# Bayesian Synthetic Control with a Soft Simplex Constraint

Yihong Xu*
Department of Economics, Texas A&M University
and
Quan Zhou
Department of Statistics, Texas A&M University

**Abstract**

The implementation of the Synthetic Control Method (SCM) in high-dimensional settings and the appropriateness of imposing the simplex constraint have been long-standing questions in empirical practice. To address both issues simultaneously, we propose a Bayesian SCM that integrates a soft simplex constraint with spike-and-slab variable selection. Our hierarchical prior structure allows researchers not only to obtain credible counterfactual estimates but also to infer the attitude of data towards the validity of the simplex constraint, offering empirical guidance on whether the constraint is appropriate in a given context. The simplex constraint, however, leads to an intractable marginal likelihood. To overcome its computational challenge, we develop a novel Markov chain Monte Carlo (MCMC) algorithm that update the regression coefficients of two predictors simultaneously, enabling efficient posterior sampling. Theoretically, we establish the consistency for the control units selection and coefficients posterior mean. Simulation studies demonstrate the nice performance of our method in various settings, and two empirical examples of economic policies are revisited to show interesting insights.

*Keywords:* average treatment effect; constrained linear regression; Gibbs sampling; Metropolis-within-Gibbs sampler; spike-and-slab prior; strong selection consistency.

*Yihong is grateful to Qi Li, Yonghong An, Angda Li and participants of 2025 Texas A&M practice job talk, 2025 Texas Econometrics Camp, 2025 Southeastern Texas Chapter of American Statistical Association for their helpful comments and suggestions.

# 1 Introduction

## 1.1 Background on Synthetic Control Methods

In many fields of social science, researchers need to identify the causal effect of an intervention or treatment of interest using only observational data, since conducting case-control experiments is either infeasible or improper. This is a common situation in economics, where interventions are implemented at an aggregate level affecting one single large unit (e.g., a country) on some aggregate outcome of interest (e.g., GDP). A classical example, which was studied in the seminal work of Abadie & Gardeazabal (2003), is how terrorist conflicts between the 1970s and 1990s in the Basque Country affected the local economy, such as the per capita GDP. Estimating how the economy would have evolved in the absence of political conflicts poses several challenges. On one hand, one should not simply compare the Basque Country with its past, as it would be confounded by other factors happening at the same time of the terrorism, such as the broader economic downturn experienced by Spain. On the other hand, the Basque Country differed from other Spanish regions in several key characteristics, and thus a straightforward comparison between the Basque Country and the rest of Spain would fail to separate the impact of terrorism from other variables affecting economic growth. To estimate the terrorism impact, Abadie & Gardeazabal (2003) proposed constructing a weighted average using the other regions in Spain to reflect the relevant economic characteristics of the Basque Country, and then this synthetic counterfactual was compared to the real GDP data of the Basque Country to assess the economic effect of terrorism. This approach is called the synthetic control method (SCM). Since its introduction, SCM has gained widespread adoption in empirical research due to its interpretability, transparency, and plausible identification assumptions. Athey & Imbens (2017) described SCM as "arguably the most important innovation in the policy evaluation literature in the last 15 years."

SCM can be formulated as follows. Consider a panel with $T = T_0 + T_1$ total observations over time for $N + 1$ units, where only the first unit receives the treatment after period $T_0$. The remaining $N$ units form the control group, which are unaffected by the intervention. Let $Z_{i,t}$ denote the outcome of interest for unit $i$ at time $t$, where $i = 1, 2, ..., N + 1$ and $t = 1, 2, ..., T$. The objective of SCM is to estimate the average treatment effect on the treated (ATT) by estimating the counterfactual outcome $Z_{1,t}^{(0)}$ for the treated unit at times $t = T_0 + 1, \ldots, T$, which represents what the outcome of unit 1 would have been in the absence of the treatment. The standard approach is to define $Z_{1,t}^{(0)}$ as a weighted combination of the outcomes of the $N$ untreated units:

$$Z_{1,t}^{(0)}(w) = \sum_{i=2}^{N+1} w_{i-1} Z_{i,t}, \tag{1}$$

where the weights $(w_i)_{i=1}^{N}$ are time-independent, non-negative and sum to one. A larger $w_i$ indicates that the $i$-th control unit better approximates the treated unit's pre-intervention characteristics. To find the optimal set of weights $(w_i)_{i=1}^{N}$ such that the synthetic control approximates the pre-treatment outcome of the treated unit as closely as possible, a common strategy is to minimize the mean squared error between the treated unit and the synthetic control in the pre-treatment period $t = 1, \ldots, T_0$:

$$\hat{w} = \underset{w \in \Delta^{N-1}}{\operatorname{argmin}} \frac{1}{T_0} \sum_{t=1}^{T_0} \left( Z_{1,t} - \sum_{i=2}^{N+1} w_{i-1} Z_{i,t} \right)^2 \tag{2}$$

where

$$\Delta^{N-1} = \left\{ u \in \mathbb{R}^N \colon u_i \geq 0 \text{ for each } i, \text{ and } \textstyle\sum_{i=1}^{N} u_i = 1 \right\}$$

denotes the $(N - 1)$-simplex. Note that for ease of presentation, we have not considered time-invariant characteristics in the above formulation, and readers are referred to Abadie (2021) for a full description of SCM. The average difference between the observed outcome

$Z_{1,t}$ and the synthetic control $Z_{1,t}^{(0)}(\hat{w})$ serves as an estimate of the ATT:

$$\widehat{\text{ATT}} = \frac{1}{T_1} \sum_{t=T_0+1}^{T} \left( Z_{1,t} - \sum_{i=2}^{N+1} \hat{w}_{i-1} Z_{i,t} \right),$$

where $T_1 = T - T_0$. A large portion of the literature on SCM methodology focuses on how to move beyond the simplex constraint and extend SCM to cases where $N$ can be very large, possibly greater than $T_0$, and below we provide a brief review.

Some researchers advocated for using regression-based methods to estimate $w$ without making the simplex assumption (Doudchenko & Imbens 2016), since they argued that the simplex constraint is not likely to hold in practice and negative weights may also be desirable in some scenarios. One influential work in this vein was Hsiao et al. (2012), who assumed the cross-sectional dependence is driven by a common factor model and estimated the counterfactual via a regression model without identifying the unobserved factors. They computed the weight $w$ in (1) as an OLS-type estimator (OLS stands for Ordinary Least Squares) without the simplex constraint, and thus the synthetic control $Z_{1,t}^{(0)}$ was no longer a convex combination of $(Z_{i,t})_{i=2}^{N+1}$. By applying the synthetic control to post-period data, they obtained counterfactual and ATT estimates for the treated unit. Xu (2017) further generalized the method of Hsiao et al. (2012) by estimating the unknown factors, and the resulting method could be applied to multiple treated units or data sets with large $N$. Note that since dimension reduction is inherent in factor model, their method does not require selecting control units when $N$ is large. On the other hand, dropping the simplex constraints would increase the chance of extrapolation, and thus lead to extreme counterfactuals. As stated in King & Zeng (2006), the conclusions based on the extreme counterfactuals with the statistic models well fitting the data may depend on some implicit convenient modeling assumptions. The causal inference based on such counterfactual predictions is no longer reliable. Meanwhile, Abadie (2021) used a geometric argument to demonstrate that the simplex constraint has the desirable sparisty-inducing property; a formal proof was given in Goh & Yu (2022). The sparsity helps interpret the economic meaning of the synthetic counterfactual constructions. Another theoretical justification for the simplex constraint was derived in the recent work of Martinez & Vives-i-Bastida (2022). Aiming to combine the strengths of both approaches, Li (2020) proposed to relax the simplex constraint by only assuming the weights are non-negative (but may not sum to one). Another point of contention is about the inclusion of the intercept term. Some researchers recommend incorporating the intercept term to allow systematic difference between the treated and control unit (Doudchenko & Imbens 2016, Goh & Yu 2022).

Application of SCM to high-dimensional data has also gained increasing attention, since in fields such as economics and political science, the number of observations can often be small due to the low frequency of data collection (often on a quarterly or even yearly basis). In certain applications, the number of pre-period observations, $T_0$, can be close to or even smaller than the number of control units (e.g., states, countries). When $N > T_0$, the optimization problem (2) is often ill-defined. In particular, OLS estimators cannot be computed since they require inverting a singular empirical covariance matrix. Similarly, though the simplex constraint significantly reduces the effective dimension of the parameter space, perfect fitting (i.e., the mean squared loss in (2) equal to zero) often occurs when $N$ is much larger than $T_0$. Consequently, methods proposed by Abadie & Gardeazabal (2003) and Hsiao et al. (2012) are considered infeasible when $N \gg T_0$, and variable selection techniques were introduced into SCM which assume that only a small number of control units should be used to construct the synthetic control. Specifically, variable selection methods that have been used in the SCM literature include elastic net (Doudchenko & Imbens 2016), Lasso (Carvalho et al. 2018, Hollingsworth & Wing 2020), and forward selection (Shi & Huang 2023), among others.

Though not the focus of the present work, we note that there is a substantial body of literature on the asymptotic theory and inference for frequentist synthetic control methods. See, for example, Abadie et al. (2010) and Firpo & Possebom (2018) for the use of permutation methods (known as "placebo tests") in ATT estimation, and Carvalho et al. (2018), Li (2020), Chernozhukov et al. (2021) for the inference with SCM under high-dimensional settings.

In this work, we adopt a Bayesian approach to construct the synthetic control, a strategy that has become popular in the recent SCM literature for its flexibility in integrating with conventional SCM, regression-based methods and high-dimensional techniques. Another important advantage of the Bayesian approach is that it provides posterior distributions for both ATT and the counterfactual outcome of the treated unit, which are more informative than point estimates. By computing posterior mean estimates, Bayesian methods make statistical inference by averaging over the posterior distributions, a feature known as model averaging (Kass & Raftery 1995), which contrasts with penalized regression methods like Lasso that focus on identifying a single best model to minimize the loss function. A potential limitation, however, is that the posterior fitting often requires the use of Markov chain Monte Carlo (MCMC) sampling, which can be very time-consuming compared to optimization-based methods. One recent work by Goh & Yu (2022) considered Bayesian linear regression where the prior distribution of the regression coefficients is uniform over a convex hull, corresponding to the simplex constraint proposed by Abadie & Gardeazabal (2003). They also discussed the Bayesian formulation of the method of Li (2020), which involves using a prior over the shifted conical hull. A similar model was studied in Martinez & Vives-i-Bastida (2022), and a Bernstein-von Mises result for the Bayes estimator was proved. Kim et al. (2020) proposed using either the horseshoe prior or the spike-and-slab prior, two well-known Bayesian approaches to variable selection, to construct synthetic control with selected control units in high-dimensional settings; the simplex constraint was dropped in their method. It was demonstrated in their paper that their method outperforms those frequentist SCM methods based on variable selection. Nevertheless, to our knowledge, there is no existing Bayesian variable selection method for constructing synthetic control that also incorporates the simplex constraint.

## 1.2 Overview of this Work

We take a neutral stance regarding the use of the simplex constraint. On one hand, this constraint makes the model easier to interpret and has demonstrated success in various real data sets. On the other hand, we agree that it can be overly restrictive in many cases. Hence, we propose a balanced approach based on a novel Bayesian model called BVS-SS (Bayesian Variable Selection with Soft Simplex constraint), which allows the data to determine whether the simplex constraint is appropriate on a case-by-case basis. BVS-SS aims to leverage the strengths of both simplex-constrained SCM and unconstrained regression-based methods, which we believe serves as an effective solution to the debate surrounding the use of this constraint. There is no similar method in the frequentist SCM literature to our knowledge.

BVS-SS utilizes a standard spike-and-slab-type prior for selecting which control units should be used for constructing the synthetic control, a technique that has gained significant success and widespread popularity in the Bayesian community (George & McCulloch 1993, 1997, Brown et al. 1998). Different from existing models, BVS-SS assumes that there is an underlying sparse "mean weight vector" $\mu$ that satisfies the simplex constraint, but the actual weights $w_1, \dots, w_N$ involved in (1) may deviate from $\mu$. The key innovation of BVS-SS lies in the introduction of a variance parameter $\tau$ quantifying the amount of this deviation. By imposing a non-informative prior on $\tau$, we let BVS-SS learn its value from the data, revealing whether the data actually agrees with the simplex assumption.

To generate samples from the posterior distribution of BVS-SS, we propose an original Metropolis-within-Gibbs sampler, where $\mu$ and error variance are updated by Gibbs schemes and $\tau$ is updated by Metropolis-Hastings steps. Our Gibbs-type updating for $\mu$ is different from existing Gibbs schemes for similar problems, as those methods typically only update one coordinate at a time, which is not feasible in our context due to the simplex constraint (i.e., given $\mu_{-j}$, we have $\mu_j = 1 - \sum_{i \neq j} \mu_i$ almost surely). We overcome this difficulty by theoretically finding the joint conditional posterior distribution of two coordinates $(\mu_i, \mu_j)$ given all the other parameters, which enables us to update two coordinates simultaneously. This full conditional posterior distribution is highly complicated due to the spike-and-slab prior we use for variable selection, but fortunately sampling from this distribution can be performed straightforwardly and efficiently. The ATT estimation for SCM contains two stages. First, we compute the posterior distribution of the weights $(w_i)_{i=1}^N$ using the pre-treatment data. Next, we estimate ATT by constructing a synthetic control counterfactual estimate using the post-treatment data. Since we have samples of weights generated from the proposed MCMC sampler, we can compute both the posterior mean estimate of ATT and its credible interval without extra computational cost.

Following the introduction of our algorithm, we also establish high-dimensional consistency results under two different regimes. First, we investigate a more challenging case where the simplex constraint is satisfied and analyze the variable selection consistency of the conditional marginal posterior $p(\gamma \mid y)$, where $\gamma$ is a binary vector indicating the model (see later for a formal definition). The presence of the simplex constraint introduces a significant complication: standard Bayesian variable selection techniques lead to marginal posteriors that involve intractable integrals, preventing closed-form expressions. This poses a major obstacle to non-asymptotic analysis. To address these challenges, we first derive upper and lower bounds for the conditional posterior $p(\gamma \mid y)$, relying on some well understood results for Gaussian / Chi-squared concentration inequalities, and then compare the ratio of posterior probabilities of the true model $\gamma^*$ versus those of underfitting and overfitting models $\gamma \neq \gamma^*$, allowing us to establish the rate of posterior concentration and consistency under the simplex constraint. Next, we consider the case where the simplex constraint is not satisfied, and the true data-generating process (DGP) coefficients are substantially different from any convex combination of candidate models—this corresponds to relatively large values of $\tau$. We show that, as $\tau \to \infty$, the posterior distribution with the simplex constraint becomes asymptotically equivalent to that without the constraint. This unconstrained case has been well studied in the literature (see, for example, Yang et al. (2016)), and thus existing results on variable selection consistency can be applied directly.

It should be noted that BVS-SS is intrinsically different from other hierarchical spike-and-slab variable selection models, such as those used by Guan & Stephens (2011), Zhou & Guan (2019) for heritability estimation in biology, since the marginal likelihood for given model and variance parameters in BVS-SS does not have a closed-form expression due to the simplex constraint (i.e., the vector $\mu$ cannot be integrated out). This is why we cannot apply the more common approach in the Bayesian variable selection literature, where one directly updates an indicator vector representing the regression model by add/delete/swap proposals (see George & McCulloch (1997), Guan & Stephens (2011), Chang & Zhou (2024), among many others). While updating the regression coefficients instead of the indicator vector may seem less efficient, this offers an elegant solution to the posterior computation of BVS-SS and yields favorable results in all of our numerical experiments. Besides, the simplex constraint encoded by BVS-SS also differs fundamentally from other constraints considered in the variable selection literature, such as the grouping, hierarchical and anti-hierarchical constraints studied by Choi et al. (2010), Farcomeni (2010) and the network constraint utilized in Li & Li (2008).

Lastly, in addition to simulated data sets, we apply our method to two empirical examples in the frequentist SCM literature. We first revisit the data set considered in Carvalho et al. (2018) to study how an anti-tax evasion program would affect inflation in Brazil. The control group consists of eight large metropolitan areas. While the method of Carvalho et al. (2018) selected the whole control group, BVS-SS selects a significantly smaller number of control units to construct the synthetic control. Then we revisit the data set studied in Shi & Huang (2023) to estimate the impact of anti-corruption policy on luxury watch imports in China. The control group contains 87 commodity categories and is larger than the number of observed treatment periods, making conventional synthetic control methods inapplicable. Our ATT estimation provides strong evidence supporting that China's anti-corruption policy is effective, which is consistent with the finding of Shi & Huang (2023), but our estimated effect is slightly smaller than that of Shi & Huang (2023).

The paper proceeds as follows. Section 2 introduces the BVS-SS model and outlines how to perform ATT estimation. Section S1 details our MCMC sampler for posterior computation with BVS-SS. Section 3 gives the theoretical results for BVS-SS under two scenarios. Section 4 presents simulation studies which illustrate the advantages of BVS-SS compared to other competing methods. Section 5 discusses ATT estimation via BVS-SS for the two empirical examples. Section 6 concludes the paper with some discussion.

# 2 Bayesian Variable Selection with Soft Simplex Constraint

## 2.1 Model and Prior

For generality and ease of notation, henceforth we use $Y \in \mathbb{R}^M$ to denote a response vector of $M$ observations and $X \in \mathbb{R}^{M \times N}$ to denote an arbitrary design matrix with $N$ explanatory variables. Assume that

$$Y = Xw + \epsilon, \quad \epsilon \sim N(0, \phi^{-1} I), \tag{3}$$

where $w \in \mathbb{R}^N$ is the vector of unknown regression coefficients, $\epsilon$ is the vector of i.i.d. Gaussian noise with variance $\phi^{-1}$, and $I$ denotes the identity matrix. For the synthetic control problem described in Section 1.1, we have $M = T_0$, $Y_t = Z_{1,t}$ and $X_{tj} = Z_{j+1,t}$ for $t = 1, \dots, T_0$ and $j = 1, \dots, N$.

We propose a hierarchical prior distribution on $w$ which induces sparsity and incorporates the simplex constraint. To describe it, introduce an indicator vector $\gamma \in \{0,1\}^N$ such that $\gamma_i = 1$ if and only if $w_i \neq 0$. Let $|\gamma| = \sum_{j=1}^N \gamma_j$ denote the number of selected predictors. Given a vector $w$, we use $w_\gamma$ to denote the subvector of $w$ with entries indexed by $\{i \colon \gamma_i = 1\}$. We consider the following prior on $(w, \phi)$,

$$\begin{aligned} \phi &\sim \text{Gamma}(\kappa_1/2, \kappa_2/2), \\ \gamma_i &\overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta), \\ \tau &\sim \text{Gamma}(a_1, a_2), \\ \mu_\gamma \mid \gamma &\sim \text{sym-Dirichlet}(\alpha), \\ \mu_i \mid \gamma_i = 0 &\sim \delta_0, \\ w_\gamma \mid \gamma, \mu_\gamma, \tau, \phi &\sim N(\mu_\gamma, \, (\tau/\phi) I), \\ w_i \mid \gamma_i = 0 &\sim \delta_0, \end{aligned} \tag{4}$$

where $\kappa_1, \kappa_2, a_1, a_2, \alpha > 0$ and $\theta \in (0,1)$ are fixed hyperparameters, $\delta_0$ denotes the Dirac measure with unit mass on 0, sym-Dirichlet($\alpha$) denotes the symmetric Dirichlet distribution on the simplex with concentration parameter $\alpha$ (the dimension of the simplex should be clear from context), and the Gamma distribution is in shape-rate parameterization. In particular, given $\alpha = 1$ and $|\gamma| = \ell$, the prior of $\mu_\gamma$ is a uniform distribution on $\Delta^{|\gamma|-1}$ with density $p(\mu_\gamma) = \Gamma(\ell)$ for each $\mu_\gamma \in \Delta^{|\gamma|-1}$, where $\Gamma(\ell) = (\ell - 1)!$ is the inverse volume of the simplex.

In plain words, we assume that given $\gamma$ and $\phi$, $w_\gamma$ follows a normal distribution with mean vector $\mu_\gamma \in \Delta^{|\gamma|-1}$ and prior variance depending on the parameter $\tau$. This conditional prior for $w$ can be viewed as an interpolation between the simplex constraint and the unconstrained setting. If the data suggests that $w$ is likely to satisfy the simplex constraint, then the posterior distribution of $\tau$ should concentrate around zero, while if the simplex constraint is significantly violated, $\tau$ should stay away from zero in the posterior. As will be demonstrated in our simulation study, $\tau$ can be reasonably estimated even with a relatively small sample size, which tells the user if the simplex constraint is appropriate for the data set being analyzed. This approach effectively reconciles the different views on the use of the simplex constraint, and thus we call our model BVS-SS (Bayesian Variable Selection with Soft Simplex constraint). We note that our model is similar to the one studied in Martinez & Vives-i-Bastida (2022); however, they did not consider variable selection, and in their implementation they assumed that $w$ satisfies the hard simplex constraint (though the soft constraint was used in their theory). Regarding the hyperparameters, $a_1, a_2$ should be chosen to reflect one's prior belief on whether simplex constraint is likely to hold, and a larger value of $\alpha$ encourages a more even distribution of the weights $\{\mu_i \colon \gamma_i = 1\}$. In practice, one often has limited information about the true weight vector $w$, but the effect of these hyperparameters quickly becomes negligible as sample size increases.

The other elements of our model are standard. The prior for $\phi$ is conjugate, and one can choose small values for $\kappa_1, \kappa_2$ as a noninformative prior. The hyperparameter $\theta$ is the marginal prior probability of an explanatory variable being included in the regression model, and a smaller value of $\theta$ represents a heavier penalty on the model complexity.

## 2.2 Posterior Distribution and ATT Estimation

Let $X_\gamma$ denote the submatrix of $X$ with columns indexed by $\{i \colon \gamma_i = 1\}$. Under the prior distribution of BVS-SS, we have

$$Y \mid \gamma, \mu_\gamma, \tau, \phi \sim N\left(X_\gamma \mu_\gamma,\, \phi^{-1}(\tau X_\gamma X_\gamma^\top + I)\right). \tag{5}$$

Since $\gamma$ can be viewed as a function of $\mu$ with $\gamma_i = \mathbb{1}_{\{\mu_i \neq 0\}}$, conditioning on $(\gamma, \mu_\gamma, \tau, \phi)$ is equivalent to conditioning on $(\mu, \tau, \phi)$. So henceforth we will denote the conditional distribution in (5) by $p(y \mid \mu, \tau, \phi)$, and $\gamma$ should be understood as $\gamma = \gamma(\mu)$. A routine calculation using Woodbury identity yields the marginal likelihood of $(\mu, \tau, \phi)$:

$$p(y \mid \mu, \tau, \phi) \propto \phi^{M/2} \tau^{-|\gamma|/2} \det(V_{\gamma,\tau})^{-1/2} \exp\left\{-\frac{\phi}{2}(y - X_\gamma \mu_\gamma)^\top \Sigma_{\gamma,\tau} (y - X_\gamma \mu_\gamma)\right\}, \tag{6}$$

where

$$V_{\gamma,\tau} = X_\gamma^\top X_\gamma + \tau^{-1} I, \quad \Sigma_{\gamma,\tau} = I - X_\gamma V_{\gamma,\tau}^{-1} X_\gamma^\top. \tag{7}$$

Further, using the conjugacy of normal-gamma prior, we find that the conditional posterior distributions of $w_\gamma$ and $\phi$ are given by

$$\phi \mid y, \mu, \tau \sim \text{Gamma}\left(\frac{M + \kappa_1}{2},\, \frac{\kappa_2 + (y - X_\gamma \mu_\gamma)^\top \Sigma_{\gamma,\tau} (y - X_\gamma \mu_\gamma)}{2}\right), \tag{8}$$

$$w_\gamma \mid y, \mu, \tau, \phi \sim N\left(V_{\gamma,\tau}^{-1}(X_\gamma^\top y + \tau^{-1} \mu_\gamma),\, \phi^{-1} V_{\gamma,\tau}^{-1}\right). \tag{9}$$

The joint posterior distribution of $(\mu, \tau, \phi)$ can be computed by

$$p(\mu, \tau, \phi \mid y) \propto p(y \mid \mu, \tau, \phi) p(\mu_\gamma \mid \gamma) p(\gamma) p(\tau) p(\phi)$$

where we have used the prior independence between $\mu, \tau, \phi$. In Section S1 we will propose an efficient MCMC sampling algorithm targeting $p(\mu, \tau, \phi \mid y)$.

Prediction with BVS-SS can be conducted as follows. Let $\tilde{X} \in \mathbb{R}^{\tilde{M} \times N}$ denote the design matrix of another $\tilde{M}$ observations, and suppose we are interested in estimating $\tilde{X}w$. Letting $\mathbb{E}_y$ denote the posterior distribution given $Y = y$ and using (9), we obtain that

$$\mathbb{E}_y[\tilde{X}w] = \mathbb{E}_y\left[\mathbb{E}_y(\tilde{X}w \mid \mu, \tau, \phi)\right] = \mathbb{E}_y\left[\tilde{X}V_{\gamma,\tau}^{-1}(X_\gamma^\top y + \tau^{-1}\mu_\gamma)\right]. \tag{10}$$

Given a sequence of $n$ MCMC samples $(\mu^{(k)}, \tau^{(k)}, \phi^{(k)})_{k=1}^n$, we can approximate the above expectation using the sample average, which yields the posterior predictive mean of $\tilde{X}w$. The covariance matrix of the posterior predictive distribution of $\tilde{X}w$ can be calculated similarly.

For applications like synthetic control, the goal is to estimate ATT among the $\tilde{M}$ observations. Denote their response under the treatment by $\tilde{Y}^{(1)}$, which is assumed observed, and denote the counterfactual response without treatment by $\tilde{Y}^{(0)}$. Let $\delta = \tilde{Y}^{(1)} - \tilde{Y}^{(0)}$, and define ATT by

$$\text{ATT} = \frac{1}{\tilde{M}}\sum_{i=1}^{\tilde{M}} \delta_i = \frac{1}{\tilde{M}}\sum_{i=1}^{\tilde{M}} \left(\tilde{Y}_i^{(1)} - \tilde{Y}_i^{(0)}\right), \tag{11}$$

which is the main parameter of interest. By replacing $\tilde{Y}^{(0)}$ using the synthetic control $\mathbb{E}_y[\tilde{X}w]$, we obtain an estimator for $\bar{\delta}$. If we have samples $(w^{(k)})_{k=1}^n$ approximating the marginal posterior distribution of $w$, we can also obtain the posterior predictive distribution of ATT by plugging $w^{(k)}$ into the estimator $\tilde{M}^{-1}\sum_{i=1}^{\tilde{M}}(\tilde{Y}_i^{(1)} - w^\top \tilde{X}_{(i)})$, where $\tilde{X}_{(i)}$ denotes the $i$-th row of $\tilde{X}$ (treated as a column vector).

### 2.3 A Metropolis-within-Gibbs Sampler

Though the Metropolis–Hastings update for $\tau$ and conjugate posterior for $\phi$ is standard, a unique challenge to generate a sequence of posterior samples $(\mu^{(k)}, \tau^{(k)}, \phi^{(k)})_{k=1}^{n}$ is that, a standard Gibbs-type updating for $\mu$ is not feasible in our context due to the simplex constraint (i.e., given $\mu_{-j}$, we have $\mu_j = 1 - \sum_{i \neq j} \mu_i$ almost surely). Therefore, we propose a novel Metropolis-within-Gibbs Sampler, with the main idea that we update two coordinates of $\mu$ simultaneously. We calculate $p(\gamma_i, \gamma_j \mid y, \mu_{-(i,j)}, \tau, \phi)$ and $p(\mu_i, \mu_j \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i, \gamma_j)$. That is, the conditional probability of selection for entries $i$ and $j$, and the conditional probability density of $\mu_i$, and $\mu_j$ given their selection status.

