelli et al. (2015a). The final phase employs the robust casewise WMS method by Bramati and Croux (2007), chosen for its asymptotic efficiency, and regression and affine equivariance properties.

To assess the models' resilience against outliers, we test the sensitivity of the estimator to small perturbations in the data, known as local contamination, through the empirical influence function. For single cellwise outliers in $\mathbf{x}_1$, the WG estimator shows a breakdown in both standardised bias and spread as the absolute value of the outlier increases. Conversely, the GY-WMS BC estimator exhibits neglectable bias across outlier levels, with bias and standard deviation spiking only when the outlier value is close to -2 or 2. This change underscores the model's initial failure to recognise such points as outliers, incorporating them into the beta estimate calculation. Once outliers exceed an absolute value of 3, they are excluded from the calculation, leading to a reduction in both bias and spread, showing the estimator's robustness against outliers.

In our analysis of the impact of single vertical outliers on beta coefficient estimation, we find that outliers in the response variable affect the estimation less severely than outliers in regressors. With the WG estimator, the variation in standardised bias increases with the size of the outlier, yet the median bias stays close to zero, highlighting the estimator's ability to mitigate one-off outlier impacts through its focus on within-unit changes.

The GY-WMS BC estimator maintains a relatively unbiased estimate across outlier values, with consistently small variation in standardised bias. This behavior indicates its robustness against vertical outliers.

Our study of the t-test's sensitivity to data changes through the EIF reveals significant differences between the WG and GY-WMS BC estimators. The WG estimator shows a pronounced sensitivity to outliers, with the reliability of the t-test significantly compromised by even a single outlying cell. On the other hand, the GY-WMS BC estimator maintains a lower, more stable t-test statistic across outlier values. We observe that the robust method significantly reduces (the EIFs of) the t-test statistics, a result of its larger standard errors for beta estimates.

These observations suggest that the WG estimator, while beneficial for its focus on within-

unit changes, may not be suitable in scenarios where cellwise outliers are present. The GY-WMS BC estimator, with its robustness to leverage outliers, offers a more reliable alternative for such cases. However, the significant spread of the EIF for both estimators indicates that outliers can still substantially affect the reliability of statistical inferences, emphasising the need for careful outlier management and robust statistical methods.

Subsequently, we performed a simulation study with both correlated and independent regressors. As expected, the WG estimator performs exceptionally well in terms of both bias and variance in a perfect, uncontaminated environment with independent variables. In our examined uncontaminated case with correlated regressors, this classical estimator also outperforms the alternatives. However, this advantage shifts when faced with cellwise outliers. The GY-WMS BC estimator stands out as the only one resilient to even 1% of cellwise outliers. While we do see an increased variance compared to the WG estimator in the uncontaminated scenario, it remains robust in the presence of 1% cellwise contamination. By the fact that the WMS BC estimator also breaks down in case of cellwise outliers, we conclude that to effectively mitigate cellwise outliers, a casewise robust estimator is not sufficient.

For the empirical study, our analysis reveals that excluding outliers via the GY filter yields estimates are similar to those obtained using the FE model in Grossman et al. (2017). For the sensitivity analysis, however, the significant key variable, N. Regional Gov pc in Grossman et al. (2017), is insignificant in the case of unaccounted-for cellwise outliers. The sensitivity analysis shows that already for a cellwise contamination of 0.5%, the contamination significantly impacts the parameter estimates. This underscores the potential issues that arise when one does not account for cellwise outliers.

6.1 Limitations and suggestions for further research

While our study introduces a promising cellwise robust methodology for handling outliers in static fixed effects panel data, there are limitations. The core of our methodology, the Gervini-Yohai (GY) filter, employs the standard normal distribution for outlier detection, a choice that aligns with our simulations derived from normal distributions. However, this assumption may not hold in real-life scenarios where data distributions deviate significantly from normality. Additionally, imputation of the coordinate-wise median might impose bias and lead to less variability in the dataset because it uses medians of the columns to impute the cells. Furthermore, our approach to computing the EIF by substituting an existing observation with an outlier could introduce additional bias, a concern not present when an outlier is added to the dataset.

To ensure computational feasibility, we constrained our study in several ways: limiting sample and bootstrap sizes, not varying the number of individuals or time periods, and solely applying the GY filter on the empirical data, without its integration into the full model. Our investigation did not extend to exploring temporal or spatial relationships between variables, except for correlated regressors using the Toeplitz matrix. We did not tackle the prevalent issue of endogeneity in panel data. Also, we did not account for correlation between entities in our bootstrapping approach.

