Robust Synthetic Control

by

Dennis Shen

B.S., University of California San Diego (2015)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2018

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
September 28, 2017

Signature redacted

Certified by .....
Devavrat Shah
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Robust Synthetic Control

by

Dennis Shen

Submitted to the Department of Electrical Engineering and Computer Science
on September 28, 2017, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

In this thesis, we present a robust generalization of the synthetic control method. 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, functions without extraneous covariates (vital to existing methods), and overcomes missing data (never been addressed in prior works). To our knowledge, we provide the first theoretical finite sample analysis for a broader class of models than previously considered in literature. Additionally, we relate the inference quality of our estimator to the amount of training data available and show our estimator to be asymptotically consistent. In order to move beyond point estimates, we introduce a Bayesian framework that not only provides practitioners the ability to readily develop different estimators under various loss functions, but also equips them with the tools to quantitatively measure the uncertainty of their model/estimates through posterior probabilities. Our empirical results demonstrate that our robust generalization yields a positive impact over the classical synthetic control method, underscoring the value of our key de-noising procedure.

Thesis Supervisor: Devavrat Shah
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

The past two years have been an amazing journey, thanks to all the people I have met along the way. I would like to begin by thanking my advisor, Devavrat Shah, for taking a gamble on me and giving me the opportunity to join his research group. Devavrat has been incredibly patient with me, giving me both time and encouragement to overcome my numerous shortcomings. In fact, rather than inundating me with research my first semester, Devavrat encouraged me to focus on my courses and attend talks to build a solid foundation, and he did so in his own unique way, asserting that "there is no use going into war with forks and knives" – anyone who knows Devavrat knows he has quite the way with words. Beyond his patience, Devavrat has morphed the way I think, teaching me how to approach and break down complex problems, and helping me realize the elegance of simplicity.

Observing that my research interests were taking a random walk, Devavrat wisely put me under the mentorship of his more senior students. Quite frankly, this thesis would definitely have not been possible without my collaborator, Jehangir Amjad. Throughout our time working together, Jehangir proved to be a tremendous mentor: helping to fix my proofs and bouncing ideas with me. I am also thankful to be living with Dogyoon Song, a walking encyclopedia. I thank Dogyoon for answering all of my math questions and for motivating me to be healthy...most of the time. Overall, I am grateful to everyone in Devavrat's SSPIN research group for both their thought-provoking and fun discussions.

Although he probably doesn't remember me (and understandably so), I am indebted to Professor Alan Oppenheim for being so kind to me before, during, and after the EECS visit days. Professor Oppenheim's genuineness is a large reason as to why I traveled across the country to pursue my graduate studies in Boston, the city that unfortunately hosts all the sports teams I loathe the most.

I am thankful to several funding agencies that supported my research, including the National Security Agency and Draper Laboratory.

I am also grateful for Boston Cares, which has provided me the opportunity to not only give back to my community, but also gain perspective of how blessed my life has been. Thank you for giving me a higher purpose and for helping me meet such wonderful and caring role models.

As with many great adventures, mine began because of a girl – my high school sweetheart, girlfriend, and best friend of 7 years, Jana. Throughout our entire time together, Jana has kept me rooted, ensuring that I maintain perspective on the most

important things in my life – besides herself. She's available when I need her most, cheers me on (even when there's not much to be proud of), and encourages me to venture beyond my comfort zones. Despite not sharing the same affinity for my field of study, she also indulges me by listening to me geek about my work. Jana is my greatest source of happiness, always.

Literally and figuratively, I would not be here today without the undying love and support of my other two best friends, my parents. At every stage of my life, my parents have undoubtedly been my most loyal and passionate fans. I can never thank them both enough for allowing me to pursue and find my own interests. Although I don't often say or show it, I deeply appreciate all of my dad's stories and advice, and for making me laugh, particularly when he knows more about what is happening around campus than I do. I am beyond thankful that I have a mom who listens to all of my pointless stories and rants, helps me rediscover my roots in art and music, and, more importantly, cooks and sends the most delicious food/care packages. My parents anchor my life and everything I accomplish is because of their love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Consider a typical comparative case-study where a legislative body is interested in measuring the impact of a policy (e.g. gun control through crime-rate) on a "treated" unit (e.g. California). Unlike the setting of "randomized control" a la A/B testing, the population of such a comparative case-study is limited to a single unit, forcing one to choose an unaffected unit as a "control" (e.g. New York). Historically, such selection was left to the discretion of domain experts. In their seminal work, Abadie and Gardeazabal [4] introduced the concept of "synthetic control", where the control unit is a convex combination of unaffected units (e.g. 80% New York, 20% Massachusetts). Theirs and various subsequent works proposed to learn the synthetic control by applying domain expertise to carefully select the candidate "donor pool" of control units, and utilizing supplementary covariates (e.g. employment rates) to learn the convex relationship.

As the main result of this work, we propose a "robust" approach to finding the synthetic control, wherein we first "de-noise" the observation data and then use the de-noised data to learn a linear relationship. The de-noising step is a distinguishing feature from prior approaches as it renders the selection of the synthetic control robust in two senses: one, it *does not* require the assistance of covariates or domain "experts"; and two, it can handle *missing* and/or *noisy* observations, an aspect that has not been previously addressed. Under a more general framework that encompasses existing models, we provide finite sample analysis and, subsequently, establish asymptotic consistency, which has been absent from literature. We also analyze the synthetic control method from a Bayesian perspective, which allows our algorithm to go beyond point estimates in expressing our uncertainties through posterior probability distributions. Using real-world datasets, we showcase the robustness of our algorithm by reproducing existing case studies without the benefits of additional covariates or domain knowledge, and in

the presence of missing information. Finally, we generate model-driven synthetic data to validate the efficacy of our algorithm.

1.1 Motivation

On November 8, 2016 in the aftermath of several high profile mass-shootings, voters in California passed Proposition 63 in to law [8]. 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" [25]. 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 (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.

1.2 Overview of Main Contributions

In this work, we revisit the study of synthetic control from a robust perspective in order to address the limitations described above. 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 does not require domain knowledge or the use of 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.

1.2.1 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 existing methods) as a luxury as opposed to a necessity.

In the spirit of combatting overfitting, we further 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 their results through posterior probabilities.

1.2.2 Theoretical performance

To the best of our knowledge, our exposition is the first to analyze both the efficacy of the synthetic control estimator with respect to the MSE and the effect of missing data on the algorithm's performance. Previously, the main theoretical result from the

synthetic control literature pertained to asymptotic unbiasedness for a linear factor model; however, the proof of the result assumed that the latent parameters, which live in the simplex, have been perfectly discovered. We provide finite sample analysis that not only highlights the value of thresholding in balancing the "bias-variance" trade-off, but also proves that the efficacy of our algorithm degrades gracefully with an increasing number of randomly missing data. Further, we show that a computationally beneficial pre-processing step allows us to establish the asymptotic consistency of our least-squares estimator in generality. Using results from the statistical learning theory literature, we provide post-intervention/generalization error bounds under the regularized (ridge regression) setting.

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 exists in the pre- and post-intervention periods. Finally, 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.

1.2.3 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.

1.3 Related Literature

The study of synthetic control (synthetic control) has received widespread attention ever since its conception by Abadie and Gardeazabal in their pioneering work [4, 1]. It has been employed in numerous case studies, ranging from criminology [26] to

health policy [23] to online advertisement to retail; other notable studies include [3, 9, 5, 7]. 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" [6]. In a somewhat different direction, Hsiao et al. introduce the panel data method [20, 21], which seems to have a close bearing with some of the approaches of this work. In particular, [20, 21] only uses data for the outcome variable and solves an ordinary least squares problem in learning synthetic control. However, [20, 21] 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, [20, 21] 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, see [19]. It should be noted that [19] also adapts the panel data method to automate the donor selection process. [15] relaxes the convexity aspect of synthetic control, and allows for an additive difference between the treated unit and donor pool, similar to the difference-in-differences (DID) method. In an effort to infer the causal impact of market interventions, [12] introduce yet another evaluation methodology based on a diffusion-regression state-space model that is fully Bayesian; similar to [1, 4, 20, 21], their model also generalizes the DID procedure. Due to the subjectivity in the choice of covariates and predictor variables, [18] provide recommendations for specification-searching opportunities in synthetic control applications.

Matrix completion and factorization approaches are well-studied problems with broad applications (e.g. compressed sensing, recommendation systems, 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 [13]. 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 [22]. For a detailed discussion on the topic, see [14]; for algorithmic implementations, see [24] 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.

Despite its popularity, there has been less theoretical work in establishing the consistency of the synthetic control method or its variants. [1] shows that the synthetic control method can produce an asymptotically unbiased estimator, but under restrictive settings; their analysis relies on the assumption that there not only exists a perfect "convex" match between the pre-treatment *noisy* outcome and covariate variables for the treated unit and donor pool, but that the algorithm has also discovered the true "convex" weights. In contrast, our analysis *does not* assume that the estimator has discovered the true set of linear weights and is truly assumption free. [17] also relaxes the strong assumption in [1], and derives conditions under which the synthetic control estimator is asymptotically unbiased. 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.

1.4 Organization of the Thesis

The rest of this work is outlined as follows: Section 2 describes our notation, setting, and proposed data model. We present the two-step algorithm in Section 3 with the corresponding theoretical and experimental results in Section 4 and Section 5, respectively. We then extend our framework to incorporate regularization methods in Section 6, and finish with a Bayesian treatment of synthetic control in Section 7. All proofs and derivations are unveiled in the appendices.

Chapter 2

Preliminaries

2.1 Setup

In this section, we define the necessary notation and describe the setting.

2.1.1 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

$$\boldsymbol{A}^\dagger = \sum_{i=1}^{k} (1/\sigma_i) y_i x_i^T, \tag{2.1}$$

where

$$\boldsymbol{A} = \sum_{i=1}^{k} \sigma_i x_i y_i^T, \tag{2.2}$$

with x_i and y_i being the left and right singular vectors of $\boldsymbol{A}$, respectively.

Let $\hat{v}$ be a random vector that is an estimate of v. Then one choice for the measure of error in estimation is the average mean-squared error, denoted as $\mathrm{MSE}(\hat{v})$, and

defined as

$$\text{MSE}(\hat{v}) = \frac{1}{n}\|v - \hat{v}\|_2^2. \tag{2.3}$$

We will denote the root mean-squared error, $\text{RMSE}(\hat{v})$, as the square root of the MSE. Since we will frequently use the ℓ_2 and spectral norms, we will adopt the shorthand notation of $\|\cdot\| \equiv \|\cdot\|_2$ for both cases by often dropping the subscript. Finally, to avoid any confusions between scalars/vectors and matrices, we will represent all matrices in bold, e.g. $\boldsymbol{A}$.

2.1.2 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 (X_1) and the donor pool ($\boldsymbol{X}$) outcomes. Suppose we observe $N \geq 2$ units across $T \geq 2$ time periods. We denote T_0 as the number of pre-intervention periods with $1 \leq T_0 < T$, 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, $2 \leq i \leq N$, are unaffected by the intervention for the entire time period $[T] = \{1, \ldots, T\}$.

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. $A_1 = [A_1^-, A_1^+]$.

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

$$X_{it} = M_{it} + \epsilon_{it}, \tag{2.4}$$

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, we further posit that for all $2 \leq i \leq N$, $t \in [T]$

$$M_{it} = f(\theta_i, \rho_t), \tag{2.5}$$

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 $d_1, d_2 \geq 1$; 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 linear function.

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

$$X_{1t} = M_{1t} + \epsilon_{1t}, \tag{2.6}$$

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 (2.6) 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 (2.6) and (2.4). 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) 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 $t \leq T_0$,

$$M_{1t} = \sum_{i=2}^{N} \beta_i M_{it}. \tag{2.7}$$

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 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

[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.

of all other entries.

Chapter 3

Algorithm

We will begin by providing intuition behind our proposed algorithm: (1) we begin by de-noising our data via singular value thresholding, a distinguishing feature from prior approaches. Since the singular values of our observation matrix, $\boldsymbol{X}$, encode both signal and noise, we attempt to find a proper 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. (2) 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 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.

Note: To simplify the exposition, we assume the entries of $\boldsymbol{X}$ are bounded by one in absolute value, i.e. $|X_{it}| \leq 1$.

3.1 Parametrized Algorithm

The algorithm utilizes the thresholding hyperparameter $\mu \geq 0$, which serves as a knob to effectively trade-off between the bias and variance of the estimator. We discuss the procedure for determining the parameter μ soon after the description of the parametrized algorithm.

Step 1. De-noising the data

1. Define $\boldsymbol{Y} = [Y_{ij}]$ with

$$Y_{ij} = \begin{cases} X_{ij}, & \text{if } X_{ij} \text{ is observed} \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

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

$$\boldsymbol{Y} = \sum_{i=1}^{N-1} s_i u_i v_i^T. \tag{3.2}$$

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

$$\hat{\boldsymbol{M}} = \frac{1}{\hat{p}} \sum_{i \in S} s_i u_i v_i^T, \tag{3.3}$$

where $\hat{p}$ denotes the fraction of observed entries in $\boldsymbol{X}$.