Fortunately, the computation of both conditional posteriors remains tractable, as we can proceed it by case by case. For $p(\mu_i, \mu_j \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i, \gamma_j)$, when $\sum_{k \neq i,j} \mu_k = 1$, it is evident that neither $i$ nor $j$ can be selected; hence, $\mu_i$ and $\mu_j$ must both be zero; when $\sum_{k \neq i,j} \mu_k < 1$, only three scenarios are possible: either $i$ is selected but not $j$, $j$ is selected but not $i$, or both are selected. In the cases where $\gamma_i = 1$ and $\gamma_j = 0$, or vice versa, the selected entry must equal $1 - \sum_{k \neq i,j} \mu_k$. It only remains $p(\mu_i, \mu_j \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i = \gamma_j = 1)$. Again, notice that the posterior is degenerate since $\mu_i + \mu_j = 1 - \sum_{k \neq i,j}$ almost surely. And for $p(\gamma_i, \gamma_j \mid y, \mu_{-(i,j)}, \tau, \phi)$, since $\gamma_j$ and $\gamma_j$ are binary, we can normalize the probability once we have a closed form expression for the posterior up to some constant.

Eventually, we derive the expression of $p(\gamma_i, \gamma_j \mid y, \mu_{-(i,j)}, \tau, \phi)$ and $p(\mu_i \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i = \gamma_j = 1)$. When $\alpha = 1$, $\mu_i \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i = \gamma_j = 1$ follows a truncated normal distribution (Lemma S1), and $p(\gamma_i, \gamma_j \mid y, \mu_{-(i,j)}, \tau, \phi)$ can always be expressed by using the CDF of the standard normal distribution, which leads to an efficient calculation. Details and formal sampling algorithm can be found in Appendix S1.

# 3 Theoretical Results for BVS-SS

In this section, we study the theoretical properties of our BVS-SS model specified by equations (3) and (4). For ease of analysis, we fix the hyperparameter $\tau$ and consider two complementary scenarios: $\tau = 0$ and $\tau \to \infty$. Following the convention in the variable selection literature (Yang et al. 2016), we interpret a model $\gamma$ as both a binary vector and a subset of $[N] := \{1, 2, \ldots, N\}$; that is, to indicate the $i$-th covariate is selected, we can write either $\gamma_i = 1$ or $i \in \gamma$. We use $| \cdot |$ to denote set cardinality, and thus $|\gamma|$ denotes the model size.

## 3.1 High-dimensional Selection Consistency with $\tau = 0$

For the first scenario, we fix the prior parameter $\tau = 0$ and assume that the true DGP satisfies the simplex constraint. Our goal is to establish, in high-dimensional regimes, the "strong selection consistency" property of BVS-SS which means that the posterior probability of the true model, denoted by $\gamma^*$, converges to one in probability under the true DGP; here, "strong" means that this property is stronger than other consistency criteria such as pairwise selection consistency (Casella et al. 2009). Since $\tau = 0$ implies $w = \mu$ almost surely, we can rewrite the model as

$$Y = X\mu + \epsilon, \quad \epsilon \sim N(0, \phi^{-1} I).$$

For technical convenience, we slightly modify the prior specification as follows:

$$p(\gamma) \propto \Gamma(|\gamma|)^{-1} \left\{ 1^\top (X_\gamma^\top X_\gamma)^{-1} 1 \right\}^{1/2} \det(X_\gamma^\top X_\gamma)^{1/2} \theta^{|\gamma|} (1 - \theta)^{N - |\gamma|}, \tag{12}$$

$$p(\phi \mid \gamma) \propto \phi^{(\kappa_1 + |\gamma| - 1)/2 - 1} e^{-\phi \kappa_2 / 2}, \tag{13}$$

$$\mu_\gamma \mid \gamma \sim \mathrm{Dir}(1),$$

where $\mathrm{Dir}(\mathbf{1})$ denotes the Dirichlet distribution with parameter vector $\mathbf{1} = (1, \ldots, 1)$, which is just the uniform distribution over the simplex $\Delta^{|\gamma|-1}$. Compared to the original prior specification given by (4), the changes are minor. First, rather than letting $\gamma_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(\theta)$, we reweight the prior probability of each $\gamma$ by three factors, where $\Gamma(|\gamma|)^{-1}$ is the volume of the simplex $\Delta^{|\gamma|-1}$, and $\mathbf{1}^\top(X_\gamma^\top X_\gamma)^{-1}\mathbf{1}$ is the variance of $\mathbf{1}^\top \hat{\mu}_\gamma$ with $\hat{\mu}_\gamma$ denoting the OLS estimator for model $\gamma$. This adjustment serves only to simplify the presentation, and we prove in the Lemma 1 below that the effect of this reweighting is typically negligible when $\theta$ decays much faster than $\sqrt{M}$. Second, we let the prior of $\phi$ depend on $\gamma$ so that for a larger model, the prior mean of $\phi$ is larger (i.e., the error variance is smaller). This modification enables a simpler comparison of the marginal likelihoods of different models after integrating out $\phi$. Note that when $\kappa_1$ is large (e.g. of order $M$), the difference between the two prior specifications for $\phi$ becomes negligible. Third, for the conditional prior for $\mu_\gamma$ given $\gamma$, we set $\alpha = 1$ in (4) to ensure that the prior density $p(\mu_\gamma \mid \gamma)$ only depends on $|\gamma|$.

**Lemma 1.** *Let $\gamma' = \gamma \cup \{j\}$ for some $j \notin \gamma$, and $p(\gamma)$ be given by (12). Then,*

$$\frac{\sqrt{M}\underline{\lambda}}{L} \leq \frac{p(\gamma')/p(\gamma)}{\theta/(1-\theta)} \leq \sqrt{\frac{LM}{\underline{\lambda}}}.$$

*Proof.* See Section S3.1. □

By integrating out $\mu_\gamma$, we can express the posterior density of $(\gamma, \phi)$ as

$$\begin{aligned} p(\gamma, \phi \mid y) &\propto p(\gamma) p(\phi \mid \gamma) \mathbb{E}_{\mu_\gamma \sim \mathrm{Dir}(\mathbf{1})} \left[ p(y \mid \gamma, \mu_\gamma, \phi) \right], \\ \text{where } p(y \mid \gamma, \mu_\gamma, \phi) &\propto \phi^{M/2} \exp\left\{ -\frac{\phi}{2} \|y - X_\gamma \mu_\gamma\|_2^2 \right\}. \end{aligned} \tag{14}$$

The marginal posterior probability of $\gamma$ can be computed as

$$p(\gamma \mid y) \propto p(\gamma) \int p(\phi \mid \gamma) p(y \mid \gamma, \phi) \mathrm{d}\phi.$$

Unlike the unconstrained Bayesian linear regression, $p(\gamma \mid y)$ is not available in closed form due to the intractable integral involved in $p(y \mid \gamma, \phi)$. We assume that $p(\gamma \mid y)$ is defined on the sparse model space $\mathbb{S}_L = \{\gamma \subset [N] \colon 1 \leq |\gamma| \leq L\}$, where $L \geq 1$ denotes the maximum model size we consider.

To establish the high-dimensional consistency result, we make the following assumptions on the design matrix, prior parameters and true DGP, where all parameters involved may vary with the sample size $M$, except the universal constants denoted by $c_\mu, c_\theta, c_M$ and $c_j$ for integer $j$. In particular, the number of variables $N$, the maximum model size $L$, and the true model size $\ell^* = |\gamma^*|$ are all allowed to go to infinity with $M$.

(A1) The design matrix $X$ satisfies that $\|X_j\|_2^2 = M$ for each $j \in [N]$, and there exists a constant $\underline{\lambda} \in (0, 1]$ such that $\lambda_{\min}(X_\gamma^\top X_\gamma) \geq M\underline{\lambda}$ for all $\gamma \in \mathbb{S}_L$, where $\lambda_{\min}$ denotes the smallest eigenvalue.

(A2) The prior distribution on $\gamma$ is given by (12) where $\theta$ satisfies $\theta/(1-\theta) = N^{-c_\theta L}$ for some universal constant $c_\theta > 0$.

(A3) The prior distribution on $\phi$ is given by (13) where hyperparameters $\kappa_1, \kappa_2$ satisfy $0 < \kappa_1 \leq M$, and $0 \leq \kappa_2 \leq \sigma^2 M/2$.

(A4) The true DGP is given by $Y \mid X \sim \mathcal{N}(X_{\gamma^*}\mu^*_{\gamma^*},\, \sigma^2 I_M)$ where $\sigma > 0$, $\gamma^*$ and $\mu^*_{\gamma^*}$ satisfy the following conditions: $\ell^* := |\gamma^*| \leq L \wedge \sqrt{L \log N}$, $\mu^*_{\gamma^*} \in \Delta^{\ell^* - 1}$, and

$$\min_{j \in \gamma^*} |\mu^*_j| \geq \frac{c_\mu \sigma \sqrt{L \log N}}{\underline{\lambda}\sqrt{M}}, \tag{15}$$

for some universal constant $c_\mu > 0$.

(A5) $L \geq 3$ and $M \geq c_M \ell^* L \log N$ for some universal constant $c_M > 0$.

Assumption (A1) requires that each column of $X$ has the same magnitude and that the minimum eigenvalue of the Gram matrix is well controlled for any model of size at most $L$. It is a mild condition, often known as "restricted eigenvalue," commonly used in the literature (Shang & Clayton 2011, Narisetty & He 2014, Yang et al. 2016). Assumption (A2) provides guidance for choosing the prior: the prior inclusion probability $\theta$ should be small enough in order to penalize large models and avoid overfitting. Similarly, Assumption (A3) requires that the choices of $\kappa_1, \kappa_2$ are not too extreme. Assumption (A4) assumes that the true DGP satisfies the simplex constraint, and (15) imposes a lower bound on the smallest nonzero coefficient of the true DGP. This is also a standard assumption and often known as the "$\beta$-min condition." Intuitively, this assumption is needed since coefficients close to zero cannot be consistently identified through variable selection. Finally, Assumption (A5) specifies the sample size needed for consistent variable selection. We will use $\mathbb{P}^*$ to denote the probability measure associated with the true DGP specified in Assumption (A4) and $\mathbb{E}^*$ to denote the corresponding expectation.

Under the above assumptions, we will prove that $p(\gamma^* \mid y)$ converges to one in probability with respect to $\mathbb{P}^*$ given sufficiently large $c_\mu, c_\theta$ and $c_M$. A key step in our proof is to establish the following posterior probability ratio bound.

**Theorem 1.** *Suppose Assumptions (A1) to (A5) hold with $c_M \geq 4$ and $c_\mu \geq 6$. With probability at least $1 - c_1 N^{-c_2}$ for some universal constants $c_1, c_2 > 0$, the following bound holds for all $\gamma \in \mathbb{S}_L$:*

$$\frac{p(\gamma \mid y)}{p(\gamma^* \mid y)} \leq 3 \left( \frac{\theta\sqrt{2\pi}}{1-\theta} \right)^{|\gamma| - \ell^*} \left( \frac{\kappa_2 + \|y - X_{\gamma^*}\hat{\mu}_{\gamma^*}\|_2^2 + \sigma^2 \ell^* \log N}{\kappa_2 + \|y - X_\gamma \hat{\mu}_\gamma\|_2^2} \right)^{(\kappa_1 + M)/2}, \qquad (16)$$

*where* $\hat{\mu}_\gamma = (X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top y$.

*Proof sketch.* The complete proof of Theorem 1 is presented in Appendix S3.2. Here we provide a sketch of the proof to highlight the key challenges and proof techniques. As shown in (14), the marginal likelihood given $(\gamma, \phi)$ can be expressed as $p(y \mid \gamma, \phi) = \mathbb{E}_{\mu_\gamma \sim \mathrm{Dir}(1)} [p(y \mid \gamma, \mu_\gamma, \phi)]$, where the expectation is an integral over the simplex. This creates a new technical difficulty, since, unlike in the unconstrained Bayesian variable selection models (Guan & Stephens 2011, Yang et al. 2016), $p(y \mid \gamma, \phi)$ is not available in closed form. To address this difficulty, we first show that, for any $\mu_\gamma \in \Delta^{|\gamma|-1}$, the likelihood $p(y \mid \gamma, \mu_\gamma, \phi)$ can be expressed as

$$p(y \mid \gamma, \mu_\gamma, \phi) \propto \phi^{M/2} \exp\left\{ -\frac{\phi}{2} \left( \|y - X_\gamma \hat{\mu}_\gamma\|_2^2 + b_\gamma + \|X_\gamma(\mu_\gamma - \check{\mu}_\gamma)\|_2^2 \right) \right\}, \tag{17}$$

where

$$\check{\mu}_\gamma = \hat{\mu}_\gamma + \frac{1 - \mathbf{1}^\top \hat{\mu}_\gamma}{\mathbf{1}^\top (X_\gamma^\top X_\gamma)^{-1} \mathbf{1}} (X_\gamma^\top X_\gamma)^{-1} \mathbf{1}, \quad b_\gamma = \frac{(1 - \mathbf{1}^\top \hat{\mu}_\gamma)^2}{\mathbf{1}^\top (X_\gamma^\top X_\gamma)^{-1} \mathbf{1}}.$$

That is, we decompose the residual sums of squares $\|y - X_\gamma \mu_\gamma\|_2^2$ into three parts. The first component, $\|y - X_\gamma \hat{\mu}_\gamma\|_2^2$, is the residual sum of squares for the least-squares estimator in the unconstrained setting. The second component, $b_\gamma \geq 0$, measures the deviation of $\hat{\mu}_\gamma$ from the simplex constraint. When $\hat{\mu}_\gamma \in \Delta^{|\gamma|-1}$, we have $b_\gamma = 0$. Note that the first two components do not depend on $\mu_\gamma$. The last component, $\|X_\gamma(\mu_\gamma - \check{\mu}_\gamma)\|_2^2$, measures the deviation of $\mu_\gamma$ from $\check{\mu}_\gamma$, which is the least-squares estimator for model $\gamma$ under the linear constraint $\mathbf{1}^\top \check{\mu}_\gamma = 1$. It has the following interpretation: $X_\gamma \check{\mu}_\gamma$ is the projection of $X_\gamma \hat{\mu}_\gamma$ onto the affine space $\{X_\gamma u \colon \mathbf{1}^\top u = 1\}$. See Appendix S2 for a brief review on the constrained least-squares estimation.

In Proposition S1, we use (17) and our modified prior distributions given in (12) and (13) to show that

$$p(\gamma, \phi \mid y) \propto f(|\gamma|)\, \phi^{(M+\kappa_1)/2}\, e^{-\frac{\phi}{2}\left(\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 + b_\gamma + \kappa_2\right)}\, \mathbb{P}\left(\tilde{U}_{\gamma,\phi} \in \tilde{\Delta}^{|\gamma|-1}\right),$$

where $f(|\gamma|)$ is a function only depending on $|\gamma|$ and prior parameter $\theta$, $\tilde{U}_{\gamma,\phi}$ is a normal random vector whose mean equals the subvector of $\check{\mu}_\gamma$ consisting of its first $|\gamma| - 1$ entries, and $\tilde{\Delta}^\ell = \{\tilde{u} \in \mathbb{R}^\ell \colon \tilde{u}_i \geq 0 \text{ for each } i, \text{ and } \sum_{i=1}^\ell \tilde{u}_i \leq 1\}$. For any $\gamma \in \mathbb{S}_L$, we have $\mathbb{P}\left(\tilde{U}_{\gamma,\phi} \in \tilde{\Delta}^{|\gamma|-1}\right) \leq 1$ and $b_\gamma \geq 0$. Hence, by integrating over $\phi$, we get an upper bound on $p(\gamma \mid y)$ that does not require any assumption on the true DGP or the design matrix; see Proposition S2.

To find a lower bound on $p(\gamma^*, \phi \mid y)$, we need to find upper bounds on $\|y - X_{\gamma^*} \hat{\mu}_{\gamma^*}\|_2^2$ and $b_{\gamma^*}$ and a lower bound on $\mathbb{P}(\tilde{U}_{\gamma^*,\phi} \in \tilde{\Delta}^{\ell^*-1})$. To this end, we construct some high-probability events in Lemma S5 to ensure that the error vector $\epsilon$ is well-behaved. This argument is standard in the high-dimensional literature and ensures that $\|y - X_{\gamma^*} \hat{\mu}_{\gamma^*}\|_2^2 = O_p(\sigma^2 M)$. For $b_{\gamma^*}$, since it measures how well $\hat{\mu}_{\gamma^*}$ satisfies the simplex constraint, we expect that $b_{\gamma^*}$ cannot be too large under our assumption on the true DGP. In Lemma S6, we prove that $b_{\gamma^*} = O_p(\sigma^2 \ell^* \log N)$. Deriving a lower bound on $\mathbb{P}(\tilde{U}_{\gamma^*,\phi} \in \tilde{\Delta}^{\ell^*-1})$ is considerably more involved. We invoke the "beta-min" condition given in Assumption (A4), which ensures that the true regression coefficients cannot be too small, to show that $\check{\mu}_{\gamma^*}$ lies in the simplex $\Delta^{\ell^*-1}$. Then, as shown in Lemma (S8), we can apply standard Gaussian concentration inequalities to get

$$1 - \mathbb{P}\left(\tilde{U}_{\gamma^*,\phi} \in \tilde{\Delta}^{\ell^*-1}\right) \leq \exp\left(-\frac{4\phi\, \sigma^2 L \log N}{(\ell^*)^2} + \frac{1}{2}\right).$$

The form of this bound enables us to integrate over $\phi$, and by using Assumptions (A3) and (A5), we obtain a lower bound on $p(\gamma^* \mid y)$; see Proposition S3. Theorem 1 then follows by combining Propositions S2 and S3. □

By Theorem 1, we can bound $p(\gamma \mid y)/p(\gamma^* \mid y)$ by analyzing the behavior of unconstrained least-squares estimators, which is well understood in the variable selection literature. This analysis then yields the strong selection consistency.

**Theorem 2.** *Suppose Assumptions (A1) to (A5) hold for sufficiently large $c_\theta, c_\mu, c_M$. Then, $\mathbb{P}^*\{p(\gamma^* \mid y) \geq 1 - c_3 N^{-c_4 L}\} \geq 1 - c_1 N^{-c_2}$ for some universal constants $c_1, c_2, c_3, c_4 > 0$.*

*Proof.* In Appendix S3.3, we prove that under our assumptions, with high probability,

$$\sup_{\gamma \in \mathbb{S}_L \setminus \{\gamma^*\}} \frac{p(\gamma \mid y)}{p(\gamma^* \mid y)} \leq 3N^{-(|\gamma \setminus \gamma^*| + |\gamma^* \setminus \gamma|)L}.$$

See Proposition S4 for the overfitted case (i.e., $\gamma^* \subset \gamma$) and Proposition S5 for the underfitted case (i.e., $\gamma^* \not\subset \gamma$). A routine calculation using $|\{\gamma \in \mathbb{S}_L : |\gamma \setminus \gamma^*| + |\gamma^* \setminus \gamma| = k\}| \leq N^k$ yields the result. □

The strong selection consistency ensures that the true model $\gamma^*$ can be identified with high probability. In addition, we can also prove the consistency for both estimating the regression coefficient vector $\mu$ and making prediction. The posterior expected $L^2$ loss for estimating $\mu$ can be expressed as

$$\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y\right] = \sum_{\gamma \in \mathbb{S}_L} p(\gamma \mid y) \left( \|\mu^*_{\gamma^c}\|_2^2 + \int \|\mu_\gamma - \mu^*_\gamma\|_2^2 \, p(\mu_\gamma, \phi \mid y, \gamma) \mathrm{d}\mu_\gamma \mathrm{d}\phi \right).$$

Similarly, given a new design matrix $\tilde{X}$, we can define the posterior expected predictive loss by $\mathbb{E}[\|\tilde{X}\mu - \tilde{X}\mu^*\|_2^2 \mid y]$. In Theorem 3, we characterize the rate at which the posterior expected loss goes to zero.

**Theorem 3.** *Under the setting of Theorem 2,*

$$\mathbb{E}\left[\|\mu-\mu^*\|_2^2 \mid y\right] = O_p\left(\max\left\{\frac{\sigma^2 L\log N}{\underline{\lambda}^2 M},\ N^{-L/(\ell^*)^2}\right\}\right).$$

*Let* $\tilde{X}\in\mathbb{R}^{\tilde{M}\times N}$ *be such that* $\max_{\gamma\in\mathbb{S}_L}\lambda_{\max}(\tilde{X}_\gamma^\top\tilde{X}_\gamma)\leq\tilde{M}\overline{\lambda}$ *for some* $\overline{\lambda}>0$*. Then,*

$$\mathbb{E}\left[\|\tilde{X}\mu-\tilde{X}\mu^*\|_2^2 \mid y\right] = O_p\left(\tilde{M}\overline{\lambda}\max\left\{\frac{\sigma^2 L\log N}{\underline{\lambda}^2 M},\ N^{-L/(\ell^*)^2}\right\}\right).$$

*Proof.* See Appendix S3.4. □

The consistency of our ATT estimator follows from the the prediction loss bound in Theorem 3. We expect that this result can be extended to more general scenarios, including cases where $[X^\top\ \tilde{X}^\top]$ is a stationary process (Li 2020).

### 3.2 Limiting Behavior when $\tau\to\infty$

Next, we investigate the performance of BVS-SS when $\tau\to\infty$. In this case, the prior places increasing mass on values of $w$ that do not satisfy the simplex constraint. We compare the behavior of the posterior distribution $p(\gamma\mid y)$ with the conventional unconstrained version, which we denote by $\tilde{p}(\gamma\mid y)$ (see Appendix S3.5). The following theorem shows that the two posterior distributions become essentially the same as $\tau\to\infty$.

**Theorem 4.** *Let* $\tilde{p}(\cdot\mid y,\tau)$ *be the posterior distribution given* $\tau$ *under the unconstrained spike-and-slab prior detailed in Section S3.5. Then*

$$\lim_{\tau\to\infty}\frac{p(\gamma\mid y,\tau)}{\tilde{p}(\gamma\mid y,\tau)}=1.$$

*Proof.* See Section S3.5 in the Appendices. □

In Theorem 4, we treat the data $(X, y)$ as fixed and let $\tau \to \infty$, but we expect that our argument can be extended to a fully non-asymptotic analysis. Such an extension may be combined with existing variable selection consistency results to show that our model also achieves selection consistency when the true DGP significantly violates the simplex constraint but $\tau$ is sufficiently large.

# 4 Simulation Studies

## 4.1 Simulation Settings

We simulate the observed training data $X \in \mathbb{R}^{M \times N}$ and $Y \in \mathbb{R}^M$ by

$$Y = Xw^* + \epsilon, \text{ where } \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1/\phi^*),$$

and $w^* \in \mathbb{R}^N$ is the true regression coefficient vector such that $w_j^* = 0$ for $j = J+1, \ldots, N$ (that is, only the first $J$ predictors have nonzero effects on $Y$). Similarly, we simulate the data under treatment $\tilde{X} \in \mathbb{R}^{\tilde{M} \times N}$ and $\tilde{Y}^{(1)} \in \mathbb{R}^{\tilde{M}}$ by

$$\tilde{Y}^{(1)} = \tilde{X}w^* + \delta + \tilde{\epsilon}, \text{ where } \delta_i \overset{\text{i.i.d.}}{\sim} N(\delta^*, 1/\nu^*),\ \tilde{\epsilon}_i \overset{\text{i.i.d.}}{\sim} N(0, 1/\phi^*).$$

The vector $\delta \in \mathbb{R}^{\tilde{M}}$ represents the treatment effects, and $\delta^* \in \mathbb{R}$ denotes the true ATT. In our simulation, we use $M = \tilde{M} \in \{25, 50, 100, 200\}$ and $N \in \{20, 50\}$. When $N = 20$, we set $J = 5$, and when $N = 50$, we set $J = 10$. Given $J$, we define a vector $\mu^* \in \Delta^{N-1}$ by

$$\mu_j^* = \begin{cases} j/S_J & \text{if } 1 \leq j \leq J, \\ 0, & \text{if } J+1 \leq j \leq N, \end{cases} \qquad \text{where } S_J = \frac{1}{2}J(J+1).$$

Then, we define $w^*$ by $w^* = \lambda\mu^*$ where $\lambda > 0$ is a scaling factor. When $\lambda = 1$, the true data-generating model also satisfies the simplex constraint. Note that, assuming $\lambda > 0$, we have $\lambda = \|w^*\|_1$. We use $\lambda \in \{1, 2, 3\}$. Given $w^*$, we set the variance parameters by $v^* = \phi^* = 4/\|w^*\|_2^2$. For each choice of $(M, N, \lambda)$, we generate 100 replicates where entries of $X, \tilde{X}$ are also sampled independently from the standard normal distribution[1]. We fix true ATT $\delta^* = 0.5$ throughout our simulation studies.

We set the hyperparameters of BVS-SS by $\kappa_1 = \kappa_2 = 1, a_1 = 0.01, a_2 = 0.1, \alpha = 1, \theta = 0.2$. The choice of $(a_1, a_2)$ guarantees that $\tau$ has prior mean 0.1 and prior variance 1. Further, for numerical stability, we truncate the prior distribution of $\tau$ by assuming that $\tau \geq 10^{-6}$. For every simulated data set, we initialize Algorithm 1 at $\tau^{(0)} = \phi^{(0)} = 1$ and $\mu^{(0)}$ sampled from its prior distribution, and we run Algorithm 1 for $1,000$ iterations and discard the first 500 iterations as burn-in.

For comparison, we have also implemented the following methods for ATT estimation. First, we use the `R` package `glmnet` to perform variable selection and ATT estimation via Lasso; the parameter of Lasso is chosen by cross-validation. Second, we compute the OLS estimator for $w^*$ using the entire data set $(X, Y)$ without variable selection, denoted by $\hat{w}_{\mathrm{LS}}$. Then ATT is estimated by

$$\widehat{\mathrm{ATT}}_{\mathrm{LS}} = \frac{1}{\tilde{M}} \sum\nolimits_{i=1}^{\tilde{M}} \left( \tilde{Y}_i^{(1)} - \hat{w}_{\mathrm{LS}}^\top \tilde{X}_{(i)} \right), \tag{18}$$

where $\tilde{X}_{(i)}$ is the $i$-th row of $\tilde{X}$ treated as a column vector. We also compute the simplex-constrained estimator $\hat{w}_{\mathrm{QP}}$ (QP stands for "quadratic programming") defined by

$$\hat{w}_{\mathrm{QP}} = \operatorname*{argmin}_{w \in \Delta^{N-1}} \|Y - Xw\|_2^2,$$

[1] See more simulations in the supplementary Section S4.

which was introduced in Abadie & Gardeazabal (2003) and can be obtained by quadratic programming (Goldfarb & Idnani 2006, Abadie et al. 2011). The corresponding ATT estimator $\widehat{\text{ATT}}_{\text{QP}}$ is obtained similarly by replacing $\hat{w}_{\text{LS}}$ with $\hat{w}_{\text{QP}}$ in (18). Finally, we consider two oracle estimators that are not implementable in practice. Denote by $\gamma^* = \gamma(w^*)$ the true indicator vector corresponding to $w^*$. We compute the OLS and QP estimators using $(X_{\gamma^*}, Y)$. That is, these two estimators have access to $\gamma^*$ (which is not possible in reality) and will be used for comparison in our simulation studies.

