Maastricht University

Finite-Sample Bias Corrections for Local Projections: Post-Estimation Corrections vs. Long-Difference Estimation

By Primo Beekhuis

This thesis investigates finite-sample bias correction methods within the local projection framework. While both Herbst and Johannsen (2024) and Piger and Stockwell (2025) propose approaches to reduce such bias, it remains unclear which method yields superior performance. Simulation results reveal that the choice of bias correction depends critically on the researcher's concern for the bias of the estimator relative to its variance. To address the high variance of the long-differenced estimator proposed by Piger and Stockwell (2025), this thesis introduces the restricted long-differenced estimator. An extensive simulation study shows that the estimator outperforms existing bias correction methods under certain bias–variance preferences and offers robust pointwise inference. Finally, an empirical reanalysis of U.S. monetary policy and information shocks, following Jarociński and Karadi (2020), demonstrates that the choice of bias correction method can substantially change the interpretation of impulse response functions, potentially even reversing the sign of the estimated effects.

Maastricht University · Supervisor: Prof. Stephan Smeekes · June 26, 2025

Contents

1 Introduction

For centuries, economists have aimed to understand the forces that drive economic activity. From the early writings of classical economists on business cycles and monetary theory to the development of modern macroeconomic models, the goal has remained unchanged: to explain how economies respond to shocks and policy interventions. Today, macroeconomists often estimate impulse response functions (IRFs) to track the dynamic effects of shocks such as interest rate hikes or increases in government spending on variables like gross domestic product (GDP) and price levels. Accurate estimation of these responses is essential in empirical macroeconomics and plays an important role in shaping policy decisions. For instance, economists at the World Bank Group have used IRFs to study the effect of tariffs on macroeconomic variables including output, productivity, and unemployment (Furceri et al., 2021). In addition, researchers at the European Central Bank have employed IRFs to analyse the impact of energy shocks on inflation in the euro area (Bobeica et al., 2025).

Impulse response functions trace the change of the dependent variable over several horizons following a one-time shock. The shock is typically constructed in such a way that it can be assumed to be exogenous, allowing researchers to separate its causal effect from other events in the economy. Importantly, impulse response functions are not forecasts. Instead, they summarize the historical average response of a variable to a given type of shock. Estimating an impulse response function accurately comes with multiple obstacles, like defining the shock variable, choosing an appropriate identification strategy, and dealing with small sample sizes that can bias the estimated responses.

Broadly speaking, there are two approaches to estimating impulse response functions. The first, and traditionally most commonly used, is the vector autoregressive (VAR) model (Sims, 1980). VAR models specify each variable as a linear function of its past values and the past values of all other variables in the system, capturing the dynamic relationships among them. The second approach is local projections (LPs), introduced by Jordà (2005), which have gained popularity in recent years. Contrary to VARs, local projections estimate impulse responses directly through horizon-by-horizon regressions. This method does not require specifying the full dynamic system, which among other desirable properties, has contributed to the widespread adoption of local projections in macroeconomic research.

Previous research has demonstrated the reduced finite-sample bias of impulse responses estimated by local projections compared to VARs (e.g., Li et al., 2024; Montiel Olea et al., 2025). Nevertheless, finite-sample bias persists and is economically meaningful as demonstrated by Herbst and Johannsen (2024). Moreover, the authors show the bias increases in magnitude with smaller sample sizes. This is concerning in the macroeconomics literature, where time series are generally short, since the available data is of low-frequency and researchers focus on limited time periods to address specific research questions. Herbst and Johannsen (2024) perform a survey of the empirical local projection literature and show that among the 71 "most relevant" papers that employ local projections empirically and cite Jordà (2005), the median time series length is around 95. Where VARs have long benefitted from bias corrections (Pope, 1990) and alternative

model specifications to reduce bias, bias corrections for local projections have only recently been proposed.

The two bias reduction methods that have recently been proposed are the Herbst and Johannsen (2024) bias correction, who introduce a recursive, post-estimation bias correction procedure, and the long-differenced (LD) regression specification formalized by Piger and Stockwell (2025). Although the long-differenced specification is simply another way of specifying the local projection regression, this thesis refers to both approaches as bias corrections, as they are designed to reduce bias relative to the standard LP estimator. Furthermore, both studies present evidence that their respective estimators provides more robust inference than the standard local projection estimator. These distinct methods raise the question which of the two bias corrections is preferred and whether this is situation-dependent. Hence, this thesis examines two aspects of bias correction methods for local projections. First, this work evaluates which of the existing bias corrections performs better. Second, a third bias-correction, referred to as the restricted long-differenced (RLD) estimator, will be proposed. This leads to the following research question:

> *How do existing bias correction methods for local projections compare in terms of inference accuracy and bias, and to what extent can the restricted long-differenced estimator improve upon them?*

In answering this research question, this thesis makes several contributions to the existing literature. First, this thesis is the first to systematically compare the two most prominent bias corrections for local projections, namely the recursive BCC correction and the long-differenced estimator, across a variety of empirically relevant data generating processes (DGPs). In doing so, it offers a comprehensive evaluation of the bias-variance trade-off inherent in these methods, which is characteristic of this literature with similar trends found in the trade-off between VARs and local projections (Li et al., 2024). Second, this work introduces the restricted long-differenced estimator, an estimator that retains the benefits of the LD specification in decreasing bias while imposing restrictions on the lag structure to reduce the variance of the estimator. Simulation results show that the RLD estimator can outperform the existing bias corrections under specific bias-variance preferences as well as provide better coverage probabilities for more robust inference. Subsequently, the findings are applied to an empirical application that revisits the macroeconomic effects of U.S. monetary policy and central bank information shocks (Jarociński and Karadi, 2020), revealing that the choice of bias correction may alter both the magnitude and even the sign of estimated impulse responses. Taken together, these results highlight the empirical relevance of finite-sample bias corrections in local projections and show that the RLD estimator is a suitable alternative.

The remainder of the thesis is structured as follows. Section 2 reviews related literature, focusing on differences between LPs and VARs, bias in local projections, and inference methods. Section 3 presents the theoretical framework, particularly focussing on the bias correction methods developed in recent literature, along with the restricted long-differenced estimator proposed in this thesis. Section 4 describes the simulation study in depth, and Section 5 discusses the associated results. Section 6 applies the bias corrections to an empirical case. Finally, Section 7

presents the conclusions of the research, describes the limitations, and outlines potential areas for further research.

2 Related Literature

Since its introduction by Jordà (2005), local projections have gained prominence in the context of time series analysis. Predominantly in macroeconomic literature, local projections are commonly employed to estimate impulse response functions. In this section, the relevant contributions in the literature that have advanced the understanding of local projections are reviewed.

2.1 VARs and LPs

Impulse response functions have long been estimated with VARs (Sims, 1980). Since Jordà (2005), local projections have become a widely used alternative. Plagborg-Møller and Wolf (2021) show that VARs and local projections estimate the same impulse response functions asymptotically, provided the number of lags tends to infinity. This result holds regardless of the identification scheme or the data generating process.

Li et al. (2024) examine the trade-offs between the two approaches through an extensive simulation study involving 6000 DGPs. Their findings indicate that local projections are preferred only if the researcher values an estimator with low bias much more than one with low variance. More generally, their results underscore a clear bias–variance trade-off: local projections tend to have lower bias but higher variance, whereas vector autoregressions display higher bias and lower variance. Montiel Olea et al. (2025) emphasize that, within this framework, achieving lower variance invariably comes at the cost of increased bias. The authors also note that the only way to ensure impulse responses estimated by a VAR-based approach have low bias is to control for so many lags that they become equivalent with the local projection. While this thesis does not compare the overall performance of the two methods, it does examine the bias–variance trade-off among bias corrections within the local projections framework.

2.2 Bias in Local Projections

Small sample bias is not novel in the estimation of impulse responses. In VAR models, this bias was identified by Nicholls and Pope (1988) and Pope (1990). Specific to local projections, Herbst and Johannsen (2024) explicitly derive an approximation for the magnitude of the bias in small samples. Using this approximation, the authors propose a post-estimation bias correction. The issue of small sample bias is also explored by Piger and Stockwell (2025), who suggest estimating impulse response functions with a long-differenced regression specification instead. The large simulation study by Li et al. (2024) conclude that local projections reduce the bias of impulse response estimations, although this advantage is only convincingly realized when the bias correction proposed by Herbst and Johannsen (2024) is applied. These authors do not examine the long-differenced specification formalized by Piger and Stockwell (2025). This thesis directly

contributes to the literature by considering the long-differenced specification, and comparing it to the standard local projection estimator and the Herbst and Johannsen (2024) correction.

Closely related is the generally jagged shape of impulse responses estimated with local projections and the accompanying higher variance compared to VARs (Li et al., 2024). Barnichon and Brownlees (2019) propose a methodology based on B-spline smoothing, which they call smooth local projections. This estimator incorporates a penalty term during estimation to push the estimated impulse response function towards a smooth function. While this tends to yield lower variance than the standard local projection estimator, it does so at the cost of increased bias over certain horizons, as shown by Li et al. (2024). Although this thesis does not evaluate smooth local projections due to its higher bias, the restricted long-differenced estimator similarly aims to reduce the variance of the long-differenced estimator.

2.3 Inference with Local Projections

Inference with local projections is an active area of research, with various methods being developed. Inoue et al. (2025) discuss how the appropriate inference method depends on the research objective. The authors consider the use of heteroskedasticity and autocorrelation consistent (HAC) estimators, as well as heteroskedasticity-robust covariance matrix estimators, for pointwise inference. Moreover, Generalized Method of Moments approaches are discussed for joint inference. Lastly, the authors introduce the concept of significance bands for hypothesis testing. This thesis focuses on pointwise inference, as it remains the most commonly used approach in the local projections literature.

Jordà (2005) proposes the use of HAC covariance estimators, such as the Newey–West estimator (Newey and West, 1987), in the local projection framework, because regression residuals are serially correlated. However, Montiel Olea and Plagborg-Møller (2021) show that inference based on heteroskedasticity-robust standard errors, such as the Eicker–Huber–White correction (White, 1980), becomes valid under lag augmentation; that is, when one more lag than necessary is included in the regression. Moreover, Herbst and Johannsen (2024) reason that even without lag augmentation, Newey-West standard errors may perform poorly since they use estimators of the autocorrelation of the regression score, which induces a bias in the standard errors in finite samples. While inference based on bootstrapped confidence intervals has been explored in the literature (Kilian and Kim, 2011), this thesis instead focuses on Newey–West and Eicker–Huber–White standard error estimators to evaluate coverage probabilities.

3 Theoretical Framework

This section provides the relevant theoretical background for local projections, their estimation, and the bias corrections. First, the standard local projection specification is introduced. Second, identification strategies are discussed within the local projection framework. Subsequently, the bias-corrected estimators proposed by Herbst and Johannsen (2024) and Piger and Stockwell (2025) will be introduced. Lastly, the restricted long-differenced specification is introduced and

its implementation is discussed.

3.1 Local Projections: Regression Specification

Local projections are estimated with horizon-by-horizon regressions by ordinary least squares (OLS). The general regression specification is,

$$y_{t+h} = \mu_h + \theta_h^{LP} x_t + \varphi_h q_t + \sum_{l=1}^{p} \gamma_{h,l}' w_{t-l} + \varepsilon_{t,h}, \tag{1}$$

for each horizon $h \in \{0, ..., h_{\max}\}$. The outcome variable of interest, h periods into the future is denoted y_{t+h} and x_t is the impulse or shock variable. The impulse response function is mapped out by θ_h^{LP}. Local projection regressions typically include lagged variables and potentially contemporaneous control variables. All lagged variables, including lags of the outcome variable and other controls, are collected in the vector w_{t-l} for lag l, while contemporaneous controls are represented by the vector q_t; it is assumed that $q_t \subseteq w_t$. Henceforth, the number of lags in a regression specification refers to the number of lags per variable, assumed equal across all variables.

A key feature of local projections is that each horizon-specific regression is estimated independently. In contrast to vector autoregressive methods, local projections do not require specifying a full dynamic system that jointly models all variables across time. This flexibility allows researchers to avoid imposing potentially restrictive assumptions on the DGP. This also makes LPs more robust to misspecification, even compared to VARs with relatively little misspecification (Montiel Olea et al., 2025). Furthermore, unlike (short-lag) VARs, local projections do not rely on extrapolating long-horizon dynamics from a small number of estimated auto-covariances, which can lead to biased impulse response estimates (Montiel Olea et al., 2025).

Herbst and Johannsen (2024) analytically derive an approximation of the bias in the standard LP specification in finite samples. This approximation shows that this bias decreases as the sample size increases. Accordingly, the simulation study is expected to reflect this property. As an illustration, Herbst and Johannsen (2024) explicitly note the approximate bias for an AR(1) DGP; this bias is an increasing function of the persistence level of the underlying process. As most macroeconomic time series are highly persistent and are short in length, this reinforces the importance of addressing bias in small samples.

3.2 Identification in Local Projections

This thesis considers two commonly used identification schemes: observed shock identification and recursive shock identification. Further methods of identification, such as instrumental variable identification in local projections (e.g., Jordà and Taylor, 2016; Ramey and Zubairy, 2018), are left to future work.

Observed shock identification applies when the shock or intervention can be measured directly and treated as exogenous (Ramey, 2016). In this case, the shock is obtained through methods external to the impulse response estimation. A prominent example is the shock constructed by

Romer and Romer (2004), who construct a monetary policy shock by isolating the unanticipated component of Federal Reserve decisions. Specifically, they measure the difference between the actual change in the federal funds rate and the predicted change. Another example is the fiscal policy shock constructed by Ramey (2011). Once the shock is identified, the impulse response at horizon h can be estimated by regressing the outcome variable at time $t+h$ on the observed shock at time t. With observed shocks, lagged control variables are not necessary for consistency, although they are generally included to improve efficiency (Li et al., 2024).

Alternatively, shocks can be identified recursively, which imposes a particular causal ordering on the variables to define the shock (see, for example, Christiano et al. (1999) for a monetary policy application). The goal is to estimate the effect of an impulse x_t on an outcome y_{t+h}. This is achieved by assuming that some variables respond to the shock contemporaneously while others do not. That is, the researcher imposes timing restrictions by deciding which variables are ordered before and after x_t in the system. The key assumption is that variables ordered before x_t do not respond contemporaneously to it, while x_t can respond immediately to them.

3.3 BCC Estimator

Herbst and Johannsen (2024) propose a bias correction method applied to standard local projection estimates. The bias-corrected estimator is computed separately for each forecast horizon. Let w_t represent the vector of control variables at time t. Note that if lagged values of the outcome variable are included as regressors in the local projection, then y_t would be in w_t. The bias correction is based on the following covariance matrix estimator:

$$\hat{\Sigma}_{w,j} = \frac{1}{T-h-j} \sum_{t=j+1}^{T-h} (w_{t-j} - \bar{w})(w_t - \bar{w})', \quad \text{for } j = 0, \ldots, h,$$

where T is the sample size and the sample mean is simply defined as:

$$\bar{w} = \frac{1}{T-h} \sum_{t=1}^{T-h} w_t.$$

The initial condition at horizon $h = 0$ is given by $\hat{\theta}_0^{BCC} = \hat{\theta}_0^{LP}$. Then the bias-corrected estimator is constructed sequentially as:

$$\hat{\theta}_h^{BCC} = \hat{\theta}_h^{LP} + \frac{1}{T-h} \sum_{j=1}^{h} \left(1 + \operatorname{tr}\left\{\hat{\Sigma}_{w,0}^{-1} \hat{\Sigma}_{w,j}\right\}\right) \hat{\theta}_{h-j,BCC}, \quad h = 1, \ldots, h_{\max}, \tag{2}$$

where $\hat{\theta}^{LP}$ refers to the standard local projection estimator from Equation 1 and $\operatorname{tr}\{\}$ is the trace operator.

The BCC estimator applies an analytical bias correction to local projection estimates at each horizon in a recursive manner. Specifically, for each $h \geq 1$, the bias correction term adds a weighted sum of the bias-corrected estimates from earlier horizons, where the earlier estimates are scaled by the trace term $1 + \operatorname{tr}\left\{\hat{\Sigma}_{w,0}^{-1} \hat{\Sigma}_{w,j}\right\}$. When the control variables are strongly

autocorrelated, the trace term is larger and thus results in a greater correction. Moreover, the scaling by $(T-h)^{-1}$ reflects that the bias correction diminishes as the effective sample size grows. Specifically, as the sample size T grows, the finite-sample bias becomes smaller, and the correction term vanishes asymptotically. Conversely, at longer horizons, the magnitude of the bias correction may increase due to the smaller effective sample size, although further assumptions are needed for this to be true. For instance, the summation term grows in absolute terms for $h \geq 1$ if the sign of the BCC estimator remains consistently positive or negative across horizons and $\operatorname{tr}\left\{\hat{\Sigma}_{w,0}^{-1}\hat{\Sigma}_{w,j}\right\} > -1$ for all $j = 1, \ldots, h$, with the latter condition generally holding in macroeconomic settings, as noted by Herbst and Johannsen (2024).

