How Sensitive Are VAR Forecasts to Prior Hyperparameters? An Automated Sensitivity Analysis*

Joshua C.C. Chan
Purdue University
and University of Technology Sydney

Liana Jacobi
Department of Economics,
University of Melbourne

Dan Zhu
Department of Business Statistics and Econometrics,
Monash University

October 2018

Abstract

Vector autoregressions combined with Minnesota-type priors are widely used for macroeconomic forecasting. The fact that strong but sensible priors can substantially improve forecast performance implies VAR forecasts are sensitive to prior hyperparameters. But the nature of this sensitivity is seldom investigated. We develop a general method based on Automatic Differentiation to systematically compute the sensitivities of forecasts—both points and intervals—with respect to any prior hyperparameters. In a forecasting exercise using US data, we find that forecasts are relatively sensitive to the strength of shrinkage for the VAR coefficients, but they are not much affected by the prior mean of the error covariance matrix or the strength of shrinkage for the intercepts.

Keywords: vector autoregression, automatic differentiation, interval forecasts

JEL classifications: C11, C53, E37

*We would like to thank Jackson Kwok for his excellent research assistance.

> The use of subjective prior beliefs must be accompanied by a local sensitivity analysis, and to the extent possible, a global sensitivity analysis.
>
> Poirier (1988, p. 130)

1 Introduction

Since the seminal work of Sims (1980), vector autoregressions (VARs) have become a workhorse model for modeling the dynamic linear interdependencies between multiple time series. In particular, VARs are widely used for macroeconomic forecasting and often serve as benchmark models for comparing the performance of new models and methods. VARs tend to have a lot of parameters, and Bayesian methods that formally incorporate strong but sensible prior information are often found to greatly improve forecast performance. Prominent examples include the Minnesota prior developed in Doan, Litterman, and Sims (1984) and Litterman (1986).

The fact that informative priors can improve forecast performance suggests VAR forecasts are sensitive to prior hyperparameters. However, the nature of this sensitivity is not well understood—it is unclear how VAR forecasts are affected by various choices of hyperparameters. Of course, the importance of sensitivity analysis has long been recognized (Leamer, 1983). It is especially so when subjective priors are used—e.g., Poirier (1988) strongly advises it be done as stated in his second pragmatic principle of model building quoted above. In practice, even when a sensitivity analysis is conducted, often only a narrow aspect is investigated. For example, forecasters might assess a specific aspect of forecast sensitivities by recomputing the forecasts using a different set of hyperparameters. But this approach is ad hoc and requires a substantial amount of computational overhead. Hence, it would be useful to have a systematic approach to assess forecast sensitivities with respect to a variety of hyperparameters as part of the Markov chain Monte Carlo (MCMC) output. This paper takes up this task.

Specifically, we develop a general framework to analyze the sensitivities of predictive outputs—such as means and quantiles of the predictive distribution—with respect to any prior hyperparameters. Our approach builds on earlier work by Jacobi, Joshi, and Zhu (2018), who introduce prior sensitivity analysis for Gibbs output based on Automatic Differentiation (AD). In a nutshell, Automatic Differentiation provides an efficient way to compute derivatives of an algorithm—i.e., local sensitivity of the outputs with respect to the inputs. It is "automatic" in the sense that for an algorithm that transforms the

input into any posterior output, there is an automatic way of deriving its complementary algorithm of computing its sensitivities.

In contrast to AD, the conventional method for assessing local sensitivities of MCMC outputs is the standard numerical finite difference method. Despite its relatively easy implementation, the first clear drawback in applying it to VARs is the computational burden. It requires at least one re-running of the whole MCMC for each parameter in the input vector for assessing first-order sensitivities, hence places a substantial amount of computational overhead. Secondly the method needs to be used with care in choosing the bumping parameter. Since its resulting sensitivities of posterior statistics are biased, there is a variance-bias trade-off in choosing the bumping parameter (Glasserman, 2013). Most importantly, however often ignored, are its subtleties in the cases of non-smooth algorithms. For example, one popular method to sample from a distribution that does not admit the application of the inverse-transform is the acceptance-rejection sampling (as used for example in Gamma random variable generation), which introduces discontinuities to the MCMC. In such cases a naive bumping of input parameters may result in very unstable estimates of the derivatives.

We illustrate our methodology using a VAR forecasting exercise that involves US GDP output growth, interest rate and unemployment rate. We assess the sensitivities of point and interval forecasts of these three variables with respect to a few key hyperparameters in a Minnesota-type prior. Our results show that point and interval forecasts are relatively sensitive to the strength of shrinkage of the VAR coefficients, but they are not much affected by the prior mean of the error covariance matrix nor the strength of shrinkage of the intercepts. In particular in the context of shorter samples, forecasts exhibit a considerable sensitivity with respect to the prior shrinkage parameter.

The rest of this paper is organized as follows. Section 2 first outlines a standard VAR and discusses the priors and the estimation. It is followed by a brief description of how point and interval forecasts from the VAR can be computed. We then introduce in Section 3 a general framework to analyze the sensitivities of the point and interval forecasts with respect to the prior hyperparameters. Section 4 considers a forecasting exercise that involves GDP output growth, interest rate and unemployment rate. Lastly, Section 5 concludes and briefly discusses some future research directions.

2 Vector Autoregressions

A vector autoregression (VAR) is a multiple-equation linear regression that aims to capture the linear interdependencies between variables over time. More specifically, let $\mathbf{y}_t$ denote a vector of observations of n variables at time t with $t = 1, \ldots, T$. Then, a p-order VAR, denoted as VAR(p), is given by:

$$\mathbf{y}_t = \mathbf{b} + \mathbf{B}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{B}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{1}$$

where $\mathbf{b}$ is an $n \times 1$ vector of intercepts, $\mathbf{B}_1, \ldots, \mathbf{B}_p$ are $n \times n$ matrices of VAR coefficients and $\boldsymbol{\Sigma}$ is a covariance matrix.

Even for small systems, VARs tend to contain a large number of parameters. For example, a VAR of $n = 4$ variables with $p = 4$ lags has $pn^2 + n = 68$ coefficients, as well as $n(n+1)/2 = 10$ free parameters in the covariance matrix. Given the typical number of quarterly observations for macroeconomic variables (e.g., less than 300), it is often hard to precisely estimate these parameters. The estimation errors in parameters in turn make forecasts based on VARs less accurate. This has motivated the development of shrinkage priors—i.e., informative priors designed to avoid over-fitting the data and to improve forecast accuracy. Prominent examples include the Minnesota prior and various variants; see, for example, Doan, Litterman, and Sims (1984), Litterman (1986) and Kadiyala and Karlsson (1997).

Given that fairly informative priors are typically used in the context of VAR forecasting, it is natural to assess how these point and interval forecasts change with respect to the strength of shrinkage (e.g., prior covariance). Below we introduce and apply an efficient approach to undertake a comprehensive prior sensitivity analysis for VAR point and interval forecasts with respect to key prior parameters.

2.1 Prior and Estimation

Before outlining the estimation method, it is convenient to rewrite the VAR in (1) in the form of a seemingly unrelated regression:

$$\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2}$$

where $\mathbf{X}_t = \mathbf{I}_n \otimes (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})$ and $\boldsymbol{\beta} = \text{vec}([\mathbf{b}, \mathbf{B}_1, \ldots, \mathbf{B}_p]')$ is the vector of intercepts and VAR coefficients stacked by rows. Here $\boldsymbol{\beta} \in \mathbb{R}^{k_\beta}$ with $k_\beta = n(np+1)$. The parameters can therefore be partitioned into two blocks: $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

Next, consider the following independent priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}}), \qquad \boldsymbol{\Sigma} \sim \mathcal{IW}(\nu_0, \mathbf{S}_0), \tag{3}$$

where $\mathcal{N}(\cdot, \cdot)$ and $\mathcal{IW}(\cdot, \cdot)$ denote respectively the normal and inverse-Wishart distributions. Here $\boldsymbol{\beta}_0$ and $\mathbf{V}_{\boldsymbol{\beta}}$ are respectively the mean vector and covariance matrix of the normal prior on $\boldsymbol{\beta}$, whereas ν_0 and $\mathbf{S}_0$ are respectively the degrees of freedom and scale matrix of the inverse-Wishart prior on $\boldsymbol{\Sigma}$.

