Two-Stage Model Averaging for Impulse Responses: Local Projections- and VARs-based Approaches

Ulrich Hounyo* **Seojin Jung†**

June 15, 2025

Abstract

We introduce a two-stage model averaging estimator for impulse response functions that flexibly combines local projection (LP) and vector autoregression (VAR) methods, as well as different statistical paradigms (Bayesian or frequentist). In the first stage, we compute weights within the LP and VAR groups. A second-stage weight, α_{LP}, then integrates the two groups by accounting for their structural differences and incorporating their respective strengths. Monte Carlo simulations illustrate a typical bias–variance trade-off: LP-based estimators tend to exhibit lower bias, whereas VAR-based estimators display lower variance. The α_{LP}-based estimators effectively balance these properties, leading to lower overall estimation error, particularly at intermediate horizons. Empirical applications further demonstrate that the proposed model averaging schemes yield smoother and more interpretable impulse response estimates than single estimators.

Keywords: *Model Averaging; Impulse Responses; Local projection; Vector Autoregression.*

1 Introduction

Impulse responses are an important tool in empirical macroeconomic analyses for estimating the effects of unanticipated structural shocks to the economy. They are often estimated by using vector autoregression (VAR) (Sims, 1980). An alternative approach, which has become increasingly popular over the past few decades, is the local projection (LP) method introduced by Jordà (2005). In general, LP directly estimates the response of future outcomes to current covariates for each forecast horizon, whereas VAR(p) extrapolates longer-run impulse responses from the first p sample autocovariances. As a result, while the estimates from the two methods tend to agree at horizons $h \leq p$, they may diverge

*Department of Economics, University at Albany – State University of New York, Albany, NY 12222, United States. E-mail address: khounyo@albany.edu

†Department of Economics, University at Albany – State University of New York, Albany, NY 12222, United States. E-mail address: sjung9@albany.edu

significantly at intermediate and long horizons. Intuitively, VARs are expected to produce lower variance but potentially higher bias relative to LPs. This trade-off is formally analyzed by Plagborg-Møller and Wolf (2021), and recent simulation-based study by Li et al. (2024) provides supporting evidence of this bias-variance trade-off between LP and VAR estimators. In addition, a growing literature has explored extensions to these estimation methods, incorporating techniques such as bias correction, shrinkage, Bayesian approaches, and model averaging. Each approach entails its own advantages and limitations.

Among the available approaches, model averaging is an effective estimation scheme that incorporates information from various estimators across different models. Researchers commonly face model specification uncertainty, particularly in macroeconomic settings where data is subject to multiple sources of uncertainty, including structural breaks and parameter instability. Model averaging can help mitigate such uncertainty by incorporating multiple models, each capturing distinct features of the economy or representing theoretical frameworks.

However, most existing studies on model averaging for impulse response estimation do not consider combining estimators derived from different estimation methods (such as LP-based and VAR-based approaches) or from different statistical paradigms (Bayesian versus frequentist). Instead, the literature typically focuses on model averaging across least squares specifications with varying lag lengths, aimed at addressing model selection within a common estimation framework. This practice likely reflects the technical difficulties that arise when combining estimators generated from fundamentally different methods. We argue that the relative scarcity of procedures that weight and combine impulse response estimators across distinct estimation methods (e.g., LP and VAR) may stem from these methodological challenges. For instance, conventional model averaging techniques such as Bayesian model averaging (BMA) are not directly applicable when one of the candidate estimators is based on LPs, as LPs are not generative models.

In contrast to this general trend, Ho et al. (2024) propose prediction pools, a model averaging scheme that can incorporate any estimation method based on predictive densities, following the approach of Geweke and Amisano (2011). Prediction pools highlight how estimators can be combined across both model classes and statistical paradigms. The key innovations of Ho et al. (2024) are conditionality and flexibility. They construct forecast densities that are conditional on a structural shock of interest, thereby tracing out model-specific responses to the shock. This feature enables impulse responses to be derived from any estimation method or framework capable of producing conditional forecast densities for a given variable at a given horizon.

Our proposed schemes differ from prediction pools in at least two dimensions. First, prediction pools treat impulse responses as conditional forecasts of the shock effect, emphasizing models that provide superior predictive performance. By contrast, we adopt traditional model averaging strategies by proposing a generalized model averaging (GMA) scheme tailored to LP-variants—constructed as a transformation of BMA—and a cross-

validation model averaging (CVA) scheme for models within the same class. Second, while both prediction pools and our approach address the sample length issue that arises when combining LP-based estimators across horizons, they do so differently. Prediction pools resolve this by estimating each horizon separately. Similarly, we estimate LP-based and VAR-based models separately at each horizon. However, we introduce an additional weight, α_{LP}, which is assigned to the LP-based estimators, with the remaining weight allocated to the VAR-based estimators. This design allows us to integrate estimators across methods within a unified impulse response analysis.

In this paper, we propose model averaging-based estimators of impulse response functions, drawing inspiration from the thick modeling approach of Granger and Jeon (2004). Granger and Jeon distinguish between thin modeling, which relies on a single preferred specification selected using criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or maximum likelihood, and thick modeling, which considers a range of plausible specifications to retain useful information that might otherwise be discarded. For example, combining forecasts from multiple models often outperforms relying on a single best forecast, just as portfolio diversification is typically superior to investing in a single asset.

Motivated by this framework, we propose impulse response estimators that combine information not only across models—as is common in the literature (see, e.g., Hansen (2016))—but, more importantly, across estimation methods, including both LP-based and VAR-based approaches. Our framework also accommodates averaging across statistical paradigms, such as Bayesian and frequentist estimators.

More specifically, we develop model averaging schemes that transform traditional methods including BMA, GMA, and CVA to allow for the integration of LP and VAR estimators. These schemes are flexible, compatible with variants of LP and VAR, and straightforward to implement in applied settings.

Our simulation study follows the settings of Li et al. (2024), employing an empirically calibrated dynamic factor model (DFM) as the data-generating process (DGP). We fit the DFM to the dataset of Stock and Watson (2016), which comprises 207 quarterly U.S. macroeconomic time series spanning a wide range of variable categories. We use the data in levels rather than transforming them to stationarity, consistent with standard practice in applied work. From the full 207-variable DFM, we draw random subsets of five variables (subject to constraints reflecting applied practice) for each monetary and fiscal shock.

Using 5,000 Monte Carlo simulations, we assess the performance of model averaging schemes: simple averaging, BMA for VAR-variants, and GMA, which is a transformed version of BMA for LP-variants, as well as CVA. These schemes are grouped into several combinations of LP and VAR estimator variants to estimate impulse response functions. The first group includes model averaging across LP-based approaches, such as least squares, bias correction, Bayesian, and shrinkage methods. The second group focuses on VAR-based approaches, including least squares, bias correction, and Bayesian methods. The final group combines all seven LP- and VAR-based estimators into a single weighted es-

timator. For each method, we consider three structural identification strategies: observed shocks, and as robustness checks, instrumental variables (IVs)/proxies and recursive identification. Performance is evaluated in terms of bias, standard deviation, and unweighted mean squared error, defined as the squared sum of the bias and standard deviation.

Our simulation results are consistent with the previous literature and confirm a clear bias-variance trade-off, particularly at intermediate horizons. LP-based estimators tend to exhibit lower bias than VAR-based estimators, while VAR-based estimators generally achieve lower variance. As a result, we find that the third group of model averaging schemes, which combines LP-based and VAR-based estimators—namely, α_{LP}-GMA and α_{LP}-CVA—achieves a favorable balance between the two properties and shows a high likelihood of outperforming the other model averaging groups.

In the empirical applications, we compare our α_{LP} schemes to other recently proposed averaging methods. In Section 5.1, we compare our impulse responses to a monetary shock with those from the Stein combination shrinkage method (VAR Avg) of Hansen (2016), using the latest Federal Reserve data. In Section 5.2, we compare our estimates for another monetary shock with those obtained from prediction pools (Ho et al., 2024), employing the Romer and Romer (2004) shocks as used in Ramey (2016). We find that the impulse responses produced by the α_{LP} schemes exhibit reduced fluctuation and generally lie between the extremes of other estimates, effectively mitigating noise and volatility.

Related Literature. Our approach is motivated by the wide range of methods available to practitioners for computing impulse responses. Researchers are no longer restricted to a single estimation method, but instead have access to multiple alternatives. In this paper, we focus on two major classes of statistical models: VARs (Sims, 1980) and LPs (Jordà, 2005). Each of these model classes includes numerous variants. For VAR-based approaches, Pope (1990) and Kilian (1998) examine the biases that may arise in the estimation of multivariate autoregressions. Giannone et al. (2015) introduce a Bayesian VAR with automatic prior selection to estimate reduced-form coefficients. On the LP side, Plagborg-Møller and Wolf (2021) propose LPs with instruments, Montiel Olea and Plagborg-Møller (2021) employ lag-augmented LPs, and Herbst and Johannsen (2024) propose a bias-corrected version of LP. In addition, Miranda-Agrippino and Ricco (2021a,b) extend a Bayesian LP framework, while Barnichon and Brownlees (2019) proposes a smooth LP based on penalized B-splines. In response to this diversity, a growing literature has emerged on averaging impulse response estimators. Hansen (2016) introduces an estimator based on least squares VARs with varying lag lengths, applying a Stein-type shrinkage approach to construct averaged impulse responses. His method can also be interpreted as implicitly averaging over least-squares LPs through the use of high-order VARs. More recently, Ho et al. (2024) propose prediction pools, a flexible model averaging framework that treats impulse responses as conditional forecasts and allows the combination of estimators from different model classes and statistical paradigms.

Outline. The remainder of the paper is organized as follows. Section 2 introduces our proposed model averaging schemes, which build upon and extend traditional approaches. Section 3 details the LP-based and VAR-based methods used to obtain weights for model averaging. Section 4 presents results from Monte Carlo simulation studies. In Section 5, we apply the proposed methods to real data. Section 6 summarizes the findings and concludes the paper. Appendix A and Appendix B describe procedures for computing optimal weights in BMA, GMA, and CVA. Appendix C reports the estimated weights used in comparisons with prediction pools in our empirical study. Online Appendices D through F present additional simulation results and robustness checks.

2 Model Averaging

When multiple competing estimation methods are available for measuring impulse response functions (IRFs), one can construct a linear combination of these estimators using a weight vector $\mathbf{w}$. To generate model-averaged IRF estimates that balance systematic error (bias) and variability (variance) across horizons, we consider a range of traditional model averaging schemes and their adapted forms.

In our framework, each model $\mathcal{M}_m$ (for $m = 1, \ldots, M$) contributes to the construction of the model-averaged IRF. For each forecast horizon h, we assign a weight vector that lies in the unit simplex:

$$\mathcal{W} = \left\{ \mathbf{w}_{mavg,h} \equiv (\omega_{1,mavg,h}, \ldots, \omega_{M,mavg,h})' \in [0,1]^M : \sum_{m=1}^{M} \omega_{m,mavg,h} = 1 \right\},$$

where $\mathbf{w}_{mavg,h}$ is the vector of model weights specific to horizon h and averaging scheme $mavg$. The notation $mavg$ denotes the particular model averaging method under consideration such as equal weighting, BMA, GMA, CVA, and their extended forms that combine LP- and VAR-based estimators (e.g., α_{LP}-GMA and α_{LP}-CVA).

2.1 Equal Averaging (abbreviated "EQ")

Equal averaging is the most straightforward form of model averaging, assigning identical weights to all M estimators. It is sometimes referred to as the simple averaging model or consensus model. For any given horizon h, the EQ weights are defined as:

$$\omega_{m,\text{EQ},h} = \frac{1}{M}, \quad \text{for } m = 1, \ldots, M. \tag{1}$$

2.2 Generalized Model Averaging (abbreviated "GMA")

BMA is a statistical method that combines predictions from multiple models, weighting each according to its posterior probability. These probabilities reflect each model's relative likelihood given the observed data. By accounting for model uncertainty, BMA often achieves more robust predictive performance than relying on a single model.

In this paper, BMA is applied to combine multiple VAR-variant estimators. When the VAR model is well-specified, its residuals are more likely to satisfy the white noise assumption, making BMA particularly well-suited to this setting. The BMA weight assigned to the impulse response from model m at horizon h is given by[1]:

$$\omega_{m,\text{BMA},h} = \frac{\exp\left(-\frac{1}{2}BIC(m)\right)}{\sum_{l=1}^{M}\exp\left(-\frac{1}{2}BIC(l)\right)}, \tag{2}$$

where $BIC(m)$ is the BIC of model m, calculated as:

$$BIC_m = -2\log(L_m) + d_m\log(T_m), \tag{3}$$

with $\log(L_m)$ representing the log-likelihood of model m, T_m the number of observations, and d_m the number of estimated parameters. Computational details are provided in Appendix A.1.

Incorporating LP-variant estimators into the BMA framework is not straightforward. LPs do not constitute a generative DGP; instead, they solve separate optimization problems at each horizon. This disconnect undermines the use of a joint likelihood and invalidates the posterior model probability interpretation central to BMA.

Moreover, the residual structures differ substantially between VAR- and LP-based methods. VARs produce horizon-invariant residuals as they extrapolate impulse responses from a fixed sample. In contrast, LPs estimate separate regressions at each horizon, resulting in horizon-specific residuals that often exhibit heteroskedasticity and autocorrelation, particularly at longer horizons. These characteristics violate the assumptions underlying standard information criteria such as the BIC and complicate the direct comparison of model fit across LP and VAR estimators.

To address the challenges, we propose a horizon-specific extension of BIC: the Heteroskedasticity and Autocorrelation-Consistent Generalized Information Criterion (HAC-GIC). It is defined as:

$$\text{HAC-GIC}_{m,h} = T_{m,h}\log(|\hat{\Sigma}^{\text{HAC}}_{m,h}|) + d_m\log(T_{m,h}), \tag{4}$$

where $\hat{\Sigma}^{\text{HAC}}_{m,h}$ is the HAC covariance estimator of the residuals for model m at horizon h.

This criterion replaces the Gaussian likelihood component in BIC with a HAC-based fit while retaining the original complexity penalty to guard against overfitting. Based on HAC-GIC, we construct model weights in the same form as equation (2). Appendix A.2 describes the implementation details.

In summary, we compute *horizon-constant* BIC values for VAR-variants and *horizon-specific* HAC-GIC values for LP-variants. This allows us to construct separate weight vectors: BMA_{VAR} among VAR-based estimators and GMA_{LP} among LP-based estimators.

[1] Although the BMA weights $\omega_{m,\text{BMA},h}$ remain constant across horizons due to the use of a horizon-invariant BIC, we retain the subscript h to maintain notational consistency with GMA, in which the weights vary by horizon.

However, this strategy does not resolve the fundamental incompatibility between the two approaches. In particular, the differing information criteria prevent a unified weighting scheme that jointly averages across both LP and VAR variants.

2.3 Cross Validation Model Averaging (abbreviated "CVA")

The second model averaging technique we consider is CVA, which emphasizes a predictive approach by employing a leave-h-out strategy. In CVA, weights are selected by minimizing a cross-validation criterion, $CV_n(\mathbf{w}_{\text{CVA},h})$, over $\mathbf{w}_{\text{CVA},h} \in \mathcal{W}^*$. This yields a weight vector that combines different estimators to minimize the squared prediction error:

$$CV_n(\mathbf{w}_{\text{CVA},h}) = \frac{1}{T}\tilde{e}(\mathbf{w}_{\text{CVA},h})'\tilde{e}(\mathbf{w}_{\text{CVA},h}), \tag{5}$$

$$\mathbf{w}_{\text{CVA},h} = \underset{\mathbf{w}_{\text{CVA},h}}{arg\min} \quad CV_n(\mathbf{w}_{\text{CVA},h}), \tag{6}$$

where $\tilde{e}(\mathbf{w}_{\text{CVA},h})$ denotes the leave-h-out residual vector used in model averaging. Additional implementation details are provided in Appendix B.

Although Jackknife Model Averaging (JMA) is widely used, it is based on a leave-one-out cross-validation criterion, which assumes that the residuals of each model are serially uncorrelated. This assumption may lead to biased weight estimates when applied to LP-based estimators, which often exhibit serial correlation. To address this, Hansen (2010) proposed using leave-h-out residuals, which are better suited to handle serial correlation in the underlying data.

In our setting, a key difficulty arises when attempting to construct a unified residual matrix $\tilde{e}$ that combines leave-h-out residuals from both LP and VAR estimators. This problem becomes particularly acute for horizons $h \geq 1$. As a result, CVA can be effectively implemented only when using either LP-based or VAR-based estimators exclusively. This limitation stems from the structural differences in the residuals generated by VAR and LP approaches, as discussed in Section 2.2.

In essence, despite addressing autocorrelation in LP-variant residuals through HAC-GIC and leave-h-out strategies, we cannot directly integrate LP- and VAR-based estimators within a single GMA or CVA framework. The fundamental differences in residual construction prevent the formulation of a unified weighting scheme that encompasses both model classes.

