mRSC: Multi-dimensional Robust Synthetic Control

Muhammad Jehangir Amjad
MIT
mamjad@mit.edu

Devavrat Shah
MIT
devavrat@mit.edu

Vishal Misra
Columbia University
vishal.misra@columbia.edu

Dennis Shen
MIT
deshen@mit.edu

ABSTRACT

When evaluating the impact of a policy (e.g., gun control) on a metric of interest (e.g., crime-rate), it may not be possible or feasible to conduct a randomized control trial. In such settings where only observational data is available, synthetic control (SC) methods [1–3] provide a popular data-driven approach to estimate a “synthetic” or “virtual” control by combining measurements of “similar” alternatives or units (called “donors”).

Recently, robust synthetic control (RSC) [4] was proposed as a generalization of SC to overcome the challenges of missing data and high levels of noise, while removing the reliance on expert domain knowledge for selecting donors. However, both SC and RSC (and its variants) suffer from poor estimation when the pre-intervention period is too short.

As the main contribution of this work, we propose a generalization of unidimensional RSC to multi-dimensional Robust Synthetic Control, mRSC. Our proposed mechanism, mRSC, incorporates multiple types of measurements (or metrics) in addition to the measurement of interest for estimating a synthetic control, thus overcoming the challenge of poor inference due to limited amounts of pre-intervention data. We show that the mRSC algorithm, when using K relevant metrics, leads to a consistent estimator of the synthetic control for the target unit of interest under any metric. Our finite-sample analysis suggests that the mean-squared error (MSE) of our predictions decays to zero at a rate faster than the RSC algorithm by a factor of K and $\sqrt{K}$ for the training (pre-intervention) and testing (post-intervention) periods, respectively. Additionally, we propose a principled scheme to combine multiple metrics of interest via a diagnostic test that evaluates if adding a metric can be expected to result in improved inference.

Our mechanism for validating mRSC performance is also an important and related contribution of this work: time series prediction. We propose a method to predict the future evolution of a time series based on limited data when the notion of time is relative and not absolute, i.e., where we have access to a donor pool that has already undergone the desired future evolution.

We conduct extensive experimentation to establish the efficacy of mRSC in three different scenarios: predicting the evolution of a

metric of interest using synthetically generated data from a known factor model, and forecasting weekly sales and score trajectories of a Walmart store and Cricket game, respectively.

CCS CONCEPTS

• **General and reference** → **Estimation**; *Evaluation*; • **Theory of computation** → *Sample complexity and generalization bounds*; • **Computing methodologies** → *Latent variable models*.

KEYWORDS

Causal Inference; Synthetic Control; Matrix Estimation; Cricket

ACM Reference Format:

Muhammad Jehangir Amjad, Vishal Misra, Devavrat Shah, and Dennis Shen. 2019. mRSC: Multi-dimensional Robust Synthetic Control. In *ACM SIGMETRICS Int'l Conference on Measurement & Modeling of Computer Systems (SIGMETRICS'19 Abstracts)*, June 24–28, 2019, Phoenix, AZ, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3326152>

1 INTRODUCTION

Quantifying the causal effect of interventions is a problem of interest across a wide array of domains. From policy making to engineering and medicine, estimating the effect of an intervention is critical for innovation and understanding existing systems. In any setting, with or without an intervention, we only observe one set of outcomes. In casual analysis, the fundamental problem is that of estimating what wasn't observed, i.e., the *counterfactual*.

In order to estimate the counterfactual, observational studies rely on the identification (or estimation) of a *control* unit. This can be achieved by relying on expert domain knowledge, or via techniques such as *matching* the target unit to existing control units (called *donors*) on covariates features or propensity scores. A popular data-driven approach to estimating the control is known as the Synthetic Control (SC) method [1–3]. SC assigns convex weights to the donors such that the resulting *synthetic* unit most closely matches the target unit according to a chosen metric of interest. A generalization of this approach, Robust Synthetic Control (RSC) [4], removes the convexity constraint and guarantees a consistent estimator that is robust to missing data and noise.

While SC, its many variants, and RSC exhibit attractive properties, they all suffer for poor estimation when the amount of training data (i.e., the length of the pre-intervention period) is small. In many of these scenarios with little pre-intervention data, there may be other data available which is related to the type of data (or metric) of interest. For example, we might be interested in crime-rate and might also have median household income and high-school

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS'19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6678-6/19/06.

<https://doi.org/10.1145/3326152>

graduation rate data available. Therefore, one remedy to the limited pre-intervention data is to utilize data from multiple metrics.

1.1 Contributions

As the main contribution of this work, we propose multi-dimensional Robust Synthetic Control (mRSC), a generalization of unidimensional RSC. Unlike standard SC-like methods, mRSC incorporates multiple types of data in a principled manner to overcome the challenge of forecasting the counterfactual from limited pre-intervention data. We show that the mRSC method is a natural consequence of the popular factor model that is commonly utilized in the field of econometrics. Through this connection, we provide a data-driven falsifiability test for mRSC, which determines whether mRSC can be expected to produce a consistent estimation of the counterfactual evolution across all metrics for a target unit (exposed to intervention) in both the pre- and post-intervention stages.

