A deep dive into the robustness properties of the regression-based synthetic control estimator

Justin Dijk - 457469

December 14, 2021

Abstract

This article examines the robustness properties of the regression-based synthetic control estimator proposed in Abadie et al. (2010). We derive that a single observation that does not adhere to the linear factor model can lead to a breakdown of the estimator and thus lead to wrong estimates of the treatment effect. Additionally, we argue that due to the non-robustness of the method, we cannot compare the pre-intervention paths of the synthetic and treated unit to determine whether the counterfactual is valid. With these findings, we construct an alternative estimator and show through simulations and an empirical application that it is more robust to outliers in the data.

Supervisor: M. Zhelonkin
Second Assessor: W. Wang

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

Contents

1 Introduction

Since its introduction in Abadie and Gardeazabal (2003) and further statistical development in Abadie et al. (2010), the synthetic control method (SCM) has become a much-used technique for causal inference. For example, in recent studies, researchers use it to quantify the effects of decriminalising prostitution on the number of rape cases (Cunningham and Shah, 2018), corporate political connections on abnormal returns (Acemoglu et al., 2016), immigration policy on labour markets (Peri and Yasenov, 2019), and COVID-19 lockdowns on air pollution (Cole et al., 2020). Due to this broad applicability in the economic and social sciences, Athey and Imbens (2017) call it "*arguably the most important innovation in the policy evaluation literature in the last 15 years.*"

The motivation behind the method is that in a setting with aggregate units, such as a country or province, there often is not an appropriate control unit that can serve as a counterfactual[1]. For example, a practitioner may use a difference-in-differences estimator, which fails because there is no untreated unit that follows the pre-intervention outcome path of the treated individual. The synthetic control method tries to solve this issue by formalising the idea that a combination of control units may serve as a suitable counterfactual instead.

Abadie and Gardeazabal (2003) propose to construct such a "synthetic" unit as a convex combination of control units. To find these convex combination weights, they solve a quadratic program, which minimises the distance between features of the treated and the synthetic individual. These features, or characteristics, are functions that depend on the pre-intervention outcomes and covariates of an individual. Additionally, because there are multiple characteristics, each receives a non-negative weight that reflects its importance. It is possible to set these importance weights according to subjective preferences. However, Abadie et al. (2010) suggest two data-driven methods that minimise the mean squared prediction error. Namely, one based on a nested optimisation and one based on multiple cross-sectional ordinary least squares regressions.

Importantly, Abadie et al. (2010) prove that the synthetic unit provides a valid counterfactual under three conditions. First off, the outcome of interest needs to follow a linear factor model such that there are no interpolation biases. Secondly, the synthetic unit must have similar pre-intervention values as the treated unit for all variables. Lastly, the number of pre-intervention periods must be large relative to the size of the transitory shocks in the factor model.

Although there is plenty of work on the methodological and statistical foundations of the SCM, it is unknown what happens to the estimator when some data points do not come from a linear factor model. In other words, in scenarios when the dataset contains some contamination or outliers. In most real-world datasets, this is the norm rather than the exception due to human error, computer malfunctions, or because a model does not describe tail behaviour adequately (Maronna et al., 2019). Consequently, it is vital to know if the estimator still gives accurate estimates in these more realistic situations, especially when the results form a foundation for future policy. Additionally, when we understand how the estimator reacts to particular observations, we

[1] Throughout this article, we use the terms "intervention" and "treatment" interchangeably, as well as "unit" and "individual".

obtain a deeper understanding of the inner workings of the SCM instead of treating it as a black box.

We address this gap in the current literature for the regression-based synthetic control method. Even though nested optimisation is a more popular alternative to find the convex combination weights, we focus on the regression-based algorithm for two reasons. The first reason is that recent literature shows that the nested-optimisation has some computational difficulties (Becker and Klößner, 2018; Klößner et al., 2018; Malo et al., 2020). For example, a simple reordering of the control units can lead to a different solution (Kuosmanen et al., 2021). Secondly, the regression-based SCM is still the default option of the *Synth* package in Stata that implements the algorithm of Abadie et al. (2010) and thus still sees a lot of use with practitioners.

With this article, we make a few contributions to the current understanding of the synthetic control method. First off, we derive the influence function of the cross-sectional ordinary least squares estimator and show that it is unbounded. In particular, it illustrates that units with features that lie far away from those of the other units have a strong influence. Consequently, characteristics that are susceptible to contamination, such as the mean of the pre-intervention outcomes, can easily lead to wrong regression coefficients. Because the non-negative importance weights for the features are a weighted mean of the standardised regression coefficients, these are subsequently not robust.

Secondly, we also derive the influence function of the quadratic program that obtains the synthetic unit given the importance weights. Similarly to the regression estimator, it is not robust to outliers when there are characteristics that are non-robust functions of the pre-intervention outcomes and covariates. Additionally, any contamination that hurts the cross-sectional relationship between the features of the units may heavily influence the convex combination weights. We show this means that it is not enough to use characteristics such as the median to get robust convex weights when the variables are not stationary.

Thirdly we demonstrate that the synthetic control method can mask outliers in the data due to its non-robustness. As a consequence, we cannot compare the pre-intervention paths of the treated and the synthetic unit to determine whether the method produces a valid counterfactual.

Lastly, we use these findings to construct an alternative approach to find the synthetic unit. If necessary, this approach first makes all variables stationary and uses a robust function, like the median, to obtain the characteristics. These features, therefore, better retain the cross-sectional relationship between the units when there is contamination. It also uses a robust regression estimator as an alternative to ordinary least squares to find the regression coefficients, standardises them with an M-dispersion estimator, and uses a different function to turn the standardised coefficients into non-negative importance weights. We show through Monte Carlo simulations and the dataset of Abadie et al. (2010) that this alternative estimator produces more robust results than the standard regression-based synthetic control method.

We organise the remainder of this article as follows. In Section 2, we briefly discuss the synthetic control method and derive the robustness properties of the regression-based estimator. Additionally, in this section, we perform Monte Carlo simulations to give further insights into how and why different types of contamination affect the importance and convex combination weights. In Section 3, we present an alternative estimator to obtain the synthetic unit, after which we compare its performance with the classical estimator in more Monte Carlo simulations. Then, in Section 4, we apply both estimators in an empirical example to study the effects of Proposition 99 on the number of smokers in California. Lastly, in Section 5, we give some concluding remarks.

2 Synthetic Control Method

2.1 Preliminaries

To motivate the use of the synthetic control method, Abadie et al. (2010) use the following setting. Suppose that we observe $i = 1, \dots, N+1$ units at time $t = 1, \dots, T$ and that we can split the observations into two intervals. Namely, a pre-intervention interval for the first $1 < T_0 < T$ observations, in which none of the units receives treatment and a post-intervention interval for the remainder of the observations. Without loss of generality, we assume that only the first unit undergoes treatment and no cross-over effects occur. That is, the intervention on the first unit does not affect the outcomes of the control units.

Suppose further, that an underlying process generates the observable outcomes y_{it} according to the linear factor model

$$y_{it}^U = \delta_t + \boldsymbol{\theta_t}^T \boldsymbol{z_i} + \boldsymbol{\mu_i}^T \boldsymbol{\lambda_t} + \epsilon_{it}, \tag{1}$$

$$y_{it} = y_{it}^U + \alpha_t D_{it}. \tag{2}$$

In the first equation, δ_t is an unobservable common factor, $\boldsymbol{z_i}$ is a $(K \times 1)$ vector of observable covariates, $\boldsymbol{\theta_t}$ is a $(K \times 1)$ vector of unknown coefficients, $\boldsymbol{\lambda_t}$ is an $(F \times 1)$ vector of unobservable common factors, $\boldsymbol{\mu_i}$ is an $(F \times 1)$ vector of unknown factor loadings, and ϵ_{it} is an idiosyncratic shock with a zero mean. In the second equation, α_t is the treatment effect at time t, and D_{it} is a dummy variable equal to one when $i = 1$ and $t > T_0$.

Lastly, assume that there is an $(N \times 1)$ vector of convex combination weights $\boldsymbol{w} = (w_2, \dots, w_{N+1})$ such that

$$y_{1t} = \sum_{i=2}^{N+1} w_i y_{it}, \quad t = 1, \dots, T_0 \quad \text{and} \tag{3}$$

$$\boldsymbol{z_1} = \sum_{i=2}^{N+1} w_i \boldsymbol{z_i}. \tag{4}$$

In this case, Abadie et al. (2010) show that when the length of the pre-intervention interval is large relative to the scale of the shocks,

$$\hat{\alpha}_t = y_{1t} - \sum_{i=2}^{N+1} w_i y_{it}, \quad t = T_0 + 1, \dots, T, \tag{5}$$

is an unbiased estimator of the treatment effect. In other words, the method estimates the causal effect with a synthetic unit that has the same pre-intervention outcome path and covariates as the treated unit.

There are two important aspects to these assumptions. First off, the estimator remains unbiased for more generic data generating processes (DGP) as long as the outcome remains a linear function of the explanatory variables. For example, the covariates may also vary over time according to an auto-regressive model. Secondly, it assumes that it can obtain appropriate synthetic control weights such that conditions (3) and (4) hold. If this is not the case, the SCM is not an appropriate method. However, we can check these conditions after the estimation of the synthetic weights.

To find the optimal combination weights Abadie et al. (2010) solve the following quadratic program

$$\begin{aligned} &\arg\min_{\boldsymbol{w}} \; (X_1 - X_0\boldsymbol{w})^T V (X_1 - X_0\boldsymbol{w}) \\ &\text{s.t.} \qquad \boldsymbol{w}^T \boldsymbol{i} = 1 \\ &\qquad\qquad w_i \geq 0, \quad i = 2, \ldots, N+1, \end{aligned} \tag{6}$$

where X_1 is a $(P \times 1)$ vector of standardised pre-intervention characteristics for the treated unit, X_0 is a $(P \times N)$ matrix with similar standardised characteristics for the control units, and $\boldsymbol{i}$ is an $(N \times 1)$ vector of ones. In practice, the features consist of linear combinations of the pre-intervention outcomes and summary statistics of the observable covariates but there are no specific rules on how to choose them. V, on the other hand, is a $(P \times P)$ diagonal and positive definite matrix with its trace equal to one. Intuitively, this matrix gives a weight to each feature that reflects how important it is.

Although it is possible to set the importance weights according to subjective preferences, Abadie et al. (2010) suggest two data-driven techniques that minimise the mean squared error (MSE). The idea of the regression-based method is to perform a cross-sectional ordinary least squares (OLS) regression of y_{it} $(i = 1, \cdots, N+1)$ on all P characteristics to obtain the standardised regression coefficients $\beta_{t,p}^{sd}$ $(p = 1, \cdots, P)$. In other words, it finds how important a characteristic is for the outcome of a period. Note that, the predictors for each period $(t = 1, \cdots, T_0)$, stays the same. Subsequently, the synthetic control method sets the importance weight of the p'th feature equal to

$$v_p = \frac{\sum_{t=1}^{T_0} {\beta_{t,p}^{sd}}^2}{\sum_{p=1}^{P} \sum_{t=1}^{T_0} {\beta_{t,p}^{sd}}^2}, \quad p = 1, \ldots, P, \tag{7}$$

which assures that the matrix is positive definite with its trace equal to one.

2.2 Theoretical robustness properties

2.2.1 Robustness of the convex combination weights

Before we discuss the robustness properties of the convex combination weights from the synthetic control method, we introduce some extra notation. Let $\boldsymbol{Y_{-1}}$ and $\boldsymbol{Z_{-1}}$ denote the matrices with the pre-intervention outcomes and covariates of all the control units. Additionally, let $h_p(\boldsymbol{y_i}, \boldsymbol{z_i})^{sd}$ denote the p'th standardised characteristic of unit i. Note that, this notation stresses the fact that a characteristic is a function of the pre-intervention outcomes or the covariates. Lastly, let

$$h_p(\boldsymbol{y_i}, \boldsymbol{z_i})^{sd}_{v_p} = \sqrt{v_p} h_p(\boldsymbol{y_i}, \boldsymbol{z_i})^{sd}, \tag{8}$$

be the weighted characteristic.

In order to study the local robustness properties of an estimator we use the influence function (IF). The IF approximates the behaviour of an estimator when there is an infinitesimal amount of contamination and is under certain conditions the asymptotic version of the sensitivity curve (Hampel, 1974). Consequently, it is desirable that the IF is bounded for an estimator.

Proposition 1 *Let $\boldsymbol{w}$ be the estimate that solves the quadratic program when the data follows a linear factor model with cumulative distribution F. Additionally, assume that the importance weights v_p $(p = 1, \cdots, P)$ are known and that there is an infinitesimal number of identical outliers with point-mass at $(\boldsymbol{Y^*}, \boldsymbol{Z^*})$. Then, the influence function at the distribution F with contamination $(\boldsymbol{Y^*}, \boldsymbol{Z^*})$ equals*

$$\text{IF}\Big((\boldsymbol{Y^*}, \boldsymbol{Z^*}); \boldsymbol{w}, F\Big) = \boldsymbol{A} \cdot \sum_{p=1}^{P} \Big(h_p(\boldsymbol{y_1^*}, \boldsymbol{z_1^*})^{sd}_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}^*}, \boldsymbol{Z_{-1}^*})^{sd^T}_{v_p} \boldsymbol{w}\Big) \boldsymbol{h_p}(\boldsymbol{Y_{-1}^*}, \boldsymbol{Z_{-1}^*})^{sd}_{v_p} + \boldsymbol{c}, \tag{9}$$

where $\boldsymbol{A}$ is a matrix and $\boldsymbol{c}$ is a vector that both do not depend on the point-mass contamination.

We show the proof of this proposition in the Technical Appendix.

The above proposition gives some important insights. First off, the influence function is unbounded when we use features that are not robust to outliers. To see this, suppose that we have a characteristic that equals the mean of the pre-intervention outcomes, and a very large outlier occurs in the outcome of one of the control units. Then, unless its importance weight is zero, this characteristic has a large influence due to its multiplication in (9). Large discrepancies between a weighted characteristic of the treated and the synthetic unit also have a large influence, which means that any contamination that distorts the cross-sectional relationship between the features can lead to large biases of the estimator. We show some more interesting consequences of Proposition 1 in the simulations in Section 2.3.

Another important robustness property is the finite-sample breakdown point (FBP). Donoho and Huber (1983) define the FBP as the largest proportion of original observations that we can replace by arbitrary outliers without making the estimator arbitrarily biased. Note that, in the context of the SCM, arbitrarily biased means that a weight equals zero or one because these are the boundaries of its parameter space. For sensible estimators, it holds that the FBP cannot exceed one half because otherwise, there are more "typical" than "atypical" points. Consequently, it is preferable that the estimator attains this upper bound.

We do not derive the breakdown point of the SCM theoretically because the quadratic program does not have an analytical solution and is not scale and regression equivariant. Nonetheless, we show in the simulations in Section 2.3 that it is zero due to the interaction between the non-negative importance weights and the synthetic weights.

2.2.2 Robustness of the non-negative characteristic weights

We now look at the robustness properties of the non-negative importance weights. As explained in Section 2.1, the SCM first performs, in each pre-intervention period, a regression of y_{it} on all the characteristics. This means that the method implicitly postulates the model

$$y_{it} = \boldsymbol{h(y_i, z_i)}^T \boldsymbol{\beta_t} + \upsilon_{it}, \quad i = 1, \dots, N+1, \tag{10}$$

where υ_{it} is an error term with mean zero and standard deviation σ_t, $\boldsymbol{h(y_i, z_i)}$ denotes the $(P \times 1)$ vector of characteristics of unit i, and $\boldsymbol{\beta_t}$ is a $(P \times 1)$ vector of unknown coefficients. We can write the corresponding solution of the above regression as the M-estimator

$$\underset{\boldsymbol{\beta_t}}{\arg\min} \sum_{i=1}^{N+1} \rho\Big(\frac{y_{it} - \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T \boldsymbol{\beta_t}}{\sigma_t}\Big), \tag{11}$$

where $\rho(x)$ is the function x^2 in case of OLS. Hence the normal equations of the M-estimator are

$$\sum_{i=1}^{N+1} \psi\Big(\frac{y_{it} - \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T \boldsymbol{\beta_t}}{\sigma_t}\Big) \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i}) = 0, \tag{12}$$

in which ψ is the derivative of the ρ function.

Because the SCM performs a standard regression, the next proposition follows immediately from Hampel (1974). However, we do still show a derivation of the IF in the Technical Appendix.

