Hidden in plain sight: Influential sets in linear models*

Nikolas Kuschnig[1] , Gregor Zens[1], Jesús Crespo Cuaresma[1,2,3,4,5]

[1] Vienna University of Economics and Business
[2] International Institute for Applied Systems Analysis
[3] Wittgenstein Centre (IIASA, OeAW, Univ. Vienna)
[4] Austrian Institute of Economic Research (WIFO)
[5] CESifo

March 23, 2021

Abstract

Assessing the robustness of the results of econometric analysis is a long standing subject of lively research. The majority of the literature focuses on sensitivity to model specification, while the quantification of sensitivity to sets of influential observations has received relatively little attention. A major obstacle in this context is masking, a phenomenon where influential observations obscure each other, which makes their identification particularly challenging. We show how inferential measures are affected by influential sets of observations and present two adaptive algorithms aimed at identifying such sets. We demonstrate the merits of these algorithms via simulation studies and empirical applications. These exercises show that masking problems and a pronounced sensitivity to influential sets are present in a wide range of scenarios. Overall, our findings suggest that increased attention to influential sets is warranted and comprehensive robustness measures for regression analysis are required.

Keywords: regression diagnostics, robustness, masking, influence

*Corresponding author: Nikolas Kuschnig, at Welthandelsplatz 1, 1020 Vienna, Austria. Email: nikolas.kuschnig@wu.ac.at. The authors gratefully acknowledge helpful comments from Ryan Giordano.

1 Introduction

Econometric methods are an important instrument of scientific discovery in the social sciences. They provide evidence-based insights into the nature of socioeconomic processes, allow us to test theories, and provide predictions. Econometric specifications act as imperfect approximations to the relationships between economic variables. Estimates of these relationships are crucial for the design of evidence-based policy measures. As a result, the assessment of the *sensitivity* of inference to changes in modelling assumptions is a particularly important topic within the field of applied econometrics. The issue of robustness in the context of regression models is a long-standing and active subject of research (Angrist and Pischke, 2010; Athey and Imbens, 2017; Leamer, 1983; Levine and Renelt, 1992; Sala-i Martin, 1997; Sala-i Martin et al., 2004; Sims, 1980; Steel, 2020).

The literature dealing with robustness in regression models tends to focus on uncertainty related to inclusion or exclusion of control variables, the functional form linking covariates with the outcome variable, as well as on the development of methods to estimate quantities of interest in the presence of such specification uncertainty. Well-known approaches such as extreme bounds analysis (Leamer, 1983), model averaging (Steel, 2020), and, more generally, regularised estimation are useful tools for covariate selection and inference in the presence of model uncertainty. These approaches are concerned with the *horizontal* dimension of the data. Robustness with respect to inclusion or exclusion of particular observations in the sample – the *vertical* dimension of the data – has received relatively little attention in the literature, beyond the identification of outliers. The focus on asymptotic analysis has been claimed responsible for this lack of interest (Leamer, 2010). In this paper, we analyse the robustness across the vertical data dimension and investigate the sensitivity of parameter estimation in linear regression models to sets of influential observations.

It is well known that small sets of influential observations may hold considerable sway over regression results (Cook, 1979). These observations deserve particular attention. Sensitivity analysis based on their exclusion provides important insights into the stability of inferential quantities and thus about the validity of conclusions drawn. In a recent contribution, Broderick, Giordano, and Meager (2020), henceforth BGM, investigate the sensitivity of regression-based inference to the exclusion of such observations. They propose a metric to approximate the largest change in quantities of interest that can be induced by dropping a given number of observations. This metric allows the computation of summary statistics, such as the share of observations that would need to be excluded to induce particular changes. For example, a switch in the sign of the estimated coefficients or a change in statistical significance.

Such statistics are valuable additions to the toolkit of applied econometrics and an important step in creating measures of robustness along the vertical dimension of the data. However, several obstacles need to be overcome in order to obtain a useful robustness measure. The approximation used by BGM builds upon initial estimates of influence and induced perturbations. As a result, its performance is sub-optimal, for instance, in the presence of *masking*, a phenomenon where influential observations obscure other influential observations (Chatterjee and Hadi, 1986). This means that sets of influential observations are identified reliably and the size of the perturbations induced is underestimated. While BGM acknowledge that their approximation provides only a lower bound of sensitivity, their metric is thus prone to convey a false sense of robustness in settings where influential sets of observations are indeed present.

We investigate two alternative approaches for assessing sensitivity to sets of observations in the context of linear regression models. Instead of relying on first order approximations of influence, these approaches are adaptive in nature. This substantially increases the reliability of the assessment at small additional computational cost. The first algorithmic approach uses the same initial approximation as BGM, but computes exact measures of perturbation. The second approach makes use of a greedy algorithm that recursively identifies influential observation, thus addressing potential masking phenomena effectively. The merits of these methods are illustrated by means of a simulation exercise and two applications to real-world examples. We re-analyse seven randomised controlled trials (RCTs) on the effectiveness of microcredit in developing countries and revisit a macroeconomic study on the existence of poverty convergence. Our findings suggest that masking phenomena pose a consistent challenge at ascertaining sensitivity. We put the proposed approach and implied measures of sensitivity into context by reconciling the recent developments in the field with a large body of statistics literature on influential observations, outliers, and robust regression. As guidance for applied researchers we reflect upon the implications and relevance of the resulting sensitivity measure as an indicator for robustness.

The remainder of the paper is structured as follows. In Section 2, the theoretical framework is established and connected to the relevant literature. The computational details are presented in Section 3. In Section 4, we illustrate the theoretical concepts, potential shortcomings, and the merits of our approach using simulation studies and real world data. Section 5 contains some discussion and concluding remarks.

2 Influential sets in linear regression models

2.1 Theoretical framework

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{1}$$

where $\mathbf{y}$ is an $N \times 1$ vector containing observations of the dependent variable, $\mathbf{X}$ is an $N \times P$ matrix of observations of explanatory variables, β is a $P \times 1$ vector of coefficients to be estimated, and ε is an $N \times 1$ vector of independent error terms with zero mean and unknown variance σ^2. Observation i, $(\mathbf{y} : \mathbf{X})_i$, corresponds to the i'th row of $\mathbf{y}$ and $\mathbf{X}$, with the notation y_i used for the i'th observation in $\mathbf{y}$ and x_i for the vector containing the covariates in $\mathbf{X}$ for the ith observation. We indicate the deletion of the rows indexed by I using the subscript (I), such that $\mathbf{y}_{(I)}$ denotes the vector corresponding to the observations of the dependent variable without the elements identified by I. A hat is used to indicate estimated quantities – that is, $\hat{\beta}$ is an estimate of β.

Our main focus lies on the sensitivity of some measure of interest λ to removing sets of influential observations from the sample. Following Belsley, Kuh, and Welsch (1980), we define influential observations as those whose omission has a large impact on λ when compared to the omission of most other observations, either individually or as a set.

A set of observations is defined as a subset $\mathcal{S}$ of the set of all observations $\tilde{\mathcal{S}} = \{s | s \in \mathbb{Z} \cap [1, N]\}$, where each element $s \in \tilde{\mathcal{S}}$ is associated with an observation of the full sample. We define $\mathcal{S}_\alpha$ for $\alpha \in [0, 1]$ as set of observations of cardinality $N_\alpha = \lceil N\alpha \rceil$. The empty set is denoted by $\varnothing$. The measure of interest, λ, is a function of the data and a set of excluded observations. Assuming that we are interested in the sensitivity of the least squares (LS) estimate of β after dropping a set of observations, this function is given by

$$\lambda(\mathcal{S}, \mathbf{y}, \mathbf{X}) = \left(\mathbf{X}'_{(\mathcal{S})}\mathbf{X}_{(\mathcal{S})}\right)^{-1}\mathbf{X}'_{(\mathcal{S})}\mathbf{y}_{(\mathcal{S})}. \tag{2}$$

In order to simplify notation, we drop the dependence of $\lambda(\mathcal{S}, \mathbf{y}, \mathbf{X})$ on the data and denote it as $\lambda(\mathcal{S})$. We concentrate on the identification of the *minimal perturbing set*, i.e. the smallest set that achieves a given change in λ, which we will call the *target perturbation*. In order to formalise this set, we first define the *maximally perturbing set*, $\mathcal{S}^*_\alpha$, which achieves the maximal perturbation for a given number of omitted observations, as

$$\mathcal{S}^*_\alpha = \underset{\mathcal{S} \in [\mathcal{S}]^\alpha}{\arg\max}\, \Delta\left(\lambda(\mathcal{S}), \lambda(\varnothing)\right) \tag{3}$$

where $\Delta(\lambda(\mathcal{S}), \lambda(\varnothing))$ is a function measuring the perturbation of interest. A standard choice for $\Delta(\lambda(\mathcal{S}), \lambda(\varnothing))$ is a norm of the difference between the values of interest for the full sample and after omitting $\mathcal{S}$, i.e. $\Delta(\lambda(\mathcal{S}), \lambda(\varnothing)) = \|\lambda(\mathcal{S}) - \lambda(\varnothing)\|$. Other forms may be important in particular applications.[1] We use Δ^*_α to indicate the perturbation associated with the set $\mathcal{S}^*_\alpha$.