3.4 Long-Differenced Specification

Piger and Stockwell (2025) approach the bias correction problem from a different angle by defining the long-differenced regression specification. The corresponding estimator of the impulse response is referred to as the long-differenced estimator. As in Equation 1, it is assumed that each variable enters the regression with the same number of lags, and w_t denotes the control variables. For illustration purposes, the regressors of the outcome variable are explicitly noted while the remaining controls are in $\tilde{w}_t$. Contemporaneous controls are omitted as in Piger and Stockwell (2025); this is the appropriate regression specification when a shock in observed directly. The long-differenced regression with p_D lags of the first difference is defined as:

$$\Delta y_{t+h} = \mu_h^D + \theta_h^{LD} x_t + \alpha_{1,h}\Delta y_{t-1} + \cdots + \alpha_{p_D,h}\Delta y_{t-p_D} + \sum_{l=1}^{p_D} \zeta_{h,l}'\Delta\tilde{w}_{t-l} + u_{t,h}, \tag{3}$$

where $\Delta y_{t+h} = y_{t+h} - y_{t-1}$, which motivates the name of the long-differenced regression. On the right-hand-side, $\Delta y_{t-l} = y_{t-l} - y_{t-l-1}$ for all $l \in \{1, \ldots, p_D\}$. The same notation holds for all other control variables. Note that the effective sample size for the standard LP with p lags is $T-h-p$, whereas the effective sample size in this specification with p_D lags of the first difference is $T-h-p_D-1$.

The derivation of this regression specification follows directly from the standard local projection specification. Specifically, one can define the first-differenced regression specification as in Equation 2 of Piger and Stockwell (2025). Then, by summing over time from t to $t+h$, as in Stock and Watson (2018), the specification in Equation 3 follows. From this derivation, it additionally follows that in the correctly specified long-differenced regression, $p_D = p + h$, where p is the number of lags in the standard LP.

Piger and Stockwell (2025) propose the long-differenced specification for settings in which shocks are either observed directly or identified using instrumental variables. Although the authors do not explicitly discuss its limitations under other identification schemes, it is clear that the specification is not appropriate for recursively identified shocks. Recursive identification imposes timing restriction by ordering the variables. Hence, variables ordered before the impulse variable must be included contemporaneously (Plagborg-Møller and Wolf, 2021). With all other control variables included in first-differenced form, there appears no natural way to include

contemporaneous control variables. Therefore, the estimates based on the LD specification are not expected to perform well under recursive identification.

3.5 Restricted Long-Differenced Specification

This thesis proposes a third estimator by imposing restrictions on the coefficients of the first-differenced regressors in the long-differenced specification. The impulse response estimator corresponding to this regression specification is referred to as the restricted long-differenced estimator. Since it builds on the long differenced specification, the restricted long differenced specification is primarily intended for observed shocks. By restricting the coefficients on the lags in the long-differenced regression, the restricted long-differenced specification has equally many lagged regressors as the standard local projection specification. Note that these lagged variables in the restricted long-differenced specification are not generally in first differences, and instead differ in the length of the difference (e.g., $y_{t-1} - y_{t-3}$ and Δy_{t-1} may be used in the same regression). For simplicity, these regressors are referred to as lags, where l lags denotes l lagged regressors per variable.

Assume that the correctly specified standard LP specification includes p lags of each variable. The correctly identified long-differenced specification then includes $p_D = p + h$ lags of the first difference. Construct the matrix R_h of dimension $p_D \times p$ for $h \in \{0, \dots, h_{\max}\}$. The entries of the matrix are defined as follows for all $j \in \{1, \dots, p_D\}$ and $i \in \{1, \dots, p\}$:

$$[R_h]_{j,i} = \begin{cases} 1 & \text{if } i \leq j \leq h + i, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\Delta Y_t = (\Delta y_{t-1}, \Delta y_{t-2}, \dots, \Delta y_{t-p_D})'$ be a $p_D \times 1$ vector of lags of the first difference. In the long-differenced regression,

$$\Delta_h y_{t+h} = \mu_h^D + \theta_h^{LD} x_t + \alpha_h' \Delta Y_t + \sum_{l=1}^{p_D} \zeta_{h,l}' \Delta \tilde{w}_{t-l} + u_{t,h},$$

replace $\alpha_h' \Delta Y_t$ with $\rho_h'(R_h' \Delta Y_t)$, where R_h' is the transpose of R_h. Let Z_t be a $p \times 1$ vector defined as $Z_t := R_h' \Delta Y_t$. This same transformation must analogously be applied to each control variable to produce a $p \times 1$ vector for each control. If $\tilde{w}_t$ contains $n_w - 1$ variables, these transformed vectors are then stacked to form the $p(n_w - 1) \times 1$ vector X_t^{RLD}. Then, estimate:

$$\Delta_h y_{t+h} = \tilde{\mu}_h^D + \theta_h^{RLD} x_t + \rho_h' Z_t + \tilde{\zeta}_h' X_t^{RLD} + \eta_{t,h}, \tag{4}$$

where θ_h^{RLD} is the impulse response at horizon h in the restricted long-differenced specification, $\rho_h = (\rho_{1,h}, \dots, \rho_{p,h})'$ is a $p \times 1$ vector, and $\tilde{\zeta}_h = (\tilde{\zeta}_{h,1}, \dots, \tilde{\zeta}_{h,p})'$ is a $p(n_w - 1) \times 1$ vector. This is the correctly specified restricted long-differenced specification.

Note that the restricted long-differenced specification directly employs the relationship between the coefficients of the lagged variables in the standard LP and the LD specifications.

Consider the standard LP specification,

$$y_{t+h} = \mu_h + \theta_h^{LP} x_t + \rho_{1,h} y_{t-1} + \cdots + \rho_{p,h} y_{t-p} + \sum_{l=1}^{p} \gamma_{h,l}' \tilde{w}_{t-l} + \varepsilon_{t,h}, \tag{5}$$

where, for illustration, lags of the outcome variable are explicitly taken out of the summation of lags. As noted by Piger and Stockwell (2025), the relationship between the coefficients on the lags of the standard LP and LD specifications is: $\alpha_{j,h} = \sum_{i=1}^{p} \rho_{i,h} \mathbf{1}(i \leq j \leq h+i)$.[1] The analogous restriction also holds for the coefficients of the lags of other included control variables in the specification.

The correctly specified restricted long-differenced specification uses p regressors per variable and sets $p_D = p + h$ to construct the restriction matrix R_h. Nevertheless, the researcher is free to choose both parameters. Let p^* denote the number of lags in the RLD specification, and let p_D^* be the parameter selected by the researcher to construct the matrix R_h. In the aforementioned description of the restricted long-differenced specification, replace p by p^* and p_D by p_D^* to construct the specification with the chosen parameters. Aligning with the goal of the RLD estimator, the choice must satisfy $p^* \leq p_D^*$. Allowing p_D^* to differ from the correctly specified value is likely beneficial, since the latter leads to a considerable reduction in the effective sample size. Specifically, the effective sample size of the RLD specification is equal to that of the LD specification with p_D^* lags of the first difference.

An astute reader may observe that in the case of a self-selected p_D^*, the resulting R_h matrix can contain trailing rows of zeros. This is suboptimal because the final entries of Z_t then correspond to these zero rows. This extends to all control variables. A simple univariate example illustrates the issue: with $p = 1$, $p_D^* = 2$, and $h = 0$, the restriction matrix is $R_0 = (1, 0)'$, implying $Z_t = \Delta y_{t-1}$. However, since ΔY_t is constructed to include two lags of the first difference rather than one, one observation is lost. In finite samples, such loss of observations can be costly. To address this, the implementation of the RLD specification omits any final zero rows in R_h before defining Z_t and X_t^{RLD}.

Furthermore, it should be noted that when $p^* = p_D^*$, the impulse response estimator of the RLD specification and the LD specification with p_D^* lags of the first difference coincide for all horizons. The proof is left to Appendix B.

This restricted long-differenced bias correction may improve on existing methods for bias reduction in local projections for two principal reasons. First, restricting coefficients in this manner reduces the number of estimators and thereby increases the efficiency of the long-differenced estimator. Piger and Stockwell (2025) point out that the econometrician can simply perform lag selection to increase the efficiency. While this is true, the absence of a widely agreed-upon lag selection procedure within the local projection framework creates uncertainty about how many lags to include. It is therefore sensible to consider such a restricted specification. Secondly, this estimator introduces an alternative lag structure by including regressors based on longer

[1] I follow this restriction as in Piger and Stockwell (2025) exactly. However, future research may consider an alternative lag structure that considers more observations close to time t. For instance, the restriction $\alpha_{j,h} = \sum_{i=1}^{p} \rho_{i,h}$ for $j = 1, ..., p$ and $\alpha_{j,h} = \sum_{i=1}^{\min(p, 1+p_D-j)} \rho_{i,h}$ for $j = p+1, \ldots, p_D$ achieves this.

differences, capturing changes over multiple periods. This allows information on longer-term dynamics to be incorporated without increasing the number of regressors, which may improve the estimation of impulse response functions across various horizons.

However, the argument in Section 3.4 regarding the unsuitability of the long-differenced specification under recursive shock identification also applies to the restricted long-differenced specification. Therefore, it is hypothesized that bias reduction will not be effective when shocks are recursively identified.

4 Methodology

In this section, the simulation study will be introduced. This thesis looks into three main models to set up data generating processes. First, a univariate autoregressive model of degree 1 is considered, followed by an empirically calibrated VAR(4) DGP. To further explore the performance of the various bias-corrections, 50 DGPs from an encompassing dynamic factor model (DFM) will be considered, inspired by Adamek et al. (2024) and Li et al. (2024). Consistent with the empirical literature, the simulation study considers sample sizes $T \in \{100, 200\}$. The effective number of observations used for estimation is smaller, however, due to the inclusion of lags and the loss of observations at longer horizons. This reflects empirical practice, where including lagged variables reduces the effective sample size.

4.1 AR(1)

Starting with an AR(1) model is useful because it is illustrative and reveals some fundamental relationships between various bias-corrected estimators. Consider the AR(1) DGP:

$$y_t = \alpha + \theta x_t + \phi y_{t-1} + u_t. \tag{6}$$

In setting up the DGP with Equation 6, the coefficients, distributions, and number of observations must be selected. To simplify as much as possible, consider the settings: $\alpha = 0$, $\theta = 1$, $x_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, and $u_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. The impulse variable x_t is treated as observed. Lastly, the values $\phi \in \{0.90, 0.95, 1.00\}$ are considered. This choice ensures that the processes have high persistence as in macroeconomic time series, with $\phi = 1$ corresponding to a unit root case. The true impulse response function in this simple case is $\theta_h = \phi^h$ for all $h = 0, \ldots, h_{\max}$, as shown in Appendix C.

4.2 VAR(4)

Beyond the theoretical fundamentals, it is of interest to test a wide range of empirically calibrated data generating processes. To start, an empirically calibrated VAR(4) is considered. Christiano et al. (2005) use quarterly U.S. data on 9 macroeconomic variables to estimate a VAR(4), which they use to analyse the effects of a monetary policy shock. The variables are: real GDP, real consumption, real investment, GDP deflator prices, real wages, labour productivity, federal funds

rate, real profits, and the growth rate of M2 (money supply). The data transformations and selected data range are followed as in the online appendix of Herbst and Johannsen (2024). All variables except the federal funds rate and the M2 growth rate enter the VAR in log levels. The data spans 30 years, from quarter 3 of 1965 to quarter 3 of 1995. This corresponds to 121 observations of each variable.

The data generating process is stationary as noted by Herbst and Johannsen (2024) and is set up by estimating the VAR(4). This allows the true impulse response function under the VAR(4) model to be computed. The variables enter the VAR in the order listed above. This ordering implies that the monetary policy shock cannot contemporaneously affect the variables ordered before it, namely real GDP, real consumption, real investment, GDP deflator prices, real wages, and labour productivity.

In the Monte Carlo simulations, a Cholesky factorization of the error covariance matrix is used to identify the shock variable. Let Σ denote the covariance matrix of the reduced-form VAR(4) residuals, and let Q be the lower-triangular matrix from the Cholesky decomposition of Σ, so that $\Sigma = QQ'$, where Q' is the transpose of Q. Each simulation, a sequence of reduced-form shocks $\{u_t\}_{t=1}^{T}$, is independently generated as:

$$u_t \sim \mathcal{N}(0, \Sigma), \quad t = 1, \ldots, T,$$

where u_t is a vector of length 9. These reduced-form shocks are rewritten in terms of structural shocks ε_t as:

$$\varepsilon_t = Q^{-1} u_t.$$

The monetary policy shock is identified as the structural shock corresponding to the federal funds rate equation, under the assumption that the federal funds rate does not contemporaneously affect the other variables ordered before it in the VAR system. Specifically, the impulse variable x_t corresponds to the seventh component of the structural shock vector ε_t. As in Herbst and Johannsen (2024), the shocks are assumed observed by the researcher. This is an important distinction: the shock is identified in a recursive manner, though treated as observed. Following these authors, the local projections are estimated for three different outcome variables, namely real GDP, GDP deflator prices and the federal funds rate.

4.3 Dynamic Factor Model

A new set of DGPs is constructed based on an empirically calibrated dynamic factor model, inspired by Li et al. (2024) and Adamek et al. (2024). As in Adamek et al. (2024), the model is estimated with monthly macroeconomic data on 122 variables from the Federal Reserve Economic Data – Monthly Database (FRED-MD; McCracken and Ng, 2016). The calibration of this DFM uses data from January 1960 to January 2020. Including data from the Great Recession allows for a longer and more recent set of data for calibrating the DFM. However, data from the COVID-19 period is excluded. This simulation study implements a stationary and a non-stationary variant of the DFM model and both observed and recursively identified monetary policy shocks are

considered.

As in the VAR(4) and in the AR(1) DGPs with $\phi < 1$, stationarity refers to cases where the statistical properties of the series, particularly the mean, are constant over time, causing the series to return to a long-run level. It is of interest to study both stationary and non-stationary DGPs to reflect different empirical modelling practices. In applied research, it is common to model data in levels (or log levels), which motivates the consideration of non-stationary DGPs. Less frequently, researchers transform the data to stationarity before estimation, suggesting that stationary DGPs should also be considered.

Dynamic factor models are a class of models that aim to capture the co-movement among a large number of time series using a small number n_f of unobserved factors that evolve over time. Both the stationary and the non-stationary variants of the DFM are considered. For the stationary variant, the FRED-MD data is transformed to stationarity as in Adamek et al. (2024). In the non-stationary variant, no such transformation is applied. Specifically, variables that were previously differenced once are retained in levels, those differenced twice are now differenced once, log-differenced variables are used in log levels, and variables originally in second log differences are instead used in first log differences. Refer to Appendix E for a more detailed explanation of the transformations. In the following explanation of the dynamic factor model, the notation of Li et al. (2024) is followed.

Consider the observed macroeconomic time series X_t of length n_X from the FRED-MD database. Note that the form of X_t (i.e., in differences or in levels) depends on whether the stationary or non-stationary variant is considered. Let the vector f_t denote the n_f latent factors and let v_t be a vector of idiosyncratic components that together drive this system of observed time series. Cointegration refers to when non-stationary variables share a common stochastic trend, such that specific linear combinations of the integrated variables are stationary (Kilian and Lütkepohl, 2017). In the non-stationary case, the factors are modelled using a Vector Error Correction Model (VECM), which captures these cointegrating relationships. In the stationary variant, a VAR(p_f) model is imposed. Note that the VECM can be transformed to VAR(p_f) form after its estimation (Kilian and Lütkepohl, 2017). Thus, in VAR(p_f) form, the factors follow:

$$f_t = \Phi(L) f_{t-1} + H\varepsilon_t, \tag{7}$$

where $\Phi(L)$ denotes a lag polynomial of order p_f, and ε_t is an $n_f \times 1$ vector of aggregate shocks. The shocks ε_t are assumed independent and identically distributed over time and are jointly normally distributed with mean zero and identity covariance matrix. The $n_f \times n_f$ matrix H determines the immediate responses of the factors due to the shocks ε_t. The series X_t are given by

$$X_t = \Lambda f_t + v_t, \tag{8}$$

where Λ is an $n_X \times n_f$ matrix and v_t is an $n_X \times 1$ vector of idiosyncratic components. Each entry of v_t corresponds to an observed variable. The idiosyncratic component of variable i is given by the AR(p_v) process,

$$v_{i,t} = \Gamma_i(L) v_{i,t-1} + \Xi_i \xi_{i,t}, \tag{9}$$

where $\xi_{i,t}$ is independent and identically distributed over time, and $\xi_t = (\xi_{1,t}, \ldots, \xi_{n_X,t})'$ is jointly normally distributed with mean zero and identity covariance matrix.

Concerning the details of the estimation of the DFM, the steps in the Supplementary Appendix of Li et al. (2024) are followed. As in Stock and Watson (2016), the model is calibrated using $n_f = 6$ factors. For Equation 7 and Equation 9, more lags are included than in Li et al. (2024), since monthly data rather than quarterly data is used. Hence, the non-stationary variant estimates the DFM with $p_f = p_v = 6$ and the stationary variant estimates the DFM with $p_f = p_v = 3$. These lag lengths are between what is preferred by AIC and BIC. The VECM model for the non-stationary variant is estimated with the URCA package in R.[2] The Johansen test with a 5% significance level selects a cointegration rank of 2 for the factor VECM. This supports the use of the VECM and indicates the presence of two linearly independent cointegrating relationships, leaving four common stochastic trends that underlie the system's dynamics (Kilian and Lütkepohl, 2017). For the stationary variant, the replication code from Adamek et al. (2024) is used. Following their code, a maximum absolute eigenvalue of 0.98 is imposed on the companion forms of the matrices Φ and Γ in Equation 7 and Equation 9, respectively, to ensure the process remains stationary.