Proposition 2 *Given a point-mass contamination $(\boldsymbol{y_j^*}, \boldsymbol{z_j}^*)$ for unit j $(1 < j < N+1)$, let $\boldsymbol{\beta_t}$ be the estimate that solves the normal equations of the general M-estimator when the data follows a linear factor model. Then it follows that*

$$\begin{aligned} \mathrm{IF_j}\Big((\boldsymbol{y_j^*}, \boldsymbol{z_j^*}); \boldsymbol{\beta_t}, F_1, \cdots, F_{N+1}\Big) = \psi\big[y_{jt}^* - \boldsymbol{h}(\boldsymbol{y_j^*}, \boldsymbol{z_j^*})^T \boldsymbol{\beta_t}\big] \boldsymbol{h}(\boldsymbol{y_j^*}, \boldsymbol{z_j^*}) \cdot \\ \left(\sum_{i=1}^{N+1} \int \psi'\big[y_{it} - \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T \boldsymbol{\beta_t}\big] \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i}) \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T dF_i(\boldsymbol{y_i}, \boldsymbol{z_i}) \right)^{-1}, \end{aligned} \tag{13}$$

where ψ' is the derivative of the ψ function and F_i is the cumulative distribution function of the observable covariates and all pre-intervention outcomes for unit i.

Because the SCM uses an unbounded ρ function, its influence function is unbounded for large residuals (Hampel et al., 2011). Additionally, leverage points have an unbounded influence, which are values of the predictor that lie far away from those of other observations. Note that, in this case, the predictors of the regression are the characteristics of the units, which means that large leverage points with an unbounded influence can appear when the characteristics are also unbounded. To make matters worse, the SCM utilises the same explanatory variables for each pre-intervention period. Consequently, when the features are very susceptible to outliers, Proposition 2 shows that it causes a large bias in each regression.

The fact that OLS is not robust is not the only reason why the importance weights are non-robust. Per definition of the standardised regression coefficients, we can rewrite (7), the function that turns the regression coefficients into the feature weights, to

$$v_p = \frac{\sum_{t=1}^{T0}(\frac{\sigma_{x_p}}{\sigma_{y_t}}\beta_{t,p})^2}{\sum_{p=1}^{P}\sum_{t=1}^{T0}(\frac{\sigma_{x_p}}{\sigma_{y_t}}\beta_{t,p})^2}, \quad p = 1, \ldots, P, \tag{14}$$

where σ_{x_p} is the standard deviation of the p'th characteristic and σ_{y_t} is the standard deviation of the outcome at time t. The synthetic control method uses the sample standard deviation to estimate these values, which has an unbounded influence function and a zero breakdown point (Maronna et al., 2019). Consequently, it is easy to see in the above equation that the importance weights also inherit these properties even when the regression estimates are robust.

2.3 Simulations

In this section, we perform multiple Monte Carlo simulations to show the effects of outliers on the solution of the standard synthetic control method. The goal of these simulations is thus not to be as realistic as possible but to confirm the theoretical findings and give some further insight into how the estimates of the SCM react to different types of contamination.

We generate the data according to the linear factor model for 50 pre-intervention periods and one post-intervention period. The model consists of three parts. First off, it contains a time-invariant unobservable common factor with a different loading for each unit. Secondly, it has five time-varying observable covariates that each follow a normal distribution with a unit-specific mean. Lastly, it incorporates a normally distributed idiosyncratic shock with zero mean and a 0.1 standard deviation.

We also make sure that there is a convex weight such that the conditions for unbiasedness hold. Consequently, we set the means and loading of the treated unit equal to a convex combination of those of the ten control units. These weights are 0.8 and 0.2 for the first two control units, respectively and zero for the remainder of the controls. Additionally, we set the treatment effect equal to one and use the means of the outcome and observable covariates as our six characteristics. For replication purposes, we provide the full details of the simulations in the Simulation Details Appendix.

2.3.1 An outlier in the outcome

The left plot of Figure 2.3.1 shows the synthetic weights when the data contains one additive outlier, of various sizes, in the outcome of the treated unit. Because the weight for most control units stays close to zero, we only display them for control units one, two and five. The plot shows that when there is no outlier, the weights of the first two control units are unbiased. Namely, they are not significantly different from 0.8 and 0.2. It also shows that the estimator is susceptible to small outliers because, in this case, it is strongly biased. However, as the (absolute) size of the outlier increases, its influence seems to reduce such that the combination weights go back to their true values. Another interesting result is that the direction of the biases depends on whether the aberrant observation is positive or negative. When it is positive, the first control receives a weight close to one, while the other two units receive a weight close to zero. Negative values, on the other hand, lead to convex weights that are too large for the second and the fifth control unit. The reason why this happens is straightforward. Because when the outlier is positive, the mean of the pre-intervention outcomes of the treated individual is closer to that of the first control. The same holds for the second and fifth control unit but only when the outlier is negative.

The right plot of Figure 2.3.1 shows the corresponding characteristic weights of all six features, in which the first feature corresponds with the mean of the pre-intervention outcomes. It confirms the theoretical finding that the estimator of these weights is not robust to outliers. For example, when the outlier is small, the first feature receives an importance weight between 0.3 and 0.6, but when the outlier is large, its importance is only around 0.05. Proposition 2 explains why the importance weight becomes smaller for the first feature. Because the mean is biased for the treated unit, it is less informational for the outcome in most periods of the treated unit. At the same time, outliers cause the characteristic of this unit to lie far away from those of other units. In other words, it is a bad leverage point.

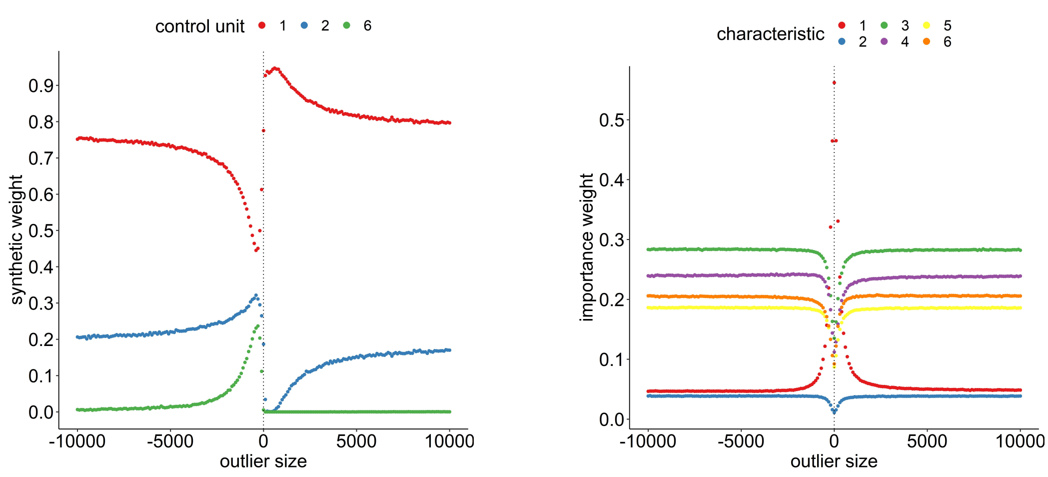

Figure 2.3.1: The estimated combination weights (left) and importance weights (right) of the synthetic control method when the data contains one additive outlier, of various sizes, in the outcome of the treated unit.

From these two plots, it is tempting to conclude that large outliers have no impact simply because the characteristic it biases receives almost no weight. However, in this case, that is not accurate. In Figure 2.3.2, we show the same two graphs but now when the outlier appears in the outcome of the first control unit instead of the treated unit. Unsurprisingly, the plot with the importance weights does not change much compared to Figure 2.3.1. However, the combination weights remain biased for all values of the contaminated data point. Therefore there must be a different reason why a large outlier in the data of the treated unit causes no bias in the optimal synthetic individual.

The reason is that the SCM standardises the characteristics before it solves the quadratic program. When a large outlier occurs in the outcome of the treated unit, it leads to an increase in the sample standard deviation of this characteristic. Hence if we standardise this feature for all individuals, the values for the control units will be close to zero. In this case, Proposition 1 shows that this feature plays no role in the estimation of the combination weights. Similar reasoning explains why outliers in the characteristic of a control unit remain influential even when they are large. In summary, although the standardisation of the feature is necessary for the quadratic program, it also has the unintentional benefit that it protects against large outliers in the treated unit when we use a non-robust characteristic.

Lastly, Figure 2.3.3 shows the estimates of the treatment effects when the outlier occurs in the treated unit or the first control unit. The estimated treatment effects in both plots are direct consequences of the convex combination weights. For example, when the outlier appears in the treated individual, the estimator becomes unbiased again as the absolute size of the outlier increases, but this does not happen when the data of the control unit has some contamination.

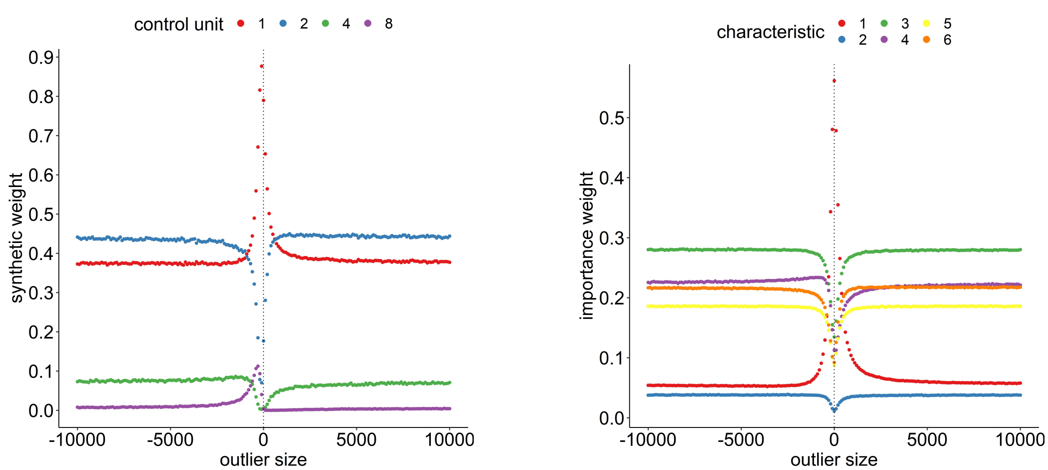

Figure 2.3.2: The estimated combination weights (left) and importance weights (right) of the synthetic control method when the data contains one additive outlier, of various sizes, in the outcome of the first control unit.

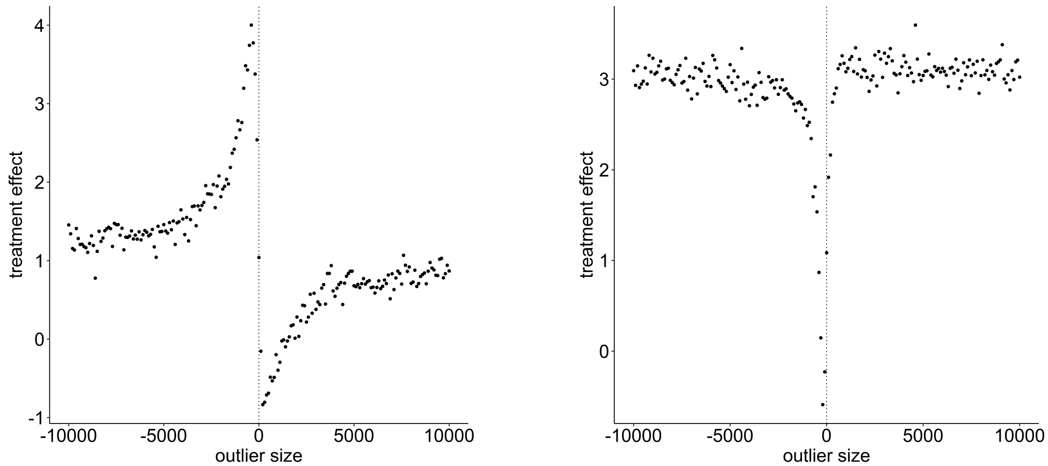

Figure 2.3.3: The estimated treatment effect when the data contains one additive outlier, of various sizes, in the outcome of the treated unit (left) or of the first control unit (right).

2.3.2 A good leverage point

We now look at what happens when the data contains a good leverage point at the individual level. That is, we generate an outlier in a covariate of a unit before we generate its outcome. This case is interesting for two reasons. First off, these types of "outliers" are often beneficial for classical methods, such as OLS, because they can reduce its variance. Secondly, in some cases, a researcher may try to remove aberrant points from each unit before using the synthetic control method. But because a good leverage point still adheres to the original linear factor model, this may not be possible.

The left plot of Figure 2.3.4 shows the effect of a good leverage point on the combination weights when we generate the outlier in the first covariate of the treated unit. Similarly to the previous figures, the combination weights are heavily biased even when the outliers are small. Additionally, in this case, the estimator pushes the combination weights to the boundaries of their parameter space when the size of the outlier is large, which is quite different from Figure 2.3.1, where the synthetic weights become unbiased again due to the standardisation of the characteristics.

The previous happens because of two reasons. In the first place, the mean of the outcome and the mean of the first covariate for the treated unit are both biased from a cross-sectional point of view. In essence, they constitute a structural break in the relationship between the features of the treated unit and the control units. Secondly, Figure 2.3.4 shows that these two features receive almost all and equal importance. Consequently, this cancels the unintentional benefit of the standardisation.

Figure 2.3.5 shows the corresponding estimated treatment effect. These estimates follow directly from the convex combination weights. For example, for large negative outliers, the treatment effect is around 8 because the synthetic unit is equal to the fifth control unit.

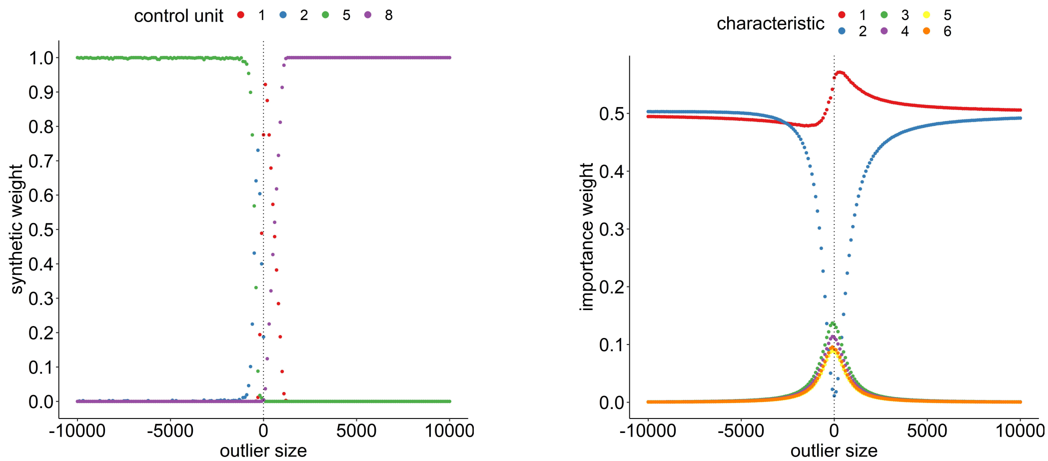

Figure 2.3.4: The estimated combination weights (left) and importance weights (right) of the synthetic control method when the data contains one outlier, of various sizes, in the first covariate of the treated unit before we generate the outcome.

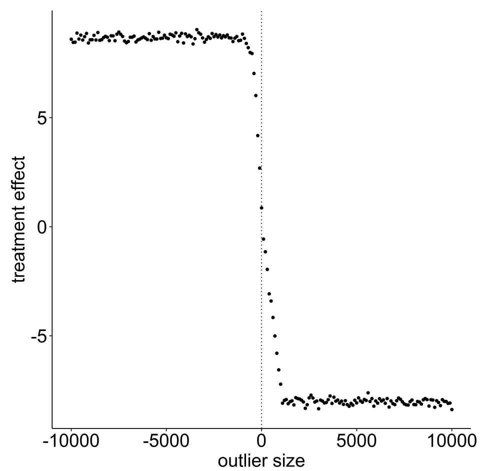

Figure 2.3.5: The estimated treatment effect of the synthetic control method when the data contains one outlier, of various sizes, in the first covariate of the treated unit before we generate the outcome.

There is another interesting type of good leverage point that can appear. Namely, a good leverage point that originates from an outlier in the common factor. Note that this type only affects the feature that involves the outcomes because we do not observe the common factor. We show the estimates of the combination and the importance weights in Figure 2.3.6. The plots show that the importance weights are significantly biased but that the combination weights remain unbiased. The reason for this is that each unit depends on the unobservable factor. Consequently, the characteristic of the outcome changes for each individual but in a way that it retains their true interrelationship. Because the synthetic unit remains unbiased, the treatment effect is also not significantly different from 1. Therefore we do not show it.