In our empirical application that involves stationary macroeconomic variables, we consider a Minnesota-type prior that shrinks the VAR coefficients to zero. Specifically, we set $\boldsymbol{\beta}_0 = \mathbf{0}$, and the covariance matrix $\mathbf{V}_{\boldsymbol{\beta}}$ is assumed to be diagonal with diagonal elements $v_{\boldsymbol{\beta},ii} = \kappa_1/(l^2 \widehat{s}_r)$ for a coefficient associated to lag l of variable r and $v_{\boldsymbol{\beta},ii} = \kappa_2$ for an intercept, where $\widehat{s}_r$ is the sample variance of an AR(4) model for the variable r. Further we set $\nu_0 = n + 3$, $\mathbf{S}_0 = \kappa_3 \mathbf{I}_n$, $\kappa_1 = 0.2^2$, $\kappa_2 = 10^2$ and $\kappa_3 = 1$. Intuitively, the coefficient associated to a lag l variable is shrunk more heavily to zero as the lag length increases, but intercepts are not shrunk to zero. These hyperparameters are standard in the literature. The key hyperparameter is κ_1 that controls the overall strength of shrinkage. For a more detailed discussion of this type of shrinkage priors, see, e.g., Koop and Korobilis (2010) or Karlsson (2013).

Given the priors in (3), the VAR can be estimated using a 2-block Gibbs sampler. To outline the Gibbs sampler, define

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}'_1 \\ \mathbf{y}'_2 \\ \vdots \\ \mathbf{y}'_T \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & \mathbf{y}'_0 & \cdots & \mathbf{y}'_{1-p} \\ 1 & \mathbf{y}'_1 & \cdots & \mathbf{y}'_{2-p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{y}'_{T-1} & \vdots & \mathbf{y}'_{T-p} \end{pmatrix}. \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}'_1 \\ \boldsymbol{\varepsilon}'_2 \\ \vdots \\ \boldsymbol{\varepsilon}'_T \end{pmatrix},$$

Then, we can rewrite (2) as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

In particular, this gives us $\mathbf{X}'\mathbf{X} \in \mathbb{R}^{k_\beta \times k_\beta}$ and $\mathbf{X}'\mathbf{Y} \in \mathbb{R}^{k_\beta \times n}$.

Let B denote the burn-in period and let G represent the number of posterior draws

required. Then, we initialize the Gibbs sampler via $\mathbf{\Sigma}^0 \in \mathbb{R}^{n\times n}$, and update for $g = 1, 2, \ldots, B+G$ via

1. Generate $(\boldsymbol{\beta}^g \,|\, \mathbf{Y}, \mathbf{\Sigma}^{g-1}) \sim \mathcal{N}(\mathbf{b}^g, \mathbf{B}^g)$, where

$$\mathbf{B}^g = (\mathbf{V}_{\boldsymbol{\beta}}^{-1} + (\mathbf{\Sigma}^{g-1})^{-1} \otimes \mathbf{X}'\mathbf{X})^{-1}, \quad \mathbf{b}^g = \mathbf{B}^g \left(\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0 + \mathrm{vec}(\mathbf{X}'\mathbf{Y}(\mathbf{\Sigma}^{g-1})^{-1})\right).$$

2. Generate $(\mathbf{\Sigma}^g \,|\, \mathbf{Y}, \boldsymbol{\beta}^g) \sim \mathcal{IW}(\nu_1, \mathbf{S}^g)$, where

$$\nu_1 = \nu_0 + T, \quad \mathbf{S}^g = \mathbf{S}_0 + \mathbf{Y}'\mathbf{Y} - (\boldsymbol{\beta}^g)'\mathbf{X}'\mathbf{Y} - ((\boldsymbol{\beta}^g)'\mathbf{X}'\mathbf{Y})' + (\boldsymbol{\beta}^g)'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^g.$$

The derivations of the two conditional distributions $(\boldsymbol{\beta} \,|\, \mathbf{Y}, \mathbf{\Sigma})$ and $(\mathbf{\Sigma} \,|\, \mathbf{Y}, \boldsymbol{\beta})$ can be found in standard Bayesian macroeconometric textbooks, such as Chapter 8 in Chan (2017) and Chapter 2 in Koop and Korobilis (2010).

2.2 Point and Interval Forecasts

In this section we describe how one can obtain point and interval forecasts from the VAR given the posterior draws of $(\boldsymbol{\beta}, \mathbf{\Sigma})$. More specifically, given data up to time t, denoted as $\mathbf{y}_{1:t}$, we use the predictive mean $\mathbb{E}(\mathbf{y}_{t+h} \,|\, \mathbf{y}_{1:t})$—where the expectation is taken with respect to the predictive density $p(\mathbf{y}_{t+h} \,|\, \mathbf{y}_{1:t})$—as the point forecast of $\mathbf{y}_{t+h}$ for forecast horizon $h > 0$. To construct interval forecasts, we use appropriate quantiles of the predictive density. For example, the 0.05 and 0.95 quantiles of the predictive density define an interval forecast with coverage probability 0.9.

Even though neither the predictive mean nor any predictive quantiles are available analytically, they can be easily estimated using simulation. Note that the predictive distribution at time $t+h$ can be expressed as

$$p(\mathbf{y}_{t+h} \,|\, \mathbf{y}_{1:t}) = \int p(\mathbf{y}_{t+h} \,|\, \mathbf{y}_{1:t}, \boldsymbol{\beta}, \mathbf{\Sigma}) p(\boldsymbol{\beta}, \mathbf{\Sigma} \,|\, \mathbf{y}_{1:t}) \mathrm{d}(\boldsymbol{\beta}, \mathbf{\Sigma}),$$

where $p(\mathbf{y}_{t+h} \,|\, \mathbf{y}_{1:t}, \boldsymbol{\beta}, \mathbf{\Sigma})$ is a Gaussian density implied by the Gaussian VAR described in equation (1). The above integral is taken with respect to the posterior distribution of the parameters. Hence, we can obtain draws from this predictive distribution by generating draws from $p(\mathbf{y}_{t+h} \,|\, \mathbf{y}_{1:t}, \boldsymbol{\beta}, \mathbf{\Sigma})$ at each iteration of the Gibbs algorithm introduced above, via the following additional step:

3. Generate $\mathbf{y}_{t+h}^g$, for $h = 1, \ldots, H$ from $(\mathbf{y}_{t+h}^g \,|\, \mathbf{y}_{1:t}, \boldsymbol{\beta}^g, \boldsymbol{\Sigma}^g) \sim \mathcal{N}(\mathbf{X}_{t+h}\boldsymbol{\beta}^g, \boldsymbol{\Sigma}^g)$.

We can then use these posterior draws $\{\mathbf{y}_{t+h}^g\}_{g=B+1}^{B+G}$ after burn-in to obtain simulation-consistent estimates of the mean and quantiles of the predictive densities.

3 Automatic Differentiation for VAR

In this section we introduce a general framework to analyze the sensitivity of the predictive outputs (e.g., predictive mean and quantiles) with respect to a set of prior hyperparameters, $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. This builds on recent work by Jacobi, Joshi, and Zhu (2018) that develops a general framework to obtain a complete set of input sensitivities for MCMC output based on Automatic Differentiation (AD), including methods to compute sensitivities for posterior statistics with respect to the full set of prior assumptions.

Automatic Differentiation is an efficient means of computing derivatives, i.e., the local sensitivity of the outputs with respect to the inputs. In a nutshell, if we have a function $g : \mathbb{R} \to \mathbb{R}$, AD translates g into its first order derivative automatically. For many applications in Bayesian MCMC, we typically have a more general mapping of the form

$$\boldsymbol{\theta}_0 \in \mathbb{R}^{m_1}, \boldsymbol{\eta}_0 \in \mathbb{R}^{n_1} \to \mathbf{G}(\boldsymbol{\theta}_0, \eta_0) \in \mathbb{R}^{m_2 \times n_2},$$

where $\boldsymbol{\eta}_0$ refers to the set of inputs in combination with $\boldsymbol{\theta}_0$ that are mapped via some MCMC algorithm $\mathbf{G}$ into posterior quantities, albeit the analyst is not interested in its relative sensitivities. For instance, the prior mean of $\boldsymbol{\beta}$ is set at zero in the Minnesota prior as an input for the MCMC algorithm, but it is not considered in our sensitivity analysis in Section 4. In general, the complementary AD computes the derivatives of the posterior output $\mathbf{G}$ with respects to the complete set of inputs. It is up to the analyst to choose which subset of inputs are included in $\boldsymbol{\theta}_0$.

