Essays in Bayesian Econometrics

Masahiro Tanaka

A thesis submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

Economics

Waseda University

2020

Reading Committee:

Hideki Konishi

Hisatoshi Tanaka

Yasuhiro Omori

Abstract

Essays in Bayesian Econometrics

Masahiro Tanaka

This thesis consists of three essays on methodological contributions to Bayesian econometrics.

The first chapter proposes a new Bayesian inferential approach to local projections. A local projection is a statistical framework that accounts for the relationship between an exogenous variable and an endogenous variable, measured at different time points. Local projections are often applied in impulse response analyses and direct forecasting. While local projections are becoming increasingly popular because of their robustness to misspecification and their flexibility, they are less statistically efficient than standard methods, such as vector autoregression. In this study, I seek to improve the statistical efficiency of local projections by developing a fully Bayesian approach that can be used to estimate local projections using roughness penalty priors. The proposed priors, which are adapted from Bayesian splines, are generated from an intrinsic Gaussian Markov random field; that is, they induce random-walk behavior on a sequence of parameters. By incorporating such prior-induced smoothness, one can use information contained in successive observations to enhance the statistical efficiency of an inference. I compare the proposed approach with the existing approaches through a series of Monte Carlo experiments. I apply the proposed approach to an analysis of monetary policy in the United States, showing that the roughness penalty priors successfully estimate the impulse response functions and improve the predictive accuracy of local projections. [*Computational Economics*, Volume 55, Issue 2, pp. 629–651.]

The second chapter develops a computational method for the Bayesian version of generalized method of moments (GMM) in difficult situations. A GMM criterion can be viewed as a quasi-likelihood, being theoretically equivalent to the Laplace approximation of the true likelihood around its mode. Exploiting this feature, one can conduct a (quasi-)Bayesian inference by replacing true likelihood with a GMM criterion. There are cases where the number of moment conditions can be large. However, a GMM estimator is unreliable when the number of moment conditions is large, that is, it is comparable or larger than the sample size. While a number of provisions for this problem is proposed in classical GMM literature, the literature on its Bayesian counterpart (i.e., Bayesian inference using a GMM criterion as a quasi-likelihood) has paid scant attention to this problem. This study fills this gap by proposing an adaptive Markov chain Monte Carlo (MCMC) approach to a GMM inference with many moment conditions. Particularly, this study focuses on the adaptive tuning of a weighting matrix on the fly. Our proposal consists of two elements. The first is the use of the nonparametric eigenvalue-regularized precision matrix estimator, which contributes to numerical stability. The second is the random update of a weighting matrix, which substantially reduces computational cost, while maintaining the accuracy of the estimation. A simulation study and real data application are then presented to illustrate the performance of the proposed approach in comparison with existing approaches. [*arXiv preprint,* arXiv:1811.00722.]

The third chapter proposes a new Bayesian approach to infer average treatment effect. The approach treats counterfactual untreated outcomes as missing observations and infers them by completing a matrix composed of realized and potential untreated outcomes using a data augmentation technique. We also develop a tailored prior that helps in the identification of parameters and induces the matrix of the untreated outcomes to be approximately low rank. While

the proposed approach is similar to synthetic control methods and other relevant methods, it has several notable advantages. Unlike synthetic control methods, the proposed approach does not require stringent assumptions. Whereas synthetic control methods do not have a statistically grounded method to quantify uncertainty about inference, the proposed approach can estimate credible sets in a straightforward and consistent manner. Our proposal approach has a better finite sample performance than the existing Bayesian and non-Bayesian approaches, as we show through a series of simulation studies. [*arXiv preprint*, arXiv:1911.01287.]

Contents

Chapter 1

Bayesian Inference of Local Projections with Roughness Penalty Priors

1.1 Introduction

Local projections introduced by Jordà (2005) provide a statistical framework that accounts for the relationship between an exogenous variable and an endogenous variable, measured at different time points. Typical applications of local projections include impulse response analyses and direct (non-iterative) forecasting (Stock and Watson, 2007). A local projection has several advantages over standard methods, such as vector autoregression (VAR). First, it does not impose a strong assumption on the data-generating process, making it robust to misspecification. Second, it can easily deal with asymmetric and/or state-dependent impulse responses (e.g., Riera-Crichton et al., 2015; Auerbach and Gorodnichenko, 2013; Ramey and Zubuairy, 2018). On the other hand, local projections have several disadvantages. First, when using a local projection, the exogenous variable must be identified beforehand. Second, a local projection is subject to more estimation risk than other methods, and typically obtains a wiggly impulse response function, (e.g., Ramey, 2016). In an impulse response analysis, the shape of an estimated impulse response function is of concern. Therefore, if an estimated impulse response function is wiggly and has wide confidence/credible intervals, it is difficult to interpret the result, and one might wrongly reject or accept a hypothesis . In this study, we address the second disadvantage of local projections.

In order to improve the statistical efficiency, we develop a fully Bayesian approach that can be used to estimate local projections using roughness penalty priors as well as B-spline basis expansions.[1] The proposed priors, which are adapted from Bayesian splines (Lang and Brezger, 2004), are generated from an intrinsic Gaussian Markov random field; that is, they induce random-walk behavior on a sequence of parameters. By incorporating such prior-induced smoothness, we can use information contained in successive observations to enhance the statistical efficiency of an inference. We compare the proposed approach with the existing approaches through a series of Monte Carlo experiments. The proposed approach is applied to an analysis of monetary policy shocks in the United States to show how the roughness penalty priors successively smooth impulse responses and improve statistical efficiency in terms of predictive accuracy. Furthermore, we show that such improvements are almost entirely attributable to the roughness penalty priors and not to the B-spline expansions.

There are three strands of studies related to this work. For the first, Barnichon and Matthes

[1] See, e.g., Geweke (2005) for a general introduction to Bayesian analysis.

(2019) approximate a moving average representation of a time series using values from Gaussian basis functions. Their approximation is simpler, but much coarser than ours. As a result, their estimated impulse responses may be excessively smoothed and vulnerable to model misspecification. For the second, to smooth an impulse response estimate, Miranda-Agrippino and Ricco (2017) penalize the estimate based on deviations from an impulse response derived from an estimated VAR. However, their approach seems not to work well in cases with asymmetric and/or state-dependent impulse responses. Furthermore, their approach uses the same dataset twice. This shortcoming can be resolved if a time series is long enough to be split into training and estimation samples, but this is not the general situation in macroeconomic studies. In contrast, our approach does not require a reference model, thus it is free from these problems.

For the third, the most relevant studies are those of Barnichon and Brownlees (2019) and El-Shagi (2019), who develop frequentist methods using roughness penalties. Although our approach can be regarded as a Bayesian counterpart to theirs, it confers four additional benefits. First, our approach is more flexible than Barnichon and Brownlees's (2019) approach: they allow a single parameter to control the smoothness of all parameter sequences, whereas we can assign different smoothing parameters to individual sequences. Second, our Bayesian approach can evaluate credible intervals in a consistent and straightforward manner, while the frequentist approaches cannot provide a theoretically grounded confidence interval. Third, in our approach, smoothing parameters are inferred from priors, implying that we can systematically consider uncertainty in the smoothness of an impulse response. In contrast, the frequentist approach prefixes smoothing parameters; Barnichon and Brownlees (2019) choose a smoothing parameter via cross-validation, while El-Shagi (2019) determines smoothing parameters on the basis of some information criteria. Fourth, our approach has better finite-sample performance than El-Shagi's (2019) approach, as shown in Section 1.5.

The rest of the paper is organized as follows. Section 1.2 introduces the model, the priors and the posterior simulation. Section 1.3 conducts a set of Monte Carlo experiments and reports the result. Section 1.4 demonstrates our approach in an analysis of the macroeconomic effects of monetary policy shocks in the United States. Section 1.5 compares the proposed approach with the existing frequentist approaches. Section 1.6 concludes this chapter.

1.2 Proposed Approach

We consider two classes of local projections: those with and those without B-spline expansions.

1.2.1 Local Projection without B-spline expansions

1.2.1.1 Model

We begin by describing a local projection (Jordà, 2005). While we consider only time series data, an extension to panel data is straightforward. A model for an individual observation is given by

$$y_{(h),t+h} = \beta_{(h)} z_t + \alpha_{(h)} + \sum_{j=1}^{J-2} \gamma_{j,(h)} w_{j,t} + u_{(h),t+h}, \ h = 0, 1, 2, ..., H; \ t = 1, ..., T,$$

where $y_{(h),t+h}$ is an endogenous variable observed at period $t+h$, $\alpha_{(h)}$ is an intercept, z_t is an exogenous variable observed at period t, $w_{1,t}, ..., w_{J-2,t}$ are covariates, which may include lags

of the endogenous and exogenous variables, $\beta_{(h)}$ and $\gamma_{j,(h)}$ are unknown coefficients, and $u_{(h),t+h}$ is a residual. The model allows asymmetric and/or state-dependent impulse responses, as in Riera-Crichton et al. (2015), Auerbach and Gorodnichenko (2013), and Ramey and Zubuairy (2018). The definition of $y_{(h),t+h}$ and z_t depends on whether the model is used for an impulse response analysis or forecasting. For the former task, $y_{(h),t+h}$ denotes a response observed h periods after shock z_t occurs at t; for the latter, z_t is one of several predictors observed at t, and $y_{(h),t+h}$ is an h-period-ahead target observation. Note that in an impulse response analysis, z_t cannot be predicted using the current and past information, which implies that z_t is uncorrelated with x_t. In what follows, we focus on impulse response analysis.

In an impulse response analysis, we seek to infer a smooth function $f_z(h)$ that represents an impulse response of y to z, namely, $f_z(h) = \partial y_{(h),t+h}/\partial z_t$. Here, we allow a sequence $\left\{\beta_{(0)}, ..., \beta_{(H)}\right\}$ to represent an impulse response of y to z, namely, $\beta_{(h)} = \partial y_{(h),t+h}/\partial z_t$. The model can be represented as

$$
\begin{aligned}
y_{(h),t+h} &= \boldsymbol{x}_t^\top \boldsymbol{\theta}_{(h)} + u_{(h),t+h}, \quad h = 0, ..., H;\ t = 1, .., T; \\
\boldsymbol{x}_t &= \left(z_t, 1, w_{1,t}, ..., w_{J-2,t}\right)^\top, \\
\boldsymbol{\theta}_{(h)} &= \left(\beta_{(h)}, \alpha_{(h)}, \gamma_{(h),1}, ..., \gamma_{(h),J-2}\right)^\top,
\end{aligned}
$$

where $\boldsymbol{x}_t$ is a J-dimensional vector of regressors and $\boldsymbol{\theta}_{(h)}$ is a vector of corresponding parameters. For notational convenience, we reindex the coefficient vector as $\boldsymbol{\theta}_{(h)} = \left(\theta_{(h),1}, ..., \theta_{(h),J}\right)^\top$. Stacking these over the projection dimension yields a representation resembling a seemingly unrelated regression (SUR): for $t = 1, ..., T$,

$$
\begin{aligned}
\boldsymbol{y}_t &= \left(\boldsymbol{I}_{H+1} \otimes \boldsymbol{x}_t^\top\right) \boldsymbol{\theta} + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim \mathcal{N}\left(\boldsymbol{0}_{H+1}, \boldsymbol{\Sigma}\right), \\
\boldsymbol{y}_t &= \left(y_{(0),t}, ..., y_{(H),t+H}\right)^\top, \quad \boldsymbol{u}_t = \left(u_{(0),t}, ..., u_{(H),t+H}\right)^\top, \\
\boldsymbol{\theta} &= \left(\boldsymbol{\theta}_{(0)}^\top, ..., \boldsymbol{\theta}_{(H)}^\top\right)^\top,
\end{aligned}
$$

where $\boldsymbol{\Sigma}$ is a covariance matrix, and $\mathcal{N}(\boldsymbol{d}, \boldsymbol{B})$ denotes a multivariate normal distribution with mean $\boldsymbol{d}$ and covariance $\boldsymbol{B}$. $\otimes$ denotes the Kronecker product.

Although multivariate time series data are considered, an LP model is not a typical time series model. The purpose of the model is to directly examine the relationship between variables measured at different time points, not to recover the underlying data generating process (DGP). Therefore, the realizations of the endogenous variable observed at different projection points $h = 0, ..., H$ are treated as different time series.

Rearranging the above representation, we express the model in matrix notation as

$$
\begin{aligned}
\boldsymbol{y} &= \left(\boldsymbol{I}_{H+1} \otimes \boldsymbol{X}\right) \boldsymbol{\theta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0}_{(H+1)T},\ \boldsymbol{\Sigma} \otimes \boldsymbol{I}_T\right) \qquad (1.1) \\
\boldsymbol{y} &= \left(\boldsymbol{y}_{(0)}^\top, ..., \boldsymbol{y}_{(H)}^\top\right)^\top, \quad \boldsymbol{X} = \left(\boldsymbol{x}_1, ..., \boldsymbol{x}_T\right)^\top, \quad \boldsymbol{u} = \left(\boldsymbol{u}_{(0)}^\top, ..., \boldsymbol{u}_{(H)}^\top\right)^\top, \\
\boldsymbol{y}_{(h)} &= \left(y_{(h),1+h}, ..., y_{(h),T+h}\right)^\top, \quad \boldsymbol{u}_{(h)} = \left(u_{(h),1+h}, ..., u_{(h),T+h}\right)^\top, \quad h = 0, ..., H.
\end{aligned}
$$

Letting $\mathcal{D}$ denote the data, the likelihood takes a standard form:

$$
\begin{aligned}
p\left(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\Sigma}\right) &= (2\pi)^{-\frac{(H+1)T}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\boldsymbol{u}^\top \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T\right) \boldsymbol{u}\right] \\
&= (2\pi)^{-\frac{(H+1)T}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{U}^\top \boldsymbol{U} \boldsymbol{\Sigma}^{-1}\right)\right], \\
\boldsymbol{U} &= \left(\boldsymbol{u}_{(0)}, ..., \boldsymbol{u}_{(H)}\right).
\end{aligned}
$$

1.2.1.2 Bayesian Inference

This section first discusses priors on the subsets of $\boldsymbol{\theta}$ and then assembles them into a prior on $\boldsymbol{\theta}$, followed by a description of the priors on the other parameters. Lastly, the posterior simulation method is discussed.

We introduce a class of roughness penalty priors for $\boldsymbol{\theta}_j = \left(\theta_{(0),j}, ..., \theta_{(H),j}\right)^\top$, $j = 1, ..., J$. Our prior construction is motivated by Lang and Brezger (2004). The prior induces an rth-order random-walk behavior on a sequence of parameters $\theta_{(0),j}, ..., \theta_{(H),j}$. When $r = 2$, the relationship between $\theta_{(h),j}$ and successive parameters is represented by

$$\theta_{(h),j} = 2\theta_{(h-1),j} - \theta_{(h-2),j} + \epsilon_{(h),j}, \quad \epsilon_{(h),j} \sim \mathcal{N}\left(0,\ \tau_j^{-1}\lambda_{(h),j}^{-1}\right),$$

for $h = r, ..., H$, where τ_j and $\lambda_{(h),j}$ are global and local smoothing parameters, respectively. Controlling local smoothness is potentially beneficial, because impulse response functions often have both strongly bent and smooth areas: for example, fast-growing responses immediately after an occurrence of shock and virtually flat responses after convergence to a long-run equilibrium. In some applications, without the adaptation for local smoothness, an estimated impulse response might be oversmoothed in some areas and undersmoothed in others. A prior on $\boldsymbol{\theta}_j$ is an improper normal prior generated by an intrinsic Gaussian Markov random field (Rue and Held, 2005), and the smoothing parameters are inferred from gamma priors, unlike in existing approaches such as Miranda-Agrippino and Ricco (2017); Barnichon and Matthes (2019). The hierarchy of the prior takes the form

$$\begin{aligned} p\left(\boldsymbol{\theta}_j|\tau_j, \boldsymbol{\Lambda}_j\right) &\propto \exp\left[-\frac{\tau_j}{2}\sum_{h=r}^{H}\lambda_{(h),j}\left(\Delta^r\theta_{(h),j}\right)^2\right] \\ &= \exp\left(-\frac{\tau_j}{2}\boldsymbol{\theta}_j^\top \boldsymbol{D}^\top\boldsymbol{\Lambda}_j\boldsymbol{D}\boldsymbol{\theta}_j\right) \end{aligned}$$

$$\lambda_{(r),j} = 1 \text{ and } \lambda_{(h),j} \sim \mathcal{G}\left(\eta_1, \eta_2\right),\ h = r+1, ..., H,$$

$$\tau_j \sim \mathcal{G}\left(\nu_1, \nu_2\right),$$

where $\boldsymbol{\Lambda}_j = \text{diag}\left(\lambda_{(r),j}, ..., \lambda_{(H),j}\right)$, $\boldsymbol{D}$ is an $(H-r)$-by-$(H+1)$ difference matrix of order r,[2] η_1, η_2, ν_1, and ν_2 are pre-fixed hyperparameters, $\mathcal{G}\left(a, b\right)$ denotes a gamma distribution with shape a and rate b (and, thus, with mean a/b and variance a/b^2), and Δ^r denotes the rth-order difference operator. By assembling the priors on the subsets of $\boldsymbol{\theta}$, the prior density for $\boldsymbol{\theta}$ conditional on $\boldsymbol{\tau} = \{\tau_1, ..., \tau_J\}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_1, ..., \boldsymbol{\Lambda}_J\}$ is represented as

$$p\left(\boldsymbol{\theta}|\boldsymbol{\tau}, \boldsymbol{\Lambda}\right) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{Q}\boldsymbol{\theta}\right),$$

$$\boldsymbol{Q} = \sum_{j=1}^{J}\left(\left(\tau_j\boldsymbol{D}^\top\boldsymbol{\Lambda}_j\boldsymbol{D}\right)\otimes\boldsymbol{E}_j\right), \tag{1.2}$$

[2] For instance, when $r = 2$,

$$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}.$$

where $\boldsymbol{E}_j$ is a J-by-J zero matrix in which the jth diagonal element is replaced by one. In what follows, the above prior is referred to as an adaptive roughness penalty (A-RP) prior. As a special case, the same prior with all local smoothing parameters set to one is called a non-adaptive roughness penalty (N-RP) prior. For the N-RP prior, (1.2) can be rewritten as

$$\boldsymbol{Q} = \boldsymbol{D}^\top \boldsymbol{D} \otimes \text{diag}\left(\tau_1, ..., \tau_J\right).$$

Choosing a prior of the covariance matrix $\boldsymbol{\Sigma}$ is non-trivial. Because of the strong correlations between the residuals, $\boldsymbol{\Sigma}$ tends to be close to a matrix of ones and almost singular. If the Jeffreys prior, a popular non-informative prior for covariance matrices, is employed, a posterior simulation easily crashes due to the singularity of the gram matrix of the realized residuals.[3] Therefore, prior-induced shrinkage is necessary to complete a posterior simulation. On the other hand, as Alvarez et al. (2014) argue, an inverse Wishart prior, another popular choice, can be unintentionally, significantly informative, resulting in significant biases. For these reasons, we use a hierarchical inverse Wishart (HIW) prior for $\boldsymbol{\Sigma}$ (Huang and Wand, 2013):

$$\begin{aligned}
\boldsymbol{\Sigma}|\boldsymbol{\Phi} &\sim \mathcal{IW}\left(2\zeta\boldsymbol{\Phi},\ \zeta + H\right), \\
\Phi &= \text{diag}\left(\phi_{(0)}, ..., \phi_{(H)}\right), \\
\phi_{(h)} &\sim \mathcal{G}\left(\frac{1}{2}, \upsilon\right), \quad h = 0, ..., H,
\end{aligned}$$

where $\phi_{(h_i)}$ is a hyperparameter to be inferred, ζ and υ are prefixed hyperparameters, and $\mathcal{IW}\left(\boldsymbol{A}, b\right)$ is an inverse Wishart distribution with scale matrix $\boldsymbol{A}$ and degrees of freedom b. This prior distribution is seen as a scale mixture of inverse Wishart distributions, and is more robust than an inverse Wishart prior. We conducted a simulation study that compares an inverse Wishart prior and the HIW prior and show that the HIW prior has better finite sample performance than an inverse Wishart prior. See Section A.1 in the Appendix for details.

We can induce this prior to be arbitrarily non-informative by setting υ to a very small value, but Huang and Wand's (2013) recommendation $\upsilon = 10^{-10}$ (in our notation) is too flat to complete the posterior simulation in this chapter. Our default choice in this chapter is $\upsilon = 0.01$. Although there is no general procedure to find a sufficiently small value of υ, the results in the subsequent sections are not sensitive to υ as long as it is chosen from a fairly large range $[10^{-4}, 10^{-1}]$ (see Section A.2 in the Appendix).

As demonstrated in Section A.2 in the Appendix, the proposed approach is not very sensitive to the choice of the hyperparameters. However, as the priors used in this chapter are not scale-invariant, a user of the proposed approach is strongly encouraged to conduct a prior sensitivity check.

A posterior simulation is conducted using the Markov chain Monte Carlo (MCMC) algorithm. Because all of the conditional posterior densities are standard, we can construct a block Gibbs sampler. The joint posterior is represented as follows:

[3] Bayesian inference using the Jeffreys prior for $\boldsymbol{\Sigma}$ almost always fails for the synthetic and real data used in the subsequent section.

Algorithm 1.1 Sampling $\boldsymbol{\theta}$ (Rue 2001)

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{P}^{-1}\right),$$
$$\boldsymbol{m} = \boldsymbol{P}^{-1}\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^{\top}\right)\boldsymbol{y}, \quad \boldsymbol{P} = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^{\top}\boldsymbol{X} + \boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^{\top}.$$

Step 1. Sample $\boldsymbol{a} \sim \mathcal{N}\left(\boldsymbol{0}_{(H+1)J}, \boldsymbol{I}_{(H+1)J}\right)$.
Step 2. Solve $\boldsymbol{L}^{\top}\boldsymbol{b} = \boldsymbol{a}$ to obtain $\boldsymbol{b}$.
Step 3. Solve $\boldsymbol{L}\boldsymbol{c} = \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^{\top}\right)\boldsymbol{y}$ to obtain $\boldsymbol{c}$.
Step 4. Solve $\boldsymbol{L}^{\top}\boldsymbol{m} = \boldsymbol{c}$ to obtain $\boldsymbol{m}$.
Step 5. Set $\boldsymbol{\theta} = \boldsymbol{b} + \boldsymbol{m}$.

$$\begin{aligned}
p\left(\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}|\mathcal{D}\right) &\propto p\left(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\Sigma}\right) p\left(\boldsymbol{\theta}|\boldsymbol{\tau}, \boldsymbol{\Lambda}\right) p\left(\boldsymbol{\tau}\right) p\left(\boldsymbol{\Lambda}\right) p\left(\boldsymbol{\Sigma}|\boldsymbol{\Phi}\right) p\left(\boldsymbol{\Phi}\right) \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{T}{2}} \exp\left[-\frac{1}{2}\boldsymbol{u}^{\top}\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_{T}\right)\boldsymbol{u}\right] \\
&\quad \times \exp\left(-\frac{1}{2}\boldsymbol{\theta}^{\top}\boldsymbol{Q}\boldsymbol{\theta}\right) \\
&\quad \times \prod_{j=1}^{J} \tau_j^{\nu_1 - 1} \exp\left(-\nu_2 \tau_j\right) \\
&\quad \times \prod_{j=1}^{J} \prod_{h=r+1}^{H} \lambda_{(h),j}^{\eta_1 - 1} \exp\left(-\eta_2 \lambda_{(h),j}\right) \\
&\quad \times |2\zeta\boldsymbol{\Phi}|^{\frac{\zeta+H}{2}} |\boldsymbol{\Sigma}|^{-\frac{\zeta+H+1}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\left(2\zeta\boldsymbol{\Phi}\right)\boldsymbol{\Sigma}^{-1}\right)\right] \\
&\quad \times \prod_{h=0}^{H} \phi_{(h)}^{-\frac{1}{2}} \exp\left(-\upsilon\phi_{(h)}\right).
\end{aligned}$$

Each sampling block is specified as follows.

Sampling $\boldsymbol{\theta}$ The conditional posterior density of $\boldsymbol{\theta}$ is given by the multivariate normal distribution:

$$\boldsymbol{\theta}|- \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{P}^{-1}\right),$$

$$\boldsymbol{m} = \boldsymbol{P}^{-1}\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^{\top}\right)\boldsymbol{y}, \tag{1.3}$$
$$\boldsymbol{P} = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^{\top}\boldsymbol{X} + \boldsymbol{Q}. \tag{1.4}$$

This block is computationally demanding, with two bottlenecks. The first is concerned with calculation of the prior precision matrix, which involves repeated high-dimensional matrix multiplications, Eq. (1.2). The computational cost declines significantly by treating $\boldsymbol{Q}$ as a sparse matrix. The second bottleneck is the inversion of $\boldsymbol{P}$. For speed and numerical stability, we apply the algorithm described in Section 2 of Rue (2001) (see Algorithm 1), which exploits a banded structure of $\boldsymbol{P}$; inverting a lower-triangular Cholesky root of $\boldsymbol{P}$, (denoted by $\boldsymbol{L}$), is faster and more numerically stable than inverting $\boldsymbol{P}$ itself.

Sampling τ and $\mathbf{\Lambda}$ The conditional posteriors of the smoothing parameters for $\boldsymbol{\theta}_j$, $j = 1, ..., J$, are specified as the following gamma distributions: for $j = 1, ..., J$,

$$\tau_j|- \sim \mathcal{G}\left(\nu_1 + \frac{H-r}{2},\ \nu_2 + \frac{1}{2}\boldsymbol{\theta}_j^\top \boldsymbol{D}^\top \mathbf{\Lambda}_j \boldsymbol{D}\boldsymbol{\theta}_j\right),$$

$$\lambda_{(h),j}|- \sim \mathcal{G}\left(\eta_1 + \frac{1}{2},\ \eta_2 + \frac{\tau_j}{2}\left(\Delta^r \theta_{(h),j}\right)^2\right),\ h = r+1, ..., H.$$

Sampling $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ The conditional posteriors of $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ are

$$\mathbf{\Sigma}|- \sim \mathcal{IW}\left(2\zeta\mathbf{\Phi} + \boldsymbol{U}^\top\boldsymbol{U},\ \zeta + H + T\right),$$
$$\phi_{(h)}|- \sim \mathcal{G}\left(\frac{\zeta + H + 1}{2},\ \upsilon + \zeta\left(\mathbf{\Sigma}^{-1}\right)_{h,h}\right), \quad h = 0, ..., H,$$

where $\left(\mathbf{\Sigma}^{-1}\right)_{h,h'}$ denotes the $(h+1, h'+1)$-element of $\mathbf{\Sigma}^{-1}$.

1.2.2 Local projection with B-spline expansions

We consider a local projection with B-spline expansions as an additional smoothing device. We intend to approximate an impulse response function $f_z(h)$ using a B-spline basis function expansion over a projection horizon[4]

$$f_z(h) = \beta_{(h)} \approx \sum_{k=1}^{K} b_k \varphi_k(h) = \boldsymbol{b}^\top \boldsymbol{\varphi}(h),$$

where K is a number of knots, $\boldsymbol{b} = (b_1, ..., b_K)^\top$ is a vector of coefficients, and $\boldsymbol{\varphi}(h) = (\varphi_1(h), ..., \varphi_K(h))^\top$ is a vector of B-spline basis functions. We define the approximations of the other coefficients in a similar fashion. Given the approximation, the model is represented as

$$\begin{aligned} y_{(h),t+h} &\approx \sum_{k=1}^{K} a_k\varphi_k(h) + \sum_{k=1}^{K} b_k\varphi_k(h) z_t + \sum_{j=1}^{J-2}\sum_{k=1}^{K} c_{j,k}\varphi_k(h) w_{j,t} + u_{(h),t+h} \\ &= \boldsymbol{a}^\top\boldsymbol{\varphi}(h) + \boldsymbol{b}^\top\boldsymbol{\varphi}(h) z_t + \sum_{j=1}^{J-2}\boldsymbol{c}_j^\top\boldsymbol{\varphi}(h) w_{j,t} + u_{(h),t+h} \\ &= \boldsymbol{\vartheta}^\top(\boldsymbol{x}_t \otimes \boldsymbol{\varphi}(h)) + u_{(h),t+h}, \end{aligned}$$

where $\boldsymbol{\vartheta} = \left(\boldsymbol{b}^\top, \boldsymbol{a}^\top, \boldsymbol{c}_1^\top, ..., \boldsymbol{c}_{J-2}^\top\right)^\top$ is a vector of corresponding parameters. We reindex $\boldsymbol{\vartheta} = \left(\boldsymbol{\vartheta}_1^\top, ..., \boldsymbol{\vartheta}_J^\top\right)^\top$ for expositional convenience. Letting $\tilde{\boldsymbol{x}}_{(h),t} = \boldsymbol{x}_t \otimes \boldsymbol{\varphi}(h)$, the model can be expressed as

$$y_{(h),t+h} \approx \boldsymbol{\vartheta}^\top\tilde{\boldsymbol{x}}_{(h),t} + u_{(h),t+h}.$$

Stacking these equations over the projection dimension yields a representation à la SUR:

$$\boldsymbol{y}_t = \tilde{\boldsymbol{X}}_t\boldsymbol{\vartheta} + \boldsymbol{u}_t,$$

[4]See, for example, De Boor (1978); Eilers and Marx (1996) for a detailed description of B-splines.

$$\boldsymbol{y}_t = \left(y_{(0,)t}, ..., y_{(H),t+H}\right)^\top, \quad \tilde{\boldsymbol{X}}_t = \left(\tilde{\boldsymbol{x}}_{(0),t}, ..., \tilde{\boldsymbol{x}}_{(H),t}\right)^\top, \quad \boldsymbol{u}_t = \left(u_{(0),t}, ..., u_{(H),t+H}\right)^\top,$$

for $t = 1, ..., T$. Rearranging the above representation delivers

$$\boldsymbol{y} = \tilde{\boldsymbol{X}}\boldsymbol{\vartheta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0}_{(H+1)T},\ \boldsymbol{\Sigma} \otimes \boldsymbol{I}_T\right), \tag{1.5}$$

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_{(0)}^\top & \cdots & \boldsymbol{y}_{(H)}^\top \end{pmatrix}^\top, \quad \tilde{\boldsymbol{X}} = \begin{pmatrix} \tilde{\boldsymbol{X}}_{(0)}^\top & \cdots & \tilde{\boldsymbol{X}}_{(H)}^\top \end{pmatrix}^\top, \quad \boldsymbol{u} = \left(\boldsymbol{u}_{(0)}^\top, ..., \boldsymbol{u}_{(H)}^\top\right)^\top,$$

$$\boldsymbol{y}_{(h)} = \begin{pmatrix} y_{(h),1} & \cdots & y_{(h),T} \end{pmatrix}^\top, \quad \tilde{\boldsymbol{X}}_{(h)} = \begin{pmatrix} \tilde{\boldsymbol{x}}_{(h),1} & \cdots & \tilde{\boldsymbol{x}}_{(h),T} \end{pmatrix}^\top, \quad h = 0, ..., H,$$

$$\boldsymbol{u}_{(h)} = \left(u_{(h),1}, ..., u_{(h),T}\right)^\top, \quad h = 0, ..., H.$$

We construct a posterior simulator for the model in a similar fashion as the model without B-spline expansions. Given the same priors for $\boldsymbol{\tau}$ and $\boldsymbol{\Lambda}$, a prior on $\boldsymbol{\vartheta}$ is constructed as

$$p\left(\boldsymbol{\vartheta}|\boldsymbol{\tau}, \boldsymbol{\Lambda}\right) \propto \exp\left(-\frac{1}{2}\boldsymbol{\vartheta}^\top \boldsymbol{Q} \boldsymbol{\vartheta}\right),$$

$$\boldsymbol{Q} = \text{blkdiag}\left(\tau_1 \boldsymbol{D}^\top \boldsymbol{\Lambda}_1 \boldsymbol{D}, , ..., \tau_J \boldsymbol{D}^\top \boldsymbol{\Lambda}_J \boldsymbol{D}\right). \tag{1.6}$$

The conditional posterior density of $\boldsymbol{\vartheta}$ is derived as the multivariate normal distribution

$$\boldsymbol{\vartheta}|- \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{P}^{-1}\right),$$

$$\begin{aligned} \boldsymbol{m} &= \boldsymbol{P}^{-1} \tilde{\boldsymbol{X}}^\top \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T\right) \boldsymbol{y}, \\ \boldsymbol{P} &= \boldsymbol{Q} + \tilde{\boldsymbol{X}}^\top \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T\right) \tilde{\boldsymbol{X}}. \end{aligned} \tag{1.7}$$

As in the model without B-spline expansions, this sampling block presents a major computational burden. On the one hand, the prior precision matrix $\boldsymbol{Q}$ can be calculated easily by virtue of its block diagonal structure (1.6) (unless the number of covariates J is not extremely large). On the other hand, the quantity $\tilde{\boldsymbol{X}}^\top \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T\right) \tilde{\boldsymbol{X}}$ in (1.7) involves a high-dimensional matrix multiplication and cannot be compressed as in (1.4), eventually making the posterior simulation more demanding than the previous case. Sampling distributions of the other parameters are derived analogously to those of a model without B-spline expansions.