The *minimal perturbing set*, $\mathcal{S}^{**}$, is given by

$$\mathcal{S}^{**} = \min_{\alpha} \mathcal{S}^*_\alpha \quad \text{s.t. } \Delta(\lambda(\mathcal{S}^*_\alpha), \lambda(\varnothing)) \geq \Delta^{**}, \tag{4}$$

where Δ^{**} denotes the *target perturbation*. The cardinality of the solution, N_{α^*} is referred to as the *minimal perturbing size*. This framework is closely related to the one put forward by BGM, with parallels between the maximally perturbing set and the 'most influential set', as well as the minimal perturbing size and the 'perturbation inducing proportion' as summary statistic.

2.2 Influence measures and influential observations

The literature is rich in methods for the identification of single influential observations, i.e. sets with cardinality $N_\alpha = 1$, and regression methods that are robust to the existence of these influential observations (see e.g. Chatterjee and Hadi, 1986; Hampel et al., 2005; Maronna et al., 2019). Measures of influence are central to this pursuit and come in a wide variety. Chatterjee and Hadi (1986) review groups of such measures and point out strong interrelationships among them. A notable group considers residuals $\mathbf{y} - \mathbf{X}\hat{\beta}$, which can readily be extended with leverage, i.e. diagonal elements of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. In the linear regression setting, an observation is generally considered influential when it exhibits both a large residual and high leverage.

Another group of measures builds on the influence function of Hampel (1974). These include Cook's distance (Cook, 1979), the Welsch-Kuh distance (Belsley et al., 1980), and modifications thereof. The notion of the confidence ellipsoid is used in other approaches for quantifying influence, including the likelihood distance (Cook and Weisberg, 1982), and robust estimation methods (Rousseeuw and Yohai, 1984; Yohai, 1987). In addition, a number of Bayesian approaches have been proposed (Box and Tiao, 1968; Pettit and Young, 1990; Verdinelli and Wasserman, 1991).

Most of these approaches consider a holistic notion of influence. However, in applied econometrics we often care about the partial influence of observations on specific inferential quantities. Specifically, the sign and significance of coefficient estimates tend to be of

[1] If we are interested in robustness to perturbations that may flip the sign of the quantity of interest, for instance, $\Delta = \mathbb{1}(\text{sign}(\lambda_{\mathcal{S}}) \neq \text{sign}(\lambda_{\varnothing}))\,\|\lambda(\mathcal{S}) - \lambda(\varnothing)\|$ could be used.

interest. Assessing the influence of particular observations on these two measures appears as a natural post-estimation step. This notion is shared by a large body of literature, which has produced a number of relevant results on sensitivity to influential observations.

Considering the case of a single influential observation, with λ as the OLS estimator of β as in Equation (2) and $\Delta\left(\lambda(\mathcal{S}), \lambda(\varnothing)\right) = \lambda(\varnothing) - \lambda(\mathcal{S})$, the perturbation of coefficients from dropping a single observation i has a closed form solution and is given by

$$\delta_i = \Delta\left(\lambda(\mathcal{S}_{\frac{1}{N}}), \lambda(\varnothing)\right) = \hat{\beta} - \hat{\beta}(i) = \frac{(\mathbf{X}'\mathbf{X})^{-1} x_i' e_i}{1 - h_i}, \tag{5}$$

where $h_i = \mathbf{H}_{ii}$ is the i'th element on the diagonal of the hat matrix. This statistic is a popular and widely available measure of influence for single observations on coefficient estimates in the context of OLS estimation.[2] As will be discussed below, quantities such as δ_i play a central role in the approach proposed by BGM. Belsley et al. (1980) suggest a scaled version of δ_i, where standard errors of the coefficients are used for normalisation. This scaling step is carried out using an estimate of the error term variance, $\hat{\sigma}^2$, that disregards observation i and is thus given by $\hat{\sigma}^2(i) = \sum_{n \neq i}^{N} \left(y_n - x_n \hat{\beta}_{(i)}\right)^2 / (N - P - 1)$.

2.3 Identifying influential sets

For sets of influential observations with $N_\alpha > 1$, identification is considerably more challenging. Despite the availability of closed-form results for many essential quantities with respect to single observations, exact solutions to the optimisation problems posed above are often intractable. For all but the simplest problems, enumeration is infeasible, requiring a total of $\binom{N}{N_{\alpha^*}}$ calculations of the influence measure. These computational requirements can be circumvented via approximate methods, but these suffer from the fact that it is possible that $\mathcal{S}_\alpha^* \not\supset \mathcal{S}_\theta^*$, as well as $\Delta_\alpha^* \not> \Delta_\theta^*$ for $\alpha > \theta$. This problem hampers the performance of approximations and implies a trade-off between computational cost and accuracy in the identification of maximally perturbing sets.[3] These difficulties are connected to the phenomenon of masking, a situation where some influential observations conceal the influence of other observations. A related concept is swamping, where observations only appear influential due to other influential observations that affect the estimate of interest. To address these problems, influential sets need to be treated as a whole, as opposed to treating them as an accumulation of single influential observations.

[2] The statistic is termed DFBETA$_i$ by Belsley et al. (1980). Its calculation is implemented in standard statistical software; see for instance R's `influence()` and related functions or Stata's `dfbeta()` function.

[3] One extreme is the (infeasible) exact solution via enumeration of all conceivable sets. Approximations based on influential sets of unit size arguably represent the other extreme.

Measures of influence may be flawed and sensitivity analyses run the risk of being severely misleading if they build upon cumulative indicators based on single observations.

To avoid these conceptual and computational issues, the task of sensitivity analysis is often approached via robustifying estimates, which is elaborated upon in a large body of literature (Hampel et al., 2005; Huber, 1964; Maronna et al., 2019; Rousseeuw and Yohai, 1984; Shotwell et al., 2011). Assessing sensitivity to influential sets has also been addressed using resampling methods, such as the the jackknife or the bootstrap (Efron and Tibshirani, 1994). However, while closely related, sensitivity measures based on resampling methods and sensitivity measures based on excluding maximally perturbing sets are not equivalent. The former yields relatively aggregated measures of sensitivity. Excluding the maximally perturbing set sheds light on the worst-case scenario.[4]

Many contributions are concerned with the related task of detecting multiple outliers. Whereas influential observations are influential with respect to some quantity, the concept of an outlier is less clear and identification is essentially an unsupervised learning task. To ensure computational feasibility, attempts have built upon sequential application of methods for single-observation statistics (Caroni and Prescott, 1992), an approach that is closely related to the algorithms outlined below. Alternatively, clustering methods to identify sets of outliers have been put forward (Hadi, 1992; Hautamaki et al., 2004; Kaufman and Rousseeuw, 2009; Kim and Krzanowski, 2007). A number of contributions suggest model selection procedures that attempt to compare relevant metrics across all possible subsets of observations (Hoeting et al., 1996; Kim et al., 2008). However, even for moderately sized samples, the cardinality of the model space to assess is prohibitive. Even efficient algorithms, such as those based on Markov chain Monte Carlo methods, only explore an extremely small fraction of the full space of specifications.

Because of these obstacles, the literature addressing the identification of influential sets of arbitrary size N_α is limited, despite the abundant methods for detecting individual influential observations. A notable exception is the recent work by BGM, who propose an approximate metric to identify observations with large impact on a quantity of interest when dropped. The metric is built on gradients of the quantity of interest – such as the one given in Equation (5) – which are computed once for the full sample. Sets of influential observations are then identified iteratively by choosing observations corresponding to the largest gradients. Perturbations are calculated by cumulatively summing up individual gradients. This implies minimal computational expense, since both the influential set and the associated perturbation are obtained in one sweep. The metric can be computed efficiently for conceivable influence measures using auto-differentiation.

[4] In particular, dropping the maximally perturbing set of size N_α reveals the extreme bounds of a delete-N_α jackknife distribution of estimates.

It is generally applicable and provides "an exact finite-sample lower bound on sensitivity for any estimator" (Broderick et al., 2020, p.1). Guaranteed computational feasibility makes this an appealing approach for sensitivity analysis.

However, minimal computational efforts imply a loss of accuracy that can be problematic when assessing sensitivity. In the absence of particularly influential observations, i.e. by extent masking and swamping, such an approximation may suffice. In fact, it will largely mirror conclusions from conventional measures, such as the one given in Equation (5). However, in the presence of these obstacles, initial measures of influence may impair the quality of the detected sets considerably. For the approximation used by BGM, this holds true not only with respect to the identification of influential sets, but also with respect to the accuracy of the reported perturbations. The latter conclusion directly results from the fact that $\delta_{\{i,j\}}$ does not equal $\delta_i + \delta_j$.

Taken together, these considerations imply that computationally cheap, but naïve measures of sensitivity to influential sets suffer precisely in the situations they are conceived to detect. In the subsequent sections, we illustrate, discuss, and address these issues in various settings. We compare three algorithms for identifying minimal perturbing sets and the associated perturbations that provide remedies for the challenges related to masking and swamping.

3 Algorithmic approaches to the identification of influential sets

In the following, three algorithms with the objective of identifying minimal perturbing sets are outlined. The first one is a naïve approximation based on influence measures for the full sample. The second and third algorithm use adaptive techniques instead of static approximations and provide substantial improvements on the naïve approximation at minimal extra cost.[5]

3.1 Algorithm 0 – Initial set and initial perturbation

The first algorithm – henceforth Algorithm 0 – is based on influence measures of single observations, computed once using the full sample. It is extremely cheap from a computational perspective, but likely to give inaccurate results when identifying influential sets. Algorithm 0, which essentially coincides with the approach proposed by BGM, approximates perturbations by accumulating influence measures derived from omitting

[5]The idea of employing adaptive algorithms to identify influential sets has been previously outlined, but not explored in detail, in Belsley et al. (1980).