2.4 α_{LP}-GMA and α_{LP}-CVA.

In this paper, we focus on combining LP- and VAR-variant estimators to obtain impulse response estimates through model averaging, as these are two of the most commonly used approaches in empirical macroeconomics with time series data. To address the structural differences in residuals between VARs and LPs, we introduce a secondary weight, $\alpha_{\text{LP}} \in [0, 1]$, which is applied to the LP weight vector obtained via GMA or CVA.

Let M_{LP} and M_{VAR} denote the number of LP-variant and VAR-variant models, respectively. The α_{LP}-GMA scheme proceeds in two steps.[2] First, we compute preliminary weights separately for LP and VAR models. For the VAR-variants, we use equation (2) with $M = M_{\text{VAR}}$ to obtain weights $\omega_{m,\text{BMA}_{\text{VAR},h}}$ for $m = 1, \ldots, M_{\text{VAR}}$. Due to the properties of VARs, these weights are constant across all horizons. For the LP-variants, we apply equation (2) using HAC-GIC values and $M = M_{\text{LP}}$ to compute $\omega_{m,\text{GMA}_{\text{LP}},h}$ for $m = 1, \ldots, M_{\text{LP}}$. These weights are horizon-specific, reflecting the separate estimation of LPs at each horizon h.

In the second step, we define the α_{LP}-GMA weights $\omega_{m,\alpha_{\text{LP}}\text{-GMA},h}$ as:

$$\omega_{m,\alpha_{\text{LP}}\text{-GMA},h} = \begin{cases} \alpha_{\text{LP}} \cdot \omega_{m,\text{GMA}_{\text{LP}},h}, & \text{for } m = 1, \ldots, M_{\text{LP}}, \\ (1 - \alpha_{\text{LP}}) \cdot \omega_{m-M_{\text{LP}},\text{BMA}_{\text{VAR}},h}, & \text{for } m = M_{\text{LP}} + 1, \ldots, M_{\text{LP}} + M_{\text{VAR}}. \end{cases} \tag{7}$$

It is straightforward to verify that the weights sum to one:

$$\begin{aligned} \sum_{m=1}^{M_{\text{LP}}+M_{\text{VAR}}} \omega_{m,\alpha_{\text{LP}}\text{-GMA},h} &= \sum_{m=1}^{M_{\text{LP}}} \omega_{m,\alpha_{\text{LP}}\text{-GMA},h} + \sum_{m=M_{\text{LP}}+1}^{M_{\text{LP}}+M_{\text{VAR}}} \omega_{m,\alpha_{\text{LP}}\text{-GMA},h} \\ &= \alpha_{\text{LP}} \cdot \underbrace{\sum_{m=1}^{M_{\text{LP}}} \omega_{m,\text{GMA}_{\text{LP}},h}}_{=1} + (1 - \alpha_{\text{LP}}) \cdot \underbrace{\sum_{m=1}^{M_{\text{VAR}}} \omega_{m,\text{BMA}_{\text{VAR}},h}}_{=1} = 1. \end{aligned}$$

We construct α_{LP}-CVA weights following a procedure parallel to that of α_{LP}-GMA. In the first step, we compute the horizon-specific weights $\omega_{m,\text{CVA}_{\text{LP}},h}$ for LP-based estimators using equation (6), and the horizon-invariant weights $\omega_{m,\text{CVA}_{\text{VAR}}}$ for VAR-based estimators. In the second step, we apply the same combination rule as in equation (7), substituting the GMA weights with the corresponding CVA weights.

The guidance of deciding α_{LP}. Researchers seeking to apply α_{LP} in impulse response analyses often face the challenge of selecting an appropriate value, as the true impulse responses are generally unobservable. Traditionally, the mean squared error (MSE) serves as a useful benchmark, capturing total estimation error by balancing bias and variance. However, in the absence of true responses, alternative evaluation metrics are necessary. In this study, we propose two data-driven approaches to guide the selection of α_{LP}.

1. Mean squared prediction error (abbreviated "MSPE")

   One approach involves using the MSPE of LP- and VAR-variant estimators as a proxy for in-sample predictive performance. MSPE evaluates the average squared difference between realized and predicted outcomes; higher MSPE indicates worse predictive accuracy. Based on this criterion, we propose a functional form for computing

[2] Although BMA is used for VAR-variant estimators and GMA is used for LP-variant estimators, GMA generalizes BMA in structure and form. For notational simplicity, we refer to the combined scheme as α_{LP}-GMA.

α_{LP}. It is normalized inverse-MSPE ratio:

$$\alpha_{\text{LP, MSPE},h} = 1 - \frac{\sum_{m=1}^{M_{\text{LP}}} \text{MSPE}_{\text{LP},m,h}}{\sum_{m=1}^{M_{\text{LP}}} \text{MSPE}_{\text{LP},m,h} + \sum_{m=1}^{M_{\text{VAR}}} \text{MSPE}_{\text{VAR},m}}. \tag{8}$$

Here, the MSPE for each estimator is computed as:

$$\text{MSPE}_{\text{LP},m,h} = \frac{1}{T-p-h} \sum_{t=p+1}^{T-h} (y_{t+h} - \hat{y}_{m,t+h|t})^2,$$

$$\text{MSPE}_{\text{VAR},m} = \frac{1}{T-p} \sum_{t=p+1}^{T} (y_t - \hat{y}_{m,t|t-h})^2,$$

where y_t denotes the realized value and $\hat{y}_{m,t+h|t}$ or $\hat{y}_{m,t|t-h}$ denotes the predicted value based on model m using information available up to time t. The prediction index reflects the structure of each estimator: LP-based forecasts are horizon-specific, whereas VAR-based forecasts are derived recursively.

2. R^2 (abbreviated "RS")

   The R^2 statistic quantifies the proportion of variance in the dependent variable y_t that is explained by the independent variables in a regression model. It serves as a measure of in-sample fit, indicating how well the model captures variation in the data. Using a common sample $\{t = 1, \ldots, T\}$, the R^2 at horizon h for model m is defined as:

$$R^2_{m,h} = 1 - \frac{SSR_{m,h}}{TSS_{m,h}},$$

   where $SSR_{m,h}$ is the sum of squared residuals and $TSS_{m,h}$ is the total sum of squares. For LP-variant estimators, R^2 is computed as:

$$R^2_{\text{LP},m,h} = 1 - \frac{\sum_{t=p+1}^{T-h} (y_{t+h} - \hat{y}_{m,t+h})^2}{\sum_{t=p+1}^{T-h} (y_{t+h} - \bar{y}_{\text{LP}})^2},$$

$$\bar{y}_{\text{LP}} = \frac{1}{T-p-h} \sum_{t=p+1}^{T-h} y_{t+h}.$$

   For VAR-variant estimators, residuals are horizon-invariant, so the R^2 value is computed once per model:

$$R^2_{\text{VAR},m} = 1 - \frac{\sum_{t=p+1}^{T} \hat{u}^2_{m,t}}{\sum_{t=p+1}^{T} (y_t - \bar{y}_{\text{VAR}})^2},$$

$$\bar{y}_{\text{VAR}} = \frac{1}{T-p} \sum_{t=p+1}^{T} y_t.$$

   Here, $\hat{u}_{m,t}$ denotes the residuals from model m's VAR estimation, which are constant

across horizons. A value of R^2 closer to 1 suggests that the model explains a substantial portion of the variance in y_t, indicating a better in-sample fit. Conversely, an R^2 near 0 implies poor explanatory power. Based on this metric, we propose determining α_{LP} as:

$$\alpha_{\mathrm{LP,RS},h} = \frac{\sum_{m=1}^{M_{\mathrm{LP}}} R^2_{\mathrm{LP},m,h}}{\sum_{m=1}^{M_{\mathrm{LP}}} R^2_{\mathrm{LP},m,h} + \sum_{m=1}^{M_{\mathrm{VAR}}} R^2_{\mathrm{VAR},m}}. \tag{9}$$

As described above, by reassigning the weights using α_{LP}, we obtain more robust impulse response estimates that benefit from the strengths of model averaging such as reduced reliance on a single estimation method and increased flexibility. Regardless of sophistication, reliance on a single method always carries the risk of model misspecification, as it is often difficult for researchers to identify the true DGP in empirical applications. In contrast, model averaging is inherently more adaptable, allowing researchers to use a common estimation framework across different lag lengths or to combine models with varying structural assumptions.

In addition, the α_{LP}-GMA and α_{LP}-CVA approaches offer two further advantages.

First, from a technical perspective, reassigning weights via α_{LP} constitutes a straightforward yet robust enhancement of traditional model averaging frameworks. In the case of α_{LP}-GMA, it obviates the need to construct a unified information criterion applicable to both LP- and VAR-variants. Instead, the BIC is used for VAR-variants, whose residuals tend to satisfy the white noise assumption, while the HAC-GIC is used for LP-variants, which may exhibit heteroskedasticity and autocorrelation. This alignment preserves consistency with the BMA logic while incorporating residual structure and model complexity appropriately through the HAC-GIC. By integrating LP- and VAR-variant estimators only after separately computing their GMA (or BMA) weights, the method circumvents the need to define universal penalty terms across model types. Furthermore, it avoids problems related to differing residual magnitudes when the forecast horizon $h > 0$. In other words, by avoiding direct comparisons between estimators based on incompatible model assumptions, the approach mitigates complications arising from structural differences between LP- and VAR-variants.

Second, introducing the secondary weight α_{LP} allows us to leverage the distinct advantages of LP and VAR approaches. As discussed previously, VARs compute impulse responses from a single model fit, maintaining consistent residual properties across horizons. This results in horizon-constant weight vectors in BMA and CVA. While VARs integrate more information by exploiting full-sample dynamics, they do not offer flexibility in assigning horizon-specific weights. In contrast, LPs estimate responses separately at each horizon and thus allow for horizon-specific weighting, though they use less information than VARs when $h > 0$. Consequently, the α_{LP} model averaging framework enhances robustness by achieving horizon-specific weighting—as in LPs—while also benefiting from the greater informational content of VARs.

In the next section, we introduce popular LP-based and VAR-based estimators, includ-

ing their least-squares variants. Section 4 then presents simulation results based on different combinations of the estimation methods introduced in Section 3.

3 Estimation Methods

In this section, we summarize popular LP-based and VAR-based estimation methods commonly used in applied macroeconomic research. These estimators serve as the inputs for our simulation study, where we construct weighted averages of impulse responses using the model averaging schemes introduced in Section 2.

LOCAL PROJECTION APPROACHES. As originally proposed by Jordà (2005), the key idea behind LPs is to estimate the effect of shocks at each horizon by directly regressing the future outcome on current covariates. Unlike VAR-based approaches, which extrapolate impulse responses recursively from a single model specification, LPs estimate a separate regression at each horizon. In our study, we consider four LP-based approaches, described below.

1. **Least-squares LP** (abbreviated "LS LP"). The LS LP does not require specification of the full multivariate dynamic system, as it directly exploits the autocovariance structure in the data up to the horizon of interest. While it typically provides a low-bias estimator of impulse responses, it may exhibit high variance, particularly for longer horizons, due to limited information in finite samples. The general form of the LS LP at horizon h is given by:

$$y_{t+h} = \mu_h + \beta_h x_t + \text{control variables} + \text{residual}_{t+h}, \tag{10}$$

   where y_{t+h} is the outcome variable, x_t is the observed shock, and the control variables typically include contemporaneous and lagged endogenous variables. The residual term captures unexplained variation at horizon h.

2. **Bias-corrected LP** (abbreviated "BC LP"). Herbst and Johannsen (2024) propose a bias-corrected version of LP, which can improve estimation accuracy, especially when the data exhibits strong persistence or the sample size is small. Although the approximate bias of the LP estimator decreases with sample size T, the proposed correction provides a meaningful adjustment for empirically relevant settings.

3. **Bayesian LP** (abbreviated "BLP"). Miranda-Agrippino and Ricco (2021a,b) propose a Bayesian LP that addresses the bias-variance trade-off of LP and VAR estimators at each horizon. They place conjugate Normal-inverse Wishart priors on the LP coefficients. The posterior distribution is obtained by combining these priors with the likelihood function conditional on the data. This approach does not explicitly modeling the autocorrelation in the projection residuals.

4. **Penalized LP** (abbreviated "Pen LP"). Barnichon and Brownlees (2019) propose a smoothed LP estimator based on penalized B-spline smoothing. This approach reduces the variance of LS LP by incorporating a penalty term that shrinks the estimated impulse responses toward a low-order polynomial, rather than zero as in traditional shrinkage methods. The penalty parameter is selected through cross-validation, allowing for a balance between bias and variance across horizons and over time.

VAR APPROACHES. As proposed by Sims (1980), the vector autoregression (VAR) model of order p estimates impulse responses out to horizon p by iterating its recursive form under the assumption that existing dynamics continue. Since the impulse response at horizon h is extrapolated from the first p autocovariances, the VAR approach typically delivers lower-variance estimates than LP, particularly when the lag structure is misspecified. However, the ordering of variables is crucial when identifying structural shocks.

1. **Least-squares VAR** (abbreviated "LS VAR"). The standard VAR model estimates reduced-form coefficients equation-by-equation using ordinary least squares (OLS). Impulse responses are then obtained by recursively applying the estimated system. The general form of a VAR(p) model is:

$$w_t = \sum_{l=1}^{p} B_l \cdot w_{t-l} + u_t, \tag{11}$$

   where w_t is a vector of observed variables, B_l are coefficient matrices for each lag l, and u_t denotes reduced-form residuals, typically assumed to be white noise.

2. **Bias-corrected VAR** (abbreviated "BC VAR"). Pope (1990) discusses the biases that can arise in estimating VAR models when residuals are non-Gaussian or heteroskedastic. Bias correction methods adjust the OLS estimates to reduce small-sample distortions, particularly in models with highly persistent time series or short sample spans.

3. **Bayesian VAR** (abbreviated "BVAR"). Giannone et al. (2015) propose a Bayesian VAR approach in which reduced-form coefficients are estimated using priors with automatic hyperparameter selection. In our implementation, we follow the procedure of Li et al. (2024), reporting posterior means of impulse responses computed from 100 posterior draws. The prior specification follows the Minnesota prior, which allows for cointegration. The degree of shrinkage—determined by the prior variance hyperparameters—is selected in a data-driven way by maximizing the marginal likelihood.

4 Monte Carlo simulations

This section presents the process and results of our simulation study. We begin by defining the empirically calibrated encompassing model in Section 4.1 and describe the correspond-

ing structural impulse response estimands in Section 4.2. Section 4.3 outlines implementation details, followed by a presentation of simulation results in Section 4.4. We adopt the approach attained by Li et al. (2024) for constructing the DGP and identifying population impulse response estimands throughout Section 4.1 to Section 4.3.[3]

4.1 Encompassing Model

DGP represents the underlying mechanism that generates the observed data. Before constructing it for simulation, we first develop a macroeconomic model that produces population impulse responses. Specifically, we use a dynamic factor model (DFM) estimated on the well-known dataset from Stock and Watson (2016). Following common applied practice, we employ a non-stationary variant of the DFM by using data in levels rather than in first differences.

In the DFM, $n_X \times 1$ vector X_t of observed macroeconomic time series (with large cross-sectional dimension) is driven by a low-dimensional $n_f \times 1$ vector of latent factors f_t and an $n_v \times 1$ vector of idiosyncratic components v_t. The latent factors are assumed to follow a non-stationary vector error correction model (VECM) with VAR(p_f) representation:

$$X_t = \Lambda f_t + v_t, \tag{12}$$

$$f_t = \Phi(L) f_{t-1} + H\varepsilon_t, \tag{13}$$

where $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n_f,t})'$ is an $n_f \times 1$ vector of aggregate shocks that are independently and identically distributed (i.i.d.) across t and mutually uncorrelated, with $\mathrm{Var}(\varepsilon_t) = I_{n_f}$. The $n_f \times n_f$ matrix H governs the contemporaneous responses of the latent factors to the structural shocks.

The idiosyncratic component $v_{i,t}$ associated with each observed variable $X_{i,t}$ follows a potentially nonstationary AR(p_v) process:

$$v_{it} = \Delta_i(L) v_{i,t-1} + \Xi_i \xi_{i,t}, \tag{14}$$

where $\xi_{i,t}$ is i.i.d. across t and i. We assume that all shocks and innovations are jointly Gaussian and homoskedastic.

4.2 DGP and Impulse Response Estimands

We now construct a lower-dimensional DGP for our simulation study based on equations (12)–(14). Specifically, we draw a random subset of $n_{\bar{w}}$ variables from the high-dimensional vector X_t, denoted by $\bar{w}_t \subset X_t$. As such, the variables in $\bar{w}_t$ follow the time series dynamics implied by the encompassing DFM and are driven by a combination of aggregate structural shocks ε_t and idiosyncratic components v_t. The specific combinations of variables used to construct $\bar{w}_t$ are selected randomly, as described in Section 4.3, resulting in a random lower-dimensional DGP.