1.3 Simulation Study

We conducted Monte Carlo simulations to investigate the performance of our proposed approach. We considered six specifications consisting of the combination of three priors, each with/without B-spline expansions. As with the N-RP and A-RP priors, we considered a weakly informative independent standard normal prior, $\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{0},\ 10^4 \boldsymbol{I}\right)$.

First, we considered a linear data generating processes (DGPs) specified by the following moving average representation:

$$y_t = \sum_{h=0}^{H} \beta_{(h)} z_{t-h} + \epsilon_t,$$

$$z_t \sim \mathcal{N}\left(0, 1\right), \quad \epsilon_t \sim \mathcal{N}\left(0, 1\right),$$

where y_t is an endogenous variable, z_t is an exogenous variable, and ϵ_t is the measurement error. A set of parameters $\beta_{(0)}, ..., \beta_{(H)}$ represents an impulse response. True parameter values are defined as a convex curve:

$$\beta_{(h)} = \frac{h \exp\left(r\left(1-h\right)\right)}{\sum_{h'=0}^{H} h' \exp\left(r\left(1-h'\right)\right)},$$

$$r \sim \mathcal{U}\left(0.1,\ 1\right),$$

where $\mathcal{U}\left(0.1,\ 1\right)$ denotes a uniform distribution with support $(0.1,\ 1)$, and r governs where the peak of the impulse response is located. Covariates are a constant and four lags of y_t and z_t. We fixed the length of the impulse response to $H = 20$ and the effective sample size to $T = 50, 100$. Hyperparameters are $\nu = \nu_1 = \nu_2 = 0.01$, $\eta = \eta_1 = \eta_2 = 0.5$, and $\upsilon = 0.01$. We choose the order of the difference matrix as $r = 2$, implying that the sequences of parameters to be inferred are reduced to straight lines. We use the B-spline basis with equidistant knots ranging from -2 to $H - 1$ with unitary increments. We set the degree of the B-spline bases to three. We generate 500 sets of synthetic data. Gibbs sampling obtained 40,000 posterior draws, after discarding the initial 10,000.[5] Each chain is initialized to an ordinary least squares estimate.

We compared the alternative approaches on the basis of four performance measures: mean squared errors (MSE), lengths of credible intervals (Length), and computational speed (Speed). MSE is the sum of mean squared errors,

$$\text{MSE} = M^{-1} \sum_{m=1}^{M} \sum_{h=0}^{H} \left(\hat{\beta}_{m,(h)} - \beta_{m,(h)}^{true}\right)^2,$$

where $\hat{\beta}_{m,(h)}$ denotes a posterior mean estimate of $\beta_{(h)}$ in the mth experiment, $\beta_{m,(h)}^{true}$ denotes the corresponding true value, and M is the total number of experiments. Length denotes the arithmetic mean of the lengths of a 90% credible interval,

$$\text{Length} = M^{-1} \left(H+1\right)^{-1} \sum_{m=1}^{M} \sum_{h=0}^{H} \left(\hat{\beta}_{m,(h)}^{95\%} - \hat{\beta}_{m,(h)}^{5\%}\right).$$

Speed is the mean computational time of posterior simulations in seconds.[6]

Table 1.1 reports results of the first experiment. With regard to MSE and Length, the N-RP and A-RP priors outperform the normal prior, while the A-RP prior performs slightly worse than the N-RP prior. Using B-spline expansions reduces MSE and Length but the magnitude is tiny. Speed depends on the prior specification and on whether a B-spline is used. The difference attributable to the choice of prior is not notably large, but the use of a B-spline imposes a significant computational burden. When B-spline expansions are emplo , approximately 95% of computational time during each MCMC cycle is spent calculating $\boldsymbol{P}$, in particular, a quantity $\tilde{\boldsymbol{X}}^{\top}\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}_T\right)\tilde{\boldsymbol{X}}$. This bottleneck is a simple matrix-matrix multiplication that is executed via a built-in mathematical routine of Matlab, so switching to a compiled language such as Fortran and C/C++ will not totally resolve the problem. We checked the sensitivity of the simulation results to choice of the hyperparameters. The results are summarized in the Appendix.

[5] In this study, the number of MCMC iterations is chosen based on pilot runs so that the minimum of the effective sample sizes for obtained posterior draws (except the warmup draws) in each experiment is no less than 15,000.

[6] All programs were written in Matlab 2016a (64 bit) and executed on an Ubuntu Desktop 16.04 LTS (64 bit), running on Intel Xeon E5-2607 v3 processors (2.6GHz).

Table 1.1: Results of the Monte Carlo simulation: linear IRF

T	Prior	B-spline	MSE	Length	Speed
50	Normal		0.542	0.976	99
	Normal	√	0.542	0.976	1174
	N-RP		0.131	0.432	105
	N-RP	√	0.130	0.423	1164
	A-RP		0.150	0.468	120
	A-RP	√	0.151	0.459	1167
100	Normal		0.243	0.599	150
	Normal	√	0.243	0.599	4296
	N-RP		0.067	0.309	152
	N-RP	√	0.067	0.301	4291
	A-RP		0.074	0.331	167
	A-RP	√	0.074	0.323	4299

Note: MSE denotes the mean squared error. Length denotes the length of the 90% credible interval. Speed denotes computational time in seconds.

It is not surprising that the B-spline function expansions have only a marginal effect, given that response variables can appear as functional data observed on an equally spaced grid.[7] Panel (a) in Figure 1.1 displays B-spline basis functions on a fine grid (2,401 points) and simulated functional data. This situation is presumed in a functional data analysis. In a local projection, however, observation points $(h = 0, ..., H)$ are sparse and invariant, as demonstrated in panel (b). As is evident from there, B-spline expansions merely allocate observed information to the fixed grids, rather than interpolating neighboring information. In our case, observed information for a single grid point is allocated to neighboring grid points with weights {1/6, 2/3, 1/6}.[8] Therefore, using a B-spline indeed smooths estimates of impulse responses, but its effectiveness is limited.

We fortified our results by considering nonlinear DGPs characterized by asymmetric and state-dependent impulse responses, respectively. The asymmetric DGP is specified by

$$y_t = \sum_{h=0}^{H} \left(\beta_{(h)}^{[1]} z_{t-h} \mathbf{1}_{\{z_{t-h}<0\}} + \beta_{(h)}^{[2]} z_{t-h} \mathbf{1}_{\{z_{t-h}\geq 0\}} \right) + \epsilon_t,$$

while the state-dependent DGP is specified by

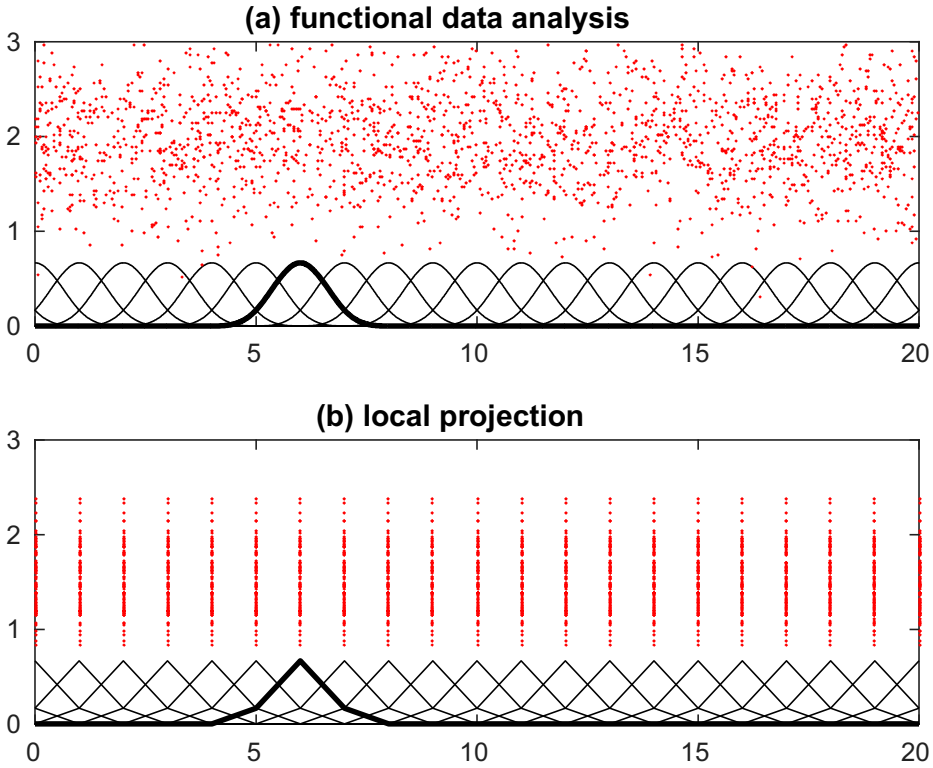
$$y_t = \sum_{h=0}^{H} \left(\beta_{(h)}^{[1]} z_{t-h} \mathbf{1}_{\{y_{t-h}<0\}} + \beta_{(h)}^{[2]} z_{t-h} \mathbf{1}_{\{y_{t-h}\geq 0\}} \right) + \epsilon_t,$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. For both cases, two sets of parameters are independently generated in the same way as the linear DGP. We set $T = 80, 160$. The other settings are exactly the same as those in the first experiment. For computational reasons, we did not consider a model with B-spline expansions. Table 1.2 presents the results, in which we largely verified the result of the first experiment. With regard to MSE and Length, the N-RP and A-RP priors consistently improve accuracy versus the normal prior.

[7] From this point of view, local projections are similar to functional data models such as Guo (2002); Morris and Carroll (2006).

[8] When the degree of the B-spline bases is increased to 5, the weight set becomes {1/120, 13/60, 33/60, 13/60,

Figure 1.1: B-spline basis



Note: The solid lines show B-spline basis functions used in a functional data analysis (panel (a)) and local projection (panel (b)), respectively. Points are simulated observations for each case. Thick lines highlight the basis functions centered at the sixth knot.

Table 1.2: Results of the Monte Carlo simulation: nonlinear IRF

T	DGP	Prior	MSE	Length
80	Asymmetric	Normal	2.489	1.439
		N-RP	0.564	0.618
		A-RP	0.643	0.670
	State-dependent	Normal	1.526	1.073
		N-RP	0.453	0.500
		A-RP	0.504	0.541
160	Asymmetric	Normal	1.211	0.911
		N-RP	0.339	0.445
		A-RP	0.373	0.477
	State-dependent	Normal	0.648	0.666
		N-RP	0.216	0.355
		A-RP	0.234	0.380

Note: MSE denotes the mean squared error. Length denotes the length of the 90% credible interval.

From the simulation study, we obtained two findings. First, our proposed approach improves the finite sample performance of local projection, while such improvements are almost entirely attributable to the roughness penalty priors and not to the B-spline expansions. Second, despite its flexibility, the A-RP prior is not superior to the N-RP prior. In conclusion, a specification with the N-RP prior and no B-spline expansion is recommendable as a first choice.

1.4 Application

To demonstrate our model, we applied our approach to an analysis of monetary policy in the United States. We use monetary policy shocks compiled by Coibion et al. (2017) which is an update of Romer and Romer (2004).[9] For the covariates and the response, we considered the following three macroeconomic variables, downloaded from the Federal Reserve Economic Data (FRED), maintained by the Federal Reserve Bank of St. Louis: the industrial production index (FRED mnemonic: INDPRO), the consumer price index for all urban consumers: all items (CPIAUCSL), and the effective federal funds rate (FEDFUNDS). We also treated all three as response variables. We included lags of monetary policy shock as covariates. All data are monthly and spans from March 1969 to December 2008. The range is limited by the availability of data for monetary policy shocks. Industrial production and the inflation rate are seasonally adjusted, and included as annualized month-to-month percentage changes (log-difference multiplied by 1,200). We included the time trend and up to four lags of covariates. We choose hyperparameters as in the previous section. The Gibbs sampler obtained a total of 40,000 posterior draws after discarding the first 10,000.

Figures 1.2, 1.3, and 1.4 show posterior estimates of the impulse responses of the macroeconomic variables to monetary policy shocks under different specifications. Figs. 1.5, 1.6, and 1.7 display credible intervals for all the specifications. The shaded areas indicate the 90% credible sets for a preferred specification using no B-spline and the N-RP prior. For all the response variables, the roughness penalty priors successfully penalize the roughness of the impulse response functions. Thus, we obtain economically plausible, smoothed estimates, and can interpret the shape of the impulse response easily, recognizing the underlining response. Use of the B-spline e lse response.

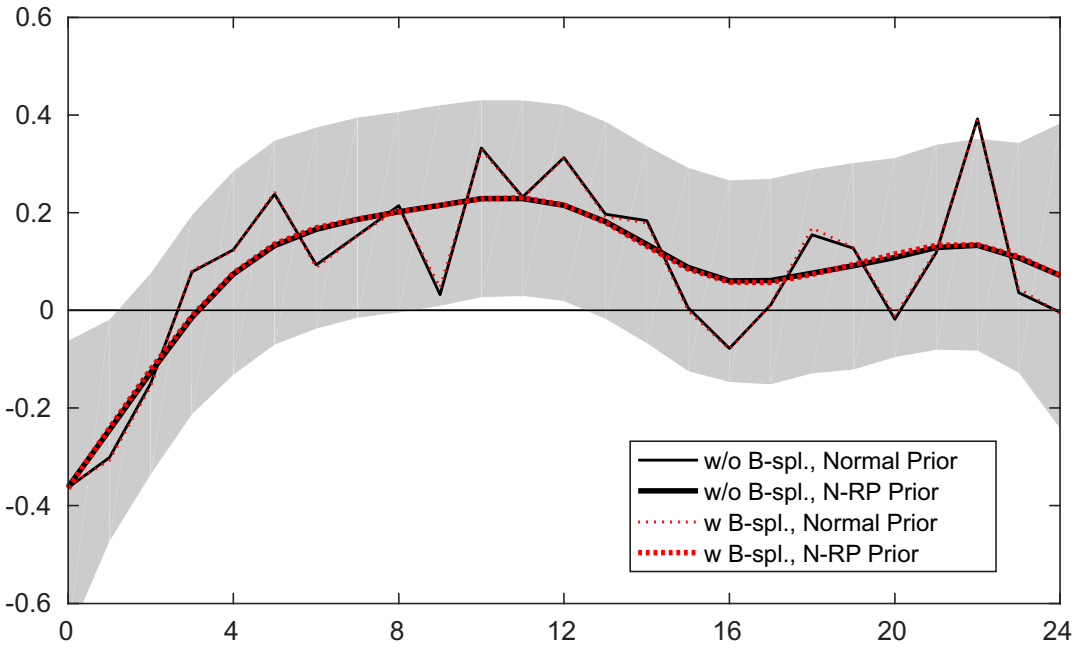
We then compared the fitness of these estimates based on the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and the Watanabe–Akaike information criterion (WAIC) (Watanabe, 2010). Table 1.3 reports on both criteria for different specifications (reported values are on the deviance scale; the smaller, the better). Specifications including the roughness penalty priors outperform the normal prior in predictive accuracy regardless of the fitness measure, while the use of a B-spline yields only limited improvement. Both B-spline and the roughness penalty prior enhance fitness, but almost all improvements originate from the latter.[10] This

1/120}. The added weights are too small to affect the estimate $(1/120 \approx 0.0083)$.

[9]The time series of monetary policy shocks is from Yuriy Gorodnichenko's website (https://eml.berkeley.edu//~ygorodni/index.htm).

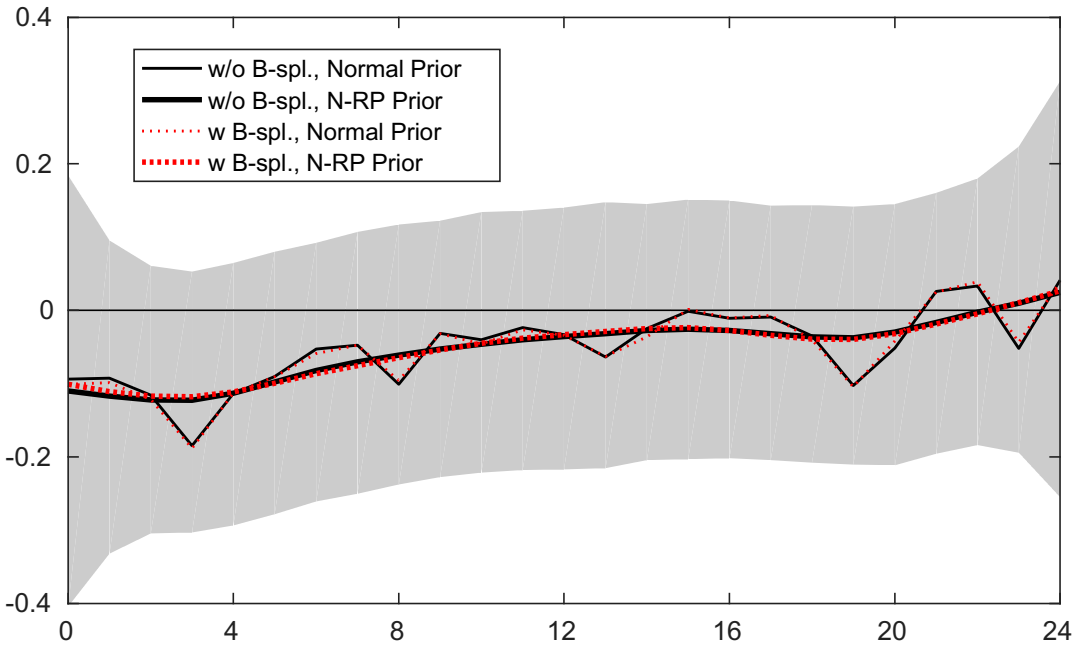
[10]Both the DIC and WAIC are asymptotically related to the AIC. Thus, one might consider evaluating the statistical significance of the difference in the values of the criteria of two models by applying a rule of thumb that is originally proposed to Bayes factor. As Burnham and Anderson (2004) describe, the AIC can be interpreted as an approximation of the log marginal likelihood of a model under a "savvy" prior that is a function of sample size and the number of model parameters. According to Jeffreys's (1961) rule of thumb, the statistical significance of the difference between two models is "weak" if the difference in the AIC/DIC/WAIC is 0-2, "positive" if 2-6, "strong" if 6-10, or "very strong" if >10 (see also Raftery, 1995). When this rule of thumb is directly applied to Table 1.3, one might be able to interpret the results as follows: the statistical significance of the differences related to the prior choice is "very strong," and the significance of the differences attributable to the use of B-splines is "weak"

Figure 1.2: Response of industrial production to monetary policy shocks



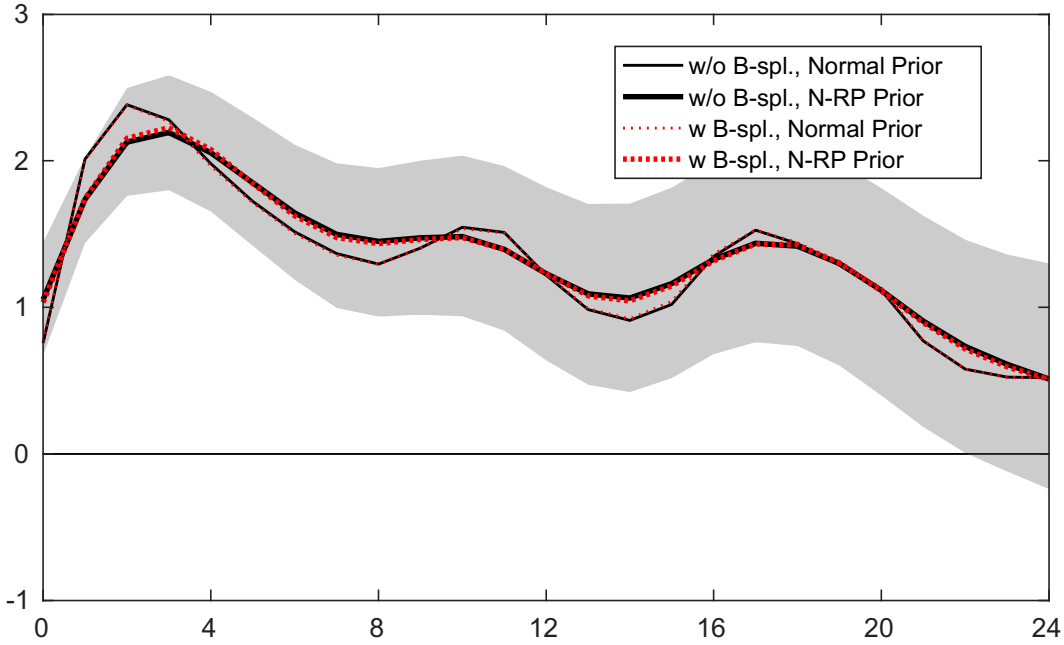
Note: The thin solid line traces the posterior mean for a specification with no B-spline and the normal prior. The thick solid line traces the posterior mean for a specification using no B-spline and the N-RP prior, and the shaded area indicates the corresponding 90% credible set. The thin dotted line traces the posterior mean for a specification with B-splines and the normal prior. The thick dotted line traces the posterior mean for a specification with B-splines and the N-RP prior.

Figure 1.3: Response of inflation to monetary policy shocks



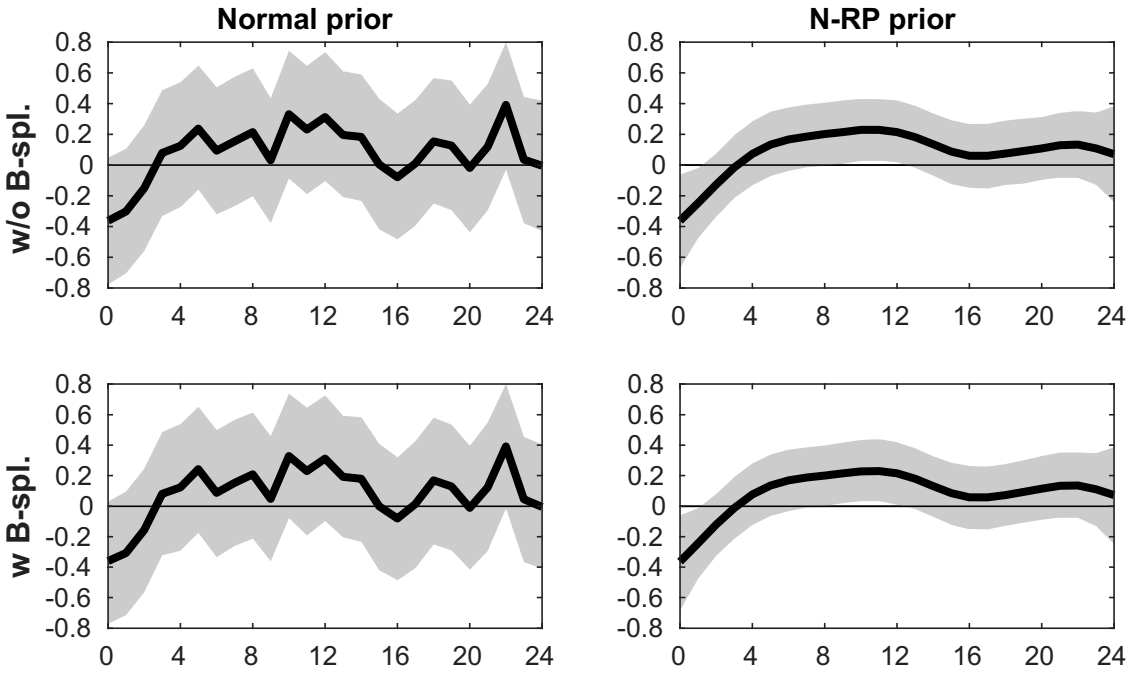
Note: The thin solid line traces the posterior mean for a specification with no B-spline and the normal prior. The thick solid line traces the posterior mean for a specification using no B-spline and the N-RP prior, and the shaded area indicates the corresponding 90% credible set. The thin dotted line traces the posterior mean for a specification with B-splines and the normal prior. The thick dotted line traces the posterior mean for a specification with B-splines and the N-RP prior.

Figure 1.4: Response of fed funds rate to monetary policy shocks



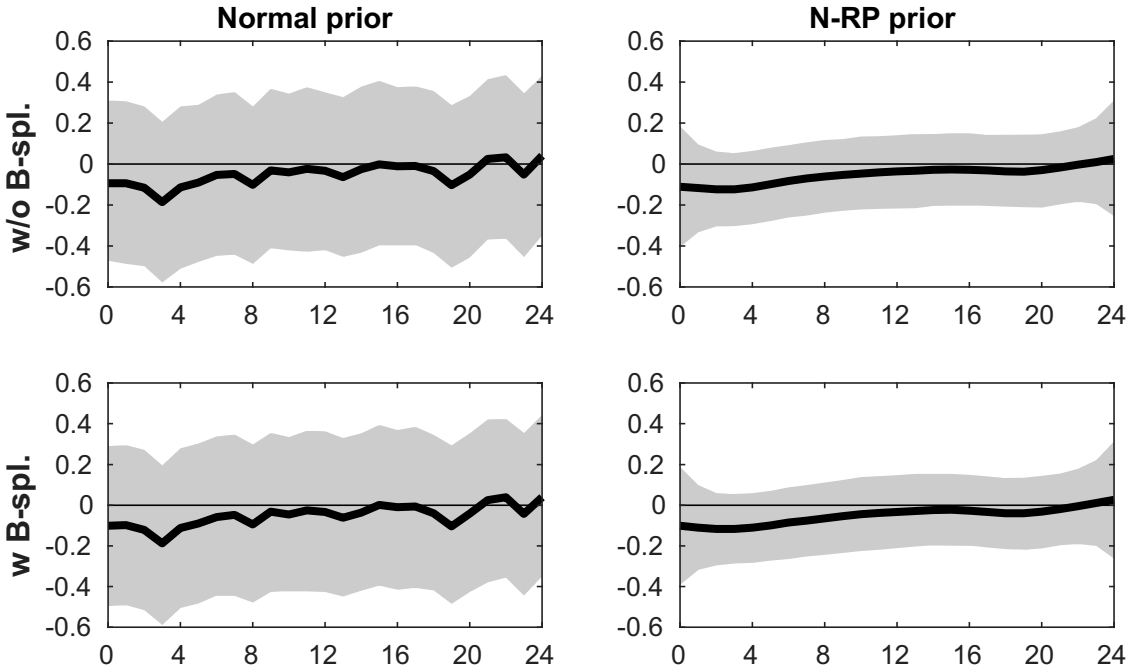
Note: The thin solid line traces the posterior mean for a specification with no B-spline and the normal prior. The thick solid line traces the posterior mean for a specification using no B-spline and the N-RP prior, and the shaded area indicates the corresponding 90% credible set. The thin dotted line traces the posterior mean for a specification with B-splines and the normal prior. The thick dotted line traces the posterior mean for a specification with B-splines and the N-RP prior.

Figure 1.5: Response of industrial production to monetary policy shocks



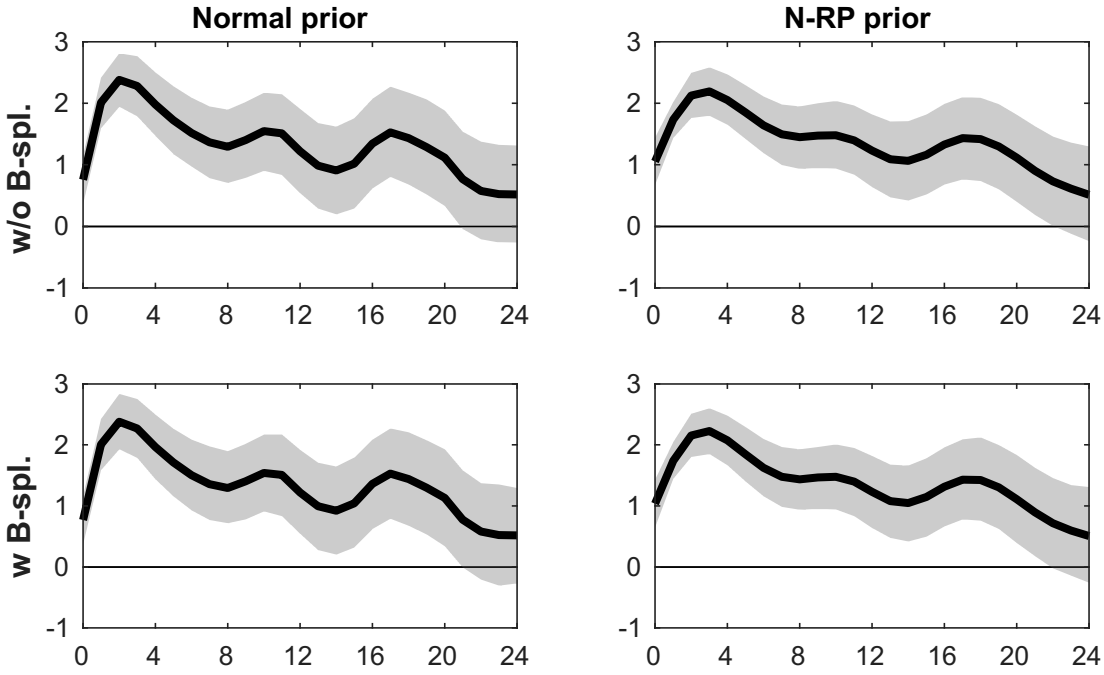
Note: The thick lines trace the posterior mean. The shaded area indicates the 90% credible set.

Figure 1.6: Response of inflation to monetary policy shocks



Note: The thick lines trace the posterior mean. The shaded area indicates the 90% credible set.

Figure 1.7: Response of the Fed funds rate to monetary policy shocks



Note: Thick lines trace the posterior mean. The shaded area indicates the 90% credible set.

finding supports our simulation results. When the B-spline is not used, the posterior simulation takes 26 minutes to generate 50,000 draws; when it is used, the same simulation takes 66 hours. Considering the higher computational cost, use of a B-spline would not be out of proportion to the benefit for many applications (Table 1.3).

1.5 Comparison with Existing Approaches

Recent frequentist approaches to estimate smooth impulse response are closely related to ours in that their objective functions have forms similar to the posterior densities we present, i.e., the sum of a log Gaussian likelihood and a penalty term. From this perspective, Barnichon and Brownlees (2019) can be seen as a frequentist counterpart to our approach with both B-spline expansions and roughness penalty priors. Their objective function is written in our notation as

$$\hat{\boldsymbol{\vartheta}} = \arg\min \left\| \boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\vartheta} \right\|^2 + \boldsymbol{\vartheta}^\top \left(\tilde{\tau} \boldsymbol{I}_J \otimes \boldsymbol{D}^\top \boldsymbol{D} \right) \boldsymbol{\vartheta}.$$

They propose to selecting a (scalar) smoothing parameter $\tilde{\tau}$ using a k-fold cross validation. Barnichon and Brownlees's (2019) approach bears only one smoothing parameter, rendering it less flexible than ours. Figures 1.8, 1.9, and 1.10 plot the posterior estimates of the (global) smoothing parameters for the real data considered in Section 1.4. As evident from these figures, the posterior estimates of the smoothing parameters are significantly different from covariate to covariate. Having single smoothing parameter seems implausible in practice.

or "positive" when the Normal prior is used while it is "very strong" when the N-RP prior is used.

