About Local Projection Impulse Response Function Reliability*

Luca Brugnolini

Central Bank of Malta & University of Rome "Tor Vergata"

Abstract

I compare the performance of the vector autoregressive (VAR) model impulse response function estimator with the Jordà (2005) local projection (LP) methodology. In a Monte Carlo experiment, I demonstrate that when the data generating process is a well-specified VAR, the standard impulse response function estimator is the best option. However, when the sample size is small, and the model lag-length is misspecified, I prove that the local projection estimator is a competitive alternative. Finally, I show how to improve the local projection performance by fixing the lag-length at each horizon.

Keywords: VAR, information criteria, lag-length, Monte Carlo

JEL: C32, C52, C53, E52

I am grateful to Giuseppe Ragusa for all his valuable suggestions, and to Lutz Kilian for sharing his code and being available to discuss the paper. Also, I acknowledge Marco Lippi for his dedication, and Oscar Jordà for is availability. Finally, I am in debt with Antonello D'Agostino, Tommaso Proietti, and Roberto Di Mari for their precious suggestions. All errors are mine. The views expressed in this paper are those of the author and do not necessarily reflect those of the Central Bank of Malta or the Eurosystem.

Corresponding author: Luca Brugnolini brugnolinil@centralbankmalta.org, senior economist at the Central Bank of Malta/Research Department, and Ph.D. candidate at The University of Rome "Tor Vergata", Department of Economics and Finance.

1 Introduction

Since Sims (1980) influential paper *Macroeconomics and reality*, vector autoregressive models (VARs) have become pervasive in the empirical economic literature. Among the tools embedded in VAR models, impulse response functions (IRFs) are undoubtedly the most actively used. In particular, macroeconomists rely on IRFs to perform causal inference, to estimate multipliers, and to study the dynamics of the main macroeconomic aggregates in stochastic models. Due to their usefulness, researchers have deeply examined statistical properties of VAR impulse response functions. Such investigations have built a general knowledge about the advantages of the estimation procedure of the VAR impulse responses. For example, a crucial advantage is that, at the first step-ahead, VAR models produce an optimal and robust to misspecification IRF (Stock and Watson, 1999). Of course, also disadvantages are presently well understood. Those are mainly related to the dependency on the *Wold's decomposition theorem*. Provided invertibility, the theorem allows recasting a p-order VAR in an infinite order vector moving average (VMA), and recover the VMA coefficients recursively as a nonlinear function of the autoregressive parameters. Due to recursiveness, standard IRFs suffer from well-known issues. For example, Pope (1990) shows that small-sample bias in the VMA coefficients stems from the bias in the estimation of the autoregressive parameters. Also, due to the nonlinear mapping between VAR and VMA coefficients, the bias increases as the horizon increases.

Recently, Jordà (2005) has introduced a novel methodology to estimate the impulse response functions, labeled *model-free* or *local projection* (LP) estimator. As the name suggests, the estimation employs nonparametric techniques. Also, the estimator is not constrained by the invertibility assumption, which allows the procedure to be computed when the VMA(∞) representation does not exist. Beside this crucial advantage, in the original paper, the author illustrates how the estimator accommodates nonlinearities, such as state and sign dependencies. Additionally, he shows how local projection can outperform a misspecified VAR model in the impulse response estimation. However, this result has been criticized, and Kilian and Kim (2011) – from now on Kilian and Kim – have opened a debate on the reliability of this novel estimator, showing that its coverage ability relative to the VAR IRFs is extremely poor.

Starting from Kilian and Kim's article, in this paper, I provide an explanation to reconcile these conflicting results, and critically assess whether VAR impulse responses are consistently better than local projection IRFs. In particular, the objective of the paper is twofold: first, it reviews the Kilian and

Kim's findings and demonstrates that the results are driven by the lag-length selection criterion. The article proves that the authors' choice deliver an unfair comparison between the local projection and the VAR impulse responses, returning a comparison between a well-specified VAR and a misspecified local projection model. To rebalance the experiment, I induce a controlled form of misspecification via the lag-length selection procedure and show that local projection is able to outperform the VAR IRFs. Secondly, the paper explores different options to select the lag-length in the LP methodology and shows that relative improvements are achieved fixing the lag-length in the projection horizons.

The rest of the paper is organized as follows. Section 2 illustrates the motivations behind the study. Section 3 summarizes the critique made by Kilian and Kim. Section 4 describe the Monte Carlo experiment and section 5 shows the simulation results. Section 6 discusses the findings and highlights some topics as potential research areas to improve the local projection methodology. Finally, section 7 concludes.

2 Motivation

Figure 1 presents meta-data on the Jordà (2005) local projection seminal paper from Google Scholar Citation. The number of citations-per-year acts as proxy-measure for the estabilishment of the methodology. A striking feature embedded in the series is the relationship between the local projection methodology and the applied macroeconomic literature; for example, Haug and Smith (2011) compare LPIRFs with standard IRFs in a small open economy. Hall et al. (2012) use LPIRFs in the estimation of a Dynamic Stochastic General Equilibrium (DSGE) model. Auerbach and Gorodnichenko (2012a,b, 2013, 2016), Owyang et al. (2013) and Ramey and Zubairy (2017) use local projection to estimate state-dependent fiscal multipliers. Hamilton (2011) exploits LPIRFs to trace out the dynamics of an oil shock 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) apply the LP estimator to assess the response of real and financial variables to a monetary policy shock[1]. 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 an established benchmark to compare the properties of their *Functional*

[1] Ramey (2016) extensively describes this novel field in a recent chapter of *The Macroeconomic Handbook*.

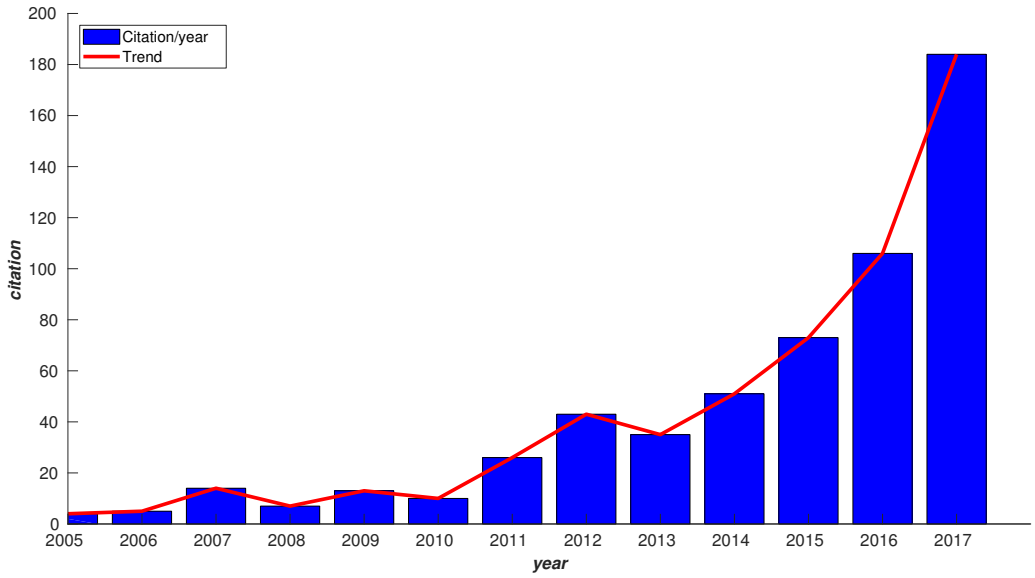

Figure 1: Jordà (2005) local projection estimator. The blue bars show the number of citations per year. The red-solid line highlights the citation trend. Data from Google Scholar Citation.

Approximation of Impulse Responses (FAIR) methodology[2].

Having a novel methodology widely embraced in the scientific community is unquestionably a positive sign. Nevertheless, such a situation requires double carefulness, and an increase in the number of tests to disclose and correct potential drawbacks. Barnichon and Brownlees (2017) take a step in this direction. In the paper, the authors show the benefits of applying a linear B-spline basis function to reduce the variance of the estimated local projection coefficients. Also, Miranda-Agrippino and Ricco (2017) propose a Bayesian estimation procedure to regularise the local projection IRFs. Both papers have introduced such methodologies based on the assumption that VAR impulse responses are more efficient than local projection, and LP is less prone to bias. In this regard, the VAR and LP IRFs should resemble the direct versus iterated forecast procedure (Marcellino et al., 2006). However, this comparison does not explain why Kilian and Kim find that local projection, besides presenting less coverage rate and more average length than VAR IRFs, also displays more bias (see note 17, p. 1463)[3]. Additionally, this study is the only published article which critically compares VAR and LP IRFs[4]. Therefore, according to their

[2]This recent methodology approximates impulse response functions with Gaussian basis functions.

[3]Provided 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 the local projection estimator.