[3] See Li et al. (2024), Section 3 and supplementary appendices for full details.

We examine three types of structural impulse response estimands commonly used to study the effects of policy shocks in applied macroeconometric analysis (Ramey, 2016; Stock and Watson, 2016). In the following notation, $\bar{w}_t$ denotes a vector of endogenous variables included in the lower-dimensional DGP, while w_t denotes the full set of observed time series. Within $\bar{w}_t$, the variable y_t denotes the outcome of interest (i.e., the response variable), and i_t denotes the policy variable used to normalize the shock when applicable. The variable $\varepsilon_{1,t}$ represents the first structural shock (when relevant), and z_t denotes an external instrument or proxy variable (when relevant), both of which are assumed observed but not endogenous to $\bar{w}_t$.

Observed shock identification. In this identification scheme, we assume that both the endogenous variables $\bar{w}_t$ and the first structural shock $\varepsilon_{1,t}$ are observed by the econometrician, so that $w_t = (\varepsilon_{1,t}, \bar{w}_t)'$. Our object of interest is the impulse response of the response variable y_t to a one-standard-deviation (i.e., one-unit) innovation in $\varepsilon_{1,t}$:

$$\theta_h \equiv \bar{\Lambda}_{l_y,\bullet}\Theta^f_{\bullet,1,h}, \quad h = 0, 1, 2, \ldots, \tag{15}$$

where $\Theta^f(L)$ denotes the impulse responses of the latent factors f_t to the structural shocks ε_t, as implied by equation (13), and $\bar{\Lambda}$ consists of those rows of Λ corresponding to $\bar{w}_t$. The index l_y indicates the location of y_t in the vector $\bar{w}_t$.

IV/proxy identification. In this case, the structural shock $\varepsilon_{1,t}$ is unobserved, and the econometrician instead observes a noisy proxy. The observed variable vector is given by $w_t = (z_t, \bar{w}_t)'$, where

$$z_t = \rho_z z_{t-1} + \varepsilon_{1,t} + \nu_t, \tag{16}$$

and ν_t is an i.i.d. process independent of all shocks and innovations in the DFM, with $\mathrm{Var}(\nu_t) = \sigma^2_\nu$. Following the unit effect normalization in Stock and Watson (2016), the impulse response of interest is:

$$\theta_h \equiv \frac{\bar{\Lambda}_{l_y,\bullet}\Theta^f_{\bullet,1,h}}{\bar{\Lambda}_{l_i,\bullet}\Theta^f_{\bullet,1,0}}, \tag{17}$$

where l_i denotes the location of the policy variable i_t in $\bar{w}_t$. This normalization defines the shock $\varepsilon_{1,t}$ such that it increases i_t by one unit on impact.

Recursive identification. Under this scheme, the econometrician observes only the endogenous variables: $w_t = \bar{w}_t$. Following the large literature on recursive identification in VARs (e.g., Christiano et al. (1999)), the estimand is the impulse response derived from a recursive (Cholesky) orthogonalization of the reduced-form (Wold) forecast errors in the VAR(∞) process for $\bar{w}_t$. The shock of interest corresponds to the orthogonalized innovation to a policy variable in w_t, which generally differs from any of the structural shocks $\varepsilon_{j,t}$

in the DFM.

4.3 Implementation

This section outlines how we calibrate the DFM using non-stationary data and estimate impulse responses under each identification scheme employed in the simulation study.

DFM Parameters. We parameterize the DFM equations (12)—(14) by estimating the model using the dataset of Stock and Watson (2016). Rather than transforming all series to stationarity, we apply a milder transformation scheme than Stock and Watson (2016): we retain variables in levels when they are differenced once in their setup, and we apply log first differences when they use log second differences.

In equation (12), the vector of observables X_t contains quarterly observations on 207 time series over the period 1959Q1–2014Q4. These series span categories including real activity, prices, productivity and earnings, interest rates and spreads, money and credit, asset and wealth indicators, oil markets, and international activity.[4]

We adopt the estimation approach of Li et al. (2024) for handling non-stationary data. Specifically, we estimate the non-stationary DFM by first extracting factors from differenced data, accumulating them, and then fitting a VECM to the accumulated factors, following Bai and Ng (2004) and Barigozzi et al. (2021). As in Stock and Watson (2016), we set the number of factors to $n_f = 6$. The cointegration rank of the factor VECM is determined using the maximum eigenvalue test of Johansen (1995), which indicates that the latent factors are driven by four common stochastic trends.

We set the lag lengths to $p_f = p_v = 4$, which is conservative in that it allows for rich dynamics and lies at the upper limit of what is typically recommended by the AIC. Given the imposed unit roots in the factor VECM and the persistence observed in many estimated idiosyncratic components, the fitted DFM exhibits substantial persistence. All DFM parameters—except for the structural impact matrix H—are estimated through this procedure.

DGP and Estimand Selection. We consider two protocols for selecting the set of observables $\bar{w}_t \subset X_t$. One protocol is designed to replicate the effects of monetary policy shocks, and the other targets the effects of fiscal policy shocks. For each type of policy shock, we randomly draw a set of macroeconomic observables $\bar{w}_t$ with $n_{\bar{w}} = 5$, resulting in a total of two DGPs.

For each DGP, we impose structural inclusion rules. In the monetary policy DGP, the effective federal funds rate is included in $\bar{w}_t$, while in the fiscal policy DGP, federal government spending is included. These variables serve dual roles as both the normalization variable i_t and the policy variable in the IV and recursive identification schemes. The remaining four variables in $\bar{w}_t$ are drawn randomly from X_t, subject to the constraint that

[4] See Stock and Watson (2016) and their data appendix for a complete list of variables and their classification.

at least one variable captures real activity and at least one reflects prices. The response variable y_t is randomly selected from the remaining four variables in $\bar{w}_t$, excluding i_t.[5]

For each DGP, we implement the three identification schemes described in Section 4.2 as follows:

1. **Observed Shock.** In equation (13), we select the structural impact matrix H to maximize the contemporaneous effect of the shock $\varepsilon_{1,t}$ on the policy variable. This optimization is performed under the constraint that H is consistent with the estimated variance-covariance matrix of reduced-form factor innovations. This ensures that policy shocks have substantial short-run effects on the corresponding policy variables.

2. **IV.** The matrix H is defined as in the observed shock experiment. In the measurement equation (16), we set $\rho_z = 0.25$. The noise variance σ_ν^2 is calibrated to yield a population IV first-stage F-statistic between 10 and 30. This range reflects empirical heterogeneity in instrument strength and ensures a realistic signal-to-noise ratio for the proxy variable.

3. **Recursive.** Following Christiano et al. (1999), we place the federal funds rate last in the ordering of variables for the monetary policy DGP. This ensures that other variables do not contemporaneously respond to monetary innovations. For the fiscal policy DGP, we follow Blanchard and Perotti (2002) by ordering government spending first, thereby allowing fiscal policy to respond to other innovations only with a lag in the recursive VAR structure.

4.4 Simulation Results

This subsection presents the results for the DGPs corresponding to fiscal and monetary policy shocks. The outcomes are reported separately by shock and identification scheme. We primarily focus on results obtained under the observed shock identification, while those based on the IV and recursive identification schemes are reported in Appendices D.2 and D.3, respectively. To approximate population impulse responses, we simulate time series of length $T = 200$ quarters and average across 5,000 Monte Carlo replications.

Models. We evaluate the performance of impulse response estimators obtained from the transformed traditional model averaging schemes introduced in Section 2. For comparison, we also include two individual estimation methods selected from Section 3, which serve as benchmarks against the model averaging approaches. All estimators are used to approximate the same population impulse responses defined in Section 4.2.

1. **Comparative group:** BC LP and BVAR.

[5]The randomly selected variables, excluding the policy variables, include the consumer price index for all urban consumers (core CPI), total nonfarm payroll employment, real gross domestic product, and the number of unemployed for 5–14 weeks. The policy variables are real government consumption expenditures and gross investment (fiscal DGP), and the effective federal funds rate (monetary DGP).

2. **Model Averaging (abbreviated "MAVG"):**

   - **MAVG**$_{\text{LP}}$: EQ$_{\text{LP}}$, GMA$_{\text{LP}}$, and CVA$_{\text{LP}}$.
   - **MAVG**$_{\text{VAR}}$: EQ$_{\text{VAR}}$, GMA$_{\text{VAR}}$, and CVA$_{\text{VAR}}$.
   - **MAVG**$_{\text{ALL}}$: EQ$_{\text{ALL}}$, $\alpha_{\text{LP},st}$-GMA, and $\alpha_{\text{LP},st}$-CVA.

Note that GMA$_{\text{VAR}}$ is identical to BMA when applied to VAR-based estimators. However, we use the label GMA$_{\text{VAR}}$ starting from this section to maintain consistency with GMA$_{\text{LP}}$, as both share the same weighting framework despite relying on different model selection criteria. Here, the subscript st refers to the specific guidance used to determine α_{LP}, such as MSPE or RS. We compare these model averaging schemes with the two benchmark estimators, as well as with each other. The benchmark estimators are chosen based on their strengths: BC LP tends to perform well in terms of bias, especially at short and intermediate horizons, while BVAR generally exhibits lower variance across horizons. Results for bias and standard deviation from all individual estimation methods are reported in Appendix D.1.

We classify the model averaging estimators into three groups based on the types of underlying estimators used in their construction. The first group, MAVG$_{\text{LP}}$, is computed using EQ, GMA, and CVA schemes applied to LP-based estimators: LS LP, BC LP, BLP, and Pen LP. The second group, MAVG$_{\text{VAR}}$, is computed using EQ, BMA, and CVA applied to VAR-based estimators: LS VAR, BC VAR, and BVAR. The final group, MAVG$_{\text{ALL}}$, is constructed using all seven LP- and VAR-based estimators under EQ, $\alpha_{\text{LP},st}$-GMA, and $\alpha_{\text{LP},st}$-CVA schemes.

Lag Length Selection. All estimation models use four lags of the data series w_t as control variables. In our DGP, the AIC almost always selects very short lag lengths, typically less than or equal to 4. Accordingly, the simulation data are generated using the rule $p = \max\{\hat{p}_{\text{AIC}}, 4\}$. We evaluate impulse response estimates at horizons $h = 0$ through $h = 20$.

Results. Figures 1 through 8 present the simulation results evaluating the average performance of impulse response estimators in response to both fiscal and monetary policy shocks, under the observed shock identification scheme. The evaluation metrics include absolute bias ($aBias$), standard deviation (SD), and unweighted root MSE at each forecast horizon. Bias measures the systematic deviation of an estimator's mean from the true response, while standard deviation quantifies the dispersion of the estimator around its mean. Together, these two components characterize the estimator's accuracy and reliability across horizons. The $aBias$ and SD are defined as follows:

$$aBias(\hat{\theta}_{m,h}) = \left| E(\hat{\theta}_{m,h}) - \theta_h \right|, \tag{18}$$

$$SD(\hat{\theta}_{m,h}) = \sqrt{Var(\hat{\theta}_{m,h})}, \tag{19}$$

where

$$Var(\hat{\theta}_{m,h}) = E\Big[\Big(\hat{\theta}_{m,h} - E[\hat{\theta}_{m,h}]\Big)^2\Big].$$

The unweighted MSE serves as a standard loss function widely used by researchers to evaluate estimator performance. It reflects the total error by assigning equal weight to both bias and variance. While this formulation is commonly adopted as a default, researchers may prefer alternative weighting schemes depending on their specific tolerance for bias or variance. The MSE is defined as:

$$\begin{aligned} \text{MSE}(\hat{\theta}_{m,h}) &= E\Big[(\hat{\theta}_{m,h} - \theta_h)^2\Big] \\ &= Bias(\hat{\theta}_{m,h})^2 + Var(\hat{\theta}_{m,h}), \end{aligned} \tag{20}$$

where $h = 0, \ldots, 20$ denotes the forecast horizon, and m indexes model abbreviations.

Figure 1 illustrates the $aBias$ and SD of model averaging estimators using only LP-based (GMA_{LP}, CVA_{LP}) or VAR-based (GMA_{VAR}, CVA_{VAR}) methods, for fiscal shocks (top panel) and monetary shocks (bottom panel). For horizons $h \leq p = 4$, most estimators perform similarly across both metrics, owing to their comparable asymptotic properties. However, beyond $h = 4$, a distinct bias-variance trade-off emerges between the MAVG_{LP} and MAVG_{VAR} groups, particularly over intermediate horizons. This trade-off is more pronounced under fiscal shocks and extends into the long-horizon period. In contrast, for monetary shocks, the divergence is milder in the long run, with VAR-based estimators demonstrating lower bias and variance beyond $h = 12$.[6]

Before adopting a model averaging scheme, researchers may consider the following three questions to evaluate its relevance. We summarize answers to each and highlight the potential benefits of combining LP- and VAR-based estimators through model averaging.

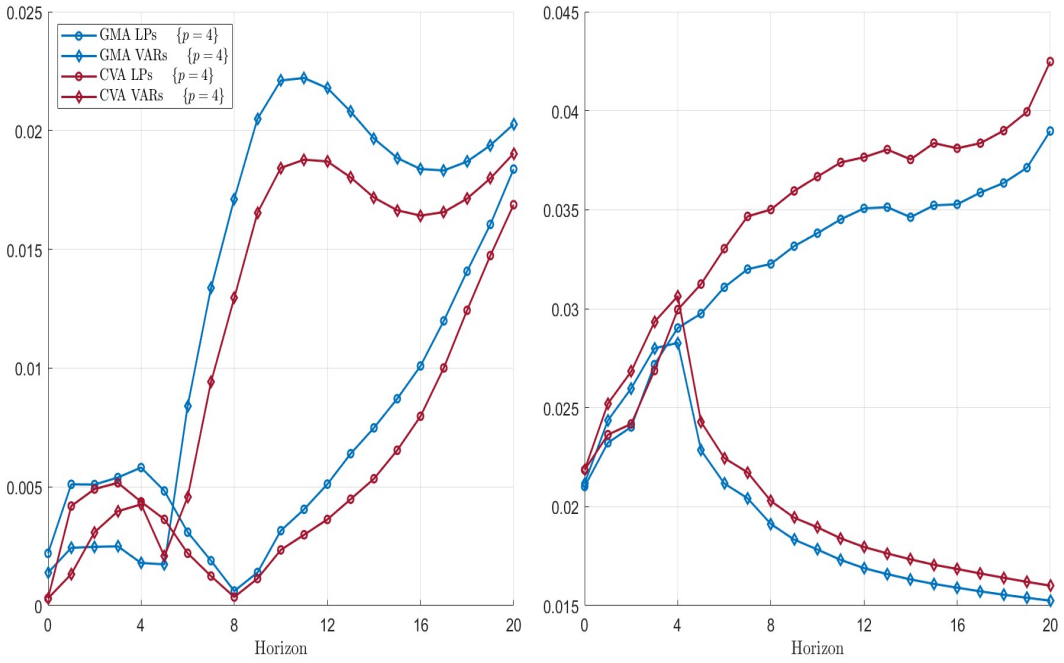
Q1: When should researchers consider MAVG estimators over single estimators?

Figures 2 through 5 compare two sophisticated single estimators with those in the MAVG_{LP}, MAVG_{VAR}, and MAVG_{ALL} groups under both fiscal and monetary shocks. For fiscal shocks (Figures 2 and 3), BC LP yields the lowest bias for short horizons ($h < 8$), while CVA_{LP} outperforms in intermediate and long horizons ($8 \leq h \leq 17$). Regarding variability, BVAR performs best up to $h = 11$ and remains competitive with MAVG_{VAR} beyond that. For monetary shocks (Figures 4 and 5), BC LP again dominates in terms of bias up to $h = 12$, after which BVAR becomes the most accurate. In terms of variance, BVAR consistently outperforms all estimators across nearly the entire horizon range.

These findings suggest that if researchers seek to minimize either bias or variance in isolation, selecting a well-performing single estimator such as BC LP or BVAR may be preferable, as they require fewer computations than model averaging methods that com-

[6] This pattern is consistent with findings in Li et al. (2024) and Kilian and Kim (2011).

