About Local Projection Impulse Response Function Reliability *

Luca Brugnolini

European Central Bank and University of Rome "Tor Vergata"

I would like to thank Giuseppe Ragusa for pointing me to the local projection methodology and for all his valuable suggestions. Lutz Kilian for sharing his code and being available to discuss the paper. Marco Lippi for his dedication. Oscar Jordà for is availability. Andrè Meier for his helpful comments and feedbacks. Antonello D'Agostino. Tommaso Proietti and Roberto Di Mari for their precious suggestions. All errors are mine.

Corresponding author: Luca Brugnolini luca.brugnolini@gmail.com, European Central Bank/Monetary Policy Strategy and PhD candidate at The University of Rome "Tor Vergata"

Abstract

I compare the performance of the VAR impulse response function (IRF) estimator with the Jordà (2005) local projection (LP) methodology. I show by a Monte Carlo exercise that when the data generating process (DGP) is a well-specified vector autoregressive model (VAR), the standard estimator is a better alternative. However, in the general case in which the sample size is small, and the lag length of the model is misspecified, the local projection estimator is a competitive alternative to the standard VAR impulse response function estimator. Along the way, I highlight some lack in the local projection literature, which can lead to potential improvement in the estimation procedure.

Keywords: VAR, information criteria, lag-length, Monte Carlo
JEL: C32, C52, C53, E52

Non-technical summary

The main working tool to investigate the effects of a policy action is the Impulse Response Function (IRF). IRFs measure the change in an endogenous variable given the shift in an exogenous one. For example, IRFs are used in monetary policy to measure the effect of an unanticipated change in the ECB interest rates on macroeconomic variables. In particular, IRFs describe the reaction of a dynamic system to an exogenous change as a discrete function of time. This characteristic feature is crucial. For example, a monetary policy action cannot have any immediate impact on real variables but can have a significant effect after a period of time. Therefore, measuring only the contemporaneous outcome might lead to misleading conclusions. This quality justifies the popularity of the IRFs as a leading working tool.

However, IRFs are estimated quantities, and as such, different methodologies exist in the literature. The most common procedure to estimate the IRFs is by the aid of an auxiliary model. In particular, researchers use a Vector Autoregressive Model (VAR) to conduct this type of analysis. However, IRFs are a quantity on its own and can also be estimated without the support of a model. In recent times, especially in macroeconomics, many authors have started applying model-free IRFs (also called Local Projection IRFs - LP). The reason is related to the flexibility of the LP methodology. Although the number of papers which employ this technique is exponentially increasing, not much is known about the performance of this estimator. Indeed, the only paper that seriously examines the issue concludes in favor of the standard methodology. Nevertheless, researchers keep using local projection without any theoretical nor empirical base.

Against this background, the present paper explores this issue and re-examines the critical findings in the previous literature. The article shows that local projection IRFs are indeed a valid option to estimate IRFs and that, in a general setting, LP may be preferable to the standard VAR IRF estimator. In particular, it shows why previous findings conclude in favor of the VAR IRFs methodologies. The reason is extremely linked to the way in which the lag-length selection is performed in both methods.

The paper has two main implications: from one point of view, by finding a crack in the previous critiques legitimates the use of LP IRFs; from the other, it shows that many details have to be deepened to exploit LP methodology at full potential. In particular, it highlights the lag-length selection procedure as the leading candidate for improvements.

1 Introduction

Since Sims (1980) influential paper *Macroeconomics and reality*, vector autoregressive models (VARs) have become pervasive in the empirical macroeconomic literature. Among the VAR tools, impulse response functions (IRFs) are undoubtedly the most employed. In general, an impulse response describes the reaction of a dynamic system to an exogenous change as a discrete function of time. In macroeconomics, impulse responses are widely used to estimate quantities of interest such as *multipliers* which are proportionality factors that weigh the impact change in an endogenous variable due to a change in an exogenous variable. For example, after the formulation of the multiplier theory by Keynes (1936) in *The General Theory of Employment, Interest and Money,* many researchers have attempted to estimate the sign and the size of the *fiscal multiplier*, which is the factor by which the national income vary due to a change in government spending. Of course, the use of impulse responses are not restricted to the fiscal policy literature but reach all economic field involving policy evaluation as changes in technology, price, taxation, lending, interest rate and in many other quantities. In macroeconomics, an unexpected shift in a well-identified variable defines a *structural shock*. The primary use of impulse responses (named *structural impulse responses*) is to estimate the effect of a structural shock to an endogenous variable. Also, impulse responses are used to trace-out the dynamic path of the endogenous variable due to the shock and in addition to computing interesting statistics as the cumulative response of the endogenous variable to the shock over time. Moreover, a researcher can use an accurate estimate of the impulse responses to test in the data the existence of theoretical economic concepts. For example, price stickiness, which is the propensity of price to react with some lag to a structural shock, or the money neutrality, which is the zero-mean reversion tendency of real variables to a monetary policy shock. In multivariate context, also the economic theory about channels through which shocks propagates can be tested by the help of impulse response functions. Finally, in modern macroeconomics, impulse responses are also used to estimate parameters in *dynamic stochastic general equilibrium* (DSGE) models. Selecting the parameters which minimize the distance from the VAR to the log-linearized DSGE model impulse responses, a researcher can determine these quantities.

For the variety of topics covered by impulse responses, their use is extensive, and due to their massive application, statistical properties of VAR impulse response functions have been studied for a long time

and by authors from very different fields. This effort has built a broad common knowledge about advantages and disadvantages of the estimation procedure of the impulse responses from a VAR model. In fact, the standard estimation technique relies on the autoregressive polynomial invertibility and the *Wold's decomposition theorem*. The theorem allows casting a p-order VAR in an infinite order vector moving average (VMA) and recover the VMA coefficients recursively as a nonlinear function of the VAR coefficients. For example, given the recursiveness in the estimation process and the mapping between the VAR and VMA model, standard IRFs suffer from some well-known issues. For example, Pope (1990) shows that small-sample bias in the impulse responses stems from the bias in VAR estimates. Also, given that impulse responses are nonlinear functions of VAR coefficients, due to nonlinearity, increasing the forecast horizons widen the bias in the estimated IRFs. However, at least at the first step ahead, VARs produce an optimal and robust to model misspecification IRF (Stock and Watson, 1999).

Recently, Jordà (2005) has introduced a new methodology to compute IRFs named *model-free* or *local projection impulse response function estimator*. The central idea resembles the direct forecast procedure. However, as the name suggests, the "model-free" estimator does not need an auxiliary model to estimate the impulse responses. Also, the local projection formulation is more general than the standard VAR procedure and allows the estimation of the impulse responses even when the VMA representation does not exist. Finally, as shown in the original paper, this methodology is robust to the different form of model misspecification and can easily accommodate non-linearities as state-dependency or size and sign dependency. For its advantages, the local projection methodology is quickly getting attention in the macroeconomic literature. In particular, it is becoming a conventional device in the economist toolbox to assess the causal relationship between exogenous shocks and endogenous macroeconomic variables.

To provide a snapshot of the popularity of this new methodology, I show in Figure 1 the number of citation per year of the local projection seminal paper by Oscar Jordà. Considering the number of citations per year as a proxy for the establishment of the methodology, I highlight in the figure how the local projection estimator is becoming successful at a rapid rate. An important feature not shown by the chart is that citations are almost entirely due to applied macroeconomic papers; for example, Haug and Smith (2011) compares LPIRFs with standard IRFs in a small open economy. Hall et al. (2012) use LPIRFs to estimate a DSGE model. Auerbach and Gorodnichenko (2012a,b, 2013, 2016), Owyang et al. (2013) and Ramey and Zubairy (2017) use local projection to determine the fiscal multiplier among the different economic states (recession/expansion). Hamilton (2011) exploit LPIRFs to trace out the dy-

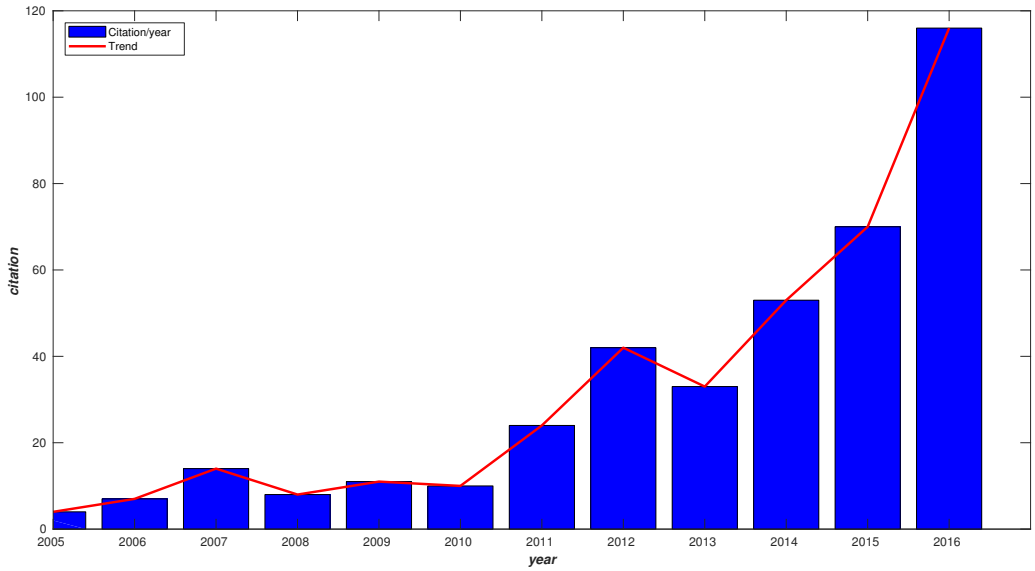

Figure 1: Jordà (2005) local projection estimator. The blue bars show the number of citations per year. The red-solid line highlights the citation trend. For 2017 the number of quotes is around 110. Data from Google Scholar Citation.

namics of an oil shock on GDP and to assess non-linearities. Ambrogio Cesa-Bianchi (2016), Tenreyro and Thwaites (2016), Caldara and Herbst (2016), Miranda-Agrippino and Ricco (2017) and Swanson (2017) have applied local projection impulse response function to assess the response of real and financial variables to a monetary policy shock. Ramey (2016) describes this brand new field in a recent chapter of *The Macroeconomic Handbook*. The impression is that this methodology is very connected to the literature of identification through external instruments. Besides, a novel series of papers by Barnichon and Matthes (Barnichon et al. 2016, Barnichon and Matthes 2017a, Barnichon and Matthes 2017b) uses local projection as a benchmark to compare the properties of the new *Functional Approximation of Impulse Responses* (FAIR) methodology. This consist in approximating impulse response function with Gaussian basis function. Having a new method widely embraced in a scientific community can be an extremely positive sign; nevertheless, the fact that is getting attention among "practitioners" is a call for "developers" to increase the number of tests and disclose potential drawbacks in the estimation procedure. Also, due to the relevance of the topics, deepening and improving the knowledge of this new methodology should be a primary concern for both statisticians and econometricians. A step in this direction is due to Barnichon and Brownlees (2017); in their paper, the authors show the benefits of applying a linear B-spline basis function to the estimated local projection coefficients to reduce their variance. Also, Miranda-Agrippino and Ricco (2017) have proposed a Bayesian estimation procedure for the local projection IRFs. Both papers have introduced such methodologies based on the assumption that VAR

IRFs are more efficient, but local projection is less prone to bias when the model is misspecified. In this regards, the two methodologies resemble the direct versus iterated forecast procedure (Marcellino et al., 2006). However, in this respect, we would like to take a step back and empirically find evidence of this functioning in the spirit of Kilian and Kim (2011) - from now on Kilian and Kim.

Kilian and Kim have opened the debate about the reliability of the local projection methodology. This paper, to the best of my knowledge, is the only critical assessment of this new estimation procedure, with a particular focus on the coverage ability of the confidence bands of both methodologies. Providing that the two methods are asymptotically equivalent, the authors test them in a small-sample context. They use a 100 data points VAR(1), a 460 data point VAR(12) and a 200 data point VARMA(1,1) and conclude that the standard VAR impulse response functions are more precise than local projection estimator. In this spirit, a working paper by Ronayne (2011) tests local projection against VAR impulse responses in an IS-LM model finding some evidence in favor of the first methodology. At the opposite, in an unpublished manuscript, Meier (2005) tests the two alternatives using as DGP the Smets and Wouters (2003) DSGE model, finding evidence in favor of the VAR methodology. Against this background, the object of the present paper is to make some clarity in this compact literature, trying to understand what drives this mixing results. Also, as a complementary objective, I highlight some gaps between the two methodologies that researchers should fill to compare them more fairly.

