Robust Synthetic Control[*]

Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen

Laboratory for Information and Decision Systems,
Statistics and Data Science Center,
Massachusetts Institute of Technology
{mamjad, devavrat, deshen}@mit.edu

Abstract

We present a robust generalization of the synthetic control method for comparative case studies. Like the classical method cf. [1, 4, 2], we present an algorithm to estimate the unobservable counterfactual of a treatment unit. A distinguishing feature of our algorithm is that of de-noising the data matrix via singular value thresholding, which renders our approach robust in multiple facets: it automatically identifies a good subset of donors for the synthetic control, overcomes the challenges of missing data, and continues to work well in settings where covariate information may not be provided. To begin with, we establish the condition under which the fundamental assumption in synthetic control-like approaches holds, i.e. when the linear relationship between the treatment unit and the donor pool prevails in both the pre- and post-intervention periods. We provide the first finite sample analysis (coupled with asymptotic results) for a broader class of models, the Latent Variable Model (LVM), in contrast to Factor Models previously considered in the literature, while also relating the interpolation and extrapolation abilities of our estimator to the amount of data available. In particular, we show that our de-noising procedure accurately imputes missing entries and filters corrupted observations in producing a consistent estimator of the underlying signal matrix, provided $p = \Omega(T^{-1+\zeta})$ for some $\zeta > 0$; here, p is the fraction of observed data and T is the time interval of interest. Under the same proportion of observations, we demonstrate that the mean-squared-error in our prediction estimation scales as $\mathcal{O}(\sigma^2/p + 1/\sqrt{T})$, where σ^2 is the variance of the inherent noise. Using a "data aggregation" method, we show that the mean-square-error can be made as small as $\mathcal{O}(T^{-1/2+\gamma})$ for any $\gamma \in (0, 1/2)$, and thus leading to a consistent estimator. In order to move beyond point estimates, we introduce a Bayesian framework that not only provides the ability to readily develop different estimators under various loss functions, but also quantifies the uncertainty of the model/estimates through posterior probabilities. Our experiments, using both synthetic and real-world datasets, demonstrate that our robust generalization yields an improvement over the classical synthetic control method, underscoring the value of our key de-noising procedure.

1 Introduction

On November 8, 2016 in the aftermath of several high profile mass-shootings, voters in California passed Proposition 63 in to law [10]. Prop. 63 "outlaw[ed] the possession of ammunition magazines that [held] more than 10 rounds, requir[ed] background checks for people buying bullets," and was proclaimed as an initiative for "historic progress to reduce gun violence" [30]. Imagine that we wanted to study the impact of Prop. 63 on the rates of violent crime in California. Randomized control trials, such as A/B testings, have been successful in establishing effects of interventions by randomly exposing segments of the population to various types of interventions. Unfortunately, a randomized control trial is not applicable in this scenario since only one California exists. Instead, a statistical comparative study could be conducted where the rates of violent crime in California are compared to a "control" state after November 2016, which we refer to as the post-intervention period. To reach a statistically valid conclusion, however, the control state must be demonstrably similar to California sans the passage of a Prop. 63 style legislation. In general, there may not exist a natural control state for California, and subject-matter experts tend to disagree on the most appropriate state for comparison.

As a suggested remedy to overcome the limitations of a classical comparative study outlined above, Abadie et al. proposed a powerful, data-driven approach to construct a "synthetic" control unit absent of intervention [1, 4, 2]. In the example above, the synthetic control method would construct a "synthetic" state of California such that the rates of violent crime of that hypothetical state would best match the rates in California before the passage of Prop. 63. This synthetic California can then serve as a data-driven counterfactual for the period after the passage of Prop. 63. Abadie et al. propose to construct such a synthetic California by choosing a convex combination of other states (donors) in the United States. For instance, synthetic California might be 80% like New York and 20% like Massachusetts. This approach is nearly entirely data-driven and appeals to intuition. For optimal results, however, the method still relies on subjective covariate information, such as employment rates, and the presence of domain "experts" to help identify a useful subset of donors. The approach may also perform poorly in the presence of non-negligible levels of noise and missing data.

1.1 Overview of main contributions.

As the main result, we propose a simple, two-step robust synthetic control algorithm, wherein the first step de-noises the data and the second step learns a linear relationship between the treated unit and the donor pool under the de-noised setting. The algorithm is robust in two senses: first, it is fully data-driven in that it is able to find a good donor subset even in the absence of helpful domain knowledge or supplementary covariate information; and second, it provides the means to overcome the challenges presented by missing and/or noisy observations. As another important contribution, we establish analytic guarantees (finite sample analysis and asymptotic consistency) – that are missing from the literature – for a broader class of models.

Robust algorithm. A distinguishing feature of our work is that of de-noising the observation data via singular value thresholding. Although this spectral procedure is commonplace in the matrix completion arena, it is novel in the realm of synthetic control. Despite its simplicity, however, thresholding brings a myriad of benefits and resolves points of concern that have not been previously addressed. For instance, while classical methods have not even tackled the obstacle of missing data, our approach is well equipped to impute missing values as a consequence of the thresholding procedure. Additionally, thresholding can help prevent the model from overfitting to the idiosyncrasies of the data, providing a knob for practitioners to tune the "bias-variance" trade-off of their model and, thus, reduce their mean square error (MSE). From empirical studies, we hypothesize that thresholding may possibly render auxiliary covariate information (vital to several existing methods) a luxury as opposed to a necessity. However, as one would expect, the algorithm can only benefit from useful covariate and/or "expert" information and we do not advocate ignoring such helpful information, if available.

In the spirit of combatting overfitting, we extend our algorithm to include regularization techniques such as ridge regression and LASSO. We also move beyond point estimates in establishing a Bayesian framework, which allows one to quantitatively compute the uncertainty of the results through posterior probabilities.

Theoretical performance. To the best of our knowledge, ours is the first to provide finite sample analysis of the MSE for the synthetic control method, in addition to guarantees in the presence of missing data. Previously, the main theoretical result from the synthetic control literature (cf. [1, 4, 2]) pertained to bounding the bias of the synthetic control estimator; however, the proof of the result assumed that the latent parameters, which live in the simplex, have a perfect pre-treatment match in the noisy predictor variables – our analysis, on the other hand, removes this assumption. We begin by demonstrating that our de-noising procedure produces a consistent estimator of the latent signal matrix (Theorems 4.1, 4.2), proving that our thresholding method accurately imputes and filters missing and noisy observations, respectively. We then provide finite sample analysis that not only highlights the value of thresholding in balancing the inherent "bias-variance" trade-off of forecasting, but also proves that the prediction efficacy of our algorithm degrades gracefully with an increasing number of randomly missing data (Theorems 4.3, 4.6, and Corollary 4.1). Further, we show that a computationally beneficial pre-processing data aggregation step allows us to establish the asymptotic consistency of our estimator in generality (Theorem 4.4).

Additionally, we prove a simple linear algebraic fact that justifies the basic premise of synthetic control, which has not been formally established in literature, i.e. the linear relationship between the treatment and donor units that exists in the pre-intervention continues to hold in post-intervention period (Theorem 4.5). We introduce a latent variable model, which subsumes many of the models previously used in literature (e.g. econometric factor models). Despite this generality, a unifying theme that connects these models is that they all induce (approximately) low rank matrices, which is well suited for our method.

Experimental results. We conduct two sets of experiments: (a) on existing case studies from real world datasets referenced in [1, 2, 4], and (b) on synthetically generated data. Remarkably, while [1, 2, 4] use numerous covariates and employ expert knowledge in selecting their donor pool, our algorithm achieves similar results without any such assistance; additionally, our algorithm detects subtle effects of the intervention that were overlooked by the original synthetic control approach. Since it is impossible to simultaneously observe the evolution of a treated unit and its counterfactual, we employ synthetic data to validate the efficacy of our method. Using the MSE as our evaluation metric, we demonstrate that our algorithm is robust to varying levels of noise and missing data, reinforcing the importance of de-noising.

1.2 Related work.

Synthetic control has received widespread attention since its conception by Abadie and Gardeazabal in their pioneering work [4, 1]. It has been employed in numerous case studies, ranging from criminology [31] to health policy [27] to online advertisement to retail; other notable studies include [3, 11, 5, 9]. In their paper on the state of applied econometrics for causality and policy evaluation, Athey and Imbens assert that synthetic control is "one of the most important development[s] in program evaluation in the past decade" and "arguably the most important innovation in the evaluation literature in the last fifteen years" [8]. In a somewhat different direction, Hsiao et al. introduce the panel data method [24, 25], which seems to have a close bearing with some of the approaches of this work. In particular, [24, 25] only uses data for the outcome variable and solves an ordinary least squares problem in learning synthetic control. However, [24, 25] restrict the subset of possible controls to units that are within the geographical or economic proximity of the treated unit. Therefore, there is still some degree of subjectivity in the choice of the donor pool. In addition, [24, 25] do not include a "de-noising" step, which is a key feature of our approach. For an empirical comparison between the synthetic control and panel data methods,

3

see [21]. It should be noted that [21] also adapts the panel data method to automate the donor selection process. [17] allows for an additive difference between the treated unit and donor pool, similar to the difference-in-differences (DID) method. Moreover, similar to our exposition, [17] relaxes the convexity aspect of synthetic control and proposes an algorithm that allows for unrestricted linearity as well as regularization. In an effort to infer the causal impact of market interventions, [14] introduce yet another evaluation methodology based on a diffusion-regression state-space model that is fully Bayesian; similar to [1, 4, 24, 25], their model also generalizes the DID procedure. Due to the subjectivity in the choice of covariates and predictor variables, [20] provides recommendations for specification-searching opportunities in synthetic control applications. The recent work of [34] extends the synthetic control method to allow for multiple treated units and variable treatment periods as well as the treatment being correlated with unobserved units. Similar to our work, [34] computes uncertainty estimates; however, while [34] obtains these measurements via a parametric bootstrap procedure, we obtain uncertainty estimates under a Bayesian framework.

Matrix completion and factorization approaches are well-studied problems with broad applications (e.g. recommendation systems, graphon estimation, etc.). As shown profusely in the literature, spectral methods, such as singular value decomposition and thresholding, provide a procedure to estimate the entries of a matrix from partial and/or noisy observations [15]. With our eyes set on achieving "robustness", spectral methods become particularly appealing since they de-noise random effects and impute missing information within the data matrix [26]. For a detailed discussion on the topic, see [16]; for algorithmic implementations, see [29] and references there in. We note that our goal differs from traditional matrix completion applications in that we are using spectral methods to estimate a low-rank matrix, allowing us to determine a linear relationship between the rows of the mean matrix. This relationship is then projected into the future to determine the counterfactual evolution of a row in the matrix (treated unit), which is traditionally not the goal in matrix completion applications. Another line of work within this arena is to impute the missing entries via a nearest neighbor based estimation algorithm under a latent variable model framework [28, 13].

There has been some recent work in using matrix norm methods in relation to causal inference, including for synthetic control. In [7], the authors use matrix norm regularization techniques to estimate counterfactuals for panel data under settings that rely on the availability of a large number of units relative to the number of factors or characteristics, and under settings that involve limited number of units but plenty of history (synthetic control). This is different from our approach, which increases robustness by "de-noising" using matrix completion methods, and then using linear regression on the de-noised matrix, instead of relying on matrix norm regularizations.

Despite its popularity, there has been less theoretical work in establishing the consistency of the synthetic control method or its variants. [1] demonstrates that the bias of the synthetic control estimator can be bounded by a function that is close to zero when the pre-intervention period is large in relation to the scale of the transitory shocks, but under the additional condition that a perfect convex match between the pre-treatment noisy outcome and covariate variables for the treated unit and donor pool exists. [19] relaxes the assumption in [1], and derives conditions under which the synthetic control estimator is asymptotically unbiased under non-stationarity conditions. To our knowledge, however, no prior work has provided finite-sample analysis, analyzed the performance of these estimators with respect to the mean-squared error (MSE), established asymptotic consistency, or addressed the possibility of missing data, a common handicap in practice.

2 Background

2.1 Notation.

We will denote $\mathbb{R}$ as the field of real numbers. For any positive integer N, let $[N] = \{1, \ldots, N\}$. For any vector $v \in \mathbb{R}^n$, we denote its Euclidean (ℓ_2) norm by $\|v\|_2$, and define $\|v\|_2^2 = \sum_{i=1}^n v_i^2$. We define its infinity norm as $\|v\|_\infty = \max_i |v_i|$. In general, the ℓ_p norm for a vector v is defined as $\|v\|_p = \left(\sum_{i=1}^n |v_i|^p \right)^{1/p}$. Similarly, for an $m \times n$ real-valued matrix $\boldsymbol{A} = [A_{ij}]$, its spectral/operator norm, denoted by $\|\boldsymbol{A}\|_2$, is defined as $\|\boldsymbol{A}\|_2 = \max_{1 \le i \le k} |\sigma_i|$, where $k = \min\{m, n\}$ and σ_i are the singular values of $\boldsymbol{A}$. The Moore-Penrose pseudoinverse $\boldsymbol{A}^\dagger$ of $\boldsymbol{A}$ is defined as

$$\boldsymbol{A}^\dagger = \sum_{i=1}^k (1/\sigma_i) y_i x_i^T, \quad \text{where} \quad \boldsymbol{A} = \sum_{i=1}^k \sigma_i x_i y_i^T, \tag{1}$$

with x_i and y_i being the left and right singular vectors of $\boldsymbol{A}$, respectively. We will adopt the shorthand notation of $\|\cdot\| \equiv \|\cdot\|_2$. To avoid any confusions between scalars/vectors and matrices, we will represent all matrices in bold, e.g. $\boldsymbol{A}$.

Let f and g be two functions defined on the same space. We say that $f(x) = \mathcal{O}(g(x))$ and $f(x) = \Omega(g(x))$ if and only if there exists a positive real number M and a real number x_0 such that for all $x \ge x_0$,

$$|f(x)| \le M|g(x)| \quad \text{and} \quad |f(x)| \ge M|g(x)|, \tag{2}$$

respectively.

2.2 Model.

The data at hand is a collection of time series with respect to an aggregated metric of interest (e.g. violent crime rates) comprised of both the treated unit and the donor pool outcomes. Suppose we observe $N \ge 2$ units across $T \ge 2$ time periods. We denote T_0 as the number of pre-intervention periods with $1 \le T_0 < T$, rendering $T - T_0$ as the length of the post-intervention stage. Without loss of generality, let the first unit represent the treatment unit – exposed to the intervention of interest at time $t = T_0 + 1$. The remaining donor units, $2 \le i \le N$, are unaffected by the intervention for the entire time period $[T] = \{1, \ldots, T\}$.

Let X_{it} denote the measured value of metric for unit i at time t. We posit

$$X_{it} = M_{it} + \epsilon_{it}, \tag{3}$$

where M_{it} is the deterministic mean while the random variables ϵ_{it} represent zero-mean noise that are independent across i, t. Following the philosophy of latent variable models [16, 28, 6, 22, 23], we further posit that for all $2 \le i \le N$, $t \in [T]$

$$M_{it} = f(\theta_i, \rho_t), \tag{4}$$

where $\theta_i \in \mathbb{R}^{d_1}$ and $\rho_t \in \mathbb{R}^{d_2}$ are latent feature vectors capturing unit and time specific information, respectively, for some $d_1, d_2 \ge 1$; the latent function $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ captures the model relationship. We note that this formulation subsumes popular econometric factor models, such as the one presented in [1], as a special case with (small) constants $d_1 = d_2$ and f as a bilinear function.

The treatment unit obeys the same model relationship during the pre-intervention period. That is, for $t \le T_0$

$$X_{1t} = M_{1t} + \epsilon_{1t}, \tag{5}$$

where $M_{1t} = f(\theta_1, \rho_t)$ for some latent parameter $\theta_1 \in \mathbb{R}^{d_1}$. If unit one was never exposed to the intervention, then the same relationship as (5) would continue to hold during the post-intervention period as well. In essence, we are assuming that the outcome random variables for *all* unaffected units follow the model relationship defined by (5) and (3). Therefore, the "synthetic control" would ideally help estimate the underlying counterfactual means $M_{1t} = f(\theta_1, \rho_t)$ for $T_0 < t \leq T$ by using an appropriate combination of the post-intervention observations from the donor pool since the donor units are immune to the treatment.

To render this feasible, we make the key operating assumption (as done in literature cf. [1, 2, 4]) that the mean vector of the treatment unit over the pre-intervention period, i.e. the vector $M_1^- = [M_{1t}]_{t \leq T_0}$, lies within the span of the mean vectors within the donor pool over the pre-intervention period, i.e. the span of the donor mean vectors $M_i^- = [M_{it}]_{2 \leq i \leq N, t \leq T_0}$ [1]. More precisely, we assume there exists a set of weights $\beta^* \in \mathbb{R}^{N-1}$ such that for all $t \leq T_0$,

$$M_{1t} = \sum_{i=2}^{N} \beta_i^* M_{it}. \tag{6}$$

This is a reasonable and intuitive assumption, utilized in literature, hypothesizing that the treatment unit can be modeled as some combination of the donor pool. In fact, the set of weights β^* are the very definition of a synthetic control.

In order to distinguish the pre- and post-intervention periods, we use the following notation for all (donor) matrices: $\boldsymbol{A} = [\boldsymbol{A}^-, \boldsymbol{A}^+]$, where $\boldsymbol{A}^- = [A_{ij}]_{2 \leq i \leq N, j \in [T_0]}$ and $\boldsymbol{A}^+ = [A_{ij}]_{2 \leq i \leq N, T_0 < j \leq T}$ denote the pre- and post-intervention submatrices, respectively; vectors will be defined in the same manner, i.e. $A_i = [A_i^-, A_i^+]$, where $A_i^- = [A_{it}]_{t \in [T_0]}$ and $A_i^+ = [A_{it}]_{T_0 < t \leq T}$ denote the pre- and post-intervention subvectors, respectively, for the ith donor. Moreover, we will denote all vectors related to the treatment unit with the subscript "1", e.g. $A_1 = [A_1^-, A_1^+]$.

In contrast with the classical synthetic control work, we allow our model to be robust to incomplete observations. To model randomly missing data, the algorithm observes each data point X_{it} in the donor pool with probability $p \in (0, 1]$, independently of all other entries. While the assumption that p is constant across all rows and columns of our observation matrix is standard in literature, our results remain valid even in situations where the probability of observation is dependent on the row and column latent parameters, i.e. $p_{ij} = g(\theta_i, \rho_j) \in (0, 1]$. In such situations, p_{ij} can be estimated as $\hat{p}_{ij}$ using consistent graphon estimation techniques described in a growing body of literature, e.g. see [13, 16, 33, 35]. These estimates can then be used in our analysis presented in Section 4.

3 Algorithm

3.1 Intuition.

We begin by exploring the intuition behind our proposed two-step algorithm: (1) *de-noising the data:* since the singular values of our observation matrix, $\boldsymbol{X} = [X_{it}]_{2 \leq i \leq N, t \in [T]}$, encode both signal and noise, we aim to discover a low rank approximation of $\boldsymbol{X}$ that only incorporates the singular values associated with useful information; simultaneously, this procedure will naturally impute any missing observations. We note that this procedure is similar to the algorithm proposed in [16]. (2) *learning β^*:* using the pre-intervention portion of the de-noised matrix, we learn the linear relationship between the treatment unit and the donor pool prior to estimating the post-intervention counterfactual outcomes. Since our objective is to produce accurate predictions, it is not obvious why the synthetic treatment unit should

---

[1]We note that this is a minor departure from the literature on synthetic control starting in [4] – in literature, the pre-intervention *noisy* observation (rather than the mean) vector X_1, is assumed to be a *convex* (rather than linear) combination of the noisy donor observations. We believe our setup is more reasonable since we do not want to fit noise.

be a convex combination of its donor pool as assumed in [1, 4, 3]. In fact, one can reasonably expect that the treatment unit and some of the donor units may exhibit negative correlations with one another. In light of this intuition, we learn the optimal set of weights via linear regression, allowing for both positive and negative elements.