## 4.2 Estimation of Average Treatment Effect

For each method considered, we compute rooted mean squared error of ATT estimation, i.e., the rooted average of $(\widehat{\text{ATT}} - \delta^*)^2$ over 100 replicates; denote it by $\text{RMSE}(\widehat{\text{ATT}})$. To measure the efficiency of an ATT estimator, we consider our BVS-SS estimator as the reference estimator, and define the relative efficiency of an estimator $\widehat{\text{ATT}}$ by

$$\text{RE}(\widehat{\text{ATT}}) = \frac{\text{RMSE}(\widehat{\text{ATT}}_{BVS-SS})}{\text{RMSE}(\widehat{\text{ATT}})}.$$

If $\text{RE}(\widehat{\text{ATT}}) = 100\%$, then this estimator is as efficient as BVS-SS.

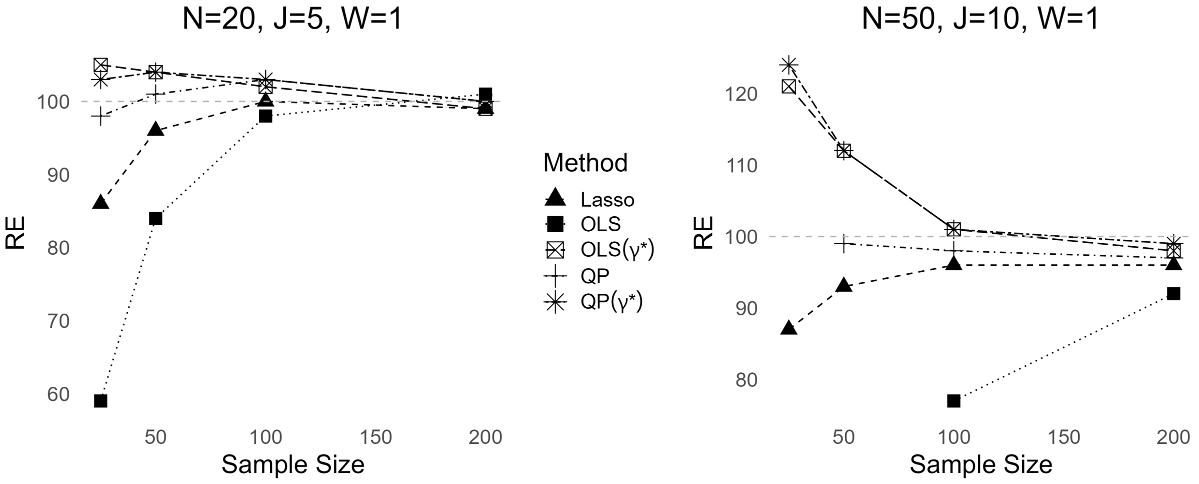

Figure 1: Relative efficiency (%) when $\|w^*\|_1 = 1$.

We report the relative efficiency of the six estimators with $\lambda = \|w^*\|_1 = 1$ in Figure 1. Our method outperforms Lasso substantially when $M = 25$ or 50, which shows that some prior knowledge about the simplex constraint can be very helpful when the sample size is small or moderate. By comparing BVS-SS with the two QP methods, we see that performing variable selection also significantly improves ATT estimation when the sample size is small. Indeed, when $N = 50, M = 25$, QP with the entire data set is not feasible since the underlying optimization problem does not have a unique solution. We also note that as sample size increases, the performance of BVS-SS quickly approaches that of the two oracle estimators, OLS($\gamma^*$) and QP($\gamma^*$).

Figure 2 and 3 present the results for $\|w^*\|_1 = 2$ or 3 (i.e., the simplex constraint is violated by the true model). Note that since the regression coefficient estimator of Lasso and OLS is scale-equivariant, the relative efficiency of Lasso and two OLS methods remain the same as in Figure 1. The most interesting observation from Figure 2 and 3 is that, though BVS-SS utilizes the simplex constraint, its performance remains robust. When $M = 25$, BVS-SS still has a significant advantage over Lasso. Indeed, Lasso almost fails to recover any signal contained in $w^*$ when $N = 50$ and $M = 25$. Once the sample size becomes large (say, $M \geq 100$), BVS-SS again achieves close-to-optimal performance, and it outperforms QP($\gamma^*$) by a wide margin, which knows the true model but assumes the simplex constraint. The robust and superior performance of BVS-SS across the three tables is largely due to the use of the soft simplex constraint, which enables BVS-SS to adapt to the unknown level of $\|w^*\|_1$ in the given data.

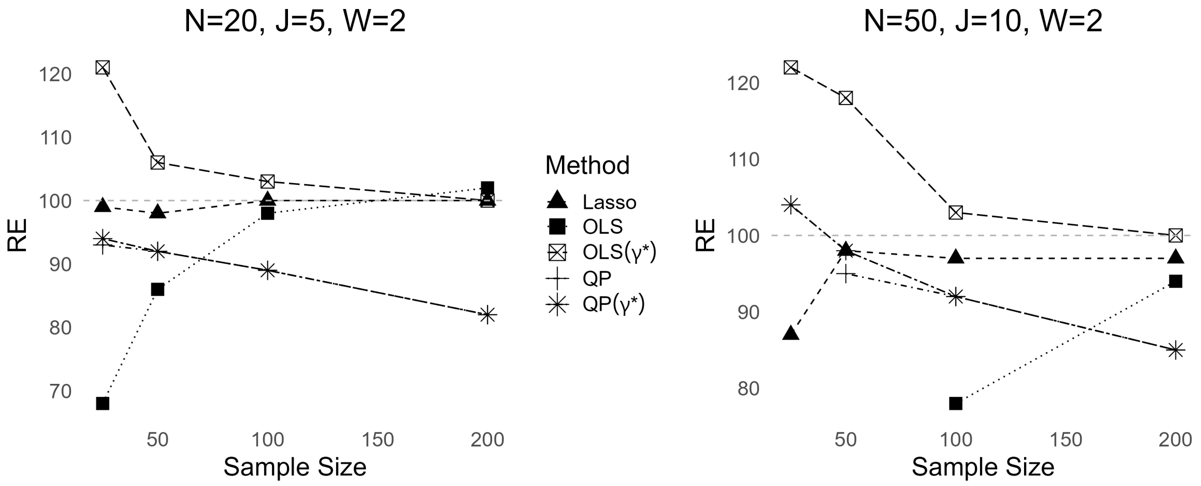

Figure 2: Relative efficiency (%) when $\|w^*\|_1 = 2$.

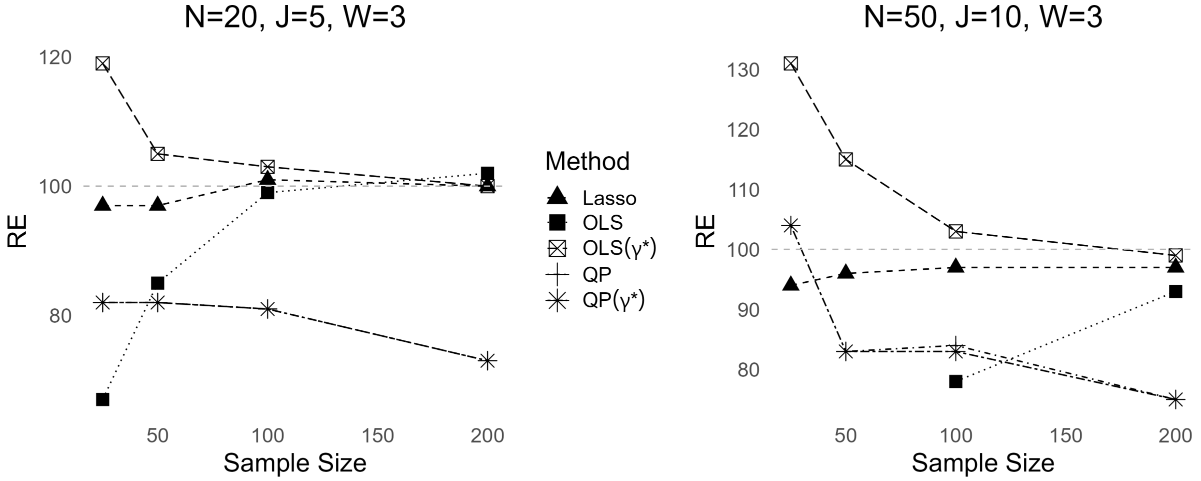

Figure 3: Relative efficiency (%) when $\|w^*\|_1 = 3$.

## 4.3 Estimation of Variance Parameters and Variable Selection

We now examine the performance of BVS-SS in terms of variable selection and the estimation of the two variance parameters $\tau, \phi$. First, we compare the variable selection result of BVS-SS with that of Lasso in Table 1, where "$\ell^1$-loss" denotes the $\ell^1$-distance between the selected model $\hat{\gamma}$ and the true model $\gamma^*$, and "model size" refers to the cardinality of $\hat{\gamma}$ (i.e., the number of nonzero entries of $\hat{w}$). For BVS-SS, the two statistics are averaged over the collected MCMC samples. It is evident that BVS-SS has a much higher accuracy than Lasso across all settings. Especially when $M$ is large, BVS-SS can correctly identify most entries of $\gamma$, while the performance of Lasso remains similar to that observed with small $M$.

|  |  |  | BVS-SS |  | Lasso |  |
|---|---|---|---|---|---|---|
| $\|w^*\|_1$ | Setting | $M$ | $\ell^1$-loss | model size | $\ell^1$-loss | model size |
| $\|w^*\|_1 = 1$ | $N = 20$<br>$J = 5$ | 25 | 3.1 | 3.9 | 6.3 | 10.2 |
|  |  | 50 | 1.9 | 4.8 | 6.7 | 11.1 |
|  |  | 100 | 1.3 | 5.1 | 6 | 10.9 |
|  |  | 200 | 0.8 | 5.4 | 6.3 | 11.3 |
|  | $N = 50$<br>$J = 10$ | 25 | 10.9 | 6.2 | 13 | 14.2 |
|  |  | 50 | 7.8 | 9.1 | 15.4 | 22.2 |
|  |  | 100 | 5.5 | 10.9 | 14.4 | 22.9 |
|  |  | 200 | 4.1 | 11.4 | 13.4 | 22.5 |
| $\|w^*\|_1 = 3$ | $N = 20$<br>$J = 5$ | 25 | 3.3 | 3.3 | 6.3 | 10.2 |
|  |  | 50 | 1.9 | 4.8 | 6.7 | 11.1 |
|  |  | 100 | 1.2 | 5 | 6 | 10.9 |
|  |  | 200 | 0.7 | 5.2 | 6.3 | 11.3 |
|  | $N = 50$<br>$J = 10$ | 25 | 10.2 | 4.5 | 13 | 14.2 |
|  |  | 50 | 8.1 | 9.2 | 15.4 | 22.2 |
|  |  | 100 | 5.6 | 11.2 | 14.4 | 22.9 |
|  |  | 200 | 3.7 | 10.8 | 13.4 | 22.5 |

Table 1: Variable selection performance of BVS-SS and Lasso averaged over 100 replicates.

Next, we visualize the distribution of the posterior mean estimate of $\phi$ (error precision) across 100 replicates in Figure 4. This statistic reflects how well the BVS-SS model fits the data. We only show the result for $N = 50$, as the distribution for $N = 20$ is similar. Observe that regardless of $\|w^*\|_1$, the posterior mean of $\phi$ increases with $M$, which is expected since a larger sample size enables BVS-SS to detect more signals in the data, resulting in a smaller estimate of the error variance. When $\|w^*\|_1 = 3$, BVS-SS appears to always underestimate $\phi$ even when the sample size is sufficiently large, which may be related to the fact that the simplex constraint is significantly violated in this case. Notably, when $\|w^*\|_1 = 2$, the posterior estimation of $\phi$ appears highly accurate for $M \geq 100$, indicating that our model achieves an optimal balance between model complexity and fitting.

Finally, we examine the posterior mean estimate of $\tau$ when $N = 50$, for which the result is presented in Figure 5. As discussed in Section 2.1, $\tau$ can be considered as an indicator of whether the simplex constraint is satisfied by the given data, and we now explain why Figure 5 provides compelling empirical evidence supporting this. When $\|w^*\|_1 = 1$, we see that the posterior mean of $\tau$ decreases as $M$ increases, since a larger sample size offers stronger evidence that the simplex constraint is satisfied, which draws $\tau$ towards zero. In contrast, when $\|w^*\|_1 = 2$ or 3, the trend reverses, since a larger sample size enables BVS-SS to detect deviations from the simplex constraint. When $M = 200$, we find that the ratio of the posterior mean estimate of $\tau$ and that of $\phi$ equals 0.0003 for $\|w^*\|_1 = 1$, 0.024 for $\|w^*\|_1 = 2$, and 0.081 for $\|w^*\|_1 = 3$. This actually aligns well with the true data-generating mechanism: since we use $J = 10$ for $N = 50$, we may estimate the "true" mean squared deviation of $w^*$ from the simplex constraint by $0.1 \times \sum_{j=1}^{10}(w_j^* - 0.1)^2$, which equals 0.021 for $\|w^*\|_1 = 2$ and 0.065 for $\|w^*\|_1 = 3$.

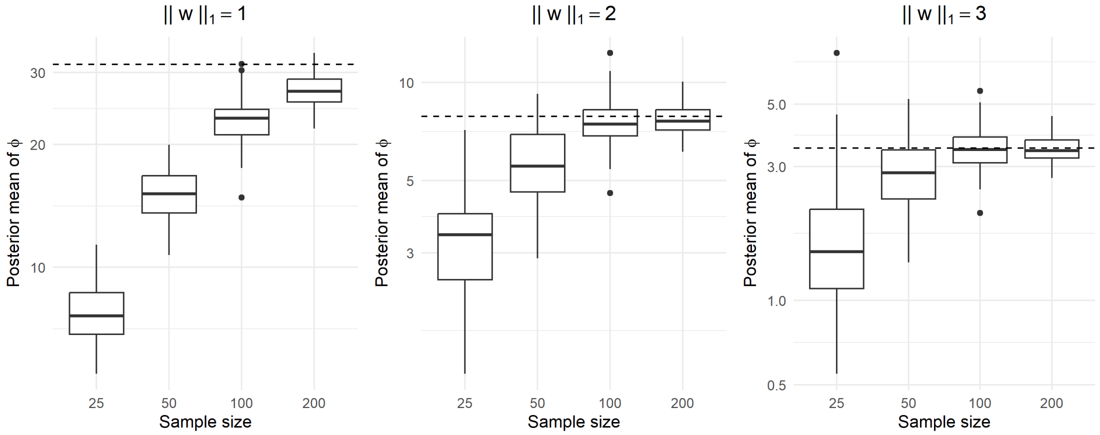

Figure 4: Distribution of the posterior mean of $\phi$ across 100 replicates with $N = 50, J = 10$. The true value $\phi^*$ is indicated by the dotted line.

# 5 Empirical Examples

In this section, we revisit two empirical examples from the SCM literature. We will use the notation introduced in Section 1.1: $T_0$ is the number of pre-treatment time periods, $T_1$

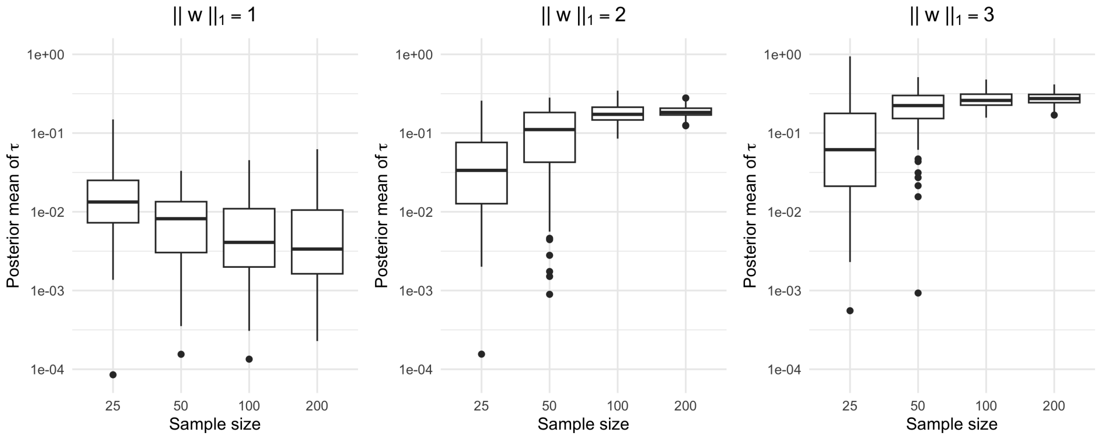

Figure 5: Distribution of the posterior mean of $\tau$ across 100 replicates with $N = 50, J = 10$.

is the number of post-treatment time periods, the total number of units is $N + 1$ where the first one is used as the response and the other $N$ units as the explanatory variables. (For BVS-SS, $M = T_0$ and $\tilde{M} = T_1$.) The first example has $N < T_0$, while the second has $N \gg T_0$, and it will be shown that our method works well in both settings.

## 5.1 Nota Fiscal Paulista: Anti-tax Evasion

An anti-tax evasion program, Nota Fiscal Paulista (NFP), was implemented in São Paulo, Brazil in October 2007 with the goal of reducing tax evasion by incentivizing consumers to request electronic receipts in exchange for participation in monthly lotteries promoted by the government, as well as chances for receiving partly tax rebates. Thus, consumers are encouraged to participate in the auditing schemes to decrease the chances of tax evasion. However, those restaurant owners and retailers may pass on part of the tax costs to consumers by raising product prices. Although the NFP was implemented sequentially across various sectors, including restaurants, bakeries, bars, and food service retailers, the literature has suggested that the food away from home (FAH) index could be used as a suitable indicator for price levels of these sectors. Carvalho et al. (2018) proposed an artificial counterfactual (ArCo) estimation procedure, by connecting Lasso to SCM without simplex constraint, to explore how the NFP affects the inflation on FAH. They also included monthly GDP growth, retail sales growth and monthly credit growth. Their ArCo estimation indicated that the ATT of NFP is 0.4478%.

We obtain the data analyzed in Carvalho et al. (2018) from the `R` package `ArCo`. The pre-treatment periods span $T_0 = 33$ months, and the post-treatment periods span $T_1 = 23$ months. Slightly different from the original study, the data includes only $N = 8$ control units, which is one fewer than that mentioned in their paper, and there is no data related to credit and retail sales. However, this does not affect the application of our method, as our framework currently does not involve covariates beyond the control units.

We run our Algorithm 1 for 1000 iterations to estimate the impact of NFP. Figure 6 shows the trajectories of the posterior samples of $\phi$ and $\log(\tau)$ (with burn-in included). Probably due to the limited number of observations, the MCMC samples of $\tau$ exhibit large variability, indicating that the simplex assumption is hard to confirm or reject. The trace plot suggests that our sampler has converged fast. Table 2 summarizes the posterior mean and 95% credible intervals of ATT, $\tau$, $\phi$ and the model size $|\gamma|$; see Figure 18 in the Appendix for histograms of the MCMC samples and Figure 20 in the Appendix for the counterfactual estimation. The first half of the samples have been dropped for burn-in. The ATT estimation is close to that of (Carvalho et al. 2018, Column 1, Table 5), but our credible interval is much smaller compared to their confidence set. While the analysis of Carvalho et al. (2018) selected all the predictors, our method only chooses around two areas for constructing the synthetic controls on average. This is consistent with the findings of our simulation study presented in Section 4, which shows that Lasso has the tendency of over-estimating the model size compared to BVS-SS.

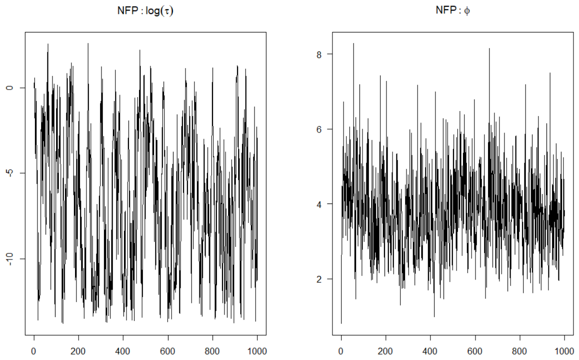

Figure 6: The trace plot of $\log(\tau)$ and $\phi$ for the NFP data set.

Table 2: NFP Impact on Food Inflation

| | ATT | $\tau$ | $\phi$ | $\lvert\gamma\rvert$ |
|---|---|---|---|---|
| Mean | 0.288 | 0.14 | 3.85 | 2.29 |
| 95% credible interval | (0.141, 0.419) | (0, 1.44) | (1.94, 6.05) | (1, 4) |

## 5.2 China's anti-corruption campaign

China launched an unprecedented anti-corruption campaign in November 2012, aimed at curbing corruption and power abuse within government and military departments. In the context of bribery, direct cash payments are often considered too blatant and are generally unpopular, and high-end imported luxury goods serve as a more discreet way to convey the value of a gift in line with cultural norms of subtlety. Empirical studies, such as Lan & Li (2018), have shown that there is a comovement between the importation of luxury watches in China and changes in the government leadership. An interesting study was conducted by Shi & Huang (2023), who investigated the effect of China's anti-corruption campaign on the importation of luxury watches. They took the monthly growth rate of luxury watch imports in US dollars as the outcome of interest and selected the control group among the growth rates of $N = 87$ commodity categories to construct the synthetic counterfactual for luxury watch imports. The study utilizes data from the United Nations Comtrade Database, focusing on the category “watches with case of, or clad with, precious metal.” January 2013 is considered the time of treatment, which was the month immediately following the announcement of the Eight-Point Policy (an anti-corruption policy). The pre-treatment period spans from February 2010 to December 2012 resulting in $T_0 = 35$ observations, and the post-treatment period covers January 2013 to December 2015, which yields $T_1 = 36$ observations. Their forward selection algorithm identified three control units: “knitted or crocheted fabric,” “cork and articles of cork,” and “salt, sulfur, earth, stone, plaster, lime, and cement.” Their estimated counterfactual predicted that without the anti-corruption campaign, luxury watch imports would have continued to increase. However, in reality, imports dropped by 42% in January 2013, while the counterfactual predicted a 1.7% increase. Their treatment effect estimation indicated a reduction of 3.09% in luxury watch imports per month over the post-treatment period. Accumulated over 36 months, the campaign reduced total luxury watch imports by approximately two-thirds.

We obtained the data of Shi & Huang (2023) from the `R` package `fdPDA`, which contains all the 88 commodity categories (including the response and $N = 87$ control categories). Figure 7 shows the MCMC trajectories of the posterior samples of $\phi$ and $\log(\tau)$. Compared to the previous real-data example, the posterior distribution of $\tau$ is still dispersed, but the ratio $\tau/\phi$ becomes much smaller, indicating that the simplex constraint is more likely to be satisfied (at least approximately) in this data set. The estimation of $\phi$ appears to be much more accurate, suggesting that the error variance is around 0.05. As detailed in Table 3, our ATT estimation is smaller than that of Shi & Huang (2023) in absolute value, but the 95% credible interval indicates that the anti-corruption has a very significant effect on the high-end watch imports. On average, our model size $|\gamma|$ is larger than that of Shi & Huang (2023) (who only selected 3 control categories), but we note that the posterior distribution of $|\gamma|$, which is provided in Figure 19 in the Appendix, still concentrates on small models with size $\leq 5$. Figure 21 in the Appendix depicts the counterfactual estimation, which indicates a dramatic fall in luxury watch importation right after the treatment.

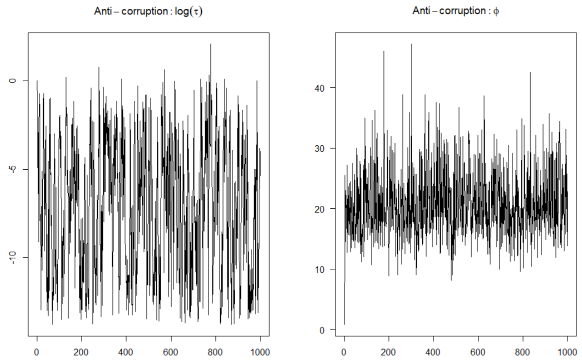

Figure 7: The trace plot of $\log(\tau)$ and $\phi$ for the anti-corruption data set.

Table 3: Anti-corruption Campaign's Impact on Luxury Watches Importation

| | ATT | $\tau$ | $\phi$ | $\lvert\gamma\rvert$ |
|---|---|---|---|---|
| Mean | -0.021 | 0.069 | 20.86 | 5.09 |
| 95% credible interval | (-0.032, -0.008) | (0, 0.641) | (12.22, 32.76) | (1, 20) |

# 6 Conclusion

We propose a novel Bayesian synthetic control method, BVS-SS, that integrates a relaxed simplex constraint with Bayesian spike-and-slab variable selection. Our approach introduces a hierarchical prior to determine how closely the data should adhere to the simplex constraint, a feature that is missing in existing frequentist or Bayesian synthetic control frameworks and helps mitigate the ongoing debate surrounding the use of simplex constraints. To efficiently compute the posterior distribution of BVS-SS, we develop a novel Metropolis-within-Gibbs sampler, which overcomes the hindrance caused by the simplex constraint by updating two regression coefficients simultaneously from the full conditional posterior. We also show that

BVS-SS consistently selects the true control units and thus provides posterior samples of coefficients with consistent means when the true DGP agrees with the simplex constraint, and BVS-SS tends to provide selection results asymptotically equivalent to the posteriors of classical spike-and-slab prior if the true DGP strongly disagrees with the simplex constraint. Simulation studies and real-data examples illustrate the advantages and usefulness of our proposed algorithm.

There are several interesting directions that future researchers may dive in. First, it is known that Gibbs sampling methods are susceptible to high collinearity among the explanatory variables, which is a common situation in SCM applications where factor models are often used as the underlying data-generating process. Though it is unclear if such issues affect the methodology proposed in this work, since our Algorithm 1 updates two coordinates at a time, it would be interesting to investigate the use of other MCMC techniques, such as pseudo-marginal sampling (Andrieu & Roberts 2009). Second, to include time-invariant, time-variant predictors or characteristics, one can extend BVS-SS by considering two types of explanatory variables, with one type following the spike-and-slab prior and the other following a regular non-informative prior without variable selection. Third, more numerical studies or model extensions can be performed to study the performance of BVS-SS when the model is mis-specified (e.g., the data is non-stationary, heterogeneous error), which would be of great interest to economists. We hope our method will serve as an inspiration for future research and facilitate the widespread use of Bayesian synthetic control methods.

# Acknowledgements

The work was supported in part by the NSF grants DMS-2311307 and DMS-2245591.

# SUPPLEMENTARY MATERIAL

## S1 A Metropolis-within-Gibbs Sampler for BVS-SS

For Bayesian variable selection involving a prior on the indicator vector $\gamma$ as we have considered, the standard approach to posterior sampling is to integrate out the regression coefficient vector and use add-delete-swap proposals to construct a Metropolis–Hastings algorithm targeting the marginal posterior distribution of $\gamma$ (or the joint distribution of $\gamma$ and some variance parameter); see, e.g., George & McCulloch (1997), Chipman et al. (2001), Guan & Stephens (2011), Zhou et al. (2022). However, this approach is not applicable to our model, since $\mu_\gamma$ cannot be integrated out in closed form from the joint posterior distribution of $(\gamma, \mu_\gamma, \tau, \phi)$. To tackle this very unique challenge, we propose a novel Metropolis-within-Gibbs sampler targeting the joint posterior distribution of $(\mu, \tau, \phi)$.

