Cross-Validating Synthetic Controls

Martin Becker, Saarland University

Stefan Klößner, Saarland University*

Gregor Pfeifer, University of Hohenheim

October 12, 2017

Abstract

While the literature on synthetic control methods mostly abstracts from out-of-sample measures, Abadie et al. (2015) have recently introduced a cross-validation approach. This technique, however, is not well-defined since it hinges on predictor weights which are not uniquely defined. We fix this issue, proposing a new, well-defined cross-validation technique, which we apply to the original Abadie et al. (2015) data. Additionally, we discuss how this new technique can be used for comparing different specifications based on out-of-sample measures, avoiding the danger of cherry-picking.

JEL Codes: C52, C22

Keywords: Synthetic Control Methods; Cross-Validation; Specification Search.

*Corresponding author: Stefan Klößner, Statistics & Econometrics, Saarland University, C3 1, D-66123, Germany; S.Kloessner@mx.uni-saarland.de; Phone +49 681 302 3179.

1 Introduction

Abadie and Gardeazabal (2003) and Abadie et al. (2010) have introduced synthetic control methods (SCM) to estimate a treated unit's development in absence of the treatment. These methods have gained a lot of popularity among applied researchers, Athey and Imbens (2017) even state that SCM "is arguably the most important innovation in the policy evaluation literature in the last 15 years". Recently, SCM have been shown to perform well against certain panel-based approaches (see Gardeazabal and Vega-Bayo, 2017), and have also been used for forecasting (see Klößner and Pfeifer, 2017).

The basic idea of SCM is to find suitable donor weights describing how the treated unit is 'synthesized' by a weighted mix of unaffected control units, building a counterfactual. Treated and synthetic unit should resemble each other as closely as possible prior to the treatment, both with respect to the outcome of interest and economic predictors. The latter are variables with predictive power for explaining the outcome. The SCM approach searches for optimal predictor weights in order to grant more importance to variables with better predictive power.

However, SCM operate merely in-sample, making it difficult to assess the counterfactual's validity. To mitigate this problem, Abadie et al. (2015) (henceforth: ADH) have expanded SCM, incorporating *cross-validation*. The pre-treatment timespan is divided into a training and a validation period, and predictor weights are selected by minimizing the out-of-sample error in the validation period. However, Klößner et al. (2017) (henceforth: KKPS) have recently shown that there is a misconception of the ADH cross-validation technique. In applications, there often exist many different solutions that minimize the out-of-sample error, rendering this technique not well-defined since predictor weights are not uniquely defined.

We fix this problem by defining unique predictor weights following two principles. Special predictors like lagged values of the outcome variable(s) are guaranteed to obtain certain minimum weights, and predictors in general shall not become irrelevant due to weights accidentally obtained too small. Applying this new cross-validation technique to ADH's original data, we exemplarily show that ADH's main finding is confirmed, while corresponding placebo exercises deliver diverging results, questioning the robustness of the main takeaway.