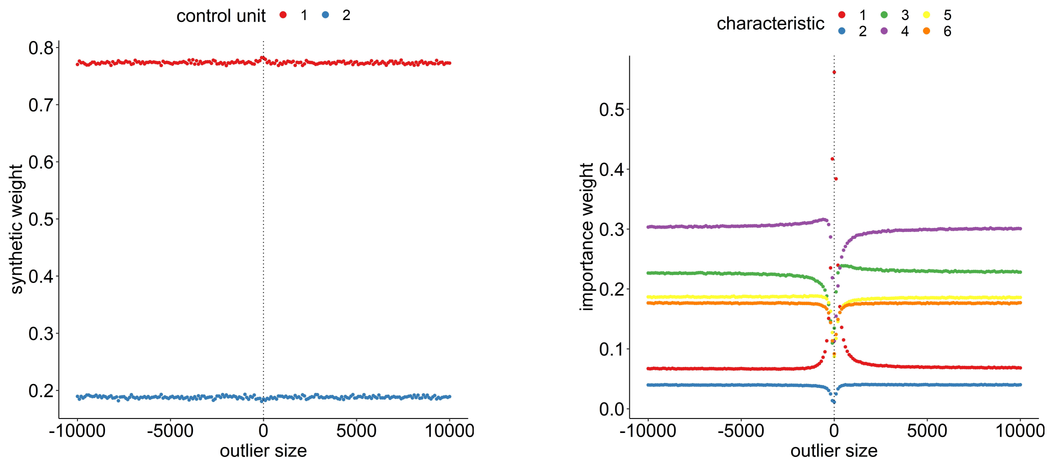

Figure 2.3.6: The estimated combination weights (left) and importance weights (right) of the synthetic control method when the data contains one outlier, of various sizes, in the common factor before we generate its outcome.

2.3.3 Pre-intervention paths

Although the previous simulations show that the SCM is susceptible to outliers and can lead to a bias of the treatment effect, we did not check whether the synthetic unit still looks like the treated unit. That is, whether they have similar pre-intervention outcomes and covariates. When these pre-intervention paths are vastly different, we deem the method inappropriate and need to use other causal inference methods. Consequently, this is an important step after estimating the convex weights. To make the visualisation more convenient, we reduce the number of observable covariates to one such that there are only two characteristics but keep the treatment effect equal to one. We provide the full details of this simulation in the Simulation Details Appendix.

Figure 2.3.7 displays the pre-intervention paths of the covariate and the outcome for the synthetic and the treated unit when there is no contamination. It shows that the pre-intervention values of these variables are almost identical for both individuals. Additionally, because there are no outliers, the treatment effect is not significantly different from one.

However, this changes when we add an additive outlier in the first period of size one to the outcome of the first control unit. Figure 2.3.8 shows that the synthetic individual no longer follows the pre-intervention paths of the treated individual. Consequently, we deem the method inappropriate because the conditions for unbiasedness do not hold and make no conclusion about the treatment effect.

Interestingly, the previous scenario is the best that can happen. Because, although the SCM is biased, we consider the method inappropriate since the synthetic unit does not look like the treated unit. But, this does not always happen. For example, we show the paths when we add an additive outlier of size one to the covariate of the treated unit in Figure 2.3.9. The pre-intervention values of both individuals look very similar such that it is likely that we do not discard the method. However, the estimated treatment effect is significantly biased because it is close to 1.5, which

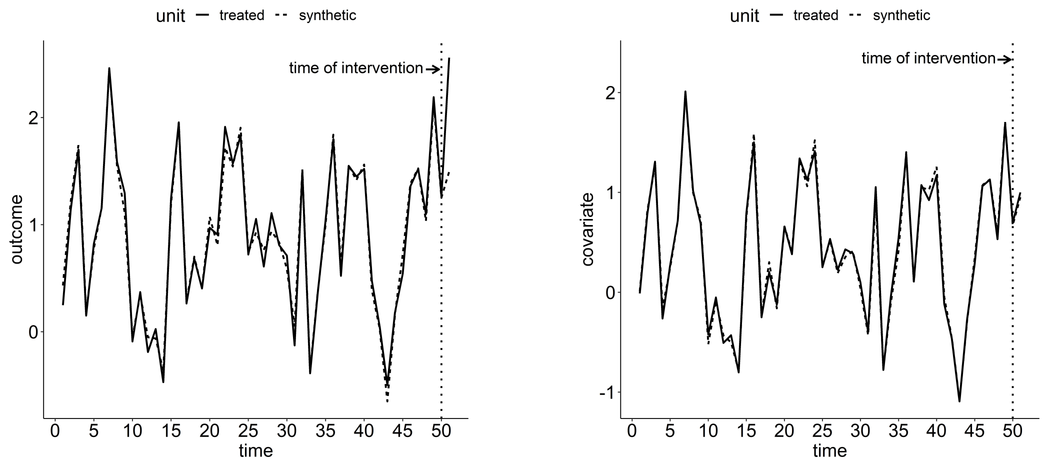

Figure 2.3.7: Pre-intervention outcome (left) and covariate (right) paths of the treated and the synthetic unit when there is no contamination.

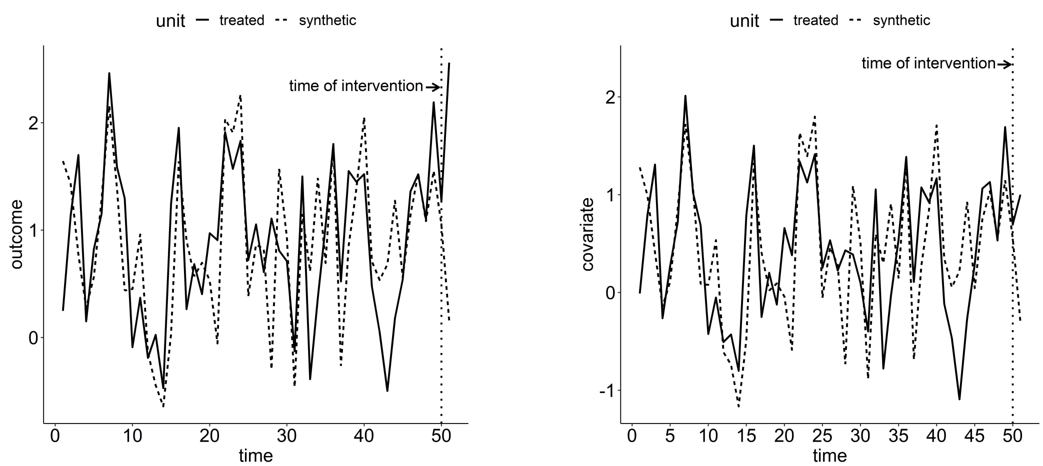

Figure 2.3.8: Pre-intervention outcome (left) and covariate (right) paths of the treated and the synthetic unit when there is an outlier in the outcome of the first control unit.

means we potentially draw an invalid conclusion.

The previous issue also happens to other non-robust estimation methods. For example, with OLS, we can not look at the residuals to detect outliers because its goal is to minimise the distance between the predicted and actual values, even when an observation is an outlier. In other words, it masks the outliers. Similarly, the SCM minimises the difference between the synthetic and the treated unit. Therefore we can not always rely on a visual inspection of the pre-intervention paths to see if the method is appropriate.

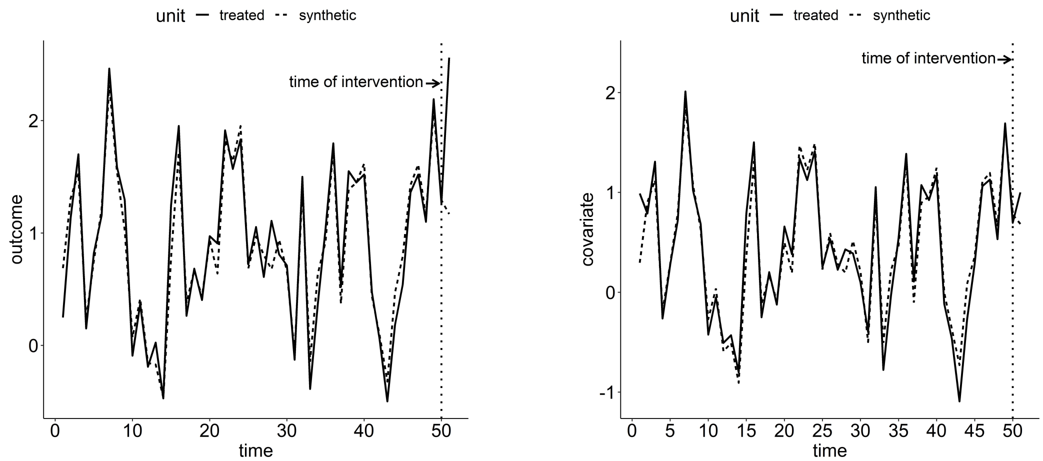

Figure 2.3.9: Pre-intervention outcome (left) and covariate (right) paths of the treated and the synthetic unit when there is an outlier in the covariate of the treated unit.

3 A Robust Estimator

In this section, we develop an alternative method to obtain the synthetic unit. We first discuss ways to estimate the non-negative importance weights robustly, after which we address the synthetic weights. Lastly, we show the strengths and weaknesses of the approach in Monte Carlo simulations.

3.1 Robust non-negative importance weights

The first step to find characteristic weights that are more robust to contamination is to use a different regression estimator. There are many alternatives to OLS that provide more robust regression estimates. However, we briefly discuss only two of these alternatives that we use in the simulation study[2].

The Mallows M-estimator is a subclass of general M-estimators which changes the normal equations of (12) to

$$\sum_{i=1}^{N+1} \psi_h\Big(y_{it} - \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T \boldsymbol{\beta_t}\Big)\eta\big(\boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})\big) = 0, \tag{15}$$

where $\psi_h(\cdot)$ is the derivative of the Huber loss function. Due to the Huber loss function, large residuals have less influence on the estimator. However, this does not solve the problem of leverage points. Therefore it also down-weights these data points through $\eta(\cdot)$, which is a function that gives smaller weights to characteristics that are far away from their mean. There are many possible ways to define such a function. But it is often a decreasing function of the Mahalanobis distance

$$D\big(\boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})\big), \boldsymbol{\mu}\big) = \sqrt{\big(\boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})\big) - \boldsymbol{\mu}\big)\boldsymbol{\Sigma}^{-1}\big(\boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})\big) - \boldsymbol{\mu}\big)}, \tag{16}$$

[2]For a complete discussion on the strengths and weaknesses of the alternative estimators in this section, see Hampel et al. (2011) and Maronna et al. (2019).

in which $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance matrix of the characteristics, respectively. Intuitively, we need to robustly estimate these with, for example, the minimum covariance determinant (MCD) (Hubert and Debruyne, 2010) or the minimum volume ellipsoid (Van Aelst and Rousseeuw, 2009). The major downside to the Mallows M-estimator is that its breakdown point cannot be larger than $1/P$, which becomes low when we utilise a lot of characteristics.

Another alternative to OLS is the MM-estimator derived by Yohai (1987), which consists of three steps.

1. Find an initial consistent estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{t},\mathbf{0}}$ with a high breakdown point.

2. Use the residuals from the first step to compute the regression scale $\hat{\sigma}_t$ robustly.

3. Find the optimal solution $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}$ of an M-estimator with the use of an iterative procedure that starts at $\hat{\boldsymbol{\beta}}_{\boldsymbol{t},\mathbf{0}}$ and keeps $\hat{\sigma}_t$ fixed.

When the M-estimator in the third step is re-descending, it inherits the robustness properties of the estimator of the initial two steps but not its efficiency. Consequently, we can utilise an estimator that finds robust and consistent estimates for $\hat{\boldsymbol{\beta}}_{\boldsymbol{t},\mathbf{0}}$ and $\hat{\sigma}_t$, after which we choose a re-descending M-estimator with high efficiency. For the first two steps, we use an S-estimator because it does not rely on a previously computed scale and has the maximum breakdown point. In the third step, we utilise an M-estimator with the Tukey loss function. The downside of the MM-estimator is that it uses an iterative algorithm to find a local or global minimum, which, on rare occasions, does not converge.

Secondly, as explained in Section 2.2.2, we need a robust estimator for the standard deviation to standardise the regression coefficients. One easy option is to utilise the median absolute deviation (MAD) estimator. However, the MAD has a very low efficiency when the variable of interest has a normal distribution. For this reason, we use an M-dispersion estimator instead. The idea behind the M-dispersion estimator is that it first uses an M-estimator to estimate the mean robustly and then applies an M-estimator of scale to find an efficient estimate of the standard deviation with a 50% breakdown point.

Lastly, we use a different function that turns the standardised coefficients into the importance weights. First off, we do not square the coefficients but take the absolute value instead. Not only does this ensure that the weight for each feature remains non-negative, but it also provides a better balance between the regression coefficients within a period. Secondly, we address another potential flaw in this function. The idea of standardising the regression coefficients $\beta_{t,p}$ $(p = 1, \cdots, P)$ is that we can compare their sizes *within* a period. However, it does not mean we can compare regression coefficients *between* periods, which Equation (7) implicitly does. Because, even when we standardise the coefficients for each period, it does not necessarily mean they are on the same scale across periods. Therefore the standardised coefficients of one regression may dominate the solution for the importance weights, which makes it is less likely that the eventual synthetic unit satisfies the conditions for unbiasedness.

Consequently, we change the function that turns the standardised regression coefficients into importance weights to

$$v_p = \frac{1}{T_0} \sum_{t=1}^{T0} k_{t,p}, \quad p = 1, \cdots, P, \tag{17}$$

where we obtain $k_{t,p}$ by solving the system of equations

$$\sum_{p=1}^{P} k_{t,p} = 1, \tag{18}$$

$$\frac{k_{t,1}}{k_{t,p}} = \frac{|\beta_{t,1}^{sd}|}{|\beta_{t,p}^{sd}|}, \quad p = 2, \cdots, P. \tag{19}$$

The above system of equations turns the absolute standardised regression coefficients into values that retain the same ratio but are comparable between different periods. Additionally, it makes sure that V is a diagonal and positive definite matrix with its trace equal to one.

3.2 Robust combination weights

It is much more difficult to obtain robust combination weights. The biggest reason for this is that we need to solve a quadratic program that is not regression and scale equivariant, and there are no robust alternatives for these types of problems in the current literature. Therefore we cannot simply replace the squared distance in the optimisation with a different loss function.

However, when we look back at the influence function of the quadratic program in Proposition 1, we see that two types of contamination heavily influence the estimator. Namely, contamination that leads to unbounded standardised characteristics of the control units and contamination that affects the cross-sectional relationship between the features.

We can solve the problem of unbounded standardised characteristics with the use of more robust location estimators like the median, a re-descending location M-estimator, or the median-of-means estimator (Lecué and Lerasle, 2020). However, even these estimators do not automatically lead to less biased combination weights in the presence of contamination.

To further explain this, we show in Figure 3.2.1 a simple example. In the example, there are three units of which their outcome depends linearly on time. When there are no outliers, the median of each unit lies in the second period, and the optimal combination weights are (0.5, 0.5). However, when we replace the second observation of the treated unit with the outlier, its median shifts to the third period and the optimal convex weights equals (0, 1). So although the median is more robust than the mean, it does not necessarily lead to more robust combination weights in the presence of non-stationarity.

The example shows that, even though the median is robust from the perspective of a single unit, it is not necessarily robust in terms of the cross-section. The reason for this is that, in the presence of contamination, the median can still shift too much. Hence it is necessary to make this shift as small as possible. For this reason, it is beneficial to make the outcome and the other covariates

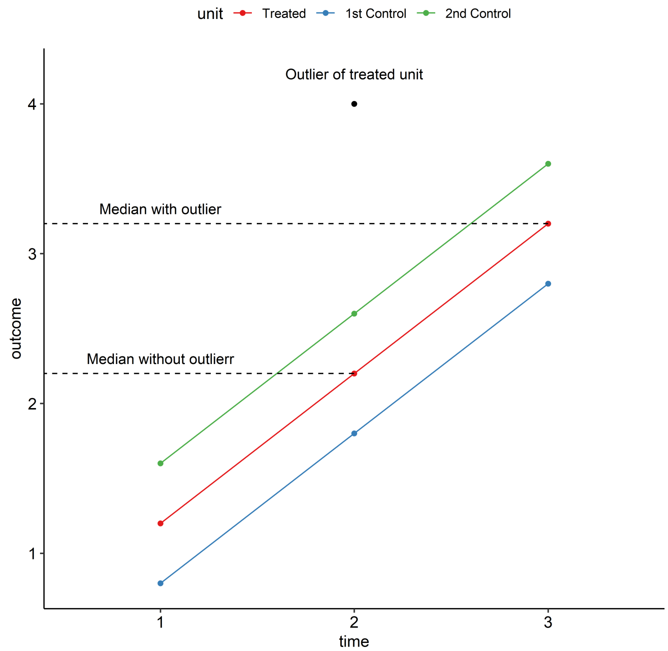

Figure 3.2.1: A simple example that shows that the median does not necessarily lead to robust combination weights.

stationary before we take the median. For instance, when we first-difference the dependent variable in the example, we obtain the original combination weights (0.5, 0.5) even when the outlier is present.