Observed Fiscal Shock
$aBias$ (Left) and SD (Right) of Estimators



Observed Monetary Shock
$aBias$ (Left) and SD (Right) of Estimators

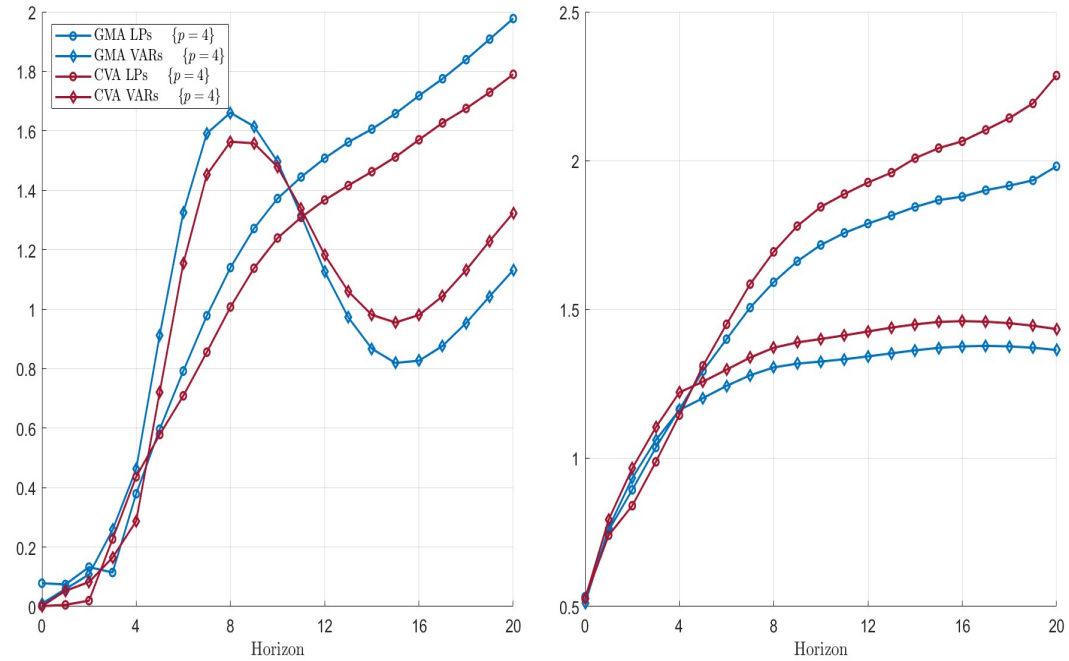


Figure 1: Average absolute bias and standard deviation when a shock is observed.

bine multiple estimators. However, these single estimators often exhibit uneven performance across horizons—particularly evident in the case of monetary shocks. Moreover, the MSE results indicate that when bias and variance are weighted equally, MAVG estimators tend to provide more balanced and robust performance. The advantages of model averaging become even more pronounced if researchers tailor their loss function by assigning specific weights to bias and variance, depending on their application.

Q2: Under what circumstances should researchers opt for MAVG_{ALL} estimators over MAVG_{LP} or MAVG_{VAR}?

After deciding to adopt a model averaging scheme to reduce total errors across horizons, researchers may face difficulty in choosing the most appropriate estimator among the available options. Insight can be gained by comparing the performance of MAVG groups, as shown in Figures 6 through 7. In the case of fiscal shocks, if minimizing bias is the primary goal, MAVG_{LP} group performs best for horizons $h > 4$. If variance reduction is the priority, MAVG_{VAR} group becomes preferable for the same horizon range. However, when bias and variance receive equal weight in the loss function, α_{LP}-GMA and α_{LP}-CVA outperform MAVG_{LP} for $h > 4$, and compete effectively with CVA_{VAR} during intermediate horizons, particularly for $8 \leq h < 13$. For monetary shocks, the pattern differs slightly. To minimize bias, MAVG_{LP} group is favored for horizons $4 < h \leq 12$, while MAVG_{VAR} group becomes more suitable thereafter. When variance reduction is the focus, MAVG_{VAR} estimators are generally preferred for $h \geq 4$. When equal weight is given to bias and variance, the $\alpha_{\text{LP,MSPE}}$-GMA and $\alpha_{\text{LP,MSPE}}$-CVA estimators offer strong performance over intermediate and long horizons, comparable to both MAVG_{LP} and MAVG_{VAR}. Likewise, $\alpha_{\text{LP,RS}}$-GMA and $\alpha_{\text{LP,RS}}$-CVA reduce overall error for $4 < h < 10$, performing similarly to MAVG_{LP}.

Overall, α_{LP}-GMA and α_{LP}-CVA deliver comparable or superior performance across shock types when unweighted MSE is used as the loss function. This suggests that these estimators are robust choices, particularly when researchers seek to balance bias and variance according to context-specific priorities.

Q3: Upon deciding to utilize MAVG_{ALL} estimators, how should researchers select the most appropriate one?

In Section 2, we introduced two approaches for determining α_{LP}—the weight that governs the combination of LP- and VAR-based estimators. These approaches generate four possible MAVG_{ALL} estimators: $\alpha_{\text{LP,MSPE}}$-GMA, $\alpha_{\text{LP,RS}}$-GMA, $\alpha_{\text{LP,MSPE}}$-CVA, and $\alpha_{\text{LP,RS}}$-CVA. Figure 8 compares their performance under both fiscal and monetary shocks. In the case of a fiscal shock, the comparison is relatively straightforward. For bias reduction, the α_{LP}-CVA estimators outperform the GMA-based alternatives, as CVA assigns weights based on direct predictive performance. However, in terms of variance, the GMA estimators yield lower standard deviations than their CVA counterparts. Under a monetary shock, the $\alpha_{\text{LP,RS}}$-CVA estimator performs best in terms of bias up to horizon $h = 11$. After

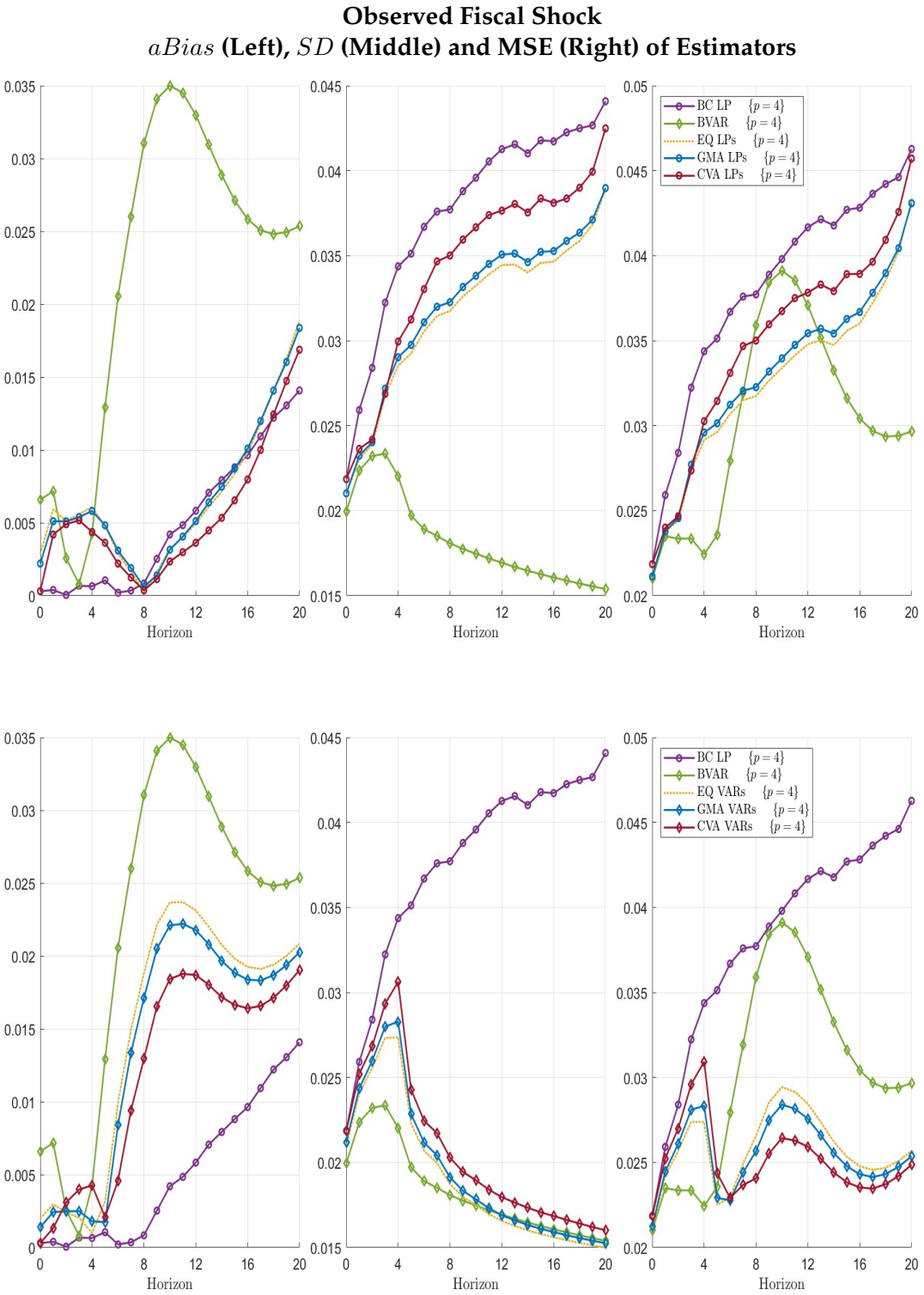


Figure 2: Average absolute bias, standard deviation and MSE when a fiscal shock is observed. The **top panel** compares estimators in the MAVG_{LP} group with BC LP and BVAR, while the **bottom panel** compares estimators in MAVG_{VAR} group with the same benchmarks.

Observed Fiscal Shock
$aBias$ (Left) , SD (Middle) and MSE (Right) of Estimators

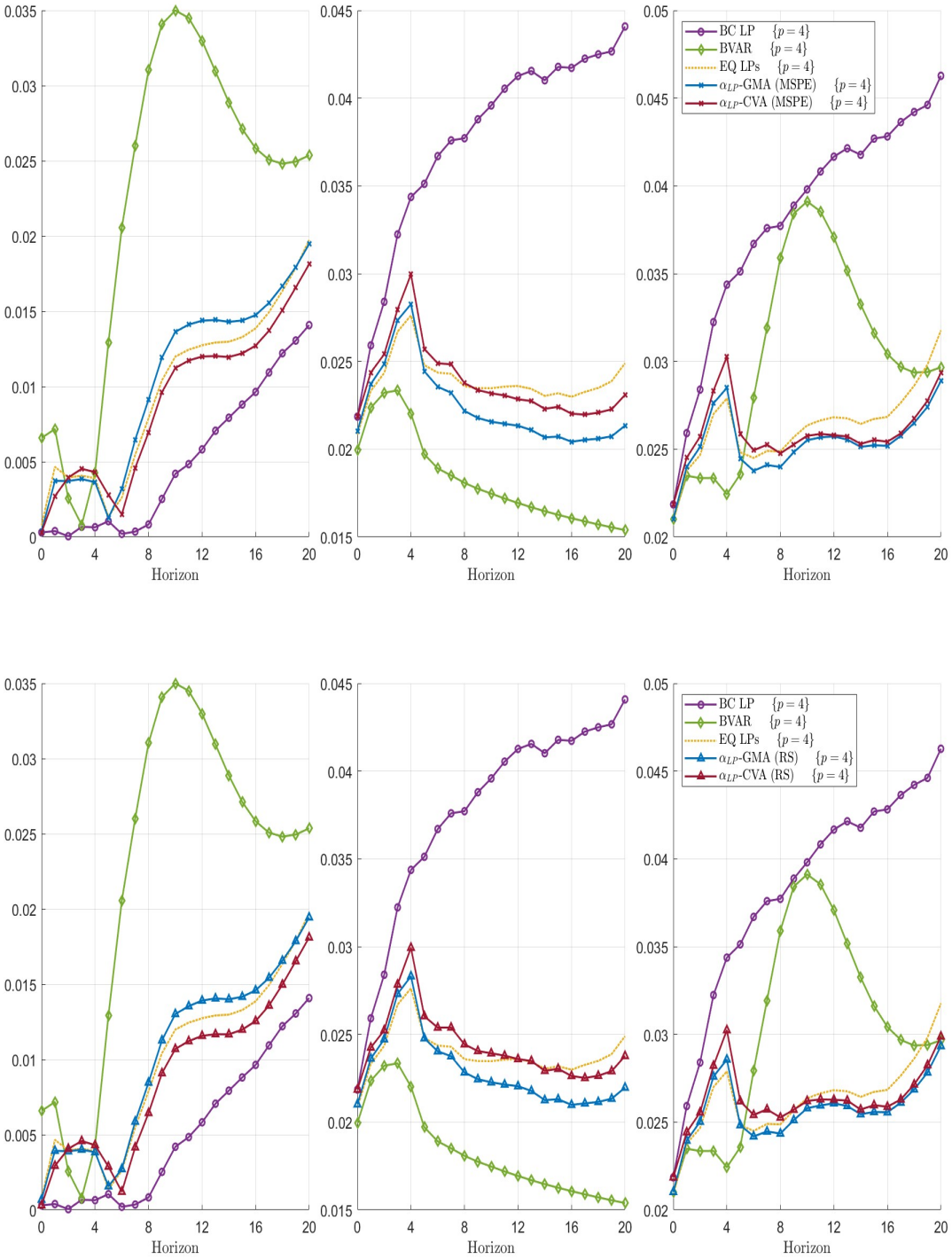


Figure 3: Average absolute bias, standard deviation and MSE when a fiscal shock is observed. The **top panel** compares estimators in the MAVG_{ALL} group using MSPE-guided α_{LP} values with BC LP and BVAR. The **bottom panel** compares estimators in the MAVG_{ALL} group using R^2-guided α_{LP} values with the same benchmarks.

Observed Monetary Shock
***aBias* (Left) , *SD* (Middle) and MSE (Right) of Estimators**

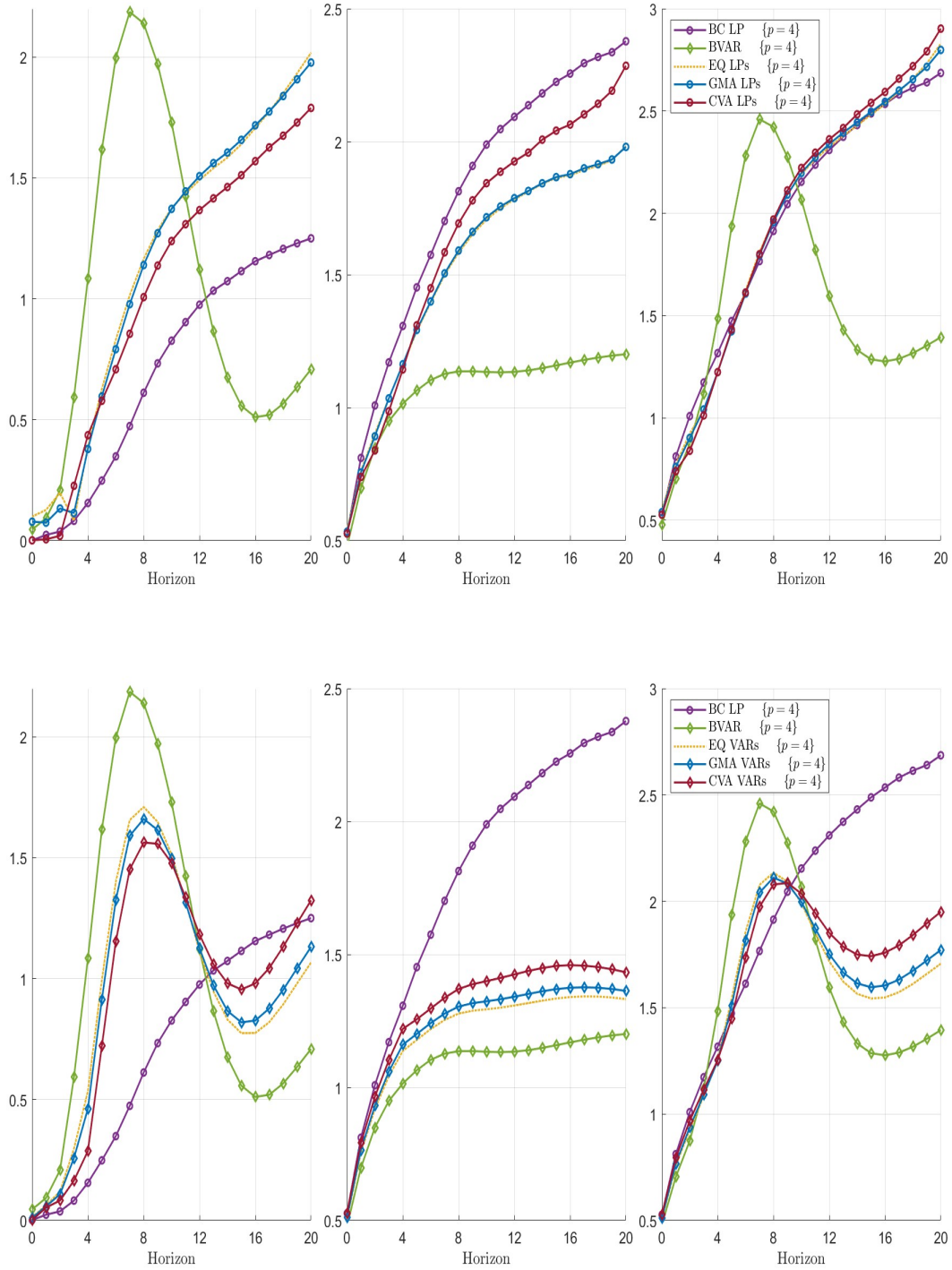


Figure 4: Average absolute bias, standard deviation and MSE when a monetary shock is observed. The **top panel** compares estimators in the MAVG_{LP} group with BC LP and BVAR, while the **bottom panel** compares estimators in MAVG_{VAR} group with the same benchmarks.