3.2 Robust algorithm (algorithm 1).

We present the details of our robust method in Algorithm 1 below. The algorithm utilizes two hyperparameters: (1) a thresholding hyperparameter $\mu \geq 0$, which serves as a knob to effectively trade-off between the bias and variance of the estimator, and (2) a regularization hyperameter $\eta \geq 0$ that controls the model complexity. We discuss the procedure for determining the hyperparameters in Section 3.4. To simplify the exposition, we assume the entries of $\boldsymbol{X}$ are bounded by one in absolute value, i.e. $|X_{it}| \leq 1$.

Algorithm 1 Robust synthetic control

Step 1. De-noising the data: singular value thresholding (inspired by [16]).

1. Define $\boldsymbol{Y} = [Y_{it}]_{2 \leq i \leq N, t \in [T]}$ with

$$Y_{it} = \begin{cases} X_{it} & \text{if } X_{it} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

2. Compute the singular value decomposition of $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \sum_{i=1}^{N-1} s_i u_i v_i^T. \tag{8}$$

3. Let $S = \{i : s_i \geq \mu\}$ be the set of singular values above the threshold μ.
4. Define the estimator of $\boldsymbol{M}$ as

$$\hat{\boldsymbol{M}} = \frac{1}{\hat{p}} \sum_{i \in S} s_i u_i v_i^T, \tag{9}$$

   where $\hat{p}$ is the maximum of the fraction of observed entries in $\boldsymbol{X}$ and $\frac{1}{(N-1)T}$.

Step 2. Learning and projecting

1. For any $\eta \geq 0$, let

$$\hat{\beta}(\eta) = \underset{v \in \mathbb{R}^{N-1}}{\arg\min} \left\| Y_1^- - (\hat{\boldsymbol{M}}^-)^T v \right\|^2 + \eta \|v\|^2. \tag{10}$$

2. Define the counterfactual means for the treatment unit as

$$\hat{M}_1 = \hat{\boldsymbol{M}}^T \hat{\beta}(\eta). \tag{11}$$

3.3 Bayesian algorithm: measuring uncertainty (algorithm 2).

In order to quantitatively assess the uncertainty of our model, we will transition from a frequentist perspective to a Bayesian viewpoint. As commonly assumed in literature, we consider a zero-mean, isotropic Gaussian noise model (i.e. $\epsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$) and use the square loss for our cost function. We present the Bayesian method as Algorithm 2. Note that we perform step one of our robust algorithm exactly as in Algorithm 1; as a result, we only detail the alterations of step two in the Bayesian version (Algorithm 2).

Algorithm 2 Bayesian robust synthetic control

Step 2. Learning and projecting

1. Estimate the noise variance via (bias-corrected) maximum likelihood, i.e.

$$\hat{\sigma}^2 = \frac{1}{T_0 - 1} \sum_{t=1}^{T_0} (Y_{1t} - \bar{Y})^2, \tag{12}$$

   where $\bar{Y}$ denotes the pre-intervention sample mean.

2. Compute posterior distribution parameters for an appropriate choice of the prior α:

$$\boldsymbol{\Sigma}_D = \left(\frac{1}{\hat{\sigma}^2} \hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T + \alpha \boldsymbol{I} \right)^{-1} \tag{13}$$

$$\beta_D = \frac{1}{\hat{\sigma}^2} \boldsymbol{\Sigma}_D \hat{\boldsymbol{M}}^- Y_1^-. \tag{14}$$

3. Define the counterfactual means for the treatment unit as

$$\hat{M}_1 = \hat{\boldsymbol{M}}^T \beta_D. \tag{15}$$

4. For each time instance $t \in [T]$, compute the model uncertainty (variance) as

$$\sigma_D^2(\hat{M}_{\cdot,t}) = \hat{\sigma}^2 + \hat{M}_{\cdot,t}^T \boldsymbol{\Sigma}_D \hat{M}_{\cdot,t}, \tag{16}$$

   where $\hat{M}_{\cdot,t} = [\hat{M}_{it}]_{2 \leq i \leq N}$ is the de-noised vector of donor outcomes at time t.

3.4 Algorithmic features: the fine print.

3.4.1 Bounded entries transformation.

Several of our results, as well as the algorithm we propose, assume that the observation matrix is bounded such that $|X_{it}| \leq 1$. For any data matrix, we can achieve this by using the following pre-processing transformation: suppose the entries of $\boldsymbol{X}$ belong to an interval $[a, b]$. Then, one can first pre-process the matrix $\boldsymbol{X}$ by subtracting $(a+b)/2$ from each entry, and dividing by $(b-a)/2$ to enforce that the entries lie in the range $[-1, 1]$. The reverse transformation, which can be applied at the end of the algorithm description above, returns a matrix with values contained in the original range. Specifically, the reverse transformation equates to multiplying the end result by $(b-a)/2$ and adding by $(a+b)/2$.

3.4.2 Solution interpretability.

For the practitioner who seeks a more interpretable solution, e.g. a convex combination of donors as per the original synthetic control estimator of Abadie et. al, we recommend using an ℓ_1-regularization penalty in the learning procedure of step 2. Due to the geometry of LASSO, the resulting estimator will be often be a sparse vector. Specifically, for any $\eta > 0$, we define the LASSO estimator to be

$$\hat{\beta}(\eta) = \arg\min_{v \in \mathbb{R}^{N-1}} \left\| Y_1^- - (\hat{\boldsymbol{M}}^-)^T v \right\|^2 + \eta \|v\|_1.$$

3.4.3 Choosing the hyperparameters.

Here, we discuss several approaches to choosing the hyperparameter μ for the singular values. If it is known a priori that the underlying model is low rank with rank at most k, then it may make sense to choose μ such that $|S| = k$. A data driven approach, however, could be implemented based on cross-validation. Precisely, reserve a portion of the pre-intervention period for validation, and use the rest of the pre-intervention data to produce an estimate $\hat{\beta}(\eta)$ for each of the finitely many choices of μ $(s_1, \ldots, s_{N-1})$. Using each estimate $\hat{\beta}(\eta)$, produce its corresponding treatment unit mean vector over the validation period. Then, select the μ that achieves the minimum MSE with respect to the observed data. Finally, [16] provides a universal approach to picking a threshold; similarly, we also propose another such universal threshold, (20), in Section 4.1. We utilize the data driven approach in our experiments in this work.

The regularization parameter, η, also plays a crucial role in learning the synthetic control and influences both the training and generalization errors. As is often the case in model selection, a popular strategy in estimating the ideal η is to employ cross-validation as described above. However, since time-series data often have a natural temporal ordering with causal effects, we also recommend employing the forward chaining strategy. Although the forward chaining strategy is similar to leave-one-out (LOO) cross-validation, an important distinction is that forward chaining does not break the temporal ordering in the training data. More specifically, for a particular candidate of η at every iteration $t \in [T_0]$, the learning process uses $[Y_{11}, \ldots, Y_{1,t-1}]$ as the training portion while reserving Y_{1t} as the validation point. As before, the average error is then computed and used to evaluate the model (characterized by the choice of η). The forward chaining strategy can also be used to learn the optimal μ.

3.4.4 Scalability.

In terms of scalability, the most computationally demanding procedure is that of evaluating the singular value decomposition (SVD) of the observation matrix. Given the ubiquity of SVD methods in the realm of machine learning, there are well-known techniques that enable computational and storage scaling for SVD algorithms. For instance, both Spark (through alternative least squares) and Tensor-Flow come with built-in SVD implementations. As a result, by utilizing the appropriate computational infrastructure, our de-noising procedure, and algorithm in generality, can scale quite well. Also note that for a low rank structure, we typically only need to compute the top few singular values and vectors. Various truncated-SVD algorithms provide resource-efficient implementations to compute the top k singular values and vectors instead of the complete-SVD.

3.4.5 Low rank hypothesis.

The factor models that are commonly used in the Econometrics literature, cf. [1, 2, 4], often lead to a low rank structure for the underlying mean matrix $\boldsymbol{M}$. When f is nonlinear, $\boldsymbol{M}$ can still be well approximated by a low rank matrix for a large class of functions. For instance, if the latent parameters

assumed values from a bounded, compact set, and if f was Lipschitz continuous, then it can be argued that $\boldsymbol{M}$ is well approximated by a low rank matrix, cf. see [16] for a very simple proof. As the reader will notice, while we establish results for low rank matrix, the results of this work are robust to low rank approximations whereby the approximation error can be viewed as "noise". Lastly, as shown in [32], many latent variable models can be well approximated (up to arbitrary accuracy ϵ) by low rank matrices. Specifically, [32] shows that the corresponding low rank approximation matrices associated with "nice" functions (e.g. linear functions, polynomials, kernels, etc.) are of log-rank.

3.4.6 Covariate information.

Although the algorithm does not appear to rely on any helpful covariate information and the experimental results, presented in Section 5, suggest that it performs on par with that of the original synthetic control algorithm, we want to emphasize that we are not suggesting that practitioners should abandon the use of any additional covariate information or the application of domain knowledge. Rather, we believe that our key algorithmic feature – the de-noising step – may render covariates and domain expertise as luxuries as opposed to necessities for many practical applications. If the practitioner has access to supplementary predictor variables, we propose that step one of our algorithm be used as a pre-processing routine for de-noising the data before incorporating additional information. Moreover, other than the obvious benefit of narrowing the donor pool, domain expertise can also come in handy in various settings, such as determining the appropriate method for imputing the missing entries in the data. For instance, if it is known a priori that there is a trend or periodicity in the time series evolution for the units, it may behoove the practitioner to impute the missing entries using "nearest-neighbors" or linear interpolation.

4 Theoretical Results

In this section, we derive the finite sample and asymptotic properties of the estimators $\hat{\boldsymbol{M}}$ and $\hat{M}_1$. We begin by defining necessary notations and recalling a few operating assumptions prior to presenting the results, with the corresponding proofs relegated to the Appendix. To that end, we re-write (3) in matrix form as $\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{E}$, where $\boldsymbol{E} = [\epsilon_{it}]_{2 \leq i \leq N, t \in [T]}$ denotes the noise matrix. We shall assume that the noise parameters ϵ_{it} are independent zero-mean random variables with bounded second moments. Specifically, for all $2 \leq i \leq N, t \in [T]$,

$$\mathbb{E}[\epsilon_{it}] = 0, \quad \text{and} \quad \text{Var}(\epsilon_{it}) \leq \sigma^2. \tag{17}$$

We shall also assume that the treatment unit noise in (5) obeys (17). Further, we assume the relationship in (6) holds. To simplify the following exposition, we assume that $|M_{ij}| \leq 1$ and $|X_{ij}| \leq 1$.

As previously discussed, we evaluate the accuracy of our estimated means for the treatment unit with respect to the deviation between $\hat{M}_1$ and M_1 measured in ℓ_2-norm, and similarly between $\hat{\boldsymbol{M}}$ and $\boldsymbol{M}$. Additionally, we aim to establish the validity of our pre-intervention linear model assumption (cf. (6)) and investigate how the linear relationship translates over to the post-intervention regime, i.e. if $M_1^- = (\boldsymbol{M}^-)^T \beta^*$ for some β^*, does M_1^+ (approximately) equal to $(\boldsymbol{M}^+)^T \beta^*$? If so, under what conditions? We present our results for the above aspects after a brief motivation of ℓ_2 regularization.

Combatting overfitting. One weapon to combat overfitting is to constrain the learning algorithm to limit the effective model complexity by fitting the data under a simpler hypothesis. This technique is known as regularization, and it has been widely used in practice. To employ regularization, we introduce a complexity penalty term into the objective function (10). For a general regularizer, the objective function takes the form

$$\hat{\beta}(\eta) = \underset{v \in \mathbb{R}^{N-1}}{\arg \min} \left\| Y_1^- - (\hat{\boldsymbol{M}}^-)^T v \right\|^2 + \eta \sum_{j=1}^{N-1} |v_j|^q, \tag{18}$$

for some choice of positive constants η and q. The first term measures the empirical error of the model on the given dataset, while the second term penalizes models that are too "complex" by controlling the "smoothness" of the model in order to avoid overfitting. In general, the impact/trade-off of regularization can be controlled by the value of the regularization parameter η. Via the use of Lagrange multipliers, we note that minimizing (18) is equivalent to minimizing (10) subject to the constraint that

$$\sum_{j=1}^{N-1} |v_j|^q \le c,$$

for some appropriate value of c. When $q = 2$, (18) corresponds to the classical setup known as *ridge regression* or *weight decay*. The case of $q = 1$ is known as the LASSO in the statistics literature; the ℓ_1-norm regularization of LASSO is a popular heuristic for finding a sparse solution. In either case, incorporating an additional regularization term encourages the learning algorithm to output a simpler model with respect to some measure of complexity, which helps the algorithm avoid overfitting to the idiosyncrasies within the observed dataset. Although the training error may suffer from the simpler model, empirical studies have demonstrated that the generalization error can be greatly improved under this new setting. Throughout this section, we will primarily focus our attention on the case of $q = 2$, which maintains our learning objective to be (convex) quadratic in the parameter v so that its exact minimizer can be found in closed form:

$$\hat{\beta}(\eta) = \left(\hat{M}^-(\hat{M}^-)^T + \eta I\right)^{-1} \hat{M}^- Y_1^-. \tag{19}$$

4.1 Imputation analysis.

In this section, we highlight the importance of our de-noising procedure and prescribe a universal threshold (similar to that of [16]) that dexterously distinguishes signal from noise, enabling the algorithm to capture the appropriate amount of useful information (encoded in the singular values of Y) while discarding out the randomness. Due to its universality, the threshold naturally adapts to the amount of structure within M in a purely data-driven manner. Specifically, for any choice of $\omega \in (0.1, 1)$, we find that choosing

$$\mu = (2 + \omega)\sqrt{T(\hat{\sigma}^2 \hat{p} + \hat{p}(1 - \hat{p}))}, \tag{20}$$

results in an estimator with strong theoretical properties for both interpolation and extrapolation (discussed in Section 4.2). Here, $\hat{p}$ and $\hat{\sigma}^2$ denote the unbiased maximum likelihood estimates of p and σ^2, respectively, and can be computed via (9) and (12).

The following Theorems (adapted from Theorems 2.1 and 2.7 of [16]) demonstrate that Step 1 of our algorithm (detailed in Section 3.2) accurately imputes missing entries within our data matrix X when the signal matrix M is either low rank or generated by an $\mathcal{L}$-Lipschitz function. In particular, Theorems 4.1 and 4.2 states that Step 1 produces a consistent estimator of the underlying mean matrix M with respect to the (matrix) mean-squared-error, which is defined as

$$\text{MSE}(\hat{M}) = \frac{1}{(N-1)T}\mathbb{E}\Big[\sum_{i=2}^{N}\sum_{j=1}^{T}(\hat{M}_{ij} - M_{ij})^2\Big]. \tag{21}$$

We say that $\hat{M}$ is a consistent estimator of M if the right-hand side of (21) converges to zero as N and T grow without bound.

The following theorem demonstrates that $\hat{M}$ is a good estimate of M when M is a low rank matrix, particularly when the rank of M is small compared to $(N - 1)p$.

Theorem 4.1. (Theorem 2.1 of [16]) *Suppose that $\boldsymbol{M}$ is rank k. Suppose that $p \geq \frac{T^{-1+\zeta}}{\sigma^2+1}$ for some $\zeta > 0$. Then using μ as defined in (20),*

$$\text{MSE}(\hat{\boldsymbol{M}}) \leq C_1 \sqrt{\frac{k}{(N-1)p}} + \mathcal{O}\Big(\frac{1}{(N-1)T}\Big), \tag{22}$$

where C_1 is a universal positive constant.

Suppose that the latent row and column feature vectors, $\{\theta_i\}$ and $\{\rho_j\}$, belong to some bounded, closed sets $K \subset \mathbb{R}^d$, where d is some arbitrary but fixed dimension. If we assume $f : K \times K \to [-1, 1]$ possesses desirable smoothness properties such as Lipschitzness, then $\hat{\boldsymbol{M}}$ is again a good estimate of $\boldsymbol{M}$.

Theorem 4.2. (Theorem 2.7 of [16]) *Suppose f is a $\mathcal{L}$-Lipschitz function. Suppose that $p \geq \frac{T^{-1+\zeta}}{\sigma^2+1}$ for some $\zeta > 0$. Then using μ as defined in (20),*

$$\text{MSE}(\hat{\boldsymbol{M}}) \leq C(K, d, \mathcal{L}) \frac{(N-1)^{-\frac{1}{d+2}}}{\sqrt{p}} + \mathcal{O}\Big(\frac{1}{(N-1)T}\Big), \tag{23}$$

where $C(K, d, \mathcal{L})$ is a constant depending on K, d, and $\mathcal{L}$.

It is important to observe that the models under consideration for both Theorems 4.1 and 4.2 encompass the mean matrices, $\boldsymbol{M}$, generated as per many of the popular Econometric factor models often considered in literature and assumed in practice. Therefore, de-noising the data serves as an important imputing and filtering procedure for a wide array of applications.

4.2 Forecasting analysis: pre-intervention regime.

Similar to the setting for interpolation, the prediction performance metric of interest is the average mean-squared-error in estimating M_1^- using $\hat{M}_1^-$. Precisely, we define

$$\text{MSE}(\hat{M}_1^-) = \frac{1}{T_0} \mathbb{E}\Big[\sum_{t=1}^{T_0} (M_{1t} - \hat{M}_{1t})^2\Big]. \tag{24}$$

If the right-hand side of (33) approaches zero in the limit as T_0 grows without bound, then we say that $\hat{M}_1^-$ is a consistent estimator of M_1^- (note that our analysis here assumes that only $T_0 \to \infty$).

In what follows, we first state the finite sample bound on the average MSE between $\hat{M}_1^-$ and M_1^- for the most generic setup (Theorem 4.3). As a main Corollary of the result, we specialize the bound in the case where we use our prescribed universal threshold. Finally, we discuss a minor variation of the algorithm where the data is pre-processed, and specialize the above result to establish the consistency of our estimator (Theorem 4.4).

4.2.1 General result.

We provide a finite sample error bound for the most generic setting, i.e. for any choice of the threshold, μ, and regularization hyperparameter, η.

Theorem 4.3. *For any $\eta \geq 0$ and $\mu \geq 0$, the pre-intervention error of the algorithm can be bounded as*

$$\text{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p^2 T_0} \mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \big\|(\hat{p} - p)\boldsymbol{M}^-\big\|\Big)^2 + \frac{2\sigma^2 |S|}{T_0} + \frac{\eta\|\beta^*\|^2}{T_0} + C_2 e^{-cp(N-1)T}. \tag{25}$$

Here, $\lambda_1, \ldots, \lambda_{N-1}$ are the singular values of $p\boldsymbol{M}$ in decreasing order and repeated by multiplicities, with $\lambda^* = \max_{i \notin S} \lambda_i$; C_1, C_2 and c are universal positive constants.

Bias-variance tradeoff. Let us interpret the result by parsing the terms in the error bound. The last term decays exponentially with $(N-1)T$, as long as the fraction of observed entries is such that, on average, we see a super-constant number of entries, i.e. $p(N-1)T \gg 1$. More interestingly, the first two terms highlight the "bias-variance tradeoff" of the algorithm with respect to the singular value threshold μ. Precisely, the size of the set S increases with a decreasing value of the hyperparameter μ, causing the second error term to increase. Simultaneously, however, this leads to a decrease in λ^*. Note that λ^* denotes the aspect of the "signal" within the matrix $\boldsymbol{M}$ that is not captured due to the thresholding through S. On the other hand, the second term, $|S|\sigma^2/T_0$, represents the amount of "noise" captured by the algorithm, but wrongfully interpreted as a signal, during the thresholding process. In other words, if we use a large threshold, then our model may fail to capture pertinent information encoded in $\boldsymbol{M}$; if we use a small threshold, then the algorithm may overfit the spurious patterns in the data. Thus, the hyperparameter μ provides a way to trade-off "bias" (first term) and "variance" (second term).

4.2.2 Goldilocks principle: a universal threshold.

Using the universal threshold defined in (20), we now highlight the prediction power of our estimator for any choice of η, the regularization hyperparameter. As described in Section 4.1, the prescribed threshold automatically captures the "correct" level of information encoded in the (noisy) singular values of $\boldsymbol{Y}$ in a data-driven manner, dependent on the structure of $\boldsymbol{M}$. However, unlike the statements in Theorems 4.1 and 4.2, the following bound does not require $\boldsymbol{M}$ to be low rank or f to be Lipschitz.