Algorithm 0: Initial set and initial perturbation approximation.

Result: An estimate of $\mathcal{S}_\alpha^{**}$ and associated Δ_α^* and α^*.

regress $y \sim X$, set initial size s.t. $N_\alpha = 1$, and maximal size $\overline{\alpha}$;

calculate the influence measure $\delta_i = \Delta\left(\lambda(\mathcal{S}_{\frac{1}{N}}), \lambda(\varnothing)\right)$ for all i;

create an ordered set $\mathcal{S}_1$ by ranking δ_i;

while $\Delta < \Delta^{**} \vee \hat{\alpha}^* > \alpha < \overline{\alpha}$ **do**

build the set $\mathcal{S}_\alpha$ using the first N_α elements of $\mathcal{S}_1$;

approximate the perturbation $\Delta = \sum_n \delta_n$ for all $n \in \mathcal{S}_\alpha$;

if $\Delta < \Delta^{**}$ **then** increase the size α;

else set $\hat{\alpha}^* = \alpha$ and decrease the size α;

end

single observations. Computation of most measures of interest is trivial, since they are often available in closed form. The desired results can usually be obtained from the same factorisation that is used to arrive at the least squares coefficient estimates. This is typically a QR factorisation, which allows the use of efficient triangular solvers, resulting in a computational complexity of $\mathcal{O}(NP^2)$.

The approximate set identification of Algorithm 0 may yield useful estimates when there is one distinct and homogeneous influential set. If there are multiple sets or observations are relatively spread out, the approximation suffers from masking and swamping problems. The quality of the perturbation approximation, $\hat{\Delta}_\alpha^*$, is restricted by the quality of the approximate maximally perturbing set. In addition, its accuracy suffers severely when there is more than one influential observation present. Taken together, Algorithm 0 is only expected to work well when there is at most one influential observation. Larger influential sets or multiple influential sets of any size will necessarily distort influence scores of any single observation and hence negatively affect approximation quality.

3.2 Algorithm 1 – Initial set and exact perturbation

Algorithm 1 is a slight adaptation of Algorithm 0 that employs the same initial approximation to identify maximally perturbing sets, but computes exact measures of perturbation. Under the assumption that perturbations are increasing with the size of the set (as implicit in Algorithm 0), the task can be interpreted along the lines of a binary search. The size is initialised and updated by halving the interval to search repeatedly, following a divide-and-conquer strategy. The computational overhead amounts to a worst case of $\mathcal{O}(\log N_{\overline{\alpha}})$ recalculations of the perturbation, where $\overline{\alpha}$ denotes the maximal allowed

Algorithm 1: Initial set approximation and exact perturbation.

Result: An estimate of $\mathcal{S}_{\alpha}^{**}$ and associated Δ_{α}^{*} and α^{*}.

regress $y \sim X$, set initial size α, and maximal size $\overline{\alpha}$;

calculate the influence measure $\delta_i = \Delta\left(\lambda(\mathcal{S}_{\frac{1}{N}}), \lambda(\varnothing)\right)$ for all i;

create an ordered set $\mathcal{S}_N$ by ranking δ_i;

while $\Delta < \Delta^{**} \vee \hat{\alpha}^{*} > \alpha < \overline{\alpha}$ **do**

build the set $\mathcal{S}_{\alpha}$ using the first N_{α} elements of $\mathcal{S}_N$;

calculate the perturbation $\Delta(\lambda(\mathcal{S}_{\alpha}), \lambda(\varnothing))$;

if $\Delta < \Delta^{**}$ **then** increase the size α;

else set $\hat{\alpha}^{*} = \alpha$ and decrease the size α;

end

size of the maximally perturbing set (which could be determined using Algorithm 0).[6] For least squares estimation, the computational cost of recalculation is hardly restrictive, making the calculation of full paths of perturbation, i.e. for α such that $N_{\alpha} = 1, \ldots, N_{\alpha^{*}}$, attractive.

Algorithm 1 addresses one major concern of the naïve approach, since re-calculating the exact perturbations for a given value of α accounts for the joint impact of individual influential sets. Nevertheless, the set approximation still suffers from masking and swamping. It is worth to note that, due to the computation of exact perturbation measures, this algorithm can actually be used to diagnose the existence of such phenomena. In the case where only one heavily influential set is present, the algorithm should be able to identify the set based on the gradient of individual influences and should therefore be able to yield the correct perturbation, which exceeds the cumulative sum of individual gradients. Assessing the perturbation path after such a highly influential set is removed entirely may be informative of masking and swamping issues. For instance, if the induced perturbations decrease considerably after the removal of a set, this can be interpreted as evidence for the existence of a masking problem.

In summary, Algorithm 1 uses a computationally efficient approximation to the minimal perturbing set and yields exact associated perturbation measures at negligible computational cost. Minimal perturbing sets are expected to be considerably smaller than those for Algorithm 0. Nevertheless, the set identification procedure may not adequately address masking and swamping issues; a flaw that is addressed in the third algorithm, outlined below.

[6] Further improvements in computational speed are possible, some of which are outlined in Subsection 3.3.

Algorithm 2: Adaptive set approximation and exact perturbation.

Result: An estimate of $\mathcal{S}_\alpha^{**}$ and associated Δ_α^* and α^*.

regress $y \sim X$, set initial size $\alpha = 0$, maximal size $\overline{\alpha}$, and step size t;

while $\Delta < \Delta^{**}$ **do**

  calculate the influence measure $\delta_i = \Delta\left(\lambda(\mathcal{S}_{\alpha+\frac{1}{N}}), \lambda(\mathcal{S}_\alpha)\right)$ for all $i \notin \mathcal{S}_\alpha$;

  build the set $\mathcal{T}_t$ from the t observations with maximal influence;

  build the set $\mathcal{S}_{\alpha+t}$ by forming the union $\mathcal{S}_\alpha \cup \mathcal{T}_t$;

  calculate the perturbation $\Delta(\lambda(\mathcal{S}_\alpha), \lambda(\varnothing))$;

  if $\alpha < \overline{\alpha}$ **then** increase the size α by t;

  else break;

end

3.3 Algorithm 2 – Adaptive set and exact perturbation

Algorithm 2 uses an adaptive procedure for identifying the minimal perturbing set. Influential observations are identified at each step in a recursive manner, facilitating the discovery of potentially masked observations. Essentially, Algorithm 2 can be characterised along the lines of a *greedy* algorithm that makes locally optimal decisions after dropping an observation. Starting with influential sets formed by single observations, the procedure adds observations with maximal influence to the set and updates the perturbed measures of interest until the target perturbation is achieved. Algorithm 2 addresses masking and swamping issues, yielding the arguably preferable method in settings with multiple influential sets.

The computational complexity of Algorithm 2 and Algorithm 1 differs primarily due to the number of recalculations that are required. The fact that we need to adaptively and step-wise increase the size of the identified set prevents us from using a divide-and-conquer strategy. That is, there is an upper bound of $\mathcal{O}(N_{\overline{\alpha}})$ on the recalculations needed. Generally, this procedure may appear to be computationally prohibitive, but efforts required are much smaller than e.g. alternative delete-d jackknife algorithms. In fact, the computational efforts are hardly restrictive in the framework of least squares estimation with moderate sample sizes.[7]

[7]For the applications presented in the following section, computation was effectively instantaneous or, at worst, a matter of seconds. Large-scale problems can be tackled by increasing the speed of the algorithm via more efficient methods to solve the linear system of equations (Shewchuk, 1994; Trefethen and Bau, 1997), various updating methods (Hammarling and Lucas, 2008; Reichel and Gragg, 1990; Sherman and Morrison, 1950), the Frisch-Waugh-Lovell theorem (Frisch and Waugh, 1933) to isolate coefficients of interest, and increased or adaptive step sizes for each iteration.

4 Illustrative empirical examples

This section presents a number of empirical examples to illustrate the performance of the algorithms discussed above and demonstrate their relative merits. We focus on perturbations of a coefficient of interest. Other perturbations and targets, such as changes in statistical significance or significant sign switches, are reported in passing, but could be pursued explicitly. First, we present examples based on simulated data. Second, we revisit seven RCTs that assess the effects of microfinance on household profits and have been studied previously in Meager (2019) and BGM. Third, we discuss the role of influential sets when assessing cross-country convergence in poverty rates. For each empirical example, we investigate and report the performance of the three algorithms outlined above.[8]

4.1 Simulated data

Consider fitting a simple linear regression model to the data depicted in Figure 1. The three observations in the right-most area of the scatter plot, marked (a), are extremely influential with respect to $\hat{\beta}$. This circumstance is reflected in common diagnostics, such as residuals, leverage, scale-location plots, and Cook's distance. Unsurprisingly, this set of observations is also identified as an influential set by all three outlined algorithms. The identified influential sets of size larger than three, however, strongly differ across algorithms. The influence of the observations marked as (b) in Figure 1 is masked by the three observations in (a). As a result, neither Algorithm 0 nor Algorithm 1 include them in approximations of the maximally perturbing set for reasonable sizes.[9] In contrast, the adaptive nature of Algorithm 2 leads to the identification of these data points as influential for sets of size four or larger.