Table 1.3: Comparison of fitness

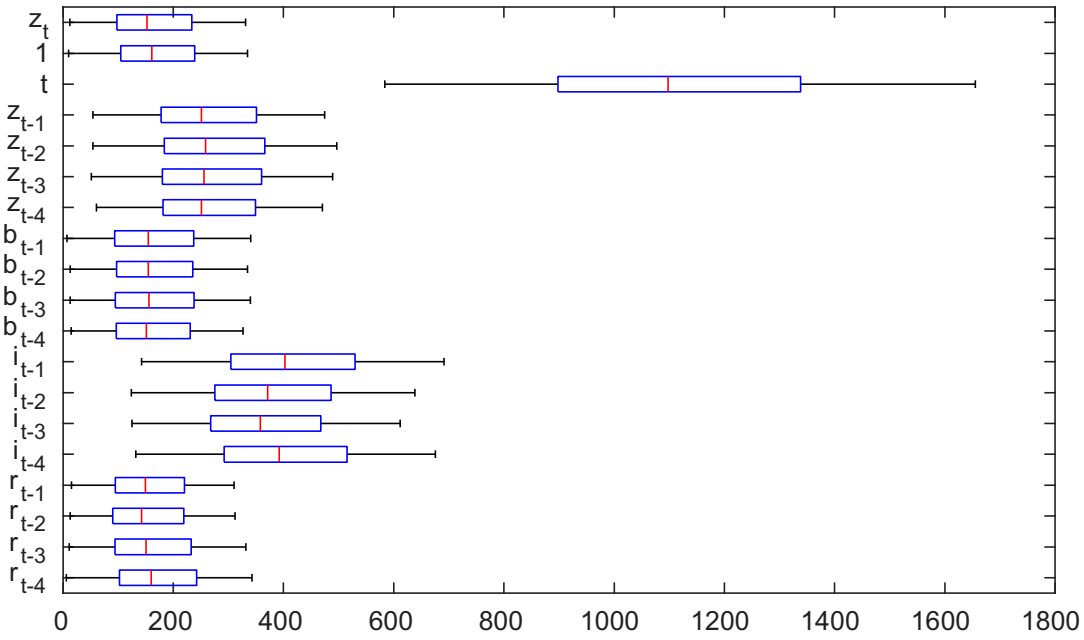
(a) Industrial production	Normal prior		N-RP prior	
	DIC	WAIC	DIC	WAIC
without B-spline	32,585	31,894	31,872	31,431
with B-spline	32,583	31,894	31,854	31,423

(b) Inflation	Normal prior		N-RP prior	
	DIC	WAIC	DIC	WAIC
without B-spline	29,955	29,073	29,321	28,675
with B-spline	29,950	29,068	29,309	28,647

(c) Fed funds rate	Normal prior		N-RP prior	
	DIC	WAIC	DIC	WAIC
without B-spline	35,457	34,901	34,756	34,455
with B-spline	35,461	34,902	34,732	34,445

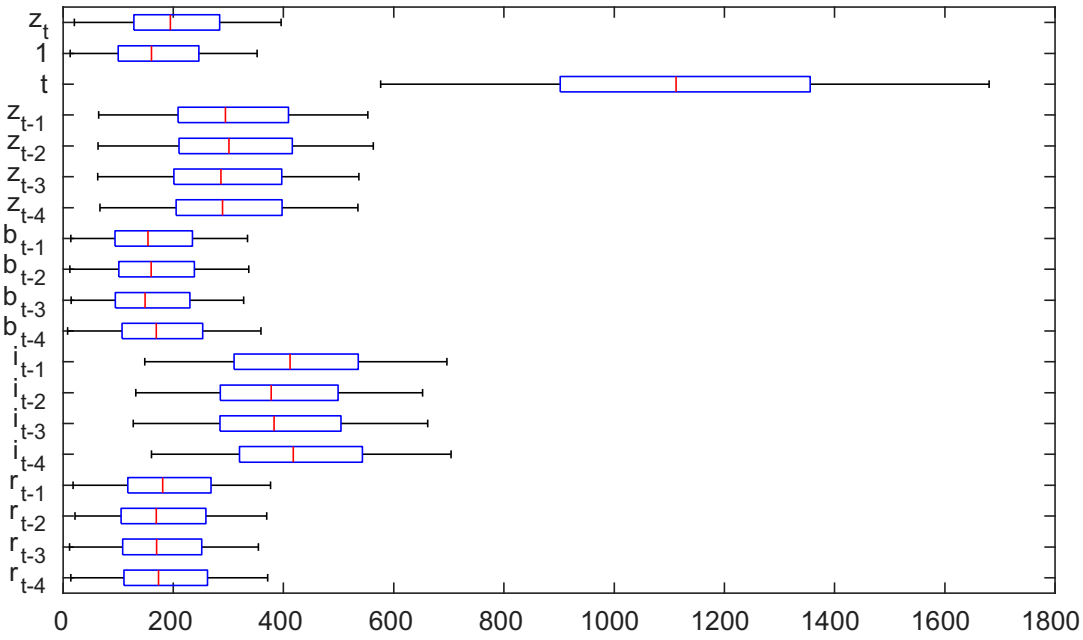
Note: Values of the DIC (deviance information criterion) and WAIC (Watanabe-Akaike information criterion) under different specifications are reported. All values are on the deviance scale.

Figure 1.8: Posterior of smoothing parameter: industrial production



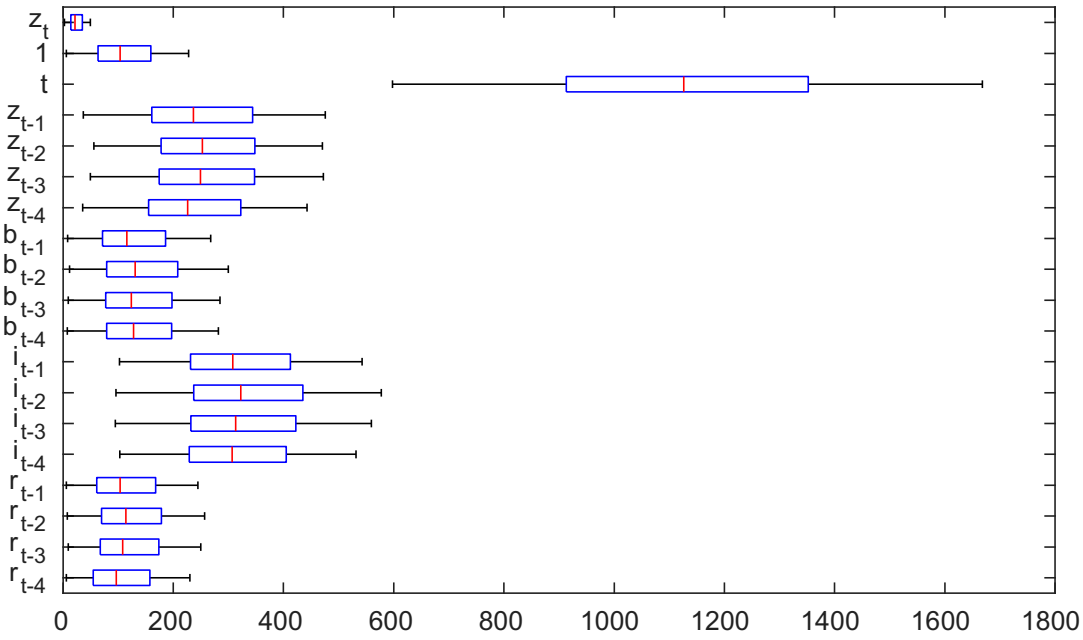
Note: The lines within the boxes denote the posterior median, the edges of the boxes denote the 25th and 75th percentiles of the posterior sample, and the end points of the solid line denote the 5th and 95th percentiles of the posterior sample.

Figure 1.9: Posterior of smoothing parameter: inflation



Note: The lines withn the boxes denote the posterior median, the edges of the boxes denote the 25th and 75th percentiles of the posterior sample, and the end points of the solid line denote the 5th and 95th percentiles of the posterior sample.

Figure 1.10: Posterior of smoothing parameter: Fed funds rate



Note: The lines withn the boxes denote the posterior median, the edges of the boxes denote the 25th and 75th percentiles of the posterior sample, and the end points of the solid line denote the 5th and 95th percentiles of the posterior sample.

El-Shagi's (2019) approach can be regarded as a frequentist version of a model using the N-RP priors and no B-spline expansion. His estimator is written in our notation as

$$\hat{\boldsymbol{\theta}} = \arg\min \left\| \boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\theta} \right\|^2 + \boldsymbol{\theta}^\top \left[\boldsymbol{D}^\top \boldsymbol{D} \otimes \text{diag}\left(\tau_1, ..., \tau_J\right) \right] \boldsymbol{\theta}.$$

This boilds down to a least squares estimator of $\boldsymbol{\theta}$ for an extended model specified by

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0}_{H-r+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_H \otimes \boldsymbol{X} \\ \boldsymbol{D} \otimes \text{diag}\left(\tau_1^{1/2}, ..., \tau_J^{1/2}\right) \end{pmatrix} \boldsymbol{\theta} + \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{u}^* \end{pmatrix},$$

where $\boldsymbol{u}^*$ is an $(H - r + 1)$-dimensional vector of pseudo residuals generated by the penalty term. Let $\boldsymbol{\Sigma}^*$ denote the covariance matrix of $\boldsymbol{u}^*$. Given $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^*$, a generalized least squares (GLS) estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \left[\boldsymbol{\Sigma}^{-1} \otimes \left(\boldsymbol{X}^\top \boldsymbol{X} \right) + \tilde{\boldsymbol{D}}^\top \left(\boldsymbol{\Sigma}^* \right)^{-1} \tilde{\boldsymbol{D}} \right]^{-1} \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^\top \right) \boldsymbol{y}, \tag{1.8}$$

$$\tilde{\boldsymbol{D}} = \boldsymbol{D} \otimes \text{diag}\left(\tau_1^{1/2}, ..., \tau_J^{1/2}\right).$$

He chooses $r = 2$, restricts $\boldsymbol{\Sigma}$ to be diagonal and set $\boldsymbol{\Sigma}^*$ to a submatrix of $\boldsymbol{\Sigma}$, that is,

$$\boldsymbol{\Sigma} = \text{diag}\left(\sigma_{0,0}^2, ..., \sigma_{H,H}^2\right), \quad \boldsymbol{\Sigma}^* = \text{diag}\left(\sigma_{1,1}^2, ..., \sigma_{H-1,H-1}^2\right) \otimes \boldsymbol{I}_J.$$

As $\boldsymbol{\Sigma}$ is unknown, the parameters are estimated through a feasible GLS procedure. First, an ordinary least squares (OLS) estimate $\hat{\boldsymbol{\theta}}_{OLS}$ is obtained, and then $\hat{\boldsymbol{\Sigma}}_{OLS}$ is computed using the realized residuals. Second, using $\hat{\boldsymbol{\Sigma}}_{OLS}$, a first-stage GLS estimate $\hat{\boldsymbol{\theta}}_{GLS,1}$ is computed as (1.8) and compute $\hat{\boldsymbol{\Sigma}}_{GLS,1}$ using the obtained realized residuals. Lastly, using $\hat{\boldsymbol{\Sigma}}_{GLS,1}$, a second-stage (final) GLS estimate $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{GLS,2}$ is obtained. He chooses $\boldsymbol{\tau} = \{\tau_1, ..., \tau_J\}$ by minimizing the finite sample corrected Akaike's information criterion (AICc) (Hurvich et al., 1998),

$$AIC_c\left(\boldsymbol{\tau}\right) = -2 \log p\left(\mathcal{D} | \hat{\boldsymbol{\theta}}_{GLS,2}, \hat{\boldsymbol{\Sigma}}_{GLS,1}\right) + 2\delta + \frac{2\delta\left(\delta + 1\right)}{T - \delta - 1},$$

or a variant of the Bayesian information criterion (BICc) analogously defined as the AICc,

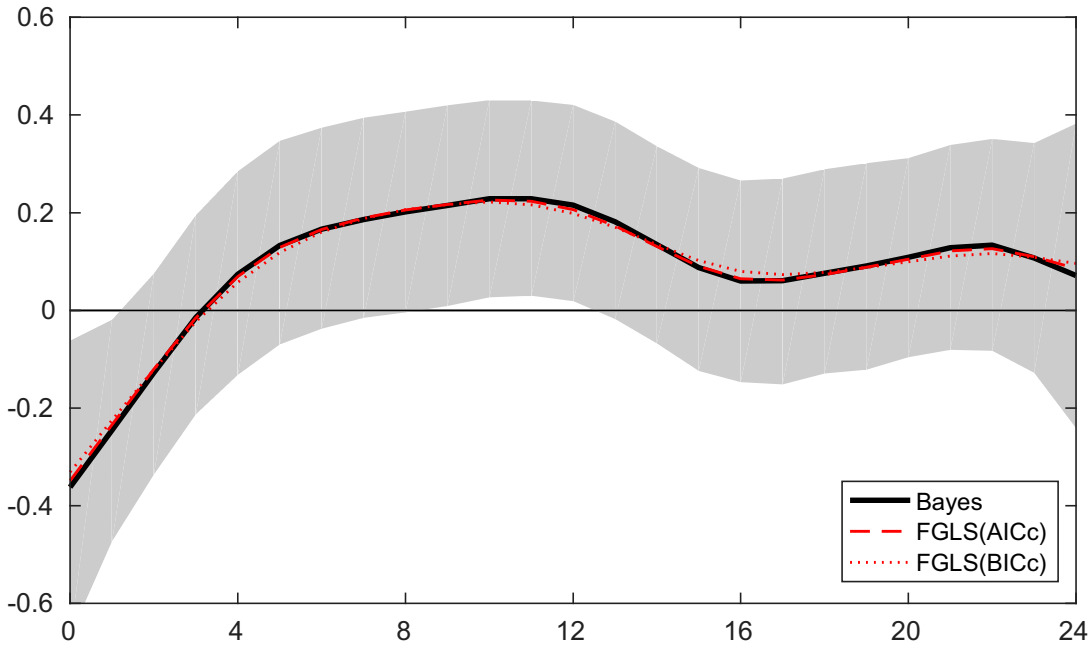
$$BIC_c\left(\boldsymbol{\tau}\right) = -2 \log p\left(\mathcal{D} | \hat{\boldsymbol{\theta}}_{GLS,2}, \hat{\boldsymbol{\Sigma}}_{GLS,1}\right) + \left(\log T\right)\delta + \frac{2\delta\left(\delta + 1\right)}{T - \delta - 1}.$$

δ denotes the effective loss of degrees of freedom (or pseudo dimension of the model) which is defined as the trace of a hat (or projection) matrix $\hat{\boldsymbol{P}}$ with $\hat{\boldsymbol{y}} = \hat{\boldsymbol{P}}\boldsymbol{y}$, that is, $\delta = \text{tr}\left\{\hat{\boldsymbol{P}}\right\}$ with

$$\hat{\boldsymbol{P}} = \left[\boldsymbol{\Sigma}^{-1} \otimes \left(\boldsymbol{X}^\top \boldsymbol{X} \right) + \left(\boldsymbol{D}^\top \left(\boldsymbol{\Sigma}^* \right)^{-1} \boldsymbol{D} \right) \otimes \text{diag}\left(\tau_1, ..., \tau_J\right) \right]^{-1} \left[\boldsymbol{\Sigma}^{-1} \otimes \left(\boldsymbol{X}^\top \boldsymbol{X} \right) \right].$$

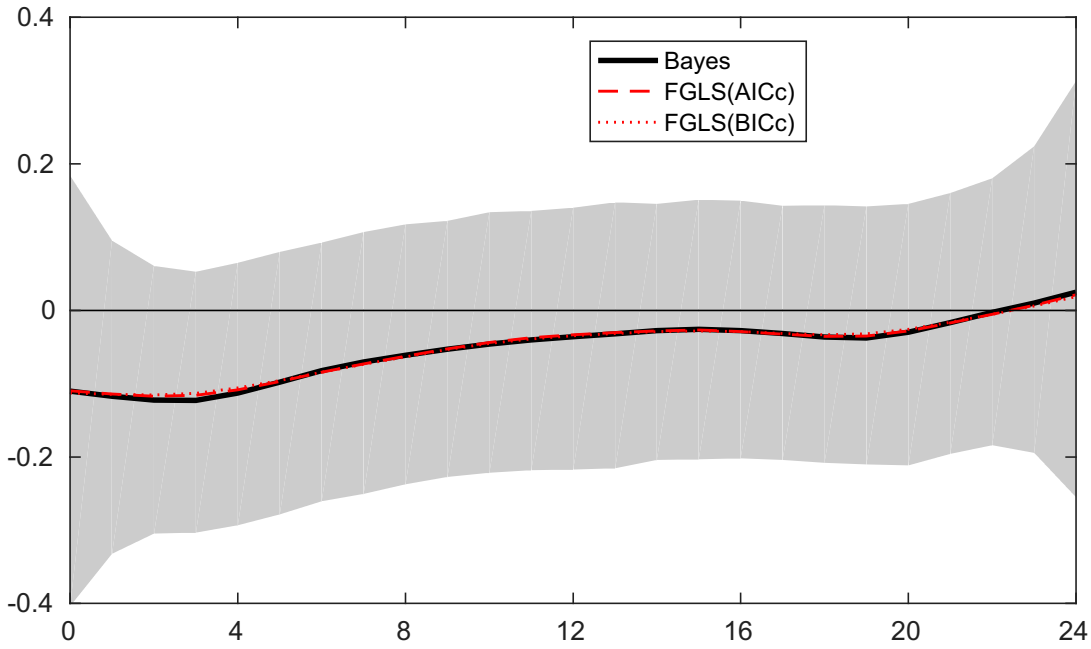
In terms of nonparametric regressions, δ measures the effective number of zero-th order polynomial bases defined over the projection horizons $\mathcal{H}$, $\boldsymbol{x}_t, ..., \boldsymbol{x}_t$. This approach can be crudely interpreted as a maximum a posteriori estimation of a local projection using a (non-adaptive) roughness penalty prior of $\boldsymbol{\theta}$, a "prior" of $\boldsymbol{\tau}$ generated from the AICc or BICc, and a non-informative prior of $\boldsymbol{\Sigma}$. Figures 1.11, 1.12, and 1.13 represent estimated IRFs of monetary policy shocks using El-Shagi's (2019) approach along with the default Bayesian estimates. The IRFs obtained by both approaches are fairly comparable.

Figure 1.11: Response of industrial production to monetary policy shocks: frequentist and Bayesian approaches



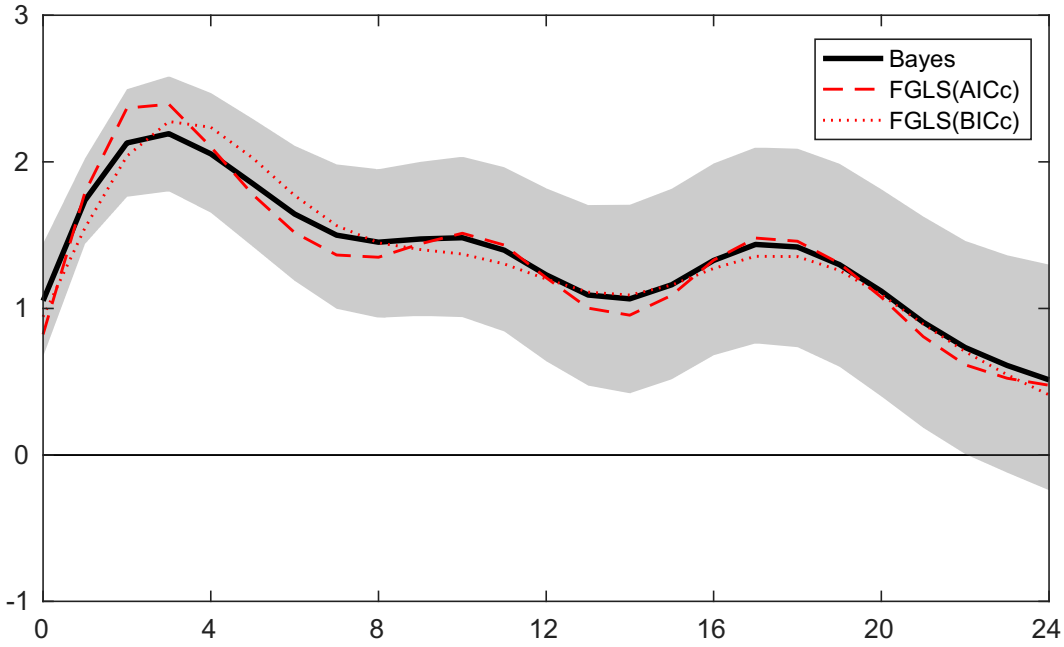
Note: The solid line traces the posterior mean for a Bayesian approach, and the shaded area indicates the corresponding 90% credible set. The dashed line traces an estimate for a frequentist approach with the AICc. The dotted line traces an estimate for a frequentist approach with the BICc.

Figure 1.12: Response of inflation to monetary policy shocks: frequentist and Bayesian approaches



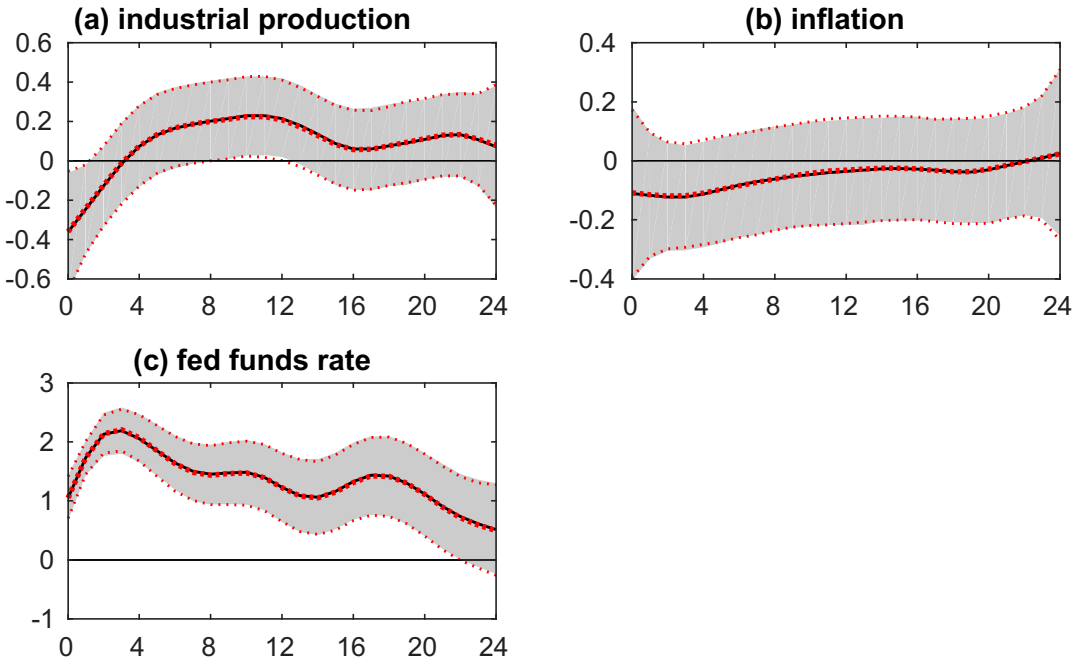
Note: The solid line traces the posterior mean for a Bayesian approach, and the shaded area indicates the corresponding 90% credible set. The dashed line traces an estimate for a frequentist approach with the AICc. The dotted line traces an estimate for a frequentist approach with the BICc.

Figure 1.13: Response of fed funds rate to monetary policy shocks: frequentist and Bayesian approaches



Note: The solid line traces the posterior mean for a Bayesian approach, and the shaded area indicates the corresponding 90% credible set. The dashed line traces an estimate for a frequentist approach with the AICc. The dotted line traces an estimate for a frequentist approach with the BICc.

Figure 1.14: Response to monetary policy shocks: inferred and fixed smoothing parameters



Note: The solid black line traces the posterior mean for a model with inferred smoothing parameters. The bold dotted line traces the posterior mean for a model with fixed smoothing parameters. The shaded area indicates a 90% credible set for a model with inferred smoothing parameters. The thin dotted line indicates 90% credible set for a model with fixed smoothing parameters.

We can identify three advantages of the proposed Bayesian approach over the frequentist approaches. First, our approach can provide credible intervals in a consistent and straightforward manner. In contrast, at this time, the frequentist approaches have no statistically grounded method to estimate confidence intervals; Barnichon and Brownlees (2019) mention a heuristic method, while El-Shagi (2019) does not discuss a method to estimate confidence intervals.

Second, while frequentist approaches fix smoothing parameters before inference by cross validation (Barnichon and Brownlees, 2019) or penalized optimization (El-Shagi, 2019), our approach infers them using priors, allowing us to systematically consider uncertainty in the smoothness of an impulse response (and other sequences of coefficients). The quantitative significance of this conceptual advantage depends on context. We re-estimated the model with the (global) smoothing parameters fixed to the posterior medians for the default specifications. As shown in Figure 1.14, for the real data in Section 1.5, there is no significant difference.

Third, the proposed approach has better finite-sample performance. We compared El-Shagi's (2019) approach to ours through a simulation study. The simulation setup is the same as that of the linear IRF in Section 1.4. [11] We examine specifications with unrestricted and diagonal covariance matrices for both El-Shagi's (2019) and our Bayesian approaches. A half-t prior is used for the diagonal elements in Σ, denoted by $\sigma^2_{(h)}$, $h = 0, ..., H$. It is derived from the HIW prior:

[11] We minimize the information criteria using the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm with bounds (Byrd et al., 1995), using a Matlab routine `minConf_TMP.m` written by Mark Schmidt. (https://www.cs.ubc.ca/~schmidtm/Software/minConf.html)

$$\sigma^2_{(h)}|\phi \sim \mathcal{IG}\left(\frac{\zeta}{2},\ \zeta\phi_{(h)}\right), \quad \phi_{(h)} \sim \mathcal{G}\left(\frac{1}{2}, \upsilon\right), \quad h = 0, ..., H.$$

We choose $\upsilon = 0.01$ as in Section 1.3. The result is summarized in Table 1.4.[12] In line with the simulation study by El-Shagi (2019), finite sample performance of the BICc is comparable to or slightly better than the AICc. The Bayesian approach obtained smaller MSE on average than the FGLS approach for both covariance specifications. For the frequentist approach, specifications with diagonal covariances obtained smaller MSEs than those with unrestricted covariances, whereas for the Bayesian approach, the situation is the opposite. It is difficult to identify a specific reason behind this twisted simulation result. The plug-in estimator of $\boldsymbol{\Sigma}$ employed in the frequentist approach might not work well for the short time series.[13] The overall winner was the Bayesian approach with unrestricted covariance. Because residuals in a local projection are strongly correlated by construction, assuming independence between them is inappropriate.

1.6 Conclusion

This study developed a fully Bayesian approach to estimate local projections using roughness penalty priors. It is also considered a specification involving a B-spline basis function expansion. Monte Carlo experiments have demonstrated that both B-splines and the roughness penalty priors improve statistical efficiency, however, almost all the improvements originate from the latter. Applying our proposed method to an analysis of monetary policy in the United States shows that the roughness penalty priors successfully smooth posterior estimates of the impulse response functions, and can improve the predictive accuracy of local projections.

This study addresses one of the two disadvantages of local projections, compared with the standard statistical framework that includes VAR, namely, that of less statistical efficiency. The other disadvantage that the exogenous variable is identified ex ante can be resolved by a two-stage regression approach, as in Aikman et al. (2016). Constructing a Bayesian counterpart to this line of approach has not been studied. In addition, it is potentially beneficial to develop more robust approaches than ours: for example, a choice of hyperparameters, heteroskedasticity and autocorrelations in errors, and so on. This study provides a first step for further developments of Bayesian local projections.

Appendix

A.1 Inverse Wishart Prior for $\boldsymbol{\Sigma}$

This section compare an inverse Wishart prior with the hierarchical inverse Wishart prior we propose. An inverse Wishart prior is specified by

$$\boldsymbol{\Sigma} \sim \mathcal{IW}\left(\boldsymbol{I}_H,\ \xi + H\right),$$

where ξ is a prefixed hyperparameter. The corresponding conditional posterior is

$$\boldsymbol{\Sigma}|- \sim \mathcal{IW}\left(\boldsymbol{\Xi} + \boldsymbol{U}^\top\boldsymbol{U},\ H + \xi + T + 1\right).$$

[12] We also considered Jeffreys prior for $\sigma^2_{(h_i)}$ and obtained almost the same result for the half-t prior (thus, it is not reported).

[13] As T increases (e.g., $T = 500$), the relative performance of the FGLS approach with unrestricted covariance becomes comparable with that with diagonal covariance (not reported).

Table 1.4: Results of the Monte Carlo simulation: comparison to the frequentist approach

T	Prior	Penalty/Prior	MSE		Length	
			full	diagonal	full	diagonal
50	FGLS	AICc	0.221	0.162	–	–
		BICc	0.166	0.110	–	–
	Bayes	N-RP	0.131	0.151	0.432	0.298
100	FGLS	AICc	0.100	0.091	–	–
		BICc	0.100	0.091	–	–
	Bayes	N-RP	0.067	0.076	0.309	0.213

Note: MSE denotes the mean squared error. Length denotes the length of the 90% credible interval.

Table 1.5: Results of the Monte Carlo simulation: inverse Wishart prior for $\boldsymbol{\Sigma}$

T	Prior	ξ	MSE	Length
50	HIW	–	0.131	0.432
	IW	0	0.227	0.299
		1	0.228	0.297
		2	0.229	0.296
100	HIW	–	0.067	0.309
	IW	0	0.090	0.217
		1	0.090	0.216
		2	0.090	0.215

Note: MSE denotes the mean squared error. Length denotes the length of the 90% credible interval.

This prior is more popular and simpler than the HIW prior. A simulation environment is the same with the linear case in Section 1.3. The simulation result is summarized in Table 1.5. Though we considered $\xi = 0, 1, 2$, the choice of ξ had almost no effect on the result. Compared to the HIW prior, the inference using the inverse Wishart prior tended to obtain a shorter length, larger MSE. Based on the simulation result, we prefer the HIW prior to the inverse Wishart prior.

A.2 Sensitivity Check

We investigated the sensitivity of the inference to the choice of hyperparameters. A simulation setting is adapted from the linear case in Section 1.3. We considered $\nu = 0.1, 0.01, 0.001, 0.0001$ for models with the N-RP prior and no B-spline. Table 1.6 shows the results, wherein two items are noteworthy. First, as long as ν is set between 0.1 and 0.0001, an estimation using the N-RP prior is more efficient than one using the normal prior. Second, we see a bias-variance trade-off: as ν increases (i.e., shrinkage toward 0), an estimator becomes more efficient but less robust. Table 1.7 reports the results for $\eta = 1, 0.5, 0.1, 0.01$. We can see that as long as it is chosen within this range, η does not significantly affect the performance. Even when η changes, the A-RP prior cannot beat the N-RP prior. Table 1.8 includes the results for $\upsilon = 0.1, 0.01, 0.001, 0.0001$. There is almost no difference between the results for the alternative specifications, which implies that $\upsilon = 0.01$ is sufficiently small for the synthetic data. The results for $r = 1, 2, 3, 4$ are shown in Table 1.9. While this experiment indicates that $r = 1$ was the best choice, the

Table 1.6: Results of the Monte Carlo simulation: sensitivity to the choice of ν

T	Prior	ν	MSE	Length
50	Normal	–	0.542	0.976
	N-RP	0.1	0.180	0.526
		0.01	0.131	0.432
		0.001	0.099	0.367
		0.0001	0.079	0.324
100	Normal	–	0.243	0.599
	N-RP	0.1	0.090	0.378
		0.01	0.067	0.309
		0.001	0.053	0.261
		0.0001	0.043	0.228

Note: MSE denotes the mean squared error. Length denotes the length of the 90% credible interval.

