Rank estimation for (approximately) low-rank matrices

Niloofar Bayat
Columbia University
niloofar.bayat@columbia.edu

Cody Morrin
Columbia University
cody.morrin@columbia.edu

Yuheng Wang
Columbia University
y.wang4@columbia.edu

Vishal Misra
Columbia University
vishal.misra@columbia.edu

ABSTRACT

In observational data analysis, e.g., causal inference, one often encounters data sets that are noisy and incomplete, but come from inherently "low rank" (or correlated) systems. Examples include user ratings of movies/products and term frequency matrices for documents amongst others. In such analysis, estimating the approximate rank of the data sets serves an important function of delineating the signal from the noise. In this paper, we propose a technique to estimate the rank of observational data matrices, compare it to previously proposed techniques, and make a specific methodological contribution of improving the algorithmic parameter estimation in the robust synthetic control method in [1].

Our most up-to-date code, model, and data can be found on `https://github.com/niloofarbayat/COVID19-synthetic-control-analysis`.

1. INTRODUCTION

Abadie and Gardeazabal [2] and Abadie et al. [3] pioneered the synthetic control method to address the problem of measuring the impact of a new regulation or a change in a region, and comparing it to the case that those changes had not happened. The idea of synthetic control stems from the classical A/B testing, where two versions of a variable are compared in the otherwise identical environment [4]. A variation of that would be where one of the variables is a placebo. In that case, the experimental units are called “treatment” and “control” (placebo) groups, and the variable changes, i.e., treatments, are applied to the treatment group [5]. Abadie et al. suggested that since in some problems we cannot have an actual control group, we construct a “synthetic” control group.

Synthetic control is a statistical method that evaluates the impact of an intervention on observational data. After the occurrence of an event or intervention (e.g., increasing tobacco tax) on a unit (e.g., a region), synthetic control estimates the evolution of some aggregate outcomes (e.g., smoking rate). To measure whether the outcomes were affected by the intervention, we need to compare them to the evolution of the outcome without the intervention in the same unit (control unit). However, if the unit under study is a unique region, we cannot have an actual control unit, but we can synthetically construct it. The key to constructing this "synthetic control" unit is using the data from other regions which did not have the intervention. [3].

In some cases, the data under study might be noisy or have missing values. The classical synthetic control method does not perform well in those cases, and robust synthetic control (RSC), a generalization of the late, has been proposed to address those limitations [1]. RSC performs “de-noising” estimations and uses the de-noised data for synthetic control analysis. It assumes a latent variable model for the generation of the data and uses hard singular value thresholding [6] to perform the de-noising.

Contributions: We find and fix a flaw of picking the algorithmic parameter of “singular values” proposed in the robust synthetic control paper [1]. We also propose a general technique to estimate the approximate rank of a matrix by analyzing the auto-correlation of the residual matrix left after extracting a low-rank approximation. Our method appears general, compares favorably with a previously proposed technique, and fixes the problem with the RSC method.

2. RELATED WORK

Abadie *et al.* first introduced synthetic control to measure the impact of political instability on economic prosperity [2]. They investigated the economic impact of conflict using the terrorist attack data in the Basque Country as a case study. They used the combination of other regions in Spain to build a “synthetic” control region that resembles the economic characteristics of the Basque Country before the outset of the terrorist attack. After that, the method has been widely applied in econometric of policy evaluations, including studying the effects of laws [7], legalized prostitution [8], and immigration policy [9], as well as biomedical disciplines [10], and social sciences [11].

Rank estimation for low-rank matrices plays a crucial role in the de-noising step of the RSC algorithm. In [12] Ubaru *et al* proposed a computationally inexpensive technique for estimating the numerical rank of a matrix. They utilized Chebyshev expansion to approximate the projector on the non-null invariant subspace of the matrix and then used a stochastic trace estimator to estimate the rank. This method is efficient, but it is not precise enough for de-noising time series due to its randomized nature.