However, there are also downsides when we make the variables stationary by transforming them. As explained in Abadie et al. (2010), the variance of the idiosyncratic shocks must be small relative to the number of pre-intervention periods. When we, for example, take the first difference to make all variables stationary, we work with the factor model

$$\Delta y_{it}^{U} = \Delta\delta_t + \Delta\boldsymbol{\theta_t}^T \boldsymbol{z_i} + \Delta\boldsymbol{\mu_i}^T \boldsymbol{\lambda_t} + \Delta\epsilon_{it}, \tag{20}$$

where Δ is the first-difference operator. Consequently, not only do we lose one pre-intervention period to estimate with, but it can also lead to an increase in the variance of the idiosyncratic shock, which becomes $\Delta\epsilon_{it}$.

Another downside is that when we transform the data, the counterfactual may not necessarily be a valid counterfactual for the untransformed data. For instance, when we first-difference the variables, the synthetic individual tries to replicate the pre-intervention changes in the outcome. However, that does not mean it is able to replicate the pre-intervention levels. In some contexts, it might be necessary that the synthetic unit reproduces the pre-intervention paths of the original data. For example, when we use the estimator in economic settings in which the level of GDP for different countries needs to converge in the absence of treatment (Abadie, 2021).

3.3 Simulations

In this section, we perform more Monte Carlo simulations. We generate the outcome for 101 periods according to the linear factor model

$$y_{it}^{U} = \boldsymbol{z_{it}}^{T}\boldsymbol{i} + \mu_i\lambda_t + \epsilon_{it}, \quad i = 1, \ldots, 15, \quad t = 1, \ldots, 101, \tag{21}$$

$$y_{it} = y_{it}^{U} + 4D_{it}, \tag{22}$$

where D_{it} is a dummy variable equal to one when $t = 101$ and $i = 1$, $\boldsymbol{z_{it}}$ is a (7×1) vector of observable covariates, $\boldsymbol{i}$ is a (7×1) vector of ones, λ_t is a common unobservable factor, μ_i is a unit-specific factor loading, and ϵ_{it} is a standard normally distributed idiosyncratic shock. To construct each observable covariate, we generate a variable according to a specific auto-regressive (AR) process and multiply it with a unit-specific coefficient. While for the unobservable common factor, we use a white noise series with a standard normal distribution such that it resembles common but unexpected shocks.

For replication purposes, we show the full details of the DGP in the Simulation Details Appendix, but two aspects are important to mention. First off, each covariate is stationary except one, which contains a linear trend. Secondly, we also make sure there is a convex weight such that the conditions for unbiasedness hold. Hence we set the AR coefficients and the loading of the treated individual equal to a convex combination of the control units, namely, 0.5, 0.3, and 0.2, for the first three controls, respectively, and zero for the remaining controls.

Lastly, we consider three different methods to find the synthetic unit:

- The "SCM" method is the standard synthetic control method and does not transform the data before it finds the combination and importance weights. Additionally, it uses the medians of the pre-intervention outcomes and covariates as characteristics.

- The "Mallows" method does transform the data before finding the convex combination and non-negative importance weights. Namely, it first differences the outcome and the covariates such that there are only stationary variables. After this, it uses the medians of the variables as characteristics. Additionally, the method utilises the Mallows M-estimator to find the regression coefficients and standardises these with an M-dispersion estimator. Lastly, it turns the standardised coefficients into importance weights with the alternative summary function (17).

- The "MM" method has the same approach as the Mallows method but uses an MM-estimator to find the regression coefficients instead of the Mallows M-estimator.

We provide the full details of the Mallows and MM methods, such as the cutoff values for the Huber and Tukey loss functions, in the Simulation Details Appendix.

3.3.1 No contamination

Table 3.3.1 shows the estimates of the three different methods when there is no contamination. Because there is no contamination, it is not surprising that the three methods result in the same (nearly) unbiased convex combination weights. Consequently, the estimated and the true treatment effects are not significantly different at a five per cent level, regardless of which technique we use. There is a stark difference in the non-negative importance weights, however. Namely, the SCM assigns all and almost equal weight to the first two characteristics, while the other two estimators give all features some importance.

Table 3.3.1: Average estimates of the Monte Carlo simulations for the three different methods when there is no contamination in the data.

	Parameter											
Method	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_2	w_3	w_4	α
SCM	0.44	0.56	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.29	0.19	3.95 (0.07)
Mallows	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.12	0.49	0.29	0.19	4.11 (0.09)
MM	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	4.04 (0.09)

Note 1: The combination weights (w_2, w_3, w_4) correspond with the first three control units and (v_1, v_2) are the importance weights for the characteristics of the outcome and the non-stationary covariate, respectively.

Note 2: The standard error of the treatment effect α is in parentheses. For the other entries, the standard error is smaller than 0.01 so we do not display them.

3.3.2 Contamination of a single unit

We now contaminate the data of a single unit. Contrary to the simulations in Section 2.3, we do not add one outlier but generate outliers according to the Tukey-Huber contamination model (Huber, 1965, 1964). Consequently, each data point of a single unit has a fixed probability of being an outlier. We consider three different simulation scenarios:

Scenario 1: There can be additive outliers in the outcome.

Scenario 2: There can be additive outliers in the non-stationary observable covariate.

Scenario 3: There can be a good leverage point, caused by an additive outlier in the non-stationary observable covariate.

We use an outlier size of 10000 such that all the outliers lie on one side of the original median. Although this is a size that we can easily observe as an outlier in a data set, it shows the properties of the alternative estimation methods in a worst-case setting. Namely, when the characteristics can shift a lot, even though we use a robust statistic like the median.

It is also important to note that a higher contamination probability does not necessarily mean more contaminated inputs for the different estimators because we contaminate the original data. In particular, when we add outliers to the non-stationary covariate, it only affects one characteristic regardless of the contamination probability. Of course, this is different when we add outliers to the outcomes because the regression step uses every single outcome.

Table 3.3.2 shows the estimates of the SCM method when the contamination appears in the treated unit for the four contamination probabilities 0.01, 0.10, 0.25, and 0.50. The table confirms the previous findings of the simulations in Section 2.3, namely that the method is not robust to outliers. For example, even when the contamination probability is 1%, outliers in the outcome already sometimes lead to heavily biased convex combination weights. Additionally, note that in the second scenario at high contamination levels, the SCM puts almost all importance on the median of the outcomes. But, because only the non-stationary covariate is biased, it still provides unbiased synthetic weights.

We display the estimates of the MM and the Mallows methods in Tables 3.3.3 and 3.3.4, respectively. Because they provide very similar estimates, we do not discuss them separately. At the lowest level of contamination, the synthetic unit does not differ from when there are no outliers. As the contamination probability increases, the convex combination weights become more and more biased. However, they remain much closer to their true values, in Scenarios 1 and 3, than those from the SCM method. The feature weights are also much more robust. For example, even when the outlier probability is 50% in Scenario 2, the MM and Mallows estimators do not set the importance weight of the second feature, which corresponds with the non-stationary covariate, to

Table 3.3.2: Average estimates of the SCM method for the different simulation scenarios when only the data of the treated unit can contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
Scenario 1	0.01	0.46	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.28	0.18	2.96 (0.12)
	0.10	0.49	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.24	0.22	-9.97 (0.44)
	0.25	0.47	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.22	0.17	-42.37 (0.75)
	0.50	0.50	0.49	0.01	0.00	0.00	0.00	0.00	0.00	0.17	0.15	0.16	-68.28 (2.13)
Scenario 2	0.01	0.47	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.48	0.29	0.18	2.28 (0.13)
	0.10	0.81	0.16	0.01	0.01	0.01	0.00	0.00	0.01	0.47	0.28	0.17	-0.19 (0.14)
	0.25	0.93	0.02	0.01	0.01	0.01	0.00	0.00	0.01	0.49	0.29	0.19	2.03 (0.14)
	0.50	0.95	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.50	0.30	0.19	3.48 (0.11)
Scenario 3	0.01	0.47	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.25	0.12	3.23 (0.20)
	0.10	0.50	0.49	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.15	0.09	0.69 (0.56)
	0.25	0.52	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.15	0.11	-0.10 (0.71)
	0.50	0.53	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.07	0.05	-362.95 (16.80)

Note 1: The Prob. column displays the probability of contamination.
Note 2: The combination weights (w_2, w_3, w_4) correspond with the first three control units and (v_1, v_2) are the importance weights for the characteristics of the outcome and the non-stationary covariate, respectively.
Note 3: The standard error of the treatment effect α is in parentheses. For the other entries, the standard error is smaller than 0.01 so we do not display them.

Table 3.3.3: Average estimates of the Mallows method for the different simulation scenarios when only the data of the treated unit can contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
Scenario 1	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.12	0.49	0.29	0.19	3.95 (0.07)
	0.10	0.19	0.22	0.08	0.10	0.13	0.09	0.08	0.12	0.47	0.27	0.18	3.91 (0.07)
	0.25	0.19	0.24	0.08	0.09	0.13	0.09	0.08	0.11	0.42	0.22	0.15	3.56 (0.07)
	0.50	0.16	0.19	0.09	0.11	0.14	0.10	0.09	0.12	0.39	0.20	0.13	3.26 (0.11)
Scenario 2	0.01	0.18	0.21	0.08	0.10	0.14	0.10	0.08	0.12	0.49	0.29	0.19	3.85 (0.07)
	0.10	0.18	0.20	0.08	0.10	0.14	0.10	0.08	0.12	0.48	0.28	0.19	3.70 (0.07)
	0.25	0.16	0.17	0.09	0.11	0.15	0.11	0.09	0.13	0.45	0.25	0.16	3.61 (0.07)
	0.50	0.15	0.04	0.10	0.13	0.18	0.13	0.11	0.15	0.43	0.24	0.15	3.44 (0.08)
Scenario 3	0.01	0.19	0.21	0.08	0.10	0.14	0.10	0.08	0.11	0.48	0.29	0.19	3.83 (0.08)
	0.10	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	3.98 (0.07)
	0.25	0.20	0.20	0.08	0.10	0.13	0.10	0.08	0.11	0.46	0.27	0.17	3.93 (0.08)
	0.50	0.18	0.16	0.09	0.11	0.14	0.10	0.09	0.13	0.43	0.24	0.15	3.87 (0.08)

Note: For more details on the entries of the table see the table notes of Table 3.3.2.

Table 3.3.4: Average estimates of the MM method for the different simulation scenarios when only the data of the treated unit can contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
Scenario 1	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	3.98 (0.07)
	0.10	0.18	0.21	0.08	0.10	0.14	0.10	0.08	0.12	0.47	0.27	0.18	3.76 (0.07)
	0.25	0.16	0.20	0.09	0.10	0.14	0.10	0.09	0.12	0.43	0.23	0.15	3.64 (0.07)
	0.50	0.16	0.19	0.09	0.11	0.14	0.11	0.09	0.12	0.38	0.20	0.13	3.47 (0.08)
Scenario 2	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	4.06 (0.09)
	0.10	0.18	0.20	0.08	0.10	0.14	0.10	0.08	0.12	0.48	0.28	0.18	3.85 (0.07)
	0.25	0.17	0.17	0.09	0.11	0.15	0.11	0.09	0.12	0.45	0.25	0.16	3.47 (0.07)
	0.50	0.15	0.04	0.11	0.13	0.18	0.13	0.11	0.15	0.43	0.24	0.15	3.43 (0.08)
Scenario 3	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.92 (0.07)
	0.10	0.18	0.19	0.08	0.10	0.14	0.10	0.09	0.12	0.48	0.29	0.19	4.06 (0.08)
	0.25	0.18	0.18	0.08	0.10	0.14	0.10	0.09	0.12	0.47	0.27	0.17	3.89 (0.07)
	0.50	0.18	0.16	0.09	0.11	0.14	0.10	0.09	0.13	0.43	0.24	0.15	3.78 (0.08)

Note: For more details on the entries of the table see the table notes of Table 3.3.2.

zero. Of course, in this case, this leads to a synthetic unit that is more biased than the synthetic unit of the SCM. However, it does show that the feature weights are more robust to outliers.

Lastly, we address two more things about these results. First off, note that the estimated treatment effects of the MM and Mallows methods are much closer to 4, even when the SCM produces a synthetic unit that is less biased such as in Scenario 2. The causes for this are the DGP of the simulation and that the two alternative methods first-difference the data. Therefore, we do not draw any conclusions about the robustness of an estimator from the estimated treatment effect but from the combination weights. Secondly, we show in the Additional Results Appendix that the MM and Mallows methods remain more robust than the SCM even when the contamination appears in the data of the first control unit instead of the treated unit.

3.3.3 Contamination of multiple units

We repeat the simulations of the previous section but now add outliers to more than one unit such that there are more biased inputs. However, when we randomly add the same outlier to numerous individuals, this does not necessarily mean that the input data is worse for the different methods. For example, recall from Section 2.3.2 that when we increase the outcome for each unit due to an outlier in the common factor, the SCM still provides unbiased synthetic units.

Therefore we use a systematic approach to contaminate more than one unit. For a given simulation scenario, we take a combination of three, five or seven individuals. Each combination consists of the treated unit and an equal number of the first three and the remainder of the control units. We add an outlier of size -10000 to an individual that belongs to one of the first three controls and an outlier of size 10000 to the others. As a reminder, note that the first three controls have a non-zero convex combination weight in the data-generating process. Consequently, the characteristics of the treated individual look less like those of the controls with a non-zero combination weight in the DGP, while the opposite holds for the other 11 control units.

We display the estimates of the SCM method when the outliers occur in the data of multiple units for the three simulation scenarios in Tables 3.3.5, 3.3.6, and 3.3.7. When we compare these tables with the results of Section 3.3.2, we can observe two things. First off, outliers in more units lead to more biased feature weights. For example, in simulation Scenario 2 with a 10% contamination probability and seven affected units, the characteristic of the non-stationary covariate only receives a 0.05 importance weight. Secondly, in each scenario, the convex combination weights are also more biased when we add outliers to the data of multiple individuals.

A different pattern arises for the MM and Mallows methods. Because they again provide similar results, we put those of the Mallows method in the Additional Results Appendix and show the estimates of the MM method in Tables 3.3.8, 3.3.9, and 3.3.10. We see that the estimated non-negative importance weights remain very stable in each scenario, regardless of how many units are affected by outliers. However, this does not hold for the convex combination weights w_1, w_2, and w_3 in the first and third simulation scenarios. Namely, in these scenarios, the synthetic units become strongly biased at a 10% contamination probability, and these biases become worse when we add outliers to more individuals.

Table 3.3.5: Average estimates of the SCM method for simulation Scenario 1 when 3, 5, or 7 units contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.47	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.27	0.17	2.63 (0.11)
	0.10	0.48	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.23	0.21	-10.56 (0.35)
	0.25	0.47	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.16	-40.82 (0.58)
	0.50	0.50	0.48	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.28	0.00	-140.69 (0.11)
5 Units	0.01	0.46	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.24	0.18	2.31 (0.11)
	0.10	0.48	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.16	-10.18 (0.11)
	0.25	0.46	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	-36.50 (0.50)
	0.50	0.49	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	-101.14 (0.87)
7 Units	0.01	0.46	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.24	0.14	2.26 (0.11)
	0.10	0.48	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	-10.42 (0.32)
	0.25	0.46	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-34.30 (0.45)
	0.50	0.50	0.46	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	-81.45 (0.39)

Note: For more details on the entries of the table see the table notes of Table 3.3.2.

Table 3.3.6: Average estimates of the SCM method for simulation Scenario 2 when 3, 5, or 7 units contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.48	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.28	0.17	2.03 (0.11)
	0.10	0.87	0.09	0.01	0.01	0.01	0.00	0.00	0.01	0.39	0.26	0.16	0.20 (0.12)
	0.25	0.94	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.39	0.27	0.17	1.97 (0.11)
	0.50	0.95	0.00	0.02	0.01	0.01	0.00	0.00	0.01	0.40	0.28	0.17	3.23 (0.11)
5 Units	0.01	0.51	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.28	0.17	1.88 (0.11)
	0.10	0.90	0.06	0.01	0.01	0.01	0.00	0.00	0.01	0.40	0.24	0.15	0.71 (0.12)
	0.25	0.95	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.40	0.24	0.16	1.97 (0.12)
	0.50	0.95	0.00	0.02	0.01	0.01	0.00	0.00	0.01	0.40	0.25	0.16	2.85 (0.11)
7 Units	0.01	0.52	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.28	0.17	1.85 (0.11)
	0.10	0.91	0.05	0.01	0.01	0.01	0.00	0.00	0.01	0.40	0.25	0.12	1.10 (0.11)
	0.25	0.95	0.01	0.02	0.01	0.01	0.00	0.00	0.01	0.40	0.25	0.12	2.44 (0.11)
	0.50	0.95	0.00	0.02	0.01	0.01	0.00	0.00	0.01	0.40	0.25	0.12	2.91 (0.11)

Note: For more details on the entries of the table see the table notes of Table 3.3.2.