### S1.1 Updating of $\tau$ and $\phi$

The updates for $\tau$ and $\phi$ are standard. We update $\phi$ by drawing it from the full conditional distribution given in (8), and we use a Metropolis–Hastings update for $\tau$ that is invariant with respect to $p(\tau \mid y, \mu, \phi)$. Explicitly, we propose a new value for $\tau^*$ by a Gaussian random walk on log-scale with variance $\eta > 0$ and accept $\tau^*$ with probability

$$\rho(\tau, \tau^*) = \min\left\{1, \frac{p(y \mid \mu, \tau^*, \phi)p(\tau^*)}{p(y \mid \mu, \tau, \phi)p(\tau)} \frac{\log(\tau^*)}{\log(\tau)}\right\}, \tag{S1}$$

where the likelihood is given in (6) and $p(\tau)$ denotes the prior density. This step can be repeated multiple times so that $\tau$ is updated to some value with larger density (conditioned on $\mu$ and $\phi$).

## S1.2 Updating of $\mu$

For $\mu$, we propose a Gibbs-type updating scheme, which, to our knowledge, is quite different from any existing Gibbs scheme for similar problems. Let $\mu_{-j}$ denote the subvector of $\mu$ with $\mu_j$ removed. Since the prior of BVS-SS assigns probability one to $\mu \in \Delta^{N-1}$, given $\mu_{-j}$, we have $\mu_j = 1 - \sum_{i \neq j} \mu_i$ almost surely, and thus the full conditional distribution $p(\mu_j \mid y, \mu_{-j}, \tau, \phi)$ is also degenerate. Hence, a Gibbs scheme that updates each $\mu_j$ from its full conditional posterior will not move at all. This motivates us to devise a more complicated updating scheme by updating $\mu_i, \mu_j$ (with $i \neq j$) simultaneously from the full conditional posterior $p(\mu_i, \mu_j \mid y, \mu_{-(i,j)}, \tau, \phi)$, where $\mu_{-(i,j)}$ denotes the subvector of $\mu$ with the $\mu_i, \mu_j$ removed.

Let $\mu_{-(i,j)}, \tau, \phi$ be given and define $s = 1 - \sum_{k \neq i,j} \mu_k$. If $s = 0$, the simplex constraint implies that $\mu_i = \mu_j = 0$ almost surely. If $s \in (0, 1]$, $p(\mu_i, \mu_j \mid y, \mu_{-(i,j)}, \tau, \phi)$ is a mixture of a discrete distribution on $\{(s, 0), (0, s)\}$ and a degenerate continuous distribution on $\{(u, s-u) \colon 0 < u < s\}$. Therefore, to find a closed-form expression for this conditional distribution, it suffices to find $p(\gamma_i, \gamma_j \mid y, \mu_{-(i,j)}, \tau, \phi)$ for $(\gamma_i, \gamma_j) \in \{(1, 0), (0, 1), (1, 1)\}$ and the conditional distribution $p(\mu_i, \mu_j \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i = \gamma_j = 1)$. To this end, we introduce the notation $\gamma^0, \gamma^i, \gamma^j, \gamma^{ij}$ defined by

$$\gamma^0_k = \mathbb{1}(\mu_k \neq 0), \quad \gamma^i_k = \mathbb{1}(\mu_k \neq 0 \text{ or } k = i), \quad \gamma^{ij}_k = \mathbb{1}(\mu_k \neq 0 \text{ or } k \in \{i, j\}). \tag{S2}$$

That is, $\gamma^i$ is the model obtained by setting $\mu_i = s, \mu_j = 0$, and $\gamma^{ij}$ is the one obtained by setting $\mu_i, \mu_j > 0$. The full conditional distribution of $(\mu_i, \mu_j)$ given other parameters and $\gamma_i = \gamma_j = 1$ is characterized in the following lemma. For simplicity, we assume $\alpha = 1$ in this section, and in Remark 1 we explain how to extend the results to other integer values of $\alpha$.

**Lemma S1.** *Assume $\alpha = 1$. Fix $i \neq j$ such that $s = 1 - \sum_{k \neq i,j} \mu_k \in (0, 1]$. The conditional* *distribution* $p(\mu_i, \mu_j \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i = \gamma_j = 1)$ *is degenerate with* $\mu_j = s - \mu_i$ *and*

$$\mu_i \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i = \gamma_j = 1 \sim N_{(0,s)} \left( \beta_{i,j},\ \frac{1}{\phi \Lambda_{i,j}} \right).$$

*In the above expression,* $N_{(a,b)}$ *denotes the univariate truncated normal distribution restricted to the interval* $(a, b)$*,*

$$\begin{aligned} \beta_{i,j} &= \frac{1}{\Lambda_{i,j}} (X_i - X_j)^\top \Sigma_{\gamma^{ij},\tau} \left( y - sX_j - \textstyle\sum_{k \neq i,j} \mu_k X_k \right), \\ \Lambda_{i,j} &= (X_j - X_i)^\top \Sigma_{\gamma^{ij},\tau} (X_j - X_i), \end{aligned}$$

*where* $\gamma^{ij}$ *is given by* (S2) *and* $\Sigma_{\gamma,\tau}$ *is given by* (7)*.*

*Proof.* We can express the conditional posterior of $(\mu_i, \mu_j)$ by

$$p(\mu_i, \mu_j \mid y, \phi, \tau, \mu_{-(i,j)}) = C\, p(y \mid \phi, \tau, \mu_i, \mu_j, \mu_{-(i,j)}) p(\mu_i, \mu_j, \mu_{-(i,j)}),$$

where $C$ is a constant that depends on $\mu_{-(i,j)}$ but not $\mu_i, \mu_j$, and the marginal likelihood $p(y \mid \phi, \tau, \mu_i, \mu_j, \mu_{-(i,j)})$ is shown in (6).

First, look at $\mu$'s prior $p(\mu_i, \mu_j, \mu_{-(i,j)})$. Recall that $\mu|\gamma \sim$ sym-Dirichlet($\alpha$). Denote $p_0(|\gamma|) = \theta^{|\gamma|}(1-\theta)^{N-|\gamma|}$ for a short notation. Given $\alpha = 1$, $\mu$ is uniform on the $\Delta^{|\gamma|-1}$. To satisfy the simplex constraint, given $\sum_{k \neq i,j} \mu_k < 1$, the density of $\mu_i = \mu_j = 0$ must be zero. Noticing that $p(\mu) = p(\mu|\gamma)p(\gamma)$, we have:

$$\begin{aligned} p(\mu_i = 0, \mu_j = 0, \mu_{-(i,j)}) &= 0, \\ p(\mu_i = s, \mu_j = 0, \mu_{-(i,j)}) &= (\ell)!\, p_0(\ell + 1), \\ p(\mu_i = t, \mu_j = s - t, \mu_{-(i,j)}) &= (\ell + 1)!\, p_0(\ell + 2), \quad \forall t \in (0, s), \end{aligned}$$

Integrating over $t$ in the last expression, we obtain that

$$\begin{aligned} p(\gamma_i = 1, \gamma_j = 1, \mu_{-(i,j)}) &= \int_0^s t \cdot p(\mu_i = t, \mu_j = s - t, \mu_{-(i,j)}) dt \\ &= s(\ell + 1)!\, p_0(\ell + 2) \end{aligned}$$

Recall the definition in Lemma S2:

$$\check{y}(u) = y - uX_i - (s - u)X_j - \textstyle\sum_{k \neq i,j} \mu_k X_k.$$

We can express the conditional posterior density of $(\mu_i, \mu_j) = (s, 0)$ by

$$p(\mu_i = s, \mu_j = 0 \mid y, \phi, \tau, \mu_{-(i,j)}) \propto A(\gamma^i, \tau) \exp\left\{-\frac{\phi}{2} \check{y}(s)^\top \Sigma(\gamma^i, \tau) \check{y}(s)\right\},$$

where

$$A(\gamma, \tau) = \det(V_{\gamma,\tau})^{-1/2} (|\gamma| - 1)!\, p_0(|\gamma|), \quad \Sigma(\gamma, \tau) = I - X_\gamma V_{\gamma,\tau}^{-1} X_\gamma^\top.$$

The conditional posterior density of $(\mu_i, \mu_j) = (0, s)$ has the same expression with $\check{y}(s)$ replaced by $\check{y}(0)$ and $\gamma^i$ replaced by $\gamma^j$. The conditional posterior density of $(\mu_i > 0, \mu_j > 0)$ is:

$$p(\mu_i = u, \mu_j = s - u \mid y, \phi, \tau, \mu_{-(i,j)}) \propto A(\gamma^{i,j}, \tau) \exp\left\{-\frac{\phi}{2} \check{y}(u)^\top \Sigma(\gamma^{i,j}, \tau) \check{y}(u)\right\}.$$

Observe that the density is proportional to the structure of an exponential of $t$'s quadratic terms, and $t$ is bounded in $(0, s)$. $t$ must follow a truncated Gaussian distribution restricted to the interval of $(0, s)$. Rewrite the above exponential part as:

$$\exp\left\{-\frac{\phi}{2}\left((X_j - X_i)^\top \Sigma(\gamma^{i,j}, \tau)(X_j - X_i)t^2 - 2((X_i - X_j)^\top \Sigma(\gamma^{i,j}, \tau)y^j) + C_3\right)\right\}.$$

where $C_3$ is some constant not containing $t$. Then it is clear to see that $\mu_i$'s conditional posterior follows a truncated Gaussian distribution as stated in Lemma S1. □

To find the posterior probabilities of $(\gamma_i, \gamma_j) = (1,0), (0,1), (1,1)$ given $y, \mu_{-(i,j)}, \tau, \phi$, we first express the full conditional posterior of $(\mu_i, \mu_j)$ by

$$p(\mu_i, \mu_j \mid y, \mu_{-(i,j)}, \tau, \phi) = C_1\, p(y \mid \mu_i, \mu_j, \mu_{-(i,j)}, \tau, \phi) p(\mu_i, \mu_j, \mu_{-(i,j)}),$$

where $C_1$ is a constant that depends on $\mu_{-(i,j)}$ but not $(\mu_i, \mu_j)$. Since $\mu_\gamma \mid \gamma \sim \text{Dirichlet}(\alpha)$, if $s = 1 - \sum_{k \neq i,j} \mu_k > 0$, we have

$$\begin{aligned} p(\mu_i = s, \mu_j = 0, \mu_{-(i,j)}) &= C_2 \frac{\Gamma((\ell+1)\alpha)}{\Gamma(\alpha)^{\ell+1}}\, s^{\alpha-1}\, \theta^{\ell+1}(1-\theta)^{N-\ell-1}, \\ p(\mu_i = u, \mu_j = s-u, \mu_{-(i,j)}) &= C_2 \frac{\Gamma((\ell+2)\alpha)}{\Gamma(\alpha)^{\ell+2}}\, u^{\alpha-1}(s-u)^{\alpha-1}\, \theta^{\ell+2}(1-\theta)^{N-\ell-2}, \quad \forall u \in (0,s), \end{aligned}$$

where $\ell = \sum_{k \neq i,j} \mathbb{1}_{\{\mu_k \neq 0\}}$ and $C_2$ is some constant that depends on $\mu_{-(i,j)}$ but not $(\mu_i, \mu_j)$. Integrating $p(y \mid \mu_i = u, \mu_j = s-u, \mu_{-(i,j)}, \tau, \phi) p(\mu_i = u, \mu_j = s-u, \mu_{-(i,j)})$ with respect to $u \in (0,s)$, we obtain the marginal conditional posterior probabilities of $\gamma^i, \gamma^j, \gamma^{ij}$.

**Lemma S2.** *Define a function $A(\gamma, \tau)$ by*

$$A(\gamma, \tau) = \frac{\Gamma(|\gamma|\alpha)}{\Gamma(\alpha)^{|\gamma|}}\, \tau^{-|\gamma|/2} \det(V_{\gamma,\tau})^{-1/2}\, \theta^{|\gamma|}(1-\theta)^{N-|\gamma|}.$$

*Fix $i \neq j$ such that $s = 1 - \sum_{k \neq i,j} \mu_k \in (0, 1]$. Define*

$$\check{y}(u) = y - uX_i - (s-u)X_j - \sum_{k \neq i,j} \mu_k X_k.$$

*Let $\gamma^0, \gamma^i, \gamma^j, \gamma^{ij}$ be given by (S2). Then,*

$$\begin{aligned}
p(\gamma^0 \mid y, \mu_{-(i,j)}, \tau, \phi) &= 0, \\
p(\gamma^i \mid y, \mu_{-(i,j)}, \tau, \phi) &= C\, s^{\alpha-1} A(\gamma^i, \tau) \exp\left\{ -\frac{\phi}{2} \check{y}(s)^\top \Sigma_{\gamma^i,\tau} \check{y}(s) \right\}, \\
p(\gamma^j \mid y, \mu_{-(i,j)}, \tau, \phi) &= C\, s^{\alpha-1} A(\gamma^j, \tau) \exp\left\{ -\frac{\phi}{2} \check{y}(0)^\top \Sigma_{\gamma^j,\tau} \check{y}(0) \right\}, \\
p(\gamma^{ij} \mid y, \mu_{-(i,j)}, \tau, \phi) &= C A(\gamma^{ij}, \tau) \int_0^s u^{\alpha-1} (s-u)^{\alpha-1} \exp\left\{ -\frac{\phi}{2} \check{y}(u)^\top \Sigma_{\gamma^{ij},\tau} \check{y}(u) \right\} \mathrm{d}u,
\end{aligned}$$

*where $C > 0$ is a constant. In particular, when $\alpha = 1$,*

$$\begin{aligned}
&p(\gamma^{ij} \mid y, \mu_{-(i,j)}, \tau, \phi) \\
&= \frac{C A(\gamma^{ij}, \tau)\sqrt{2\pi}}{\sqrt{\phi \Lambda_{i,j}}} e^{-\frac{\phi}{2}\left\{ \check{y}(0)^\top \Sigma_{\gamma^{ij},\tau} \check{y}(0) - \beta_{i,j}^2 \Lambda_{i,j} \right\}} \left\{ \Phi\left( (s - \beta_{i,j}) \sqrt{\phi \Lambda_{i,j}} \right) - \Phi\left( -\beta_{i,j} \sqrt{\phi \Lambda_{i,j}} \right) \right\},
\end{aligned}$$

*where $\beta_{i,j}, \Lambda_{i,j}$ are as given in Lemma S1 and $\Phi$ denotes the CDF of $N(0,1)$.*

*Proof.* For simplicity, we assume $\alpha = 1$ as in Lemma S1. $A(\gamma, \tau)$ collapses to $\det(V_{\gamma,\tau})^{-1/2} (|\gamma| - 1)!\, p_0(|\gamma|)$. The extension is almost trivial. Given $\mu_{-(i,j)}$, $\mu_i = s$ and $\mu_j = 0$, $\gamma_i = 1$ and $\gamma_j = 0$ with probability equal to one. Thus,

$$\begin{aligned}
p(\gamma^i \mid y, \mu_{-(i,j)}, \tau, \phi) &= p(\gamma_i = 1, \gamma_j = 0 \mid y, \mu_{-(i,j)}, \tau, \phi) \\
&= p(\mu_i = s, \mu_j = 0 \mid y, \mu_{-(i,j)}, \tau, \phi) \\
&= A(\gamma^i, \tau) \exp\left\{ -\frac{\phi}{2} \check{y}(s)^\top \Sigma_{\gamma^i,\tau} \check{y}(s) \right\}
\end{aligned}$$

Similarly,

$$
\begin{aligned}
&p(\gamma^0 \mid y, \mu_{-(i,j)}, \tau, \phi) = 0, \\
&p(\gamma^j \mid y, \mu_{-(i,j)}, \tau, \phi) \propto A(\gamma^j, \tau) \exp\left\{ -\frac{\phi}{2} \check{y}(0)^\top \Sigma_{\gamma^j,\tau} \check{y}(0) \right\},
\end{aligned}
$$

To find the conditional posterior of $\gamma^{ij}$, we need to integrate out $u$:

$$
\begin{aligned}
p(\gamma_i = 1, \gamma_j = 0 \mid y, \phi, \tau, \mu_{-(i,j)}) \propto& \int_0^s A(\gamma^{i,j}, \tau) \exp\left\{ -\frac{\phi}{2} \check{y}(u)^\top \Sigma_{\gamma^{i,j},\tau} \check{y}(u) \right\} du \\
&= A(\gamma^{i,j}, \tau) \exp\left\{ -\frac{\phi}{2} \left( \check{y}(u)^\top \Sigma_{\gamma^{i,j},\tau} \check{y}(u)^\top - \beta_{i,j}^2 \Lambda_{i,j} \right) \right\} \\
&\quad \frac{\sqrt{2\pi}}{\sqrt{\phi \Lambda_{i,j}}} \int_0^s \frac{\sqrt{\phi \Lambda_{i,j}}}{\sqrt{2\pi}} \exp\left\{ -\frac{\phi \Lambda_{i,j}}{2} (u - \beta_{i,j})^2 \right\} \mathrm{d}u,
\end{aligned}
$$

Observe that the integration part equals to an integral of a truncated Gaussian variable, and thus can be considered as a normalizing constant:

$$
\int_0^s \frac{\sqrt{\phi \Lambda_{i,j}}}{\sqrt{2\pi}} \exp\left\{ -\frac{\phi \Lambda_{i,j}}{2} (u - \beta_{i,j})^2 \right\} \mathrm{d}u = \ \Phi\left( (s - \beta_{i,j}) \sqrt{\phi \Lambda_{i,j}} \right) - \Phi\left( -\beta_{i,j} \sqrt{\phi \Lambda_{i,j}} \right),
$$

where $\Phi$ denotes the CDF of $N(0, 1)$. □

**Remark 1.** For any positive integer $\alpha$, the integral involved in $p(\gamma^{ij} \mid y, \mu_{-(i,j)}, \tau, \phi)$ can always be expressed by using the CDF of the standard normal distribution, and thus the conditional posterior probabilities of $\gamma^i, \gamma^j, \gamma^{ij}$ can be computed very efficiently. For non-integer-valued $\alpha$, numerical integration is needed. The conditional posterior distribution of $(\mu_i, \mu_j)$ given $\gamma^{ij}$ is

$$
p(\mu_i = u, \mu_j = s - u \mid y, \mu_{-(i,j)}, \tau, \phi, \gamma_i = \gamma_j = 1) \ \propto \ u^{\alpha - 1} (s - u)^{\alpha - 1} e^{-\frac{\phi}{2} \check{y}(u)^\top \Sigma_{\gamma^{ij},\tau} \check{y}(u)},
$$

for $u \in (0, s)$. When $\alpha = 1$, $u$ follows the truncated normal distribution, as shown in Lemma S1, and this allows for straightforward sampling of $(\mu_i, \mu_j)$. For other integer values of $\alpha$, one can use rejection sampling to generate samples of $u$, where the truncated normal distribution can be used as a reference distribution.

## S1.3 Algorithm

We can now formally describe our Metropolis-within-Gibbs sampler targeting $p(\mu, \tau, \phi \mid y)$, which is given in Algorithm 1. To emphasize its application to synthetic control, we assume that we have access to another data set $(\tilde{X}, \tilde{Y}^{(1)})$, and we take the approach outlined in Section 2.2 to estimating ATT defined in (11).

The most computationally intensive part of Algorithm 1 is the evaluation of expressions of the form $V_{\gamma,\tau}^{-1} z$ for a given vector $z$, which may seem highly time-consuming if $|\gamma|$ is large. But this can be implemented efficiently by computing and updating the Cholesky decomposition of $V_{\gamma,\tau}$, a technique commonly employed in Bayesian spike-and-slab variable selection (Smith & Kohn 1996, George & McCulloch 1997). More specifically, at the beginning of the $t$-th iteration, we compute the Cholesky decomposition of $V_{\gamma,\tau^{(t-1)}}$, where $\gamma$ is the model selected in the last iteration. Then, each time we draw $(\mu_i, \mu_j)$ from its full conditional posterior, we need to change one or two coordinates of $\gamma$, and the resulting Cholesky decomposition of $V_{\gamma,\tau^{(t-1)}}$ can be obtained by standard updating algorithms (Golub & Van Loan 2013).

Certain variations of Algorithm 1 are also straightforward to implement. For example, one can replace the fixed-scan Gibbs updating of $\mu$ in Algorithm 1 by a random-scan update: in the $t$-th iteration, we uniformly sample a pair $i < j$ such that $\mu_i^{(t-1)} + \mu_j^{(t-1)} > 0$ and update $(\mu_i^{(t)}, \mu_j^{(t)})$ from its conditional posterior distribution (an acceptance-rejection step is needed to correct for the proposal bias), and then draw $\phi^{(t)}$ from the full conditional posterior and

**Algorithm 1**. Metropolis-within-Gibbs sampling for BVS-SS.

```
Input: training data set $(X, Y)$, hyperparameters $\kappa_1, \kappa_2, a_1, a_2, \alpha, \theta$, algorithm
       parameters $n_\tau \in \mathbb{N}, \eta > 0$ for updating $\tau$, number of iterations $n$, initial
       state $(\mu^{(0)}, \tau^{(0)}, \phi^{(0)})$, data set under treatment $(\tilde{X}, \tilde{Y}^{(1)})$.
for $t = 1, \ldots, n$ do
    Set $\mu \leftarrow \mu^{(t-1)}$ ;
    for $i = 1, \ldots, N-1$ do
        for $j = i+1, \ldots, N$ do
            Calculate $s \leftarrow 1 - \sum_{k \neq i,j} \mu_k$ ;
            if $s = 0$ then
                Set $\mu_i \leftarrow 0$, $\mu_j \leftarrow 0$;
            else
                Calculate $p(\gamma \mid y, \mu_{-(i,j)}, \tau, \phi)$ for $\gamma = \gamma^i, \gamma^j, \gamma^{ij}$ using Lemma S2;
                Draw $\gamma \in \{\gamma^i, \gamma^j, \gamma^{ij}\}$ with probability $p(\gamma \mid y, \mu_{-(i,j)}, \tau, \phi)$;
                if $\gamma = \gamma^i$ then
                    Set $\mu_i \leftarrow s$, $\mu_j \leftarrow 0$;
                else if $\gamma = \gamma^j$ then
                    Set $\mu_i \leftarrow 0$, $\mu_j \leftarrow s$;
                else
                    Draw $u \sim N_{(0,s)}(\beta_{i,j}, (\phi\Lambda_{i,j})^{-1})$ by Lemma S1;
                    Set $\mu_i \leftarrow u$, $\mu_j \leftarrow s - u$;
    Set $\mu^{(t)} \leftarrow \mu$;
    Draw $\phi^{(t)}$ from $p(\phi \mid y, \mu^{(k)}, \tau^{(k-1)})$ by (8);
    Set $\tau \leftarrow \tau^{(t-1)}$;
    for $i = 1, \ldots, n_\tau$ do
        Propose $\tau^*$ by $\log \tau^* = \log \tau + N(0, \eta)$;
        Calculate acceptance probability $\rho(\tau, \tau^*)$ by (S1);
        Set $\tau \leftarrow \tau^*$ with probability $\rho(\tau, \tau^*)$;
    Set $\tau^{(t)} \leftarrow \tau$;
    Draw $w^{(t)}$ from the conditional distribution $p(w \mid y, \mu^t, \tau^t, \phi^t)$ given in (9);
    Calculate $\tilde{Y}^{(0,t)} \leftarrow \tilde{X} w^{(t)}$ ;
    Set $\widehat{\mathrm{ATT}}^{(t)} \leftarrow \mathrm{mean}(\tilde{Y}^{(1)} - \tilde{Y}^{(0,t)})$;
Set $\widehat{\mathrm{ATT}} \leftarrow n^{-1} \sum_{t=1}^{n} \widehat{\mathrm{ATT}}^{(t)}$;
Output: Samples $(\mu^{(t)}, \tau^{(t)}, \phi^{(t)}, w^{(t)})_{t=1}^{n}$, ATT estimate $\widehat{\mathrm{ATT}}$.
```

update $\tau^{(t)}$ by one Metropolis–Hastings step. In this case, the iterative complex factorization algorithm proposed by Zhou & Guan (2019) can be used to efficiently solve the linear system $V_{\gamma,\tau}^{-1} z$.

**Remark 2.** In Algorithm 1, we draw $w^{(t)}$ from the conditional distribution $p(w \mid y, \mu^t, \tau^t, \phi^t)$ so that the samples $(w^{(t)})_{t=1}^n$ can be used to approximate the posterior distribution of $w$. However, if the objective is to compute the posterior mean estimate for ATT, one can further reduce the variance of the estimator by directly setting $w^{(t)}$ to its conditional posterior mean as in (10).

# S2 Constrained Least-squares Estimators

In this section, we review some known results about the least-squares estimator under a linear constraint (Amemiya 1985). We present the results in the general form although for our analysis, we only need to consider the constraint $\mathbf{1}^\top\mu = 1$, which is part of the simplex constraint.

**Theorem S1.** *Let $y \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n\times p}$, $Q \in \mathbb{R}^{p\times k}$ for some $k < p$, and $v \in \mathbb{R}^k$. Assume that $Z^\top Z$ is invertible. Let $\mathcal{U} = \{\mu \in \mathbb{R}^p \colon Q^\top\mu = v\}$. Define the unconstrained least-squares estimator $\hat{\mu}$ and the linearly constrained least-squares estimator $\check{\mu} \in \mathcal{U}$ by*

$$\hat{\mu} = \operatorname*{argmin}_{\mu\in\mathbb{R}^p} \|y - Z\mu\|_2^2, \quad \check{\mu} = \operatorname*{argmin}_{\mu\in\mathcal{U}} \|y - Z\mu\|_2^2.$$

*Let $R \in \mathbb{R}^{p\times(p-k)}$ be any matrix such that $R^\top Q = 0$ and $[Q\ R]$ is invertible. Then, the following results hold.*

*(i) $\check{\mu}$ can be expressed by*

$$\begin{aligned}\check{\mu} &= \hat{\mu} - (Z^\top Z)^{-1}Q\left[Q^\top(Z^\top Z)^{-1}Q\right]^{-1}(Q^\top\hat{\mu} - v)\\ &= R(R^\top Z^\top ZR)^{-1}R^\top Z^\top y + \left[I - R(R^\top Z^\top ZR)^{-1}R^\top Z^\top Z\right]Q(Q^\top Q)^{-1}v.\end{aligned}$$

*(ii) For any $\mu \in \mathcal{U}$, we have $\|Z(\hat{\mu} - \mu)\|^2 = \|Z(\hat{\mu} - \check{\mu})\|^2 + \|Z(\check{\mu} - \mu)\|^2$. Hence,*

$$\|y - Z\mu\|_2^2 = \|y - Z\hat{\mu}\|_2^2 + \|Z(\hat{\mu} - \check{\mu})\|_2^2 + \|Z(\check{\mu} - \mu)\|_2^2.$$

*(iii) If $y = Z\mu^* + \epsilon$ for some $\mu^* \in \mathcal{U}$, then*

$$\check{\mu} = \mu^* + R(R^\top Z^\top ZR)^{-1}R^\top Z^\top\epsilon.$$

*Proof.* The claimed expressions for $\check{\mu}$ given in part (i) are proved in Chapters 1.4.1 and 1.4.2 of Amemiya (1985). Part (iii) is proved in Chapter 1.4.3 of Amemiya (1985). To prove part (ii), we use part (i) to get that, for any $w$ such that $Q^\top w = 0$,

$$w^\top Z^\top Z(\check{\mu} - \hat{\mu}) = -w^\top Q \left[Q^\top (Z^\top Z)^{-1} Q\right]^{-1} (Q^\top \hat{\mu} - v) = 0.$$

Since $Q^\top(\check{\mu} - \mu) = v - v = 0$ for any $\mu \in \mathcal{U}$, this implies that $\|Z(\hat{\mu} - \mu)\|^2 = \|Z(\hat{\mu} - \check{\mu})\|^2 + \|Z(\check{\mu} - \mu)\|^2$. The claimed decomposition for $\|y - Z\mu\|_2^2$ then follows from that $Z\hat{\mu}$ is the orthogonal projection of $y$ onto the column space of $Z$. □

Next, we prove some linear algebra results that will be useful for analyzing the constrained least-squares estimator.