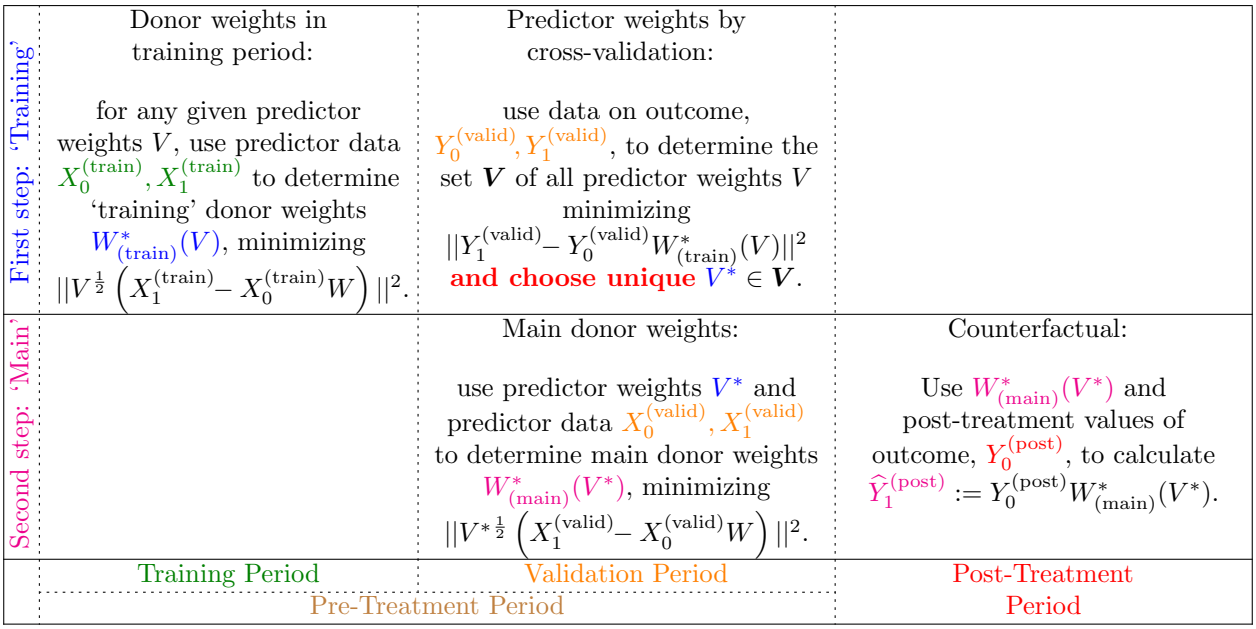

Figure 1: Schematic Overview of Cross-Validation Technique. This is a refined version of KKPS's Figure 1.

2 SCM and Cross-Validation

For the synthetic control method, one uses two types of data—the variable of interest, Y, and predictor variables, X. The latter consist of M linear combinations of pre-treatment values of Y as well as r other covariates with explanatory power for Y. Both Y and X are considered for a treated unit and for so-called donors, i.e., non-treated units, denoted by subscripts "$_1$" and "$_0$", respectively. In the example discussed by ADH and KKPS and also throughout this paper, the treated unit is the 1990 reunified Germany, while ($J = 16$) Western OECD countries serve as donors. The variable of interest (Y) is GDP per capita, and the $k = M + r$ predictors (X) are the average of lagged GDP values ($M = 1$) and the covariates trade openness, inflation rate, industry share of value added, amount of schooling, and the investment rate ($r = 5$).[1]

The corresponding quantities as well as a schematic overview of the cross-validation technique are provided in Figure 1. The SCM cross-validation approach consists of two steps. During the first step, called 'training', so-called predictor weights V^* are determined using cross-validation techniques, during the second step, these weights V^* are used to estimate the variable of interest's counterfactual development in absence of the treatment. For determining V^*, the training step decomposes the pre-treatment timespan (1971-1990) into a training (1971-1980) and a validation period (1981-1990). In the training period, one makes use of the

[1]For more details on variables as well as donor choice, see Abadie et al. (2015, p. 509).

$k \times J$ matrix $X_0^{(\text{train})}$ and the k-dimensional vector $X_1^{(\text{train})}$, containing time averages of the predictors' data for the donor units and the treated unit, respectively. Given these, one considers, for any given positive predictor weights $V = (v_1, \ldots, v_k)' \in \mathbb{R}^k_{++}$, the so-called training weights $W^*_{(\text{train})}(V) \in \mathbb{R}^J$, which are defined as the solution of

$$\min_W ||V^{\frac{1}{2}} \left(X_1^{(\text{train})} - X_0^{(\text{train})} W \right) ||^2 = \min_W \sum_{m=1}^{k} v_m \left(X_{1m}^{(\text{train})} - X_{0m}^{(\text{train})} W \right)^2 \quad \text{s.t. } W \geq 0, \mathbb{1}'W = 1, \tag{1}$$