From the DFM estimated on all 122 variables, we construct 50 distinct DGPs, each based on a sample of 5 variables. As in Li et al. (2024), the federal funds rate is included in every selection and is always ordered last in the recursive identification scheme. Moreover, each selection includes at least one price measure and at least one measure of economic activity.[3] The remaining two variables are randomly selected, ensuring five distinct variables are chosen. The outcome variable is ordered first and is randomly chosen from the four selected variables, where the federal funds rate is not considered as an outcome variable.

In the case of an observed shock, the matrix H is selected to maximize the impact of the first component of the structural shock vector ε_t on the federal funds rate, subject to the constraint that H is a valid decomposition of the reduced-form error covariance matrix (Li et al., 2024). Let H_1 denote the first column of H.

Note that the estimands differ between the observed and recursively identified shock cases. For observed shocks in the non-stationary case, the estimated VECM coefficients must first be converted into VAR form. These VAR coefficients are then transformed into their moving average representation (Kilian and Lütkepohl, 2017) and using the vector H_1, the impulse responses of the latent factors to the structural shock are computed. The estimand is then obtained by projecting these responses onto the observed outcome variable, using the corresponding row of the factor loading matrix Λ from Equation 8. For recursively identified shocks, the dynamic factor model must first be rewritten in its ABCD form (Fernández-Villaverde et al., 2007) in order to compute the impulse response functions. Further details on this procedure can be found in the Supplementary Appendix of Li et al. (2024).

Considering 50 DGPs permits a wide variety of empirically relevant impulse responses to be

[2]The *ca.jo* function in the URCA package estimates a VECM and performs a Johansen test. The employed settings are: *type = eigen, ecdet = trend, spec = transitory.*

[3]One price measure is randomly selected from the variables listed in Table S.8, and one activity variable is randomly selected from those listed in Tables S.2 and S.4 in the Supplementary Appendix of Adamek et al. (2024).

considered. Inspired by the summary statistics in Li et al. (2024), Table 1 presents similar and additional summary statistics of the stationary and non-stationary DGPs under observed shock identification. The summary statistics for the non-stationary DGP with recursively identified shocks are left to Appendix A.6. For the non-stationary DGPs, the number of interior local maxima indicates that the impulse responses are largely monotonic. This also aligns with the median horizon of the maximum absolute value being 20. This is likely due to the use of monthly data and thus considering impulse responses less than two years ahead ($h_{\max} = 20$). There appears to be more variation in the stationary DGPs. Notably, the horizon of the maximum absolute value is considerably smaller in the stationary case. Across both variants of the DFM, the (scaled) initial response indicates variables immediately move in both positive and negative directions in response to the monetary policy shock.

Non-stationary DGPs				
	Min	Median	Mean	Max
Sign persistence	7	21	20.44	21
Average/(max absolute value)	-0.89	0.70	0.47	0.88
Number of interior local extrema	0	0	0.48	3
Horizon of max absolute value	11	20	19.26	20
Initial response	-0.56	0.31	0.24	0.71

Stationary DGPs				
	Min	Median	Mean	Max
Sign persistence	0	12	10.70	20
Average/(max absolute value)	-0.89	-0.14	-0.08	0.83
Number of interior local extrema	0	1	1.84	5
Horizon of max absolute value	0	0.5	2.80	20
Initial response	-3.69	0.53	-0.05	3.39

Table 1: Summary statistics of the 50 non-stationary and 50 stationary DGPs for observed shocks and horizons $h = 0, \ldots, 20$. Sign persistence refers to the number of horizons with the same sign as θ_0. Average/(max absolute value) is given by $\left(\frac{1}{21}\sum_{h=0}^{20}\theta_h\right) / \max_h |\theta_h|$. Initial response is $\theta_0 / \sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$.

4.4 Estimation of Local Projections

Once the shock is identified, either observed or recursively identified, the impulse responses can be estimated with the four estimators. First, some commonalities of all estimation procedures are described before explaining differences in specifications depending on how the shock is identified.

First, the maximum horizon, $h_{\max}$, is set to 20 throughout. In the AR(1) case, this implies 20 periods, in the VAR(4) case this is 20 quarters, and for the DFM this corresponds to 20 months. This ensures a far enough estimation of shocks into the future and is consistent with the literature on local projections in simulation studies (e.g., Montiel Olea et al., 2025; Herbst

and Johannsen, 2024). The considered sample sizes are $T \in \{100, 200\}$ for every simulation. Moreover, in drawing the data in each simulation, 100 burn-in observations are performed to remove initialization bias.

With respect to the estimation of impulse response functions, the approach taken tries to mirror that of an econometrician without prior knowledge of the DGP. Therefore, an intercept is included in every specification performed. Moreover, the standard local projection estimates always include a linear time trend. This is unnecessary for the long-differenced specification due to the nature of the regressors used. Following this, the RLD specification does not include a linear time trend either. All regressions are performed by OLS and lag selection is discussed in Section 4.5.

Regarding the set-up of the local projection specifications, the standard and long-differenced specifications need to be make explicit, with the BCC and RLD estimators following as described in Section 3. Consider the specifications with observed shock identification and let ε_t be the observed shock. Then, in Equation 1 and Equation 3, set $x_t = \varepsilon_t$. For the standard LP specification, lags of all variables, including the outcome variable, are included in levels. For the long-differenced specification, the same variables are included as lags, although now in first differences. As in Herbst and Johannsen (2024) and Piger and Stockwell (2025), but unlike Li et al. (2024), no lags of the observed shock are included. Robustness checks clearly indicate that the difference between including and excluding such lags is negligible.

Now consider the specifications in the case of a recursively identified shock, which is only considered in the DFM. For the DFM, only monetary policy shocks are considered, hence the impulse variable is defined as the federal funds rate (Li et al., 2024). Lagged controls are included for all variables (including the impulse variable) in levels for the standard LP and in first differences for the LD specification. In addition, the specifications for recursively identified shocks must include contemporaneous controls for consistency. In the standard LP specification, contemporaneous regressors of all variables ordered before the policy variable in the identification scheme are included (Plagborg-Møller and Wolf, 2021). The federal funds rate is ordered last, so all other variables are included contemporaneously. The long-differenced specification does not have an established way of dealing with recursively identified shocks, since Piger and Stockwell (2025) claim that the long-differenced estimator is only suitable for identification with observed shocks or instrumental variables. Therefore, they do not propose a long-differenced regression specification for recursively identified shocks. To provide a complete evaluation of the existing bias corrections, an appropriate specification of the long-differenced specification must be defined. In adjusting the specification as little as possible, a regressor Δz_t is added to the set of regressors for all those variables z_t ordered before x_t.

4.5 Lag Selection

Critically, the number of lags must be selected in setting up the standard, long-differenced, and restricted long-differenced specifications. This is not necessary for the BCC estimator by definition of the estimator. There is no widely agreed-upon method for lag selection in the

local projection literature, particularly for the long-differenced specification. This hampers the evaluation of the restricted long-differenced estimator because one of its main benefits potentially lies in the fewer regressors. Nevertheless, two types of lag selection are considered.

The first approach builds on the equivalence between local projections and VARs at horizon $h = 1$ (Jordà and Taylor, 2025). This involves the estimation of an auxiliary VAR that includes the outcome variable, impulse variable, and other controls that are to be included in the local projection. Information criteria can then be applied to the VAR and then the selected lag length p is used for all horizons. This thesis uses the Akaike Information Criterion (AIC) with a maximum lag length of 15. This approach is referred to as the VAR-AIC lag selection henceforth. This lag selection for the standard LP has been used in recent work (e.g., Jordà and Taylor, 2025; Li et al., 2024; Montiel Olea et al., 2025). Importantly, the direct relation between the standard local projection and long-differenced specifications then suggests $p_D = p + h$ lags should be imposed in the long-differenced specification. Moreover, the RLD can be specified with $p^* = p$ lags and with $p_D^* = p + h$. The clear downside to this approach for the long-differenced specification is that it includes up to $p + h_{\max}$ lags of the first difference, and consequently may lead to more coefficients to estimate than available observations. Therefore, results for this approach are only presented for the AR(1) DGP.

Piger and Stockwell (2025) employ an alternative lag selection procedure for both the standard LP and LD specifications. In this procedure, both the number of lags for the LP specification, p, and the number of lags of the first difference for the LD specification, p_D, are chosen based on a test-down procedure. Such a test-down procedure is implemented by starting with a maximum lag length and sequentially testing the joint significance of the last lag of all variables using an F-test. If the final lag is not jointly significant at the 5% level, the lag length is reduced by one, and the test is repeated. This process continues until the last lags are jointly significant, at which point the corresponding lag length is selected. This procedure is applied at the first horizon where the true impulse response differs from zero: at $h = 0$ for observed shocks (and for the VAR(4) simulations with the federal funds rate as the outcome), and at $h = 1$ for simulations with recursively identified shocks. The selected lag length is then used for all horizons. The initial maximum lag length is set to the smaller of (i) 15, and (ii) the largest lag length such that the total number of coefficients to estimate remains smaller than the number of observations. There appears to be no analogous manner for lag selection in the RLD specification for two reasons. First, the lag structure of the RLD specification differs over horizons, making a test-down procedure inappropriate. Second, not only the number of lags p^*, but also the parameter p_D^* must be selected. As an alternative, the number of lags is set equal to the minimum of the two selected lag lengths, $p^* = \min(p, p_D)$, and $p_D^* = p_D$. This approach is sensible because it limits the number of lags in the RLD specification to be no greater than in the LD specification, aligning with the goal of reducing the variance of the LD estimator. Moreover, it ensures that the effective sample sizes of the RLD and LD specifications coincide under the test-down procedure. Note that, by the argument in Appendix B, the estimated impulse responses by the RLD and LD estimators coincide when $p^* = p_D^*$, or equivalently when $p \geq p_D$.

The simulation study primarily employs the second of the lag selection procedures. Specifically,

for the AR(1) DGP both lag selections are considered, while for the VAR(4) and DFM data generating processes, only the test-down procedure is considered.

4.6 Metrics

The four estimators of the impulse response functions will be evaluated on three main measures that align with the existing literature. These metrics are the bias, standard deviation, and coverage probability of the respective estimators. The Monte Carlo mean of the absolute bias and standard deviation are taken across simulations and divided by the root mean square of the true impulse response, $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$. This normalization is particularly important in the simulations for the DFM, since this relies on a comparison across DGPs. For the DFM, the median is taken across DGPs to reduce the effect of potential outliers that may arise from randomly selecting combinations of five variables to set up the DGPs.

Additionally, inspired by Li et al. (2024), the loss function is considered for each of the estimators. Let $\hat{\theta}_h$ be the estimator of θ_h. Then, the loss function is defined as,

$$\mathcal{L}\left(\theta_h, \hat{\theta}_h\right) = \lambda\left(\mathbb{E}\left[\hat{\theta}_h - \theta_h\right]^2\right) + (1-\lambda)\mathrm{Var}\left(\hat{\theta}_h\right),$$

where the expectation and variance in this expression refer to the Monte Carlo mean and variance, and λ is simply a weight. Based on the bias and variance measures alone, the estimator with the lowest loss function value is preferred for that horizon and assigned weight. In essence, this is the mean squared error with different weights. For λ closer to 1, the researcher is primarily concerned with bias, whereas with λ closer to zero, the researcher is concerned with estimator variance. As local projections should only be preferred over VARs if the researcher is primarily concerned with bias, the weights $\lambda \in [0.5, 1]$ are considered.

Importantly, the loss function is normalized in the DFM simulations to allow for comparison across DGPs. The scaling factor is adjusted to account for the squaring of the bias and standard deviation in the loss function, resulting in a normalization factor $\frac{1}{21}\sum_{h=0}^{20}\theta_h^2$. Then, the mean loss function across DGPs is considered as in Li et al. (2024) to determine the optimal estimator for each weight-horizon combination. Using the mean is appropriate, since potential outliers in either bias or standard deviation can be offset by the other in the loss function.

Coverage is computed based on pointwise 95% confidence intervals. The standard errors are estimated with both the Newey-West and Eicker–Huber–White standard errors. The relevance of computing coverage based on two difference standard error estimators follows from the ongoing research into this aspect of inference with local projections. In particular, Jordà (2005) suggests the use of HAC standard errors, such as Newey-West standard errors, whereas Herbst and Johannsen (2024) provide evidence that the Eicker–Huber–White standard errors perform better in small samples. The use of alternative standard error estimators such as the equally weighted cosine estimator proposed by Lazarus et al. (2018), as well as the implementation of joint inference methods discussed in Inoue et al. (2025), is left to future work.

For the Newey-West standard errors, the lag truncation parameter (or lag length) must be selected. Jordà (2005) suggests that the selected lag length must increase with the horizon of the

considered impulse response. Using the AR(1) DGP and the correctly specified standard local projection specification, the calculation of the auto-covariance shows that $h+1$ lags must be included for the auto-covariance to be zero. Thus, all computations of Newey-West standard errors employ a lag truncation parameter of $h+1$. Refer to Appendix D for the proof in the AR(1) case.

5 Results

In this section, the main findings from the Monte Carlo simulations are discussed. The main results for the sample size $T = 200$ are displayed, while the corresponding results for $T = 100$ are left to the appendices. Section 5.1 distinguishes between the results generated based on the VAR-AIC lag selection and the test-down lag selection. All further results are based on the test-down lag selection.

5.1 Baseline Results

The results to the AR(1) DGP are considered as a baseline, because they give insights into some of the theoretical properties of local projections and the closely related bias-variance trade-off. Figure 1 shows the mean normalized absolute bias and standard deviation across 5000 simulations for $T = 200$. The results for both types of lag selection are presented for $\phi \in \{0.95, 1\}$. As expected, both the bias and in particular the standard deviation increase with the horizon for each estimator. The increase in standard deviation is likely caused by the growing temporal distance between the shock and the outcome variable, which reduces the precision of the estimates.

Most noticeably, there is a large disparity in the bias of the four estimators. The standard LP specification has the largest bias and lowest standard deviation at all horizons and for both degrees of persistence. The BCC estimator eliminates some bias at the cost of a slightly higher standard deviation. The long-differenced and restricted long-differenced specifications reduce the bias further, to near-zero levels for $\phi = 0.95$, again at the cost of higher standard deviations. This bias-variance trade-off directly relates to a core element of this literature. Montiel Olea et al. (2025) and Li et al. (2024) demonstrate that local projections are preferred over VARs only when the researcher strongly prioritizes low bias over variance. These initial results demonstrate that such a trade-off also exists for bias corrections for local projections: a lower bias comes at the cost of increased variance in the estimator. An exception to this trade-off is observed for $\phi = 0.95$ in the test-down selected lag length, since the LD and RLD estimations show both smaller bias and standard deviation than the BCC estimations. Moreover, the trade-off is not necessarily a linear one. In particular, while the bias is considerably different at all horizons $h \geq 2$, the standard deviations only tend to diverge at longer horizons.

For the lag length selected with the VAR-AIC approach, a lag length of one is almost always chosen. The median lag length is 1 for all values of ϕ and, for instance, the mean lag length selected for $\phi = 1$ is 1.02. This implies that in almost all simulations $h+1$ lags of the first difference are included in the LD specification and 1 lag in the RLD specification. For $\phi = 1$, the

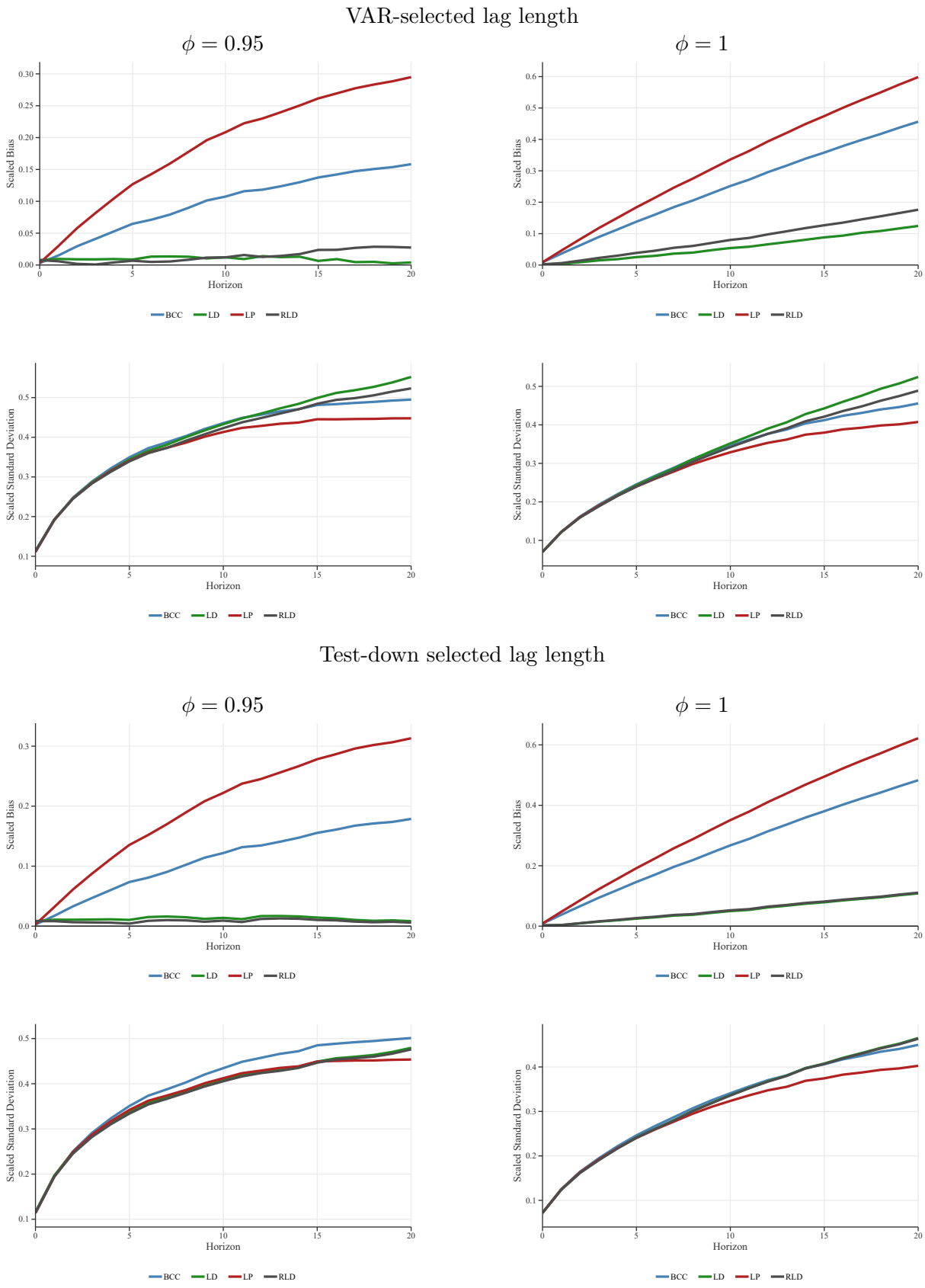


Figure 1: Mean absolute bias and standard deviation of different estimators, normalized by $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$, across two lag selection methods, for $\phi \in \{0.95, 1\}$. All results based on 5000 simulations.

estimates from the long-differenced specification consistently shows lower bias than the RLD estimates, although this cannot be generalized to the $\phi = 0.95$ instance. Moreover, at long horizons, the increased number of regressors in the LD specification leads to a higher standard deviation than for the other estimators. This is particularly true for the smaller sample size $T = 100$ in Appendix A.2.