[4]There are also two unpublished articles which compare the performance of the two estimators. However, results are mixed. On one side, Ronayne (2011) tests local projection against VAR impulse responses in an IS-LM framework finding evidence in favor of the first methodology. On the other hand, Meier (2005) tests the two alternatives generating data from the linearised

conclusions, researchers should abandon local projection in favor of the VAR methodology.

Against this background, the objective of the paper is to make clarity on this compact literature, starting reviewing the Kilian and Kim's article, and understanding the drivers of these mixing results.

3 Local projection critique

Kilian and Kim compare the VAR and LP impulse response estimators in a small sample setting, selecting the *effective coverage rate* (ECR) and the *average length* (AL) of the impulse response confidence bands as performance criteria.

Equation 1 presents the effective coverage rate. This is a $(H \times 1)$ vector reporting for each horizon $h = 1, \ldots, H$, the proportion of time in which the interval contains the true value of the IRFs in percentage points.

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

Where $m = 1, ..., M$ is the number of repetitions in the Monte Carlo simulation, $IRF_{true}(h)$ is the true impulse response generated from a Data Generating Process (DGP), and $IRF_r^{(m)}(h)$, with $r = \{L, H\}$, is the upper/lower bound of the estimated confidence intervals 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. Therefore $0 \leq ECR(h) \leq 1$. When the effective coverage rate is equal to zero, on average, the estimated confidence bands never contain the true impulse response. On the other hand, when $ECR = 1$, the confidence bands include the true impulse response with probability 1. Depending on the significance level α, as closer the estimated $ECR(h)$ to its complement $(1 - \alpha)$, as precise the estimator. For example, when $\alpha = 0.05$, the proportion of time in which the interval contains the true values of the impulse response function should be on average 95%. Deviations from this value induce over/under coverage of the impulse responses.

Equation 2 describes the average length of the confidence bands. This is a $(H \times 1)$ vector which measures the absolute distance between confidence bands. Shorter average length implies more precise

version of Smets and Wouters (2003) model, finding evidence in favor of the VAR methodology.

estimates.

$$AL(h) = \frac{1}{M}\sum_{m=1}^{M}\left|IRF_H^{(m)}(h) - IRF_L^{(m)}(h)\right|, \quad h = 1, ..., H \tag{2}$$

Both the ECR and the AL are functions of the estimated confidence bands. Therefore, to mimimise the probability of having results depending on the particular confidence band estimator, the autors compare four different possibilities. For VAR IRFs, they select the Lutkepohl (1990) *delta method* and Kilian's *bias-corrected bootstrap* (Kilian, 1998a,b,c)[5]. For local projection, they choose the asymptotic procedure developed by Jordà (Jordà 2009, 2005) and the block-bootstrap developed within their article. The main results of the paper is that VAR IRFs always have more coverage rate and less average length than the LP counterpart.

4 Simulation

In this section, I perform and extend the Monte Carlo simulation presented by Kilian and Kim. In particular, I focus on the application of the VAR(12) data generating process. This is the four-variate model ($K = 4$) used in Christiano et al. (1999) to identify the US monetary policy shock. The variables involved are the CFNAI index of US real activities[6], 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 in levels, while the CPI and the commodity prices in log-differences multiplied by 1200. All the variables are demeaned. Theoretically, the monetary policy shock is the only exogenous source identified in the model, as Christiano et al. (1999) exploit contemporaneous restriction to determine the Taylor rule used by the FED. However, assuming that the impact matrix A_0 is identified, implies that all the shocks in the system are also identified[7]. Accordingly, it is possible to evaluate the ECR and the AL on the K exogenous sources in the system. The Monte Carlo experiment is designed as follows:

1. Fitting a VAR(12) on the data-set;

[5]The bias-corrected bootstrap is a generalization of the nonparametric bootstrap procedure developed by Runkle (1987), and tailored to account for bias and skewness in the impulse responses.

[6]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.

[7]Appendix A presents a review of the VAR and the LP methodology to derive reduced form and structural impulse response functions.

2. Generating simulated data from the estimated model ($T = b + 456$, where $b = 300$ is the burn-in);

3. Selecting the lag-length p via information criteria (IC), and fitting a VAR(p) on the simulated data;

4. Computing VAR impulse response functions;

5. Selecting the LP lag-length p^{lp} via IC – this is a $(H \times 1)$ vector;

6. Computing local projection impulse response functions;

7. Computing 95% confidence bands using the four described procedures[8];

At the end of the $M = 1000$ Monte Carlo repetitions, the effective coverage rate and the average length are computed. The significance level is set to $\alpha = 0.05$[9].

5 Results

5.1 Simulation performing AIC model selection

Figure 2 presents the effective coverage rate and average length computed at the end of the Monte Carlo simulation (top row, first and second panels). In particular, we compute these statistics for each structural IRF by assuming a known impact matrix A_0. This assumption also implies that the K^2 IRFs in the system are identified. Therefore, to summarise the results, the figure displays the averages along the K^2 impulse responses, as described in equation 3[10].

$$\bar{C}(h) = \frac{1}{K^2} \sum_{j=1}^{K^2} C_j, \quad C = \{AL, ECR\} \tag{3}$$

Solid and dash-dotted blue lines present the results for the confidence bands computed with the VAR delta method and the bootstrap procedure. Dashed and dotted green lines display the results for the confidence bands computed with LP asymptotic and bias-corrected bootstrap procedures. In the simulation, the lag-length p and p^{lp} are selected with Akaike (1974) information criterion (AIC) with an upper bound

[8] For the bootstrap and block-bootstrap procedures, I repeat the algorithm 500 times for each Monte Carlo simulation. Using up to 2000 repetitions does not affect the results

[9] Given the complexity of the experiment (especially considering the number of bootstrap repetitions), I implemented the original Kilian and Kim's code in Julia Language (Bezanson et al. 2017). My VAR Julia package is still a work in progress. However, it is available and freely downloadable on my Github webpage.

[10] For completeness, appendix B shows the results for the single IRFs related to the monetary policy shock, as presented in Kilian and Kim's paper. For consistency with figure 2, I also report the IRFs bias, mean squared error, and standard deviation.

$\overline{p} = 12$ as in the Kilian and Kim's paper. The solid red line highlights the reference value for each statistic. Additionally, the figure reports the estimated bias, mean squared error (MSE), and standard deviation (STD) of the impulse responses, as described in equation 4 to 6 (second row, first to third panels)[11].

$$BIAS\,(h) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{IRF}^{(m)}(h) - IRF_{true}(h)\right), \quad h = 1, ..., H \tag{4}$$

$$MSE\,(h) = \frac{1}{M}\sum_{m=1}^{M}\left(IRF_{true}(h) - \hat{IRF}^{(m)}(h)\right)^2, \quad h = 1, ..., H \tag{5}$$

$$STD\,(h) = \sqrt{MSE(h) - BIAS(h)^2}, \quad h = 1, ..., H \tag{6}$$

Figure 2 displays significative advantages in using the VAR IRF estimator. For horizons $h = 1, \ldots, 25$, VAR IRFs present higher effective coverage rate and lower average length. Additionally, the procedure shows a lower MSE and STD. The results for the bias are mixed. However, the LP estimator does not present any visible advantage. The overall assessment is extremely negative for the local projection estimator, and the results highlight characteristics which should discourage the use of such methodology. However, this result is particularly difficult to reconcile with the Jordà (2005) original findings, which

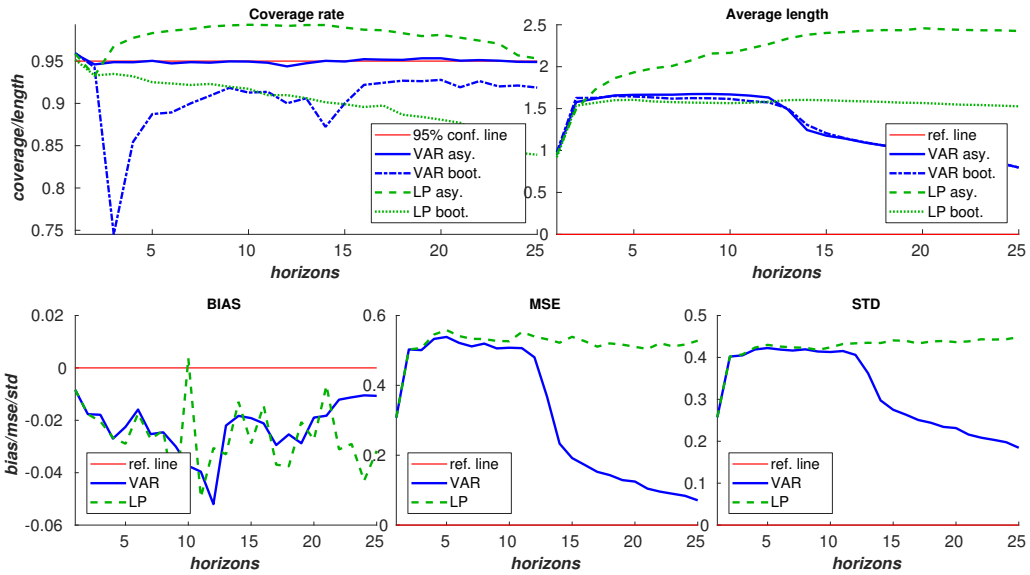

Figure 2: IRFs results from the Monte Carlo experiment. We average all the statistics along all the shocks and all the variables in the system (the number of impulse responses in a $K = 4$ variable VAR is $K^2 = 16$). VAR asy. denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). 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.