Figure 2 depicts the identified influential set of increasing size for the three algorithms. In addition, the least squares regression lines after removing influential sets of size three and seven are provided. The panels of Figure 2 reveal how masking and swamping can affect the identification of influential sets. New members of the set for sizes above three are constrained to the lower left quadrant when using Algorithms 0 or 1. These observations are relatively inconsequential with respect to the least squares fit and their membership in the influential set is the product of swamping effects. The perturbation

[8]For comparability reasons, we use the implementation of Algorithm 0 in the R package `zaminfluence` (Broderick et al., 2020), which is available at `github.com/rgiordan/zaminfluence`. Given the targeted perturbations, this should be equivalent to using Algorithm 0 with the measure of influence given in Equation (5). In practice, the BGM metric falls slightly behind, as documented in Figure A4 of the Appendix.

[9]The observations are removed at sizes N_α of 36 and 40, out of a total of 50 observations.

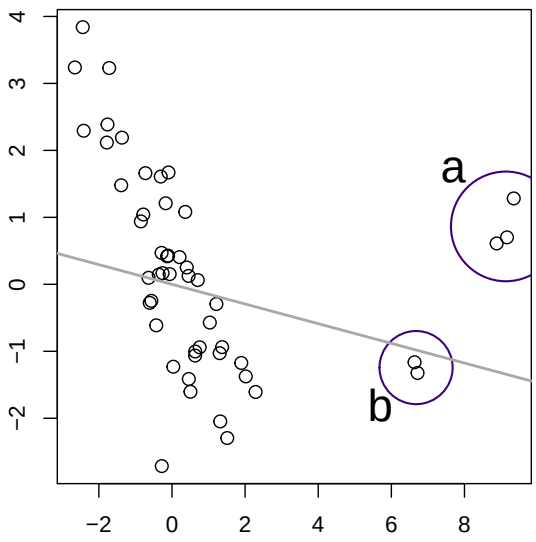

Figure 1: Scatter plot of the dependent against the explanatory variable ($N = 50$). The least-squares fit is indicated by the gray line, influential sets are labelled (a) and (b).

achieved by Algorithm 0 and 1 reaches a relative peak after removing three observations. Estimates of the regression line which are essentially indistinguishable when larger sets are considered (see the Appendix, Figure A1 for more details). Moreover, the approximation in Algorithm 0 fails to account for the full influence of the first influential set, yielding a significantly lower perturbation after its removal. Algorithm 2 does not suffer from these problems and yields considerably more accurate sets for $N_\alpha > 3$. The additional perturbations induced remain relevant and the regression lines differ clearly.

In order to assess whether differences across algorithms are systematic, we perform a simulation exercise with four different data generating processes (DGPs). DGP1 is a linear regression model where observations of the dependent variable are linked to a covariate drawn from a standard normal distribution using a slope parameter β=1. DGP2 is similar, but ten percent of the draws of the covariate correspond to a normal distribution with higher variance, resulting in a subsample with increased leverage. DGP3 introduces a mixture component, where ten percent of the observations of the covariate have high leverage and a coefficient above unity. These observations can be seen as a (potentially disjoint) influential set with $\alpha = 0.1$. In DGP4, two five percent mixture components are simulated to contaminate the dataset instead. The first component mirrors the mixture component from DGP3. The second component has slightly higher leverage and a coefficient which is further increased. This process induces two sets of influential observations which have the potential to mask each other. A formal description of the DGPs is given in the caption of Figure 3.

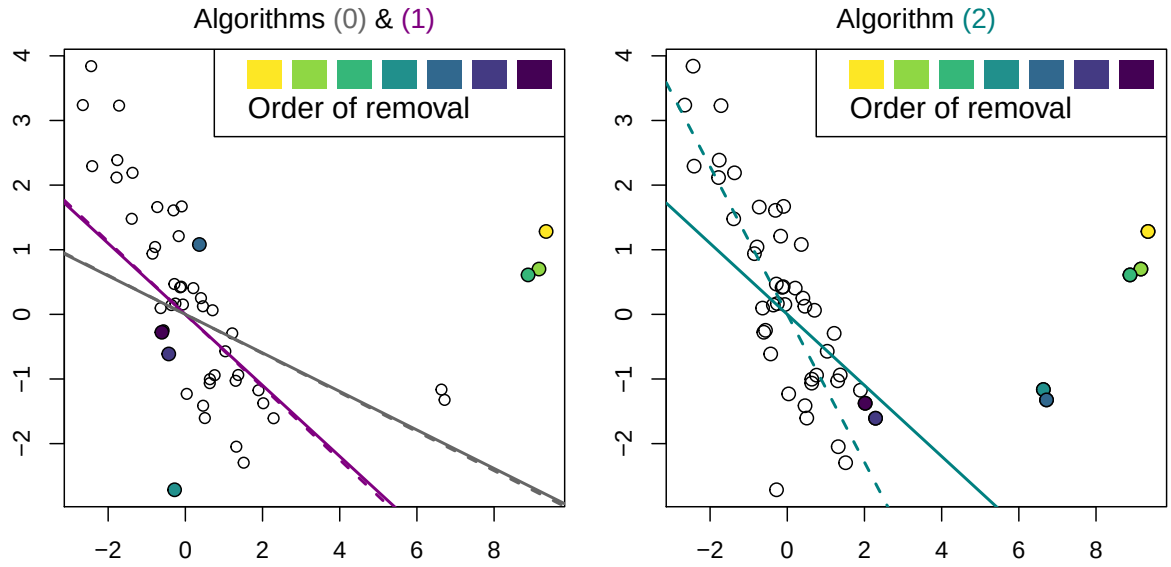

Figure 2: Influential set identification with Algorithms 0 and 1 (left panel, identical sets) and Algorithm 2 (right panel). Regression lines with the approximated maximal perturbations for sets of size three and seven are indicated as solid and dashed lines, respectively. Lines for Algorithm 0 are held in gray, Algorithm 1 in purple, and Algorithm 2 in teal.

We simulate 1,000 data sets for $N = 100$ and a single explanatory variable and collect the perturbations detected by Algorithms 0, 1, and 2.[10] The average coefficient paths as influential sets are removed are reported in Figure 3. It can be seen that Algorithm 1 improves upon the coefficient paths obtained by Algorithm 0 due to the computation of exact perturbations. Algorithm 2 dominates the other two alternative methods whenever more than one observation is removed, and its relative performance increases with the number of influential observations. This result extends to statistics based on individual runs for the simulation setting presented (see Appendix, Table A1).

In the top-left panel of Figure 3, with data based on DGP1, we observe similar performance of Algorithms 1 and 2, while Algorithm 0 lags behind slightly. When introducing high leverage observations in DGP2 (top-right panel), the gap between search algorithms based on initial and adaptive influential set identification schemes widens. This can be explained by the fact that Algorithms 0 and 1 are not able to account for masking of influential observations as others are removed.

In the bottom-left panel, with data based on DGP3, the problems created by masking of influential observations are even more striking. The initial approximation employed

[10]Other simulation settings for model size $P = 5$, where a single covariate is distorted, and with influential sets of reduced size are qualitatively similar. Results are presented in the Appendix (see Figures A2 and A3). Additional simulation designs for $N = \{100, 500, 1000\}$, reduced influential sets corresponding to $\alpha = 0.05$, and different number of covariates were also implemented and led to qualitatively similar results. These results are available from the authors upon request.

by Algorithms 0 and 1 suffers from masking problems, even for observations within the influential set, which is not identified consistently. The results presented in this panel also highlight the fact that the total perturbation induced by the set exceeds the sum of perturbations induced by its parts (as would be implied by initial estimates). This leads to a characteristic break point in the coefficient paths when ten observations (the true number of highly influential observations) are removed for Algorithms 1 and 2. This break is absent for Algorithm 0, which coincides with a considerable underestimation of the magnitude of the perturbation.

These effects are even more pronounced under the DGP4, for which results are presented in the bottom-right panel. Algorithm 0 identifies the first influential set but vastly underestimates its induced perturbation. Algorithm 1 reliably finds the first influential set and yields good estimates of the size of its attached perturbation, but fails to identify the second set accurately. Algorithm 2 correctly identifies both sets, as can be recognised by the distinct breaks at influential set sizes of 5 and 10.

4.2 Seven microfinance studies

Much work in development economics has aimed at quantifying the impact of microfinance on poverty in developing countries. While a general interest in microfinance as a development policy tool has a relatively long history in economics (Morduch, 1999), large-scale studies based on experimental designs have become available only recently. Meager (2019) discusses and summarises seven of these studies in the context of an external validity assessment. These studies analyse data from RCTs that took place in Bosnia & Herzegovina (Augsburg et al., 2015), Ethiopia (Tarozzi et al., 2015), India (Banerjee et al., 2015), Mexico (Angelucci et al., 2015), Mongolia (Attanasio et al., 2015), Morocco (Crépon et al., 2015), and the Philippines (Karlan and Zinman, 2011). BGM assess these studies and focus on robustness of the average treatment effect (ATE) estimates when excluding a small number of influential observations. In general, their findings imply that the effect of microcredit on household level business profits is not particularly robust and that the size of ATE estimates are usually driven by very small sets of observations.