Appendix A.7 reports the summary statistics of the test-down lag selection procedure, which generally selects more lags than the VAR-AIC approach. Figure 1 shows that the biases and standard deviations of the LD and RLD estimators are almost identical under this lag selection, which can be explained by two related factors. First, the two specifications are very similar at small lag lengths. Second, whenever the selected lag lengths for the LP estimator, p, is greater than or equal to the lag length for the LD specification, p_D, the RLD and LD impulse response estimators coincide. This is in contrast with the results based on the VAR-AIC lag selection, where the RLD estimator shows lower variance than the LD estimator.

Beyond the infeasibility of the VAR-AIC approach for the long-differenced specification with more variables, Figure 1 indicates the test-down selection reduces the variance of the LD estimator at almost no additional cost of bias. For instance, for $\phi = 1$ and at horizon $h = 20$, both the normalized standard deviation of the LD estimator is lower (0.465 vs. 0.524) and the normalized bias is lower (0.108 vs 0.124) when the test-down procedure is used. This suggests that the LD specification should only be used in a lag-restricted setting and should not be implemented as the direct derivation of the LD specification implies.

Figure 1 also illustrates that the bias and variance increase with the level of persistence of the underlying DGP. The results for $\phi = 0.90$, relegated to Appendix A.1, reinforce this observation and also follow a similar pattern among estimators as observed in Figure 1. Furthermore, the complete set of results with sample size $T = 100$ are presented in Appendix A.2. Notably, the bias-variance trade-off becomes relatively larger; that is, the differences in bias and standard deviation between estimators is greater for the sample size $T = 100$.

5.2 Results from the Empirically Calibrated VAR(4)

While the results from the AR(1) model provide useful insights, they lack empirical relevance. In contrast, the empirically calibrated VAR(4) DGP enables the estimation of impulse responses to a monetary policy shock. Figure 2 displays the normalized bias and standard deviation, as well as the coverage probabilities of the four bias-correction methods. The variables of interest are real GDP, inflation, and the federal funds rate, with a sample size of $T = 200$. Coverage probabilities are computed using both Newey-West standard errors (solid lines) and Eicker-Huber-White standard errors (dashed lines).

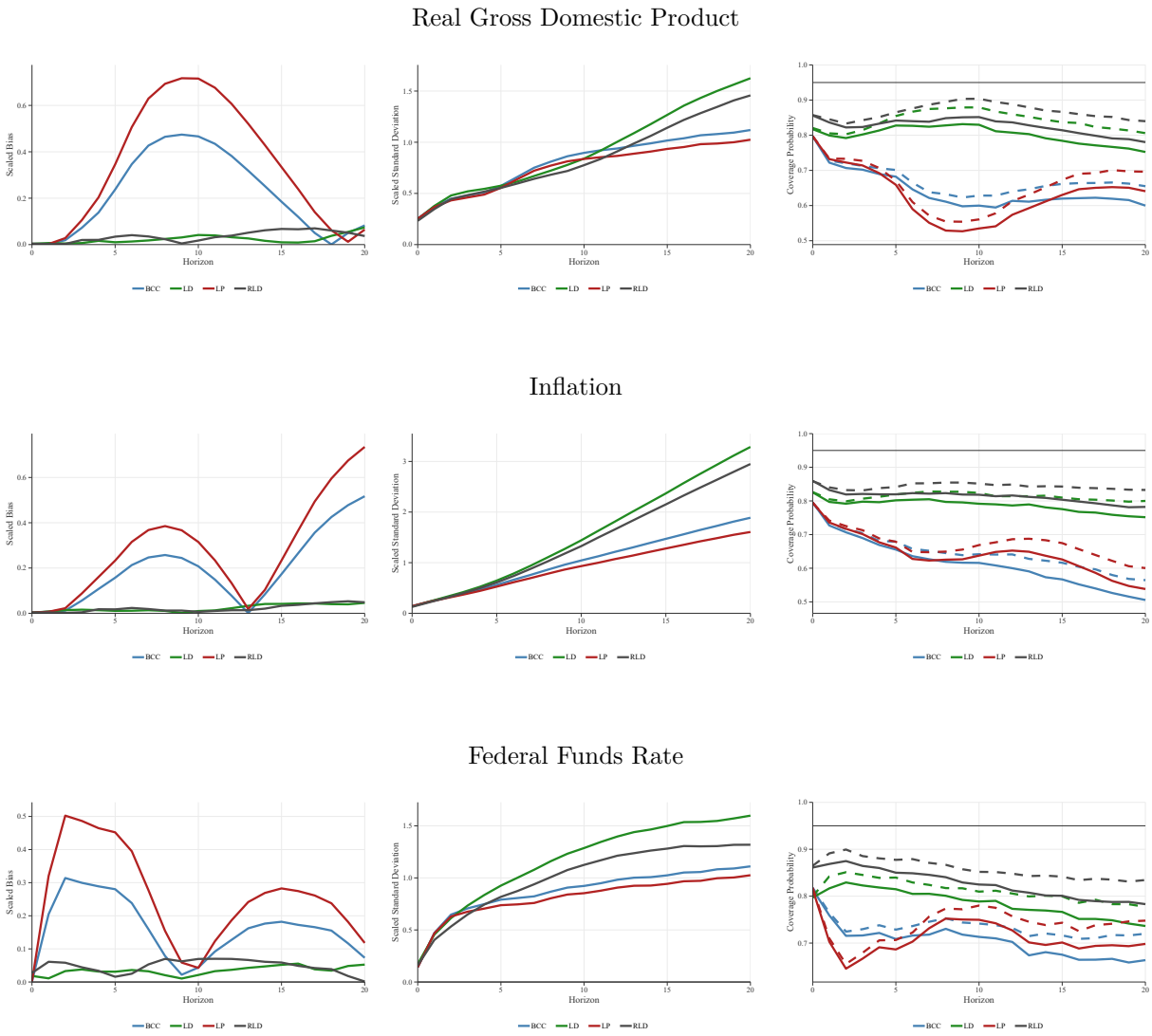
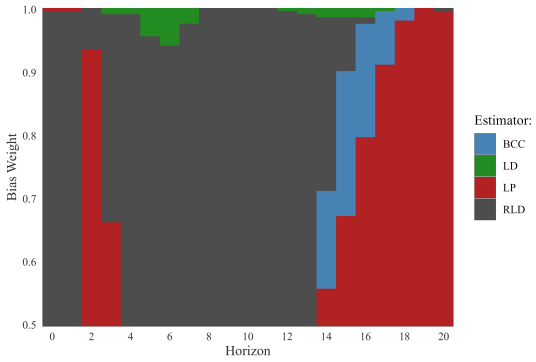


Figure 2: Mean absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$, along with coverage probabilities. Confidence intervals are constructed at the 95% level using Eicker–Huber–White (dashed) and Newey–West (solid) standard errors. Results are shown for three U.S. macroeconomic variables: real GDP, inflation, and the federal funds rate. All figures are based on 5000 simulations and sample size $T = 200$.

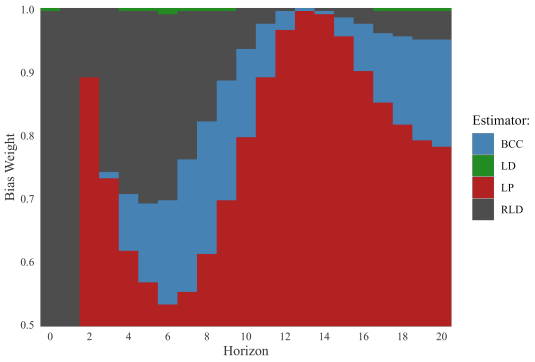
In contrast to the AR(1) case, the bias in the VAR(4) setting does not increase monotonically over horizons, illustrating that theoretical expectations regarding estimator performance may diverge from empirical results. Nonetheless, standard deviations generally rise with the horizon, and coverage probabilities tend to decrease. As in the AR(1) Monte Carlo simulations, the standard local projection estimator has the highest bias across most horizons and outcome variables, while simultaneously showing the lowest standard deviation. The other estimators reduce bias but at the expense of higher variance. Moreover, at longer horizons, particularly when inflation is the dependent variable, the standard deviations of the LD and RLD estimators increases substantially relative to that of the standard LP. Lastly, the RLD estimator shows a relatively larger decrease in standard deviation compared to the LD estimator than in the AR(1)

simulations.

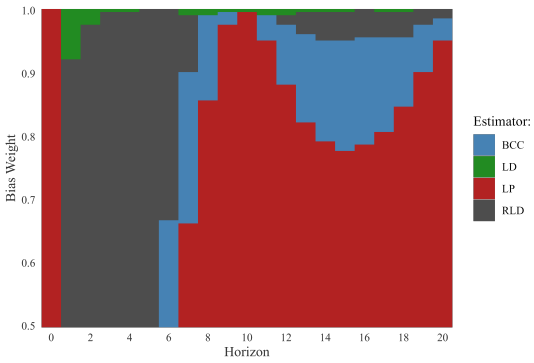
The coverage probabilities shown in Figure 2 suggest that confidence intervals constructed using the restricted long-differenced estimator most closely match the intended 95% confidence level across all horizons and outcome variables. Although the differences are slightly smaller in the case of the $T = 100$ sample size, as shown in Appendix A.3, the overall conclusion remains unchanged. This result is interesting given that the RLD estimator tends to have slightly higher bias than the LD estimator, suggesting that its confidence intervals may, on average, be wider, and thereby more accurately reflect the true estimation uncertainty. The BCC and standard LP estimators exhibit similar coverage patterns, varying by horizon and variable, but their coverage levels are consistently lower than those of the LD and RLD estimators. This is primarily due to the considerably larger bias of these estimators. Furthermore, confidence intervals based on Eicker-Huber-White standard errors provide higher coverage than those based on Newey-West standard errors. This indicates that Newey-West estimates underestimate the standard errors, resulting in confidence intervals that are too narrow. The robustness of this finding across all horizons, estimators, and outcome variables supports the use of heteroscedasticity-robust standard errors over HAC standard errors.



(a) Real GDP



(b) Inflation



(c) Federal Funds Rate

Figure 3: Method that minimizes the loss function across outcome variables real GDP, inflation and the federal funds rate, with weights $\lambda \in [0.5, 1]$, at 0.005 intervals. Results are based on 5000 simulations and a sample size $T = 200$.

Figure 3 presents a direct comparison of the estimators across horizons and different values of the loss function weight λ. When two estimators yield identical loss function values, the preference order is: standard LP, BCC, LD, RLD. The standard LP is favoured in the case of ties due to its computational simplicity. The lower standard deviation of the LP estimator makes it appealing when the bias has a lower weight in the loss function. However, as shown by Li et al. (2024), VAR-based approaches tend to be more suitable for low bias weights. At higher bias weights, estimators with bias corrections generally perform better. The LD estimator is preferred only at certain horizons and when bias minimization is a researcher's only concern. Both the BCC and RLD estimators cover substantial regions of the figures, with the RLD often being the more favourable choice at intermediate horizons and when the bias weight is large.

For the sample size $T = 100$, Appendix A.7 shows that the number of selected lags is similar for the LP and LD specifications. The higher variance of the LD estimator, as observed in Appendix A.3, therefore arises from the structure of the regression itself, rather than from the inclusion of additional controls. As expected, the biases and standard deviations are larger, while coverage probabilities are lower compared to when the sample size is $T = 200$. Despite the differences in magnitude, the overall patterns across horizons remain similar. The estimators that minimize the loss function are also displayed. These is largely similar to those in Figure 3, though the BCC estimator covers a larger portion of the parameter space when $T = 100$.

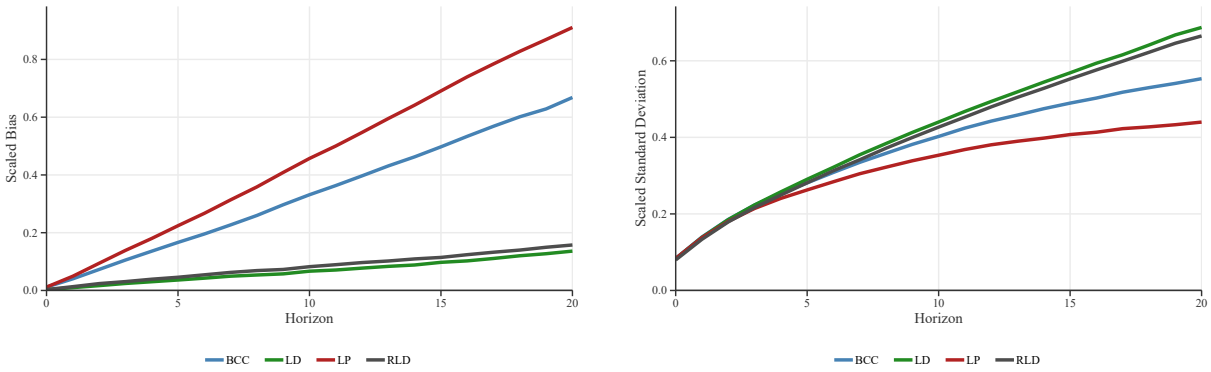
5.3 DFM Results

This section presents the results of the simulation study based on the dynamic factor model introduced in Section 4.3. In contrast to earlier sections, the analysis here covers both stationary and non-stationary DGPs. Moreover, it considers both observed and recursively identified shocks. These extensions allow for a more complete evaluation of estimator performance across a range of empirically relevant DGPs.

5.3.1 Non-Stationary DGPs and Observed Shocks

Consider the case where the data is not initially differenced and a VECM model is imposed in Equation 7 so that the time series may be non-stationary. Moreover, as in previous Monte Carlo simulations, the shocks are observed. Figure 4 shows the median bias and standard deviation across the 50 non-stationary DGPs. The bias is substantially lower for the LD and RLD estimators than for the standard LP. This does however come at the cost of higher standard deviations. As observed in previous results, the BCC estimator lies somewhere between the extremes, both in terms of bias and standard deviation.

Combining the two measures, Figure 5 illustrates the method that minimizes the mean loss function across DGPs, at each horizon and for different weights on the bias term in the loss function. The figure shows that the standard LP estimator performs best at short horizons and for low bias weights. The BCC estimator covers a large region of the parameter space, while the LD estimator is only preferred when the bias weight is close to 1. The RLD estimator is primarily favoured in the region with high bias weights and long horizons.



(a) Bias (b) Standard Deviation

Figure 4: Median across 50 non-stationary DGPs of the (mean) absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$. Shocks are observed. Results are based on 1000 simulations per DGP and sample size $T = 200$.

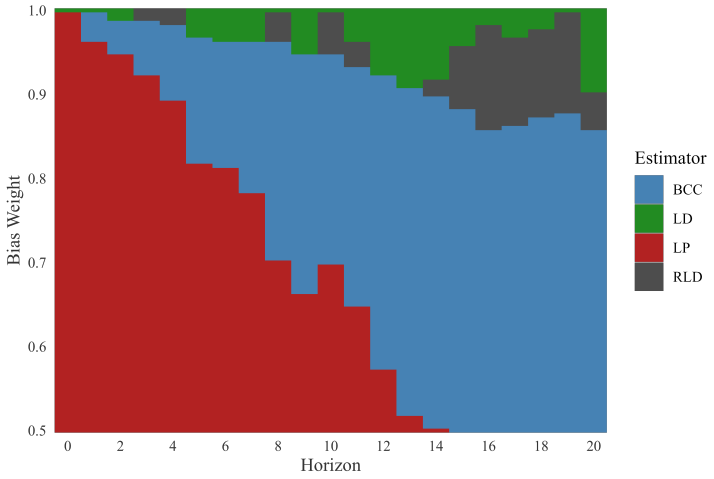


Figure 5: Method that minimizes the mean loss function across 50 non-stationary DGPs, with weights $\lambda \in [0.5, 1]$, at 0.005 intervals. The loss functions are normalized by $\frac{1}{21}\sum_{h=0}^{20}\theta_h^2$. Shocks are observed and the sample size is $T = 200$.