The application of AD is the translation of $\mathbf{G}$ into a set of first order derivatives

$$\frac{\partial \mathbf{G}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_0} = \frac{\partial \operatorname{vec}(\mathbf{G}'(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0))}{\partial \boldsymbol{\theta}_0} = \begin{bmatrix} \frac{\partial G_{1,1}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,1}} & \frac{\partial G_{1,1}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,2}} & \cdots & \frac{\partial G_{1,1}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,m_1}} \\ \frac{\partial G_{1,2}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,1}} & \frac{\partial G_{1,2}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,2}} & \cdots & \frac{\partial G_{1,2}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,m_1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial G_{m_2,n_2}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,1}} & \frac{\partial G_{m_2,n_2}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,2}} & \cdots & \frac{\partial G_{m_2,n_2}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_{0,m_1}} \end{bmatrix},$$

where $\boldsymbol{\theta}_{0,j}$ is the jth element of $\boldsymbol{\theta}_0$, this Jacobian matrix of dimension $m_2 n_2 \times p$.

While AD methods have been widely used to undertake input sensitivity analysis in the context of less computationally intensive classical simulation methods, particularly in financial mathematics (see Giles and Glasserman, 2006; Joshi and Yang, 2011), it has only been recently introduced in the context of MCMC simulation by Jacobi, Joshi, and Zhu (2018). In particular, the paper develops an AD approach and AD based methods for a comprehensive prior robustness and convergence analysis of MCMC output and shows how the Forward mode of differentiation can be applied to compute Jacobian matrices of first order derivatives for MCMC based statistics in various standard models.

AD is "automatic" in the sense that for an algorithm that transforms the input vector $\boldsymbol{\theta}_0$ into the posterior output vector, there is an automatic way of deriving its complementary algorithm of computing its sensitivities without manually deriving the symbolic formula of the derivatives. It is derived by first decomposing the original algorithm $\mathbf{G}$ into simpler operations $\mathbf{G}_1, \ldots, \mathbf{G}_k$:

$$\mathbf{G} = \mathbf{G}_k \circ \mathbf{G}_{k-1} \circ \cdots \circ \mathbf{G}_1.$$

Then, the derivative of $\mathbf{G}$ can be obtained via the chain-rule (that is implemented automatically in the compute program)

$$\frac{\partial \mathbf{G}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_0} = J\mathbf{G}_k \times J\mathbf{G}_{k-1} \times \cdots \times J\mathbf{G}_1,$$

where $J\mathbf{G}_i, i = 1, \ldots, k$ are the intermediate Jacobians of the simpler operations. While the end result $\frac{\partial \mathbf{G}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)}{\partial \boldsymbol{\theta}_0}$ is a dense matrix, the $J\mathbf{G}_i$'s are typically very sparse matrices because each operation $\mathbf{G}_i$ typically only updates one or two variables.

In the context of MCMC, sensitivities can often be derived using information about model dynamics in simulation—i.e., the dependence of the posterior distribution on the set of prior assumptions. AD accomplishes this by differentiating the evolution of the underlying state variables along each path. In comparison to the widely used numerical finite difference methods, AD requires additional model analysis and programming, but this additional effort is often justified by the improvement in the quality and comprehensiveness of calculated local sensitivities. Due to the computational burden of numerical finite difference methods, typically only a very limited prior robustness analysis is implemented.

Symbolic differentiation is another method widely implemented in computer software including Matlab and Mathematica. In general, if the original function of mapping is

simple enough, a symbolic differentiation package can be applied which outputs symbolic descriptions of the derivatives. Many of our derivation in the following section uses this symbolic derivative idea for illustration purposes, but the exact implementation is done via AD which simplifies long symbolic derivative expressions by sharing some intermediate results between the main original algorithm, $\mathbf{G}_i$ and the complementary derivative algorithm, $J\mathbf{G}_i$.

3.1 Sensitivities for Prior Shrinkage

For the VAR model introduced in Section 2, we are interested in the sensitivities of the forecasts with respect to the prior hyperparameters $(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}}, \nu_0, \mathbf{S}_0)$. Therefore, let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \text{vec}(\mathbf{V}_{\boldsymbol{\beta}}), \nu_0, \text{vec}(\mathbf{S}_0))$ denote the vector of all input parameters of interest. AD offers an efficient numerical approach to compute the complete set of first order derivatives of a wide range of MCMC output with respect to $\boldsymbol{\theta}_0$. Of particular interest here are the prior robustness of the mean and interval predictions. In the context of the Minnesota prior the prior variances are specified in terms of the scale parameters κ_1, κ_2 and κ_3 which respectively "scale" the prior variances for the lag effects, the intercept and the variance parameters. For our empirical analysis in Section 4 we set $\boldsymbol{\theta}_0 = (\kappa_1, \kappa_2, \kappa_3)$ as these present the key parameters for our forecast sensitivity analysis of prior shrinkage.

Our focus is on the sensitivities of point and interval forecasts. As discussed in Section 2, the MCMC algorithm first generate the model parameters and then the forecast values. Since the predictions depend on the model parameters, we first discuss the AD approach to obtain the first order derivatives for the model parameters with respect to $\boldsymbol{\theta}_0$. These are obtained by differentiating through the 2-step Gibbs algorithm that generates the parameter draws. Next we show how to obtain the first order derivatives for the point and interval forecasts.

We have also provided Matlab and R code to implement the AD based prior sensitivity analysis described below.[1] For some more technical points on the implementation, interested readers are referred to the Technical Appendix at the end of this paper as well as to the discussion in Jacobi, Joshi, and Zhu (2018).

[1]The code can be downloaded at `http://joshuachan.org/code.html`.

3.2 Sensitivities of Model Parameters

In the first step, $\boldsymbol{\beta}$ is updated from a multivariate Normal as $\boldsymbol{\beta}^g = \mathbf{b}^g + \text{chol}(\mathbf{B}^g)\mathbf{Z}$, for $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{k_\beta})$, where $k_\beta = n(np+1)$ is the dimension of $\boldsymbol{\beta}$. Therefore, we need to apply the chain rule and product rule and employ matrix and vector calculus to differentiate the following expression:

$$\frac{\partial \boldsymbol{\beta}^g}{\partial \boldsymbol{\theta}_0} = \frac{\partial \mathbf{b}^g}{\partial \boldsymbol{\theta}_0} + \left(\mathbf{Z}' \otimes \mathbf{I}_{k_\beta}\right) \frac{\partial \,\text{chol}(\mathbf{B})}{\partial \mathbf{B}}\bigg|_{\mathbf{B}=\mathbf{B}^g} \frac{\partial \mathbf{B}^g}{\partial \boldsymbol{\theta}_0},$$

which requires the derivatives of the Cholesky decomposition as well as the derivatives of the mean $\frac{\partial \mathbf{b}^g}{\partial \boldsymbol{\theta}_0}$ and covariance matrix $\frac{\partial \mathbf{B}^g}{\partial \boldsymbol{\theta}_0}$. While all these derivatives can be obtained via standard AD schemes, depending on the software implementation, some operations here involve large matrices, which require particular care for an efficient implementation. Nevertheless, many of these terms show up repeatedly, and this allows us to dramatically reduce the computation effort to compute the final derivatives, which contributes to the high efficiency of AD methods.

In the second step of the Gibbs sampler, we obtain a draw $\boldsymbol{\Sigma}^g \sim \mathcal{IW}(\nu, \mathbf{S}^g)$, where $\nu = \nu_0 + T$. This can be done by first sampling $\mathbf{S} \sim \mathcal{W}(\nu, (\mathbf{S}^g)^{-1})$, and returning $\boldsymbol{\Sigma}^g = \mathbf{S}^{-1}$. Hence, to compute the derivatives for $(\boldsymbol{\Sigma}^g \,|\, \mathbf{Y}, \boldsymbol{\beta}^g) \sim \mathcal{IW}(\nu, \mathbf{S}^g)$, $\frac{\partial \Sigma^g}{\partial \boldsymbol{\theta}_0}$, we consider the Bartlett decomposition of the Wishart distribution $\mathcal{W}(\nu, (\mathbf{S}^g)^{-1})$. Let $\mathbf{L}$ denote the Cholesky factor of $(\mathbf{S}^g)^{-1}$ so that $(\mathbf{S}^g)^{-1} = \mathbf{L}\mathbf{L}'$ and let $\mathbf{A} = (a_{ij})$ denote a lower triangular matrix such that the diagonal elements are distributed as χ^2 random variables $a_{ii}^2 \sim \chi^2_{\nu-i+1}$ and the lower triangular elements are standard Gaussian random variables $a_{ij} \sim \mathcal{N}(0,1)$ for $i > j$. Then, $\mathbf{S} = \mathbf{L}\mathbf{A}\mathbf{A}'\mathbf{L}'$ has the Wishart distribution $\mathcal{W}(\nu, (\mathbf{S}^g)^{-1})$, and we return $\boldsymbol{\Sigma}^g = (\mathbf{L}\mathbf{A}\mathbf{A}'\mathbf{L}')^{-1}$. Note that we can avoid the explicit evaluation of $(\mathbf{S}^g)^{-1}$ which simplifies the computation to obtain $\frac{\partial \Sigma^g}{\partial \boldsymbol{\theta}_0}$.