2.1 Robust Synthetic Control

The robust synthetic control (RSC) method [1], as mentioned earlier, is a generalization of the classical synthetic control method. It makes the synthetic control estimation robust to randomly missing data and high variance noise. This generalization estimates the synthetic control weights using the unobserved mean values instead of the noisy observations. The estimation is done by “de-noising” the data matrix using matrix completion and then using regression to determine the synthetic control weights, and has been shown to be equivalent to principal component regression. Furthermore, the counterfactual outcome can be estimated by *any* linear combination of the donor units, relaxing the convex constraints on the weights in classical synthetic control.

3. DENOISING AND ESTIMATING THE APPROXIMATE RANK OF DATA MATRIX

The theory of robust synthetic control and synthetic interventions is built upon the critical assumption that the observational data is coming from a “nice” (Lipschitz) latent variable function [1, 13], which uses the so-called “latent tensor factor model”. The assumption of the Lipschitzness is equivalent to the observational tensor (or matrix) being low rank. This rank plays a crucial role in the denoising step of the robust synthetic control algorithm and its offshoots ([14]). Restricting the case to a two-dimensional (i.e. matrix) factor model, the observations are:

$$X_{it} = M_{it} + \epsilon_{it}$$

where M_{it} is coming from the low-rank system, and ϵ_{it} is iid noise. The "denoising" step of RSC estimates M_{it} via a universal singular value thresholding mechanism proposed in [6]. Since apriori we do not know the system’s rank, the rank needs to be estimated via the observation matrix. If the number of singular values retained is too low, the denoised estimate $\hat{M}$ does not capture the complexity of the underlying system. Conversely, if the number of singular values retained is too high, the estimate overfits on noise, causing large errors in the prediction/counterfactual estimation step. If one knows the variance of the noise σ^2, and if $\sigma^2 \leq 1$, then [6] suggests retaining the singular values that satisfy

$$S := \left\{ i : s_i \geq (2 + \eta)\sqrt{n\hat{q}} \right\} \qquad (1)$$

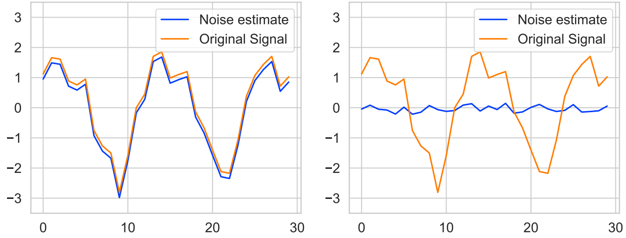
where $\hat{q} := \hat{p}\sigma^2 + \hat{p}(1 - \hat{p})(1 - \sigma^2)$, with $\hat{p}$ being the (iid) probability of observing a value, n the nominal rank of the matrix, σ^2 the variance of the noise, and η a small constant in the range $[0.1, 1]$. In [1], this threshold is extended to the case where the noise variance is unknown. They suggest the following estimator for the noise variance:

$$\hat{\sigma}^2 = \frac{1}{T_0 - 1} \sum_{t=1}^{T_0} (Y_{1t} - \bar{Y})^2 \qquad (2)$$

where T_0 is the number of pre-intervention sample points and $\bar{Y}$ denotes the pre-intervention sample mean. Y_{it} is simply X_{it} if it is observed with probability p, and 0 if it is unobserved with probability $1-p$. This estimate $\hat{\sigma}$ however suffers from a *significant* flaw. It assumes that $\bar{Y}$ is the estimate of the expected value of *all* Y_{it}. In general, $\mathbf{Y_i}$ is a vector, coming from a low dimensional space plus an additive noise term, and there is no requirement or assumption that the expected value of observations is constant over every element of the vector. Even taking the sample average along the column dimension produces no improvement, since again, there is no assumption that the expected value of Y_{it} is independent of t.