Step 2. Learning and projecting

1. Let $\hat{\beta}$ be the estimate of β obtained by solving the least-squares problem

$$\hat{\beta} = \underset{v \in \mathbb{R}^{N-1}}{\arg\min} \left\| Y_1^- - (\hat{\boldsymbol{M}}^-)^T v \right\|^2. \tag{3.4}$$

2. Define the counterfactual means for the treatment unit as

$$\hat{M}_1 = \hat{\boldsymbol{M}}^T \hat{\beta}. \tag{3.5}$$

3.1.1 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 $(a + b)/2$.

3.1.2 Choosing the hyperparameter, μ

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 $\hat{\beta}$ for each of the finitely many choices of μ $(s_1, \ldots, s_{N-1})$. Using each estimate $\hat{\beta}$, 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, [14] provides a universal approach to picking a threshold. As discussed in Section 5, we utilize the data driven approach for producing our results.

3.1.3 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.

3.1.4 Remarks on 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 [14] 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 [27], many latent variable models can be well approximated (up to arbitrary accuracy ϵ) by low-rank matrices. Specifically, [27] shows that the corresponding low-rank approximation matrices associated with "nice" functions (e.g. linear functions, polynomials, kernels, etc.) are of log-rank.

Chapter 4

Summary of Main Results

In this section, we derive the finite sample and asymptotic properties of the estimator, $\hat{M}_1$. 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 (2.4) 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 $2 \leq i \leq N, t \in [T]$,

$$\mathbb{E}[\epsilon_{it}] = 0, \quad \text{and} \quad \text{Var}(\epsilon_{it}) \leq \sigma^2. \tag{4.1}$$

We shall also assume that the treatment unit noise in (2.6) obeys (4.1). Further, we assume the relationship in (2.7) holds.

As previously discussed, we wish to evaluate the accuracy of our estimated means for the treatment unit with respect to the MSE, i.e. the deviation between $\hat{M}_1^-$ and M_1^- measured in ℓ_2-norm. Additionally, we aim to establish the validity of our pre-intervention linear model assumption (cf. (2.7)) 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 $(\boldsymbol{M}^+)^T \beta$ and if so, under what conditions? We now present our results for the above two aspects.

4.1 Pre-intervention analysis

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

$$\text{MSE}(\hat{M}_1^-) = \frac{1}{T_0}\mathbb{E}\Big[\sum_{t=1}^{T_0}(M_{1t} - \hat{M}_{1t})^2\Big]. \tag{4.2}$$

We say that $\hat{M}_1^-$ is a consistent estimator if (4.2) approaches 0 as $T_0 \to \infty$. In what follows, we first state the finite sample bound on (4.2) for the most generic setup (Theorem 4.1.1). As a main Corollary of the result, we specialize the bound in the case where $\boldsymbol{M}$ is low-rank. (Corollary 4.1.1). 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 4.1.2).

4.1.1 General result

We provide a finite sample error bound for the most generic setting.

Theorem 4.1.1. *The pre-intervention error of the algorithm can be bounded as*

$$\text{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p^2 T_0}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \big\|(\hat{p} - p)\boldsymbol{M}^-\big\|\Big)^2 \|\beta\|^2 + \frac{2\sigma^2|S|}{T_0} \tag{4.3}$$

$$+ C_2(N-1)\|\beta\|^2 e^{-cp(N-1)T}. \tag{4.4}$$

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.

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).

4.1.2 Goldilocks Principle

With an appropriate choice for the hyperparameter μ (and hence S), we state the following result for the specialized setting whereby the signal matrix $\boldsymbol{M}$ is low rank.

Corollary 4.1.1. *Let* $\text{rank}(\boldsymbol{M}) = k$ *for some* $1 \leq k < N - 1$. *Let the choice of* μ *be such that* $|S| = k$. *Suppose* $\sigma^2 p + p(1-p) \geq T^{-1+\zeta}$ *for some* $\zeta > 0$. *Let* $T \leq \alpha T_0$ *for some constant* $\alpha > 1$. *Then*

$$\lim_{T_0 \to \infty} \text{MSE}(\hat{M}_1^-) \leq \frac{C_1 \|\beta\|^2}{p}(\sigma^2 + (1-p)). \tag{4.5}$$

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.

4.1.3 Asymptotic Consistency

We present a straightforward pre-processing step that leads to the asymptotic consistency of our algorithm. The pre-processing step simply involves replacing the columns of $\boldsymbol{X}$ by the averages 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 smaller size matrix.

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

$$\bar{X}_{ij} = \frac{1}{\tau} \sum_{t \in B_j} X_{it}. \tag{4.6}$$

Let $\bar{\boldsymbol{M}}^- = [\bar{M}_{ij}]_{2\le i\le N, 1\le j\le \tau}$ with

$$\bar{M}_{ij} = \frac{1}{\tau}\sum_{t\in B_j} M_{it}. \tag{4.7}$$

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

$$\hat{M}_{1t} = \sum_{i=2}^{N} \hat{\beta}_i X_{it}. \tag{4.8}$$

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

$$\hat{\bar{M}}_{1j} = \sum_{i=2}^{N} \hat{\beta}_i \hat{\bar{M}}_{ij}. \tag{4.9}$$

Our measure of estimation error is defined as

$$\mathrm{MSE}(\hat{\bar{M}}_1^-) = \frac{1}{\tau}\mathbb{E}\Big[\sum_{1\le j\le \tau} (\bar{M}_{1j} - \hat{\bar{M}}_{1j})^2\Big]. \tag{4.10}$$

We state the following result.

Theorem 4.1.2. *Let* $\mathrm{rank}(\bar{\boldsymbol{M}}^-) = k$ *for some* $1 \le k < N-1$. *Let the choice of* μ *be such that* $|S| = k$. *Then*

$$\lim_{T_0\to\infty} \mathrm{MSE}(\hat{\bar{M}}_1^-) = 0.$$

We note that the method of [4, Sec 2.3] learns the weights (here $\hat{\beta}$) 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.

[1] In practice, one can first de-noise $\boldsymbol{X}^+$ via step one of Section 3, and use the entries of $\hat{\boldsymbol{M}}^+$ in (4.8).

4.2 Post-intervention analysis (static rank)

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 (2.7). The question still remains, however, 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 entire 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 4.2.1. *Let* (2.7) *hold for some* β. *Let* $\text{rank}(\boldsymbol{M}^-) = \text{rank}(\boldsymbol{M})$. *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 4.2.2. *Assuming* $\text{rank}(\boldsymbol{M}^-) = \text{rank}(\boldsymbol{M})$, *the post-intervention error is bounded above by*

$$\text{RMSE}(\hat{M}_1^+) \leq \frac{C_1\sqrt{T_0}}{p\mu\sqrt{T-T_0}}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \big\|(\hat{p}-p)\boldsymbol{M}^+\big\|\Big) + \frac{\|\boldsymbol{M}^+\|}{\sqrt{T-T_0}}\mathbb{E}\Big\|\hat{\beta}-\beta\Big\| + \frac{C_2\sqrt{T_0(N-1)}}{\mu}e^{-cp(N-1)T}.$$

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.*

Let us interpret the post-intervention error bound by decomposing the RMSE into two error terms (we will ignore the third expression since it decreases exponentially fast with the size of the training set): the first error term derives from the de-noising/estimation error from Step one of our robust algorithm, and the second term captures the learning algorithm's error (in this case, linear regression) from Step two. Similar to the pre-intervention error, there is a trade-off between "bias" and "variance", which is dictated by the choice of the threshold value μ. To see this, we analyze the key information ratio λ^*/μ within the first term. As μ increases, our de-noising process uses less singular values (smaller set S), rendering λ^* – the signal not captured in the thresholding process – to also increase. On the flip side, if we use a small threshold, then we are utilizing most of the data matrix's singular values, yielding λ^*

to also decrease. In either case, there is a tension that exists due to the thresholding procedure since λ^* and μ are positively correlated.

The second error term, which is controlled by the expression $\left\|\hat{\beta} - \beta\right\|$, is a function of the learning algorithm used to estimate $\hat{\beta}$. As we will shortly see, using regularization can decrease the MSE between $\hat{\beta}$ and the true, underlying β, thus reducing the overall post-intervention error.

Chapter 5

Experimental Results

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. For the purposes of this section, we refer to the algorithm presented in Section 3 as *robust synthetic control (linear)*. Additionally, we introduce a variation to our proposed algorithm by restricting $\hat{\beta}$ to have non-negative components that sum to one; we refer to this variation as *robust synthetic control (convex)* [1].

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 the Basque

[1] In the Econometrics literature, an emphasis has been placed on having $\hat{\beta}$ being "convex" as it provides an intuitive interpretation: the treatment unit is *proportionately like* the donor units.

Country and California Prop. 99 case studies below to assess the significance of our estimates.

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.

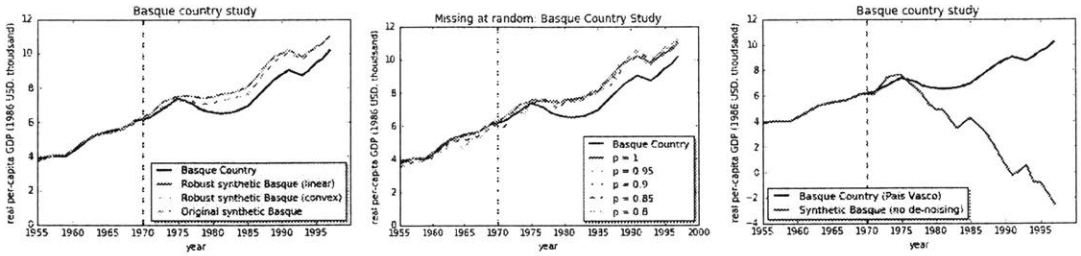
5.1.1 Results

Figure 5-1a shows that our method (both linear and convex) 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 (linear and convex) synthetic control 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.

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 5-1b suggesting the robustness of our (linear) algorithm. Finally, we produce Figure 5-1c 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.

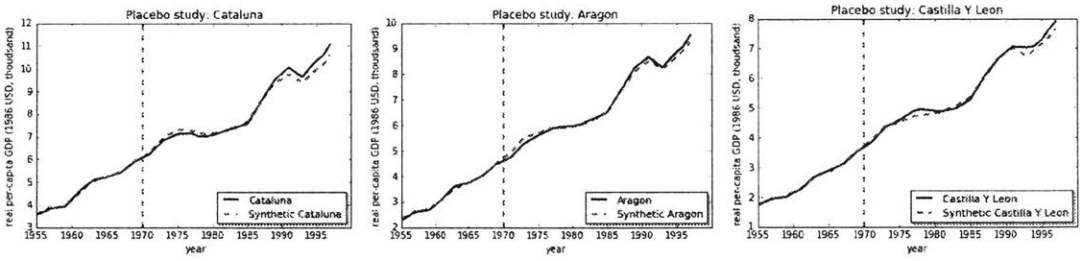


(a) Comparison of methods. **(b)** Missing data. **(c)** Impact of de-noising.

Figure 5-1: Trends in per-capita GDP between Basque Country vs. synthetic Basque Country.

5.1.2 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 5-2a, 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 5-2b and 5-2c.

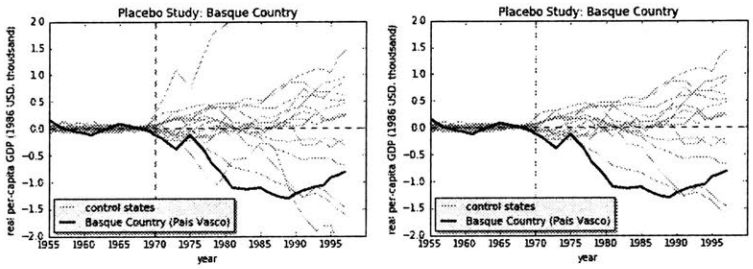


(a) Cataluna. **(b)** Aragon. **(c)** Castilla Y Leon.

Figure 5-2: Trends in per-capita GDP for placebo regions.

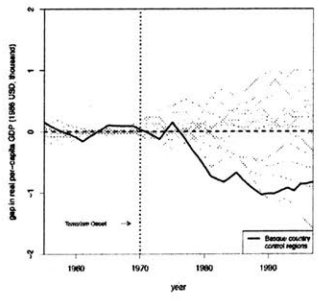
Additionally, we performed the exact inferential test on all control regions and plotted the resulting per-capita GDP gaps in Figures 5-3a and 5-3b, whereby Figure 5-3b excluded two control regions; the purpose behind this action will be made clear in the following paragraph. The resulting figures suggest that there is a low probability of obtaining a large economic divergence similar to that of Basque Country, when we reassign the intervention to the donor regions.