Further, we demonstrate that prediction power of mRSC improves over RSC as the number of included relevant data types increases. Specifically, the mean-squared error (MSE) vanishes at a rate scaling with K and $\sqrt{K}$ for pre- and post-intervention estimates, respectively, where K denotes the number of metrics; we highlight that $K = 1$ reduces to the vanilla RSC framework. Additionally, desirable properties of the RSC algorithm, namely robustness to missing data and noise, carry over. Finally, we conduct extensive experimentation to establish the efficacy of this generalized method in comparison to RSC via synthetically generated datasets and two real-world case-studies: product sales in retail and score trajectory forecasting in the game of Cricket. We now summarize these contributions in greater detail.

Model. We consider a natural generalization of the factor (or latent variable) model considered in the literature cf. [4], where the measurement associated with a given unit at a given time for a given metric is a function of the features associated with the particular unit, time, and metric. A special case of such models, the linear factor models, are the *bread and butter* models within the Econometrics literature, and are assumed in [1–3]. We argue that as long as the function is Lipschitz and the features belong to a compact domain, then the resulting factor model can be well approximated by low-rank third-order tensor with the dimensions of the tensor corresponding to the unit, time, and metric, respectively. Therefore, to simplify this exposition, we focus on the setup where the underlying model is indeed a low-rank tensor. We assume that our observations are a corrupted version of the ground truth low-rank tensor with potentially missing values. More specifically, we assume we have access to pre-intervention data for all units and metrics, but only observe the post-intervention data for the donor units. Our goal is to estimate the “synthetic control” using all donor units and metrics in the pre-intervention period. The estimated synthetic control can then help estimate the *future* (post-intervention) ground-truth measurements for all metrics associated with the treatment unit.

Algorithm and analysis. The current SC literature [1–4] relies on the following key assumption: for any metric of interest, the treatment unit’s measurements across time can be expressed as a linear combination of the donor units’ measurements. In this work, we start by arguing that we need not make this assumption as it

holds naturally under the low-rank tensor model. Furthermore, the low-rank tensor model suggests that the identical linear relationship holds across all metrics simultaneously with high probability. That is, in order to utilize measurements from other metrics to estimate the synthetic control for a given metric and unit of interest, we can effectively treat the measurements for other metric as additional measurements for our metric of interest! In other words, the number of pre-intervention measurements is essentially multiplied by the number of available metrics. The resulting mRSC algorithm is natural extension of RSC but with multiple measurements. Using a recently developed method for analyzing error-in-variable regression in the high-dimensional regime, we derive finite-sample guarantees on the performance of the mRSC algorithm. We conclude that the pre- and post-intervention MSE for mRSC decays faster than that of RSC by a factor K and $\sqrt{K}$, respectively, where K is the number of available metrics. This formalizes the intuition that data from other metrics can be treated as belonging to the original metric of interest. In summary, mRSC provides a way to overcome limited pre-intervention data by simply utilizing pre-intervention data across other metrics.

Experiments. We validate our finite sample analyses via empirical studies. To begin, using synthetically generated data, we argue that flattening the third-order tensor of measurement data into a matrix (by stacking the unit by time slices across metrics) is valid if the resulting matrix exhibits a similar singular value spectrum (rank) as that of the matrix with only a single metric of interest. Our results demonstrate that data generated as per our model pass the diagnostic test; the test, however, fails when the data does not come from our model.

Next, we study the efficacy of mRSC in a real-world case-study regarding weekly sales data across multiple departments at various Walmart stores. Across all our experiments, we consistently find that mRSC significantly outperforms RSC when the pre-intervention data is small. Specifically, we observe that the test prediction error of mRSC is 5–7x better compared to vanilla RSC when very limited pre-intervention data is available. Further, even if the pre-intervention data is significant, mRSC’s test prediction error continues to outperform that of RSC. Finally, we demonstrate the prediction ability of mRSC via the task of forecasting score trajectories in the game of cricket. Using examples of actual cricket games, we also show that the mRSC algorithm successfully captures and reacts to nuances of cricket.

ACKNOWLEDGMENTS

Supported by NSF grants CMMI-1462158, CMMI-1634259, CNS-1523546 and CNS-1618911; Draper Fellowship; Google FRA.

REFERENCES

- [1] A. Abadie, A. Diamond, and J. Hainmueller. 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program. *J. Amer. Statist. Assoc.* (2010).
- [2] A. Abadie, A. Diamond, and J. Hainmueller. 2011. Synth: An R Package for Synthetic Control Methods in Comparative Case Studies. *Journal of Statistical Software* (2011).
- [3] A. Abadie and J. Gardeazabal. 2003. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review* (2003).
- [4] Muhammad Amjad, Devavrat Shah, and Dennis Shen. 2018. Robust Synthetic Control. *Journal of Machine Learning Research* 19, 17 (2018), 1–50.