We address the problem of assessing perturbations in ATE estimates in these studies using the replication data accompanying the study of Meager (2019). The model considered is a simple treatment effect model of the form

$$y_i = \alpha + D_i\beta + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \tag{6}$$

where y_i corresponds to household business profits in household $i = 1, \ldots, N$ and D_i is a randomised treatment dummy indicating whether household i has been assigned to the

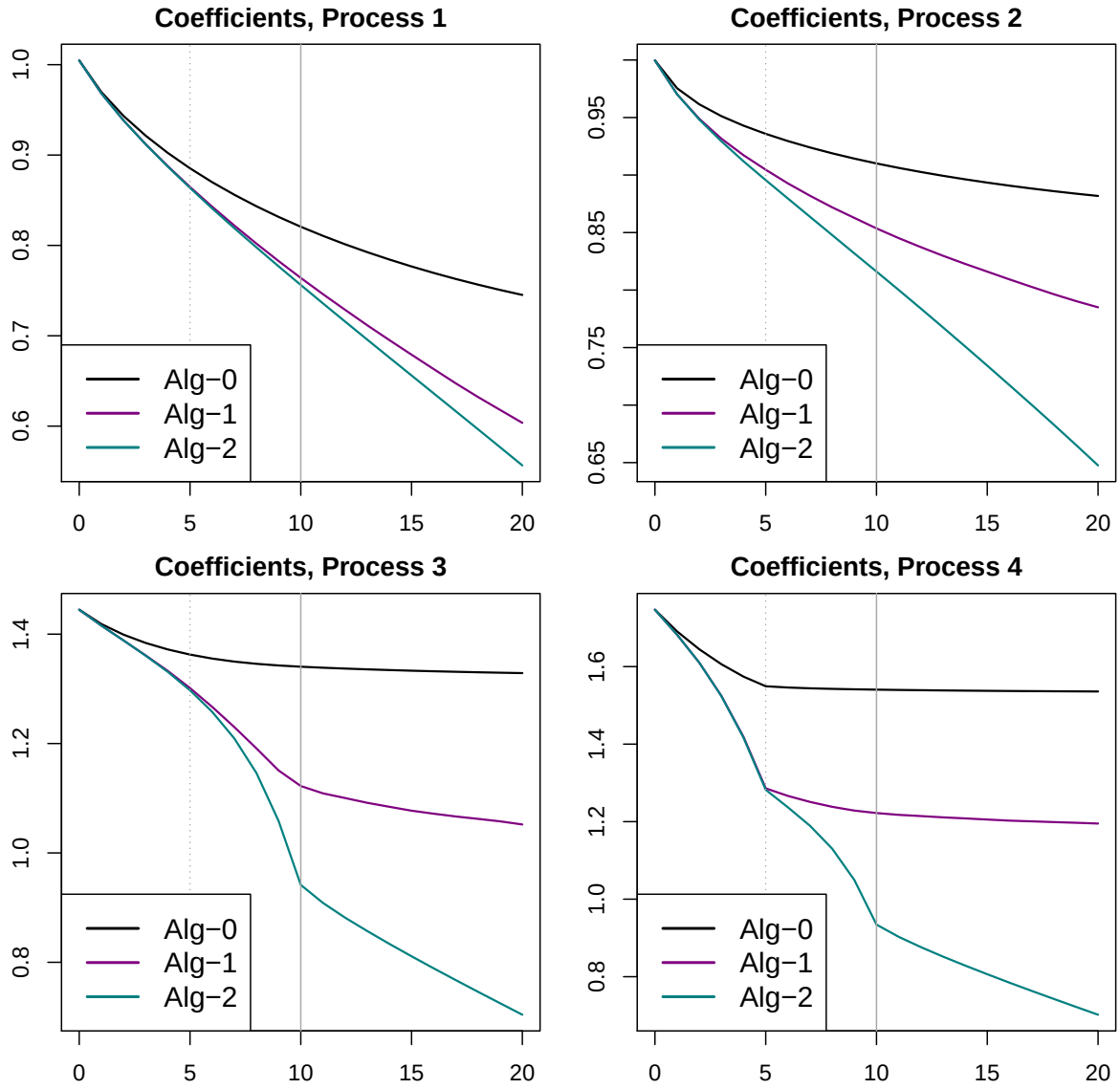

Figure 3: Average coefficient paths along the number of observations removed, calculated using Algorithms 0, 1, and 2. Results are based on averages across 1000 simulation runs in four different settings with $N = 100$. Process 1 is a linear model with $\beta = \sigma^2 = 1$, where the only control is drawn from $\mathcal{N}(0, 1)$. In process 2, ten observations of the covariate are drawn from $\mathcal{N}(0, 5)$ instead. For process 3, a set of ten observations has high leverage (being drawn from $\mathcal{N}(7, 1)$) and a different coefficient ($\beta^a \sim \mathcal{N}(1.5, 0.5)$). Process 4 splits this set of ten observations into two, one of which has even higher leverage (being drawn from $\mathcal{N}(10, 1)$) and a different coefficient ($\beta^b \sim \mathcal{N}(2, 0.5)$).

treatment group that benefits from easier access to microfinance compared to the control group. Under assumptions that are standard in experimental contexts, β measures the ATE. This simple model is estimated separately for each dataset, with sample sizes that range from around 1,000 to around 16,500 observations. To ensure comparability to the

Study region	BIH		MON		ETH		MEX		MOR		PHI		IND	
Algorithm	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)
Sign-switch	14	13	16	15	1	1	1	1	11	11	9	9	6	6
Significant (99%)	49	39	43	37	117	13	20	12	35	33	74	54	41	35
Sample size	1,195		961		3,113		16,560		5,498		1,113		6,863	

Table 1: Number of observations needed to induce a sign-switch and a significant sign-switch as well as sample size for each of the seven microfinance RCTs. BIH = Bosnia & Herzegovina, MON = Mongolia, ETH = Ethiopia, MEX = Mexico, MOR = Morocco, PHI = Philippines, IND = India.

results in BGM and Meager (2019), control variables and fixed effects are omitted from the econometric specification.[11]

Table 1 presents the number of observations that suffices to reach target perturbations to the full sample results when excluded. Specifically, we report the number of observations necessary to switch the sign of the point estimate $\hat{\beta}$ and to switch the sign and achieve a significant estimate of $\hat{\beta}$ at the 1% significance level. For the sake of brevity, we focus on comparing Algorithm 0 to Algorithm 2 and omit the results of Algorithm 1, which are close to those of Algorithm 2 in this case.

The first take-away from Table 1 is that a relatively small number of observations drive the full-sample estimate of the ATE, a result which is in line with the findings of BGM. In the studies based on data for Ethiopia and Mexico, the exclusion of a single observation is enough to change the sign of $\hat{\beta}$. In addition, for the data from Mexico, dropping around 0.07% of the observations leads to a change in the sign of the ATE and a resulting parameter which is statistically significant. The second insight from Table 1 is that Algorithm 2 is more efficient in detecting influential sets compared to Algorithm 0, in line with the simulation studies outlined above. Focusing on the data for Ethiopia, Algorithm 0 suggests that the exclusion of 117 observations (3.8% of the data set) changes the sign of the ATE estimate and leads to statistical significance with an opposite sign of the estimate obtained with the full sample. Using Algorithm 2 reveals that a much smaller set of 13 observations (0.4% of the data set) is enough to achieve this level of perturbation. Algorithm 2 performs particularly well in scenarios where highly influential observations mask the presence of other influential observations.

Note that, if the possibility of measurement error can be ruled out, small sets of influential observations driving the ATE may actually imply the presence of heterogeneous

[11] As argued in BGM and Meager (2019), excluding the set of controls has only minor effects on inference in these studies.

treatment effects. A small number of observations driving the treatment effect is equivalent to a small share of the population reacting to the treatment. Assuming that 1% of the population reacts to the treatment, excluding merely ten observations is enough to overthrow the full sample results when $N = 1000$. In the context of microcredit, this is in line with previous literature that has shown that households with previous business experience are more sensitive to microcredit interventions when compared to the average household (Meager, 2019; Crépon et al., 2015). A key finding of Meager (2020) is that the effects of microcredit interventions are most likely zero for a large proportion of the population and that significant (positive) effects are concentrated in the right tails of the outcome distribution. Some of the influential sets identified are far below 1% or even 0.1% of the sample size. From a policy perspective, this may cast doubt on the effectiveness of microcredit as a tool for poverty reduction and inclusive development. From an econometrician's viewpoint, this implies that aggregating evidence from many (potentially underpowered) studies may be necessary to draw robust conclusions on the effect of interest. Evidence aggregation is attempted in Meager (2019) and Meager (2020), where Bayesian hierarchical models are used to combine effect estimates of the seven studies discussed above.

4.3 Poverty convergence

A large number of empirical studies assess the patterns of cross-country convergence in living standards, as measured by GDP per capita (Barro and Sala-i Martin, 1992; Johnson and Papageorgiou, 2020). Convergence in absolute poverty rates, however, has been examined less often. Ravallion (2012) addresses this question in the theoretical framework given by the combination of two facts: higher average incomes tend to lead to lower poverty rates (Bourguignon, 2003) and mean incomes tend to converge across countries. While these two concepts taken together clearly point into the direction of convergence in poverty rates, Ravallion (2012) is not able to detect poverty convergence in a sample of 89 countries. The proposed econometric specification takes the form

$$T_i^{-1}(\ln H_{it} - \ln H_{it-1}) = \alpha + \beta \ln H_{it-1} + \varepsilon_{it}, \tag{7}$$

where H_{it} denotes the poverty headcount ratio in country i in time period t, T_i is the country-specific observation period in years and ε_{it} is an error term assumed to fulfil the standard assumptions of the linear regression model. This specification relates the annualised growth rate of the poverty headcount ratio to the log of the initial poverty headcount index. Ravallion (2012) estimates β and finds it to be slightly positive and statistically insignificant.