Figure 6 shows the median coverage probabilities across 50 DGPs. The dashed lines represent coverage based on Eicker–Huber–White standard errors, whereas the solid lines correspond to Newey–West standard errors. Coverage is significantly higher for the LD and RLD estimators compared to the standard local projection specification. This is likely due to the substantial bias of the standard LP estimator observed in Figure 4. The BCC estimator also exhibits low coverage, particular at longer horizons. The LD and RLD specifications yield high and nearly identical coverage. Additionally, the figure shows that Eicker–Huber–White standard errors lead to confidence intervals with better coverage than those based on Newey–West, further supporting the use of heteroskedasticity-robust standard errors over HAC estimators in local projections.

The results for the sample size $T = 100$ are presented in Appendix A.4. While the biases

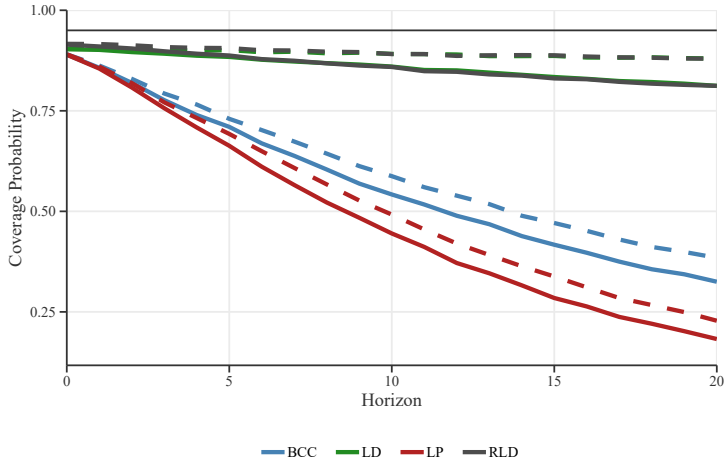


Figure 6: Median coverage probabilities across 50 non-stationary DGPs. Confidence intervals are constructed at the 95% level using Eicker–Huber–White (dashed lines) and Newey–West (solid lines) standard errors. Results are based on 1000 simulations per DGP and sample size $T = 200$.

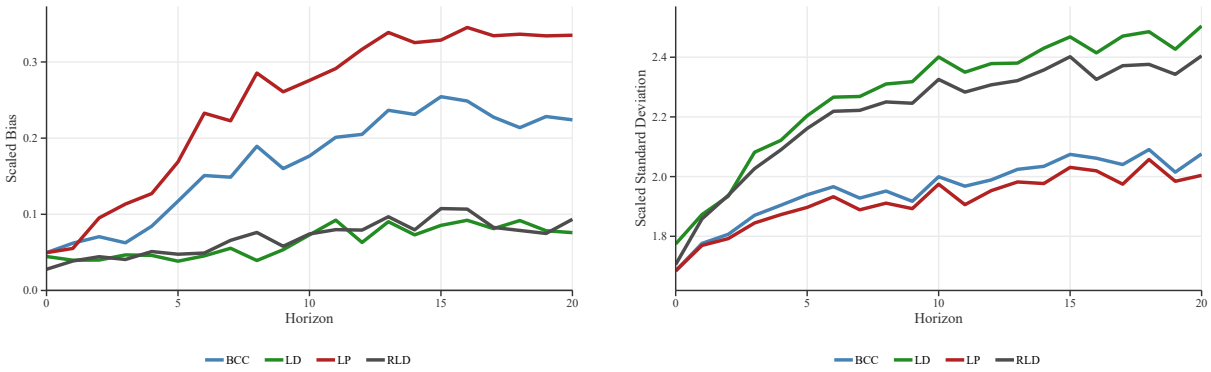
and standard deviations are higher across all estimators and the coverage probability is lower, the results are qualitatively similar in terms of bias and standard deviation. Notably, the RLD estimator is selected as the optimal estimator over a much larger region of the parameter space, especially at longer horizons. Its coverage is also slightly higher than that of the LD estimator.

5.3.2 Stationary DGPs

Researchers may transform data prior to estimation to ensure stationarity, so the stationary variant of the DFM is also considered. Figure 7 shows the median bias and standard deviation across 50 stationary DGPs, where the shock is observed. Interestingly, the lines are more jagged than in the non-stationary cases. This is likely due to the lower magnitude of the bias in the stationary DGPs compared to previous simulations. Robustness checks show that instead of a consistent directional bias, the estimated impulse response functions tend to alternate around the true IRF. This may explain the fluctuations in the figure. Limited computer power prevents more Monte Carlo simulations from being performed to potentially smoothen the plots.

Nevertheless, Figure 7 displays trends that are qualitatively similar to those in the non-stationary case. Once again, the standard LP and LD specifications lie on opposite ends of the bias-variance spectrum. Both the LD and RLD estimators have lower bias but higher variance compared to the BCC and LP estimators. The median standard deviations of the LD and RLD estimators are substantially larger than those of the LP and BCC estimators, making the BCC estimator the most appealing across most horizons and weights of the loss function.

Figure 8 shows the median coverage across DGPs using Eicker-Huber-White (dashed) and Newey-West (solid) standard errors. As expected with the lowered bias, coverage probabilities are higher across all estimators compared to the non-stationary case. It is striking to observed the higher coverage achieved by the RLD estimator compared to any of the other estimators at all horizons and with both types of standard errors. This provides strong evidence of the RLD



(a) Bias (b) Standard Deviation

Figure 7: Median across 50 stationary DGPs of the (mean) absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$. Shocks are observed. Results are based on 1000 simulations per DGP and sample size $T = 200$.

estimator's performance for inference under stationary DGPs. This is reinforced by the results for the sample size $T = 100$ in Appendix A.5. Interestingly, the BCC estimator shows slightly lower coverage than the standard LP estimator over most horizons. Since both estimators use the same standard errors to construct confidence intervals and the BCC estimator has lower median bias across DGPs, this suggests that the higher variance of the BCC estimator negatively affects its coverage probabilities.

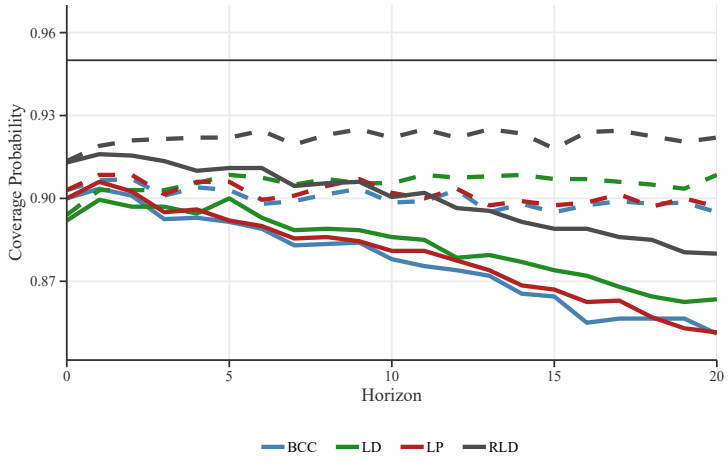
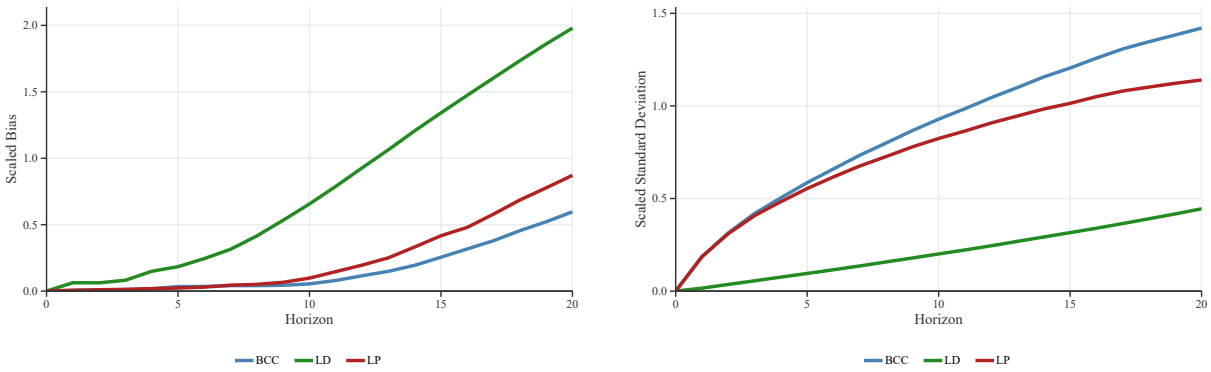


Figure 8: Median coverage probabilities across 50 stationary DGPs. Confidence intervals are constructed at the 95% level using Eicker–Huber–White (dashed lines) and Newey–West (solid lines) standard errors. Results are based on 1000 simulations per DGP and sample size $T = 200$.

5.3.3 Recursively Identified Shocks

Figure 9 displays the median bias and standard deviation across 50 non-stationary DGPs for recursively identified shocks and with a sample size of $T = 200$. Results for the RLD estimator are omitted, as they are nearly indistinguishable from those of the LD estimator. This is because

of the relatively large scale of the bias and standard deviation axes, which makes the differences between LD and RLD negligible compared to those among the other estimators.



(a) Bias (b) Standard Deviation

Figure 9: Median across 50 non-stationary DGPs of the (mean) absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$. Shocks are recursively identified. Results are based on 1000 simulations and sample size $T = 200$.

As argued in Section 3.4, the long-differenced specification is inappropriate when shocks are identified recursively. The same limitation applies to the restricted long-differenced specification. The results presented in Figure 9 corroborate this point: the long-differenced estimator displays substantially higher bias compared to both the standard local projection and BCC estimators. This stands in stark contrast to the small bias observed in earlier results based on observed shocks. Interestingly, the long-differenced estimator shows lower standard deviation than the other methods; however, given that bias is the primary concern in local projection applications, this reduction in variance does not compensate for the considerable bias. Therefore, the long-differenced approach is unsuitable under recursively identified shocks. Robustness checks show that the inclusion of contemporaneous controls in levels rather than in first differences in the LD specification results in similar conclusions. Instead, researchers should prefer the BCC or standard LP estimator, with the BCC estimator offering lower bias.

The corresponding results for $T = 100$ are reported in Appendix A.6. A direct comparison with the $T = 200$ sample size is possible, as the scaling term in the denominator remains unchanged across sample sizes. For the LD estimator, the bias remains nearly unchanged between the two sample sizes. This reinforces the unsuitability for estimating impulse responses with the LD and RLD estimators under recursive identification, since it indicates that the bias persists across sample sizes. The LP and BCC estimators, however, show larger biases at $T = 100$, as expected due to the smaller sample size.

5.4 A Note on Test-Down Lag Selection

The presented results have largely been based on the test-down lag selection procedure. Appendix A.7 summarizes the number of lags selected in each of the DGPs using this procedure. Generally, more lags are selected for $T = 200$ than for $T = 100$. Whether more lags are chosen for

the standard LP or the LD specification depends on the DGP. For the AR(1) and VAR(4) DGPs, a comparison of the median selected lag length with the correctly specified lag length indicates that the test-down procedure yields closer results at $T = 100$ than at $T = 200$. Specifically, for $T = 200$, the chosen lag length tends to exceed the lag length of the DGP. Overall, the test-down procedure works approximately as it should and the initially selected 5% significance level is appropriate.

The restricted long-differenced specification must also be discussed in the context of the test-down lag selection. The RLD specification relies on the selection of two parameters, p^* and p_D^*, where p^* is the number of lags and p_D^* is required to set up the restriction matrix R_h. This complicates matters in terms of the implementation of the RLD independently of other specifications in empirical settings. Recall that p and p_D represent the lag lengths selected by the test-down procedure for the standard LP and LD specifications, respectively. If using the test-down procedure, this implies that the researcher needs to perform two lag selections and find $p < p_D$ for the RLD specification to show results differing from the LD specification. For some DGPs, this is observed consistently, such as in the stationary DFM and VAR(4) simulations. Consequently, the results indicate larger disparities between LD and RLD estimations and the benefits, such as lower variance and higher coverage, associated with the RLD specification are more pronounced.

The issues surrounding the selection of p^* and p_D^* also pertain to the discussed simulation results. Given that the outcomes are reported in averages over Monte Carlo simulations, the performance of the LD and RLD specifications may seem more similar than would be observed in empirical applications. This is because in the simulations where $p^* = p_D^*$, the LD and RLD estimations of the impulse response coincide, causing the absolute differences in bias, standard deviation, and coverage between the two estimators to potentially be larger in empirical practice. Note that this is strictly referring to empirical practice and the theoretical properties are well-represented by the Monte Carlo simulations. An empirical researcher will not employ the RLD estimator with $p^* = p_D^*$ alongside the LD estimator as it provides no new information. Consider a DGP where the RLD estimator consistently shows greater bias than the LD estimator, then the bias observed in empirical applications is hypothesized to be slightly larger than in the simulations. Similarly, the lower standard deviation of the RLD estimator is likely to be more pronounced. Nonetheless, robustness checks with the VAR(4) in which $p^* < p_D^*$ is artificially imposed show that the changes to the original results are marginal, and the conclusions, including those regarding coverage probabilities, are unchanged. Note that this artificial restriction is not supported by a formal lag selection procedure. A well-established method for lag selection in local projections more generally would be valuable for revisiting and formalizing this discussion.

6 Empirical Case: The Effect of Monetary Policy and Information Shocks

This section applies the various estimators studied throughout this thesis to an empirical question, thereby demonstrating that the choice of bias-correction is empirically relevant. Jarociński and Karadi (2020) investigate the effects of Federal Reserve monetary policy shocks and information shocks on five U.S. macroeconomic and financial variables. As in Piger and Stockwell (2025), this re-examination focusses on two outcomes, namely industrial production and the consumer price index (CPI).

6.1 Empirical Design and Identification Framework

Jarociński and Karadi (2020) combine sign restrictions and high-frequency identification to separately identify monetary policy and central bank information shocks. The authors then estimate the impulse response functions using a Bayesian structural VAR. Sign restrictions involve imposing theoretically grounded constraints on the direction of the response of certain variables to structural shocks. High-frequency identification exploits the financial market's immediate reaction to monetary policy announcements (e.g., within a 30-minute window), treating these high-frequency changes as exogenous indicators of monetary policy shocks. Within the identification mechanism, the authors further distinguish between "Poor Man's sign restrictions" and a median rotation approach that imposes sign restrictions. Unlike the second approach, the first assumes that each policy announcement can be classified as either being a monetary shock or an information shock, and not both. While both methods yield similar impulse responses in Jarociński and Karadi (2020), the median rotation approach is preferred in this empirical application due to its weaker assumptions. The identified monetary policy shock refers to an unexpected change in the policy interest rate; specifically, a positive monetary policy shock corresponds to an increase in the interest rate. Second, central bank information shocks reflect changes in the central bank's economic outlook revealed through policy announcements.

As in Piger and Stockwell (2025), this empirical application can be addressed with local projections. In particular, it is of interest whether the choice of bias-corrected estimator is empirically relevant. The analysis uses monthly data of five variables, in addition to the aforementioned shock. The variables include 100 times the logarithm of the consumer price index, 100 times the logarithm of industrial production, and the monthly average one-year Treasury bond yield. This data is drawn from the Federal Reserve Economic Data, with the variable codes CPIAUSL, INDPRO, and DGS1, respectively. Missing values from the one-year Treasury bond yield data are excluded from the computation of the mean. Furthermore, the Gilchrist and Zakrajšek (2012) excess bond premium is included and can be found in the dataset provided for Bauer and Swanson (2023).[4] Daily closing prices of the S&P 500 index are taken from Investing.com and converted to monthly observations by averaging. Finally, the updated values of the aforementioned monthly shocks are available through Marek Jarociński's website.[5]

[4] `www.frbsf.org/research-and-insights/data-and-indicators/monetary-policy-surprises/`
[5] `https://marekjarocinski.github.io/jkshocks/jkshocks.html`

Data is taken from February 1990, which is the first date that the Jarociński and Karadi (2020) shock series is available, up to December 2019. Given that no lags of the observed shock are included in the specifications, data on lagged controls prior to February 1990 are used so that early observations of the shock series do not need to be excluded from the estimation. This gives us a sample size of 359 observations. Compared to the sample sizes of 100 and 200 used in the Monte Carlo simulations and to the smaller samples typically seen in the literature as noted by Herbst and Johannsen (2024), this represents a relatively favourable empirical setting. Local projections are performed for up to three years (36 months). The minimum effective sample size thus consists of 323 observations.