Since there is no ground truth, we continue to use the seminal results of [2] as a baseline. We begin by noting that [2] removed the plots of all five regions that had a poor pre-treatment period fit (regions with a mean-squared error, with respect to some pre-intervention validation period, that is five times greater than that for Basque Country); we display their resulting figure for the 12 remaining regions in Figure 5-4a as a visual reference. As a result, we removed the two regions – Balearic Islands and Madrid – that were mentioned in [2]. Thus, Figure 5-3a represents the result of our inferential test on all control regions while 5-3b excludes the Balearic Islands and Madrid. Even though [2] used 13 additional covariates and excluded more "bad" regions from their permutation placebo test, we observe that our results are nearly identical. This reinforces the robustness of our algorithm, highlighting the profound impact of de-noising.



(a) Includes all control regions. **(b)** Excludes 2 regions.

Figure 5-3: Per-capita GDP gaps for Basque Country and control regions.



(a) Excludes 5 regions.

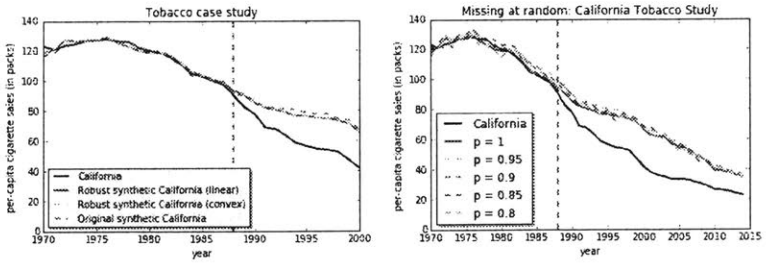
Figure 5-4: Per-capita GDP gaps for Basque Country and control regions: results by [2].

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.

5.2.1 Results

As shown in Figure 5-5a, in the pre-intervention period of 1970-88, our control (linear and convex) 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 qualitatively similar to that of [4]. As seen in Figure 5-5b, our (linear) algorithm is again robust to randomly missing data.



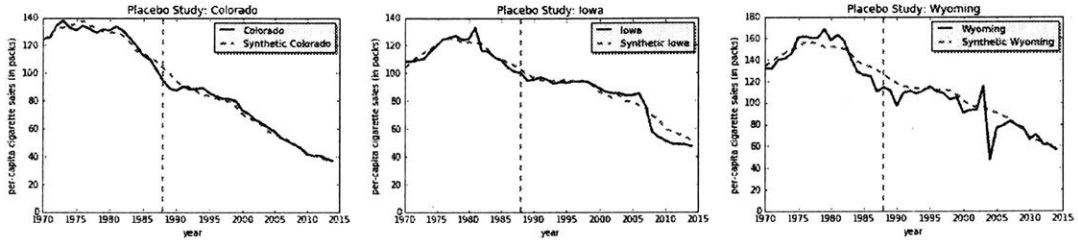
(a) Comparison of methods. **(b)** Missing data.

Figure 5-5: Trends in per-capita cigarette sales between California vs. synthetic California.

5.2.2 Placebo tests

We now proceed to apply the same placebo tests to the California Prop 99 dataset. Figures 5-6a, 5-6b, and 5-6c are three examples of the applied placebo tests on the

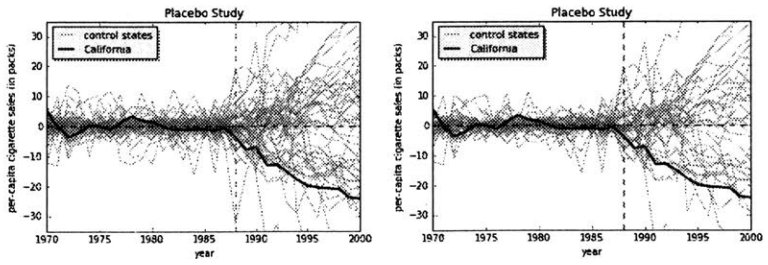
remaining states (including District of Columbia) within the United States.



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

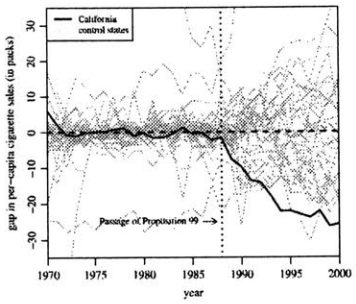
Figure 5-6: Placebo Study: trends in per-capita cigarette sales for Colorado, Iowa, and Wyoming.

We again apply the iterative inferential technique to all 50 states and the District of Columbia. Unlike the case of [1], our estimated effects shown in Figures 5-7a and 5-7b are produced without the benefits of any covariates and without the elimination of "bad" states or years post-2000. Note that we plot the predicted effects for all donor units in Figure 5-7a, but we exclude the twelve states that were discarded during the learning process of [1] in Figure 5-7b. For comparison, we display the resulting estimated effects for the 38 states used in the estimation process of [1] in Figure 5-8. We find that our inferential placebo test results are again similar to that of [1]. However, even though our algorithm was "handicapped" by using only the time series of data for the outcome variable (per-capita cigarette sales), we observe a noticeable difference in estimated effects during the pre-intervention period of 1970-1988. In particular, in both 5-7a and 5-7b, our estimated gaps are bounded roughly between $[-15, 10]$, while some of the estimated gaps of [1] diverge greatly outside of that interval, indicating a poor pre-treatment period fit.



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

Figure 5-7: Per-capita cigarette sales gaps in California and control regions.



(a) Excludes 12 states.

Figure 5-8: Per-capita cigarette sales gaps in California and control regions: results by [1].

5.3 Discussion

Although the experimental results suggest that our robust algorithm 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.

5.4 Synthetic simulations

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

5.4.1 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: $f_1 = \rho_t \bmod (360), f_2 = \rho_t \bmod (180), f_3 = 2 \cdot \rho_t \bmod (360), f_4 = 2.0 \cdot \rho_t \bmod (180)$. 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 $t = 1600$.

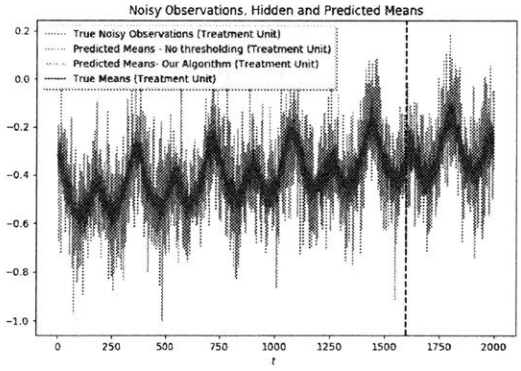


Figure 5-9: Treatment unit: noisy observations (gray) and true means (blue) and the estimates from our algorithm (red) and one where no singular value thresholding is performed (green). The plots show all entries normalized to lie in range $[-1, 1]$. Notice that the estimates in red generated by our model are much better at estimating the true underlying mean (blue) when compared to an algorithm which performs no singular value thresholding.

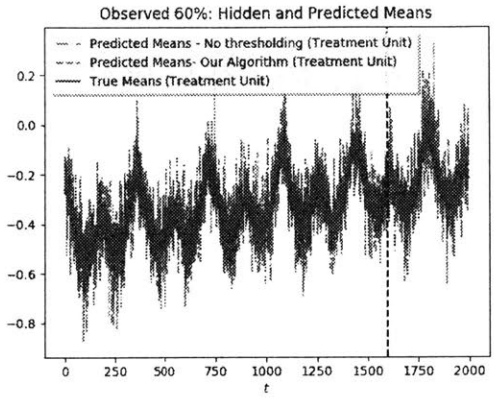


Figure 5-10: Same dataset as shown in Figure 5-9 but with 40% data missing at random. Treatment unit: not showing the noisy observations for clarity; plotting true means (blue) and the estimates from our algorithm (red) and one where no singular value thresholding is performed (green). The plots show all entries normalized to lie in range $[-1, 1]$.

5.4.2 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 5.1, 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 5-9 and 5-10 plot the estimates of algorithm with no missing data (Figure 5-9) and with 40% randomly missing data (Figure 5-10) 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 5.1: 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

5.4.3 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 5.2 show that for range of parameters the generalization error with de-noising is consistency better than that without de-noising.

Table 5.2: 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

Chapter 6

Regularization

"Suppose there exist two explanations for an occurrence. In this case the simpler one is usually better."

Occam's Razor.

6.1 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 (3.4), redefining the learning procedure in Step two of our algorithm. For a general regularizer, the objective function now takes the form

$$\hat{\beta}_\eta = \underset{v \in \mathbb{R}^{N-1}}{\arg\min} \left\| Y_1^- - (\hat{\boldsymbol{M}}^-)^T v \right\|^2 + \eta \sum_{j=1}^{N-1} |v_j|^q, \tag{6.1}$$

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. Note that if $\eta = 0$, then the complexity penalty is nullified and our objective returns to its original form. In general, the impact/trade-off of regularization can be controlled by the value of the regularization parameter η – the choice of this parameter will be

discussed in a later subsection. Via the use of Lagrange multipliers, we note that minimizing (6.1) is equivalent to minimizing (3.4) subject to the constraint that

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

for some appropriate value of c. When $q = 2$, (6.1) corresponds to the classical setup known as *ridge regression*[1]. 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.

6.2 Ridge Rigression

We will now focus our attention on the quadratic regularizer, $q = 2$, also known as ridge regression. This particular form of regularization encourages the learning algorithm to reduce the size of the coefficients to decay towards zero, unless supported by the data. Although the quadratic ℓ_2 penalty adds some bias, the penalty also reduces the variance of the produced estimator. Additionally, ridge regression possesses the advantage of maintaining the objective function to be (convex) quadratic in the parameter, v, so that its exact minimizer can be found in closed form:

$$\hat{\beta}_\eta = \left(\hat{\boldsymbol{M}}^-(\hat{\boldsymbol{M}}^-)^T + \eta \boldsymbol{I}\right)^{-1} \hat{\boldsymbol{M}}^- Y_1^-, \tag{6.2}$$

where the subscript denotes the dependency on the choice of the regularization parameter η. We note that $\hat{\boldsymbol{M}}^-(\hat{\boldsymbol{M}}^-)^T + \eta \boldsymbol{I}$ is a positive definite matrix for any $\eta > 0$, thus its inverse always exists. Consequently, the quadratic regularizer requires no rank (or dimension) assumptions on the matrix $\hat{\boldsymbol{M}}^-$ [11]. This highlights another reason why regularization is a popular heuristic as adding regularization often makes the problem easier to solve numerically.

[1]Due to its popularity, the regularization setting of $q = 2$ has many other names in literature, including Tikhonov regression and weight decay.

6.2.1 Pre-intervention analysis

Let us study the finite sample pre-intervention error bound when we substitute ridge regression in place of ordinary linear regression in learning the regression coefficients, β.

Theorem 6.2.1. *For any $\eta > 0$, the pre-intervention error of the algorithm can be bounded as*

$$\text{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p^2 T_0}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \big\|(\hat{p}-p)\boldsymbol{M}^-\big\|\Big)^2 \|\beta\|^2 + \frac{2\sigma^2|S|}{T_0} \tag{6.3}$$

$$+ \frac{\eta\|\beta\|^2}{T_0} + C_2(N-1)\|\beta\|^2 e^{-c(N-1)Tp}. \tag{6.4}$$

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.

As evident by (C.7), the upper bound on the pre-intervention (training) error includes an additional error term, $\eta\|\beta\|^2/T_0$, derived from regularization. Therefore, as η increases, the impact of the regularization also magnifies, driving the learning algorithm to reduce the model complexity at the expense of a increased bias and potentially larger training error. However, as frequently demonstrated by empirical studies, the incorporation of regularization also reduces the the generalization error, which is the key quantity of interest we aim to reduce.

6.2.2 Post-intervention analysis (static rank)

Before we analyze the general case for the post-intervention (generalization) error, let us first study the static rank scenario.

Theorem 6.2.2. *Assuming* $\text{rank}(\boldsymbol{M}^-) = \text{rank}(\boldsymbol{M})$, *the post-intervention root mean-square error (RMSE) is bounded above by*

$$\text{RMSE}(\hat{M}_1^+) \leq \frac{C_1\sqrt{T_0}}{p\mu\sqrt{T-T_0}}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \big\|(\hat{p}-p)\boldsymbol{M}^+\big\|\Big) + \frac{\|\boldsymbol{M}^+\|}{\sqrt{T-T_0}}\mathbb{E}\Big\|\hat{\beta}_\eta - \beta\Big\|$$
$$+ \frac{C_2\sqrt{T_0(N-1)}}{\mu}e^{-cp(N-1)T}.$$

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.

The post-intervention RMSE bound under the ridge regression setting is identical to that of linear regression (see Lemma 4.2.2), with the exception of the second error term, i.e. $\left\|\hat{\beta}_\eta - \beta\right\|$. Recall that this error discrepancy arises from the fact that we are implementing two different algorithms, i.e. linear regression versus (quadratic) regularized linear regression, to learn the synthetic control. Interestingly, [16] demonstrates that there exists a regularization hyperparameter $\eta > 0$ such that

$$\left\|\hat{\beta}_\eta - \beta\right\| \leq \left\|\hat{\beta} - \beta\right\|,$$

without any assumptions on the rank of $\hat{M}^-$.