Table 1.7: Results of the Monte Carlo simulation: sensitivity to the choice of η

T	Prior	η	MSE	Length
50	Normal	–	0.542	0.976
	N-RP	–	0.131	0.432
	A-RP	1.0	0.143	0.454
		0.5	0.150	0.468
		0.1	0.163	0.494
		0.01	0.157	0.481
100	Normal	–	0.243	0.599
	N-RP	–	0.067	0.309
	A-RP	1.0	0.072	0.323
		0.5	0.074	0.331
		0.1	0.078	0.344
		0.01	0.074	0.332

Note: MSE denotes the mean squared error. Length denotes the length of the 90% credible interval.

approach using the N-RP prior consistently outperformed that using the normal prior.

Table 1.8: Results of the Monte Carlo simulation: sensitivity to the choice of υ

T	Prior	υ	MSE	Length
50	Normal	–	0.542	0.976
	N-RP	0.1	0.130	0.438
		0.01	0.131	0.432
		0.001	0.131	0.432
		0.0001	0.131	0.432
100	Normal	–	0.243	0.599
	N-RP	0.1	0.067	0.314
		0.01	0.067	0.309
		0.001	0.067	0.309
		0.0001	0.067	0.309

Note: MSE denotes the mean squared error. Length denotes the length of the 90% credible interval.

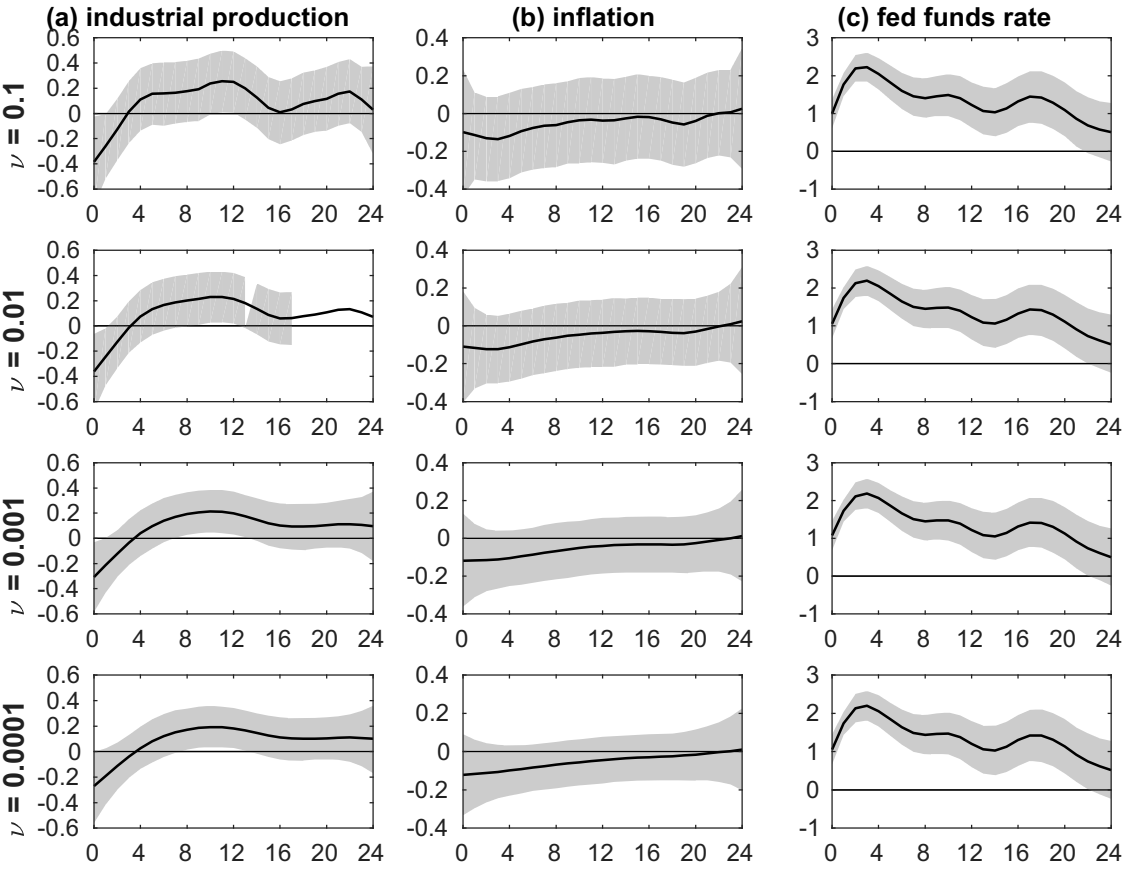
Table 1.9: Results of the Monte Carlo simulation: sensitivity to choice of r

T	Prior	r	MSE	Length
50	Normal	–	0.542	0.976
	N-RP	1	0.077	0.381
		2	0.131	0.432
		3	0.169	0.471
		4	0.197	0.503
100	Normal	–	0.243	0.599
	N-RP	1	0.046	0.287
		2	0.067	0.309
		3	0.081	0.328
		4	0.090	0.344

Note: MSE denotes the mean squared error. Length denotes the length of the 90% credible interval.

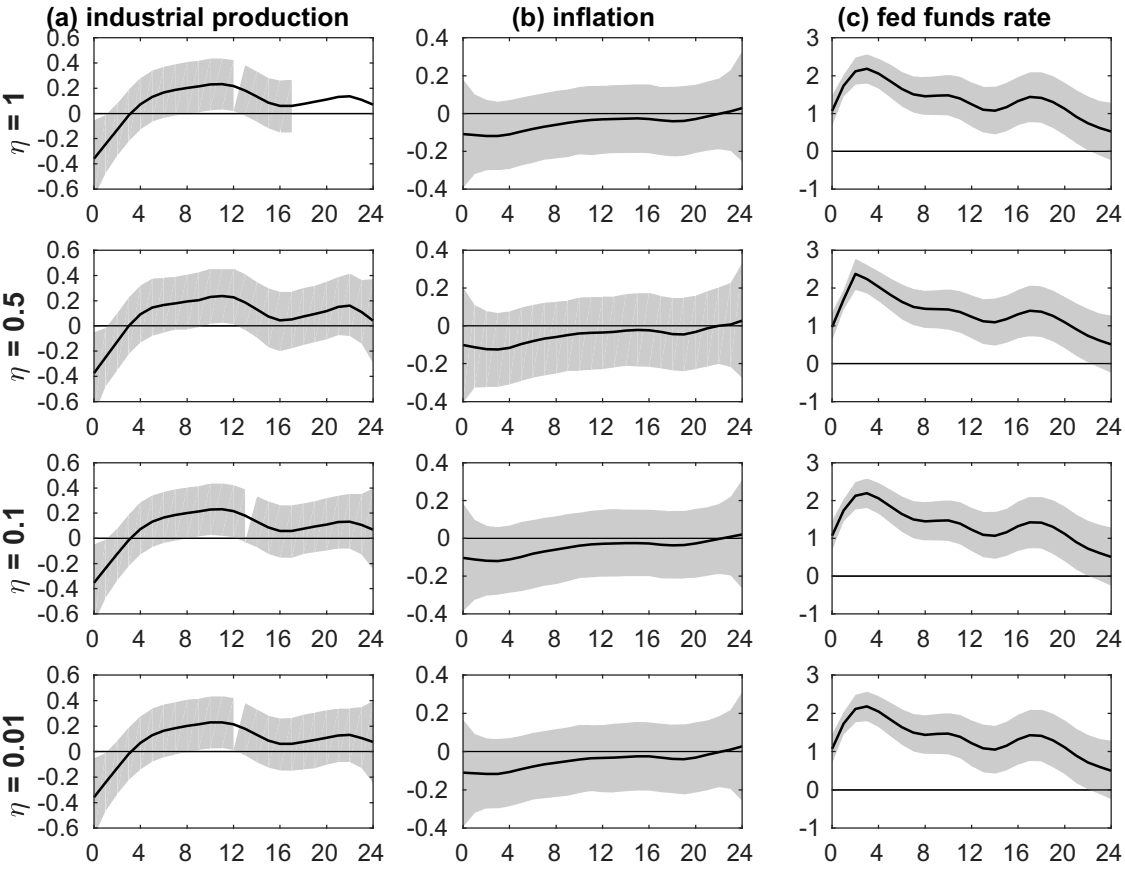
We also conducted a series of sensitivity checks for the real data application in Section 1.5 using the same alternative hyperparameter values above. Figures 1.15 to 1.18 depict the results for ν, η, υ, and r, respectively. We see that choice of the hyperparameters did not affect the shape of the impulse response functions.

Figure 1.15: Response to monetary policy shocks: sensitivity to ν



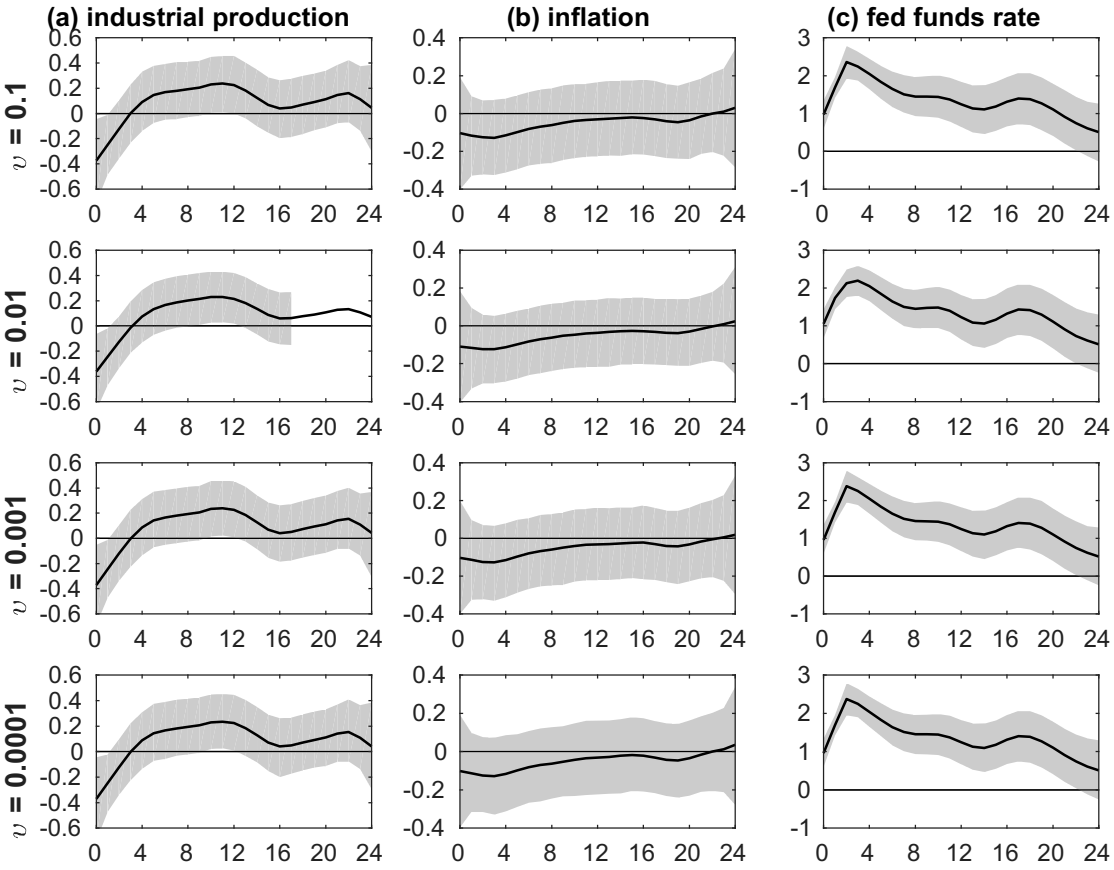
Note: Thick lines trace the posterior mean. Shaded area indicates 90% credible set.

Figure 1.16: Response to monetary policy shocks: sensitivity to η



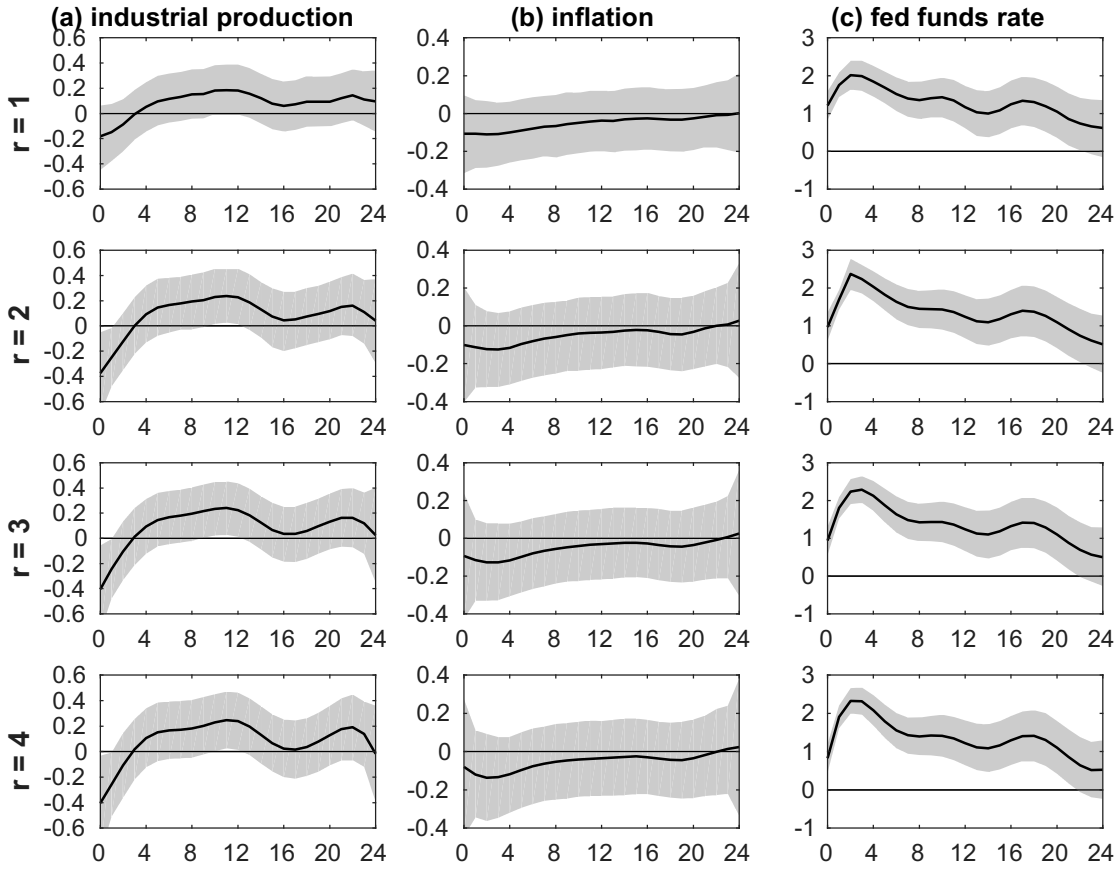
Note: Thick lines trace the posterior mean. Shaded area indicates 90% credible set.

Figure 1.17: Response to monetary policy shocks: sensitivity to υ



Note: Thick lines trace the posterior mean. Shaded area indicates 90% credible set.

Figure 1.18: Response to monetary policy shocks: sensitivity to r



Note: Thick lines trace the posterior mean. Shaded area indicates 90% credible set.

Chapter 2

Adaptive MCMC for Generalized Method of Moments with Many Moment Conditions

2.1 Introduction

The generalized method of moments (GMM) is a widely used statistical framework (Hansen, 1982; Hall, 2005). Under GMM, unknown parameters are estimated via a set of moment conditions. A parameter estimate is obtained by minimizing a GMM criterion constructed as a quadratic form and composed of the sample mean of a vector-valued function that represents the moment conditions and a weighting matrix. While GMM uses only lower-order moments, thus being statistically less efficient than full-information methods such as the maximum likelihood method, it has many advantages, including robustness to model misspecification, nonparametric treatment of heteroskedasticity, and computational simplicity.

This study focuses on the Bayesian version of GMM. A GMM criterion can be viewed as a quasi-likelihood, being theoretically equivalent to the Laplace approximation of the true likelihood around its mode (Chernozhukov and Hong, 2003). Exploiting this feature, one can conduct a (quasi-)Bayesian inference by replacing true likelihood with a GMM criterion, as discussed by, for example, Kim (2002); Yin (2009).[1] Posterior draws from a quasi-posterior density (product of quasi-likelihood and prior density) can be simulated using standard Bayesian Markov Chain Monte Carlo (MCMC) techniques, such as the Metropolis-Hastings algorithm. In this study, we call this inferential approach Bayesian GMM, in contradistinction to classical GMM.

A GMM criterion has many moment conditions for applications, making the estimator considerably unreliable. There are cases where the number of moment conditions can be large, including dynamic panel models (e.g., Arellano and Bond, 1991; Blundell and Bond, 1998; Roberts and Rosenthal, 2009; Vieira et al., 2012), instrumental variable methods (e.g., Chernozhukov and Hansen, 2005, 2013), and identification through heteroskedasticity (Lewbel, 2012).

The literature on classical GMM proposes several provisions to the problem such as systematic moment selection (Andrews, 1999; Andrews and Lu, 2001; Hall and Peixe, 2003; Hall et al., 2007; Okui, 2009; Donald et al., 2009; Canay, 2010; DiTraglia, 2016; Chang and DiTraglia, 2018), averaging (Chen et al., 2016), and shrinkage estimation (Liao, 2013; Fan and Liao, 2014; Cheng and Liao, 2015; Caner et al., 2018). On the contrary, the literature on Bayesian GMM has paid scant attention to the problem, although remedies tailored to clas-

[1] See also Belloni and Chernozhukov (2009); Li and Jiang (2016) for a discussion of theoretical properties.

sical GMM are not straightforwardly applicable to Bayesian GMM for two reasons. First, they are two-stage procedures in which the final estimate is computed based on the first estimate with the identity weighting matrix. However, such a strategy is not feasible in Bayesian GMM; because the relative contributions of a GMM criterion (quasi-likelihood) and a prior density to the quasi-posterior depend on the weighting matrix, the mode of a quasi-posterior under the identity weighting matrix is not consistent with that under the optimal weighting matrix. Therefore, in Bayesian GMM, a weighting matrix has to be estimated along with the unknown parameters of interest. Second, Bayesian GMM is often used for cases where numerical optimization does not work well for the reason that a GMM criterion is discontinuous in parameters or has many local optima. Therefore, even when a non-informative prior is employed, there are cases where a first-step estimate is not readily available. The purpose of this study is to bridge this gap by proposing a novel method to deal with Bayesian GMM with many moment conditions.

For both classical and Bayesian GMM, choosing a good weighting matrix is a significant issue. It is theoretically optimal to set a weighting matrix to the precision matrix (i.e., the inverse of the covariance matrix) of moment conditions, evaluated based on true parameter values. Since this approach is infeasible in practice, two-step and continuously updated estimators are commonly used in classical GMM (Hansen, 1982; Hansen et al., 1996). By contrast, the literature on Bayesian GMM has paid less attention to the weighting matrix choice. Chernozhukov and Hong (2003), who use the random-walk Metropolis-Hasting algorithm, suggest recomputing the weighting matrix each time a parameter proposal is drawn; a posterior mean of the weighting matrix is supposed to be optimal on average. In this approach, the unknown parameters and a weighting matrix are updated concurrently. Consequently, the surface of the quasi-posterior becomes complicated, making the MCMC algorithm inefficient and unstable. To tackle this problem, Yin et al. (2011) propose an approach they call stochastic GMM, where unknown parameters are updated one by one and the corresponding weighting matrix is also updated accordingly. Their approach improves the numerical stability of the posterior simulator by suppressing changes in the posterior in a single cycle. However, this approach requires so many matrix inversions of the weighting matrix that it is not practical for models with many moment conditions.

There are two difficulties in setting a weighting matrix when the number of moment conditions is large. First, as in classical GMM, the sample estimate of the covariance matrix of the moment conditions is unreliable, and the inversion of the covariance matrix can amplify estimation errors. Second, it is computationally demanding because the inversion of the sample covariance matrix is repeatedly computed. This problem is specific to Bayesian GMM.

In this study, we develop an adaptive MCMC approach to deal with the problem of many moment conditions in Bayesian GMM. The proposal consists of two main contributions. First, we propose estimating the precision matrix of the moment conditions using the nonparametric eigenvalue-regularized precision matrix estimator developed by Lam (2016). This estimator is more numerically stable than the standard estimator, the inverse of a sample covariance matrix. Through a series of Monte Carlo experiments, we show that the proposed approach outperforms existing ones in terms of both statistical and computational efficiency. Second, we propose a random updating of a weighting matrix using the recursive mean of the posterior samples. In our approach, adaptation probabilities are set to decrease exponentially, which ensures the validity of the MCMC algorithm, and significantly saves computational cost.

The rest of this chapter is structured as follows. Section 2.2 introduces the proposed approach. Section 2.3 conducts a simulation study. In Section 2.4, we apply the approach to a real data problem as an example. Section 2.5 concludes the paper with a discussion.

2.2 Method

2.2.1 Setup and challenges

We consider the Bayesian inference of a statistical model by means of a set of moment conditions. Assume that a likelihood function can be approximated by a quasi-likelihood based on a GMM criterion (Hansen, 1982). We call this inferential approach Bayesian GMM (Kim, 2002; Yin, 2009). Given data $\mathcal{D}$ and an L-dimensional parameter $\boldsymbol{\theta}$, a statistical model is estimated through a set of moment conditions represented by a K-dimensional vector of moment functions $\boldsymbol{m}_n(\boldsymbol{\theta}) = (m_{n,1}(\boldsymbol{\theta}), ..., m_{n,K}(\boldsymbol{\theta}))$:

$$E[\boldsymbol{m}_n(\boldsymbol{\theta})] = \mathbf{0}_K.$$

A GMM criterion $v(\boldsymbol{\theta})$ is defined as the quadratic form of the sample mean of $\boldsymbol{m}_n(\boldsymbol{\theta})$, denoted by $\bar{\boldsymbol{m}}(\boldsymbol{\theta})$, and a symmetric positive definite weighting matrix $\boldsymbol{W}$:

$$v(\boldsymbol{\theta}) = \bar{\boldsymbol{m}}(\boldsymbol{\theta})^\top \boldsymbol{W} \bar{\boldsymbol{m}}(\boldsymbol{\theta}),$$

$$\bar{\boldsymbol{m}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{m}_n(\boldsymbol{\theta}),$$

where N is the sample size. For notational convenience, we omit the dependence on $\mathcal{D}$ from functions $\boldsymbol{m}_n(\boldsymbol{\theta})$, $\bar{\boldsymbol{m}}(\boldsymbol{\theta})$, and $v(\boldsymbol{\theta})$. A quasi-lik
as

$$\begin{aligned}
(\quad) &= \left(\frac{2}{N}\right)^{-\frac{K}{2}} \det(\boldsymbol{W})^{\frac{1}{2}} \exp\left[-\frac{N}{2} v(\boldsymbol{\theta})\right] \\
&= \left(\frac{2\pi}{N}\right)^{-\frac{K}{2}} \det(\boldsymbol{W})^{\frac{1}{2}} \exp\left[-\frac{N}{2} \bar{\boldsymbol{m}}(\boldsymbol{\theta})^\top \boldsymbol{W} \bar{\boldsymbol{m}}(\boldsymbol{\theta})\right].
\end{aligned}$$

A GMM criterion can be seen as the Laplace approximation of the negative true likelihood evaluated around the mode (Chernozhukov and Hong, 2003). Given a prior density $p(\boldsymbol{\theta})$, the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$ is approximated as

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx \frac{q(\boldsymbol{\theta}|\mathcal{D})\, p(\boldsymbol{\theta})}{\int q(\boldsymbol{\theta}'|\mathcal{D})\, p(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'}, \tag{2.1}$$

where the denominator is generally unknown but constant. The posterior samples $\boldsymbol{\theta}_{[j]} = \left(\theta_{[j],1}, ..., \theta_{[j],L}\right)^\top$ are drawn from this target density (evaluated up to the normalizing constant) using Bayesian simulation techniques. In this study, we consider using Metropolis-Hastings (MH) algorithm as in previous studies (e.g., Chernozhukov and Hong, 2003; Yin, 2009). Given a current state $\boldsymbol{\theta}$, a single step of a MH algorithm is specified as follows.

1. A proposal $\boldsymbol{\theta}'$ is generated from a proposal kernel $p(\boldsymbol{\theta}'|\boldsymbol{\theta})$.
2. Compute the MH ratio $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta})$ as

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta}'|\mathcal{D})\, p(\boldsymbol{\theta}')\, p(\boldsymbol{\theta}'|\boldsymbol{\theta})}{q(\boldsymbol{\theta}|\mathcal{D})\, p(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{\theta}')}.$$

3. Set the next state to the proposal $\boldsymbol{\theta}^* = \boldsymbol{\theta}'$ with probability of $\alpha\left(\boldsymbol{\theta}', \boldsymbol{\theta}\right) \vee 1$, or set the next state to the current state $\boldsymbol{\theta}^* = \boldsymbol{\theta}$ with probability of $1 - \left(\alpha\left(\boldsymbol{\theta}', \boldsymbol{\theta}\right) \vee 1\right)$.

4. Return the next state $\boldsymbol{\theta}^*$.

As in classical GMM, the statistical efficiency of the Bayesian GMM critically depends on the choice of the weighting matrix $\boldsymbol{W}$. $\boldsymbol{W}$ is optimal when it is set to the precision matrix of the moment conditions based on true parameter values $\boldsymbol{\theta}_0$. This choice is optimal in that it minimizes the Kullback-Leibler divergence of the true data generating process to the set of all asymptotically less restrictive distributions (Li and Jiang, 2016). Let $\boldsymbol{M}\left(\boldsymbol{\theta}\right) = \left(\boldsymbol{m}_1\left(\boldsymbol{\theta}\right), ..., \boldsymbol{m}_N\left(\boldsymbol{\theta}\right)\right)^\top$ denote an N-by-K matrix of the moment functions. The optimal choice of weighting matrix in finite sample is

$$\boldsymbol{W}\left(\boldsymbol{\theta}_0\right) = \left[N^{-1}\boldsymbol{M}\left(\boldsymbol{\theta}_0\right)^\top \boldsymbol{M}\left(\boldsymbol{\theta}_0\right)\right]^{-1}.$$

In classical GMM, it is a common practice to employ the two-step (Hansen, 1982) or continuously updating estimators (Hansen et al., 1996). The two-step estimation method obtains a first-stage estimate using an arbitrary weighting matrix (e.g., an identity matrix), then obtains a second-stage estimate using a weighting matrix to a precision matrix of the moment conditions based on the first-stage estimate. The continuously updating estimation method repeats the two-step estimation for more than one time.

Despite its critical importance, the practical choice of $\boldsymbol{W}$ in the context of Bayesian GMM has received rather scant attention. A straightforward approach to choosing $\boldsymbol{W}$, which is employed by, for instance, Chernozhukov and Hong (2003); Yin (2009), can be described as follows. At the jth MCMC iteration, given the current parameters $\boldsymbol{\theta}_{[j-1]}$, a proposal $\boldsymbol{\theta}'$ is simulated for a proposal density $p\left(\boldsymbol{\theta}'|\boldsymbol{\theta}_{[j-1]}\right)$. For simplicity, we assume the density is symmetric, e.g., a normal distribution. The weighting matrix is set to the precision matrix of the moment condition based on $\boldsymbol{\theta}'$, that is, the parameter vector and weighting matrix are concurrently proposed and updated (i.e., accepted or rejected). We call this approach the concurrent GMM. The MH ratio is calculated as

$$\begin{aligned}
\alpha\left(\boldsymbol{\theta}', \boldsymbol{\theta}_{[j-1]}\right) &= \frac{q\left(\boldsymbol{\theta}'|\mathcal{D}\right) p\left(\boldsymbol{\theta}'\right) p\left(\boldsymbol{\theta}'|\boldsymbol{\theta}_{[j-1]}\right)}{q\left(\boldsymbol{\theta}_{[j-1]}|\mathcal{D}\right) p\left(\boldsymbol{\theta}_{[j-1]}\right) p\left(\boldsymbol{\theta}_{[j-1]}|\boldsymbol{\theta}'\right)} \\
&= \frac{\det\left(\boldsymbol{W}\left(\boldsymbol{\theta}'\right)\right)^{\frac{1}{2}} \exp\left[-\frac{N}{2}\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}'\right)^\top \boldsymbol{W}\left(\boldsymbol{\theta}'\right)\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}'\right)\right] p\left(\boldsymbol{\theta}'\right)}{\det\left(\boldsymbol{W}\left(\boldsymbol{\theta}_{[j-1]}\right)\right)^{\frac{1}{2}} \exp\left[-\frac{N}{2}\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}_{[j-1]}\right)^\top \boldsymbol{W}\left(\boldsymbol{\theta}_{[j-1]}\right)\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}_{[j-1]}\right)\right] p\left(\boldsymbol{\theta}_{[j-1]}\right)}.
\end{aligned}$$

This approach is motivated by setting a weighting matrix to an optimal one on average. Note that uncertainty about $\boldsymbol{W}$ is inherently different from that about $\boldsymbol{\theta}$; $\boldsymbol{W}$ is not inferred using a prior but it is crudely tuned along the posterior simulation.

Yin et al. (2011) argue this approach is slow to converge, because the concurrent updating of $\boldsymbol{\theta}$ and $\boldsymbol{W}$ complicates the surface of the target density, resulting in an inefficient move of the MH sampler. They propose an alternative approach, named stochastic GMM, where the elements of $\boldsymbol{\theta}$ are updated one by one, keeping $\boldsymbol{W}$ unchanged. This approach is designed to update $\boldsymbol{\theta}$ and $\boldsymbol{W}$ gradually, suppressing instantaneous changes in the shape of the target density. Let $\boldsymbol{\theta}_{[j,l]} = \left(\theta_{[j,l],1}, ..., \theta_{[j,l],l}, \theta_{[j,l-1],l+1}, ..., \theta_{[j,l-1],L}\right)^\top$ denote a state at the jth MCMC iteration after the lth parameter was updated. Once a proposed value of $\theta'_{[j,l],l}$ is simulated, a proposal is constructed as $\boldsymbol{\theta}'_{[j,l]} = \left(\theta_{[j,l],1}, ..., \theta_{[j,l],l-1}, \theta'_{[j,l],l}, \theta_{[j.l-1],l+1}, ..., \theta_{[j,l-1],L}\right)^\top$, and the MH ratio is

given by

$$\alpha\left(\boldsymbol{\theta}'_{[j,l]}, \boldsymbol{\theta}_{[j,l-1]}\right) = \frac{\exp\left[-\frac{N}{2}\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}'_{[j,l]}\right)^{\top}\boldsymbol{W}\left(\boldsymbol{\theta}_{[j,l-1]}\right)\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}'_{[j,l]}\right)\right] p\left(\boldsymbol{\theta}'_{[j,l]}\right)}{\exp\left[-\frac{N}{2}\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}_{[j,l-1]}\right)^{\top}\boldsymbol{W}\left(\boldsymbol{\theta}_{[j,l-1]}\right)\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}_{[j,l-1]}\right)\right] p\left(\boldsymbol{\theta}_{[j,l-1]}\right)}.$$

The underlying justification of this approach is exactly the same as the concurrent GMM. When the number of moment conditions is large, this approach is computationally heavy, because it requires many matrix inversions.

There are two challenges with regard to the choice of the weighting matrix for Bayesian GMM, especially when the number of moment conditions K is large, that is, K is comparable to or even larger than the sample size N. The first is that when K is large, the covariance of the moment functions is ill-estimated, and estimation errors are amplified through matrix inversions. As mentioned in Section 2.1, remedies in classical GMM literature cannot be directly imported to Bayesian GMM. Using the Moore-Penrose generalized inverse is a simple solution, but it does not work well, as shown by the simulation study reported in Section 2.3.[2] The second challenge is the computational cost. The existing approaches require repeated inversion of the sample covariance of the moment functions, thus imposing severe computational loads.