Corollary 4.1. *Suppose $p \geq \frac{T^{-1+\zeta}}{\sigma^2+1}$ for some $\zeta > 0$. Let $T \leq \alpha T_0$ for some constant $\alpha > 1$. Then for any $\eta \geq 0$ and using μ as defined in (20), the pre-intervention error is bounded above by*

$$\mathrm{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p}(\sigma^2 + (1-p)) + \mathcal{O}(1/\sqrt{T_0}), \tag{26}$$

where C_1 is a universal positive constant.

As an implication, if $p = (1+\vartheta)\sqrt{T_0}/(1+\sqrt{T_0})$ and $\sigma^2 \leq \vartheta$, we have that $\mathrm{MSE}(\hat{M}_1^-) = \mathcal{O}(1/\sqrt{T_0})$. More generally, Corollary 4.1 shows that by adroitly capturing the signal, the resulting error bound simply depends on the variance of the noise terms, σ^2, and the error introduced due to missing data. Ideally, one would hope to overcome the error term when T_0 is sufficiently large. This motivates the following setup.

4.2.3 Consistency.

We present a straightforward pre-processing step that leads to the consistency of our algorithm. The pre-processing step simply involves replacing the columns of $\boldsymbol{X}$ by the averages of subsets of its columns. This admits the same setup as before, but with the variance for each noise term reduced. An implicit side benefit of this approach is that required SVD step in the algorithm is now applied to a matrix of smaller dimensions.

To begin, partition the T_0 columns of the pre-intervention data matrix $\boldsymbol{X}^-$ into Δ blocks, each of size $\tau = \lfloor T_0/\Delta \rfloor$ except potentially the last block, which we shall ignore for theoretical purposes; in practice, however, the remaining columns can be placed into the last block. Let $B_j = \{(j-1)\tau + \ell : 1 \leq \ell \leq \tau\}$ denote the column indices of $\boldsymbol{X}^-$ within partition $j \in [\Delta]$. Next, we replace the τ columns within each partition by their average, and thus create a new matrix, $\bar{\boldsymbol{X}}^-$, with Δ columns and $N-1$ rows. Precisely, $\bar{\boldsymbol{X}}^- = [\bar{X}_{ij}]_{2 \leq i \leq N, j \in [\Delta]}$ with

$$\bar{X}_{ij} = \frac{1}{\tau} \sum_{t \in B_j} X_{it} \cdot D_{it}, \tag{27}$$

13

where

$$D_{it} = \begin{cases} 1 & \text{if } X_{it} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

For the treatment row, let $\bar{X}_{1j} = \frac{\hat{p}}{\tau} \sum_{t \in B_j} X_{1t}$ for all $j \in [\Delta]^2$. Let $\bar{\boldsymbol{M}}^- = [\bar{M}_{ij}]_{2 \leq i \leq N, j \in [\Delta]}$ with

$$\bar{M}_{ij} = \mathbb{E}[\bar{X}_{ij}] = \frac{p}{\tau} \sum_{t \in B_j} M_{it}. \tag{28}$$

We apply the algorithm to $\bar{\boldsymbol{X}}^-$ to produce the estimate $\hat{\bar{\boldsymbol{M}}}^-$ of $\bar{\boldsymbol{M}}^-$, which is sufficient to produce $\hat{\beta}(\eta)$. This $\hat{\beta}(\eta)$ can be used to produce the post-intervention synthetic control means $\hat{M}_1^+ = [\hat{M}_{1t}]_{T_0 < t \leq T}$ in a similar manner as before [3]: for $T_0 < t \leq T$,

$$\hat{M}_{1t} = \sum_{i=2}^{N} \hat{\beta}_i(\eta) X_{it}. \tag{29}$$

For the pre-intervention period, we produce the estimator $\hat{\bar{M}}_1^- = [\hat{\bar{M}}_{1j}]_{j \in [\Delta]}$: for $j \in [\Delta]$,

$$\hat{\bar{M}}_{1j} = \sum_{i=2}^{N} \hat{\beta}_i(\eta) \hat{\bar{M}}_{ij}. \tag{30}$$

Our measure of estimation error is defined as

$$\text{MSE}(\hat{\bar{M}}_1^-) = \frac{1}{\Delta} \mathbb{E}\Big[\sum_{j=1}^{\Delta} (\bar{M}_{1j} - \hat{\bar{M}}_{1j})^2 \Big]. \tag{31}$$

For simplicity, we will analyze the case where each block contains at least one entry such that $\bar{\boldsymbol{X}}^-$ is completely observed. We now state the following result.

Theorem 4.4. *Fix any* $\gamma \in (0, 1/2)$ *and* $\omega \in (0.1, 1)$. *Let* $\Delta = T_0^{\frac{1}{2} + \gamma}$ *and* $\mu = (2 + \omega)\sqrt{T_0^{2\gamma}(\hat{\sigma}^2 \hat{p} + \hat{p}(1 - \hat{p}))}$. *Suppose* $p \geq \frac{T_0^{-2\gamma}}{\sigma^2 + 1}$ *is known. Then for any* $\eta \geq 0$,

$$\text{MSE}(\hat{\bar{M}}_1^-) = \mathcal{O}(T_0^{-1/2 + \gamma}). \tag{32}$$

We note that the method of [4, Sec 2.3] learns the weights (here $\hat{\beta}(0)$) by pre-processing the data. One common pre-processing proposal is to also aggregate the columns, but the aggregation parameters are chosen by solving an optimization problem to minimize the resulting prediction error of the observations. In that sense, the above averaging of column is a simple, data agnostic approach to achieve a similar effect, and potentially more effectively.

4.3 Forecasting analysis: post-intervention regime.

For the post-intervention regime, we consider the average root-mean-squared-error in measuring the performance of our algorithm. Precisely, we define

$$\text{RMSE}(\hat{M}_1^+) = \frac{1}{\sqrt{T - T_0}} \mathbb{E}\Big[\Big(\sum_{t > T_0}^{T} (M_{1t} - \hat{M}_{1t})^2 \Big)^{1/2} \Big]. \tag{33}$$

---

[2]Although the statement in Theorem 4.4 assumes that an oracle provides the true p, we prescribe practitioners to use $\hat{p}$ since $\hat{p}$ converges to p almost surely by the Strong Law of Large Numbers.

[3]In practice, one can first de-noise $\boldsymbol{X}^+$ via step one of Section 3, and use the entries of $\hat{\boldsymbol{M}}^+$ in (29).

The key assumption of our analysis is that the treatment unit signal can be written as a linear combination of donor pool signals. Specifically, we assume that this relationship holds in the pre-intervention regime, i.e. $M_1^- = (\boldsymbol{M}^-)^T \beta^*$ for some $\beta^* \in \mathbb{R}^{N-1}$ as stated in (6). However, the question still remains: does the same relationship hold for the post-intervention regime and if so, under what conditions does it hold? We state a simple linear algebraic fact to this effect, justifying the approach of synthetic control. It is worth noting that this important aspect has been amiss in the literature, potentially implicitly believed or assumed starting in the work by [4].

Theorem 4.5. *Let* (6) *hold for some* β^*. *Let* $\mathrm{rank}(\boldsymbol{M}^-) = \mathrm{rank}(\boldsymbol{M})$. *Then* $M_1^+ = (\boldsymbol{M}^+)^T \beta^*$.

If we assume that the linear relationship prevails in the post-intervention period, then we arrive at the following error bound.

Theorem 4.6. *Suppose* $p \geq \frac{T^{-1+\varsigma}}{\sigma^2+1}$ *for some* $\varsigma > 0$. *Suppose* $\left\| \hat{\beta}(\eta) \right\|_\infty \leq \psi$ *for some* $\psi > 0$. *Let* $\alpha' T_0 \leq T \leq \alpha T_0$ *for some constants* $\alpha', \alpha > 1$. *Then for any* $\eta \geq 0$ *and using* μ *as defined in* (20), *the post-intervention error is bounded above by*

$$\mathrm{RMSE}(\hat{M}_1^+) \leq \frac{C_1}{\sqrt{p}}(\sigma^2 + (1-p))^{1/2} + \frac{C_2 \|\boldsymbol{M}\|}{\sqrt{T_0}} \cdot \mathbb{E}\left\| \hat{\beta}(\eta) - \beta^* \right\| + \mathcal{O}(1/\sqrt{T_0}),$$

where C_1 *and* C_2 *are universal positive constants.*

Benefits of regularization. In order to motivate the use of regularization, we analyze the error bounds of Theorems 4.3 and 4.6 to observe how the pre- and post-intervention errors react to regularization. As seen from Theorem 4.3, the pre-intervention error *increases* linearly with respect to the choice of η. Intuitively, this increase in pre-intervention error derives from the fact that regularization reduces the model complexity, which biases the model and handicaps its ability to fit the data. At the same time, by restricting the hypothesis space and controlling the "smoothness" of the model, regularization prevents the model from overfitting to the data, which better equips the model to generalize to unseen data. Therefore, a larger value of η *reduces* the post-intervention error. This can be seen by observing the second error term of Theorem 4.6, which is controlled by the expression $\left\| \hat{\beta}(\eta) - \beta^* \right\|$. In words, this error is a function of the learning algorithm used to estimate β^*. Interestingly, [18] demonstrates that there exists an $\eta > 0$ such that

$$\left\| \hat{\beta}(\eta) - \beta^* \right\| \leq \left\| \hat{\beta}(0) - \beta^* \right\|,$$

without any assumptions on the rank of $\hat{\boldsymbol{M}}^-$. In other words, [18] demonstrates that regularization can decrease the MSE between $\hat{\beta}(\eta)$ and the true β^*, thus reducing the overall error. Ultimately, employing ridge regression introduces extraneous bias into our model, yielding a higher pre-intervention error. In exchange, regularization reduces the post-intervention error (due to smaller variance).

4.4 Bayesian analysis.

We turn our attention to a Bayesian treatment of synthetic control. By operating under a Bayesian framework, we allow practitioners to naturally encode domain knowledge into prior distributions while simultaneously avoiding the problem of overfitting. In addition, rather than making point estimates, we can now quantitatively express the uncertainty in our estimates with posterior probability distributions.

We begin by treating β^* as a random variable as opposed to an unknown constant. In this approach, we specify a prior distribution, $p(\beta)$, that expresses our apriori beliefs and preferences about the underlying parameter (synthetic control). Given some new observation for the donor units, our goal is to make predictions for the counterfactual treatment unit on the basis of a set of pre-intervention (training)

data. For the moment, let us assume that the noise parameter σ^2 is a known quantity and that the noise is drawn from a Gaussian distribution with zero-mean; similarly, we temporarily assume $\boldsymbol{M}^-$ is also given. Let us denote the vector of donor estimates as $M_{.t} = [M_{it}]_{2 \leq i \leq N}$; we define $X_{.t}$ similarly. Denoting the pre-intervention data as $D = \{(Y_{1t}, M_{.t}) : t \in [T_0]\}$, the likelihood function $p(Y_1^- \mid \beta, \boldsymbol{M}^-)$ is expressed as

$$p(Y_1^- \mid \beta, \hat{\boldsymbol{M}}^-) = \mathcal{N}((\boldsymbol{M}^-)^T \beta, \sigma^2 \boldsymbol{I}), \tag{34}$$

an exponential of a quadratic function of β. The corresponding conjugate prior, $p(\beta)$, is therefore given by a Gaussian distribution, i.e. $\beta \sim \mathcal{N}(\beta \mid \beta_0, \boldsymbol{\Sigma}_0)$ with mean β_0 and covariance Σ_0. By using a conjugate Gaussian prior, the posterior distribution, which is proportional to the product of the likelihood and the prior, will also be Gaussian. Applying Bayes' Theorem (derivation unveiled in the Appendix), we have that the posterior distribution is $p(\beta \mid D) = \mathcal{N}(\beta_D, \boldsymbol{\Sigma}_D)$ where

$$\boldsymbol{\Sigma}_D = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{M}^- (\boldsymbol{M}^-)^T \right)^{-1} \tag{35}$$

$$\beta_D = \boldsymbol{\Sigma}_D \left(\frac{1}{\sigma^2} \boldsymbol{M}^- Y_1^- + \boldsymbol{\Sigma}_0^{-1} \beta_0 \right). \tag{36}$$

For the remainder of this section, we shall consider a popular form of the Gaussian prior. In particular, we consider a zero-mean isotropic Gaussian with the following parameters: $\beta_0 = 0$ and $\boldsymbol{\Sigma}_0 = \alpha^{-1} \boldsymbol{I}$ for some choice of $\alpha > 0$. Since $\boldsymbol{M}^-$ is unobserved by the algorithm, we use the estimated $\hat{\boldsymbol{M}}^-$, computed as per step one of Section 3, as a proxy; therefore, we redefine our data as $D = \{(Y_{1t}, \hat{M}_{.t}) : t \in [T_0]\}$. Putting everything together, we have that $p(\beta \mid D) = \mathcal{N}(\beta_D, \boldsymbol{\Sigma}_D)$ whereby

$$\boldsymbol{\Sigma}_D = \left(\alpha \boldsymbol{I} + \frac{1}{\sigma^2} \hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T \right)^{-1} \tag{37}$$

$$\beta_D = \frac{1}{\sigma^2} \boldsymbol{\Sigma}_D \hat{\boldsymbol{M}}^- Y_1^- \tag{38}$$

$$= \frac{1}{\sigma^2} \left(\frac{1}{\sigma^2} \hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T + \alpha \boldsymbol{I} \right)^{-1} \hat{\boldsymbol{M}}^- Y_1^-. \tag{39}$$

4.4.1 Maximum a posteriori (MAP) estimation.

By using the zero-mean, isotropic Gaussian conjugate prior, we can derive a point estimate of β^* by maximizing the log posterior distribution, which we will show is equivalent to minimizing the regularized objective function of (10) for a particular choice of η. In essence, we are determining the optimal $\hat{\beta}$ by finding the most probable value of β^* given the data and under the influence of our prior beliefs. The resulting estimate is known as the maximum a posteriori (MAP) estimate.

We begin by taking the log of the posterior distribution, which gives the form

$$\ln p(\beta \mid D) = -\frac{1}{2\sigma^2} \left\| Y_1^- - (\hat{\boldsymbol{M}}^-)^T \beta \right\|^2 - \frac{\alpha}{2} \|\beta\|^2 + \text{const.}$$

Maximizing the above log posterior then equates to minimizing the quadratic regularized error (10) with $\eta = \alpha \sigma^2$. We define the MAP estimate, $\hat{\beta}_{\text{MAP}}$, as

$$\hat{\beta}_{\text{MAP}} = \underset{\beta \in \mathbb{R}^{N-1}}{\arg \max} \ln p(\beta \mid D)$$

$$= \underset{\beta \in \mathbb{R}^{N-1}}{\arg \min} \frac{1}{2} \left\| Y_1^- - (\hat{\boldsymbol{M}}^-)^T \beta \right\|^2 + \frac{\alpha \sigma^2}{2} \|\beta\|^2$$

$$= \left(\hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T + \alpha \sigma^2 \boldsymbol{I} \right)^{-1} \hat{\boldsymbol{M}}^- Y_1^-. \tag{40}$$

With the MAP estimate at hand, we then make predictions of the counterfactual as

$$\hat{M}_1 = \hat{\boldsymbol{M}}^T \hat{\beta}_{\text{MAP}}. \tag{41}$$

Therefore, we have seen that the MAP estimation is equivalent to ridge regression since the introduction of an appropriate prior naturally induces the additional complexity penalty term.

4.4.2   Fully Bayesian treatment.

Although we have treated β^* as a random variable attached with a prior distribution, we can venture beyond point estimates to be fully Bayesian. In particular, we will make use of the posterior distribution over β^* to marginalize over all possible values of β^* in evaluating the predictive distribution over Y_1^-. We will decompose the regression problem of predicting the counterfactual into two separate stages: the *inference* stage in which we use the pre-intervention data to learn the predictive distribution (defined shortly), and the subsequent *decision* stage in which we use the predictive distribution to make estimates. By separating the inference and decision stages, we can readily develop new estimators for different loss functions without having to relearn the predictive distribution, providing practitioners tremendous flexibility with respect to decision making.

Let us begin with a study of the inference stage. We evaluate the predictive distribution over Y_{1t}, which is defined as

$$
\begin{aligned}
p(Y_{1t} \mid \hat{M}_{\cdot t}, D) &= \int p(Y_{1t} \mid \hat{M}_{\cdot t}, \beta) \, p(\beta \mid D) \, d\beta \\
&= \mathcal{N}(\hat{M}_{\cdot t}^T \beta_D, \sigma_D^2),
\end{aligned}
\tag{42}
$$

where

$$\sigma_D^2 = \sigma^2 + \hat{M}_{\cdot,t}^T \boldsymbol{\Sigma}_D \hat{M}_{\cdot,t}. \tag{43}$$

Note that $p(\beta \mid D)$ is the posterior distribution over the synthetic control parameter and is governed by (37) and (39). With access to the predictive distribution, we move on towards the decision stage, which consists of determining a particular estimate $\hat{M}_{1t}$ given a new observation vector $X_{\cdot t}$ (used to determine $\hat{M}_{\cdot t}$). Consider an arbitrary loss function $L(Y_{1t}, g(\hat{M}_{\cdot t}))$ for some function g. The expected loss is then given by

$$
\begin{aligned}
\mathbb{E}[L] &= \int \int L(Y_{1t}, g(\hat{M}_{\cdot t})) \cdot p(Y_{1t}, \hat{M}_{\cdot t}) \, dY_{1t} \, d\hat{M}_{\cdot t} \\
&= \int \left(\int L(Y_{1t}, g(\hat{M}_{\cdot t})) \cdot p(Y_{1t} \mid \hat{M}_{\cdot t}) \, dY_{1t} \right) p(\hat{M}_{\cdot t}) \, d\hat{M}_{\cdot t},
\end{aligned}
\tag{44}
$$

and we choose our estimator $\hat{g}(\cdot)$ as the function that minimizes the average cost, i.e.,

$$\hat{g}(\cdot) = \underset{g(\cdot)}{\arg \min} \, \mathbb{E}[L(Y_{1t}, g(\hat{M}_{\cdot t}))]. \tag{45}$$

Since $p(\hat{M}_{\cdot t}) \geq 0$, we can minimize (44) by selecting $\hat{g}(\hat{M}_{\cdot t})$ to minimize the term within the parenthesis for each individual value of Y_{1t}, i.e.,

$$
\begin{aligned}
\hat{M}_{1t} &= \hat{g}(\hat{M}_{\cdot t}) \\
&= \underset{g(\cdot)}{\arg \min} \int L(Y_{1t}, g(\hat{M}_{\cdot t})) \cdot p(Y_{1t} \mid \hat{M}_{\cdot t}) \, dY_{1t}.
\end{aligned}
\tag{46}
$$

As suggested by (46), the optimal estimate $\hat{M}_{1t}$ for a particular loss function depends on the model only through the predictive distribution $p(Y_{1t} \mid \hat{M}_{\cdot t}, D)$. Therefore, the predictive distribution summarizes all of the necessary information to construct the desired Bayesian estimator for any given loss function L.

4.4.3 Bayesian least-squares estimate.

We analyze the case for the squared loss function (MSE), a common cost criterion for regression problems. In this case, we write the expected loss as

$$\mathbb{E}[L] = \int \left(\int (Y_{1t} - g(\hat{M}_{\cdot t}))^2 \cdot p(Y_{1t} \mid \hat{M}_{\cdot t}) \, dY_{1t} \right) p(\hat{M}_{\cdot t}) \, d\hat{M}_{\cdot t}.$$

Under the MSE cost criterion, the optimal estimate is the mean of the predictive distribution, also known as the Bayes' least-squares (BLS) estimate:

$$\begin{aligned}
\hat{M}_{1t} &= \mathbb{E}[Y_{1t} \mid \hat{M}_{\cdot t}, D] \\
&= \int Y_{1t} \, p(Y_{1t} \mid \hat{M}_{\cdot t}, D) dY_{1t} \\
&= \hat{M}_{\cdot t}^T \beta_D.
\end{aligned} \tag{47}$$

Remark 4.6.1. Since the noise variance σ^2 is usually unknown in practice, we can introduce another conjugate prior distribution $p(\beta, 1/\sigma^2)$ given by the Gaussian-gamma distribution. This prior yields a Student's t-distribution for the predictive probability distribution. Alternatively, one can estimate σ^2 via (12).