[11] Kilian and Kim compute these statistics in the original code, but those were not reported in the paper. In the figures, we report the average bias, mean square error, and standard deviation for the K^2 impulse responses, computed as in equation 3.

highlight the robustness of the local projection estimator. Therefore, to deepen this controversial result, I analyze the Kilian and Kim's and Jordà's Monte Carlo experiments, focusing on their differences. In particular, I examine the lag-length selection procedure. The reason behind this choice is related to the different information criteria selected by Jordà (2005) and Kiliam and Kim in their articles. Jordà (2005) uses the Schwarz (1978) information criterion (BIC), while Kilian and Kim employ the AIC.

Figure 3 is the starting point of the analysis. It presents the distributions of the lag-length p and p^{lp} using the AIC in the $M = 1000$ repetitions of the Monte Carlo algorithm. This is the information criterion employed in the Kilian and Kim's experiment. The upper panel (blue bars) shows the distribution of the VAR lag-length, while the lower panels (green bars) displays the distributions of the local projection lag-length for $h = [1, 5, 10, 20]$. The figure shows some clear patterns. First, the lag-length p

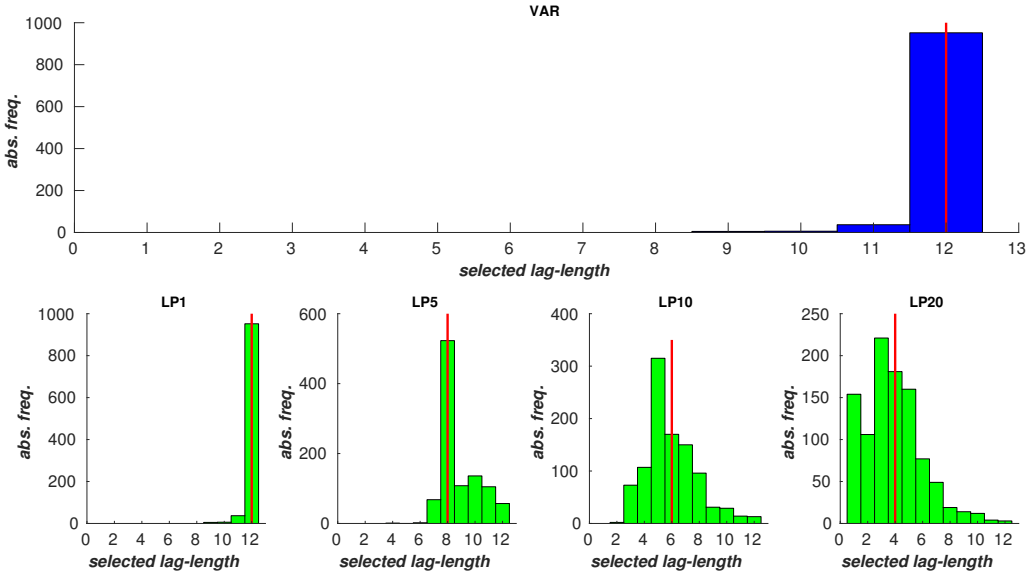

Figure 3: Selected lag-length distribution for VAR and LP IRFs using AIC as in Kilian and Kim. The four lower panels show the distribution for the local projection selected lag-length for the horizons $h = [1, 5, 10, 20]$. The red-solid vertical line shows the median lag-length.

selected for the VAR(p) model in each round of the Monte Carlo simulation is $p = 12$ with probability approaching one. Twelve is also the correct lag-length selected for the VAR data generating process. This finding explains why the lag-length selection procedure induces an unfair comparison between the local projection and the 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[12]. In this setting, it is implausible for any model to outperform the VAR, as the model should outperform the data generating process itself.

There is a second issue related to the lag-length selection procedure: in a VAR model, the lag-length is a scalar, while in the local projection methodology, the lag-length is a $(H \times 1)$ vector. For the VAR lag-length selection procedure there are articles suggesting how to select this parameter optimally. For example, the choice of the AIC in the Kilian and Kim's experiment is guided by Ivanov and Kilian (2005), which is a study on maximizing the performance of the estimated VAR IRFs. For the local projection methodology there are no studies on this critical topic, and having H parameters to be selected has generated in the literature at least two different practices.

1. Selecting the lag-length for each projection using an information criteria (Jordà, 2005, and Kilian and Kim).

2. Selecting the lag-length at the first step-ahead ($h = 1$), and using it for the H projections (this procedure has been employed in many empirical papers as Auerbach and Gorodnichenko, 2012a, Owyang et al. 2013, Tenreyro and Thwaites 2016, Caldara and Herbst 2016, Ambrogio Cesa-Bianchi, 2016).

A major implication of the lag-length selection procedure is visible in the second row of figure 3. The lag-length distributions of the local projection methodology are remarkably wild, and two main features emerge. On one side, at the first step-ahead, the local projection and the VAR impulse responses have identical distributions. This result is well-known, and due to Jordà (2005). Secondly, starting from the second step-ahead (not shown in the figure), the distributions begin widening, and the mode shifting toward more parsimonious models. This feature is evident in panel 2 to 4, where the mode is equal to eight, six and four respectively. This phenomenon is likely due to the increasing order of the moving average (MA) component of the local projection error. The MA component contributes capturing the variation in the data while fitting the local projection model and resulting in a more parsimonious autoregressive component. However, in the present setting, this feature contributes shifting the local projection model away from the true VAR(12) data generating process. Therefore, while the VAR IRFs is based on a well-specified model, as it selects the lag-length of the true DGP, the local projection estimator corresponds to the true data generating process only at the first horizon. Additionally, it turns increasingly misspecified

[12] Applying the LP estimator, the sample-size reduces with both lags p and leads h.

as the horizon increases. This controversial situation makes the comparison between the two methodologies unfair and is likely to produce misleading results.

Thirdly, a further source of distortion can be identified in the specification of the sample size T and the maximum lag-length $\overline{p}$ of the Kilian and Kim's Monte Carlo experiment. These two parameters play a prominent role in the lag-length selection procedure, and their choice, combined with the use of the Akaike's information criterion, supports the VAR in selecting the correct lag-length. In fact, compared to other criteria, the Akaike's criterion tends to prefer over-parametrized models. For VAR models, this feature implies a preference for specifications with higher lag-length. Other information criteria have a different tendency. For example, the Hurvich and Tsai (1993) criterion (AICC) introduces a correction term to attenuate this behavior. On the other hand, the Schwarz's criterion heavily penalizes over-parametrization, resulting in more parsimonious models. Finally, the Hannan and Quinn (1979) criterion (HQC) penalizes over-parametrization, but not as heavily as BIC.

To understand the implications of this point, table 1 reports the results of steps 1 to 5 of the Monte Carlo experiment described in section 4 by twisting the parameters $\overline{p}$ and T and selecting the lag-length with AIC, AICC, BIC, and HQC[13]. The table shows the relative frequency of the selected lag-length in percentage points. When the sample size is 400 and $\overline{p} = 12$, the AIC chooses the correct lag-length in

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 $\overline{p} = 12$.

[13] Appendix D reports the results for the entire Monte Carlo experiment.

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. Employing more parsimonious criteria always results in a misspecified VAR model. For example, the AIC and AICC distributions overlap as T increases. However, when T is small, the AICC distribution shrinks around $p = 1$. The distributions for the BIC and HQC are more concentrated around parsimonious models, with modes shifting from $p = 1$ to $p = 3$. 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 models[14].

To rebalance the experiment and having results reflecting a more fair comparison between the two procedures, in the next section I consider the case in which both methodologies present a controlled form of lag-length misspecification. In particular, I simulate the Monte Carlo experiment described in section 4 by choosing the lag-length with the most parsimonious criteria among the four previously described.