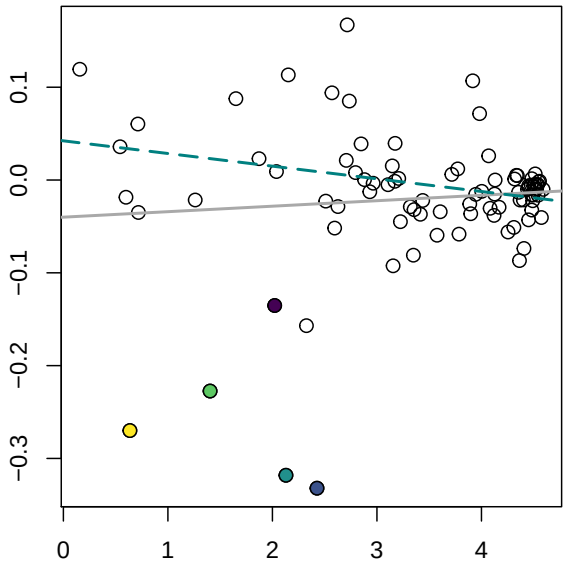

Figure 4: Data and regression line for Ravallion (2012) before (solid line) and after (dashed line) removing the minimal perturbing set $\hat{\mathcal{S}}_5^*$ (highlighted in colour). The horizontal axis holds the logarithm of initial poverty headcount index; the vertical axis annualised log differences of poverty headcount ratios.

In a response, Crespo Cuaresma et al. (2016) point out that the original, non-significant finding is likely due to a number of Eastern European countries exhibiting very low initial poverty headcount ratios. The log-transformation in Equation (7) implies that small absolute changes translate into large growth rates in poverty headcount ratios for these economies. This makes the experience of these countries very influential when estimating the parameters in Equation (7). Crespo Cuaresma et al. (2016) show that when explicitly controlling for the poverty trajectories of Eastern European countries, there is indeed empirical evidence for cross-country convergence in poverty rates.

We first revisit the problem using the same data set as used in both Ravallion (2012) and Crespo Cuaresma et al. (2016). Figure 4 presents the convergence scatter plot, with coloured observations for the countries that belong to the minimal perturbing set as identified by Algorithm 2. It suffices to remove only five countries from the data set to achieve statistically significant poverty convergence (a negative and significant estimate of β) using the specification given by Equation (7). These five countries are Belarus, Latvia, Ukraine, Poland, and the Russian Federation. This result stresses the need to take into account the different experience of Eastern European countries when analysing cross-country poverty dynamics.

In an additional exercise, we investigate an alternative specification suggested in Crespo Cuaresma et al. (2016) that takes the form

$$T_i^{-1}(H_{it} - H_{it-1}) = \alpha + \beta H_{it-1} + \varepsilon_{it} \tag{8}$$

where variable definitions are the same as in Equation 7. This econometric specification is based on the concept of a semi-elastic relationship between poverty reduction and economic growth from Klasen and Misselhorn (2008). It relates changes in poverty headcount ratios to the initial level of poverty instead of growth rates in poverty headcount ratios to the initial log level of poverty. Using this alternative specification, Crespo Cuaresma et al. (2016) find clear empirical evidence for poverty convergence using the original data from Ravallion (2012).

To assess how robust this finding is to excluding sets of observations, we re-estimate the specification given by Equation (8) using an updated dataset sourced from PovCalNet.[12] Starting with the full sample of poverty headcount observations (using a poverty line of \$2 a day), a number of data quality filters are applied. First, observations that are not based on household surveys are excluded. Second, countries where the longest observation period is below ten years are excluded. Finally, when both income and consumption based poverty rates are available, consumption based data are preferred. This procedure leaves us with a sample of 124 countries. For each country, the longest time span available is used to compute annualised changes in poverty rates.

The full sample LS estimate $\hat{\beta} = -0.019$ with a standard error of 0.002 implies significant poverty convergence. Algorithm 0 does not detect any set of observations that overthrows this result, even after excluding more than 50% of the sample. Algorithms 1 and 2 detect sets of observations that nullify the significant result when excluded. These sets include 64 (Algorithm 1) and 26 (Algorithm 2) countries. In other words, one has to exclude at least 20% of observations to achieve a sign change in the significantly negative full sample estimate. As a result of these exercises, we conclude that poverty convergence is a relatively robust empirical regularity.

5 Concluding remarks

In this paper, we investigated the sensitivity of inferential statistics in linear regression models to sets of influential observations. We showed how masking issues may hamper naïve approaches to summary measures of sensitivity, which in turn yielded misleading indications of robustness. This was particularly problematic in the actual presence of highly influential sets. We proposed two algorithms that are more useful in such scenarios and outperform existing approaches considerably at little additional cost. These

[12] PovCalNet data can be obtained from http://iresearch.worldbank.org/PovcalNet/home.aspx.

algorithms allow for a more precise assessment of the degree of sensitivity of estimates to sets of influential observations. Intuitive and directly interpretable summary statistics, such as the number of observations needed to achieve a sign-flip or change in significance of coefficients (as done in Broderick et al., 2020), can be easily derived.

While the achievable perturbations may seem alarming in certain applications, interpretation requires careful contextualisation. Sensitivity to removal of observations is decidedly not a conclusive indicator of a lack of validity. This is particularly the case for the relative number of removals – a great deal of interesting phenomena are exceedingly rare and insights hinge on few influential observations, with examples including rare diseases, economic crises or policy interventions that only affect outcomes for a small part of the population. Nevertheless, caution should be exercised if the sensitivity of regression results to influential observations is high. Sensitivity to the omission of very few observations may indicate sampling bias. Assuming that internal validity is given, the sample size is unlikely to accurately represent the population of interest. This may be a sign of an underpowered study and of a lack of external validity.

Several pathways for future work that builds upon our results can be envisaged. There is room for more comprehensive statistics that summarise robustness of inferential statistics to influential sets. Further methodological improvements in terms of computational efficiency and accuracy, for instance via sampling-based approaches, are possible as well. Lastly, performing robustness analysis to other empirical phenomena different from those presented here may deliver valuable insights that are relevant for academics and policy makers.

References

Manuela Angelucci, Dean Karlan, and Jonathan Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by compartamos banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.

Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30, 2010. doi:10.1257/jep.24.2.3.

Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017. doi:10.1257/jep.31.2.3.

Orazio Attanasio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart. The impacts of microfinance: Evidence from joint-liability lending in Mongolia. *American Economic Journal: Applied Economics*, 7(1):90–122, 2015.

Britta Augsburg, Ralph De Haas, Heike Harmgart, and Costas Meghir. The impacts of microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics*, 7(1):183–203, 2015.

Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1):22–53, 2015.

Robert J Barro and Xavier Sala-i Martin. Convergence. *Journal of Political Economy*, 100(2): 223–251, 1992.

David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 1980. doi:10.1002/0471725153.

François Bourguignon. The growth elasticity of poverty reduction: explaining heterogeneity across countries and time periods. *Inequality and growth: Theory and policy implications*, 1 (1), 2003.

George EP Box and George C Tiao. A Bayesian approach to some outlier problems. *Biometrika*, 55(1):119–129, 1968. doi:10.1093/biomet/55.1.119.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*, 2020.

Chrys Caroni and Philip Prescott. Sequential application of wilks's multivariate outlier test. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):355–364, 1992.

Samprit Chatterjee and Ali S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986. doi:10.1214/ss/1177013622.

Ralph Dennis Cook. Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174, 1979. doi:10.2307/2286747.

Ralph Dennis Cook and Sanford Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

Bruno Crépon, Florencia Devoto, Esther Duflo, and William Parienté. Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in morocco. *American Economic Journal: Applied Economics*, 7(1):123–50, 2015.

Jesús Crespo Cuaresma, Stephan Klasen, and Konstantin M Wacker. There is poverty convergence. *Available at SSRN 2718720*, 2016.

Bradley Efron and Robert J Tibshirani. *An introduction to the Bootstrap*. CRC Press, 1994.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933. doi:10.2307/1907330.

Ali S. Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):761–771, 1992. doi:10.1111/j.2517-6161.1992.tb01449.x.

Sven Hammarling and Craig Lucas. Updating the QR factorization and the least squares problem, 2008. URL `http://eprints.maths.manchester.ac.uk/id/eprint/1192`.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi:10.1080/01621459.1974.10482962.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: The approach based on influence functions*, volume 196. John Wiley & Sons, 2005. doi:10.1002/9781118186435.

Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 430–433. IEEE, 2004. doi:10.1109/ICPR.2004.1334558.

Jennifer Hoeting, Adrian E Raftery, and David Madigan. A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, 22(3):251–270, 1996. doi:10.1016/0167-9473(95)00053-4.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi:10.1214/aoms/1177703732.

Paul Johnson and Chris Papageorgiou. What remains of cross-country convergence? *Journal of Economic Literature*, 58(1):129–75, 2020.

Dean Karlan and Jonathan Zinman. Microcredit in theory and practice: Using randomized credit scoring for impact evaluation. *Science*, 332(6035):1278–1284, 2011.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: An introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

Sung-Soo Kim and Wojtek J Krzanowski. Detecting multiple outliers in linear regression using a cluster method combined with graphical visualization. *Computational Statistics*, 22(1): 109–119, 2007.