Observed Monetary Shock
$aBias$ (Left) , SD (Middle) and MSE (Right) of Estimators

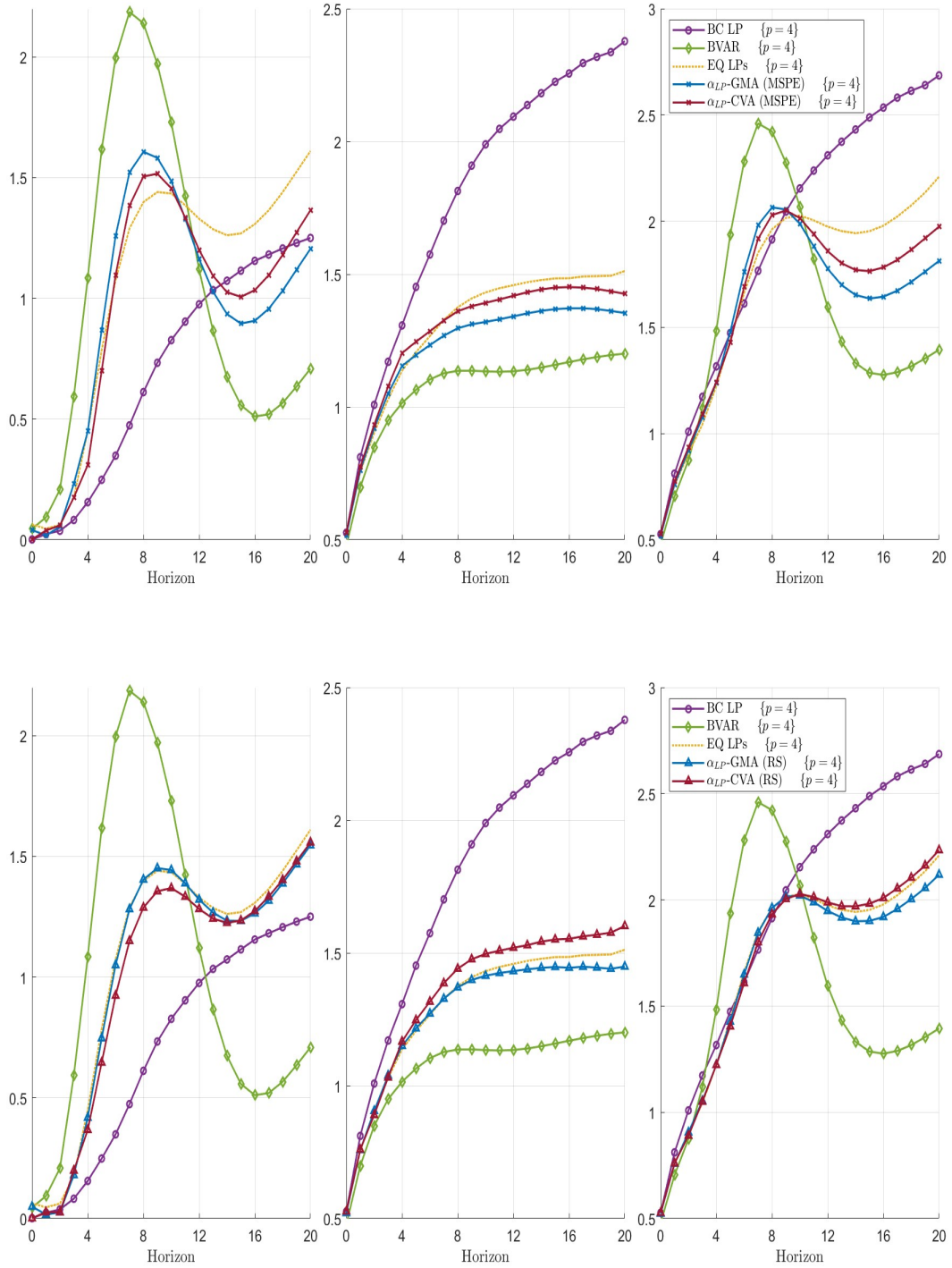


Figure 5: Average absolute bias, standard deviation and MSE when a monetary shock is observed. The **top panel** compares estimators in the MAVG_{ALL} group using MSPE-guided α_{LP} values with BC LP and BVAR. The **bottom panel** compares estimators in the MAVG_{ALL} group using R^2-guided α_{LP} values with the same benchmarks.

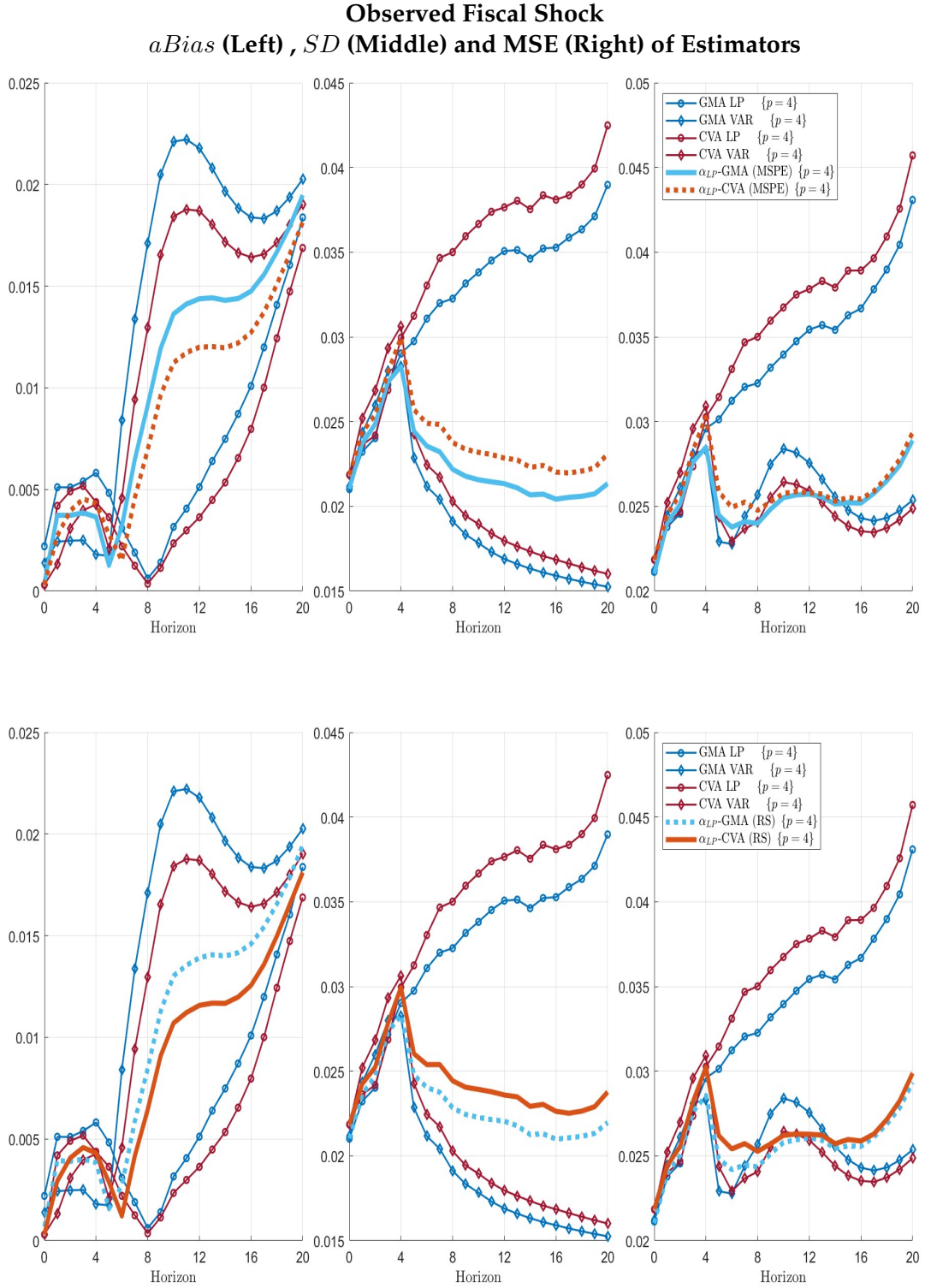


Figure 6: Average absolute bias, standard deviation, and MSE when a fiscal shock is observed. The **top panel** compares estimators in the MAVG_{LP} and MAVG_{VAR} groups with those in the MAVG_{ALL} group using MSPE-guided α_{LP} values. The **bottom panel** compares the same groups using R^2-guided α_{LP} values. All MAVG groups exclude EQ-based estimators.

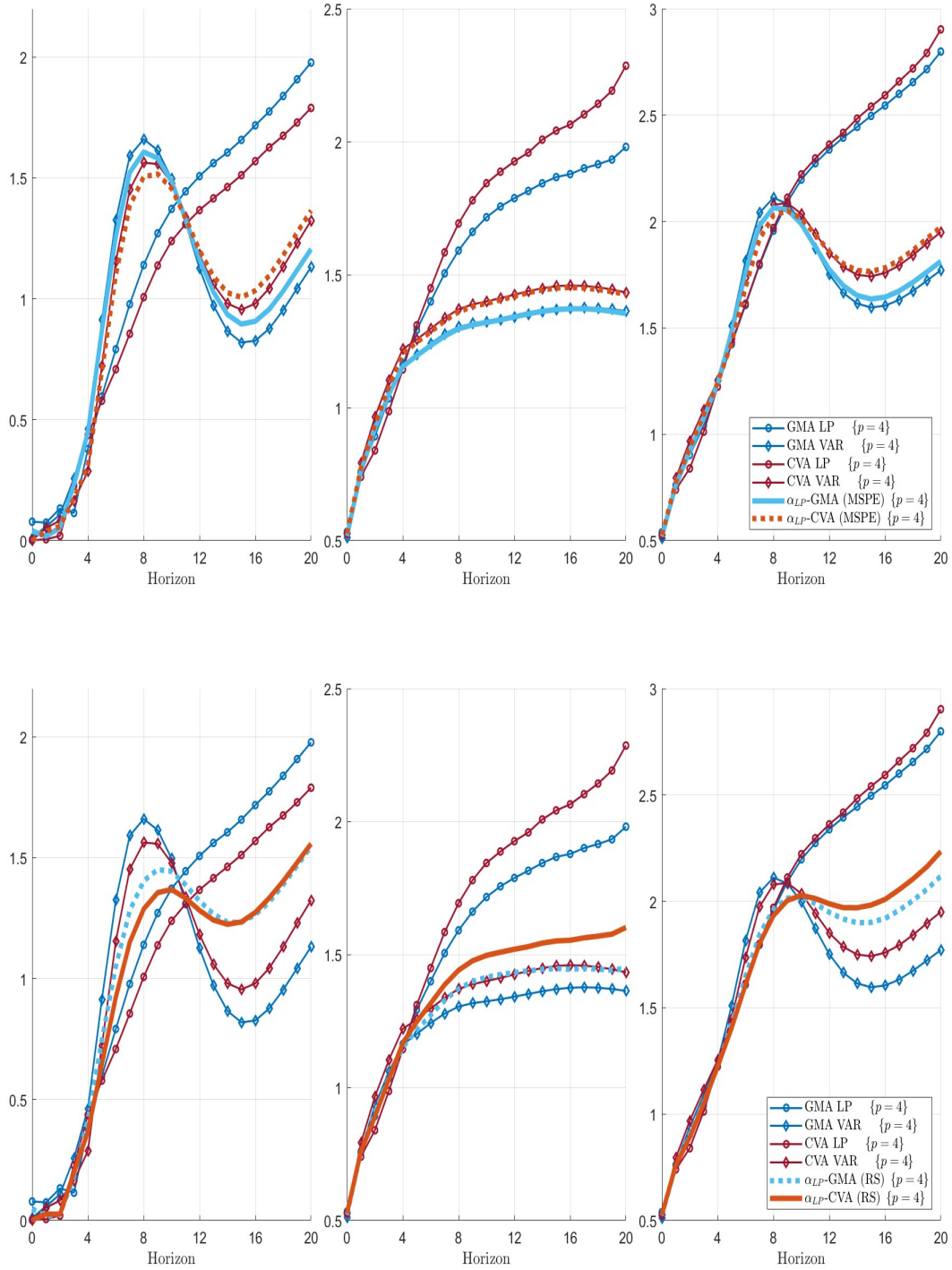


Figure 7: Average absolute bias, standard deviation, and MSE when a monetary shock is observed. The **top panel** compares estimators in the MAVG_{LP} and MAVG_{VAR} groups with those in the MAVG_{ALL} group using MSPE-guided α_{LP} values. The **bottom panel** compares the same groups using R^2-guided α_{LP} values. All MAVG groups exclude EQ-based estimators.

this point, $\alpha_{\text{LP,MSPE}}$-GMA exhibits lower bias. Regarding variability, $\alpha_{\text{LP,MSPE}}$-GMA consistently achieves the lowest standard deviation beyond $h = 4$.

Notably, estimators that incorporate both model fit and prediction performance in their weighting schemes tend to outperform those that apply the same criterion for both weight assignments. Accordingly, the MSE of the $\alpha_{\text{LP,MSPE}}$-GMA or $\alpha_{\text{LP,RS}}$-CVA estimators is consistently lower across horizons after $h = 4$.

Advantages of using α_{LP}-based estimators.

The simulation results confirm that incorporating the second-stage weight, α_{LP}, to combine LP- and VAR-based estimators in model averaging leads to improved estimators that reduce total error and display less fluctuation and greater consistency across horizons. In Section 2, we outlined the theoretical benefits of model averaging schemes that incorporate α_{LP}. Here, we revisit those advantages in light of the simulation evidence.

Superior Intermediate Horizon Performance. Model averaging with α_{LP} exhibits stable performance in terms of total error by leveraging both LP- and VAR-based estimators, which inherently exhibit a bias-variance trade-off. As a result, these estimators perform particularly well at intermediate horizons when evaluated using MSE. Regardless of how sophisticated a single estimator may be, relying on a single method cannot consistently yield stable performance across all horizons. For instance, the BVAR model shows extreme bias at certain horizons when the shock is monetary, substantially increasing its total error.

Horizon-Specific Responses with Extended Information. The combined use of LP- and VAR-based estimators enables the strengths of both approaches to be utilized. These combined estimators deliver horizon-specific impulse responses, offering greater flexibility than MAVG_{VAR}, while also incorporating longer-term information compared to MAVG_{LP}. Simulation results confirm that such combinations yield more robust estimates, especially at intermediate horizons, regardless of the shock type.

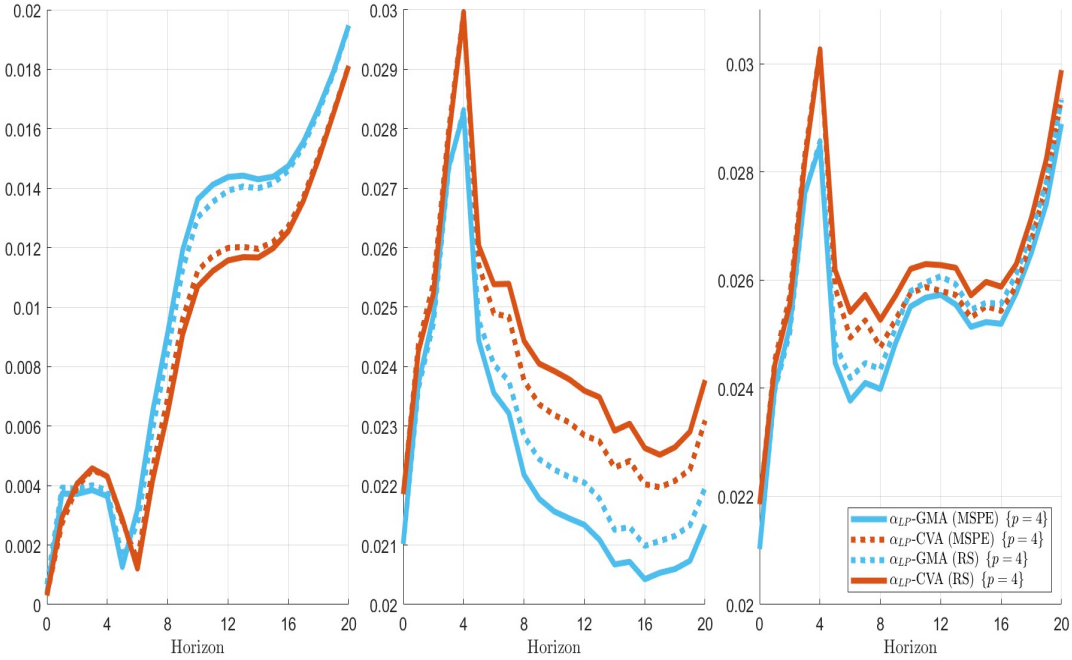
Enhanced Performance through Balanced Model Fit and Prediction Accuracy. α_{LP} model averaging allows researchers to consider both model fit and predictive performance through two-stage weighting. For instance, in $\alpha_{\text{LP,MSPE}}$-GMA, GMA assigns initial weights based on model fit and complexity, while $\alpha_{\text{LP,MSPE}}$ reflects predictive performance. Conversely, in $\alpha_{\text{LP,RS}}$-CVA, the initial CVA weights are based on predictive accuracy, and $\alpha_{\text{LP,RS}}$ reflects model fit. Notably, one of these hybrid estimators—either $\alpha_{\text{LP,MSPE}}$-GMA or $\alpha_{\text{LP,RS}}$-CVA—consistently outperforms the others across the full horizon range, regardless of the relative emphasis placed on bias or variance in the loss function.