where $V^{\frac{1}{2}}$ is the k-dimensional diagonal matrix with the roots of V's elements on the diagonal, while $X_{1m}^{(\text{train})}$ and $X_{0m}^{(\text{train})}$ denote the m-th component and row of $X_1^{(\text{train})}$ and $X_0^{(\text{train})}$, respectively, and $\mathbb{1}$ is the vector of ones. The training weights $W^*_{(\text{train})}(V)$ describe to what extent each donor country is used during the training step to produce a 'synthetic', i.e., counterfactual, Germany, given that the predictors are weighted according to V. As the training weights $W^*_{(\text{train})}(V)$ depend on the predictor weights V, one aims at finding those predictor weights V^* that produce the best forecast. This is done in the second part of the training step, making use of the data in the validation period, the $L \times J$ matrix $Y_0^{(\text{valid})}$ and the L-dimensional vector $Y_1^{(\text{valid})}$, containing the variable of interest's data for the validation period. In particular, ADH define predictor weights $V^* = (v_1^*, \ldots, v_k^*)$ as *the* predictor weights that minimize the out-of-sample error $||Y_1^{(\text{valid})} - Y_0^{(\text{valid})} W^*_{(\text{train})}(V)||^2$ over V, i.e., V^* is *the* solution of

$$\min_V ||Y_1^{(\text{valid})} - Y_0^{(\text{valid})} W^*_{(\text{train})}(V)||^2 \text{ s.t. } V \geq 0, \mathbb{1}'V = 1, \tag{2}$$

where the predictor weights, without loss of generality, have been normalized to sum to unity.

After V^* has been determined, one proceeds to the 'main' step, calculating the donor weights $W^*_{(\text{main})}(V^*)$ as the minimizer of

$$\min_W \sum_{m=1}^{k} v_m^* \left(X_{1m}^{(\text{valid})} - X_{0m}^{(\text{valid})} W \right)^2 \text{ s.t. } W \geq 0, \mathbb{1}'W = 1, \tag{3}$$

where the $k \times J$ matrix $X_0^{(\text{valid})}$ and the k-dimensional vector $X_1^{(\text{valid})}$ contain the predictors' data for the validation period. Eventually, counterfactual values $\widehat{Y}_1^{(\text{post})}$ for comparing with actual values $Y_1^{(\text{post})}$ are given by $Y_0^{(\text{post})} W^*_{(\text{main})}(V^*)$, where $Y_0^{(\text{post})}$ contains the donors' post-treatment outcome data.

However, KKPS show that this approach often leads to ambiguous counterfactual values because V^* is not well-defined due to Equation (2) not having a *unique* solution, but many different minimizers:

$$\boldsymbol{V} := \{V : V \text{ is a minimizer of Equation (2)}\} \tag{4}$$

denotes the corresponding set of minimizers, which often is not a singleton. Thus, in order to come up with a well-defined cross-validation technique, it is necessary to single out one specific, uniquely defined element of $\boldsymbol{V}$. In order to do so, we first prevent predictors related with the outcome from being attributed too small predictor weights, as otherwise the dependent variable may be fitted very poorly in the 'main' step. Therefore, we restrict $\boldsymbol{V}$ to

$$\widetilde{\boldsymbol{V}} := \left\{ V \in \boldsymbol{V} : \frac{\frac{1}{M}\sum_{m=1}^{M} v_m}{\max_{m=1,\ldots,k} v_m} \geq \frac{1}{2} \max_{\widetilde{v} \in \boldsymbol{V}} \frac{\frac{1}{M}\sum_{m=1}^{M} \widetilde{v}_m}{\max_{m=1,\ldots,k} \widetilde{v}_m} \right\},$$