In this paper, I start from Kilian and Kim Monte Carlo study to critically assess whether VAR impulse responses are consistently better than local projection IRFs. The objective of the paper is to understand under which conditions local projection is a convincing alternative. Due to the marked connection of this methodology to the empirical macroeconomic literature, my focus is on a realistic DGP as the VAR(12) model developed by Christiano et al. (1999); this model is designed to estimate the effects of a monetary policy shock in The US, meaning an unexpected raise or cut in the Federal Reserve interest rate. Kilian and Kim use the same model as DGP. Following the two authors, in this article, I abstract from the identification problem and assume that the identification procedure in the data generating process is known. The reason is that the identification in the VAR class of models is a thorny issue, and given that the focus of the paper is to assess *ceteris paribus* the performance of local projection against VAR impulse responses, assuming an unknown identification procedure can potentially include a new source of bias and contaminating the estimation. The standard method adopted in the applied literature is to estimate a time-invariant impact matrix aided by a VAR model and weight the local projection coeffi-

cients by the appropriate impact matrix coefficients. However, even if Jordà uses this procedure in the original GAUSS code of the paper, he does not provide any theoretical justification (as also noted in Kilian and Kim, see note 4). Also, a more recent procedure consists in estimating a proxy series for the unobserved shock series and plug it directly into the local projection framework (Tenreyro and Thwaites, 2016; Ramey, 2016; Swanson, 2017; Owyang et al., 2013). My object is to remain as close as possible to their analysis to have a fair comparison between the two procedure. For this reason, I apply equal performance measures and generate for each repetition an equivalent number of data points. In the analysis, I prove that some assumptions about the lag-length selection procedure deliver an unfair comparison between the local projection and VAR impulse responses, returning a comparison between a well-specified VAR and a misspecified local projection model. Using a well-specified VAR returns correctly specified IRFs where the only source of distortion arises from small-sample bias. On the contrary, local projection suffers from an augmented form of small-sample bias plus the model misspecification bias. Therefore, in the following simulation, I release the implicit assumption that the VAR model is well specified. Results show a very different picture, with the local projection being on average preferred to VAR impulse responses.

Although the results do not strongly conclude in favor of one methodology, the massive use of the Jordà's estimator in the applied macroeconomic literature is a fact that researchers should cautiously consider. For this reason, it is extremely important that econometrician and statisticians quickly gather with this new methodology to investigate under which conditions local projection is valid, but also under which is not. This paper goes in this direction, attempting to find a crack in the powerful critique by Kilian and Kim and contributing to the existing literature by further testing the local projection methodology under the assumption of misspecified leg-length.This scenario seems natural given that in applied experiments the true DGP is always unknown. Moreover, a voluminous literature suggests that the data generating process of many macroeconomic variables are vector autoregressive moving average (VARMAs) and not VAR (Zellner and Palm, 1974; Wallis, 1977; Cooley and Dwyer, 1998 and Chan and Eisenstat, 2017). Thus, it is possible to state that standard IRFs estimated from a VAR are misspecified by construction. Then, for its properties, the local projection methodology is an excellent candidate to be a valid and more general alternative to standard IRFs estimator.

The rest of the paper is organized as follows. Section 2 describes the VAR and LP methodologies. Section 3 describes the critiques by Kilian and Kim. Section 4 describes the Monte Carlo experiment

performed in this paper. Section 5 discusses the results and highlights some gaps between the literature of the two methodologies. Finally, Section 5 concludes.

2 Impulse response function estimation

In this section, I report the standard procedure to estimate impulse responses recursively from a VAR model and the direct estimation procedure used in local projection.

2.1 VAR impulse response functions

The standard procedure to recover impulse responses is to map the estimated VAR coefficient to VMA coefficients recursively. The Wold representation theorem is the mapping device. It states that any covariance-stationary time series can be rewritten as a sum of present and past innovation. Therefore, the first step in the impulse response estimation is to estimate the VAR autoregressive coefficients via ordinary least square (OLS). The OLS in the autoregressive model is the best linear unbiased estimator (BLUE) and corresponds to the conditional maximum likelihood estimator [1]. Provided that all the roots of the autoregressive polynomial lie outside the unit circle, a VAR(p) is always invertible and can be rewritten as a VMA(∞). This estimation procedure is theoretically justified when the model corresponds to the underlined DGP. Equation 1 shows a *reduced form* VAR(p):

$$y_t = B_1 y_{t-1} + B_2 y_{t-2} + ... + B_p y_{t-p} + e_t \tag{1}$$

where $t = p+1, ..., T$, $y_t \equiv (y_{1t}, y_{2t}, ..., y_{Kt})'$ is $(K \times 1)$ random vector, B_i, $i = 1, ..., p$, are $(K \times K)$ matrices of VAR coefficients and $e_t \equiv (e_{1t}, e_{2t}..., e_{Kt})'$ is $(K \times 1)$ vector of independent and identically distributed white noise with $\mathbb{E}(e_t) = 0$, $\mathbb{E}(e_s e_t') = 0$, for $s \neq t$ and $(K \times K)$ variance-covariance matrix $\mathbb{E}(e_t e_t) = \Sigma_e$. The VAR($p$) process can always be rewritten in *structural form* as in Equation 2,

$$A_0 y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + \varepsilon_t \tag{2}$$

where the $(K \times K)$ variance-covariance matrix of ε_t, Σ_ε is diagonal and positive-definite. A_0 is

[1] Hamilton 1994 chapter 8 for the proof.

$(K \times K)$ and in economic parlance is called the *impact matrix* and contains the contemporaneous effect of the increase of each endogenous variable on the others. The relationship between structural shocks ε_t and reduce form shocks e_t is given by Equation 3:

$$A_0 e_t = \varepsilon_t \tag{3}$$

According to the identification scheme chosen, A_0 might be any matrix imposing $\frac{1}{2}(K^2 - K)$ restrictions. As we abstract to any identification issue in this paper, we simply assume that the DGP identification scheme is known and accomplished by imposing timing restriction, such as $\Sigma_e = A_0^{-1}\varepsilon_t\varepsilon_t' A_0^{-1'} = A_0^{-1}\Sigma_\varepsilon A_0^{-1'} = A_0^{-1}A_0^{-1'}$, and $\Sigma_\varepsilon = \mathbb{I}_K$, where $\mathbb{I}_K$ is a $(K \times K)$ identity matrix. Given that Σ_e is Hermitian, than A_0^{-1} can be retrieved from its Cholesky factorization. Also, due to positive-definiteness of Σ_e, the decomposition is unique. By applying the Wold representation theorem, the VAR model in Equation 1 can be rewritten as Equation 4:

$$y_t = \Phi(L)e_t \tag{4}$$

Where $\Phi(L) = (\mathbb{I}_k + \Phi_1 L + \Phi_2 L^2 - ...)$ is the VMA polynomial, Φ_j, $j = 1, ..., \infty$, are $(K \times K)$ matrices of VMA coefficients and L is the lag operator such that $Ly_t \equiv y_{t-1}$. In order to estimate the VAR impulse responses and showing the mapping between VAR coefficients and IRFs, we use the conditional forecast difference definition as in Hamilton (1994) in Equation 5:

$$IRF(t, h, a_i) = \mathbb{E}\left[y_{t+h}|e_t = a_i\right] - \mathbb{E}\left[y_{t+h}|e_t = 0_k\right]$$

$$IRF(t, h, a_i) = \Phi_h a_i \tag{5}$$

Where 0_K is a $(K \times 1)$ zero column vector and the impulse responses are a function of time t, horizons h and a $(K \times 1)$ column vector of the impact matrix A_0^{-1} (namely a_i). By using this definition, we can recover the IRFs coefficients from the relationship linked VAR to VMA polynomial $\Phi(L) \equiv B(L)^{-1}$, $B(L)^{-1} \equiv (\mathbb{I}_k - B_1 L - B_2 L^2 - ... - B_p L^p)^{-1}$. Given this relationship, it is possible to show that the set of reduced form and structural IRFs Θ_h are given by Equations 6 and 7:

$$\Phi_h = \sum_{l=1}^{h} \Phi_{h-l} B_l, \quad h = 1, 2, ..., H \tag{6}$$

$$\Theta_h = \Phi_h A_0^{-1}, \quad h = 1, 2, ..., H \tag{7}$$

where $\Phi_0 = \mathbb{I}_K$, $B_l = 0$ for $l > p$ and A_0^{-1} satisfy the triangular representation $A_0^{-1}(A_0^{-1})' = \Sigma_e$. From the last two equations, it is clear that IRFs are functions of VMA parameters, which in turns are non-linear functions of the VAR parameters.

2.2 Local projection impulse response functions

Differently, from the recursive procedure, the local projection procedure consists of directly estimating the autoregressive coefficients locally at each step-ahead, regressing the dependent variable on its past as in Equation 8.

$$\begin{cases} y_{t+1} = B_1^1 y_t + B_2^1 y_{t-1} + ... + B_p^1 y_{t-p} + e_{t+1}, \quad e_{t+1} \sim MA(1) \\ y_{t+2} = B_1^2 y_t + B_2^2 y_{t-1} + ... + B_p^2 y_{t-p} + e_{t+2}, \quad e_{t+2} \sim MA(2) \\ \quad \vdots \\ y_{t+H} = B_1^H y_t + B_2^H y_{t-1} + ... + B_p^H y_{t-p} + e_{t+H}, \quad e_{t+H} \sim MA(H) \end{cases} \tag{8}$$

As shown by Jordà (2005), directly estimating the $(K \times K)$ autoregressive coefficients B_1^h, $h = 1, ..., H$, correspond to estimating the IRFs without casting the Wold representation theorem. He also shows that the errors arising from this projections are a VMA process of order h. Due to this issue, the author suggests estimating the variance-covariance matrix using the Newey and West (1987) and Andrews (1991) *heteroskedasticity and autocorrelation consistent estimator* (HAC).

Jordà shows that local projection is pointwise more robust to model misspecification than VAR impulse responses due to the direct estimation versus the VAR iterated procedure. According to the author, the iterated method compounds the error due to misspecification as a function of the horizon. Also, as shown by the author, the local projection can be easily adapted to a nonlinear framework by including nonlinear regressors (*flexible local projection*). In this context, a researcher can simultaneously or individually estimate all the equations. This critical twist is particularly appreciated in the applied macroeconomic literature, giving the ability to practitioners to easily embed in the model crucial economic feature as a state, size and sign dependencies (Owyang et al. 2013, Tenreyro and Thwaites 2016).

Secondly, an essential advantage of local projection over VAR IRFs is in the direct estimation of the impulse response coefficients. However, to determine the structural impulse responses, a researcher needs to estimate the impact matrix in a first-step aided by an auxiliary model as a VAR, and then plug the estimated matrix in the local projection equations. Standard VAR identification procedure can be used. The only alternative is when there is a proxy series for the shocks available (see for example Swanson, 2017). Equation 9 shows structural impulse response function estimated via local projection:

$$\begin{cases} y_{t+1} = \hat{A}_1^1 y_t + \hat{A}_2^1 y_{t-1} + ... + \hat{A}_p^1 y_{t-p} + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim MA(1) \\ y_{t+2} = \hat{A}_1^2 y_t + \hat{A}_2^2 y_{t-1} + ... + \hat{A}_p^2 y_{t-p} + \varepsilon_{t+2}, \quad \varepsilon_{t+2} \sim MA(2) \\ \quad \vdots \\ y_{t+H} = \hat{A}_1^H y_t + \hat{A}_2^H y_{t-1} + ... + \hat{A}_p^H y_{t-p} + \varepsilon_{t+H}, \quad \varepsilon_{t+H} \sim MA(H) \end{cases} \tag{9}$$

Where $\hat{A}_p^h = \hat{A}_0^{-1} B_p^h$ and $\hat{A}_1^h = \hat{A}_0^{-1} B_1^h$ are the structural local projection IRFs. Both Jordà and Kilian and Kim use this methodology in their paper, even if the fact that the structural local projection impulse response functions are found using an auxiliary model does not have any theoretical treatment. This is because both the authors are abstracting from the identification issue. Given that my objective is to assess the local projection procedure ceteris paribus to the other studies, in the present paper, I do not take any further step in this direction and assume that the identification scheme is known and that the impact matrix is projection invariant.