Table 3.3.7: Average estimates of the SCM method for simulation Scenario 3 when 3, 5, or 7 units contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.47	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.22	0.11	2.81 (0.16)
	0.10	0.50	0.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.04	-12.58 (0.53)
	0.25	0.51	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.03	-43.43 (0.79)
	0.50	0.33	0.31	0.18	0.01	0.06	0.06	0.03	0.04	0.00	0.01	0.00	-156.38 (2.37)
5 Units	0.01	0.47	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.17	0.10	2.46 (0.11)
	0.10	0.48	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	-7.21 (0.46)
	0.25	0.44	0.51	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	-32.16 (0.68)
	0.50	0.27	0.32	0.16	0.03	0.08	0.07	0.03	0.04	0.00	0.00	0.03	-134.59 (0.87)
7 Units	0.01	0.47	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.18	0.08	2.26 (0.11)
	0.10	0.47	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-6.73 (0.46)
	0.25	0.41	0.54	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	-30.87 (0.65)
	0.50	0.30	0.37	0.11	0.03	0.04	0.05	0.05	0.06	0.00	0.00	0.00	-117.09 (1.36)

Note: For more details on the entries of the table see the table notes of Table 3.3.2.

Table 3.3.8: Average estimates of the MM method for simulation Scenario 1 when 3, 5, or 7 units contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	3.87 (0.07)
	0.10	0.18	0.21	0.08	0.10	0.13	0.10	0.09	0.11	0.42	0.25	0.17	3.66 (0.07)
	0.25	0.17	0.22	0.08	0.10	0.13	0.10	0.09	0.12	0.25	0.16	0.10	3.60 (0.07)
	0.50	0.17	0.17	0.09	0.11	0.14	0.11	0.09	0.12	0.01	0.08	0.00	3.09 (0.10)
5 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.28	0.19	3.86 (0.07)
	0.10	0.18	0.20	0.08	0.10	0.14	0.10	0.08	0.12	0.39	0.20	0.16	3.63 (0.07)
	0.25	0.18	0.20	0.08	0.10	0.14	0.10	0.08	0.12	0.16	0.05	0.09	3.37 (0.08)
	0.50	0.18	0.16	0.08	0.11	0.15	0.11	0.09	0.12	0.00	0.00	0.00	3.04 (0.10)
7 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.28	0.19	3.83 (0.08)
	0.10	0.17	0.19	0.08	0.10	0.14	0.10	0.09	0.12	0.36	0.18	0.14	3.89 (0.07)
	0.25	0.18	0.19	0.08	0.10	0.14	0.10	0.08	0.12	0.17	0.05	0.05	3.31 (0.08)
	0.50	0.24	0.14	0.07	0.11	0.14	0.09	0.08	0.12	0.02	0.00	0.00	3.06 (0.08)

Note: For more details on the entries of the table see the table notes of Table 3.3.2.

Table 3.3.9: Average estimates of the MM method for simulation Scenario 2 when 3, 5, or 7 units contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.98 (0.07)
	0.10	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	4.15 (0.07)
	0.25	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	4.00 (0.07)
	0.50	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.97 (0.07)
5 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.09	0.08	0.11	0.49	0.29	0.19	3.91 (0.07)
	0.10	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.98 (0.07)
	0.25	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	4.00 (0.07)
	0.50	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.78 (0.07)
7 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.09	0.08	0.11	0.49	0.29	0.19	3.91 (0.07)
	0.10	0.19	0.21	0.08	0.10	0.13	0.09	0.08	0.11	0.49	0.29	0.19	3.87 (0.08)
	0.25	0.18	0.21	0.08	0.10	0.14	0.10	0.08	0.11	0.48	0.29	0.19	3.83 (0.07)
	0.50	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	4.08 (0.07)

Note: For more details on the entries of the table see the table notes of Table 3.3.2.

Table 3.3.10: Average estimates of the MM method for simulation Scenario 3 when 3, 5, or 7 units contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	3.92 (0.07)
	0.10	0.18	0.20	0.08	0.10	0.14	0.10	0.09	0.12	0.43	0.26	0.18	3.93 (0.07)
	0.25	0.18	0.19	0.08	0.10	0.13	0.10	0.09	0.12	0.16	0.10	0.10	3.44 (0.08)
	0.50	0.15	0.13	0.11	0.11	0.15	0.12	0.10	0.13	0.01	0.00	0.03	1.55 (0.11)
5 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.09	0.08	0.11	0.49	0.29	0.19	3.79 (0.08)
	0.10	0.15	0.18	0.09	0.11	0.15	0.11	0.09	0.13	0.39	0.22	0.16	3.78 (0.07)
	0.25	0.14	0.18	0.09	0.11	0.15	0.11	0.09	0.13	0.08	0.02	0.07	3.36 (0.08)
	0.50	0.09	0.11	0.11	0.14	0.16	0.13	0.12	0.14	0.01	0.00	0.01	2.90 (0.10)
7 Units	0.01	0.18	0.21	0.08	0.10	0.14	0.10	0.08	0.12	0.48	0.29	0.19	3.99 (0.08)
	0.10	0.13	0.18	0.09	0.11	0.15	0.11	0.09	0.13	0.37	0.20	0.12	3.87 (0.07)
	0.25	0.13	0.19	0.09	0.11	0.15	0.11	0.09	0.13	0.06	0.01	0.01	3.42 (0.08)
	0.50	0.11	0.09	0.09	0.16	0.15	0.13	0.13	0.13	0.01	0.00	0.00	2.93 (0.08)

Note: For more details on the entries of the table see the table notes of Table 3.3.2.

Although the MM method remains more robust than the standard synthetic control method, its difference in performances between simulation Scenarios 1, 2, and 3 are quite large. Table 3.3.11 shows the average standard deviations of the first-differenced but non-contaminated observable variables. It is immediately clear from this table that the variance of the outcome is much higher than the variances of the covariates. In particular, the standard deviation of the non-stationary variable is only 1.13, while it is 4.49 for the outcome. Consequently, when we add outliers, the median of the outcome shifts, on average, a lot more than the median of the covariates. Because the solution of the quadratic program is susceptible to changes in the characteristics, this explains the difference in performances of the MM method between the simulation scenarios.

Table 3.3.11: The average standard deviations of the first-differenced observable variables when there is no contamination.

	Variable							
	Δy	Δz_1	Δz_2	Δz_2	Δz_4	Δz_5	Δz_6	Δz_7
SD	4.49	1.13	1.28	1.18	1.98	1.33	1.10	1.43

Note: z_1 corresponds with the observable covariate that is non-stationary when we do not first-difference.

4 Empirical Application

In this section, we compare the behaviour of the standard regression-based synthetic control estimator with its robust alternative in a real-world example. For this, we repeat the analysis of Abadie et al. (2010) and then perform a sensitivity analysis.

4.1 Proposition 99

In 1988, California passed Proposition 99, a piece of legislation that had the intention to reduce the number of smokers. Due to this statute, from 1989 onwards, the state imposed a 25 cent excise tax on each pack of cigarettes and other tobacco-related products. Additionally, to further deter people from smoking, California used the tax revenue on healthcare programs and anti-tobacco advertisements to educate people about the dangers of smoking.

To estimate the causal effect of Proposition 99 on the number of cigarette sales, we use the same dataset as Abadie et al. (2010), which consists of annual state-level data between 1970 and 2000 for 39 states[3]. Consequently, we have 19 pre-intervention observations and 38 control units. These control units are states that did not introduce any anti-tobacco measures during the sample period. For each state, we observe five variables, namely per capita cigarette sales (in packs), per capita

[3] For a complete discussion on the sources of the dataset, we refer the reader to Abadie et al. (2010).

income (logged), per capita beer consumption (in gallons), the average retail price of cigarettes (in cents per pack), and the percentage of the population aged between 15 and 24.

First, we repeat the implementation of Abadie et al. (2010) to estimate the causal effect of Proposition 99 on the number of cigarette sales. Hence we utilise the standard regression-based synthetic control estimator with seven characteristics, namely the cigarette sales of 1975, 1980 and, 1985 and the average over the period 1980-1988 of each explanatory variable. Table 4.1.1 gives the importance weights and the values of the features for both California and its synthetic state. It shows that only the cigarette sales in 1980 and 1988 are important determinants of yearly cigarette sales. Therefore the counterfactual matches the sales of California in 1980 and 1988 closely. On the other hand, it fails to replicate some of the other features of the treated unit, such as the cigarette sales in 1975, the average beer consumption and the average retail price of cigarettes.

Additionally, the synthetic unit is sparse because only four states have a non-zero combination weight, namely Colorado (0.494), Connecticut (0.063), Nevada (0.146), and Utah (0.297). As we show in Figure 4.1.1, it manages to match the pre-intervention sales path of California, especially between 1980 and 1988. However, after the passage of Proposition 99, the sales of the two states diverge such that smoking was much less prevalent in California. Therefore these results indicate that Proposition 99 caused a large decrease in tobacco consumption.

Table 4.1.1: Importance weights and feature values of California and its synthetic state from the standard synthetic control estimator.

		State	
Characteristics	Importance weight	California	Synthetic
Cigarette sales 1975	0.04	127.10	124.10
Cigarette sales 1980	0.75	120.20	120.28
Cigarette sales 1988	0.21	90.10	90.37
Income (logged)	0.00	10.08	9.91
Beer consumption	0.00	24.28	23.05
Retail price of cigarettes	0.00	89.42	87.51
Percent aged 15-24	0.00	17.35	17.57

Note: The logged income, beer consumption, retail price, and percent aged 15-24 are averages of 1980-1988.

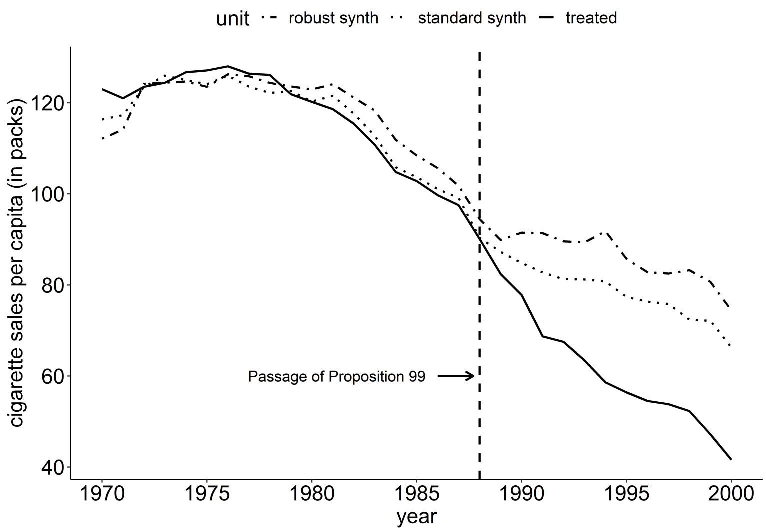

Figure 4.1.1: Cigarette sales of California and the synthetic states from both the standard and the robust estimator.

We repeat the previous analysis but use the robust estimator to obtain the synthetic unit. Consequently, we use the MM-estimator to find the regression coefficients, standardise these with an M-dispersion estimator and use the alternative summary function to turn the standardised coefficients into feature weights. Furthermore, it is insufficient to replace the mean of the variables with the median to obtain robust characteristics because the variables display strong trends in the pre-intervention interval. For example, most states saw a yearly decline in their cigarette sales after 1980 and an increasing personal income due to strong economic growth. Therefore, we take the natural log of each variable and first difference them, which makes them stationary and effectively turns them into growth rates. After this, we take the medians in the pre-intervention period of the retail price of cigarettes, the percentage of the population aged 15-25, income, and beer consumption as the characteristics of the explanatory variables. The other two features that we use are the medians of cigarette sales in 1975-1980 and 1981-1988.

Table 4.1.2 gives the importance weights and the values of the features for California and the synthetic unit from the robust estimator. When we compare the feature weights to those of the standard synthetic control method in Table 4.1.1, we notice that they are more balanced because each characteristic is now important instead of only the cigarette sales. Additionally, because the robust method uses different characteristics with different weights, the optimal synthetic unit also looks different. Namely, now Colorado (0.375), Idaho (0.332), Pennsylvania (0.171), and Wisconsin (0.122) are the states that receive non-zero combination weights.

Although the synthetic unit from the robust alternative manages to match the growth rates of the pre-intervention cigarette sales in California, this does not necessarily hold for the actual level of sales. Namely, when we compare the pre-intervention paths of the synthetic units in Figure 4.1.1, it is clear that the state from the standard estimator fits much better. The reason for this is twofold. First off, the robust estimator tries to match the change in sales of California and not

Table 4.1.2: Importance weights and feature values of California and its synthetic state from the robust estimator.

		State	
Characteristics	Importance weight	California	Synthetic
Cigarette sales 1975-1980	0.22	0.74	0.81
Cigarette sales 1981-1988	0.10	-2.28	-2.37
Income	0.11	6.70	7.25
Beer consumption	0.13	-2.28	-2.04
Retail price	0.20	-0.75	-0.62
Percent aged 15-24	0.23	-2.90	-2.88

Note: Each characteristic is the median growth rate (in percentages) over the period 1980-1988, except for cigarette sales, for which the table shows the specific time frame.

necessarily its level of sales. Secondly, recall from Section 2.3.3 that we need to be careful with judging the pre-intervention path of the standard estimator because it may mask outliers.

Lastly, we compare the post-intervention growth rates of California and its synthetic state from the robust estimator. Figure 4.1.2 shows that the growth rates of both states were the same in 1988. After the anti-smoking legislation, Californians reduced, on average, their cigarette consumption yearly by 6.55 per cent between 1989 and 2000. On the contrary, citizens from the synthetic state lowered their smoking much less, namely only 2.44 per cent each year. Consequently, from these results, we can conclude that Proposition 99 increased the rate at which tobacco consumption in California declined.

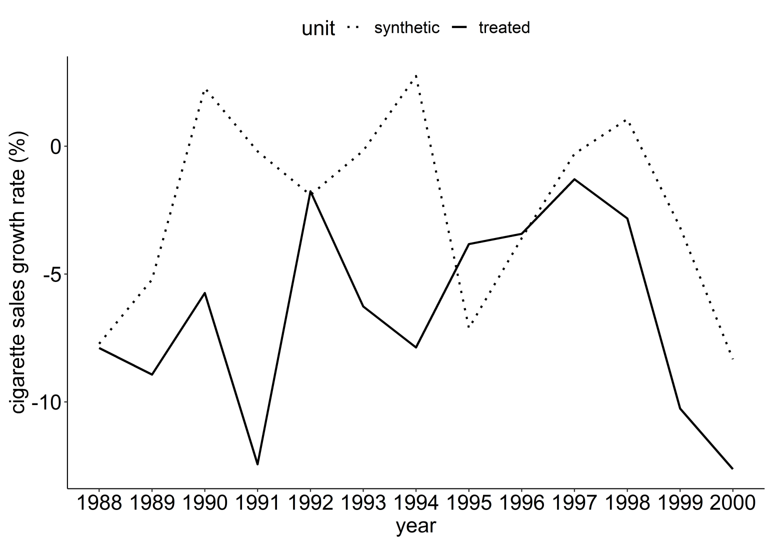

Figure 4.1.2: The post-intervention growth rates of cigarette sales for California and the synthetic counterfactual from the robust estimator.

4.2 Consequences of illicit tobacco trade

The number of cigarette sales in the dataset comes from state-wide tax revenue. Therefore it does not account for any illicit tobacco trade, such as smuggling packs from states with low excise taxes to states with higher excise taxes. Consequently, cigarette sales are only a proxy of the true smoking prevalence in a state. Abadie et al. (2010) address this issue by arguing that this effect has to be massive to fully explain the gap between the sales of California and its synthetic state in the post-treatment period.

There is, however, a possible flaw with this argument. As Table 4.2.1 indicates, even before Proposition 99, the price of cigarettes varied a lot between the states in our sample. For example, in 1980, citizens from Connecticut paid almost fifty per cent more for a pack of cigarettes than people from Kentucky. Therefore there is also uncertainty about the true smoking prevalence in the pre-treatment period, which we use to estimate the synthetic unit. Additionally, multiple studies show that, in certain states, more than twenty per cent of the cigarettes come from another state (Stehr, 2005; Lovenheim, 2008). Hence the effect of illicit tobacco trade is not tiny.