ensuring that the outcome-related predictors' relative importance must not fall below half of what it could maximally be. Second, no economic predictor should accidentally become essentially irrelevant due to an extremely small relative weight. Thus, within $\widetilde{\boldsymbol{V}}$, the ratio of the smallest over the largest predictor weight should be as large as possible, i.e., $\frac{\min_{m=1,\ldots,k} v_m}{\max_{m=1,\ldots,k} v_m}$ should be maximal within $\widetilde{\boldsymbol{V}}$. If there exists more than one element of $\widetilde{\boldsymbol{V}}$ with that property, we can among those choose the one for which the ratio $\frac{v_{(2)}}{\max_{m=1,\ldots,k} v_m}$ of the second-smallest predictor weight, $v_{(2)}$, over the largest predictor weight becomes maximal. If, again, there are several solutions to this maximization problem, we may maximize among those the ratio $\frac{v_{(3)}}{\max_{m=1,\ldots,k} v_m}$, and so on. Proposition 1 in the appendix shows that this procedure results in uniquely defined predictor weights V^*.

3 Estimating the Effect of the German Reunification

Equipped with our new, properly defined cross-validation approach, we now compare the results of ADH and KKPS to the results our new method delivers.[2]

The unique predictor weights delivered by our cross-validation technique are 80.94% for

[2]Calculations were done using the statistics software R-3.3.3 (see R Core Team, 2016) in combination with package MSCMT (see Becker and Klößner, 2017b).

GDP per capita, 5.82% for trade openness, 1.11% for inflation, 1.11% for industry share of value added, 4.77% for amount of schooling, and 6.25% for investment rate. The corresponding estimated gap due to the reunification, the difference between actual and counterfactual values, is displayed in Figure 2 (black, solid line, labeled 'cv'). In line with ADH and KKPS, we find a loss of ca. 3,000 USD. This loss is due to Germany's reunification, as the upper part of Figure 3 reveals, which shows the ratio of root mean squared differences between actual and counterfactual GDP after and before the reunification for a so-called placebo study. Here, each donor country is artificially assigned as the treated unit, while Germany moves to the donor pool. The post-pre-ratio of Germany is much larger than all placebo countries' ratios, indicating that the measured loss in GDP can actually be considered statistically significant.

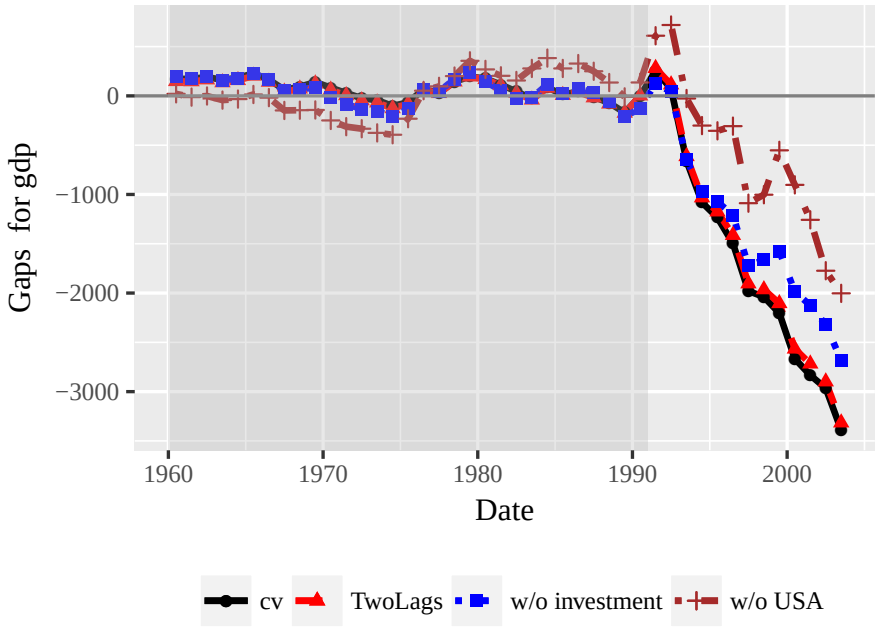

Figure 2: Gaps between GDP per capita of actual and synthetic Germany, estimated using cross-validation with specifications of ADH ('cv'), using last two outcome lags ('TwoLags'), without investment ('w/o investment'), and without U.S. data ('w/o USA').