2.2.2 Proposed approach

The proposal of this study is comprised of two elements: regularized precision matrix estimation and random update of weighting matrix. The former is aimed at improving the numerical stability in update of $\boldsymbol{W}$, while the latter is introduced to reduce the computational cost.

First, we propose to compute $\boldsymbol{W}$ using the nonparametric eigenvalue-regularized (NER) precision matrix estimator (Lam, 2016), in which the eigenvalues of a sample covariance matrix are regularized through splitting of the data.[3] The estimator has several favorable properties. First, it is asymptotically optimal with respect to the Stein's loss (Proposition 2 in Lam, 2016, p. 937). Second, it is optimization-free and thus computationally less demanding than the other shrinkage covariance/precision matrix estimators.[4]

Given $\boldsymbol{\theta}$, the moment functions are partitioned as $\boldsymbol{M}(\boldsymbol{\theta}) = \left(\boldsymbol{M}_1(\boldsymbol{\theta})^{\top}, \boldsymbol{M}_2(\boldsymbol{\theta})^{\top}\right)^{\top}$, where the sizes of $\boldsymbol{M}_1(\boldsymbol{\theta})$ and $\boldsymbol{M}_2(\boldsymbol{\theta})$ are N_1-by-K and N_2-by-K, respectively. The covariance matrices of the sub-samples are computed in a standard manner: $\tilde{\boldsymbol{\Sigma}}_i = N_i^{-1}\boldsymbol{M}_i(\boldsymbol{\theta})^{\top}\boldsymbol{M}_i(\boldsymbol{\theta})$, $i = 1, 2$. Let N^* denote the sample size of the first sub-sample, or the splitting location, $N_1 = N^*$, and then $N_2 = N - N^*$. The eigenvalue decomposition of $\tilde{\boldsymbol{\Sigma}}_i$ is represented by $\tilde{\boldsymbol{\Sigma}}_i = \boldsymbol{P}_i\boldsymbol{D}_i\boldsymbol{P}_i^{\top}$, $i = 1, 2$, where $\boldsymbol{D}_i = \text{diag}\left(d_{i,1}, ..., d_{i,K}\right)$ is a diagonal matrix containing the eigenvalues of $\tilde{\boldsymbol{\Sigma}}_i$, $d_{i,1} \geq \cdots \geq d_{i,K}$, and $\boldsymbol{P}_i = \left(\boldsymbol{p}_{i,1}, ..., \boldsymbol{p}_{i,K}\right)$ is a matrix composed of the corresponding eigenvectors. Following Lam (2016), the sample covariance matrix of the moment functions is estimated as

$$\tilde{\boldsymbol{\Sigma}}_{NER} = \boldsymbol{P}_1\left[\left(\boldsymbol{P}_1^{\top}\tilde{\boldsymbol{\Sigma}}_2\boldsymbol{P}_1\right) \odot \boldsymbol{I}_K\right]\boldsymbol{P}_1^{\top},$$

where $\boldsymbol{I}_K$ is a K-dimensional identity matrix and $\odot$ denotes the Hadamard product. Therefore, the corresponding precision matrix is given by

$$\boldsymbol{W}(\boldsymbol{\theta}) = \boldsymbol{P}_1\left[\left(\boldsymbol{P}_1^{\top}\tilde{\boldsymbol{\Sigma}}_2\boldsymbol{P}_1\right) \odot \boldsymbol{I}_K\right]^{-1}\boldsymbol{P}_1^{\top}. \tag{2.2}$$

[2] See Satchachai and Schmidt (2008) on this point for classical GMM.

[3] Abadir et al. (2014) consider a closely related covariance estimator.

[4] See, e.g., Pourahmadi (2011); Fan et al. (2016); Lam (2020) for a survey of the literature of covariance/precision matrix estimation.

Lam (2016) suggests improving this estimator by averaging many (e.g., 50) estimates using different sets of partitioned data that are generated via random permutation. For robustness, we also randomly permute $\boldsymbol{m}_n(\boldsymbol{\theta})$, $n = 1, ..., N$, for each computation of $\boldsymbol{W}$.

The choice of the split location N^* is non-trivial. Theorem 5 of Lam (2016, p. 941) suggests that when $K/N \to c$, it is asymptotically efficient to choose $N^* = N - aN^{1/2}$, with some constants $c, a > 0$. However, this poses two difficulties. First, this asymptotic property is not applicable when N^*/N goes to a constant smaller than 1. Second, there is no practical guidance for setting a. Lam (2016) proposes to choose N^* to minimize the following criterion by means of a grid search:

$$g(N^*) = \left\| \sum_{m=1}^{M} \left(\tilde{\boldsymbol{\Sigma}}_{NER}^{(m)} - \tilde{\boldsymbol{\Sigma}}_{2}^{(m)} \right) \right\|_F^2, \tag{2.3}$$

where the superscripts for $\boldsymbol{\Sigma}_{NER}^{(m)}$ and $\boldsymbol{\Sigma}_{2}^{(m)}$ denote indices for different permutations, M is a number of permutations executed, and $\|\cdot\|_F$ denotes the Frobenius norm. Lam (2016) considers the following grid as a set of candidates for N^*:

$$\left\{ 2N^{1/2},\ 0.2N,\ 0.4N,\ 0.6N,\ 0.8N,\ N - 2.5N^{1/2},\ N - 1.5N^{1/2} \right\}. \tag{2.4}$$

In our framework, one might consider tuning N^* adaptively based on the above criterion as well. However, we do not adopt such a strategy, because the criterion is not informative enough to pin down the optimal choice of N^*, as shown in the subsequent section. A default choice in this study is $N^* = 0.6N$, that is, the median of Lam's (2016) grid. As shown in the next section, simulated posteriors are not sensitive to N^*, as long as N^* is within a moderate range.

In classical GMM, Doran and Schmidt (2006) suggest using principal components of a weighting matrix. From the author's experience, a strategy using the standard principal component analysis to estimate the weighting matrix does not work well for Bayesian GMM, which is not considered in this study.

Next, we consider randomly updating a weighting matrix $\boldsymbol{W}$. We explicitly treat $\boldsymbol{W}$ as a tuning parameter, and update it on the fly as in adaptive MCMC algorithms (Haario et al., 2001; Andrieu and Thoms, 2008; Roberts and Rosenthal, 2009). Our adaptation procedure is motivated by Bhattacharya and Dunson (2011). At the jth MCMC iteration, the adaptation of $\boldsymbol{W}$ occurs with probability $s(j) = \exp(\alpha_0 + \alpha_1 j)$, regardless of the previous proposal being accepted or rejected. Throughout the study, we chose $\alpha_0 = -1$ and $\alpha_1 = -10/J_{warmup}$, where J_{warmup} denotes the number of warmup iterations. If an adaptation occurs, $\boldsymbol{W}$ is updated using the mean of the previous sample obtained; at the jth iteration, $\bar{\boldsymbol{\theta}}_{[j-1]} = (j-1)^{-1} \sum_{j'=1}^{j-1} \boldsymbol{\theta}_{[j']}$. After warmup iterations, $\boldsymbol{W}$ is fixed to the end. This adaptation strategy satisfies the convergence condition in Theorem 5 of Roberts and Rosenthal (2007). In our implementation, at every jth iteration, a random variable is simulated from a standard uniform distribution, $u_j \sim \mathcal{U}(0,1)$, and $\boldsymbol{W}$ is updated if $u_j < s(j)$, where $\mathcal{U}(a,b)$ denotes a uniform distribution with support on interval (a,b). At the jth iteration, given a proposal $\boldsymbol{\theta}'$, the MH ratio is calculated as

$$\alpha\left(\boldsymbol{\theta}', \boldsymbol{\theta}_{[j-1]}\right) = \frac{\exp\left[-\frac{N}{2}\bar{\boldsymbol{m}}(\boldsymbol{\theta}')^\top \boldsymbol{W}\left(\bar{\boldsymbol{\theta}}_{[j-1]}\right)\bar{\boldsymbol{m}}(\boldsymbol{\theta}')\right] p(\boldsymbol{\theta}')}{\exp\left[-\frac{N}{2}\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}_{[j-1]}\right)^\top \boldsymbol{W}\left(\bar{\boldsymbol{\theta}}_{[j-1]}\right)\bar{\boldsymbol{m}}\left(\boldsymbol{\theta}_{[j-1]}\right)\right] p\left(\boldsymbol{\theta}_{[j-1]}\right)}.$$

This treatment of $\boldsymbol{W}$ does not conflict with the theoretical results of Bayesian GMM (e.g., Kim 2002; Chernozhukov and Hong 2003; Belloni and Chernozhukov 2009; Li and Jiang 2016). Whilst in the existing literature there is a discrepancy between theory and practical computation in how a weighting matrix is treated, our treatment of $\boldsymbol{W}$ is in agreement with the theoretical results than existing approaches.

2.3 Simulation Study

We compare the proposed approach with alternatives.[5] We compare the NER estimator given by (2.2) with the standard estimators specified by

$$\boldsymbol{W}(\boldsymbol{\theta}) = \begin{cases} \left[N^{-1}\boldsymbol{M}(\boldsymbol{\theta})^{\top}\boldsymbol{M}(\boldsymbol{\theta})\right]^{-1}, & K \leq N, \\ \left[N^{-1}\boldsymbol{M}(\boldsymbol{\theta})^{\top}\boldsymbol{M}(\boldsymbol{\theta})\right]^{+}, & K > N, \end{cases}$$

where $\boldsymbol{A}^{+}$ denotes the Moore-Penrose generalized inverse of a matrix $\boldsymbol{A}$. Six adaptation strategies are considered. The first is fixing the weighting matrix of the moment conditions based on the true parameter value (*Oracle*), the second is the concurrent Bayesian GMM (*Concurrent*) (Chernozhukov and Hong, 2003; Yin, 2009), and the third is the stochastic GMM (*Stochastic*) (Yin et al., 2011). The fourth is a MCMC version of the continuously updating GMM estimator (Hansen et al., 1996) (*Continuous*), that is, $\boldsymbol{W}$ is updated in each cycle based on the current recursive means of the sampled parameters. The fifth is the random update strategy we propose (*Random*).

We adopt an instrumental variable (IV) regression as laboratory. A true data generating process is specified by the following equations, for $n = 1, ..., N$,

$$x_n = \boldsymbol{z}_n^{\top}\boldsymbol{\delta} + w_n, \quad w_n \sim \mathcal{N}\left(0, \sigma_x^2\right), \tag{2.5}$$

$$y_n = \gamma x_n + \varphi\left(x_n - \boldsymbol{z}_n^{\top}\boldsymbol{\delta}\right) + u_n, \quad u_n \sim \mathcal{N}\left(0, \sigma_y^2\right), \tag{2.6}$$

where y_n is a response variable, x_n is an endogenous covariate, $\boldsymbol{z}_n$ is a K-dimensional vector of instruments, u_n and w_n are normally distributed errors, and $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. $\gamma = 0.5$ is a coefficient to be inferred. $\varphi = 0.2$ is a fixed parameter. The instruments are generated from a latent factor model: for $n = 1, ..., N$,

$$\boldsymbol{z}_n = \boldsymbol{B}\boldsymbol{\nu}_n + \boldsymbol{\epsilon}_n,$$

$$\boldsymbol{\nu}_n \sim \mathcal{N}(\boldsymbol{0}_S, \boldsymbol{I}_S), \quad \boldsymbol{\epsilon}_n \sim \mathcal{N}\left(\boldsymbol{0}_K, \boldsymbol{\Psi}^2\right),$$

where S is the number of latent factors, $\boldsymbol{\nu}_n$ is an S-dimensional vector of latent factors, $\boldsymbol{\epsilon}_n$ is a K-dimensional vector of idiosyncratic errors with covariance $\boldsymbol{\Psi}^2$, and $\boldsymbol{B}$ is a K-by-S matrix of factor loadings. The distribution of $\boldsymbol{z}_n$ is written as

$$\boldsymbol{z}_n \sim \mathcal{N}\left(\boldsymbol{0}_K,\ \boldsymbol{B}\boldsymbol{B}^{\top} + \boldsymbol{\Psi}^2\right), \quad n = 1, ..., N.$$

$\boldsymbol{\Psi}^2$ are $\boldsymbol{B}$ are set as follows:

$$\boldsymbol{\Psi}^2 = \text{diag}\left(\psi_1^2, ..., \psi_K^2\right), \quad \psi_k \sim \mathcal{U}(2, 4), \quad k = 1, ..., K,$$
$$\boldsymbol{B} = (b_{k,s}), \quad b_{k,s} \sim \mathcal{U}(0, 1), \quad k = 1, ..., K;\ s = 1, ..., S.$$

The coefficients of $\boldsymbol{z}_n$ are generated as

$$\boldsymbol{\delta} = \boldsymbol{A}^{\top}\boldsymbol{\eta}, \quad \boldsymbol{A} = \boldsymbol{B}^{\top}\left(\boldsymbol{B}\boldsymbol{B}^{\top} + \boldsymbol{\Psi}^2\right)^{-1},$$
$$\boldsymbol{\eta} = (\eta_1, ..., \eta_S)^{\top}, \quad \eta_s \sim \mathcal{U}(0, 1), \quad s = 1, ..., S.$$

[5]The programs in this study are written in Matlab 2019b (64bit), and executed on an Ubuntu Desktop 18.04 LTS (64bit), running on AMD Ryzen Threadripper 1950X (4.2GHz).

We consider three scenarios with different numbers of instruments $K = \{50, 150, 250\}$ and factors $S = \{K, K/2, 3\}$. The standard deviations of the errors, σ_x and σ_y, are chosen so that the ratios of the standard deviations of the errors to those of the signals, denoted by q_x and q_y, are ς_x and ς_y, respectively:

$$\sigma_x = \varsigma_x q_x, \quad \sigma_y = \varsigma_y q_y,$$
$$q_x = \sqrt{\boldsymbol{\delta}^\top \left(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Psi}^2\right) \boldsymbol{\delta}},$$
$$q_y = \sqrt{\gamma^2 \left(1 + \varsigma_x^2\right) + \varphi^2 \varsigma_x^2} q_x.$$

We fix $\varsigma_x = \varsigma_y = 2$. Unknown parameter γ is inferred through a set of moment conditions,

$$E\left[\left(y_n - \gamma x_n\right) \boldsymbol{z}_n\right] = \mathbf{0}_K.$$

We assign a flat prior on γ, $p(\gamma) \propto 1$. The sample size is fixed at $N = 200$. For posterior sampling, we employ an adaptive MH sampler of Vihola (2012), which automatically tunes the covariance of a proposal density. The tuning parameters of the sampler are chosen as in Vihola (2012). For all experiments, we simulate a total of 70,000 draws: the initial 20,000 draws are used for warmup and the subsequent 50,000 for posterior estimates.[6] The initial value of γ is randomly generated from a uniform distribution with interval $(-2.5, 3.5)$. $\boldsymbol{W}$ is initialized to an identity matrix.

We evaluate the results of inference of γ according to four measures. The first is the failure rate (Fail): when the estimated inter-quantile range of a target posterior density is larger than 1 or smaller than 0.01, we regard the MCMC run as failed. The second is the mean squared error of the posterior mean estimate (MSE). The third is the inter-quantile range of the posterior density (IQR). The fourth is the total computation time measured in seconds (Speed). We conduct 500 experiments.

We compare the results for the precision matrix estimators. The left halves of Tables 2.1-2.3 show the results for the standard precision matrix estimator and the right halves show those for the NER estimator. The upper parts of Tables 2.1-2.3 report the results for $K = 50$, the middle parts for $K = 150$, and the lower parts for $K = 250$. We see a similar pattern from the tables regardless of the number of latent factors S relative to K.

When $K > N$, the numbers of Fail for the standard estimator exceeded a half of the number of experiments (500), and the posterior simulations using the standard estimator are unsuccessful. For instance, when $S = 3$, $K = 250$, and *Random* is used, the standard estimator failed 485 of 500 experiments (the last row of Table 2.3). In contrast, even with $K > N$, unless using *Concurrent*, the numbers of Fail for the NER estimator are zero, which means that the NER estimator provides reasonable posterior estimates. Therefore, when $K > N$, only the NER estimator is a viable option.

For most cases, the MSEs for the NER estimator are smaller than those for the standard estimator. For instance, when $S = 3$, $K = 150$, and *Random* is used, MSE for the standard estimator was 0.0809, while that for the NER estimator was 0.0155. Thus, in terms of the estimation accuracy, the NER estimator outperforms the standard estimator. While the advantage of the NER estimator over the standard estimator in terms of MSE is not so large for relative easy cases, i.e., K and/or S are small, even when the number of moment conditions K is smaller than the sample size N, the NER estimator is likely to obtain a more accurate posterior mean estimate than the standard precision estimator. It is also worth mentioning that, when $K > N$, the

[6] In this study, the number of MCMC iterations is chosen based on pilot runs so that minimum of the effective sample sizes for obtained posterior draws (except the warmup draws) in each experiment is no less than 15,000.

posterior simulation using the NER estimator is almost as precise as the cases with $K < N$. For instance, when $S = 3$ and *Random* is used, the standard estimator had MSEs of 0.247, 0.0809, and N/A (all the experiments failed) for $K = 50$, 150, and 250, respectively. On the other hand, for the same cases, the NER estimator obtained MSEs of 0.0215, 0.0155, and 0.0166 for $K = 50$, 150, and 250, respectively. A comparison between the results for the *Oracle* cases with different precision estimators and $K = 50, 150$ reveals that the NER estimator is not better than the standard one if the true value of $\boldsymbol{\theta}$ is known. For instance, when $S = 3$ and *Random* is used, MSEs for the standard estimator are 0.0104 and 0.0012 for $K = 50$ and 150, respectively, while those for the NER estimator are 0.0185 and 0.0121 for $K = 50$ and 150, respectively. However, as suggested by a comparison between MSEs for cases using updating procedures other than Oracle, the gain from the numerical stability of the NER estimator outweighs its efficiency loss in practical situations.

We also investigate the sensitivity of the above results to the choice of split location N^*. We conduct Monte Carlo experiments using different N^* and two preferred adaptation strategies, *Stochastic* and *Random*. Following Lam (2016), we consider the grid of (2.4) (each N^* is rounded to the nearest integer). Table 2.4 shows that the NER estimator consistently outperforms the standard estimator, irrespective of the split location choice. In terms of MSE, a moderate value of N^* is preferred. To investigate how much this result is in agreement with the criteria based on the Frobenius norm (2.3), we simulate the values of (2.3) for different random permutations of the moment conditions using the true parameter. Panel (a) of Figure 2.1 reports the median and 90 percentile intervals of the simulated values for a fine grid $\{0.1N,\ 0.15N, ..., 0.9N\}$. We only report the results for $K = 250$, as those for $K = 50, 150$ are qualitatively similar. As evident from the panel, an extremely high N^* is not preferred, but the criterion is not informative enough to select a good N^* from a considerably large range. The variability of the criterion is not attributable to the small sample size. We conduct the same simulation as in panel (a) but the sample size increases to $N = 5,000$. Panel (b) of Figure 2.1 shows the results. As is the case of $N = 200$, the values of the criterion based on the Frobenius norm are almost indifferent for a large range. As such, we recommend setting N^* to approximately half the sample size as default.

Next, we compare the results for the adaptation strategies. There are five points worth mentioning. First, *Concurrent* does not work despite high computational cost. Second, the relative advantage of *Stochastic* to *Concurrent* in terms of numerical stability is in line with Yin et al. (2011). Third, in terms of MSE, all *Stochastic*, *Continuous,* and *Random* work well. *Stochastic* is better than *Continuous* and *Random*, while *Continuous* and *Random* are comparable. Fourth, as shown by the IQR estimates, *Continuous* and *Random* are more optimistic than *Stochastic*. Fifth, *Random* is much faster than *Stochastic* and *Continuous*. Figure 2.2 provides a typical example of recursive posterior mean and the occurrence of random adaptation (NER estimator, $K = 150$). From this figure, a posterior mean is fairly fast to converge, which indicates that most updates of the weighting matrix in *Continuous* are essentially redundant. We find *Random* has a good balance between statistical and computational efficiency, so that it is recommendable for a test run. Compared with *Random*, although *Stochastic* is computationally demanding, it is more accurate and conservative. Therefore, it is suitable for a final estimate.

Ta le 2.1: Comparison of different approaches (1): $S = K$

	Estimator	Standard				NER			
K	Adaptation	Fail	MSE	IQR	Time	Fail	MSE	IQR	Time
50	Oracle	0/500	0.0089	0.1292	2.2	0/500	0.0166	0.1349	2.2
	Concurrent	201/500	–	–	9.9	486/500	–	–	35.9
	Stochastic	0/500	0.0205	0.1305	4.6	0/500	0.0184	0.1529	11.2
	Continuous	0/500	0.0210	0.1289	4.9	0/500	0.0215	0.1348	12.4
	Random	0/500	0.0210	0.1287	2.3	0/500	0.0215	0.1348	2.6
150	Oracle	0/500	0.0012	0.0480	5.7	/500	0.0114	0.0939	5.7
	Concurrent	383/500	–	–	56.4	500/500	–	–	247.5
	Stochastic	0/500	0.0336	0.0689	19.9	/500	0.0137	0.1209	67.8
	Continuous	0/500	0.0686	0.0480	21.8	/500	0.0162	0.0939	76.1
	Random	0/500	0.0711	0.0481	6.3	/500	0.0166	0.0936	8.3
250	Oracle	375/500	–	–	5.4	0/500	0.0115	0.0772	5.4
	Concurrent	500/500	–	–	458.5	500/500	–	–	259.1
	Stochastic	395/500	–	–	126.1	0/500	0.0142	0.1075	76.4
	Continuous	480/500	–	–	138.4	0/500	0.0162	0.0782	81.7
	Random	450/500	–	–	10.9	0/500	0.0176	0.0769	8.6

Notes: The column labeled Fail reports the number of failed runs. Column MSE reports the mean squared errors of posterior mean estimates. Column IQR reports inter-quantile ranges of posterior densities. Column Time reports averages of computation time measured in seconds.

Table 2.2: Comparison of different approaches (2): $S = K/2$

	Estimator	Standard				NER			
K	Adaptation	Fail	MSE	IQR	Time	Fail	MSE	IQR	Time
50	Oracle	0/500	0.0085	0.1290	2.2	0/500	0.0151	0.1346	2.2
	Concurrent	212/500	–	–	9.9	481/500	–	–	35.7
	Stochastic	0/500	0.0217	0.1308	4.6	0/500	0.0179	0.1513	11.1
	Continuous	0/500	0.0228	0.1288	4.9	0/500	0.0205	0.1335	12.3
	Random	0/500	0.0228	0.1287	2.3	0/500	0.0204	0.1343	2.6
150	Oracle	0/500	0.0012	0.0476	5.7	0/500	0.0112	0.0937	5.7
	Concurrent	402/500	–	–	56.4	500/500	–	–	247.8
	Stochastic	0/500	0.0380	0.0701	19.9	0/500	0.0135	0.1187	68.0
	Continuous	0/500	0.0619	0.0477	21.7	0/500	0.0169	0.0931	76.2
	Random	0/500	0.0673	0.0482	6.3	0/500	0.0160	0.0935	8.3
250	Oracle	359/500	–	–	10.4	0/500	0.0095	0.0787	10.4
	Concurrent	500/500	–	–	1036.2	500/500	–	–	757.8
	Stochastic	398/500	–	–	226.2	0/500	0.0113	0.1076	200.1
	Continuous	475/500	–	–	307.7	0/500	0.0130	0.0787	225.5
	Random	467/500	–	–	21.4	0/500	0.0135	0.0782	18.3

Notes: The column labeled Fail reports the number of failed runs. Column MSE reports the mean squared errors of posterior mean estimates. Column IQR reports inter-quantile ranges of posterior densities. Column Time reports averages of computation time measured in seconds.

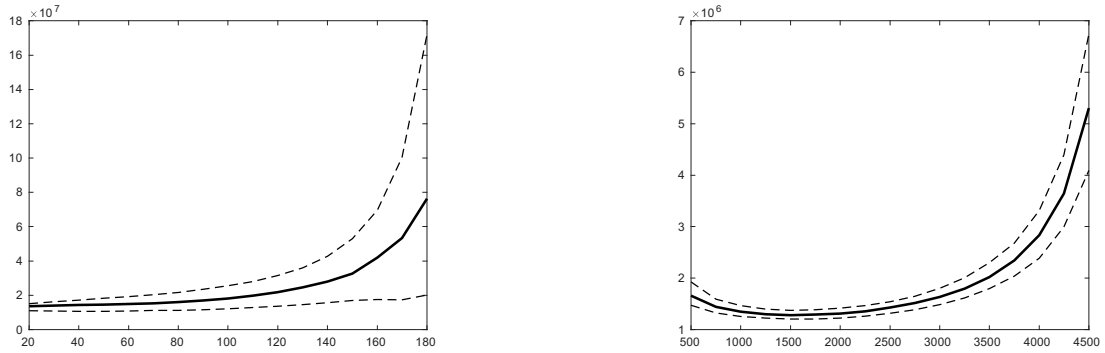
Table 2.3: Comparison of different approaches (3): $S = 3$

	Estimator	Standard				NER			
K	Adaptation	Fail	MSE	IQR	Time	Fail	MSE	IQR	Time
	Oracle	0/500	0.0104	0.1288	2.6	0/500	0.0185	0.1258	2.6
	Concurrent	208/500	–	–	11.2	480/500	–	–	40.8
50	Stochastic	0/500	0.0242	0.1303	5.4	0/500	0.0207	0.1352	12.8
	Continuous	0/500	0.0247	0.1283	5.7	0/500	0.0219	0.1251	14.2
	Random	0/500	0.0247	0.1282	2.8	0/500	0.0215	0.1267	3.1
	Oracle	0/500	0.0012	0.0479	6.2	0/500	0.0121	0.0868	6.2
	Concurrent	398/500	–	–	71.3	500/500	–	–	285.1
150	Stochastic	0/500	0.0379	0.0689	24.1	0/500	0.0139	0.1055	77.7
	Continuous	0/500	0.0761	0.0487	26.5	0/500	0.0165	0.0863	87.3
	Random	0/500	0.0809	0.0488	7.0	0/500	0.0155	0.0867	9.2
	Oracle	363/500	–	–	10.8	0/500	0.0110	0.0706	10.8
	Concurrent	500/500	–	–	1157.6	500/500	–	–	829.3
250	Stochastic	480/500	–	–	295.7	0/500	0.0142	0.0937	218.7
	Continuous	493/500	–	–	343.0	0/500	0.0161	0.0706	246.3
	Random	485/500	–	–	23.1	0/500	0.0166	0.0707	19.4

Notes: The column labeled Fail reports the number of failed runs. Column MSE reports the mean squared errors of posterior mean estimates. Column IQR reports inter-quantile ranges of posterior densities. Column Time reports averages of computation time measured in seconds.

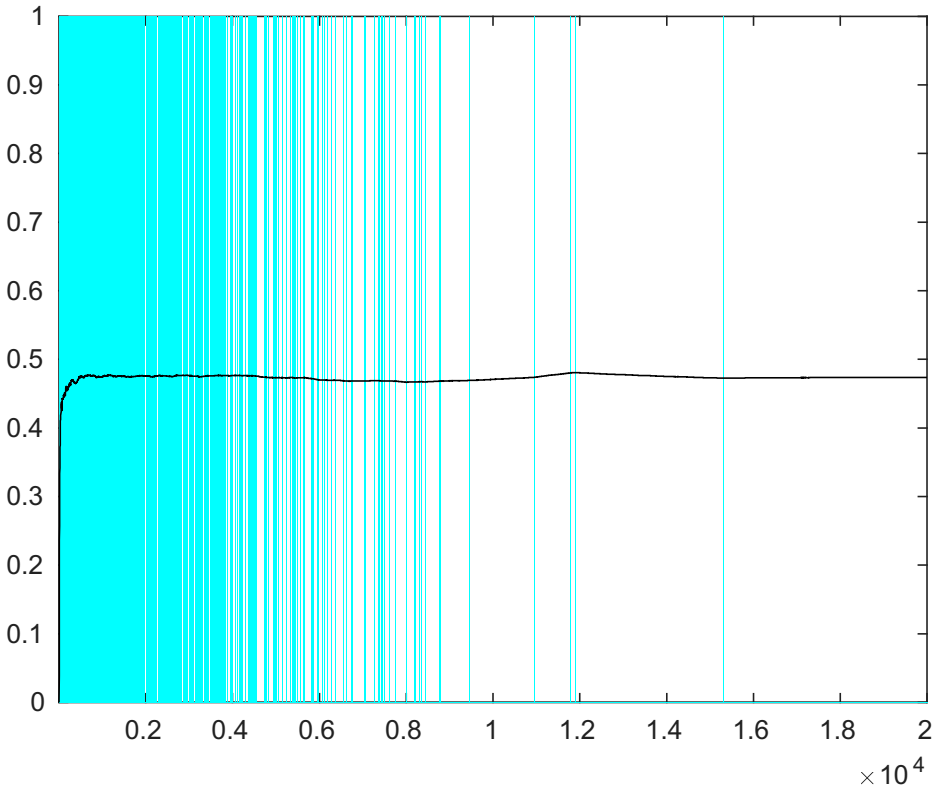
Figure 2.1: The Frobenius norm criterion for different permutations

(a) $N = 200$ (b) $N = 5000$



Notes: Solid lines denote the median, and dashed lines denote the 90 percentile interval. $K = 250$. Moment conditions are calculated based on the true parameter value.

Figure 2.2: An example of random adaptation



Notes: The x-axis denotes MCMC iterations and the y-axis denotes parameter values. A thin solid vertical line denotes the occurrence of adaptation. A bold solid line denotes a recursive mean of posterior samples.

Table 2.4: Comparison of different choices of N^*

		Adaptation	Stochastic		Random	
K	Estimator	N^*	MSE	IQR	MSE	IQR
50	Standard		0.0242	0.1303	0.0247	0.1282
	NER	$28\ (=\lceil 2N^{1/2}\rceil)$	0.0223	0.1174	0.0237	0.1109
		$40\ (=0.2N)$	0.0220	0.1209	0.0225	0.1131
		$80\ (=0.4N)$	0.0211	0.1299	0.0229	0.1201
		$120\ (=0.6N)$	0.0207	0.1354	0.0220	0.1261
		$160\ (=0.8N)$	0.0206	0.1382	0.0218	0.1266
		$164\ (=\lceil N-2.5N^{1/2}\rceil)$	0.0205	0.1386	0.0223	0.1252
		$178\ (=\lceil N-1.5N^{1/2}\rceil)$	0.0206	0.1408	0.0239	0.1233
150	Standard		0.0379	0.0689	0.0809	0.0488
	NER	$28\ (=\lceil 2N^{1/2}\rceil)$	0.0171	0.0854	0.0182	0.0733
		$40\ (=0.2N)$	0.0163	0.0904	0.0177	0.0753
		$80\ (=0.4N)$	0.0145	0.1016	0.0165	0.0816
		$120\ (=0.6N)$	0.0139	0.1055	0.0160	0.0864
		$160\ (=0.8N)$	0.0143	0.1051	0.0161	0.0887
		$164\ (=\lceil N-2.5N^{1/2}\rceil)$	0.0144	0.1054	0.0158	0.0882
		$178\ (=\lceil N-1.5N^{1/2}\rceil)$	0.0147	0.1083	0.0165	0.0847
250	Standard		–	–	–	–
	NER	$28\ (=\lceil 2N^{1/2}\rceil)$	0.0177	0.0743	0.0184	0.0591
		$40\ (=0.2N)$	0.0168	0.0796	0.0187	0.0611
		$80\ (=0.4N)$	0.0148	0.0913	0.0173	0.0667
		$120\ (=0.6N)$	0.0142	0.0938	0.0163	0.0707
		$160\ (=0.8N)$	0.0148	0.0916	0.0166	0.0710
		$164\ (=\lceil N-2.5N^{1/2}\rceil)$	0.0149	0.0916	0.0175	0.0724
		$178\ (=\lceil N-1.5N^{1/2}\rceil)$	0.0153	0.0939	0.0176	0.0693

Notes: The column labeled Fail reports the number of failed runs. Column MSE reports the mean squared errors of posterior mean estimates. Column IQR reports inter-quantile ranges of posterior densities.