3.3 Sensitivities for Point Forecasts

In this section we derive explicit expressions of the first-order derivatives of the point forecasts with respect to the input vector $\boldsymbol{\theta}_0$. To make the discussion concrete, we set $p = 2$. Recall that at each iteration, we draw $\mathbf{Z}_h \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ for $h = 1, 2, \ldots, H$ and

construct the point forecasts $\mathbf{y}^g_{T+1}, \ldots, \mathbf{y}^g_{T+h}$ as follows:

$$
\begin{aligned}
\mathbf{y}^g_{T+1} &= \mathbf{b}^g + \mathbf{B}^g_1 \mathbf{y}_T + \mathbf{B}^g_2 \mathbf{y}_{T-1} + \operatorname{chol}(\boldsymbol{\Sigma}^g)\mathbf{Z}_1, \\
\mathbf{y}^g_{T+2} &= \mathbf{b}^g + \mathbf{B}^g_1 \mathbf{y}^g_{T+1} + \mathbf{B}^g_2 \mathbf{y}_T + \operatorname{chol}(\boldsymbol{\Sigma}^g)\mathbf{Z}_2, \\
\mathbf{y}^g_{T+h} &= \mathbf{b}^g + \mathbf{B}^g_1 \mathbf{y}^g_{T+h-1} + \mathbf{B}^g_2 \mathbf{y}^g_{T+h-2} + \operatorname{chol}(\boldsymbol{\Sigma}^g)\mathbf{Z}_h, \quad h = 3, \ldots, H.
\end{aligned}
$$

Their corresponding Jacobians are therefore

$$
\begin{aligned}
\frac{\partial \mathbf{y}^g_{T+1}}{\partial \boldsymbol{\theta}_0} =& \frac{\partial \mathbf{b}^g}{\partial \boldsymbol{\theta}_0} + (\mathbf{y}'_T \otimes \mathbf{I}_n)\frac{\partial \mathbf{B}^g_1}{\partial \boldsymbol{\theta}_0} + (\mathbf{y}'_{T-1} \otimes \mathbf{I}_n)\frac{\partial \mathbf{B}^g_2}{\partial \boldsymbol{\theta}_0} + (\mathbf{Z}'_1 \otimes \mathbf{I}_n)\left.\frac{\partial \operatorname{chol}(\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^g} \frac{\partial \boldsymbol{\Sigma}^g}{\partial \boldsymbol{\theta}_0}, \\
\frac{\partial \mathbf{y}^g_{T+2}}{\partial \boldsymbol{\theta}_0} =& \frac{\partial \mathbf{b}^g}{\partial \boldsymbol{\theta}_0} + ((\mathbf{y}^g_{T+1})' \otimes \mathbf{I}_n)\frac{\partial \mathbf{B}^g_1}{\partial \boldsymbol{\theta}_0} + \mathbf{B}^g_1 \frac{\partial \mathbf{y}^g_{T+1}}{\partial \boldsymbol{\theta}_0} + (\mathbf{y}'_T \otimes \mathbf{I}_n)\frac{\partial \mathbf{B}^g_2}{\partial \boldsymbol{\theta}_0} \\
&+ \left(\mathbf{Z}^T_2 \otimes \mathbf{I}_n\right)\left.\frac{\partial \operatorname{chol}(\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^g} \frac{\partial \boldsymbol{\Sigma}^g}{\partial \boldsymbol{\theta}_0}, \\
\frac{\partial \mathbf{y}^g_{T+h}}{\partial \boldsymbol{\theta}_0} =& \frac{\partial \mathbf{b}^g}{\partial \boldsymbol{\theta}_0} + ((\mathbf{y}^g_{T+h-1})' \otimes \mathbf{I}_n)\frac{\partial \mathbf{B}^g_1}{\partial \boldsymbol{\theta}_0} + \mathbf{B}^g_1 \frac{\partial \mathbf{y}^g_{T+h-1}}{\partial \boldsymbol{\theta}_0} + ((\mathbf{y}^g_{T+h-2})' \otimes \mathbf{I}_n)\frac{\partial \mathbf{B}^g_2}{\partial \boldsymbol{\theta}_0} \\
&+ \mathbf{B}^g_2 \frac{\partial \mathbf{y}^g_{T+h-2}}{\partial \boldsymbol{\theta}_0} + (\mathbf{Z}'_h \otimes \mathbf{I}_n)\left.\frac{\partial \operatorname{chol}(\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^g} \frac{\partial \boldsymbol{\Sigma}^g}{\partial \boldsymbol{\theta}_0}.
\end{aligned}
$$

While the sensitivities of the one-step-ahead point forecast only depend on the draws of the model parameters and thus the sensitivities of the model parameters, sensitivities of further forecast horizons depend on both the sensitivities of the model parameters and the previous periods' forecasts. There is also a clear autoregressive structure in the first-order sensitivities, i.e., $\mathbf{B}^g_1 \frac{\partial \mathbf{y}^g_{T+h-1}}{\partial \boldsymbol{\theta}_0}$ and $\mathbf{B}^g_1 \frac{\partial \mathbf{y}^g_{T+h-2}}{\partial \boldsymbol{\theta}_0}$. The sample mean of these sensitivity draws gives an estimate of the sensitivity for the point forecasts, when the model parameters are drawn from the posterior distribution.

3.4 Sensitivities for Predictive Quantiles

Interval forecasts are appropriate quantiles of the predictive distributions. Quantile sensitivities are more difficult to compute than those of the point forecasts. This is because algorithmically quantiles are estimated using the corresponding sample order statistics by sorting the sample, and such operations are not continuously differentiable. Hence, the derivative operator cannot be directly applied in a finite sample. Moreover, the predictive distribution is not available analytically and we therefore cannot apply distributional derivatives directly to obtain quantile sensitivities. However, there is some progress in estimating quantiles sensitivities in the classical simulation literature, see e.g., the batched

infinitesimal estimator in Hong (2009) and the Kernel-smooth estimator in Liu and Hong (2009).

Below we present a consistent method for estimating quantile sensitivities in the context of MCMC, which shares some similarities with Fu, Hong, and Hu (2009). Suppose our forecast random variable Y is absolutely continuous with the distribution $F_Y(\cdot;\boldsymbol{\theta}_0)$. For a given $\alpha \in (0,1)$, the α-quantile, denoted as Y^*, is defined implicitly by

$$F_Y(Y^*;\boldsymbol{\theta}_0) = \alpha.$$

By the implicit function theorem, we have

$$\frac{\partial Y^*}{\partial \boldsymbol{\theta}_0} = -\frac{\frac{\partial F_Y(y;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}}{f_Y(y;\boldsymbol{\theta}_0)}\bigg|_{y=Y^*},$$

where $f_Y(\cdot;\boldsymbol{\theta}_0)$ is the associated density function, which is unfortunately unknown. However, suppose there exists a latent random vector $\mathbf{Z} \sim f_{\mathbf{Z}}(\cdot;\boldsymbol{\theta}_0)$ such that

$$F_Y(y;\boldsymbol{\theta}_0) = \mathbb{E}\left[G\left(y;\mathbf{Z}(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0\right)\right]$$

for a function $G\left(y;\mathbf{Z}(\boldsymbol{\theta}_0),\boldsymbol{\theta}_0\right)$ that is absolutely continuous in y, and differentiable almost surely in $\boldsymbol{\theta}_0$.[2] Then, we can approximate $\frac{\partial Y^*}{\partial \boldsymbol{\theta}_0}$ via

$$-\frac{\sum_{i=1}^{N} \frac{\partial G(y;\mathbf{Z}(\boldsymbol{\theta}_0)^i,\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}}{\sum_{i=1}^{N} g(y;\mathbf{Z}(\boldsymbol{\theta}_0)^i,\boldsymbol{\theta}_0)}\bigg|_{y=Y^*} \tag{4}$$

where $\mathbf{Z}(\boldsymbol{\theta}_0)^i \sim f_{\mathbf{Z}}(\cdot;\boldsymbol{\theta}_0), i = 1,\ldots,N$ and g is the derivative of G with respect to y.