Our new cross-validation technique is also useful for comparing different specifications according to an objective out-of-sample measure, without the danger of cherry-picking.[3] For instance, instead of using the lagged outcomes' average, one might choose the two most recent lagged outcome values as predictors, as in Montalvo (2011). This specification performs actually slightly better than ADH's specification, with an RMSPE in the validation period of 65.616 compared to 67.678. As Figures 2 and 3 show, results for this specification are very

[3]Ferman et al. (2017) discuss the dangers of cherry-picking when the standard SCM approach is used.

similar to those for ADH's specification. Alternatively, when removing the investment rate from the predictor set or excluding the U.S. from the donor pool, the cross-validation criterion rises to 70.198 or 84.728, respectively. The corresponding timelines in Figure 2 still show a considerable estimated loss in GDP per capita due to the reunification. However, estimated losses are much smaller than for the other specifications, especially those derived without data on the U.S. Correspondingly, Figure 3 shows that these reductions are no longer significant according to the placebo study. When investment is discarded, Germany's post-pre-ratio is only the second-largest, while it is only the fifth-largest when the U.S. data are removed from the sample. Thus, confirming the findings of KKPS and in contrast to ADH, we find that the U.S. data are essential for detecting a significant economic gap due to Germany's reunification.

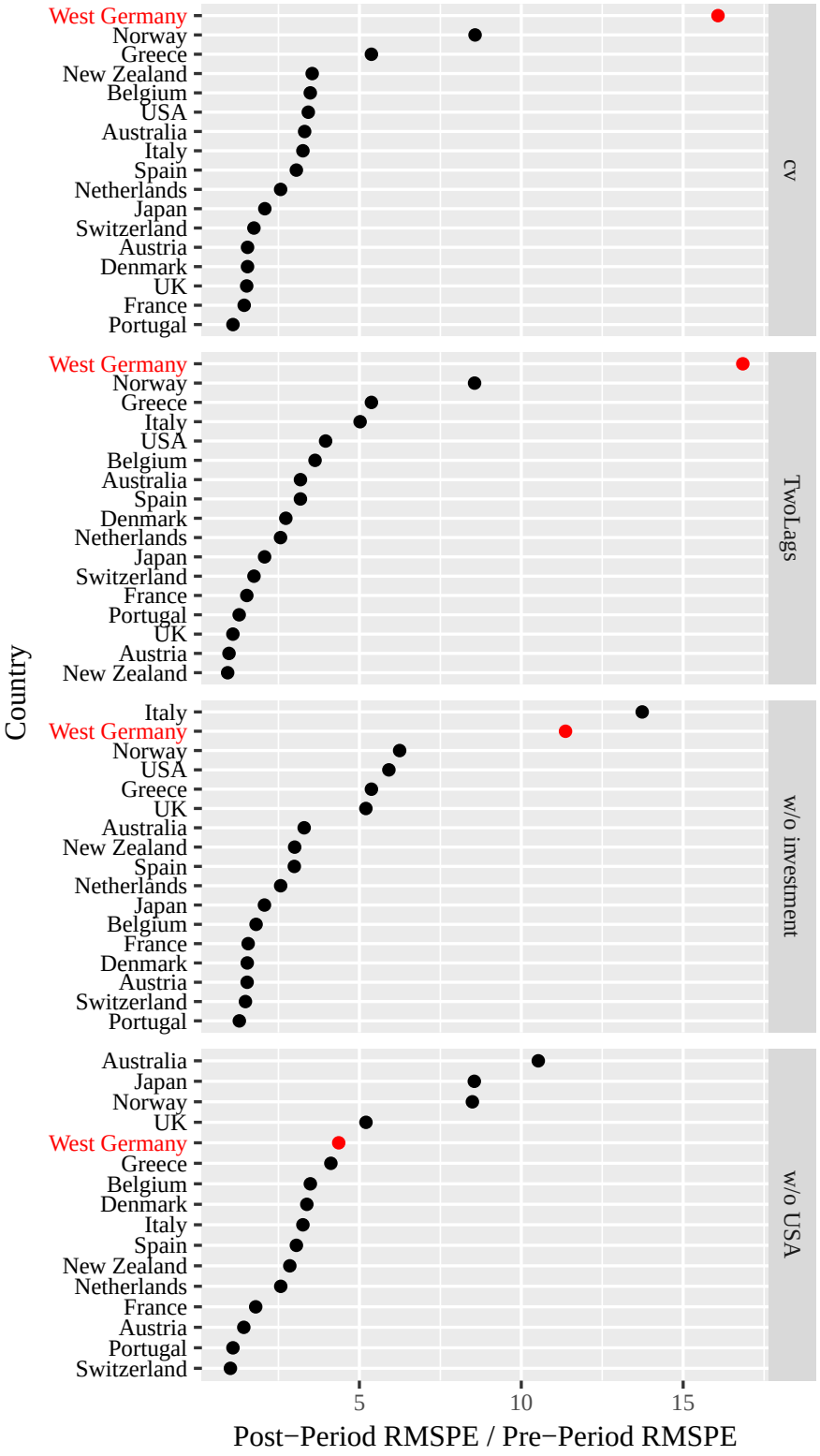

Figure 3: Ratios of post-treatment over pre-treatment root mean square prediction error (RMSPE) for in-space placebos, estimated using cross-validation with specifications of ADH ('cv'), using last two outcome lags ('TwoLags'), without investment ('w/o investment'), and without U.S. data ('w/o USA').

References

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., Diamond, A., and Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *The American Economic Review*, 93(1):113–132.

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. (2017a). Fast and Reliable Computation of Generalized Synthetic Controls. *Econometrics and Statistics*, pages n/a–n/a.

Becker, M. and Klößner, S. (2017b). *MSCMT: Multivariate Synthetic Control Method Using Time Series*. R package version 1.3.0.