4.5 Robustness Check

This section reinforces the main conclusions presented in Section 4.4 by showing that they remain robust to various modifications of our baseline simulation design. For the robust-

Observed Fiscal Shock

***aBias* (Left) , *SD* (Middle) and MSE (Right) of Estimators**



Observed Monetary Shock

***aBias* (Left) , *SD* (Middle) and MSE (Right) of Estimators**

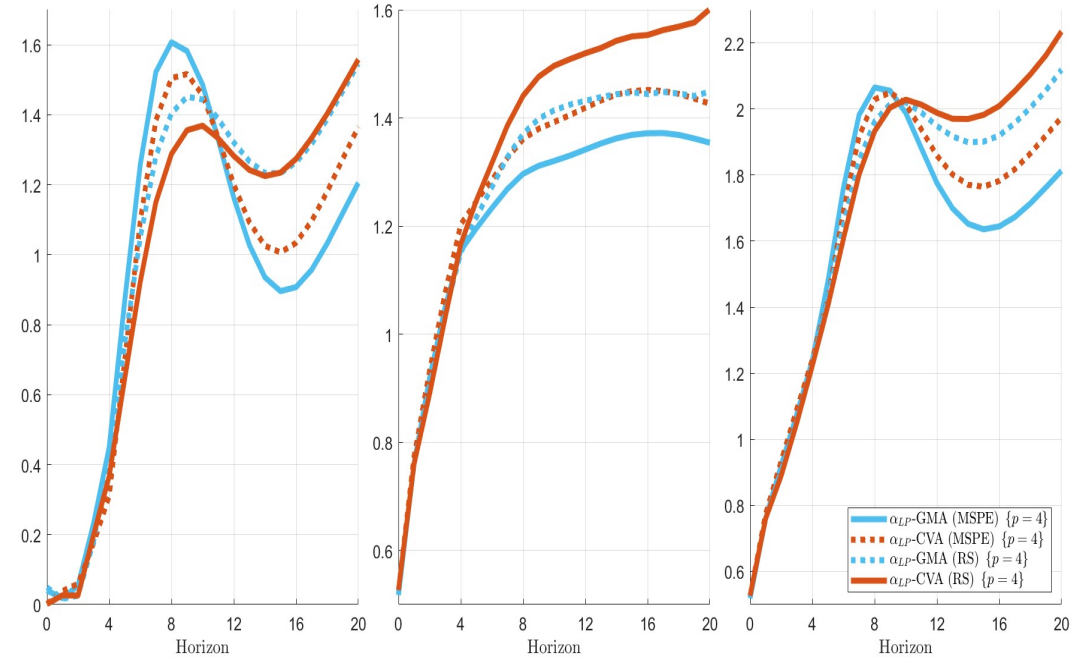


Figure 8: Average absolute bias, standard deviation and MSE when a shock is observed.

ness simulations, we employ trimmed IRFs, except for those examining alternative lag lengths. Trimming is used because the variance of some estimators can be disproportionately affected by outliers, potentially obscuring bias patterns. Specifically, we winsorize each row of simulated IRFs at the 1st and 99th percentiles by replacing extreme values with the corresponding thresholds.

IV/Proxy Identification. The results under IV/proxy identification closely mirror those obtained under observed shock identification. When the shock is fiscal, the bias-variance trade-off between LP-based and VAR-based estimators is again evident, especially at intermediate horizons. When the shock is monetary, the superior performance of MAVG_{VAR} estimators observed in the baseline setting becomes even more pronounced. Accordingly, α_{LP}-based estimators that combine LP and VAR approaches continue to achieve lower MSE than MAVG_{LP} and are comparable to or more stable than MAVG_{VAR} alone, particularly for researchers concerned with both bias and variance. Since no single estimator consistently outperforms others in terms of bias across all horizons, and MAVG_{VAR} estimators generally exhibit lower variability, these findings suggest that MAVG estimators can offer more robust performance by mitigating the respective weaknesses of LP- and VAR-based estimators. Finally, consistent with the main findings, α_{LP} estimators that incorporate both model fit and predictive accuracy continue to outperform those relying on a single criterion. Detailed results and figures are reported in Appendix D.2.

Recursive Identification. The results under recursive identification are broadly consistent with those from observed shock identification, with the exception of the bias-variance trade-off among model averaging estimators. This trade-off is less evident when the shock is fiscal. In contrast, when the shock is monetary, MAVG_{VAR} estimators outperform others in both matrices from the intermediate horizon onward, in line with the main results. Additionally, regardless of the shock type, no single estimator outperforms across all horizons, and single estimators tend to exhibit larger fluctuations compared to the baseline results. This suggests a higher likelihood that model averaging techniques can lower loss function values for researchers. This advantage is especially clear in the monetary shock case, where MAVG estimators display lower bias and variability than the selected single estimators.

Stationary Data. Although most empirical studies estimate VARs and LPs using a combination of stationary and non-stationary variables in levels (e.g., Ramey (2016)), some researchers prefer to transform all variables to achieve stationarity prior to analysis. We replicate the analysis using the stationary DFM estimated by Stock and Watson (2016) as the basis for the DGPs. We generate impulse responses using the same estimation procedures as in the baseline, with the exception of the BVAR estimator. For BVAR, the prior is adjusted to converge toward white noise rather than random walks. The full results and figures are reported in Appendix E.

While some differences emerge, the main conclusions are generally preserved under the stationary DGPs. The bias-variance trade-off between LP- and VAR-based estimators becomes less distinct, as shrinkage and Bayesian estimators display greater fluctuation in bias, while LS LP and BC LP maintain relatively low and stable bias but exhibit higher variability across horizons. These patterns continue to support the usefulness of model averaging techniques, which offer more robust estimators by balancing these trade-offs. This is clearly visible in the associated figures. However, when the shock is monetary, the benefit of combining model fit and prediction accuracy in the α_{LP} weighting schemes appears somewhat diminished.

Other lag lengths. The results using shorter or longer lag lengths under observed shock identification remain consistent with those in the main analysis, although the bias and variance of LP and VAR estimators are somewhat improved or worsened depending on the specification. As a result, the bias-variance trade-off between MAVG_{LP} and MAVG_{VAR} estimators is still observed at intermediate horizons, though its magnitude changes slightly when the shock is fiscal. When the shock is monetary, the trade-off becomes less pronounced, but MAVG_{VAR} estimators continue to outperform in both bias and variance at relatively longer horizons, in line with the main results. These findings reinforce that no single estimator dominates across all horizons, and the presence of the bias-variance trade-off between LP- and VAR-based estimators supports the robustness of the α_{LP} schemes. Our simulation results confirm that these schemes effectively lower researchers' loss across different lag specifications, even though their relative performance may vary slightly depending on the lag length used.

5 Empirical Applications

We now compare our transformed model averaging estimators, which combine both LP-based and VAR-based methods, with other model averaging techniques recently proposed in the literature. In Section 5.1, we evaluate the impulse responses to a monetary shock obtained from our approach against those produced by the Stein combination shrinkage method of Hansen (2016), using updated Federal Reserve data. In Section 5.2, we compare our estimates to those generated by a more recent model averaging scheme: prediction pools (Ho et al., 2024).

5.1 Assessment of Comparisons: Stein Combination vs. α_{LP}-GMA, α_{LP}-CVA

Hansen (2016) proposes a data-driven model averaging approach that assigns weights to minimize an estimated MSE of a vector-valued parameter of interest. The method is designed to improve estimation accuracy and address limitations of multivariate autoregressive models.

Data. We estimate the seven-variable medium-scale models following Giannone et al. (2015), as applied in Hansen (2016), using the latest Federal Reserve data covering the period from 1959:Q1 to 2023:Q4.

Table 1: Data description (1959:1Q-2023:4Q)

Description	FRED	Transformation
Real Gross Domestic Product	GDPC1	$4\log$
GDP Implicit Price Deflator	GDPDEF	$4\log$
Real Personal Consumption Expenditure	PCECC96	$4\log$
Real Gross Private Domestic Investment	GPDIC1	$4\log$
Hours Worked: Nonfarm Business Sector	HOANBS	$4\log$
Real Compensation per Hour	COMPRNFB	$4\log$
Federal Funds Rate	FF	$\div 10$

Models. We consider six models in this empirical application:

1. Single estimation models: BC LP and BVAR.
2. Hansen's Stein combination (abbreviated "VAR Avg"):
    - AR and LS VAR.
3. MAVG_{ALL}: $\alpha_{\text{LP,MSPE}}$-GMA, $\alpha_{\text{LP,RS}}$-GMA, $\alpha_{\text{LP,MSPE}}$-CVA, and $\alpha_{\text{LP,RS}}$-CVA.
    - LP-based approaches: LS LP, BC LP, BLP, and Pen LP.
    - VAR-based approaches: LS VAR, BC VAR, and BVAR.

In line with the simulation setup, we include BC LP and BVAR to facilitate a more comprehensive comparison. The VAR Avg estimator is constructed as a weighted average of ten specifications: univariate AR(1) to AR(5) models and multivariate VAR(1) to VAR(5) models. Under our classification, VAR Avg falls under MAVG_{VAR}, since it exclusively relies on VAR-based approaches. However, it has been noted that VAR Avg estimators implicitly incorporate LS LP behavior through the use of long-lag VARs.[7]

Our primary comparison focuses on the impulse responses obtained from MAVG_{ALL} versus those from VAR Avg, with BC LP and BVAR included as benchmarks. As in the simulation study, the weights in our model averaging schemes are computed using all seven estimators introduced in Section 3. All models, except for VAR Avg, control for five lags and include a constant term, in accordance with standard practice in the literature. The shock is identified via recursive ordering, under the assumption that the monetary shock is unobserved. For LS VAR models, we estimate the equation in (11), placing the policy variable (FF) last in the vector w_t, which contains all variables listed in Table 1. For LS LP models, we estimate the regression in (10), where x_t is the policy variable, and control variables include contemporaneous and lagged values of w_t.

[7]VARs with long lag lengths can approximate the behavior of LP estimators with several lagged controls. See Miranda-Agrippino and Ricco (2021b)and Plagborg-Møller and Wolf (2021)for details.

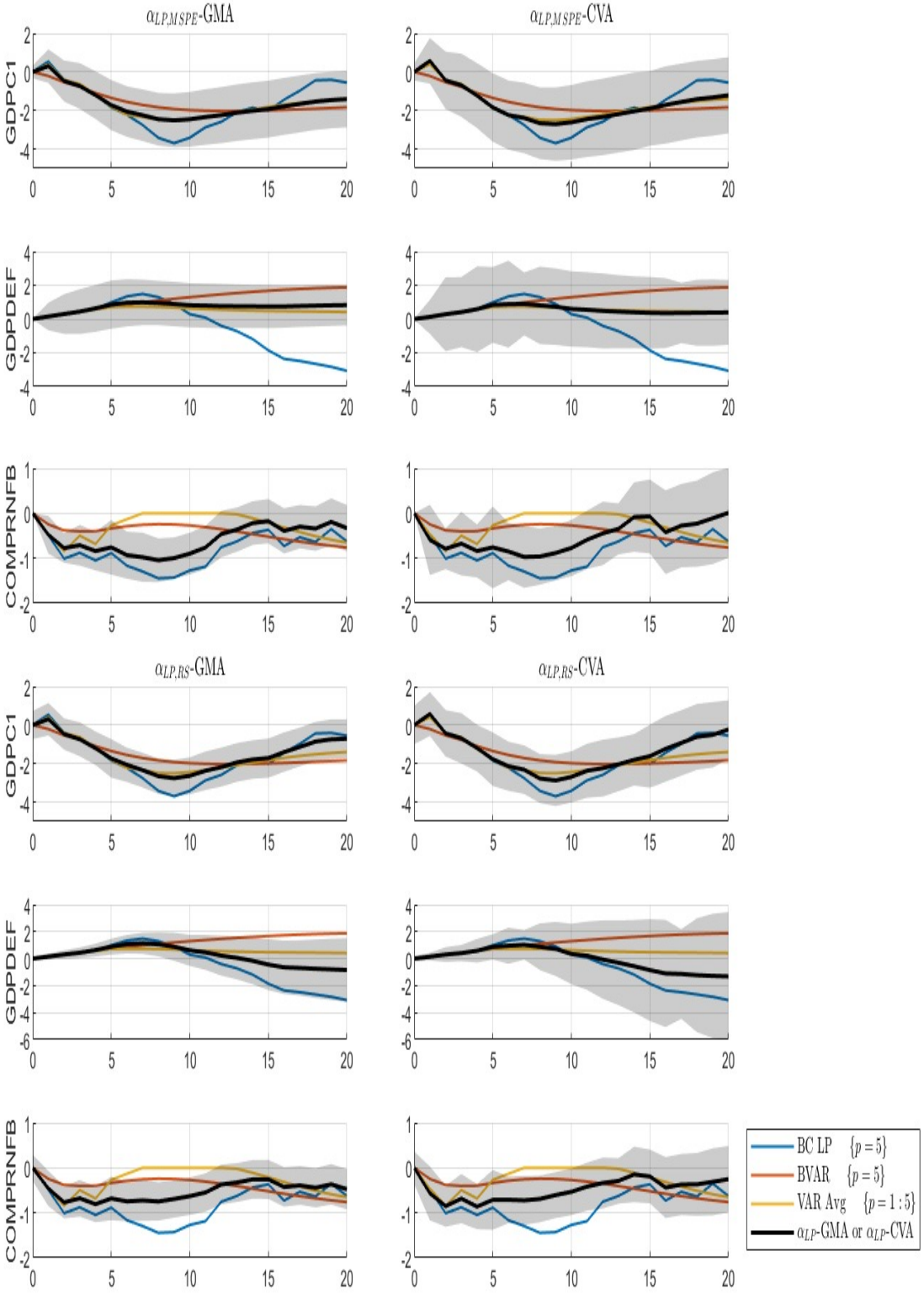


Figure 9: Impulse Response Estimates Due to a Monetary (Fed Funds) Shock. Shaded regions indicate 95% error bands. Note that error-band widths are rescaled as follows: Real GDP (GDPC1) by a factor of 5; GDP Deflator (GDPDEF) by 30 for the $\alpha_{\text{LP,MSPE}}$ estimator and by 10 for the $\alpha_{\text{LP,RS}}$ estimator; and Real Compensation per Hour (COMPRNFB) by 3.

Results. Figure 9 displays the impulse response estimates for all models described above. A common pattern across all variables except for real compensation per hour is that model averaging schemes, including VAR Avg, help smooth the fluctuations observed in single estimators beyond horizon $h = 5$. For real compensation per hour, estimators show high volatility across methods. Among the estimators, the α_{LP} estimators yield stable and smoothed trajectories, mitigating the sharp fluctuations seen in both single estimators and VAR Avg.

When comparing the $\alpha_{\text{LP,MSPE}}$-based and $\alpha_{\text{LP,RS}}$-based approaches, the MSPE-based variants tend to produce estimates closer to those of VAR Avg when no significant divergence is present. This is likely because MSPE-based methods assign relatively high weights to LS VAR. (See Appendix C.1.) Moreover, the MSPE-based responses are generally smoother than their RS-based counterparts, except in the case of real compensation per hour. This pattern suggests that weighting based on prediction accuracy leads to more conservative (i.e., less noisy) responses.

Comparing GMA-based and CVA-based approaches in terms of their first-stage weight assignment, CVA estimators appear slightly more responsive to local dynamics at medium and longer horizons (after $h = 5$), whereas GMA-based estimates tend to be more parsimonious, producing tighter impulse response trajectories. Also, GMA-based estimators exhibit narrower 95 % confidence intervals. Under the usual regularity conditions, these tighter bands reflect superior finite-sample efficiency rather than an understatement of uncertainty relative to CVA-based estimators.