**Lemma S3.** *Let $Q \in \mathbb{R}^{p\times k}$ and $R \in \mathbb{R}^{p\times(p-k)}$ be such that $R^\top Q = 0$ and $[Q\ R]$ is invertible. For any positive definite matrix $A \in \mathbb{R}^{p\times p}$,*

$$Q(Q^\top A^{-1} Q)^{-1} Q^\top = A - AR(R^\top AR)^{-1} R^\top A, \tag{S3}$$

$$\lambda_{\max}(R(R^\top AR)^{-1}R^\top) \le \lambda_{\max}(A^{-1}). \tag{S4}$$

*Proof.* The first identity is well known in the statistical literature; see, e.g., Smyth & Verbyla (1996). For completeness, we provide a proof here. Since $A$ is positive definite, it has a real-valued square root $A^{1/2}$. Consider the block matrix $B = [A^{-1/2}Q \quad A^{1/2}R] \in \mathbb{R}^{p\times p}$, where the two component matrices are orthogonal due to the assumption $R^\top Q = 0$. Hence, the orthogonal projection on the column space of $B$ can be written as $P_1 + P_2$ where

$$P_1 = A^{-1/2} Q \left(Q^\top A^{-1} Q\right)^{-1} Q A^{-1/2},$$

$$P_2 = A^{1/2} R \left(R^\top A R\right)^{-1} R A^{1/2}.$$

Since $\operatorname{rank}(B) = \operatorname{rank}(A^{-1/2}Q) + \operatorname{rank}(A^{1/2}R)$ and $[Q\ R]$ is invertible, the matrix $B$ also has full rank. It follows that $P_1 + P_2 = I$, which yields (S3).

To prove (S4), it suffices to note that for any vector $z \in \mathbb{R}^p$,

$$z^\top R(R^\top AR)^{-1}R^\top z = z^\top A^{-1/2}P_2A^{-1/2}z \leq \|A^{-1/2}z\|^2 \leq \lambda_{\max}(A^{-1})\|z\|^2,$$

where the first equality follows from the fact that $P_2$ is an orthogonal projection matrix. $\square$

# S3 Proofs for Section 3

## S3.1 Proof of Lemma 1

*Proof.* It suffices to prove that under Assumption (A1), for any $\gamma, \gamma' \in \mathbb{S}_L$ such that $\emptyset \neq \gamma \subseteq \gamma'$, we have

$$\frac{|\gamma'|}{M\underline{\lambda}} \geq \mathbf{1}^\top (X_{\gamma'}^\top X_{\gamma'})^{-1}\mathbf{1} \geq \mathbf{1}^\top (X_{\gamma}^\top X_{\gamma})^{-1}\mathbf{1} \geq \frac{1}{M}, \tag{S5}$$

$$(M\underline{\lambda})^{|\gamma'|-|\gamma|} \leq \frac{\det(X_{\gamma'}^\top X_{\gamma'})}{\det(X_{\gamma}^\top X_{\gamma})} \leq M^{|\gamma'|-|\gamma|}. \tag{S6}$$

Consider (S5) first. Let $S = X_{\gamma'\setminus\gamma}^\top (I - H_\gamma) X_{\gamma'\setminus\gamma}$. Since $X_{\gamma'} = [X_\gamma \; X_{\gamma'\setminus\gamma}]$, applying the block matrix inversion formula shows that for any $z^\top = [z_1^\top \; z_2^\top]$,

$$z^\top (X_{\gamma'}^\top X_{\gamma'})^{-1} z = z_1^\top (X_\gamma^\top X_\gamma)^{-1} z_1 + \tilde{z}^\top S^{-1} \tilde{z} \geq z_1^\top (X_\gamma^\top X_\gamma)^{-1} z_1,$$

where $\tilde{z} = z_2 - X_{\gamma'\setminus\gamma}^\top X_\gamma (X_\gamma^\top X_\gamma)^{-1} z_1$. In particular, $\mathbf{1}^\top (X_{\gamma'}^\top X_{\gamma'})^{-1}\mathbf{1} \geq \mathbf{1}^\top (X_\gamma^\top X_\gamma)^{-1}\mathbf{1}$. When $\gamma$ only contains one covariate, Assumption (A1) implies that $X_\gamma^\top X_\gamma = M$, which yields the asserted lower bound. For the asserted upper bound, by the min-max theorem for eigenvalues,

$$\mathbf{1}^\top (X_{\gamma'}^\top X_{\gamma'})^{-1}\mathbf{1} \leq \|\mathbf{1}\|_2^2 \, \lambda_{\max}\left((X_{\gamma'}^\top X_{\gamma'})^{-1}\right) = \frac{|\gamma'|}{\lambda_{\min}(X_{\gamma'}^\top X_{\gamma'})} \leq \frac{|\gamma'|}{M\underline{\lambda}},$$

where the last step follows from Assumption (A1).

To prove (S6), we first assume $|\gamma'| - |\gamma| = 1$ and let $\gamma' \setminus \gamma = \{j\}$. Applying the block matrix inversion formula, we obtain that

$$\det(X_{\gamma'}^\top X_{\gamma'}) = \det(X_\gamma^\top X_\gamma)\det(S).$$

Since $S = X_j^\top (I - H_\gamma) X_j$ and $I - H_\gamma$ is a projection matrix, we have $S \leq M$ under Assumption (A1). Moreover, by Lemma S4 below, we have $S \geq M\underline{\lambda}$. Hence,

$$M\underline{\lambda} \leq \frac{\det(X_{\gamma'}^\top X_{\gamma'})}{\det(X_\gamma^\top X_\gamma)} \leq M.$$

To conclude the proof, observe that for any $\gamma \subseteq \gamma'$, $\det(X_{\gamma'}^\top X_{\gamma'})/\det(X_\gamma^\top X_\gamma)$ can be written as the product of $(|\gamma'| - |\gamma|)$ ratios, where the $k$-th ratio is $\det(X_{\gamma_{k+1}}^\top X_{\gamma_{k+1}})/\det(X_{\gamma_k}^\top X_{\gamma_k})$ for some $\gamma_k, \gamma_{k+1}$ such that $\gamma_{k+1} \setminus \gamma_k = \{j\}$ for some $j$. □

**Lemma S4.** *Under Assumption (A1), for any $\gamma, \gamma' \in \mathbb{S}_L$ such that $\gamma \subseteq \gamma'$,*

$$\lambda_{\min}(X_{\gamma'\setminus\gamma}^\top (I - H_\gamma) X_{\gamma'\setminus\gamma}) \geq M\underline{\lambda}.$$

*Proof.* See Yang et al. (2016, pp. 2522). □

## S3.2 Proof of Theorem 1

We divide the proof into multiple parts. First, in Section S2, we construct the events that happen with probability at least $1 - c_1 N^{-c_2}$ under the true data-generating probability measure $\mathbb{P}^*$, where $c_1, c_2 > 0$ are some universal constants. In Section S3.2.2, we derive an explicit expression for $p(\gamma, \phi \mid y)$ by using the property of the constrained least-squares estimators. In Section S3.2.3, we find a lower bound on $p(\gamma^*, \phi \mid y)$ by analyzing both the unconstrained and constrained least-squares estimators under $\gamma^*$, In Section S3.2.4, we prove Theorem 1 by integrating over $\phi$. Throughout our proof, we use

$$H_\gamma := X_\gamma (X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top$$

to denote the projection matrix onto the column space of $X_\gamma$.

### S3.2.1 High-probability Events

We begin by constructing the events that are essential to our non-asymptotic analysis of the posterior distribution. They enable us to control the behavior of the error vector $\epsilon$. The proof is based on standard concentration inequalities for normal distributions.

**Lemma S5.** *Suppose Assumption (A1) holds, $M \geq \log N$, $N \geq 2$ and $L \geq 3$. Then,*

$$\mathbb{P}^*(E_1 \cap E_2 \cap E_3 \cap E_4) \geq 1 - c_1 N^{-c_2},$$

*for some universal constants $c_1, c_2 > 0$, where*

$$\begin{aligned}
E_1 &= \left\{ \max_{\substack{(\gamma_1,\gamma_2)\in\mathbb{S}_L\times\mathbb{S}_L \\ \gamma_2\subset\gamma_1}} \frac{\epsilon^\top (H_{\gamma_1} - H_{\gamma_2})\epsilon}{\ell_1 - \ell_2} \leq 3\sigma^2 L \log N \right\}, \\
E_2 &= \left\{ \frac{1}{M} \max_{1\leq j\leq N} |\epsilon^\top X_j| \leq \frac{3\sigma\sqrt{\log N}}{\sqrt{M}} \right\}, \\
E_3 &= \left\{ \epsilon^\top H_{\gamma^*}\epsilon \leq \sigma^2 \ell^* \log N \right\}, \\
E_4 &= \left\{ \frac{7}{8} \leq \frac{\epsilon^\top \epsilon}{\sigma^2 M} \leq \frac{5}{4} \right\}.
\end{aligned}$$

*Proof.* Given any $\gamma_2 \subset \gamma_1$, we can write $H_{\gamma_1} - H_{\gamma_2} = \sum_k (H_{\gamma_k} - H_{\gamma_{k+1}})$ where $\gamma_k = \gamma_{k+1} \cup \{j\}$ for some $j \notin \gamma_{k+1}$. Hence, $E_1 = \tilde{E}_1$, where

$$\tilde{E}_1 = \left\{ \max_{\substack{\gamma\in\mathbb{S}_{L-1} \\ j\in[N], j\notin\gamma}} \epsilon^\top (H_{\gamma\cup\{j\}} - H_\gamma)\epsilon \leq 3\sigma^2 L \log N \right\}.$$

The rank of $H_{\gamma\cup\{j\}} - H_\gamma$ is 1, and thus $\sigma^{-2}\epsilon^\top(H_{\gamma\cup\{j\}} - H_\gamma)\epsilon \sim \chi^2_1$. The concentration inequality for normal distribution yields

$$\mathbb{P}^*\left\{\epsilon^\top(H_{\gamma\cup\{j\}} - H_\gamma)\epsilon \geq 2\sigma^2 t\right\} \leq 2e^{-t}, \quad \forall\, t \geq 0.$$

Choose $t = (3/2)L\log N$. Applying the union bound and using $L \geq 3$, we get

$$\begin{aligned}\mathbb{P}^*(E_1) = \mathbb{P}^*(\tilde{E}_1) &\geq 1 - \sum_{\substack{\gamma\in\mathbb{S}_{L-1}\\ j\in[N], j\notin\gamma}} \mathbb{P}^*\left\{\epsilon^\top(H_{\gamma\cup\{j\}} - H_\gamma)\epsilon \geq 3\sigma^2 L\log N\right\}\\ &\geq 1 - 2N^{L+1}\cdot N^{-\frac{3L}{2}}\\ &\geq 1 - 2N^{-\frac{1}{2}}.\end{aligned}$$

Consider $E_2$. Define

$$d(\epsilon) = \max_{1\leq j\leq N} |\epsilon^\top X_j|.$$

A routine argument using (A1) shows that $d(\epsilon)$ is Lipschitz continuous in $\epsilon$ with Lipschitz constant $\sqrt{M}$. Since $\epsilon \sim N(0, \sigma^2 I_M)$ under $\mathbb{P}^*$, by the concentration inequality for Lipschitz functions of Gaussian vectors, we have

$$\mathbb{P}^*\left(|d(\epsilon) - \mathbb{E}^*[d(\epsilon)]| \geq t\sigma\sqrt{M}\right) \leq 2e^{-t^2/2}. \tag{S7}$$

Since each $X_j^\top\epsilon \sim N(0, \sigma^2 M)$, a standard result for the maximum of (possibly dependent) Gaussian random variables yields that

$$\mathbb{E}^*\left[d(\epsilon)\right] \leq \sigma\sqrt{2M\log(2N)} \leq 2\sigma\sqrt{2M\log N}, \tag{S8}$$

whenever $N > 1$. Letting $t = \sqrt{\log N}$ in (S7) and using (S8), we obtain that

$$
\begin{aligned}
\mathbb{P}^*(E_2) &= 1 - \mathbb{P}^*\left(d(\epsilon) > 3\sigma\sqrt{M \log N}\right) \\
&\geq 1 - \mathbb{P}^*\left(|d(\epsilon) - \mathbb{E}^*[d(\epsilon)]| > \sigma\sqrt{M \log N}\right) \\
&\geq 1 - 2N^{-1/2}
\end{aligned}
$$

An application of the union bound concludes the proof.

For $E_3$, the standard concentration inequality for chi-squared distributions yields that $\mathbb{P}^*(E_3) \geq 1 - c_1 N^{-c_2 \ell^*}$ for some universal constants $c_1, c_2 > 0$ (Laurent & Massart 2000). Similarly, one can also verify that $\mathbb{P}^*(E_4) \geq 1 - c_1 e^{-c_2 M}$ for some $c_1, c_2 > 0$. □

### S3.2.2 Posterior Density of $(\gamma, \phi)$

The presence of the simplex constraint makes the technical analysis substantially different from that for unconstrained variable selection models. The main challenge is to bound the marginal likelihood

$$p(y \mid \gamma, \phi) = \mathbb{E}_{\mu_\gamma \sim \mathrm{Dir}(1)}\left[p(y \mid \gamma, \mu_\gamma, \phi)\right].$$

To address this, we begin by deriving an equivalent expression for $p(y \mid \gamma, \mu_\gamma, \phi)$ using Theorem S1. For each $\gamma \in \mathbb{S}_L$, define the OLS estimator for the model $\gamma$ by

$$\hat{\mu}_\gamma = (X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top y.$$

Let $\check{\mu}_\gamma$ denote the least-squares estimator under the constraint $\mathbf{1}^\top \mu_\gamma = 1$. By Theorem S1,

$$\check{\mu}_\gamma = \hat{\mu}_\gamma + \frac{1 - \mathbf{1}^\top \hat{\mu}_\gamma}{\mathbf{1}^\top (X_\gamma^\top X_\gamma)^{-1} \mathbf{1}} (X_\gamma^\top X_\gamma)^{-1} \mathbf{1}.$$

Further, for any $u \in \Delta^{|\gamma|-1}$, we have

$$\left\| y - X_\gamma u \right\|_2^2 = \left\| y - X_\gamma \hat{\mu}_\gamma \right\|_2^2 + \left\| X_\gamma (u - \check{\mu}_\gamma) \right\|_2^2 + b_\gamma, \tag{S9}$$

where

$$b_\gamma = \left\| X_\gamma (\hat{\mu}_\gamma - \check{\mu}_\gamma) \right\|_2^2 = \frac{(1 - \mathbb{1}^\top \hat{\mu}_\gamma)^2}{\mathbb{1}^\top (X_\gamma^\top X_\gamma)^{-1} \mathbb{1}}. \tag{S10}$$

Equation (S9) decomposes the residual sum of squares for any $u$ satisfying the simplex constraint into three parts. The first term, $\|y - X_\gamma \hat{\mu}_\gamma\|_2^2$, is the residual sum of squares for the least-squares estimator in the unconstrained setting. The second term, $\|X_\gamma (u - \check{\mu}_\gamma)\|_2^2$, measures the deviation of $u$ from the constrained least-squares estimator $\check{\mu}_\gamma$. Note that $X_\gamma \check{\mu}_\gamma$ is the projection of $X_\gamma \hat{\mu}_\gamma$ onto the affine space $\{X_\gamma u \colon \mathbb{1}^\top u = 1\}$; in particular, we always have $\mathbb{1}^\top \check{\mu}_\gamma = 1$. The last quantity, $b_\gamma \geq 0$, measures the deviation of $\check{\mu}_\gamma$ from $\hat{\mu}_\gamma$, which captures how well the model satisfies the simplex constraint. When $\hat{\mu}_\gamma \in \Delta^{|\gamma|-1}$, we have $b_\gamma = 0$.

For each $\gamma$, we also define two matrices $D_\gamma \in \mathbb{R}^{|\gamma| \times (|\gamma|-1)}$ and $\Pi_\gamma \in \mathbb{R}^{(|\gamma|-1) \times |\gamma|}$ by

$$D_\gamma = \begin{bmatrix} I_{|\gamma|-1} & (-\mathbb{1}) \end{bmatrix}^\top, \quad \Pi_\gamma = \begin{bmatrix} I_{|\gamma|-1} & 0 \end{bmatrix}. \tag{S11}$$

Clearly $D_\gamma$ is a full-rank matrix whose column space is orthogonal to the one vector $\mathbb{1}$, and $\Pi_\gamma u$ yields the subvector of $u$ containing the first $|\gamma| - 1$ entries. Since the dimension of $D_\gamma$ or $\Pi_\gamma$ is always clear from context, we will omit the subscript and simply write $D$ and $\Pi$.

In Proposition S1, we show that the decomposition given in (S9) enables us to compare the marginal likelihood $p(y \mid \gamma, \phi)$ with that in the unconstrained setting. The expectation with respect to $u \sim \mathrm{Dir}(\mathbb{1})$ is converted to the probability mass (up to a normalizing constant) of a $(|\gamma| - 1)$-dimensional Gaussian distribution centered at $\Pi \check{\mu}_\gamma$.

**Proposition S1.** *Let $D, \Pi$ beas defined in (S11). Define*

$$f(\ell) = \theta^{\ell}(1-\theta)^{N-\ell}(2\pi)^{(\ell-1)/2}.$$

*For any $\gamma \in \mathbb{S}_L$ and $\phi > 0$, the posterior density $p(\gamma, \phi \mid y)$ given in (14) can be expressed as*

$$p(\gamma, \phi \mid y) = C f(|\gamma|)\, \phi^{(M+\kappa_1)/2}\, e^{-\frac{\phi}{2}\left(\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 + b_\gamma + \kappa_2\right)}\, \mathbb{P}\left(\tilde{U}_{\gamma,\phi} \in \tilde{\Delta}^{|\gamma|-1}\right),$$

*where $C > 0$ is the unknown normalizing constant, $\tilde{U}_{\gamma,\phi} \sim \mathcal{N}(\Pi\check{\mu}_\gamma, (\phi D^\top X_\gamma^\top X_\gamma D)^{-1})$, and*

$$\tilde{\Delta}^{\ell} = \left\{\tilde{u} \in \mathbb{R}^{\ell} \colon \tilde{u}_i \geq 0 \text{ for each } i, \text{ and } \textstyle\sum_{i=1}^{\ell} \tilde{u}_i \leq 1\right\}.$$

*Proof.* Observe that for any vector $w$ with $\mathbf{1}^\top w = 0$, we have $w = D\Pi w$. Hence, for any $u \in \Delta^{|\gamma|-1}$,

$$\|X_\gamma(u - \check{\mu}_\gamma)\|_2^2 = \|X_\gamma D\, \Pi(u - \check{\mu}_\gamma)\|_2^2 = \|X_\gamma D(\tilde{u} - \Pi\, \check{\mu}_\gamma)\|_2^2. \tag{S12}$$

where $\tilde{u} = \Pi u \in \tilde{\Delta}^{|\gamma|-1}$. By equations (S9) and (S12) and letting $u = \mu_\gamma$, we find that

$$\begin{aligned}
& p(y \mid \gamma, u, \phi) \\
& \propto \phi^{M/2} \exp\left\{-\frac{\phi}{2}\|y - X_\gamma u\|_2^2\right\} \\
& = \phi^{M/2} \exp\left\{-\frac{\phi}{2}\left(\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 + b_\gamma + \|X_\gamma D(\tilde{u} - \Pi\check{\mu}_\gamma)\|_2^2\right)\right\} \\
& = \frac{\phi^{(M-|\gamma|+1)/2}(2\pi)^{(|\gamma|-1)/2}}{\sqrt{\det(D^\top X_\gamma^\top X_\gamma D)}} e^{-\frac{\phi}{2}\left(\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 + b_\gamma\right)}\, \mathcal{N}(\tilde{u};\, \Pi\check{\mu}_\gamma,\, (\phi D^\top X_\gamma^\top X_\gamma D)^{-1}),
\end{aligned} \tag{S13}$$

where $\tilde{u} = \Pi u$ and $\mathcal{N}(\cdot; m, V)$ denotes the density function of the normal distribution with mean $v$ and covariance matrix $V$. We claim that

$$\det(D^\top X_\gamma^\top X_\gamma D) = \det(X_\gamma^\top X_\gamma)\mathbf{1}^\top(X_\gamma^\top X_\gamma)^{-1}\mathbf{1}. \tag{S14}$$

To prove this, define $\tilde{D} = [D\ 1]$. A routine calculation yields $\det(\tilde{D}\tilde{D}^\top) = |\gamma|^2$. Hence,

$$\det(X_\gamma^\top X_\gamma) = \frac{1}{|\gamma|^2} \det(\tilde{D}^\top X_\gamma^\top X_\gamma \tilde{D})$$
$$= \frac{1}{|\gamma|^2} \det(D^\top X_\gamma^\top X_\gamma D) \det\left(1^\top X_\gamma^\top X_\gamma 1 - 1^\top X_\gamma^\top X_\gamma D \left[D^\top X_\gamma^\top X_\gamma D\right]^{-1} D^\top X_\gamma^\top X_\gamma 1\right),$$

where the second step follows from the standard block matrix determinant formula. By the identity (S3) in Lemma S3, the matrix involved in the second determinant term can be simplified to

$$1^\top X_\gamma^\top X_\gamma 1 - 1^\top X_\gamma^\top X_\gamma D \left[D^\top X_\gamma^\top X_\gamma D\right]^{-1} D^\top X_\gamma^\top X_\gamma 1 = \frac{|\gamma|^2}{1^\top (X_\gamma^\top X_\gamma)^{-1} 1},$$

which yields (S14).

Using (S13), (S14) and the prior specification given in (12) and (13), we get

$$p(\gamma, u, \phi \mid y) = Cp(\gamma)p(\phi \mid \gamma)p(u \mid \gamma)p(y \mid \gamma, u, \phi)$$
$$= Cf(|\gamma|)\, \phi^{(M+\kappa_1)/2}\, e^{-\frac{\phi}{2}\left(\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 + b_\gamma + \kappa_2\right)} \mathcal{N}(\tilde{u};\, \Pi\check{\mu}_\gamma,\, (\phi D^\top X_\gamma^\top X_\gamma D)^{-1}) \tag{S15}$$

Since integrating in $u$ over $\Delta^{|\gamma|-1}$ is the same as integrating in $\tilde{u} = \Pi u$ over $\tilde{\Delta}^{|\gamma|-1}$, we get

$$p(\gamma, \phi \mid y) = \int_{\tilde{\Delta}^{|\gamma|-1}} p(\gamma, u, \phi \mid y) \mathrm{d}\tilde{u},$$

which yields the claimed identity. □

#### S3.2.3 Lower Bound on the Posterior Density of $(\gamma^*, \phi)$

Next, we consider the true model $\gamma^*$. By Proposition S1, to lower bound $p(\gamma^*, \phi \mid y)$, we need to find an upper bound on $b_\gamma^*$ and a lower bound on $\mathbb{P}(\tilde{U}_{\gamma^*,\phi} \in \tilde{\Delta}^{\ell^*-1})$ where

$\tilde{U}_{\gamma^*,\phi} \sim \mathcal{N}(\Pi\, \check{\mu}_{\gamma^*}, (\phi D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1})$. Consider $b_\gamma^*$ first.

**Lemma S6.** *Suppose $1^\top \mu_{\gamma^*}^* = 1$. On the event $E_3$ defined in Lemma S5, $b_{\gamma^*} \leq \sigma^2 \ell^* \log N$.*

*Proof.* For the true model $\gamma^*$, $\hat{\mu}_{\gamma^*}$ can be written as

$$\hat{\mu}_{\gamma^*} = (X_{\gamma^*}^\top X_{\gamma^*})^{-1} X_{\gamma^*}^\top (X_{\gamma^*} \mu_{\gamma^*}^* + \epsilon) = \mu_{\gamma^*}^* + (X_{\gamma^*}^\top X_{\gamma^*})^{-1} X_{\gamma^*}^\top \epsilon.$$

Using $1^\top \mu_{\gamma^*}^* = 1$ we find that

$$1 - 1^\top \hat{\mu}_{\gamma^*} = -1^\top (X_{\gamma^*}^\top X_{\gamma^*})^{-1} X_{\gamma^*}^\top \epsilon.$$

By Theorem S1 and the definition of $b_\gamma$ given in (S10),

$$b_{\gamma^*} = \left\| X_{\gamma^*} (\hat{\mu}_{\gamma^*} - \check{\mu}_{\gamma^*}) \right\|_2^2 \leq \left\| X_{\gamma^*} (\hat{\mu}_{\gamma^*} - \mu_{\gamma^*}^*) \right\|_2^2 = \epsilon^\top H_{\gamma^*} \epsilon \leq \sigma^2 \ell^* \log N,$$

on the event $E_3$. □

Bounding $\mathbb{P}(\tilde{U}_{\gamma^*,\phi} \in \tilde{\Delta}^{\ell^*-1})$ is more difficult, and we divide the argument into two lemmas.