An interesting example of a state with low cigarette prices in 1980 is Colorado. During this period, its citizens paid 54.6 cents for a pack of cigarettes, while some of its neighbouring states, such as Nebraska (59.5 cents) and Oklahoma (62.9 cents), were more expensive. To account for the possible effect of cross-border smuggling, which means the true smoking prevalence in Colorado is lower than its cigarette sales, we contaminate the original data by subtracting a percentage of the sales in 1980.

Table 4.2.1: The price in 1980 and the average price between 1975 and 1988 for a pack of cigarettes per state.

State	Price 1980	Price 1975-1988	State	Price 1980	Price 1975-1988
Alabama	60.6	106.8	Nevada	63.1	119.2
Arkansas	61.5	110.2	New Hampshire	55.3	107.1
California	62.1	119.9	New Mexico	62.6	107.6
Colorado	54.6	104.5	North Carolina	47.3	90.8
Connecticut	67.0	127.3	North Dakota	59.6	112.5
Delaware	62.7	108.6	Ohio	58.7	103.1
Georgia	59.3	101.9	Oklahoma	62.9	108.4
Idaho	56.4	106.2	Pennsylvania	61.3	109.6
Illinois	60.0	115.5	Rhode Island	60.0	120.6
Indiana	53.7	97.5	South Carolina	52.3	95.0
Iowa	58.8	113.1	South Dakota	58.8	107.7
Kansas	58.3	107.6	Tennessee	60.3	103.1
Kentucky	46.3	88.3	Texas	63.7	117.4
Louisiana	60.0	108.4	Utah	57.2	112.7
Maine	59.0	119.2	Vermont	58.9	111.8
Minnesota	63.0	123.7	Virginia	48.5	95.2
Mississippi	59.7	105.6	West Virginia	64.3	105.7
Missouri	57.3	100.8	Wisconsin	61.2	118.8
Montana	56.7	103.7	Wyoming	55.3	99.6
Nebraska	59.5	110.5	-	-	-

Figure 4.2.1 shows the importance and synthetic weights from the standard synthetic estimator for various reductions in cigarette sales. When we look at the importance weights, we see that the importance of cigarette sales in 1980 gradually decreases as the smuggling effect becomes stronger. A different pattern arises for the combination weights, however. Because even at small decreases, the states that make up the synthetic unit change substantially. For example, when the smoking prevalence is 5 per cent lower than the reported sales, the synthetic state consists almost solely of Colorado. Also note, that at higher sales reductions, states such as Wyoming and Texas suddenly have non-zero weights. We see in Figure 4.2.2 that the non-robustness of the estimator also affects the post-intervention paths and thus the causal effect of Proposition 99. In general, the higher the smuggling effect in 1980 Colorado, the higher the increase of the causal effect.

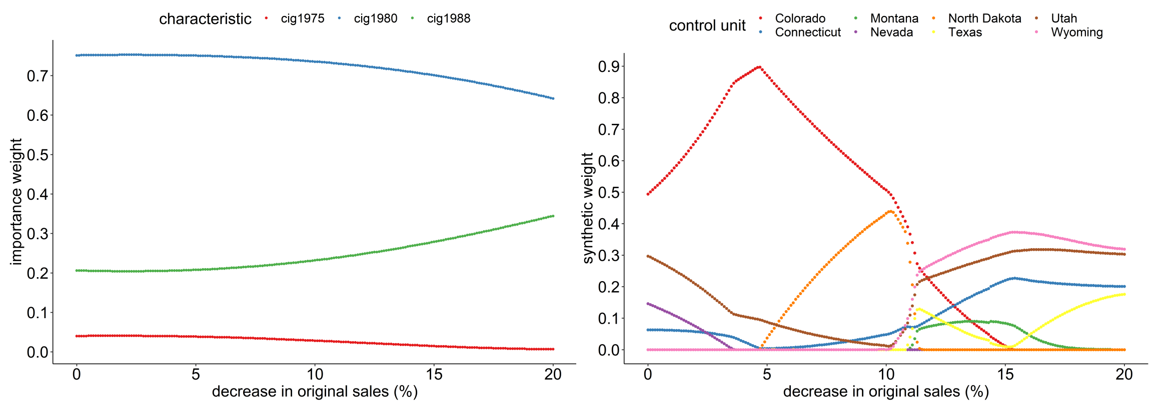

Figure 4.2.1: The non-zero feature (left) and combination weights (right) from the standard synthetic control estimator when we decrease the cigarette sales of Colorado in 1980 by between 0 and 20 per cent.

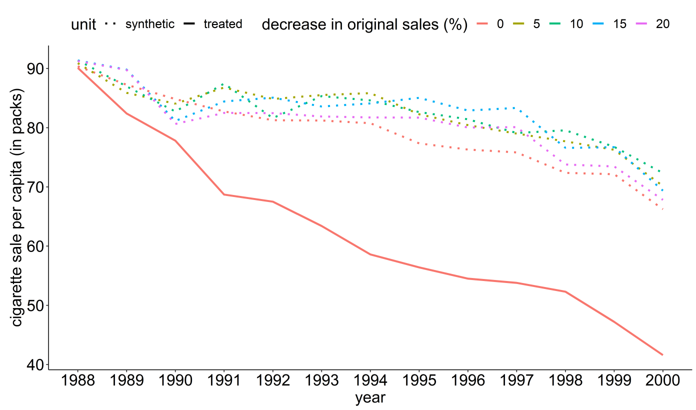

Figure 4.2.2: The post-intervention cigarette sales from the standard synthetic control estimator when we decrease the original sales of Colorado in 1980 by various percentages.

Next, we repeat the previous analysis but use the robust estimator. Figure 4.2.3 shows that the importance weights from this estimator are (almost) not affected by changes in sales, even when the smuggling effect is strong. Additionally, the combination weights are more stable than those of the standard synthetic control method. For example, when we reduce the sales by five per cent, the role of Colorado in the counterfactual now only changes by roughly 0.03, compared to 0.42 in Figure 4.2.1. Because the optimal synthetic unit is more robust to small changes in the dataset, the post-intervention paths also vary less. As we can see in Figure 4.2.4, the various outcome paths are visually indistinguishable from each other.

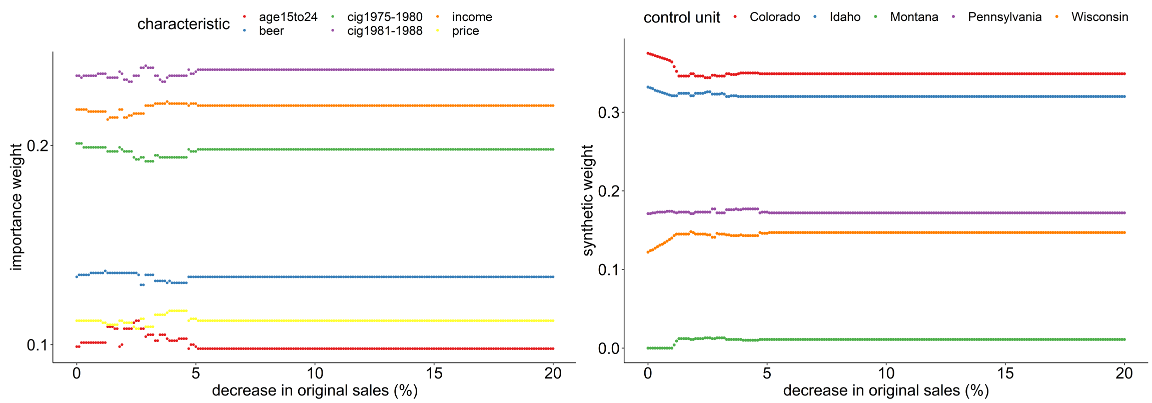

Figure 4.2.3: The non-zero feature (left) and combination weights (right) from the robust estimator when we decrease the cigarette sales of Colorado in 1980 by between 0 and 20 per cent.

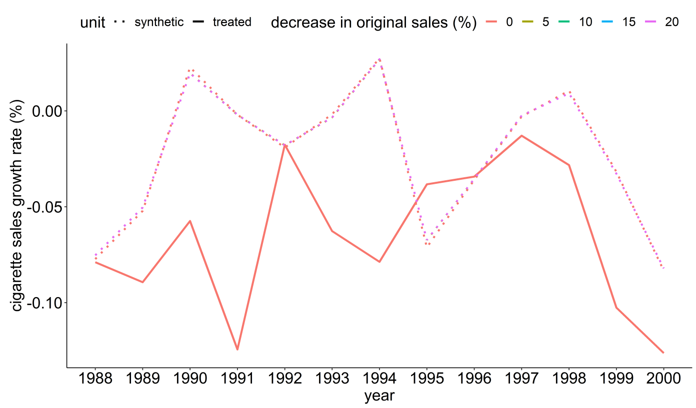

Figure 4.2.4: The post-intervention cigarette sale growth rates from the robust estimator when we decrease the original sales of Colorado in 1980 by various percentages.

So far, we assumed that the smuggling effect was only present in 1980. However, during the whole pre-intervention period, Colorado had a lower pack price than most of its neighbouring states. For this reason, we now reduce the sales for Colorado in each year between 1975 and 1988 by various percentages. We show the results of the standard synthetic control method in Figures 4.2.5 and 4.2.6. Note that in this situation, the weights for the characteristics do not change, while the synthetic unit does change when we account for smuggling, albeit differently than in Figure 4.2.5. For example, we now observe that the combination weight of Colorado steadily increases when the sales decrease by between 0 and 17 per cent and that no new states enter the synthetic unit. Lastly, note that the change in the composition of the counterfactual still impacts the causal effect of Proposition 99.

An interesting thing happens when we use the robust estimator instead of the standard estimator. As we show in Figure 4.2.7, the optimal combination weights do not change at all. Additionally, the same holds for the feature weights and, of course, the post-intervention paths. Hence we do

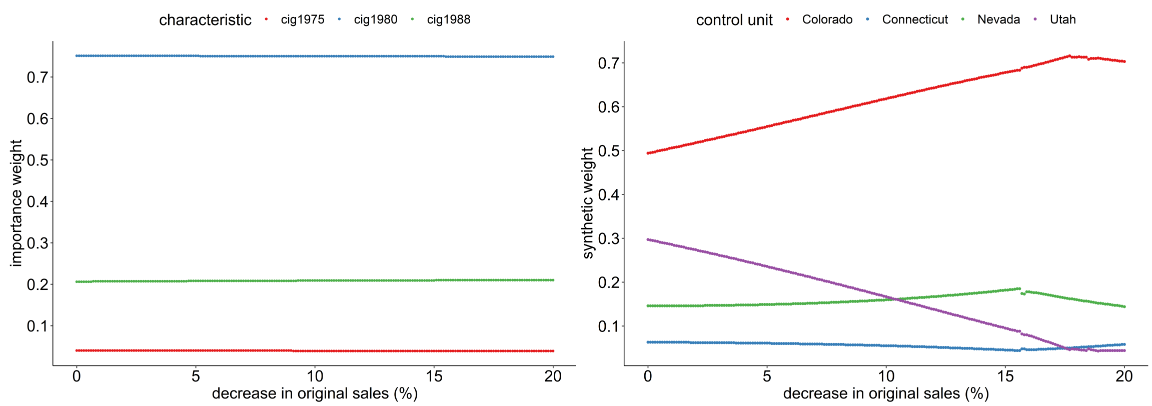

Figure 4.2.5: The non-zero feature (left) and combination weights (right) from the standard synthetic control estimator when we decrease the cigarette sales of Colorado between 1975 and 1988 by various percentages

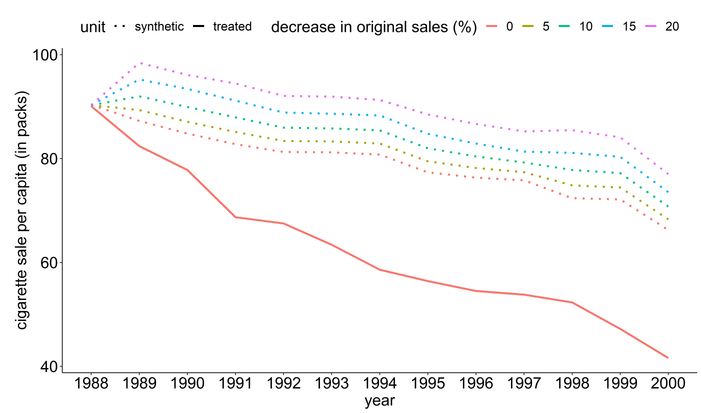

Figure 4.2.6: The post-intervention cigarette sales when we decrease the original sales data of Colorado between 1975 and 1988 by between 0 and 20 per cent.

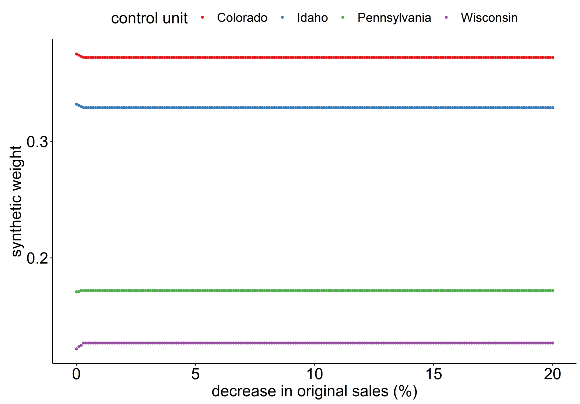

Figure 4.2.7: The non-zero synthetic weights from the robust method when we decrease the original sales of Colorado in each year between 1975 and 1988 by various percentages.

not present them. There is, however, an easy explanation for why this happens. Namely, since we use the growth rates in sales instead of levels, a reduction in the sales of Colorado during the entire pre-intervention period does not change the inputs for the robust estimator.

Lastly, we want to address the fact that multiple states may have been affected by cross-border cigarette smuggling. Nevertheless, the main takeaway from these examples is that even in a realistic setting with sensible uncertainties, the counterfactual of the standard regression-based synthetic control estimator is not robust to slight changes in the dataset, which in turn can lead to uncertain causal effects. Therefore we need to be careful with basing policy decisions on the results of this estimator.

5 Conclusion

In this paper, we discussed the robustness properties of the regression-based synthetic control estimator of Abadie et al. (2010). We showed that because the estimator uses a standard ordinary least squares estimator, the importance weights for the characteristics are susceptible to outliers in the data, especially when the features are non-robust functions of the pre-intervention variables. We also derived the influence function of the quadratic program, which revealed that the synthetic unit is not robust to outliers that affect the cross-sectional relationship between the characteristics. Consequently, small amounts of contamination can already lead to a strongly biased causal effect. Additionally, because the estimator is non-robust, it can mask outliers. Hence we cannot look at the differences between the pre-intervention paths of the treated and synthetic units to determine whether the counterfactual is valid.

With these findings, we suggested an alternative estimator. If necessary, this approach first makes each variable stationery, after which it uses a robust function to form the characteristics. Therefore, it keeps the effect of outliers on the cross-sectional relationship between the features to a minimum. Furthermore, it uses a robust regression estimator, such as the MM-estimator or the Mallows M-estimator, to find the regression coefficients and subsequently standardises these with an M-dispersion estimator. Moreover, it utilises a more balanced function instead of the weighted mean to turn the standardised coefficients into importance weights. Both Monte Carlo simulations and an empirical example showed that this alternative had a much better performance in terms of robustness than the classical method.

Although our approach to finding the synthetic unit is more robust than the classical approach, it can still happen that it produces poor estimates at low levels of contamination when the variances of the variables are high. The cause of this is that it still relies on solving a quadratic program, which is non-robust. In the current literature, there are no robust alternatives for the quadratic program. Therefore this is a possible subject for future research. We also did not address the consequences of outliers on the placebo test of Abadie et al. (2010), which determines whether the causal effect is significant or not. Because the test relies on the synthetic unit itself, it may not be robust and lead to a wrong conclusion. Consequently, this is also an interesting aspect of the synthetic control method to examine. Lastly, it could be worthwhile to repeat previous empirical research with our alternative approach and see how it alters the results and conclusions.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1):113–132.

Acemoglu, D., Johnson, S., Kermani, A., Kwak, J., and Mitton, T. (2016). The value of connections in turbulent times: Evidence from the United States. *Journal of Financial Economics*, 121(2):368–391.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Becker, M. and Klößner, S. (2018). Fast and reliable computation of generalized synthetic controls. *Econometrics and Statistics*, 5:1–19.

Cole, M. A., Elliott, R. J., and Liu, B. (2020). The impact of the Wuhan COVID-19 lockdown on air pollution and health: A machine learning and augmented synthetic control approach. *Environmental and Resource Economics*, 76(4):553–580.