2.4 Application

To demonstrate the proposed method, we apply it to a demand analysis for automobiles. Berry et al. (1995) consider an IV regression model of demand for automobiles specified by

$$y_{i,t} = \gamma p_{i,t} + \boldsymbol{\delta}^\top \boldsymbol{x}_{i,t} + u_{i,t},$$

$$y_{i,t} = \log(s_{i,t}) - \log(s_{0,t}).$$

$s_{i,t}$ denotes the market share of product i on market t, with subscript 0 denoting the outside option. The treatment $p_{i,t}$ is the product price. $u_{i,t}$ is an error term, and γ and $\boldsymbol{\delta}$ are the parameters to be estimated. The primary focus of this application is the inference of γ.

We consider two specifications.[7] The first specification coincides with Berry et al. (1995) as follows. A vector of covariates $\boldsymbol{x}_n$ includes four covariates, namely, air conditioning dummy, horsepower to weight ratio, miles per dollar, and vehicle size. A set of instruments contains the

[7] All data are extracted from R package `hdm` (version 0.2.3).

four covariates and ten variables, namely, sum of each covariate taken across models made by product t's firm, sum of each covariate taken across competitor firms' products, total number of models produced by product t's firm, and total number of models produced by the firm's competitors. The second specification is an extension of the first, which is considered in Chernozhukov et al. (2015). $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$ are extended from the first case by incorporating a time trend, quadratic and cubic terms of all continuous covariates, and first-order interaction terms. The numbers of the instruments in the first and second specifications are 10 and 48, respectively. The sample size is $N = 2,217$, being larger than the numbers of instruments. Nevertheless, because of ill-posedness of the data set, the covariance of a classical estimator is nearly singular. We use a constant prior; thus, if the relationship between the instruments and the treatment is linear and the distributions of residuals are normal, a posterior estimate coincides with a two-stage least square estimate. The posterior estimate is obtained using different combinations of precision matrix estimators and the adaptation of proposal density. A total of 250,000 posterior draws are sampled with the last 200,000 drawn for posterior analysis.

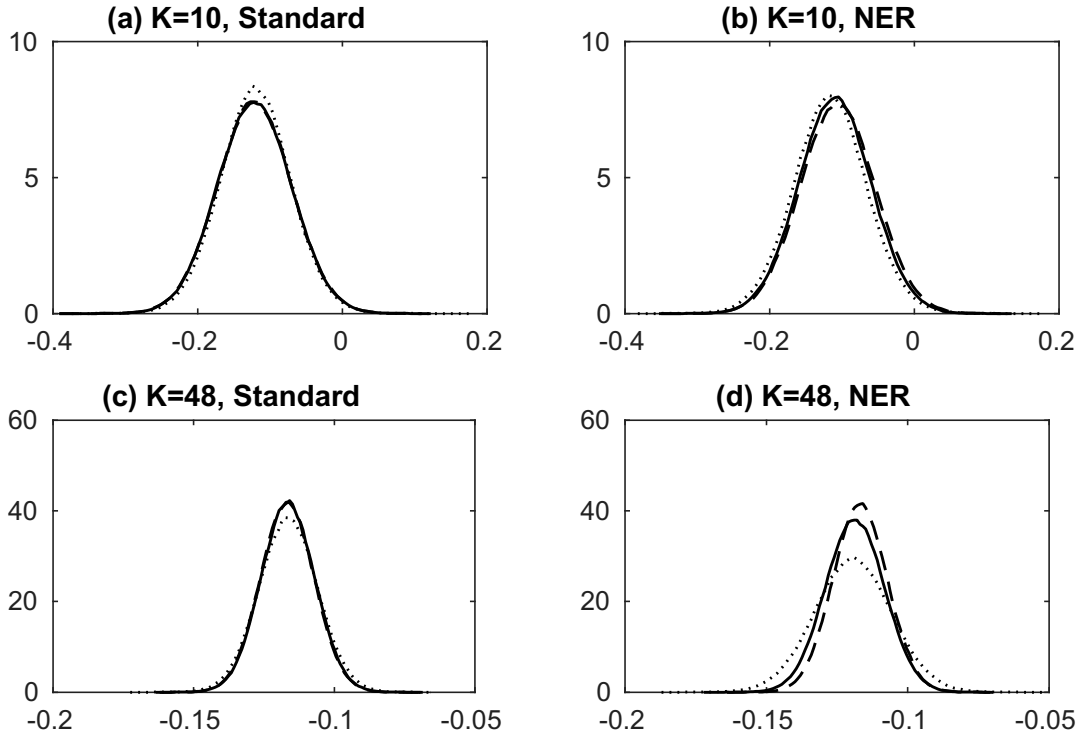
Table 2.5 summarizes the results of the posterior estimate for the coefficient on price. Although the number of moment conditions is fairly smaller than the sample size, MCMC runs using *Concurrent* fails to converge. By contrast, MCMC runs using the NER estimator obtain sensible posterior samples, irrespective of the adaptation strategy. For comparison, Table 2.5 also includes the estimates obtained using four alternative methods. The first two are conventional: ordinary least squares (OLS) and two-stage least squares (2SLS) methods. The second two are state-of-the-art: IV with instrument selection based on a least absolute shrinkage and selection operator (Chernozhukov et al., 2015) (LASSO-IV), and Bayesian IV with a factor shrinkage prior (Hahn et al., 2018) (HS-IV). LASSO-IV is designed to select fewer relevant instruments, while HS-IV is designed to compress observed information into few latent factors. The two methods assume a linear relationship between instruments and the endogenous variable and Gaussianity of the error terms, while our method does not impose such assumptions. These alternative methods obtain larger estimates than the conventional ones, and the estimates considerably depend on a set of (potential) instruments. By contrast, our method estimated the coefficient to be intermediate between OLS and 2SLS, nearly irrespective of the choice of instruments. As shown in Figure 2.3, the posterior densities of γ for alternative approaches (excluding *Concurrent* adaptation) are quite similar.

2.5 Discussion

We propose a new adaptive MCMC approach to infer Bayesian GMM with many moment conditions. Our proposal consists of two elements. The first is the use of a nonparametric eigenvalue-regularized precision matrix estimator (Lam, 2016) for estimating the weighting matrix. This prevents us from ill-estimating the weighting matrix. The second is the use of random adaptation. By setting adaptation probability as exponentially decreasing, it can significantly reduce the computational burden, while retaining statistical efficiency. We show the superiority of the proposed approach over existing approaches through simulation, and demonstrate the approach by applying it to a demand analysis for automobiles.

There are several promising research areas stem from this study. First, a theoretical investigation of the effects of tuning/estimation of a weighting matrix on the posterior density is needed, which is absent in the literature. Second, while the proposed approach seems to be fairly robust to N^*, there is room for improvement by finding a better N^*. Third, while this study addresses only problems caused by many moment conditions, problems caused by many

Figure 2.3: Posterior distribution of γ



Notes: Solid lines trace the mean estimates for *Stochastic*. Dashed lines trace the mean estimates for *Continuous*. Dotted lines trace the mean estimates for *Random*.

Table 2.5: Posterior estimates of γ

		Standard			NER		
K		Mean	Std	Time	Mean	Std	Time
10	Concurrent	–	–	1214.4	–	–	1213.6
	Stochastic	-0.120	0.049	334.3	-0.117	0.051	508.3
	Continuous	-0.122	0.051	363.3	-0.106	0.051	439.9
	Random	-0.122	0.051	215.7	-0.110	0.050	220.8
	OLS	-0.089	0.004				
	2SLS	-0.142	0.012				
	LASSO-IV	-0.185	0.014				
48	Concurrent	–	–	2606.9	–	–	3711.2
	Stochastic	-0.116	0.011	1230.7	-0.119	0.014	1613.8
	Continuous	-0.117	0.010	1071.4	-0.117	0.010	1432.4
	Random	-0.117	0.010	698.3	-0.119	0.010	705.6
	LASSO-IV	-0.221	0.015				
	HS-IV	-0.275	0.018				

Notes: The column labeled Mean reports mean estimates. Column Std reports standard errors. Column Time reports computation time measured in seconds.

unknown parameters are also important. The proposed method should serve as a stepping stone for further development of inferential methods for high-dimensional Bayesian GMM. Finally, it is worth conducting a thorough comparison between the proposed approach and existing classical and Bayesian approaches tailored to a specific class of models such as IV regressions and dynamic panel models.

Chapter 3

Bayesian Matrix Completion Approach to Causal Inference with Panel Data

3.1 Introduction

Program/policy evaluations and comparative case studies using observational data are pervasive in social and natural sciences and in government and business practice. In particular, causal inference is an integral part of social sciences, where randomized experiments are usually infeasible. For instance, although Abadie et al. (2015) analyzed the economic cost of the German reunification in 1990, we cannot repeat such a political event many times in a controlled fashion.

The primary interest of this study is inference of causal effects of a treatment, such as average treatment effect and average treatment effect on treated (ATET). Suppose we have panel data with J units and T time periods. An outcome of unit j at period t is denoted by $y_{j,t}\left(s_{j,t}\right)$, where $s_{j,t}=1$ when the unit is exposed to treatment and $s_{j,t}=0$ otherwise. Let $\mathcal{I}_1$ and $\mathcal{I}_0$ be sets of indices for treated and untreated observations, respectively. Then, for instance, ATET is defined as

$$\varphi=\frac{1}{|\mathcal{I}_1|}\sum_{(j,t)\in\mathcal{I}_1}\left(y_{j,t}\left(1\right)-y_{j,t}\left(0\right)\right),$$

where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$. Inference of causal effects amounts to inference of counterfactual untreated outcomes $y_{j,t}\left(0\right)$, $(j,t)\in\mathcal{I}_1$, or the "potential outcome" in terms of Neyman–Rubin's causal model (Im - ous challenge to statisticians, and numerous approaches have been proposed: the difference-in-differences estimator, regression discontinuity design, matching-based methods, etc.[1]

In this study, we propose a new Bayesian approach for inferring the causal effect of a binary treatment with panel data. We transform a statistical problem of causal inference into a matrix completion problem, an extensively studied issue in machine learning (e.g., Keshavan et al., 2010). Our approach implements in two steps. First, the potential outcomes are inferred via a Bayesian matrix completion method. Then, a causal effect is inferred based on the posterior draws of the potential outcomes.

We model the sum of a matrix of outcomes using two-component factorization and a matrix of covariate effects. The potential outcomes are treated as missing observations and simulated from the posterior predictive distribution,

$$\int p\left(y_{j,t}\left(0\right),(j,t)\in\mathcal{I}_1|\mathcal{D},\mathbf{\Theta}\right)p\left(\mathbf{\Theta}|\mathcal{D}\right)\mathrm{d}\mathbf{\Theta},$$

[1] See, e.g., Imbens and Rubin (2015).

where $\mathcal{D}$ denotes a set of observations including untreated outcomes $y_{j,t}(0)$, $(j,t) \in \mathcal{I}_0$ and exogenous covariates and Θ denotes a set of parameters and random variables to be sampled. The other unknown parameters, such as coefficients on the covariates and the variance of measurement error, are simulated from the conditional posterior distribution. Leaving aside the covariate effects, the proposed approach can be thought of as treating inference of the potential outcomes as multiple imputation of a matrix of panel data that is probably rank deficient.

Given the posterior draws of potential outcomes, we can infer a causal effect of interest. For instance, when we have a total of N_{post} posterior draws of potential outcomes, the posterior mean estimate of the ATET is given by

$$\widehat{\varphi} = \frac{1}{N_{post}} \sum_{i=1}^{N_{post}} \frac{1}{|\mathcal{I}_1|} \sum_{(j,t)\in\mathcal{I}_1} \left(y_{j,t}(1) - y_{j,t}^{(i)}(0) \right),$$

where $y_{j,t}^{(i)}(s_{j,t})$ denotes the ith posterior draw of the potential outcome of unit j at period t.

To facilitate this task, we develop a tailored prior that induces the model to be lower rank, adapting a cumulative shrinkage process prior (Legramanti et al., 2020). With this prior specification, there is no need to specify the rank of the outcome matrix because the prior pushes insignificant columns of one of the factorizations toward zero.

Our Bayesian approach has two notable advantages. First, it can provide credible intervals in a consistent and straightforward manner, while the existing non-Bayesian approaches have difficulty quantifying uncertainty. As hypothesis testing is an essential component of scientific research, this advantage is a strong reason to use a Bayesian method. Second, our approach has better finite sample performance than that of the existing approaches. By means of a series of simulation studies, we show that our proposal is competitive with the existing approaches in terms of the precision of the prediction of potential outcomes.

Three strands of the literature are particularly relevant to this study. First, the proposed approach is related to a class of synthetic control methods (SCMs) (e.g., Abadie and Gardeazabal, 2003; Abadie et al., 2010).[2] This class of methods is aimed at obtaining "synthetic" observations of untreated outcomes as weighted sums of the outcomes of the control units. Despite its increasing popularity, the original SCM (Abadie et al., 2010) has two notable shortcomings. The first shortcoming is that it imposes a strong assumption that the weights of synthetic observations are nonnegative and sum to one. This assumption implies that the treated unit falls in the convex hull of the control units and that synthetic observations are positively correlated with the control units, which is not plausible in many real situations. While some alternative approaches (Doudchenko and Imbens, 2017; Kim et al., 2020; Amjad et al., 2018) do not require these assumptions, our approach has better finite sample performance under various data generating processes, as shown in the simulation studies. The second shortcoming is that the original SCM does not have an effective method for assessing the uncertainty of the obtained estimates. Abadie et al. (2010) conduct a series of placebo studies, but the approach incurs size distortion (Hahn and Shi, 2017). Recently, Li (forthcoming) proposes a subsampling method to obtain confidence intervals for SCMs, but our Bayesian approach can obtain credible intervals simply as a byproduct of posterior simulation.

Second, an approach developed by Athey et al. (2018) is particularly related to our proposal. They also treat potential outcomes as missing data and estimate them via matrix completion with the nuclear norm penalty (Mazumder et al., 2010). However, Athey et al.'s (2018) non-Bayesian approach does not have an estimator for confidence intervals.

[2] See Abadie (forthcoming) for a recent overview of the literature on SCMs.

Finally, our proposal is conceptually similar to approaches proposed by Brodersen et al. (2015); Ning et al. (2019) in that all of them infer potential outcomes as missing observations in Bayesian manners. On the other hand, their approaches rely on the fit of a time-series model, while our approach exploits the factor structure of panel data. Therefore, our proposal is better suited for typical panel data covering short time periods where it is difficult to estimate a time-series model.

The remainder of this study is structured as follows. In Section 3.2, we introduce a new Bayesian approach to causal inference with panel data and compare it with the existing alternatives. In Section 3.3, we illustrate the proposed approach by applying it to simulated and real data. We conduct a simulation study and show that our proposed method is competitive with the existing approaches in terms of the precision of the predictions of potential outcomes. Then, the proposed approach is applied to the evaluation of the tobacco control program implemented in California in 1988. The last section concludes the study.

3.2 Proposed Approach

3.2.1 Framework

An individual outcome is modeled as follows: for $j = 1, ..., J$; $t = 1, ..., T$,

$$y_{j,t}\left(s_{j,t}\right) = \gamma_{j,t} + \boldsymbol{x}_{j,t}^{\top}\boldsymbol{\beta} + u_{j,t}, \quad u_{j,t} \sim \mathcal{N}\left(0, \tau^{-1}\right),$$

where $\gamma_{j,t}$ is a unit- and time-specific intercept, $\boldsymbol{x}_{j,t}$ is an L-dimensional vector of covariates that may contain unit- and/or time-specific effects, $\boldsymbol{\beta}$ is the corresponding coefficient vector, and $u_{j,t}$ is an error term that is distributed according to a normal distribution with precision τ. $s_{j,t}$ is a treatment indicator: $s_{j,t} = 0$ if untreated and $s_{j,t} = 1$ if treated. A set of the treatment indicators is denoted by $\boldsymbol{S}$=$\{s_{j,t}\}$. The covariates $\boldsymbol{x}_{j,t}$ are completely observed for all the units and periods.

$\mathcal{I}_1$ and $\mathcal{I}_0$ are sets of indices for treated and untreated observations, respectively. $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_0$ denotes a set of all the indices for the observations. Let $\boldsymbol{Y}$ be a J-by-T matrix composed of (actually) untreated outcomes, $y_{j,t}\left(0\right)$, $(j,t) \in \mathcal{I}_0$, and counterfactual untreated outcomes, $y_{j,t}\left(0\right)$, $(j,t) \in \mathcal{I}_1$. The latter elements are also called "potential outcomes" (e.g., Im
. We define sets of observed and unobserved untreated outcomes respectively as

$$\boldsymbol{Y}^{obs} = \left\{y_{j,t}\left(0\right),\ (j,t) \in \mathcal{I}_0\right\}, \quad \boldsymbol{Y}^{miss} = \left\{y_{j,t}\left(0\right),\ (j,t) \in \mathcal{I}_1\right\}.$$

In this study, we treat $\boldsymbol{Y}^{miss}$ as missing and infer it as a set of unknown parameters via matrix completion, using a Markov chain Monte Carlo (MCMC) method. In other words, we transform the inferential problem into a matrix completion problem and infer the potential outcomes by imputing them via data augmentation (Tanner and Wong, 1987) (or, more generally, Gibbs sampling). The responses under treatment, $y_{j,t}\left(1\right)$, $(j,t) \in \mathcal{I}_1$, are observed, but they are not used for inference of $\boldsymbol{Y}^{miss}$.

Let $\boldsymbol{X} = \{\boldsymbol{x}_{j,t}\}$ be a set of covariates. Define a J-by-T matrix of the covariate effects as $\boldsymbol{\Xi} = (\xi_{j,t})$ with $\xi_{j,t} = \boldsymbol{x}_{j,t}^{\top}\boldsymbol{\beta}$. The model can be posed in a matrix representation as

$$\boldsymbol{Y} = \boldsymbol{\Gamma} + \boldsymbol{\Xi} + \boldsymbol{U},$$
$$\boldsymbol{\Gamma} = \left(\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_T\right), \quad \text{with } \boldsymbol{\gamma}_t = \left(\gamma_{1,t}, ..., \gamma_{J,t}\right)^{\top},$$

where $\boldsymbol{\Gamma}$ is a J-by-T matrix and $\boldsymbol{U} = (u_{j,t})$ is a J-by-T matrix of the error terms.

We have several assumptions in the model. First, we make the standard stable unit treatment value assumption (STUVA) (e.g., Imbens and Rubin, 2015): there is no interference between units, there is a single type of treatment, and each unit has two potential outcomes. In addition, the assignment mechanism is assumed to be unconfounded, i.e., ignorable, conditional on the covariates $\boldsymbol{X}$:

$$p\left(\boldsymbol{S}|\boldsymbol{Y}^{obs},\boldsymbol{Y}^{miss},\boldsymbol{X}\right)=p\left(\boldsymbol{S}|\boldsymbol{X}\right).$$

In contrast to the difference-in-differences estimator, the treated and untreated units are not supposed to have parallel trends in outcome.

A matrix of untreated outcomes $\boldsymbol{Y}$ can be structured flexibly. For instance, when only the Jth unit is affected by the treatment for the last $T-T_0$ periods as in the standard synthetic control method (SCM) (Abadie and Gardeazabal, 2003; Abadie et al., 2010), $\boldsymbol{Y}$ is specified as

$$\boldsymbol{Y}=\begin{pmatrix} y_{1,1}\left(0\right) & \cdots & y_{1,T_0}\left(0\right) & y_{1,T_0+1}\left(0\right) & \cdots & y_{1,T}\left(0\right) \\ \vdots & & \vdots & \vdots & & \vdots \\ y_{J-1,1}\left(0\right) & \cdots & y_{J-1,T_0}\left(0\right) & y_{J-1,T_0+1}\left(0\right) & \cdots & y_{J-1,T}\left(0\right) \\ y_{J,1}\left(0\right) & \cdots & y_{J,T_0}\left(0\right) & \surd & \cdots & \surd \end{pmatrix},$$

where $\surd$ denotes a missing entry. It is possible to allow more than one treated unit:

$$\boldsymbol{Y}=\begin{pmatrix} y_{1,1}\left(0\right) & \cdots & y_{1,T_0}\left(0\right) & y_{1,T_{0+1}}\left(0\right) & \cdots & y_{1,T_0+1}\left(0\right) \\ \vdots & & \vdots & \vdots & & \vdots \\ y_{J_0,1}\left(0\right) & \cdots & y_{J_0,T_0}\left(0\right) & y_{J_0,T_{0+1}}\left(0\right) & \cdots & y_{J_0,T_0+1}\left(0\right) \\ y_{J_0+1,1}\left(0\right) & \cdots & y_{J_0+1,T_0}\left(0\right) & \surd & \cdots & \surd \\ \vdots & & & \vdots & & \vdots \\ y_{J,1}\left(0\right) & \cdots & y_{J,T_0}\left(0\right) & \surd & \cdots & \surd \end{pmatrix}.$$

Furthermore, it is possible to handle a more complex structure:

$$\boldsymbol{Y}=\begin{pmatrix} \surd & \cdots & \surd & y_{1,t'}\left(0\right) & \cdots & \cdots & \cdots & y_{1,T}\left(0\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ y_{j^\sharp,1}\left(0\right) & \cdots & y_{j^\sharp,t'+1}\left(0\right) & \surd & y_{j^\sharp,t'+1}\left(0\right) & \cdots & \cdots & y_{j^\sharp,T}\left(0\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{j^\flat,1}\left(0\right) & \cdots & y_{j^\flat,t'+1}\left(0\right) & \surd & \surd & y_{j^\flat,t'+2}\left(0\right) & \cdots & y_{j^\flat,T}\left(0\right) \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ y_{J,1}\left(0\right) & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & y_{J,T}\left(0\right) \end{pmatrix}.$$

From a theoretical perspective, Bayesian inference of $\boldsymbol{Y}^{miss}$ is specified as follows. Let $\boldsymbol{\Theta}=\{\boldsymbol{\Gamma},\boldsymbol{\beta},\tau\}$ denote the set of all unknown parameters. If $\boldsymbol{Y}^{obs}$, $\boldsymbol{Y}^{miss}$, and $\boldsymbol{S}$ are given, the complete-data likelihood is represented as

$$\begin{aligned} p\left(\boldsymbol{Y}^{obs},\boldsymbol{Y}^{miss}|\boldsymbol{S},\boldsymbol{X},\boldsymbol{\Theta}\right) &= \left(2\pi\right)^{-\frac{JT}{2}}\tau^{\frac{JT}{2}}\exp\left\{-\frac{\tau}{2}\text{tr}\left(\boldsymbol{U}^\top\boldsymbol{U}\right)\right\} \\ &= \left(2\pi\right)^{-\frac{JT}{2}}\tau^{\frac{JT}{2}}\exp\left\{-\frac{\tau}{2}\text{vec}\left(\boldsymbol{U}\right)^\top\text{vec}\left(\boldsymbol{U}\right)\right\}, \\ \boldsymbol{U} &= \boldsymbol{Y}-\boldsymbol{\Gamma}-\boldsymbol{\Xi}, \end{aligned}$$

where $\text{vec}\left(\cdot\right)$ denotes the column-wise vectorization operator. The joint posterior distribution of the missing observations of the responses and the unknown parameters is proportional to

the product of the prior density of $\boldsymbol{\Theta}$ and the joint likelihood of all the potential outcomes $\{y_{j,t}(0), y_{j,t}(1), (j,t) \in \mathcal{I}\}$, the treatment indicators $\boldsymbol{S}$, and the covariates $\boldsymbol{X}$:

$$p\left(\boldsymbol{Y}^{miss}, \boldsymbol{\Theta} | \boldsymbol{Y}^{obs}, \boldsymbol{S}, \boldsymbol{X}\right) \propto p(\boldsymbol{\Theta}) \prod_{(j,t)\in\mathcal{I}} p\left(y_{j,t}(0), y_{j,t}(1), \boldsymbol{S}, \boldsymbol{X} | \boldsymbol{\Theta}\right),$$

where $p(\boldsymbol{\Theta})$ denotes the prior density of $\boldsymbol{\Theta}$. Given $\boldsymbol{Y}^{obs}$ and $\boldsymbol{\Theta}$, the conditional posterior distribution of the missing responses is given by

$$\begin{aligned} p\left(\boldsymbol{Y}^{miss} | \boldsymbol{Y}^{obs}, \boldsymbol{S}, \boldsymbol{X}, \boldsymbol{\Theta}\right) \propto & \prod_{(j,t)\in\mathcal{I}} p\left(s_{j,t} | y_{j,t}(0), y_{j,t}(1), \boldsymbol{x}_{j,t}, \boldsymbol{\Theta}\right) \\ & \times p\left(y_{j,t}(0), y_{j,t}(1) | \boldsymbol{x}_{j,t}, \boldsymbol{\Theta}\right) p\left(\boldsymbol{x}_{j,t} | \boldsymbol{\Theta}\right). \end{aligned}$$

By the unconfoundedness assumption, the assignment mechanism $p\left(s_{j,t} | y_{j,t}(0), \boldsymbol{x}_{j,t}, \boldsymbol{\Theta}\right)$ and the covariate distribution $p\left(\boldsymbol{x}_{j,t} | \boldsymbol{\Theta}\right)$ are ignorable:

$$\begin{aligned} p\left(\boldsymbol{Y}^{miss} | \boldsymbol{Y}^{obs}, \boldsymbol{S}, \boldsymbol{X}, \boldsymbol{\Theta}\right) \propto & \prod_{(j,t)\in\mathcal{I}_1} p\left(y_{j,t}(0), y_{j,t}(1) | \boldsymbol{x}_{j,t}, \boldsymbol{\Theta}\right) \\ & \times \prod_{(j,t)\in\mathcal{I}_0} p\left(y_{j,t}(0), y_{j,t}(1) | \boldsymbol{x}_{j,t}, \boldsymbol{\Theta}\right) \\ \propto & \prod_{(j,t)\in\mathcal{I}_1} p\left(y_{j,t}(0) | y_{j,t}(1), \boldsymbol{x}_{j,t}, \boldsymbol{\Theta}\right) \\ & \times \prod_{(j,t)\in\mathcal{I}_0} p\left(y_{j,t}(1) | y_{j,t}(0), \boldsymbol{x}_{j,t}, \boldsymbol{\Theta}\right) \\ \propto & \prod_{(j,t)\in\mathcal{I}_1} p\left(y_{j,t}(0) | y_{j,t}(1), \boldsymbol{x}_{j,t}, \boldsymbol{\Theta}\right) \\ = & \; p\left(\boldsymbol{Y}^{miss} | \boldsymbol{Y}^{obs}, \boldsymbol{X}, \boldsymbol{\Theta}\right). \end{aligned}$$

Thus, the conditional posterior of $\boldsymbol{Y}^{miss}$ depends only on the observed information $\left(\boldsymbol{Y}^{obs}, \boldsymbol{X}\right)$ and the parameters $\boldsymbol{\Theta}$, and it can be derived from the complete-data likelihood. In turn, the conditional posterior of $\boldsymbol{\Theta}$ is proportional to the product of the complete-data likelihood and the prior of $\boldsymbol{\Theta}$:

$$p\left(\boldsymbol{\Theta} | \boldsymbol{Y}^{miss}, \boldsymbol{Y}^{obs}, \boldsymbol{S}, \boldsymbol{X}\right) \propto p(\boldsymbol{\Theta}) p\left(\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{miss} | \boldsymbol{S}, \boldsymbol{X}, \boldsymbol{\Theta}\right).$$

Therefore, as the conditional posteriors of $\boldsymbol{Y}^{miss}$ and $\boldsymbol{\Theta}$ are simulable, we can conduct a posterior simulation using a Gibbs sampler: $\boldsymbol{Y}^{miss}$ and $\boldsymbol{\Theta}$ are alternately simulated from the corresponding conditional posterior distributions.

Once the approximation of the posterior distribution of $\boldsymbol{Y}^{miss}$ is obtained, we can evaluate treatment effects straightforwardly. For instance, the posterior density of ATET, φ, can be represented as

$$\begin{aligned} E\left[\varphi | \boldsymbol{Y}^{obs}, \boldsymbol{X}\right] = & \int\int \left[\frac{1}{|\mathcal{I}_1|} \sum_{(j,t)\in\mathcal{I}_1} \left(y_{j,t}(1) - y_{j,t}(0)\right)\right] \\ & \times p\left(\boldsymbol{Y}^{miss}, \boldsymbol{\Theta} | \boldsymbol{Y}^{obs}, \boldsymbol{X}\right) \mathrm{d}\boldsymbol{Y}^{miss} \mathrm{d}\boldsymbol{\Theta}. \end{aligned}$$

Given the posterior draws of the potential outcomes, the posterior mean estimate of ATET is computed as

$$\widehat{\varphi} = \frac{1}{N_{post}} \sum_{i=1}^{N_{post}} \left[\frac{1}{|\mathcal{I}_1|} \sum_{(j,t) \in \mathcal{I}_1} \left(y_{j,t}(1) - y_{j,t}^{(i)}(0) \right) \right],$$

where $y_{j,t}^{(i)}(0)$ denotes the ith posterior draw of the potential outcome of unit j at period t and N_{post} is the number of posterior draws used for the posterior analysis. The posterior estimates of the variance/quantiles of the posterior of ATET are obtained analogously.

3.2.2 Priors

As the structure of the model indicates, unless some restrictions are imposed, we cannot identify $\mathbf{\Gamma}$ and $\boldsymbol{Y}^{miss}$. We induce $\mathbf{\Gamma}$ to be low rank and decompose it into two parts as

$$\mathbf{\Gamma} = \mathbf{\Phi}\mathbf{\Psi}^{\top},$$

$$\mathbf{\Phi} = \left(\boldsymbol{\phi}_{(1)}, ..., \boldsymbol{\phi}_{(J)}\right)^{\top} \in \mathbb{R}^{J \times H}, \quad \text{with } \boldsymbol{\phi}_{(j)} = \left(\phi_{j,1}, ..., \phi_{j,H}\right)^{\top},$$

$$\mathbf{\Psi} = \left(\boldsymbol{\psi}_{(1)}, ..., \boldsymbol{\psi}_{(T)}\right)^{\top} \in \mathbb{R}^{T \times H}, \quad \text{with } \boldsymbol{\psi}_{(t)} = \left(\psi_{t,1}, ..., \psi_{t,H}\right)^{\top},$$

where $H < \min(J, T)$. Although this decomposition is not unique, as $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are not identified, exact parameter identification is not necessary for our purpose: we require the identification of the convolution, $\mathbf{\Gamma}$, not that of its factorization, $\mathbf{\Phi}$ and $\mathbf{\Psi}$.

Nevertheless, when $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are not identified, the posterior simulation can diverge, which is computationally inefficient. We use a prior motivated by singular value decomposition (SVD). When the SVD of $\mathbf{\Gamma}$ is represented as $\mathbf{\Gamma} = \boldsymbol{E}_1 \boldsymbol{D} \boldsymbol{E}_2^{\top}$, we interpret $\mathbf{\Psi} = \boldsymbol{E}_2$ as the right orthonormal matrix and $\mathbf{\Phi} = \boldsymbol{E}_1 \boldsymbol{D}$ as the product of the left orthonormal matrix $\boldsymbol{E}$ and the diagonal matrix having the eigenvalues in its principal diagonal $\boldsymbol{D}$. Two types of priors introduced in what follows correspond to the interpretation of $\mathbf{\Phi}$ and $\mathbf{\Psi}$.