Ferman, B., Pinto, C., and Possebom, V. (2017). Cherry picking with synthetic controls. Working Paper.

Gardeazabal, J. and Vega-Bayo, A. (2017). An Empirical Comparison Between the Synthetic Control Method and Hsiao et al.'s Panel Data Approach to Program Evaluation. *Journal of Applied Econometrics*, 32(5):983–1002.

Klößner, S., Kaul, A., Pfeifer, G., and Schieler, M. (2017). Comparative politics and the synthetic control method revisited: A note on Abadie et al. (2015). *Swiss Journal of Economics and Statistics*, pages n/a–n/a.

Klößner, S. and Pfeifer, G. (2017). Outside the box: Using synthetic control methods as a forecasting technique. *Applied Economics Letters*, pages n/a–n/a.

Montalvo, J. G. (2011). Voting after the bombings: A natural experiment on the effect of terrorist attacks on democratic elections. *The Review of Economics and Statistics*, 93(4):1146–1154.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

A Appendix

Let $\mathcal{V}$ be a blunt convex cone in N-dimensional space, i.e., a convex cone not containing 0. For vectors $v = (v_1, \ldots, v_N)' \in \mathcal{V}$, we denote by $v_{(\cdot)} := (v_{(1)}, \ldots, v_{(N)})'$ the 'ordered' version of v with $v_{(1)} \leq v_{(2)} \leq \ldots \leq v_{(N)}$.

Proposition 1. *There exists an up to scaling unique element v^* of $\mathcal{V}$ for which, for all $v \in \mathcal{V}$, $(\frac{v^*_{(1)}}{v^*_{(N)}}, \ldots, \frac{v^*_{(N-1)}}{v^*_{(N)}})$ is lexicographically at least as large as $(\frac{v_{(1)}}{v_{(N)}}, \ldots, \frac{v_{(N-1)}}{v_{(N)}})$.*

Proof. First of all, we show that such v^* exists: if there is an up to scaling *unique* maximizer of $\frac{v_{(1)}}{v_{(N)}} = \frac{\min(v)}{\max(v)}$, then this gives the vector we are looking for. If the solution to maximizing $\frac{v_{(1)}}{v_{(N)}}$ is not unique, even up to scaling, then we can look among all these maximizers for those with maximal $\frac{v_{(2)}}{v_{(N)}}$. Again, if this procudes a solution which is unique up to scaling, this is the vector we are looking for. If there are still several solutions, even after scaling, we can proceed by maximizing $\frac{v_{(3)}}{v_{(N)}}$ among these, and so on. In the end, this procedure will terminate with a vector v^* for which $(\frac{v^*_{(1)}}{v^*_{(N)}}, \ldots, \frac{v^*_{(N-1)}}{v^*_{(N)}})$ is lexicographically maximal.