Cunningham, S. and Shah, M. (2018). Decriminalizing indoor prostitution: Implications for sexual violence and public health. *The Review of Economic Studies*, 85(3):1683–1715.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, 157184.

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

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: The approach based on influence functions*, volume 196. New York, United States: John Wiley & Sons.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):492–518.

Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758.

Hubert, M. and Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):36–43.

Klößner, S., Kaul, A., Pfeifer, G., and Schieler, M. (2018). Comparative politics and the synthetic control method revisited: A note on Abadie et al.(2015). *Swiss Journal of Economics and Statistics*, 154(1):1–11.

Kuosmanen, T., Zhou, X., Eskelinen, J., and Malo, P. (2021). Design flaw of the synthetic control method. *MPRA Paper 106390*.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906–931.

Lovenheim, M. F. (2008). How far to the border?: The extent and impact of cross-border casual cigarette smuggling. *National Tax Journal*, 61(1):7–33.

Malo, P., Eskelinen, J., Zhou, X., and Kuosmanen, T. (2020). Computing synthetic controls using bilevel optimization. *MPRA Paper 104058*.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust statistics: Theory and Methods (with R)*. Hoboken, United States: John Wiley & Sons.

Peri, G. and Yasenov, V. (2019). The labor market effects of a refugee wave: Synthetic control method meets the mariel boatlift. *Journal of Human Resources*, 54(2):267–309.

Peña, D. and Yohai, V. (1999). A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association*, 94(446):434–445.

Stehr, M. (2005). Cigarette tax avoidance and evasion. *Journal of Health Economics*, 24(2):277–297.

Van Aelst, S. and Rousseeuw, P. (2009). Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):71–82.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656.

A Technical Appendix

A.1 Proof of Proposition 1

To find the $(N \times 1)$ vector of combination weights $\boldsymbol{w} = (w_2, \ldots, w_{N+1})$ given the non-negative weights v_p $(p = 1, \cdots, P)$, we solve the quadratic program

$$\begin{aligned} \arg\min_{\boldsymbol{w}} \quad & \sum_{p=1}^{P} \Big(h_p(\boldsymbol{y_1}, \boldsymbol{z_1})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}}, \boldsymbol{Z_{-1}})_{v_p}{}^T \boldsymbol{w} \Big)^2 \\ \text{s.t.} \quad & \boldsymbol{w}^T \boldsymbol{i} = 1 \\ & w_i \geq 0, \quad i = 2, \ldots, N+1, \end{aligned} \tag{23}$$

where $h_p(\cdot)_{v_p}$ is a function that maps the pre-intervention outcomes $\boldsymbol{y_i}$ and predictors $\boldsymbol{z_i}$ to the p'th weighted standardised characteristic for unit i (for ease of notation we suppress the superscript sd), $\boldsymbol{Y_{-1}}$ is a $(T_0 \times N)$ matrix with outcomes of the control units, and $\boldsymbol{Z_{-1}}$ is a $(K \times N)$ matrix with the characteristics of the control units.

By use of a quadratic penalty function we can rewrite the above as the bounded optimisation

$$\arg\min_{\boldsymbol{w}} \sum_{p=1}^{P} \Big(h_p(\boldsymbol{y_1}, \boldsymbol{z_1})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}}, \boldsymbol{Z_{-1}})_{v_p}{}^T \boldsymbol{w} \Big)^2 + \lambda(\boldsymbol{w}^T \boldsymbol{i} - 1)^2 \tag{24}$$

$$\text{s.t.} \quad w_i \geq 0, \quad i = 2, \ldots, N+1, \tag{25}$$

when $\lambda \to \infty$. Consequently. the first order conditions given λ are

$$\sum_{p=1}^{P} \Big(h_p(\boldsymbol{y_1}, \boldsymbol{z_1})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}}, \boldsymbol{Z_{-1}})_{v_p}{}^T \boldsymbol{w} \Big) \boldsymbol{h_p}(\boldsymbol{Y_{-1}}, \boldsymbol{Z_{-1}})_{v_p} + \lambda(\boldsymbol{w}^T \boldsymbol{i} - 1)\boldsymbol{i} = 0 \tag{26}$$

$$\text{s.t.} \quad w_i \geq 0, \quad i = 2, \cdots, N+1. \tag{27}$$

We can write (26) as a functional, namely

$$\begin{aligned} \sum_{p=1}^{P} \int \Big(h_p(\boldsymbol{y_1}, \boldsymbol{z_1})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}}, \boldsymbol{Z_{-1}})_{v_p}{}^T T(F) \Big) \boldsymbol{h_p}(\boldsymbol{Y_{-1}}, \boldsymbol{Z_{-1}})_{v_p} dF(\boldsymbol{Y}, \boldsymbol{Z}) + \\ \lambda(T(F)^T \boldsymbol{i} - 1)\boldsymbol{i} = 0, \end{aligned} \tag{28}$$

in which $T(\cdot)$ is the functional and $F(\boldsymbol{Y}, \boldsymbol{Z})$ is the cumulative distribution function (CDF) when the outcomes of all units follow a linear factor model. For ease of notation we suppress the bound constraints, because they only limit the parameter space and have no other effect with respect to the influence function.

Now assume that the observations come from the Tukey-Huber contamination model

$$F^{\epsilon} = (1-\epsilon)F + \epsilon\Delta_{(\boldsymbol{Y^*},\boldsymbol{Z^*})}. \tag{29}$$

Consequently, there is a chance of ϵ that the data come from the CDF $\Delta_{(\boldsymbol{Y^*},\boldsymbol{Z^*})}$, which is a distribution that puts all its mass at the point $(\boldsymbol{Y^*}, \boldsymbol{Z^*})$. When we substitute this distribution into (28) and separate the integrals we get

$$\begin{aligned}(1-\epsilon)\sum_{p=1}^{P}\int\Big(h_p(\boldsymbol{y_1},\boldsymbol{z_1})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}{}^{T}T(F^{\epsilon})\Big)\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}dF(\boldsymbol{Y},\boldsymbol{Z}) +\\ \epsilon\sum_{p=1}^{P}\int\Big(h_p(\boldsymbol{y_1},\boldsymbol{z_1})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}{}^{T}T(F^{\epsilon})\Big)\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}d\Delta(\boldsymbol{Y^*},\boldsymbol{Z^*}) +\\ \lambda(T(F^{\epsilon})^{T}\boldsymbol{i}-1)\boldsymbol{i} = 0.\end{aligned} \tag{30}$$

We now take the derivative with respect to ϵ and apply the chain rule to get

$$\begin{aligned}-(1-\epsilon)\frac{\partial T(F^{\epsilon})}{\partial\epsilon}\sum_{p=1}^{P}\int\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}{}^{T}dF(\boldsymbol{Y},\boldsymbol{Z}) -\\ \sum_{p=1}^{P}\int\Big(h_p(\boldsymbol{y_1},\boldsymbol{z_1})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}{}^{T}T(F^{\epsilon})\Big)\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}dF(\boldsymbol{Y},\boldsymbol{Z}) -\\ \epsilon\frac{\partial T(F^{\epsilon})}{\partial\epsilon}\sum_{p=1}^{P}\int\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}{}^{T}d\Delta(\boldsymbol{Y^*},\boldsymbol{Z^*}) +\\ \sum_{p=1}^{P}\int\Big(h_p(\boldsymbol{y_1},\boldsymbol{z_1})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}{}^{T}T(F^{\epsilon})\Big)\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}d\Delta(\boldsymbol{Y^*},\boldsymbol{Z^*}) +\\ \lambda\boldsymbol{i}\boldsymbol{i}^{T}\frac{dT(F^{\epsilon})}{\partial\epsilon} = 0.\end{aligned} \tag{31}$$

If we let $\epsilon \to 0^{+}$, the third term above vanishes and the second term equals

$$-\lambda(T(F)^{T}\boldsymbol{i}-1)\boldsymbol{i} \tag{32}$$

because of (28). When we use this along with the definition of the influence function we get

$$\begin{aligned}-\mathrm{IF}\Big((\boldsymbol{Y^*},\boldsymbol{Z^*});T,F,\lambda\Big)\sum_{p=1}^{P}\int\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}\boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}{}^{T}dF(\boldsymbol{Y},\boldsymbol{Z}) +\\ \lambda(T(F)^{T}\boldsymbol{i}-1)\boldsymbol{i} + \sum_{p=1}^{P}\Big(h_p(\boldsymbol{y_1^*},\boldsymbol{z_1^*})_{v_p} - \boldsymbol{h_p}(\boldsymbol{Y_{-1}^*},\boldsymbol{Z_{-1}^*})_{v_p}{}^{T}T(F)\Big)\boldsymbol{h_p}(\boldsymbol{Y_{-1}^*},\boldsymbol{Z_{-1}^*})_{v_p} +\\ \mathrm{IF}\Big((\boldsymbol{Y^*},\boldsymbol{Z^*});T,F,\lambda\Big)\lambda\boldsymbol{i}\boldsymbol{i}^{T} = 0.\end{aligned} \tag{33}$$

Consequently, we find that

$$\mathrm{IF}\Big((\boldsymbol{Y^*},\boldsymbol{Z^*});T,F,\lambda\Big)=\Bigg(\sum_{p=1}^{P}\int \boldsymbol{h_p}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}\boldsymbol{h}(\boldsymbol{Y_{-1}},\boldsymbol{Z_{-1}})_{v_p}{}^T dF(\boldsymbol{Y},\boldsymbol{Z})+\lambda I_p\Bigg)^{-1}\cdot$$

$$\Bigg(\lambda(1-T(F)^T\boldsymbol{i})\boldsymbol{i}+\sum_{p=1}^{P}\Big(h_p(\boldsymbol{y_1^*},\boldsymbol{z_1^*})_{v_p}-\boldsymbol{h_p}(\boldsymbol{Y_{-1}^*},\boldsymbol{Z_{-1}^*})_{v_p}{}^T(F)\Big)\boldsymbol{h_p}(\boldsymbol{Y_{-1}^*},\boldsymbol{Z_{-1}^*})_{v_p}\Bigg),\quad (34)$$

where we assume that the first term on the right hand side is invertible. Note that, the above still depends on λ, hence the true influence function equals

$$\lim_{\lambda\to\infty}\mathrm{IF}\Big((\boldsymbol{Y^*},\boldsymbol{Z^*});T,F,\lambda\Big).\quad (35)$$

In case this limit exists, we have an influence function of the form

$$\mathrm{IF}\Big((\boldsymbol{Y^*},\boldsymbol{Z^*});F\Big)=\boldsymbol{A}\cdot\sum_{p=1}^{P}\Big(h_p(\boldsymbol{y_1^*},\boldsymbol{z_1^*})_{v_p}-\boldsymbol{h_p}(\boldsymbol{Y_{-1}^*},\boldsymbol{Z_{-1}^*})_{v_p}{}^T T(F)\Big)\boldsymbol{h_p}(\boldsymbol{Y_{-1}^*},\boldsymbol{Z_{-1}^*})_{v_p}+\boldsymbol{c},\quad (36)$$

where $\boldsymbol{A}$ and $\boldsymbol{c}$ are a matrix and vector that do not depend on the point-mass contamination.

A.2 Proof of Proposition 2

The normal equations of the M-estimator for the cross-sectional regression step at time t equal

$$\sum_{i=1}^{N+1}\psi\Big(\frac{y_{it}-\boldsymbol{h}(\boldsymbol{y_i},\boldsymbol{z_i})^T\boldsymbol{\beta_t}}{\sigma_t}\Big)\boldsymbol{h}(\boldsymbol{y_i},\boldsymbol{z_i})=0,\quad (37)$$

where y_{it} is the outcome of unit i at time t, $\boldsymbol{h}(\boldsymbol{y_i},\boldsymbol{z_i})$ is the $(P\times 1)$ vector of characteristics for unit i, σ_t is the standard deviation of the idiosyncratic shock, and β_t is the vector of regression coefficients.

We can rewrite the normal equations in terms of a functional

$$\sum_{i=1}^{N+1}\int\psi\big[y_{it}-\boldsymbol{h}(\boldsymbol{y_i},\boldsymbol{z_i})^T T(F_1,\cdots,F_N)\big]\boldsymbol{h}(\boldsymbol{y_i},\boldsymbol{z_i})dF_i(\boldsymbol{y_i},\boldsymbol{z_i})=0,\quad (38)$$

in which $T(\cdot)$ is the estimator and F_i is the CDF of the data for unit i when its outcomes follow a linear factor model.

Now assume that there is a chance ϵ_i that the data of unit i comes from the Tukey-Huber contamination model

$$F_i^\epsilon=(1-\epsilon_i)F_i+\epsilon_i\Delta_{(\boldsymbol{y_i^*},\boldsymbol{z_i^*})},\quad (39)$$

in which $\Delta_{(\boldsymbol{y_i^*},\boldsymbol{z_i^*})}$ is a CDF that puts all it mass at the point $(\boldsymbol{y_i^*},\boldsymbol{z_i^*})$. Without loss of generality, suppose that only for the j'th unit $(1<j<N)$ the chance of contamination is larger than zero.

Additionally, for ease of notation, let $T_j^{\epsilon}(\cdot) = T(F_1, \ldots, F_j^{\epsilon}, \ldots, F_{N+1})$. Then the normal equations of the M-estimator under contamination equal

$$\begin{aligned}\sum_{i \neq j} \int \psi\big[y_{it} - \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T T_j^{\epsilon}(\cdot)\big] \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i}) dF_i(\boldsymbol{y_i}, \boldsymbol{z_i}) +& \\ (1 - \epsilon_j) \int \psi\big[y_{jt} - \boldsymbol{h}(\boldsymbol{y_j}, \boldsymbol{z_j})^T T_j^{\epsilon}(\cdot)\big] \boldsymbol{h}(\boldsymbol{y_j}, \boldsymbol{z_j}) dF_j +& \\ \epsilon_j \int \psi\big[y_{jt} - \boldsymbol{h}(\boldsymbol{y_j}, \boldsymbol{z_j})^T T_j^{\epsilon}(\cdot)\big] \boldsymbol{h}(\boldsymbol{y_j}, \boldsymbol{z_j}) d\Delta_{(\boldsymbol{y_j^*}, \boldsymbol{z_j^*})}(\boldsymbol{y_j}, \boldsymbol{z_j}) = 0.&\end{aligned} \tag{40}$$

If we now take the derivative with respect to ϵ_j, let $\epsilon_j \to 0^+$, and use the definition of the influence function like in the proof of Proposition 1, we get

$$\begin{aligned}\mathrm{IF_j}\big((\boldsymbol{y_j^*}, \boldsymbol{z_j^*}); F_1, \cdots, F_N\big) \sum_{i=1}^{N} \int \psi'\big[y_{it} - \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T T(F_1, \cdots, F_N)\big] \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i}) \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T dF_i(\boldsymbol{y_i}, \boldsymbol{z_i}) -& \\ \psi\big[y_{jt}^* - \boldsymbol{h}(\boldsymbol{y_j^*}, \boldsymbol{z_j^*})^T T(F_1, \cdots, F_N)\big] \boldsymbol{h}(\boldsymbol{y_j^*}, \boldsymbol{z_j^*}) = 0,&\end{aligned} \tag{41}$$

hence the influence function is given by

$$\begin{aligned}\mathrm{IF_j}\big((\boldsymbol{y_j^*}, \boldsymbol{z_j^*}); F_1, \cdots, F_N\big) = \psi\big[y_{jt}^* - \boldsymbol{h}(\boldsymbol{y_j^*}, \boldsymbol{z_j^*})^T T(F_1, \cdots, F_N)\big] \boldsymbol{h}(\boldsymbol{y_j^*}, \boldsymbol{z_j^*}) \cdot& \\ \left(\sum_{i=1}^{N} \int \psi'\big[y_{it} - \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T T(F_1, \cdots, F_N)\big] \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i}) \boldsymbol{h}(\boldsymbol{y_i}, \boldsymbol{z_i})^T dF_i(\boldsymbol{y_i}, \boldsymbol{z_i}) \right)^{-1}.&\end{aligned} \tag{42}$$

B Simulation Details Appendix

B.1 Simulation details Sections 2.3.1 and 2.3.2

For the simulations in Sections 2.3.1 and 2.3.2. we generate the data according to the linear factor model

$$\begin{aligned}y_{it}^U &= \sum_{k=1}^{5} \mathcal{N}(\mu_{ik}, 1) + \kappa_i \lambda_t + \mathcal{N}(0, 0.1), \quad i = 1, \cdots, 11, \quad t = 1, \cdots, 51, \\ y_{it} &= y_{it}^U + D_{it},\end{aligned}$$

where λ_t is the unobserved factor that follows an auto-regressive (AR) model without a constant and slope parameter 0.6. The innovations of the AR model come from a standard normal distribution and we use 65 burn-in periods. Additionally, D_{it} is a dummy variable equal to one when $i = 1$ and $t = 51$, μ_{ik} is a unit and covariate-specific mean, and κ_i is a unit-specific factor loading. We show the values for these unit-specific parameters in Table B.1.1. Lastly, we repeat the experiment 500 times.