Switching the focus from the advantages to the drawbacks of the local projection procedure, the data consuming nature appears as its first limit. In fact, increasing the horizons of the impulse responses reduces the sample available for the estimation itself. Therefore, while VAR consumes data only along the lag dimension (p), local projection consumes data along both the lag (p) and the lead (h) dimension.

3 Critique: Kilian and Kim

The more influential critique of the local projection method is due to Kilian and Kim. The authors compare the two methodologies in a small sample size using as a performance criterion the *effective coverage rate* (ECR) of the impulse response confidence bands and the *average length* (AL). The effective cover-

age rate, as defined in Equation 10, is a $(H \times 1)$ vector which describes how many times the true IRFs lie in the confidence interval of the estimated impulse response functions in repeated trials[2].

$$ECR(h) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}\left(IRF_{true}(h) \in \left[IRF_L^{(m)}(h), IRF_H^{(m)}(h)\right]\right), \quad h = 1, ..., H \tag{10}$$

Where $m = 1, ..., M$ is the number of repetition in the Monte Carlo simulation, $IRF_{true}(h)$ is the true impulse response generated from the DGP and $IRF_r^{(m)}(h)$, with $r = \{L, H\}$, is the upper/lower bound of the estimated confidence interval for horizon h. $\mathbb{I}$ is an indicator function which assigns value 1 when the true impulse response belongs to the estimated confidence bands and 0 otherwise, implying $0 \leq ECR(h) \leq 1$. When the coverage rate is equal to zero, the true impulse response, on average, never lies inside the estimated confidence band. At the opposite, when it is 1, the true impulse response lies in the confidence interval with probability 1. Depending on the significance level α chosen by the researcher, as the closer the estimated $ECR(h)$ to its complement $(1 - \alpha)$, as precise the estimator. For example, when $\alpha = 0.05$ the true value of the impulse response function should lie on average 95% of the times within the confidence bands. Deviations from this value induce over/under coverage of the impulse responses.

As a complementary criterion, Equation 11 describes the average length of the confidence bands. The AL results in a $(H \times 1)$ vector which measures the confidence band wideness .

$$AL(h) = \frac{1}{M}\sum_{m=1}^{M} \left|IRF_H^{(m)}(h) - IRF_L^{(m)}(h)\right|, \quad h = 1, ..., H \tag{11}$$

Shorter average length implies more precise estimate on average. The two criteria together give some complementary information. For example, when an estimator has very high coverage rate but very high average length, it means that the higher coverage comes from more substantial uncertainty. A second case is when the estimator has meager coverage rate and very low AL. In this instance, the two criteria suggest that even if the confidence bands are precisely estimated, the true impulse response is out of them most of the time. Naturally, the best situation is when an estimator has high coverage rate and low average length. The worst is the opposite.

Using these criteria for evaluating the performance of the local projection and VAR estimators, im-

[2]In what follows, I am saving on notation when it is clear from the context.

pose the researcher to take a stand concerning the methodology used to estimate the confidence bands of the IRFs. Given that this choice may influence the performance, the authors choose two different criteria for each method and report the results for all the four. The compared methods to estimate confidence bands for VAR are Lutkepohl (1990) *delta method* and Kilian's *bias-corrected bootstrap* method (Kilian, 1998a,b,c). The former is a generalization of the nonparametric bootstrap confidence interval by Runkle (1987), which is tailored to account also for bias and skewness in IRFs. For local projection, we compare the asymptotic procedure developed by Jordà (Jordà 2009, 2005) and the block-bootstrap method developed by Kilian and Kim.

The main results of the paper show that the VAR IRFs always have more coverage rate and less average length than the LP counterpart.

4 Monte Carlo experiment

Focusing on the Kilian and Kim VAR(12) DGP, we briefly describe the steps involved in the Monte Carlo experiment. I analyze the simulation of $K = 4$ variable VAR(12) model which uses as DGP the model by Christiano et al. (1999) to identify the monetary policy shock. The reason for relying on a VAR(12) DGP is that the bivariate VAR(1) model is a too simple example to describe realistic models often used by economists, while VARMA models are not a popular choice among practitioners. Thus, given that the paper is tailored to be helpful to practitioners, I point the analysis to the most relevant case in their respect. The variables involved are the CFNAI index of US real activities[3], the US CPI inflation, US commodity price inflation and the effective FED fund rate. The sample covers the period from January 1970 to December 2007 (the sample size is around 460 observations), and the model is specified as in Kilian and Kim to enhance comparability. Therefore, the CFNAI index and the FED fund rate are taken in level, while the CPI and the commodity prices in log difference multiplied by 1200. All the variables are demeaned. Theoretically, the monetary policy shock is the only one identified in this model, as the authors exploit contemporaneous restriction to determine the Taylor rule used by the FED. The authors also include the commodity price index to account for inflation expectation and address the "price puzzle". The former is the tendency of the IRFs to show price growing after an interest rate hike.

[3]This is a measure of real output gap produced by the Federal Reserve Bank of Chicago. More details can be found at the following link: CFNAI index.

However, assuming that the impact matrix A_t is identified, implies that all the shocks in the system are also identified. Thus, also the impulse produced by the VAR and LP for other shocks can be used to assess the ECR and AL. The Monte Carlo experiment is designed as follows:

1. Fitting a VAR(12) on the original data-set;
2. Estimating and saving the coefficients by an OLS regression;
3. Generating a new series of simulated data from the estimated model ($T = 456 + b$, where $b = 300$ is the burn-in);
4. Selecting the lag-length p via IC;
5. Fitting a VAR(p) on the simulated data and compute VAR impulse response function;
6. Computing local projection impulse responses by selecting the lag-length p^{lp} via IC (this is a $(H \times 1)$ vector);
7. Computing 95% confidence bands by delta method and double-bootstrap using 500 repetitions of the algorithm[4];
8. Storing the results.

At the end of the $M = 1000$ Monte Carlo repetition of step 1 to 8, I compute the ECR and AL by averaging the stored results. In comparing the outcome of the coverage rate, the significance level is set to $\alpha = 0.05$. Given the complexity of the experiment and the high number of repetition (especially considering the bootstrap methodology), I implement the original code in Julia Language (Bezanson et al. 2017), which is a dynamic and high-performance programming language[5].

Figure 2 shows the main results from the Monte Carlo experiment selecting the lag-length with the AIC with an upper bound $\overline{p} = 12$ as in the Kilian and Kim paper. The first two panels presents the ECR and AL for four different methodologies to estimate the impulse responses confidence banda. These are the asymptotic delta method for VAR impulse responses (Lutkepohl 1990), the asymptotic interval for local projection (Jordà, 2005, 2009) and bias-corrected bootstrap (Kilian 1998a,b,c; Kilian and Kim 2011) for both methodologies. The solid red line acts as a reference line for each statistics. For

[4] Using 2000 repetitions does not affect the results

[5] The package I am currently building for the VAR and LP estimation is available at this link.

completeness, I extend the original paper figures including also the estimated *BIAS*, *Mean-Squared-Error* (MSE) and *Standard Deviation* (STD) of the impulse responses as described in Equation 12 to 14[6].

$$BIAS\,(h) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{IRF}^{(m)}(h) - IRF_{true}(h)\right), \quad h = 1, ..., H \tag{12}$$

$$MSE\,(h) = \frac{1}{M}\sum_{m=1}^{M}\left(IRF_{true}(h) - \hat{IRF}^{(m)}(h)\right)^2, \quad h = 1, ..., H \tag{13}$$

$$STD\,(h) = \sqrt{MSE(h) - BIAS(h)^2}, \quad h = 1, ..., H \tag{14}$$

Conversely to the original paper, the figure shows the average result along with all the K^2 IRFs as described in Equation 15 instead of only the IRFs for the monetary policy shock.

$$\bar{C}\,(h) = \frac{1}{K^2}\sum_{j=1}^{K^2} C_j, \quad C = \{AL, BIAS, ECR, MSE, STD\} \tag{15}$$

Given that a known identification scheme is chosen, all the shocks can be considered identified. However, for completeness, I also show in the appendix the single IRFs for the monetary policy shock as presented in the original paper. From the figure, it is clear that VAR IRFs always have higher coverage rate, lower average length, but also less MSE and STD. The results for the BIAS are mixed, but the LP estimator varies widely. Those are all characteristics which should lead a researcher to fully discard the local projection methodology in favor of the VAR procedure. This result is difficult to reconcile with the Jordà (2005) original paper findings, which states that the main advantages of the local projection against VAR IRFs are in a misspecified context. To deepen the cause of this result, I analyze the way in which the entire Monte Carlo exercise is built and in particular the lag-length selection procedure. Figure 3 is the starting point of this analysis. The upper panel shows the distribution of the lag-length selected using the AIC for the VAR model in all the $M = 1000$ replication of the Monte Carlo simulation (this is also the same simulation performed by Kilian and Kim), and the lower-panels presents the distribution for the $h = 1, 5, 10, 20$ horizons of the LPIRFs estimator. From the figure, it is easy to see that the lag-length p selected by the AIC to fit the VAR(p) model in each repetition of the Monte Carlo experiment is almost always $p = 12$. This is also the correct lag-length of the data generating process. In this way the lag-length selection procedure delivers an unfair comparison between the local projection and

[6]The authors computed these statistics in the original code, but those were not reported in the paper.

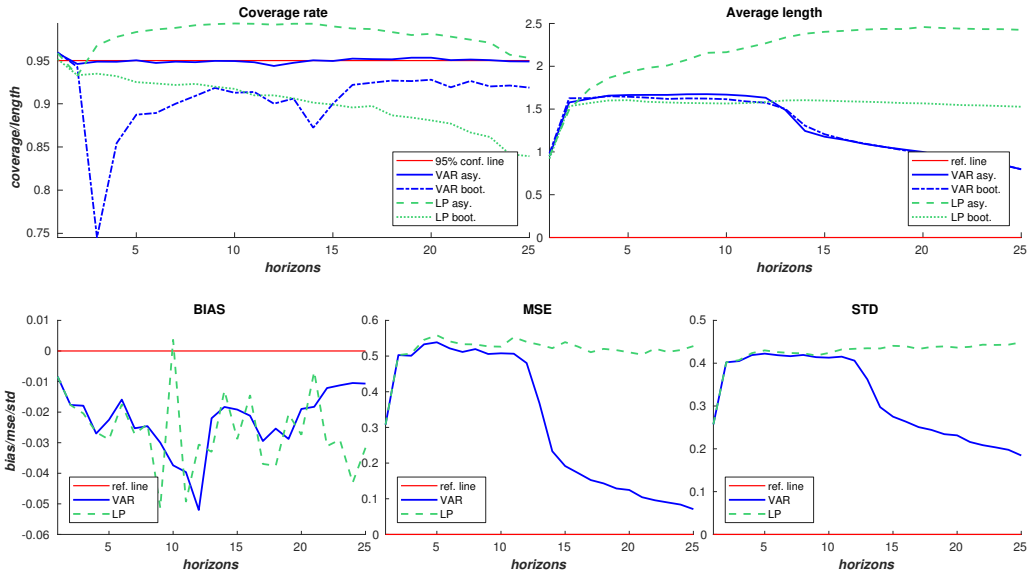

Figure 2: IRFs results from the Monte Carlo experiment. We average all the statistics along all the shocks and all the variables in the system (the number of impulse responses in a $K = 4$ variable VAR is $K^2 = 16$). VAR asy. denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 1998a,b,c). LP asy. denotes the asymptotic interval for LPs. LP boot. refers to the bias-corrected block bootstrap interval for LPs. The AIC selects all lag orders with an upper bound $\overline{p} = 12$. The solid red line acts as a reference line for each statistics.

VAR impulse responses, returning a comparison between a well-specified VAR and a misspecified local projection model. Using a well-specified VAR returns correctly specified IRFs where the only source of distortion arising from small-sample bias. On the contrary, local projection suffers from an augmented form of small-sample bias plus the model misspecification bias. In this setting, it is implausible for any model to outperform the VAR.

There is a second issue, which is not of minor importance. LPIRFs is a very young methodology and still is not clear which is the best way to chose its lag-length. The choice of the AIC follows Ivanov and Kilian (2005), which is a study that gives guidelines for selecting the lag-length criterion to maximize the performance of the estimated VAR IRFs. Moreover, given that the local projection lag-length can be chosen for each projection, a researcher has at least two different alternatives to consider:

- Selecting the lag-length once, and keeping it fixed for all the horizons. This method is the one used in all the empirical papers working with local projections as Auerbach and Gorodnichenko, 2012a, Owyang et al. 2013, Tenreyro and Thwaites 2016, Caldara and Herbst 2016, Ambrogio Cesa-Bianchi, 2016.