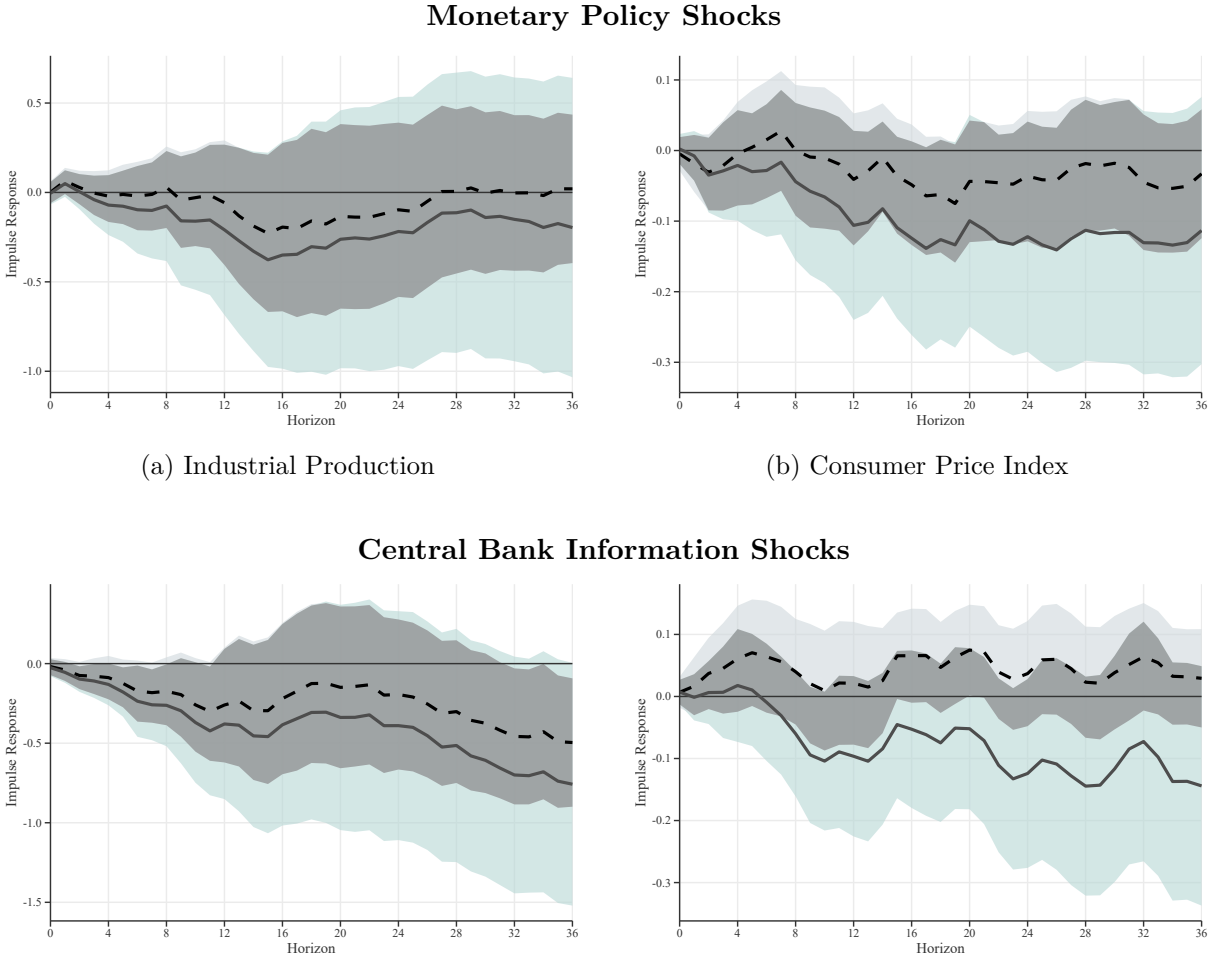
Local projection estimations include an intercept, and the standard specification additionally includes a linear time trend. Jarociński and Karadi (2020) include 12 lags in their VAR specification and Piger and Stockwell (2025) follow this by including equally many lags in the local projection specifications. Matching the focus of the thesis and the preference for the test-down procedure, the same test-down procedure will be performed as in the simulation study, with $p_{\max} = 12$ and a 5% threshold level of significance. Given that the data can be extended to dates before February 1990 for all variables except the shock, the test-down procedure is performed accordingly, where in each iteration with a new lag length, the full shock series is used. In contrast to the Monte Carlo simulations, this implies longer lag lengths do not reduce the effective sample size.

6.2 Empirical Results: Bias Corrections in Practice

This empirical application examines the impulse responses estimated by the three bias-corrected estimators (BCC, LD and RLD) in an empirical setting. The results for the standard LP are omitted, as they closely resemble those of the BCC estimator. Although their relative performance cannot be assessed in terms of Monte Carlo bias or variance, empirical differences in the estimated responses to monetary policy and central bank information shocks can be identified.

Figure 10 presents the impulse response functions of monthly industrial production and CPI to a one standard deviation monetary policy shock and a one standard deviation central bank information shock. The dashed and solid lines indicate the BCC and LD impulse response estimates, respectively. For pointwise inference, the 95% confidence intervals are computed based on Eicker-Huber-White standard errors as these performed better in the simulation study. The darkest shade areas are those where the confidence intervals overlap, whereas the blue-tinted and light-grey areas indicate confidence intervals unique to the long-differenced and BCC estimations, respectively.

Several important observations should be noted from Figure 10. Given a sample size exceeding 300 observations, one would expect the bias corrections to estimate similar impulse responses. Hence, it is striking to observe a substantial and economically meaningful difference between the impulse responses estimated by the two bias-correction methods. In particular, Figure 10d shows that the signs of the estimated impulse responses differ, completely altering the interpretation of the effect of the shock depending on which bias-correction is chosen. Moreover, the magnitude of



(a) Industrial Production

(b) Consumer Price Index

(c) Industrial Production

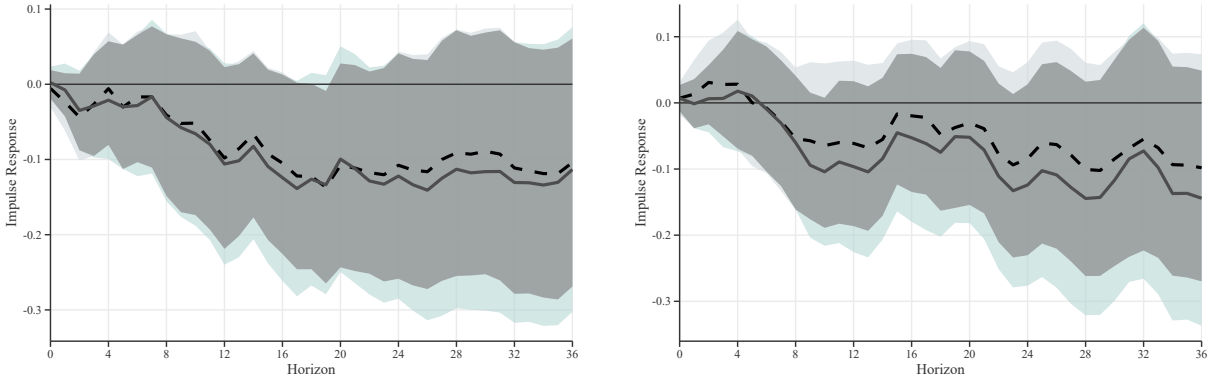
(d) Consumer Price Index

Figure 10: Impulse responses of U.S. industrial production and CPI to a one standard deviation Jarociński and Karadi (2020) monetary policy and central bank information shock. The dashed line shows the BCC estimations and the solid line shows the LD estimations. Light blue and light grey areas indicate 95% confidence intervals, using Eicker-Huber-White standard errors, unique to the long-differenced specification and BCC estimator, respectively; the dark grey area shows their overlap.

the impulse response estimated by the long-differenced local projection tends to be larger than that estimated by the BCC estimator. Finally, the confidence intervals based on Eicker–Huber–White standard errors are generally wider under the long-differenced specification, consistent with the higher coverage observed in the simulation study.

Although the confidence intervals overlap, the opposing signs of the impulse responses in Figure 10d highlight the crucial role of the choice of bias correction in empirical analysis, even in determining the direction of a response. This raises the question of how these estimates compare to the results in Appendix C3 of Jarociński and Karadi (2020), where a Bayesian VAR is used instead of local projections. For monetary policy shocks, it is unclear which estimator has impulse responses more closely matching those presented by the authors. However, for central bank information shocks, the BCC estimator more closely resembles the Bayesian VAR results.

Specifically, Jarociński and Karadi (2020) hypothesize and find that information shocks raise the consumer price index. The same sign and approximate magnitude are also captured by the BCC estimator, whereas the LD estimator suggests a negative effect. The BCC estimator also better aligns with the results from the Bayesian VAR for industrial production. However, this comparison to the Bayesian VAR estimates is not conclusive about which estimator is more suitable in an empirical setting.



(a) Monetary Policy Shock

(b) Central Bank Information Shock

Figure 11: Impulse responses of U.S. CPI to a one standard deviation Jarociński and Karadi (2020) monetary policy and central bank information shock. The dashed line shows the RLD estimations and the solid line shows the LD estimations. Light blue and light grey areas indicate 95% confidence intervals, using Eicker-Huber-White standard errors, unique to the long-differenced specification and restricted long-differenced estimator, respectively; the dark grey area shows their overlap.

Figure 11 compares the impulse responses estimated using the long-differenced and restricted long-differenced specifications. The plots follow the same structure as before, with the restricted long-differenced specification now replacing the BCC estimator. In the test-down procedure, the standard local projection selects fewer lags than the long-differenced specification only when the consumer price index is the outcome of interest. While the estimated responses differ slightly between specifications, the differences are smaller than those observed in Figure 10. For the monetary policy shock, the long-differenced specification yields slightly wider confidence intervals, though this pattern does not clearly hold for the central bank information shock. When industrial production is the outcome, the LD specification selects fewer lags than the standard LP specification, leading the RLD and LD estimates to coincide.

7 Conclusion

Local projections consistently estimate impulse response (Jordà and Taylor, 2025). In small samples, however, bias can lead to economically meaningful deviations of the estimated impulse response from the true impulse response. This thesis performs an extensive simulation study to evaluate the performance of several bias correction methods. These methods consistently reduce bias relative to the standard local projection estimator, though often at the cost of

increased variance. The second contribution of this thesis is the introduction of the restricted long-differenced estimator. The empirical case in Section 6 highlights the importance of selecting an appropriate bias correction. This is most notable in estimating the effect of central bank information shocks on the consumer price index, since the estimated sign of the impulse response varies depending on the bias correction method.

Li et al. (2024) discuss the difficulty associated with choosing an estimator based on the macroeconomic data used in local projections. This emphasises the importance of a researcher's position on the bias–variance spectrum when choosing a bias reduction method. If local projections are preferred over VARs, bias should be the researcher's primary concern (Li et al., 2024). In this view, the results from the simulation study suggest that the standard local projection estimator is generally not optimal, and a bias correction method is preferred. In particular, the long-differenced estimator is preferred for researchers that are almost only concerned with bias, while the BCC estimator is optimal for those who also place some weight on variance. The restricted long-differenced estimator provides a middle ground between the two by offering a lower variance than the LD estimator, but giving up some of the bias reduction to achieve this. Coverage probabilities are generally highest for the restricted long-differenced estimator and lowest for the standard LP. Notably, the Eicker-Huber-White standard errors ensure more accurate coverage than the Newey-West standard errors, suggesting they are preferred in small-sample local projections. This finding generalizes the results of Herbst and Johannsen (2024) to all four bias correction methods and provides additional evidence based on the large set of empirically calibrated DGPs considered.

The simulation study additionally examines the use of recursively identified shocks. The results clearly indicate that the BCC correction is preferred, unless the researcher places a strong emphasis on minimizing variance, in which case the standard local projection estimator performs better. However, as shown by Li et al. (2024), VAR-based approaches may be more appropriate if the researcher aims to keep the variance low.

The Monte Carlo simulations indicate several desirable properties of the restricted long-differenced estimator. The smaller number of regressors reduce the variance of the estimator consistently, while retaining most of the bias-reduction provided by the LD estimator. Moreover, compared to the three other estimators, the RLD estimator achieves equal or higher coverage probabilities across the considered DGPs. The extent of this advantage varies across DGPs, with the RLD estimator showing particularly high coverage in stationary settings, including the VAR(4) and DGPs based on the stationary DFM. For the non-stationary DFM, the RLD estimator performs particularly well compared to the other bias corrections when the sample size is $T = 100$.

Despite its advantages, the restricted long-differenced estimator also presents challenges in implementation. In particular, the researcher must choose two lag parameters, p^* and p_D^*. Setting $p_D^* = p + h$, where p is chosen with the VAR-AIC lag selection corresponding to the standard LP, inefficiently uses the available sample at longer horizons. A more suitable alternative is the test-down procedure, which tends to select lag lengths that are closer to those of the true underlying DGP. However, this method is still not optimal, as the selection is performed at short

horizons ($h = 0$ or $h = 1$) which may not capture the dynamics relevant at longer horizons. Future research should focus on developing more robust and theoretically justified approaches to lag selection in the local projection framework, particularly in the context of the restricted long-differenced estimator.

This research is subject to several further limitations that offer promising areas for future research. The true impulse response functions generated from the 50 non-stationary DFM-based DGPs are relatively similar and have few interior extrema. Expanding the set of DGPs and incorporating quarterly data may lead to greater variation in the true impulse responses, improving the evaluation of estimator performance. Furthermore, the presented results do not focus on identification approaches with instrumental variables. With the use of instrumental variables under the LP framework becoming increasingly popular in applied work, investigating the performance of bias corrections in such settings may improve the applicability of the results. In addition, while this thesis considers two standard error estimators, it does not examine alternatives such as the equally-weighted cosine estimator. Herbst and Johannsen (2024) show that this standard error estimator has superior coverage probabilities for the BCC estimator; whether this extends to other bias corrections is yet to be determined. A final avenue for future work is evaluating bias corrections in settings with non-linear local projections.

References

Adamek, R., Smeekes, S., and Wilms, I. (2024). Local projection inference in high dimensions. *The Econometrics Journal*, 27(3):323–342.

Barnichon, R. and Brownlees, C. (2019). Impulse response estimation by smooth local projections. *The Review of Economics and Statistics*, 101(3):522–530.

Bauer, M. D. and Swanson, E. T. (2023). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, 37(1):87–155.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120(1):387–422.

Bobeica, E., Holton, S., Huber, F., and Martínez Hernández, C. (2025). Beware of large shocks! A non-parametric structural inflation model. Working Paper, European Central Bank.

Christiano, L., Eichenbaum, M., and Evans, C. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy*, 113(1):1–45.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? In *Handbook of Macroeconomics*, volume 1, pages 65–148. Elsevier.

Fernández-Villaverde, J., Rubio-Ramírez, J. F., Sargent, T. J., and Watson, M. W. (2007). ABCs (and Ds) of understanding VARs. *American Economic Review*, 97(3):1021–1026.

Furceri, D., Hannan, S. A., Ostry, J. D., and Rose, A. K. (2021). The macroeconomy after tariffs. Working Paper WPS9854, World Bank Group.

Gilchrist, S. and Zakrajšek, E. (2012). Credit spreads and business cycle fluctuations. *American Economic Review*, 102(4):1692–1720.

Herbst, E. P. and Johannsen, B. K. (2024). Bias in local projections. *Journal of Econometrics*, 240(1):105655.

Inoue, A., Jordà, , and Kuersteiner, G. M. (2025). Inference for local projections. *The Econometrics Journal*, page utaf004.

Jarociński, M. and Karadi, P. (2020). Deconstructing monetary policy surprises— The role of information shocks. *American Economic Journal: Macroeconomics*, 12(2):1–43.

Jordà, O. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182.

Jordà, and Taylor, A. M. (2016). The time for austerity: Estimating the average treatment effect of fiscal policy. *The Economic Journal*, 126(590):219–255.

Jordà, and Taylor, A. M. (2025). Local projections. *Journal of Economic Literature*, 63(1):59–110.

Kilian, L. and Kim, Y. J. (2011). How reliable are local projection estimators of impulse responses? *The Review of Economics and Statistics*, 93(4):1460–1466.

Kilian, L. and Lütkepohl, H. (2017). *Structural vector autoregressive analysis*. Cambridge University Press, 1 edition.

Lazarus, E., Lewis, D. J., Stock, J. H., and Watson, M. W. (2018). HAR inference: Recommendations for practice. *Journal of Business & Economic Statistics*, 36(4):541–559.

Li, D., Plagborg-Møller, M., and Wolf, C. K. (2024). Local projections vs. VARs: Lessons from thousands of DGPs. *Journal of Econometrics*, 244(2):105722.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Montiel Olea, J. L. and Plagborg-Møller, M. (2021). Local projection inference is simpler and more robust than you think. *Econometrica*, 89(4):1789–1823.

Montiel Olea, J. L., Plagborg-Møller, M., Qian, E., and Wolf, C. (2025). Local projections or VARs? A primer for macroeconomists. Working Paper w33871, National Bureau of Economic Research.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703.

Nicholls, D. F. and Pope, A. L. (1988). Bias in the estimation of multivariate autoregressions. *Australian Journal of Statistics*, 30A(1):296–309.

Piger, J. M. and Stockwell, T. (2025). Differences from differencing: Should local projections with observed shocks be estimated in levels or differences? *Journal of Applied Econometrics*. Forthcoming.

Plagborg-Møller, M. and Wolf, C. K. (2021). Local projections and VARs estimate the same impulse responses. *Econometrica*, 89(2):955–980.

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 *Handbook of Macroeconomics*, volume 2, pages 71–162. Elsevier.

Ramey, V. A. (2011). Can government purchases stimulate the economy? *Journal of Economic Literature*, 49(3):673–685.

Ramey, V. A. and Zubairy, S. (2018). Government spending multipliers in good times and in bad: Evidence from US historical data. *Journal of Political Economy*, 126(2):850–901.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.

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

Stock, J. and Watson, M. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of Macroeconomics*, volume 2, pages 415–525. Elsevier.