5.2 Simulation performing BIC model selection

In this section, I perform the Monte Carlo experiment described in section 4, choosing the lag-length for VAR and LP IRFs using the Schwarz information criterion. The BIC ensures selecting a model far from the true DGP in terms of lag-length. Figure 4 shows the lag-length distribution of the $M = 1000$ repetitions of the Monte Carlo simulation for both the VAR and the local projection procedure. The figure displays a scenario remarkably different from the one showed in figure 3. The distributions of the selected lag-length are entirely shifted toward more parsimonious models (as in table 1). In this context, the mode of the VAR lag-length distribution is $p = 2$, symmetrically the local projection distribution presented in the second row (first-left panel) 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. The first panel highlights the sensitivity of the two methodologies with respect to lag-length misspecification. In fact, the coverage rate decreases up to 65%-85% for local projection and till 40% for the VAR impulse responses. This is a large drop compared to what figure 2 shows. However, considering the relative performance of the two methodologies, the results are reverted. In fact, the true value of the IRFs is not contained in the VAR confidence bands more than 50% of the times. The same ratio is extremely lower for LP (around 20%). However, VAR

[14] Table C.1 and C.2 in appendix C show the cases in which $\overline{p} = 6$ and $\overline{p} = 18$.

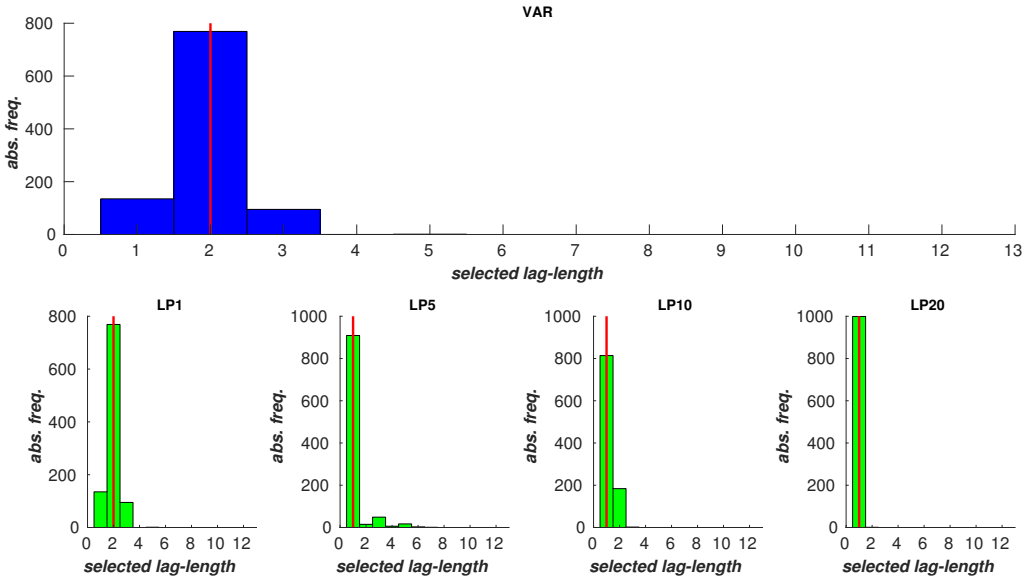

Figure 4: Selected lag-length distribution for VAR and LP IRFs using BIC criterion. The four lower panels show the distribution for the local projection selected lag-length for the horizons $h = [1, 5, 10, 20]$. The red-solid vertical line shows the median lag-length.

IRFs, due to lower STD, also have shorter average length, which can hint an explanation for the result. Nevertheless, VAR IRFs seems more biased on average, and less prone to errors in MSE terms, at least for further horizons. The latter feature is likely due to shorter STD and in turn, to the evidence that LP is a data consuming approach.

Motivated by these findings, in the next section, I perform the Monte Carlo experiment described in section 4 by selecting the local projection lag-length once for all the projections. Adopting this approach, for all the horizons h, the distribution of the LP and VAR lag-length are equal. Additionally, this twist allows to better control the induced misspecification by precisely selecting the desired lag-length. Finally, this case is an interesting one to consider because many empirical studies are using this procedure 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 the next section, I test this option.

5.3 Simulation with a fixed lag-length

Motivated by the previous results, in this section, I perform the Monte Carlo experiment described in section 4 by fixing the lag-length for both methodologies. Fixing the lag-length implies that local pro-

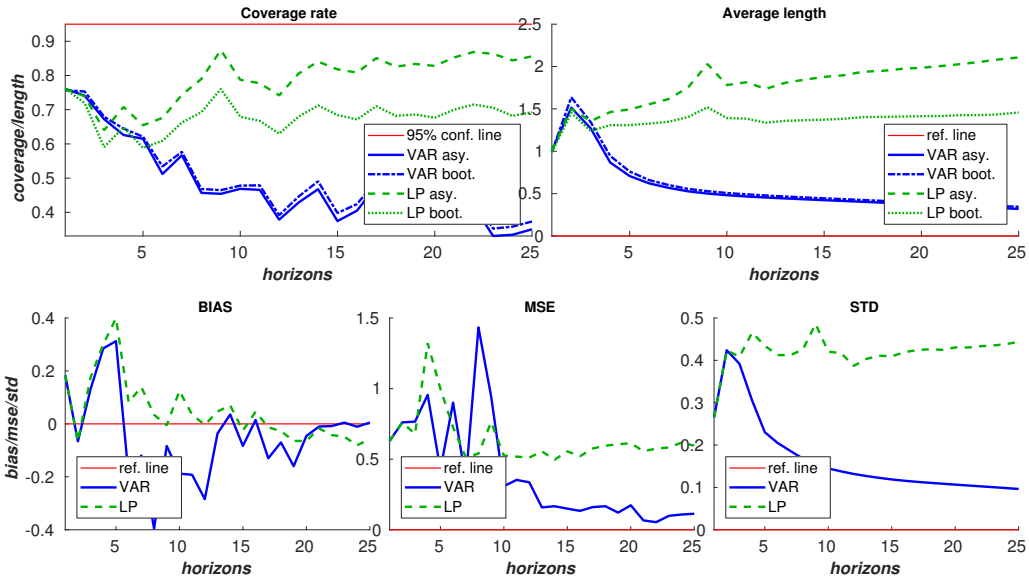

Figure 5: Average of statistics along all the shocks and all the variables in the system (the number of impulse responses in a $K = 4$ variable VAR is $K^2 = 16$). VAR asy.denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 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.

jection always presents a fixed order autoregressive component, while the order of the moving average increases with the horizon. For example, choosing $p = 1$ at horizon $h = 10$ returns a projection with an autoregressive component of order one and an MA error of order ten. I examine the cases in which $p \equiv [1, 3, 6, 9, 12]$, and compare the average criteria along the VAR and local projection impulse responses. On one side, the objective of the simulation is to understand which estimator performs better. On the other, by varying the degree of misspecification, the exercise allows assessing the existence of particular patterns. Therefore, to map the performance of the two methodologies, for each horizon h, I derive a measure to compare the difference between the average of the ECR^i, $i = \{LP, VAR\}$, absolute distance from the 95% correct acceptance rate (ΔECR), as presented in equation 7.

$$\Delta ECR\,(h) = \frac{1}{K^2}\sum_{j=1}^{K^2}\left(\left|0.95 - ECR_j^{LP}\right| - \left|0.95 - ECR_j^{VAR}\right|\right) \tag{7}$$

Assuming symmetric losses by deviating from the 0.95 threshold, when $\Delta ECR\,(h) > 0$, the VAR methodology displays higher coverage than local projection, on the contrary, when $\Delta ECR\,(h) < 0$, the reverse is true. When $\Delta ECR\,(h) = 0$, the two methodologies present equal 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 the Monte Carlo simulations with fixed lag-length $p \equiv [1, 3, 6, 9, 12]$. The

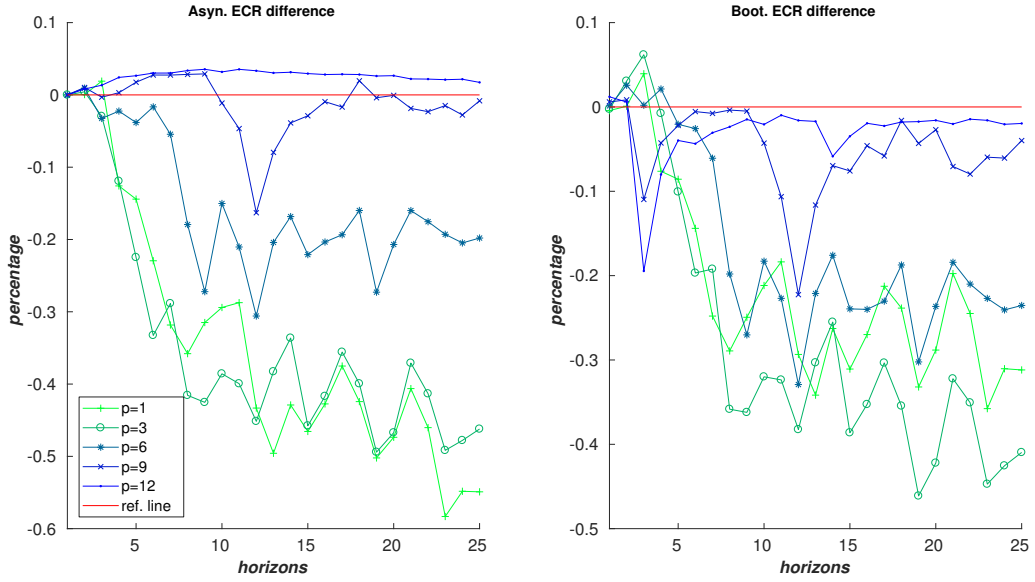

Figure 6: The figure shows the ECR for different Monte Carlo simulations using asymptotic and bootstrap confidence interval and selecting a different lag-length.

left panel shows the ΔECR computed with the VAR and LP asymptotic procedures. On the other hand, the right panel shows the difference between bootstrap procedures. Both panels display a clear advantage in terms of ECR of the local projection estimator. The coverage rate increases up to a maximum of 60% for the local projection estimator with respect to the VAR impulse responses. In particular, three striking features emerge; first, the relative performance of the local projection increases as the horizon increases. The reason is linked to the increasing wideness of the local projection confidence bands which enhance 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 (figure 6, left panel). For example, the ΔECR line corresponding to $p = 12$ is entirely above zero.