5 Experiments

We begin by exploring two real-world case studies discussed in [1, 2, 4] that demonstrate the ability of the original synthetic control's algorithm to produce a reliable counterfactual reality. We use the same case-studies to showcase the "robustness" property of our proposed algorithm. Specifically, we demonstrate that our algorithm reproduces similar results even in presence of missing data, and without knowledge of the extra covariates utilized by prior works. We find that our approach, surprisingly, also discovers a few subtle effects that seem to have been overlooked in prior studies. In the following empirical studies, we will employ three different learning procedures as described in the robust synthetic control algorithm: (1) linear regression ($\eta = 0$), (2) ridge regression ($\eta > 0$), and (3) LASSO ($\zeta > 0$).

As described in [1, 2, 3], the synthetic control method allows a practitioner to evaluate the reliability of his or her case study results by running placebo tests. One such placebo test is to apply the synthetic control method to a donor unit. Since the control units within the donor pool are assumed to be unaffected by the intervention of interest (or at least much less affected in comparison), one would expect that the estimated effects of intervention for the placebo unit should be less drastic and divergent compared to that of the treated unit. Ideally, the counterfactuals for the placebo units would show negligible effects of intervention. Similarly, one can also perform exact inferential techniques that are similar to permutation tests. This can be done by applying the synthetic control method to every control unit within the donor pool and analyzing the gaps for every simulation, and thus providing a distribution of estimated gaps. In that spirit, we present the resulting placebo tests (for only the case of linear regression) for the Basque Country and California Prop. 99 case studies below to assess the significance of our estimates.

We will also analyze both case studies under a Bayesian setting. From our results, we see that our predictive uncertainty, captured by the standard deviation of the predictive distribution, is influenced by the number of singular values used in the de-noising process. Therefore, we have plotted the eigenspectrum of the two case study datasets below. Clearly, the bulk of the signal contained within the datasets is encoded into the top few singular values – in particular, the top two singular values. Given that the validation errors computed via forward chaining are nearly identical for low-rank settings (with the exception of a rank-1 approximation), we shall use a rank-2 approximation of the data matrix. In

order to exhibit the role of thresholding in the interplay between bias and variance, we also plot the cases where we use threshold values that are too high (bias) or too low (variance). For each figure, the dotted blue line will represent our posterior predictive means while the shaded light blue region spans one standard deviation on both sides of the mean. As we shall see, our predictive uncertainty is smallest in the neighborhood around the pre-intervention period. However, the level of uncertainty increases as we deviate from the the intervention point, which appeals to our intuition.

In order to choose an appropriate choice of the prior parameter α, we first use forward-chaining for the ridge regression setting to find the optimal regularization hyperparameter η. By observing the expressions of (19) and (40), we see that $\eta = \alpha\sigma^2$ since ridge regression is closely related to MAP estimation for a zero-mean, isotropic Gaussian prior. Consequently, we choose $\alpha = \eta/\hat{\sigma}^2$ where η is the value obtained via forward chaining.

(a) Eigenspectrum of Basque data. **(b)** Eigenspectrum of California data.

5.1   Basque Country

The goal of this case-study is to investigate the effects of terrorism on the economy of Basque Country using the neighboring Spanish regions as the control group. In 1968, the first Basque Country victim of terrorism was claimed; however, it was not until the mid-1970s did the terrorist activity become more rampant [4]. To study the economic ramifications of terrorism on Basque Country, we only use as data the per-capita GDP (outcome variable) of 17 Spanish regions from 1955-1997. We note that in [4], 13 additional predictor variables for each region were used including demographic information pertaining to one's educational status, and average shares for six industrial sectors.

Results. Figure 2a shows that our method (for all three estimators) produces a very similar qualitative synthetic control to the original method even though we do not utilize additional predictor variables. Specifically, the synthetic control resembles the observed GDP in the pre-treatment period between 1955-1970. However, due to the large-scale terrorist activity in the mid-70s, there is a noticeable economic divergence between the synthetic and observed trajectories beginning around 1975. This deviation suggests that terrorist activity negatively impacted the economic growth of Basque Country.

One subtle difference between our synthetic control – for the case of linear and ridge regression – and that of [4] is between 1970-75: our approach suggests that there was a small, but noticeable economic impact starting just prior to 1970, potentially due to first terrorist attack in 1968. Notice, however, that the original synthetic control of [4] diverges only after 1975. Our LASSO estimator's trajectory also agrees with that of the original synthetic control method's, which is intuitive since both estimators seek sparse solutions.

To study the robustness of our approach with respect to missing entries, we discard each data point uniformly at random with probability $1 - p$. The resulting control for different values of p is presented in

Figure 2b suggesting the robustness of our (linear) algorithm. Finally, we produce Figure 2c by applying our algorithm without the de-noising step. As evident from the Figure, the resulting predictions suffer drastically, reinforcing the value of de-noising. Intuitively, using an appropriate threshold μ equates to selecting the correct model complexity, which helps safeguard the algorithm from potentially overfitting to the training data.
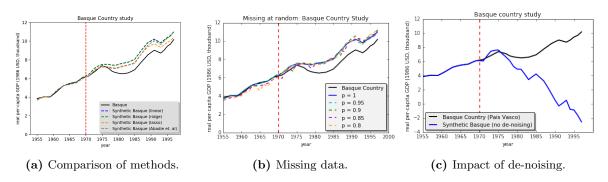

| **(a)** Comparison of methods. | **(b)** Missing data. | **(c)** Impact of de-noising. |

Figure 2: Trends in per-capita GDP between Basque Country vs. synthetic Basque Country.

Placebo tests. We begin by applying our robust algorithm to the Spanish region of Cataluna, a control unit that is not only similar to Basque Country, but also exposed to a much lower level of terrorism [2]. Observing both the synthetic and observed economic evolutions of Cataluna in Figure 3a, we see that there is no identifiable treatment effect, especially compared to the divergence between the synthetic and observed Basque trajectories. We provide the results for the regions of Aragon and Castilla Y Leon in Figures 3b and 3c.

| **(a)** Cataluna. | **(b)** Aragon. | **(c)** Castilla Y Leon. |

Figure 3: Trends in per-capita GDP for placebo regions.

Finally, similar to [2], we plot the differences between our estimates and the observations for Basque Country and all other regionals, individually, as placebos. Note that [2] excluded five regions that had a poor pre-intervention fit but we keep all regions. Figure 4a shows the resulting plot for all regions with the solid black line being Basque Country. This plot helps visualize the extreme post-intervention divergence between the predicted means and the observed values for Basque. Up until about 1990, the divergence for Basque Country is the most extreme compared to all other regions (placebo studies) lending credence to the belief that the effects of terrorism on Basque Country were indeed significant. Refer to Figure 4b for the same test but with Madrid and Balearic Islands excluded, as per [2]. The conclusions drawn should remain the same, pointing to the robustness of our approach.
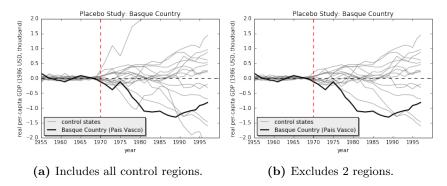
(a) Includes all control regions.   **(b)** Excludes 2 regions.

Figure 4: Per-capita GDP gaps for Basque Country and control regions.

Bayesian approach. We plot the resulting Bayesian estimates in the figures below under varying thresholding conditions. It is interesting to note that our uncertainty grows dramatically once we include more than two singular values in the thresholding process. This confirms what our theoretical results indicated earlier: choosing a smaller threshold, μ, would lead to a greater number of singular values retained which results in higher variance. On the other hand, notice that just selecting 1 singular value results in an apparently biased estimate which is overestimating the synthetic control. It appears that selecting the top two singular values balance the bias-variance tradeoff the best and is also agrees with our earlier finding that the data matrix appears to be of rank 2 or 3. Note that in this setting, we would find it hard to reject the null-hypothesis because the observations for the treated unit lie within the uncertainty band of the estimated synthetic control.
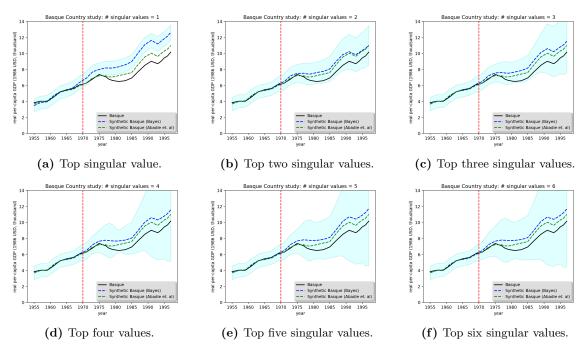

(a) Top singular value.   **(b)** Top two singular values.   **(c)** Top three singular values.

(d) Top four values.   **(e)** Top five singular values.   **(f)** Top six singular values.

Figure 5: Trends in per-capita GDP between Basque Country vs. synthetic Basque Country.

21

5.2 California Anti-tobacco Legislation

We study the impact of California's anti-tobacco legislation, Proposition 99, on the per-capita cigarette consumption of California. In 1988, California introduced the first modern-time large-scale anti-tobacco legislation in the United States [1]. To analyze the effect of California's anti-tobacco legislation, we use the annual per-capita cigarette consumption at the state-level for all 50 states in the United States, as well as the District of Columbia, from 1970-2015. Similar to the previous case study, [4] uses 6 additional observable covariates per state, e.g. retail price, beer consumption per capita, and percentage of individuals between ages of 15-24, to predict their synthetic California. Furthermore, [4] discarded 12 states from the donor pool since some of these states also adopted anti-tobacco legislation programs or raised their state cigarette taxes, and discarded data after the year 2000 since many of the control units had implemented anti-tobacco measures by this point in time.

Results. As shown in Figure 6a, in the pre-intervention period of 1970-88, our control matches the observed trajectory. Post 1988, however, there is a significant divergence suggesting that the passage of Prop. 99 helped reduce cigarette consumption. Similar to the Basque case-study, our estimated effect is similar to that of [4]. As seen in Figure 6b, our algorithm is again robust to randomly missing data.
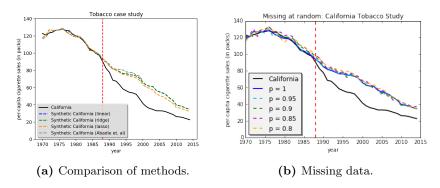

(a) Comparison of methods.   **(b)** Missing data.

Figure 6: Trends in per-capita cigarette sales between California vs. synthetic California.

Placebo tests. We now proceed to apply the same placebo tests to the California Prop 99 dataset. Figures 7a, 7b, and 7c are three examples of the applied placebo tests on the remaining states (including District of Columbia) within the United States.
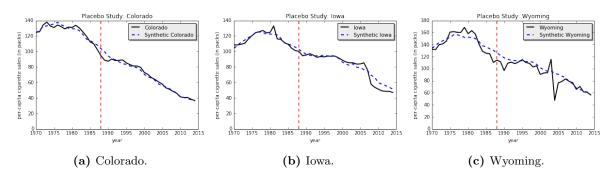

(a) Colorado.   **(b)** Iowa.   **(c)** Wyoming.

Figure 7: Placebo Study: trends in per-capita cigarette sales for Colorado, Iowa, and Wyoming.

Finally, similar to [1], we plot the differences between our estimates and the actual observations for California and all other states, individually, as placebos. Note that [1] excluded twelve states but we keep all states. Figure 8a shows the resulting plot for all states with the solid black line being California.

This plot helps visualize the extreme post-intervention divergence between the predicted means and the observed values for California. Up until about 1995, the divergence for California was clearly the most significant and consistent outlier compared to all other regions (placebo studies) lending credence to the belief that the effects of Proposition 99 were indeed significant. Refer to Figure 8b for the same test but with the same twelve states excludes as in [1]. Just like the Basque Country case study, the exclusion of states should not affect the conclusions drawn.

(a) Includes all donors.   **(b)** Excludes 12 states.

Figure 8: Per-capita cigarette sales gaps in California and control regions.

Bayesian approach. Similar to the Basque Country case study, our predictive uncertainty increases as the number of singular values used in the learning process exceeds two. In order to gain some new insight, however, we will focus our attention to the resulting figure associated with three singular values, which is particularly interesting. Specifically, we observe that our predictive means closely match the counterfactual trajectory produced by the classical synthetic control method in both the pre- and post-intervention periods (up to year 2000), and yet our uncertainty for this estimate is significantly greater than our uncertainty associated with the estimate produced using two singular values. As a result, it may be possible that the classical synthetic control method overestimated the effect of Prop. 99, even though the legislation did probably discourage the consumption of cigarettes – a conclusion reached by both our robust approach and the classical approach.
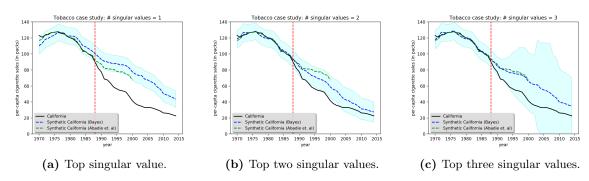

(a) Top singular value.   **(b)** Top two singular values.   **(c)** Top three singular values.

Figure 9: Trends in per-capita cigarette sales between California vs. synthetic California.

Remark 5.0.1. We note that in [3], the authors ran two robustness tests to examine the sensitivity of their results (produced via the original synthetic control method) to alterations in the estimated convex weights – recall that the original synthetic control estimator produces a β^* that lies within the simplex. In particular, the authors first iteratively reproduced a new synthetic West Germany by removing one of the countries that received a positive weight in each iteration, demonstrating that their synthetic model is fairly robust to the exclusion of any particular country with positive weight. Furthermore, [3] examined the trade-off between the original method's ability to produce a good estimate and the sparsity

of the given donor pool. In order to examine this tension, the authors restricted their synthetic West Germany to be a convex combination of only four, three, two, and a single control country, respectively, and found that, relative to the baseline synthetic West Germany (composed of five countries), the degradation in their goodness of fit was moderate.

5.3 Synthetic simulations

We conduct synthetic simulations to establish the various properties of the estimates in both the pre- and post-intervention stages.

Experimental setup. For each unit $i \in [N]$, we assign latent feature θ_i by drawing a number uniformly at random in $[0, 1]$. For each time $t \in [T]$, we assign latent variable $\rho_t = t$. The mean value $m_{it} = f(\theta_i, \rho_t)$. In the experiments described in this section, we use the following:

$$f(\theta_i, \rho_t) = \theta_i + (0.3 \cdot \theta_i \cdot \rho_t/T) * (\exp^{\rho_t/T}) +$$
$$\cos(f_1\pi/180) + 0.5\sin(f_2\pi/180) + 1.5\cos(f_3\pi/180) - 0.5\sin(f_4 * \pi/180)$$

where f_1, f_2, f_3, f_4 define the periodicities: $f_1 = \rho_t \mod (360), f_2 = \rho_t \mod (180), f_3 = 2 \cdot \rho_t \mod (360), f_4 = 2.0 \cdot \rho_t \mod (180)$. The observed value X_{it} is produced by adding i.i.d. Gaussian noise to mean with zero mean and variance σ^2. For this set of experiments, we use $N = 100, T = 2000$, while assuming the treatment was performed at $t = 1600$.
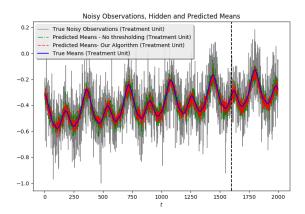

Figure 10: Treatment unit: noisy observations (gray) and true means (blue) and the estimates from our algorithm (red) and one where no singular value thresholding is performed (green). The plots show all entries normalized to lie in range $[-1, 1]$. Notice that the estimates in red generated by our model are much better at estimating the true underlying mean (blue) when compared to an algorithm which performs no singular value thresholding.
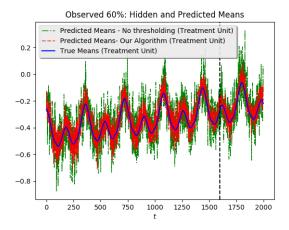
Figure 11: Same dataset as shown in Figure 10 but with 40% data missing at random. Treatment unit: not showing the noisy observations for clarity; plotting true means (blue) and the estimates from our algorithm (red) and one where no singular value thresholding is performed (green). The plots show all entries normalized to lie in range $[-1, 1]$.

Training error approximates generalization error. For the first experimental study, we analyze the relationship between the pre-intervention MSE (training error) and the post-intervention MSE (generalization error). As seen in Table 1, the post-intervention MSE closely matches that of the pre-intervention MSE for varying noise levels, σ^2. Thus suggesting efficacy of our algorithm. Figures 10 and 11 plot the estimates of algorithm with no missing data (Figure 10) and with 40% randomly missing data (Figure 11) on the same underlying dataset. All entries in the plots were normalized to lie within $[-1, 1]$. These plots confirm the robustness of our algorithm. Our algorithm outperforms the algorithm with no singular value thresholding under all proportions of missing data. The estimates from the algorithm which performs no singular value thresholding (green) degrade significantly with missing data while our algorithm remains robust.

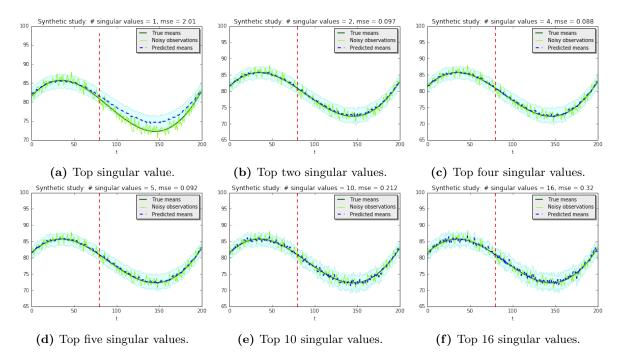**Table 1:** Training vs. generalization error

Noise	Training error	Generalization error
3.1	0.48	0.53
2.5	0.31	0.34
1.9	0.19	0.22
1.3	0.09	0.1
0.7	0.027	0.03
0.4	0.008	0.009
0.1	0.0005	0.0006

Benefits of de-noising. We now analyze the benefit of de-noising the data matrix, which is the main contribution of this work compared to the prior work. Specifically, we study the generalization error of method using de-noising via thresholding and without thresholding as in prior work. The results summarized in Table 2 show that for range of parameters the generalization error with de-noising is consistency better than that without de-noising.

25

Table 2: Impact of thresholding

Noise	De-noising error	No De-noising error
3.1	0.122	0.365
2.5	0.079	0.238
1.9	0.046	0.138
1.6	0.032	0.098
1	0.013	0.038
0.7	0.006	0.018
0.4	0.002	0.005

Bayesian approach. From the synthetic simulations (figures below), we see that the number of singular values included in the thresholding process plays a crucial role in the model's prediction capabilities. If not enough singular values are used, then there is a significant loss of information (high bias) resulting in a higher MSE. On the other hand, if we include too many singular values, then the model begins to overfit to the dataset by misinterpreting noise for signal (high variance). As emphasized before, the goal is to find the simplest model that both fits the data and is also plausible, which is achieved when four singular values are employed.

(a) Top singular value.

(b) Top two singular values.

(c) Top four singular values.

(d) Top five singular values.

(e) Top 10 singular values.

(f) Top 16 singular values.

6 Conclusion

The classical synthetic control method is recognized as a powerful and effective technique for causal inference for comparative case studies. In this work, we motivate a robust synthetic control algorithm, which attempts to improve on the classical method in the following regimes: (a) randomly missing data and (b) large levels of noise. We also demonstrate that the algorithm performs well even in the absence of covariate or expert information, but do *not* propose ignoring information which may eliminate "bad" donors. Our data-driven algorithm, and its Bayesian counterpart, uses singular value thresholding

to impute missing data and "de-noise" the observations. Once "de-noised", we use regularized linear regression to determine the synthetic control. Motivating our algorithm is a modeling framework, specifically the Latent Variable Model, which is a generalization of the various factor models used in related work. We establish finite-sample bounds on the MSE between the estimated "synthetic" control and the latent *true* means of the treated unit of interest. In situations with plentiful data, we show that a simple data aggregation method can lead to an asymptotically consistent estimator. Experiments on synthetically generated data (where the *truth* is known) and on real-world case-studies allow us to demonstrate the promise of our approach, which is an improvement over the classical method.