First, we restrict $\mathbf{\Psi}$ to be unitary, i.e., $\mathbf{\Psi}^{\top}\mathbf{\Psi} = \boldsymbol{I}_H$, and assign a uniform Haar prior to $\mathbf{\Psi}$, $p(\mathbf{\Psi}) \propto \mathbb{I}(\mathbf{\Psi} \in \mathcal{M}_{T \times H})$, where $\mathcal{M}_{T \times H}$ denotes a Stiefel manifold with dimensions of $T \times H$ and $\mathbb{I}(\cdot)$ denotes the indicator function. This restriction implies that the covariance of the rows of $\mathbf{\Psi}$ is $T^{-1}\boldsymbol{I}_H$. Then, $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_T)$ can be regarded as being generated from a static factor model as

$$\boldsymbol{\gamma}_t = \mathbf{\Phi}\boldsymbol{\psi}_t, \quad t = 1, ..., T,$$

where $\boldsymbol{\psi}_t$ is interpreted as a vector of independently distributed "latent factors" and $\mathbf{\Phi}$ is interpreted as a matrix of "factor loadings".

Second, we arrange the relative magnitudes of the columns of $\mathbf{\Phi}$ in descending order. For this purpose, we adapt a cumulative shrinkage process prior (Legramanti et al., 2020) to our context. A prior of $\mathbf{\Phi}$ is specified by the following hierarchy:

$$\begin{aligned}
\phi_{j,h}|\lambda_h &\sim \mathcal{N}\left(0, \lambda_h^2\right), \quad j = 1, ..., J;\ h = 1, ..., H, \\
\lambda_h|\pi_h &\sim \pi_h \delta_{\lambda_\infty} + (1 - \pi_h)\, \mathcal{IG}(\kappa_1, \kappa_2), \quad h = 1, ..., H, \\
\pi_h &= \sum_{l=1}^{h} \omega_l, \quad \text{with } \omega_l = \zeta_l \prod_{m=1}^{l-1} (1 - \zeta_m), \quad h = 1, ..., H, \\
\zeta_h &\sim \mathcal{B}(1, \eta), \quad h = 1, ..., H - 1, \\
\zeta_H &= 1,
\end{aligned}$$

where $\mathcal{IG}(a, b)$ is an inverse gamma distribution with shape parameter a and rate parameter b, and $\mathcal{B}(a, b)$ is a beta distribution (of the first kind) with scale parameters a and b. The prior of $\phi_{j,h}$ is a scale mixture of normal distributions. The prior distribution of the variances λ_h belongs to a class of spike-and-slab priors (e.g., Ishwaran et al., 2005), in that the prior consists of spike δ_{λ_∞} and slab $\mathcal{IG}(\kappa_1, \kappa_2)$. Although δ_{λ_∞} can be zero, we set it to a small nonzero value for the ease of posterior simulation (Ishwaran et al., 2005; Legramanti et al., 2020). The prior distribution of the weights π_h exploits the stick-breaking construction of the Dirichlet process (Ishwaran and James, 2001). As h grows, the distribution of λ_h concentrates around δ_{λ_∞} since $\lim_{h\to\infty} \pi_h = 1$ almost surely.

In turn, for the remaining parameters, we employ standard priors. For $\boldsymbol{\beta}$, we use an independent normal prior with mean zero and precision α, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}_L,\ \alpha^{-1}\boldsymbol{I}_L)$. The prior distribution of τ is specified by a gamma distribution with shape parameter ν_1 and rate parameter ν_2, $\tau \sim \mathcal{G}(\nu_1, \nu_2)$.

Although we do not consider them in this study, many alternative priors can be used for $\boldsymbol{\Theta}$. Bhattacharya and Dunson (2011) consider a prior similar to the cumulative shrinkage process prior, called the multiplicative gamma process prior. This prior cannot simultaneously control the rate of shrinkage and the prior for the active elements; thus, it readily overshrinks the model. See Durante (2017) and Legramanti et al. (2020) for further discussion. In addition, many fully Bayesian approaches exist for estimating or completing low-rank matrices (e.g., Salakhutdinov and Mnih, 2008; Ding et al., 2011). However, these approaches do not consider parameter identification. The only exception is Tang et al. (2019). They factorize a possibly rank-deficient matrix $\boldsymbol{\Gamma}$ into three parts as in SVD, $\boldsymbol{\Gamma} = \boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{\Psi}^\top$, where $\boldsymbol{D}$ is diagonal. While they suppose $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ to be unitary, as in this study, the diagonal elements of $\boldsymbol{D}$ are not restricted: the ordering of rows of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ and the diagonal elements of $\boldsymbol{D}$ are freely permuted along the posterior simulation.

3.2.3 Posterior simulation

For posterior simulation, we develop an MCMC sampler. We conduct posterior simulations using a hybrid of two algorithms. To address the unitary constraint, we sample $\boldsymbol{\Psi}$ using the geodesic Monte Carlo on embedded manifolds (Byrne and Girolami, 2013). As the conditionals of the remaining parameters are standard, the remaining parameters are updated via Gibbs steps. See the Appendix for the computational details.

While Legramanti et al. (2020) adaptively tune the rank of a matrix of interest, we prefix the rank of $\boldsymbol{\Gamma}$, H, for several reasons. First, the unitary constraint on $\boldsymbol{\Psi}$ makes it difficult to change H adaptively. Second, as our prior pushes $\boldsymbol{\Gamma}$ to be low rank, it is unnecessary to exactly specify the true rank of $\boldsymbol{\Gamma}$: if the h'th eigenvalue of $\boldsymbol{\Gamma}$ is negligible, the prior standard deviation of the h'th row of $\boldsymbol{\Phi}$ is inclined to be δ_∞ (spike part). Therefore, we recommend choosing a conservative value for H or tuning H based on test runs.

3.2.4 Extensions

We mention some simple extensions. First, we can make the model more robust to outliers by modeling the measurement errors using a distribution with heavier tails than those of a normal distribution. For instance, following Geweke (1993), the generalized Student's t error is modeled

as

$$\begin{aligned} u_{j,t}|\tau,\omega_{j,t} &\sim \mathcal{N}\left(0,\ \tau^{-1}\omega_{j,t}^{-1}\right), \quad j=1,...,J;\ t=1,...,T, \\ \omega_{j,t}|\upsilon &\sim \mathcal{G}\left(\frac{\upsilon}{2},\frac{\upsilon}{2}\right), \quad j=1,...,J;\ t=1,...,T, \\ \upsilon &\sim f(\upsilon), \end{aligned}$$

where $\omega_{j,t}$ is an auxiliary random variable, υ is the number of degrees of freedom of $u_{j,t}$, and $f(\upsilon)$ is a prior distribution of υ.

Second, to allow serial correlations in the error terms, their distribution can be modeled as

$$\boldsymbol{u}_j = (u_{j,1},...,u_{j,T})^\top |\tau,\rho \sim \mathcal{N}\left(\mathbf{0}_T,\ \tau^{-1}\boldsymbol{R}\right),$$

$$\boldsymbol{R} = (r_{t,t'}),\ \text{with } r_{t,t'} = \rho^{|t-t'|},$$

where $\boldsymbol{R}$ is a correlation matrix whose generic element $r_{t,t'}$ is specified as a function of an autocorrelation parameter $\rho \in (-1,1)$. As the conditional posterior of ρ is not standard, ρ is sampled using, e.g., the random-walk Metropolis-Hastings algorithm.

3.2.5 Comparison with existing approaches

The class of SCMs (Abadie and Gardeazabal, 2003; Abadie et al., 2010) is closely related to the proposed approach. In SCMs, "synthetic" untreated outcomes are estimated as weighted sums of the untreated units. This approach imposes three strong assumptions: no intercept, nonnegativity of the weights, and weights that sum to one. However, none of these assumptions appears plausible in many real cases. The proposed approach is free from such restrictions. Doudchenko and Imbens (2017) propose an approach that does not impose any of these restrictions on the weights and use a penalty similar to the elastic net estimator (Zou and Hastie, 2005). Amjad et al. (2018) propose a robust synthetic control method (RSCM). The difference between RSCM and the abovementioned SCMs is that RSCM constructs a design matrix using the SVD of a matrix composed of the outcomes of untreated units: SVD is used for dimension reduction and denoising. Xu (2017) also considers a similar modeling strategy.

All the existing non-Bayesian approaches, including Abadie et al. (2010), Doudchenko and Imbens (2017), Amjad et al. (2018), and Xu (2017), share the same caveat: they cannot evaluate confidence intervals straightforwardly. Abadie et al. (2010) conduct a series of placebo studies, which can be interpreted as permutation tests to quantify the uncertainty of an inference, but the size of the permutation tests may be distorted as shown by Hahn and Shi (2017). No statistically sound method has been developed to estimate confidence intervals of synthetic control methods. Recently, Li (forthcoming) proposes a subsampling method to obtain confidence intervals. In contrast, our Bayesian approach can estimate credible intervals as a byproduct of posterior simulation.

Kim et al. (2020) develop a Bayesian version of Doudchenko and Imbens's (2017) approach. Instead of the elastic net penalty, they propose the use of a shrinkage prior, e.g., the horseshoe prior (Carvalho et al., 2010). As with Bayesian inference, their approach can consistently obtain credible intervals. Our fully Bayesian approach also enjoys the same advantage. Amjad et al. (2018) also mention a Bayesian version of RSCM, but the method is not fully Bayesian in that the SVD of an outcome matrix is treated as given, and uncertainty about the decomposition is ignored.

Our proposal is closely related to Athey et al. (2018), where an estimation problem is treated as a matrix completion problem with a nuclear norm penalty. Athey et al. (2018) call their estimator the matrix completion with a nuclear norm minimization estimator (MC-NNM). The prior of $\mathbf{\Gamma}$ used in our approach plays a similar role to the nuclear norm penalty because the nuclear norm is a convex relaxation of the rank constraint (Fazel et al., 2001). This family of approaches involving matrix completion has two notable advantages over SCMs. First, treatment is allowed to occur arbitrarily, not consecutively. Second, while SCMs use only pretreatment observations for estimation, this family exploits all the observations, including the treated periods (except treated outcomes). Therefore, this class is likely to be statistically more efficient than SCMs, as shown in the simulation study below. Similar to Amjad et al.'s (2018) approach, the matrix completion approaches intend to capture the underlying factor structure of panel data. While in Amjad et al.'s (2018) approach, a threshold for truncating the eigenvalues of an outcome matrix must be specified (hard thresholding), our approach and Athey et al.'s (2018) approach do not because the cumulative shrinkage process prior and the nuclear norm penalty automatically push the model to be low rank (soft thresholding). Indeed our approach prefixed H, but H is merely an upper bound of the rank of $\mathbf{\Gamma}$; our approach can infer the causal effects without presupposing/estimating the rank of $\mathbf{\Gamma}$. As with other non-Bayesian approaches, Athey et al.'s (2018) approach provides only a point estimation, while our proposal readily estimates credible intervals.

Finally, Brodersen et al. (2015) and Ning et al. (2019) also develop Bayesian approaches to causal inference that use structural time series models, more specifically, state-space models. Brodersen et al.'s (2015) approach relies on a univariate state-space model, while Ning et al.'s (2019) approach uses a multivariate state-space model that allows spatial correlations between units. These two approaches are similar to ours in that both tend to obtain potential outcomes using Bayesian methods. On the other hand, there is a notable difference between their approaches and ours: their approaches rely on the fit of a state-space model, while our approach exploits the factor structure of panel data. As a consequence, our proposal is better suited for typical panel data where due to the short sample, it is difficult to recover the dynamics of the potential outcomes from the observations.

3.3 Application

3.3.1 Simulated data

We conduct a simulation study to demonstrate the proposed approach. In our experimental setting, only the Jth unit is treated, and it is exposed to the treatment during the last T_1 periods of T. Let T_0 denote the number of untreated periods; thus, $T = T_0 + T_1$. The realized treated outcomes are specified by the sums of hypothetical untreated outcomes $y_{J,t}(0)$ and the average treatment effect on treated denoted by φ:

$$y_{J,t}(1) = y_{J,t}(0) + \varphi, \quad t = T_0 + 1, ..., T.$$

We consider three types of data-generating processes (DGPs). In the first two types, $y_{j,t}(0)$ is generated from a factor model: for $j = 1, ..., J$; $t = 1, ..., T$,

$$(y_{1,t}(0), ..., y_{J,t}(0))^\top = \boldsymbol{y}_t(0) = \mathbf{\Phi}\boldsymbol{\psi}_t + \boldsymbol{u}_t,$$
$$\mathbf{\Psi} = (\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_T)^\top, \quad \text{with } \boldsymbol{\psi}_t = (\psi_{1,t}, \psi_{2,t}, \psi_{3,t})^\top$$

$$\boldsymbol{u}_t = (u_{1,t}, ..., u_{J,t})^\top \sim \mathcal{N}(\mathbf{0}_J, \boldsymbol{I}_J),$$

where $\boldsymbol{\psi}_t$ is a vector of latent factors, $\boldsymbol{\Phi}$ is a matrix of factor loadings, and $\boldsymbol{u}_t$ is a vector of error terms. We do not include any covariates. Entries in $\boldsymbol{\Phi}$ are generated independently from a standard normal distribution:

$$\boldsymbol{\Phi} = (\phi_{j,t}), \quad \text{with } \phi_{j,t} \sim \mathcal{N}(0,1).$$

In the first case, called DGP-independent, latent factors are independently distributed according to a normal distribution specified as

$$\boldsymbol{\psi}_t \sim \mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3), \quad t = 1, ...T.$$

The second case is called DGP-dependent, where the row of motion of $\boldsymbol{\psi}_t$ is specified by the following process:

$$\psi_{j,1} = \epsilon_{j,1}, \quad j = 1, 2, 3,$$

$$\begin{cases} \psi_{1,t} = 0.6\psi_{1,t-1} + \epsilon_{1,t} \\ \psi_{2,t} = 0.4\psi_{2,t-1} + \epsilon_{2,t}, \quad t = 2, 3, ..., T, \\ \psi_{3,t} = 0.2\psi_{3,t-1} + \epsilon_{3,t} \end{cases}$$

$$\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \epsilon_{2,t}, \epsilon_{3,t})^\top \sim \mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3), \quad t = 1, ..., T.$$

The third case, DGP-weighted, is motivated by a simulation study in Kim et al. (2020). In this setting, outcomes of untreated units are generated from a multivariate normal distribution, and outcomes of treated units are constructed as weighted sums of untreated units:

$$y_{J,t} = \sum_{j=1}^{J-1} \alpha_j y_{j,t} + u_{J,t}, \quad u_{J,t} \sim \mathcal{N}\left(0, \sigma^2\right), \quad t = 1, ..., T,$$

$$(y_{1,t}(0), ..., y_{J-1,t}(0))^\top = \boldsymbol{y}_{1:J-1,t}(0) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad t = 1, ..., T,$$

$$\boldsymbol{\alpha} = (\alpha_j), \quad \alpha_j = \begin{cases} 3 & j = 1 \\ 2 & j = 2 \\ 1 & j = 3 \\ 0 & j = 4, ..., J-1 \end{cases},$$

$$\boldsymbol{\Sigma} = (\sigma_{i,j}), \quad \sigma_{i,j} = \begin{cases} 10 & i = j \\ 0.5 & i \neq j \end{cases}.$$

$\boldsymbol{\mu}$ is specified as follows:

$$\boldsymbol{\mu} = (\mu_j), \quad \mu_j = \begin{cases} 10 & j = 1 \\ 20 & j = 2 \\ 30 & j = 3 \\ 40 & j = 4 \\ 15 & j = 5, ..., 9 \end{cases}, \text{ for } J = 10,$$

and

$$\boldsymbol{\mu} = (\mu_j)\,, \quad \mu_j = \begin{cases} 10 & j = 1 \\ 20 & j = 2 \\ 30 & j = 3 \\ 40 & j = 4 \\ 15 & j = 5, ..., 10 \\ 25 & j = 11, ..., 20 \\ 35 & j = 21, ..., 30 \\ 45 & j = 31, ..., 39 \end{cases}, \text{ for } J = 40.$$

We compare six alternative approaches.

1. The first is the original synthetic control method described in Abadie et al. (2010) (SCM-ABD).

2. The second is a method proposed by Doudchenko and Imbens (2017) (SCM-DI).

3. The third is a Bayesian approach developed by Kim et al. (2020). According to their simulation study, specifications with the horseshoe (Carvalho et al., 2010) and spike-and-slab (Ishwaran et al., 2005) priors outperform other alternatives. While the performances of these two priors are comparable, posterior simulation using the horseshoe prior is faster. Thus, we consider the horseshoe prior for Kim et al.'s (2019) approach and refer to this specific approach as Bayesian synthetic control method (BSCM). We sample weighting parameters in the observation model using the elliptical slice sampler (Hahn et al., 2019) and obtain the remaining parameters (noise variance and shrinkage parameters) using a Gibbs sampler, as in Makalic and Schmidt (2016).

4. The fourth is the robust synthetic control method introduced by Amjad et al. (2018) (RSCM). Specifically, we consider their primary choice described in Algorithm 1 of the original paper (p. 8).

5. The fifth is the matrix completion with a nuclear norm minimization estimator Athey et al. (2018) (MC-NNM).

6. The sixth is the proposed approach, Bayesian matrix completion with the cumulative shrinkage process prior (BMC-CSP). The prefixed hyperparameters for the cumulative shrinkage process prior are chosen following Legramanti et al. (2020) as $\eta = 5$ and $\kappa_1 = \kappa_2 = 2$. While Legramanti et al. (2020) use $\delta_{\lambda_\infty} = 0.05$, we use a smaller value, $\delta_{\lambda_\infty} = 0.01$. The maximum rank of $\boldsymbol{\Theta}$ is set to $H = \min(J, T)$.

To ensure a fair comparison, we use the same prior for the error variance in BSCM as in the proposed approach. We choose the hyperparameters as $\nu_1 = \nu_2 = 0.001$, inducing the prior of τ to be fairly noninformative. For MC-NNM, we choose tuning parameters via five-fold cross-validation, where the training samples are randomly chosen without replacement. For SCM-DI and RSCM, the tuning parameters are determined by forward chaining: the tuning parameters are chosen by minimizing the mean squared errors of one-step-ahead out-of-sample predictions, and the training sample is initially set to five and expanded sequentially to $T_0 - 1$. For BSCM and BMC-CSP, we obtain 40,000 draws after discarding the initial 10,000.[3] All the posterior simulations pass Geweke's (1992) convergence test at a significance level of 5%.

[3]The number of MCMC iterations is chosen based on pilot runs so that the minimum of the effective sample sizes for obtained posterior draws (except the warmup draws) in each experiment is no less than 10,000.

We consider four types of sample size, namely, combinations of $J \in \{5, 20\}$ and $T_0 \in \{10, 40\}$, and the length of the treated periods is fixed to $T_1 = 20$. A total of 200 experiments are conducted for each case. As noted earlier, an estimation of treatment effects amounts to an estimation of potential outcomes. Therefore, we evaluate the alternatives based on the precision of the estimates of $y_{j,t}(0)$, $t = T_0 + 1, ..., T$, measured by the mean of the sum of the squared errors (MSE) and the mean of the sum of the absolute errors (MAE). For the Bayesian approaches, we compute posterior means of predicted potential outcomes. We also report the mean computation time measured in seconds (Time).[4] For each experiment, the MSE and MAE are normalized by the corresponding values for SCM-ABD.

Table 3.1 summarizes the results of the simulation study for DGP-independent. In terms of MSE and MAE, irrespective of the combination of (J, T_0), the proposed approach consistently outperforms the others. The recently proposed alternatives are comparable to or worse than SCM-ABD. In this setting, RSCM is consistently inferior to the original SCM-ABD. In terms of computational time, as expected, the Bayesian approaches are slower than the non-Bayesian options. Indeed, BMC-CSP is computationally heavy, but the computational cost is not prohibitive. In our simulation study, SCM-DI is slow to converge, possibly due to nonsmooth objective functions. The simulation results for DGP-dependent are reported in Table 3.2. In terms of MSE and MAE, RSCM, MC-NNM, and BMC-SCP perform best. Table 3.3 summarizes the results for DGP-weighted. SCM-DI and BSCM perform very well because this DGP is exactly consistent with the DGPs of the models. In contrast to the other DGPs, the predictive accuracy of RSCM and MC-NNM is much worse than that of the others, including SCM-ABD. Although BMC-SCP performs worse than SCM-DI and BSCM, it consistently outperforms the remaining approaches. In summary, while the relative finite sample performance of the alternative approaches depends on the DGP, the proposed approach, BMC-SCP, is fairly competitive under various circumstances.

3.3.2 Real data

As an illustration, we apply the proposed approach to evaluate California's tobacco control program implemented in 1988. We replicate Abadie et al.'s (2010) study using the same data, annual state-level panel data spanning periods from 1970 to 2000.[5] The first 19 years are the pretreatment period. Only California is treated, while the other 38 states are used as control units. We include seven time-invariant covariates: log of gross domestic product per capita, percentage share of 15–24–year–old people in the population, retail price, beer consumption per capita, and cigarette sales per capita in 1980 and 1975; see Abadie et al. (2010) for further details. We use the same hyperparameters as in the simulation study. We draw 100,000 posterior samples and use the last 80,000 samples for posterior analysis.[6]

Figure 3.1 compares the realized per capita cigarette sales in California (solid black line), the potential per capita cigarette sales in "synthetic California" obtained using the original SCM (Abadie et al., 2010) (dashed black line), and the posterior mean estimates of the corresponding potential outcomes obtained by the proposed method (solid red line). The estimates obtained using the proposed method are in line with the estimates obtained using the original SCM.

[4]We wrote all the programs in Matlab R2019b (64 bit) and executed them on an Ubuntu Desktop 18.04 LTS (64 bit), running on AMD Ryzen Threadripper 1950X (4.2GHz).

[5]The data and the Matlab program were downloaded from Jens Hainmueller's personal website. (https://web.stanford.edu/~jhain/synthpage.html)

[6]We also conduct a posterior simulation where the unitary constraint on $\mathbf{\Psi}$ is removed and $\mathbf{\Psi}$ is sampled via a standard Gibbs step, but this approach is unsuccessful because the Markov chains diverge, resulting in numerical error.

Table 3.1: Results of simulation study (1): DGP-independent

(J, T_0)	Approach	MSE	MAE	Time
$(5, 10)$	(1) SCM-ABD	1.00	1.00	0.1
	(2) SCM-DI	1.48	1.17	24.7
	(3) BSCM	1.60	1.23	11.3
	(4) RSCM	1.22	1.08	6.5
	(5) MC-NNM	0.99	0.98	0.8
	(6) BMC-CSP	0.95	0.96	92.7
$(5, 40)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	1.54	1.21	96.0
	(3) BSCM	1.53	1.21	28.3
	(4) RSCM	1.27	1.11	8.1
	(5) MC-NNM	1.09	1.04	1.6
	(6) BMC-CSP	0.90	0.95	507.8
$(20, 10)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	0.96	0.97	175.5
	(3) BSCM	0.97	0.97	14.9
	(4) RSCM	1.29	1.10	53.2
	(5) MC-NNM	1.00	0.98	0.8
	(6) BMC-CSP	0.90	0.93	96.1
$(20, 40)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	1.16	1.06	728.6
	(3) BSCM	1.68	1.27	46.1
	(4) RSCM	1.59	1.24	74.3
	(5) MC-NNM	1.09	1.04	2.1
	(6) BMC-CSP	0.94	0.97	514.5

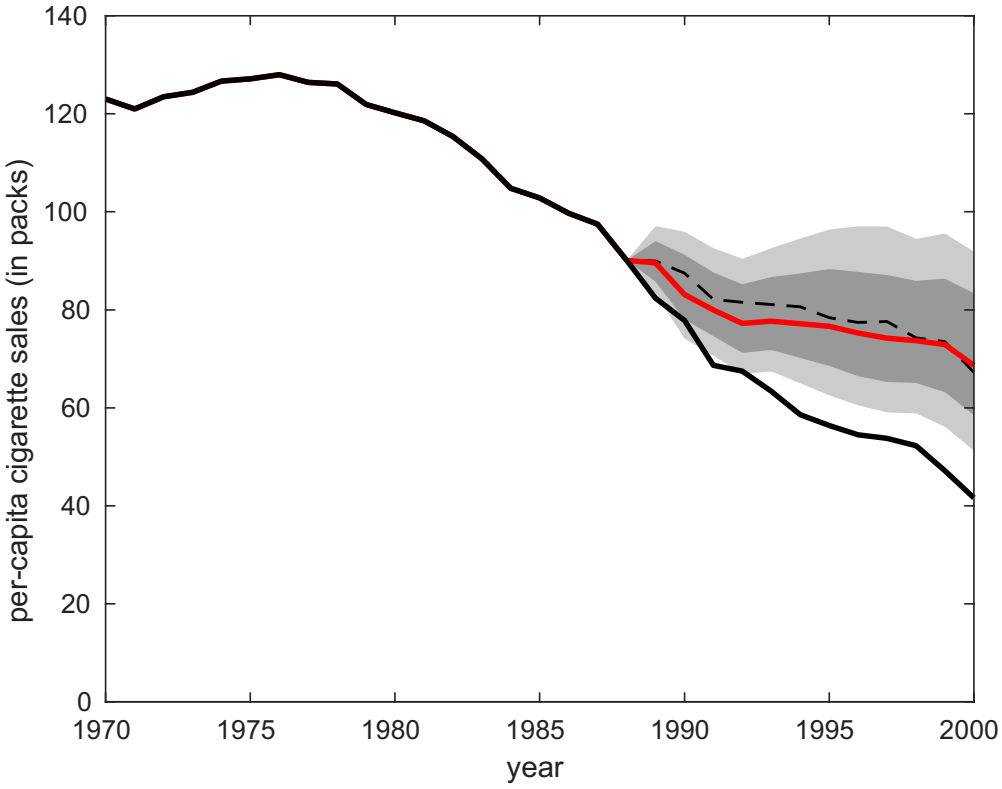
Table 3.2: Results of simulation study (2): DGP-dependent

(J, T_0)	Approach	MSE	MAE	Time
$(5, 10)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	1.44	1.15	23.9
	(3) BSCM	1.57	1.21	10.3
	(4) RSCM	0.71	0.84	6.1
	(5) MC-NNM	0.73	0.85	0.6
	(6) BMC-CSP	0.71	0.84	93.8
$(5, 40)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	1.53	1.22	97.9
	(3) BSCM	1.42	1.18	26.0
	(4) RSCM	0.81	0.89	7.6
	(5) MC-NNM	0.88	0.93	1.1
	(6) BMC-CSP	0.79	0.89	508.7
$(20, 10)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	0.86	0.92	170.6
	(3) BSCM	0.87	0.92	14.0
	(4) RSCM	0.75	0.86	52.0
	(5) MC-NNM	0.76	0.86	0.5
	(6) BMC-CSP	0.75	0.86	97.0
$(20, 40)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	1.47	1.19	698.4
	(3) BSCM	1.75	1.30	42.3
	(4) RSCM	0.87	0.93	65.9
	(5) MC-NNM	0.90	0.95	1.5
	(6) BMC-CSP	0.88	0.94	518.6

Table 3.3: Results of simulation study (3): DGP-weighted

(J, T_0)	Approach	MSE	MAE	Time
$(5, 10)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	0.03	0.17	25.4
	(3) BSCM	0.03	0.17	17.3
	(4) RSCM	23.28	5.70	12.7
	(5) MC-NNM	15.18	4.62	3.2
	(6) BMC-CSP	0.63	0.78	90.4
$(5, 40)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	0.39	0.60	70.1
	(3) BSCM	0.43	0.63	54.1
	(4) RSCM	14.67	4.60	23.9
	(5) MC-NNM	13.88	4.44	6.3
	(6) BMC-CSP	0.71	0.83	507.3
$(20, 10)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	0.02	0.13	211.2
	(3) BSCM	0.02	0.13	21.0
	(4) RSCM	26.56	6.17	86.7
	(5) MC-NNM	11.74	3.96	3.5
	(6) BMC-CSP	0.05	0.19	92.2
$(20, 40)$	(1) SCM-ABD	1.00	1.00	<0.1
	(2) SCM-DI	0.03	0.17	622.1
	(3) BSCM	0.03	0.17	84.1
	(4) RSCM	22.77	5.74	185.6
	(5) MC-NNM	9.70	3.63	7.4
	(6) BMC-CSP	0.41	0.63	511.9

Figure 3.1: Trends in per-capita cigarette sales



Notes: The solid black line traces the realized per capita cigarette sales in California. The dashed black line and the solid red line trace the estimated potential per capita cigarette sales using SCM-ABD and BMC-CSP, respectively. The light and dark shaded areas indicate the 90% and 70% credible sets, respectively.

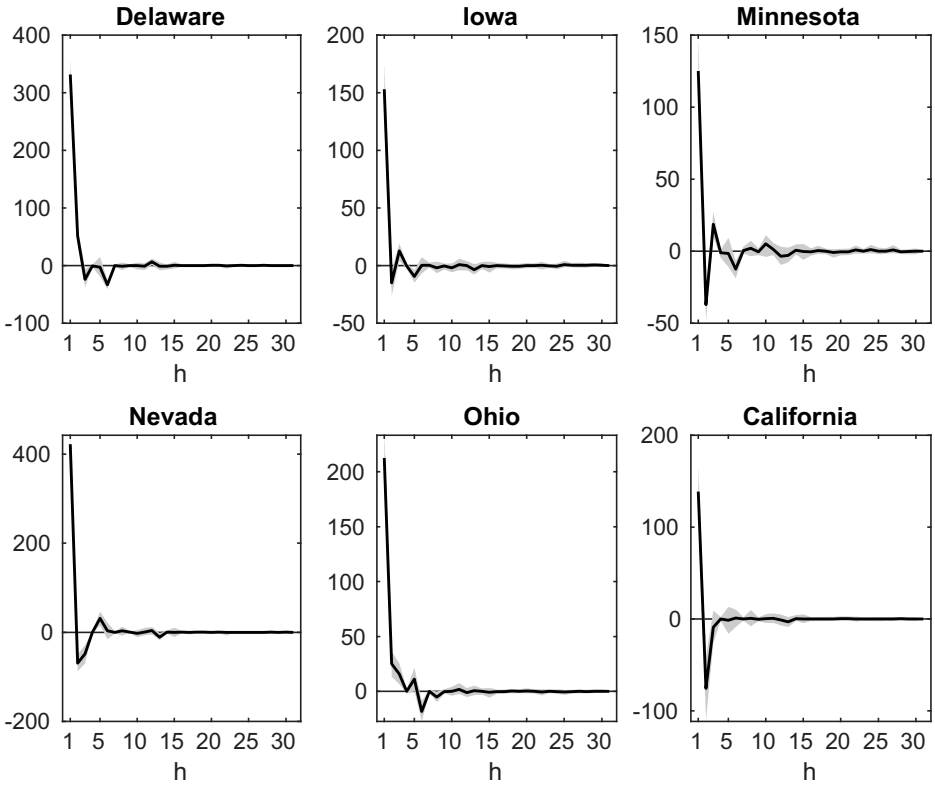
Posterior estimates of 90% and 70% credible sets are also reported (shaded areas). As the credible sets do not include the realized California, the program has statistically significant effects on tobacco consumption in California, confirming the conclusion in the original paper.

Figure 3.2 depicts the posterior estimates of some rows of $\mathbf{\Phi}$, which can be interpreted as state-specific loadings. While the estimates have different patterns, reflecting heterogeneity in US states, their magnitude is roughly decreasing with $h = 1, ..., H$ as intended by the prior. Figure 3.3 plots the posterior mean estimates of the eigenvalues of $\mathbf{\Gamma}$, which suggests that approximately half of the eigenvalues are not essential.