As shown in the previous subsections, the main reason why the coverage rate of the local projection is higher than the VAR impulse responses is connected to the wideness of the confidence bands. Figure 7 confirms this result, reporting in the second panel the STD ratio between the LP and VAR IRFs and in the third the ratio for the MSE. When the ratios are larger the one, the STD/MSE for the LP is larger than the VAR IRFs. In practice, these ratios are extremely large. The average STD ratio ranges from one to

seven, while the MSE ratio from one to sixteen. A clear pattern emerges from the figures; first, as lower the lag-length p (meaning as more misspecified the model), as wider the LP confidence bands. Secondly, as higher the horizon h as larger the ratio between the two procedure. In particular, such a difference in the STD ratio entirely enters in the computation of the MSE ratio between the two methodologies. This explains why the MSE ratio rapidly increases up to fifteen. From this perspective, large confidence bands imply a misspecification error due to the lag-length selection procedure. Naturally, the first best estimator would be unbiased and present small 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 explore the LP and VAR estimator bias level, I derive an average absolute measure of "bias distance" between the two procedures computed as in equation 8:

$$\Delta BIAS\,(h) = \frac{1}{K^2}\sum_{j=1}^{K^2}(\left|BIAS_j^{LP}\right| - \left|BIAS_j^{VAR}\right|) \tag{8}$$

$\Delta BIAS$ is positive, when $\left|BIAS^{LP}\,(h)\right| > \left|BIAS^{VAR}\,(h)\right|$, and negative in the opposite situation. When $\Delta BIAS = 0$, the two procedures presents the same degree of bias. The first panel of figure 7

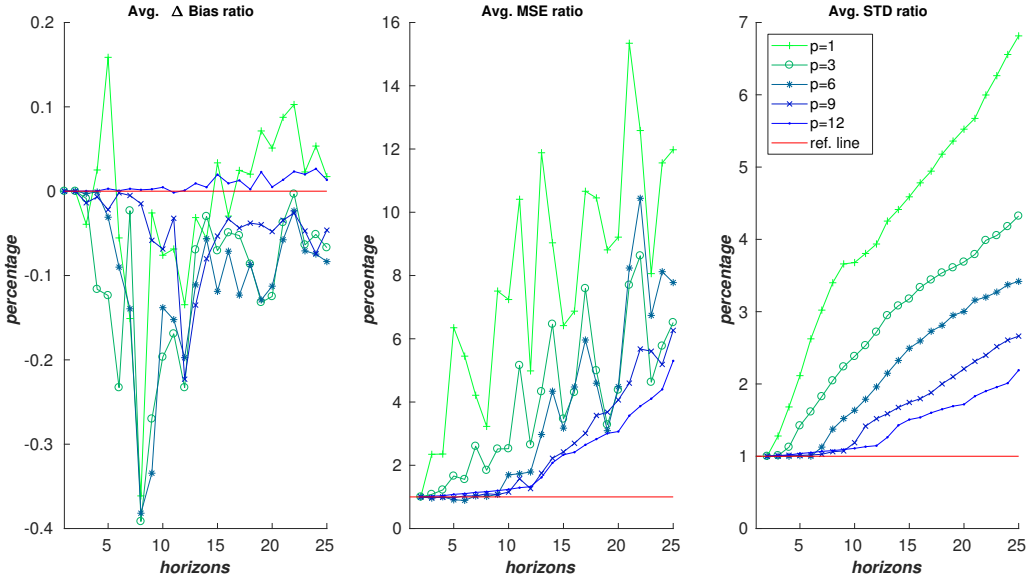

Figure 7: The figure shows the average bias, MSE, and STD for different Monte Carlo. The chart reports simulations results using asymptotic and bootstrap confidence intervals by selecting different lag-lengths.

displays the $\Delta BIAS$. The figure highlights the advantage of the local projection procedure, as $\Delta BIAS$ is more frequently in negative territory. This result is starkly in contrast with the Kilian and Kim's

finding. However, when $p = 12$ (solid blue line), meaning that the VAR model is correctly specified, $\Delta BIAS \geq 0$, and it is possible to retrieve the authors' result.

The central message that emerges from figure 6 and 7 is reasonably clear. As the model is misspecified, the VAR impulse responses return a vector of points far from the true ones. Nevertheless, confidence bands are narrow around these points and unable of signaling any uncertainty. On the other hand, the local projection estimator returns points closer to the correct ones 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 parameters.

6 Discussion: what is missing in the local projection literature?

The Monte Carlo experiments presented in section 5 show some interesting results about the coverage ability of the VAR and the LP methodology. In particular, they legitimate 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 paragraph, building on section 5, I summarize what are in my opinion the results that are still missing to reach this goal.

First, as highlighted in section 5, the local projection estimator has by construction an MA(h) error term, whose order increases with the horizon h. Jordà (2005) suggests estimating the variance-covariance matrix with 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's estimator in favor of methods which produce fewer distortions. The author cites voluminous literature starting from Kiefer et al. (2000). 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 routes to cover. Also Jordà, in its original paper, made a proposal which goes in the direction of improving the estimator efficiency (p. 165, last block, lines 5 to 9)[15]. The idea is simply to include in the projection at

[15] Also Miranda-Agrippino and Ricco (2017) and Barnichon and Brownlees (2017) methodologies follow this route.

step $h+1$ the estimate of the residual at step h. This should reduce the estimation uncertainty and shrink the confidence bands. However, a study to assess this procedure and possible drawbacks is needed to confirm this guess, at least in a small sample set.

Secondly, from the Monte Carlo simulations emerge 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 figure 2–5 show how the outcome can differ in a small sample setting. The reader is pointed to the practitioners' guide to best select the lag-length in a VAR model by Ivanov and Kilian (2005) to have a better idea of how the lag-length influences the impulse response function estimation (figure 1, p. 9). Moreover, the authors argue that the lag-length is not per se important, but it is only in the way it affects the precision of the IRF estimates. For example, by choosing as a performance metric the MSE, 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 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-sets with different frequencies and sample size T[16]. For the local projection, it is still an open issue how to select the best lag-length to use in each projection. Given that each projection may have different 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), which 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 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 param-

[16] In particular, the study 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.

eters as the Schwarz information criterion. However, for local projection, the same reasoning is not so straightforward given that a researcher can estimate the regression coefficients 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.

Indeed, further studies are necessary to deepen these crucial topics. However, a correct choice of the lag-length appears as the leading route to follow.

7 Conclusion

Starting from Kilian and Kim (2011), I compare the performance of the local projection estimator against the standard VAR impulse response function estimator. I replicate the Monte Carlo simulation performed by the two authors highlighting the arguments that lead to conclude in favor of VAR IRFs. In particular, I show that the Akaike (1974) information criterion 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 investigations.

In the analysis, I perform a Monte Carlo experiment with a controlled form of misspecification. Therefore, I compare VAR and LP IRFs in a context in which both models are misspecified. The results change completely and are no longer in favor of the VAR IRFs. Motivated by these findings, I perform a third Monte Carlo experiment fixing the lag-length for each equation of the local projection procedure. The results show that when the model is misspecified, the VAR impulse responses return a vector of points with narrow confidence bands far from the true ones. By contrast, the local projection estimator returns points closer to the correct ones together with large confidence bands. Additionally, the simulations display evidence that the performance of LP IRFs improves by selecting the lag-length once for all the projections.

Finally, based on these findings, I present a discussion on the missing ingredients in the local projection literature which can considerably improve the estimation performance. There is certainly much to

be done.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

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.

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.

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

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.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–195.

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.

Hurvich, C. M. and Tsai, C.-L. (1993). A corrected akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis*, 14(3):271–279.

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.

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. *mimeo*.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

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.

A Review of VAR and LP Impulse response function

A.1 VAR impulse response functions

The standard procedure to recover impulse responses consist in mapping 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 innovations. 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 [17]. 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. The *reduced form* VAR(p) is shown in equation A.1.