- Selecting the lag-length for each projection.

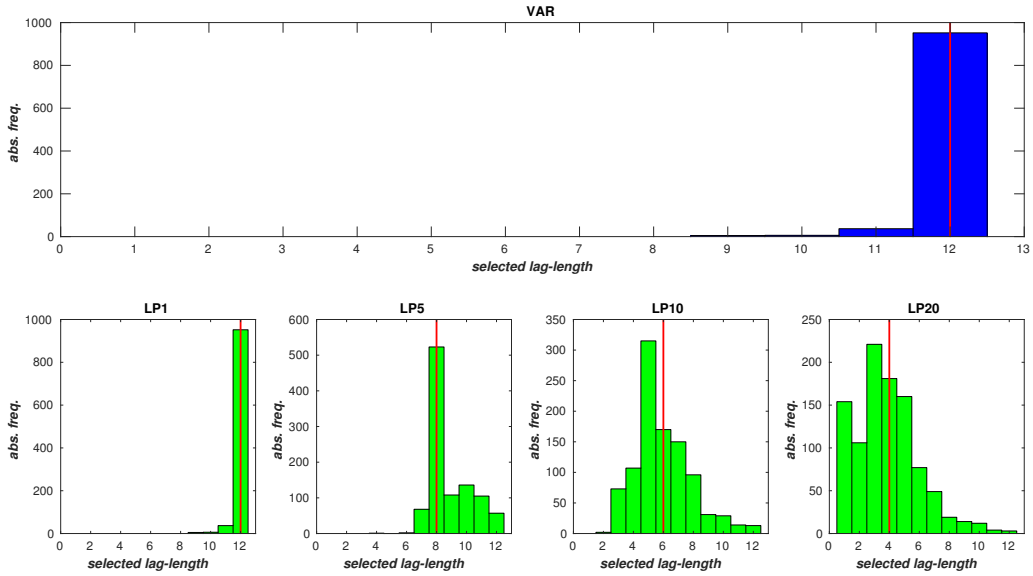

Figure 3: Selected lag-length distribution for VAR and LP IRFs using AIC as in Kilian and Kim. The four lower panels show the distribution for the local projection selected lag-length for the horizons $h = [1, 5, 10, 20]$. The red-solid vertical line shows the median lag-length.

It is easy to notice that at the first horizon, the scenario is identical to the VAR lag-length selection (bottom figure, first panel). However, starting from the second panel, the distribution begins widening, and the mode shifting from right to left[7]. Therefore, as shown in the figure, while the VAR is a well-specified model, given that always select the correct lag-length of the true DGP, the local projection estimator it is only at the first horizon, and get increasingly misspecified as the horizon increases. This situation makes the comparison between the two methodologies highly unfair and can produce misleading results.

Secondly, a second source of unfairness in the comparison can be identified. The use of the Akaike information criterion for both the methodologies can be taken as a fair choice. However, some features of the Monte Carlo experiment, as the sample size T and the maximum lag-length $\overline{p}$ equal to the correct lag-length of the VAR DGP helps the information criterion to select the exact lag-length only for the VAR model. In fact, given that the AIC tends to penalize less over-parametrized model, for VAR this means that specifications with higher lag-length will usually be preferred (overestimation). This is opposite to what, for example, the BIC criterion does, often resulting in under-parametrized models. For clarity, in Table 1 we report some Monte Carlo example changing $\overline{p}$ and T and using the most popular information

[7]This phenomenon is likely due to the increasing order of the MA part of the local projection IRFs.

criteria to select the lag-length for the same DGP used in the paper[8]. The table shows how many time the information criteria select some specific lag-length in percentage points. When the sample size is

Table 1: Lag-length selected in a Monte Carlo exercise.

Horizon	T = 100				T = 200				T = 400			
	AIC	BIC	HQC	AICC	AIC	BIC	HQC	AICC	AIC	BIC	HQC	AICC
1	13.7	99.4	83.1	62.6	0.2	93.7	31.8	1.6	0	31.6	0	0
2	20	0.6	15.5	33.8	6.4	6.3	54.3	36.7	0	64.8	26.9	0
3	10	0	1.4	3.4	13	0	13.4	46.9	0	3.6	53.8	0.2
4	1.5	0	0	0.2	3	0	0.4	6.5	0	0	3.7	0.1
5	0.8	0	0	0	6.1	0	0.1	5.8	0	0	9.2	0.7
6	0.7	0	0	0	1.9	0	0	1.2	0	0	0.6	0.3
7	0.9	0	0	0	2.2	0	0	0.8	0	0	0.9	0.3
8	1.4	0	0	0	3.6	0	0	0.2	0	0	0.7	1.1
9	2.4	0	0	0	13.1	0	0	0.3	1.2	0	3.4	23.2
10	2.6	0	0	0	7.2	0	0	0	2.1	0	0.2	8.7
11	6.2	0	0	0	8.9	0	0	0	4.5	0	0.1	10.1
12	39.8	0	0	0	34.4	0	0	0	92.2	0	0.5	55.3

Note: the table shows the results from a Monte Carlo exercise which simulates data from the four variable VAR(12) by Christiano et al. (1999) and select the lag-length using AIC, BIC, AICC, and HQC. We repeat the process $M = 1000$ times, and we report the relative frequency of the lag-length selected by each of the procedure in percentage points. We repeat the process for sample of size $T = 100, 200, 400$. The maximum lag-length allowed to be selected in the procedure is $\overline{p} = 12$.

400 and $\overline{p} = 12$, the AIC chooses the correct lag-length more than 90% of the cases. However, as the sample size decreases to $T = 200$ and $T = 100$, the correct lag-length is selected less than 40% of the time. Using more parsimonious criteria result in all but one case to pick the wrong lag-length, ending in a misspecified VAR model. As expected the BIC chooses a much more parsimonious model while HQC has a much broader distribution. The corrected Akaike criterion has a distribution which is very close to the standard AIC. Reducing or increasing the maximum lag-length to test $\overline{p}$, always results in a worse scenario. Selecting $\overline{p} < 12$ leads never to choose the correct lag-length for any criteria while selecting $\overline{p} > 12$ leads to a shift in the probability towards higher order model[9].

Having verified that dealing with a well-specified model is a rare case, even when the true DGP is a VAR, a way to have a more fair comparison between the two procedure is to consider the case in which both methodologies deal with the lag-length misspecification and shift away from the true DGP in a controlled manner. The misspecified case is also the general case to consider in the empirical analysis when the true DGP is always unknown. Therefore, to shed some lights on the performance of the VAR impulse responses against the local projection estimator in a truly misspecified context, I rerun the Kilian and Kim's Monte Carlo simulation selecting the lag-length with a more parsimonious information

[8] Those are the Akaike (AIC), the Schwarz (BIC), the Hurvic-Tsai (AICC) and the Hannan-Quinn (HQC).

[9] We show in the appendix the cases in which $\overline{p} = 6$ and $\overline{p} = 18$.

criterion. This twist allows inducing some lag-length misspecification in both the models.

4.1 Simulation performing BIC model selection

In this section, I rerun the previous experiment, choosing the lag-length for both methodologies using the BIC information criterion. Due to a stronger penalization, the BIC criterion ensures us to select a model with a lag-length far from the true DGP lag-length. Figure 4 shows the lag-length distribution of the 1000 repetition of the Monte Carlo simulation for both the VAR and the local projection procedure. The

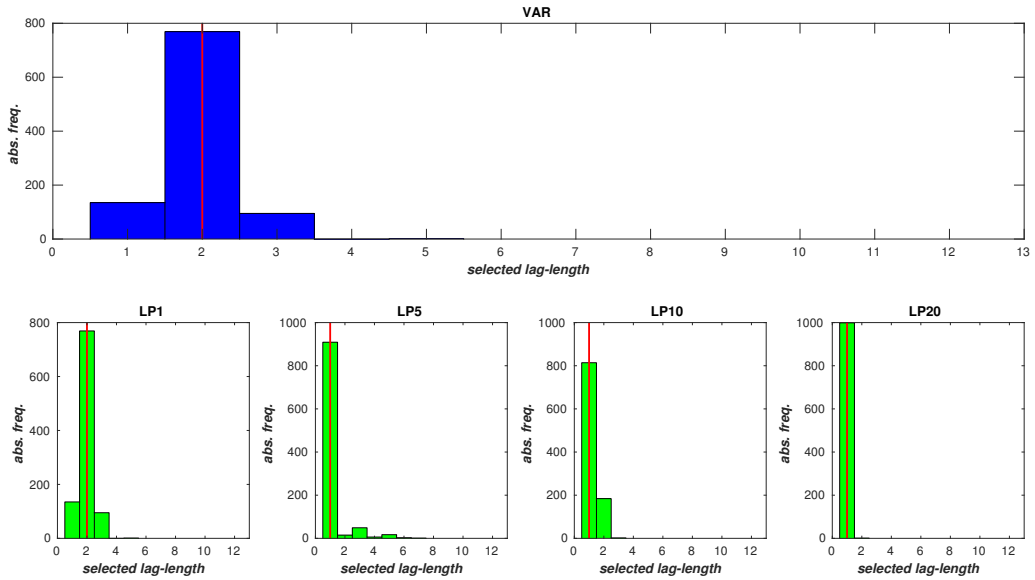

Figure 4: Selected lag-length distribution for VAR and LP IRFs using BIC criterion. The four lower panels show the distribution for the local projection selected lag-length for the horizons $h = [1, 5, 10, 20]$. The red-solid vertical line shows the median lag-length.

figure displays a very different scenario, with the distribution of the selected lag-length entirely shifted to the left. Now, the mode of the VAR lag-length distribution is $p = 2$, symmetrically the local projection distribution presented in the first-left panel in the second row is exactly equal to the VAR distribution, while as the horizon increases, the mass tends to shrink on a very parsimonious model with $p = 1$. Figure 5 reports the average result for all K^2 IRFs generated from the model. It is easy to notice from the first panel that, overall, the two methodologies are very sensible to lag-length misspecification, with the coverage rate decreasing to 65%-85% for local projection and till 40% for the VAR impulse responses. Also, the scale of all the other criteria is much larger than in the previous case. However, considering the relative performance, the situation is now very different. In fact, the true value of the IRFs is on

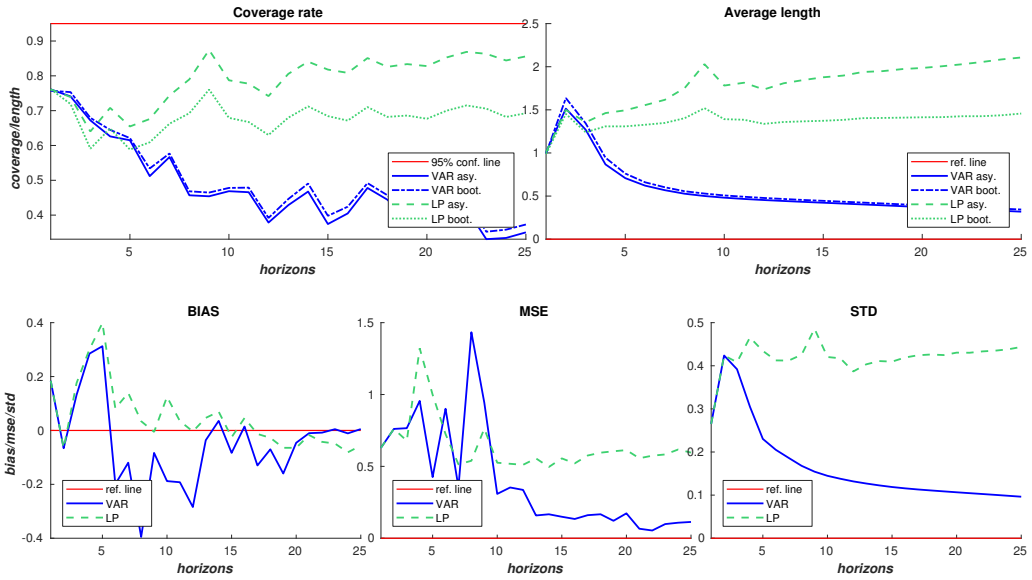

Figure 5: Average of statistics along all the shocks and all the variables in the system (the number of impulse responses in a $K = 4$ variable VAR is $K^2 = 16$). VAR asy.denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 1998a,b,c). LP asy. denotes the asymptotic interval for LPs. LP boot. refers to the bias-corrected block bootstrap interval for LPs. The BIC selects all lag orders with an upper bound $\overline{p} = 12$. The solid red line acts as a reference line for each statistics.

average outside the VAR confidence bands more than 50% of the times. This ratio is much lower for LP (around 20%). However, VAR IRFs have, on average, shorter average length, which could explain why the confidence bands miss so many times the target (this is due to lower STD). Nevertheless, VAR IRFs seems more biased on average, but better in MSE terms, at least for further horizons. This fact is due to shorter STD and in turn, to the evidence that LP is a data consuming approach. Thus, increasing the IRFs horizons, we are reducing the sample size available for the estimation. In any case, even if results are not conclusive, they shed some light on the possibility to use LP IRFs for empirical application in place of the VAR IRFs and inevitably call for further investigations.