Stock, J. H. and Watson, M. W. (2018). Identification and estimation of dynamic causal effects in macroeconomics using external instruments. *The Economic Journal*, 128(610):917–948.

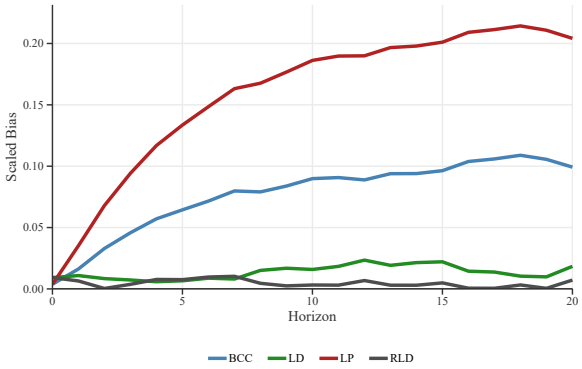
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817.

Appendix A Further Results

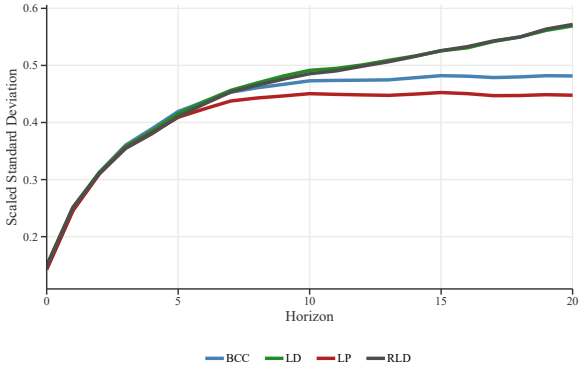
This appendix presents additional results from all the conducted Monte Carlo simulations. The majority of these results are simply the corresponding results for the sample size $T = 100$. For the AR(1) DGP, the results for persistence $\phi = 0.90$ are also included here. Section A.7 summarizes the lag selection with the test-down procedure.

A.1 AR(1): Results for $\phi = 0.90$ with $T = 200$

VAR-selected lag length

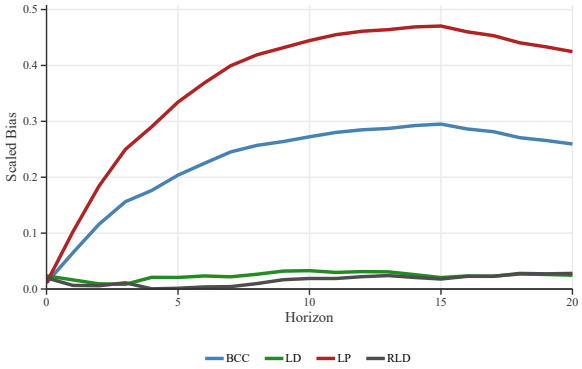


(a) Industrial Production

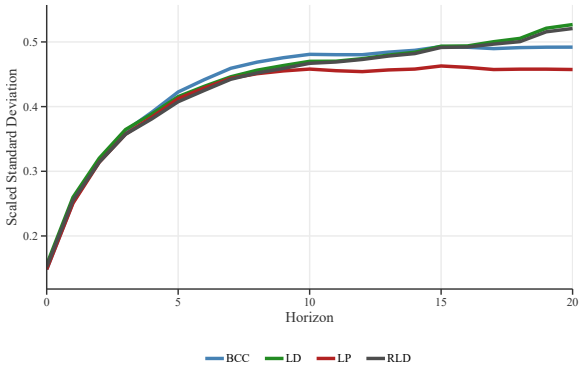


(b) Consumer Price Index

Test-down selected lag length



(c) Industrial Production



(d) Consumer Price Index

Figure 12: Mean absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$, across two lag selection methods, for $\phi = 0.90$. All results based on 5000 simulations and sample size $T = 200$.

A.2 AR(1): Results for $T = 100$

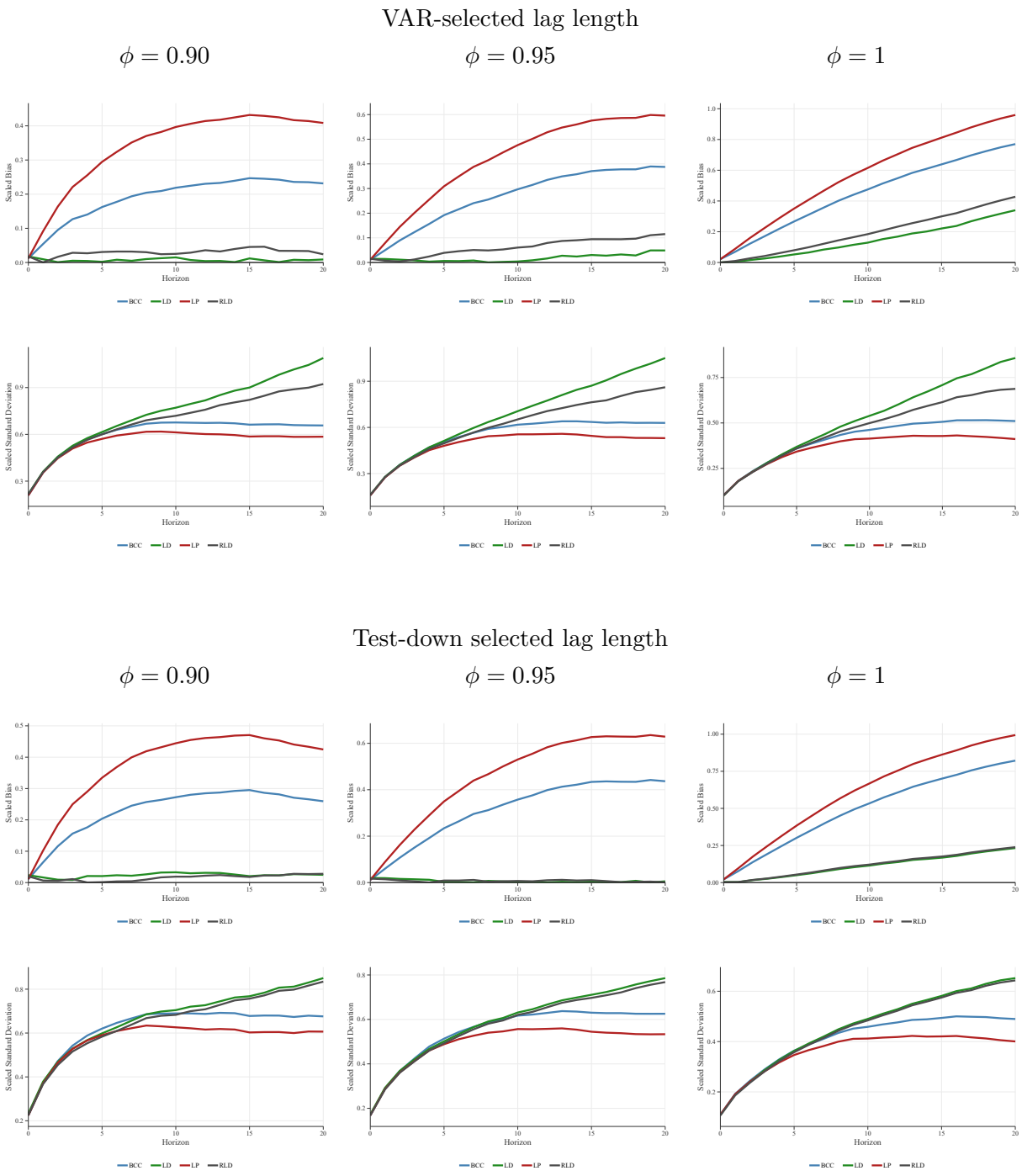


Figure 13: Mean absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$, across two lag selection methods, for $\phi \in \{0.90, 0.95, 1\}$. All results based on 5000 simulations and sample size $T = 100$.

A.3 VAR(4): Results for $T = 100$

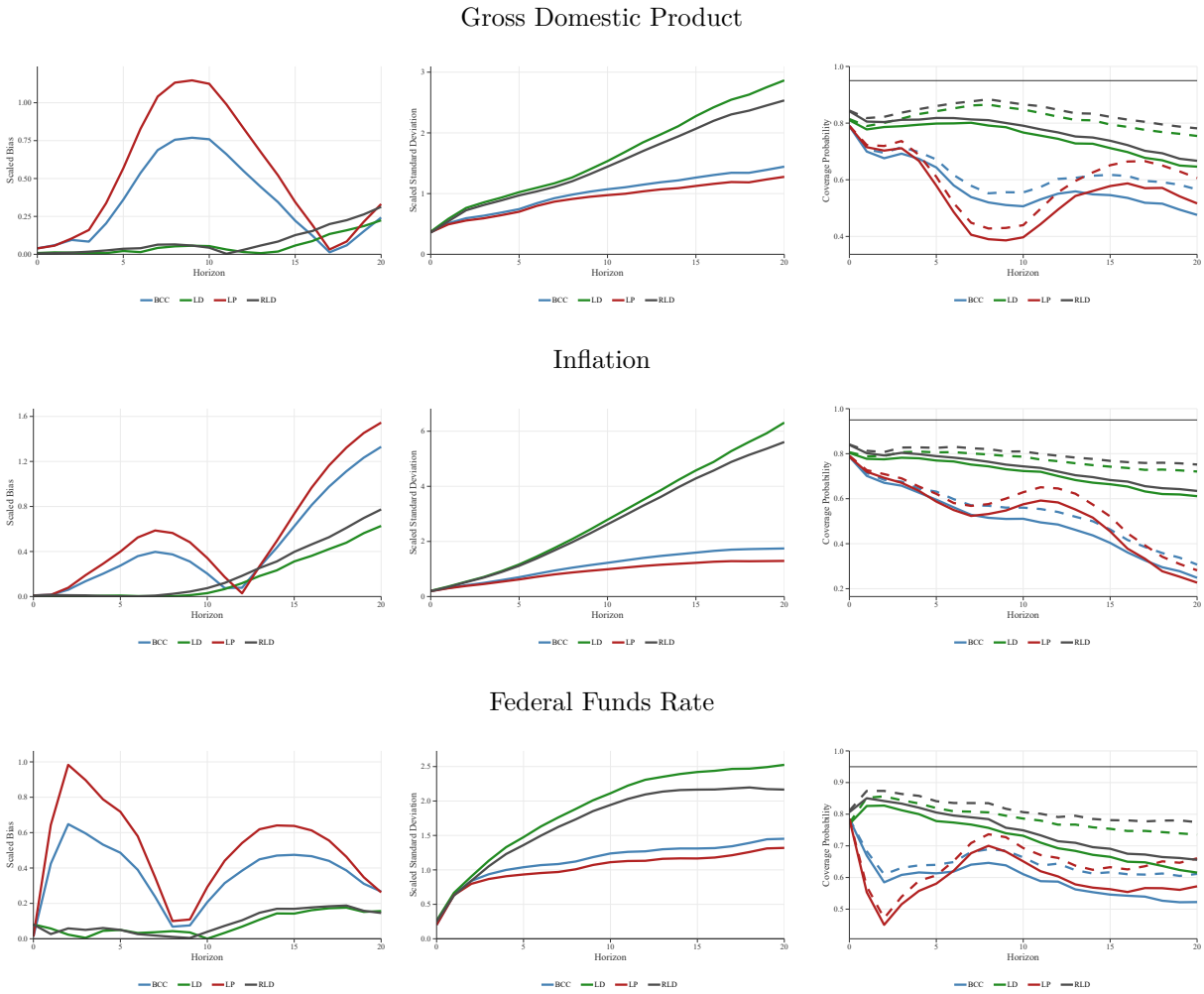
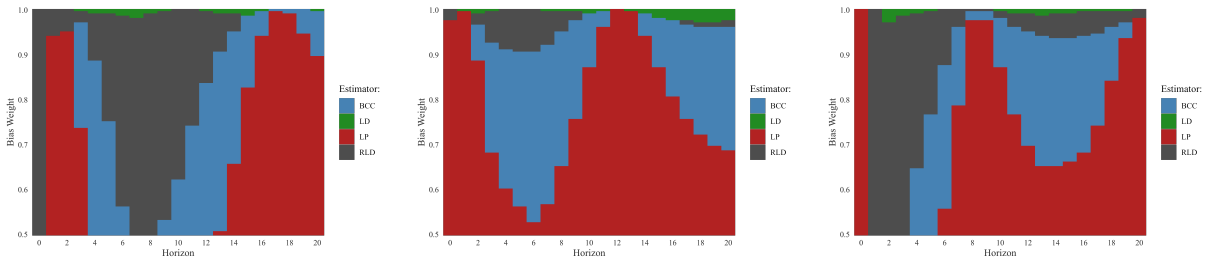


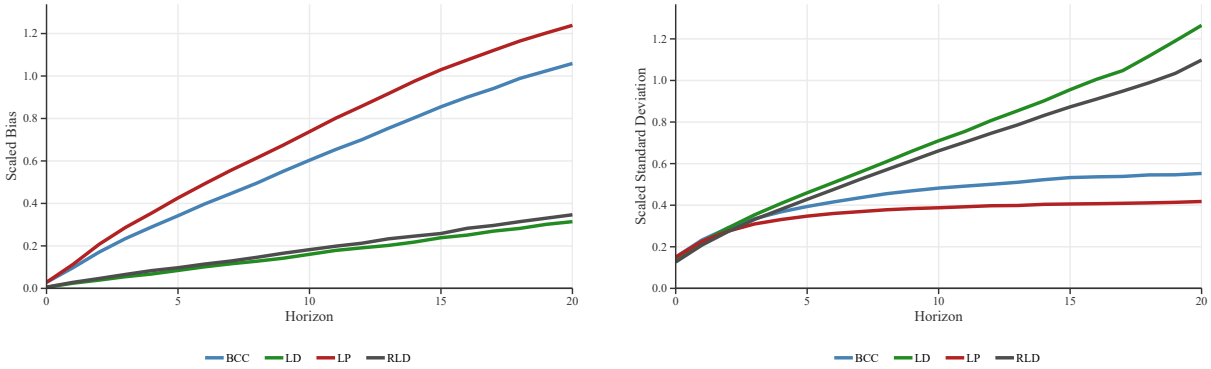
Figure 14: Mean absolute bias and standard deviation of the estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$, along with coverage probabilities. Confidence intervals are constructed at the 95% level using Eicker–Huber–White (dashed) and Newey–West (solid) standard errors. Results for real GDP, inflation, and the federal funds rate are based on 5,000 simulations with $T = 100$.



(a) Real GDP (b) Inflation (c) Federal Funds Rate

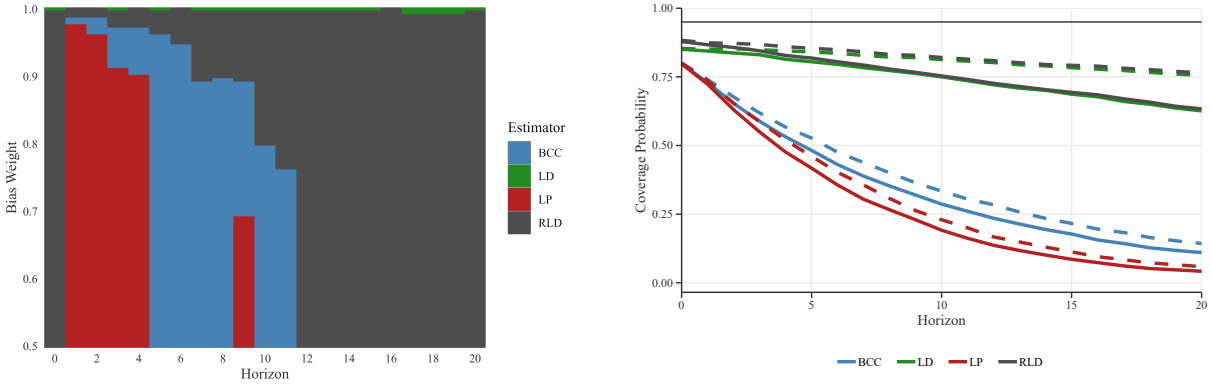
Figure 15: Method that minimizes the loss function across outcome variables real GDP, inflation and the federal funds rate, with weights $\lambda \in [0.5, 1]$ at 0.005 intervals. Results are based on 5000 simulations and a sample size $T = 100$.

A.4 Observed Shocks and Non-Stationary DGPs ($T = 100$)



(a) Bias (b) Standard Deviation

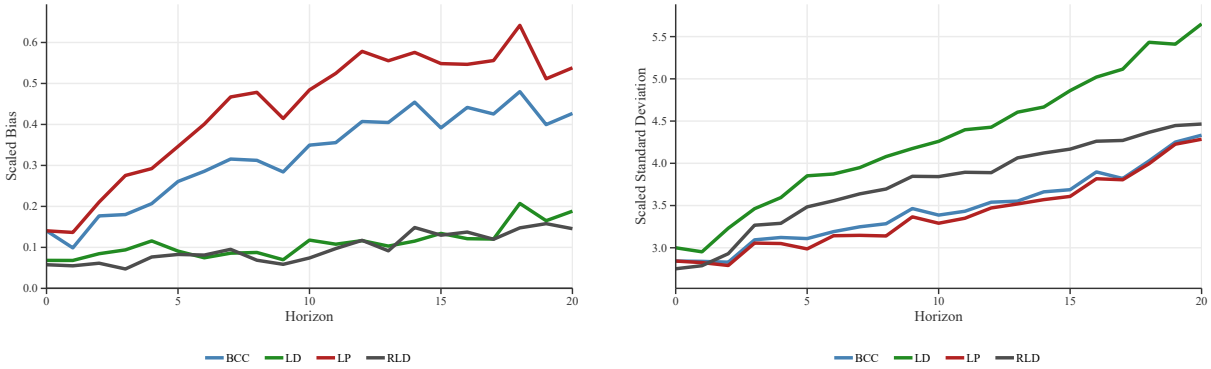
Figure 16: Median across 50 non-stationary DGPs of the (mean) absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$. Shocks are observed. Results are based on 1000 simulations per DGP and sample size $T = 100$.