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

where $t = p+1, ..., T$, $y_t \equiv (y_{1t}, y_{2t}, ..., y_{Kt})'$ is a $(K \times 1)$ 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 a $(K \times 1)$ vector of independent and identically distributed white noises with $\mathbb{E}(e_t) = 0$, $\mathbb{E}(e_s e_t') = 0$, for $s \neq t$ and a $(K \times K)$ 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 A.2.

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

where the $(K \times K)$ variance-covariance matrix of ε_t, Σ_ε is diagonal and positive-definite. A_0 is $(K \times K)$ and in economic parlance is called the *impact matrix* and contains the contemporaneous effects 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 A.3.

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

According to the identification scheme chosen, A_0 might be any matrix imposing $\frac{1}{2}(K^2 - K)$ restrictions. As we abstract from identification issues in this paper, we simply assume that the DGP identification scheme is known and accomplished by imposing timing restriction, such as $\Sigma_e = A_0^{-1} \varepsilon_t \varepsilon_t' A_0^{-1'} =$

[17] Hamilton 1994 chapter 8 for the proof.

$A_0^{-1}\Sigma_\varepsilon A_0^{-1'} = A_0^{-1}A_0^{-1'}$, 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 A.1 can be rewritten as in equation A.4.

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

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 $Ly_t \equiv y_{t-1}$. 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 A.5:

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

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

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 linking the VAR to the 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 sets of reduced form and structural IRFs Θ_h are given by equations A.6 and A.7:

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

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

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.

A.2 Local projection impulse response functions

The local projection procedure is used to estimate the autoregressive coefficients directly at each h-step-ahead, regressing the dependent variable on its past as in equation A.8.

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

As shown by Jordà (2005), directly estimating the $(K \times K)$ autoregressive coefficients B_1^h, $h = 1, ..., H$, corresponds to estimating the IRFs without casting the Wold representation theorem. He also shows that the errors arising from this projections are VMA processes 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). Also, 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 errors due to misspecification as a function of the horizon. Also, as shown by the author, the local projection estimator 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 a model crucial economic features as state, size and sign dependency (Owyang et al. 2013, Tenreyro and Thwaites 2016).

Secondly, an essential advantage over VAR IRFs is in the direct estimation of the impulse response coefficients. However, to determine the structural impulse responses, a researcher still needs to estimate the impact matrix in a first-step aided by an auxiliary model as a VAR model. Equation A.9 shows the

structural impulse response functions estimated via local projection.

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

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. However, structural local projection impulse response functions computed with an auxiliary model and a projection invariant matrix is a topic which does not have any theoretical treatment in the literature. Given that the objective of the paper is to assess the local projection procedure ceteris paribus to the other studies, in this article, I do not take any further step in this direction and assume that the identification scheme is known. The only known alternative to this procedure is when a proxy series for the shock is available (see for example Swanson, 2017). In fact, the series can be used as an external instrument to achieve the identification.

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) dimensions.

B Results for individual IRFs

Figure B.1 to B.3 presents the results of the Monte Carlo simulation for individual impulse response functions, as in the original Kilian and Kim's article. 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 solid red line acts as the reference value for each statistic. The figures confirm the findings highlighted in section 5. In particular, when the lag-length is selected using the AIC, the VAR is a well-specified model, and the VAR IRFs outperform local projection. On the other hand, when the lag-length is selected using the BIC, as in figure B.4 to B.6, both the models are misspecified, and the LP impulse response functions outperform the VAR methodology.

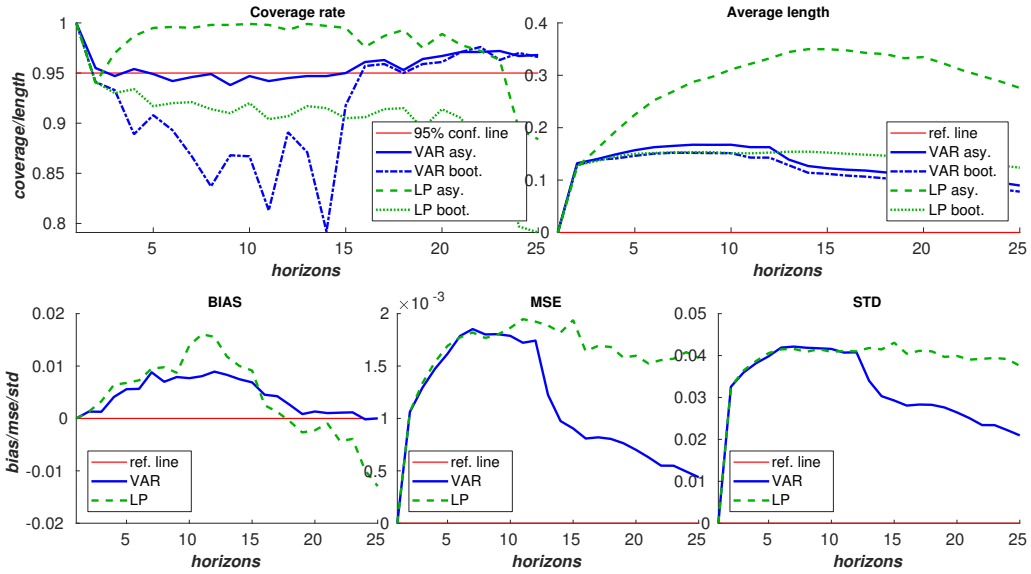

Figure B.1: IRFs to a monetary policy shock on CFNAI index of US real activity. VAR asy. denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 1998a,b,c). LP asy. denotes the asymptotic interval for LPs. LP boot. refers 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.

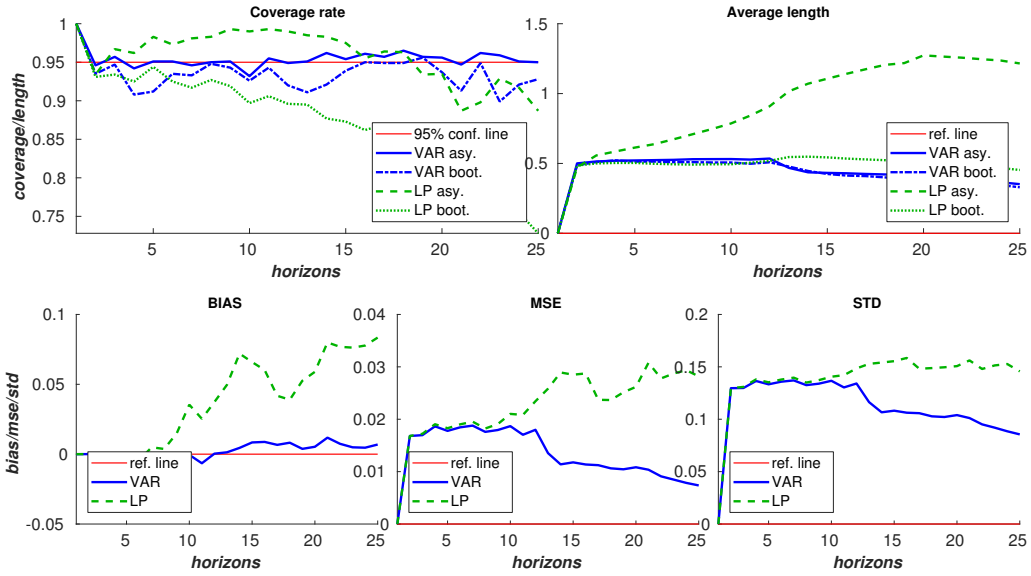

Figure B.2: Note as in Figure B.1. The reference variable is now US CPI inflation.

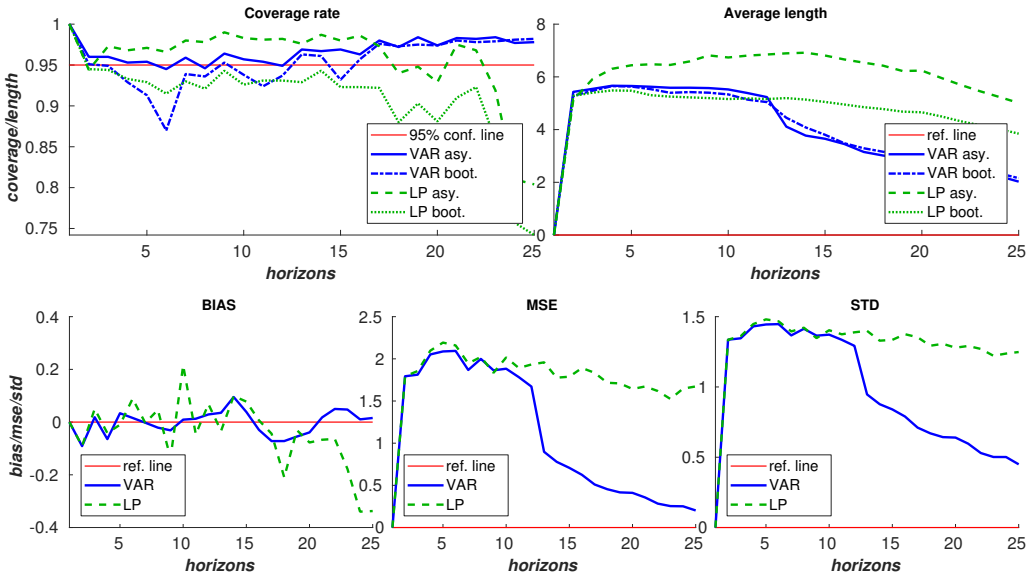

Figure B.3: Note as in Figure B.1. The reference variable is now US commodity price inflation.