Ultimately, employing ridge regression introduces extraneous bias into our model, yielding a higher pre-intervention error. However, the sacrifice in the pre-intervention error returns to us the benefit of a smaller post-intervention error bound (due to smaller variance), the quantity of interest that we truly care about.

6.3 Ridge Regression Generalization Error

We will now develop generalization (post-intervention) error results under a generic setting without any assumptions on the relationship between the rank of $\boldsymbol{M}^-$ and $\boldsymbol{M}$. Throughout this section in introducing definitions and theorems, we will temporarily adopt the notation established in the statistical learning theory literature before connecting the borrowed notation to our own framework at the very end. In particular, we make use of the notations, definitions, and results from [10].

6.3.1 Notations

Let $\mathcal{X}$ and $\mathcal{Y} \subset \mathbb{R}$ denote the input and output spaces, respectively. We denote our training dataset of size m as

$$D = \{z_1 = (x_1, y_1), \ldots, z_m = (x_m, y_m)\},$$

where each datapoint $z_i \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is drawn i.i.d. from an unknown probability distribution $\mathcal{P}$. A learning algorithm is defined to be a function A that maps from $\mathcal{Z}^m$ into $\mathcal{F} \in \mathcal{Y}^{\mathcal{X}}$; in other words, a learning algorithm is a mapping from a training set D onto a function A_D (where the subscript makes explicit the dependency of the mapping on the given dataset), which is itself a mapping from the input space $\mathcal{X}$ to the output space $\mathcal{Y}$.

From D, we can construct the following datasets by (1) removing the i-th element

$$D^{\setminus i} = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{T_0}\},$$

and (2) replacing the i-th element

$$D^{i} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_{T_0}\},$$

where the replacement element z_i' is also assumed to be drawn from the same distribution $\mathcal{P}$ and is independent from D.

We measure the accuracy of the algorithm by defining a loss function; specifically, for a hypothesis $f \in \mathcal{F}$ and a datapoint $z \sim \mathcal{P}$, we denote the associated loss as $\ell(f, z)$. In order to accurately assess the performance of our algorithm, we will study the *generalization/testing error*, which is defined as

$$R(A, D) = \mathbb{E}_z[\ell(A_D, z)], \tag{6.5}$$

where the subscript z denotes that the expectation is taken with respect to the randomness in the example z. Since $\mathcal{P}$ is unknown, we cannot compute the generalization error without making unrealistically strong assumptions on the form of $\mathcal{P}, \ell$, or f. As is often the case, we use the *empirical/training error*

$$R_{emp}(A, D) = \frac{1}{m} \sum_{i=1}^{m} \ell(A_D, z_i) \tag{6.6}$$

as a simple estimator of the generalization error. Therefore, our goal is to use the empirical error to approximate the true generalization error. To simply matters, we will often use the following shorthand notations, $R \equiv R(A, D)$ and $R_{emp} \equiv R_{emp}(A, D)$.

We will make use of a generalization error bound that depends on the stability of the algorithm. Since one source of randomness an algorithm has to overcome is the sampling mechanism by which the data is generated, a way to quantify stability is to observe how changes in the training set can influence the hypothesis produced by the algorithm. With this intuition in mind, we now define one particular notion of stability that can be applied to a large class of algorithms, including regularization based algorithms.

Definition 6.3.1. (Uniform Stability) An algorithm A has uniform stability α

with respect to the loss function ℓ if $\forall S \in \mathcal{Z}^m, \forall i \in [m]$ the following holds:

$$\|\ell(A_D, \cdot) - \ell(A_{D^{\setminus i}}, \cdot)\|_\infty \leq \alpha.$$

Definition 6.3.2. A loss function ℓ defined on $\mathcal{F} \times \mathcal{Y}$ is σ-admissible with respect to $\mathcal{F}$ if the associated cost function c is convex with respect to its first argument and the following condition holds $\forall y_1, y_2 \in \mathcal{R}$, $\forall y' \in \mathcal{Y}$,

$$|c(y_1, y') - c(y_2, y')| \leq \sigma |y_1 - y_2|,$$

where $\mathcal{R} = \{y : \exists f \in \mathcal{F}, \exists x \in \mathcal{X}, f(x) = y\}$ is the domain of the first argument of c.

Remark 6.3.0.1. In the case of a quadratic loss function, for instance, this condition is verified if $\mathcal{Y}$ is bounded and $\mathcal{F}$ is totally bounded; i.e., there exists an $M < \infty$ such that

$$\forall f \in \mathcal{F}, \|f\|_\infty \leq M$$

and

$$\forall y \in \mathcal{Y}, |y| \leq M.$$

6.3.2 Results

In order for an algorithm to better generalize when given unseen data, regularization is often employed to reduce the complexity of the learned function at the expense of a larger training error. Although uniform stability may appear to be a strict condition, ridge regression has been shown to exhibit uniform stability, which is controlled by the regularization parameter. We begin, however, with a result from [10] on the uniform stability of reproducing kernel Hilbert spaces (RKHS) learning.

Theorem 6.3.1. *Let $\mathcal{F}$ be a reproducing kernel Hilbert space with kernel k such that $\forall x \in \mathcal{X}$, $k(x, x) = \langle \Phi(x), \Phi(x) \rangle \leq \kappa^2 < \infty$. Let ℓ be σ-admissible with respect to $\mathcal{F}$. The learning algorithm A defined by*

$$A_D = \arg\min_{g \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \ell(g, z_i) + \eta \|g\|_k^2,$$

has uniform stability α with respect to ℓ with

$$\alpha = \frac{\sigma^2 \kappa^2}{2\eta m}.$$

In the special case of ridge regression (quadratic complexity penalty term), [10] provide the following result.

Corollary 6.3.1. *The regularized least squares algorithm is defined by*

$$A_D = \arg\min_{g \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \ell(g, z_i) + \eta \|g\|_k^2,$$

where $\ell(f, z) = (f(x) - y)^2$. *The stability bound for this algorithm is*

$$\alpha \leq \frac{2\kappa^2 B^2}{\eta m}$$

so that for any $\delta \in (0, 1)$, *the generalization error bound holds with probability at least* $1 - \delta$,

$$R \leq R_{emp} + \frac{4\kappa^2 B^2}{\eta m} + \left(\frac{8\kappa^2 B^2}{\eta} + 2B \right) \sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{6.7}$$

6.3.3 Our setting

Returning to our setup, we have that $\Phi(x) = x$, yielding $k(x, x) = \|x\|^2$ and, thus, $\kappa^2 \leq N - 1$. Since we assumed that our entries are bounded by one in absolute value, we have that $B = 2$. Plugging in our parameters into Corollary 6.3.1, we obtain the following proposition:

Proposition 6.3.1. *For any* $\delta \in (0, 1)$, *the generalization/post-intervention error bound holds with probability at least* $1 - \delta$,

$$R \leq R_{emp} + \frac{16(N-1)}{\eta T_0} + \left(\frac{32(N-1)}{\eta} + 4 \right) \sqrt{\frac{\ln(1/\delta)}{2T_0}}, \tag{6.8}$$

where $R = \mathbb{E}[(Y_{1t} - \hat{M}_{1t})^2]$ *for any* $t > T_0$, *and* $R_{emp} = (1/T_0) \sum_{t=1}^{T_0} (Y_{1t} - \hat{M}_{1t})^2$.

From (6.8), we observe that the generalization error decreases as the regularization parameter increases. Again, our gain in the post-intervention regime comes at the expense of a greater pre-intervention error. However, since our objective is to analyze the impact of a policy by comparing the post-intervention observed and counterfactual outcomes, we should prioritize a smaller generalization error over a smaller training error.

Moreover, with the exception of the training error term, R_{emp}, of (6.8), all other terms asymptotically decay to 0. Therefore, our estimator is consistent if the training error also converges to 0. Following the proof of Theorem 4.1.1 (relegated to the appendix), we have that

$$\begin{aligned} R_{emp} &= \frac{1}{T_0}\left\|Y_1^- - \hat{M}_1^-\right\|^2 \\ &\leq \frac{1}{T_0}\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\|^2 \|\beta\|^2 + \frac{\eta}{T_0}\|\beta\|^2 + \frac{1}{T_0}\left\|\epsilon_1^-\right\|^2. \end{aligned}$$

However, if we apply our pre-processing procedure, then our training error converges to 0 in expectation, i.e. $\mathbb{E}[R_{emp}] \to 0$ as $T_0 \to \infty$.

Remark 6.3.1.1. For completeness, we note that our input and output spaces are $\mathcal{X} = [-1, 1]^{N-1}$ and $\mathcal{Y} = [-1, 1]$, respectively. Similarly, our training data, of size $m = T_0$, takes the form

$$D = \{z_t = (Y_{1t}, X_{it}) : 2 \leq i \leq N, t \in [T_0]\}.$$

Recall, however, that we will use Step one of our robust algorithm (described in Section 3) to transform X_{it} into $\hat{M}_{it}$ for all $2 \leq i \leq N$ and $t \in [T_0]$.

6.4 Choosing the Regularization Hyperparameter, η

From Theorem 6.2.1 and Theorem 6.2.2, we recognize that the regularization parameter 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 the regularization hyperparameter, η, is to employ cross-validation. Under the simplest hold-out cross-validation scenario, an "appropriate" proportion of the training data (in this case, the pre-intervention data) is set aside as the validation set, and is not used in the learning process as to prevent data leakage. Using only the training data not included in the validation set, an estimate $\hat{\beta}_\eta$ is produced for finitely many choices of η. The choice of η that minimizes the MSE with respect to the observed values in the validation set is subsequently determined to be the optimal regularization value, and is used to learn the final $\hat{\beta}_\eta$ with all of the given training data (validation set included). A simple, but powerful variant to the described cross-validation method is k-fold cross validation, where the training data is now

partitioned into k subsets. For each of the finitely many candidates of η, the model selection process now uses $k-1$ of the subsets as the training data and the k-th subset as the validation set. This process is applied until all k subsets have been used as the validation set (essentially applying the hold-out method k times), whereby the validation error is then averaged over all k subsets. When the training data is small, it is commonplace to choose $k = T_0$ (the number of datapoints in the training set) – this method is known as leave-one-out (LOO) cross-validation.

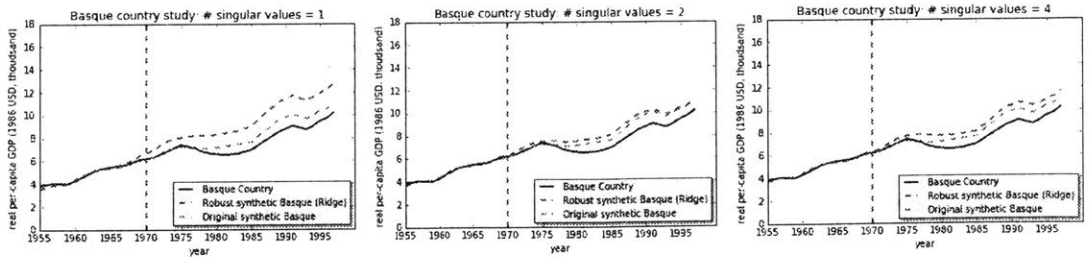
However, since time-series data often have a natural temporal ordering and causal effect, we recommend employing the forward chaining strategy. Although the forward chaining strategy is similar to LOO cross-validation, an important distinction is that forward chaining does not break the temporal ordering 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).

6.5 Experimental Results

We investigate the Basque Country case study through the lens of regularization. Throughout the experiments, we employ the forward chaining strategy to learn the regularization parameter η.

6.5.1 Ridge regression

We display the resulting figures after applying ridge regression under varying thresholding scenarios.

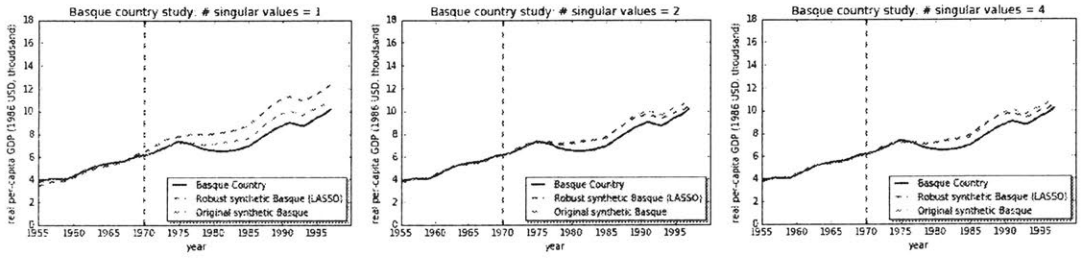


(a) Top singular value. **(b)** Top two singular values. **(c)** Top four singular values.

Figure 6-1: Trends in per-capita GDP between Basque Country vs. synthetic Basque Country.