In our context for estimating quantile sensitivities of Y_{T+h}, a quantile Y^* is approximated via the associated sample order statistic. The natural candidate for the latent vector $\mathbf{Z}$ is $(Y_{T+h-1},\boldsymbol{\beta},\boldsymbol{\Sigma})$, and the corresponding G is simply the density of Y_{T+h} conditional of $\mathbf{Z}$, which is Gaussian. Hence, both the density function as well as the derivatives of the distribution function are easy to evaluate.

[2]Note that we make the dependence of $\mathbf{Z}$ on $\boldsymbol{\theta}_0$ explicit.

3.5 Efficient Implementation

We can improve the speed and reduce the amount of memory required in computing the above sensitivities by using a more compact representation of the VAR. More specifically, we rewrite the VAR in (1) as:

$$\mathbf{y}_t' = \mathbf{z}_t'\mathbf{B} + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{z}_t' = (1, \mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}')$ and $\mathbf{B} = (\mathbf{b}, \mathbf{B}_1, \ldots, \mathbf{B}_p)'$. Then, stacking the equations over $t = 1, \ldots, T$, we have

$$\mathbf{y} = \mathbf{Z}\mathbf{B} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is a $T \times n$ matrix of observations, and $\mathbf{Z}$ and $\boldsymbol{\varepsilon}$ are defined similarly. Since each row of $\boldsymbol{\varepsilon}$ is conditionally independent and normally distributed with mean zero and covariance $\boldsymbol{\Sigma}$, it follows the matric-variate normal distribution $\boldsymbol{\varepsilon} \sim \mathcal{MN}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_T)$ (see, e.g., Bauwens et al., 1999, p. 301-303). For a more detailed discussion on the estimation using this compact representation of the VAR, see the review in Karlsson (2013) and Woźniak (2016).

4 Empirical Application

In our empirical application we consider a 3-variable VAR that involves US quarterly data on the unemployment rate, interest rate and real GDP from 1954:Q3 to 2017:Q4. These three variables are commonly used in forecasting (e.g., Banbura, Giannone, and Reichlin, 2010; Koop, 2013; Chan, 2018). The real GDP series is transformed to annualized growth rates, whereas the unemployment rate and interest rate are not transformed. All data are sourced from the Federal Reserve Bank of St. Louis economic database. The data are plotted in Figure 1.

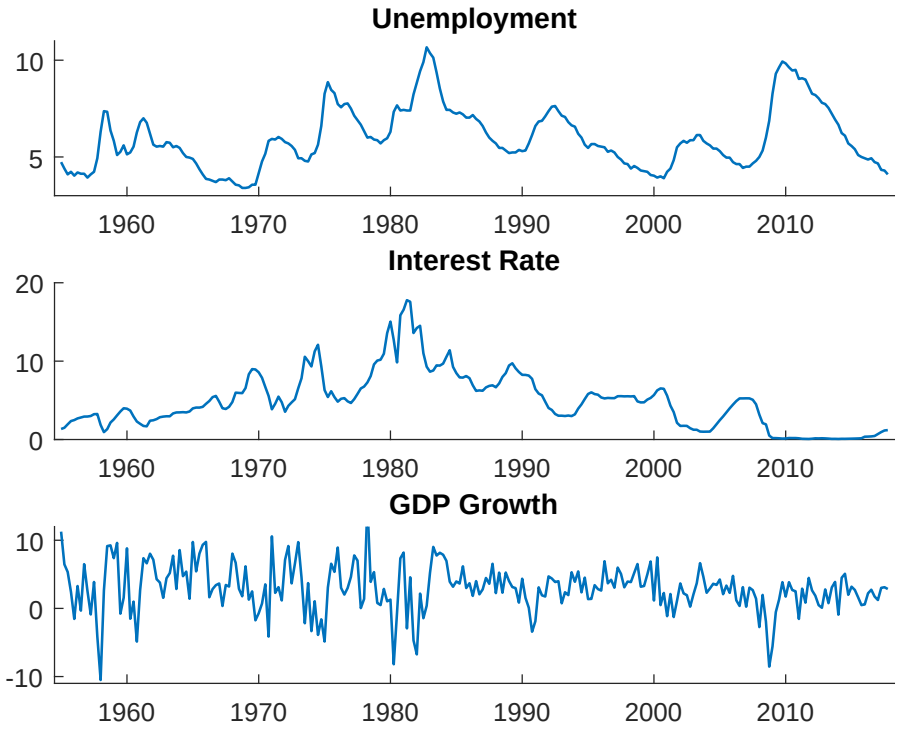

Figure 1: Plots of the US unemployment rate, interest rate and real GDP growth.

Our main objects of interest are the point and interval forecasts for these three variables. In particular, we follow the common practice to use the 16- and 84-percentiles to define the 68% interval forecasts. These forecasts are presented in Figure 2. The VAR forecasts that real GDP growth will remain relatively flat at around 3% over the next 20 quarters whereas interest rate will gradually increase from about 1.4% in 2018:Q1 to about 3.3% at the end of the forecast period. In contrast, the VAR predicts that unemployment rate will halt its decline since the Great Recession and will gradually increase 5.5% toward the end of the forecast period. Our results also indicate that there is a lot of uncertainty around these point forecasts. In particular, the 68% interval forecasts of the three variables are all fairly wide, even for relatively short forecast horizons.

These forecasts are somewhat in line with the projections provided by the US Congressional Budget Office. For example, in their report *The Budget and Economic Outlook: 2018 to 2028* published in April 2018 (CBO, 2018), the Federal funds rate is forecasted to rise from about 1.9% in 2018 to 3.8% in 2022, whereas real GDP growth will drop from about 3% in 2018 to only about 1.5% in 2022.

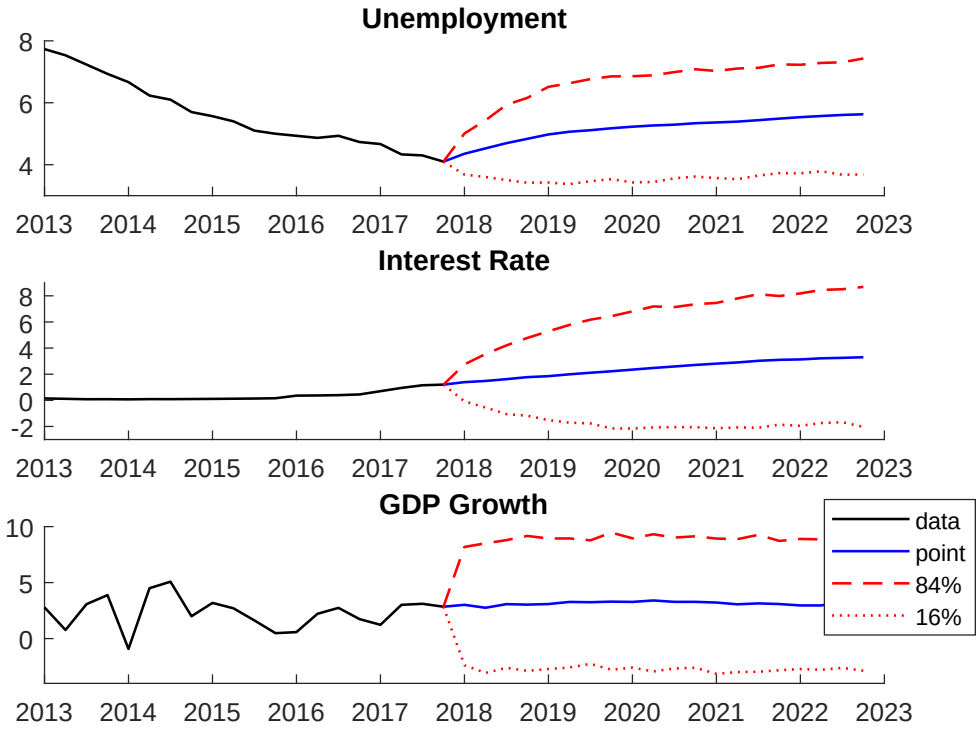

Figure 2: Point and interval forecasts for unemployment rate, interest rate and real GDP growth.