Motivated by these findings, I rerun the Monte Carlo experiment selecting the local projection lag-length once for all the projection. In this way, the distribution of LP lag-length chosen is equal to the VAR lag-length distribution. This case is an interesting one to consider because many empirical studies are using this twist without any theoretical nor empirical basis. However, whether this is better than selecting the lag-length for each projection is a practical question, and in this paper, I test whether this could be the case.

4.2 Simulation with fixed lag-length

Motivated by the previous exercise, in this section, I rerun the last experiment, fixing the lag-length for both methodologies. Keeping the lag-length fixed implies that the local projection has always a fixed order autoregressive component, while the order of the moving average part is increasing for each horizon by construction. For example, choosing $p = 1$ at horizon $h = 10$ returns a local projection with an autoregressive component of order 1 and a moving average error component of order 10. I examine the cases in which p range from 1 to 12 ($p \equiv [1, 3, 6, 9, 12]$). I compare the average criteria along the VAR and local projection impulse responses to understand which one performs better. Also, it is interesting to learn whether increasing the model misspecification gives some clear pattern. Figure 6 shows the ECR for different Monte Carlo replication with different models. To understand the performance of the two methodologies, for each horizon h I compare the difference between the average of the ECR distance from the 95% correct acceptance rate.

$$\Delta ECR\,(h) = \frac{1}{K^2}\sum_{j=1}^{K^2}(\left|0.95 - ECR_j^{LP}\right| - \left|0.95 - ECR_j^{VAR}\right|) \tag{16}$$

Assuming symmetric losses by deviating from the 0.95 threshold, when $\Delta ECR\,(h) > 0$, the VAR methodology has more coverage than local projection, on the contrary, when $\Delta ECR\,(h) < 0$, the reverse is true. When $\Delta ECR\,(h) = 0$ the two methodology has the same ECR. As a relative measure, $\Delta ECR\,(h)$ gives information only on the relative performance of the two procedure. Figure 6 shows the $\Delta ECR\,(h)$ for different Monte Carlo procedure with fixed lag-length $p \equiv [1, 3, 6, 9, 12]$ and each horizon h. The charts show a clear average advantage in using local projection against VAR impulse responses. The coverage rate increase by a maximum of 60% for the local projection estimator with respect to the VAR impulse responses. Three striking features emerge; first, the relative performance in the local projection increases as the horizon increases. The reason is linked to the increasing wideness of the local projection confidence bands which increases the local projection coverage rate; also, it is connected to the deteriorating performance of the var coverage rate as the horizon increases. Second, as the lag-length distance from the true lag-length $p = 12$ increases, the LP coverage rate is higher relative to the VAR ECR. Third, the asymptotic ECR difference for the VAR confidence bands seems to perform better than the bootstrap counterpart. This feature is evident from the $p = 12$ line, which is entirely above zero. Also when $p = 9$ the relative performance is mixed.

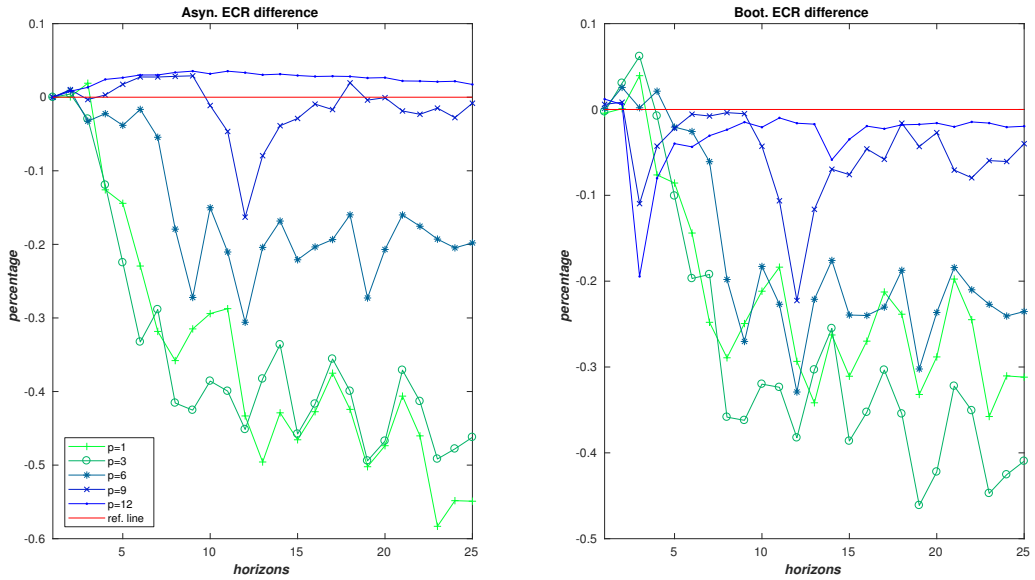

Figure 6: The figure shows the ECR for different Monte Carlo simulations using asymptotic and bootstrap confidence interval and selecting a different lag-length.

As showed in the previous results, the main reason why the coverage rate of local projection is higher than VAR impulse responses is mainly linked to the wideness of the confidence bands. Figure 7 shows this result, reporting in the second and third panel the average standard deviation ratio between the LP and VAR IRFs. The ratio between the standard deviation of the two procedure is substantial. The average STD of LP confidence bands ranges from one to seven times the STD of the VAR intervals. A clear pattern emerges from the plot; First, as shorter the lag-length (meaning as more misspecified the model), as wider the LP confidence bands. Secondly, as higher the horizon as bigger the ratio between the two procedure. Such a significant difference entirely enters in the computation of the MSE ratio between the two methodologies. In fact, the MSE ratio rapidly increases up to fourteen. Then, large confidence bands are clear symptoms of some misspecification occurring. Of course, the first best would be having a unbias estimator with thinly confidence bands. However, as a second best, one might argue that the efficiency of an estimator is a second-order property, and having an unbiased estimator with larger confidence bands would be better than having a bias one with shorter intervals. To dig in that, we report in Figure 7 also the average absolute bias difference between the two procedures computed as follows:

$$\Delta BIAS\,(h) = \frac{1}{K^2}\sum_{j=1}^{K^2}(\left|BIAS_j^{LP}\right| - \left|BIAS_j^{VAR}\right|) \tag{17}$$

Equation 17 returns a positive $\Delta BIAS$ when $\left|BIAS^{LP}(h)\right| > \left|BIAS^{VAR}(h)\right|$, and a negative one when the opposite is true. From the first panel of Figure 7, it is clear that on average local projection

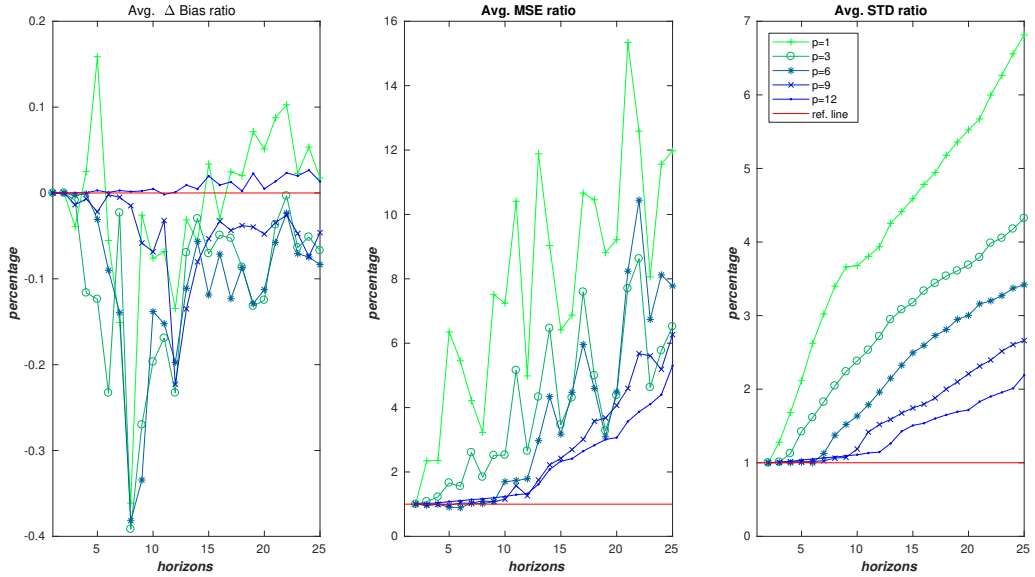

Figure 7: The figure shows the average BIAS, MSE, and STD for different Monte Carlo. The chart reports simulations results using asymptotic and bootstrap confidence intervals by selecting different lag-lengths.

is fairly more precise than the VAR counterpart. This result is in contrast with Kilian and Kim findings (see note 17). However, differently from the STD and MSE counterpart, for BIAS is not possible to find an evident ranking due to higher lag-length and horizons.

The central message that emerges from the last two pictures is reasonably clear. As the model is misspecified, the VAR impulse responses return a vector of points which are far from the true ones. Nevertheless, confidence bands are narrow around these points unable to signaling any uncertainty. By contrast, the local projection estimator returns a point which is closer to the correct one together with large confidence bands. The wide intervals, from one side signal the uncertainty in the estimation, from the other, with high probability, contain the true parameter.

5 Discussion

The Monte Carlo experiment shows some charming result about the coverage ability of the VAR and the LP methodology. In particular, it legitimates the use of local projection also when the true DGP is a

VAR. Of course, in applied works, a researcher never knows the nature of the DGP, and for this reason, more flexible alternatives are usually preferred to more rigid models. The local projection estimator is not an exception. Moreover, this methodology is still not exploited at full capacity, and in this section, I summarize some of the results that are still missing in the literature to improve on the LP estimation procedure.

First, as highlighted in the previous section, the local projection estimator has by construction an MA(h) error term, which builds up as the horizon h increases. Jordà, in the original paper suggests estimating the variance-covariance matrix using the Newey and West (1987) and Andrews (1991) *heteroskedasticity and autocorrelation consistent estimator* (HAC). However, in a recent paper by Stock and Watson (2017), there is a clear call for abandoning the Newey-West estimator in favor of methods which produce fewer distortions. The author cites a voluminous literature starting from Kiefer et al. (2000) and the following literature. As one of the main issue highlighted by the Monte Carlo experiment is the confidence band wideness of the impulse response functions, exploiting different methodologies for estimating the variance-covariance matrix would be one of the first route to cover. A second possibility, which still goes in this direction, would be to follow a suggestion given by Jordà in its original paper which could improve the estimator efficiency. The idea is simply to include in the projection at step $h+1$ the estimate of the residual at step h. This should reduce the estimation uncertainty and shrink the bands. However, a study to assess this procedure and possible drawback is needed to confirm this guess, at least in a small sample set.

Secondly, from the Monte Carlo simulation emerges a crucial issue for both the VAR and the LP estimator: the lag-length selection procedure. In fact, the lag-length influences both the shape and the magnitude of the impulse responses and my simulation shows how different can be the outcome in a small sample setting. To give some more flavor to what happens to VAR IRFs by changing the lag-length, the reader is pointed to the practitioners' guide to best select the lag-length in a VAR model by Ivanov and Kilian (2005). The first figure of the paper shows how crucial is the correct choice of the lag-length for precisely estimate the impulse responses. However, the paper conclusion leads to further complications. In particular, they argue that the lag-length is not per se important, but it is in the way it affects the precision of the IRFs. For example, by choosing as a performance metric the MSE, as it is a common practice, the accuracy of the IRFs become a nonlinear function of the estimated bias and variance. It's important to notice that this two statistics may move in opposite directions. A practical case

is when a researcher underestimates the true lag-length of the VAR model. In that case, the result will be a biased impulse response but with variance smaller than the true one. Such reduction may more than offset the bias, leading to more precise IRFs in MSE sense. The study concludes suggesting the use of different criteria for data-set with different frequencies and sample size T. In particular, it recommends AIC for monthly data, HQC for quarterly data with $T > 120$ and BIC for quarterly data $T < 120$. These findings also justify why Kilian and Kim, in choosing the lag-length in their paper uses the AIC.