5.2 Assessment of Comparisons: Prediction Pools vs. α_{LP}-GMA, α_{LP}-CVA

Ho et al. (2024) propose a model averaging method by extending the prediction pool framework originally introduced by Geweke and Amisano (2011). In this approach, the weights are chosen to maximize the predictive performance of the averaged model at each horizon. Each candidate model contributes a forecast density conditional on both the model's parameters and the specified shock. By interpreting impulse responses as conditional forecasts, this framework can be adapted to construct model-averaged impulse responses.

Data. This empirical application follows the analysis of the Romer and Romer (2004) monetary policy shocks, as employed in Ramey (2016), covering the period from March 1969 to December 1996.

Models. We compare our transformed traditional model averaging schemes with the prediction pools proposed by Ho et al. (2024), evaluating both impulse response estimates and model weights.

1. Prediction pools (abbreviated "*Pool*").
2. MAVG$_{\text{ALL}}$: $\alpha_{\text{LP,MSPE}}$-GMA, $\alpha_{\text{LP,RS}}$-GMA, $\alpha_{\text{LP,MSPE}}$-CVA, and $\alpha_{\text{LP,RS}}$-CVA.

Table 2: Data description (1969:Mar.-1996:Dec.)

Description	
Log of Industrial Production	LIP
Unemployment Rate	UNEMP
Log of Consumer Price Index	LCPI
Real Gross Private Domestic Investment	LPCOM
Federal Funds Rate	FFR
Romer and Romer instrument	RRORIG

Within these model averaging schemes, we consider three candidate models. VAR(p) and LP(p) estimators include a constant term, and p indicates the lag length, following the same specification conventions. All impulse responses are estimated under IV/proxy identification, using the Romer-Romer shock series as the instrument.

- **Internal instrument VAR(12)** We estimate a VAR model using equation (11), with the Romer and Romer (2004) instrument z_t placed first in the observable set $w_t = \{z_t, \bar{w}_t\}$, where $\bar{w}_t$ includes all remaining variables. The monetary policy shock is identified as the first innovation from a Cholesky decomposition of the reduced-form residuals.

- **LP(2) with Contemporaneous Controls** For the LP models, we estimate:

$$y_{t+h} = \mu_h + \beta_h z_t + \text{control variables} + \text{residulal}_{t+h}, \tag{21}$$

  where the control variables include contemporaneous values of endogenous variables and lags of both the instrument z_t and endogenous variables.

- **LP(2) without Contemporaneous Controls** This specification is identical to the previous LP model but excludes contemporaneous endogenous variables from the set of controls.

Results. The impulse response estimates are summarized in Figure 10. Following a contractionary monetary policy shock, output and inflation decline, while the unemployment rate gradually rises. Across short horizons, most estimators yield similar responses; however, they begin to diverge at intermediate and longer horizons.

Comparing the $\alpha_{\text{LP,MSPE}}$-based and $\alpha_{\text{LP,RS}}$-based approaches, we find that the MSPE-based variants closely resemble the VAR response, particularly beyond the initial horizons. This stems from the secondary weights in the MSPE-based estimators assign substantial weight to the VAR model. For the same reason, these estimators tend to align closely with the $Pool$ estimator, as both rely on prediction accuracy when constructing weights. Indeed, $Pool$ also assigns most of its weight to the VAR model (see Appendix C.2). If the $\alpha_{\text{LP,MSPE}}$ approaches had employed a comparable out-of-sample prediction criterion, their responses would likely be even more similar to those of $Pool$. By contrast, the $\alpha_{\text{LP,RS}}$-based

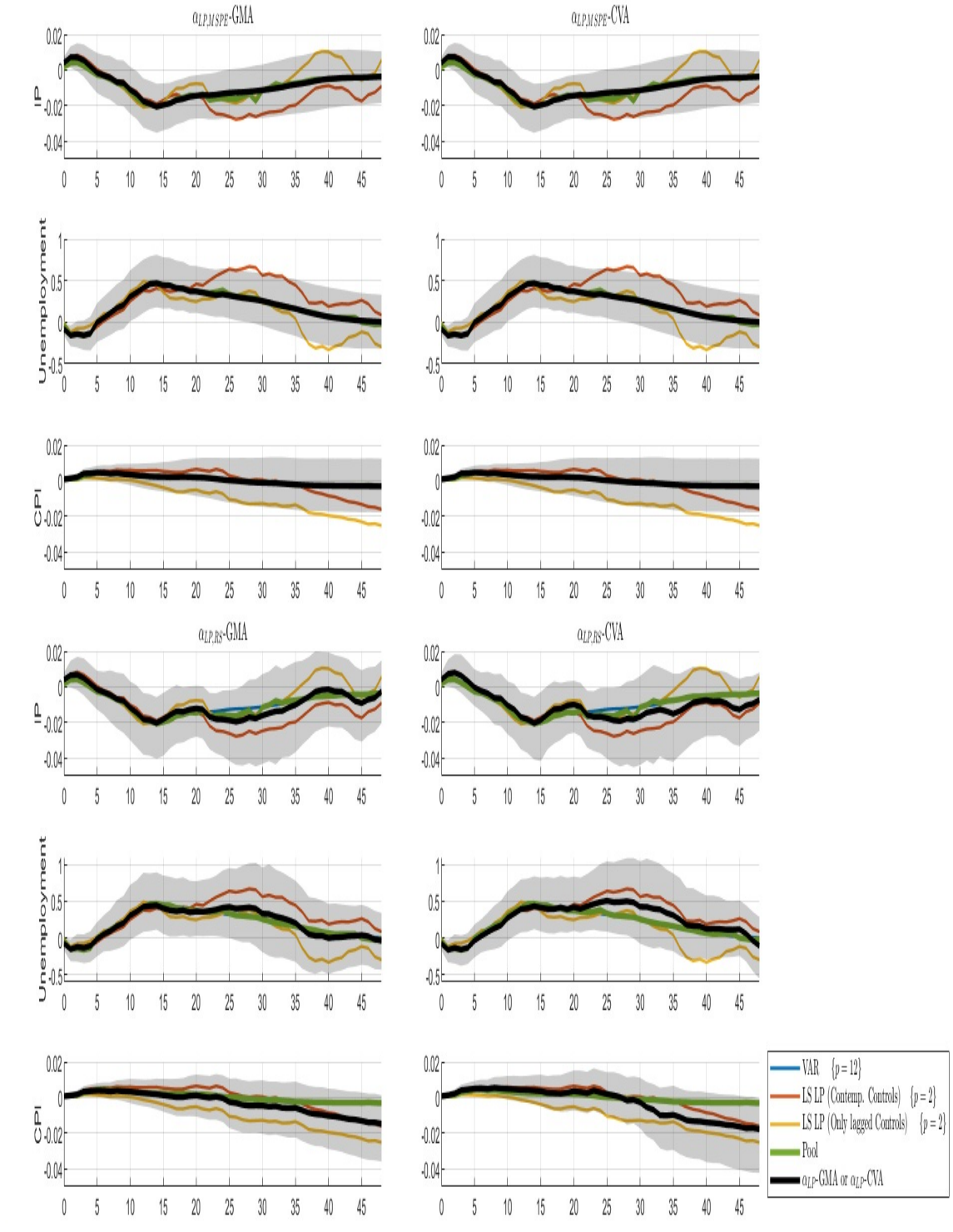


Figure 10: Impulse Response Estimates Due to a Monetary (Romer and Romer) Shock. Shaded regions indicate 95% error bands.

approaches tend to exhibit slightly steeper responses. Because R^2-based weights are based on model fit, they may be more sensitive to noise or overfitting, especially in volatile series.

Turning to the comparison between GMA- and CVA-based approaches, we observe

that GMA produces smoother and more centralized impulse response paths. In contrast, CVA is more responsive to local dynamics, particularly at intermediate and longer horizons. This pattern is consistent with findings from the first empirical study.

In summary, across both empirical applications, model averaging estimators typically span a middle ground between extremes. They yield more moderate estimates than single models, highlighting the value of combining information when responses are noisy or volatile. Moreover, we confirm that $\alpha_{\text{LP,MSPE}}$-based estimators are slightly more conservative and flatter than their $\alpha_{\text{LP,RS}}$-based counterparts. Meanwhile, GMA estimators tend to be slightly more parsimonious than CVA-based approaches.

6 Conclusion

In this paper, we propose a two-stage model averaging scheme based on a transformed traditional model averaging framework for estimating IRFs, referred to as α_{LP}-based model averaging. Our approach introduces a secondary weight, α_{LP}, which enables the combination of estimators across different models and, more importantly, across fundamentally different estimation methods such as LP and VAR. The scheme also accommodates both Bayesian and frequentist paradigms. Notably, the proposed methods are straightforward to implement and consistent with established practices in applied macroeconometrics.

The key contribution of this study is to extend traditional model averaging by recognizing and reflecting the structural differences between LP- and VAR-based estimators. In the first stage, we compute weights within each group using either model fit or prediction accuracy. In the second stage, we assign a weight α_{LP} to combine LP- and VAR-based estimators, effectively circumventing the incompatibilities that typically hinder joint averaging across these frameworks. This design allows researchers to exploit the strengths of both approaches while avoiding the need for a universal residual structure or penalty function.

Our simulation study confirms the advantages of this design. The α_{LP}-based estimators consistently reduce total estimation error, particularly at intermediate horizons where the bias-variance trade-off between LP and VAR is most pronounced. These results are robust across different identification schemes, lag specifications, and assumptions about stationarity. Moreover, we find that combining distinct criteria in the two stages—for example, using model fit in the first stage and prediction accuracy in the second—can further improve performance by reducing volatility in the estimated responses.

Empirical applications further demonstrate the practical benefits of the proposed approach. Across different macroeconomic variables, α_{LP}-based estimators generally produce smoother and more interpretable impulse response paths than single estimators. While their trajectories often fall between those of existing model averaging methods, such as Stein combination shrinkage and prediction pools, the α_{LP}-based estimators offer a flexible structure that accommodates both LP- and VAR-based models. This flexibility allows researchers to generate more robust responses, particularly when dealing with noisy or

volatile data.

Overall, our two-stage model averaging framework provides a practical strategy for combining LP- and VAR-based methods. By integrating multiple estimation approaches while accounting for their structural differences, α_{LP}-based estimators improve empirical reliability and provide a flexible tool for structural macroeconomic analysis.

References

Jushan Bai and Serena Ng. A PANIC attack on unit roots and cointegration. *Econometrica*, 72(4):1127–1177, 2004.

Matteo Barigozzi, Marco Lippi, and Matteo Luciani. Large-dimensional dynamic factor models: Estimation of impulse–response functions with I (1) cointegrated factors. *Journal of Econometrics*, 221(2):455–482, 2021.

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

Olivier Blanchard and Roberto Perotti. An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *Quarterly Journal of Economics*, 117(4):1329–1368, 2002.

Fabio Canova. *Methods for applied macroeconomic research*, volume 13. Princeton University Press, 2007.

Xu Cheng and Bruce E. Hansen. Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280–293, 2015.

Lawrence J. Christiano, Martin Eichenbaum, and Charles L. Evans. Monetary policy shocks: What have we learned and to what end? In John B. Taylor and Michael Woodford, editors, *Handbook of Macroeconomics*, volume 1, pages 65–148. Elsevier, 1999.

John Geweke and Gianni Amisano. Optimal prediction pools. *Journal of Econometrics*, 164 (1):130–141, 2011.

Domenico Giannone, Michele Lenza, and Giorgio E. Primiceri. Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2):436–451, 2015.

Clive W.J. Granger and Youngil Jeon. Thick modeling. *Economic Modelling*, 21(2):323–343, 2004.

Bruce E. Hansen. Multi-step forecast model selection. In *20th Annual Meetings of the Midwest Econometrics Group*, pages 1–2, April 2010.

Bruce E. Hansen. Stein combination shrinkage for vector autoregressions. Manuscript, University of Wisconsin-Madison, 2016.

Bruce E. Hansen and Jeffrey S. Racine. Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46, 2012.

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

Paul Ho, Thomas A. Lubik, and Christian Matthes. Averaging impulse responses using prediction pools. *Journal of Monetary Economics*, page 103571, 2024.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial (with comments and rejoinder). *Statistical Science*, 14(4): 382–417, 1999.

Søren Johansen. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, Oxford, 1995.

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

Lutz Kilian. Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics*, 80(2):218–230, 1998.

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

Dan Li, Mikkel Plagborg-Møller, and Christian K. Wolf. Local projections vs. VARs: Lessons from thousands of DGPs. *Journal of Econometrics*, page 105722, 2024.

Silvia Miranda-Agrippino and Giovanni Ricco. Bayesian local projections. 2021a.

Silvia Miranda-Agrippino and Giovanni Ricco. The transmission of monetary policy shocks. *American Economic Journal: Macroeconomics*, 13(3):74–107, 2021b.

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

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

A. Lawrence Pope. Biases of estimators in multivariate non-Gaussian autoregressions. *Journal of Time Series Analysis*, 11(3):249–258, 1990.

Valerie A. Ramey. Macroeconomic shocks and their propagation. In John B. Taylor and Harald Uhlig, editors, *Handbook of Macroeconomics*, volume 2, pages 71–162. Elsevier, 2016.

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

Christopher A. Sims. Macroeconomics and reality. *Econometrica*, pages 1–48, 1980.

James H. Stock and Mark W. Watson. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In John B. Taylor and Harald Uhlig, editors, *Handbook of Macroeconomics*, volume 2, pages 415–525. Elsevier, 2016.

Appendix A Model averaging based on GMA

As discussed in Section 2, VAR and LP variants differ in their residual structures, which complicates direct comparisons using the identical information criterion. When VAR variants are well-specified, their residuals are more likely to satisfy the white noise assumption. Therefore, we apply traditional BMA using the BIC for VAR-based estimators.

By contrast, applying the same BIC to LP variants can lead to elevated autocorrelation and potential heteroskedasticity in the residuals as the horizon increases, which hinders valid comparisons. To address this, we employ HAC-GIC for model averaging among LP variants. This criterion accounts for the more complex residual structure inherent in LP estimation. By adopting HAC-GIC within a BMA-style framework, we construct what we refer to as GMA for LP-based estimators.

Appendix A.1 Bayesian Model Averaging (BMA)

Let the m-th VAR-based model (for $m = 1, ..., M_{\text{VAR}}$) be expressed in matrix form as

$$Y = X\boldsymbol{\beta} + \boldsymbol{e}, \tag{A.1}$$

where $Y \in \mathbb{R}^{T\times N}$, $X \in \mathbb{R}^{T\times J}$, $\boldsymbol{\beta} \in \mathbb{R}^{J\times N}$, and $\boldsymbol{e} \in \mathbb{R}^{T\times N}$. Here, T is the number of time-series observations, N denotes the number of observable variables in the VAR system, and $J = 1 + N \cdot p$, where p is the lag length. The matrix X collects a constant and all lagged regressors.

More specifically, let $Y = [y_1, y_2, \ldots, y_N]$, where each $y_n \in \mathbb{R}^{T\times 1}$ denotes the time series for variable n, with $n = 1, \ldots, N$. The matrix of regressors is given by $X = [x_1', x_2', \ldots, x_T']'$, where each row vector $x_i' \in \mathbb{R}^{1\times J}$, with $i = 1, \ldots, T$. The coefficient matrix is $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_N]$, where each $\beta_n \in \mathbb{R}^{J\times 1}$ contains the coefficients for the n-th observable. Finally, the error matrix is $\boldsymbol{e} = [e_1, e_2, \ldots, e_N]'$, with each $e_n \in \mathbb{R}^{T\times 1}$ denoting the residual vector for equation n.

From now on, for simplicity, we omit the subscription n from all variables.[8] Then, for each n,

$$y = X\beta + e, \tag{A.2}$$

where $\mathbb{E}(e_i|x_i) = 0$, and $\sigma_i^2 = \mathbb{E}(e_i^2|x_i)$ to allow for heteroscedasticity.

[8] For example, $y = y_n$, $\beta = \beta_n$, $e = e_n$, $e_i^2 = e_{i,n}^2$, and $\sigma_i^2 = \sigma_{i,n}^2$.

The log-likelihood function of each model for the variable of interest is

$$\log L(\hat{\beta}, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_T^2 | y, X) \tag{A.3}$$

$$= -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{T}\log\hat{\sigma}_i^2 - \frac{1}{2}\sum_{i=1}^{T}\frac{(y_i - x_i\hat{\beta})^2}{\hat{\sigma}_i^2} \tag{A.4}$$

$$= -\frac{1}{2}\left(T\log(2\pi) + \sum_{i=1}^{T}\log\hat{\sigma}_i^2\right) - \frac{1}{2}\sum_{i=1}^{T}\frac{(y_i - x_i\hat{\beta})^2}{\hat{\sigma}_i^2}. \tag{A.5}$$

Let $D = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_T^2)$. Then the likelihood simplifies to

$$\log L = -\frac{T}{2}\log|D| - \frac{1}{2}\hat{e}' D^{-1}\hat{e}. \tag{A.6}$$

Ignoring constants that are invariant across models, the simplified log-likelihood becomes

$$\log L \propto -\frac{T}{2}\log|D|. \tag{A.7}$$

The BIC of the m is then given by:

$$BIC_m = T_m \log|D_m| + d_m \log(T_m), \tag{A.8}$$

where D_m is the estimated forecast error covariance matrix, and d_m is the number of parameters in model m.