References

[1] A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californiaâs tobacco control program. *Journal of the American Statistical Association*, 2010.

[2] A. Abadie, A. Diamond, and J. Hainmueller. Synth: An r package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 2011.

[3] A. Abadie, A. Diamond, and J. Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 2014.

[4] A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 2003.

[5] B. Adhikari and J. Alm. Evaluating the economic effects of flat tax reforms using synthetic control methods. *Southern Economic Association*, 2016.

[6] D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

[7] S. Athey, M. Bayati, N. Doudchenko, and G. Imbens. Matrix completion methods for causal panel data models. 2017.

[8] S. Athey and G. Imbens. The state of applied econometrics - causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32, 2016.

[9] H. Aytug, M. Kutuk, A. Oduncu, and S. Togan. Twenty years of the eu-turkey customs union: A synthetic control method analysis. *Journal of Common Market Studies*, 2016.

[10] BallotPedia. California proposition 63, background checks for ammunition purchases and large-capacity ammunition magazine ban (2016). *www.ballotpedia.org*, 2016.

[11] A. Billmeier and T. Nannicini. Assessing economic liberalization episodes: A synthetic control approach. *The Review of Economics and Statistics*, 2013.

[12] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] C. Borgs, J. Chayes, C. E. Lee, and D. Shah. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. *Advances in Neural Information Processing Systems*, 2017.

[14] K. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. Scott. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 2015.

[15] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *CoRR*, abs/0805.4471, 2008.

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

[17] N. Doudchenko and G. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. *NBER Working Paper No. 22791*, 2016.

[18] R. Farebrother. Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):248–250, 1976.

[19] B. Ferman and C. Pinto. Revisiting the synthetic control estimator. 2016.

[20] B. Ferman, C. Pinto, and V. Possebom. Cherry picking with synthetic controls. 2016.

[21] J. Gardeazabal and A. Vega-Bayo. An empirical comparison between the synthetic control method and hsiao et al.'s panel data approach to program evaluation. *Journal of Applied Econometrics*, 2016.

[22] D. N. Hoover. Relations on probability spaces and arrays of random variables. 1979.

[23] D. N. Hoover. Row-columns exchangeability and a generalized model for exchangeability. *Exchangeability in probability and statistics*, (281-291), 1981.

[24] C. Hsiao. *Analysis of panel data.* Cambridge University Press, 2014.

[25] C. Hsiao, H. Ching, and S. Wan. A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 2011.

[26] Jha, S. K., and R. D. S. Yadava. Denoising by singular value decomposition and its application to electronic nose data processing. *IEEE Sensors Journal*, 11:35–44, June 2010.

[27] N. Kreif, R. Grieve, D. Hangartner, A. Turner, S. Nikolova, and M. Sutton. Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, 2015.

[28] C. E. Lee, Y. Li, D. Shah, and D. Song. Blind regression via nearest neighbors under latent variable models. *Advances in Neural Information Processing Systems*, 2016.

[29] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[30] P. McGreevy. California voters approve gun control measure proposition 63. *Los Angeles Times*, Nov. 2016.

[31] J. Saunders, R. Lundberg, A. Braga, G. Ridgeway, and J. Miles. A synthetic control approach to evaluating place-based crime interventions. *Journal of Quantitative Criminology*, 2014.

[32] M. Udell and A. Townsend. Nice latent variable models have log-rank. 2017.

[33] P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *https://arxiv.org/abs/1309.5936*.

[34] Y. Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1), 2017.

[35] J. Yang, Q. Han, and E. M. Airoldi. Nonparametric estimation and testing of exchangeable graph models. *Journal of Machine Learning Research, Conference and Workshop Proceedings*, 33:1060–1067.

A Useful Theorems

We present useful theorems that we will frequently employ in our proofs.

Theorem A.1. Perturbation of singular values.
Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $m \times n$ matrices. Let $k = \min\{m, n\}$. Let $\lambda_1, \ldots, \lambda_k$ be the singular values of $\boldsymbol{A}$ in decreasing order and repeated by multiplicities, and let $\tau_1, \ldots, \tau_k$ be the singular values of $\boldsymbol{B}$ in decreasing order and repeated by multiplicities. Let $\delta_1, \ldots, \delta_k$ be the singular values of $\boldsymbol{A} - \boldsymbol{B}$, in any order but still repeated by multiplicities. Then,

$$\max_{1 \leq i \leq k} |\lambda_i - \tau_i| \leq \max_{1 \leq i \leq k} |\delta_i|.$$

References for the proof of the above result can be found in [16], for example.

Theorem A.2. Poincaré separation Theorem.
Let $\boldsymbol{A}$ be a symmetric $n \times n$ matrix. Let $\boldsymbol{B}$ be the $m \times m$ matrix with $m \leq n$, where $\boldsymbol{B} = \boldsymbol{P}^T \boldsymbol{A} \boldsymbol{P}$ for some orthogonal projection matrix $\boldsymbol{P}$. If the eigenvalues of $\boldsymbol{A}$ are $\sigma_1 \leq \ldots \leq \sigma_n$, and those of $\boldsymbol{B}$ are $\tau_1 \leq \ldots \leq \tau_m$, then for all $j < m + 1$,

$$\sigma_j \leq \tau_j \leq \sigma_{n-m+j}.$$

Remark A.2.1. In the case where $\boldsymbol{B}$ is the principal submatrix of $\boldsymbol{A}$ with dimensions $(n-1) \times (n-1)$, the above Theorem is also known as Cauchy's interlacing law.

Theorem A.3. Bernstein's Inequality.
Suppose that $X_1, \ldots, X_n$ are independent random variables with zero mean, and M is a constant such that $|X_i| \leq M$ with probability one for each i. Let $S := \sum_{i=1}^n X_i$ and $v := Var(S)$. Then for any $t \geq 0$,

$$\mathbb{P}(|S| \geq t) \leq 2 \exp\left(-\frac{3t^2}{6v + 2Mt}\right).$$

Theorem A.4. Hoeffding's Inequality.
Suppose that $X_1, \ldots, X_n$ are independent random variables that are strictly bounded by the intervals $[a_i, b_i]$. Let $S := \sum_{i=1}^n X_i$. Then for any $t > 0$,

$$\mathbb{P}(|S - \mathbb{E}[S]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Theorem A.5. Theorem 3.4 of [16]
Take any two numbers m and n such that $1 \leq m \leq n$. Suppose that $\boldsymbol{A} = [A_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$ is a matrix whose entries are independent random variables that satisfy, for some $\delta^2 \in [0, 1]$,

$$\mathbb{E}[A_{ij}] = 0, \quad \mathbb{E}[A_{ij}^2] \leq \delta^2, \quad and \quad |A_{ij}| \leq 1 \quad a.s.$$

Suppose that $\delta^2 \geq n^{-1+\zeta}$ for some $\zeta > 0$. Then, for any $\omega \in (0, 1)$,

$$\mathbb{P}(\|\boldsymbol{A}\| \geq (2 + \omega)\delta\sqrt{n}) \leq C(\zeta)e^{-c\delta^2 n},$$

where $C(\zeta)$ depends only on ω and ζ, and c depends only on ω. The same result is true when $m = n$ and A is symmetric or skew-symmetric, with independent entries on and above the diagonal, all other assumptions remaining the same. Lastly, all results remain true if the assumption $\delta^2 \geq n^{-1+\zeta}$ is changed to $\delta^2 \geq n^{-1}(\log n)^{6+\zeta}$.

Remark A.5.1. The proof of Theorem A.5 can be found in [16] under Theorem 3.4.

B   Useful Lemmas

We begin by proving (and providing) a series of useful lemmas that we will frequently use to derive our desired results.

Lemma B.1. *Suppose C is an $m \times n$ matrix composed of an $m \times p$ submatrix A and an $m \times (n-p)$ submatrix B, i.e., $C = \begin{bmatrix} A \mid B \end{bmatrix}$. Then, the spectral (operator) norms of A and B are bounded above by the spectral norm of C,*

$$\max\{\|A\|, \|B\|\} \le \|C\|.$$

Proof. Without loss of generality, we prove the case for $\|A\| \le \|C\|$, since the same argument applies for $\|B\|$. By definition,

$$C^T C = \begin{bmatrix} A^T A & A^T B \\ B^T A & B^T B \end{bmatrix}.$$

Let $\sigma_1, \ldots, \sigma_n$ be the eigenvalues of $C^T C$ in increasing order and repeated by multiplicities. Let $\tau_1, \ldots, \tau_p$ be the eigenvalues of $A^T A$ in increasing order and repeated by multiplicities. By the Poincaré separation Theorem A.2, we have for all $j < p+1$,

$$\sigma_j \le \tau_j \le \sigma_{n-p+j}.$$

Thus, $\tau_p \le \sigma_n$. Since the eigenvalues of $C^T C$ and $A^T A$ are the squared singular values of C and A respectively, we have

$$\sqrt{\tau_p} = \|A\| \le \|C\| = \sqrt{\sigma_n}.$$

We complete the proof by applying an identical argument for the case of $\|B\|$.   ∎

Lemma B.2. *Let A be any m by n matrix, and let $A^\dagger$ be its corresponding pseudoinverse. Then, the matrices $P_1 = AA^\dagger$ and $P_2 = A^\dagger A$ are projection matrices.*

Proof. We first prove that P_1 is a projection matrix. In order to show P_1 is a projection matrix, we must demonstrate that P_1 satisfies two properties: namely, (1) P_1 is symmetric, i.e. $P_1^T = P_1$, and (2) P_1 is idempotent, i.e. $P_1^2 = P_1$.

Let $A = Q_1 \Sigma Q_2^T$ represent the SVD of A, with the pseudoinverse expressed as $A^\dagger = Q_2 \Sigma^+ Q_1^T$. As a result,

$$\begin{aligned} P_1 &= AA^\dagger \\ &= Q_1 \Sigma Q_2^T Q_2 \Sigma^+ Q_1^T \\ &= Q_1 \Sigma \Sigma^+ Q_1^T. \end{aligned}$$

Note that

$$\begin{aligned} P_1^T &= (Q_1 \Sigma \Sigma^+ Q_1^T)^T \\ &= Q_1 \Sigma \Sigma^+ Q_1^T \\ &= P_1, \end{aligned}$$

which proves that P_1 is symmetric. Furthermore,

$$\begin{aligned} P_1^2 &= (Q_1 \Sigma \Sigma^+ Q_1^T) \cdot (Q_1 \Sigma \Sigma^+ Q_1^T) \\ &= Q_1 \Sigma \Sigma^+ \Sigma \Sigma^+ Q_1^T \\ &= Q_1 \Sigma \Sigma^+ Q_1^T \\ &= P_1, \end{aligned}$$

which proves that P_1 is idempotent. The same argument can be applied for P_2.   ∎

Lemma B.3. *The eigenvalues of a projection matrix are 1 or 0.*

Proof. Let λ be an eigenvalue of the projection matrix $\boldsymbol{P}$ for some eigenvector v. Then, by definition of eigenvalues,

$$\boldsymbol{P}v = \lambda v.$$

However, by the idempotent property of projection matrices ($\boldsymbol{P}^2 = \boldsymbol{P}$), if we multiply the above equality by $\boldsymbol{P}$ on the left, then we have

$$\boldsymbol{P}(\boldsymbol{P}v) = \boldsymbol{P}(\lambda v)$$
$$= \lambda^2 v.$$

Since $v \neq 0$, the eigenvalues of $\boldsymbol{P}$ can only be members $\mathbb{R}$ whereby $\lambda^2 = \lambda$. Ergo, we must have that $\lambda \in \{0, 1\}$. ∎

Lemma B.4. *Let $\boldsymbol{A} = \sum_{i=1}^{m} \sigma_i x_i y_i^T$ be the singular value decomposition of $\boldsymbol{A}$ with $\sigma_1, \ldots, \sigma_m$ in decreasing order and with repeated multiplicities. For any choice of $\mu \geq 0$, let $S = \{i : \sigma_i \geq \mu\}$. Define*

$$\hat{\boldsymbol{B}} = \sum_{i \in S} \sigma_i x_i y_i^T.$$

Let $\tau_1, \ldots, \tau_m$ be the singular values of $\boldsymbol{B}$ in decreasing order and repeated by multiplicities, with $\tau^* = \max_{i \notin S} \tau_i$. Then

$$\left\| \hat{\boldsymbol{B}} - \boldsymbol{B} \right\| \leq \tau^* + 2\|\boldsymbol{A} - \boldsymbol{B}\|.$$

Proof. By Theorem A.1, we have that $\sigma_i \leq \tau_i + \|\boldsymbol{A} - \boldsymbol{B}\|$ for all i. Applying triangle inequality, we obtain

$$\left\| \hat{\boldsymbol{B}} - \boldsymbol{B} \right\| \leq \left\| \hat{\boldsymbol{B}} - \boldsymbol{A} \right\| + \|\boldsymbol{A} - \boldsymbol{B}\|$$
$$= \max_{i \notin S} \sigma_i + \|\boldsymbol{A} - \boldsymbol{B}\|$$
$$\leq \max_{i \notin S} \left(\tau_i + \|\boldsymbol{A} - \boldsymbol{B}\| \right) + \|\boldsymbol{A} - \boldsymbol{B}\|$$
$$= \tau^* + 2\|\boldsymbol{A} - \boldsymbol{B}\|.$$

∎

Lemma B.5. *Let $\boldsymbol{A} = \sum_{i=1}^{m} \sigma_i x_i y_i^T$ be the singular value decomposition of $\boldsymbol{A}$. Fix any $\delta > 0$ such that $\mu = (1 + \delta)\|\boldsymbol{A} - \boldsymbol{B}\|$, and let $S = \{i : \sigma_i \geq \mu\}$. Define*

$$\hat{\boldsymbol{B}} = \sum_{i \in S} \sigma_i x_i y_i^T.$$

Then

$$\left\| \hat{\boldsymbol{B}} - \boldsymbol{B} \right\| \leq (2 + \delta)\|\boldsymbol{A} - \boldsymbol{B}\|.$$

Proof. By the definition of μ and hence the set of singular values S, we have that

$$\left\| \hat{\boldsymbol{B}} - \boldsymbol{B} \right\| \leq \left\| \hat{\boldsymbol{B}} - \boldsymbol{A} \right\| + \|\boldsymbol{A} - \boldsymbol{B}\|$$
$$= \max_{i \notin S} \sigma_i + \|\boldsymbol{A} - \boldsymbol{B}\|$$
$$\leq (1 + \delta)\|\boldsymbol{A} - \boldsymbol{B}\| + \|\boldsymbol{A} - \boldsymbol{B}\|$$
$$= (2 + \delta)\|\boldsymbol{A} - \boldsymbol{B}\|.$$

∎

Lemma B.6. *Lemma 3.5 of [16]* Let $\boldsymbol{A} = \sum_{i=1}^{m} \sigma_i x_i y_i^T$ *be the singular value decomposition of* $\boldsymbol{A}$. *Fix any* $\delta > 0$ *and define* $S = \{i : \sigma_i \geq (1 + \delta) \| \boldsymbol{A} - \boldsymbol{B} \| \}$ *such that*

$$\hat{\boldsymbol{B}} = \sum_{i \in S} \sigma_i x_i y_i^T.$$

Then

$$\left\| \hat{\boldsymbol{B}} - \boldsymbol{B} \right\|_F \leq K(\delta) (\| \boldsymbol{A} - \boldsymbol{B} \| \| \boldsymbol{B} \|_*)^{1/2},$$

where $K(\delta) = (4 + 2\delta)\sqrt{2/\delta} + \sqrt{2 + \delta}$.

Proof. The proof can be found in [16]. ∎

C Preliminaries.

To simplify the following exposition, we assume that $|M_{ij}| \leq 1$ and $|X_{ij}| \leq 1$. Recall that all entries of the pre-intervention treatment row are observed such that $Y_1^- = X_1^- = M_1^- + \epsilon_1^-$. On the other hand, every entry within the pre- and post-intervention periods for the donor units are observed independently of the other entries with some arbitrary probability p. Specifically, for all $2 \leq i \leq N$ and $j \in [T]$, we define $Y_{ij} = X_{ij}$ if X_{ij} is observed, and $Y_{ij} = 0$ otherwise. Consequently, observe that for all $i > 1$ and j,

$$\mathbb{E}[Y_{ij}] = pM_{ij}$$

and

$$\begin{aligned}
\mathrm{Var}(Y_{ij}) &= \mathbb{E}[Y_{ij}^2] - (\mathbb{E}[Y_{ij}])^2 \\
&= p\mathbb{E}[X_{ij}^2] - (pM_{ij})^2 \\
&\leq p(\sigma^2 + M_{ij}^2) - (pM_{ij})^2 \\
&= p\sigma^2 + pM_{ij}^2(1 - p) \\
&\leq p\sigma^2 + p(1 - p).
\end{aligned}$$

Recall that $\hat{p}$ denotes the proportion of observed entries in $\boldsymbol{X}$ and $\hat{\sigma}^2$ represents the (unbiased) sample variance computed from the pre-intervention treatment row (12). Given the information above, we define, for any $\omega \in (0.1, 1)$, three events $E_1, E_2,$ and E_3 as

$$\begin{aligned}
E_1 &:= \{|\hat{p} - p| \leq \omega p/z\}, \\
E_2 &:= \{|\hat{\sigma}^2 - \sigma^2| \leq \omega \sigma^2/z\}, \\
E_3 &:= \{\|\boldsymbol{Y} - p\boldsymbol{M}\| \leq (2 + \omega/2)\sqrt{Tq}\},
\end{aligned}$$

where $q = \sigma^2 p + p(1 - p)$; for reasons that will be made clear later, we choose $z = 60(\frac{\sigma^2 + 1}{\sigma^2})$. By Bernstein's Inequality, we have that

$$\mathbb{P}(E_1) \geq 1 - 2e^{-c_1(N-1)Tp},$$

for appropriately defined constant c_1. By Hoeffding's Inequality, we obtain

$$\mathbb{P}(E_2) \geq 1 - 2e^{-c_2 T\sigma^2}$$

for some positive constant c_2. Moreover, by Theorem A.5,

$$\mathbb{P}(E_3) \geq 1 - Ce^{-c_3 Tq}$$

as long as $q = \sigma^2 p + p(1-p) \geq T^{-1+\zeta}$ for some $\zeta > 0$. In other words,

$$p(\sigma^2 + 1) \geq p(\sigma^2 + (1-p))$$
$$\geq T^{-1+\zeta}.$$

Consequently, assuming the event E_3 occurs, we require that $p \geq \frac{T^{-1+\zeta}}{\sigma^2+1}$ for some $\zeta > 0$.

Finally, as previously discussed, we will assume that both N and T grow without bound in our imputation analysis. However, in our forecasting analysis, only $T_0 \to \infty$.

D   Imputation Analysis

In this section, we prove that our key de-noising procedure produces a consistent estimator of the underlying mean matrix, thereby adroitly imputing the missing entries and filtering corrupted observations within our data matrix.

Lemma D.1. *Let $\boldsymbol{M} = [M_{ij}]$ be defined as before. Suppose f is a Lipschitz function with Lipschitz constant $\mathcal{L}$ and the latent row and column feature vectors come from a compact space K of dimension d. Then for any small enough $\delta > 0$,*

$$\|\boldsymbol{M}\|_* \leq \delta(N-1)\sqrt{T} + C(K, d, \mathcal{L})\sqrt{(N-1)T\delta^{-d}},$$

where $C(K, d, L)$ is a constant that depends on K, d, and $\mathcal{L}$.