6.5.2 LASSO

Similarly, we display the resulting figures after applying LASSO under varying thresholding scenarios. Since the LASSO strategy seeks sparse solutions, it is not surprising that we find that the resulting estimates derived from our regularized robust setting are nearly identical to that of original synthetic control estimates since the latter method indirectly learns sparse solutions by enforcing the parameter values to lie within the simplex. Due to its sparsity, the LASSO solution, also provides an interpretable solution, which some practitioners may find valuable.



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

Figure 6-2: Trends in per-capita GDP between Basque Country vs. synthetic Basque Country.

Chapter 7

Bayesian Synthetic Control

7.1 A Bayesian Perspective

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 our uncertainty of our model 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, $p(\beta)$, over β 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 $M_{\cdot t} = [M_{it}]_{2 \leq i \leq N}$; we define $X_{\cdot t}$ similarly. Denoting the pre-intervention data as $D = \{(Y_{1t}, M_{\cdot t}) : t \in [T_0]\}$, the likelihood function $p(Y_1^- \mid \beta, \boldsymbol{M}^-)$ is expressed as

$$p(Y_1^- \mid \beta, \hat{\boldsymbol{M}}^-) = \mathcal{N}((\boldsymbol{M}^-)^T \beta, \sigma^2 \boldsymbol{I}), \tag{7.1}$$

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 over β is $p(\beta \mid D) = \mathcal{N}(\beta_D, \boldsymbol{\Sigma}_D)$ where

$$\boldsymbol{\Sigma}_D = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{M}^{-}(\boldsymbol{M}^{-})^T\right)^{-1} \tag{7.2}$$

$$\beta_D = \boldsymbol{\Sigma}_D\left(\frac{1}{\sigma^2}\boldsymbol{M}^{-}Y_1^{-} + \boldsymbol{\Sigma}_0^{-1}\beta_0\right). \tag{7.3}$$

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 $D = \{(Y_{1t}, \hat{M}_{\cdot t}) : t \in [T_0]\}$. Putting everything together, we have that $p(\beta \mid D) = \mathcal{N}(\beta_D, \boldsymbol{\Sigma}_D)$ whereby

$$\boldsymbol{\Sigma}_D = \left(\alpha\boldsymbol{I} + \frac{1}{\sigma^2}\hat{\boldsymbol{M}}^{-}(\hat{\boldsymbol{M}}^{-})^T\right)^{-1} \tag{7.4}$$

$$\beta_D = \frac{1}{\sigma^2}\boldsymbol{\Sigma}_D\hat{\boldsymbol{M}}^{-}Y_1^{-} \tag{7.5}$$

$$= \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\hat{\boldsymbol{M}}^{-}(\hat{\boldsymbol{M}}^{-})^T + \alpha\boldsymbol{I}\right)^{-1}\hat{\boldsymbol{M}}^{-}Y_1^{-}. \tag{7.6}$$

7.1.1 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 (**??**) 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 regular-

ized error (??) with $\eta = \alpha\sigma^2$. We define the MAP estimate, $\hat{\beta}_{\text{MAP}}$, as

$$\begin{aligned}
\hat{\beta}_{\text{MAP}} &= \arg\max_{\beta \in \mathbb{R}^{N-1}} \ln p(\beta \mid D) \\
&= \arg\min_{\beta \in \mathbb{R}^{N-1}} \frac{1}{2} \left\| Y_1^- - (\hat{\boldsymbol{M}}^-)^T \beta \right\|^2 + \frac{\alpha\sigma^2}{2} \|\beta\|^2 \\
&= \left(\hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T + \alpha\sigma^2 \boldsymbol{I} \right)^{-1} \hat{\boldsymbol{M}}^- Y_1^-.
\end{aligned}$$

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

$$\hat{M}_1 = \hat{\boldsymbol{M}}^T \hat{\beta}_{\text{MAP}}. \tag{7.7}$$

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.

7.1.2 Fully Bayesian treatment

Although we have treated β as a random variable attached with a prior distribution, we must go beyond point estimates in order 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

$$p(Y_{1t} \mid \hat{M}_{\cdot t}, D) = \int p(Y_{1t} \mid \hat{M}_{\cdot t}, \beta)\ p(\beta \mid D)\, d\beta \tag{7.8}$$

$$= \mathcal{N}(\hat{M}_{\cdot t}^T \beta_D, \sigma_D^2), \tag{7.9}$$

where

$$\sigma_D^2 = \sigma^2 + \beta_D^T \mathbf{\Sigma}_D \beta_D. \tag{7.10}$$

Note that $p(\beta \mid D)$ is the posterior distribution over the synthetic control parameter and is governed by (7.4) and (7.6). 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

$$\mathbb{E}[L] = \int \int L(Y_{1t}, g(\hat{M}_{\cdot t})) \cdot p(Y_{1t}, \hat{M}_{\cdot t})\, dY_{1t}\, d\hat{M}_{\cdot t} \tag{7.11}$$

$$= \int \left(\int L(Y_{1t}, g(\hat{M}_{\cdot t})) \cdot p(Y_{1t} \mid \hat{M}_{\cdot t})\, dY_{1t} \right) p(\hat{M}_{\cdot t})\, d\hat{M}_{\cdot t}, \tag{7.12}$$

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

$$\hat{g}(\cdot) = \arg\min_{g(\cdot)} \mathbb{E}[L(Y_{1t}, g(\hat{M}_{\cdot t}))]. \tag{7.13}$$

Since $p(\hat{M}_{\cdot t}) \geq 0$, we can minimize (7.12) by selecting $\hat{g}(\hat{M}_{\cdot t})$ to minimize the term within the parenthesis for each individual value of Y_{1t}, i.e.,

$$\hat{M}_{1t} = \hat{g}(\hat{M}_{\cdot t}) \tag{7.14}$$

$$= \arg\min_{g(\cdot)} \int L(Y_{1t}, g(\hat{M}_{\cdot t})) \cdot p(Y_{1t} \mid \hat{M}_{\cdot t})\, dY_{1t}. \tag{7.15}$$

As suggested by (7.15), 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.

7.1.3 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

$$\mathbb{E}[L] = \int \left(\int (Y_{1t} - g(\hat{M}_{\cdot t}))^2 \cdot p(Y_{1t} \mid \hat{M}_{\cdot t})\, dY_{1t} \right) p(\hat{M}_{\cdot t})\, d\hat{M}_{\cdot t}. \tag{7.16}$$

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

$$\hat{M}_{1t} = \mathbb{E}[Y_{1t} \mid \hat{M}_{\cdot t}, D] \tag{7.17}$$

$$= \int Y_{1t}\, p(Y_{1t} \mid \hat{M}_{\cdot t}, D) dY_{1t} \tag{7.18}$$

$$= \hat{M}_{\cdot t}^T \beta_D. \tag{7.19}$$

To see why the BLS estimate is the minimizer of a quadratic loss criterion, we analyze a simple scalar case where we denote x as the given feature vector and y as the target value. As a refresher, recall that

$$\hat{y}_{\text{BLS}}(x) = \arg\min_a \int (y-a)^2 p(y \mid x) dy.$$

Differentiating the above integral, we obtain

$$\begin{aligned}\frac{\partial}{\partial a}\int (y-a)^2 p(y \mid x) dy &= \int \frac{\partial}{\partial a}(y-a)^2 p(y \mid x) dy \\ &= -2\int (y-a) p(y \mid x) dy.\end{aligned}$$

Setting the expression to zero at $a = \hat{y}_{\text{BLS}}(x)$ gives us our desired result:

$$\begin{aligned}\Big[\int (y-a) p(y \mid x) dy\Big]_{a=\hat{y}_{\text{BLS}}(x)} &= \int y p(y \mid x) dy - \hat{y}_{\text{BLS}}(x) \int p(y \mid x) dy \\ &= \mathbb{E}[y \mid x] - \hat{y}_{\text{BLS}}(x) \int p(y \mid x) dy \\ &= \mathbb{E}[y \mid x] - \hat{y}_{\text{BLS}}(x) = 0.\end{aligned}$$

A similar derivation applies to vector case.

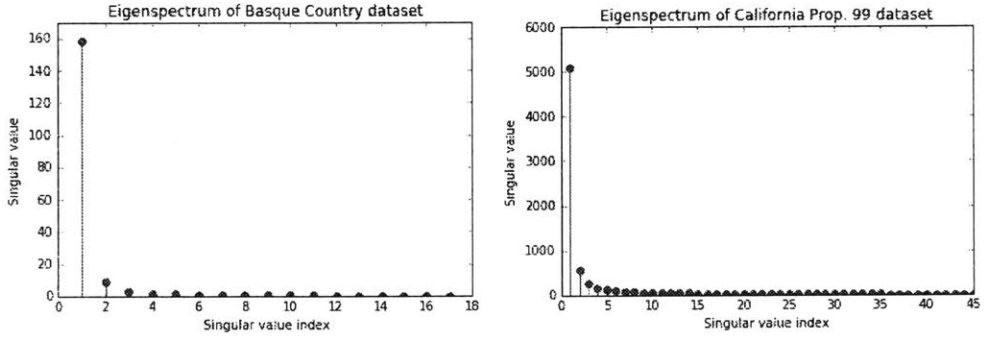
Remark 7.1.0.1. Since the noise variance σ^2 is often unknown in practice, we can introduce another conjugate prior distribution $p(\beta, 1/\sigma^2)$ given by the Gaussian-gamma distribution. This prior yields a Student's t-distribution for the predictive probability distribution.

7.2 Experimental Results

We will now study both the Basque Country and California Prop. 99 case studies under a Bayesian setting. We estimate the noise variance by using the unbiased correction

of the maximum likelihood estimate, i.e. $\hat{\sigma}^2 = 1/(T_0 - 1)\sum_{t=1}^{T_0}(Y_{1t} - \bar{Y})^2$, where $\bar{Y}$ denotes the sample mean. From our results, we will 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 denoising process. Therefore, we have plotted the eigenspectrum of the singular values 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.



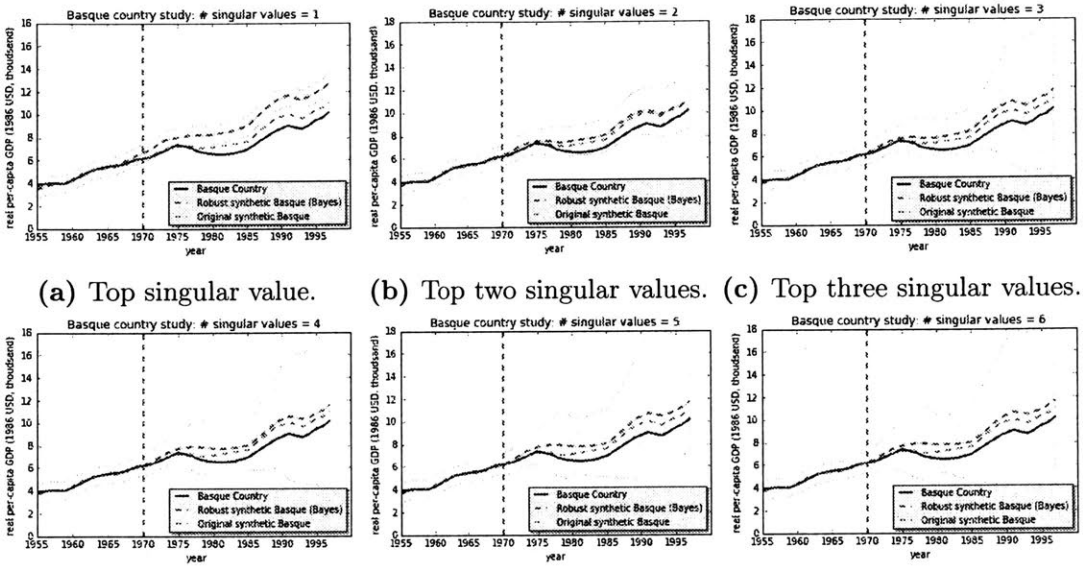
(a) Eigenspectrum of Basque data. **(b)** Eigenspectrum of California data.

7.2.1 Basque Country

We plot the resulting Bayesian estimates in the figures below under varying thresholding conditions. From previous discussions, we know that a Gaussian prior for the latent parameter β amounts to estimating $\hat{\beta}$ under a ridge regression setting for a particular choice of η. Therefore, it is not surprising that the posterior mean of our predictive distribution closely resembles the counterfactual trajectory derived for ridge regression.

Furthermore, it is interesting to note that our uncertainty grows dramatically once

we include more than two singular values in the thresholding process.



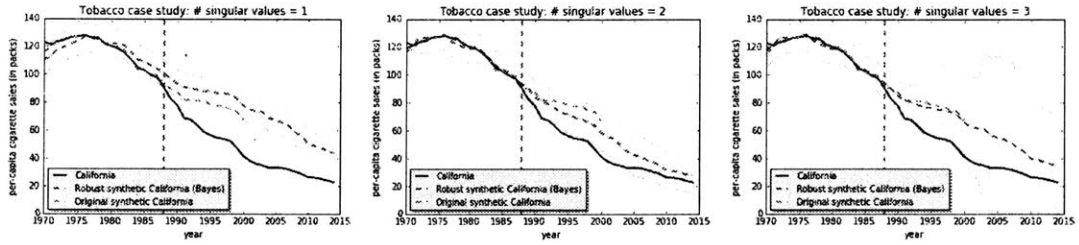
(a) Top singular value. (b) Top two singular values. (c) Top three singular values.