**Lemma S7.** *Suppose Assumptions (A1) and (A4) hold. On the event $E_2$ defined in Lemma S5, $\check{\mu}_{\gamma^*} \in \Delta^{\ell^*-1}$, and the smallest element of $\check{\mu}_{\gamma^*}$, denoted by $\check{\mu}_{\min}^*$, satisfies*

$$\check{\mu}_{\min}^* \geq (c_\mu - 3) \frac{\sigma \sqrt{L \log N}}{\underline{\lambda} \sqrt{M}}. \tag{S16}$$

*Proof.* Since $1^\top \check{\mu}_\gamma = 1$, we only need to prove the lower bound (S16). By part (iii) of Theorem S1,

$$\check{\mu}_{\gamma^*} = \mu_{\gamma^*}^* + D(D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1} D^\top X_{\gamma^*}^\top \epsilon.$$

Let $e_j$ denote the vector whose $j$-th entry is one and other entries are all zero. By the inequality (S4) given in Lemma S3, on the event $E_2$,

$$\begin{aligned}
\|\check{\mu}_{\gamma^*} - \mu^*_{\gamma^*}\|_\infty &= \left\| D(D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1} D^\top X_{\gamma^*}^\top \epsilon \right\|_\infty \\
&= \max_j \left| e_j^\top D(D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1} D^\top X_{\gamma^*}^\top \epsilon \right| \\
&\leq \lambda_{\max}\left( D(D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1} D^\top \right) \left( \max_j \|e_j\|_2 \right) \|X_{\gamma^*}^\top \epsilon\|_2 \\
&\leq \lambda_{\max}\left( (X_{\gamma^*}^\top X_{\gamma^*})^{-1} \right) \sqrt{\ell^*} \|X_{\gamma^*}^\top \epsilon\|_\infty \\
&\leq \frac{3\sigma\sqrt{\ell^* \log N}}{\underline{\lambda}\sqrt{M}}, \qquad \text{(S17)}
\end{aligned}$$

where the last step follows from Assumption (A1) and the definition of $E_2$. By Assumption (A4),

$$\|\check{\mu}_{\gamma^*} - \mu^*_{\gamma^*}\|_\infty \leq \frac{3}{c_\mu} \min_{j \in \gamma^*} |\mu^*_j|,$$

Hence,

$$\check{\mu}^*_{\min} \geq \min_{j \in \gamma^*} |\mu^*_j| - \|\check{\mu}_{\gamma^*} - \mu^*_{\gamma^*}\|_\infty \geq \left(1 - \frac{3}{c_\mu}\right) \min_{j \in \gamma^*} |\mu^*_j|,$$

from which the claimed bound follows. □

**Lemma S8.** *Suppose Assumptions (A1) and (A4) hold with $c_\mu > 6$. Let*

$$r = \frac{c_\mu - 3}{\sqrt{\ell^* - 1}} \frac{\sigma\sqrt{L \log N}}{\underline{\lambda}\sqrt{M}}.$$

*On the event $E_2$ defined in Lemma S5, we have*

$$\mathcal{B}_r(\Pi\, \check{\mu}_{\gamma^*}) \coloneqq \{\tilde{u} \in \mathbb{R}^{\ell^* - 1} \colon \|\tilde{u} - \Pi\, \check{\mu}_{\gamma^*}\|_2 \leq r\} \subset \tilde{\Delta}^{\ell^* - 1}.$$

*Further, for $\tilde{U}_{\gamma^*,\phi} \sim \mathcal{N}(\Pi\,\check{\mu}_{\gamma^*}, (\phi D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1})$, we have*

$$\mathbb{P}\left(\tilde{U}_{\gamma^*,\phi} \in \tilde{\Delta}^{\ell^*-1}\right) \geq \mathbb{P}(\tilde{U}_{\gamma^*,\phi} \in \mathcal{B}_r(\Pi\,\check{\mu}_{\gamma^*})) \geq 1 - e^{1/2} \exp\left(-\frac{4\phi\,\sigma^2 L \log N}{(\ell^*)^2}\right).$$

*Proof.* Since $\Pi\,\check{\mu}_{\gamma^*}$ is the subvector of $\check{\mu}_{\gamma^*}$, Lemma S7 implies that $\Pi\,\check{\mu}_{\gamma^*} \in \tilde{\Delta}^{\ell^*-1}$ when $c_\mu > 3$. We first find the minimum distance from $\Pi\,\check{\mu}_{\gamma^*}$ to the hyperplanes of $\tilde{\Delta}^{\ell^*-1}$, which are determined by $\{\tilde{u} \colon \tilde{u}_i = 0\}$ for $i = 1, \dots, \ell^* - 1$ and $\{\tilde{u} \colon \sum_{i=1}^{\ell^*-1} \tilde{u}_i = 1\}$. The distance from $\Pi\,\check{\mu}_{\gamma^*}$ to $\{\tilde{u} : \tilde{u}_i = 0\}$ equals the $i$-th entry of $\Pi\,\check{\mu}_{\gamma^*}$, and the distance to $\{\tilde{u} \colon \sum_{i=1}^{\ell^*-1} \tilde{u}_i = 1\}$ is $(1 - \mathbf{1}^\top \Pi\,\check{\mu}_{\gamma^*})/\sqrt{\ell^* - 1}$. Since $\mathbf{1}^\top \check{\mu}_{\gamma^*} = 1$, $1 - \mathbf{1}^\top \Pi\,\check{\mu}_{\gamma^*}$ equals the last entry of $\check{\mu}_{\gamma^*}$, which proves that the ball $\mathcal{B}_r(\Pi\,\check{\mu}_{\gamma^*})$ is contained in the simplex $\tilde{\Delta}^{\ell^*-1}$ whenever $r \leq \check{\mu}^*_{\min}/\sqrt{\ell^* - 1}$. By Lemma S7, this condition is satisfied for the given choice of $r$. Clearly, this implies that $\mathbb{P}(\tilde{U} \in \tilde{\Delta}^{\ell^*-1}) \geq \mathbb{P}(\tilde{U}_{\gamma^*,\phi} \in \mathcal{B}_r(\Pi\,\check{\mu}_{\gamma^*}))$.

By a well-known result for Gaussian concentration (Vershynin 2018, Theorem 5.1.4), for the random vector $\tilde{U} = \tilde{U}_{\gamma^*,\phi}$, we have

$$\mathbb{P}\left(\|\tilde{U} - \Pi\,\check{\mu}_{\gamma^*}\|_2 - \mathbb{E}\|\tilde{U} - \Pi\,\check{\mu}_{\gamma^*}\|_2 > t\right) \leq \exp\left\{-\frac{\phi t^2}{2}\lambda_{\min}(D^\top X_{\gamma^*}^\top X_{\gamma^*} D)\right\}, \quad \text{(S18)}$$

for every $t \geq 0$. A direct calculation using $D^\top D = I_{\ell^*} + \mathbf{1}\mathbf{1}^\top$ shows that $\lambda_{\max}(D^\top D) = \ell^*$ and $\lambda_{\min}(D^\top D) = 1$. Hence, under Assumption (A1),

$$\lambda_{\min}(D^\top X_{\gamma^*}^\top X_{\gamma^*} D) \geq \lambda_{\min}(X_{\gamma^*}^\top X_{\gamma^*})\lambda_{\min}(D^\top D) \geq M\underline{\lambda}. \quad \text{(S19)}$$

The expectation term on the left-hand side of (S18) then can be bounded by

$$\mathbb{E}\|\tilde{U} - \Pi\,\check{\mu}_{\gamma^*}\|_2 \leq \sqrt{\lambda_{\max}((\phi D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1})}\,\mathbb{E}\chi_{\ell^*-1} \leq \frac{\sqrt{\ell^*}}{\sqrt{\phi M \underline{\lambda}}}, \quad \text{(S20)}$$

where $\chi_{\ell^*-1}$ denotes a random variable following the $\chi$-distribution with $(\ell^* - 1)$ degrees of freedom. Applying (S18) with (S19) and (S20) yields

$$\begin{aligned}
\mathbb{P}\left(\tilde{U} \in \mathcal{B}_r(\Pi\,\check{\mu}_{\gamma^*})\right) &= 1 - \mathbb{P}\left(\|\tilde{U} - \Pi\,\check{\mu}_{\gamma^*}\|_2 - \mathbb{E}\|\tilde{U} - \Pi\,\check{\mu}_{\gamma^*}\|_2 > r - \mathbb{E}\|\tilde{U} - \Pi\,\check{\mu}_{\gamma^*}\|_2\right) \\
&\geq 1 - \mathbb{P}\left(\|\tilde{U} - \Pi\,\check{\mu}_{\gamma^*}\|_2 - \mathbb{E}\|\tilde{U} - \Pi\,\check{\mu}_{\gamma^*}\|_2 > r - \frac{\sqrt{\ell^*}}{\sqrt{\phi M \underline{\lambda}}}\right) \\
&\geq 1 - \exp\left\{-\frac{\phi M \underline{\lambda}}{2}\left(r - \frac{\sqrt{\ell^*}}{\sqrt{\phi M \underline{\lambda}}}\right)^2\right\}.
\end{aligned}$$

Using $2ab \leq ca^2 + c^{-1}b^2$ with $a = r$, $b = \sqrt{\ell^*}/\sqrt{\phi M \underline{\lambda}}$ and $c = \ell^*/(\ell^* + 1)$, we get

$$\begin{aligned}
\left(r - \frac{\sqrt{\ell^*}}{\sqrt{\phi M \underline{\lambda}}}\right)^2 &\geq r^2 + \frac{\ell^*}{\phi M \underline{\lambda}} - \frac{\ell^*}{\ell^* + 1} r^2 - \frac{\ell^* + 1}{\ell^*} \frac{\ell^*}{\phi M \underline{\lambda}} \\
&= \frac{r^2}{\ell^* + 1} - \frac{1}{\phi M \underline{\lambda}} \\
&= \frac{(\check{\mu}^*_{\min})^2}{(\ell^* - 1)(\ell^* + 1)} - \frac{1}{\phi M \underline{\lambda}} \\
&\geq (c_\mu - 3)^2 \frac{\sigma^2 L \log N}{(\ell^*)^2 M \underline{\lambda}^2} - \frac{1}{\phi M \underline{\lambda}} \\
&\geq \frac{9\sigma^2 L \log N}{(\ell^*)^2 M \underline{\lambda}} - \frac{1}{\phi M \underline{\lambda}},
\end{aligned}$$

where the last step follows from $c_\mu \geq 6$ and $\underline{\lambda} \leq 1$. Therefore,

$$\mathbb{P}\left(\tilde{U} \in \mathcal{B}_r(\Pi\,\check{\mu}_{\gamma^*})\right) \geq 1 - \exp\left\{-\frac{9\phi\sigma^2 L \log N}{2(\ell^*)^2} + \frac{1}{2}\right\}.$$

The stated bound thus follows. □

### S3.2.4 Proof of Theorem 1

*Proof of Theorem 1.* This follows from Propositions S2 and S3 that we prove below, where we further bound $b_{\gamma^*}$ using Lemma S6. □

**Proposition S2.** *For any $\gamma \in \mathbb{S}_L$ and $\phi > 0$,*

$$p(\gamma, \phi \mid y) \leq Cf(|\gamma|)\,\Gamma\left(\frac{M+\kappa_1}{2}\right)\left\{\frac{\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 + \kappa_2}{2}\right\}^{-(\kappa_1+M)/2},$$

*where the constant $C$ and function $f$ are as given in Proposition S1.*

*Proof.* Since $\mathbb{P}(\tilde{U}_{\gamma,\phi} \in \tilde{\Delta}^{|\gamma|-1}) \in [0,1]$ and $b_\gamma \geq 0$, by Proposition S1,

$$p(\gamma, \phi \mid y) \leq Cf(|\gamma|)\phi^{(M+\kappa_1)/2} e^{-\frac{\phi}{2}\left(\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 + \kappa_2\right)}.$$

Integrating over $\phi$ yields the claimed bound. □

**Proposition S3.** *Suppose Assumptions (A1) and (A5) hold with $c_\mu > 6$ and $c_M \geq 4$. On the event $E_2 \cap E_3 \cap E_4$ defined in Lemma S5, we have*

$$p(\gamma^* \mid y) \geq \frac{C}{3} f(\ell^*)\Gamma\left(\frac{M+\kappa_1}{2}\right)\left\{\frac{\|y - X_{\gamma^*}\hat{\mu}_{\gamma^*}\|_2^2 + b_{\gamma^*} + \kappa_2}{2}\right\}^{-(M+\kappa_1)/2},$$

*where the constant $C$ and function $f$ are as given in Proposition S1.*

*Proof.* Define

$$K = \frac{1}{2}\left(\kappa_2 + \|y - X_{\gamma^*}\hat{\mu}_{\gamma^*}\|_2^2 + b_{\gamma^*}\right), \quad \tilde{K} = \frac{4\sigma^2 L \log N}{(\ell^*)^2}. \tag{S21}$$

By Proposition S1 and Lemma S8, we have

$$p(\gamma^*, \phi \mid y) \geq Cf(\ell^*)\phi^{(M+\kappa_1)/2} e^{-K\phi}\left(1 - e^{1/2} e^{-\tilde{K}\phi}\right).$$

Integrating over $\phi$ yields

$$
\begin{aligned}
p(\gamma^* \mid y) &\geq Cf(\ell^*) \int \phi^{(M+\kappa_1)/2} e^{-K\phi} \left(1 - e^{1/2} e^{-\tilde{K}\phi}\right) \mathrm{d}\phi \\
&= Cf(\ell^*) \left[\int \phi^{(M+\kappa_1)/2} e^{-K\phi} \mathrm{d}\phi - e^{1/2} \int \phi^{(M+\kappa_1)/2} e^{-(K+\tilde{K})\phi} \mathrm{d}\phi\right] \\
&= Cf(\ell^*) \Gamma\left(\frac{M+\kappa_1}{2}\right) \left[K^{-(M+\kappa_1)/2} - e^{1/2}(K+\tilde{K})^{-(M+\kappa_1)/2}\right]. \qquad \text{(S22)}
\end{aligned}
$$

It remains to show that the term involving $K + \tilde{K}$ in (S22) is negligible. First, we bound $K$ and $\tilde{K}$. On the event $E_4$,

$$
\|y - X_{\gamma^*} \hat{\mu}_{\gamma^*}\|_2^2 = \epsilon^\top (I - H_{\gamma^*}) \epsilon \leq \epsilon^\top \epsilon \leq \frac{5}{4} \sigma^2 M.
$$

Using Assumption (A3), Lemma S6 and $M \geq 4L \log N$, we get

$$
2K = \kappa_2 + \|y - X_{\gamma^*} \hat{\mu}_{\gamma^*}\|_2^2 + b_{\gamma^*} \leq \frac{1}{2}\sigma^2 M + \frac{5}{4}\sigma^2 M + \sigma^2 \ell^* \log N \leq 2\sigma^2 M. \qquad \text{(S23)}
$$

Meanwhile, the conditions $M \geq 4L \log N$ and $\ell^* \leq \sqrt{L \log N}$ imply that

$$
4\sigma^2 \leq \tilde{K} \leq \sigma^2 M. \qquad \text{(S24)}
$$

Hence, we can use the inequality $\log(1+x) \geq x/(1+x)$ for $x > -1$ to get

$$
\begin{aligned}
\left(\frac{K+\tilde{K}}{K}\right)^{-(M+\kappa_1)/2} &\overset{\text{by (S23)}}{\leq} \left(1 + \frac{\tilde{K}}{\sigma^2 M}\right)^{-(M+\kappa_1)/2} \leq \exp\left\{-\frac{M\tilde{K}}{2(\sigma^2 M + \tilde{K})}\right\} \\
&\overset{\text{by (S24)}}{\leq} \exp\left(-\frac{\tilde{K}}{4\sigma^2}\right) \overset{\text{by (S24)}}{\leq} e^{-1}. \qquad \text{(S25)}
\end{aligned}
$$

Plugging this bound into (S22) yields

$$
\begin{aligned}
p(\gamma^* \mid y) &\geq Cf(\ell^*)\Gamma\left(\frac{M+\kappa_1}{2}\right)K^{-(M+\kappa_1)/2}\left\{1-e^{1/2}\left(\frac{K+\tilde{K}}{K}\right)^{-(M+\kappa_1)/2}\right\} \\
&\geq (1-e^{-1/2})Cf(\ell^*)\Gamma\left(\frac{M+\kappa_1}{2}\right)K^{-(M+\kappa_1)/2}.
\end{aligned}
$$

The claimed bound then follows by using $1-e^{-1/2} \geq 1/3$. □

## S3.3 Proof of Theorem 2

To prove Theorem 2, we first use Theorem 1 to identify the rate at which $p(\gamma \mid y)/p(\gamma^* \mid y)$ goes to zero for any single $\gamma \neq \gamma^*$. We consider two different cases separately: $\gamma^* \subset \gamma$ (overfitted) and $\gamma^* \not\subset \gamma$ (underfitted).

### S3.3.1 Bounds for Overfitted Models

**Proposition S4.** *Suppose Assumptions (A1) to (A5) hold and $c_\theta$ is sufficiently large. On the event $E_1 \cap E_2 \cap E_3 \cap E_4$ defined in Lemma S5,*

$$
\sup_{\gamma \in \mathbb{S}_L \,:\, \gamma^* \subsetneq \gamma} \frac{p(\gamma \mid y)}{p(\gamma^* \mid y)} \leq 3N^{-|\gamma \setminus \gamma^*|_L}.
$$

*Proof.* Fix an overfitted model $\gamma \supsetneq \gamma^*$ and let $\ell = |\gamma| > \ell^*$. Applying the inequality $(1+x)^n \leq e^{nx}$ for $n > 0$ and using

$$
\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 - \|y - X_{\gamma^*}\hat{\mu}_{\gamma^*}\|_2^2 = y^\top (H_{\gamma^*} - H_\gamma) y,
$$

we obtain from Theorem 1 that

$$
\frac{p(\gamma \mid y)}{p(\gamma^* \mid y)} \leq 3\left(\frac{\theta\sqrt{2\pi}}{1-\theta}\right)^{\ell-\ell^*} \exp\left\{\frac{\kappa_1 + M}{2}\frac{y^\top (H_\gamma - H_{\gamma^*})y + \sigma^2 \ell^* \log N}{\kappa_2 + \|y - X_\gamma \hat{\mu}_\gamma\|_2^2}\right\}.
$$

Since $\gamma^* \subset \gamma$, we have $H_{\gamma^*} X_{\gamma^*} = H_\gamma X_{\gamma^*} = X_{\gamma^*}$, which implies that, on the event $E_1$,

$$y^\top (H_\gamma - H_{\gamma^*}) y = \epsilon^\top (H_\gamma - H_{\gamma^*}) \epsilon \leq 3(\ell - \ell^*) \sigma^2 L \log N.$$

This further yields that $\sigma^2 \ell^* \log N + y^\top (H_\gamma - H_{\gamma^*}) y \leq 4(\ell - \ell^*) \sigma^2 L \log N$, as $\ell - \ell^* \geq 1$ and $L \geq \ell$. Similarly, on the event $E_1 \cap E_4$, we can use $M \geq 4L \log N$ to get

$$\|y - X_\gamma \hat{\mu}_\gamma\|_2^2 = \epsilon^\top (I - H_\gamma) \epsilon \geq \frac{7}{8} \sigma^2 M - 3\sigma^2 L \log N \geq \frac{1}{8} \sigma^2 M.$$

It thus follows from Theorem 1 and Assumptions (A2) and (A3) that

$$\begin{aligned}
\frac{p(\gamma \mid y)}{p(\gamma^* \mid y)} &\leq 3 \left( \frac{\theta \sqrt{2\pi}}{1 - \theta} \right)^{\ell - \ell^*} \exp \left\{ \frac{\kappa_1 + M}{2} \frac{y^\top (H_\gamma - H_{\gamma^*}) y + \sigma^2 \ell^* \log N}{\kappa_2 + \|y - X_\gamma \hat{\mu}_\gamma\|_2^2} \right\} \\
&\leq 3 \left( \frac{\theta \sqrt{2\pi}}{1 - \theta} \right)^{\ell - \ell^*} \exp \left\{ \frac{4(\ell - \ell^*) \sigma^2 L \log N}{(\kappa_2 + \|y - X_\gamma \hat{\mu}_\gamma\|_2^2)/M} \right\} \\
&\leq 3 \left( \frac{\theta \sqrt{2\pi}}{1 - \theta} \right)^{\ell - \ell^*} \exp \left\{ 32(\ell - \ell^*) L \log N \right\} \\
&= 3(2\pi)^{(\ell - \ell^*)/2} \exp \left\{ (32 - c_\theta)(\ell - \ell^*) L \log N \right\}.
\end{aligned}$$

Letting $c_\theta$ be sufficiently large, we obtain the claimed bound. □

### S3.3.2 Bounds for Underfitted Models

**Proposition S5.** *Suppose Assumptions (A1) to (A5) hold for sufficiently large $c_\theta, c_\mu, c_M$. On the event $E_1 \cap E_2 \cap E_3 \cap E_4$ defined in Lemma S5,*

$$\sup_{\gamma \in \mathbb{S}_L :\, \gamma^* \not\subset \gamma} \frac{p(\gamma \mid y)}{p(\gamma^* \mid y)} \leq 3N^{-(|\gamma \setminus \gamma^*| + |\gamma^* \setminus \gamma|)L}.$$

*Proof.* The overall proof strategy parallels the argument for Proposition S4. Let $\ell = |\gamma|$,

$\tilde{\gamma} = \gamma \cup \gamma^*$ and $\tilde{\ell} = |\tilde{\gamma}|$. We rewrite the inequality (16) given in Theorem 1 as

$$\frac{p(\gamma \mid y)}{p(\gamma^* \mid y)} \leq R_{\text{over}} R_{\text{under}},$$

where

$$\begin{aligned} R_{\text{over}} &= 3 \left( \frac{\theta\sqrt{2\pi}}{1-\theta} \right)^{\tilde{\ell}-\ell^*} \left( \frac{\kappa_2 + \|y - X_{\gamma^*}\hat{\mu}_{\gamma^*}\|_2^2 + \sigma^2 \ell^* \log N}{\kappa_2 + \|y - X_{\tilde{\gamma}}\hat{\mu}_{\tilde{\gamma}}\|_2^2} \right)^{(\kappa_1+M)/2}, \\ R_{\text{under}} &= \left( \frac{\theta\sqrt{2\pi}}{1-\theta} \right)^{\ell-\tilde{\ell}} \left( \frac{\kappa_2 + \|y - X_{\tilde{\gamma}}\hat{\mu}_{\tilde{\gamma}}\|_2^2}{\kappa_2 + \|y - X_{\gamma}\hat{\mu}_{\gamma}\|_2^2} \right)^{(\kappa_1+M)/2}. \end{aligned}$$

Since $\gamma^* \subset \tilde{\gamma}$, Proposition S4 can be applied to bound $R_{\text{over}}$ by $R_{\text{over}} \leq 2N^{-L(\tilde{\ell}-\ell^*)}$.

It only remains to bound $R_{\text{under}}$. As in the overfitted case, applying the inequality $(1+x)^n \leq e^{nx}$ and using $\kappa_1 \leq M$ in Assumption (A3), we get

$$R_{\text{under}} \leq \left( \frac{\theta\sqrt{2\pi}}{1-\theta} \right)^{\ell-\tilde{\ell}} \exp\left\{ -\frac{y^\top (H_{\tilde{\gamma}} - H_\gamma) y}{(\kappa_2 + \|y - X_\gamma \hat{\mu}_\gamma\|_2^2)/M} \right\}. \tag{S26}$$

The difference in residual sum of squares can be bounded as

$$\begin{aligned} y^\top (H_{\tilde{\gamma}} - H_\gamma) y &= \|(H_{\tilde{\gamma}} - H_\gamma)(X_{\gamma^*}\mu^*_{\gamma^*} + \epsilon)\|_2^2 \\ &= \|(I - H_\gamma) X_{\gamma^*}\mu^*_{\gamma^*} + (H_{\tilde{\gamma}} - H_\gamma)\epsilon\|_2^2 \\ &\geq \left( \|(I - H_\gamma) X_{\gamma^*}\mu^*_{\gamma^*}\|_2 - \|(H_{\tilde{\gamma}} - H_\gamma)\epsilon\|_2 \right)^2, \end{aligned}$$

where on the second line we have used $H_{\tilde{\gamma}} X_{\gamma^*} = X_{\gamma^*}$ since $\tilde{\gamma}$ is overfitted, and the last step follows from the reverse triangle inequality. On the event $E_1$, $\|(H_{\tilde{\gamma}} - H_\gamma)\epsilon\|_2^2 \leq$

$3(\tilde{\ell}-\ell)\sigma^2 L\log N$. Under Assumptions (A1) and (A4), we apply Lemma S4 to get

$$\begin{aligned}\|(I-H_\gamma)X_{\gamma^*}\mu^*_{\gamma^*}\|_2^2 &= \|(I-H_\gamma)X_{\gamma^*\setminus\gamma}\mu^*_{\gamma^*\setminus\gamma}\|_2^2\\ &\geq M\underline{\lambda}\|\mu^*_{\gamma^*\setminus\gamma}\|_2^2\\ &\geq (\tilde{\ell}-\ell)c_\mu^2\sigma^2 L\log N.\end{aligned}$$

For sufficiently large $c_\mu$, we thus find that

$$y^\top(H_{\tilde{\gamma}}-H_\gamma)y \geq \frac{1}{2}\|(I-H_\gamma)X_{\gamma^*}\mu^*_{\gamma^*}\|_2^2.$$

To bound the denominator in the exponent, we apply the triangle inequality to get

$$\kappa_2+\|y-X_\gamma\hat{\mu}_\gamma\|_2^2 \leq \frac{1}{2}\sigma^2 M+2\epsilon^\top\epsilon+2\|(I-H_\gamma)X_{\gamma^*}\mu_{\gamma^*}\|_2^2 \leq 3\sigma^2 M+2\|(I-H_\gamma)X_{\gamma^*}\mu_{\gamma^*}\|_2^2,$$

on the event $E_4$. Define $S=\|(I-H_\gamma)X_{\gamma^*}\mu_{\gamma^*}\|_2^2$. Using the elementary inequality $(x_1+x_2)^{-1}\geq(2x_1)^{-1}\wedge(2x_2)^{-1}$, we can bound the exponent in (S26) as

$$\begin{aligned}\frac{y^\top(H_{\tilde{\gamma}}-H_\gamma)y}{(\kappa_2+\|y-X_\gamma\hat{\mu}_\gamma\|_2^2)/M} &\geq \frac{S/2}{3\sigma^2+2S/M}\\ &\geq \frac{S}{12\sigma^2}\wedge\frac{M}{8}\\ &\geq \left(\frac{c_\mu^2}{12}\wedge\frac{c_M}{8}\right)(\tilde{\ell}-\ell)L\log N,\end{aligned}$$

where in the last step we have used Assumptions (A4) and (A5) and the fact $\tilde{\ell}-\ell\leq\ell^*$. It is now clear from (S26) that as long as $c_\mu$ and $c_M$ are sufficiently large relative to $c_\theta$, $R_{\text{under}}\leq N^{-L(\tilde{\ell}-\ell)}$. Hence, we conclude that

$$\frac{p(\gamma\mid y)}{p(\gamma^*\mid y)}\leq R_{\text{over}}R_{\text{under}}\leq 3N^{-L(2\tilde{\ell}-\ell^*-\ell)}.$$

This yields the claimed bound upon since $\tilde{\ell} - \ell^* = |\gamma \setminus \gamma^*|$ and $\tilde{\ell} - \ell = |\gamma^* \setminus \gamma|$. □

## S3.4 Proof of Theorem 3

*Proof.* Consider $\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y\right]$ first. For any $\mu$ satisfying the simplex constraint, it is easy to show that $\|\mu - \mu^*\|_2^2 \leq 2$. Hence, for any $\gamma \in \mathbb{S}_L$, $\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y, \gamma\right] \leq 2$ since $p(\mu_\gamma \mid y, \gamma)$ has support $\Delta^{|\gamma|-1}$. Applying Theorem 2, we get

$$\begin{aligned}\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y\right] &= \sum_{\gamma \in \mathbb{S}_L} p(\gamma \mid y)\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y, \gamma\right] \\ &\leq p(\gamma^* \mid y)\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y, \gamma^*\right] + 2\sum_{\gamma \in \mathbb{S}_L \setminus \{\gamma^*\}} p(\gamma \mid y) \\ &\leq \mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y, \gamma^*\right] + 2c_3 N^{-c_4 L}, \qquad \text{(S27)}\end{aligned}$$

where $c_3, c_4 > 0$ are universal constants introduced in Theorem 2.