Proof. The proof is a straightforward adaptation of the arguments from [[16], Lemma 3.6]; however, we provide it here for completeness. By the Lipschitzness assumption, every entry in $\boldsymbol{M} = [M_{ij}] = [f(\theta_i, \rho_j)]$ is Lipschitz in both its arguments, space (i) and time (j). For any $\delta > 0$, it is not hard to see that one can find a finite covering $P_1(\delta)$ and $P_2(\delta)$ of K so that for any $\theta, \rho \in K$, there exists $\theta' \in P_1(\delta)$ and $\rho' \in P_2(\delta)$ such that

$$|f(\theta, \rho) - f(\theta', \rho')| \leq \delta.$$

Without loss of generality, let us consider the case where $P(\delta) = P_1(\delta) = P_2(\delta)$. For every latent row feature θ_i, let $p_1(\theta_i)$ be the unique element in $P(\delta)$ that is closest to θ_i. Similarly, for the latent column feature ρ_j, find the corresponding closest element in $P(\delta)$ and denote it by $p_2(\rho_j)$. Let $\boldsymbol{B} = [B_{ij}]$ be the matrix where $B_{ij} = f(p_1(\theta_i), p_2(\rho_j))$. Using the arguments from above, we have that for all i and j,

$$\|\boldsymbol{M} - \boldsymbol{B}\|_F^2 = \sum_{i,j}(f(\theta_i, \rho_j) - f(p_1(\theta_i), p_2(\rho_j)))^2 \leq (N-1)T\delta^2.$$

Therefore,

$$\|\boldsymbol{M}\|_* \leq \|\boldsymbol{M} - \boldsymbol{B}\|_* + \|\boldsymbol{B}\|_*$$
$$\overset{(a)}{\leq} \sqrt{N-1}\|\boldsymbol{M} - \boldsymbol{B}\|_F + \|\boldsymbol{B}\|_*$$
$$\leq \delta(N-1)\sqrt{T} + \|\boldsymbol{B}\|_*,$$

where (a) follows from the fact that $\|\boldsymbol{Q}\|_* \leq \sqrt{\text{rank}(\boldsymbol{Q})}\|\boldsymbol{Q}\|_F$ for any real-valued matrix $\boldsymbol{Q}$. In order to bound the nuclear norm of $\boldsymbol{B}$, note that (by its construction) for any two columns, say $j, j' \in [N-1]$, if

$p_2(\rho_j) = p_2(\rho'_j)$ then it follows that the columns of j and j' of $\boldsymbol{B}$ are identical. Thus, there can be at most $|P(\delta)|$ distinct columns (and rows) of $\boldsymbol{B}$. In other words, $\text{rank}(\boldsymbol{B}) \leq |P(\delta)|$. Ergo,

$$\|\boldsymbol{B}\|_* \leq \sqrt{|P(\delta)|}\|\boldsymbol{B}\|_F$$
$$\leq \sqrt{|P(\delta)|}\sqrt{(N-1)T}.$$

Due to the Lipschitzness property of f and the compactness of the latent space, it can be shown that $|P(\delta)| \leq C(K, d, \mathcal{L})\delta^{-d}$ where $C(K, d, \mathcal{L})$ is a constant that depends only on K, d, and $\mathcal{L}$ (the Lipschitz constant of f). ∎

Lemma D.2. (Theorem 1.1 of [16]) *Let $\hat{\boldsymbol{M}}$ and $\boldsymbol{M}$ be defined as before. Suppose that $p \geq \frac{T^{-1+\zeta}}{\sigma^2+1}$ for some $\zeta > 0$. Then using μ as defined in (20)*

$$\text{MSE}(\hat{\boldsymbol{M}}) \leq \frac{C_1\|\boldsymbol{M}\|_*}{(N-1)\sqrt{Tp}} + \mathcal{O}\Big(\frac{1}{(N-1)T}\Big),$$

where C_1 is a universal positive constant.

Proof. Let $\delta > 0$ be defined by the relation

$$(1+\delta)\|\boldsymbol{Y} - p\boldsymbol{M}\| = (2+\omega)\sqrt{T\hat{q}},$$

where $\hat{q} = \hat{\sigma}^2\hat{p} + \hat{p}(1-\hat{p})$. Observe that if E_1, E_2, and E_3 happen, then

$$
\begin{aligned}
1+\delta &\geq \frac{(2+\omega)\sqrt{T(\hat{\sigma}^2\hat{p} + \hat{p}(1-\hat{p}))}}{(2+\omega/2)\sqrt{T(\sigma^2 p + p(1-p))}} \\
&\geq \frac{(2+\omega)\sqrt{1-\omega/z}\sqrt{(1-\omega/z)\sigma^2 p + p(1-p-\omega p/z)}}{(2+\omega/2)\sqrt{\sigma^2 p + p(1-p)}} \\
&= \frac{(2+\omega)\sqrt{1-\omega/z}}{2+\omega/2}\sqrt{1 - \frac{\omega}{z}\Big(\frac{\sigma^2+p}{\sigma^2+1-p}\Big)} \\
&\geq \frac{(2+\omega)\sqrt{1-\omega/z}}{2+\omega/2}\sqrt{1 - \frac{\omega}{z}\Big(\frac{\sigma^2+1}{\sigma^2}\Big)} \\
&= \frac{(2+\omega)\sqrt{1-\omega/z}}{2+\omega/2}\sqrt{1 - \frac{\omega}{60}} \\
&\geq \frac{2+\omega}{2+\omega/2}\Big(1 - \frac{\omega}{60}\Big) \\
&\geq \Big(1 + \frac{\omega}{5}\Big)\Big(1 - \frac{\omega}{60}\Big) \\
&\geq 1 + \frac{\omega}{5} - \frac{1}{50}.
\end{aligned}
$$

Let $K(\delta)$ be the constant defined in Lemma B.6. Since $\omega \in (0.1, 1)$, $\delta \geq \frac{10\omega-1}{50} > 0$ and

$$
\begin{aligned}
K(\delta) &= (4+2\delta)\sqrt{2/\delta} + \sqrt{2+\delta} \\
&\leq 4\sqrt{1+\delta}\sqrt{2/\delta} + 2\sqrt{2(1+\delta)} + \sqrt{2(1+\delta)} \\
&= (4\sqrt{2/\delta} + 3\sqrt{2})\sqrt{1+\delta} \\
&\leq C_1\sqrt{1+\delta}
\end{aligned}
$$

where C_1 is a constant that depends only on the choice of ω. By Lemma B.6, if E_1, E_2 and E_3 occur, then

$$
\begin{aligned}
\left\|\hat{p}\hat{M} - pM\right\|_F^2 &\le C_2(1+\delta)\|Y - pM\|\|pM\|_* \\
&\le C_3\sqrt{T\hat{q}}\|pM\|_* \\
&\le C_4\sqrt{Tq}\|pM\|_*
\end{aligned}
$$

for an appropriately defined constant C_4. Therefore,

$$
\begin{aligned}
p^2\left\|\hat{M} - M\right\|_F^2 &\le C_5\hat{p}^2\left\|\hat{M} - M\right\|_F^2 \\
&\le C_5\left\|\hat{p}\hat{M} - pM\right\|_F^2 + C_5(\hat{p}-p)^2\|M\|_F^2 \\
&\le C_6\sqrt{Tq}\|pM\|_* + C_5(\hat{p}-p)^2(N-1)T,
\end{aligned}
$$

where the last inequality follows from the boundedness assumption of M. In general, since $|M_{ij}|$ and $|Y_{ij}| \le 1$,

$$
\begin{aligned}
\left\|\hat{M} - M\right\|_F &\le \left\|\hat{M}\right\|_F + \|M\|_F \\
&\le \sqrt{|S|}\left\|\hat{M}\right\| + \|M\|_F \\
&= \frac{\sqrt{|S|}}{\hat{p}}\|Y\| + \|M\|_F \\
&\le (N-1)^{3/2}T\|Y\| + \|M\|_F \\
&\le (N-1)^{3/2}T\sqrt{(N-1)T} + \sqrt{(N-1)T} \\
&\le 2(N-1)^2T^{3/2}.
\end{aligned}
$$

Let $E := E_1 \cap E_2 \cap E_3$. Applying DeMorgan's Law and the Union Bound,

$$
\begin{aligned}
\mathbb{P}(E^c) &= \mathbb{P}(E_1^c \cup E_2^c \cup E_e^c) \\
&\le \mathbb{P}(E_1^c) + \mathbb{P}(E_2^c) + \mathbb{P}(E_3^c) \\
&\le C_7 e^{-c_8 T(p(N-1)+\sigma^2+q)} \\
&= C_7 e^{-c_8\phi T}, \tag{48}
\end{aligned}
$$

where we define $\phi := p(N-1) + \sigma^2 + q$ and C_7, c_8 are appropriately defined. Observe that $\mathbb{E}(\hat{p}-p)^2 = \frac{p(1-p)}{(N-1)T}$. Thus, by the law of total probability and noting that $\mathbb{P}(E) \le 1$ (for appropriately defined constants),

$$
\begin{aligned}
\mathbb{E}\left\|\hat{M} - M\right\|_F^2 &\le \mathbb{E}\left[\left\|\hat{M} - M\right\|_F^2 \mid E\right] + \mathbb{E}\left[\left\|\hat{M} - M\right\|_F^2 \mid E^c\right]\mathbb{P}(E^c) \\
&\le C_6 p^{-1}\sqrt{Tq}\|M\|_* + C_5 p^{-1}(1-p) + C_9(N-1)^4 T^3 e^{-c_8\phi T} \\
&= C_6 p^{-1/2}T^{1/2}(\sigma^2 + (1-p))^{1/2}\|M\|_* + C_5 p^{-1}(1-p) + C_9(N-1)^4 T^3 e^{-c_8\phi T}.
\end{aligned}
$$

Normalizing by $(N-1)T$, we obtain

$$
\text{MSE}(\hat{M}) \le \frac{C_{12}\|M\|_*}{(N-1)\sqrt{T}p} + \frac{C_5(1-p)}{(N-1)Tp} + C_{10}e^{-c_{11}\phi T}.
$$

The proof is complete assuming constants are re-named. ∎

D.1 Proof of Theorem 4.1

Theorem (4.1). (**Theorem 2.1 of [16]**) *Suppose that M is rank k. Suppose that $p \geq \frac{T^{-1+\zeta}}{\sigma^2+1}$ for some $\zeta > 0$. Then using μ as defined in* (20),

$$\mathrm{MSE}(\hat{M}) \leq C_1 \sqrt{\frac{k}{(N-1)p}} + \mathcal{O}\Big(\frac{1}{(N-1)T}\Big),$$

where C_1 is a universal positive constant.

Proof. By the low rank assumption of M, we have that

$$\|M\|_* \leq \sqrt{\mathrm{rank}(M)}\|M\|_F$$
$$\leq \sqrt{k(N-1)T}.$$

The proof follows from a simple application of Lemma D.2. ∎

D.2 Proof of Theorem 4.2

Theorem (4.2). (**Theorem 2.7 of [16]**) *Suppose f is a $\mathcal{L}$-Lipschitz function. Suppose that $p \geq \frac{T^{-1+\zeta}}{\sigma^2+1}$ for some $\zeta > 0$. Then using μ as defined in* (20),

$$\mathrm{MSE}(\hat{M}) \leq C(K,d,\mathcal{L})\frac{(N-1)^{-\frac{1}{d+2}}}{\sqrt{p}} + \mathcal{O}\Big(\frac{1}{(N-1)T}\Big),$$

where $C(K,d,\mathcal{L})$ is a constant depending on K, d, and $\mathcal{L}$.

Proof. Since f is Lipschitz, we invoke Lemmas D.1 and D.2 and choose $\delta = (N-1)^{-1/(d+2)}$. This completes the proof. ∎

E Forecasting Analysis: Pre-Intervention Regime

Here, we will bound the pre-intervention ℓ_2 error of our estimator in order to measure its prediction power.

E.1 Linear Regression

In this section, we will analyze the performance of our algorithm when learning β^* via linear regression, i.e. $\eta = 0$. As a result, we will temporarily drop the dependency on η in this subsection such that $\hat{\beta} = \hat{\beta}(0)$. To ease the notational complexity of the following Lemma E.1 proof, we will make use of the following notations for **only** in this subsection:

$$Q := (M^-)^T \tag{49}$$
$$\hat{Q} := (\hat{M}^-)^T \tag{50}$$

such that

$$M_1^- := Q\beta^* \tag{51}$$
$$\hat{M}_1^- := \hat{Q}\hat{\beta}. \tag{52}$$

Lemma E.1. *Suppose $Y_1^- = M_1^- + \epsilon_1^-$ with $\mathbb{E}[\epsilon_{1j}] = 0$ and $Var(\epsilon_{1j}) \leq \sigma^2$ for all $j \in [T_0]$. Let β^* be defined as in (6) and let $\hat{\beta}$ be the minimizer of (10). Then for any $\mu \geq 0$ and $\eta = 0$,*

$$\mathbb{E}\left\|M_1^- - \hat{M}_1^-\right\|^2 \leq \mathbb{E}\left\|(\boldsymbol{M}^- - \hat{\boldsymbol{M}}^-)^T \beta^*\right\|^2 + 2\sigma^2 |S|. \tag{53}$$

Proof. Recall that for the treatment row, $Y_1^- = M_1^- + \epsilon_1^-$ with $M_1^- = \boldsymbol{Q}\beta^*$. Since $\hat{\beta}$, by definition, minimizes $\left\|Y_1^- - \hat{\boldsymbol{Q}}v\right\|$ for any $v \in \mathbb{R}^{N-1}$, we subsequently have

$$
\begin{aligned}
\left\|M_1^- - \hat{M}_1^-\right\|^2 &= \left\|(Y_1^- - \epsilon_1^-) - \hat{\boldsymbol{Q}}\hat{\beta}\right\|^2 \\
&= \left\|(Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}) + (-\epsilon_1^-)\right\|^2 \\
&= \left\|Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}\right\|^2 + \left\|\epsilon_1^-\right\|^2 + 2\langle -\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}\rangle \\
&\leq \left\|Y_1^- - \hat{\boldsymbol{Q}}\beta^*\right\|^2 + \left\|\epsilon_1^-\right\|^2 + 2\langle -\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}\rangle \\
&= \left\|(\boldsymbol{Q}\beta^* + \epsilon_1^-) - \hat{\boldsymbol{Q}}\beta^*\right\|^2 + \left\|\epsilon_1^-\right\|^2 + 2\langle -\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}\rangle \\
&= \left\|(\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^* + \epsilon_1^-\right\|^2 + \left\|\epsilon_1^-\right\|^2 + 2\langle -\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}\rangle \\
&= \left\|(\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\right\|^2 + 2\left\|\epsilon_1^-\right\|^2 + 2\langle \epsilon_1^-, (\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\rangle + 2\langle -\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}\rangle.
\end{aligned}
$$

Taking expectations, we arrive at the inequality

$$\mathbb{E}\left\|\hat{M}_1^- - M_1^-\right\|^2 \leq \mathbb{E}\left\|(\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\right\|^2 + 2\mathbb{E}\left\|\epsilon_1^-\right\|^2 + 2\mathbb{E}[\langle \epsilon_1^-, (\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\rangle] + 2\mathbb{E}[\langle -\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}\rangle]. \tag{54}$$

We will now deal with the two inner products on the right hand side of equation (54). First, observe that

$$
\begin{aligned}
\mathbb{E}[\langle \epsilon_1^-, (\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\rangle] &= \mathbb{E}[(\epsilon_1^-)^T]\boldsymbol{Q}\beta^* - \mathbb{E}[(\epsilon_1^-)^T\hat{\boldsymbol{Q}}]\beta^* \\
&= -\mathbb{E}[(\epsilon_1^-)^T]\mathbb{E}[\hat{\boldsymbol{Q}}]\beta^* \\
&= 0,
\end{aligned}
$$

since the additive noise terms are independent random variables that satisfy $\mathbb{E}[\epsilon_{ij}] = 0$ for all i and j by assumption, and $\hat{\boldsymbol{Q}} := (\hat{\boldsymbol{M}}^-)^T$ depends only on the noise terms for $i \neq 1$; i.e., the construction of $\hat{\boldsymbol{Q}} := (\hat{\boldsymbol{M}}^-)^T$ excludes the first row (treatment row), and thus depends solely on the donor pool.

For the other inner product term, we begin by recognizing that $(\epsilon_1^-)^T\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger\epsilon_1^-$ is a scalar random variable, which allows us to replace the random variable by its own trace. This is useful since the trace operator is a linear mapping and is invariant under cyclic permutations, i.e., $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$. As a result,

$$
\begin{aligned}
\mathbb{E}[(\epsilon_1^-)^T\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger\epsilon_1^-] &= \mathbb{E}[\mathrm{tr}((\epsilon_1^-)^T\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger\epsilon_1^-)] \\
&= \mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger\epsilon_1^-(\epsilon_1^-)^T)] \\
&= \mathrm{tr}\left(\mathbb{E}[\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger\epsilon_1^-(\epsilon_1^-)^T]\right) \\
&= \mathrm{tr}\left(\mathbb{E}[\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger]\mathbb{E}[\epsilon_1^-(\epsilon_1^-)^T]\right) \\
&\leq \mathrm{tr}\left(\mathbb{E}[\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger]\sigma^2 I\right) \\
&= \sigma^2\mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger)] \\
&\overset{(a)}{=} \sigma^2\mathbb{E}[\mathrm{rank}(\hat{\boldsymbol{Q}})] \\
&\leq \sigma^2 |S|,
\end{aligned}
$$

where (a) follows from the fact that $\hat{Q}\hat{Q}^\dagger$ is a projection matrix by Lemma B.2. As a result, $\hat{Q}\hat{Q}^\dagger$ has rank$(\hat{Q})$ eigenvalues equal to 1 and all other eigenvalues equal to 0 (by Lemma B.3), and since the trace of a matrix is equal to the sum of its eigenvalues, $\mathrm{tr}(\hat{Q}\hat{Q}^\dagger) = \mathrm{rank}(\hat{Q})$. Simultaneously, by the definition of $\hat{Q} := (\hat{M}^-)^T$, we have that the rank of $\hat{Q} := (\hat{M}^-)^T$ is at most $|S|$. Returning to the second inner product and recalling $\hat{\beta} = \hat{Q}^\dagger Y_1^-$,

$$
\begin{aligned}
\mathbb{E}[\langle -\epsilon_1^-, Y_1^- - \hat{Q}\hat{\beta}\rangle] &= \mathbb{E}[(\epsilon_1^-)^T\hat{Q}\hat{\beta}] - \mathbb{E}[(\epsilon_1^-)^T Y_1^-] \\
&= \mathbb{E}[(\epsilon_1^-)^T\hat{Q}\hat{Q}^\dagger Y_1^-] - \mathbb{E}[(\epsilon_1^-)^T]M_1^- - \mathbb{E}[(\epsilon_1^-)^T\epsilon_1^-] \\
&= \mathbb{E}[(\epsilon_1^-)^T\hat{Q}\hat{Q}^\dagger]M_1^- + \mathbb{E}[(\epsilon_1^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_1^-] - \mathbb{E}[(\epsilon_1^-)^T\epsilon_1^-] \\
&\overset{(a)}{=} \mathbb{E}[(\epsilon_1^-)^T]\mathbb{E}[\hat{Q}\hat{Q}^\dagger]M_1^- + \mathbb{E}[(\epsilon_1^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_1^-] - \mathbb{E}[(\epsilon_1^-)^T\epsilon_1^-] \\
&= \mathbb{E}[(\epsilon_1^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_1^-] - \mathbb{E}\|\epsilon_1^-\|^2 \\
&\leq \sigma^2|S| - \mathbb{E}\|\epsilon_1^-\|^2,
\end{aligned}
$$

where (a) follows from the same independence argument used in evaluating the first inner product. Finally, we incorporate the above results to (54) to arrive at the inequality

$$
\begin{aligned}
\mathbb{E}\left\|\hat{M}_1^- - M_1^-\right\|^2 &\leq \mathbb{E}\left\|(Q - \hat{Q})\beta^*\right\|^2 + 2\mathbb{E}\|\epsilon_1^-\|^2 + 2(\sigma^2|S| - \mathbb{E}\|\epsilon_1^-\|^2) \\
&= \mathbb{E}\left\|(Q - \hat{Q})\beta^*\right\|^2 + 2\sigma^2|S|.
\end{aligned}
$$

∎

Lemma E.2. *For $\eta = 0$ and any $\mu \geq 0$, the pre-intervention error of the algorithm can be bounded as*

$$
\mathrm{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p^2 T_0}\mathbb{E}\left(\lambda^* + \|Y - pM\| + \left\|(\hat{p} - p)M^-\right\|\right)^2 + \frac{2\sigma^2|S|}{T_0} + C_2 e^{-cp(N-1)T}. \tag{55}
$$

Here, $\lambda_1, \ldots, \lambda_{N-1}$ are the singular values of pM in decreasing order and repeated by multiplicities, with $\lambda^* = \max_{i\notin S}\lambda_i$; C_1, C_2 and c are universal positive constants.