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

Figure 7-2: Trends in per-capita GDP between Basque Country vs. synthetic Basque Country.

7.2.2 California Anti-tobacco Legislation

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.

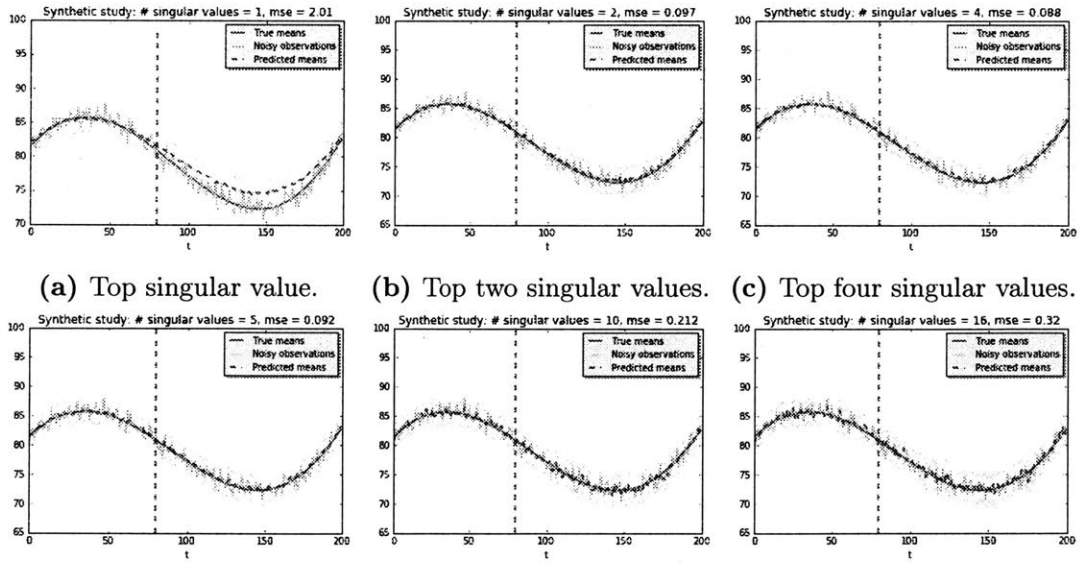


(a) Top singular value. **(b)** Top two singular values. **(c)** Top three singular values.

Figure 7-3: Trends in per-capita cigarette sales between California vs. synthetic California.

7.2.3 Synthetic data

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.

Bibliography

[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] S. Athey and G. Imbens. The state of applied econometrics - causality and policy evaluation. 2016.

[7] 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.

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

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

[10] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, pages 499–526, 2002.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[12] 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.

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

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

[15] N. Doudchenko and G. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. 2016.

[16] 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.

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

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

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

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

Appendix A

Useful Theorems

We present useful theorems that we will frequently employ in the following sections to prove our desired results.

Theorem A.0.1. 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,

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

Remark A.0.1.1. See [14] for references to the proof of the statement.

Theorem A.0.2. 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.0.2.1. 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.0.3. Theorem 3.4 of [14]
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 $\delta^2 \in [0,1]$,

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

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

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

where $C(\zeta)$ *depends only on* η *and* ζ, *and* c *depends only on* η. *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.0.3.1. The proof of Theorem A.0.3 can be found in [14] under Theorem 3.4.

Appendix B

Linear Regression

Throughout the proofs in this chapter (and the appendix in general), we denote C_1, C_2, and c as universal positive constants that depend on the choice of $\eta \in (0,1)$, if applicable. The values for C_1, C_2, and c may change from line to line or even within a line.

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 $Y_{ij} = X_{ij}\mathbb{1}_{(X_{ij} \text{ observed})}$, where $\mathbb{1}$ is the indicator function. Under this model, for all units in the donor pool and across all time periods,

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

Recall that $\hat{p}$ denotes the proportion of observed entries in the data matrix $\boldsymbol{X}$. We define the event E_1 as

$$E_1 := \{|\hat{p} - p| \leq \eta p/z\}, \tag{B.1}$$

for some choice of $\eta \in (0,1)$ and $z \geq 0$. By Bernstein's inequality, for any $t \geq 0$,

$$\mathbb{P}(|\hat{p} - p| \geq t) \leq 2\exp\left\{-\frac{3(N-1)Tt^2}{6p(1-p) + 2t}\right\}.$$

For concreteness, we arbitrarily choose $z = 20$ in the proofs of the Lemmas and

Theorem 4.1.1. As a result, we have that

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

Finally, we assume that the total number of units N, and hence the size of the donor pool, is sublinear in T, i.e. $N = o(T)$. However, for the sake of analytical simplicity, we proceed with our analysis under the assumption that N is fixed. In other words, the only dimension that increases is the number of pre-intervention periods T_0.

We begin by proving a useful lemma that allows us to bound the spectral norm of a *child* submatrix, e.g. $\boldsymbol{A}$, by the spectral norm of the larger, *parent* matrix, e.g. $\boldsymbol{C} = \left[\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{B} \end{array} \right]$.

Lemma B.0.1. *Suppose $\boldsymbol{C}$ is an $m \times n$ matrix composed of an $m \times p$ submatrix $\boldsymbol{A}$ and an $m \times (n-p)$ submatrix $\boldsymbol{B}$, i.e., $\boldsymbol{C} = \left[\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{B} \end{array} \right]$. Then, the spectral (operator) norms of $\boldsymbol{A}$ and $\boldsymbol{B}$ are bounded above by the spectral norm of $\boldsymbol{C}$,*

$$\max\{\|\boldsymbol{A}\|, \|\boldsymbol{B}\|\} \leq \|\boldsymbol{C}\|.$$

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

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

Let $\sigma_1, \ldots, \sigma_n$ be the eigenvalues of $\boldsymbol{C}^T\boldsymbol{C}$ in increasing order and repeated by multiplicities. Let $\tau_1, \ldots, \tau_p$ be the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ in increasing order and repeated by multiplicities. By the Poincaré separation Theorem A.0.2, we have for all $j < p + 1$,

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

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

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

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

B.1 Pre-intervention analysis

In this section we prove Theorem 4.1.1 and Corollary 4.1.1 (restated below).

We now prove the following two key Lemmas, which, when amalgamated, provide us with a universal upper bound on the pre-intervention MSE for any general noise model that satisfies the conditions described in section **??**. Moreover, the following Lemmas allow us to express the pre-intervention MSE in a way that highlights the inherent bias-variance tradeoff of the algorithm with respect to the choice of μ.

Remark B.1.0.1. To ease the notational complexity of the following Lemma B.1.1 proof, we will make use of the following notations for **only** this derivation:

$$\boldsymbol{Q} := (\boldsymbol{M}^{-})^{T} \tag{B.2}$$

$$\hat{\boldsymbol{Q}} := (\hat{\boldsymbol{M}}^{-})^{T} \tag{B.3}$$

such that

$$M_1^{-} := \boldsymbol{Q}\beta \tag{B.4}$$

$$\hat{M}_1^{-} := \hat{\boldsymbol{Q}}\hat{\beta}. \tag{B.5}$$

Lemma B.1.1. *Let β be defined as the vector of weights such that $M_1^{-} = (\boldsymbol{M}^{-})^{T}\beta$ and has minimum norm (since β may not be unique). Then, the universal, unnormalized pre-intervention MSE is bounded above as*

$$\mathbb{E}\left\|\hat{M}_1^{-} - M_1^{-}\right\|^2 \leq \mathbb{E}\left\|\hat{\boldsymbol{M}}^{-} - \boldsymbol{M}^{-}\right\|^2 \|\beta\|^2 + 2\sigma^2|S|. \tag{B.6}$$

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

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

where the last inequality from the submultiplicative property of induced norms. Taking the expectation, we arrive at the inequality

$$\mathbb{E}\left\|\hat{M}_1^- - M_1^-\right\|^2 \leq \mathbb{E}\left\|\boldsymbol{Q} - \hat{\boldsymbol{Q}}\right\|^2 \|\beta\|^2 + 2\mathbb{E}\left\|\epsilon_1^-\right\|^2 + 2\mathbb{E}[\langle \epsilon_1^-, (\boldsymbol{Q} - \hat{\boldsymbol{Q}})\beta\rangle] + 2\mathbb{E}[\langle -\epsilon_1^-, Y_1^- - \hat{\boldsymbol{Q}}\hat{\beta}\rangle]. \tag{B.7}$$

We will now deal with the two inner products on the right hand side of equation (B.7). 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 $\hat{\boldsymbol{Q}} := \hat{\boldsymbol{M}}^-$ depends only on the noise terms for $i \neq 1$; i.e., the construction of $\hat{\boldsymbol{Q}} := \hat{\boldsymbol{M}}^-$ 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., $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$. As a result,

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

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

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

where (a) follows from the same independence argument used in evaluating the first inner product. Finally, we incorporate the above results to (B.7) to arrive at the inequality

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

■

Lemma B.1.2. *Let $\lambda_1, \ldots, \lambda_{N-1}$ be the singular values of $p\boldsymbol{M}$ in decreasing order and repeated by multiplicities, with $\lambda^* = \max_{i \notin S} \lambda_i$. Then assuming E_1 occurs,*

$$\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\| \leq \frac{C_1}{p}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \left\|(\hat{p} - p)\boldsymbol{M}^-\right\|\Big). \tag{B.8}$$

Proof. Recall that $s_1, \ldots, s_{N-1}$ are the singular values of $\boldsymbol{Y}$ in decreasing order with repeated multiplicities. By Theorem A.0.1, we have $s_i \leq \lambda_i + \|\boldsymbol{Y} - p\boldsymbol{M}\|$ for all i. Thus, assuming E_1 occurs

$$\begin{aligned}
p\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\| &\leq C_1\hat{p}\left\|\hat{\boldsymbol{M}}^- - \boldsymbol{M}^-\right\| \\
&\leq C_1\left\|\hat{p}\hat{\boldsymbol{M}}^- - p\boldsymbol{M}^-\right\| + C_1\left\|(\hat{p} - p)\boldsymbol{M}^-\right\| \\
&\leq C_1\left\|\boldsymbol{Y}^- - \hat{p}\hat{\boldsymbol{M}}^-\right\| + C_1\left\|\boldsymbol{Y}^- - p\boldsymbol{M}^-\right\| + C_1\left\|(\hat{p} - p)\boldsymbol{M}^-\right\| \\
&\overset{(a)}{\leq} C_1\left\|\boldsymbol{Y} - \hat{p}\hat{\boldsymbol{M}}\right\| + C_1\left\|\boldsymbol{Y}^- - p\boldsymbol{M}^-\right\| + C_1\left\|(\hat{p} - p)\boldsymbol{M}^-\right\| \\
&= C_1 \max_{i \notin S} s_i + C_1\left\|\boldsymbol{Y}^- - p\boldsymbol{M}^-\right\| + C_1\left\|(\hat{p} - p)\boldsymbol{M}^-\right\| \\
&\leq C_1 \max_{i \notin S}\Big(\lambda_i + \|\boldsymbol{Y} - p\boldsymbol{M}\|\Big) + C_1\left\|\boldsymbol{Y}^- - p\boldsymbol{M}^-\right\| + C_1\left\|(\hat{p} - p)\boldsymbol{M}^-\right\| \\
&\overset{(b)}{\leq} C_1 \max_{i \notin S} \lambda_i + (C_1 + 1)\|\boldsymbol{Y} - p\boldsymbol{M}\| + C_1\left\|(\hat{p} - p)\boldsymbol{M}^-\right\| \\
&\leq C_1\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \left\|(\hat{p} - p)\boldsymbol{M}^-\right\|\Big),
\end{aligned}$$

where (a) and (b) follow from Lemma B.0.1. Note that we have absorbed $(N-1)\|\beta\|^2$ into the universal constant C_2 since N is assumed to be fixed. The resulting form is also more aethestically appealing to display, hence the absorption. ■