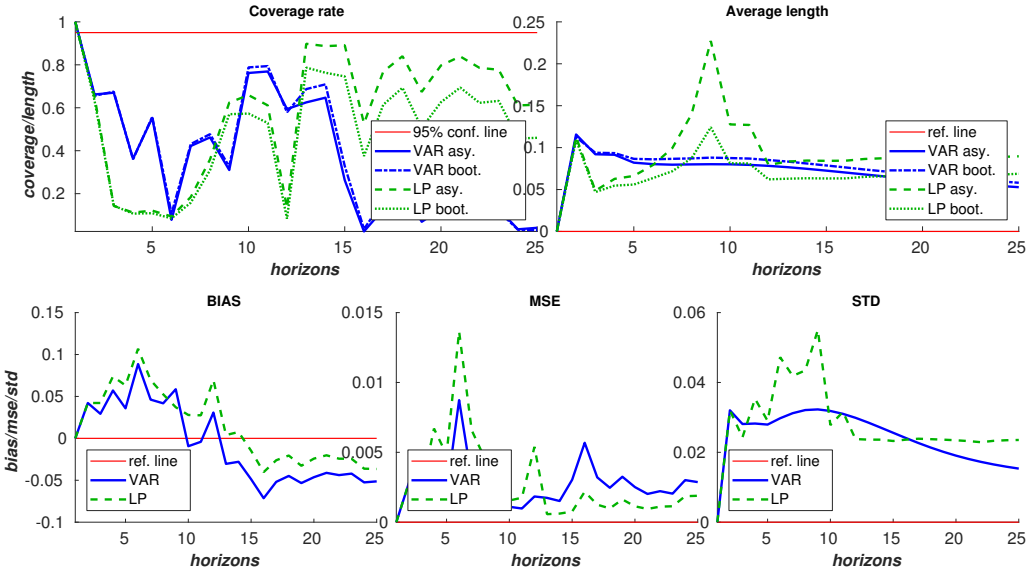

Figure B.4: IRFs to a monetary policy shock on CFNAI index of US real activity. VAR asy. denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (Kilian 1998a,b,c). LP asy. denotes the asymptotic interval for LPs. LP boot. refers 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.

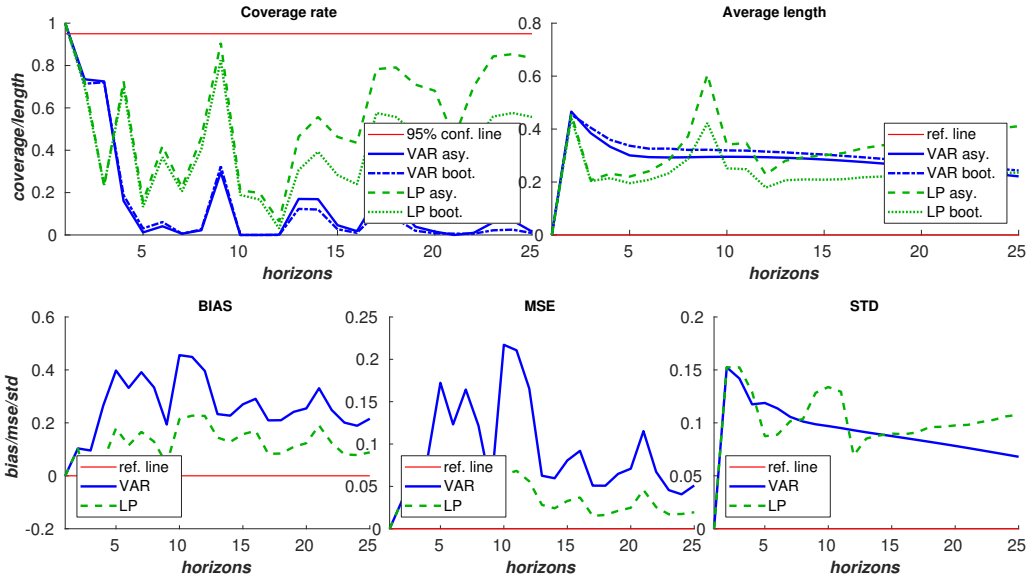

Figure B.5: Note as in Figure B.4. The reference variable is now US CPI inflation

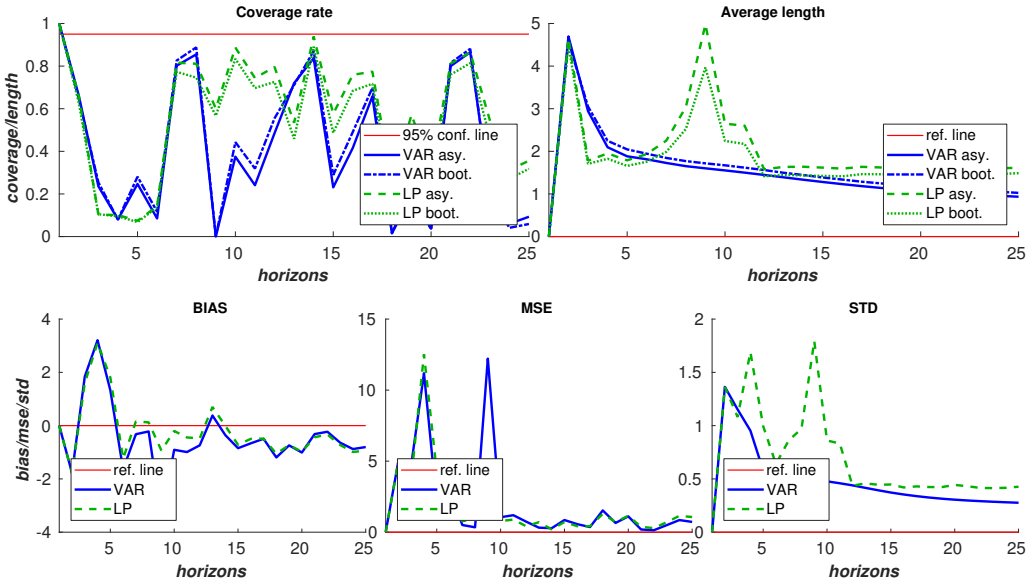

Figure B.6: Note as in Figure B.4. The reference variable is now US commodity price inflation.

C Additional tables

C.1 Results for $\bar{p} = 6$ and $\bar{p} = 18$

Table C.1: Lag-length selected in a Monte Carlo exercise. $\bar{p} = 6$

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 $\bar{p} = 6$.

Table C.2: Lag-length selected in a Monte Carlo exercise. $\bar{p} = 18$

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 $\bar{p} = 18$.

D Additional simulations

In this section, I report the results of the Monte Carlo simulation using the corrected Akaike criterion (AICC) and the Hannan-Quinn criterion (HQC). These are popular model selection criteria extensively used in the empirical macroeconomics and time-series literature.

D.1 Simulation performing model selection using HQC

Figure D.1 presents the distributions of the lag-length p and p^{lp} in the $M = 1000$ repetitions of the Monte Carlo algorithm using the HQC. The upper panel (blue bars) shows the distribution of the VAR lag-length, while the lower panels (green bars) displays the distribution of the local projection lag-length for $h = [1, 5, 10, 20]$. The results are mixed between the results of the Monte Carlo simulation performed using the AIC and the BIC. The main difference concerns the support of distributions, which have broader coverage. The VAR distribution has mode $p = 3$, while the LP distribution mode shifts between one and six. Figure D.2 presents the average of the effective coverage rate, average length, mean squared error,