Sung-Soo Kim, Sung H Park, and WJ Krzanowski. Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model. *Journal of Applied Statistics*, 35(3):283–291, 2008.

Stephan Klasen and Mark Misselhorn. Determinants of the growth semi-elasticity of poverty reduction. Technical report, IAI Discussion Papers, 2008.

Edward E Leamer. Let's take the con out of econometrics. *American Economic Review*, 73(1): 31–43, 1983. URL https://www.jstor.org/stable/1803924.

Edward E Leamer. Tantalus on the road to asymptopia. *Journal of Economic Perspectives*, 24 (2):31–46, 2010. doi:10.1257/jep.24.2.31.

Ross Levine and David Renelt. A sensitivity analysis of cross-country growth regressions. *American Economic Review*, pages 942–963, 1992. URL https://www.jstor.org/stable/2117352.

Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: Theory and methods (with R)*. John Wiley & Sons, 2019. doi:10.1002/9781119214656.

Rachael Meager. Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, 2019.

Rachael Meager. Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. *Manuscript: MIT*, 2020.

Jonathan Morduch. The microfinance promise. *Journal of Economic Literature*, 37(4): 1569–1614, 1999.

Lawrenece I Pettit and Karen DS Young. Measuring the effect of observations on Bayes factors. *Biometrika*, 77(3):455–466, 1990. doi:10.1093/biomet/77.3.455.

Martin Ravallion. Why don't we see poverty convergence? *American Economic Review*, 102 (1):504–23, 2012.

Lothar Reichel and William B Gragg. Algorithm 686: FORTRAN subroutines for updating the QR decomposition. *ACM Transactions on Mathematical Software (TOMS)*, 16(4):369–377, 1990. doi:10.1145/98267.98291.

Peter Rousseeuw and Victor Yohai. Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer, 1984. doi:10.1007/978-1-4615-7821-5_15.

Xavier Sala-i Martin, Gernot Doppelhofer, and Ronald I Miller. Determinants of long-term growth: A Bayesian averaging of classical estimates (bace) approach. *American Economic Review*, pages 813–835, 2004. doi:10.1257/0002828042002570.

Xavier X Sala-i Martin. I just ran two million regressions. *The American Economic Review*, pages 178–183, 1997. doi:10.3386/w6252.

Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1): 124–127, 1950. doi:10.1214/aoms/1177729893.

Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. *Carnegie-Mellon University. Department of Computer Science*, 1994.

Matthew S Shotwell, Elizabeth H Slate, et al. Bayesian outlier detection with dirichlet process mixtures. *Bayesian Analysis*, 6(4):665–690, 2011. doi:10.1214/11-BA625.

Christopher A Sims. Macroeconomics and reality. *Econometrica*, pages 1–48, 1980. doi:10.2307/1912017.

Mark FJ Steel. Model averaging and its use in economics. *Journal of Economic Literature*, 58 (3):644–719, 2020.

Alessandro Tarozzi, Jaikishan Desai, and Kristin Johnson. The impacts of microcredit: Evidence from ethiopia. *American Economic Journal: Applied Economics*, 7(1):54–89, 2015.

Lloyd N Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997. doi:10.1137/1.9780898719574.

Isabella Verdinelli and Larry Wasserman. Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing*, 1(2):105–117, 1991.

Victor J Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, pages 642–656, 1987. doi:10.1214/aos/1176350366.

A Additional tables and figures

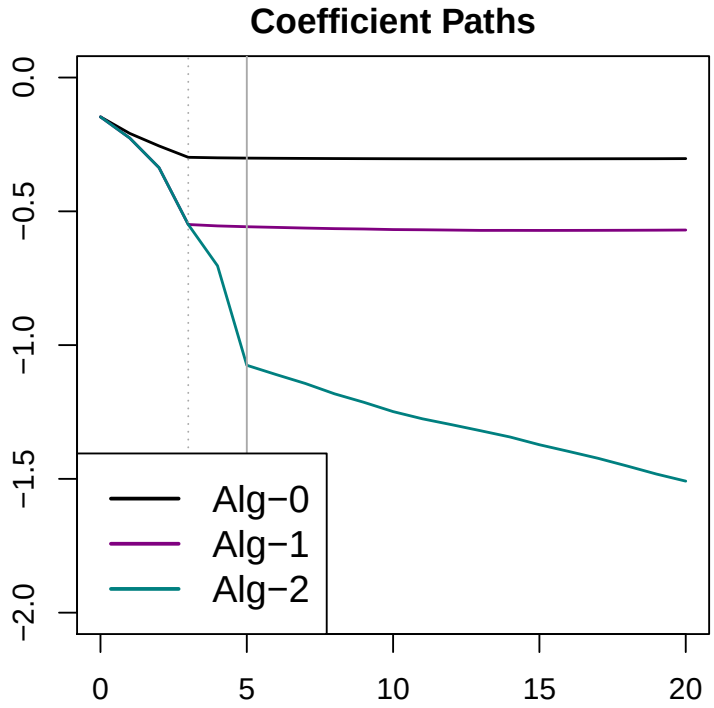

Figure A1: Coefficient paths for the data in Figure 1 obtained from Algorithms 0, 1, and 2.

Process	Algorithm	2 removed	5 removed	10 removed	20 removed
DGP1	A (2)	0.064	0.427	0.897	0.999
	A (1) & A (2)	0.936	0.573	0.103	0.001
	A (0) or A (1)	0	0	0	0
DGP2	A(2)	0.176	0.715	0.968	1.000
	A (1) & A (2)	0.824	0.285	0.032	0.000
	A (0) or A (1)	0	0	0	0
DGP3	A (2)	0.079	0.543	0.948	1.000
	A (1) & A (2)	0.921	0.457	0.052	0.000
	A (0) or A (1)	0	0	0	0
DGP4	A (2)	0.086	0.226	0.996	1.000
	A (1) & A (2)	0.914	0.774	0.004	0.000
	A (0) or A (1)	0	0	0	0

Table A1: Performance measured by the percentage of maximal perturbations between algorithms achieved at $\{2, 5, 10, 20\}$ removed observations over 1000 simulations. As can be seen, Algorithm 2 strictly dominates Algorithm 0 and weakly dominates Algorithm 1 in the simulations when more than one observation is removed. Processes are the same as in Figure 3.

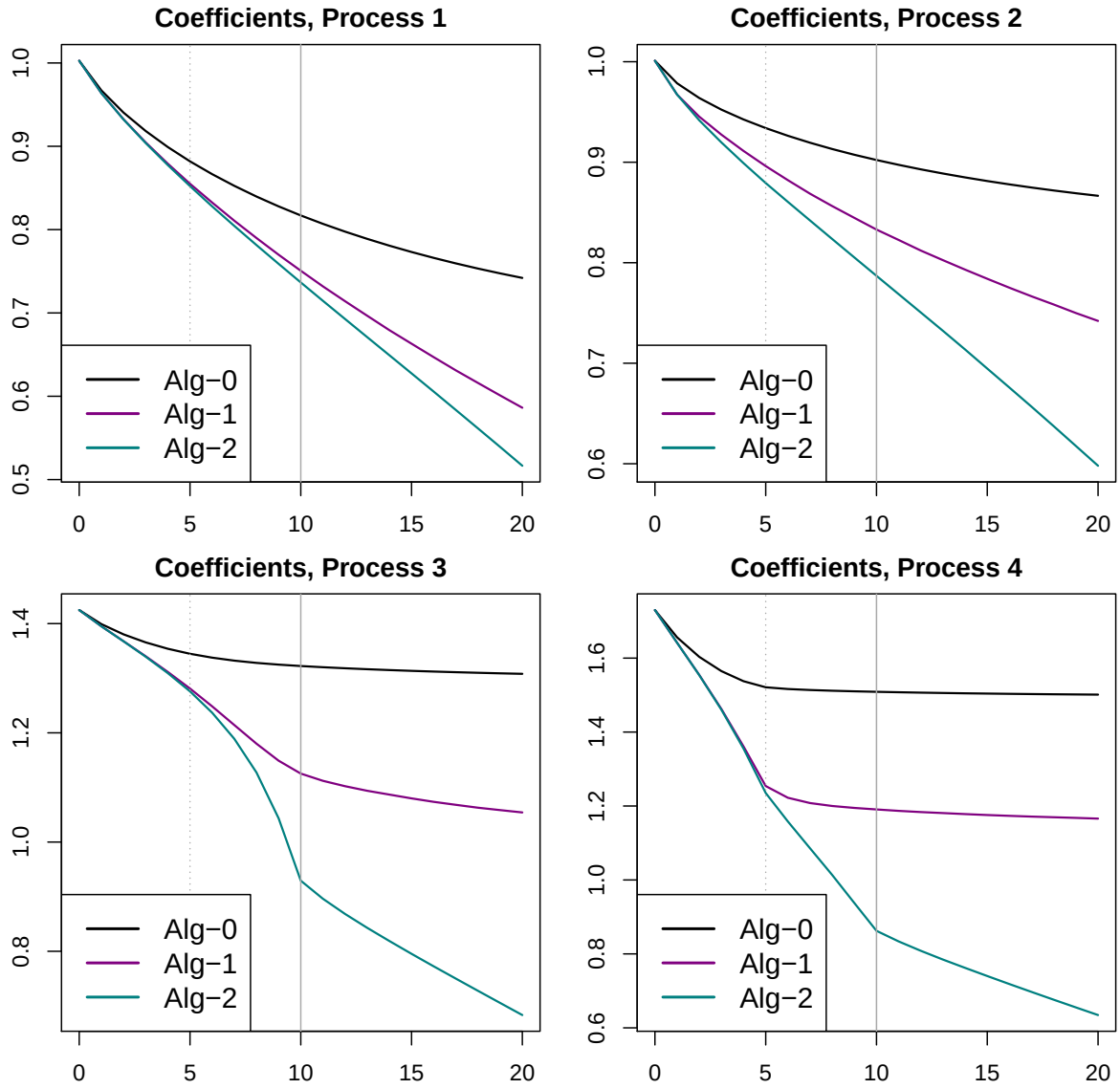

Figure A2: Average coefficient paths for four different DGPs over a thousand simulations calculated with Algorithms 0, 1, and 2. Processes are the same as in Figure 3, but the simulation includes four additional variables, where both data and coefficients drawn from a standard normal.