The BMA weight assigned to model m is horizon-constant and given by

$$\omega_{m,\text{BMA},h} = \frac{exp(-BIC_m/2)}{\sum_{l=1}^{M} exp(-BIC_l/2)}. \tag{A.9}$$

Thus, the BMA weights are normalized likelihood-based scores penalized by model complexity.

Appendix A.2 Generalized Model Averaging (GMA)

Let the m-th LP-based model (for $m = 1, \ldots, M_{\text{LP}}$) be expressed in matrix form as

$$y = X\beta + e, \tag{A.10}$$

where $y \in \mathbb{R}^{T\times 1}$, $X \in \mathbb{R}^{T\times J}$, $\beta \in \mathbb{R}^{J\times 1}$, and $e \in \mathbb{R}^{T\times 1}$. Here, T is the number of time-series observations, and J is the number of regressors including a constant term and lagged variables as per the LP specification.

To account for possible heteroskedasticity and autocorrelation in the residuals, we adopt the HAC covariance matrix estimator. The HAC-adjusted covariance matrix is de-

fined as

$$\hat{\Sigma}^{\text{HAC}} = \hat{\Gamma}_0 + \sum_{j=1}^{q} \left(1 - \frac{j}{q+1}\right) \left(\hat{\Gamma}_j + \hat{\Gamma}_j'\right), \tag{A.11}$$

where $\hat{\Gamma}_j = \frac{1}{T} \sum_{t=j+1}^{T} \hat{e}_t \hat{e}_{t-j}'$ is the sample autocovariance at lag j, and q is the bandwidth parameter. The bandwidth parameter q plays a crucial role in constructing the HAC-adjusted covariance matrix. The Newey-West rule is commonly used in practice:

$$q = \text{round}\left(4\left(\frac{T}{100}\right)^{2/9}\right), \tag{A.12}$$

An alternative is an automatic, data-dependent method that adapts to the autocorrelation structure of the residuals, which we adopt in this study. Specifically, for each regressor's residual series S_1, we compute the first-order autocorrelation estimate:

$$\hat{\rho} = \frac{S_1(1:T-1)' S_1(2:T)}{S_1(1:T-1)' S_1(1:T-1)}. \tag{A.13}$$

Then, across k regressors, the effective bandwidth is calculated as

$$q = 1.8171 \cdot \left(\frac{\sum_{i=1}^{k} \frac{\hat{\rho}_i^2}{(1-\hat{\rho}_i)^4}}{\sum_{i=1}^{k} \frac{(1-\hat{\rho}_i)^2}{(1-\hat{\rho}_i)^4}}\right)^{1/3} T^{1/3}. \tag{A.14}$$

This adaptive method offers greater flexibility by tailoring q to the estimated persistence in the residuals and is especially useful in small or moderate sample sizes or when the residual autocorrelation varies across models or horizons.

To evaluate model fit while accounting for the HAC structure, we define the HAC-Generalized Information Criterion (HAC-GIC) for model m at horizon h as:

$$\text{HAC-GIC}_{m,h} = T_{m,h} \log\left(\left|\hat{\Sigma}_{m,h}^{\text{HAC}}\right|\right) + d_m \log(T_{m,h}), \tag{A.15}$$

where $\hat{\Sigma}_{m,h}^{\text{HAC}}$ is the HAC-adjusted covariance matrix of residuals, d_m is the number of parameters in model m, and $T_{m,h}$ is the number of observations available at horizon h.

The GMA weight assigned to model m at horizon h is given by

$$\omega_{m,\text{GMA},h} = \frac{\exp(-\text{HAC-GIC}_{m,h}/2)}{\sum_{l=1}^{M} \exp(-\text{HAC-GIC}_{l,h}/2)}. \tag{A.16}$$

Thus, GMA weights are normalized scores based on HAC-adjusted model fit and penalized by model complexity. Unlike BMA, GMA weights vary by horizon h, allowing for horizon-specific model averaging.

Appendix B Cross Validation Averaging (CVA)

We modify the CVA method proposed by Hansen (2010) and Cheng and Hansen (2015). Let $y \in \mathbb{R}^{T\times 1}$ denote the outcome vector, and let $X^m \in \mathbb{R}^{T\times J_m}$ denote the regressor matrix for model m, for $m = 1, \ldots, M(M_{\text{LP}}$ or $M_{\text{VAR}})$, where J_m may vary by model.

For simplicity, we drop any variable-specific subscript (e.g., n) from all notation.[9] We consider the conditional mean representation:

$$y_i = \mu_i + e_i, \qquad \mathbb{E}(e_i \mid x_i) = 0, \tag{B.1}$$

where $\mu = (\mu_1, \ldots, \mu_T)'$ and $e = (e_1, \ldots, e_T)'$ are $T \times 1$ vectors. The conditional variance $\sigma_i^2 = \mathbb{E}(e_i^2 \mid x_i)$ is allowed to vary across i.

For each model m, we estimate μ via least squares:

$$\hat{\mu}^m = X^m \hat{\beta}^m, \quad \text{where} \quad \hat{\beta}^m = \left(X^{m\prime} X^m\right)^{-1} X^{m\prime} y. \tag{B.2}$$

The model averaging estimator is then:

$$\hat{\mu}(\omega) = \sum_{m=1}^{M} \omega_m \hat{\mu}^m. \tag{B.3}$$

At each time $i = h, h+1, \ldots, T-h+1$, where $h > 1$, we compute model estimates using a symmetric subsample centered at i. Specifically, let

$$\mathcal{I}_i = \{i - h + 1, \ldots, i + h - 1\}$$

denote the subsample used for estimation, and let

$$\mathcal{T}_i = \{1, \ldots, T\} \setminus \mathcal{I}_i$$

denote the set of observations omitted.

Let $X^m_{-\mathcal{I}_i}$ and $y_{-\mathcal{I}_i}$ denote the regressor matrix and outcome vector for model m restricted to the subsample $\mathcal{T}_i$. Then, the leave-h-out OLS estimator at time i is given by:

$$\tilde{\beta}_t^m = \left(X^{m\prime}_{-\mathcal{I}_t} X^m_{-\mathcal{I}_t}\right)^{-1} X^{m\prime}_{-\mathcal{I}_t} y_{-\mathcal{I}_t},$$

and the corresponding in-sample fitted value at time i is

$$\tilde{\mu}_i^m = x_i^m \tilde{\beta}_i^m.$$

The leave-h-out residual for model m at time i is defined as:

$$\tilde{e}_i^m = y_i - \tilde{\mu}_i^m. \tag{B.4}$$

[9] For example, $y = y_n$, $\mu = \mu_n$, and $e = e_n$.

Let $\tilde{e}^m \in \mathbb{R}^{T\times 1}$ be the vector of leave-h-out residuals stacked over i, and define $\tilde{e} = [\tilde{e}^1, \ldots, \tilde{e}^M] \in \mathbb{R}^{T\times M}$. Then, model-averaged predictions and residuals are given by:

$$\tilde{\mu}(\mathbf{w}_{\text{CVA},h}) = \tilde{\mu}\mathbf{w}_{\text{CVA},h}, \tag{B.5}$$

$$\tilde{e}(\mathbf{w}_{\text{CVA},h}) = y - \tilde{\mu}(\mathbf{w}_{\text{CVA},h}) = \tilde{e}\mathbf{w}_{\text{CVA},h}, \tag{B.6}$$

where $\tilde{\mu} = [\tilde{\mu}^1, \ldots, \tilde{\mu}^M] \in \mathbb{R}^{T\times M}$ stacks fitted values across models.

The cross-validation loss function is:

$$CV_h(\mathbf{w}_{\text{CVA},h}) = \frac{1}{T}\tilde{e}(\mathbf{w}_{\text{CVA},h})'\tilde{e}(\mathbf{w}_{\text{CVA},h}) = \mathbf{w}'_{\text{CVA},h}S_h\mathbf{w}_{\text{CVA},h}, \tag{B.7}$$

where $S_h = \frac{1}{T}\tilde{e}'\tilde{e}$ is an $M \times M$ matrix constructed from the cross-validated residuals.

The optimal CVA weights minimize this loss:

$$\mathbf{w}_{\text{CVA},h} = \arg\min_{\mathbf{w}\in\mathcal{W}^*} CV_h(\mathbf{w}). \tag{B.8}$$

Appendix C Empirical Application: Model Averaging Weights

We report the weights assigned to each estimator within the model averaging schemes used in Sections 5.1 and 5.2.

Appendix C.1 Stein Combination vs. α_{LP}-GMA, α_{LP}-CVA

Figure C.1 presents the weights assigned to each estimator in the model averaging schemes used in Section 5.1, reported by response variable and forecast horizon.

Appendix C.2 Prediction Pools vs. α_{LP}-GMA, α_{LP}-CVA

Figure C.2 presents the weights assigned to each estimator in the model averaging schemes used in Section 5.2, reported by response variable and forecast horizon.

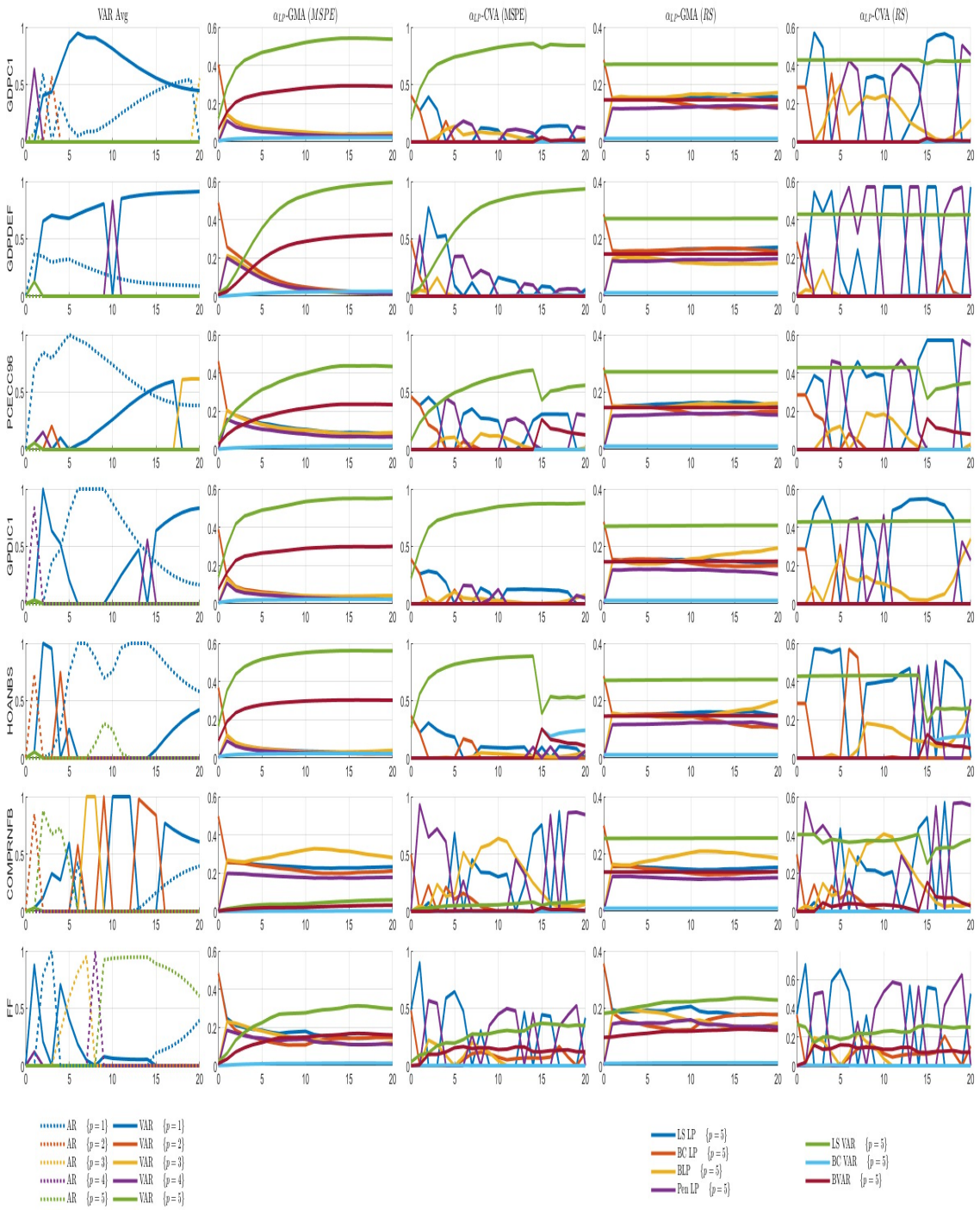


Figure C.1: Weights of Model Averaging: Stein combination (VAR Avg) and α_{LP} schemes. The **left** legend displays the weights from VAR Avg (10 estimators); the two **right** legends show the weights assigned in the α_{LP} schemes (7 estimators).

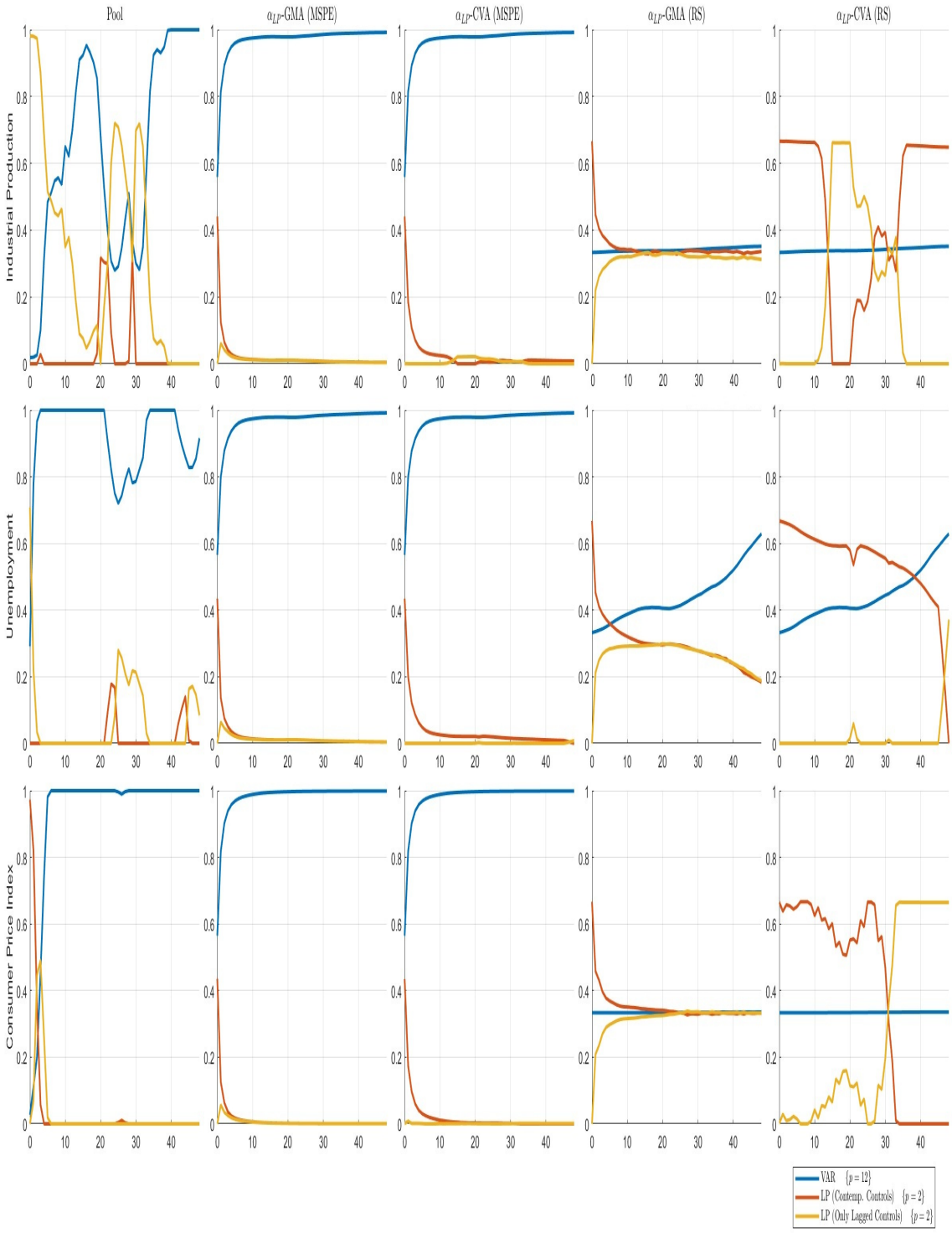


Figure C.2: Weights of Model Averaging: Prediction Pools ($Pool$), and α_{LP} schemes.