We now prove uniqueness of v^* up to scaling. To this end, assume that another vector $\widetilde{v}$ is given for which $(\frac{\widetilde{v}_{(1)}}{\widetilde{v}_{(N)}}, \ldots, \frac{\widetilde{v}_{(N-1)}}{\widetilde{v}_{(N)}})$ is also lexicographically maximal. Then, $(\frac{\widetilde{v}_{(1)}}{\widetilde{v}_{(N)}}, \ldots, \frac{\widetilde{v}_{(N-1)}}{\widetilde{v}_{(N)}})$ and $(\frac{v^*_{(1)}}{v^*_{(N)}}, \ldots, \frac{v^*_{(N-1)}}{v^*_{(N)}})$ must coincide. Assuming w.l.o.g. that $\widetilde{v}$ and v^* are scaled such that $\widetilde{v}_{(N)} = \max(\widetilde{v}) = 1 = \max(v^*) = v^*_{(N)}$, this simplifies to $(\widetilde{v}_{(1)}, \ldots, \widetilde{v}_{(N-1)}) = (v^*_{(1)}, \ldots, v^*_{(N-1)})$, showing that $\widetilde{v}$ and v^* are permutations of each other. We denote by $N_{\min,\widetilde{v}} := \{n \in \{1, \ldots, N\} : \widetilde{v}_j = \widetilde{v}_{(1)} = v^*_{(1)}\}$ the set of all components where $\widetilde{v}$ takes its minimum. Analogously, $N_{\min,v^*} := \{n \in \{1, \ldots, N\} : v^*_j = v^*_{(1)} = \widetilde{v}_{(1)}\}$ denotes the set of all components where v^* takes its minimum. $N_{\min,\widetilde{v}}$ and $N_{\min,v^*}$ have a non-empty intersection, because otherwise we would have for $v := \widetilde{v} + v^* \in \mathcal{V}$: $v_{(N)} = \max(v) \leq 2$ and $v_{(1)} = \min(v) > \min(\widetilde{v}) + \min(v^*) = 2\widetilde{v}_{(1)} = 2v^*_{(1)}$, implying that $\frac{v_{(1)}}{v_{(N)}} > \widetilde{v}_{(1)} = v^*_{(1)}$, in contradiction to the optimality of $\widetilde{v}$ and v^*. Thus, $N_{\min,\widetilde{v}} \cap N_{\min,v^*} \neq \varnothing$. In particular, therefore, $\widetilde{v}$ and v^* coincide for all components $n \in N_{\min,\widetilde{v}} \cap N_{\min,v^*}$. From here, we can proceed iteratively, by considering $\{1, \ldots, N\} \setminus (N_{\min,\widetilde{v}} \cap N_{\min,v^*})$ and showing that there are further components where $\widetilde{v}$ and v^* coincide, both taking the value $\widetilde{v}_{(2)} = v^*_{(2)}$, and so on. Overall, then, $\widetilde{v}$ and v^* must coincide completely. □

Klößner et al. (2017, Lemma 1) and Becker and Klößner (2017a, Proposition 3) show that $\boldsymbol{V}$ as defined in Equation (4) is a convex set which can be described by finitely many linear (in-)equalities. Thus, the same holds true for $\widetilde{\boldsymbol{V}}$. Applying Proposition 1 to $\widetilde{\boldsymbol{V}}$ shows that V^* is uniquely defined. V^* can be calculated by solving a series of linear programs.