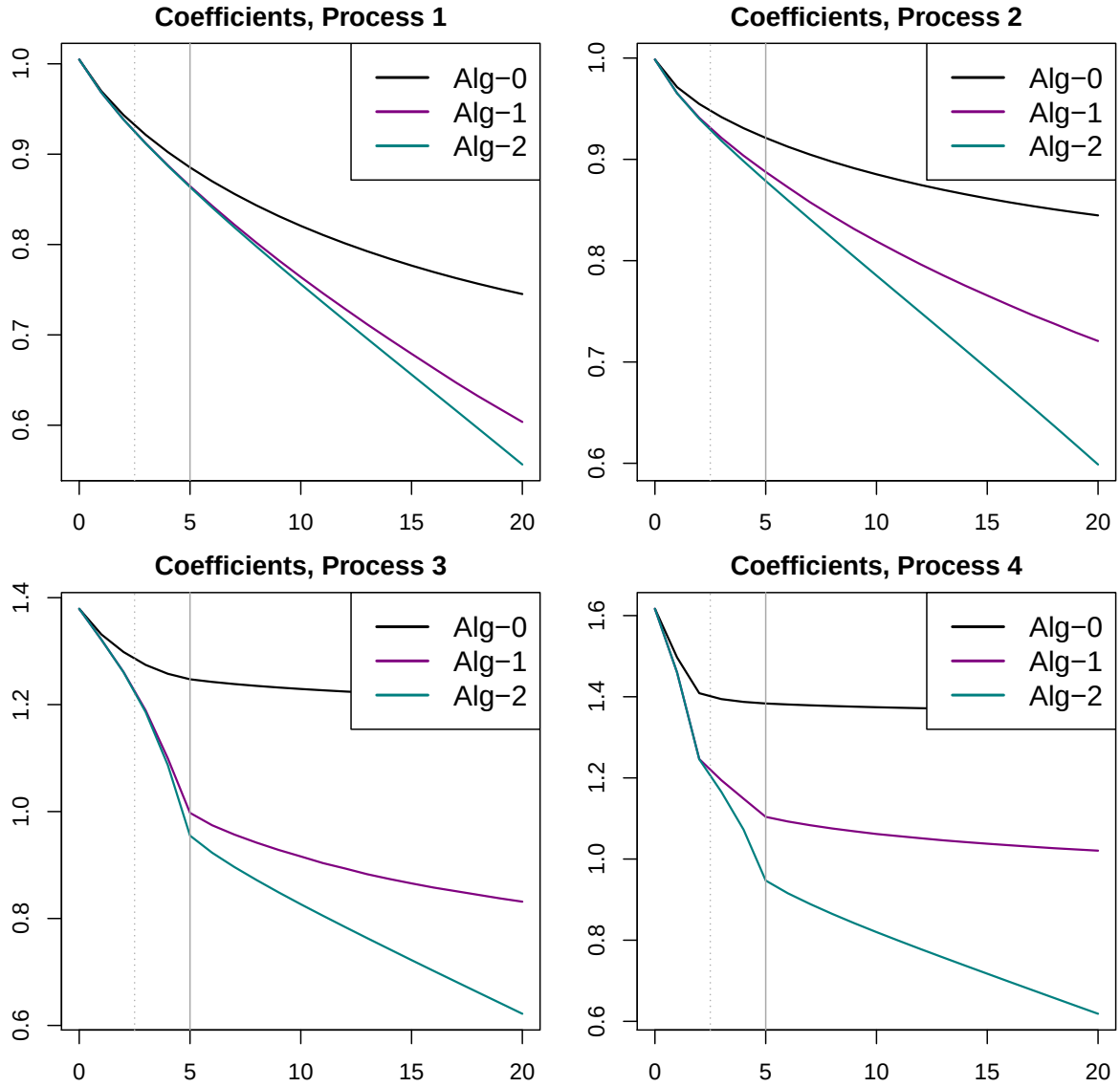

Figure A3: Average coefficient paths for four different DGPs over a thousand simulations calculated with Algorithms 0, 1, and 2. Processes are the same as in Figure 3, but the cardinality of changed sets is adapted. Process 2 and 3 now include five observations of particular interest. The mixture in process 4 is now composed of three observations mirroring process 3 and two additional ones.

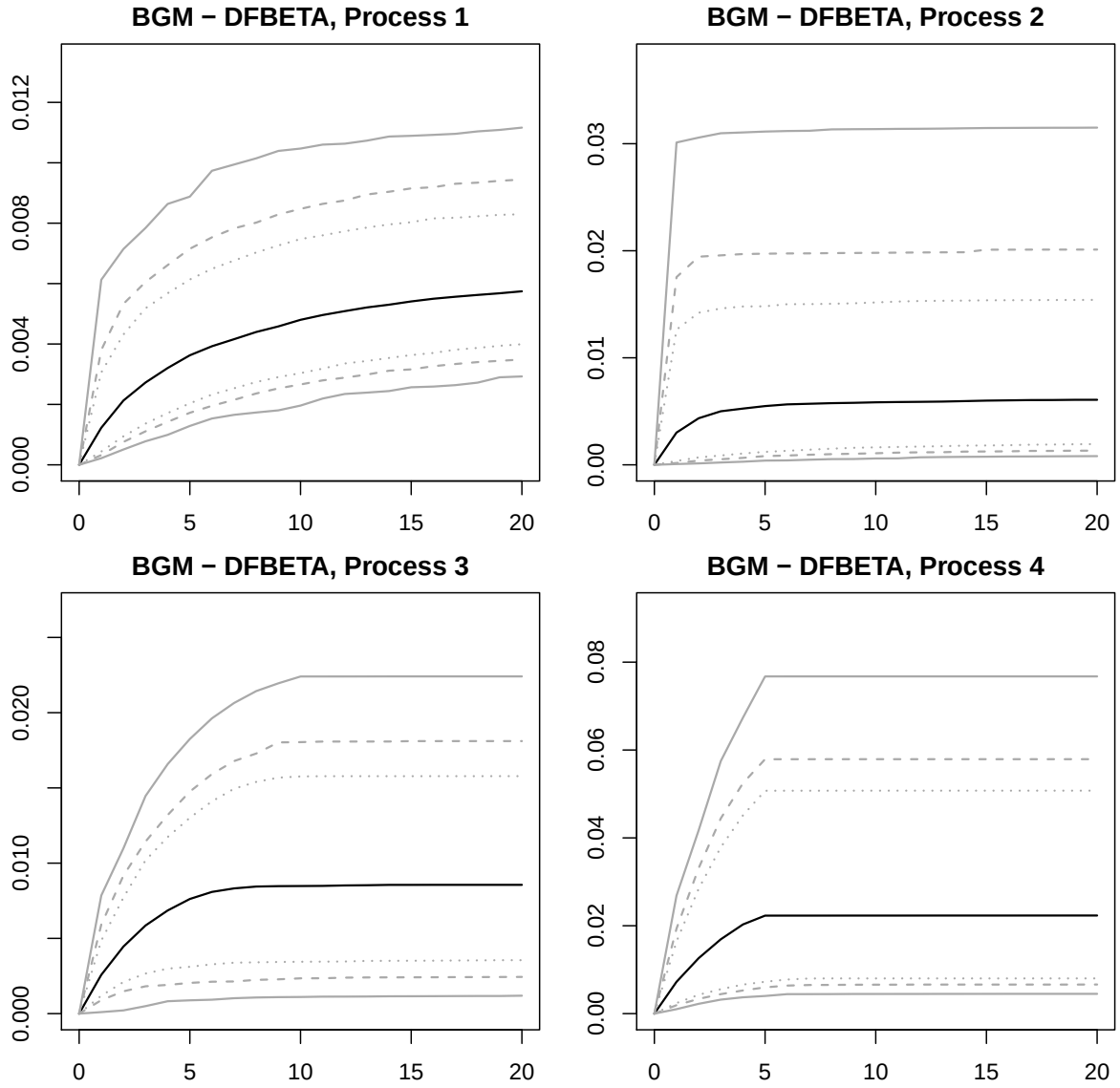

Figure A4: Difference between perturbed coefficients retrieved with Algorithm 0 using the BGM implementation and DFBETA (see Equation (5) for the measure) as measures of influence. Plotted are the median in black as well as 1%, 5%, and 10% quantiles of the difference for four different DGPs over a thousand simulations. Processes are the same as in Figure 3. Note the small but consistent deviations. The BGM approximation already falls behind slightly on the first removal, for which DFBETA is exact.