Proof. Recall that $E_1 := \{|\hat{p} - p| \leq \frac{\omega p}{z}\}$ for some choice of $\omega \in (0.1, 1)$. Thus, under the event E_1,

$$
\begin{aligned}
p\left\|\hat{M}^- - M^-\right\| &\leq C_1\hat{p}\left\|\hat{M}^- - M^-\right\| \\
&\leq C_1\left(\left\|\hat{p}\hat{M}^- - pM^-\right\| + \left\|(\hat{p} - p)M^-\right\|\right) \\
&\overset{(a)}{\leq} C_1\left(\left\|\hat{p}\hat{M} - pM\right\| + \left\|(\hat{p} - p)M^-\right\|\right) \\
&\overset{(b)}{\leq} C_1\left(\lambda^* + 2\|Y - pM\| + \left\|(\hat{p} - p)M^-\right\|\right)
\end{aligned}
$$

where (a) follows from Lemma B.1 and (b) follows from Lemma B.4. In general, since $|M_{ij}|$ and $|Y_{ij}| \leq 1$,

$$
\begin{aligned}
\left\|\hat{M}^- - M^-\right\| &\overset{(a)}{\leq} \left\|\hat{M}\right\| + \left\|M^-\right\| \\
&= \frac{1}{\hat{p}}\|Y\| + \left\|M^-\right\| \\
&\leq (N-1)T\|Y\| + \left\|M^-\right\| \\
&\leq (N-1)T\sqrt{(N-1)T} + \sqrt{(N-1)T_0} \\
&\leq 2((N-1)T)^{3/2}. \tag{56}
\end{aligned}
$$

(a) follows from a simple application of Lemma B.1 and the triangle inequality of operator norms. By the law of total probability and $\mathbb{P}(E_1) \leq 1$,

$$\mathbb{E}\left\|(\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-)^T \beta^*\right\|^2 \leq \mathbb{E}\left[\left\|(\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-)^T \beta^*\right\|^2 \mid E_1\right] + \mathbb{E}\left[\left\|(\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-)^T \beta^*\right\|^2 \mid E_1^c\right]\mathbb{P}(E_1^c)$$

$$\overset{(a)}{\leq} \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\|^2 \mid E_1\right]\|\beta^*\|^2 + \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\|^2 \mid E_1^c\right]\|\beta^*\|^2 \mathbb{P}(E_1^c)$$

$$\leq \frac{C_2}{p^2}\mathbb{E}\left[\left(\lambda^* + 2\|\boldsymbol{Y} - p\boldsymbol{M}\| + \|(\hat{p}-p)\boldsymbol{M}^-\|\right)^2 \mid E_1\right] + C_3((N-1)T)^{3/2}e^{-cp(N-1)T},$$

where (a) follows because the spectral norm is an induced norm, and the last inequality makes use of the results from above. Note that C_2 and C_3 are appropriately defined to depend on β^*. Moreover, for any non-negative valued random variable X and event E with $\mathbb{P}(E) \geq 1/2$,

$$\mathbb{E}[X \mid E] \leq \frac{\mathbb{E}[X]}{\mathbb{P}(E)} \leq 2\mathbb{E}[X]. \tag{57}$$

Using the fact that $\mathbb{P}(E_1) \geq 1/2$ for large enough T, N, we apply Lemma E.1 to obtain (with appropriately defined constants C_4, C_5, c_6)

$$\text{MSE}(\hat{M}_1^-) \leq \frac{1}{T_0}\mathbb{E}\left\|(\boldsymbol{M}^- - \hat{\boldsymbol{M}}^-)^T \beta^*\right\|^2 + \frac{2\sigma^2|S|}{T_0}$$

$$\leq \frac{C_4}{p^2 T_0}\mathbb{E}\left(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \|(\hat{p}-p)\boldsymbol{M}^-\|\right)^2 + \frac{2\sigma^2|S|}{T_0} + C_5 e^{-c_6 p(N-1)T}. \tag{58}$$

The proof is completed assuming we re-label constants C_4, C_5, c_6 as C_1, C_2, and c, respectively. ∎

E.2 Ridge Regression

In this section, we will prove our results for the ridge regression setting, i.e. $\eta > 0$. Let us begin by deriving the closed form expression of $\hat{\beta}(\eta)$.

Derivation of $\hat{\beta}(\eta)$. We derive the closed form solution for $\hat{\beta}(\eta)$ under the new objective function with the additional complexity penalty term:

$$\left\|Y_1^- - (\hat{\boldsymbol{M}}^-)^T v\right\|^2 + \eta\|v\|^2 = (Y_1^-)^T Y_1^- - 2v^T \hat{\boldsymbol{M}}^- Y_1^- + v^T \hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T v + \eta v^T v.$$

Setting the gradient of the above expression to zero and solving for v, we obtain

$$\nabla_v \left\{\left\|Y_1^- - (\hat{\boldsymbol{M}}^-)^T v\right\|^2 + \eta\|v\|^2\right\}_{v = \hat{\beta}(\eta)} = -2\hat{\boldsymbol{M}}^- Y_1^- + 2\hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T v + 2\eta v = 0.$$

Therefore,

$$\hat{\beta}(\eta) = \left(\hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T + \eta \boldsymbol{I}\right)^{-1} \hat{\boldsymbol{M}}^- Y_1^-.$$

Remark E.0.1. To ease the notational complexity of the following Lemmas E.3 and E.5 proofs, we will make use of the following notations for **only** this derivation: Let

$$\boldsymbol{Q} := (\boldsymbol{M}^-)^T \tag{59}$$

$$\hat{\boldsymbol{Q}} := (\hat{\boldsymbol{M}}^-)^T \tag{60}$$

such that

$$M_1^- := \boldsymbol{Q}\beta^* \tag{61}$$

$$\hat{M}_1^- := \hat{\boldsymbol{Q}}\hat{\beta}. \tag{62}$$

40

Lemma E.3. *Let $\boldsymbol{P}_\eta = \hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T$ denote the projection matrix under the quadratic regularization setting. Then, the non-zero singular values of $\boldsymbol{P}_\eta$ are $s_i^2/(s_i^2 + \eta)$ for all $i \in S$.*

Proof. Recall that the singular values of $\boldsymbol{Y}$ are s_i, while the singular values of $\hat{\boldsymbol{Q}}$ are those $s_i \geq \mu$. Let $\hat{\boldsymbol{Q}} = \boldsymbol{U\Sigma V}^T$ be the singular value decomposition of $\hat{\boldsymbol{Q}}$. Since $\boldsymbol{VV}^T = \boldsymbol{I}$, we have that

$$
\begin{aligned}
\boldsymbol{P}_\eta &= \hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T \\
&= \boldsymbol{U\Sigma V}^T(\boldsymbol{V\Sigma}^2\boldsymbol{V}^T + \eta\boldsymbol{I})^{-1}V\boldsymbol{\Sigma U}^T \\
&= \boldsymbol{U\Sigma V}^T(\boldsymbol{V\Sigma}^2\boldsymbol{V}^T + \eta\boldsymbol{VV}^T)^{-1}V\boldsymbol{\Sigma U}^T \\
&= \boldsymbol{U\Sigma V}^T\boldsymbol{V}(\boldsymbol{\Sigma}^2 + \eta\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{V\Sigma U}^T \\
&= \boldsymbol{U\Sigma}(\boldsymbol{\Sigma}^2 + \eta\boldsymbol{I})^{-1}\boldsymbol{\Sigma U}^T \\
&= \boldsymbol{UDU}^T,
\end{aligned}
$$

where

$$
\boldsymbol{D} = \operatorname{diag}\left(\frac{s_1^2}{s_1^2 + \eta}, \ldots, \frac{s_{|S|}^2}{s_{|S|}^2 + \eta}, 0, \ldots, 0\right).
$$

$\blacksquare$

Lemma E.4. *Suppose $Y_1^- = M_1^- + \epsilon_1^-$ with $\mathbb{E}[\epsilon_{1j}] = 0$ and $Var(\epsilon_{1j}) \leq \sigma^2$ for all $j \in [T_0]$. Let β^* be defined as in (6), i.e. $M_1^- = (\boldsymbol{M}^-)^T\beta^*$, and let $\hat{\beta}(\eta)$ be the minimizer of (10). Then for any $\mu \geq 0$ and $\eta > 0$,*

$$
\mathbb{E}\left\|M_1^- - \hat{M}_1^-\right\|^2 \leq \mathbb{E}\left\|(\boldsymbol{M}^- - \hat{\boldsymbol{M}}^-)^T\beta^*\right\|^2 + \eta\|\beta^*\|^2 - \eta\mathbb{E}\left\|\hat{\beta}(\eta)\right\|^2 + 2\sigma^2|S|. \tag{63}
$$

Proof. The following proof is a slight modification for the proof of Lemmas E.1. In particular, observe that $\hat{\beta}(\eta)$ minimizes $\left\|Y_1^- - \hat{\boldsymbol{Q}}v\right\| + \eta\|v\|^2$ for any $v \in \mathbb{R}^{N-1}$. As a result,

$$
\begin{aligned}
&\left\|M_1^- - \hat{M}_1^-\right\|^2 + \eta\left\|\hat{\beta}(\eta)\right\|^2 \\
&= \left\|(Y_1^- - \epsilon_1^-) - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)\right\|^2 + \eta\left\|\hat{\beta}(\eta)\right\|^2 \\
&= \left\|(Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)) + (-\epsilon_1^-)\right\|^2 + \eta\left\|\hat{\beta}(\eta)\right\|^2 \\
&= \left\|Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)\right\|^2 + \eta\left\|\hat{\beta}(\eta)\right\|^2 + \left\|\epsilon_1^-\right\|^2 + 2\langle-\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)\rangle \\
&\leq \left\|Y_1^- - \hat{\boldsymbol{Q}}\beta^*\right\|^2 + \eta\|\beta^*\|^2 + \left\|\epsilon_1^-\right\|^2 + 2\langle-\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)\rangle \\
&= \left\|(\boldsymbol{Q}\beta^* + \epsilon_1^-) - \hat{\boldsymbol{Q}}\beta^*\right\|^2 + \eta\|\beta^*\|^2 + \left\|\epsilon_1^-\right\|^2 + 2\langle-\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)\rangle \\
&= \left\|(\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^* + \epsilon_1^-\right\|^2 + \eta\|\beta^*\|^2 + \left\|\epsilon_1^-\right\|^2 + 2\langle-\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)\rangle \\
&= \left\|(\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\right\|^2 + \eta\|\beta^*\|^2 + 2\left\|\epsilon_1^-\right\|^2 + 2\langle\epsilon_1^-, (\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\rangle + 2\langle-\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)\rangle
\end{aligned}
$$

Taking expectations, we have

$$
\begin{aligned}
&\mathbb{E}\left\|\hat{M}_1^- - M_1^-\right\|^2 \\
&\leq \mathbb{E}\left\|(\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\right\|^2 + \eta\left(\|\beta^*\|^2 - \mathbb{E}\left\|\hat{\beta}(\eta)\right\|^2\right) + 2\mathbb{E}\left\|\epsilon_1^-\right\|^2 + 2\mathbb{E}\langle\epsilon_1^-, (\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\rangle + 2\mathbb{E}\langle-\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}(\eta)\rangle.
\end{aligned}
$$

As before, we have that $\mathbb{E}\langle\epsilon_1^-, (\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta^*\rangle = 0$ by the zero-mean and independence assumptions of the noise random variables. Similarly, note that

$$
\begin{aligned}
\mathbb{E}[(\epsilon_1^-)^T \hat{\boldsymbol{Q}}\hat{\beta}(\eta)] &= \mathbb{E}[(\epsilon_1^-)^T \hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T Y_1^-] \\
&= \mathbb{E}[(\epsilon_1^-)^T \hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T]M_1^- + \mathbb{E}[(\epsilon_1^-)^T \hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^T + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T\epsilon_1^-] \\
&= \mathbb{E}[(\epsilon_1^-)^T \hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T\epsilon_1^-] \\
&= \mathbb{E}[\mathrm{tr}((\epsilon_1^-)^T \hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T\epsilon_1^-)] \\
&= \mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T\epsilon_1^-(\epsilon_1^-)^T)] \\
&= \mathrm{tr}(\mathbb{E}[\hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T\epsilon_1^-(\epsilon_1^-)^T]) \\
&= \mathrm{tr}(\mathbb{E}[\hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T]\mathbb{E}[\epsilon_1^-(\epsilon_1^-)^T]) \\
&\leq \sigma^2\mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{Q}}(\hat{\boldsymbol{Q}}^T\hat{\boldsymbol{Q}} + \eta\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}^T)] \\
&\overset{(a)}{\leq} \sigma^2\mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger)] \\
&\overset{(b)}{=} \sigma^2\mathrm{rank}(\hat{\boldsymbol{Q}}) \\
&\leq \sigma^2|S|,
\end{aligned}
$$

where (a) follows from Lemma E.3, and as before, (b) follows because $\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger$ is a projection matrix. ∎

Lemma E.5. *For any $\eta > 0$ and $\mu \geq 0$, the pre-intervention error of the regularized algorithm can be bounded as*

$$
\mathrm{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p^2 T_0}\mathbb{E}\left(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \|(\hat{p}-p)\boldsymbol{M}^-\|\right)^2 + \frac{2\sigma^2|S|}{T_0} + \frac{\eta\|\beta^*\|^2}{T_0} + C_2 e^{-cp(N-1)T}.
$$

Here, $\lambda_1, \ldots, \lambda_{N-1}$ are the singular values of $p\boldsymbol{M}$ in decreasing order and repeated by multiplicities, with $\lambda^* = \max_{i\notin S}\lambda_i$; C_1, C_2 and c are universal positive constants.

Proof. The proof follows the same arguments as that of Lemma E.2. ∎

E.3 Combining linear and ridge regression.

E.3.1 Proof of Theorem 4.3

Theorem (4.3). *For any $\eta \geq 0$ and $\mu \geq 0$, the pre-intervention error of the algorithm can be bounded as*

$$
\mathrm{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p^2 T_0}\mathbb{E}\left(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \|(\hat{p}-p)\boldsymbol{M}^-\|\right)^2 + \frac{2\sigma^2|S|}{T_0} + \frac{\eta\|\beta^*\|^2}{T_0} + C_2 e^{-cp(N-1)T}.
$$

Here, $\lambda_1, \ldots, \lambda_{N-1}$ are the singular values of $p\boldsymbol{M}$ in decreasing order and repeated by multiplicities, with $\lambda^* = \max_{i\notin S}\lambda_i$; C_1, C_2 and c are universal positive constants.

Proof. The proof follows from a simple amalgamation of Lemmas E.2 and E.5. ∎

E.3.2 Proof of Corollary 4.1

Corollary (4.1). *Suppose $p \geq \frac{T^{-1+\zeta}}{\sigma^2 + 1}$ for some $\zeta > 0$. Let $T \leq \alpha T_0$ for some constant $\alpha > 1$. Then for any $\eta \geq 0$ and using μ as defined in (20), the pre-intervention error is bounded above by*

$$\mathrm{MSE}(\hat{M}_1^-) \le \frac{C_1}{p}(\sigma^2 + (1-p)) + \mathcal{O}(1/\sqrt{T_0}),$$

where C_1 is a universal positive constant.

Proof. Since the singular value threshold $\mu = (2+\omega)\sqrt{T\hat{q}}$, let us define δ so that

$$(1+\delta)\|\boldsymbol{Y} - p\boldsymbol{M}\| = (2+\omega)\sqrt{T\hat{q}},$$

where $\hat{q} = \hat{\sigma}^2\hat{p} + \hat{p}(1-\hat{p})$; recall that $q = \sigma^2 p + p(1-p)$. If E_3 happens, then we know that $\delta \ge 0$. Therefore, assuming E_1, E_2, and E_3 happens, Lemma B.5 states that

$$\begin{aligned}
\left\|\hat{p}\hat{\boldsymbol{M}} - p\boldsymbol{M}\right\| &\le (2+\delta)\|\boldsymbol{Y} - p\boldsymbol{M}\| \\
&\le 2(1+\delta)\|\boldsymbol{Y} - p\boldsymbol{M}\| \\
&= (4+2\omega)\sqrt{T\hat{q}} \\
&\le C_1\sqrt{Tq}
\end{aligned} \tag{64}$$

for an appropriately defined constant C_1. Therefore,

$$\begin{aligned}
p\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\| &\le C_2\hat{p}\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\| \\
&\le C_2\left(\left\|\hat{p}\hat{\boldsymbol{M}}^- - p\boldsymbol{M}^-\right\| + \left\|(\hat{p}-p)\boldsymbol{M}^-\right\|\right) \\
&\overset{(a)}{\le} C_2\left(\left\|\hat{p}\hat{\boldsymbol{M}} - p\boldsymbol{M}\right\| + \left\|(\hat{p}-p)\boldsymbol{M}^-\right\|\right) \\
&\le C_2\left(C_1\sqrt{Tq} + \left\|(\hat{p}-p)\boldsymbol{M}^-\right\|\right)
\end{aligned} \tag{65}$$

where (a) follows from Lemma B.1. Applying the logic that led to (56), we have that, in general,

$$\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\| \le 2((N-1)T)^{3/2}. \tag{66}$$

Let $E := E_1 \cap E_2 \cap E_3$. Further, using the same argument that led to (48), we have

$$\mathbb{P}(E^c) \le C_3 e^{-c_4\phi T}$$

where we define $\phi := p(N-1) + \sigma^2 + q$ and C_3, c_4 are appropriately defined. Thus, by the law of total probability and noting that $\mathbb{P}(E) \le 1$,

$$\begin{aligned}
\mathbb{E}\left\|(\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-)^T\beta^*\right\|^2 &\le \mathbb{E}\left[\left\|(\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-)^T\beta^*\right\|^2 \mid E\right] + \mathbb{E}\left[\left\|(\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-)^T\beta^*\right\|^2 \mid E^c\right]\mathbb{P}(E^c) \\
&\overset{(a)}{\le} \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\|^2 \mid E\right]\|\beta^*\|^2 + \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\|^2 \mid E^c\right]\|\beta^*\|^2\mathbb{P}(E^c) \\
&\le \frac{C_5}{p^2}\mathbb{E}\left[\left(\sqrt{Tq} + \left\|(\hat{p}-p)\boldsymbol{M}^-\right\|\right)^2 \mid E\right] + C_6((N-1)T)^{3/2}e^{-c_4\phi T},
\end{aligned} \tag{67}$$

where (a) follows because the spectral norm is an induced norm and the last inequality makes use of the results from above. Note that C_5 and C_6 are appropriately defined to depend on β^*. Using the fact that $\mathbb{P}(E) \ge 1/2$ for large enough T, N, we apply Lemmas E.1 and E.5 as well as (57) to obtain (with appropriately defined constants C_7, C_8, c_9)

$$\begin{aligned}
\mathrm{MSE}(\hat{M}_1^-) &\le \frac{1}{T_0}\mathbb{E}\left\|(\boldsymbol{M}^- - \hat{\boldsymbol{M}}^-)^T\beta^*\right\|^2 + \frac{2\sigma^2|S|}{T_0} + \frac{\eta\|\beta^*\|^2}{T_0} \\
&\le \frac{C_7}{p^2 T_0}\mathbb{E}\left(\sqrt{Tq} + \left\|(\hat{p}-p)\boldsymbol{M}^-\right\|\right)^2 + \frac{2\sigma^2(N-1)}{T_0} + \frac{\eta\|\beta^*\|^2}{T_0} + C_8 e^{-c_9\phi T}.
\end{aligned} \tag{68}$$