It remains to bound

$$\begin{aligned}\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y, \gamma^*\right] &= \int \|\mu_{\gamma^*} - \mu^*_{\gamma^*}\|_2^2 \, p(\mu_{\gamma^*}, \phi \mid y, \gamma^*) \mathrm{d}\mu_{\gamma^*} \mathrm{d}\phi \\ &= \frac{1}{p(\gamma^* \mid y)} \int \|\mu_{\gamma^*} - \mu^*_{\gamma^*}\|_2^2 \, p(\gamma^*, \mu_{\gamma^*}, \phi \mid y) \mathrm{d}\mu_{\gamma^*} \mathrm{d}\phi. \qquad \text{(S28)}\end{aligned}$$

Since when $\ell^* = 1$, $\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y, \gamma^*\right] = 0$ due to the simplex constraint, we assume $\ell^* \geq 2$ henceforth. We use the notation introduced in Proposition S1 and write $u = \mu^*_\gamma$, $\tilde{u} = \Pi u$. The integral in (S28) is understood as an integral over $(\tilde{u}, \phi) \in \tilde{\Delta}^{\ell^*-1} \times [0, \infty)$. Let $\mathcal{B}_r(\Pi\, \check{\mu}_{\gamma^*})$ be as defined in Lemma S8, and define

$$B_1 = \{(\tilde{u}, \phi) \colon \tilde{u} \in \mathcal{B}_r(\Pi\, \check{\mu}_{\gamma^*}),\ \phi > 0\}, \quad B_2 = \{(\tilde{u}, \phi) \colon \tilde{u} \in \tilde{\Delta}^{\ell^*-1} \setminus \mathcal{B}_r(\Pi\, \check{\mu}_{\gamma^*}),\ \phi > 0\}.$$

Since $\mathcal{B}_r(\Pi\, \check{\mu}_{\gamma^*}) \subset \tilde{\Delta}^{\ell^*-1}$ by Lemma S8, we can split the integral in (S28) into integrals over

$B_1$ and $B_2$. We claim that for some univeral constant $c_5 > 0$,

$$\|u - \mu^*_{\gamma^*}\| \leq \frac{c_5 \sigma \sqrt{L \log N}}{\underline{\lambda}\sqrt{M}} \text{ on } B_1, \text{ and } \|u - \mu^*_{\gamma^*}\| \leq 2 \text{ on } B_2. \tag{S29}$$

To prove the claim for $B_1$, we note that by the definition of $\mathcal{B}_r(\Pi\, \check{\mu}_{\gamma^*})$, if $(\tilde{u}, \phi) \in B_1$,

$$\begin{aligned}\|u - \check{\mu}_{\gamma^*}\|_2^2 &= \|\tilde{\mu} - \Pi\, \check{\mu}_{\gamma^*}\|_2^2 + \|1 - \mathbf{1}^\top \tilde{u} - (1 - \mathbf{1}^\top \Pi\, \check{\mu}_{\gamma^*})\|^2 \\ &\leq r^2 + \|\mathbf{1}^\top (\tilde{u} - \Pi\, \check{\mu}_{\gamma^*})\|^2 \leq \ell^* r^2,\end{aligned}$$

in which the last step follows from Cauchy-Schwarz inequality. Using the definition of $r$ given in Lemma S7 and the assumption $\ell^* \geq 2$, we find that

$$\|u - \check{\mu}_{\gamma^*}\|_2 \leq \sqrt{2}(c_\mu - 3)\frac{\sigma\sqrt{L \log N}}{\underline{\lambda}\sqrt{M}}, \tag{S30}$$

Meanwhile, repeating the argument for (S17), we get

$$\|\mu^*_{\gamma^*} - \check{\mu}_{\gamma^*}\|_2 \leq \frac{3\sigma\sqrt{\ell^* \log N}}{\underline{\lambda}\sqrt{M}}, \tag{S31}$$

on the event $E_2$. Combining (S30) and (S31) proves (S29). Plugging the bounds in (S29) back into (S28), we get

$$\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y, \gamma^*\right] \leq \frac{c_5^2 \sigma^2 L \log N}{\underline{\lambda}^2 M} + \frac{2}{p(\gamma^* \mid y)} \int_{B_2} p(\gamma^*, u, \phi \mid y) \mathrm{d}\tilde{u}\, \mathrm{d}\phi. \tag{S32}$$

By (S15) and using the notation introduced in equation (S21), we have

$$p(\gamma^*, u, \phi \mid y) = C f(\ell^*)\, \phi^{(M+\kappa_1)/2}\, e^{-K\phi}\, \mathcal{N}(\tilde{u};\, \Pi\check{\mu}_{\gamma^*},\, (\phi D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1}).$$

Applying Proposition S3, we get

$$\begin{aligned}\frac{\int_{B_2} p(\gamma^*, u, \phi \mid y)\mathrm{d}\tilde{u}\,\mathrm{d}\phi}{p(\gamma^* \mid y)} &\le \frac{3\int_{B_2} \phi^{(M+\kappa_1)/2}\, e^{-K\phi}\,\mathcal{N}(\tilde{u};\, \Pi\check{\mu}_{\gamma^*},\, (\phi D^\top X_{\gamma^*}^\top X_{\gamma^*} D)^{-1})\mathrm{d}u\,\mathrm{d}\phi}{\Gamma\left(\frac{M+\kappa_1}{2}\right) K^{-(M+\kappa_1)/2}} \\ &= \frac{3\int \phi^{(M+\kappa_1)/2}\, e^{-K\phi}[1-\mathbb{P}(\tilde{U}_{\gamma^*,\phi} \in \tilde{\Delta}^{|\gamma|-1})]\mathrm{d}\phi}{\Gamma\left(\frac{M+\kappa_1}{2}\right) K^{-(M+\kappa_1)/2}}, \end{aligned} \tag{S33}$$

where $\tilde{U}_{\gamma^*,\phi} \sim \mathcal{N}(\Pi\check{\mu}_\gamma,\, (\phi D^\top X_\gamma^\top X_\gamma D)^{-1})$ is as defined in Proposition S1. Using Lemma S8 and the notation $\tilde{K}$ introduced in equation (S21),

$$\begin{aligned}\int \phi^{(M+\kappa_1)/2}\, e^{-K\phi}[1-\mathbb{P}(\tilde{U}_{\gamma^*,\phi} \in \tilde{\Delta}^{|\gamma|-1})]\mathrm{d}\phi &\le e^{1/2}\int \phi^{(M+\kappa_1)/2}\, e^{-(K+\tilde{K})\phi}\mathrm{d}\phi \\ &= e^{1/2}\Gamma\left(\frac{M+\kappa_1}{2}\right)(K+\tilde{K})^{-(M+\kappa_1)/2}.\end{aligned}$$

By (S25),

$$\left(\frac{K+\tilde{K}}{K}\right)^{-(M+\kappa_1)/2} \le \exp\left(-\frac{\tilde{K}}{4\sigma^2}\right) = \exp\left(-\frac{L\log N}{(\ell^*)^2}\right).$$

Since $e^{1/2} \le 2$, it follows from (S33) that

$$\frac{\int_{B_2} p(\gamma^*, u, \phi \mid y)\mathrm{d}\tilde{u}\,\mathrm{d}\phi}{p(\gamma^* \mid y)} \le 6\exp\left(-\frac{L\log N}{(\ell^*)^2}\right). \tag{S34}$$

Combining (S27), (S32) and (S34) yields the claimed bound for $\mathbb{E}\left[\|\mu-\mu^*\|_2^2 \mid y\right]$.

For the posterior expected prediction loss, observe that given any model $\gamma$ and $\mu_\gamma \in \Delta^{|\gamma|-1}$, we can write $\tilde{X}\mu = \tilde{X}_{\gamma^*}\mu_{\gamma^*} + \tilde{X}_{\gamma\setminus\gamma^*}\mu_{\gamma\setminus\gamma^*}$. Hence,

$$\begin{aligned}\|\tilde{X}\mu - \tilde{X}\mu^*\|_2^2 = \|\tilde{X}_{\gamma^*}(\mu_{\gamma^*} - \mu_{\gamma^*}^*) + \tilde{X}_{\gamma\setminus\gamma^*}\mu_{\gamma\setminus\gamma^*}\|_2^2 &\le 2\|\tilde{X}_{\gamma^*}(\mu_{\gamma^*} - \mu_{\gamma^*}^*)\|_2^2 + 2\|\tilde{X}_{\gamma\setminus\gamma^*}\mu_{\gamma\setminus\gamma^*}\|_2^2 \\ &\le 2\tilde{M}\overline{\lambda}\|\mu_{\gamma^*} - \mu_{\gamma^*}^*\|_2^2 + 2\tilde{M}\overline{\lambda}\|\mu_{\gamma\setminus\gamma^*}\|_2^2 = 2\tilde{M}\overline{\lambda}\|\mu-\mu^*\|_2^2.\end{aligned}$$

Therefore, the claimed bound on $\mathbb{E}\left[\|\tilde{X}\mu - \tilde{X}\mu^*\|_2^2 \mid y\right]$ immediately follows from the bound on $\mathbb{E}\left[\|\mu - \mu^*\|_2^2 \mid y\right]$. □

## S3.5 Proof of Theorem 4

Consider the model $Y = Xw + \epsilon$ with $\epsilon \sim N(0, \phi^{-1}I)$. We use $\tilde{p}(\cdot \mid y)$ to denote the posterior distribution under the following prior:

$$\begin{aligned}
\phi &\sim \text{Gamma}(\kappa_1/2, \kappa_2/2), \\
\gamma_i &\overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta), \\
\tau &\sim \text{Gamma}(a_1, a_2), \\
w_\gamma \mid \gamma, \tau, \phi &\sim N(0,\, (\tau/\phi)I), \\
w_i \mid \gamma_i = 0 &\sim \delta_0.
\end{aligned}$$

This is the classical setup for the spike-and-slab Bayesian variable selection, except that we place an additional prior distribution on $\tau$ so that it aligns with our BVS-SS model.

***Proof of Theorem 4.*** The posterior probability of $\gamma$ given $\tau$ under the two models can be expressed as

$$p(\gamma \mid y, \tau) = C_\tau\, p_0(\gamma \mid y, \tau), \quad \tilde{p}(\gamma \mid y, \tau) = \tilde{C}_\tau\, \tilde{p}_0(\gamma \mid y, \tau), \tag{S35}$$

where $C_\tau, \tilde{C}_\tau$ denote the normalizing constants, and $p_0, \tilde{p}_0$ denote the un-normalized posterior distributions given by

$$\begin{aligned}
p_0(\gamma \mid y, \tau) &= p(\gamma) \int p(y \mid \gamma, \mu_\gamma, \phi, \tau) p(\mu_\gamma \mid \gamma) p(\phi) \mathrm{d}\mu_\gamma\, \mathrm{d}\phi, \\
\tilde{p}_0(\gamma \mid y, \tau) &= \tilde{p}(\gamma) \int \tilde{p}(y \mid \gamma, \phi, \tau) \tilde{p}(\phi) \mathrm{d}\phi,
\end{aligned}$$

Note that the two models share the same prior distribution on $(\gamma, \phi)$: $p(\gamma) = \tilde{p}(\gamma)$ and $p(\phi) = \tilde{p}(\phi)$. For the unconstrained spike-and-slab model, we have

$$\tilde{p}(y \mid \gamma, \phi, \tau) \ = \left(\frac{\phi}{2\pi}\right)^{M/2} \tau^{-\ell/2} \mathrm{det}(V_{\gamma,\tau})^{-1/2} \exp\left\{-\frac{\phi}{2} y^\top \Sigma_{\gamma,\tau} y\right\},$$

where $\ell = |\gamma|$, $V_{\gamma,\tau}$ and $\Sigma_{\gamma,\tau}$ are as defined in (7). Hence,

$$\tilde{p}_0(\gamma \mid y, \tau) = G(\gamma, \tau) \left(y^\top \Sigma_{\gamma,\tau} y + \kappa_2\right)^{-(M+\kappa_1)/2}, \tag{S36}$$
$$\text{where } G(\gamma, \tau) = p(\gamma) \frac{\Gamma((M+\kappa_1)/2)\,(\kappa_2/2)^{\kappa_1/2}}{\Gamma(\kappa_1/2)\,(2\pi)^{M/2}} \tau^{-\ell/2} \mathrm{det}(V_{\gamma,\tau})^{-1/2}\, 2^{(M+\kappa_1)/2}.$$

Recall our likelihood given by (6):

$$p(y \mid \mu_\gamma, \gamma, \phi, \tau) \ = \left(\frac{\phi}{2\pi}\right)^{M/2} \tau^{-\ell/2} \mathrm{det}(V_{\gamma,\tau})^{-1/2} \exp\left\{-\frac{\phi}{2} (y - X_\gamma \mu_\gamma)^\top \Sigma_{\gamma,\tau} (y - X_\gamma \mu_\gamma)\right\}.$$

By integrating out $\mu_\gamma$ and $\phi$, we get

$$p_0(\gamma \mid y, \tau) = G(\gamma, \tau) \mathbb{E}_{\mu_\gamma \sim \mathrm{Dir}(\alpha)} \left[\left\{(X_\gamma \mu_\gamma - y)^\top \Sigma_{\gamma,\tau} (X_\gamma \mu_\gamma - y) + \kappa_2\right\}^{-(M+\kappa_1)/2}\right], \tag{S37}$$

where $\mathrm{Dir}(\alpha)$ denotes the Dirichlet distribution with parameter vector $\alpha\mathbf{1}$ (i.e., symmetric Dirichlet distribution with concentration parameter $\alpha$).

Combining (S35), (S36) and (S37), we get

$$\frac{p(\gamma \mid y, \tau)}{\tilde{p}(\gamma \mid y, \tau)} = \frac{C_\tau}{\tilde{C}_\tau} \frac{p_0(y \mid \gamma, \tau)}{\tilde{p}_0(y \mid \gamma, \tau)} = \frac{C_\tau}{\tilde{C}_\tau} \mathbb{E}_{\mu_\gamma \sim \mathrm{Dir}(\alpha)} \left[F(\mu_\gamma; \gamma, \tau)\right], \tag{S38}$$

where

$$F(\mu_\gamma; \gamma, \tau) = \left( \frac{(X_\gamma \mu_\gamma - y)^\top \Sigma_{\gamma,\tau} (X_\gamma \mu_\gamma - y) + \kappa_2}{y^\top \Sigma_{\gamma,\tau} y + \kappa_2} \right)^{-\frac{M+\kappa_1}{2}} \leq \left( \frac{\kappa_2}{y^\top \Sigma_{\gamma,\tau} y + \kappa_2} \right)^{-\frac{M+\kappa_1}{2}}.$$

The inequality follows from that $\Sigma_{\gamma,\tau} = (I + \tau X_\gamma X_\gamma^\top)^{-1}$ is always non-negative definite. Hence, we can apply dominated convergence theorem to get

$$\lim_{\tau \to \infty} \mathbb{E}_{\mu_\gamma \sim \mathrm{Dir}(\alpha)} \left[ F(\mu_\gamma; \gamma, \tau) \right] = \mathbb{E}_{\mu_\gamma \sim \mathrm{Dir}(\alpha)} \left[ \lim_{\tau \to \infty} F(\mu_\gamma; \gamma, \tau) \right] = 1. \tag{S39}$$

To see that $F(\mu_\gamma; \gamma, \tau)$ converges to 1 as $\tau \to \infty$, we note that

$$\begin{aligned} &\lim_{\tau \to \infty} \lambda_{\max} \left( I - X_\gamma (X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top - \Sigma_{\gamma,\tau} \right) \\ = &\lim_{\tau \to \infty} \lambda_{\max} \left( X_\gamma (X_\gamma^\top X_\gamma + \tau^{-1} I)^{-1} X_\gamma^\top - X_\gamma (X_\gamma^\top X_\gamma)^{-1} X_\gamma^\top \right) = 0, \end{aligned}$$

which can be proved by a routine calculation using the singular value decomposition of $X_\gamma$. Note that when $X_\gamma$ does not have full rank, we only need to change the limit to $I - X_\gamma (X_\gamma^\top X_\gamma)^+ X_\gamma^\top$ where $A^+$ denotes the Moore-Penrose pseudoinverse of the matrix $A$. Hence, $X_\gamma^\top \Sigma_{\gamma,\tau}$ converges to 0 in $L^2$-norm, from which it follows that $F(\mu_\gamma; \gamma, \tau) \to 1$.

By (S39) and (S38), we can prove the claim provided that $\lim_{\tau \to \infty} C_\tau / \tilde{C}_\tau = 1$. To this end, recall that $C_\tau, \tilde{C}_\tau$ are normalizing constants and thus can be expressed as

$$C_\tau^{-1} = \sum_{\gamma \in \mathbb{S}_L} p_0(\gamma \mid y, \tau), \quad \tilde{C}_\tau^{-1} = \sum_{\gamma \in \mathbb{S}_L} \tilde{p}_0(\gamma \mid y, \tau).$$

Since the set $\mathbb{S}_L$ is finite and we have shown that $p_0(y \mid \gamma, \tau) / \tilde{p}_0(y \mid \gamma, \tau) \to 1$ for each $\gamma$, we have $C_\tau / \tilde{C}_\tau \to 1$ as well. □

# S4 Additional Simulations

We present the simulation results for a sparse regression model in the main text, where each entry of $X$ follows an i.i.d standard normal distribution. In this section, we run additional simulations to check the variable selection performance when (1) the covariates in $X$ are correlated, and (2) the data $(Y, X)$ follows a non-sparse model. We consider some commonly used settings in the literature (Hsiao et al. 2012, Shi & Huang 2023, etc.), assuming that the DGPs are driven by factor models.

## S4.1 Non-sparse Models

We first consider non-sparse coefficients and make the data generated by a factor model.

$$(Y, X) = F^\top \Lambda^\top + e$$

where $Y$ is the first column of the matrix on the right hand side, $X$ is the rest of that matrix, and $e$ is an $(M + \tilde{M}) \times (N + 1)$ error matrix with each entry $e_{ij} \overset{\text{i.i.d.}}{\sim} N(0, 0.5^2)$. We simulate the treatment by

$$\tilde{Y}_i^{(1)} = Y_i + \delta_i + \tilde{\epsilon}_i, \;\; i = M + 1, \ldots, \tilde{M}.$$

where $\delta_i$ and $\tilde{\epsilon}_i$ is the same as in Section 4.1. The four common factors are

$$\begin{aligned}
&f_{1,i} \sim \mathcal{N}(0, 1), \\
&f_{2,i} = 0.9 f_{2,i-1} + e_{1,i}, \\
&f_{3,i} = 0.5 f_{3,i-1} + e_{2,i} + 0.5 e_{2,i-1} \\
&f_{4,i} = e_{3,i} + 0.8 e_{3,i-1} + 0.4 e_{3,i-2}, \quad i = 1, \ldots, M + \tilde{M}.
\end{aligned}$$

and the loadings are

$$\lambda_{j,l} \overset{\text{i.i.d.}}{\sim} \begin{cases} Unif[1,2]; & j = 0, \dots, J \\ Unif[-\frac{2}{M+\tilde{M}}, -\frac{2}{M+\tilde{M}}], & j > J. \end{cases}$$

All the units in the control group is correlated to the treated unit through the common factors, but the correlation remains only in the first $J$ units, while that for others diminish when the sample size grows.

We report the RMSEs of the six estimators in Figure 8. Notice that the simplex constraint should commonly be violated under a factor model. Our method outperforms three widely used SCM estimators or variants consistently, and close to the two oracle estimators. The performance of OLS also quickly approaches to the oracle performances, which aligns with the results in the SCM by OLS literature, such as Hsiao et al. (2012). QP's performance is relatively worse due to the violation of simplex constraint. It agrees with our concepts in the main text that, the robust and superior performance of BVS-SS across the three tables is largely due to the use of the soft simplex constraint, even when the true DGP follows some unknown non-sparse models.

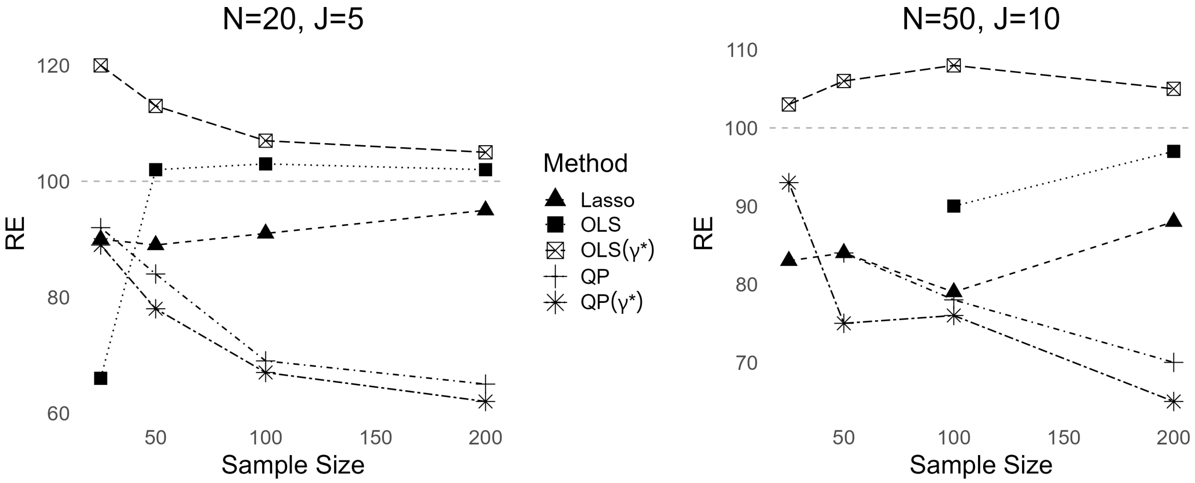

Figure 8: Relative efficiency (%) for a factor model.

The performance of BVS-SS in terms of variable selection is shown in Table 4. Unlike the results for the sparse models, since all the control units are correlated, and the factor loadings are non-zero, the minimal elements assumption may not be satisfied, and therefore BVS-SS may select a larger model.

Table 4: Variable selection performance of BVS-SS and Lasso averaged over 100 replicates.

| | | BVS-SS | | Lasso | |
|---|---|---|---|---|---|
| Setting | $M$ | $\ell^1$-loss | model size | $\ell^1$-loss | model size |
| | 25 | 5.4 | 5.9 | 4.8 | 6 |
| $N = 20$ | 50 | 5.4 | 6.5 | 4.3 | 6.3 |
| $J = 5$ | 100 | 5.4 | 7.2 | 6.2 | 8.5 |
| | 200 | 4.9 | 7.4 | 7.2 | 10.1 |
| | 25 | 15.8 | 14.2 | 9.3 | 8 |
| $N = 50$ | 50 | 16.9 | 17.1 | 9.1 | 9 |
| $J = 10$ | 100 | 17.4 | 19.4 | 11.4 | 11.9 |
| | 200 | 17.3 | 21.3 | 16 | 17.8 |

Unlike in the sparse model, the posterior mean of $\phi$ does not exhibit a clear decreasing trend. This is likely due to violations of the simplex constraint, with the degree of violation varying across each iteration. Notably, the true model does not contain a $\phi$ as estimated in our analysis in this case. Meanwhile, the trend of $\tau$ remains relatively stable, exhibiting only a very slight decrease as $M$ increases. Moreover, the magnitude of the posterior mean of $\tau$ is very close to that in the sparse model when $||w||_1 = 2$ in the main text. This further supports the notion that the factor model violates the simplex constraint.

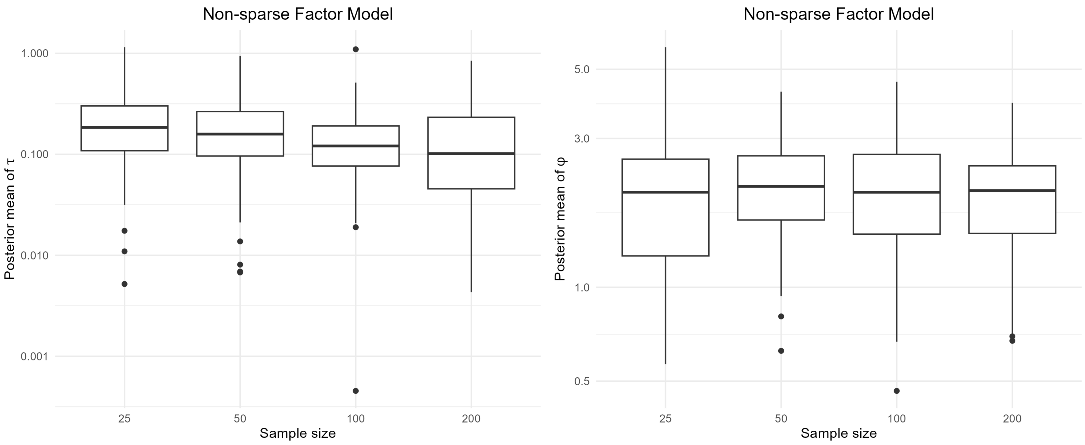

Figure 9: Distribution of the posterior mean of $\tau$ and $\phi$ across 100 replicates with $N = 20, J = 5$.

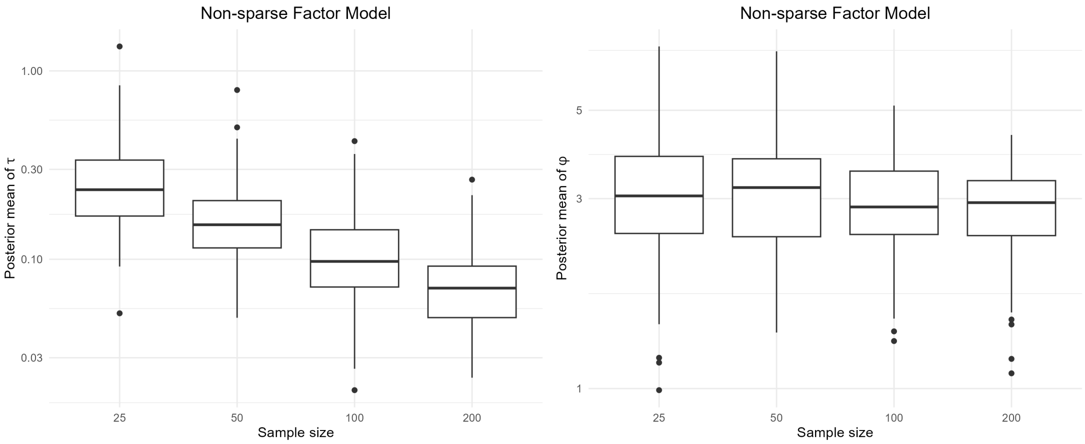

Figure 10: Distribution of the posterior mean of $\tau$ and $\phi$ across 100 replicates with $N = 50, J = 10$.

## S4.2 Sparse Models

Next, we consider a sparse model, such that $X$ is generated with a factor model, and the potential outcome of $Y$ that never receives a treatment is a convex combination of $X$ with sparse weights,

$$X = F^\top \Lambda^\top + e,$$

$$Y = Xw^* + \epsilon.$$

The treatment and sparse weights are formulated same as in Section 4.1. $X$ is generated from the same four factors as in Section S4.1, while the loadings $\lambda_{j,l} \overset{\text{i.i.d.}}{\sim} N(1,1), j = 1, \ldots, N, l = 1, 2, 3, 4$. The sparsity still remains through $w^*$, but all units in the control group are linked by some factors that do not directly affect the treated unit.

We report the RMSEs of the six estimators with $\lambda = \|w^*\|_1 = 1, 2, 3$ in the following three figures. The relative performance of these six estimators is very similar to that in the main text, indicating that even when control units are correlated, BVS-SS can still provide robust and superior performance for ATT estimation.

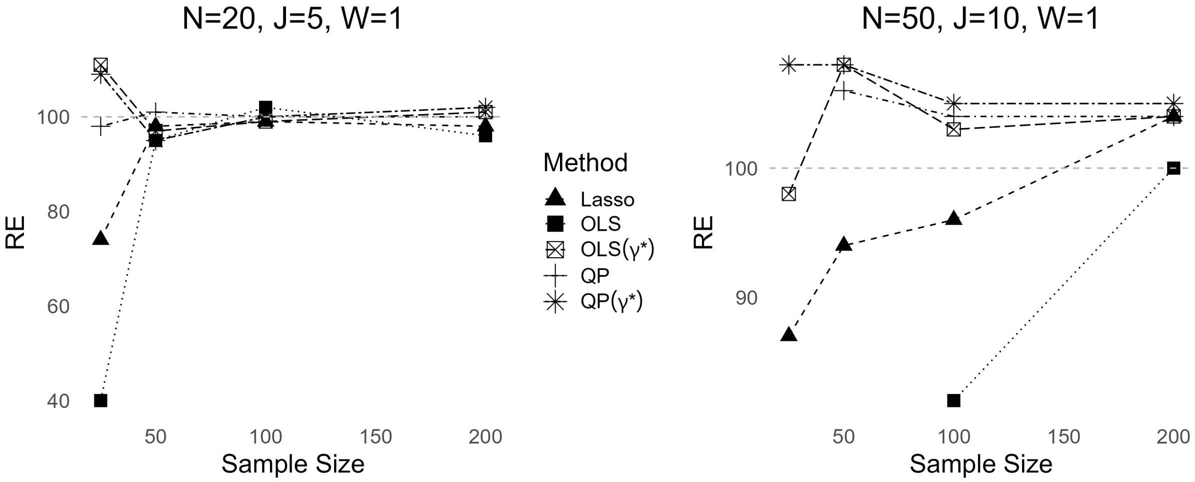

Figure 11: Relative efficiency (%) when $\|w^*\|_1 = 1$ (a sparse factor model).