(a) Optimal Estimator (b) Coverage Probability

Figure 17: Panel (a) shows the method that minimizes the mean loss function across 50 non-stationary DGPs, with weights $\lambda \in [0.5, 1]$, at 0.005 intervals. The loss functions are normalized by $\frac{1}{21}\sum_{h=0}^{20}\theta_h^2$. Panel (b) shows the median coverage probabilities across 50 non-stationary DGPs. Confidence intervals are constructed at the 95% level using Eicker–Huber–White (dashed lines) and Newey–West (solid lines) standard errors. Results are based on 1000 simulations per DGP and sample size $T = 100$.

A.5 Stationary DGPs with Observed Shocks ($T = 100$)



(a) Bias (b) Standard Deviation

Figure 18: Median across 50 stationary DGPs of the (mean) absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$. Shocks are observed. Results are based on 1000 simulations per DGP and sample size $T = 100$.

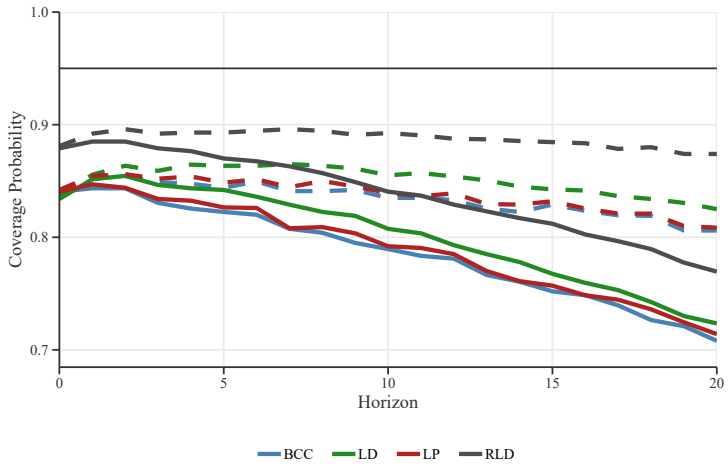


Figure 19: Median coverage probabilities across 50 stationary DGPs. Confidence intervals are constructed at the 95% level using Eicker–Huber–White (dashed lines) and Newey–West (solid lines) standard errors. Results are based on 1000 simulations per DGP and sample size $T = 100$.

A.6 Recursively Identified Shocks

	Min	Median	Mean	Max
Sign persistence	8	19	18.14	19
Average/(max absolute value)	-0.69	0.41	0.30	0.68
Number of interior local extrema	0	2	1.16	4
Horizon of max absolute value	20	20	20	20
Initial response	-0.48	0.64	0.10	1.70

Table 2: Summary statistics of the 50 non-stationary DGPs with recursively identified shocks. Sign persistence refers to the number of horizons with the same sign as θ_1 (so maximum sign persistence is 19). Average/(max absolute value) is given by $\left(\frac{1}{21}\sum_{h=0}^{20}\theta_h\right) / \max_h |\theta_h|$. Initial response is $\theta_1 / \sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$.

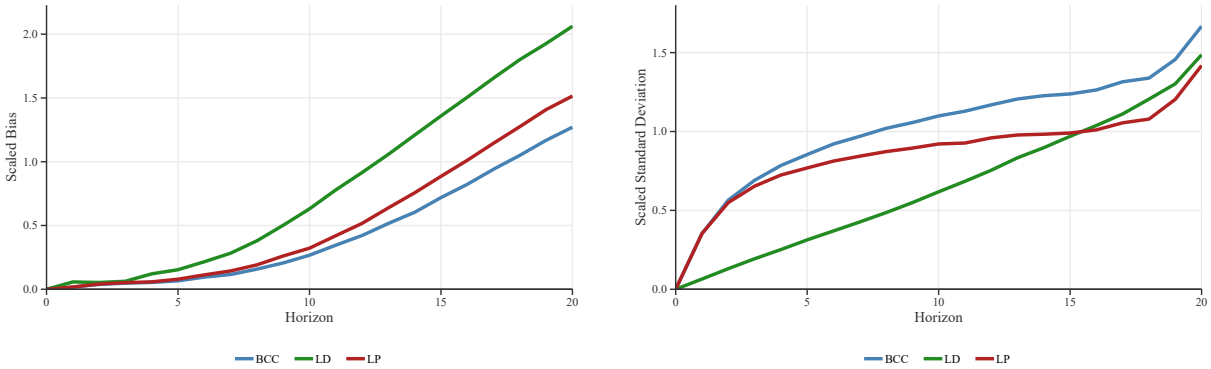


Figure 20: Median across 50 DGPs of the (mean) absolute bias and standard deviation of different estimators, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$. Results are based on 1000 simulations and sample size $T = 100$.

A.7 Selected Lag Lengths

This section presents the results of the test-down lag selection procedure across all DGPs. Table 3 reports the distribution of selected lag lengths across different DGPs, estimation methods (LP and LD), and sample sizes ($T = 100$ and $T = 200$).

Table 3: Distribution of Selected Lag Lengths Across DGPs and Estimation Methods

DGP	LP/LD	Min	20	40	50	60	80	Max
T = 100								
AR(1)								
$\phi = 0.90$	LP	1	1	1	1	5	11	15
	LD	1	1	2	5	7	12	15
$\phi = 0.95$	LP	1	1	1	1	5	11	15
	LD	1	1	1	2	6	11	15
$\phi = 1$	LP	1	1	1	2	5	11	15
	LD	1	1	1	1	5	10	15
VAR(4)								
Real GDP	LP	1	3	4	4	4	6	7
	LD	1	2	4	4	4	5	7
Inflation	LP	1	3	4	4	5	6	7
	LD	1	4	4	4	5	6	7
FFR	LP	3	4	4	4	4	5	7
	LD	2	3	4	4	5	6	7
DFM								
Observed (NS)	LP	1	1	2	4	6	10	12
	LD	1	1	1	1	4	8	12
Observed (S)	LP	1	1	2	2	4	9	12
	LD	1	2	2	3	5	9	12
Recursive	LP	1	1	1	4	6	10	12
	LD	1	1	1	1	3	8	12
T = 200								
AR(1)								
$\phi = 0.90$	LP	1	1	1	2	5	11	15
	LD	1	2	6	8	10	13	15
$\phi = 0.95$	LP	1	1	1	2	6	11	15
	LD	1	1	2	5	8	12	15
$\phi = 1$	LP	1	1	1	1	5	11	15
	LD	1	1	1	1	5	11	15
VAR(4)								
Real GDP	LP	3	4	7	10	11	14	15
	LD	4	6	7	8	9	12	15
Inflation	LP	3	4	7	10	11	14	15
	LD	4	6	7	8	9	12	15
FFR	LP	4	4	4	5	8	13	15
	LD	3	7	8	9	10	12	15
DFM								
Observed (NS)	LP	1	2	2	5	7	12	15
	LD	1	1	3	5	7	11	15
Observed (S)	LP	1	2	3	3	6	11	15
	LD	1	3	6	7	9	12	15
Recursive	LP	1	1	2	5	7	12	15
	LD	1	1	2	4	6	11	15

Note: This table reports the minimum, selected quantiles (20th, 40th, 50th, 60th, 80th), and maximum of the lag lengths selected using the test-down procedure across different DGPs, estimation methods (LP and LD), and sample sizes ($T = 100$ and $T = 200$). FFR denotes the federal funds rate; NS is non-stationary; S is stationary. For the DFM, results are based on 50 distinct DGPs; quantiles are computed within each DGP, and the median across DGPs is reported. The recursive lag length refers to lags added beyond the contemporaneous controls.

Appendix B Equivalence of LD and RLD Impulse Responses

Let p^* be the number of lags of each variable in the RLD specification, and let p_D^* be the parameter used to construct the restriction matrix. This appendix proves that, under observed shock identification, if $p^* = p_D^*$, the RLD estimator coincides with the LD estimator based on a specification with p_D^* lags of the first difference, for all horizons h.

Define a sub-diagonal of a square matrix to be any diagonal below the main diagonal. For example, in the following 3×3 matrix, the ones immediately below the main diagonal form the first sub-diagonal:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Proof. Let p^* denote the number of desired lags in the RLD specification. Assume the parameter $p_D^* = p^*$.

Now consider the matrix R_h as defined in Section 3.5. For all $j \in \{1, \dots, p^*\}$ and $i \in \{1, \dots, p^*\}$, its entries are given by

$$[R_h]_{j,i} = \begin{cases} 1 & \text{if } i \le j \le h + i, \\ 0 & \text{otherwise.} \end{cases}$$

This defines a square matrix of dimension $p^* \times p^*$. This matrix has $p^* - 1$ sub-diagonals. For $h = 0$, we have $R_0 = I_{p^*}$. By construction of the matrix, for $h > 0$, the matrix contains ones on the main diagonal and the h sub-diagonals directly below it. Hence, for $h \ge p^* - 1$, R_h is a lower-triangular matrix of ones.

Consider the LD specification with p_D^* lags of the first difference. Consider a variable y_t for which its lags are included in the LD specification. That is, the LD specification includes the vector of regressors $\Delta Y_t = (\Delta y_{t-1}, \dots, \Delta y_{t-p^*})'$. Then, in the RLD specification, the regressor vector is $Z_t = R_h' \Delta Y_t$. The structure of R_h' ensures it has full rank, or equivalently, a non-zero determinant. This implies R_h' is invertible. Hence, in the LD specification, each regressor in ΔY_t can be recovered from the RLD regressors $Z_t = R_h' \Delta Y_t$ via linear transformation, i.e., $\Delta Y_t = (R_h')^{-1} Z_t$. This argument implies that each control variable in the LD specification can be represented as a linear combination of regressors in the RLD specification. Then, given other specifications (e.g., inclusion of intercept) coincide, the properties of OLS imply that the estimated impulse responses much also coincide. Therefore, if $p_D^* = p^*$, we conclude $\theta_h^{LD} = \theta_h^{RLD}$ for all h. □

Note that when $p_D^* > p^*$, the matrix R_h' of dimension $p^* \times p_D^*$ is no longer full column rank, since the rank is at most p^*. Therefore, the RLD regressors do not span the full LD regressor space. In this case, the estimators do not generally coincide.

Appendix C AR(1) True Impulse Response Derivation

From Equation 6, with $\alpha = 0$ and $\theta = 1$, a proof by induction will show that, for all $h = 0, \ldots, h_{\max}$, y_{t+h} can be expressed as:

$$y_{t+h} = \phi^{h+1} y_{t-1} + \sum_{j=0}^{h} \phi^j x_{t+h-j} + \sum_{j=0}^{h} \phi^j u_{t+h-j}. \tag{10}$$

For $h = 0$, we have:

$$y_t = \phi y_{t-1} + x_t + u_t,$$

which coincides with Equation 10. Assume that for a horizon $h = k - 1$, Equation 10 holds. Then, for $h = k$,

$$\begin{aligned} y_{t+k} &= \phi y_{t+k-1} + x_{t+k} + u_{t+k} \\ &= \phi(\phi^k y_{t-1} + \sum_{j=0}^{k-1} \phi^j x_{t+(k-1)-j} + \sum_{j=0}^{k-1} \phi^j u_{t+(k-1)-j}) + x_{t+k} + u_{t+k} \\ &= \phi^{k+1} y_{t-1} + \sum_{j=0}^{k} \phi^j x_{t+k-j} + \sum_{j=0}^{k} \phi^j u_{t+k-j}. \end{aligned}$$

This is the required result. By the Principle of Mathematical Induction, Equation 10 holds for all $h = 0, \ldots, h_{\max}$. Then,

$$\frac{\partial y_{t+h}}{\partial x_t} = \phi^h \quad \text{for all } h = 0, \ldots, h_{\max}.$$

This is the impulse response for a one-unit shock in x_t. As x_t is i.i.d. and drawn from a standard normal distribution, this is equivalent to the response for a one standard deviation shock.

Appendix D Truncation Lag for the Newey–West Estimator

This appendix motivates the use of $h+1$ lags in the Newey-West estimator. The AR(1) DGP is used to show that the auto-covariance of the error term in the correctly specified local projection regression is zero with a lag parameter $h+1$. Consider the AR(1) as in Equation 6:

$$y_t = \alpha + \theta x_t + \phi y_{t-1} + u_t.$$

Without loss of generality and as in Piger and Stockwell (2025), assume $x_t \overset{\text{i.i.d.}}{\sim} (0, \sigma_x^2)$ and $u_t \overset{\text{i.i.d.}}{\sim} (0, \sigma_u^2)$. Additionally, assume $\mathbb{E}[x_t u_{t+j}] = 0$ for all j and t (i.e., they are uncorrelated) and assume x_t and x_{t+j} are uncorrelated for all t and all $j \neq 0$. In the following, replace the original θ_h^{LP} notation with θ_h for simplicity. Then, the correctly specified standard local projection specification is given by:

$$y_{t+h} = \mu_h + \theta_h x_t + \rho_h y_{t-1} + \varepsilon_{t,h}, \tag{11}$$

where, by extension of Appendix C and from Piger and Stockwell (2025), $\theta_h = \theta\phi^h$, $\rho_h = \phi^{h+1}$, and

$$\varepsilon_{t,h} = \sum_{i=0}^{h-1} \theta_i x_{t+h-i} + \sum_{i=0}^{h} \phi^i u_{t+h-i}.$$

To simplify notation, fix h and let $\varepsilon_{t,h} = \varepsilon_t$ for all t. Sufficiently many lags k should be selected such that $E[\varepsilon_t \varepsilon_{t-k}] = 0$. The unconditional auto-covariance is given by the expectation:

$$\mathbb{E}[\varepsilon_t \varepsilon_{t-k}] = \mathbb{E}\left[\left(\sum_{i=0}^{h-1} \theta_i x_{t+h-i} + \sum_{i=0}^{h} \phi^i u_{t+h-i}\right)\left(\sum_{j=0}^{h-1} \theta_j x_{t-k+h-j} + \sum_{j=0}^{h} \phi^j u_{t-k+h-j}\right)\right].$$

As $\mathbb{E}[x_t u_{t+j}] = 0$ for all t and j, the cross terms cancel out, so:

$$\mathbb{E}[\varepsilon_t \varepsilon_{t-k}] = \mathbb{E}\left[\sum_{i=0}^{h-1}\sum_{j=0}^{h-1} \theta_i \theta_j x_{t+h-i} x_{t-k+h-j}\right] + \mathbb{E}\left[\sum_{i=0}^{h}\sum_{j=0}^{h} \phi^i \phi^j u_{t+h-i} u_{t-k+h-j}\right].$$

Note that $\mathbb{E}[x_{t+h-i} x_{t-k+h-j}] \neq 0 \Leftrightarrow t+h-i = t-k+h-j \Leftrightarrow j = i-k$. Similarly, $\mathbb{E}[u_{t+h-i} u_{t-k+h-j}] \neq 0 \Leftrightarrow j = i-k$. Hence:

$$\mathbb{E}[\varepsilon_t \varepsilon_{t-k}] = \sum_{i=k}^{h-1} \theta_i \theta_{i-k} \sigma_x^2 + \sum_{i=k}^{h} \phi^i \phi^{i-k} \sigma_u^2.$$

In the first term, $i = k, \ldots, h-1$, and in the second term, $i = k, \ldots, h$; at the boundary, only the second term contributes. For the overall expression to be non-zero, it is necessary that $k \leq h$. Equivalently, the residuals ε_t exhibit autocorrelation up to lag h. Consequently, the Newey-West standard error estimator must include $h+1$ lags to account for serial correlation in the regression errors. This result holds for all $h = 0, \ldots, h_{\max}$.

The same argument applies to the correctly specified long-differenced specification of the AR(1) model. Refer to Piger and Stockwell (2025) for the corresponding expression for the error

term, from which the derivation follows analogously. Extending this argument, the same result can be derived for the RLD specification.

Appendix E Transformation of FRED-MD Data

For the complete list of variables in the FRED-MD database, refer to the Online Appendix of Adamek et al. (2024) or McCracken and Ng (2016). The variables for the stationary DFM are transformed as in Adamek et al. (2024), which is based on the transformations of Bernanke et al. (2005). Table 4 provides the transformations for the stationary data. In the non-stationary DFM variant, the codes change as follows: first difference (2) becomes levels (1); second difference (3) becomes first difference (2); first log difference (5) becomes log in levels (4); second log difference (6) becomes first log difference (5); percentage change (7) becomes levels (1).

Table 4: Transformations for the Stationary DFM

Code	**Transformation**
1	$f(x_t) = x_t$
2	$f(x_t) = x_t - x_{t-1}$
3	$f(x_t) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$
4	$f(x_t) = \log(x_t)$
5	$f(x_t) = \log(x_t) - \log(x_{t-1})$
6	$f(x_t) = (\log(x_t) - \log(x_{t-1})) - (\log(x_{t-1}) - \log(x_{t-2}))$
7	$f(x_t) = (x_t - x_{t-1})/x_{t-1}$