From Jensen's Inequality, $\mathbb{E}|\hat{p} - p| \leq \sqrt{\mathrm{Var}(\hat{p})}$ where $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{(N-1)T}$. Therefore,

$$\mathbb{E}\Big(\sqrt{Tq}\big\|(\hat{p} - p)\boldsymbol{M}^-\big\|\Big) \leq \frac{q^{1/2}\sqrt{p(1-p)}}{\sqrt{N-1}}\big\|\boldsymbol{M}^-\big\|$$

$$\leq \sqrt{qp(1-p)T_0}.$$

At the same time,

$$\mathbb{E}\big\|(\hat{p} - p)\boldsymbol{M}^-\big\|^2 = \mathbb{E}(\hat{p} - p)^2 \cdot \big\|\boldsymbol{M}^-\big\|^2$$

$$\leq \frac{p(1-p)T_0}{T}$$

$$\leq p(1-p).$$

Putting everything together, we arrive at the inequality

$$\mathrm{MSE}(\hat{M}_1^-) \leq \frac{C_7}{p^2 T_0}\Big(qT + p(1-p) + 2\sqrt{qp(1-p)T_0}\Big) + \frac{2\sigma^2(N-1)}{T_0} + \frac{\eta\|\beta^*\|^2}{T_0} + C_8 e^{-c_9\phi T}$$

$$= \frac{C_{10}q}{p^2} + \frac{C_7(1-p)}{pT_0} + \frac{C_{11}(q(1-p))^{1/2}}{p^{3/2}\sqrt{T_0}} + \frac{2\sigma^2(N-1)}{T_0} + \frac{\eta\|\beta^*\|^2}{T_0} + C_8 e^{-c_9\phi T}$$

$$= \frac{C_{10}}{p}(\sigma^2 + (1-p)) + \mathcal{O}(1/\sqrt{T_0}).$$

The proof is complete assuming we re-label C_{10} as C_1. ∎

E.3.3 Proof of Theorem 4.4

Theorem (4.4). *Fix any $\gamma \in (0, 1/2)$ and $\omega \in (0.1, 1)$. Let $\Delta = T_0^{\frac{1}{2}+\gamma}$ and $\mu = (2+\omega)\sqrt{T_0^{2\gamma}(\hat{\sigma}^2\hat{p} + \hat{p}(1-\hat{p}))}$. Suppose $p \geq \frac{T_0^{-2\gamma}}{\sigma^2+1}$ is known. Then for any $\eta \geq 0$,*

$$\mathrm{MSE}(\hat{\bar{M}}_1^-) = \mathcal{O}(T_0^{-1/2+\gamma}).$$

Proof. To establish Theorem 4.4, we shall follow the proof of Corollary 4.1, using the block partitioned matrices instead. Recall that $\tau = T_0/\Delta$ where $\Delta = T_0^{1/2+\gamma}$. For analytical simplicity, we define the random variable

$$D_{it} = \begin{cases} 1 & \text{w.p. } p, \\ 0 & \text{otherwise}, \end{cases}$$

whose definition will soon prove to be useful. As previously described in Section 4, for all $i > 1$ and $j \in [\Delta]$, we define

$$\bar{X}_{ij} = \frac{1}{\tau}\sum_{t \in B_j} X_{it} \cdot D_{it}$$

and

$$\bar{M}_{ij} = \frac{p}{\tau}\sum_{t \in B_j} M_{it}.$$

Let us also define $\bar{\boldsymbol{E}}^- = [\bar{\epsilon}_{ij}]_{2 \leq i \leq N, j \leq \Delta}$ with entries

$$\bar{\epsilon}_{ij} = \frac{1}{\tau}\sum_{t \in B_j} \epsilon_{it} \cdot D_{it}. \tag{69}$$

44

For the first row (treatment unit), since we know p by assumption, we define for all $j \in [\Delta]$

$$\bar{X}_{1j} = \frac{p}{\tau} \sum_{t \in B_j} X_{1t} \tag{70}$$

$$= \frac{p}{\tau} \sum_{t \in B_j} (M_{1t} + \epsilon_{1t})$$

$$= \frac{p}{\tau} \sum_{t \in B_j} M_{1t} + \frac{p}{\tau} \sum_{t \in B_j} \epsilon_{1t}$$

$$= \bar{M}_{1j} + \bar{\epsilon}_{1j}, \tag{71}$$

whereby $\bar{M}_{1j} = \frac{p}{\tau} \sum_{t \in B_j} M_{1t}$ and $\bar{\epsilon}_{1j} = \frac{p}{\tau} \sum_{t \in B_j} \epsilon_{1t}$. Under these constructions, the noise entries remain zero-mean random variables for all i, j, i.e. $\mathbb{E}[\bar{\epsilon}_{ij}] = 0$. However, the variance of each noise term is now rescaled, i.e. for $i = 1$

$$\text{Var}(\bar{\epsilon}_{1j}) = \frac{p^2}{\tau^2} \sum_{t \in B_j} \text{Var}(\epsilon_{1t})$$

$$\leq \frac{\sigma^2}{\tau},$$

and for $i > 1$,

$$\text{Var}(\bar{\epsilon}_{ij}) = \frac{1}{\tau^2} \sum_{t \in B_j} \text{Var}(\epsilon_{it} \cdot D_{it})$$

$$\overset{(a)}{\leq} \frac{1}{\tau^2} \sum_{t \in B_j} (\sigma^2 p(1-p) + \sigma^2 p^2)$$

$$\leq \frac{\sigma^2}{\tau}.$$

(a) used the fact that for any two independent random variables, X and Y, $\text{Var}(XY) = \text{Var}(X)\text{Var}(Y) + \text{Var}(X)(\mathbb{E}[Y])^2 + \text{Var}(Y)(\mathbb{E}[X])^2$. Thus, for all i, j, $\text{Var}(\bar{\epsilon}_{ij}) \leq \sigma^2/\tau := \bar{\sigma}^2$.

We now show that the key assumption of (6) still holds under this setting with respect to the newly defined variables. In particular, for every partition $j \in [\Delta]$ of row one,

$$\bar{M}_{1j} = \frac{p}{\tau} \sum_{t \in B_j} M_{1t}$$

$$= \frac{p}{\tau} \sum_{t \in B_j} \left(\sum_{k=2}^{N} \beta_k^* M_{kt} \right)$$

$$= \sum_{k=2}^{N} \beta_k^* \left(\frac{p}{\tau} \sum_{t \in B_j} M_{kt} \right)$$

$$= \sum_{k=2}^{N} \beta_k^* \bar{M}_{kj}.$$

As a result, we can express $\bar{M}_1^- = (\bar{M}^-)^T \beta^*$ for the same β^* as in (6).

Following a similar setup as before, we define the matrix $\bar{Y}^- = [\bar{Y}_{ij}]_{2 \leq i \leq N, j \leq \Delta}$. Since we have assumed that each block contains at least one observed entry, we subsequently have that $\bar{Y}_{ij} = \bar{X}_{ij}$ for all i and j. We now proceed with our analysis in the exact same manner with the only difference being

45

our newly defined set of variables and parameters. For completeness, we will highlight certain details below.

To begin, observe that $\mathbb{E}[\bar{Y}_{ij}] = \bar{M}_{ij}$ while

$$\mathrm{Var}(\bar{Y}_{ij}) \leq \frac{\sigma^2 p + p(1-p)}{\tau}.$$

Consequently, we redefine the event $E_3 := \{\|\bar{Y}^- - \bar{M}^-\| \leq (2+\omega)\sqrt{\Delta\bar{q}}\}$ for some choice $\omega \in (0.1, 1)$ and for $\bar{q} = \frac{\sigma^2 p + p(1-p)}{\tau}$. By Theorem A.5, it follows that $\mathbb{P}(E_3) \geq 1 - C'e^{-c\bar{q}\Delta}$.

Similar to before, let δ be defined by the relation

$$(1+\delta)\|\bar{Y}^- - \bar{M}^-\| = (2+\omega)\sqrt{\Delta\hat{\bar{q}}},$$

where $\hat{\bar{q}} = \frac{\hat{\sigma}^2 \hat{p} + \hat{p}(1-\hat{p})}{\tau}$. Letting $E = E_1 \cap E_2 \cap E_3$ and using arguments ((64), (65), (56)) that led us to (67), we obtain

$$\mathbb{E}\left\|(\hat{\bar{M}}^- - \bar{M}^-)^T \beta^*\right\|^2 \leq \mathbb{E}\left[\left\|(\hat{\bar{M}}^- - \bar{M}^-)^T \beta^*\right\|^2 \mid E\right] + \mathbb{E}\left[\left\|(\hat{\bar{M}}^- - \bar{M}^-)^T \beta^*\right\|^2 \mid E^c\right]\mathbb{P}(E^c)$$
$$\leq C_1 \Delta\bar{q} + C_2 e^{-c_3\phi\Delta},$$

where $\phi := p(N-1) + \sigma^2 + \bar{q}$. Utilizing Lemmas E.1 and E.5 gives us (for appropriately defined constants and defining $q = \sigma^2 p + p(1-p)$ as before such that $\bar{q} = q/\tau$)

$$\mathrm{MSE}(\hat{\bar{M}}_1^-) \leq C_1 \bar{q} + \frac{2\bar{\sigma}^2 k}{\Delta} + \frac{\eta\|\beta^*\|^2}{\Delta} + C_4 e^{-c_5\phi\Delta}.$$
$$= \frac{C_1 q}{\tau} + \frac{2\sigma^2 k}{\tau\Delta} + \frac{\eta\|\beta^*\|^2}{\Delta} + C_4 e^{-c_5 \frac{q}{\tau}\Delta}$$
$$= \frac{C_1 q}{T_0^{1/2-\gamma}} + \frac{2\sigma^2 k}{T_0} + \frac{\eta\|\beta^*\|^2}{T_0^{1/2+\gamma}} + C_4 e^{-c_5 q T_0^{2\gamma}}$$
$$= \mathcal{O}(T_0^{-1/2+\gamma}).$$

This concludes the proof. ∎

F   Forecasting Analysis: Post-Intervention Regime

We now bound the post-intervention ℓ_2 error of our estimator.

F.1   Proof of Theorem 4.5

Theorem (4.5). *Let* (6) *hold for some* $\beta^* \in \mathbb{R}^{N-1}$. *Let* $\mathrm{rank}(M^-) = \mathrm{rank}(M)$. *Then* $M_1^+ = (M^+)^T\beta^*$.

Proof. Suppose we begin with only the matrix M^-, i.e. $M = M^-$. From the assumption that $M_1^- = (M^-)^T\beta^*$, we have for $t \leq T_0$

$$M_{1t} = \sum_{j=2}^{N} \beta_j^* M_{jt}.$$

Suppose that we now add an extra column to $\boldsymbol{M}^-$ so that $\boldsymbol{M}$ is of dimension $N \times (T_0 + 1)$. Since $\text{rank}(\boldsymbol{M}^-) = \text{rank}(\boldsymbol{M})$, we have for $j \in [N]$

$$M_{j,T_0+1} = \sum_{t=1}^{T_0} \pi_t M_{jt},$$

for some set of weights $\pi \in \mathbb{R}^{T_0}$. In particular, for the first row we have

$$
\begin{aligned}
M_{1,T_0+1} &= \sum_{t=1}^{T_0} \pi_t M_{1t} \\
&= \sum_{t=1}^{T_0} \pi_t \Big(\sum_{j=2}^{N} \beta_j^* M_{jt} \Big) \\
&= \sum_{j=2}^{N} \beta_j^* \Big(\sum_{t=1}^{T_0} \pi_t M_{jt} \Big) \\
&= \sum_{j=2}^{N} \beta_j^* M_{j,T_0+1}.
\end{aligned}
$$

By induction, we observe that for any number of columns added to $\boldsymbol{M}^-$ such that $\text{rank}(\boldsymbol{M}^-) = \text{rank}(\boldsymbol{M})$, we must have $M_1^+ = (\boldsymbol{M}^+)^T \beta^*$ where $\boldsymbol{M}^+ = [M_{it}]_{2 \leq i \leq N, T_0 < t \leq T}$. ∎

F.2 Proof of Theorem 4.6

Theorem (4.6). *Suppose $p \geq \frac{T^{-1+\zeta}}{\sigma^2+1}$ for some $\zeta > 0$. Suppose $\left\| \hat{\beta}(\eta) \right\|_\infty \leq \psi$ for some $\psi > 0$. Let $\alpha' T_0 \leq T \leq \alpha T_0$ for some constants $\alpha', \alpha > 1$. Then for any $\eta \geq 0$ and using μ as defined in (20), the post-intervention error is bounded above by*

$$\text{RMSE}(\hat{M}_1^+) \leq \frac{C_1}{\sqrt{p}}(\sigma^2 + (1-p))^{1/2} + \frac{C_2 \|\boldsymbol{M}\|}{\sqrt{T_0}} \cdot \mathbb{E}\left\| \hat{\beta}(\eta) - \beta^* \right\| + \mathcal{O}(1/\sqrt{T_0}),$$

where C_1 and C_2 are universal positive constants.

Proof. We will prove Theorem 4.6 by drawing upon techniques and results from prior proofs. We begin by applying triangle inequality to obtain

$$
\begin{aligned}
\left\| \hat{M}_1^+ - M_1^+ \right\| &= \left\| (\hat{\boldsymbol{M}}^+)^T \hat{\beta}(\eta) - (\boldsymbol{M}^+)^T \beta^* \right\| \\
&= \left\| (\hat{\boldsymbol{M}}^+)^T \hat{\beta}(\eta) - (\boldsymbol{M}^+)^T \beta^* + (\boldsymbol{M}^+)^T \hat{\beta}(\eta) - (\boldsymbol{M}^+)^T \hat{\beta}(\eta) \right\| \\
&\leq \left\| (\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+)^T \hat{\beta}(\eta) \right\| + \left\| (\boldsymbol{M}^+)^T (\hat{\beta}(\eta) - \beta^*) \right\|.
\end{aligned}
$$

Taking expectations and using the property of induced norms gives

$$
\begin{aligned}
\mathbb{E}\left\| \hat{M}_1^+ - M_1^+ \right\| &\leq \mathbb{E}\left[\left\| \hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+ \right\| \cdot \left\| \hat{\beta}(\eta) \right\| \right] + \left\| \boldsymbol{M}^+ \right\| \cdot \mathbb{E}\left\| \hat{\beta}(\eta) - \beta^* \right\| \\
&\leq \sqrt{N}\psi \cdot \mathbb{E}\left\| \hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+ \right\| + \left\| \boldsymbol{M}^+ \right\| \cdot \mathbb{E}\left\| \hat{\beta}(\eta) - \beta^* \right\|, \quad (72)
\end{aligned}
$$

where the last inequality uses the boundedness assumption of $\hat{\beta}(\eta)$. Observe that the first term on the right-hand side of (72) is similar to that of (53) and (63) with the main difference being (72) uses the

47

post-intervention submatrices, $\hat{M}^+$ and M^+, as opposed to the pre-intervention submatrices, $\hat{M}^-$ and M^-, in (53) and (63). Therefore, using (57) and the arguments that led to (67), it follows that (with appropriate constants C_1, C_2, c_3)

$$\mathbb{E}\left\|\hat{M}^+ - M^+\right\| \leq \frac{C_1}{p}\mathbb{E}\left(\sqrt{Tq} + \left\|(\hat{p}-p)M^+\right\|\right) + C_2((N-1)T)^{3/2}e^{-c_3\phi T},$$

where the slight modification arises due to the fact that we are now operating in the post-intervention regime. In particular, $\|M^+\| \leq \sqrt{(N-1)(T-T_0)}$ and $\left\|\hat{M}^+\right\| \leq ((N-1)T)^{3/2}$. Further, note that q and ϕ are defined exactly as before, i.e. $q = \sigma^2 p + p(1-p)$ and $\phi = p(N-1) + \sigma^2 + q$. Following the proof of Corollary 4.1, we apply Jensen's Inequality to obtain

$$\mathbb{E}\left\|(\hat{p}-p)M^+\right\| = \mathbb{E}|\hat{p}-p| \cdot \left\|M^+\right\|$$
$$\leq \sqrt{\frac{p(1-p)}{(N-1)T}} \cdot \sqrt{(N-1)(T-T_0)}$$
$$\leq \sqrt{p(1-p)}.$$

Putting the above results together, we have (for appropriately defined constants)

$$\text{RMSE}(\hat{M}_1^+) \leq \frac{C_1\sqrt{N}\psi}{p\sqrt{T-T_0}}\left(\sqrt{Tq} + \sqrt{p(1-p)}\right) + \frac{\|M^+\|}{\sqrt{T-T_0}} \cdot \mathbb{E}\left\|\hat{\beta}(\eta) - \beta^*\right\| + C_4 e^{-c_5\phi T}$$
$$\overset{(a)}{\leq} \frac{C_6\sqrt{q}}{p} + \frac{C_7\sqrt{1-p}}{\sqrt{pT_0}} + \frac{C_8\|M\|}{\sqrt{T_0}} \cdot \mathbb{E}\left\|\hat{\beta}(\eta) - \beta^*\right\| + C_4 e^{-c_5\phi T}$$
$$= \frac{C_6}{\sqrt{p}}(\sigma^2 + (1-p))^{1/2} + \frac{C_8\|M\|}{\sqrt{T_0}} \cdot \mathbb{E}\left\|\hat{\beta}(\eta) - \beta^*\right\| + \mathcal{O}(1/\sqrt{T_0}),$$

where (a) follows from Lemma B.1. Renaming constants would provide the desired result. ∎

G   A Bayesian Perspective

Derivation of posterior parameters.

The following is based on the derivation presented in Section 2.2.3 of [12], and is presented here for completeness. Suppose we are given a multivariate Gaussian marginal distribution $p(x)$ paired with a multivariate Gaussian conditional distribution $p(y \mid x)$ – where x and y may have differing dimensions – and we are interested in computing the posterior distribution over x, i.e. $p(x \mid y)$. We will derive the posterior parameters of $p(x \mid y)$ here. Without loss of generality, suppose

$$p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$$
$$p(y \mid x) = \mathcal{N}(y \mid Ax + b, \Sigma^{-1}),$$

where μ, A, and b are parameters that govern the means, while Λ and Σ are precision (inverse covariance) matrices.

We begin by finding the joint distribution over x and y. Ignoring the terms that are independent of

x and y and encapsulating them into the "const." expression, we obtain

$$\begin{aligned}
\ln p(x, y) &= \ln p(x) + \ln p(y \mid x) \\
&= -\frac{1}{2}(x - \mu)^T \mathbf{\Lambda}(x - \mu) - \frac{1}{2}(y - \mathbf{A}x - b)^T \mathbf{\Sigma}(y - \mathbf{A}x - b) + \text{const.} \\
&= -\frac{1}{2}x^T(\mathbf{\Lambda} + \mathbf{A}^T\mathbf{\Sigma}\mathbf{A})x - \frac{1}{2}y^T\mathbf{\Sigma}y + \frac{1}{2}x^T\mathbf{A}^T\mathbf{\Sigma}y + \text{const.} \\
&= -\frac{1}{2}\begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} \mathbf{\Lambda} + \mathbf{A}^T\mathbf{\Sigma}\mathbf{A} & -\mathbf{A}^T\mathbf{\Sigma} \\ -\mathbf{\Sigma}\mathbf{A} & \mathbf{\Sigma} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \text{const.} \\
&= -\frac{1}{2}z^T\mathbf{Q}z + \text{const.},
\end{aligned}$$

where $z = [x, y]^T$, and

$$\mathbf{Q} = \begin{bmatrix} \mathbf{\Lambda} + \mathbf{A}^T\mathbf{\Sigma}\mathbf{A} & -\mathbf{A}^T\mathbf{\Sigma} \\ -\mathbf{\Sigma}\mathbf{A} & \mathbf{\Sigma} \end{bmatrix}$$

is the precision matrix. Applying the matrix inversion formula, we have that the covariance matrix of z is

$$\text{Var}(z) = \mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^T \\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{\Sigma}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T \end{bmatrix}.$$

After collecting the linear terms over z, we find that the mean of the Gaussian distribution over z is defined as

$$\mathbb{E}[z] = \mathbf{Q}^{-1} \begin{bmatrix} \mathbf{\Lambda}\mu - \mathbf{A}^T\mathbf{\Sigma}b \\ \mathbf{\Sigma}b \end{bmatrix}.$$

Now that we have the parameters over the joint distribution of x and y, we find that the posterior distribution parameters over x are

$$\begin{aligned}
\mathbb{E}[x \mid y] &= (\mathbf{\Lambda} + \mathbf{A}^T\mathbf{\Sigma}\mathbf{A})^{-1}\{\mathbf{A}^T\mathbf{\Sigma}(y - b) + \mathbf{\Lambda}\mu\} \\
\text{Var}(x \mid y) &= (\mathbf{\Lambda} + \mathbf{A}^T\mathbf{\Sigma}\mathbf{A})^{-1}.
\end{aligned}$$