For the local projection it is still an open issue how to select the best lag-length to use in each projection. Given each projection may have different numbers of lags, finding a right criterion can lead to significant improvements in the estimation procedure.

Finally, it is essential to remark a significant result for VAR impulse responses due to Kilian (2001) that can be translated into the LP framework. In the paper, the author shows that underestimating and overestimating the lag-length does not produce symmetric errors in a small sample. Due to the mapping between VAR and VMA coefficients, underestimating the lag-length of the impulse response functions implies a cut in the polynomial order of the IRFs which translate in a reduction of the curvature in the shape of the impulse responses. On the contrary, overestimating the lag-length adds curvature to the impulse responses. This valuable insight has some practical consequences, and translate in a preference for more parametrized VAR models when the research focus on impulse responses. In turn, this leads to avoid information criteria for the selection of the lag length which penalize too much for the number of parameters as the Schwarz information criterion. However, for the local projection, the same reasoning is not so straightforward given that a researcher can estimate the coefficient at each horizon. In this case, the autoregressive part may support mixed lag-length. Also, the VMA(h) part will increase with the horizons resulting in more complex models.

6 Conclusion

Starting from Kilian and Kim (2011), I analyze the performance of local projection against vector autoregressive impulse response functions. I replicate the Monte Carlo exercise performed by the two authors highlighting some arguments that lead to conclude in favor of VAR IRFs strongly. In particular, I show that the lag-length selected in the experiment returns a comparison between a well-specified VAR and a

misspecified local projection model. Using a well-specified VAR returns correctly specified IRFs where the only source of distortion arising from small-sample bias. On the contrary, local projection suffers from an augmented form of small-sample bias plus the model misspecification bias. This features in the construction of the Monte Carlo experiment leave some space for further investigation to understand whether the result can be experiment dependent.

In the analysis, I perform an analogous Monte Carlo inducing some misspecification through the modification of the model selection criteria. Therefore, I compare VAR and LP IRFs in a context in which both models are misspecified. The results are very different and are no longer solidly in favor of the VAR IRFs. Due to the importance of IRFs, especially in empirical macroeconomic studies, and given the ambiguity of the result, further studies are necessary to determine which methodology dominates and in which context a researcher should use it. Therefore, motivated by these findings, I rerun a third Monte Carlo experiment fixing the lag-length for each equation of the local projection procedure. The results show that as the model is misspecified, the VAR impulse responses return a vector of points which are far from the true ones and with narrow confidence bands. By contrast, the local projection estimator returns a point which is closer to the correct one together with large confidence bands. Also, we have some evidence that the performance of LP IRFs improves considerably by selecting the lag-length once for all the projection. Of course, in the light of the discussion presented in the paper, this results deserves further testings.

References

Ambrogio Cesa-Bianchi, Gregory Thwaites, A. V. (2016). Monetary Policy Transmission in an Open Economy: New Data and Evidence from the United Kingdom. *SSRN Electronic Journal*.

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817–858.

Auerbach, A. J. and Gorodnichenko, Y. (2012a). Fiscal Multipliers in Recession and Expansion. In Alesina, A. and Giavazzi, F., editors, *Fiscal Policy after the Financial Crisis*, pages 63–98. University of Chicago Press.

Auerbach, A. J. and Gorodnichenko, Y. (2012b). Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy*, 4(2):1–27.

Auerbach, A. J. and Gorodnichenko, Y. (2013). Output spillovers from fiscal policy. *The American Economic Review*, 103(3):141–146.

Auerbach, A. J. and Gorodnichenko, Y. (2016). Effects of fiscal shocks in a globalized world. *IMF Economic Review*, 64(1):177–215.

Barnichon, R. and Brownlees, C. T. (2017). Impulse Response Estimation by Smooth Local Projections. *SSRN Electronic Journal*.

Barnichon, R. and Matthes, C. (2017a). Functional Approximations of Impulse Responses (FAIR): New Insights into the Asymmetric Effects of Monetary Policy. *Manuscript, Federal Reserve Bank of San Francisco*.

Barnichon, R. and Matthes, C. (2017b). Understanding the size of the government spending multiplier: It is in the sign. *SSRN Electronic Journal*.

Barnichon, R., Matthes, C., and Ziegenbein, A. (2016). Theory Ahead of Measurement? Assessing the Nonlinear Effects of Financial Market Disruptions. *Federal Reserve Bank of Richmond Working Paper, No. 16-15*.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98.

Caldara, D. and Herbst, E. (2016). Monetary Policy Real Activity, and Credit Spreads: Evidence from Bayesian Proxy SVARs. *Finance and Economics Discussion Series*, 2016(049):1–51.

Chan, J. C. and Eisenstat, E. (2017). Efficient estimation of bayesian varmas with time-varying coefficients. *Journal of Applied Econometrics*, 32(7):1277–1297.

Christiano, L., Eichenbaum, M., and Evans, C. (1999). Monetary Policy Shocks: What Have We Learned and to What End? In (ed.), J. B. T. . M. W., editor, *Hanbook of Macroeconomics*, volume 1, chapter 2, pages 65–148. Elsevier, 1 edition.

Cooley, T. F. and Dwyer, M. (1998). Business cycle analysis without much theory A look at structural VARs. *Journal of Econometrics*, 83(1-2):57–88.

Hall, A. R., Inoue, A., Nason, J. M., and Rossi, B. (2012). Information criteria for impulse response function matching estimation of DSGE models. *Journal of Econometrics*, 170(2):499–518.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton.

Hamilton, J. D. (2011). Nonlinearities and the Macroeconomic Effect of Oil Price. *Macroeconomic Dynamics*, 15(S3):364–378.

Haug, A. A. and Smith, C. (2011). Local linear impulse responses for a small open economy. *Oxford Bulletin of Economics and Statistics*, 74(3):470–492.

Ivanov, V. and Kilian, L. (2005). A Practitioner's Guide to Lag Order Selection For VAR Impulse Response Analysis. *Studies in Nonlinear Dynamics & Econometrics*, 9:article 2.

Jordà, O. (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, 95(1):161–182.

Jordà, O. (2009). Simultaneous Confidence Regions for Impulse Responses. *Review of Economics and Statistics*, 91(3):629–647.

Keynes, J. M. (1936). *The General Theory of Employment Interest and Money*. Harcourt, Brace and Company, New York.

Kiefer, N. M., Vogelsang, T. J., and Bunzel, H. (2000). Simple Robust Testing of Regression Hypotheses. *Econometrica*, 68(3):695–714.

Kilian, L. (1998a). Accounting for Lag Order Uncertainty in Autoregressions: the Endogenous Lag Order Bootstrap Algorithm. *Journal of Time Series Analysis*, 19(5):531–548.

Kilian, L. (1998b). Confidence intervals for impulse responses under departures from normality. *Econometric Reviews*, 17(1):1–29.

Kilian, L. (1998c). Small-sample Confidence Intervals for Impulse Response Functions. *Review of Economics and Statistics*, 80(2):218–230.

Kilian, L. (2001). Impulse response analysis in vector autoregressions with unknown lag order. *Journal of Forecasting*, 20(3):161–179.

Kilian, L. and Kim, Y. J. (2011). How Reliable Are Local Projection Estimators of Impulse Responses? *Review of Economics and Statistics*, 93(4):1460–1466.

Lutkepohl, H. (1990). Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models. *The Review of Economics and Statistics*, 72(1):116–125.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526.

Meier, A. (2005). How Big is the Bias in Estimated Impulse Responses? A Horse Race between VAR and Local Projection Methods. Unpublished.

Miranda-Agrippino, S. and Ricco, G. (2017). The Transmission of Monetary Policy Shocks. *Bank of England Working Paper, No. 657*.

Newey, W. K. and West, K. D. (1987). A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

Owyang, M. T., Ramey, V. A., and Zubairy, S. (2013). Are Government Spending Multipliers Greater during Periods of Slack? Evidence from Twentieth-Century Historical Data. *American Economic Review*, 103(3):129–134.

Pope, A. L. (1990). Biases of Estimators in Multivariate Non-Gaussian Autoregressions. *Journal of Time Series Analysis*, 11(3):249–258.

Ramey, V. (2016). Macroeconomic Shocks and Their Propagation. In Taylor, J. B. and Uhlig, H., editors, *Handbook of Macroeconomics*, volume 2A, chapter 3, pages 71–162. Elsevier.

Ramey, V. A. and Zubairy, S. (2017). Government spending multipliers in good times and in bad: evidence from us historical data. Fortcoming *Journal of Political Economy*.

Ronayne, D. (2011). Which Impulse Response Functions? *Warwick Economic Research Paper, No. 971*.

Runkle, D. E. (1987). Vector Autoregressions and Reality. *Journal of Business & Economic Statistics*, 5(4):437.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1):1–48.

Smets, F. and Wouters, R. (2003). An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area. *Journal of the European Economic Association*, 1(5):1123–1175.

Stock, J. and Watson, M. (1999). A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series. *NBER Working Paper 6607*.

Stock, J. H. and Watson, M. W. (2017). Twenty Years of Time Series Econometrics in Ten Pictures. *Journal of Economic Perspectives*, 31(2):59–86.

Swanson, E. (2017). Measuring the Effects of Federal Reserve Forward Guidance and Asset Purchases on Financial Markets. *NBER Working Paper, No. 23311*.

Tenreyro, S. and Thwaites, G. (2016). Pushing on a String: US Monetary Policy Is Less Powerful in Recessions. *American Economic Journal: Macroeconomics*, 8(4):43–74.

Wallis, K. F. (1977). Multiple Time Series Analysis and the Final Form of Econometric Models. *Econometrica*, 45(6):1481–1497.

Zellner, A. and Palm, F. C. (1974). Time series analysis and simultaneous equation econometric models. *Journal of Econometrics*, 2(1):17–54.

A Additional Figure

A.1 Simulation performing model selection with AIC and BIC

Figure 8 to 10 show the results of the Monte Carlo experiment. In particular, they show the IRFs to a monetary policy shock on the CFNAI index of US real activity, US CPI inflation and US real commodity price inflation. The figures present the results for the asymptotic delta method for VAR impulse responses (Lutkepohl 1990), the asymptotic interval for local projection and bias-corrected bootstrap (Kilian 1998a,b,c) for both methodologies. The AIC selects all lag orders with an upper bound $\overline{p} = 12$. The solid red line acts as a reference line for each statistics. I extend the figures presented in the original paper including also the estimated *BIAS*, *mean-squared-error* (MSE) and *standard deviation* (STD) of the impulse responses as described in Equation 12 to 14. The figures confirm the findings we presented in Section 4 for the average criteria for the K^2 impulse responses. In particular, it is evident that as the lag-length is selected using the AIC criteria (and thus only the VAR resembles the DGP), the VAR IRFs are superior regarding ECR and AL. However, as the lag-length is selected in a way that both the models misspecify the DGP, the LP impulse response functions have more ECR and less BIAS.