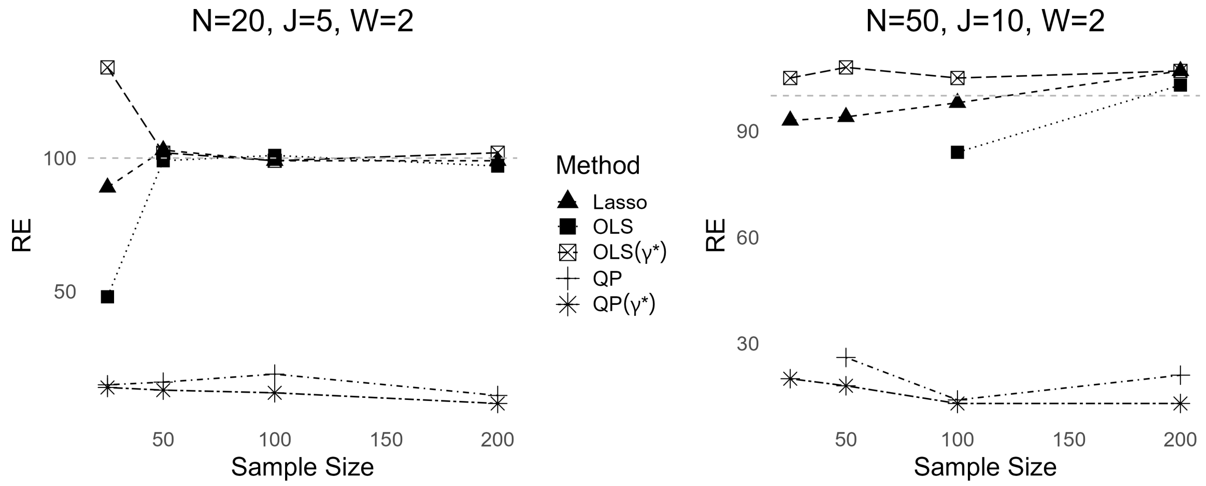

Figure 12: Relative efficiency (%) when $\|w^*\|_1 = 2$ (a sparse factor model).

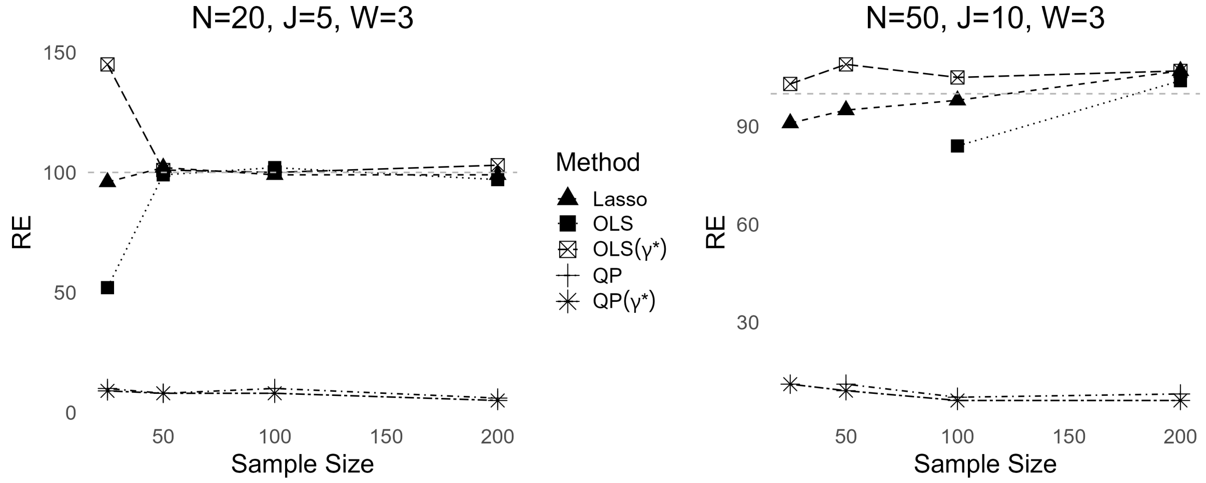

Figure 13: Relative efficiency (%) when $\|w^*\|_1 = 3$ (a sparse factor model).

When the underlying DGP follows a factor model, it is no longer possible to uniformly verify whether the simplex constraint is satisfied. Moreover, the assumption on the minimum signal strength (A4) may fail in this setting. As a result, we cannot guarantee that BVS-SS consistently selects the "correct" model in the sense established by our theorem. Nevertheless, we observe that, compared with Lasso, BVS-SS typically selects a model of smaller size, which facilitates interpretation of the counterfactual construction by highlighting the control units that contribute most to the counterfactual.

Table 5: Variable selection performance of BVS-SS and Lasso averaged over 100 replicates.

|  |  |  | BVS-SS |  | Lasso |  |
|---|---|---|---|---|---|---|
| $\|w^*\|_1$ | Setting | $M$ | $\ell^1$-loss | model size | $\ell^1$-loss | model size |
| $\|w^*\|_1 = 1$ | $N = 20$<br>$J = 5$ | 25 | 5.6 | 7.2 | 5.7 | 8.9 |
|  |  | 50 | 3.8 | 6.4 | 5.2 | 8.9 |
|  |  | 100 | 2.2 | 5.8 | 5.5 | 9.9 |
|  |  | 200 | 1.2 | 5.4 | 5.1 | 9.8 |
|  | $N = 50$<br>$J = 10$ | 25 | 19.5 | 19.1 | 12.7 | 10.9 |
|  |  | 50 | 15.5 | 16.7 | 12.5 | 13.3 |
|  |  | 100 | 10 | 13.3 | 13.9 | 20.3 |
|  |  | 200 | 6.1 | 11.2 | 14.1 | 21.2 |
| $\|w^*\|_1 = 3$ | $N = 20$<br>$J = 5$ | 25 | 3.8 | 3.2 | 5.7 | 8.9 |
|  |  | 50 | 2.4 | 3.8 | 5.2 | 8.9 |
|  |  | 100 | 1.6 | 4.1 | 5.5 | 9.9 |
|  |  | 200 | 1 | 4.4 | 5.1 | 9.8 |
|  | $N = 50$<br>$J = 10$ | 25 | 10.7 | 4.6 | 12.7 | 10.9 |
|  |  | 50 | 8.4 | 5.1 | 12.5 | 13.3 |
|  |  | 100 | 6 | 6.1 | 13.9 | 20.3 |
|  |  | 200 | 3.9 | 7 | 14.1 | 21.2 |

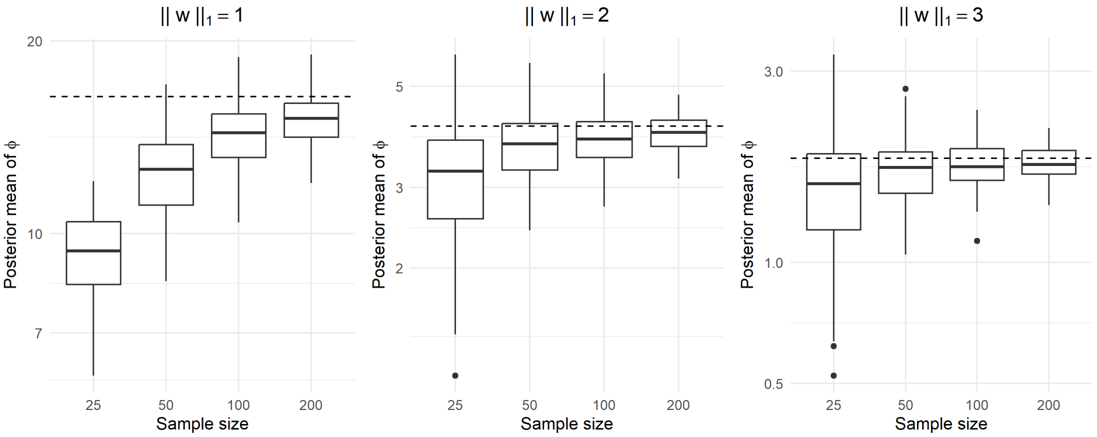

Figure 14: Distribution of the posterior mean of $\phi$ across 100 replicates with $N = 20, J = 5$. The true value $\phi^*$ is indicated by the dotted line.

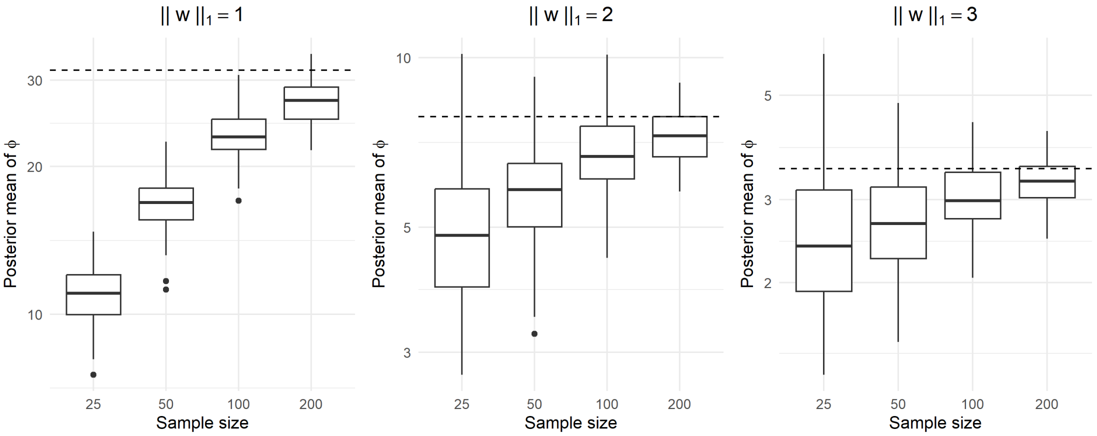

Figure 15: Distribution of the posterior mean of $\phi$ across 100 replicates with $N = 50, J = 10$. The true value $\phi^*$ is indicated by the dotted line.

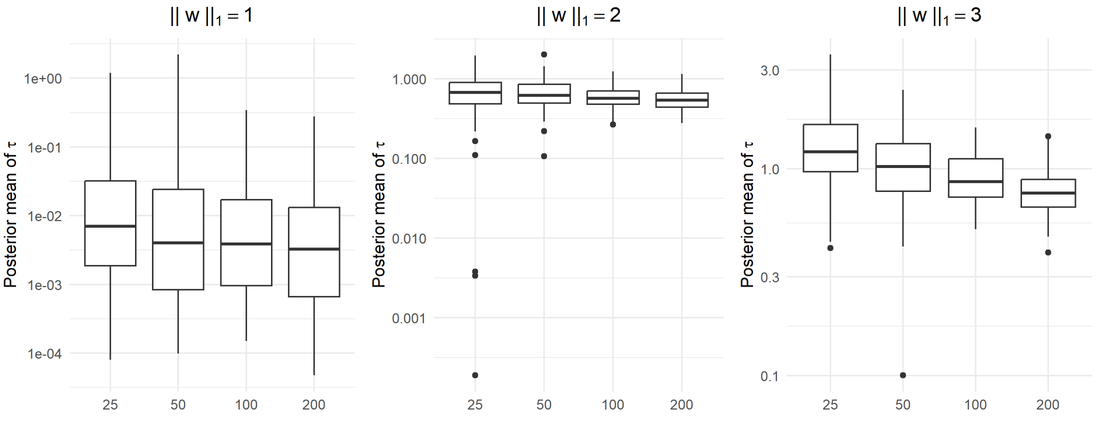

Figure 16: Distribution of the posterior mean of $\tau$ across 100 replicates with $N = 20, J = 5$.

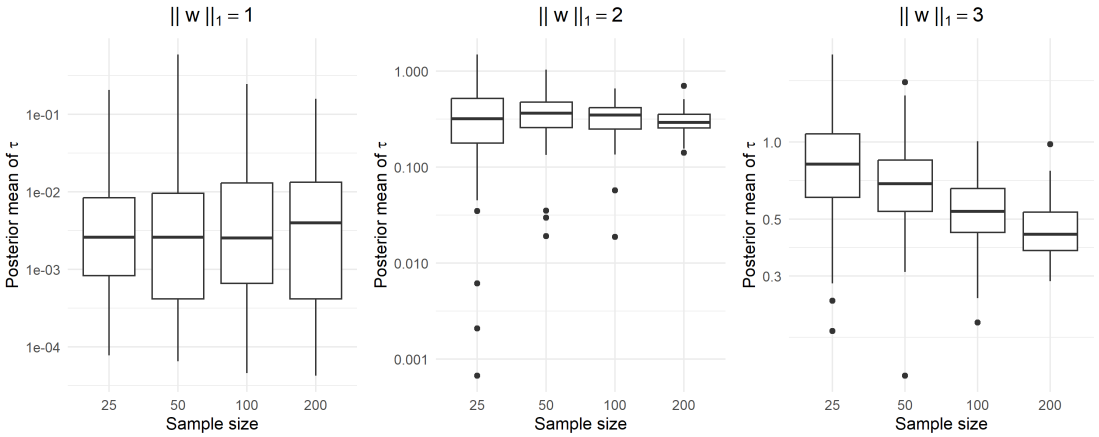

Figure 17: Distribution of the posterior mean of $\tau$ across 100 replicates with $N = 50, J = 10$.

# S5 Supplementary Empirical Examples Results

In Figures 18 and 19 we visualize the posterior distributions of $\tau$, $\phi$, model size and ATT estimate for the two real data sets considered in Section 5. Figure 20 and 21 provide the counterfactual plots for the two empirical examples in our main text.

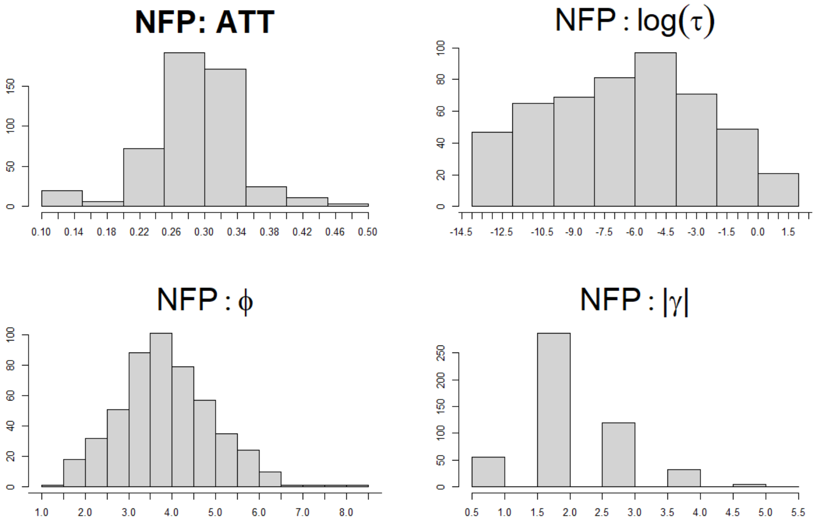

Figure 18: Posterior distributions for the Nota Fiscal Paulista data set.

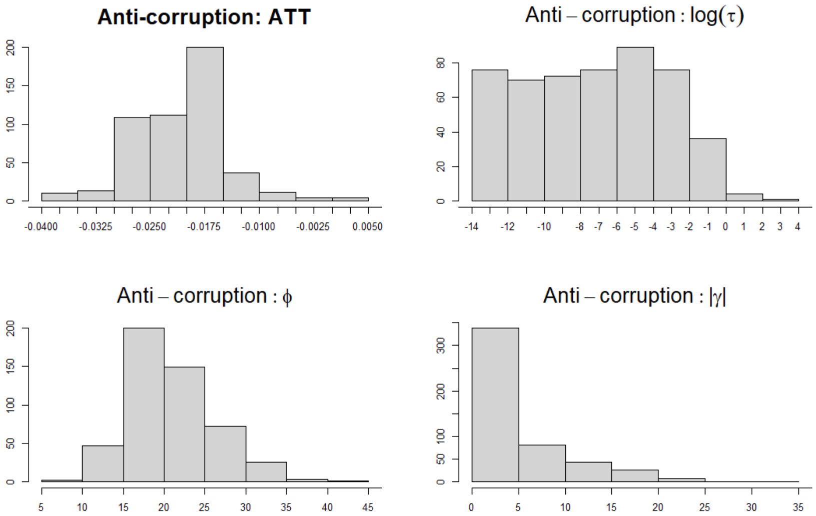

Figure 19: Counterfactual Estimations for the China's Anti-corruption Campaign data set.

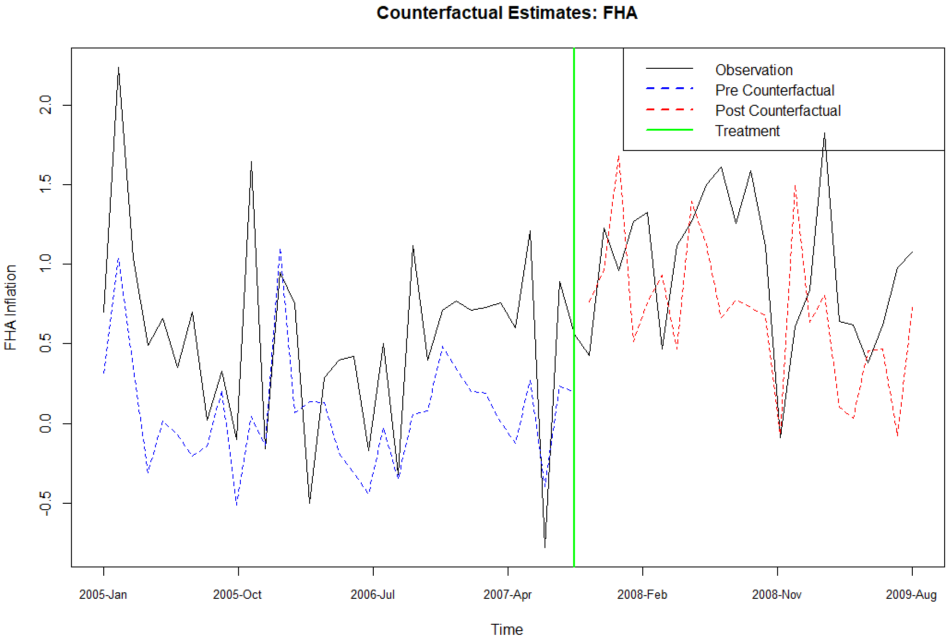

Figure 20: Counterfactual estimation for the Nota Fiscal Paulista data set.

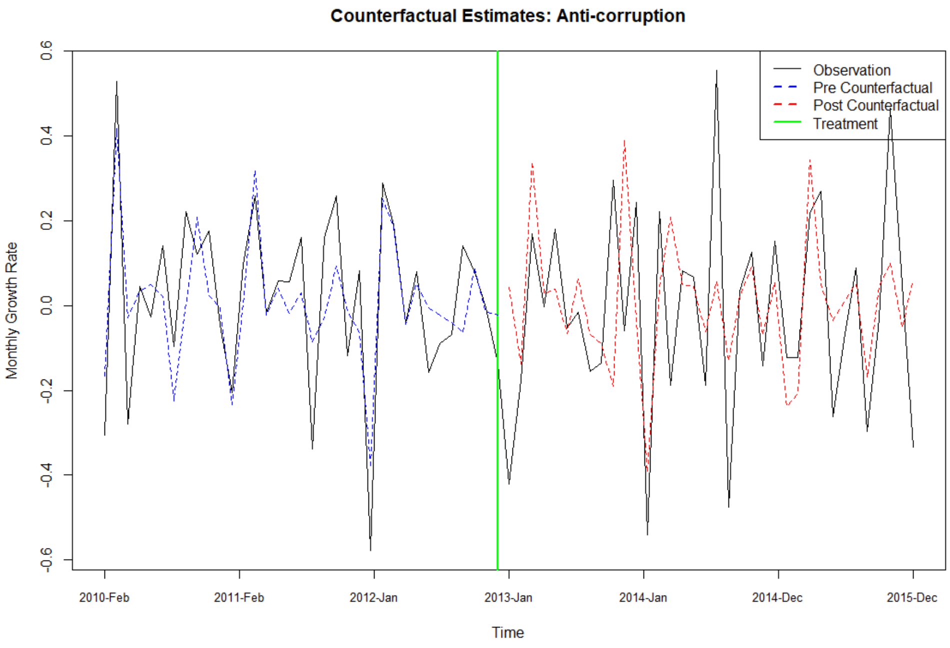

Figure 21: Counterfactual estimation for the China's Anti-corruption Campaign data set.

# References

Abadie, A. (2021), 'Using synthetic controls: Feasibility, data requirements, and methodological aspects', *Journal of Economic Literature* **59**(2), 391–425.

Abadie, A., Diamond, A. & Hainmueller, J. (2010), 'Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program', *Journal of the American statistical Association* **105**(490), 493–505.

Abadie, A., Diamond, A. & Hainmueller, J. (2011), 'Synth: An R package for synthetic control methods in comparative case studies', *Journal of Statistical Software* **42**(13).

Abadie, A. & Gardeazabal, J. (2003), 'The economic costs of conflict: A case study of the basque country', *American economic review* **93**(1), 113–132.

Amemiya, T. (1985), *Advanced econometrics*, Harvard university press.

Andrieu, C. & Roberts, G. O. (2009), 'The pseudo-marginal approach for efficient Monte Carlo computations', *The Annals of Statistics* **37**(2), 697–725.

Athey, S. & Imbens, G. W. (2017), 'The state of applied econometrics: Causality and policy evaluation', *Journal of Economic perspectives* **31**(2), 3–32.

Brown, P. J., Vannucci, M. & Fearn, T. (1998), 'Multivariate Bayesian variable selection and prediction', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(3), 627–641.

Carvalho, C., Masini, R. & Medeiros, M. C. (2018), 'Arco: An artificial counterfactual approach for high-dimensional panel time-series data', *Journal of econometrics* **207**(2), 352–380.

Casella, G., Girón, F. J., Martínez, M. L. & Moreno, E. (2009), 'Consistency of Bayesian procedures for variable selection', *The Annals of Statistics* **37**(3), 1207–1228.

Chang, H. & Zhou, Q. (2024), 'Dimension-free relaxation times of informed MCMC samplers on discrete spaces', *arXiv preprint arXiv:2404.03867* .

Chernozhukov, V., Wüthrich, K. & Zhu, Y. (2021), 'An exact and robust conformal inference method for counterfactual and synthetic controls', *Journal of the American Statistical Association* **116**(536), 1849–1864.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P. & Stine, R. A. (2001), 'The practical implementation of Bayesian model selection', *Lecture Notes-Monograph Series* pp. 65–134.

Choi, N. H., Li, W. & Zhu, J. (2010), 'Variable selection with the strong heredity constraint and its oracle property', *Journal of the American Statistical Association* **105**(489), 354–364.

Doudchenko, N. & Imbens, G. W. (2016), Balancing, regression, difference-in-differences and synthetic control methods: A synthesis, Technical report, National Bureau of Economic Research.

Farcomeni, A. (2010), 'Bayesian constrained variable selection', *Statistica Sinica* **20**(3), 1043–1062.

**URL:** *https://www.jstor.org/stable/24309479*

Firpo, S. & Possebom, V. (2018), 'Synthetic control method: Inference, sensitivity analysis and confidence sets', *Journal of Causal Inference* **6**(2), 20160026.

George, E. I. & McCulloch, R. E. (1993), 'Variable selection via Gibbs sampling', *Journal of the American Statistical Association* **88**(423), 881–889.

George, E. I. & McCulloch, R. E. (1997), 'Approaches for Bayesian variable selection', *Statistica sinica* pp. 339–373.

Goh, G. & Yu, J. (2022), 'Synthetic control method with convex hull restrictions: a bayesian maximum a posteriori approach', *The Econometrics Journal* **25**(1), 215–232.

Goldfarb, D. & Idnani, A. (2006), Dual and primal-dual methods for solving strictly convex quadratic programs, *in* 'Numerical Analysis: Proceedings of the Third IIMAS Workshop Held at Cocoyoc, Mexico, January 1981', Springer, pp. 226–239.

Golub, G. H. & Van Loan, C. F. (2013), *Matrix computations*, JHU press.

Guan, Y. & Stephens, M. (2011), 'Bayesian variable selection regression for genome-wide association studies and other large-scale problems', *The Annals of Applied Statistics* **5**(3), 1780.

Hollingsworth, A. & Wing, C. (2020), 'Tactics for design and inference in synthetic control studies: An applied example using high-dimensional data', *Available at SSRN 3592088* .

Hsiao, C., Steve Ching, H. & Ki Wan, S. (2012), '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* **27**(5), 705–740.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Kim, S., Lee, C. & Gupta, S. (2020), 'Bayesian synthetic control methods', *Journal of Marketing Research* **57**(5), 831–852.

King, G. & Zeng, L. (2006), 'The dangers of extreme counterfactuals', *Political analysis* **14**(2), 131–159.

Lan, X. & Li, W. (2018), 'Swiss watch cycles: Evidence of corruption during leadership transition in china', *Journal of Comparative Economics* **46**(4), 1234–1252.

Laurent, B. & Massart, P. (2000), 'Adaptive estimation of a quadratic functional by model selection', *Annals of Statistics* pp. 1302–1338.

Li, C. & Li, H. (2008), 'Network-constrained regularization and variable selection for analysis of genomic data', *Bioinformatics* **24**(9), 1175–1182.

Li, K. T. (2020), 'Statistical inference for average treatment effects estimated by synthetic control methods', *Journal of the American Statistical Association* **115**(532), 2068–2083.

Martinez, I. & Vives-i-Bastida, J. (2022), 'Bayesian and frequentist inference for synthetic controls', *arXiv preprint arXiv:2206.01779* .
**URL:** *https://arxiv.org/abs/2206.01779*

Narisetty, N. N. & He, X. (2014), 'Bayesian variable selection with shrinking and diffusing priors', *The Annals of Statistics* **42**(2), 789–817.

Shang, Z. & Clayton, M. K. (2011), 'Consistency of bayesian linear model selection with a growing number of parameters', *Journal of Statistical Planning and Inference* **141**(11), 3463–3474.

Shi, Z. & Huang, J. (2023), 'Forward-selected panel data approach for program evaluation', *Journal of Econometrics* **234**(2), 512–535.

Smith, M. & Kohn, R. (1996), 'Nonparametric regression using Bayesian variable selection', *Journal of Econometrics* **75**(2), 317–343.

Smyth, G. K. & Verbyla, A. P. (1996), 'A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(3), 565–572.

Vershynin, R. (2018), *High-dimensional probability: An introduction with applications in data science*, Vol. 47, Cambridge university press.

Xu, Y. (2017), 'Generalized synthetic control method: Causal inference with interactive fixed effects models', *Political Analysis* **25**(1), 57–76.

Yang, Y., Wainwright, M. J. & Jordan, M. I. (2016), 'On the computational complexity of high-dimensional Bayesian variable selection', *The Annals of Statistics* **44**(6), 2497–2532.

Zhou, Q. & Guan, Y. (2019), ‘Fast model-fitting of Bayesian variable selection regression using the iterative complex factorization algorithm’, *Bayesian analysis* **14**(2), 573.

Zhou, Q., Yang, J., Vats, D., Roberts, G. O. & Rosenthal, J. S. (2022), ‘Dimension-free mixing for high-dimensional Bayesian variable selection’, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**(5), 1751–1784.