The focus of this paper is to assess how sensitive these point and interval forecasts are with respect to three key hyperparameters: κ_1, κ_2 and κ_3. Recall that κ_1 and κ_2 control, respectively, the strength of shrinkage of the VAR coefficients and the intercepts, whereas κ_3 controls the prior mean of $\boldsymbol{\Sigma}$.

We use AD to compute the derivatives of the forecasts with respect to κ_1, κ_2 and κ_3, and the results are presented in Figure 3. More specifically, the first row shows the derivatives of the three point forecasts, namely, real GDP growth (GDP), interest rate (i) and unemployment rate (u) with respect to κ_1, κ_2 and κ_3. The second and third rows show respectively the corresponding derivatives of the 84- and 16-percentiles.

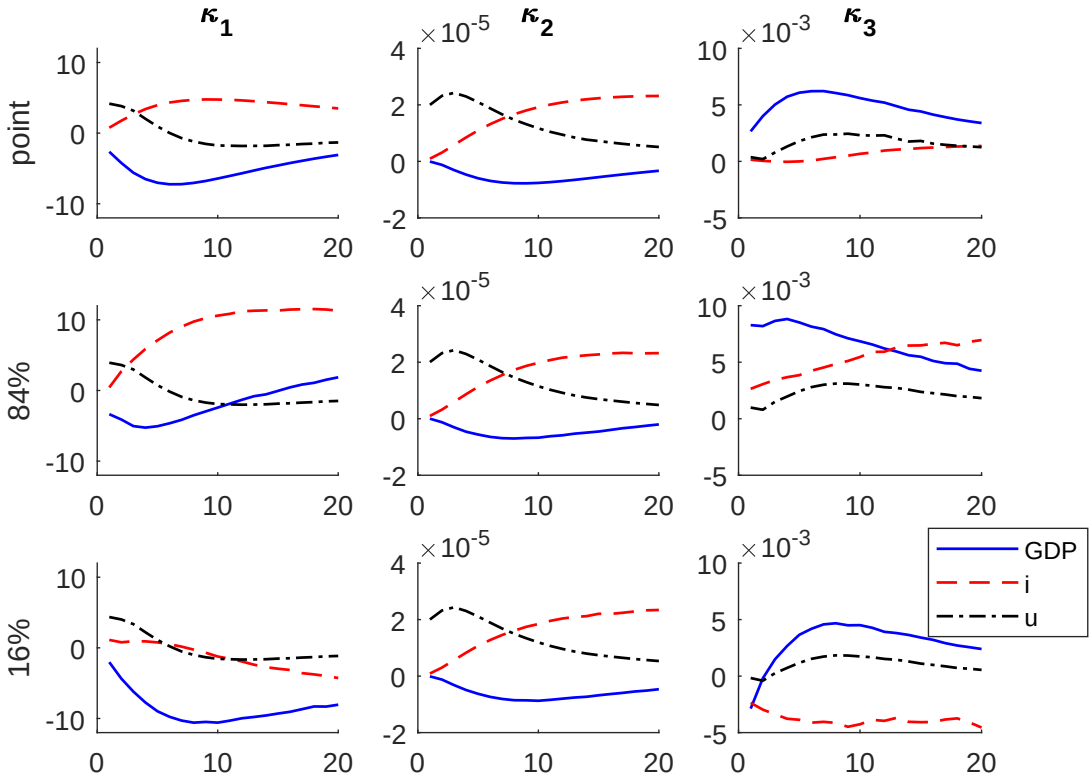

Figure 3: Derivatives of the point forecasts (top row), 84-percentiles (middle row) and 16-percentiles (bottom row) of the variables real GDP growth (GDP), interest rate (i) and unemployment rate (u) with respect to κ_1 (left column), κ_2 (middle column) and κ_3 (right column); full sample.

Our results indicate that the point and interval forecasts are relatively insensitive to both κ_2 and κ_3. For example, if we change the value of κ_2 by one unit, the changes of point and interval forecasts are of the order of 10^{-5}. The impact of κ_3 is larger, but it is still inconsequential: the changes of all forecasts are of the order of 10^{-3} for each unit change of κ_3. Interestingly, the effects on the point and interval forecasts are of the same order. *A priori*, one might expect that, say, κ_3 would have a smaller impact on the point forecast than the interval forecasts as κ_3 controls the prior mean of the error covariance matrix $\boldsymbol{\Sigma}$. This is apparently not the case.

In contrast, the forecasts are all much more sensitive to the value of κ_1. This is perhaps not surprising as κ_1 controls the strength of shrinkage of the VAR coefficients, and it is well-known that appropriate shrinkage can substantially improve forecast performance. In addition to confirming the important role of shrinkage, our results also allow us to

calculate how small changes in κ_1 affect the forecasts. For example, if we increase κ_1 by 0.01 (recall that we set $\kappa_1 = 0.2^2 = 0.04$), the one-quarter-ahead point forecasts of real GDP growth, interest rate and unemployment rate will change by -0.026, 0.008 and 0.04, respectively.[3]

Figure 3 also suggests that sensitivities tend to decrease as forecast horizon increases. This could be due to the fact that long-horizon forecasts depend mainly on a particular combination of the VAR coefficients, but not on individual coefficients. To elaborate, recall that long-horizon forecasts converge to the unconditional mean of the system $\boldsymbol{\mu} = (\mathbf{I}_n - \sum_{i=1}^{p} \mathbf{B}_p)^{-1}\mathbf{b}$, which in general can be more precisely estimated than individual VAR coefficients. Consequently, the estimated $\boldsymbol{\mu}$ tend to be less sensitive to prior hyperparameters compared to individual VAR coefficients.

Next, we redo our sensitivity analysis using a shorter sample. This is motivated by the observation that many forecasters apply VAR to time series with significantly fewer observations than our full sample—either because of data availability issues or because they expect structural changes in their data and past observations might be less relevant. We re-estimate our model using data from 1989:Q3 to 2017Q4, and the estimated sensitivities are reported in Figure 4.

Despite the shorter sampler, the point and interval forecasts remain relatively insensitive to the hyperparameters κ_2 and κ_3. In contrast, the sensitivities of short-horizon forecasts with respect to κ_1 can be an order of magnitude larger. For example, if we increase κ_1 by 0.01, the one-quarter-ahead point forecasts of real GDP growth, interest rate and unemployment rate will change by -0.24, 0.075 and 0.4, respectively, (compared to -0.026, 0.008 and 0.04 when the full sample is used). Hence, it is especially important to conduct a sensitivity analysis when a relatively short sample is used.

Compared to the full sample results, the decrease in sensitivities as forecast horizon increases is more visible here. Using a shorter sample, it might be more difficult to pin down individual VAR coefficients than the unconditional mean. Consequently, short-horizon forecasts that depend more on individual coefficients would be more sensitive to hyperparameter values.

[3]To check these estimates, we rerun the sampler with $\kappa_1 = 0.05$, while keeping other hyperparameters exactly the same. The changes in the one-quarter-ahead point forecasts of real GDP growth, interest rate and unemployment rate are respectively -0.021, 0.007 and 0.034, which are very close to the original estimates.