3.4 Concluding Remarks

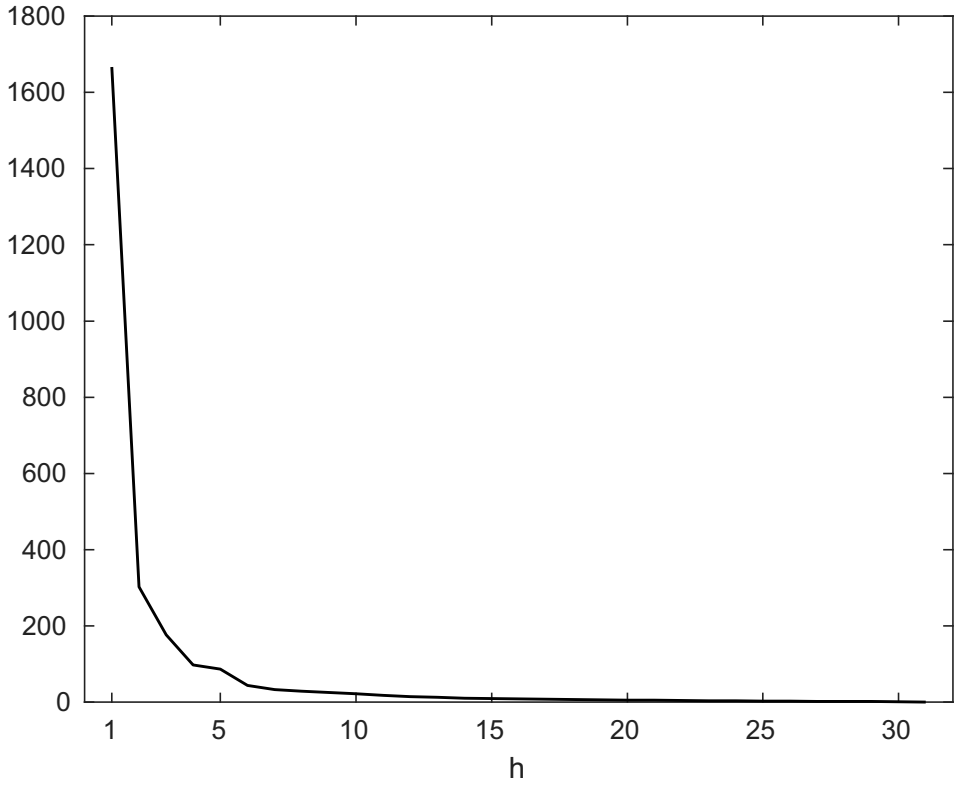
This study develops a novel Bayesian approach to causal analysis using panel data. We treat the problem of inferring a treatment effect as a matrix completion problem: counterfactual untreated outcomes are inferred using a data augmentation technique. We also propose a prior structured to help identification and to obtain a low-rank approximation of the panel data. In contrast to existing non-Bayesian methods, the proposed Bayesian approach can estimate credible intervals straightforwardly. By means of a series of simulation studies, we show that the

Figure 3.2: Posterior estimates of the rows of Φ.



Notes: The solid black lines trace the posterior mean estimates of the rows of Φ. The shaded areas indicate the 90% credible sets.

Figure 3.3: Posterior mean estimates of the eigenvalues of $\mathbf{\Gamma}$.



proposed approach outperforms the existing ones in terms of the prediction of hypothetical untreated outcomes, that is, the accuracy of the treatment effect estimates.

While asymptotic argument is not absolutely necessary for Bayesian analysis, there is a need to investigate frequentist (asymptotic) properties of the proposed approach, such as posterior consistency and Bernstein-von Mises theorem. However, to the best of the author's knowledge, there is no published work on frequentist properties of Bayesian matrix factorization/completion, except Mai and Alquier (2015).[7] The author hopes that this study stimulates further theoretical studies in the related research horizons.

[7]Mai and Alquier (2015) propose a Bayesian estimator for the matrix completion method and provide an oracle inequality for this estimator. However, they employ a uniform prior, and the proof critically depends on this prior choice; thus, their discussion is not easily extended to other environments.

Appendix: Computational Details

This appendix describes the computational details of the posterior simulation of the proposed approach. The joint posterior is specified as

$$\begin{aligned}
p\left(\boldsymbol{Y}^{miss}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\zeta}, \boldsymbol{\Lambda} | \boldsymbol{Y}^{obs}, \boldsymbol{X}\right) \propto & \; p\left(\boldsymbol{Y}^{obs} | \boldsymbol{Y}^{miss}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\beta}, \tau; \boldsymbol{X}\right) p\left(\boldsymbol{Y}^{miss}\right) p\left(\tau\right) \\
& \times p\left(\boldsymbol{\beta}\right) p\left(\boldsymbol{\Psi}\right) p\left(\boldsymbol{\Phi} | \boldsymbol{\Lambda}\right) p\left(\boldsymbol{\Lambda} | \boldsymbol{\zeta}\right) p\left(\boldsymbol{\zeta}\right) \\
\propto & \; \tau^{\frac{JT}{2}} \exp\left\{-\frac{\tau}{2} \text{tr}\left(\boldsymbol{U}^{\top} \boldsymbol{U}\right)\right\} \times \tau^{\nu_1 - 1} \exp\left(-\nu_2 \tau\right) \\
& \times \exp\left\{-\frac{\alpha}{2} \boldsymbol{\beta}^{\top} \boldsymbol{\beta}\right\} \times \mathbb{I}\left(\boldsymbol{\Psi} \in \mathcal{M}_{T \times H}\right) \\
& \times \prod_{j=1}^{J} \exp\left\{-\frac{1}{2} \boldsymbol{\phi}_{(j)}^{\top} \text{diag}\left(\lambda_1^{-1}, ..., \lambda_H^{-1}\right) \boldsymbol{\phi}_{(j)}\right\} \\
& \times \prod_{h=1}^{H} \left[\left(\sum_{l=1}^{h} \zeta_l \prod_{m=1}^{l-1} \left(1 - \zeta_m\right)\right) \delta_{\lambda_\infty} \right. \\
& + \left(1 - \sum_{l=1}^{h} \zeta_l \prod_{m=1}^{l-1} \left(1 - \zeta_m\right)\right) \\
& \left. \times \left\{\frac{\kappa_2^{\kappa_1}}{\Gamma\left(\kappa_1\right)} \lambda_h^{-\kappa_1 - 1} \exp\left(-\frac{\kappa_2}{\lambda_h}\right)\right\}\right] \\
& \times \prod_{h=1}^{H-1} \left(1 - \zeta_h\right)^{\eta - 1},
\end{aligned}$$

where $\boldsymbol{\zeta} = \left(\zeta_1, ..., \zeta_{H-1}\right)^{\top}$ and $\boldsymbol{\Lambda} = \text{diag}\left(\lambda_1, ..., \lambda_H\right)$. Each sampling block is specified in what follows.

Sampling $\boldsymbol{\Phi}$ Each row of $\boldsymbol{\Phi}$ is sampled from a multivariate normal distribution. For $j = 1, ..., J$,

$$\begin{aligned}
\boldsymbol{\phi}_{(j)} | \text{rest} &\sim \mathcal{N}\left(\boldsymbol{m}_{\boldsymbol{\phi}_{(j)}}, \boldsymbol{P}_{\boldsymbol{\phi}_{(j)}}^{-1}\right), \\
\boldsymbol{m}_{\boldsymbol{\phi}_{(j)}} &= \boldsymbol{P}_{\boldsymbol{\phi}_{(j)}}^{-1} \boldsymbol{\Psi}^{\top} \left(\boldsymbol{y}_{(j)} - \boldsymbol{\xi}_{(j)}\right), \\
\boldsymbol{P}_{\boldsymbol{\phi}_{(j)}} &= \boldsymbol{\Lambda}^{-1} + \boldsymbol{\Psi}^{\top} \boldsymbol{\Psi}, \\
\boldsymbol{Y} = \left(\boldsymbol{y}_{(1)}, ..., \boldsymbol{y}_{(J)}\right)^{\top}, & \quad \boldsymbol{\Xi} = \left(\boldsymbol{\xi}_{(1)}, ..., \boldsymbol{\xi}_{(J)}\right)^{\top}.
\end{aligned}$$

Sampling the shrinkage parameters Define an indicator function z_h with probability mass function $Pr\left(z_h = 1\right) = \omega_l$, $l = 1, ..., H$ and

$$\lambda_h | z_h \sim \mathbb{I}\left(z_h \leq h\right) \delta_{\lambda_\infty} + \left(1 - \mathbb{I}\left(z_h \leq h\right)\right) \mathcal{IG}\left(\kappa_1, \kappa_2\right).$$

Then the conditional posterior mass function of z_h is specified as

$$p\left(z_h = l | \text{rest}\right) \propto \begin{cases} \omega_l \text{N}\left(\boldsymbol{\phi}_h | \boldsymbol{0}_J, \; \lambda_\infty \boldsymbol{I}_J\right), & l = 1, ..., h, \\ \omega_l \text{t}_{2\kappa_1}\left(\boldsymbol{\phi}_h | \boldsymbol{0}_J, \; \frac{\kappa_2}{\kappa_1} \boldsymbol{I}_J\right), & l = h+1, ..., H, \end{cases}$$

where $\mathrm{N}(\boldsymbol{x}|\boldsymbol{a},\boldsymbol{B})$ is the PDF of a multivariate normal distribution with mean $\boldsymbol{a}$ and covariance $\boldsymbol{B}$ evaluated at $\boldsymbol{x}$ and $\mathrm{t}_c(\boldsymbol{x}|\boldsymbol{a},\boldsymbol{B})$ is the PDF of a multivariate t distribution with location parameter $\boldsymbol{a}$, scale parameter $\boldsymbol{B}$, and c degrees of freedom. The sampling distributions of ζ_l and λ_h are

$$\begin{aligned}
\zeta_h|\text{rest} &\sim \mathcal{B}\left(1+\sum_{l=1}^{H}\mathbb{I}(z_l=h),\quad \eta+\sum_{l=1}^{H}\mathbb{I}(z_l>h)\right),\quad h=1,...,H-1,\\
\lambda_h|\text{rest} &\sim \mathbb{I}(z_h\leq h)\,\delta_{\lambda_\infty}+(1-\mathbb{I}(z_h\leq h))\,\mathcal{IG}\left(\kappa_1+\frac{J}{2},\ \kappa_2+\frac{1}{2}\sum_{j=1}^{J}\phi_{j,h}^2\right),\quad h=1,...,H.
\end{aligned}$$

Sampling $\boldsymbol{\Psi}$ To sample $\boldsymbol{\Psi}$, we employ the geodesic Monte Carlo on embedded manifolds developed by Byrne and Girolami (2013). The algorithm for sampling $\boldsymbol{\Psi}$ is summarized in Algorithm 3.1. Let $\pi(\boldsymbol{\Psi})$ be the posterior density of $\boldsymbol{\Psi}$ conditional on the other parameters. Then we have

$$\log\pi(\boldsymbol{\Psi})=\text{(constant)}-\frac{\tau}{2}\text{tr}\left\{\left(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{\Psi}^\top-\boldsymbol{\Xi}\right)^\top\left(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{\Psi}^\top-\boldsymbol{\Xi}\right)\right\},$$

and the gradient with respect to $\boldsymbol{\Psi}$ is derived as

$$\nabla_{\boldsymbol{\Psi}}\log\pi(\boldsymbol{\Psi})=\tau\left(\boldsymbol{Y}-\boldsymbol{\Xi}\right)^\top\boldsymbol{\Phi}-\tau\boldsymbol{\Psi}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}.$$

The step size ε is adaptively tuned to maintain the average acceptance rate near a target value a^*. In the ith iteration, ε is updated according to the following rule which is motivated by the Robbins-Monro algorithm (Robbins and Monro, 1951):[8]

$$\log(\varepsilon)\leftarrow\log(\varepsilon)+i^{-1/\varsigma}(a^*-\bar{a}_i),$$

where $\bar{a}_i$ is the average acceptance rate in the ith iteration and $\varsigma\in(0.5,1)$ is a tuning parameter. We choose $a^*=0.6$ and $\varsigma=0.6$.[9] The number of steps is fixed to five, $N_{step}=5$, based on pilot runs.

Sampling $\boldsymbol{\beta}$ $\boldsymbol{\beta}$ is simulated from a multivariate normal distribution:

$$\begin{aligned}
\boldsymbol{\beta}|\text{rest}&\sim\mathcal{N}\left(\boldsymbol{m}_{\boldsymbol{\beta}},\boldsymbol{P}_{\boldsymbol{\beta}}^{-1}\right),\\
\boldsymbol{m}_{\boldsymbol{\beta}}&=\tau\boldsymbol{P}_{\boldsymbol{\beta}}^{-1}\boldsymbol{X}^\top\text{vec}\left(\boldsymbol{Y}-\boldsymbol{\Theta}\right),\\
\boldsymbol{P}_{\boldsymbol{\beta}}&=\alpha\boldsymbol{I}_L+\tau\boldsymbol{X}^\top\boldsymbol{X},\\
\boldsymbol{X}&=\begin{pmatrix}\boldsymbol{X}_1\\ \vdots\\ \boldsymbol{X}_T\end{pmatrix},\ \text{with }\boldsymbol{X}_t=\begin{pmatrix}\boldsymbol{x}_{1,t}^\top\\ \vdots\\ \boldsymbol{x}_{J,t}^\top\end{pmatrix}.
\end{aligned}$$

Sampling τ τ is updated via the following gamma distribution:

$$\tau|\text{rest}\sim\mathcal{G}\left(\nu_1+\frac{JT}{2},\ \nu_2+\frac{1}{2}\text{tr}\left(\boldsymbol{U}^\top\boldsymbol{U}\right)\right).$$

[8] Similar rules are considered for random-walk Metropolis-Hastings algorithms (e.g., Atchadé and Rosenthal, 2005; Andrieu and Thoms, 2008; Vihola, 2011) and an adaptive version of the Metropolis adjusted Langevin algorithm (Atchadé, 2006).

[9] The target acceptance rate a^* is chosen based on a multivariate effective sample size (Vats et al., 2019).

Algorithm 3.1 Geodesic Monte Carlo

Input: $\boldsymbol{\Psi}_0$ (current state), $\pi(\boldsymbol{\Psi})$ (target kernel), ε (step size), N_{step} (number of steps)

$\text{vec}(\boldsymbol{V}) \sim \mathcal{N}(\mathbf{0}_{TH}, \boldsymbol{I}_{TH})$

$\boldsymbol{V} \leftarrow \boldsymbol{V} - \frac{1}{2}\boldsymbol{\Psi}_0\left(\boldsymbol{\Psi}_0^\top \boldsymbol{V} + \boldsymbol{V}^\top \boldsymbol{\Psi}_0\right)$

$\mathcal{H}_0 \leftarrow \log \pi(\boldsymbol{\Psi}_0) - \frac{1}{2}\text{vec}(\boldsymbol{V})^\top \text{vec}(\boldsymbol{V})$

$\boldsymbol{\Psi}_1 \leftarrow \boldsymbol{\Psi}_0$

for $i = 1, ..., N_{step}$ do

  $\boldsymbol{V} \leftarrow \boldsymbol{V} + \frac{\varepsilon}{2}\nabla_{\boldsymbol{\Psi}_1} \log \pi(\boldsymbol{\Psi}_1)$

  $\boldsymbol{V} \leftarrow \boldsymbol{V} - \frac{1}{2}\boldsymbol{\Psi}_1\left(\boldsymbol{\Psi}_1^\top \boldsymbol{V} + \boldsymbol{V}^\top \boldsymbol{\Psi}_1\right)$

  $\boldsymbol{A} \leftarrow \boldsymbol{\Psi}_1^\top \boldsymbol{V}$, $\boldsymbol{S} \leftarrow \boldsymbol{V}^\top \boldsymbol{V}$

  $\begin{pmatrix} \boldsymbol{\Psi}_1 & \boldsymbol{V} \end{pmatrix} \leftarrow \begin{pmatrix} \boldsymbol{\Psi}_1 & \boldsymbol{V} \end{pmatrix} \exp\left(\varepsilon \begin{pmatrix} \boldsymbol{A} & -\boldsymbol{S} \\ \boldsymbol{I}_H & \boldsymbol{A} \end{pmatrix}\right) \begin{pmatrix} \exp(-\varepsilon \boldsymbol{A}) & \boldsymbol{O}_{H\times H} \\ \boldsymbol{O}_{H\times H} & \exp(-\varepsilon \boldsymbol{A}) \end{pmatrix}$

  $\boldsymbol{V} \leftarrow \boldsymbol{V} + \frac{\varepsilon}{2}\nabla_{\boldsymbol{\Psi}_1} \log \pi(\boldsymbol{\Psi}_1)$

  $\boldsymbol{V} \leftarrow \boldsymbol{V} - \frac{1}{2}\boldsymbol{\Phi}_1\left(\boldsymbol{\Psi}_1^\top \boldsymbol{V} + \boldsymbol{V}^\top \boldsymbol{\Psi}_1\right)$

end for

$\mathcal{H}_1 \leftarrow \log \pi(\boldsymbol{\Psi}_1) - \frac{1}{2}\text{vec}(\boldsymbol{V})^\top \text{vec}(\boldsymbol{V})$

$w \sim \mathcal{U}(0, 1)$

if $w < \exp(\mathcal{H}_1 - \mathcal{H}_0)$ then

  $\boldsymbol{\Psi} \leftarrow \boldsymbol{\Psi}_1$

end if

return $\boldsymbol{\Psi}$

Note: $\exp(\boldsymbol{A})$ denotes the matrix exponential operator.

Sampling $\boldsymbol{Y}^{miss}$ The conditional posterior distribution of a missing observation of unit j in time period t is a normal distribution,

$$y_{j,t}(0) \,|\text{rest} \sim \mathcal{N}\left(\gamma_{j,t} + \xi_{j,t},\ \tau^{-1}\right), \quad (j,t) \in \mathcal{I}_1.$$

Bibliography

A , and J. Hainmueller (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505.

——— (2015), "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510.

Abadie, A. (forthcoming), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Apects," *Journal of Economic Literature*.

Abadie, A. and J. Gardeazabal (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132.

Abadir, K. M., W. Distaso, and F. Žikeš (2014), "Design-free Estimation of Variance Matrices," *Journal of Econometrics*, 181, 165–180.

Aikman, D., O. Bush, and A. M. Taylor (2016), "Monetary Versus Macroprudential Policies: Causal Impacts of Interest Rates and Credit Controls in the Era of the UK Radcliffe Report," NBER Working Paper 22380.

Alvarez, I., J. Niemi, and M. Simpson (2014), "Bayesian Inference for a Covariance Matrix," in *26th Annual Conference on Applied Statistics in Agriculture*.

Amjad, M., D. Shah, and D. Shen (2018), "Robust Synthetic Control," *Journal of Machine Learning Research*, 19, 802–852.

Andrews, D. W. (1999), "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67, 543–563.

Andrews, D. W. and B. Lu (2001), "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, 101, 123–164.

Andrieu, C. and J. Thoms (2008), "A Tutorial on Adaptive MCMC," *Statistics and Computing*, 18, 343–373.

Arellano, M. and S. Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and An Application to Employment Equations," *Review of Economic Studies*, 58, 277–297.

Atchadé, Y. F. (2006), "An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift," *Methodology and Computing in Applied Probability*, 8, 235–254.

Atchadé, Y. F. and J. S. R l (2005), "On Adaptive Markov Chain Monte Carlo Algorithms," *Bernoulli*, 11, 815–828.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2018), "Matrix Completion Methods for Causal Panel Data Models," arxiv preprint, arXiv:1710.10251.

Auerbach, A. J. and Y. Gorodnichenko (2013), "Fiscal Multipliers in Recession and Expansion," in A. Alesina and F. Giavazzi eds. *Fiscal Policy after the Financial Crisis*: University of Chicago Press, 63–98.

Barnichon, R. and C. Matthes (2019), "Functional Approximations of Impulse Responses," *Journal of Monetary Economics*, 99, 41–55.

Barnichon, R. and C. Brownlees (2019), "Impulse Response Estimation By Smooth Local Projections," *Review of Economics and Statistics*, 101, 522–230.

Belloni, A. and V. Chernozhukov (2009), "On the Computational Complexity of MCMC-based Estimators in Large Samples," *Annals of Statistics*, 37, 2011–2055.

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Bhattacharya, A. and D. B. Dunson (2011), "Sparse Bayesian Infinite Factor Models," *Biometrika*, 291–306.

Blundell, R. and S. Bond (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115–143.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, and S. L. Scott (2015), "Inferring Causal Impact Using Bayesian Structural Time-series Models," *Annals of Applied Statistics*, 9, 247–274.

Burnham, K. P. and D. R. Anderson (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261–304.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995), "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, 16, 1190–1208.

Byrne, S. and M. Girolami (2013), "Geodesic Monte Carlo on Embedded Manifolds," *Scandinavian Journal of Statistics*, 40, 825–845.

Canay, I. A. (2010), "Simultaneous Selection and Weighting of Moments in GMM Using a Trapezoidal Kernel," *Journal of Econometrics*, 156, 284–303.

Caner, M., X. Han, and Y. Lee (2018), "Adaptive Elastic Net GMM Estimation with Many Invalid Moment Conditions: Simultaneous Model and Moment Selection," *Journal of Business and Economic Statistics*, 36, 24–46.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480.

Chang, M. and F. J. DiTraglia (2018), "A Generalized Focused Information Criterion for GMM," *Journal of Applied Econometrics*, 33, 378–397.

Chen, X., D. T. Jacho-Chávez, and O. Linton (2016), "Averaging of an Increasing Number of Moment Condition Estimators," *Econometric Theory*, 32, 30–70.

Cheng, X. and Z. Liao (2015), "Select the Valid and Relevant Moments: An Information-based LASSO for GMM with Many Moments," *Journal of Econometrics*, 186, 443–464.

Chernozhukov, V. and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

——— (2013), "Quantile Models with Endogeneity," *Annual Review of Economics*, 5, 57–81.

Chernozhukov, V., C. Hansen, and M. Spindler (2015), "Post-selection and Post-regularization Inference in Linear Models with Many Controls and Instruments," *American Economic Review*, 105, 486–90.

Chernozhukov, V. and H. Hong (2003), "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293–346.

Coibion, O., Y. Gorodnichenko, L. Kueng, and J. Silvia (2017), "Innocent Bystanders? Monetary Policy and Inequality," *Journal of Monetary Economics*, 88, 70–89.

De Boor, C. (1978), *A Practical Guide to Splines*, 27: Springer-Verlag New York.

Ding, X., L. He, and L. Carin (2011), "Bayesian Robust Principal Component Analysis," *IEEE Transactions on Image Processing*, 20, 3419–3430.

DiTraglia, F. J. (2016), "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM," *Journal of Econometrics*, 195, 187–208.

Donald, S. G., G. W. Imbens, and W. K. Newey (2009), "Choosing Instrumental Variables in Conditional Moment Restriction Models," *Journal of Econometrics*, 152, 28–36.

Doran, H. E. and P. Schmidt (2006), "GMM Estimators with Improved Finite Sample Properties Using Principal Components of the Weighting Matrix, with an Application to the Dynamic Panel Data Model," *Journal of Econometrics*, 133, 387–409.

Doudchenko, N. and G. W. Imbens (2017), "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," arxiv preprint, arXiv:1610.07748.

Durante, D. (2017), "A Note on the Multiplicative Gamma Process," *Statistics and Probability Letters*, 122, 198–204.

Eilers, P. H. and B. D. Marx (1996), "Fle le Smoothing with B-splines and Penalties," *Statistical Science*, 11, 89–121.

El-Shagi, M. (2019), "A Simple Estimator for Smooth Local Projections," *Applied Economics Letters*, 26, 830–834.

Fan, J. and Y. Liao (2014), "Endogeneity in High Dimensions," *Annals of Statistics*, 42, 872–917.

Fan, J., Y. Liao, and H. Liu (2016), "An Overview of the Estimation of Large Covariance and Precision Matrices," *Econometrics Journal*, 19, C1–C32.

Fazel, M., H. Hindi, and S. P. Boyd (2001), "A Rank Minimization Heuristic with Application to Minimum Order System Approximation," in *Proceedings of the American Control Conference*, 6, 4734–4739.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-based Approaches to the Calculations of Posterior Moments," *Bayesian Statistics*, 4, 641–649.

——— (1993), "Bayesian Treatment of the Independent Student-t Linear Model," *Journal of Applied Econometrics*, 8, S19–S40.

——— (2005), *Contemporary Bayesian Econometrics and Statistics*, 537: John Wiley & Sons.

Guo, W. (2002), "Functional Mixed Effects Models," *Biometrics*, 58, 121–128.

Haario, H., E. Saksman, and J. Tamminen (2001), "An Adaptive Metropolis Algorithm," *Bernoulli*, 7, 223–242.

Hahn, J. and R. Shi (2017), "Synthetic Control and Inference," *Econometrics*, 5, 52.

Hahn, P. R., J. He, and H. Lopes (2018), "Bayesian Factor Model Shrinkage for Linear IV Regression with Many Instruments," *Journal of Business and Economic Statistics*, 36, 278–287.

Hahn, P. R., J. He, and H. F. Lopes (2019), "Efficient Sampling for Gaussian Linear Regression with Arbitrary Priors," *Journal of Computational and Graphical Statistics*, 28, 142–154.

Hall, A. R. (2005), *Generalized Method of Moments*: Oxford University Press.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007), "Information in Generalized Method of Moments Estimation and Entropy-based Moment Selection," *Journal of Econometrics*, 138, 488–512.

Hall, A. R. and F. P. Peixe (2003), "A Consistent Method for the Selection of Relevant Instruments," *Econometric Reviews*, 22, 269–287.

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P., J. Heaton, and A. Yaron (1996), "Finite-sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14, 262–280.

Huang, A. and M. P. Wand (2013), "Simple Marginally Noninformative Prior Distributions for Covariance Matrices," *Bayesian Analysis*, 8, 439–452.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society, Series B*, 60, 271–293.

Imbens, G. W. and D. B. Rubin (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*: Cambridge University Press.

Ishwaran, H. and L. F. James (2001), "Gibbs Sampling Methods for Stick-breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Ishwaran, H., J. S. Rao et al. (2005), "Spike and Sla le Selection: Frequentist and Bayesian Strategies," *Annals of Statistics*, 33, 730–773.

Jeffreys, H. (1961), *Theory of Probability*: Oxford University Press, 3rd edition.

Jordà, Ò. (2005), "Estimation and Inference of Impulse Responses Local Projections," *American Economic Review*, 95, 161–182.

Keshavan, R. H., A. Montanari, and S. Oh (2010), "Matrix Completion from Noisy Entries," *Journal of Machine Learning Research*, 11, 2057–2078.

Kim, J.-Y. (2002), "Limited Information Likelihood and Bayesian Analysis," *Journal of Econometrics*, 107, 175–193.

Kim, S., C. Lee, and S. Gupta (2020), "Bayesian Synthetic Control Methods," *Journal of Marketing Research*, 57, 831–852.

Lam, C. (2016), "Nonparametric Eigenvalue-regularized Precision or Covariance Matrix Estimator," *Annals of Statistics*, 44, 928–953.

——— (2020), "High-dimensional Covariance Matrix Estimation," *Wiley Interdisciplinary Reviews: Computational Statistics*, 12, e1485.

Lang, S. and A. Brezger (2004), "Bayesian P-splines," *Journal of Computational and Graphical Statistics*, 13, 183–212.

Legramanti, S., D. Durante, and D. B. Dunson (2020), "Bayesian Cumulative Shrinkage for Infinite Factorizations," *Biometrika*, 107, 745–752.

Lewbel, A. (2012), "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models," *Journal of Business and Economic Statistics*, 30, 67–80.

Li, C. and W. Jiang (2016), "On Oracle Property and Asymptotic Validity of Bayesian Generalized Method of Moments," *Journal of Multivariate Analysis*, 145, 132–147.

Li, K. T. (forthcoming), "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," *Journal of the American Statistical Association*.

Liao, Z. (2013), "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection," *Econometric Theory*, 29, 857–904.

Mai, T. T. and P. Alquier (2015), "A Bayesian Approach for Noisy Matrix Completion: Optimal Rate under General Sampling Distribution," *Electronic Journal of Statistics*, 9, 823–841.

Makalic, E. and D. F. Schmidt (2016), "A Simple Sampler for the Horseshoe Estimator," *IEEE Signal Processing Letters*, 23, 179–182.

Mazumder, R., T. Hastie, and R. Tibshirani (2010), "Spectral R larization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322.

Miranda-Agrippino, S. and G. Ricco (2017), "The Transmission of Monetary Policy Shocks," Staff Working Paper 657, Bank of England.

Morris, J. S. and R. J. Carroll (2006), "Wavelet-based Functional Mixed Models," *Journal of the Royal Statistical Society, Series B*, 68, 179–199.

Ning, B., S. Ghosal, and J. Thomas (2019), "Bayesian Method for Causal Inference in Spatially-correlated Multivariate Time Series," *Bayesian Analysis*, 14, 1–28.

Okui, R. (2009), "The Optimal Choice of Moments in Dynamic Panel Data Models," *Journal of Econometrics*, 151, 1–16.

Pourahmadi, M. (2011), "Covariance Estimation: The GLM and Regularization Perspectives," *Statistical Science*, 26, 369–387.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research," *Sociological Methodology*, 25, 111–163.

Ramey, V. A. (2016), "Macroeconomic Shocks and Their Propagation," in J. B. Taylor and H. Uhlig eds. *Handbook of Macroeconomics*, 2A: Elsevier, Chap. 2, 71–162.

Ramey, V. A. and S. Zubuairy (2018), "Government Spending Multipliers in Good Times and in Bad: Evidence from U.S. Historical Data," *Journal of Political Economy*, 126, 850–901.

Riera-Crichton, D., C. A. Vegh, and G. Vuletin (2015), "Procyclical and Countercyclical Fiscal Multipliers: Evidence from OECD Countries," *Journal of International Money and Finance*, 52, 15–31.

Robbins, H. and S. Monro (1951), "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, 22, 400–407.

Roberts, G. O. and J. S. Rosenthal (2007), "Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms," *Journal of Applied Probability*, 44, 458–475.

——— (2009), "Examples of Adaptive MCMC," *Journal of Computational and Graphical Statistics*, 18, 349–367.

Romer, C. D. and D. H. Romer (2004), "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 94, 1055–1084.

Rue, H. (2001), "Fast Sampling of Gaussian Markov Random Fields," *Journal of the Royal Statistical Society, Series B*, 63, 325–338.

Rue, H. and L. Held (2005), *Gaussian Markov Random Fields: Theory and Applications*: CRC press.

Salakhutdinov, R. and A. Mnih (2008), "Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo," in *Proceedings of the 25th International Conference on Machine Learning*, 880–887.

Satchachai, P. and P. Schmidt (2008), "GMM with More Moment Conditions Than Observations," *Economics Letters*, 99, 252–255.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002), "Bayesian Measures of Model Complexity and fit," *Journal of the Royal Statistical Society, Series B*, 64, 583–639.

Stock, J. H. and M. W. Watson (2007), "Why Has US Inflation Become Harder to Forecast?" *Journal of Money, Credit and Banking*, 39, 3–33.

Tang, K., Z. Su, J. Zhang, L. Cui, W. Jiang, X. Luo, and X. Sun (2019), "Bayesian Rank Penalization," *Neural Networks*, 116, 246–256.

Tanner, M. A. and W. H. Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Vats, D., J. M. Flegal, and G. L. Jones (2019), "Multivariate Output Analysis for Markov Chain Monte Carlo," *Biometrika*, 106, 321–337.

Vieira, F., R. MacDonald, and A. Damasceno (2012), "The Role of Institutions in Cross-section Income and Panel Data Growth Models: A Deeper Investigation on the Weakness and Proliferation of Instruments," *Journal of Comparative Economics*, 40, 127–140.

Vihola, M. (2011), "On the Stability and Ergodicity of Adaptive Scaling Metropolis Algorithms," *Stochastic Processes and Their Applications*, 121, 2839–2860.

——— (2012), "Robust Adaptive Metropolis Algorithm with Coerced Acceptance Rate," *Statistics and Computing*, 22, 997–1008.

Watanabe, S. (2010), "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory," *Journal of Machine Learning Research*, 11, 3571–3594.

Xu, Y. (2017), "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 25, 57–76.

Yin, G. (2009), "Bayesian Generalized Method of Moments," *Bayesian Analysis*, 4, 191–207.

Yin, G., Y. Ma, F. Liang, and Y. Yuan (2011), "Stochastic Generalized Method of Moments," *Journal of Computational and Graphical Statistics*, 20, 714–727.

Zou, H. and T. Hastie (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.