Table B.1.1: Values of the parameters for the simulations in Sections 2.3.1 and 2.3.2

	Parameter					
Unit	μ_{i1}	μ_{i2}	μ_{i3}	μ_{i4}	μ_{i5}	κ_i
Treated	0.90	-2.40	1.60	3.40	5.70	0.46
1st Control	1.00	-2.00	1.50	4.00	6.00	0.50
2nd Control	0.50	-4.00	2.00	1.00	4.50	0.30
3th Control	0.20	1.00	1.80	3.00	2.00	0.80
4th Control	0.40	-3.00	2.00	7.20	1.40	0.30
5th Control	0.80	-3.00	0.50	2.40	1.80	-0.50
6th Control	1.20	0.30	2.70	0.40	3.00	1.20
7th Control	1.20	-0.80	0.44	3.20	1.00	1.50
8th Control	0.70	2.40	9.10	5.30	0.80	0.40
9th Control	1.80	0.33	-1.00	0.40	0.22	0.90
10th Control	2.20	-3.00	0.40	3.40	2.10	0.00

Note: The parameters for the treated unit are a convex combination of the first two control units, namely (0.8, 0.2).

B.2 Simulation details Section 2.3.3

For the simulations in Section 2.3.3 we first generate an observable covariate for each unit that follows an AR(1) process with unit-specific constant c_i and slope parameter ϕ_i. The innovations of this process come from a standard normal distribution. Then, we generate the outcome as a linear factor model

$$
\begin{aligned}
y_{it}^U &= z_{it} + k_i \lambda_t, \quad i = 1, \cdots, 11, \quad t = 1, \cdots, 51, \\
y_{it} &= y_{it}^U + D_{it},
\end{aligned}
$$

where z_{it} is the unit-specific and observable AR(1) process and the other parameters are explained in Section B.1. The values for the unit-specific parameters are given in Table B.2.1.

Table B.2.1: Values of the parameters for the simulations in Section 2.3.3.

	Parameter		
Unit	c_i	ϕ_i	κ_i
Treated	0.60	0.34	0.47
1st Control	0.60	0.20	0.50
2nd Control	0.40	0.40	0.20
3th Control	0.90	0.60	0.80

Note: The parameters for the treated unit are a convex combination of the control units, namely (0.5, 0.3, 0.2).

B.3 Simulation details Section 3.3

As mentioned in Section 3.3, we first generate seven time series that follow the auto-regressive (AR) model

$$x_{j,t} = a_j + b_j t + \phi_{j,1} y_{j,t-1} + \phi_{j,2} y_{j,t-2} + \phi_{j,3} y_{j,t-3} + \epsilon_{j,t}, \quad j = 1, \ldots, 7, \tag{43}$$

where $\epsilon_{j,t}$ is the innovation and follows a standard normal distribution. After this, we turn each time-series into the observable covariate

$$z_{i,j,t} = m_{i,j} x_{j,t}, \; j = 1, \ldots, 7, \tag{44}$$

where $m_{i,j}$ is a unit and time-series specific constant. We display the parameters for the seven time-series and the unit-specific coefficients in Tables B.3.1 and B.3.2, respectively and repeat the experiment 500 times.

The implementation details for the Mallows method are as follows:

- It uses a Huber loss function with 1.35 as the cutoff value such that it has a 95% efficiency at the normal.
- The weighting function $\eta(\cdot)$ is the inverse of the Mahalanobis distance based on the Minimum Covariance Determinant.
- For the standardisation of the regression coefficients, it uses an optimal (99% efficiency and 50% breakdown point) dispersion M-estimator as explained in Maronna et al. (2019).

Lastly, the details of the MM method are:

- It uses the S-estimator with a 50% breakdown point for the first two steps of the MM algorithm.
- As starting values for the S-estimator it utilizes the deterministic approach of Peña and Yohai (1999).
- For the third step in the MM algorithm, it uses an M-estimator with the Tukey loss function and the cut-off value 4.685.
- For the standardisation of the regression coefficients, it uses an optimal (99% efficiency and 50% breakdown point) dispersion M-estimator as explained in Maronna et al. (2019).

Table B.3.1: Parameter values for the seven time-series of Section 3.3.

	Parameter				
Time-series	a_j	b_j	$\phi_{j,1}$	$\phi_{j,2}$	$\phi_{j,3}$
1	0.00	2.00	0.70	0.00	0.00
2	2.00	0.00	0.20	0.40	0.00
3	0.20	0.00	0.55	0.00	0.00
4	0.40	0.00	0.15	-0.70	0.18
5	-0.43	0.00	-0.18	0.00	0.00
6	0.00	0.00	0.50	0.00	0.00
7	0.00	0.00	0.33	0.27	0.00

Note: Only the first time-series is non-stationary because it has a time trend.

Table B.3.2: Parameter values of each unit for the simulations in Section 3.3

	Parameter							
Unit	κ_i	$m_{j,1}$	$m_{j,2}$	$m_{j,3}$	$m_{j,4}$	$m_{j,5}$	$m_{j,6}$	$m_{j,7}$
Treated	0.237	-1.461	-0.588	0.483	-0.075	-0.178	-0.035	0.613
1st Control	0.350	-1.220	-1.860	0.580	0.570	-1.270	1.370	1.490
2nd Control	0.840	-1.910	1.200	1.890	-1.620	1.430	-1.180	-1.580
3rd Control	-0.950	-1.390	-0.090	-1.870	0.630	0.140	-1.830	1.710
4th Control	0.640	0.080	-1.640	1.680	1.640	0.310	-0.820	0.150
5th Control	-0.440	-0.480	-1.610	-1.510	1.960	0.300	-1.130	0.550
6th Control	-0.700	-0.790	-0.690	1.620	-0.290	-0.790	-1.350	-1.080
7th Control	-0.910	-0.670	0.180	-0.790	-1.990	0.670	0.800	-1.690
8th Control	0.090	-0.170	-1.300	0.500	0.410	1.720	0.510	-1.690
9th Control	0.860	-1.810	-0.050	-1.330	-1.270	-1.350	-1.650	-1.810
10th Control	0.220	0.530	-0.760	0.630	0.430	1.290	0.300	0.590
11th Control	0.700	-1.290	-1.060	-0.660	-0.340	0.400	0.200	0.790
12th Control	0.990	-1.520	1.560	-0.870	-1.190	1.480	-0.300	-1.840
13th Control	0.200	-0.620	-0.440	1.970	1.840	1.330	1.670	0.550
14th Control	-0.340	1.740	-1.270	-0.480	-0.610	-0.070	-1.160	1.660

Note: The parameters of the treated unit are a convex combination of the first three control units, namely (0.5, 0.3, 0.2).

C Additional Results Appendix

C.1 Additional results Section 3.3.2.

Table C.1.1: Average estimates of the SCM method for the different simulation scenarios when only the data of the first control unit can contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
Scenario 1	0.01	0.46	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.28	0.17	4.28 (0.08)
	0.10	0.49	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.22	0.06	4.76 (0.10)
	0.25	0.48	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.20	0.04	4.61 (0.11)
	0.50	0.50	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	2.93 (0.14)
Scenario 2	0.01	0.44	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.48	0.30	0.19	4.70 (0.10)
	0.10	0.69	0.28	0.01	0.01	0.00	0.01	0.00	0.01	0.41	0.29	0.18	6.11 (0.11)
	0.25	0.90	0.06	0.01	0.01	0.01	0.00	0.00	0.01	0.38	0.28	0.16	4.49 (0.12)
	0.50	0.95	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.36	0.27	0.16	3.15 (0.12)
Scenario 3	0.01	0.47	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.41	0.28	0.17	4.35 (0.10)
	0.10	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.21	0.07	4.16 (0.12)
	0.25	0.51	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.20	0.06	3.71 (0.12)
	0.50	0.52	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.07	1.19 (0.20)

Note 1: The Prob. column displays the probability of contamination.
Note 2: The combination weights (w_2, w_3, w_4) correspond with the first three control units and (v_1, v_2) are the importance weights for the characteristics of the outcome and the non-stationary covariate, respectively.
Note 3: The standard error of the treatment effect α is in parentheses. For the other entries, the standard error is smaller than 0.01 so we do not display them.

Table C.1.2: Average estimates of the Mallows method for the different simulation scenarios when only the data of the first control unit can contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
Scenario 1	0.01	0.18	0.21	0.08	0.10	0.14	0.10	0.08	0.12	0.49	0.29	0.19	3.99 (0.09)
	0.10	0.19	0.22	0.08	0.10	0.13	0.10	0.09	0.11	0.47	0.29	0.19	3.90 (0.09)
	0.25	0.19	0.22	0.08	0.09	0.13	0.10	0.08	0.11	0.32	0.24	0.15	3.91 (0.09)
	0.50	0.17	0.19	0.08	0.10	0.14	0.11	0.09	0.12	0.05	0.06	0.07	3.53 (0.10)
Scenario 2	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	4.11 (0.09)
	0.10	0.18	0.21	0.08	0.10	0.14	0.10	0.08	0.12	0.48	0.28	0.19	3.87 (0.09)
	0.25	0.18	0.20	0.08	0.10	0.14	0.10	0.08	0.12	0.47	0.29	0.19	3.95 (0.09)
	0.50	0.15	0.07	0.10	0.13	0.17	0.12	0.11	0.15	0.42	0.27	0.17	3.90 (0.08)
Scenario 3	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	4.05 (0.09)
	0.10	0.19	0.20	0.08	0.10	0.13	0.10	0.09	0.11	0.45	0.27	0.18	3.98 (0.09)
	0.25	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.30	0.19	0.14	3.92 (0.09)
	0.50	0.18	0.17	0.09	0.11	0.14	0.11	0.09	0.12	0.05	0.05	0.04	2.17 (0.14)

Note: For more details on the entries of the table see the table notes of Table C.1.1.

Table C.1.3: Average estimates of the MM method for the different simulation scenarios when only the data of the first control unit can contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
Scenario 1	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	4.09 (0.09)
	0.10	0.18	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.47	0.29	0.19	3.97 (0.10)
	0.25	0.18	0.20	0.08	0.10	0.14	0.10	0.08	0.12	0.47	0.30	0.19	3.98 (0.09)
	0.50	0.16	0.18	0.09	0.10	0.14	0.11	0.09	0.12	0.05	0.06	0.07	3.34 (0.10)
Scenario 2	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.82 (0.09)
	0.10	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	3.81 (0.09)
	0.25	0.19	0.22	0.08	0.09	0.13	0.10	0.08	0.11	0.32	0.24	0.15	3.80 (0.09)
	0.50	0.16	0.07	0.10	0.13	0.17	0.13	0.11	0.15	0.43	0.27	0.18	4.00 (0.09)
Scenario 3	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	3.86 (0.09)
	0.10	0.18	0.20	0.08	0.10	0.14	0.10	0.08	0.11	0.45	0.26	0.18	4.06 (0.08)
	0.25	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.30	0.19	0.14	3.77 (0.09)
	0.50	0.18	0.17	0.09	0.11	0.14	0.11	0.09	0.12	0.05	0.05	0.04	2.28 (0.14)

Note: For more details on the entries of the table see the table notes of Table C.1.1.

C.2 Additional results Section 3.3.3.

Table C.2.1: Average estimates of the Mallows method for simulation Scenario 1 when 3, 5, or 7 units contain contamination.

		Parameter											
	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.19	0.21	0.08	0.10	0.14	0.10	0.08	0.11	0.48	0.29	0.19	4.01 (0.07)
	0.10	0.18	0.20	0.08	0.09	0.14	0.10	0.09	0.12	0.42	0.24	0.17	3.71 (0.07)
	0.25	0.17	0.22	0.08	0.10	0.13	0.10	0.08	0.12	0.24	0.16	0.10	3.44 (0.08)
	0.50	0.15	0.17	0.09	0.11	0.15	0.11	0.10	0.13	0.01	0.08	0.01	2.92 (0.10)
5 Units	0.01	0.19	0.21	0.08	0.10	0.14	0.10	0.08	0.11	0.48	0.29	0.19	4.07 (0.07)
	0.10	0.18	0.20	0.08	0.10	0.14	0.10	0.09	0.12	0.39	0.20	0.16	3.76 (0.07)
	0.25	0.17	0.20	0.08	0.10	0.14	0.10	0.08	0.12	0.17	0.05	0.09	3.54 (0.08)
	0.50	0.17	0.17	0.08	0.12	0.15	0.10	0.08	0.13	0.01	0.00	0.01	3.00 (0.10)
7 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.12	0.48	0.29	0.19	3.96 (0.07)
	0.10	0.17	0.19	0.08	0.10	0.14	0.10	0.09	0.12	0.37	0.19	0.14	3.72 (0.07)
	0.25	0.18	0.20	0.08	0.10	0.14	0.10	0.08	0.12	0.17	0.05	0.05	3.50 (0.08)
	0.50	0.24	0.15	0.07	0.11	0.14	0.09	0.08	0.12	0.02	0.00	0.00	3.06 (0.10)

Note: For more details on the entries of the table see the table notes of Table C.1.1.

Table C.2.2: Average estimates of the Mallows method for simulation Scenario 2 when 3, 5, or 7 units contain contamination.

	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.19	0.21	0.08	0.10	0.14	0.10	0.08	0.11	0.49	0.29	0.19	3.95 (0.08)
	0.10	0.18	0.21	0.08	0.10	0.14	0.10	0.08	0.11	0.49	0.29	0.19	3.98 (0.07)
	0.25	0.19	0.21	0.08	0.10	0.13	0.09	0.08	0.11	0.49	0.29	0.19	4.00 (0.08)
	0.50	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	4.01 (0.07)
5 Units	0.01	0.19	0.21	0.08	0.10	0.14	0.10	0.08	0.11	0.49	0.29	0.19	3.95 (0.07)
	0.10	0.18	0.21	0.08	0.10	0.13	0.10	0.08	0.12	0.49	0.29	0.19	3.90 (0.08)
	0.25	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	3.91 (0.08)
	0.50	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.94 (0.08)
7 Units	0.01	0.19	0.21	0.08	0.10	0.14	0.10	0.08	0.11	0.49	0.29	0.19	3.97 (0.07)
	0.10	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.94 (0.07)
	0.25	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.98 (0.07)
	0.50	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.48	0.29	0.19	3.99 (0.07)

Column group heading: Parameter (v_1 through α).

Note: For more details on the entries of the table see the table notes of Table C.1.1.

Table C.2.3: Average estimates of the Mallows method for simulation Scenario 3 when 3, 5, or 7 units contain contamination.

	Prob.	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	w_1	w_2	w_3	α
3 Units	0.01	0.18	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.92 (0.07)
	0.10	0.17	0.20	0.08	0.10	0.14	0.10	0.09	0.12	0.43	0.26	0.18	3.89 (0.07)
	0.25	0.17	0.18	0.08	0.11	0.14	0.10	0.09	0.12	0.15	0.10	0.10	3.43 (0.07)
	0.50	0.10	0.11	0.12	0.11	0.15	0.14	0.13	0.15	0.12	0.07	0.07	2.21 (0.11)
5 Units	0.01	0.19	0.21	0.08	0.10	0.13	0.10	0.08	0.11	0.49	0.29	0.19	3.91 (0.08)
	0.10	0.15	0.18	0.09	0.11	0.15	0.11	0.09	0.13	0.39	0.22	0.16	3.91 (0.07)
	0.25	0.13	0.18	0.09	0.12	0.15	0.11	0.09	0.13	0.08	0.02	0.07	3.47 (0.08)
	0.50	0.09	0.10	0.10	0.15	0.17	0.13	0.12	0.14	0.18	0.10	0.07	2.84 (0.11)
7 Units	0.01	0.18	0.21	0.08	0.10	0.14	0.10	0.08	0.12	0.49	0.29	0.19	3.79 (0.08)
	0.10	0.13	0.18	0.09	0.11	0.15	0.11	0.09	0.13	0.37	0.20	0.12	3.84 (0.07)
	0.25	0.13	0.19	0.09	0.12	0.15	0.11	0.09	0.13	0.07	0.01	0.01	3.32 (0.08)
	0.50	0.11	0.09	0.10	0.16	0.15	0.12	0.13	0.14	0.20	0.10	0.06	3.05 (0.08)

Column group heading: Parameter (v_1 through α).

Note: For more details on the entries of the table see the table notes of Table C.1.1.