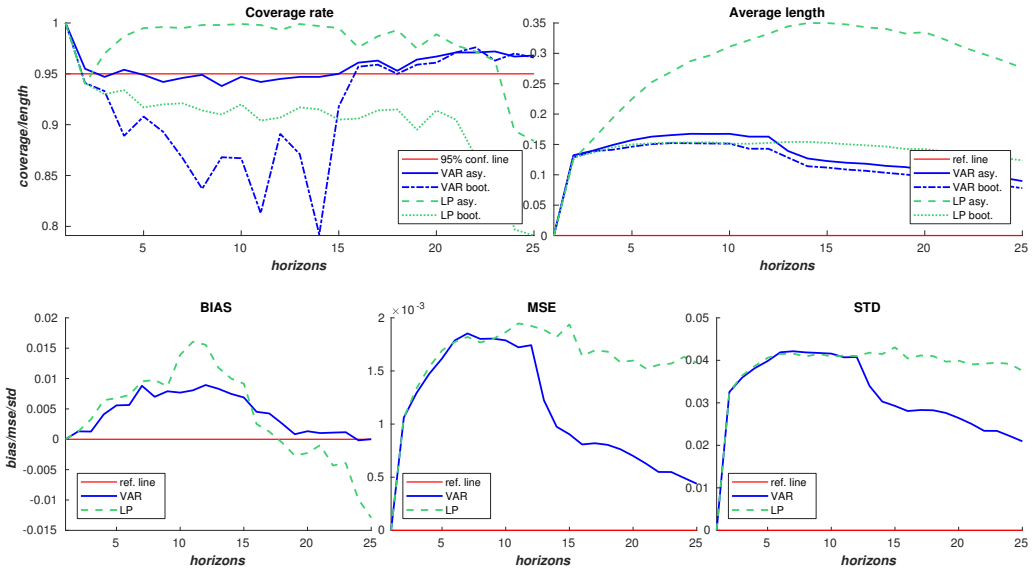

Figure 8: IRFs to a monetary policy shock on CFNAI index of US real activity. VAR asy. denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 1998a,b,c). LP asy. denotes the asymptotic interval for LPs. LP boot. refers to the bias-corrected block bootstrap interval for LPs. The AIC selects all lag orders with an upper bound $\overline{p} = 12$. The solid red line acts as a reference line for each statistics.

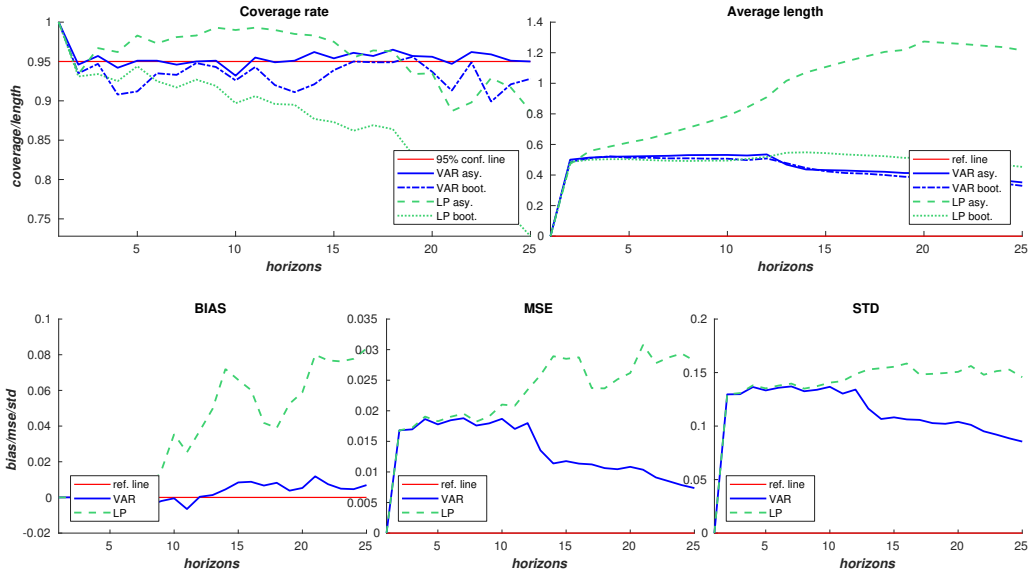

Figure 9: Note as in Figure 8. The reference variable is now US CPI inflation.

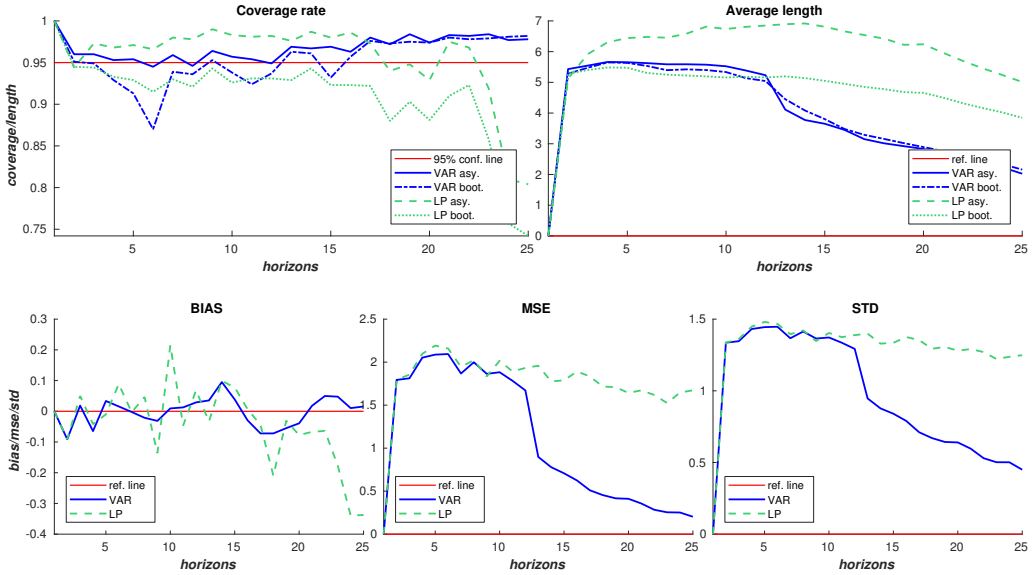

Figure 10: Note as in Figure 8. The reference variable is now US commodity price inflation.

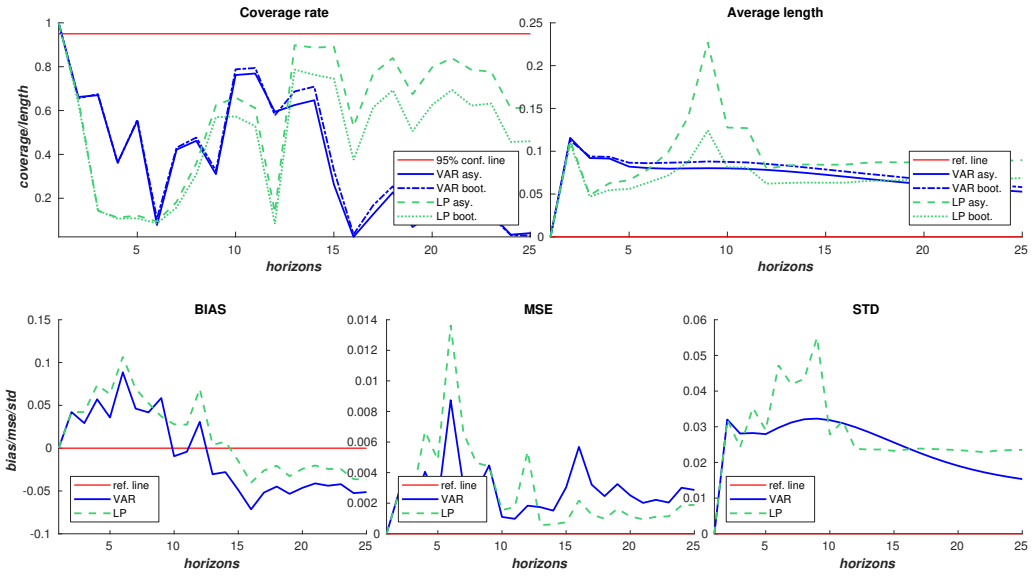

Figure 11: IRFs to a monetary policy shock on CFNAI index of US real activity. VAR asy. denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 1998a,b,c). LP asy. denotes the asymptotic interval for LPs. LP boot. refers to the bias-corrected block bootstrap interval for LPs. The BIC selects all lag orders with an upper bound $\overline{p} = 12$. The solid red line acts as a reference line for each statistics.

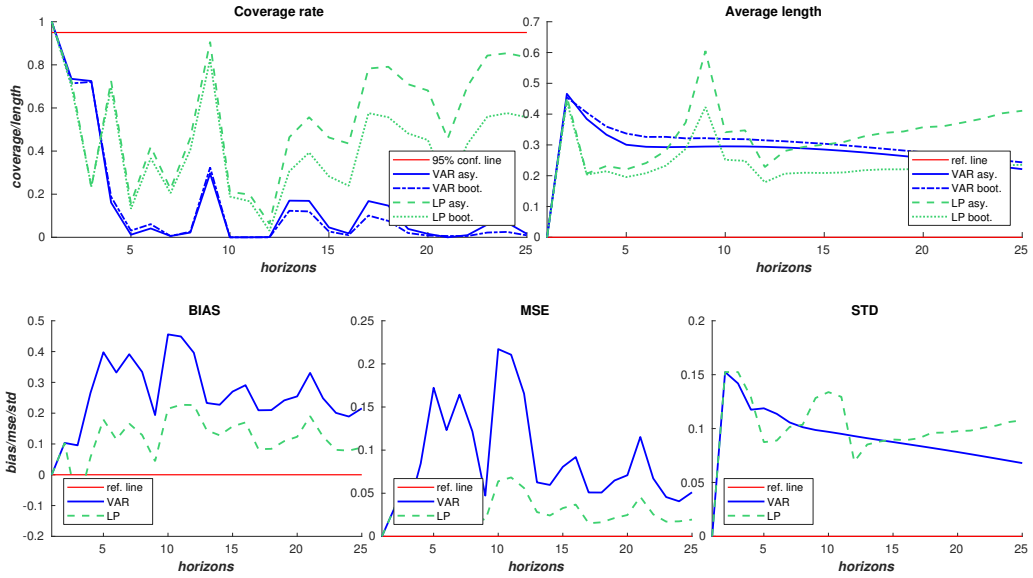

Figure 12: Note as in Figure 11. The reference variable is now US CPI inflation

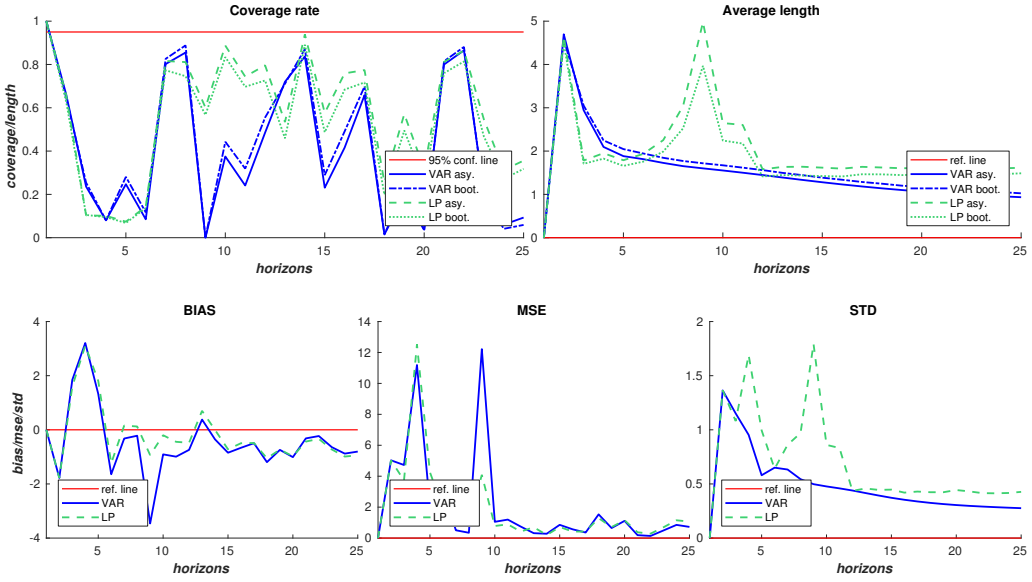

Figure 13: Note as in Figure 11. The reference variable is now US commodity price inflation.

A.2 Additional Figure: simulation performing model selection with HQC and AICC

For completeness, in this section, I also report the result of the simulations using the AICC or HQC. Those are two popular criteria often used by practitioners besides the AIC and BIC. Figure 14 and 15 show the lag-length distribution for the VAR and LP IRFs using the AICC and HQC. For the AICC we have a situation which is very similar to the one described using the AIC, the only difference is that the distribution is less concentrated on the correct lag-length. Using the HQC instead, the results are in between the AIC and BIC, with a distribution which has a broader support. For the VAR case, the HQC has mode $p = 3$ while the AICC has $p = 12$. For the LP distribution again we notice the same features I stressed for the AIC and BIC distribution, meaning that the first plot in the second row resembles the VAR case, while as the projection horizon increases, the mass shifts toward a parsimonious specification.