Theorem (4.1.1). *The pre-intervention error of the algorithm can be bounded as*

$$\mathrm{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p^2 T_0}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \left\|(\hat{p} - p)\boldsymbol{M}^-\right\|\Big)^2\|\beta\|^2 + \frac{2\sigma^2|S|}{T_0} \tag{B.9}$$

$$+ C_2(N-1)\|\beta\|^2 e^{-cp(N-1)T}. \tag{B.10}$$

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. We invoke Lemmas B.1.2 and B.1.1 and apply the total law of probability to

arrive at the inequality

$$\begin{aligned}
\mathbb{E}\left\|\hat{\boldsymbol{M}}^{-}-\boldsymbol{M}^{-}\right\|^{2} &= \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^{-}-\boldsymbol{M}^{-}\right\|^{2} \mid E_1\right]\mathbb{P}(E_1)+\mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^{-}-\boldsymbol{M}^{-}\right\|^{2} \mid E_1^c\right]\mathbb{P}(E_1^c) \\
&\leq \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^{-}-\boldsymbol{M}^{-}\right\|^{2} \mid E_1\right]+\mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^{-}-\boldsymbol{M}^{-}\right\|^{2} \mid E_1^c\right]\mathbb{P}(E_1^c) \\
&\leq \frac{C_1}{p^2}\mathbb{E}\Big(\lambda^{*}+\|\boldsymbol{Y}-p\boldsymbol{M}\|+\left\|(\hat{p}-p)\boldsymbol{M}^{-}\right\|\Big)^{2}+\mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^{-}-\boldsymbol{M}^{-}\right\|^{2} \mid E_1^c\right]\mathbb{P}(E_1^c) \\
&\leq \frac{C_1}{p^2}\mathbb{E}\Big(\lambda^{*}+\|\boldsymbol{Y}-p\boldsymbol{M}\|+\left\|(\hat{p}-p)\boldsymbol{M}^{-}\right\|\Big)^{2}+C_2(N-1)T_0\mathbb{P}(E_1^c) \\
&\leq \frac{C_1}{p^2}\mathbb{E}\Big(\lambda^{*}+\|\boldsymbol{Y}-p\boldsymbol{M}\|+\left\|(\hat{p}-p)\boldsymbol{M}^{-}\right\|\Big)^{2}+C_2(N-1)T_0e^{-c(N-1)Tp},
\end{aligned}$$

where E_1^c denotes the complementary event of E_1. Dividing throughout by T_0 gives the desired bound:

$$\begin{aligned}
\mathrm{MSE}(\hat{M}_1^{-}) &\leq \frac{1}{T_0}\mathbb{E}\left\|\hat{\boldsymbol{M}}^{-}-\boldsymbol{M}^{-}\right\|^{2}\|\beta\|^{2}+\frac{2\sigma^2|S|}{T_0} \\
&\leq \frac{C_1}{p^2T_0}\mathbb{E}\Big(\lambda^{*}+\|\boldsymbol{Y}-p\boldsymbol{M}\|+\|(\hat{p}-p)\boldsymbol{M}\|\Big)^{2}\|\beta\|^{2}+\frac{2\sigma^2|S|}{T_0}+C_2e^{-c(N-1)Tp}.
\end{aligned}$$

Note that we have absorbed $(N-1)\|\beta\|^2$ into the universal constant C_2 since N is assumed to be fixed. The resulting form is also more aethestically appealing to display, hence the absorption. ■

Corollary (4.1.1). *Let* $\mathrm{rank}(\boldsymbol{M})=k$ *for some* $1\leq k<N-1$. *Let the choice of* μ *be such that* $|S|=k$. *Suppose* $\sigma^2p+p(1-p)\geq T^{-1+\zeta}$ *for some* $\zeta>0$. *Let* $T\leq\alpha T_0$ *for some constant* $\alpha>1$. *Then*

$$\lim_{T_0\to\infty}\mathrm{MSE}(\hat{M}_1^{-})\leq\frac{C_1\|\beta\|^2}{p}(\sigma^2+(1-p)).$$

Proof. For the rest of this proof, let $z=N$ in (B.1), which is assumed to be a fixed constant. As a result, note that the exponent in the rightmost term of (B.12) is now C_2e^{-cpT}, where the $1/N$ factor has been absorbed into c. Further, recall that λ_i are the singular values of $p\boldsymbol{M}$ in decreasing order with repeated multiplicities. Thus, assuming $\mathrm{rank}(\boldsymbol{M})=k$ and $|S|=k$,

$$\lambda^{*}=\max_{i\notin S}\lambda_i=0.$$

Consequently, we have that

$$\mathrm{MSE}(\hat{M}_1^-) \leq \frac{C_1\|\beta\|^2}{p^2T_0}\mathbb{E}\Big(\|\boldsymbol{Y}-p\boldsymbol{M}\|+\big\|(\hat{p}-p)\boldsymbol{M}^-\big\|\Big)^2+\frac{2\sigma^2|S|}{T_0}+C_2e^{-cpT} \quad \text{(B.11)}$$

$$= \frac{C_1\|\beta\|^2}{p^2T_0}\mathbb{E}\Big(\|\boldsymbol{Y}-p\boldsymbol{M}\|+\big\|(\hat{p}-p)\boldsymbol{M}^-\big\|\Big)^2+\frac{2\sigma^2k}{T_0}+C_2e^{-cpT}. \quad \text{(B.12)}$$

Observe that for $2 \leq i \leq N$ and $j \in [T]$,

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

Consequently, we define the event $E_2 := \{\|\boldsymbol{Y}-p\boldsymbol{M}\| \leq (2+\eta/2)\delta\sqrt{T}\}$ where we also define $\delta^2 = p\sigma^2 + p(1-p)$. By Theorem A.0.3, we have that $\mathbb{P}(E_2) \geq 1 - 2e^{-c\delta^2 T}$ for $\delta^2 \in [0,1]$.

Assume E_1 and E_2 occur. Since $\|\boldsymbol{M}^-\| \leq \sqrt{(N-1)T_0}$, we have that

$$\begin{aligned}
\mathbb{E}\Big(\|\boldsymbol{Y}-p\boldsymbol{M}\| \cdot \big\|(\hat{p}-p)\boldsymbol{M}^-\big\|\Big) &\leq (2+\eta/2)\delta\sqrt{T}\cdot\mathbb{E}\big\|(\hat{p}-p)\boldsymbol{M}^-\big\| \\
&\leq \frac{(2+\eta/2)\delta\sqrt{T}\eta p}{N}\|\boldsymbol{M}^-\| \\
&\leq \frac{(2+\eta/2)\delta\sqrt{T}\eta p}{N}\sqrt{(N-1)T_0} \\
&\leq \frac{(2+\eta/2)\delta\eta pT}{\sqrt{N-1}} \\
&= \frac{(2\eta+\eta^2/2)\delta pT}{\sqrt{N-1}}.
\end{aligned}$$

Simultaneously, note that $\mathbb{E}(\hat{p}-p)^2 = p(1-p)/(N-1)T$. As a result,

$$\begin{aligned}
\mathbb{E}\big\|(\hat{p}-p)\boldsymbol{M}^-\big\|^2 &= \big\|\boldsymbol{M}^-\big\|^2\,\mathbb{E}(\hat{p}-p)^2 \\
&\leq \frac{p(1-p)T_0}{T} \\
&\leq p(1-p).
\end{aligned}$$

Applying the above inequalities into the first term on the RHS of (B.12) and using

the total of probability, we have that

$$\mathbb{E}\Big(\|\boldsymbol{Y}-p\boldsymbol{M}\|+\big\|(\hat{p}-p)\boldsymbol{M}^{-}\big\|\Big)^{2}\leq C_{3}\delta^{2}T+p(1-p)+\frac{(4\eta+\eta^{2})\delta pT}{\sqrt{N-1}}+C_{4}(N-1)Te^{-c\delta^{2}T},$$

whereby C_3 depends on the choice of η, and C_4 depends on both ζ and η. Thus, dividing by T_0 gives

$$\frac{1}{T_{0}}\mathbb{E}\Big(\|\boldsymbol{Y}-p\boldsymbol{M}\|+\big\|(\hat{p}-p)\boldsymbol{M}^{-}\big\|\Big)^{2}\leq C_{3}\delta^{2}+\frac{(4\eta+\eta^{2})\delta p}{\sqrt{N-1}}+\frac{p(1-p)}{T_{0}}+C_{4}(N-1)e^{-c\delta^{2}T}.$$

Clearly, the last two terms go to 0 in the limit as $T_0 \to \infty$. Returning to (B.12), we subsequently have

$$\lim_{T_{0}\to\infty}\mathrm{MSE}(\hat{M}_{1}^{-})\leq\frac{C_{1}\delta^{2}\|\beta\|^{2}}{p^{2}}+\frac{C_{1}(4\eta+\eta^{2})\delta p\|\beta\|^{2}}{p^{2}\sqrt{N-1}} \tag{B.13}$$

$$=\frac{C_{1}(\sigma^{2}p+p(1-p))\|\beta\|^{2}}{p^{2}}+\frac{C_{1}(4\eta+\eta^{2})\sqrt{\sigma^{2}p+p(1-p)}\|\beta\|^{2}}{p\sqrt{N-1}} \tag{B.14}$$

$$=\frac{C_{1}(\sigma^{2}+(1-p))\|\beta\|^{2}}{p}+\frac{C_{1}(4\eta+\eta^{2})\sqrt{\sigma^{2}+(1-p)}\|\beta\|^{2}}{\sqrt{p(N-1)}} \tag{B.15}$$

$$\leq\frac{C_{1}\|\beta\|^{2}}{\sqrt{p}}\Big(\frac{\sigma^{2}+(1-p)}{\sqrt{p}}+\frac{(4\eta+\eta^{2})\sqrt{\sigma^{2}+(1-p)}}{\sqrt{N-1}}\Big). \tag{B.16}$$

Since the second term of (B.16) depends on the choice of $\eta \in (0,1)$, we can essentially render it as a negligibly small quantity by choosing a small enough η. Therefore, the asymptotic error bound of (B.16) is dominated by the first term. In addition, if we let $N = o(T)$ grow without bound, then the term also disappears in the asymptotic regime. Note that the exponential term of (B.12) still decays to 0 when we choose $z = N$ and let $N \to \infty$, so long as $N = o(T)$. As a result, we only display the first term in the theorem above purely for aesthetic purposes. ■

B.2 Consistency: block partitioning

Theorem (4.1.2). *Let* $\mathrm{rank}(\bar{\boldsymbol{M}}^{-}) = k$ *for some* $1 \leq k < N-1$. *Let the choice of* μ *be such that* $|S| = k$. *Then*

$$\lim_{T_{0}\to\infty}\mathrm{MSE}(\hat{\bar{M}}_{1}^{-})=0.$$

Proof. We prove Theorem 4.1.2 following the proofs of Theorem 4.1.1 and Corollary 4.1.1, using the block partitioned matrices instead. We first define $\bar{\boldsymbol{E}}^{-} = [\bar{\epsilon}_{ij}]_{2 \le i \le N, j \le \tau}$ with entries

$$\bar{\epsilon}_{ij} = \frac{1}{\tau} \sum_{t \in B_j} \epsilon_{it}. \tag{B.17}$$

We define $\bar{\epsilon}_{1j}$ for $j \in [\tau]$ in the same manner as (B.17). Consequently, for all $i \in [N]$ and $j \in [\tau]$, we maintain the generalized factor model relationship, $\bar{X}_{ij} = \bar{M}_{ij} + \bar{\epsilon}_{ij}$. As a byproduct, we have in matrix form, $\bar{\boldsymbol{X}}^{-} = \bar{\boldsymbol{M}}^{-} + \bar{\boldsymbol{E}}^{-}$. Under this construction, the noise entries remain zero-mean random variables: $\mathbb{E}[\bar{\epsilon}_{ij}] = 0$. However, the variance of each noise term is now rescaled by $1/\tau$,

$$\mathrm{Var}(\bar{\epsilon}_{ij}) = \frac{1}{\tau^2} \sum_{t \in B_j} \mathrm{Var}(\epsilon_{it}) = \frac{\sigma^2}{\tau}.$$

For notational purposes, let $\bar{\sigma}^2 = \mathrm{Var}(\bar{\epsilon}_{ij})$. We now show that the key assumption of (2.7) still holds under this setting with respect to the newly defined variables. In particular, for every partition $j \in [\tau]$ of row one,

$$\begin{aligned}
\bar{M}_{1j} &= \frac{1}{\tau} \sum_{t \in B_j} M_{1t} \\
&= \frac{1}{\tau} \sum_{t \in B_j} \Big(\sum_{k=2}^{N} \beta_k M_{kt} \Big) \\
&= \sum_{k=2}^{N} \beta_k \Big(\frac{1}{\tau} \sum_{t \in B_j} M_{kt} \Big) \\
&= \sum_{k=2}^{N} \beta_k \bar{M}_{kj}.
\end{aligned}$$

As a result, we can express $\bar{M}_1^{-} = (\bar{\boldsymbol{M}}^{-})^T \beta$ for the same β as in (2.7).

Following the same setup as before, we define $\bar{\boldsymbol{Y}}^{-} = [\bar{Y}_{ij}]_{2 \le i \le N, j \le \tau}$ where $\bar{Y}_{ij} = \bar{X}_{ij}$ if $\bar{X}_{ij}$ is observed and 0 otherwise. In most practical cases, the proposed averaging pre-processing step would produce a matrix $\bar{\boldsymbol{X}}^{-}$ without any missing entries. However, for the sake of completeness, we will analyze the case where $\bar{X}_{ij}$ is observed with some arbitrary probability $\bar{p}$. Concretely, we define $\bar{p} = 1 - (1-p)^{\tau}$ since $\bar{X}_{ij}$ is unobserved only if all X_{it} for $t \in B_j$ are unobserved. From the model setup, we assume that each X_{it} is observed, independently of all other entries, with probability p; hence, the

definition of $\bar{p}$. We now proceed with our analysis in the exact same manner with the only difference being our newly defined set of variables and parameters.