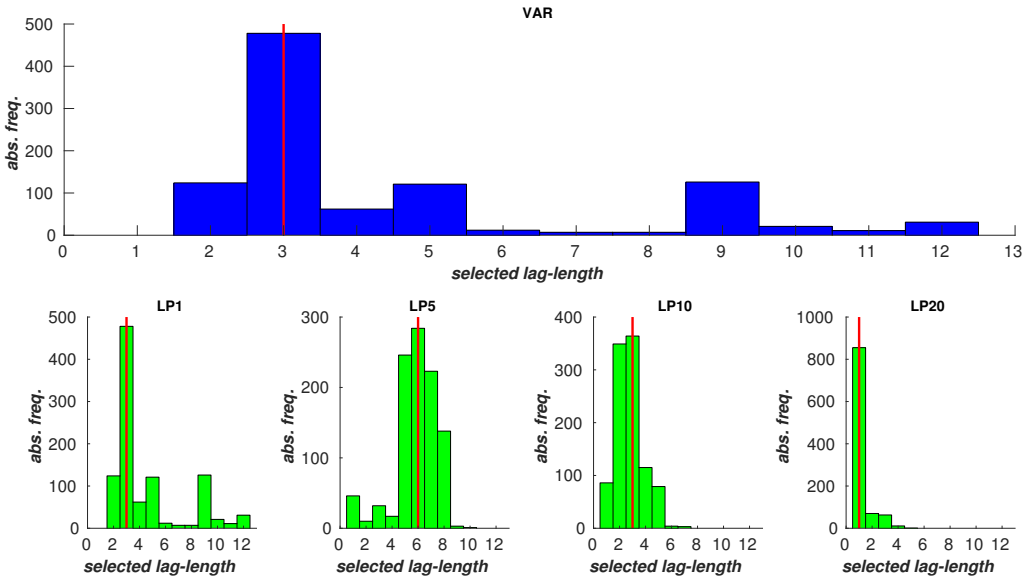

Figure D.1: Selected lag-length distribution for VAR and LP IRFs using HQC. The four lower panels show the distribution for the local projection selected lag-length for the horizons $h = [1, 5, 10, 20]$. The red-solid vertical line shows the median lag-length.

standard deviation, and bias of the impulse responses. The results of using the HQC are extremely close to the results presented using the BIC.

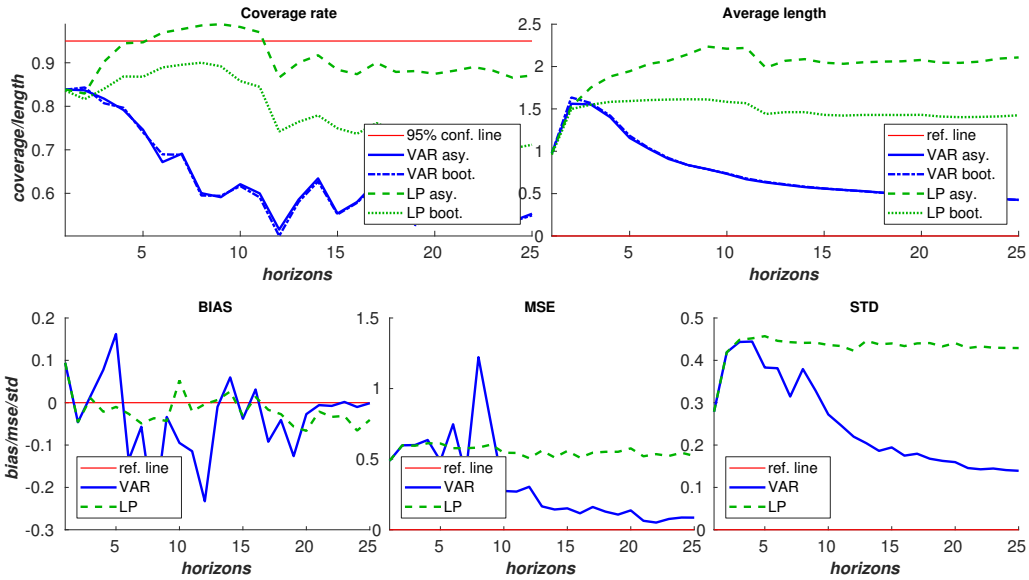

Figure D.2: Average of statistics along all the shocks and all the variables in the system (the number of impulse responses in a $K = 4$ variable VAR is $K^2 = 16$). VAR asy.denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (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.

D.2 Simulation performing model selection using AICC

Figure D.3 presents the distributions of the lag-length p and p^{lp} in the $M = 1000$ repetitions of the Monte Carlo algorithm using the AICC. The upper panel (blue bars) shows the distribution of the VAR lag-length, while the lower panels (green bars) displays the distribution of the local projection lag-length for $h = [1, 5, 10, 20]$. The results resemble the results of the Monte Carlo simulation performed using the AIC. The main difference is that the distribution is less concentrated around the correct lag-length. The VAR distribution has mode $p = 12$, while the LP distribution mode shifts from twelve to three. Figure D.4 presents the average of the effective coverage rate, average length, mean squared error, standard deviation, and bias of the impulse responses. As expected, due to the similarities between AIC and AICC, the results are extremely close.

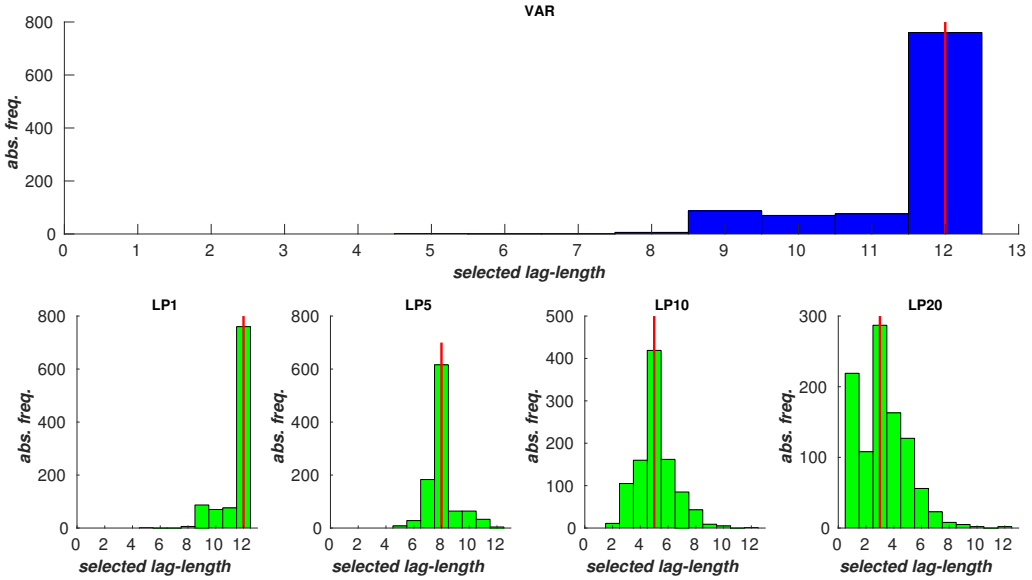

Figure D.3: Selected lag-length distribution for VAR and LP IRFs using AICC. The four lower panels show the distribution for the local projection selected lag-length for the horizons $h = [1, 5, 10, 20]$. The red-solid vertical line shows the median lag-length.

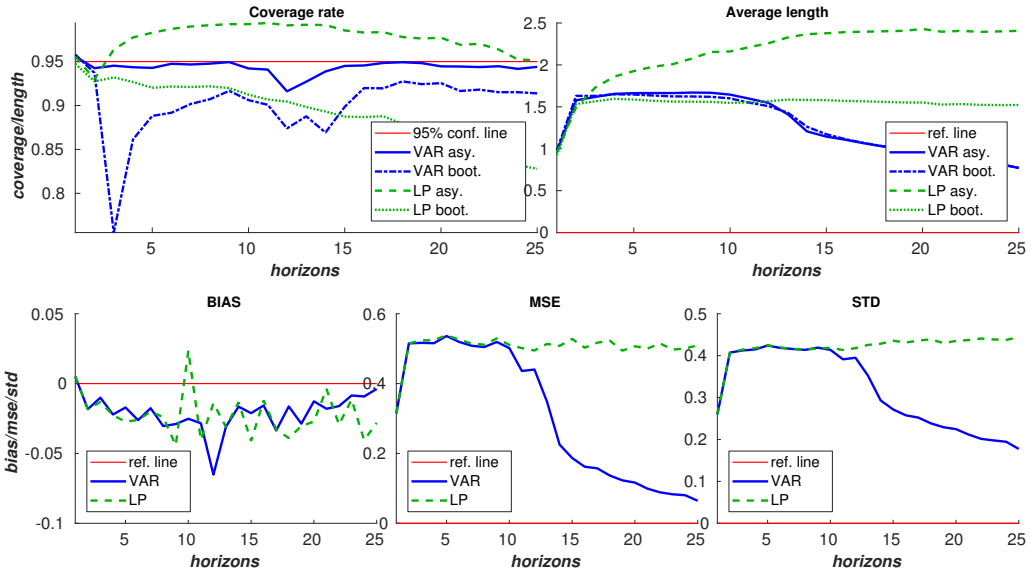

Figure D.4: Average of statistics along all the shocks and all the variables in the system (the number of impulse responses in a $K = 4$ variable VAR is $K^2 = 16$). VAR asy.denotes the asymptotic delta method for VAR impulse responses (Lutkepohl 1990). VAR boot refers to the bias-corrected bootstrap (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.