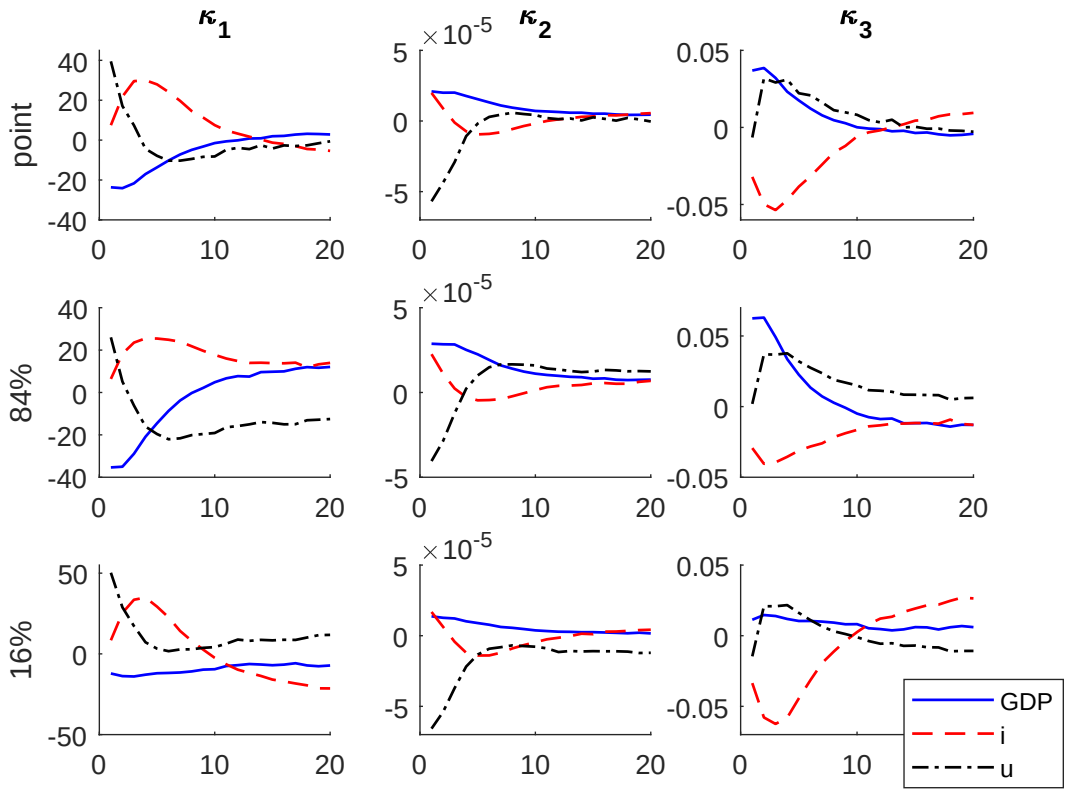

Figure 4: Derivatives of the point forecasts (top row), 84-percentiles (middle row) and 16-percentiles (bottom row) of the variables real GDP growth (GDP), interest rate (i) and unemployment rate (u) with respect to κ_1 (left column), κ_2 (middle column) and κ_3 (right column); subsample from 1989:Q3 to 2017:Q4.

5 Concluding Remarks

We have developed a general method based on Automatic Differentiation to assess how sensitive VAR forecasts are with respect to various key hyperparameters in a Minnesota-type prior. Using a US dataset, we have found that both point and density forecasts are relatively sensitive to the shrinkage strength of the VAR coefficients, but are not affected by that of the intercepts. Moreover, one could use our sensitivity estimates to obtain forecasts under slightly different hyperparameters. Hence, our approach provides an automatic way to assess the robustness of the forecasts.

In future work, it would be useful to develop similar automated sensitivity analysis of forecasts from more flexible models. This is motivated by recent findings that flexible

models such as time-varying parameter VARs developed in Cogley and Sargent (2001, 2005) and Primiceri (2005) tend to forecast substantially better, as demonstrated in Clark (2011), D'Agostino, Gambetti, and Giannone (2013) and Cross and Poon (2016).

Technical Appendix: Illustration of Automatic Differentiation

In general, Automatic Differentiation (AD) translates an algorithm of turning inputs into outputs, into its complementary algorithm of computing derivatives of the outputs with respect to the inputs. It heavily depends on decomposing the base algorithm into simpler operations such as addition, subtraction and multiplication, and updates the derivatives using chain-rule-based techniques based on the composition of these simpler operations.

To illustrate the idea, consider at the beginning of the gth iteration of the MCMC, we have already obtained $\left(\boldsymbol{\Sigma}^{g-1}\right)^{-1}$. If AD is applied complementarily, we have also its derivatives with respect to $\boldsymbol{\theta}_0$, denoted $\frac{\partial \text{vec}\left((\boldsymbol{\Sigma}^{g-1})^{-1}\right)}{\partial \boldsymbol{\theta}_0}$. Recall that in the MCMC algorithm, we first compute

$$\mathbf{K}^g = \mathbf{V}_{\boldsymbol{\beta}}^{-1} + (\boldsymbol{\Sigma}^{g-1})^{-1} \otimes \mathbf{X}'\mathbf{X}, \quad \mathbf{B}^g = (\mathbf{K}^g)^{-1}$$

and

$$\boldsymbol{\alpha}^g = \mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0 + \text{vec}(\mathbf{X}'\mathbf{Y}(\boldsymbol{\Sigma}^{g-1})^{-1}), \quad \mathbf{b}^g = \mathbf{B}^g\boldsymbol{\alpha}^g$$

Then, we sample the vector of coefficients as

$$\boldsymbol{\beta}^g = \mathbf{b}^g + \text{chol}(\mathbf{B}^g)\mathbf{Z}^g.$$

In what follows we apply AD to obtain $\frac{\partial \text{vec}(\boldsymbol{\beta}^g)}{\partial \boldsymbol{\theta}_0}$.

Consistent with the MCMC algorithm, terms such as

$$\frac{\partial \mathbf{V}_{\boldsymbol{\beta}}^{-1}}{\partial \boldsymbol{\theta}_0} = -((\mathbf{V}_{\boldsymbol{\beta}}^{-1})' \otimes \mathbf{V}_{\boldsymbol{\beta}}^{-1})\frac{\partial \mathbf{V}_{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}_0} \tag{5}$$

and

$$\frac{\partial \mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0}{\partial \boldsymbol{\theta}_0} = \mathbf{V}_{\boldsymbol{\beta}}^{-1}\frac{\partial \boldsymbol{\beta}_0}{\partial \boldsymbol{\theta}_0} + (\boldsymbol{\beta}_0' \otimes \mathbf{I}_{k_\beta})\frac{\partial \mathbf{V}_{\boldsymbol{\beta}}^{-1}}{\partial \boldsymbol{\theta}_0}$$

are pre-computed to save the computational cost.[4] [5] Given $\frac{\partial\left((\boldsymbol{\Sigma}^{g-1})^{-1}\right)}{\partial\boldsymbol{\theta}_0}$, then we have the complementary algorithm of computing derivatives

$$\frac{\partial\mathbf{K}^g}{\partial\boldsymbol{\theta}_0}=\frac{\partial\mathbf{V}_{\boldsymbol{\beta}}^{-1}}{\partial\boldsymbol{\theta}_0}+\frac{\partial\left((\boldsymbol{\Sigma}^{g-1})^{-1}\right)}{\partial\boldsymbol{\theta}_0}\otimes\text{vec}(\mathbf{X}'\mathbf{X}),\quad \frac{\partial\mathbf{B}^g}{\partial\boldsymbol{\theta}_0}=-(\mathbf{B}^{g'}\otimes\mathbf{B}^g)\frac{\partial\mathbf{K}^g}{\partial\boldsymbol{\theta}_0},$$

$$\frac{\partial\boldsymbol{\alpha}^g}{\partial\boldsymbol{\theta}_0}=\frac{\partial\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0}{\partial\boldsymbol{\theta}_0}+(\mathbf{I}_n\otimes\mathbf{X}'\mathbf{Y})\frac{\partial\left((\boldsymbol{\Sigma}^{g-1})^{-1}\right)}{\partial\boldsymbol{\theta}_0},\quad \frac{\partial\mathbf{b}^g}{\partial\boldsymbol{\theta}_0}=\mathbf{B}^g\frac{\partial\boldsymbol{\alpha}^g}{\partial\boldsymbol{\theta}_0}+(\boldsymbol{\alpha}^{g'}\otimes\mathbf{I}_{k_\beta})\frac{\partial\mathbf{B}^g}{\partial\boldsymbol{\theta}_0}.$$

Consequently, we can compute

$$\frac{\partial\boldsymbol{\beta}^g}{\partial\boldsymbol{\theta}_0}=\frac{\partial\mathbf{b}^g}{\partial\boldsymbol{\theta}_0}+\left(\mathbf{Z}'\otimes\mathbf{I}_{k_\beta}\right)\left.\frac{\partial\,\text{chol}(\mathbf{B})}{\partial\mathbf{B}}\right|_{\mathbf{B}=\mathbf{B}^g}\frac{\partial\mathbf{B}^g}{\partial\boldsymbol{\theta}_0},$$

where the term $\left.\frac{\partial\,\text{chol}(\mathbf{B})}{\partial\mathbf{B}}\right|_{\mathbf{B}=\mathbf{B}^g}$ can be found in Jacobi et al. (2018).