In that spirit, we define the event $E_1 := \{|\hat{p} - \bar{p}| \leq \eta\bar{p}/N\}$ where $\hat{p}$ now refers to the proportion of observed entries in $\bar{\boldsymbol{X}}^-$. Let $\bar{\lambda}_i$ denote the singular values of $\bar{p}\bar{\boldsymbol{M}}^-$ in decreasing order and with repeated multiplicities, whereby $\bar{\lambda}^* = \max_{i \notin S} \bar{\lambda}_i$. With minor variations to the proofs of Lemmas B.1.1 and B.1.2, we arrive at the inequality

$$\text{MSE}(\hat{\bar{M}}_1^-) \leq \frac{C_1\|\beta\|^2}{\bar{p}^2\tau}\mathbb{E}\Big(\bar{\lambda}^* + \|\bar{\boldsymbol{Y}}^- - \bar{p}\bar{\boldsymbol{M}}^-\| + \|(\hat{p} - \bar{p})\bar{\boldsymbol{M}}^-\|\Big)^2 + C_2 e^{-c\tau\bar{p}/N} + \frac{2\bar{\sigma}^2 k}{\tau}.$$

Since $\text{rank}(\bar{\boldsymbol{M}}^-) = k$ and $|S| = k$, we have that $\bar{\lambda}^* = 0$. Thus,

$$\text{MSE}(\hat{\bar{M}}_1^-) \leq \frac{C_1\|\beta\|^2}{\bar{p}^2\tau}\mathbb{E}\Big(\|\bar{\boldsymbol{Y}}^- - \bar{p}\bar{\boldsymbol{M}}^-\| + \|(\hat{p} - \bar{p})\bar{\boldsymbol{M}}^-\|\Big)^2 + C_2 e^{-c\tau\bar{p}/N} + \frac{2\bar{\sigma}^2 k}{\tau}.$$

Similarly, we define the event $E_2 := \{\|\bar{\boldsymbol{Y}}^- - \bar{p}\bar{\boldsymbol{M}}^-\| \leq (2 + \eta/2)\delta\sqrt{\tau}\}$ where we define $\delta^2 = \bar{p}\bar{\sigma}^2 + \bar{p}(1 - \bar{p})$. After a careful massaging of the proofs, we arrive at the familiar inequality:

$$\begin{aligned}
\lim_{\tau\to\infty} \text{MSE}(\hat{\bar{M}}_1^-) &\leq \lim_{\tau\to\infty} \frac{C_1\|\beta\|^2}{\sqrt{\bar{p}}}\Big(\frac{\bar{\sigma}^2 + (1 - \bar{p})}{\sqrt{\bar{p}}} + \frac{(4\eta + \eta^2)\sqrt{\bar{\sigma}^2 + (1 - \bar{p})}}{\sqrt{N - 1}}\Big) \\
&\leq \lim_{\tau\to\infty} \frac{C_1\|\beta\|^2}{\sqrt{\bar{p}}}\Big(\frac{\sigma^2/\tau + (1 - \bar{p})}{\sqrt{\bar{p}}} + \frac{(4\eta + \eta^2)\sqrt{\sigma^2/\tau + (1 - \bar{p})}}{\sqrt{N - 1}}\Big) \\
&\leq \lim_{\tau\to\infty} \frac{C_1\|\beta\|^2}{\sqrt{\bar{p}}}\Big(\frac{\sigma^2/\tau + (1 - \bar{p})}{\sqrt{\bar{p}}} + \frac{\sqrt{\sigma^2/\tau + (1 - \bar{p})}}{\sqrt{N - 1}}\Big) \\
&= 0,
\end{aligned}$$

since $\bar{p} = 1$ and $\sigma^2/\tau = 0$ as $\tau \to \infty$. Therefore, by pre-processing the data in the proposed manner, our estimator is asymptotically consistent. ■

B.3 Post-intervention analysis (static rank)

Theorem (4.2.1). *Let* (2.7) *hold for some* $\beta \in \mathbb{R}^{N-1}$. *Let* $\text{rank}(\boldsymbol{M}^-) = \text{rank}(\boldsymbol{M})$. *Then* $M_1^+ = (\boldsymbol{M}^+)^T\beta$.

Proof. Suppose we begin with only the matrix $\boldsymbol{M}^-$, i.e. $\boldsymbol{M} = \boldsymbol{M}^-$. From the

assumption that $M_1^- = (\boldsymbol{M}^-)^T\beta$, we have for $t \le T_0$

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

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 $\boldsymbol{M}^+ = [M_{it}]_{2 \le i \le N, T_0 < t \le T}$. ■

Lemma B.3.1. *Assuming E_1 occurs, the (un-normalized) post-intervention error is bounded above by*

$$\left\| \hat{M}_1^+ - M_1^+ \right\| \le \frac{C_1\sqrt{T_0}}{p\mu} \Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \left\|(\hat{p} - p)\boldsymbol{M}^+\right\| \Big) + \left\| (\boldsymbol{M}^+)^T(\hat{\beta} - \beta) \right\|.$$

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. Observe that $\left\| (\hat{\boldsymbol{M}}^+)^\dagger \right\| = 1/(\min_{i \in S} s_i)$. However, by definition of the set S,

all singular values within S satisfy $s_i \geq \mu$, yielding $\left\|(\hat{\boldsymbol{M}}^+)^\dagger\right\| \leq 1/\mu$. Therefore,

$$\begin{aligned}
\left\|\hat{M}_1^+ - M_1^+\right\| &= \left\|(\hat{\boldsymbol{M}}^+)^T\hat{\beta} - (\boldsymbol{M}^+)^T\beta\right\| \\
&= \left\|(\hat{\boldsymbol{M}}^+)^T\hat{\beta} - (\boldsymbol{M}^+)^T\beta + (\boldsymbol{M}^+)^T\hat{\beta} - (\boldsymbol{M}^+)^T\hat{\beta}\right\| \\
&\leq \left\|(\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+)^T\hat{\beta}\right\| + \left\|(\boldsymbol{M}^+)^T(\hat{\beta} - \beta)\right\| \\
&\leq \left\|\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+\right\|\left\|\hat{\beta}\right\| + \left\|(\boldsymbol{M}^+)^T(\hat{\beta} - \beta)\right\| \\
&\leq \left\|\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+\right\|\left\|(\hat{\boldsymbol{M}}^+)^\dagger\right\|\|Y_1^-\| + \left\|(\boldsymbol{M}^+)^T(\hat{\beta} - \beta)\right\| \\
&\leq \frac{\sqrt{T_0}}{\mu}\left\|\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+\right\| + \left\|(\boldsymbol{M}^+)^T(\hat{\beta} - \beta)\right\| \\
&\overset{(a)}{\leq} \frac{C_1\sqrt{T_0}}{p\mu}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \|(\hat{p} - p)\boldsymbol{M}^+\|\Big) + \left\|(\boldsymbol{M}^+)^T(\hat{\beta} - \beta)\right\|,
\end{aligned}$$

where (a) follows from a minor adaptation of Lemma B.1.2. ■

Theorem (4.2.2). *The (unnormalized) post-intervention root mean-square error (RMSE) is bounded above by*

$$\begin{aligned}
\mathrm{RMSE}(\hat{M}_1^+) \leq{}& \frac{C_1\sqrt{T_0}}{p\mu\sqrt{T - T_0}}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \|(\hat{p} - p)\boldsymbol{M}^+\|\Big) + \frac{\|\boldsymbol{M}^+\|}{\sqrt{T - T_0}}\mathbb{E}\left\|\hat{\beta} - \beta\right\| \\
&+ \frac{C_2\sqrt{T_0(N-1)}}{\mu}e^{-cp(N-1)T}.
\end{aligned}$$

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 application of B.3.1 and the total law of probability. Specifically,

$$\begin{aligned}
\mathbb{E}\left\|\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+\right\| &\leq \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+\right\| \mid E_1\right] + \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+\right\| \mid E_1^c\right]\mathbb{P}(E_1^c) \\
&\leq \frac{C_1}{p}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \|(\hat{p} - p)\boldsymbol{M}^+\|\Big) + \mathbb{E}\left[\left\|\hat{\boldsymbol{M}}^+ - \boldsymbol{M}^+\right\|^2 \mid E_1^c\right]\mathbb{P}(E_1^c) \\
&\leq \frac{C_1}{p}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \|(\hat{p} - p)\boldsymbol{M}^+\|\Big) + C_2\sqrt{(N-1)(T - T_0)}e^{-cp(N-1)T}.
\end{aligned}$$

Merging the above result with B.3.1 gives our desired result. ■

Appendix C

Regularization

In this chapter, we will prove our results for the ridge regression setting.

C.1 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:

$$\left\|Y_1^- - (\hat{\boldsymbol{M}}^-)^T v\right\|^2 + \eta\|v\|^2 = (Y_1^-)^T Y_1^- - 2v^T \hat{\boldsymbol{M}}^- Y_1^- + v^T \hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T v + \eta v^T v. \tag{C.1}$$

Setting the gradient of (6.1) to zero, and solving for v, we obtain

$$\begin{aligned} \nabla_v \left\{ \left\|Y_1^- - (\hat{\boldsymbol{M}}^-)^T v\right\|^2 + \eta\|v\|^2 \right\}_{v=\hat{\beta}_\eta} &= -2\hat{\boldsymbol{M}}^- Y_1^- + 2\hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T v + 2\eta v = 0 \\ \Rightarrow \hat{\beta}_\eta &= \left(\hat{\boldsymbol{M}}^- (\hat{\boldsymbol{M}}^-)^T + \eta \boldsymbol{I}\right)^{-1} \hat{\boldsymbol{M}}^- Y_1^-. \end{aligned}$$

C.2 Pre-intervention analysis

Remark C.2.0.1. To ease the notational complexity of the following Lemma C.2.1 and Theorem 6.2.1 proofs, we will make use of the following notations for **only** this derivation: Let

$$\boldsymbol{Q} := (\boldsymbol{M}^-)^T \tag{C.2}$$

$$\hat{\boldsymbol{Q}} := (\hat{\boldsymbol{M}}^-)^T \tag{C.3}$$

such that

$$M_1^- := \boldsymbol{Q}\beta \tag{C.4}$$

$$\hat{M}_1^- := \hat{\boldsymbol{Q}}\hat{\beta}. \tag{C.5}$$

Lemma C.2.1. *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 $\hat{\boldsymbol{Q}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the singular value decomposition of $\hat{\boldsymbol{Q}}$. Since $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$, we have that

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

where

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

■

Theorem (6.2.1). *For any $\eta > 0$, the pre-intervention error of the algorithm can be bounded as*

$$\operatorname{MSE}(\hat{M}_1^-) \leq \frac{C_1}{p^2T_0}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \big\|(\hat{p}-p)\boldsymbol{M}^-\big\|\Big)^2\|\beta\|^2 + \frac{2\sigma^2|S|}{T_0} \tag{C.6}$$

$$+ \frac{\eta\|\beta\|^2}{T_0} + C_2(N-1)\|\beta\|^2 e^{-c(N-1)Tp}. \tag{C.7}$$

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 following proof is a slight modification for the proofs of Lemmas B.1.1 and Theorem 4.1.1. 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,

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

Taking expectations, we have

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

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 C.2.1, and as before, (b) follows because $\hat{\boldsymbol{Q}}\hat{\boldsymbol{Q}}^\dagger$ is a

projection matrix. The rest of the proof follows as in the proof of Lemma B.1.1 by employing Lemma B.1.2. ■

C.3 Post-intervention analysis (static rank)

Theorem (6.2.2). *Assuming* $\text{rank}(\boldsymbol{M}^-) = \text{rank}(\boldsymbol{M})$, *the post-intervention root mean-square error (RMSE) is bounded above by*

$$\text{RMSE}(\hat{M}_1^+) \leq \frac{C_1\sqrt{T_0}}{p\mu\sqrt{T-T_0}}\mathbb{E}\Big(\lambda^* + \|\boldsymbol{Y} - p\boldsymbol{M}\| + \big\|(\hat{p}-p)\boldsymbol{M}^+\big\|\Big) + \frac{\|\boldsymbol{M}^+\|}{\sqrt{T-T_0}}\mathbb{E}\Big\|\hat{\beta}_\eta - \beta\Big\| + \frac{C_2\sqrt{T_0(N-1)}}{\mu}e^{-cp(N-1)T}.$$

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 exactly from the proofs of Lemma B.3.1 and Theorem 4.2.2, and by observing that $\big\|\hat{\beta}_\eta\big\| \leq \big\|\hat{\beta}\big\|$ due to the complexity penalty term. ■

Appendix D

A Bayesian Perspective

D.1 Derivation of posterior parameters

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

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

where μ, $\boldsymbol{A}$, and b are parameters that govern the means, while $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$ 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

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

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

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

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

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

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

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

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

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