This flaw can be illustrated by the following toy example. To represent a dataset with 50 units and 100 time points, we generate 4 independent sinusoidal vectors of size 100, and construct an observation matrix using a linear combination of those vectors and a small additive noise to generate 50 rows, resulting in a matrix of size 50×100. The nominal rank of this matrix is 50, but from construction, the (approximate) rank is 4. We then estimate the noise as outlined in [1]. As we can see in Figure 1(a), the noise estimate is *nearly indistinguishable* from the signal. This is not surprising because the sample mean of our observation matrix is ≈ 0.

If instead, we estimate the noise by subtracting a rank-4 approximation from the observation matrix (obtained via singular value thresholding), then the estimate that we obtain, which is shown in Figure 1(b), is a much better estimate of the noise and its variance.



(a) Robust Synthetic Control (b) rank-4 approximation

Figure 1: Noise estimate obtained via subtracting a rank-4 approximation of the observation matrix

3.1 Estimating approximate rank via an autocorrelation test

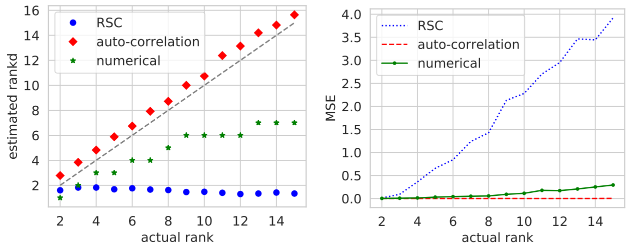
We now present our technique of computing the approximate rank building on the assumption that the noise is iid.

The problem with the approach we just presented, is that it is circular - to estimate the noise (variance), one needs the low-rank approximation, and to estimate the rank of the matrix, one needs the noise variance as in Equation 1. To fix this problem, we compute the noise vector for different singular values. Since the noise vector is a time series, we can compute the auto-correlation (ACF) of the resulting rows of the matrix, which represent individual time series for the units. If the noise ACF is insignificant for different time lags, then the noise does not have any observable patterns, and we can assume that our estimation of the iid noise is correct.

We implement the Breusch–Godfrey test where the null hypothesis is that the noise has an insignificant ACF. We then compute the ACF for different time lags, choosing the cut-off for the significance of each lag to be a 95% interval, i.e., we conclude the ACF at each lag is non-zero if it is greater in magnitude than the boundary of the say the 95% confidence interval. We then compute the p-values; if $p - value > 0.05$, we do not have enough statistical evidence to reject the null hypothesis, and we cannot assume that our noise values are dependent or carry any pattern from the signal. Conversely, for $p - value < 0.05$, we can reject the null hypothesis and conclude that the residual time series is correlated.

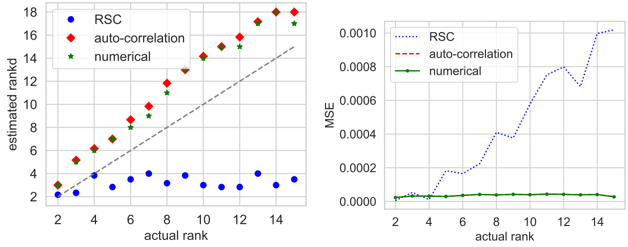
To perform a low-rank estimation of the signal, we start by rank 1, compute the noise ACF, and increment the rank until we reach an iid noise using the aforementioned method. If no rank results in an iid noise, the method returns the nominal rank. We repeat this process for every row of the signal and compute the minimum rank for which the noise vector for that row has an insignificant ACF. Then we compute the average rank among them and round that value to the nearest integer as our rank estimate. We compare our rank estimation method with the work of [12] which focuses on fast methods of numerical rank estimation. We build 50 random signals of each rank $k = 2$ to $k = 15$ as mentioned in the construction of our toy example and compute the rank of the matrix using the two methods. The average MSE of our method is 0.0461, and the estimated rank is correct 25% of times with the average difference of 1.68. While the average MSE using [12] is 0.1128, and the estimated rank is correct 25% of times with the average difference of 3.22.