Despite the slower computational performance of the cellwise robust model, the standalone application of the GY filter is practical and offers a valuable preliminary robustness check. This

aspect underscores the method's utility even in its simplest form.

Our research, while comprehensive, does not fully capture the diversity of real-world data, particularly in terms of dynamic panel data structures or non-linear effects. The effectiveness of the GY filter and subsequent imputation strategy, along with the reliance on coordinate-wise medians for imputation, may vary across different data structures and outlier patterns, potentially limiting the generalisability of our findings. Additionally, the GY filter does not account for temporal correlations or inter-entity correlations, commonly found in panel data. Also, the computational and interpretative complexity introduced by the casewise robust WMS method poses further challenges.

Future studies could also explore the influence functions for cellwise robust estimators analytically and examine their breakdown points. Additionally, exploring bootstrap techniques tailored for generating standard deviation estimates from samples with cellwise outliers presents a promising avenue for research. Finally, one could investigate an empirical dataset containing cellwise outliers, and apply cellwise robust methods against that.

References

Agostinelli, C., Leung, A., Yohai, V. J. & Zamar, R. H. (2015a). Rejoinder on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, *24*, 484–488.

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

Agostinelli, C. & Yohai, V. J. (2016). Composite robust estimators for linear mixed models. *Journal of the American Statistical Association*, *111*(516), 1764–1774.

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

Aquaro, M. & Čížek, P. (2013). One-step robust estimation of fixed-effects panel data models. *Computational Statistics & Data Analysis*, *57*(1), 536–548.

Bakar, N. M. A. & Midi, H. (2015). Robust centering in the fixed effect panel data model. *Pakistan Journal of Statistics*, *31*(1).

Baltagi, B. H. & Baltagi, B. H. (2008). *Econometric analysis of panel data* (Vol. 4). Springer.

Beyaztas, B. H. & Bandyopadhyay, S. (2022). Data driven robust estimation methods for fixed effects panel data models. *Journal of Statistical Computation and Simulation*, *92*(7), 1401-1425.

Bramati, M. C. & Croux, C. (2007). Robust estimators for the fixed effects panel data model. *The Econometrics Journal*, *10*(3), 521–540.

Čížek, P. (2013). Reweighted least trimmed squares: an alternative to one-step estimators. *Test*, *22*(3), 514–533.

Farcomeni, A. (2014). Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, *56*(1), 102–111. doi: 10.1080/00401706.2013.826148

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

Gnanadesikan, R. & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81–124.

Grossman, G., Pierskalla, J. H. & Boswell Dean, E. (2017). Government fragmentation and public goods provision. *The Journal of Politics*, *79*(3), 823–840.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation.* University of California, Berkeley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393.

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

Huber, P. J. & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). Wiley.

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

Leung, A., Zhang, H. & Zamar, R. (2016). Robust regression estimation and inference in the presence of cellwise and casewise contamination. *Computational Statistics & Data Analysis*, *99*, 1–11. doi: 10.1016/j.csda.2016.01.004

Lopuhaa, H. P. & Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 229–248.

Maronna, R. A. & Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, *89*(1-2), 197–214.

Muhammad, S., Shamshuddeen, S. & Baoku, I. G. (2021). Robust parameter estimation for random effect panel data model in the presence of heteroscedasticity and influential observations. *FUDMA Journal of Sciences*, *5*(3), 93–100.

Raymaekers, J. & Rousseeuw, P. J. (2019). Handling cellwise outliers by sparse regression and robust covariance. *arXiv preprint arXiv:1912.12446*.

Raymaekers, J. & Rousseeuw, P. J. (2023). Challenges of cellwise outliers. *arXiv preprint arXiv:2302.02156*.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871–880.

Rousseeuw, P. J. & Leroy, A. M. (2005). *Robust regression and outlier detection*. John wiley & sons.

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

Saraceno, G. & Agostinelli, C. (2021). Robust multivariate estimation based on statistical depth filters. *Test*, *30*(4), 935–959.

Štefelová, N., Alfons, A., Palarea-Albaladejo, J., Filzmoser, P. & Hron, K. (2021). Robust regression with compositional covariates including cellwise outliers. *Advances in Data Analysis and Classification*, *15*(4), 869–909.

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

Víšek, J. Á. (2015). Estimating the model with fixed and random effects by a robust method. *Methodology and Computing in Applied Probability*, *17*(4), 999–1014.

Walach, J., Filzmoser, P., Kouřil, Š., Friedeckỳ, D. & Adam, T. (2020). Cellwise outlier detection and biomarker identification in metabolomics based on pairwise log ratios. *Journal of Chemometrics*, *34*(1), e3182.