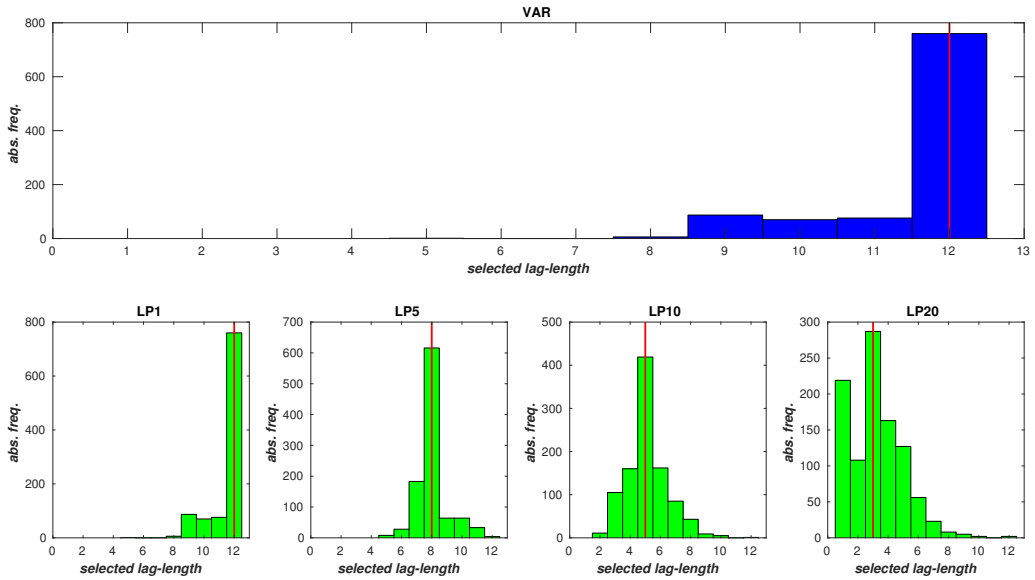

Figure 14: Selected lag-length distribution for VAR and LP IRFs using AICC. The four lower panels show the distribution for the local projection selected lag-length for the horizons $h = [1, 5, 10, 20]$. The red-solid vertical line shows the median lag-length.

Figure 16 and 17 show the average ECR, AL, MSE, STD and BIAS for the AICC and HQC. As expected, due to the similarities between the AIC and AICC, the results are very similar. Instead, the results of the HQC are much closer to those presented for the BIC.

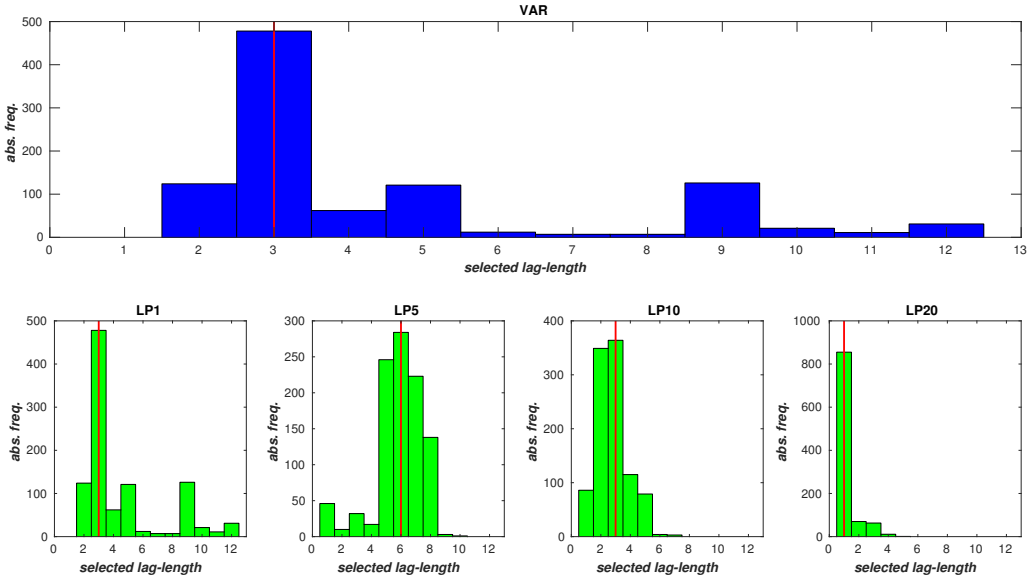

Figure 15: Selected lag-length distribution for VAR and LP IRFs using HQC. The four lower panels show the distribution for the local projection selected lag-length for the horizons $h = [1, 5, 10, 20]$. The red-solid vertical line shows the median lag-length.

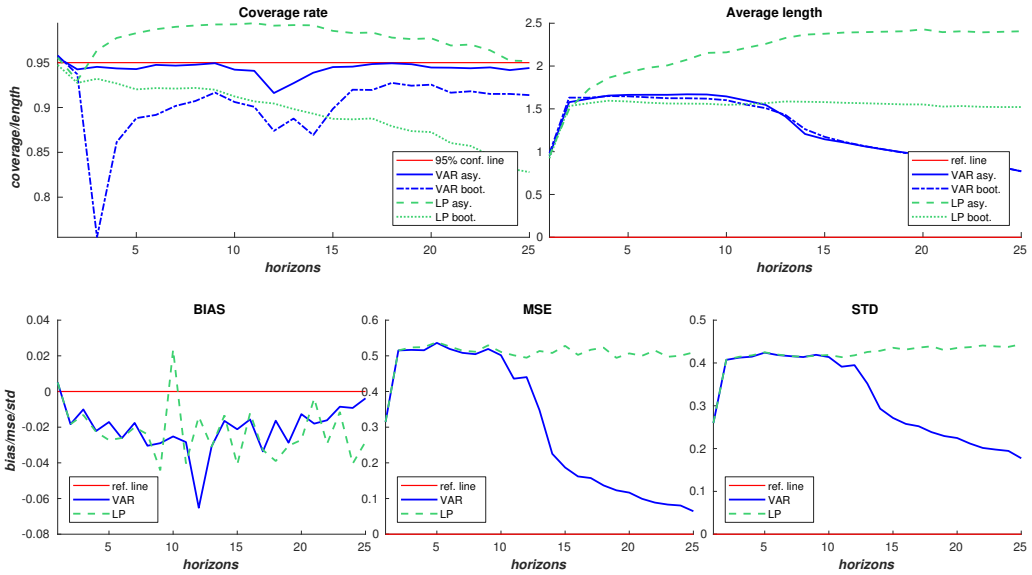

Figure 16: Average of statistics along all the shocks and all the variables in the system (the number of impulse responses in a $K = 4$ variable VAR is $K^2 = 16$). VAR asy.denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 1998a,b,c). LP asy. denotes the asymptotic interval for LPs. LP boot. refers to the bias-corrected block bootstrap interval for LPs. The AICC selects all lag orders with an upper bound $\overline{p} = 12$. The solid red line acts as a reference line for each statistics.

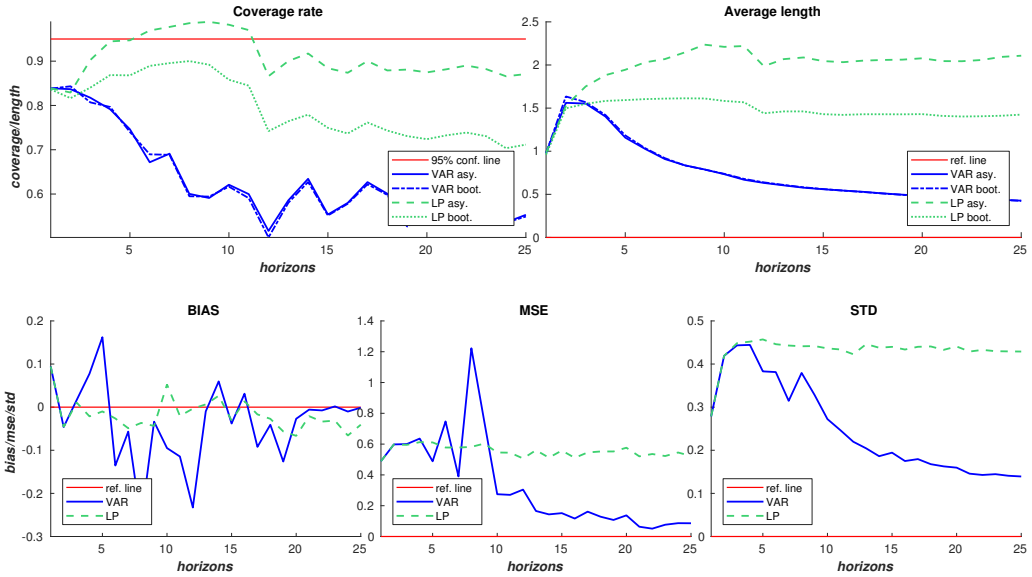

Figure 17: Average of statistics along all the shocks and all the variables in the system (the number of impulse responses in a $K = 4$ variable VAR is $K^2 = 16$). VAR asy.denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 1998a,b,c). LP asy. denotes the asymptotic interval for LPs. LP boot. refers to the bias-corrected block bootstrap interval for LPs. The HQC selects all lag orders with an upper bound $\overline{p} = 12$. The solid red line acts as a reference line for each statistics.

B Additional Table

Table 2: Lag-length selected in a Monte Carlo exercise.

Horizon	T = 100				T = 200				T = 400			
	AIC	BIC	HQC	AICC	AIC	BIC	HQC	AICC	AIC	BIC	HQC	AICC
1	19.5	99	79.3	55.9	0.2	94.7	28.9	1.7	0	26.2	0	0
2	38.2	1	18.8	38.7	14.8	5.3	55.8	32.6	0.1	68.4	24.3	0.2
3	24.2	0	1.9	5.4	34.7	0	14.7	48.3	2.4	5.4	59.2	6.9
4	6.5	0	0	0	9.6	0	0.4	9.1	2	0	5.7	3.6
5	7.5	0	0	0	22.2	0	0.2	6.5	19.7	0	9.6	31.8
6	4.1	0	0	0	18.5	0	0	1.8	75.8	0	1.2	57.5

Note: the table shows the results from a Monte Carlo exercise which simulates data from the four variable VAR(12) by Christiano et al. (1999) and select the lag-length using AIC, BIC, AICC, and HQC. We repeat the process $M = 1000$ times, and we report the relative frequency of the lag-length selected by each of the procedure in percentage points. We repeat the process for sample of size $T = 100, 200, 400$. The maximum lag-length allowed to be selected in the procedure is $\overline{p} = 6$.

Table 3: Lag-length selected in a Monte Carlo exercise.

Horizon	T = 100				T = 200				T = 400			
	AIC	BIC	HQC	AICC	AIC	BIC	HQC	AICC	AIC	BIC	HQC	AICC
1	0	99.4	15	68.5	0.3	95.7	32.1	3	0	32.6	0	0
2	0	0.5	2.4	29.6	7.6	4.3	54.5	38	0	64.4	29.3	0.1
3	0	0	0.1	1.8	14.8	0	12.6	44.3	0	3	53.1	0.3
4	0	0	0	0.1	4.1	0	0.6	6.8	0	0	4.1	0.2
5	0	0	0	0	5.5	0	0.2	5.7	0	0	8.2	0.5
6	0	0	0	0	2	0	0	1.1	0	0	0.6	0.2
7	0	0	0	0	3	0	0	0.5	0.1	0	0.5	1.1
8	0	0	0	0	3.3	0	0	0.2	0	0	0	2.1
9	0	0	0	0	12.4	0	0	0.4	2.1	0	3.5	21.1
10	0	0	0	0	5.3	0	0	0	1.3	0	0	9.3
11	0	0	0	0	9.5	0	0	0	5.7	0	0.1	9.1
12	0	0	0	0	23	0	0	0	86.7	0	0.6	56
13	0	0	0	0	4.4	0	0	0	3.3	0	0	0
14	0	0	0	0	1.2	0	0	0	0.8	0	0	0
15	0	0	0	0	1.7	0	0	0	0	0	0	0
16	0	0	0	0	1	0	0	0	0	0	0	0
17	0.1	0	0	0	0.6	0	0	0	0	0	0	0
18	99.9	0.1	82.5	0	0.3	0	0	0	0	0	0	0

Note: the table shows the results from a Monte Carlo exercise which simulates data from the four variable VAR(12) by Christiano et al. (1999) and select the lag-length using AIC, BIC, AICC, and HQC. We repeat the process $M = 1000$ times, and we report the relative frequency of the lag-length selected by each of the procedure in percentage points. We repeat the process for sample of size $T = 100, 200, 400$. The maximum lag-length allowed to be selected in the procedure is $\overline{p} = 18$.