We observe that the work of [12] always underestimates the rank in this case of periodic signals, so we can further improve the performance of our method by setting the rank estimation of [12] as the lower bound for our method. Integrating these two methods, the average MSE would be improved to 0.00062, and the estimated



(a) Rank Estimation (b) MSE

Figure 2: Average values of rank estimation and MSE across different models for our example with periodic signals.



(a) Rank Estimation (b) MSE

Figure 3: Average values of rank estimation and MSE across different models for sklearn low rank matrix generator.

rank is correct 23% of times with the average difference of 1.3.

Figure 2 displays a comparison of our method's average rank estimation and MSE with those of RSC, and the numerical rank estimation method in [12]. We ran the simulation 50 times for each rank and computed the average among them. We observe that our method's rank estimation closely follows the actual rank of the signal, and our MSE is the lowest among the three methods. We also observe that RSC fails to detect the rank of the signal as expected since the signal has a periodic non-constant expected value. In Figure 3 we did the same comparison but this time we generated the low-rank matrices using a function from the scikit-learn library in python [15]. For these classes of low-rank matrices, with bell-shaped singular values, the numerical approximation technique works well and has performance similar to our technique, but the RSC method again fails to capture the true rank of the matrices.

4. CONCLUSION

In this paper, we provide a way to estimate the approximate rank of observational data matrices by looking at the auto-correlation of the residual signal. Our technique has similar performance to a previously proposed numerical technique for fast estimation of matrix rank and outperforms it when the nature of the (low rank) basis vectors is periodic. Both those techniques outperform the method proposed in the RSC paper, where it severely underestimates the true rank of the data matrix and thereby misses out in capturing the structure of the underlying signal. Our contribution fixes this important flaw in the RSC mechanism.

5. REFERENCES

[1] M. Amjad, D. Shah, and D. Shen. Robust synthetic control. *J. Mach. Learn. Res.*, 19(1):802–852, January 2018.

[2] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

[3] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 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, 2010.

[4] Scott WH Young. Improving library user experience with a/b testing: Principles and process. *Weave: Journal of Library User Experience*, 1(1), 2014.

[5] Steve Chaplin. The placebo response: an important part of treatment. *Prescriber*, 17(5):16–22, 2006.

[6] S. Chatterjee. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43:177–214, 2015.

[7] John J Donohue, Abhay Aneja, and Kyle D Weber. Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis. *Journal of Empirical Legal Studies*, 16(2):198–247, 2019.

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

[9] Sarah Bohn, Magnus Lofstrom, and Steven Raphael. Did the 2007 legal arizona workers act reduce the state's unauthorized immigrant population? *Review of Economics and Statistics*, 96(2):258–269, 2014.

[10] Noémi Kreif, Richard Grieve, Dominik Hangartner, Alex James Turner, Silviya Nikolova, and Matt Sutton. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health economics*, 25(12):1514–1528, 2016.

[11] Boris Heersink, Brenton D Peterson, and Jeffery A Jenkins. Disasters and elections: Estimating the net effect of damage and relief in historical perspective. *Political Analysis*, 25(2):260–268, 2017.

[12] Shashanka Ubaru and Yousef Saad. Fast methods for estimating the numerical rank of large matrices. In *International Conference on Machine Learning*, pages 468–477. PMLR, 2016.

[13] Anish Agarwal, Abdullah Alomar, Romain Cosson, Devavrat Shah, and Dennis Shen. Synthetic Interventions. `https://arxiv.org/pdf/2006.07691.pdf`.

[14] M. Amjad, V. Misra, D. Shah, and D. Shen. mRSC: Multi-dimensional robust synthetic control. In *Proceedings of the ACM on Measurement and Analysis of Computing Systems (Sigmetrics 2019)*, volume 3, page 37, June 2019.

[15] Make low rank matrix.