Now given $\frac{\partial\boldsymbol{\beta}^g}{\partial\boldsymbol{\theta}_0}$, we can then apply the same logic to obtain $\frac{\partial\left((\boldsymbol{\Sigma}^{g})^{-1}\right)}{\partial\boldsymbol{\theta}_0}$. Recall that

$$\nu_1=\nu_0+T,\quad \mathbf{S}^g=\mathbf{S}_0+\mathbf{Y}'\mathbf{Y}-(\boldsymbol{\beta}^g)'\mathbf{X}'\mathbf{Y}-((\boldsymbol{\beta}^g)'\mathbf{X}'\mathbf{Y})'+(\boldsymbol{\beta}^g)'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^g$$

$$\mathbf{L}=\text{chol}((\mathbf{S}^g)^{-1})$$

$$(\boldsymbol{\Sigma}^g)^{-1}=\mathbf{L}\mathbf{A}\mathbf{A}'\mathbf{L}',$$

where we consider the Bartlett decomposition of the Wishart distribution. It follows that

$$\frac{\partial\mathbf{S}^g}{\partial\boldsymbol{\theta}_0}=\frac{\partial\mathbf{S}_0}{\partial\boldsymbol{\theta}_0}+((\boldsymbol{\beta}^g)'\mathbf{X}'\mathbf{X}-\mathbf{Y}'\mathbf{X})\frac{\partial\boldsymbol{\beta}^g}{\partial\boldsymbol{\theta}_0}+((\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^g-\mathbf{Y}'\mathbf{X})\otimes\mathbf{I}_n)\left.\frac{\partial\mathbf{B}'}{\partial\mathbf{B}}\right|_{\mathbf{B}=\boldsymbol{\beta}^g}\frac{\partial\boldsymbol{\beta}^g}{\partial\boldsymbol{\theta}_0}$$

$$\frac{\partial\mathbf{L}}{\partial\boldsymbol{\theta}_0}=\left.\frac{\partial\text{chol}(\mathbf{B})}{\partial\mathbf{B}}\right|_{\mathbf{B}=(\mathbf{S}^g)^{-1}}(-(\mathbf{S}^g)^{-1'}\otimes(\mathbf{S}^g)^{-1})\frac{\partial\mathbf{S}^g}{\partial\boldsymbol{\theta}_0}$$

[4] The formula in (5) holds also for symmetric matrix $\mathbf{V}_{\boldsymbol{\beta}}$ if we differentiate it with respect to a vector $\boldsymbol{\theta}_0$ with independent components. To prove this formula, consider a symmetric, invertible $q\times q$ matrix $\mathbf{C}$. Since $\mathbf{C}^{-1}\mathbf{C}=\mathbf{I}_q$, taking derivative of both sides gives

$$\mathbf{C}'\otimes\mathbf{I}_q\frac{\partial\mathbf{C}^{-1}}{\partial\boldsymbol{\theta}_0}+\mathbf{I}_q\otimes\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\boldsymbol{\theta}_0}=\mathbf{0}.$$

Using $\mathbf{C}'=\mathbf{C}$ and re-arranging terms, we obtain

$$\frac{\partial\mathbf{C}^{-1}}{\partial\boldsymbol{\theta}_0}=-\left(\mathbf{C}\otimes\mathbf{I}_q\right)^{-1}\mathbf{I}_q\otimes\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\boldsymbol{\theta}_0}=-\mathbf{C}^{-1}\otimes\mathbf{I}_q\mathbf{I}_q\otimes\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\boldsymbol{\theta}_0}=-\mathbf{C}^{-1}\otimes\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\boldsymbol{\theta}_0}.$$

[5] Here $\frac{\partial\mathbf{V}_{\boldsymbol{\beta}}}{\partial\boldsymbol{\theta}_0}$ and $\frac{\partial\boldsymbol{\beta}_0}{\partial\boldsymbol{\theta}_0}$ depend on the specification of prior and the set of prior parameters of interest.

$$\frac{\partial\left((\boldsymbol{\Sigma}^g)^{-1}\right)}{\partial\boldsymbol{\theta}_0} = \mathbf{L}\mathbf{A}\mathbf{A}'\frac{\partial\mathbf{B}'}{\partial\mathbf{B}}\bigg|_{\mathbf{B}=\mathbf{L}}\frac{\partial\mathbf{L}}{\partial\boldsymbol{\theta}_0} + (\mathbf{L}\mathbf{A}\mathbf{A}'\otimes\mathbf{I}_n)\frac{\partial\mathbf{L}}{\partial\boldsymbol{\theta}_0}.$$

Here $\frac{\partial\mathbf{B}'}{\partial\mathbf{B}}$ is the commutation matrix associated with the matrix transpose operation (Magnus and Neudecker, 1979). We have also ignored the sensitivities with respect to ν_0 in this paper, and refer interested readers to Jacobi, Joshi, and Zhu (2018).

Typically, an efficient AD package does not evolve this complementary algorithm symbolically as we just demonstrated above, but passes the original algorithm by reference. It recognises simple operations, e.g. $\mathbf{V}_{\boldsymbol{\beta}}^{-1}$ is a matrix inversion operation of the form $\mathbf{A}^{-1}|_{\mathbf{V}_{\boldsymbol{\beta}}}$, then translates to its derivative counterpart $-(\mathbf{A}^{-1'}\otimes\mathbf{A}^{-1})|_{\mathbf{A}=\mathbf{V}_{\boldsymbol{\beta}}}$. This is done numerically, such that there is no need for the above symbolic derivations at all.

References

M. Banbura, D. Giannone, and L. Reichlin. Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

L. Bauwens, M. Lubrano, and J. Richard. *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, New York, 1999.

CBO. The budget and economic outlook: 2018 to 2028. Technical report, The US Congressional Budget Office, 2018.

J. C. C. Chan. Notes on Bayesian Macroeconometrics. 2017. Available at: `http://joshuachan.org/notes_BayesMacro.html`.

J. C. C. Chan. Large Bayesian VARs: A flexible Kronecker error covariance structure. *Journal of Business and Economic Statistics*, 2018. Forthcoming.

T. E. Clark. Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *Journal of Business and Economic Statistics*, 29(3):327–341, 2011.

T. Cogley and T. J. Sargent. Evolving post-world war II US inflation dynamics. *NBER Macroeconomics Annual*, 16:331–388, 2001.

T. Cogley and T. J. Sargent. Drifts and volatilities: Monetary policies and outcomes in the post WWII US. *Review of Economic Dynamics*, 8(2):262–302, 2005.

J. Cross and A. Poon. Forecasting structural change and fat-tailed events in Australian macroeconomic variables. *Economic Modelling*, 58:34–51, 2016.

A. D'Agostino, L. Gambetti, and D. Giannone. Macroeconomic forecasting and structural change. *Journal of Applied Econometrics*, 28:82–101, 2013.

T. Doan, R. Litterman, and C. Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100, 1984.

M. C. Fu, L. J. Hong, and J.-Q. Hu. Conditional Monte Carlo estimation of quantile sensitivities. *Management Science*, 55(12):2019–2027, 2009.

M. Giles and P. Glasserman. Smoking adjoints: Fast Monte Carlo Greeks. *Risk*, 19(1): 88–92, 2006.

P. Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer Science & Business Media, 2013.

L. J. Hong. Estimating quantile sensitivities. *Operations Research*, 57(1):118–130, 2009.

L. Jacobi, M. S. Joshi, and D. Zhu. Automated sensitivity analysis for Bayesian inference via Markov chain Monte Carlo: Applications to Gibbs sampling. 2018. Available at SSRN: `http://dx.doi.org/10.2139/ssrn.2984054`.

M. Joshi and C. Yang. Algorithmic Hessians and the fast computation of cross-gamma risk. *IIE Transactions*, 43(12):878–892, 2011.

K. Kadiyala and S. Karlsson. Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, 12(2):99–132, 1997.

S. Karlsson. Forecasting with Bayesian vector autoregressions. In G. Elliott and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 2 of *Handbook of Economic Forecasting*, pages 791–897. Elsevier, 2013.

G. Koop. Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, 28(2):177–203, 2013.

G. Koop and D. Korobilis. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3(4):267–358, 2010.

E. E. Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

R. Litterman. Forecasting with Bayesian vector autoregressions — five years of experience. *Journal of Business and Economic Statistics*, 4:25–38, 1986.

G. Liu and L. J. Hong. Kernel estimation of quantile sensitivities. *Naval Research Logistics*, 56(6):511–525, 2009.

J. R. Magnus and H. Neudecker. The commutation matrix: some properties and applications. *The Annals of Statistics*, pages 381–394, 1979.

D. J. Poirier. Frequentist and subjectivist perspectives on the problems of model building in economics. *Journal of Economic Perspectives*, 2(1):121–144, 1988.

G. E. Primiceri. Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72(3):821–852, 2005.

C. A. Sims. Macroeconomics and reality. *Econometrica*, 48:1–48, 1980.

T. Woźniak. Bayesian vector autoregressions. *Australian Economic Review*, 49(3):365–380, 2016.